Big data and return on marketing investment in a South African insurer

Darryl Grater

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 | Strategic Marketing

Johannesburg 2021

ABSTRACT

The purpose of the study centres around the enablement of the information and knowledge harvested from big data into increasing the return on marketing investment (ROMI). Many financial services providers offer a wide variety of financial products which include life insurance, health insurance, investment products, wellness and short-term insurance – most of these markets are saturated with multiple product providers all competing for share of potential customers' attention and ultimately share of wallet. Although there are various studies involving the use and interpretation of big data in terms of propensity models and product recommender systems, cross-sell propensity modelling using personalisation based on existing financial products enjoyed by a customer has not been sufficiently or adequately researched - particularly for a short-term insurer within a Southern African context

Considering this was a big data study, the research was quantitative whereby 566,758 customer profiles from a South African financial services provider (FSP) were extracted from the FSP's databases, and thereafter scrutinised. The data was analysed using statistical methods which included *Extreme Gradient Boosting* to investigate and assess the impact of various existing customer characteristics and the acceptance of a personal lines short-term insurance quote when offered. The study embarked to understand if certain characteristics of existing customers could be identified which indicate materiality to short-term insurance product acceptance, and how one by using this personalisation information can create ring-fenced segments of existing customers for focused insurance product cross-sell campaigns, with better sales results versus standard busines development.

Results identified in this research conclude that the theoretical framework supports the results exposed in this study. Certain customer characteristics from their product utilisation from life insurance, health insurance, banking, wellness and investment products indicate traits which are highly material to

their conversion rate and short-term insurance product acceptance. By using this personalisation information and creating segments which portray the best conversion rates, focused sales campaigns result in higher sales ratios and therefore marketers and business development executives see better sales conversion rates versus mass-market, broad-based initiatives (for example random blind, cold calling). Sales resources can therefore be deployed to quote short-term insurance products to these customer segments and will render higher sales and thus a higher return on marketing investment.

In conclusion, the contribution of this research study enables the financial services industry to focus on cross-sell and upsell campaigns with higher sales performances by using the personalisation insights from their existing customer base. Furthermore, marketers and business development executives can use similar frameworks to develop similar propensity models for other financial products. The research also opens a path for future similar research to be conducted across various other financial products (not only on short-term insurance impact) and across geographies extending further than a South African customer base.

DECLARATION

I, Darryl Andrew Grater, 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 | Strategic Marketing at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other university.

Darryl Andrew Grater

Signed at Cape Town, on the 30th of April 2021

DEDICATION

I dedicate this dissertation to my late father, Kennard Leon Grater, who was so very proud to hear I had been accepted into the Masters Programme at Wits Business School – I will forever treasure his memory.

ACKNOWLEDGEMENTS

First and foremost, I must commend my supervisor, Laurence Beder, who has been a never-ceasing source of direction and mentorship throughout this dissertation's journey – I remain grateful for the opportunity he provided to me and the constant encouragement in my search for broader knowledge through this research cycle.

A special thank you to Wits Business School and all their staff; their selfless devotion to academia, student development and growing the African knowledge base was a privilege to experience.

I was most fortunate to work with Brett Rowland, a data scientist and professional statistician whom I consulted with during this research process – his insights and suggestions were invaluable.

To my wife, Tumisho Grater, who supported me in my academic journey from the very beginning, and continued to provide warm meals and the necessary space to complete the various academic objectives over the years – thank you!

