Bias in data used to train sales-based decision-making algorithms in a South African retail bank

Alice Wong

Student Number: 2303668

Supervisor: Ayanda Magida

A research report submitted to the Faculty of Commerce, Law and Management, University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business

Johannesburg, 2021

ABSTRACT

Banks are increasingly using algorithms to drive informed and automated decision-making. Due to algorithms being reliant on training data for the model to learn the correct outcome, banks must ensure that the customer data is securely and fairly used when creating product offerings as there is a risk of perpetuating intentional and unintentional bias. This bias can result from unrepresentative and incomplete training data or inherently biased data due to past social inequalities. This study aimed to understand the potential bias found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings.

The research adopted a qualitative approach and was conducted through ten virtual one-on-one interviews with semi-structured questions. Purposive sampling was used to select banking professionals from data science teams in a particular South African retail bank across demographics and levels of seniority. The data collected from the participants in the interviews were then thematically analysed to draw a conclusion based on the findings.

Key findings included:

An inconsistent understanding across data science teams in a South African retail bank around the prohibition of using the gender variable. This could result in certain developers using proxy variables for gender to inform certain product offerings. A potential gap in terms of the potential usage of proxy variables for disability (due to non-collection of this demographic attribute) to inform certain product offerings. Although disability was not identified as a known biased variable, it did, however, raise the question of whether banks should be collecting the customer's disability data and doing more in terms of social responsibility to address social inequalities and enable disabled individuals to contribute as effectively as abled individuals. As algorithms tend to generalise based on the majority's requirements, this would result in a higher error rate of under-represented groups of individuals or minority groups. This could result in financial exclusion or incorrect products being offered to certain groups of customers

which, if not corrected, would lead to the continued subordination of certain groups of customers based on demographic attributes.

Keywords: Algorithms, Bias, Customer Product Offerings, Data, Decision-making, Retail Banking

DECLARATION

I, Alice Wong, declare that this research report is my own work except as indicated in the references and acknowledgements. It is submitted in partial fulfilment of the requirements for the degree of Master of Management in the field of Digital Business at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in this or any other university.

Name: Alice Wong Signature: 

Signed at Rondebosch

On the 23rd day of February 2022

DEDICATION

This research paper is dedicated to my parents who have always believed in me and encouraged me to become a better version of myself continuously and to stay focused on climbing my own mountain. I am truly thankful to have you in my corner and for your unconditional love.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Ms Ayanda Magida, for her knowledge and guidance throughout this process. Your support and expertise pushed me and challenged me to elevate my work and thinking level.

I would like to thank all the lecturers at Wits Business School for their passionate delivery of knowledge throughout the Master's programme.

Thank you to my employer for allowing me to take time off to complete my studies and constant support. I appreciate all the stimulating conversations and words of encouragement to my mentor and colleagues.

Thank you for the endless support and encouragement throughout this process (and for all the fun distractions) to my family and friends. I could not have completed this research report without all of you.

Finally, to my Master's group and friends, thank you for all of the laughs and interactions. This journey would not have been as enjoyable without all of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AI	Artificial Intelligence
NCA	National Credit Act
PEPUDA	The Promotion of Equality and Prevention of Unfair Discrimination Act
POPIA	Protection of Personal Information Act
TCF	Treating Customers Fairly

1. INTRODUCTION

1.1 Purpose of the study

The purpose of this qualitative study was to gain a deeper understanding of the potential bias that can be found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings.

1.2 Background of the study

According to PricewaterhouseCoopers, several factors are influencing the banking industry, namely: customer expectations, technological capabilities, regulatory requirements, demographics and economics (Sullivan et al., 2020). This research predominantly focused on the customer expectations and technological capabilities factors.

With the rapid advancement in technological capabilities and competition (local, global, within and outside the banking sector), customer expectations continuously rise (MoneyMarketing, 2019). Customers have become accustomed to receiving personalised experiences due to the customised services they receive from the larger digital disruptors and, as a result, expect personalised experiences from their banks (and other industries) in the delivery method of their choice (Marous, 2021). Banks are now tasked with the challenge of keeping up with the constantly evolving customer expectations or run the risk of having their customers switch to another bank or FinTech (Bellens & Meekings, 2020).

Banks need to focus efforts on meeting customer personalisation requirements and moving away from traditional segmenting customers. Accenture found in its survey that customers want offerings that address their core needs and do not only focus on traditional banking products. The survey alluded to customers' expectations of receiving personalised advice from their banks based on their personal needs (W.UP, 2018).

Banks are in a prime position to fulfil customers' personalisation requirements as they continue to collect and store large volumes of data on what and where customers are willing to spend their money (Daly, 2020) that can be used to inform an individual's personalised experiences. The end goal for banks should be to create a segment that encompasses personalised and contextual offerings, content and customer experience (W.UP, 2018). This can be achieved by banks applying the right algorithmic technologies (Merlicek, 2018) to extract value from data (Martin, 2019).

To fulfil customer personalisation expectations, banks need to have a foundation of data and analytics capabilities (Marous, 2021). The use of data and analytics capabilities is not a new concept to the South African retail banks as they have been continuously trying to reap the benefits of applying the capabilities to automate previously highly resource-intensive and repetitive tasks (Chuprina & Kovalenko, 2021) as well as to reduce human error (Adams et al., 2020). This can be evidenced across the South African retail banks. TymeBank successfully streamlines the onboarding process by applying machine learning and advanced algorithms to perform verification checks on a customer against official sources (Malinga, 2019). This has resulted in a turnaround time of fewer than five minutes to open an account with TymeBank. Standard Bank implemented a WhatsApp-based chatbot to address specific concerns and questions by customers 24/7 (Nienaber, 2020). Absa's investment robo-advisor, "Virtual Investor", makes use of artificial intelligence and algorithms to make recommendations to customers based on their financial risk profiles (Tobor, 2017). These are just a few examples of how some South African retail banks have leveraged algorithmic technologies (including artificial intelligence (AI) based on algorithmic models (Joseph, 2020)) to enhance services offered to their customers.

With banks continuing to take advantage of the benefits that arise with the use of algorithms, some challenges must be addressed to prevent legal claims, fines and penalties from regulators due to non-compliance to relevant legislation (for example, Protection of Personal Information Act (POPIA) and Treating Customers Fairly (TCF)) and reputational damage (Scanlon, 2020). These challenges include bias in algorithmic decision-making that can systemically

disadvantage certain groups of individuals by gender or ethnicity (Dickson, 2020). Banks must be conscious of these challenges and proactively prevent them from materialising.

This study aimed to understand the potential bias found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings.

1.3 Research problem

Banks are increasingly using algorithms to drive informed and automated decision-making as they continue to collect and store large volumes of data. Algorithms rely on training data (such as historical data collected on its customers) for the model to learn the correct outcome that should be applied to other customers (Turner Lee & Resnick, 2019). One of the main concerns arising from the advancement of algorithmic usage is around the secure and fair usage of customer's data when creating product offerings, as mismanagement could result in intentional and unintentional biases (Bryant et al., 2019). There is evidence that certain decisions are less favourable towards specific groups of individuals by deploying these models. This bias can result from unrepresentative and incomplete training data or inherently biased data due to past social inequalities.

The impact of unethical algorithms creates numerous challenges. For certain South African customers, repercussions include further marginalisation and financial exclusion of specific groups of individuals (Moosajee, 2019). It is imperative that banks end-to-end protect customers' data (including the requirement around anonymisation of customers' data) and fairness when using customers' data (Bryant et al., 2019). For banks, the repercussions include legal claims, fines and penalties from regulators due to non-compliance to relevant legislation (for example, POPIA and TCF), and reputational damage (Scanlon, 2020). Retail banks are required to be compliant with many pieces of legislation. They are required to comply with the National Credit Act (NCA) to offer credit products to customers. The NCA includes requirements around protection

against discrimination, directly or indirectly, in line with The Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) (Republic of South Africa, 2006). PEPUDA explicitly highlights three categories where unfair discrimination is prohibited: race, gender and disability (Republic of South Africa, 2000).

This study focused on three potential biases found in the datasets used to train sales-based decision-making algorithms based on the abovementioned. These factors are gender bias, disability bias and imbalanced data. This research focused on gender and disability bias as racial bias has been widely researched.

1.4 Research questions

1.4.1 *Main research question*

How do underlying biases in training datasets used to train sales-based decision-making algorithms used by South African retail banks influence customer product offerings?

1.4.2 *Sub-research questions*

1) How does gender influence product offerings' algorithmic decisions in retail banking?
2) How does disability influence product offerings' algorithmic decisions in retail banking?
3) In what way does imbalanced data influence product offerings' algorithmic decisions in retail banking?

1.5 Significance of the study

The insights drawn from this research could be useful in identifying current gaps in the South African banking industry relating to the fair and ethical use of algorithms. In identifying these gaps, recommendations can be made in terms of workable steps to mitigate the risk of bias in datasets used to train algorithms. It may also lay the foundation for future research.

The insights drawn from this research can also provide a different perspective for the South African banking industry in terms of shifting their focus from being predominantly profit-driven to being more customer-centric. This is due to the representational harms that algorithms can reinforce by perpetuating stereotypes if adequate controls are not implemented to mitigate these harms as data can repeat further and transfer these social realities onto other individuals and generations (Scholz, 2017).

1.6 Delimitations of the study

- This study was confined to algorithms used for credit specific sales-based decision-making by South African retail banks that create credit product offerings (for example, loan pre-approvals for customers).
- This study focuses on bias in the training data (specifically gender bias, disability bias and imbalanced data) used to train sales-based decision-making algorithms and excludes all other types of bias in algorithmic decision-making.
- This study excluded algorithms used for autonomous systems.

1.7 Definition of terms

Algorithm An algorithm is a tool that transforms data (Scholz, 2017). Algorithms use data to create a specific decision that generates an action that may have ethical consequences (Tsamados et al., 2021).

Bias Bias is ‘a deviation from a standard’ (Danks & London, 2017:4692). Danks and London (2017) expand on bias in terms of moral bias and statistical bias. Moral bias refers to a judgement that deviates from a moral norm. Statistical bias refers to representing a certain group that deviates

	from the overall population. Statistical bias can aid in identifying moral biases.
Data	Data refers to a digital resource that can be generated and stored in a relational database (Scholz, 2017). De Mauro et al. (2016) expand on this definition: data is an information asset that requires specific technology and analytical methods to transform data into value. This study will focus on customer data generated and stored by South African retail banks that require specific technology and analytical methods to transform data into value.
Decision-making	The process of choosing the correct action from two or more alternatives with the intent of achieving a desired result (Sharma, n.d.).
Disability bias	Disability-specific moral bias where disability-based judgement deviates from a moral norm. This definition lends from Danks and London's (2017) definition of ‘moral bias’. This can include discrimination against an individual or group based on their disability status.
Gender bias	Gender-specific moral bias is where gender-based judgement deviates from a moral norm. This definition lends from Danks and London's (2017) definition of ‘moral bias’. This can include discrimination against an individual or group based on their gender.
Imbalanced data	Datasets that are an unequal representation of groups. There are two ways in which imbalanced data occurs during the data collection phase: customer distribution and nature of the event (Zhang & Zhou, 2019).

Retail bank	A financial institution that provides a banking product and service offerings to individual customers.
Sales-based decision-making	The process of choosing the correct action from two or more alternatives (Sharma, n.d.) with the intent of making a sale.

1.8 Assumptions

- Bias in algorithms created by South African retail banks occur predominantly as a result of training data bias.
- The datasets used to train sales-based decision-making algorithms include bias due to past social inequalities or financial exclusion factors. Further, these datasets have not been corrected prior to training the algorithms.
- South African retail banks create algorithms for decision-making. Further, these algorithms are used to generate sales-based customer offerings.
- South African retail banks use aggregated customer data to offer products to customers in similar role profiles or based on similar customer behaviours, such as gender and age groupings.
- South African retail banks are compliant with legislation applicable to the South African banking industry.

1.9 Chapter outline

The research report consists of six chapters and is structured as follows:

- Chapter 1 introduces and contextualises the research problem. The research problem and questions will form the basis for the following chapters.
- Chapter 2 provides a review of current, relevant academic literature relating to data, algorithms, bias in algorithmic decision-making focusing on training data bias, consequences of bias and controls that mitigate

training bias. Propositions that address the research questions will be stated in line with the literature review. The analytical framework underpinning this study will also be unpacked in this chapter.

- Chapter 3 outlines the research methodology used in this study to collect and analyse the data.
- Chapter 4 presents the findings from the data collected.
- Chapter 5 discusses and interprets the findings from chapter 4 to answer the research questions.
- Chapter 6 provides a conclusion to the study and provides recommendations for future studies.

2. LITERATURE REVIEW

2.1 Introduction

This chapter incorporates the current body of knowledge relating to this study. The research aimed to understand the potential bias found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings.

This chapter covers the various themes relating to the study. It begins with defining data and the advantages around its usage, defining algorithms, describing the various uses of algorithms across industries and the drivers behind the increased usage of algorithms. It then defines bias and describes the types of bias within an algorithmic value chain. Training data bias is then explored in terms of gender bias, disability, and imbalanced data. It then gives an overview of the consequences of bias followed by controls that could be put in place to mitigate bias. The chapter ends with the theoretical and conceptual frameworks applicable for this study.

2.2 An overview of Data

2.2.1 *Data definition*

This study focused on customer data generated (internally and externally) and stored (or used) by South African retail banks that require specific technology and analytical methods to transform data into value. This used Scholz's (2017) definition of data, a digital resource that can be generated and stored in a relational database. De Mauro et al. (2016) expand on this definition where data is an information asset that requires specific technology and analytical methods to transform data into value.

2.2.2 *Data lifecycle*

Khan et al. (2014) propose a general data lifecycle that incorporates big data terminologies that describe the different stages in a general data lifecycle, depicted in Figure 2-1.

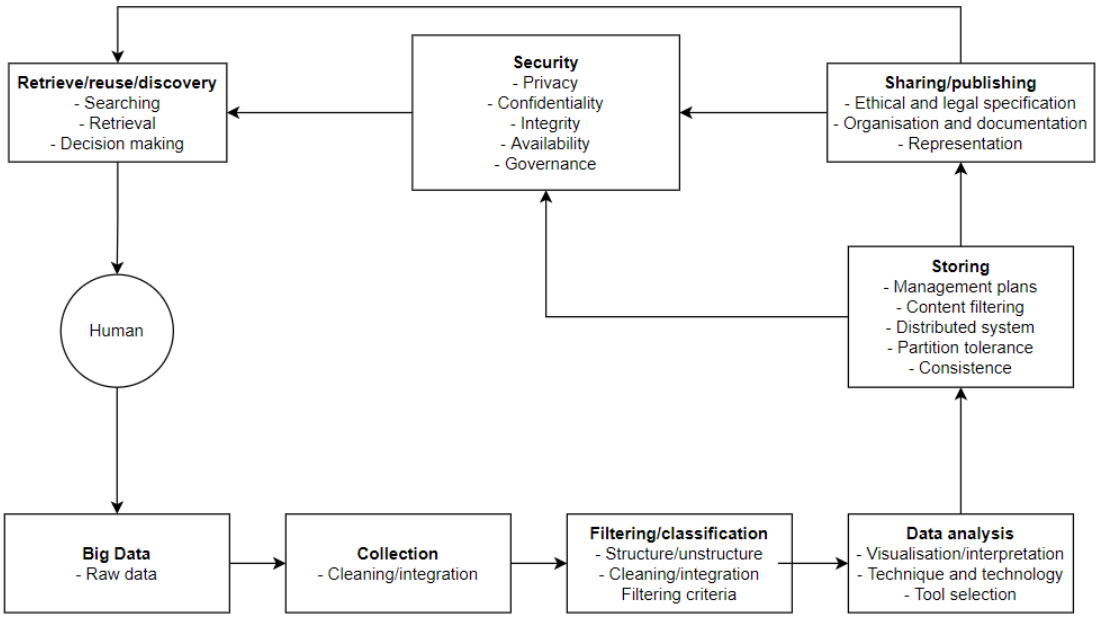

Figure 2-1: Proposed data lifecycle (Khan et al., 2014:9)

Khan et al. (2014) describe the different stages as follows:

- **Big data**: Where data are transformed and stored in a value-added state.
- **Collection**: The first stage of any data lifecycle is typically data collection or data generation. Depending on the type of data, different techniques can be used to get raw data from a specific environment.
- **Filtering/classification**: An important stage in managing data is the transition of raw data to published data processes.
- **Data analysis**: This stage enables organisations to manage the abundance of information that can impact their business. Data analysis can be challenging due to the data complexity that must be analysed and the scalability of the algorithms that support these processes. The two main objectives of data analysis are understanding the relationships

between the data attributes and developing effective data mining techniques that will enable the accurate prediction of future observations.

- **Storing**: The storage of data must meet the requirements of reliability, availability and easy accessibility.
- **Sharing/publishing**: The data must be analysed and shared or published to benefit the necessary audiences.
- **Security**: This stage encompasses various security considerations such as data security, governance requirements and agendas. An important consideration is to determine the appropriateness of the data type and its use when organisations are developing data, systems and tools, and policies and procedures to protect privacy, confidentiality and intellectual property.
    - **Privacy**: Organisations must safeguard and protect the privacy of their customers' information by preventing any possible identification of personal data (De Mauro et al., 2016).
- **Retrieve/reuse/discover**: The data retrieval stage enables the re-usage of currently existing data to discover new and valuable information. Once the data is published, users can validate the existing data and create new data points that align with their interests and needs to support their desired results (Khan et al., 2014).

2.2.3 *Data usage opportunities*

A larger amount of data has been created, shared and used recently. However, according to the data-information-knowledge-wisdom hierarchy, the information that results from structuring data to become useful and relevant with a specific purpose becomes knowledge that creates value for organisations (De Mauro et al., 2016). The basic competitive strategy for organisations is to capitalise on valuable knowledge.

By capitalising on this knowledge, organisations can gain advantages in operational efficiency, strategic direction, customer service, product development, and customers and markets (Khan et al., 2014). Many industries

make use of big data, such as, but not limited to: banking, insurance, telecommunications, healthcare, education and hospitality (Smalec, 2021).

As more data are being collected, the most important step is to process that data and extract valuable information to make the right decisions for the practical use of its conclusions. Therefore, it is important to understand possible data processing techniques.

2.3 An overview of Algorithms

2.3.1 *Algorithm definition*

This study focused on algorithms as the analytical method that is used to transform data into value. Cormen et al. (2022) describe an algorithm as a computational procedure that has been defined to take some value(s) as an input and produce some value(s) as an output. This is achieved through a sequence of computational steps to transform the input into the output. Therefore, this research report used Scholz's (2017) definition of an algorithm which is a tool that transforms data. Tsamados et al. (2021) expand on this definition: algorithms use data to create a specific decision that generates an action that may have ethical consequences.

2.3.2 *Algorithm development process*

Algorithms are reliant on training data to specify the correct outcomes for individuals. The algorithms learn from the training data to be applied to other individuals and generate predictions in terms of what the correct outputs or outcomes should be for these other individuals (Turner Lee et al., 2019). Barocas et al. (2017) use a simplified model to describe the different stages of a typical algorithmic system that produces outputs, depicted in Figure 2-2.

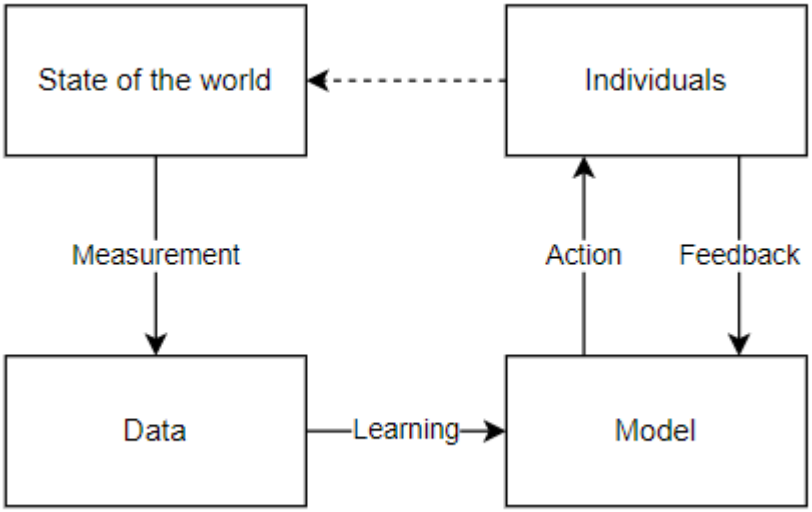

Figure 2-2: Stages of a typical algorithmic system that produces outputs (Barocas et al., 2017:15)

Barocas et al. (2017) describe the different stages as follows:

- **Measurement**: The "state of the world" is converted or translated into a dataset.
- **Learning**: The data are converted into a model, and the model summarises the training data patterns to make generalisations.
- **Action**: The model's predictions are used to trigger an action based on the model being inputted with new or unseen data. The underlying training data patterns that the model learned from can alter the "state of the world".
    - **Accuracy**: Typically, models are built to maximise accuracy by minimising both types of errors, the false positives and false negatives. However, there is the possibility that the errors may impact demographic groups differently, which organisations may not have considered.
- **Feedback**: Customers provide feedback based on their reaction to the action, which should be and is often used to refine the model.