TABLE OF CONTENTS

ABSTRACT **2**

DECLARATION **4**

DEDICATION **5**

ACKNOWLEDGEMENTS **6**

LIST OF FIGURES **11**

LIST OF TABLES **13**

LIST OF ABBREVIATIONS **14**

1 INTRODUCTION **16**

1.1 PURPOSE OF THE STUDY 16

1.2 CONTEXT OF THE STUDY 18

1.3 PROBLEM STATEMENT 19

1.3.1 MAIN PROBLEM 19

1.3.2 SUB-PROBLEMS 19

1.4 PURPOSE OF THE STUDY 19

1.5 RESEARCH OBJECTIVES 20

1.5.1 THEORETICAL OBJECTIVES 20

1.5.2 EMPIRICAL OBJECTIVES 21

1.6 RESEARCH QUESTIONS 21

SUB-PROBLEM 1 - 21

RESEARCH QUESTION 1: 21
SUB-PROBLEM 2 - 21
RESEARCH QUESTION 2: 21
1.7 RESEARCH GAP AND JUSTIFICATION OF THE STUDY 21
1.8. SIGNIFICANCE OF THE STUDY 21
1.9. DELIMITATIONS OF THE STUDY 22
1.10 DEFINITION OF TERMS 23
1.11 ASSUMPTIONS 24

2 LITERATURE REVIEW 26

2.1 INTRODUCTION 26
2.2 THEORETICAL GROUNDING 27
2.3 EMPIRICAL REVIEW 27
2.3.1 DATA, AND ITS LINK TO INFORMATION AND KNOWLEDGE 27
2.3.2 DATA AND PERSONALISATION 30
2.3.3 BIG DATA AND PREDICTIVE ANALYTICS 33
2.3.3.1 RESEARCH QUESTION 1: 36
2.3.4 TIGHT MARGINS AND A SOFT MARKET 36
2.3.5 RETURN ON MARKETING INVESTMENT (ROMI) 38
2.3.6 ROMI CONSTRAINTS 40
2.3.5.2 RESEARCH QUESTION 2: 42
2.3.7 HYPOTHESIS 42
2.4 CONCLUSION OF LITERATURE REVIEW 43

3. RESEARCH METHODOLOGY 45

3.1 RESEARCH PHILOSOPHY 45
3.2 RESEARCH STRATEGY 45
3.3 RESEARCH DESIGN 47

3.4 RESEARCH PROCEDURE AND METHODS 48
3.5 TARGET POPULATION AND SAMPLING 50
3.6 ETHICAL CONSIDERATIONS WHEN COLLECTING DATA 53
3.7 DATA COLLECTION RIGOUR AND STORAGE 53
3.8 DATA PROCESSING AND ANALYSIS 55
3.9 DESCRIPTION OF THE RESPONDENTS 59
3.10 VALIDITY AND RELIABILITY 60

4 PRESENTATION OF RESULTS 68

4.1 INTRODUCTION 68
4.2 DESCRIPTIVE STATISTICS AND CORRELATIONS 69
4.3 MODEL ARCHITECTURE 76
4.4 MODEL EVALUATION 77

5 DISCUSSION OF RESULTS 80

5.1 INTRODUCTION 80
5.2 DESCRIPTION OF RESULTS 80
5.3 CONCLUSION 88

6. CONCLUSION AND RECOMMENDATIONS 89

6.1 INTRODUCTION 89
6.2 CONCLUSION OF THE STUDY 89
6.3 RECOMMENDATIONS AND MANAGERIAL IMPLICATIONS 90
6.4 RECOMMENDATIONS FOR FURTHER RESEARCH 93
6.5 LIMITATIONS OF THE STUDY 96
6.6 CONCLUSION 97

7 BIBLIOGRAPHY 99

8 APPENDIX 112

8.1 APPENDIX A 112

8.2 APPENDIX B 112

8.3 APPENDIX C 113

8.4 APPENDIX D 113

8.5 APPENDIX E 114

8.6 APPENDIX F 115

8.7 APPENDIX G 116

8.8 APPENDIX H 117

8.9 APPENDIX I 117

8.10 APPENDIX J 118

8.11 APPENDIX K 119

8.12 APPENDIX L 120

8.13 APPENDIX M 120

8.14 APPENDIX N 121

8.15 APPENDIX O 121

8.16 APPENDIX P 121

LIST OF FIGURES

Figure 1: Hypothesis 42
Figure 2: Model performance on test and train datasets vs. training rounds 64
Figure 3: Probability calibration plot 65
Figure 4: Discrimination threshold sensitivity plot 67
Figure 5: Business channel from where quotes are received 69
Figure 6: Health insurance product indicator 70
Figure 7: Health plan type indicator 70
Figure 8: Health plan type indicator 70
Figure 9: Hospital gap cover indicator 71
Figure 10: Wellness product indicator 71
Figure 11: Wellness product engagement indicator 71
Figure 12: Wellness product duration 72
Figure 13: Life product indicator 72
Figure 14: Credit card product indicator 72
Figure 15: Credit card product type indicator 73
Figure 16: Credit card product health indicator 73
Figure 17: Credit card duration indicator 73
Figure 18: Investment product indicator 74
Figure 19: Investment product indicator 74
Figure 20: SES Indicator 74
Figure 21: Family size indicator 75
Figure 22: Number of products added in last 6 months indicator 75
Figure 23: Conversion model performance 78
Figure 24: Conversion model confusion matrix 79
Figure 25: SHAP Analysis 84
Figure 26: SHAP partial-dependence plot on age of customer 85
Figure 27: Analysis of conversion rates 86
Figure 28: Hypothesis proven 87
Figure 29: Data to wisdom relationship (Bellinger et al., 2004) 112
Figure 30: Modelling and analytics (Evans & Lindner, 2012) 112
Figure 31: Data mining and predictive analytics process chain (Hair, 2007) 113
Figure 32: (Statistics-South-Africa, 2020) 113

Figure 33: Ethical Clearance (Discovery Ltd) 114
Figure 34: Research population split (Author) 115
Figure 35: Decision tree for choosing a correct statistical technique (Black et al., 2010) 116
Figure 36: Construct function to maximise validation performance 117
Figure 37: Evaluation on model on test, train and validation sets 117
Figure 38: Bayesian hyperparameter tuning 118
Figure 39: Probability calibration plot 120
Figure 40: Threshold plot for the XGBClassifier 120
Figure 41: SHAP global feature importance code 121
Figure 42: SHAP partial-dependence plot of "age" feature 121
Figure 43: Code of customised customer segment for 57.98% conversion propensity 121

LIST OF TABLES

Table 1: Dataset description 60
Table 2: Reliability and validity summary 61
Table 3: Iterations of the Bayesian optimisation algorithm 119

LIST OF ABBREVIATIONS

B2C: Business to consumer

CRM: Customer Relationship management

CSV: Comma-separated-values

DV: Dependent variable

FN: False negative

FNR: False negative Rate

FP: False positive

FPR: False positive rate

FSP: Financial services provider

IMC: Integrated marketing communication

IoT: Internet-of-things

IV1: Independent variable 1

IV2: Independent variable 2

MIS: Management information system

NPTB: Next-product-to-buy

POPI: Protection of Personal Information Act, No 4 of 2013 (RSA)

PA: Predictive analytics

PPV: Positive predictive value

PRC: Precision recall curve

ROC: Receiver operating characteristics

ROI: Return on investment

ROMI: Return on marketing investment

RSA: Republic of South Africa

SES: Social economic score

SHAP: **SH**apley **A**dditive ex**P**lanations

SQL: Structured Query Language

TN: True negative

TP: True positive

TPR: True Positive rate

UM: User Model

XGBoost: Extreme Gradient Boosting Package

1 INTRODUCTION

1.1 Purpose of the study

Financial services providers often spend substantial amounts of money on marketing and associated advertising campaigns, particularly in the short-term insurance sector. Very often this marketing communication is via mass media and a “spray and pray” type approach, where marketing initiatives primarily focus on television, radio and digital advertising. These platforms are expensive and not necessarily effective. Blind cold-calling is often even less effective. This study plans on researching how the use of big data can be used to identify critical variables linked to short-term insurance purchase propensity, enabling a marketer to micro-segment the target audience, thereafter allowing individualised and tailored busines development to render a higher return on marketing investment. Research shows that companies that use big data can see up to a 20% increase ROI (Perrey, Spillecke, & Umblijs, 2013).