2.3.3 *Algorithm uses*

Data can be used for a whole host of applications where benefits such as the potential for improved decision-making can be extracted. Algorithms are used to extract information from the data. This information can generate an informed to a fully automated decision (Olhede & Wolfe, 2018). Decision-making algorithms are increasingly being used in various departments and industries, which range from simple to complex models (Mittelstadt et al., 2016) to make key decisions (Tsamados et al., 2021).

Some examples of algorithmic systems include predictive policing systems that predict crime location and people involved (Richardson et al., 2019); automated journalism to create and disseminate news stories (Diakopoulos & Koliska, 2017). transactional fraud detection flags anomalies to normal customer behaviour (Rajaei, 2017); churn predictions that predict the intention of a customer to leave the bank (Sayed et al., 2018); and healthcare decision support systems that recommend diagnoses to physicians (Dey & Rautaray, 2014).

2.3.4 *Algorithm increased usage drivers*

According to Delen (2014), the key drivers behind the increasing use of algorithms can be grouped into three categories: need, availability and affordability, and cultural change.

Delen and Ram (2018) describe each of the drivers as follows:

- As customer expectations continue to increase, organisations demand to move towards relevant, data-driven, quicker and actionable insights.
- With the constantly increasing accumulation and creation of data inside and outside of the organisation, coupled with enhanced data collection technologies and data processing technologies, organisations can process large and complex data much faster and often in real-time.
- Pricing has also decreased due to cloud enablers allowing organisations to rent analytics capabilities and pay per usage.
- The organisational shift from subjective to data-driven decision-making.

As the use of algorithms continues to rise, consequential decisions will increasingly be made about individuals based on algorithmic outputs. Therefore, it is important to understand the ethical implications of algorithms and the potential risks posed to individuals and societies.

2.4 An overview of Bias

2.4.1 *Bias definition*

This study focused on the ethical concerns of bias and used Danks and London's (2017:4692) understanding where bias refers to ‘a deviation from a standard’. Danks and London (2017) argue that there are multiple categories of bias, such as moral and statistical biases. Moral bias refers to a judgement that deviates from a moral norm. Statistical bias refers to the representation of a certain group that deviates from the overall population. Statistical bias can aid in identifying moral biases. For example, many professions may exhibit gender disparities, for instance, a lower percentage of women in aerospace engineering but a higher percentage in speech and language (Danks & London, 2017). According to Danks and London (2017), statistical bias is exhibited in the professions but acts as a proxy to identify moral bias, which is the under-representation of women in certain professions. Barocas et al. (2017) expand on bias, referring to the demographic disparities in algorithms that should not be tolerated for societal reasons.

2.4.2 *Types of bias within an algorithmic value chain*

Silva and Kenney (2019) adopted a value chain by Danks and London (2017) that identifies sources of conscious or unconscious bias within an algorithmic value chain, depicted in Figure 2-3.

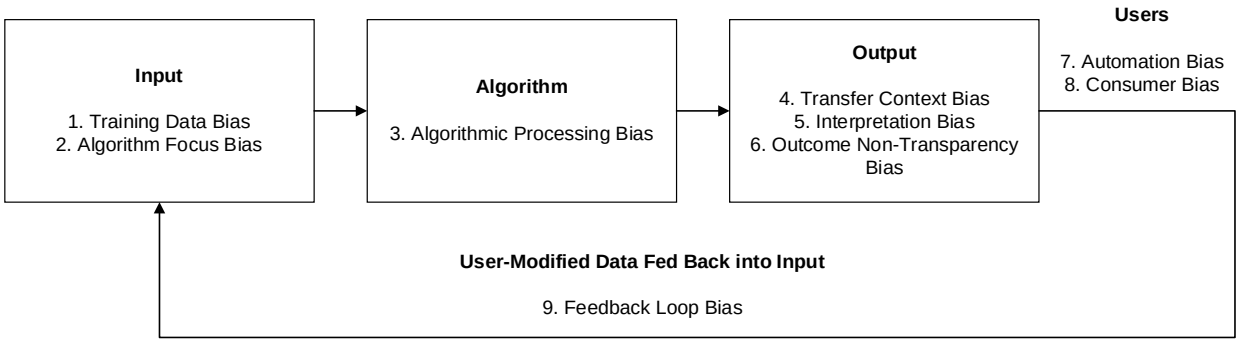

Figure 2-3: Potential biases in the algorithmic value chain (Silva & Kenney, 2019:38)

Silva and Kenney (2019) describe the different types of bias as follows:

1. **Training data bias**: datasets used to train the algorithms with underlying bias reflected in the algorithms.
2. **Algorithm focus bias**: the inclusion and exclusion of specific characteristics such as gender and race resulting in bias.
3. **Algorithmic processing bias**: bias that is programmed into the algorithm. For example, by placing a higher weighting on a specific characteristic.
4. **Transfer context bias**: where an algorithmic output is used outside of its original context. For example, by hiring candidates based on their credit scores.
5. **Interpretation bias**: when humans apply their own biases to an algorithm's output.
6. **Outcome non-transparency bias**: when humans do not understand the reasons behind an algorithm's decision, especially in AI and machine learning, due to lack of transparency.
7. **Automation bias**: when algorithm decisions are automated, without human intervention, due to trust being placed on the output to be factual as opposed to a prediction that contains a certain margin of error.
8. **Consumer bias**: human bias captured based on their daily online interactions. For example, conscious or unconscious discrimination of another human based on specific identifiable characteristics.
9. **Feedback loop bias**: human behaviour data trails created that get learned by algorithms. For instance, predictive policing based on historical crime

data to determine criminal hotspots will justify dispatching more police who will make more arrests creating a feedback loop.

Some examples of bias in algorithmic decision-making include, but are not limited to, gender-based discrimination in online recruitment tools, gender and race-based word association biases, race-based online advertisement discriminations, race-based facial recognition technologies discrimination and race-based criminal justice discriminations (Turner Lee et al., 2019).

This research focused on training data bias as bias in algorithms arises mainly from the data used to train algorithms (Shah, 2018) that may include past social inequalities (Richardson et al., 2019). According to Turner Lee et al. (2019), there are many bias causes, but two significant contributors include historical human biases and unrepresentative data.

2.5 Training data bias reflecting demographic disparities

The data used to train algorithms are rarely collected for a specific design and continue to be used despite having the potential to be biased and unrepresentative of the population (Olhede & Wolfe, 2018). Thus, the training data will reflect the demographic disparities in society. Numerous contributing factors could lead to these demographic disparities, such as historical discrimination, indirect views and stereotypes about certain demographics, and the distribution of certain demographics (Barocas et al., 2017). This can be exacerbated in machine learning algorithms that will learn from data that potentially encompass underlying patterns of discrimination (Binns, 2018) and unequal representation of the population (Zhang & Zhou, 2019), continuing the cycle of discrimination through the feedback loop (Barocas et al., 2017). This can lead to certain groups of individuals being unfairly denied access to certain financial products such as loans (Binns, 2018) or individuals being treated less favourably for having a particular characteristic (Castelnovo et al., 2020).

2.5.1 *Training data bias in banking*

Financial data is susceptible to bias and imbalance due to financial institutions collecting the data from their customers (Zhang & Zhou, 2019). Measurements about customers' demographic attributes are subjective and challenging due to how the demographic attributes are perceived and the changing perceptions of the demographic attributes, resulting in changes to what organisations measure. Alternative data points to customer-specific demographic attributes may be considered. However, most attributes tend to be correlations of or representations for demographic variables (Barocas et al., 2017).

As algorithms are created for a specific purpose, new categories must be created to define the target variable, which is the outcome that the organisation is trying to predict. Biases in a target variable's training data are guaranteed to bias the predictions as they are specifically created to inform the specific purpose or problem; for example, banks determine a customer's "creditworthiness", which is not a customer's inherent attribute (Barocas et al., 2017).

Algorithms have no general way of distinguishing between patterns in the training data that represent knowledge that organisations want to pursue and patterns in the training data that represent stereotypes that organisations want to avoid. Without intervention, the algorithms will extract both the knowledge-based patterns as well as the incorrect and harmful patterns (Barocas et al., 2017).

Retail banks are required to be compliant with many pieces of legislation. They are required to comply with the National Credit Act (NCA) to offer credit products to customers. The NCA includes requirements around protection against discrimination, directly or indirectly, in line with The Promotion of Equality and Prevention of Unfair Discrimination Act (PEPUDA) (Republic of South Africa, 2006). PEPUDA explicitly highlights three categories where unfair discrimination is prohibited: race, gender and disability (Republic of South Africa, 2000). This research focused on gender and disability bias as racial bias has been widely researched.

2.5.2 *Gender bias*

In the past, and ongoing in some cultural contexts, females and males have performed different roles in economies and societies. Female employment has typically been concentrated in unpaid or paying household and caretaking work perceived as lower-skilled work in the public sector. In contrast, male employment has typically been paid outside the home, either in the private sector or higher-skilled positions in the public sector (van Staveren, 2001).

Gelb (2003) presented past statistics from Stats SA showing that, between 1995 and 1999, females earned less than males and participated less in the labour force in South Africa. With females today earning higher incomes and playing a more significant role in household financial decisions (Zhang & Zhou, 2019), it is important to understand the impact of gender bias in retail banking algorithms when creating product offerings.

Proposition 1: The inclusion of the gender variable as an input to South African retail banks' sales-based decision-making algorithms creates gender biased product offerings that will not meet the customers' needs.

2.5.3 *Disability bias*

The World Health Organization estimates that approximately 15% of the world's population lives with some form of disability (World Health Organisation, 2020). According to Bunbury (2019), there is a perception that a disabled person's autonomy is limited due to their disability. The consequence of this perception is a limitation on their ability to participate in society if not cured by a medical professional.

However, a lot can be learned from disability social models, which views a predicted difference in a disabled individual's ability to excel in their role without the appropriate accommodations instead of their inherent capabilities. The individual is only disabled because the physical environment has not been catered for or appropriate policies have not yet been adopted to enable equal participation. Changes to a dataset in terms of using an individual's disability

attribute will not address the injustice of conditions that prevent certain individuals from contributing as effectively as others (Barocas et al., 2017).

Anastasiou and Kauffman (2013) expand that capitalism and the demands for increased productivity have contributed to the marginalisation of disabled people due to the perception that they are not economically productive. This creates both financial and social inequality (Bunbury, 2019). Therefore, it is important to understand the impact of disability bias in retail banking algorithms when creating product offerings.

Proposition 2: The inclusion of the disability variable as an input to South African retail banks' sales-based decision-making algorithms creates disability biased product offerings that will not meet the customers' needs.

2.5.4 *Imbalanced data*

Imbalanced data describes an unequal representation of groups. There are two ways in which imbalanced data occurs during the data collection phase: customer distribution and nature of the event. An example of "nature of the event" includes areas such as fraud detection where the number of true fraudulent events is significantly smaller than valid events resulting in imbalanced data. An example of "customer distribution" includes financial institutions having a greater number of male customers than female customers (Zhang & Zhou, 2019). This study focused on customer distribution due to the focus on sales-based decision-making algorithms that retail banks use when creating product offerings.

Wentzel et al. (2016) found that five demographic attributes are most significantly associated with the financial exclusion (not having a bank account) for South Africans with a Living Standards Measure score of less than four: education level, the primary source of income, age, home language, and the number of dependents supported by the respondent.

Under-representation of a certain group of individuals, irrespective of demographic factors, could result in algorithms generating outputs based on the majority's requirements (Zhang & Zhou, 2019), leading to algorithmic outputs that

are not aligned with customers' needs. This is due to algorithms performing better when there is more balanced data. Therefore, the algorithms would not perform as well for customers in minority groups as algorithms will generalise based on the majority resulting in a higher error rate for the minority group (Barocas et al., 2017). Turner Lee et al. (2019) mention the inverse where algorithms with an overrepresentation dataset would also skew the algorithmic prediction towards a specific outcome. Therefore, it is important to understand the impact of imbalanced data in retail banking algorithms when creating product offerings.

Proposition 3: The inclusion of imbalanced data as an input to South African retail banks' sales-based decision-making algorithms creates irrelevant product offerings that will not meet the customers' needs or will lead to financial exclusion of minority group customers.

2.5.5 *Additional forms of training data bias*

There is a phenomenon referred to as **data drift**. This phenomenon describes population characteristics shifting over time. Subpopulations changing at a different rate over time could lead to disparities (Barocas et al., 2017).

There is another phenomenon referred to as **data leakage**. This phenomenon describes an inflated estimate of the algorithm's performance. This could be due to the training data and test data not being independent of one another resulting in the algorithm being overfitted to the test data (Barocas et al., 2017).

2.6 Consequences of bias

Algorithms are deployed with the intention of organisations acting on the outputs. A significant limitation of algorithms is that they reveal correlations; however, organisations act on the algorithm's outputs as though they revealed causation. The outputs, therefore, can impact the outcomes, future training datasets and/or society. The more pervasive algorithms become, the stronger the impact (Barocas et al., 2017). Exploring the intentional and unintentional algorithmic consequences is crucial as current policies may not be sufficient in identifying, mitigating and remedying the impact on the consumer (Turner Lee et al., 2019).

In terms of impact on individuals, Barocas et al. (2017) refer to two types of consequences:

- **Allocative harms**: The decision-making systems are discriminatory, resulting in opportunities being withheld for certain groups of individuals. These effects would be realised immediately.
- **Representational harms**: This is where decision-making systems reinforce the subordination of certain groups of individuals based on demographic attributes by perpetuating stereotypes. These effects have long-term impacts.

In terms of impact on the organisation, Barocas et al. (2017) mention reduced profits, and Bryant et al. (2019) argues brand and reputational damage.

2.7 Controls to mitigate training data bias

There are techniques available to mitigate societal biases; however, there are fundamental limitations regarding what can be achieved when considering algorithms as a socio-technical system instead of a mathematical abstraction. Textbooks refer to the ideal being training and test data that is independent and identically distributed as a simplification but might be practically unachievable (Barocas et al., 2017).

Turner Lee et al. (2019) identify a few bias detection and intervention techniques:

- Identifying variables that are proxies for sensitive attributes. For example, the usage of a postal code as a proxy for race. Using these proxy variables in the algorithm's training data could lead to incorrect and discriminatory outcomes.
    - A technique that could be used to identify these proxy variables is to use the sensitive attributes to detect and mitigate the usage of proxy variables.
- Testing the algorithmic outputs for anomalous results, including comparing outcomes for different groups of individuals. This can be done systematically (i.e., via simulations) prior to deploying the algorithms into

production. Investigations can then be performed after the testing, if required, to determine if the outcome is truly unfair. Generally, it is not possible to have equal error rates for all groups of individuals. Therefore, interventions may be required in terms of which error rates should be equalised to establish fair outcomes across all groups of individuals.

- Developers to identify ways to reduce disparities between groups of individuals without sacrificing the algorithm's overall performance; for example, adding in additional training data for datasets that may be under-representative of certain groups of individuals.
- Presence of ethical frameworks and governance standards for the ethical use of algorithms. The European Union mentions seven governance principles in their ethics guidelines, including human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; environmental and societal well-being; and accountability.
    - The limitation is that legislators often become aware of shortcomings after the event has happened with the majority of laws written prior to the advent of the internet. Legislative action could trigger an action to create the best practices that organisations need to adhere to to prevent algorithms from exacerbating explicit discriminations.
- Reliance on regular independent audits for bias in the end-to-end process, from the input data to the output decisions.
- Reliance on cross-functional teams and expertise in the organisation for reviews on decisions to identify potential bias prior to deploying the models into production, especially if the decisions could result in risks that could harm certain groups of individuals if acted upon. Experts can be from different departments, disciplines and possibly sectors to facilitate discussions around accountability standards and strategies for mitigating bias. These cross-functional teams can be both internal to the organisation as well as external experts.

- Continued reliance placed on human judgement to moderate the decisions made by algorithms, to identify and correct biased outcomes, especially in edge cases.

Barocas et al. (2017) identified a few algorithmic interventions for the use of biased variables or variables with high correlations to biased variables:

- Data cleaning techniques such as omitting the demographic variable or the variable with correlation to demographic variables as a predictor. The limitation to this is that it could have a negative impact on the model's accuracy. It is also noted that in real datasets, most variables tend to be proxies for demographic variables, so omitting these variables may not be a feasible approach.
- Another data cleaning technique mentioned is the inclusion of an additional diversity criterion. However, the limitation is that the model could weight this irrelevant attribute as significant, impacting the model's accuracy.
- Another intervention technique is to select different cutoffs or thresholds for different groups of individuals so that there are equal probabilities for positive or favourable outcomes. The limitation is that there is no mathematical method to determine which cutoff to select. Certain situations are reliant on legislation as a guide.

2.8 Analytical framework

2.8.1 *Theoretical framework*

The social constructivism theory supports this study. Social constructivism plays a role in shaping data. The social constructivism theory argues that particular social groups can shape social reality if others adopt new agreed-upon norms through social interaction. Data can repeat further and transfer these social realities onto other individuals and generations (Scholz, 2017). According to Barocas et al. (2017), algorithms are deployed with the intention of organisations acting on the outputs. The outputs, therefore, can impact the outcomes, future

training datasets and/or society. This implies that as algorithms become more pervasive, the stronger the impact. One of the consequences referred to by Barocas et al. (2017) is representation harms, where decision-making systems reinforce the subordination of certain groups of individuals based on demographic attributes by perpetuating stereotypes. These effects will have long-term impacts.

This research aimed to understand how underlying bias in historical datasets used to train decision-making algorithms used in South African retail banks influence customer product offerings.

Social constructivism is relevant to this study as it relates to the potential to perpetuate further bias against certain groups of individuals due to past social inequalities or financial exclusion factors.

2.8.2 *Conceptual framework*

The two components of this framework are data bias and imbalanced data; both will influence the algorithmic output that generates the South African retail banking product offering (Figure 2-4).

The data bias comprises gender bias and disability bias (Figure 2-4). These data bias factors are internal to the customer and influence sales-based product offering decisions. These will be interpreted from the bank's perspective.

The imbalanced data can be made up of various characteristics but will be looked at holistically in terms of the under-representation of a population. The under-representation of a population can be linked to banking processes and rules that have previously resulted in the financial exclusion of individuals. These will be interpreted from the bank's perspective.

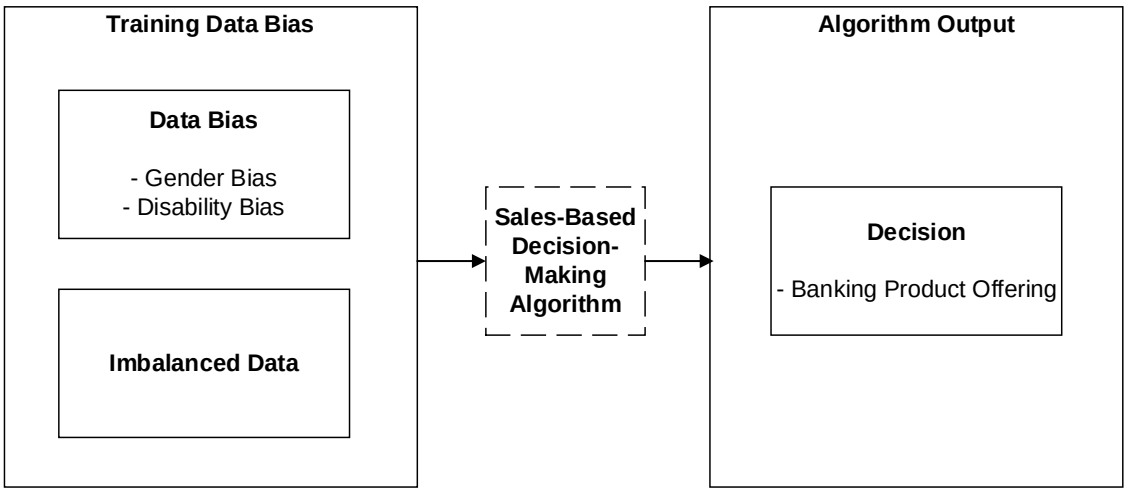

Figure 2-4: Banking product offerings' algorithmic decisions analytical framework

2.9 Conclusion of literature review

With increased data being collected, organisations need to structure this data so that it becomes valuable for the organisation. The most important step is to process the data and extract this valuable information to make the right decisions for the practical use of its conclusions.