The literature review interrogates the main themes, commencing with data, and its link to information, knowledge and resultant wisdom. Personalisation, and how its integral link and dependences on data is explored, after which business analytics (focusing on predictive analytics) is discussed - important considering this is a key theme of the intended research. Thereafter, the saturated, short-term insurance market is discussed (SAIA, 2016), along with the market’s resultant tight financial margins (KPMG, 2014). This then finally leads to budget impacts, where particularly a marketing budget is discussed, and how marketers are needing to identify initiatives which cost less while still delivering the required (if not increased) sales figures (Danaher & Rust, 1996). This is important considering big data has the capability of shifting competition by amending corporate eco-systems, enabling innovation and renovating processes (Brown, Chui, & Manyika, 2011). These themes are also linked to the relevant research questions for context in chapter 2.

The research methodology is quantitative, with the research design being longitudinal and cross-sectional. This is required as the data instrument (an existing insurer's database) was analysed during the various points in time (a longitudinal view) and across numerous cases linking multiple variables (a cross-sectional view). A stratified sampling technique (Bryman, 2012) was utilised to identify the target population. Thereafter, the actual analysis was conducted with *Python* programming language, where *Python* is very well suited to tackle the challenges in the financial services sector specifically regarding financial and data analytics (Hilpisch, 2014). Extreme Gradient Boosting (*XGBoost*) is a popular gradient boosting machine learning algorithm (T. Chen et al., 2015), and was selected with a binary logistic objective function to frame as a classification task.

To assess validity of the XGBoost model, an upfront test-train and validation split, Bayesian hyperparameter tuning, and a post-hoc probability calibration check was conducted, with reliability being confirmed by a threshold discrimination analysis assessing the statistical stability of the validation metrics.

The research investigates if increased short-term insurance sales from a set number of quotes can be achieved when focusing on customised segments of potential customers. These customised segments are based on customer features (hence the literature review covering personalisation) which are identified from big data associated to the customers' existing financial products they currently enjoy. The research focuses on exposing the benefit of a cross-sell campaign using big data to create a propensity model that renders a higher conversion rate, meaning the insurer can see higher sales with the same quotes' cost; keeping quotes numbers stable and rendering higher sales would see a higher revenue against the same cost, and hence render a higher return on marketing investment.

1.2 Context of the study

Financial services providers, more specifically short-term insurance, long-term insurance, health insurance, investments and banking services providers play in a saturated South African market. An example of this is the short-term insurance sector, where the South African Insurance Association has 59 members (SAIA, 2016), many of them offering similar products. Due to the many insurers and resultant saturated market increasing price pressure (with insurers often competing on the lowest price), these financial services providers often have tight margins and resultant budgets are under pressure; this is evident by the same short-term insurance industry's often low underwriting margins, for example 0.8% in 2013 (KPMG, 2014).

Besides budget pressures, marketing specialists within this industry are having to work harder and find smarter ways to attract customers due to a variety of factors. These include:

- the internet and the ability for potential customers to shop online for insurance, often with the purchase incentive being price and not necessarily value or loyalty (PWC, 2016)
- the entrance of new direct market players, who often market lowest price offers (KingPrice, 2012), (OUTsurance, 1998), (Moneyweb, 2017), increasing the levels of competition in an already tough market.

With tight advertising budgets in most corporates (Danaher & Rust, 1996), fierce competition within the short-term insurance sector (Ziady, 2016), and the rapid increase in available data (worldwide data is estimated to double every forty months (Sicular, 2013)), this research aims to illustrate that personalised marketing activity using the same (or lesser) budget, through available big data, renders higher sales conversions, and therefore a higher return on marketing investment. This research intends to address a potential opportunity for marketers within the South African financial services, more specifically within the short-term insurance sector.