Many industries are using algorithms to reap the benefits of faster decision-making at scale to help meet increasing customer expectations. However, algorithms are not immune to bias introduced at various stages in the algorithmic value chain, with the most significant contributor being inherently biased data used to train algorithms.

South African retail banks must understand the consequences that could arise from the increased use of algorithms to drive sales-based decision-making based on biased training data, as this could further marginalise certain groups of individuals.

There are techniques available to mitigate societal biases; however, there are fundamental limitations in terms of what can be achieved. Failing to incorporate these techniques into the banks' processes or addressing biased training data could result in the risk of banks further perpetuating discrimination at scale.

Proposition 1: The inclusion of the gender variable as an input to South African retail banks' sales-based decision-making algorithms creates gender biased product offerings that will not meet the customers' needs.

Proposition 2: The inclusion of the disability variable as an input to South African retail banks' sales-based decision-making algorithms creates disability biased product offerings that will not meet the customers' needs.

Proposition 3: The inclusion of imbalanced data as an input to South African retail banks' sales-based decision-making algorithms creates irrelevant product offerings that will not meet the customers' needs or will lead to financial exclusion of minority group customers.

3. RESEARCH METHODOLOGY

This chapter describes the research methodology followed to test the propositions made in Chapter 2. It explains the research approach, design, and instruments used for data collection. The end of the chapter discusses the research's transferability, credibility, and dependability.

3.1 Research approach

A qualitative study was adopted for this research to address the research questions. Qualitative research involves questions and responses that are open-ended (Creswell & Creswell, 2017). This approach was appropriate for the study as it allowed for a deeper understanding (Queirós et al., 2017) on the impact of bias on product offering algorithmic outputs.

The research was exploratory to probe the data science teams' understanding of data bias and imbalanced data's influence on sales-based decision-making algorithms' outputs. This was done to understand the influence of data bias and imbalanced data and gain deeper insights into the processes followed, controls in place, and gaps present.

3.2 Research design

This study used a qualitative approach that is underpinned by an interpretivist philosophy as it allowed for understanding from participants' perspectives (Merriam & Tisdell, 2015). An interpretive design was appropriate for this research as it helped in deriving greater understanding and deeper insights into the impact of bias on sales-based algorithmic outputs.

The research was conducted through virtual one-on-one interviews with semi-structured questions to understand better the participants' perspectives on the impact of bias on sales-based algorithmic outputs. This method allowed for additional perspectives to be discovered relating to bias in algorithmic decision-making.

The particular South African retail bank's complaints data was also collected to further explore the impact of bias from the customer's perspective. The complaints data was used to support the findings from the participants, representing triangulation.

3.3 Data collection methods

The data was collected through virtual one-on-one in-depth interviews with participants in the sample. The interview questions were pre-defined prior to the participant's interview; however, the nature of the interview was conversational. Greater insights can be derived from a semi-structured interview as it provides the opportunity for follow-up questions, probing for additional information and requesting justification for answers provided (Queirós et al., 2017). The interview guide was iteratively enhanced after each participant's interview as new information and insights were introduced.

This research intended to gain a deeper understanding of the potential bias found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings. Therefore, having open-ended questions and allowing for follow-up questions, probing questions and justifications for answers provided was essential to enhance understanding.

The particular South African retail bank's complaints data was also collected and analysed. From January 2019 to December 2021, the complaints data was analysed to further explore the impact of bias from the customer's perspective. The keywords used to filter the unstructured data included "bias", "disability", "discrimination", "gender" and "racist". The complaints were specific to customers who were declined credit product offerings.

3.4 Population and sample

The research focused on banking professionals from data science teams employed by a particular bank. The study population comprised skilled professionals across age groups, demographics and gender. Interviews were conducted with several analysts, managers and data scientists.

3.4.1 *Population*

The study population comprised banking professionals from data science teams in business units across three South African retail bank (Figure 3-1).

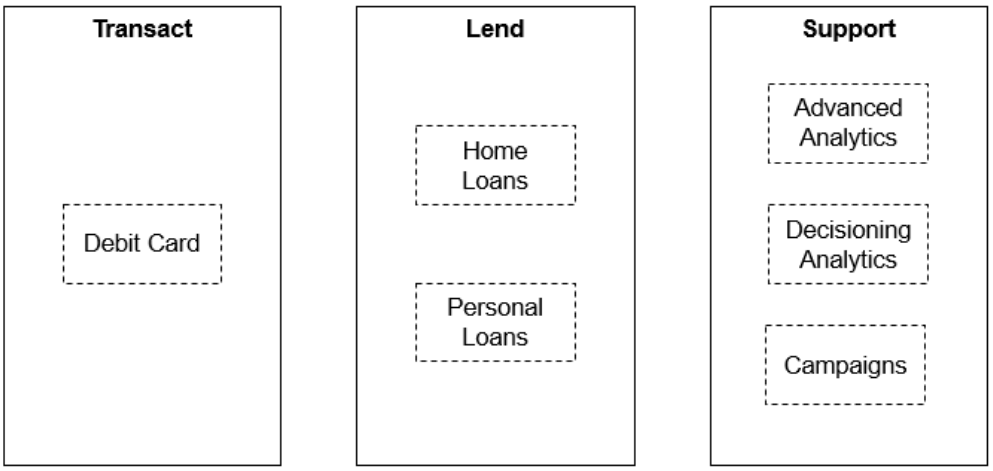

Figure 3-1: Retail banking in South Africa pillar view with product examples

3.4.2 *Sample and sampling method*

The purposive sampling method was used as it enabled the selection of a sample where the greatest insights would be discovered, understood and gained due to their expertise and experience (Merriam & Tisdell, 2015).

The study sample comprised banking professionals from data science teams in Product House Business Units across the different pillars in a particular South African retail bank. Therefore, purposive sampling allowed for the selection of these professionals. The remaining banking professionals were excluded from this study.

Reaching saturation is described as the most significant criterion in terms of sample size adequacy. Saturation is reached once no new insights or issues emerge during the data collection process (Hennink et al., 2017). Green and Thorogood (2018) suggest that little new information emerges after 20 interview participants. Therefore, the intended sample size for this study was 20 people. However, ten participants were interviewed due to the observation of saturation,

where no new insights or issues emerged relating to the research question and objectives (Hennink et al., 2017).

Permission was obtained to interview banking professionals from data science teams in Product House business units for a particular South African retail bank (Appendices E and F) from the Retail Segment Chief Risk Officer and Head of Legal. Email requests were sent to the participants. Participants who responded were interviewed.

3.5 The research instrument

An interview schedule was used to conduct the semi-structured interviews and gather responses from participants. The interview schedule comprised of four main sections, which addressed some general questions followed by gender bias, disability bias and imbalanced data that influence sales-based algorithmic outputs (Appendix (A)).

The questions on the interview schedule included a spread between open-ended and specific questions. The semi-structured interview allowed for additional questions if further clarification was needed or new insights were introduced during the interview.

The interview guide was iteratively enhanced after each participant's interview as new information and insights were introduced.

3.6 Procedure for data collection

The particular bank's permission was obtained to interview the participants from data science teams within Product House Business Units (Appendix (G)). Email requests were sent to participants. MS Teams meetings were set up to conduct the virtual one-on-one interviews with the willing participants.

The context was provided to the participant in terms of the research purpose and what the data would be used for (Appendix (D)). The participant was notified that the interview session would be recorded so that the interview could be

transcribed. Participants were asked questions as described in the interview schedule. Additional questions outside of the interview schedule were asked based on their responses.

The interview data was recorded using the data collection platform, and the recordings were transcribed and used for thematic analysis and coding. The voice recordings were not published. Notes were also taken during the sessions.

3.7 Data analysis and interpretation

Participant responses were analysed using inductive thematic analysis. The analysis was conducted following guidelines by Braun and Clarke (2006):

- The interviews were voice recorded and transcribed into text.
- The transcribed text was summarised and analysed to create an initial list of codes.
- Codes were grouped if similar, used as-is or discarded if found to be irrelevant to the study.
- Themes and sub-themes emerged through the grouping of similar codes. The themes and sub-themes were reviewed for relevancy, leading to instances of further separating, combining, creating or deleting themes and sub-themes.
- Once the final list of themes and sub-themes had been finalised, there were instances where some were re-named to convey the context of the theme and sub-themes in relation to the study.
- These themes and sub-themes were used to complete the study's data analysis to discuss the bias factors influencing sales-based algorithmic decisions.

3.8 Limitations of the study

- Due to the banks' confidentiality clauses, some of the answers provided may be inaccurate.

- The sample size was not large enough to be representative of the population.
- Geographic bias due to all participants being based in Johannesburg, Gauteng.
- Obtaining one-on-one time with data science participants in the organisation.
- Potential gap in the data collected due to the population interviewed only including data science team participants.
- The data collected could have been impacted by how certain questions were asked, including some leading questions.

3.9 Transferability and dependability

3.9.1 *Transferability*

Transferability refers to the extent to which the study results can be transferred into other contexts with different participants. This can be achieved by contextualising the research and using purposive sampling (Anney, 2014). The research context was specific to gender bias, disability bias and imbalanced data bias in training data to train sales-based decision-making algorithms used in South African retail banks to create customer product offerings.

The sampling method used was purposive sampling targeting banking professionals from data science teams in Product House Business Units across the different pillars in a particular South African retail bank.

3.9.2 *Credibility*

According to Anney (2014), credibility determines whether the research findings reflect the participants' original data and whether the interpretation of their views has been made so correctly. This can be achieved through peer debriefing, where the researcher seeks support and guidance from other scholarly professionals to improve the quality of their inquiry findings (Anney, 2014).

This research was able to maintain its credibility by leveraging the support and guidance provided by the research committee and an appointed research supervisor.

3.9.3 *Dependability*

Dependability refers to the findings' stability over time (Anney, 2014). This can be achieved by establishing an audit trail. All participant interview transcripts were retained, and the research steps were transparently reported.

3.10 Demographic profile of respondents

High-level demographic information was collected from the participants, as it was believed that they would influence the results, as follows:

- Gender
- Business unit
- Role title
- Employment equity assignment
- Tenure at the organisation
- Duration of data sciences or data and analytics experience

3.11 Ethical considerations

This research adhered to the ethical research standards set out by the University of the Witwatersrand, Wits Business School.

Permission to conduct the study was obtained from the South African retail bank. Written consent (Appendix (C)) or recorded verbal consent was obtained from each interview participant. Each interview participant was informed of the scope, the purpose and the outcomes of the study. Each interview participant was given the option to pull out of the research at any point and all information gathered from that participant would be destroyed. The interview transcripts were not published. Participant names and personal identifiers were not published. The bank and its participants' anonymity and confidentiality were maintained.

4. PRESENTATION OF RESULTS / FINDINGS

4.1 Introduction

Ten one-on-one interviews were conducted with banking professionals in data science teams who build models for retail in a South African retail bank. The interviews' duration ranged between 25 minutes and 1 hour 23 minutes.

The participants were spread across the lend, transact and support pillars (see Figure 3-1 in section 3.4.1). The participants' roles were Quantitative Analysts, Data Scientists and Analytics Managers. The profiling of the participants is graphically represented in Figures 4-1, 4-2, 4-3 and 4-4.

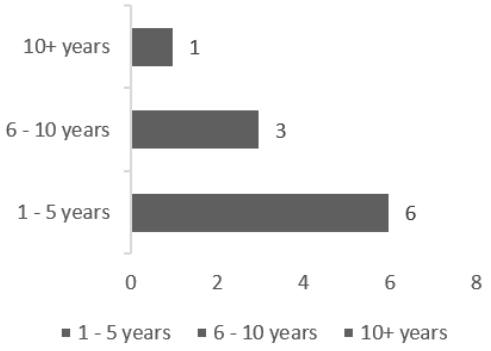

Figure 4-1: Participants' data science experience

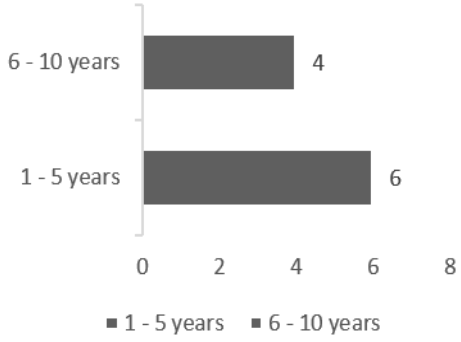

Figure 4-3: Participants' tenure in the organisation

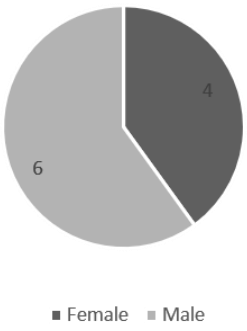

Figure 4-2: Participants' gender

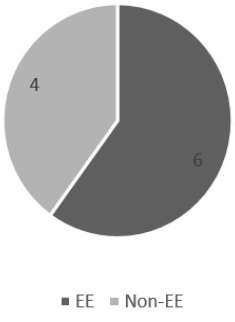

Figure 4-4: Participants' employment equity status

The interviews aimed to explore how underlying bias in training datasets used to train sales-based decision-making algorithms used by South African retail banks influence customer product offerings. The participant responses were analysed using inductive thematic analysis. A consistency matrix (Appendix (B)) was used to ensure the data collected, research questions, interview questions and analysis method was consistent. The thematic framework for how the research questions will be answered is depicted in Figure 4-5. It should be noted that the dark grey shaded blocks focus on the main research question (i.e., the themes for this study), whereas the light grey shaded blocks are additional insights that emerged during the interviews.

The following themes emerged from the data and will be discussed:

- Data-driven approaches to creating (personalised) customer products
- Causes of biased algorithmic outputs
- Consequences of biased algorithmic outputs
- Controls that mitigate biased algorithmic outputs

4.2 Data-driven approaches to creating (personalised) customer products

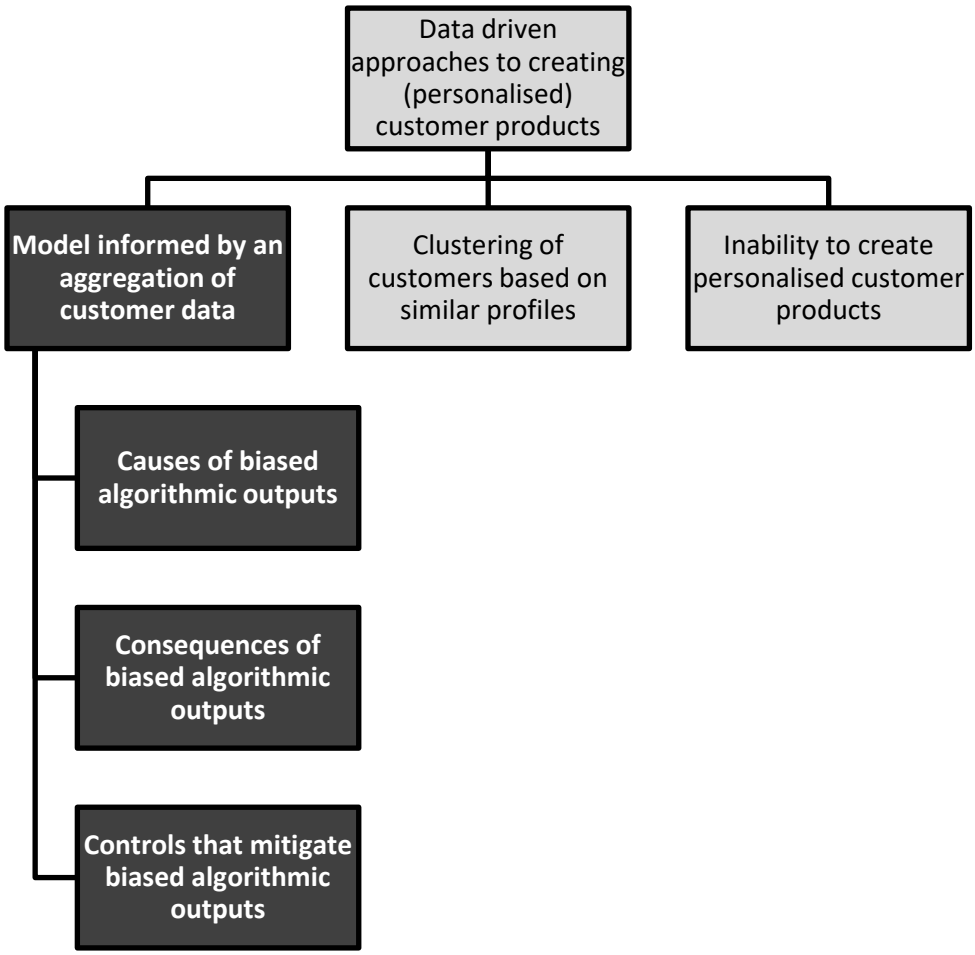

Figure 4-5: Thematic framework

Despite not being a specific theme for this study, it was important to contextualise how the particular South African retail bank used **data-driven approaches to create customer products**, which were found to be:

- Models informed by an aggregation of customer data
- Clustering of customers based on similar profiles
- Inability to create personalised products

As the focus of this study is on **algorithms that offer products to customers**, models informed by an aggregation of customer data were unpacked further to gain a deeper understanding of potentially biased algorithmic outputs and their impact. Therefore, it was necessary to understand the causes, the consequences and the controls that mitigate biased algorithmic outputs.

The causes of biased algorithmic outputs were found to be:

- Training data selection bias
- Organisational-related control gaps

It was important to understand the direct impact of biased algorithmic outputs and their resulting impact externally and internally. Therefore, the consequences of biased algorithmic outputs are categorised as:

- Incorrect predictions or biased algorithmic outputs
- External: customer impact
- Internal: bank impact

It was also important to identify the controls that were in place to mitigate training data selection bias, which are categorised as:

- Testing variables for correlation to other variables
- Dataset being representative of populations (including inclusivity of populations)
- Data cleaning
- Reliance placed on individuals
- Systematic monitoring
- Legislation and policies

Regarding approaches to creating data-driven personalised customer products, the interviews revealed that the business units seemed to differ in their approaches to creating these 'push to customer' personalised customer offerings. The approaches were models informed by an aggregation of customer data, clustering of customers based on profiles and the inability to create personalised offers (due to not obtaining the customer's marketing consent).

4.2.1 *Models informed by an aggregation of customer data*

Four out of ten participants mentioned that one of the approaches involved **using historic aggregated customer data** to train an algorithm **to predict for future customers**. Once the algorithm has been trained, individual customer data is then fed into the algorithm to create an offering for that specific customer based on similar customer behaviours. Example responses were:

> *'… to build algorithms, [the banks] generally use an entire like say customer base. But, for a customer, [the banks] use specifically those customer details.' (Participant 5)*

> *'... [models] get trained with an aggregation of a bunch of customer data... to train this model and hope is by doing that you can generalise [the model] to all customers there when it comes to the final data prediction stage.' (Participant 6)*

All the participants mentioned that the use of algorithms in the bank is increasing. The participants mentioned that some of the reasons or drivers behind the increased algorithm usage included automation (and the reduction or removal of human error), the complexity of problems, improvements in technology, increased data availability, competitive landscape (and maintaining market leadership) and profit-driven (including cost efficiencies). Example responses were:

> *'I think it's just because technology is improving. Technology is improving and the various types of algorithms in the accuracy is also improving, right? … So, I think with technology increasing inevitably*

your algorithms will increase as well because I think people rely a lot on machine learning and artificial intelligence as they are increasing because it removes that human error element.' (Participant 1)

'Wouldn't that always be like profit-driven, to make more money? Well, more I would say healthy money if that makes sense.' (Participant 3)

'I think it's the drive of the bank to always compete in the market and either continue their leadership in the market … Because, there's increased problems that's coming through, plus enhanced problems, so there's always new things to solve. So, maybe before you'd have a basic problem and [the bank will] try to solve [the problem] with one algorithm. But, with enhanced problems coming through, [the problem] requires, like maybe two or three algorithms to solve [the problem]. So, with the increased number of problems results in more algorithms.' (Participant 4)

'I think obviously the whole world is moving towards like automation and less human-based decisions. And it's obviously a lot more cost effective, a lot more efficient. So, as a result, they [are] investing a bit more into the use of algorithms to make these decisions.' (Participant 5)

'I think it's just the scope and the increase of data availability. So, because [the banks] have so much more data available that previously wasn't being used specifically … People are starting to realise the value that they can glean from data that previously, they weren't able to sufficiently analyse.' (Participant 6)

'I think the reason for the increases is the increase in the demand for data-driven solutions … In the current times that we're living in, the data tells [the bank] or gives [the bank] a third dimension of the customer today. It's a dimension that [the banks] don't necessarily know offhand. Yes, [the banks] know your customer age and all of these things because that's given to [the banks]. But when [the banks] look at [the individual] from a data perspective and not from a customer

perspective, then [the data] gives [the bank] a different dimension. [The data] tells [the bank] the type of customer and who and what type of customers are. Like [an individual], [the data] gives [the bank] the risk of this customer and gives [the bank] a lot of generic features that [the bank] would not necessarily know by foresight of just looking at an individual customer. And that's one of the biggest drivers for using data-driven solutions like this.' (Participant 10)

4.2.2 *Clustering of customers based on similar profiles*

Four out of ten participants mentioned another approach involving clustering of customers based on common attributes or behaviours. **Offers are then created based on similar products** taken up by customers with common attributes or behaviours. Example responses were:

'… [the banks] group customers in terms of [the banks] segment them(sic) to see if [customers] have the same behaviours or [customers] show the same spending patterns or whatever. You know, the same income groups… I mean, you don't know what you need to improve on or what you can offer to customers that [customers] don't currently have, that you need to see what your current base is already using.' (Participant 2)

'So, if a group of people happen to all fall within the same age range, or fall within the same area, or have the same products, or have the same spending habits, or have the same travelling habits. Uhm, then you'll cluster them to that particular group, and then you will then say, ok, this particular group can be sold this product.' (Participant 7)

4.2.3 *Inability to create personalised products*

Some participants mentioned that there are also instances where the bank is **unable to create 'push to customer' offers for customers due to not obtaining their marketing consent**. These customers would need to apply for

products the traditional way. This has resulted in a decrease in sales made by the bank.