1.3 Problem statement

1.3.1 Main problem

Assess the value of big data on customized business development activity in order to increase return on marketing investment for a South African short-term insurer.

1.3.2 Sub-problems

1. The first sub-problem is to determine the predictive customer features by modelling on big data to embark on customised business development for a South African short-term insurer.
2. The second sub-problem is to evaluate the effectiveness of the chosen elements applicable to a South African short-term insurer in the form of ROMI.

1.4 Purpose of the study

There is research available that covers the importance of personalisation in marketing (Goldsmith, 1999). There is also research which has confirmed that there are greater successes in marketing campaigns which are personalised (Kim & Jun, 2008). Much of the research conducted has delved into digital distribution, primarily via cellular phones (Conti, Jennett, Maestre, & Sasse, 2012), (Shan, Chin, Sulaiman, & Muharam, 2016). This research paper bridges the gap in ascertaining links between big data and personalised marketing activity for South African-based financial services, particularly within short-term insurance.

Many banks and insurance companies hold vast amounts of data relating to their customers and their interactions with the company. This data is enlarged when the insurer provides products in various risk markets, for example in long-term insurance, short-term insurance, health insurance, credit card, investments

and wellness (Discovery, 2017) – this allows the insurer a broad and holistic view of the customer, their purchase behaviour, their lifetime value and effectively their customer equity.

The aim of the research is to understand the elements within an insurer's database which reflect predictive tendencies for their customers to purchase short-term insurance. From the analysis, the ability to thereafter implement individualized business development activity is outlined, with the intention of understanding how big data utilization can be applied in the design process of individualized marketing activity.

Finally, the research investigates the difference in traditional, generic business development outcomes versus individualized marketing campaigns and the comparative return on marketing investment (ROMI). ROMI in this context would be measured via conversion rates (the % of quotes that convert into active policies). The research endeavors to expose the financial incentive for the insurer to focus on campaigns that offer such higher ROMI results.

1.5 Research objectives

1.5.1 Theoretical objectives

The theoretical objectives of this research were:

- To review literature on big data and its current application into better understanding potential consumers,
- To explore research conducted on the impact of personalization and marketing,
- To discuss the marketing budget pressure linked to the tight margins within the short-term insurance industry, and
- To explore literature and research on the cost and returns of traditional, mass-media marketing campaigns – the research offers insight into personalized marketing campaign ROMI.

1.5.2 Empirical objectives

The empirical objectives are to statistically examine the links between the elements identified in the research in order to justify the arguments proposed within the research (Myers, Well, & Lorch, 2010). These objectives are realised through research of the following research questions:

1.6 Research questions

Sub-problem 1 -

Research question 1:
How can big data be used to expose customer-related features from non-short term insurance products which have the biggest impact on a client's short-term insurance purchasing propensity?

Sub-problem 2 -

Research question 2:
What is the ROMI impact of applying these big-data driven insights into customised business development activities?

1.7 Research gap and justification of the study

The link between personalized business development and higher ROMI within South African FSPs remains unclear and has not been conclusively resolved by academic literature in a structured manner – herein lies the research gap.

1.8. Significance of the study

This research study should add value to senior marketers who are accountable for budgets, marketing strategies and associated sales figures. The research

highlights the importance of using big data to understand individual consumer sales propensity based on existing financial products they enjoy, and the impact of personalised marketing activity relevant to an individualised audience. With up to 92% of advertising professionals believing in personalisation as a factor of sales activity success (Kim & Jun, 2008), this research will show that a tailored marketing approach can offer higher ROMI when compared to historical broad-based marketing campaigns, which are often expensive considering their mass media, above the line approach. Cold-calling which is a far cheaper form of business development also renders some of the lowest sales result (Rumbauskas Jr, 2010).

The research also illustrates the importance of turning big data into smart (useful) data, applicable and relevant to the market the South African insurer operates within, to maximise business development (sales) figures.

1.9. Delimitations of the study

The scope of this research does not include examining the underlying sub-elements of consumer variables (for example the sub-elements of a social-economic-status score). The study also only focuses on consumer data within a single FSP holdings company, and not multiple financial services providers within different umbrella holding companies.

This study does research the magnitude of various consumer variables which impact propensity to purchase a product, however there may be other variables not identified by this study. The aim of this study is not to investigate consumer data from customer bases outside of South Africa.

Common among research studies involving large datasets is the power problem, where significance of findings can be diluted due to the sheer size of the sample being researched (Fan, Han, & Liu, 2014). Statistical significance may be therefore small in some cases, however tiny correlations in vast datasets can translate into meaningful revenue streams.

1.10 Definition of terms

A list of terms frequently used in this research include:

- **Combinatorial explosion**: this is the drastic increase in complexity of a problem.
- **Comma-separated-values**: is a plain text file which contains lists of data
- **Conversion**: this is the ratio of when a potential customer is quoted for a product and accepts that said quote, thus becoming a customer, versus all customers quoted. For example, if there are 4 customers quoted and 2 accept the said quotes, then the conversion is 50%.
- **Decision Tree**: this is where the user migrates from observations on a factor (which are represented in branches) and then to conclusions regarding the factor's target value (which is represented in leaves).
- **False Negative**: number of quoted customers who converted but have a "did not convert" indicator test result.
- **False Negative Rate**: is the proportion of quoted customers who did convert but have a "did not convert" indicator test result.
- **False Positive**: is the number of quoted customers who didn't convert but have a converted indicator test result.
- **False Positive Rate**: is the proportion of quoted customers who didn't convert but have a converted indicator test result.
- **Loss ratio**: is the percentage of claims costs of an insurer versus the premiums collected.
- **Predictive Analytics**: is a range of business intelligence technologies which identifies patterns and relationships within large datasets which can be utilised to predict events and behaviour (Eckerson, 2007)
- **Positive Predictive Value**: is the proportion of quoted customers with a converted test result who did indeed convert when quoted.
    - This is identical to "precision" in the model.
- **Precision-Recall Curve**: this is the plot of precision (= PPV) versus recall (= sensitivity) for all potential cut-offs for a test.

- **Prevalence** – the proportion of quoted customers who did convert when quoted.
- **Recall**: this is identical to sensitivity (noted below)
- **Receiver Operating Characteristics**: is the plot of true positive proportion (= sensitivity) versus the false positive proportion (= 1 – sensitivity) for all potential cut-off for the specific test.
- **Sensitivity**: this is the proportion of quoted customers who did convert with the assay in question; this is calculated as TP / (TP + FN).
- **SHapley Additive exPlanations**: is method developed to explain individual predictions.
- **Social economic score**: proprietary information which is a standardised index of socio-economic status
- **Specificity**: is the proportion of quoted customers who did not convert with the assay in question; this is calculated as TN / (TN + FP)
- **True Negative**: this is the number of quoted customers who did not convert when quoted who also have a negative test result for the assay in question.
- **True Positive**: this is the number of quoted customers who did convert when quoted who also had a positive result for the assay in question.
- **True Positive Rate**: this is (similar to recall) identical to sensitivity.
- **Extreme Gradient Boosting Package**: is a software library that provides the regularizing gradient-boosting algorithm for this study.

1.11 Assumptions

The following assumptions were significant to establish a baseline for the research study:

- The sample of consumers used for the purpose of this research is representative of the rest of the population which purchase financial services provider products
- The financial services provider has logged and recorded pertinent customer interactions within the CRM

- The research findings should be applicable to other categories within the South African financial services sector
- The research results should not reflect bias as data was analysed based on actual consumer characteristics and choices, and not from questionnaires.

2 LITERATURE REVIEW

2.1 Introduction

This research study investigates the differences between data, information and resulting knowledge. The review thereafter offers context on why data and information is so important in the business world today, especially after the emergence of the data revolution, and the link between data and competitive advantage. Thereafter, the emergence of personalization as the new marketing responsibility is discussed, essential in today's market. Personalization's link and dependence on data is presented, along with the connection of enhanced marketing results attributed to personalization.