> *'… customers are actually opting out, [the bank] actually reduced [their] offers so that actually reduces sales.' (Participant 4)*

4.3 Causes of biased algorithmic outputs

Several causes emerged that could lead to biased algorithmic outputs. The main cause revealed by the interviews, and mentioned by all of the participants, was that the **data selected to train the algorithms was the biggest contributor** to bias in the algorithm's decision-making when deciding whether to offer products to a customer or not. There were different issues revealed during the data selection process that could lead to the training data being biased, and these will be presented in-depth:

- The inclusion of testing data in the training dataset
- The use of biased variables, or variables with correlations to biased variables
- The use of imbalanced data
- Environmental changes that cause data to shift, once a model is deployed into production

In addition, organisational-related control gaps that emerged during the interviews could contribute to a biased algorithmic output. The causes that could lead to biased algorithmic outputs are depicted in Figure 4-6. It should be noted that the dark grey shaded blocks focus on the research sub-questions' specific components being studied, and the light grey shaded blocks are additional insights that emerged during the interviews.

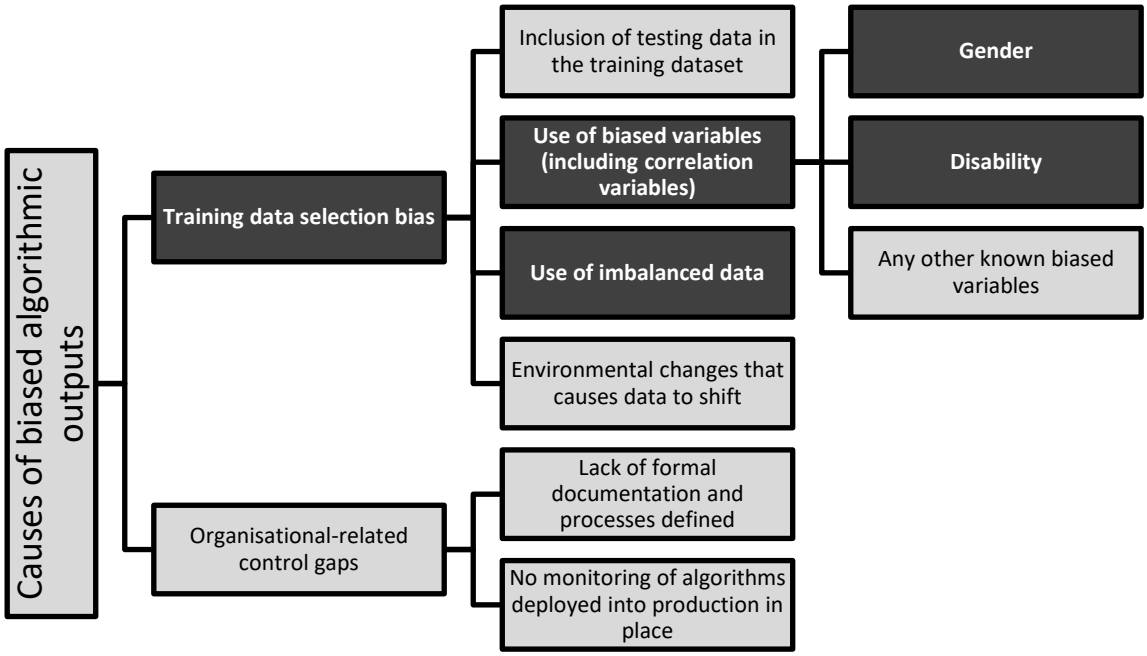

Figure 4-6: Causes of biased algorithmic outputs

4.3.1 *Training data selection bias*

Inclusion of testing data in the training dataset

Some participants mentioned that the inclusion of testing data in the training dataset, used to train algorithms, could result in biased algorithmic outputs when selecting data. The concern around including testing data in the training dataset is that the model's prediction accuracy would be questionable as **the model has already learned from the data that is being used to test** the algorithm's prediction accuracy. Some participant's responses were:

> *'Also, when training models and algorithms and all of that, typically you don't use the entire customer base. You kind of split it up into your training sets and testing sets and all of those things as well. So, you need to make those splits to make sure that you don't have any bleeding of data points from one data group to the other. Just make sure that your results are valid.' (Participant 6)*

> *'… you get data biasness that comes in based on the fact that somebody is using data that [the model] trained on to test data and*

then that causes a lot of biasness because the model is now biased. The model has learned everything on that train data and now you go and take the same data to test it; it's pointless. Which causes a biasness in the results.' (Participant 10)

Use of biased variables

Eight of the ten participants mentioned using biased variables and variables with high correlations to biased variables was one of the issues that could result in biased algorithmic outputs when selecting data. The concern raised with the use of biased variables was that **the model would focus on these trends and use it as a contributor** to inform the model's prediction, **which could favour or disfavour certain groups of individuals**. Some participant's responses were:

'[Bias] could also be when you create the rules, there could be bias there… when you look at certain trends to determine decisioning or rules, and if you don't make sure that those variables, for example, aren't bias or correlated to known bias variables, bias could be introduced there as well.' (Participant 3)

'… there's something [the data science teams] test which is called structural bias, which is [the data science teams] include variables like age, gender, race, etc. and then [the data science teams] test to see is it biasing to any of those sub-cohorts.' (Participant 4)

'So, I think obviously when you're building a model, you definitely need to make sure that there's no bias in your data that you're building it on. I think you could introduce unexpected bias when you know you could have variables that actually correlated to, for example, a specific race or specific gender, which may cause this unnecessary bias. So, it's always important to understand what sort of variables you're using in your model.' (Participant 5)

Several known biased variables were mentioned by the participants, including gender, race, age and geographical location.

Gender

Seven of the ten participants referred to gender as a known biased variable that should not be used to inform decision-making by the bank. None of the participants indicated any specific reason for the exclusion of the gender variable other than governance that was preventing them from using the variable. The participants elaborated:

> *'… it's the red tape that prevents us from using gender. There was always underlying understanding that you never used it before.' (Participant 7)*

> *'… that's just something I knew from early on, or even before POPIA, that you know, gender and race are variables that I have to stay away from.' (Participant 8)*

> *'[The data science teams are] not allowed to use specific variables which are gender-based or race-based or stuff like that. So, even in the current modelling, [the data science teams] don't necessarily use things like that.' (Participant 10)*

Five of the ten participants highlighted that **other variables could lean towards determining or identifying gender and that these need to be investigated and addressed** if found to have a correlation to gender. For example:

> *'There could be other variables that [the bank] might not be aware of, but I am aware that [other variables] could also come back to be a bit biased in like gender bias … It's possible that in its nature might come back to actually, sort of, you know, without using the gender variable, but [other variables] might come back to actually lead more towards one gender than another… [the data science teams] would have to actually investigate to see if you know, that was actually the case. Then, if that's the case… it might mean that you need to remove that [other variable] and use something else.' (Participant 8)*

One of the ten participants mentioned an opposing view around **bypassing the governance and red tape** that the bank has in place to **use another variable that can lead to a customer's gender:**

> *'You can work around that. You can get certain variables that are able to identify a particular person X as this particular gender … are you saying that I can't use those underlying datasets that lead me to say this particular group belongs to genders… I don't think there's anything that's put in place to prevent you from doing that. No, from my experience, no.' (Participant 7)*

Additional questions were posed to the participants relating to the gender variable with regards to the use of gender in the algorithms and gender-based complaints (refer to section 4.4.2 for the analysis of complaints). Only one of the participants indicated that gender was used in an algorithm that offers products to customers due to gender being a significant feature that differentiates customers' behaviour. However, the commentary provided by the participant indicated that the participant was referring to the retail industry and not necessarily banking:

> *'So, I think from expert opinion, and even from models, you will pick up [gender is] one of the significant features that differentiate between behaviour between people in terms of their sales behaviour, the things that they buy, the things that they focus on, all of those things… I think there is certain divergence in terms of interest and the data there.' (Participant 6)*

<u>Disability</u>

None of the participants mentioned disability as a known biased variable. Further, eight of the ten participants mentioned that **disability is not a data point collected from customers**, for example:

> *'To be honest, I don't even think there's a disability indicator in [the bank's] data. I haven't ever seen that.' (Participant 1)*

'… [the banks] don't have a variable that identifies the disability.' (Participant 8)

Therefore, it would also **not be possible to determine if variables were being used that could be correlated to disability**, as one participant proffered:

'Technically, [the bank] wouldn't be able to check against [disability]… Technically, if [the banks] don't capture the information, [the bank] can't check [disability] against anything, whether it's correlation or predictive power or something like that. So theoretically, there could be a variable that is correlated to [disability], but, because [the banks] don't even have that information, [the bank] won't be able to check.' (Participant 3)

Additional questions were posed to the participants relating to the disability variable with regards to the use of disability in the algorithms and disability-based complaints (refer to section 4.4.2 for the analysis of complaints). None of the participants indicated that disability was used in an algorithm that offers products to customers due to disability not being captured from customers, as confirmed by participant 2:

'Yeah, [disability is] definitely not an input field.' (Participant 2)

Any other known biased variables

Six of the ten participants referred to age and geographical location (including the derivatives, i.e., postal code and area code) as known biased variables, while five of the ten participants mentioned age as a known biased variable. However, one participant indicated that it was a **necessary bias as age is being used as a protective variable for older customers:**

'Age most definitely is a biased type of indicator or variable… So, if a customer is above 60, [the bank is] not likely to offer them a personal loan, right? So, and now, there is the conversation of, is that the right thing to do? Because that's bias. [The bank] won't offer [customers] a

product because you're a bit too old because there's a risk that you might not pay it back.' (Participant 1)

'It would be age … The younger you are, the better offer you get. It would be almost to protect your customers that are on [a] pension, you know, so it's more like a protective variable.' (Participant 3)

Four of the ten participants mentioned geographical location as a known biased variable that should not be used as **it can lead to customers' specific demographics**. For example:

'Sometimes [customers'] postal code can be biased, so sometimes [the bank] will explicitly tell us, check the correlation between these variables with any no go variables, if I can call it that. (Participant 3)

'So, [the bank has] this thing called area rating. So, they rate specific suburbs. The problem with that is it creates a bias where you know certain demographics live in a specific suburb. So, it's almost an unwanted bias … that can come up in the model building process and then, I think obviously, as you produce these offers, you create unwanted bias in the market where you know you might offer it to [a] specific demographic.' (Participant 5)

Use of imbalanced data

Six of the ten participants mentioned that using imbalanced data was one of the issues that could result in biased algorithmic outputs when selecting data. The concern raised with the use of imbalanced data was that **the model would be uncertain when responding to or predicting for individuals who were under-represented** in the training data, which could result in incorrect product offerings being made. Some participant's responses were:

'I would say [the biggest contributor to bias is] the data that's available. So, in particular cases, there might be a group that's also very overrepresented versus a very small minority group that is not as represented … When you're making a prediction model, a lot of the

time when the model is uncertain, it might end up just providing the same answer it would give for the majority rather than now for this specific edge case.' (Participant 6)

'Bias could be introduced where maybe some customers, [the bank doesn't] have enough data on some customers. So, maybe those that [the banks] don't have enough data on or enough insights might not be offered more like the same products as those that we do have a lot of data on.' (Participant 8)

An additional question was posed to the participants relating to the under representation-based complaints (refer to section 4.4.2 for the analysis of complaints).

An interesting suggestion was raised by one of the participants around the potential **use or creation of different algorithms for different population groups** as a potential **solution for inclusivity of minority groups or the groups of individuals with insufficient data:**

'[Customers] where there's no, you know. Maybe, they're new to bank customers where you know there isn't enough data on and stuff. Try and, you know, try and have certain maybe algorithms that are specific to those [customers]. So, you know, have different algorithms for different population groups.' (Participant 8)

Environmental changes that cause data to shift

Five of the ten participants mentioned that environmental changes that cause data used in the training dataset to train algorithms to shift were one of the issues that could result in biased algorithmic outputs when selecting data to inform models deployed into production. The participants mentioned the introduction of new populations (that wasn't previously included in the training data), customer financial, lifestyle behaviour changes and COVID-19. These **natural occurrences should be expected and addressed** as these **changes impact the algorithm's prediction accuracy**. For example:

‘So, you might find that maybe your out of time sample, there's certain populations, or subpopulations that have shifted and your model might not be suitable for those populations… so, you might find that over time [the banks have] actually through different offerings, [the banks have] now brought in this new population into [the bank].’ (Participant 5)

‘There is a phenomenon called data drift. Where basically, the data changes that your model was based on. So, because the data is now different to what [the model] was originally trained on, the model isn't performing as well, and this is something that just naturally happens… so, because your base data is changing, your model is also not necessarily training as well unless it's been trained on more data … It's quite important to kind of pick up those change points.’ (Participant 6)

‘The way you set up your data here was once off using the period that was used to test and train. But, going forward, the code has to be changed to pull in today's data to predict for tomorrow or this month’s data to predict for next month… I think it is essential given the day and age we’re living in and the change of data given COVID-19. It's the numbers that [the banks] see now. You know, it's not entirely the same: people’s lifestyles, spending patterns, living patterns have changed drastically, so [the banks] should be expecting the data to change drastically.’ (Participant 10)

4.3.2 *Organisational-related control gaps*

Lack of formal documentation or processes defined

Three of the ten participants mentioned that specific documentation around the prohibition to use certain data points or customer attributes was missing. There is a concern that this **might result in inconsistent processes being followed and controls being implemented** by data science teams in different business units in the same bank:

'… it's not in the process to say, listen, [the business units] need to consider gender to campaign to these specific customers, but there's also not something in place to say. You know, I've never seen or heard anyone say [the business units] can't prioritise gender. So, I think there can be work done in terms of putting proper controls in place.' (Participant 1)

'You can work around that. You can get certain variables that are able to identify a particular person X as this particular gender … are you saying that I can't use those underlying datasets that lead me to say this particular group belongs to genders… I don't think there's anything that's put in place to prevent you from doing that. No, from my experience no.' (Participant 7)

One of the participants raised **uncertainty around roles and responsibilities regarding monitoring algorithms and whether that should be a business responsibility or a data science team's** responsibility to execute:

'I'm also not sure where does [monitoring] needs to sit. Would it be something that needs to sit with us? I guess.' (Participant 9)

Incorrect model selection due to potentially insufficient guidelines

Four of the ten participants mentioned that incorrect model selection could result in biased algorithmic outputs **where the model could favour a certain group of customers over the other population groups**. One of the participants indicated that there are procedures that data science teams can follow; however, noted that **there would need to be awareness around these procedures:**

'… if you choose the wrong model. If your parameters make you select the model selection process wrong. Uhm, the model you may favour, may tend towards a particular group, particular cohort, or particular answer, relative to what a more mutually inclusive or not mutually inclusive, [the] model would work, so to speak.'(Participant 7)

'Statistical biasness comes about when you don't follow assumptions for parametric models, as an example. Which causes multicollinearity and things like that, which causes biasness in estimation … the person needs to understand what they're doing … There are procedures in place that, you know, take care of it, but obviously one has to be aware of it.' (Participant 10)

No monitoring of algorithms deployed into production in place

One of the participants mentioned that there was a **lack of implementation of post-production algorithm monitoring** due to the initial satisfaction from the business based on the initial set of prediction results **without considering the possibility of incorrection predictions in the future**. This is a **cause for concern due to the environmental changes that occur naturally over time**, as mentioned in section 4.3.1 (on environmental changes that cause data to shift). The participant mentioned that **due to increased demands for other models to be created, the creation of post-production monitoring is not prioritised:**

'The solution was built and productionalised and whatnot, [business] was happy with the performance. Then, [monitoring] didn't seem as though [monitoring is] a big need that needs to be put in place. Then, other things come up [the data science teams] should prioritise then, you know, [the monitoring requirement] just falls away … So, no one is truly concerned about those situation[s] or the cases that [the models] miss and stuff.' (Participant 9)

One of the participants mentioned that post-production algorithm monitoring would not be in place unless a systematic monitoring system was created:

'In essence, it should be happening but, a lot of the time in the bank it doesn't happen unless there are monitoring system has been set up' (Participant 10)

Lack of awareness from business in terms of production algorithm monitoring

One of the participants mentioned that there might be **a lack of awareness from business regarding the importance of monitoring production algorithms**, resulting in instances of no monitoring of post-production algorithms, **despite being recognised as a known need** by the data science teams:

> *'Perhaps there hasn't been that urgency or hasn't been realised there's a need to start assessing. I'd say that's probably why it's not in place … I think it's a definite requirement. It has to be in place, and so if [monitoring is] recognised as a requirement like we say, then [monitoring is] something then that will have to be put in, you know, prioritised.' (Participant 9)*

4.4 Consequences of biased algorithmic outputs

Incorrect predictions or biased algorithmic outputs was the main consequence revealed by the interviews as presented in section 4.4.1. These incorrect predictions or biased algorithmic outputs would lead to internal and external consequences, as presented in sections 4.4.2 and 4.4.3.

4.4.1 *Incorrect predictions or biased algorithmic outputs*

All the participants mentioned that underlying bias in the training data would result in **incorrect predictions or biased algorithmic outputs**, leading to internal and external consequences, as indicated in sections 4.4.2 and 4.4.3. For example:

> *'If your model is built on a biased population that's not representative of this true walking population. Let's say that, or true customer base. That can have a negative impact on the decision-making if you can say that. So, [the biased dataset] could result in the algorithm making the incorrect decisions on those customers.' (Participant 5)*

> *'… your data set is not clean if your data set is not distributed. If it favours one particular group or ideal, you'll always have a biased model.' (Participant 7)*

4.4.2 *External: customer impact*

Eight of the ten participants mentioned that biased algorithmic outputs would have customer-related consequences. The participants mentioned **financial exclusion of customers, disadvantaging certain groups of customers and not addressing the customers' needs:**

> *'… a lot of these times, [algorithms] bring about inequalities that you already understood, but to have models and AI that builds and uses that information can be dangerous in the sense that a lot of the business' decision-making could tend to favour one particular group versus another … You could find that you're disproportionately disfavouring a particular group or gender or race simply because an AI told you.' (Participant 7)*

> *' So, where maybe [customers] that [the banks] don't have enough data on or enough insights might not be offered more like the same products as those that [the banks] do have a lot of data on … You not really addressing a customer's needs, so there might be some customers that you sort of, you know, automatically, you know, discriminate against. So, you might not really offer them the products, even though [customers] do have a need for them.' (Participant 8)*

> *'As a bank want to take the least amount of risk. And for those that don't have a steady income, yes, [current metrics used for scoring are] going to be a disadvantage for [certain customers] because [the banks] don't offer credit products and certain products at specific risk level.' (Participant 10)*

<u>Bias related customer complaints</u>

As mentioned in section 4.3.1, under gender, disability and use of imbalanced data, additional questions were posed to the participants around awareness of customer complaints relating to an unfair outcome due to their gender, disability or under-representation of the populations. All of the participants indicated that

there were **no complaints from customers regarding an unfair outcome** due to their gender, disability or imbalanced data.

The bank's complaints data, between the periods of January 2019 to December 2021, was analysed to further explore the impact of bias from the customer's perspective. The keywords used to filter the unstructured data included "bias", "disability", "discrimination", "gender" and "racist". The complaints were specific to customers who were declined credit product offerings.

In most instances, customer complaints that made reference to any of the keywords used were **declined due to the affordability criteria used by the bank**. An insight that emerged was around **the South African retail banks making use of different affordability criteria and potential use of different credit bureau data**. This provides an insight into why customers could be declined a credit product offering from one South African retail bank but approved by another.

Another interesting insight that emerged post-resolution of one customer's complaint was around a customer's admission that **they did not believe the bank to be biased;** however, it was **influenced by social media posts** around South African retail banks being biased against certain customers' demographics.

One of the participants raised an interesting point around incorrect predictions being aligned to the business' risk appetite to incur financial losses due to customers' inability to pay. The **risk appetite is informed by the number of false positives the bank is willing to accept in terms of financial losses**. However, there appears to be **no indication in terms of consideration of the impact on customers who would receive incorrect predictions**. One participant proffered:

> *'So, the cutoff will be determined on the risk appetite that the business is willing to take, right? Ideally, [the business units] look for the highest true positive rate and the lowest false positive rate, right? … because [the data science teams] are in business, [the data science teams]*

would ask [the] business what's the appetite of false positives based on the use cases that [the data science teams] have.' (Participant 10)

4.4.3 *Internal: bank impact*

Five of ten of the participants mentioned that biased algorithmic outputs would have bank-related consequences. The participants mentioned **reputational damage and financial impact**, including regulatory fines, profit decline and inaccurate allocation of funds towards organisational process improvements:

'You could get, obviously like regulatory fines. And then, I also think it could lead to bad business.' (Participant 3)

'If there's bias and it's publicly known, even if [bias] wasn't the intention of [the bank]. You might get customers that are subject to the bias not necessarily demanding stuff in [the bank], so you'll actually be reducing your sales.' (Participant 4)

'It would hurt you[r] business, you know, monetary and whatnot. But, also reputation-wise.' (Participant 9)

'The biggest concern for me would be financial constraints. Well, financial impact on the business … [The bank is] investing optimisation costs, operational costs to productivity costs. All of those costs attached to an incorrect prediction. To me, that's the biggest impact of biasness.' (Participant 10)

4.5 Controls that mitigate biased algorithmic outputs

Several controls were revealed from the interviews, all of which mitigate training data selection bias (refer to section 4.3.1) as presented in section 4.5.2.

4.5.1 *Controls that mitigate organisational-related control gaps*

There were no controls mentioned by the participants for organisational-related control gaps as the name implies; these are control gaps.

4.5.2 *Controls that mitigate training data selection bias*

A Sankey relational diagram was created on Atlas.ti that indicated overlapping controls against training data selection bias (Figure 4-7). These controls included testing variables for correlation, systematic monitoring, reliance placed on individuals, legislation and policies, dataset is the representation of populations (including inclusivity of population groups) and data cleaning techniques.

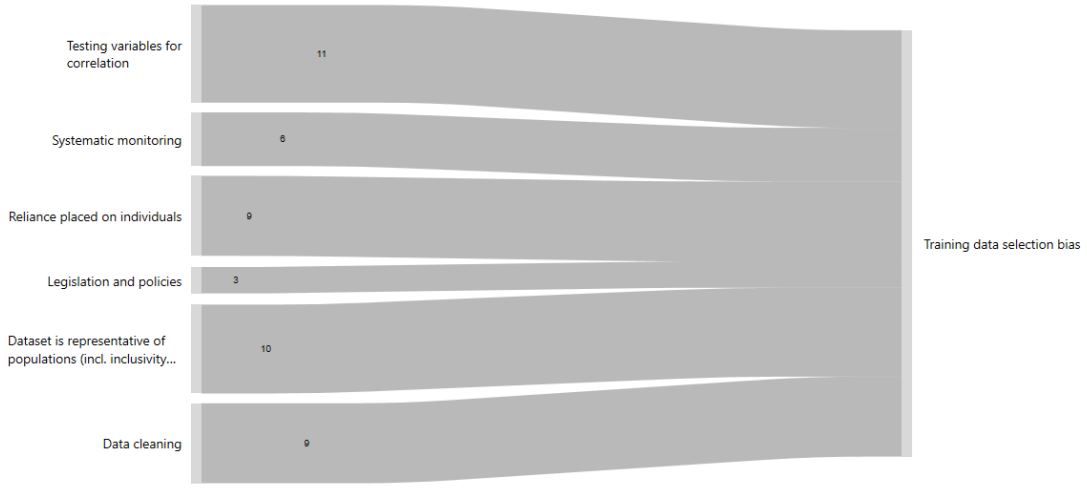

Figure 4-7: Relational diagram from Atlas.ti indicating overlapping controls against causes of biased algorithmic outputs

Testing variables for correlation

Six of the ten participants mentioned testing variables used in the training data for correlation to known biased variables and other variables as a control to mitigate data selection bias, specifically the use of biased variables (refer to section 4.3.1). The participants also mentioned that the algorithm's output is also tested for potential bias:

> *'What [the data science teams] do is when [the data science teams] build an algorithm, [the data science teams] include variables like race and gender, but [the data science teams] don't use it in the algorithm build. Why [the data science teams] keep it in is just at the end when [the data science teams are] testing, [the data science teams] just want to see is it biasing towards male or female for example. So yeah, so that's how [the data science teams] control it.' (Participant 4)*

'You need to make sure one, your data is not biased and or two, make sure the variables used in that algorithm are not biased. And yeah, that comes through testing and obviously testing that the output isn't biased.' (Participant 5)

'So, then you do a test for a correlation coefficient test. To test that the two variables are not correlated to each other.' (Participant 10)

Dataset is representative of populations (including inclusivity of population groups)

Nine of the ten participants mentioned the importance of representation and inclusivity of the populations in the training data as a control to mitigate data selection bias, specifically the use of imbalanced data (refer to section 4.3.1). The participants indicated that the representation of the populations required **is dependent on the algorithm's target output** such that **an adequate representation of the populations in the training dataset would need to be aligned to the distributions of the customer base that the model is intended to be used on** as this would enable the algorithm to make more accurate predictions. However, the participants emphasised **the importance of including sufficient data for all customer groups** so that **the model can be trained on sufficient data to predict for all groups of customers:**

'If it's representative of what you're experiencing and it's related to the target. So, for example, if I know that on average, all [the business unit's] customers default at a 2% rate, then when I'm selecting my population, I'll target that the population on average that I'm selecting also has a default rate of 2%. So, I'd say it's relating to your target and seeing if the population meets that target.' (Participant 4)

'The population must be representative of the population that you're going to use the model on … else you could cause issues in the way that the prediction happens or the accuracy of the prediction. So, I'd say it should, at worse, be representative of the population that the model is being used on; at best, it should represent the entire population in the industry and still be accurate.' (Participant 5)

'Where you have very minority groups, for example, they only make up about like 1% of the population that you're trying to represent. You might end up oversampling from that population just to make sure that they actually make a dent in the model being trained. The model needs to be presented with enough examples of something to be able to make a prediction on [customers] properly.' (Participant 6)

Data cleaning

The interviews revealed that there were many data cleaning techniques being used by the bank as a control to mitigate data selection bias (refer to section 4.3.1). The participants mentioned that these data cleaning techniques are dependent on the algorithm's problem statement:

'Yeah, so [data cleaning] depends on the problem.' (Participant 3)

'So, [data cleaning] all just depends on the type of model that you're building.' (Participant 5)

The participants mentioned data cleaning techniques that would solve for customers with missing data, outlier data, ensuring representativeness of the populations in the dataset and variables with correlation to known biased variables.

<u>Missing customer data</u>

Participants mentioned that other data points could be used to fill or determine missing customer values if a relationship exists between the two variables. Alternatively, the participants mentioned that missing customer data would be excluded from the dataset if the data was missing for an invalid reason.

'How can [the bank] maybe leverage a different variable to clean up that variable? So, I think a very easy example would be if age isn't populated properly, you could always leverage ID number to calculate it.' (Participant 3)

'… customer has a whole bunch of missing information. Uh, so generally, you'd want to exclude those fields or those customers from your base. Because, and [the data is] missing because [the reason is] for some invalid reason. Maybe [the reason] was an application loaded onto the workflow and [the application] was a duplicate application, so they then stop the process there.' (Participant 5)

'In some instances, if you don't have data for all the months. You might fill in the missing gaps by using, maybe you know, taking the previous value or the next value after that. In some instances, you might just fill the gaps with, we just average it, so the mean or the mode. So, there are instances like that where you have to deal with missing data.' (Participant 8)

Outlier data

Some of the participants mentioned that **outlier data would be removed if the data was not expected or realistic**. Some of the participants also mentioned that **it was necessary to focus on the use of ranges as opposed to a singular value**.

'When you get the data in its raw format, you probably want to analyse something that won't bring about any distortion to your results. So, for example, let's say I want to analyse the average loan amount customers come for and suddenly my data, there's a typo where somebody entered a customer wanted 20 billion rand. For example, if I average that out, I'm going to say, you know [the outlier data is] going to distort and bring my average really high. So, to avoid things like that, we need to clean the data to make [the data used] realistic. And, what [the bank] actually see for customers.' (Participant 4)

'Yeah, so just in some cases, for example, where you have an outlier value. So, a lot of the time, when you're training an algorithm or model or anything. You would try not to focus too much on age cases. So, you would like to have things within a certain range rather than just

focusing on one value in there that might skew your algorithm.' (Participant 6)

'So, I think from a personal experience, the cleaning of the data involves removing any white noise. So, if you [are] working with numerical data. If you have extreme values that do not necessarily conform to what is the statistical medians, statistical mean, that could throw off information, data set may not be normally distributed because of that. So, that's where the cleaning will be taking place then. It's also the most basic cleaning in terms of removing any weird or funny characters depending on the data set you're using.' (Participant 7)

Ensuring representativeness of the populations in the dataset

Some of the participants mentioned two scenarios where populations would not be adequately represented in the dataset: first, where the bank has no data on certain population groups who are applying for banking products; second, where the bank has minimal data on certain minority population groups. Where the bank has no data on certain population groups, some of the participants mentioned using methods such as reject inference to **use alternative data sources**, such as external credit bureau data, **to introduce data on the 'unseen' customer base**. Where the bank has minimal data on certain minority population groups, some of the participants mentioned the use of methods such as stratified sampling and oversampling **to ensure that there is adequate data on minority population groups so that the algorithm has enough data on those customers to make a prediction**. These data cleaning techniques would mitigate the training data selection bias, specifically the use of imbalanced data referred to in section 4.3.1. Responses from the participants include:

'Your out of time sample is slightly different to your training data set and what you might have to do then is, there's something called reject inference. So, where you find that there's a certain population that's not in your training data sets that you can bring in through this reject inference process. Or, you develop a new training set? Yeah, and

there's a lot of techniques that you can use to do that so you can make sure in your training set, you have a bigger population of this new customer base through things like stratified sampling. Or you can do something called oversampling as well.' (Participant 5)

'In the case where you have very minority groups, for example, they only make up about 1% of the population that you're trying to represent. You might end up oversampling from that population just to make sure that they actually make a dent in the model being trained. The model needs to be presented with enough examples of something to be able to make a prediction on [customers] properly.' (Participant 6)

'For the full population, a subset of the population that you're trying to predict generally very hard to predict if [the group of customers' dataset is] 5% and below and [the data science teams] need to use oversampling or stratified sampling to help with that.' (Participant 10)

Variables with correlation to known biased variables

Two of the ten participants mentioned specific techniques such as **the dilution or grouping of certain categories** being used to ensure that other variables in the training data, with correlations to known biased variables, **cannot be used to identify specific customers' demographics**. The participants indicated that the alternative technique is **to remove the variables with correlation to known biased variables:**

'Besides removing that variable, let's say [the variable is] a merchant spend category, you know you might be able to group [the variable] where you sort of dilute, you know, if you had, you know, a category that looked at like hardware stores or something like that, or DIY stores. Maybe, you might just throw in, I don't know, retail or something like that, where it might dilute the category a bit. Reduce the number of categories that you have, so [the variable] doesn't necessarily just point to a particular gender.' (Participant 8)

Reliance placed on individuals

All the participants mentioned instances where reliance is placed on individuals as a control to mitigate data selection bias (refer to section 4.3.1). The participants mentioned business unit and committee reviews, internal audit reviews, and advice or knowledge from senior management or internal networking connections.

All of the participants mentioned that the algorithms would need to be reviewed by executive committees or management committees, legal risk and compliance teams, and requiring formal presentation at committees **to receive expert opinion or feedback on the algorithms developed** by panel members. The participants indicated that **the formal presentation at committees is not a once-off occurrence but rather a requirement before proceeding to the next phase**. Participants also indicated that **new algorithm developments would require more approvals than algorithm enhancements**. The algorithms would also be documented and reviewed for final sign off. The expectation is that these individuals are experts in the bank who would be able **to use their expertise and vast institutional knowledge to assess the models being built** and provide the necessary feedback. For example:

> *'Everything goes through EXCO and MANCO and stuff like that. And usually, that's where they check the risk drivers and you know, the sales impact and stuff like that. Combined with that, [the data science teams] also had a meeting with legal risk and compliance where they also gave their input and concerns. And then, someone in my team had to take the changes through a technical committee. So, [the committee members] would be almost like the scoring and algorithm experts that had a look at what [the business units] want to do and does it make sense and does it align with everything the bank does? And then, other than that, I think it's just like documenting everything and getting sign off.' (Participant 3)*

> *'When something's existing and you want to improve it, you probably have to take your method across to various stakeholders for approval*

that this will work. Then, you then get their sign off that, okay, this is something [the business units] like and [the business units] want to test. So, you then develop the business case for [the model] and actually do the work to show the results of [the model,] and then you take that back to them to show them the results, and then [the various stakeholders] approve or not. When you're starting something new. It's [a] very similar approach, but yeah, it's a lot more forums you need to go to for approval.' (Participant 4)

Two of the participants mentioned that specific internal audits are performed to review and **replicate the model development process followed** by the data science teams as well as **assess the data used to raise or flag any potential issues identified**. According to Participant 5:

'All of the data and the whole model build actually gets audited. So, there's that. So, there's a whole audit process that these models go through. So, [the internal audit team is] a model validations team, then replicate your process, check your data, make sure there's no issues in there.' (Participant 5)

The participants mentioned that **the knowledge** from other data science teams as well as from managers **would be shared informally and on an ad hoc basis**. However, **it would only be effective if that knowledge was present in the bank**. There is a concern that **this form of communication may not always be effective** in reaching all data science teams **and may not be sustainable in terms of continuity**. This control is potentially linked to the organisational-related control gaps, specifically lack of formal documentation and processed defined (refer to section 4.3.2). Participants proffered:

'Someone from [another business unit] did raise for a different variable that we need to just double check it against any bias … so, sometimes they will explicitly tell us, check the correlation between these variables with any no go variables … but, I think because bias has been so important in the last few years or decades with data analysis you do have people that could guide you in the right direction. You know to

mitigate [bias], but I feel like almost that knowledge needs to be in your team or company.' (Participant 3)

'I would say that's where a lot of previous knowledge also comes in. because, the senior management or even people that work with [data] a lot, which are like, now [the data] looks a bit weird and I know this sounds terrible because it's not statistically sound.' (Participant 3)

'I think there are obvious biases that [the banks] are aware of … I just remember some of them because of the way [specific biased variables] were communicated to me; I guess it just came via my manager.' (Participant 9)

An interesting insight was raised by one of the participants around a potential gap that has resulted from over-reliance on individuals or experts in the bank with longstanding institutional knowledge. The participant indicated that there is no formal documentation or listings of models already created in the organisation and that **the only control to determine whether a similar model has been created is in a committee with the panel of experts who are expected to be able to recollect or be aware of all models executed in the bank based on memory**. The participant also indicated that there was no pre-approval committee prior to commencing the model building. There is a concern that this could result in wasted efforts:

'I've seen a solution not get approved because, you know, the sentiment from the committee was that there's already something that addresses this in part or entirely … you just rely on the people you talk to. If no one you know alerts you to the fact that something like [the model being built] exists, you definitely can spend a lot of time building something and then only down the line once you start going to these different committees, only discover that. Yeah, so there isn't a formal setting where stuff is pre-approved, no … I didn't get the sense that [the committee panel members were] picking it up from an official list. [The model inventory information is] just based on what the people in the room are aware of.' (Participant 9)

Systematic monitoring

Seven of the ten participants mentioned systematic monitoring in place as a control to mitigate data selection bias. The participants mentioned continuous re-calibration of the algorithm's training data with current 'up-to-date' data, presentation of the algorithm outputs' performance at business unit committees, and pre- and post-model checks.

The participants mentioned that **there would need to be continuous re-calibration of the algorithm's training data with the latest data** to mitigate the issue of environmental changes (refer to section 4.3.1, specifically on environmental changes that cause data to shift) that would include market changes and population tests. One participant proffered:

> *'So, every month [the data science teams] run a rolling three-month worth of data to almost, sort of, re-calibrate this model. And, the reason for that is because obviously like the market changes quite quickly. So, [the data science team] always want to refresh [the model] with the latest data … for all our models, yeah, to monitor all of them. So, population shifts, accuracy tests, you know." (Participant 5)*

The participants mentioned that the performance of the production algorithms' outputs is presented and discussed at business unit committees. **The outputs are reviewed in terms of their alignment with the results that the committee panel members expect**.

> *'[The business unit] have monthly committees where [the data science team] take the algorithm outputs each month to show how is it performing. If it's expected or not expected.'(Participant 4)*

One participant mentioned pre- and post-model checks in place. The pre-checks will **check whether the algorithm's input data fits certain parameters**. The post-check will **check whether the algorithms have run and whether it has run correctly**. The participant indicated that if an issue is encountered with the pre- and post-checks, this **issue or failure would trigger an investigation:**

'[The team who looks after centralised models are] basically monitoring whether or not the model has run and whether it has run as it should. And, they might, there are certain post-model checks that [the team who looks after centralised models] run that [the data science teams] have given them, but [the data science teams have] also installed pre-model checks as well, so that's another form of monitoring [the data science teams] do to say okay cool. Before [the data science teams] even run the model, can [the data science teams] check whether or not the information that [the data science teams] require that goes into the model itself, whether [the information would] fit certain parameters? If those parameters change then [the data science teams] know not to run the model because there's something wrong with the underlying information. Same thing with the post-model check. [The data science teams] have set standards within which the post-model [the data science teams] check runs. If, in some shape or form, the model fails either one of those checks, then [the data science teams] know something is wrong towards the model and [the data science teams] go back. So, in terms of monitoring and maintenance, [monitoring is] usually done with pre- and post-check.' (Participant 7)

One participant provided their view on why continuous monitoring is not yet in place. The participant mentioned that the main issue stemmed from **algorithm prediction outputs not being saved in such a way that the predictions can be audited or verified**. There is a concern that if this issue is not resolved, **this could potentially impact future requirements around the 'explainability' of model predictions**.

'I think the main problem that a lot of models suffer from is that their outputs aren't necessarily being saved in such a way that it can be checked later on where somebody can go back and do an audit of all the predictions on things that were made by model previously … There's a big push in the data science and machine learning communities at the moment for the explainability of models, where you need to be able to show what it is that your model is predicting on. So,

what features is it taking into account when making a final prediction so that you can see whether your model has bias or not and whether [the model is] focusing on the correct thing.' (Participant 6)

Legislation and policies

POPIA

All participants mentioned POPIA and the Act's significant role as a control to mitigate data selection bias (refer to section 4.3.1). The participants mentioned **several POPIA-related process changes that bolstered the bank's control environment,** including the teams are more conservative when sharing data, the obtaining of the customer's marketing consent, the requirement to anonymise the aggregated customer data used in the algorithms, the impact of adhering to retention periods for storing of customer data and requiring absolute justifiable reasons to use customer's attributes.

One of the participants mentioned that with the implementation of POPIA, the teams are more conservative when sharing data:

'People are more conservative in the way that they share data.' (Participant 1)

Some of the participants mentioned the requirement to obtain a customer's marketing consent to create 'push to customer' products offers. The **customers who opt-out of receiving marketing consent would be automatically excluded from the creation of personalised product offers**, as mentioned in section 4.2.3 (specifically, inability to create personalised offers). Among the participant's responses were:

'Unless the customer opted out for marketing consent from [the bank's] side, then [the business unit] definitely need to take that into consideration. But, I mean, if the customer is fine with [the bank] selling him products, then [the bank] can push that, [the bank] can analyse the data as much as possible to try and get the best solution for the customer.' (Participant 2)

> *'[The business units] actually sometimes have to give offers to customers … But, before doing this [the banks] actually have to ask the customer. Do [the customers] even want offers? And, can [the bank] use the information for offers? But, because of [POPIA], the customers are actually opting out.' (Participant 4)*

Some of the participants mentioned that the aggregated customer data used to train the algorithms must be anonymised. The participants indicated that **there would be very little identifiable information that would lead to a specific customer**. Additionally, the participants indicated that **the volume of customer data is so large that the data is, in most instances, viewed as aggregated values and not individual data points**. According to participant 6:

> *'In terms of just how it is being used and all of that, [the aggregated customer data] is very anonymised in the sense that most of the data that you use, you would not trace back to a specific person. So, a lot of the things that are being stored also is basically just random data points. So, it has values for the various features and things that you look at, but very little identifiable information where [the data] will lead you back to a person. And also, in a lot of cases, there is just so much data if you're aggregating it into large groups that you almost never take a look at every single item in isolation at any point. You only look at almost like aggregated values at the end of the day.' (Participant 6)*

One of the participants mentioned the impact around adhering to retention periods specified by POPIA and its impact on algorithms' prediction accuracy. POPIA specifies requirements or timelines around how long the banks are allowed to store a customer's data. The participant mentioned that **these restrictions have resulted in less data being available to use as training data** for the algorithms. The participant indicated that algorithms perform better with more data used to train the algorithm; however, **these legislative restrictions could have a potential negative impact on the algorithm's prediction accuracy:**

'I think it has a big impact in terms of historical data that [the banks] have available … The best thing to do is to look at the past and [customers'] historical behaviour in terms of certain things. But, with POPIA, [the banks] are very restricted in how much data [the banks] can store and how long and all of those things. So, I think it does have an impact in terms of [the banks] have less data at your disposal that you can use to train models.' (Participant 6)

Some of the participants have mentioned POPIA's requirement around the non-usage of customer personal information. Participant 7 indicated that after POPIA's implementation, there is a requirement around **the need to provide justifiable cause and a valid use case to use customers' personal information and special personal information:**

'There was always [an] underlying understanding that you never used [customer personal information] before, but with POPIA, [the use of customer personal information] has become, you know, hard-lined and there's a stake in the ground … Saying you can't use [customer personal information]. And if you can, you need justifiable reasons as to why.' (Participant 7)

Other legislation and internal policies

Only one of the participants mentioned reliance placed on regulation and internal regulation as a control to mitigate data selection bias (refer to section 4.3.1). The lack of participants referencing regulation and internal policies could be due to poor communication and potentially links to the organisational-related control gap regarding lack of formal documentation and processes defined (refer to section 4.3.2). As proffered by participant 1:

'Yeah, so I think that's where regulation like internal regulation plays a big role. So, that needs to be regulated and communicated.' (Participant 1)

4.6 Summary of results/findings

Three data-driven approaches to creating personalised ‘push to customer’ product offerings were identified from the interviews, namely models informed by an aggregation of customer data, clustering of customers based on similar profiles and the inability to create personalised products. As the research focused on algorithms that offer products to customers, a deep dive was performed on causes, consequences and controls that mitigate biased algorithmic outputs.

There were two main causes of biased algorithmic outputs: training data selection bias and organisational-related control gaps. The training data selection bias issues identified included including testing data in the training dataset, use of biased variables, use of imbalanced data and environmental changes that cause data to shift.

The interviews revealed that the consequences of biased algorithmic outputs included incorrect predictions or biased algorithmic outputs, an external customer impact and an internal bank impact. The controls mentioned in the interviews were identified to mitigate training data selection bias, including testing variables for correlation, systematic monitoring, reliance placed on individuals, legislation and policies, usage of a representative dataset and data cleaning techniques.

The research themes that emerged from the data linked to research questions and propositions are tabulated in Table 1.

Table 1: Tabulated research themes

RQ #	RQ	Prop #	Proposition	Themes
1	How do underlying bias in training datasets used to train sales-based decision-making algorithms used by South African Retail banks influence customer product offerings?			• Data-driven approaches to creating (personalised) customer products • Causes of biased algorithmic outputs • Consequences of biased algorithmic outputs • Controls that mitigate biased algorithmic outputs
1.1	How does gender influence product offerings' algorithmic decisions in Retail banking?	1.1	The inclusion of the gender variable as an input to South African retail banks' sales-based decision-making algorithms creates gender biased product offerings that will not meet the customers' needs.	• Causes of biased algorithmic outputs • Consequences of biased algorithmic outputs • Controls that mitigate biased algorithmic outputs

RQ #	RQ	Prop #	Proposition	Themes
1.2	How does disability influence product offerings' algorithmic decisions in Retail banking?	1.2	The inclusion of the disability variable as an input to South African retail banks' sales-based decision-making algorithms creates disability biased product offerings that will not meet the customers' needs.	• Causes of biased algorithmic outputs • Consequences of biased algorithmic outputs • Controls that mitigate biased algorithmic outputs
1.3	In what way does imbalanced data influence product offerings' algorithmic decisions in Retail banking?	1.3	The inclusion of imbalanced data as an input to South African retail banks' sales-based decision-making algorithms creates irrelevant product offerings that will not meet the customers' needs or will lead to financial exclusion of minority group customers.	• Causes of biased algorithmic outputs • Consequences of biased algorithmic outputs • Controls that mitigate biased algorithmic outputs

In the next chapter, these findings will be discussed in relation to existing research.

5. DISCUSSION OF RESULTS / FINDINGS

5.1 Introduction

The purpose of the study was to gain a deeper understanding of the potential bias found in the training data used to train sales-based decision-making algorithms used by South African retail banks to create customer product offerings.

With the rapid advancement of technological capabilities and competition, customer expectations have heightened (MoneyMarketing, 2019) in wanting personalised experiences from banks (Marous, 2021). Customers want offerings that address their core needs and focus on traditional banking products (W.UP, 2018). For banks to meet these customer needs, they must apply algorithms to meet these personalisation needs at scale (Martin, 2019). However, the training data used to train these decision-making algorithms could result in less favourable decisions made towards certain groups of individuals (Dickson, 2020). This could be due to unrepresentative and incomplete training data or inherently biased data (Turner Lee et al., 2019). The impact of unethical algorithms would result in a profit impact as well as reputational damage for the bank (Scanlon, 2020); the impact on the customer would be further marginalisation or financial exclusion for certain groups of individuals (Moosajee, 2019). The banks must address these challenges before deploying algorithms.

The specific results as depicted in an updated thematic framework (Figure 5-1) based on the previous chapter will be discussed and interpreted in relation to existing research to answer the main and sub-research questions:

Main research question: How do underlying bias in training datasets used to train sales-based decision-making algorithms used by South African retail banks influence customer product offerings?

Sub-research question 1: How does gender influence product offerings' algorithmic decisions in retail banking?

Sub-research question 2: How does disability influence product offerings' algorithmic decisions in retail banking?

Sub-research question 3: In what way does imbalanced data influence product offerings' algorithmic decisions in retail banking?

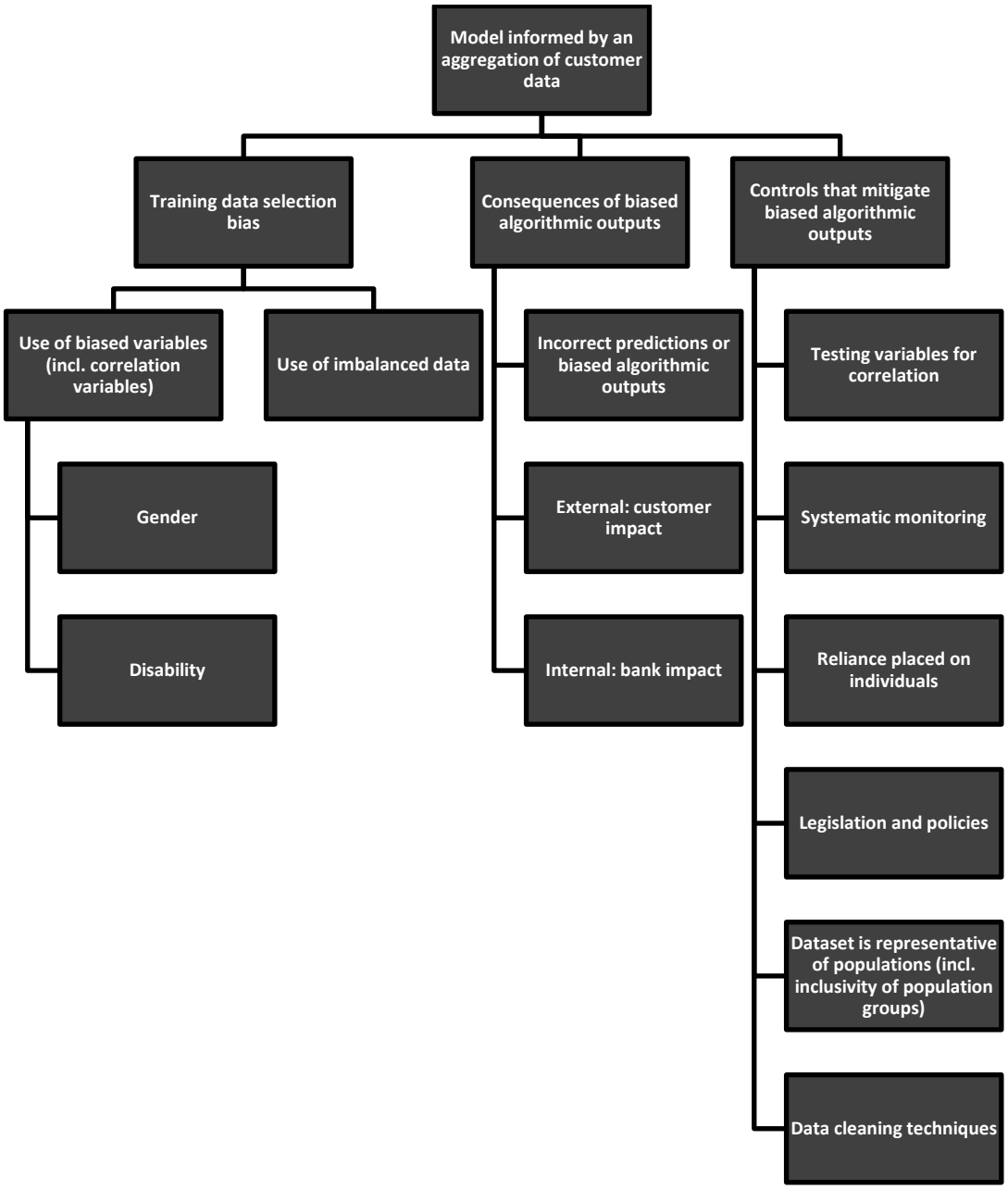

Figure 5-1: Thematic framework for discussion

5.2 Data-driven approaches to creating (personalised) customer products

The data suggested that historic aggregated customer data are used to train an algorithm to predict future customers. After which, a customer's data will be fed into the algorithm to create an offering specific to that customer, based on similar customer behaviours. The data support the theory, which indicates that algorithms learn from the training data to be applied to other individuals and generate predictions in terms of the correct outputs or outcomes for these other individuals (Turner Lee et al., 2019).

However, the banks need to ensure that this use of customer data is in line with the relevant legislation (for example, POPIA) in that the data used to inform the algorithms maintains the anonymity of the customers. This anonymising of customer data in the training data is a control in place in the bank, as discussed in section 5.5.1. According to De Mauro et al. (2016), organisations must safeguard and protect the privacy of their customers' information by preventing any possible identification of personal data.

The analysis identified automation, problem complexity, technological improvements, increased data availability, competitive landscape and profit as drivers behind the increased use of algorithms, which can be grouped in the same three categories as put forward by Delen and Ram (2018): need, availability and affordability, and cultural change. The details of the drivers from the analysis follow.

5.2.1 *Need*

The findings suggested that due to the financial institution being in a competitive environment, they must continuously evolve to meet customer needs to continue their market leadership. However, due to the changes in customer needs, the complexity of the problems has enhanced as well as the number of problems that need to be solved for customers; therefore, requiring more algorithms than would be required for basic problems.

According to Delen and Ram (2018), as customer expectations continue to increase, organisations demand to move towards relevant, data-driven, quicker and actionable insights. The theory supports this. Khan et al. (2014) also mention that the basic competitive strategy for organisations is to capitalise on valuable knowledge. By capitalising on this knowledge, organisations gain advantages in operational efficiency, strategic direction, customer service, product development, customers and markets.

The findings suggested that using algorithms would also increase healthy profits for the organisation. This is contributed to organisations using data to gain a third dimension of the customer (i.e., by creating customer attributes to gain insight on the customer, e.g., calculating a customer's "risk" score; as opposed to merely using one-dimensional generic customer attributes, e.g., the specific attributes that the customer gives the organisation such as their age, etc.). This third-dimensional view of the customer will enhance the bank's understanding of a customer and their needs and allow banks to offer products to the customer in a way that would not be harmful to the bank or the customer (e.g., reckless lending).

However, banks must be conscious of the bias in the training data as the training data would need to be specifically formulated to align to the target variable that the organisations are aiming to create. According to Barocas et al. (2017), the bias in the training dataset is due to organisations specifically creating target variables, for example, "creditworthiness", that are not inherent customer attributes. The banks should potentially explore this in greater detail regarding the bias that could occur.

5.2.2 *Availability and affordability*

The findings suggested that technology improvements aided improvements in algorithmic output accuracy, resulting in an increase in reliance on algorithms, machine learning and AI because of reducing or removing human error.

The findings also suggested increased data availability, including previously collected data but not specifically used or analysed. With the growing field of data and analytics, more individuals are being exposed to more knowledge in terms of

the value of data combined with the improvement of technologies; organisations are now shifting towards converting their data into knowledge.

This is supported by the theory, according to Delen and Ram (2018), with the constantly increasing accumulation and creation of data inside and outside of the organisation, coupled with enhanced data collection technologies and data processing technologies, organisations can process large and complex data much faster and often in real-time. De Mauro et al. (2016) also mentioned that a larger amount of data has been created, shared and used recently. However, argue that the information from structuring data to become useful and relevant with a specific purpose becomes knowledge that creates value for organisations.

There is one availability and affordability driver that Delen and Ram (2018) mention: pricing decreasing due to cloud enablers who enable organisations to rent analytics capabilities and pay per usage. However, the population selected in the sample may not be the same individuals involved in these types of discussions.

5.2.3 *Cultural change*

The data suggested that there has been an observable shift towards automation and reduced human-based decision-making. The shift towards automation also leads to greater efficiencies bringing about greater cost efficiencies, reducing cost to income and improving profits. This is supported by the theory, according to Delen and Ram (2018), of the organisational shift from subjective to data-driven decision-making. Mittelstadt et al. (2016) and Tsamados et al. (2021) also mention that decision-making algorithms are increasingly being used across various departments and industries, ranging from simple to complex models to make key decisions.

5.3 Training data selection bias

The data identified the use of biased variables, including other variables with high correlations to biased variables and the use of imbalanced data as factors that could lead to training data selection bias. This supports the theory as indicated

by Turner Lee et al. (2019), who mention there are many causes of bias, but two significant contributors include historical human biases and unrepresentative data.

Findings from the current study also revealed other factors that could introduce bias when selecting training data, include data leakage where the model is overfitted with the test data due to lack of or insufficient independence between the training and test datasets as well as data drift, which is a natural occurrence resulting from a shift in the data causing the model to produce incorrect predictions (though, the latter is more an issue for models deployed into production). This supports the theory that two additional training data selection bias causes are data leakage and data drift (Barocas et al., 2017).

Another phenomenon that Barocas et al. (2017) describe is the bias in the training dataset due to organisations specifically creating target variables, for example, “creditworthiness” that are not inherent customer attributes. The banks should potentially explore this phenomenon in greater detail as the participants discussed target variables. However, the biased training data selected to inform these target variables were not identified as concerns.

5.3.1 *Gender*

The data suggested that the gender variable had been identified as a known biased variable and that the bank had governance requirements in place that prevented data science teams from using the gender variable to inform decision-making algorithms. However, the data did not reveal any additional explanation or motivation for why the gender variable should not be used outside of governance requirements. This led to opposing views regarding using other variables to lead to the customers’ gender. This could stem from a lack of awareness of why gender should not be used.

According to Barocas et al. (2017), numerous contributing factors lead to demographic disparities, including historical discrimination, indirect views and stereotypes about certain demographics, and the distribution of certain demographics. van Staveren (2001) explains that females and males have

historically performed different roles in economies and societies where female employment has typically been concentrated in unpaid household and caretaking work perceived as lower-skilled work in the public sector. In contrast, male employment has typically been paid outside of the home, either in the private sector or higher-ranking positions in the public sector. According to Zhang and Zhou (2019), females today earn higher incomes and play a more significant role in household financial decisions. Thus, if the dataset remains imbalanced for gender, the algorithms could provide results that have a higher error rate for females due to algorithms generalising based on the majority's requirements (Barocas et al., 2017).

5.3.2 *Disability*

The findings suggested that disability was not considered a known biased variable. Additionally, the disability variable is not even a demographic attribute variable collected from customers. Therefore, the disability variable is not used to inform decision-making algorithms. However, due to not having the disability variable, the bank cannot test other variables used in the training data for correlation to a disability, resulting in the bank unknowingly biasing against disabled individuals by using a proxy variable.

According to Anastasiou and Kauffman (2013), capitalism and the demands for increased productivity have contributed to the marginalisation of disabled people due to the perception that they are not economically productive. However, Barocas et al. (2017) expand a predicted difference in a disabled individual's ability to excel in their role without the appropriate accommodations instead of their inherent capabilities. The individual is only disabled in the sense that the physical environment has not been catered for or appropriate policies have not yet been adopted to enable equal participation. Any changes to a dataset in terms of using an individual's disability attribute will not address the injustice of conditions that prevent certain individuals from contributing as effectively as others.

Therefore, it raises the question of whether banks should be collecting this information and doing more in terms of social responsibility to address these

social inequalities to help enable these individuals, i.e., offers to make their homes more suitable to enable productivity so that they can contribute as effectively as others.

5.3.3 *Use of imbalanced data*

The data suggested that the use of imbalanced data was one of the known issues that could result in incorrect algorithmic predictions made on certain groups of individuals under-represented in the training data. This is due to the algorithm providing the same answer it would for the majority as opposed to a specific edge case when the algorithm is uncertain.

This is supported by Zhang and Zhou (2019), who mention that under-representation of a certain group of individuals could result in algorithms generating outputs based on the majority's requirements and not aligning to the customer's needs. According to Barocas et al. (2017), algorithms perform better when there is more data. Therefore, the algorithms will generalise based on the majority, resulting in a higher error rate for the minority group. This could result in financial exclusion or incorrect products being created for certain groups of customers which, if not corrected, would lead to continued subordination of certain groups of customers based on demographic attributes.

The data also revealed the potential for using different algorithms for different groups of individuals to solve for minority groups. However, there would be a potential bias if algorithms are not consistently aligned when required post-production maintenance or updates. To support this claim, Barocas et al. (2017) mention that this is a technique that can be used, noting that there are technical and ethical challenges associated with this technique.

5.4 Consequences of biased algorithmic outputs

5.4.1 *Incorrect predictions or biased algorithmic outputs*

The data suggested that underlying bias in the training data would result in incorrect predictions or biased algorithmic outputs, leading to internal and

external consequences. This is supported by the theory of using biased variables and imbalanced data. In terms of using biased variables, according to Barocas et al. (2017), algorithms cannot distinguish between patterns in the training data that represent stereotypes that organisations want to avoid. In terms of imbalanced data, the algorithms would generate outputs based on the majority's requirements. These scenarios would lead to incorrect predictions or biased algorithmic outcomes for certain individuals.

Interestingly, despite the data suggesting that the use of algorithms is increasing, it did not reveal concerns around algorithms identifying correlations; however, the organisations are acting on these correlations as though they were equivalent to causation. Nor did the data reveal concerns around the continuous perpetuating of bias resulting from feedback loops. According to Binns (2018), Zhang and Zhou (2019), and Barocas et al. (2017), demographic disparities and under-representation can be exacerbated in machine learning algorithms learning from data that potentially encompass underlying patterns of discrimination and unequal representation of the population and continue the cycle of discrimination through feedback loops. According to Barocas et al. (2017), the more pervasive algorithms become, the stronger the impact.

5.4.2 *External: customer impact*

The data suggested that biased algorithmic outputs would have customer-related consequences: financial exclusion, disadvantaging certain groups of customers and not addressing customers' needs. This is supported by theory in terms of allocative harms, where the algorithms withhold opportunities for certain groups of individuals (Barocas et al., 2017).

The data revealed an interesting insight regarding incorrect predictions aligned to the bank's risk appetite to incur financial losses due to the customer's inability to pay. The risk appetite is informed by the number of false positives the bank is willing to accept in terms of these financial losses. However, there was no indication in terms of consideration of the impact on customers who would receive incorrect predictions. As mentioned in section 5.4.1, the data did not reveal representational harms where the algorithms reinforce the subordination of

certain groups of individuals based on demographic attributes by perpetuating stereotypes, resulting in long-term impact. This is a concern as the data suggested that the use of algorithms is increasing, and in line with Barocas et al. (2017), the more pervasive algorithms become, the stronger the impact. As per the social constructivism framework, particular social groups are able to shape social reality if others adopt new agreed-upon norms through social interaction. Data can repeat further and transfer these social realities onto other individuals and generations (Scholz, 2017). Banks should reconsider their risk appetite statements' considerations to potentially factor in the impact on customers and the representational harms of incorrect predictions.

5.4.3 *Internal: bank impact*

The data suggested that biased algorithmic outputs would have bank-related consequences, including reputational damage and financial impact. This is supported by theory as put forward by Barocas et al. (2017) in terms of reduced profits and Bryant et al. (2019) in terms of brand and reputational damage.

5.5 Controls that mitigate biased algorithmic outputs

The data revealed several controls that mitigated training data selection bias, including testing variables for correlation, systematic monitoring, reliance placed on individuals, legislation and policies, dataset is representative of the populations and data cleaning techniques.

5.5.1 *Testing variables for correlation to known biased variables*

The data revealed that this could include including certain known biased variables (such as gender and race) in the final stages of the algorithm's development to identify if other variables being used in the training data have high correlations to these known biased variables. The data revealed that two data cleaning techniques could be used to resolve the issue of variables with correlation to known biased variables. The first involves diluting or grouping certain categories

so that the variable cannot be used to identify specific customers' demographics. The second involves removing the variables.

The theory supports this; according to Turner Lee et al. (2019), a technique that could be used to identify these proxy variables is to use the sensitive attributes to detect and mitigate the use of proxy variables (for example, postal code, which could be a proxy for race). However, there is a limitation, as mentioned by Barocas et al. (2017), where most variables tend to be proxies for demographic variables; thereby rendering this technique infeasible.

5.5.2 *Systematic monitoring*

The data revealed that this would include presenting the algorithm outputs' performance at committees where the committee panel experts would review the outputs to determine if there is alignment with expected results. This is in line with Turner Lee et al. (2019), who suggest testing the algorithmic outputs for anomalous results through simulations. However, what the data did not reveal was comparing these outcomes for different groups of individuals. Though, Turner Lee et al. (2019) argue that it is not always possible to have equal error rates for all groups of individuals. Therefore, interventions may be required in terms of which error rates should be equalised to establish fair outcomes across all groups of individuals.

5.5.3 *Reliance placed on individuals*

The data revealed that there were different instances where reliance would be placed on individuals, including committee reviews, internal audit reviews, and advice or knowledge from senior management or through internal networking connections. The committee reviews would provide developers with expert opinions or feedback on algorithms throughout the algorithm development phase.

The theory supports this; Turner Lee et al. (2019) propose relying on the organisation's cross-functional teams and expertise for reviews on decisions to identify potential bias prior to deploying the models into production, especially if the decisions could result in risks that could harm certain groups of individuals if

acted upon. The internal audit team would review the end-to-end algorithm development process to flag potential issues. Turner Lee et al. (2019) further suggest relying on regular independent audits for bias in the end-to-end process, from the input data to the output decisions. It would be interesting to see what skills are present in the bank's internal audit teams.

The experiential knowledge from other data science teams in other business units, as well as from managers, would get shared informally. There was a concern about continuity or that the communication method was potentially insufficient to create the necessary awareness to all impacted developers. This is supported by the theory, similar to the committee reviews. According to Turner Lee et al. (2019), a technique is to place reliance on cross-functional teams and expertise in the organisation for reviews on decisions to identify potential bias prior to deploying the models into production, especially if the decisions could result in risks that could harm certain groups of individuals if acted upon. However, due to these informal interactions, not all developers may receive the same communication, resulting in unintentional or intentional biases slipping through in production without being picked up by experts or colleagues.

5.5.1 *Legislation and policies*

The data revealed that there was a big emphasis on POPIA from a legislation and policies perspective and other legislation but to a much lesser extent. The emphasis on POPIA was significantly around the anonymising of the aggregated customer data used in algorithms and requiring absolute justifiable reasons to use customer attributes, especially in committee reviews. According to Turner Lee et al. (2019), two of the governance principles put forward by the European Union in their ethics guidelines is privacy and data governance, as well as non-discrimination and fairness.

In terms of other legislation, there was no evidence in the data referencing other legislations and internal policies, which could be due to poor communication or lack of formal standards and guidelines documentation that would be deemed necessary to have in place in the bank. However, as argued by Turner Lee et al. (2019), the limitation is that legislators only become aware of shortcomings after

the event has happened. The majority of laws were written prior to the advent of the internet. Despite legislative action triggering an action to create the best practices to which organisations would need to adhere to prevent algorithms from exacerbating explicit discriminations. The banks do not necessarily need to wait for legislation and could be proactive in creating these best practices and sharing this knowledge with other banks and other organisations in other industries.

5.5.1 *Dataset is representative of populations (including inclusivity of population groups)*

The data revealed that often, the representation of the populations in the training data is dependent on the algorithm's target output; therefore, the training data would be manipulated to align to the algorithm's target output.

Barocas et al. (2017) argue that there are often scenarios where new categories must be created to define the target variable, which is the outcome that the organisation is trying to predict. This automatically biases the training data, which is guaranteed to bias the predictions due to the training data being specifically created to inform the specific target variable, for example, determining a customer's "creditworthiness", which is not an inherent attribute that customers have. The bank should investigate the potential impact of creating these training datasets to align to the target variable to ensure that they are not introducing unintentional or intentional biases.

The data also revealed the importance of including sufficient data for all customer groups so that the algorithms could be trained on sufficient data. Intervention techniques included introducing additional data points to the training dataset to ensure adequate data on minority population groups so that the algorithm has enough data on those customers to make a prediction. Turner Lee et al. (2019) point out that developers can identify ways to reduce disparities between groups of individuals without sacrificing the algorithm's overall performance, such as adding in additional training data for datasets that may be under representative of certain groups of individuals.

5.5.1 *Data cleaning techniques*

The specific data cleaning techniques that would resolve issues of using variables with correlations to known biased variables and imbalanced data have been discussed under section 5.9.1 and section 5.9.5.

5.6 Chapter conclusion

The bank uses historic aggregated customer data to train algorithms to predict for future customers. As a result of the implementation of POPIA on 1 July 2021, the bank's data science teams ensure that this aggregated customer data in the training data is anonymised to maintain the customer's privacy. Training data selection bias was identified as one of the main causes of biased algorithmic outputs, specifically using biased variables (including proxy variables for biased variables) and imbalanced data.

In terms of gender: The gender demographic attribute was identified as a known biased variable that the bank had prohibited data science teams from using to inform algorithmic decision-making. However, it was unclear in terms of whether the data science teams fully understood the reasons behind the variable's exclusion. This could lead to certain developers using proxy variables that could lean towards a customer's gender, introducing gender-based bias into some of the bank's algorithmic decision-making, impacting or influencing certain product offerings made to certain genders.

Another issue that could result in incorrect offers being made for females is dependent on the income disparities between males and females. Due to algorithms generalising based on the majority's requirements (Barocas et al., 2017), if females who earn higher incomes are not adequately represented in the training data and are a minority group, this could result in algorithms providing results that have a higher error rate for females.

In terms of disability: The disability demographic attribute was not identified as a known biased variable and was not a variable the bank was collecting from the customers. However, due to the bank not collecting customers' disability data,

the bank can also not test other variables for correlation to a disability, leading to potential unintentional bias against disabled customers.

There is a perception that disabled individuals are not as economically productive (Anastasiou & Kauffman, 2013); however, Barocas et al. (2017) argue that the challenge for disabled individuals to contribute as effectively as abled individuals are due to appropriate physical environment accommodations not being in place rather than their inherent capabilities. Therefore, it raises the question of whether banks should collect customer disability data and do more in terms of social responsibility to address these social inequalities and enable disabled individuals to contribute as effectively as abled individuals (i.e., creating equal participation in society).

In terms of imbalanced data: Imbalanced data was a known issue in the bank due to algorithms' tendency to generalise based on the majority's requirements resulting in a higher error rate for under-represented groups of individuals or minority groups.

The consequence of using known biased variables or imbalanced data is incorrect predictions or biased algorithmic outputs that would have an external and internal impact.

In terms of external impact: Incorrect predictions or biased algorithmic outputs could result in financial exclusion for certain groups of individuals, disadvantaging certain groups of customers and not addressing customer needs. The bank needs to ensure that the impact of representational harms, where the algorithms reinforce the subordination of certain groups of individuals based on demographic attributes by perpetuating stereotypes, is understood, and controls are implemented to mitigate these harms. As per the social constructivism theory, data can repeat further and transfer these social realities onto other individuals and generations (Scholz, 2017).

In terms of internal impact: Incorrect predictions or biased algorithmic outputs could result in reputational damage and decreased profits for the bank.

The bank has several controls that mitigate training data selection bias, including testing variables for correlation, systematic monitoring, reliance placed on individuals, legislation and policies, dataset is representative of the populations and data cleaning techniques. However, despite having these controls in place, there are fundamental limitations in terms of what can be achieved when considering algorithms as a socio-technical system (Barocas et al., 2017).

The next and final chapter concludes the study and provides recommendations and suggestions for future studies.

6. CONCLUSIONS & RECOMMENDATIONS

6.1 Introduction

With the rapid advancement in technological capabilities, and local and global competition within and outside of the banking sector, customer expectations are continuously rising (MoneyMarketing, 2019). Customers are now expecting personalised advice from their banks based on their needs (W.UP, 2018). Banks can achieve this by applying algorithmic technologies (Merlicek, 2018) to extract value from data (Martin, 2019). However, the use of algorithms does not come without its limitations.

According to Barocas et al. (2017), these limitations include the following:

- Unbiased measurement being infeasible.
- Decision-makers not generally being involved in the measurement phase.
- Observational data being insufficient in identifying the disparity causes which would be required to design impactful interventions.
- Current literature assuming simplistic mathematical systems yet ignoring the effect of these interventions on individuals as well as the long-term effects on society.

Despite these limitations, Barocas et al. (2017) argue that there are two reasons to be optimistic:

- Data-driven decision-making allows for greater transparency when compared to human-based decision-making. It requires algorithm creators to articulate the objectives and allows the creators to understand the trade-offs between what is wanted or needed.
- The shift towards automated decision-making forces creators to be explicit about what needs to be achieved with decision-making. Creators need to formally state the objectives and intentions behind algorithms, enabling organisations to have formal discussions around fairness. These discussions could contribute towards the creation or updating of

legislation, policies, procedures and guidelines towards fair usage of algorithms.

This chapter integrates the findings of the three propositions that were made in the original research question and objectives from Chapter 1. The chapter concludes with recommendations and suggestions for future research.

6.2 Conclusion

This research, as discussed in Chapter 1, aimed to explore how underlying bias in training datasets used to train sales-based decision-making algorithms used by South African retail banks influence customer product offerings.

This explorative and qualitative study achieved the research objective using three sub-research questions, as discussed in Chapter 3:

Main Research Question: How do underlying bias in training datasets used to train sales-based decision-making algorithms used by South African retail banks influence customer product offerings?

Sub-Research Question 1: How does gender influence product offerings' algorithmic decisions in retail banking?

Sub-Research Question 2: How does disability influence product offerings' algorithmic decisions in retail banking?

Sub-Research Question 3: In what way does imbalanced data influence product offerings' algorithmic decisions in retail banking?

The results of the study were presented in Chapter 4 and discussed in Chapter 5. Key themes relating to the impact of underlying bias in training datasets used to train sales-based decision-making algorithms used by South African retail banks were:

- Data-driven approaches to creating (personalised) customer products
- Causes of biased algorithmic outputs
- Consequences of biased algorithmic outputs

- Controls that mitigate biased algorithmic outputs

6.2.1 *Gender factors' influence on product offerings' algorithmic decisions in retail banking*

The gender demographic attribute was identified as a known biased variable that the bank had prohibited data science teams from using to inform algorithmic decision-making. However, it was unclear if the data science teams fully understood the reasons behind the variable's exclusion. This could lead to certain developers using proxy variables that could lean towards a customer's gender, introducing gender-based bias into some of the bank's algorithmic decision-making, impacting or influencing certain product offerings made to certain genders.

This highlights the need for the organisation to create awareness or education/knowledge programmes and to create and communicate policies, guidelines and frameworks around fair and ethical usage of algorithms and the consequences that could emanate as a result of underlying bias in training data used to train decision-making algorithms. This includes documenting institutional knowledge present in the bank. By implementing these controls, the organisation should see more consistency in the application by developers, and customers would experience more consistent and fair outcomes from the organisation.

Another issue that could result in incorrect offers being made for females is dependent on the income disparities between males and females. Due to algorithms generalising based on the majority's requirements (Barocas et al., 2017), if females who earn higher incomes are not adequately represented in the training data and are a minority group, this could result in algorithms providing results that have a higher error rate for females.

This highlights the need for the organisation to ensure that there are adequate data points in the training data so that the algorithm can predict so that the customers' needs are met. This may require the organisation to review outcomes for different groups of individuals (i.e., based on different income segmentations) separately.

6.2.2 *Disability factors' influence on product offerings' algorithmic decisions in retail banking*

The disability demographic attribute was not identified as a known biased variable and was not a variable the bank was collecting from the customers. Thus, the bank can also not test other variables for correlation to a disability, which open the possibility of unintentional bias against disabled customers.

There is a perception that disabled individuals are not as economically productive (Anastasiou & Kauffman, 2013); however, Barocas et al. (2017) argue that the challenge for disabled individuals to contribute as effectively as abled individuals are due to appropriate physical environment accommodations not being in place rather than their inherent capabilities.

This is important for banks as they are viewed as trusted partners that enable customers financially. Therefore, it raises the question of whether banks should collect customer disability data and do more in terms of social responsibility to address these social inequalities and enable disabled individuals to contribute as effectively as abled individuals (i.e., creating equal participation in society).

6.2.3 *Imbalanced data factors' influence on product offerings' algorithmic decisions in retail banking*

Imbalanced data was a known issue in the bank due to algorithms' tendency to generalise based on the majority's requirements resulting in a higher error rate for under-represented groups of individuals or minority groups. This could result in financial exclusion or incorrect products being created for certain groups of customers which, if not corrected, would lead to continued subordination of certain groups of customers based on demographic attributes.

This highlights the need for the organisation to ensure that there are adequate data points in the training data so that the algorithm can predict so that the customers' needs are met. This may require the organisation to review outcomes for different groups of individuals separately.

6.3 Recommendations

As South African retail banks continue to increase their use of algorithms for decision-making, the banks must remain cognisant of the potential to perpetuate societal biases through the algorithms.

The banking industry in South Africa should collaborate efforts to create best practices for fair and ethical algorithm use. The banks should feel incentivised to address bias in algorithmic decision-making proactively. The banks which contribute towards the fair and ethical use of algorithms should be recognised and acknowledged by South African legislators and the customers for executing on or contributing to creating best practices.

There should also be greater awareness or communication around fair and ethical algorithm use and the consequences that could emanate as a result of underlying bias in training data used to train decision-making algorithms throughout the organisation. This awareness or knowledge should also be shared with customers, as customers play an important role in providing feedback to the banks. This feedback should be used to iteratively make algorithms' decision-making more accurate, fair and ethical.

The bank should also document the institutional knowledge present in the bank. This will enable greater and more effective communication to all impacted bank staff and ensure continuity should certain individuals with institutional knowledge leave the bank. This would also result in more consistent processes being followed by developers in different business units.

The bank should also implement a process where the outcomes for different groups of individuals (i.e., based on different income segmentations) are reviewed separately in terms of error rates to ensure that the needs of customers are met.

The bank should also ensure that the impact of representational harms, where the algorithms reinforce the subordination of certain groups of individuals based on demographic attributes by perpetuating stereotypes, is understood, and controls are implemented to mitigate these harms. As per the social

constructivism theory, data can repeat further and transfer these social realities onto other individuals and generations (Scholz, 2017).

6.4 Suggestions for further research

This research report focused on bias in the training data used to train algorithms. Future studies could be expanded to include other types of bias within an algorithmic value chain, as documented in section 2.4.2. This research report also only focused on gender, disability and imbalanced data. Future studies could be expanded to other training data biases such as age, geographical location, and data drift in models deployed into production.

- In terms of age, it would be important to see whether algorithms are discriminating against customers for being too old to take up credit-based products or whether there is discrimination against younger customers who are potentially newly employed. This study would be necessary as it could contribute to a best practice guideline in using age to inform algorithmic decision-making.
- In terms of geographical location, especially in South Africa, it would be important to see whether algorithms discriminate against individuals living in certain suburbs, as this variable could be used as a proxy for race.
- In terms of data drift in models deployed into production, these natural occurrences cause populations to shift (especially in light of COVID-19). This study would be important as it could contribute to a best practice guideline in terms of identifying and mitigating data drift.

This research focused on data science teams in a South African retail bank. Future studies could be expanded to include the three lines of defence (business, risk and audit) and customers for greater end-to-end insights on bias. Future studies could also be expanded to the other banking institutions in South Africa, such as commercial banks and investment banks. Future studies could incorporate these additional services with South African banks diversifying their services to include telecommunications and insurance. Future studies could also

expand on the customer products offered as this study focused on credit specific banking product offerings.

An interesting case study for future studies could also include the demographic breakdown of complaints and product take-ups to identify potential underlying biases. This could also be expanded to social media data.

Another interesting case study could include results from robo-advisor tools that have been deployed by banks using the same income and expense type information but changing demographic attributes to identify if there is any bias in algorithmic decision-making in these tools.

Furthermore, the impact of banks' risk appetite for financial losses due to incorrect customer predictions can be investigated. This could be assessed in terms of types of customers impacted (i.e., across different demographic attributes) and its short-term and long-term effects in terms of representational harms.

Analysis using retail data (e.g., "lay-by" or loyal programme data) to introduce minority group customer transactional data can be performed. Such a study could include a comparison between banks that leverage this data against those that don't and explore the influence in terms of accuracy of decisions made or impact on inclusivity and customer social sentiment.

An unanswered question is how are organisations (or even countries) being incentivised for being transparent about disclosing work towards, for example, ethics guidelines, or identifying issues in their organisations and sharing solutions/best practices implemented to mitigate the issues (or by sharing the root cause of the issues and sharing solutions)? Research in this regard can extend to multiple industries, e.g., the impact of South Africa disclosing research on the Omnicron COVID-19 variant and the consequences of this disclosure or knowledge sharing. The same could be done in the banking industry regarding customer perception and legislative response (e.g., regulatory penalties issued) due to disclosures around the identification of customer-specific biases.

Finally, it is recommended that the influence of social media on customer's perception of bias and the outcomes of their product offerings from their banks be investigated. This could be comparative across multiple banks as banks would have different risk appetites in terms of lending criteria and usage of credit bureau reports. Additionally, such a study could explore consumer knowledge of these different components.

REFERENCES

Adams, R., Fourie, W., Marivate, V., & Platinga, P. (2020). *Introducing The Series: Can AI and Data Support a More Inclusive and Equitable South Africa*. Policy Action Network. https://policyaction.org.za/sites/default/files/PAN_TopicalGuide_AIData1_IntroSeries_Elec.pdf

Anastasiou, D., & Kauffman, J. M. (2013). The social model of disability: Dichotomy between impairment and disability. *The Journal of Medicine and Philosophy: A forum for bioethics and philosophy of medicine*, *38*(4), 441-459. https://doi.org/10.1093/jmp/jht026

Anney, V. N. (2014). Ensuring the quality of the findings of qualitative research: Looking at trustworthiness criteria. *Journal of Emerging Trends in Educational Research and Policy Studies*, *5*(2), 272-281.

Barocas, S., Hardt, M., & Narayanan, A. (2017). *Fairness in machine learning*. https://fairmlbook.org/pdf/fairmlbook.pdf

Bellens, J., & Meekings, K. (2020). *How banks can stay relevant as customer preferences change*. EY. https://www.ey.com/en_gl/banking-new-decade/how-banks-can-stay-relevant-as-customer-preferences-change

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In S. A., Friedler & C. Wilson (Eds.) *Proceedings of Machine Learning Research: Vol. 81 Conference on Firness, Accountability and Transparency* (pp. 149-159). http://proceedings.mlr.press/v81/binns18a/binns18a.pdf

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101.

Bryant, R., Cintas, C., Wambugu, I., Kinai, A., & Weldemariam, K. (2019). Analysis bias in sensitive personal information used to train financial models. *IBM Research*. https://arxiv.org/pdf/1911.03623.pdf

Bunbury, S. (2019). Unconscious bias and the medical model: How the social model may hold the key to transformative thinking about disability discrimination. *International Journal of Discrimination and the Law, 19*(1), 26-47. https://doi.org/10.1177%2F1358229118820742

Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., & Gonzalez, B. S. M. (2020, Dec 10-13). *BeFair: Addressing Fairness in the Banking Sector* [Conference Session]. 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA. https://doi.org/10.1109/BigData50022.2020.9377894

Chuprina, R., & Kovalenko, O. (2021). *Machine Learning in Banking – Opportunities, Risks, Use Cases*. SPD Group. https://spd.group/machine-learning/machine-learning-in-banking/

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2022). *Introduction to algorithms*. MIT press.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage Publications.

Daly, R. (2020). *The growth of data in banking is unstoppable*. Accenture. https://bankingblog.accenture.com/growth-of-data-in-banking-is-unstoppable

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In C. Sierra (Ed.) *Proceedings of the Twenty-Sixth International Joint Conference of Artificial Intelligence* (pp. 4691-4697). https://doi.org/10.24963/ijcai.2017/654

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review, 65*(3), 122-135. https://doi.org/http://dx.doi.org/10.1108/LR-06-2015-0061

Delen, D. (2014). *Real-world data mining: Applied business analytics and decision making*. FT Press.

Delen, D., & Ram, S. (2018). Research challenges and opportunities in business analytics. *Journal of Business Analytics, 1*(1), 2-12. https://doi.org/10.1080/2573234X.2018.1507324

Dey, M., & Rautaray, S. S. (2014). Study and analysis of data mining algorithms for healthcare decision support system. *International Journal of Computer Science and Information Technologies, 5*(1), 470-477.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism, 5*(7), 809-828.

Dickson, B. (2020). *How banks use AI to catch criminals and detect bias*. The Next Web. https://thenextweb.com/news/how-banks-use-ai-to-catch-criminals-and-detect-bias

Gelb, S. (2003). *Inequality in South Africa: Nature, causes and responses*. The EDGE Institute. http://wolpetrust.org.za/dialogue2004/CT052004gelb_report.pdf

Green, J., & Thorogood, N. (2018). *Qualitative methods for health research*. SAGE Publications Ltd.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research, 27*(4), 591-608. https://doi.org/10.1177%2F1049732316665344

Joseph, R. (2020). *Fact sheet: The role of algorithms, automation and artificial intelligence in human resources management*. S. B. f. P. Practices. https://cdn.ymaws.com/www.sabpp.co.za/resource/resmgr/siphiwe_2020/fact_sheet_july_2020.pdf

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., & Gani, A. (2014). Big data: Survey, technologies, opportunities, and challenges. The *Scientific World Journal, 2014*, 712826. https://doi.org/10.1155/2014/712826

Malinga, S. (2019). *How AI helped TymeBank gain 670K customers*. ITWeb. https://www.itweb.co.za/content/WnxpE74DX9K7V8XL/j5alrvQgkY1MpYQk

Marous, J. (2021). Lack of personalization puts banks at odds with consumer expectations. The Financial Brand. https://thefinancialbrand.com/106537/banking-customer-personalization-experience-trends/

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835-850.

Merlicek, R. (2018). *Algorithms are the future of retail banking*. FinTech Business. https://www.fintechbusiness.com/blogs/1074-algorithms-are-the-future-of-retail-banking

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation* (4th ed.). John Wiley & Sons.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. https://doi.org/10.1177%2F2053951716679679

MoneyMarketing (2019). *2019:The year of rapidly rising customer expectations in the banking sector*. https://moneymarketing.co.za/2019-the-year-of-rapidly-rising-customer-expectations-in-the-banking-sector/

Moosajee, N. (2019). *Fix AI's racist, sexist bias*. Mail & Guardian. https://mg.co.za/article/2019-03-14-fix-ais-racist-sexist-bias/

Nienaber, M. (2020). *Leveraging technology to enhance the client experience*. The Media Online. https://themediaonline.co.za/2020/06/leveraging-technology-to-enhance-the-client-experience/

Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170364. https://doi.org/10.1098/rsta.2017.0364

Queirós, A., Faria, D., & Almeida, F. (2017). Strengths and limitations of qualitative and quantitative research methods. *European Journal of*

Education Studies 3(9), 369-387. http://dx.doi.org/10.46827/ejes.v0i0.1017

Rajaei, A. (2017). Identification of fraud in banking data and financial institutions using classification algorithms. *International Journal of Information, Security and Systems Management, 6*(2), 663-667.

Republic of South Africa (2000) Promotion of Equality and Prevention of Unfair Discrimination Act, 2000. https://www.gov.za/sites/default/files/gcis_document/201409/a4-001.pdf

Republic of South Africa, (2006). National Credit Act, 2005 https://www.justice.gov.za/mc/vnbp/act2005-034.pdf

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review. Online, 192*. https://www.nyulawreview.org/wp-content/uploads/2019/04/NYULawReview-94-Richardson-Schultz-Crawford.pdf

Sayed, H., Abdel-Fattah, M. A., & Kholief, S. (2018). Predicting potential banking customer churn using apache spark ML and MLlib packages: a comparative study. *International Journal of Advanced Computer Science and Applications, 9*(11), 674-677.

Scanlon, L. (2020). *AI in financial services: addressing the risk of bias*. Pinsent Masons. https://www.pinsentmasons.com/out-law/analysis/ai-financial-services-risk-of-bias

Scholz, T. M. (2017). *Big data in organizations and the role of human resource management: A complex systems theory-based conceptualization*. Peter Lang Academic Research.

Shah, H. (2018). Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2128), 20170362. https://doi.org/10.1098/rsta.2017.0362

Sharma, P. (n.d.) *Decision-making: Definition, importance and principles | Management*. YourArticelLibrary. https://www.yourarticlelibrary.com/management/decision-making-management/decision-making-definition-importance-and-principles-management/70038

Silva, S., & Kenney, M. (2019). Algorithms, platforms, and ethnic bias. *Communications of the ACM, 62*(11), 37-39. https://doi.org/10.1145/3318157

Smalec, A. (2021). Big Data as a tool helpful in communication management. *Procedia Computer Science*, *192*(2021), 5156-5165. https://doi.org/10.1016/j.procs.2021.09.293

Sullivan, B., Garvey, J., Alcocer, J., & Eldridge, A. (2020). *Retail banking 2020 Evolution or revolution?* PwC. Retail Banking. https://www.pwc.com/gx/en/banking-capital-markets/banking-2020/assets/pwc-retail-banking-2020-evolution-or-revolution.pdf

Tobor, N. (2017). *South Africa's Absa has launched an investment robo-advisor*. iAfrikan. https://www.iafrikan.com/2017/08/03/south-africas-absa-launches-new-online-self-service-investment-platform/

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, 37, 215-230. https://doi.org/10.1007/s00146-021-01154-8

Turner Lee, N., & Resnick, P. Barton, G. (2019). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms*. Brookings. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

van Staveren, I. (2001). Gender biases in finance. *Gender & Development, 9*(1), 9-17. https://doi.org/10.1080/13552070127734

Wentzel, J. P., Diatha, K. S., & Yadavalli, V. S. S. (2016). An investigation into factors impacting financial exclusion at the bottom of the pyramid in South Africa. *Development Southern Africa, 33*(2), 203-214. https://doi.org/10.1080/0376835X.2015.1120648

World Health Organisation (2021). *Disability and health*. https://www.who.int/news-room/fact-sheets/detail/disability-and-health

W.UP (2018). Personalisation is the new segmentation in banking. https://wup.digital/blog/personalisation-vs-segmentation/

Zhang, Y., & Zhou, L. (2019, Dec 8-14). Fairness assessment for artificial intelligence in financial industry. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. https://arxiv.org/pdf/1912.07211.pdf

APPENDIX (A) Instrument

Table 2: Interview Questions

Question #	Research Question(s)
A. General: Algorithms, Bias, Data Bias	
A1	What is your understanding of an algorithm?
A1a	a. When would you need an algorithm?
A2	In your opinion, is the usage of algorithms increasing or decreasing by the bank?
A2a	b. Why do you think that is?
A2b	c. In your opinion, what are the reasons behind the drivers for the increase?
A3	What is your involvement in the development of the sales specific algorithm that offers products to customers?

Question #	Research Question(s)
A3a	a. If answer is “the development of product rules”: Are you involved in testing the algorithm and validating its outputs? If not, why not? If so, are there any insights you would like to share from that process?
A3b	b. If the answer is “the development of the algorithm”: What is the process that gets followed to develop the algorithm?
A3c	c. Does the bank use an aggregation of existing customer data to create product offerings for customers?
A3d	d. Are there steps taken to cleanse the data? If yes, please elaborate on the steps and rationales behind the steps.
A3e	e. Where is your data stored?
A3f	f. How real-time is the data?
A4	Are the algorithmic outputs that get developed decisions that a human must action, or do they result in automated decision-making?

Question #	Research Question(s)
A4a	a. If answer is “a bit of both” or “where a human must action”: is there the intention to move towards fully automated decision-making?
A4b	b. If answer is “automated”: what is the impact of POPIA on this development in terms of its requirement around automated decision-making?
A4c	c. Do we use machine learning or artificial intelligence on the algorithms?
A5	How frequently does your team re-test the algorithm to ensure that it is still relevant or correct?
A5a	a. What monitoring or testing does the team perform?
A5b	b. Do you think any enhanced monitoring or testing is required?
A6	What software tool(s) does the team use to build the models?

Question #	Research Question(s)
A7	What is your understanding of bias?
A7a	a. In your opinion, where do you think bias can be introduced during the process (ie where the algorithm offers products or doesn't offer products to customers)?
A7b	b. If more than one type of bias was mentioned, which bias do you believe is the biggest contributor out of the ones you've listed?
A7c	c. In your opinion, what do you think are some of the consequences of bias previously mentioned?
A7d	d. What are some of the controls in place to mitigate these types of bias?
A8	How has POPIA's implementation on 1 July 2021 changed any of the processes that were previously followed?
A9	Where are the algorithm rules developed and presented?
B. Proposition 1: The inclusion of the gender variable as an input to South African retail banks' sales-based decision-making algorithms creates gender biased product offerings that will not meet the customers' needs	

Question #	Research Question(s)
B1	Is gender one of the characteristics used to create product offers to customers?
B1a	a. If the answer is “yes”: In your opinion, do you believe that it is necessary to use gender to offer products to customers?
B1b	b. Have there been any instances where females or males have complained about an unfair outcome that they believe to be unfair based on their gender? If so, please elaborate.
B1c	c. What processes are in place to ensure that the algorithm’s prediction is not biased against females or males?
B1d	d. What controls are in place to prevent gender-biased outcomes from the algorithm?
B1e	e. Are there data cleansing processes specifically relating to gender characteristic? In your opinion, do you believe that the data cleansing process introduces further bias?
B2	Is gender one of the characteristics used to score or price customers?
B2a	a. If the answer is “yes”: in your opinion, do you believe that it is necessary to use gender to score or price customers?

Question #	Research Question(s)
B3	b. To what extent is only the individual's data used to determine the outcome of the product being offered to the customer?
C. Proposition 2: The inclusion of the disability variable as an input to South African retail banks' sales-based decision-making algorithms creates disability biased product offerings that will not meet the customers' needs	
C1	Is disability one of the characteristics used to create offers for customers?
C1a	a. If the answer is "yes": in your opinion, do you believe that it is necessary to use disability to offer products to customers?
C1b	b. Have there been any instances where disabled or abled individuals have complained about an unfair outcome that they believe to be unfair based on their disability or lack thereof? If so, please elaborate.
C1c	c. What processes are in place to ensure that the algorithm's prediction is not biased against disabled people or abled people?
C1d	d. What controls are in place to prevent disability-biased outcomes from the algorithm?
C1e	e. Are there data cleansing processes specifically relating to the disability characteristic? In your opinion, do you believe that the data cleansing process introduces further bias?

Question #	Research Question(s)
C2	Is disability one of the characteristics used to score or price customers?
C2a	a. If the answer is “yes”: in your opinion, do you believe that it is necessary to use disability to score or price customers?
C3	To what extent is only the individual’s data used to determine the outcome of the product being offered to the customer?
D. Proposition 3: The inclusion of imbalanced data as an input to South African retail banks’ sales-based decision-making algorithms creates irrelevant product offerings that will not meet the customers’ needs or will lead to financial exclusion of minority group customers	
D1	In your opinion, what is an adequate representation of a population required in a dataset?
D1a	a. What are checks performed to ensure that there is an adequate representation of populations? If no checks are performed, why and elaborate?
D1b	b. Have there been any instances where individuals have complained about an unfair outcome that they believe to be unfair based on the under-representation of the population? If so, please elaborate.

Question #	Research Question(s)
D1c	c. What processes are in place to ensure that the algorithm’s prediction is not biased against individuals due to the under-representation of the population?
D1d	d. What controls are in place to prevent under-representation-based outcomes from the algorithm?
D2	In your opinion, what characteristics would be necessary to ensure fair product offerings, scoring, or pricing?
D3	Do you believe that the metrics used for scoring advantage or disadvantage certain groups of individuals based on any of the key inputs used ie whether it's scoring, pricing or product offers?
D3a	a. Do you think it's necessary to include these characteristics?
D4	Is there a higher weighting on certain characteristics?
D5	To what extent is only the individual’s data used to determine the outcome of the product being offered to the customer?

APPENDIX (B) Consistency Matrix

Table 3: Consistency Matrix

RQ #	Research question	Prop #	Proposition	Data collection detail	Data analysis method
1	How do underlying bias in training datasets used to train sales-based decision-making algorithms used by South African Retail banks influence customer product offerings?				
1.1	How does gender influence product offerings' algorithmic decisions in Retail banking?	1.1	The inclusion of the gender variable as an input to South African retail banks' sales-based decision-making algorithms creates gender biased product offerings that will not meet the customers' needs.	Interview guide questions B1, B1a, B1b, B1c, B1d, B1e, B2, B2a, B3	Thematic analysis

RQ #	Research question	Prop #	Proposition	Data collection detail	Data analysis method
1.2	How does disability influence product offerings' algorithmic decisions in Retail banking?	1.2	The inclusion of the disability variable as an input to South African retail banks' sales-based decision-making algorithms creates disability biased product offerings that will not meet the customers' needs.	Interview guide questions C1, C1a, C1b, C1c, C1d, C1e, C2, C2a, C3	Thematic analysis
1.3	In what way does imbalanced data influence product offerings' algorithmic decisions in Retail banking?	1.3	The inclusion of imbalanced data as an input to South African retail banks' sales-based decision-making algorithms creates irrelevant product offerings that will not meet the customers' needs or will lead to financial exclusion of minority group customers.	Interview guide questions D1, D1a, D1b, D1c, D1d, D2, D3, D3a, D4, D5	Thematic analysis

APPENDIX (C) Consent Form

Title of project: Bias in data used to train sales-based decision-making algorithms in a South African Retail bank

Name of researcher: Alice Wong

I, …….. (name of participant), agree to participate in this research project. The research has been explained to me and I understand what my participation will involve. I agree to the following:

(Please circle the relevant options below).

I agree that my participation will remain anonymous	YES	NO
I agree that the researcher may use anonymous quotes in her research report	YES	NO
I agree that the interview may be audio recorded	YES	NO

_____ (signature)

_____ (name of participant)

_____ (date)

_____ (signature)

_____ (name of person seeking consent)

_____ (date)

APPENDIX (D) Participation Information Sheet

Good day,

My name is Alice Wong, and I am a Masters student at the Wits Business School, University of the Witwatersrand, Johannesburg. As part of my studies, I am investigating the influence of data bias on algorithmic outputs under the supervision of Miss Ayanda Magida. This research project aims to gain a deeper understanding of the potential bias that can be found in the training data used to train sales-based decision-making algorithms used by a South African Retail bank to create customer product offerings.

As part of this project, I would like to invite you to take part in an interview. This activity will involve approximately 50 questions and will take around 60 minutes. With your permission, I would also like to audio record the interview using MS Teams. This recording will be stored on my personal laptop and only the researcher will have access to this recording. It will be deleted after 5 years.

There will be no personal costs to you if you participate in this project. You will not receive any direct benefits from participation but there are no disadvantages or penalties if you do not choose to participate or if you withdraw from the study. You may withdraw at any time or not answer any question if you do not want to. The interview will be completely confidential and anonymous as I will not be asking for your name or any identifying information, and the information you give to me will be held securely and not disclosed to anyone else. If you experience any distress or discomfort at any point in this process, we will stop the interview or resume another time.

If you have any questions during or afterwards about this research, feel free to contact me on the details listed below. This study will be written up as a research report which will be available online through the university library website. If you wish to receive a summary of this report, I will be happy to send it to you. If you have any concerns or complaints regarding the ethical procedures of this study, you are welcome to contact the University Human Research Ethics Committee (Non-Medical), telephone +27(0) 11 717 1408, email hrecnon-medical@wits.ac.za

Thank you in advance, your participation is immensely appreciated.

Yours sincerely,
Alice Wong

Researcher: Alice Wong,
Email: 2303668@students.wits.ac.za | phone number: +27 72 370 9279

Supervisor: Ayanda Magida
Email: Ayanda.Magida@wits.ac.za | phone number: +27 11 717 3953

APPENDIX (E) Ethics Approval Letter

Graduate School of Business Administration
University of the Witwatersrand, Johannesburg

Wits Business School Ethics Committee
Constituted under the University Human Research Ethics Committee (Non-Medical)

Ethics Clearance Certificate

Ethics protocol number: WBS/DB2303668/246
This certificate is only valid with a legitimate ethics protocol number and signed by the Researcher (below,

This certificate is only valid if accompanied by formal permission from the relevant stakeholder(s).

Project title	Bias in data used to train sales-based decision-making algorithms in a South African retail bank
Investigator / Researcher	Miss Alice Wong
Nature of Project	MM (Digital Business)
Decision of the Committee	Approved, provided stakeholders and participants are guaranteed confidentialty.
Issue Date of Certificate	2021-09-26
Expiry date	Date of submission of the project report
Chairperson	Prof Anthony Stacey ☎ +27 11 717 3587 +27 82 880 4531 anthony.stacey@wits.ac.za

---

Declaration by Researcher

One copy must be signed by the Researcher and returned to the Chairperson of the Wits Business School Ethics Committee.

I fully understand the conditions under which I am authorized to carry out the abovementioned research and I guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I undertake to resubmit the protocol to the Committee.

Signature

27 September 2021
Date:

APPENDIX (F) Organisation: request for approval

5 July 2021

Att: *<removed for confidentiality>*

Re: Permission to conduct research at *<removed for confidentiality>*

The purpose of this letter is to request permission to conduct a research study within *<removed for confidentiality>* as part of my academic qualification for a Masters of Management in the field of Digital Business at Wits Business School.

My research paper's title is: Bias in data used to train sales-based decision-making algorithms in South African Retail banks.

I am conducting research on the potential bias that can be found in the training data used to train sales-based decision-making algorithms, used by South African Retail banks, to create customer product offerings. The specific biases that will be investigated are: gender bias, disability bias and imbalanced data. The purpose of the study is to determine the influence of these biases within *<removed for confidentiality>* sales-based algorithmic outputs, as it is one of the largest banks in the country as well as one of the leaders in terms of digitalisation.

Please note, as follows:

- I will invite individuals from *<removed for confidentiality>* Product House Business Units' Data Science teams (across analyst, manager and head roles) to participate in this study. If they agree, they will be asked to be interviewed. These interviews will be audio recorded and transcribed, however, these transcripts will not be published as part of the study.
- Willing participants will be asked to give their written or verbal consent before the research begins. Their responses will be treated confidentially, and identities (their names and the name of the organisation) will be anonymous. Individual privacy will be maintained in all published and written data resulting from the study.
- The research participants will not be advantaged or disadvantaged in any way. They will be reassured that they can withdraw their permission at any time during this project without any penalty. There are no foreseeable risks in participating in this study. The participants will not be paid for this study.

- Any sensitive information provided by the participants regarding the Bank's specific processes or strategies will be treated as confidential and will not be published in the final report.
- The name of the organisation will not be included in the final report. The final report will refer to a traditional bank.
- All the data collected will be anonymised and kept confidential.
- The results will be used for academic purposes only and may be published in an academic journal. The final report will be published in the Wits library and online repository. *<removed for confidentiality>* will be provided with a summary of the findings, on request.
- All research data will be destroyed after 5 years after completion of the final report.

I therefore request permission in writing to conduct my research at *<removed for confidentiality>*. The permission letter should be on your organisation's headed paper, signed and dated, and specifically referring to myself by name and the title of my study.

Please let me know if you require any further information. I look forward to your response as soon as is convenient.

Yours sincerely,

_____

Alice Wong
+27 72 370 9279
2303668@students.wits.ac.za

APPENDIX (G) Organisation Approval Letter

The organisation approval letter has been obtained and provided to the research supervisor and ethics committee. The organisational approval letter has been requested to remain anonymous.