In addition, the short-term insurance industry competes in a soft market where many competitors undercut each other's rates, and resultant tighter business margins dictate the need for more cost-effective marketing campaigns which still need to render the same or even higher returns. It is common knowledge that add-on selling is more cost-efficient as customer acquisition costs are lower when inducing purchases from an existing client base (R. Blattberg, Getz, & Thomas, 2001); prioritizing marketing activities which cost less and render higher returns are thus strategically relevant in the current market environment.

The low margins within the short-term insurance sector, exacerbated by the soft market reducing revenue, places severe pressure on operational budgets, with marketing being no exception. With the South African economy and economic growth under strain, worsened by recent political events, disposable income is expected to remain tight within South African households, and insurance spend is expected to see continued pressure. Customers in search for better deals (often being only cheaper premium) will maintain the high churn cycles in the market, and initiatives to strengthen relationships with clients (for example, cross-selling) empowers the organization to learn more about their customers and consequently enables them to more offer competitive products. This also

has the positive effects of increasing an insurer's share of wallet, customer lifetime value, and resultant customer equity.

Finally, business analytics is shown to play a defining and integral role in marketing and business efficiency, with descriptive, prescriptive and predictive analytics explained. Various previous studies are summarized which researched cross-selling, database selling, next-product-to-buy models, recommender systems and propensity models.

2.2 Theoretical grounding

Big data allows an organisation to develop new offerings, while still reducing time and costs (Davenport & Dyché, 2013). Although marketers for a long time have been interested in identifying the deal-prone segment (R. Blattberg, Buesing, Peacock, & Sen, 1978), big data enables a marketing specialist to mould individualised marketing campaigns and thereafter improve the customer's experience (Goller & Hoffmann, 2013). Previous research has reiterated the importance of customisation and personalisation, noting this as the most important new idea in marketing, and further suggesting it should become a standard part of the marketing mix, forming the "8^{th} P" (Goldsmith, 1999). In addition, marketing budgets are under strain (Danaher & Rust, 1996) which is exacerbated by slowed economic growth within South Africa (Stats-SA, 2015). The study has theoretical roots in business analytics, specifically predictive analytics (with a big data foundation), individualisation, and financial domains.

2.3 Empirical review

2.3.1 Data, and its link to information and knowledge

Looking back over the last twenty years, data has been increasing on a large scale in various and different fields. The International Data Corporation in 2011 reported that the volume of data created and copied across the world was approximately 1.8zb (10^{21} bytes), which then increased by around nine times

over five years; it is expected to double at a minimum of every two years in the very near future (M. Chen, Mao, & Liu, 2014). Across an extensive variety of fields, data is in the process of being collected and amassed at an intensive pace - there is a crucial need to extract valuable information from the fast growing masses of digital data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

All companies have data, whether is it computed and stored; or it could simply be known and stored within the minds of employees. It's important to indicate that data is not necessarily knowledge that can be used for meaningful decisions (Ackoff, 1989). Bellinger et al., 2004 contends however that the structure is less involved than defined by Ackoff, where a diagram illustrates the transition from data, to information, to knowledge, and lastly to wisdom (refer to Appendix A). Bellinger contends that understanding supports the transition from each phase to the next, and furthermore understanding is not a detached, independent level. A definition of data, information and knowledge is (Beamish & Armistead, 2001):

- data = points of reality
- information = organised data
- knowledge = information, context and experience.

Companies must be able to convert data into knowledge to best leverage this into competitive advantage, especially if they plan to use this information to tailor communication or customise sale activity to prospective clients; the 1985 Harvard Business Review examined the importance of the information revolution, specifically in three essential ways (Porter & Millar, 1985):

- information transforms the industry structure, and in so doing, amends the rules of competition
- knowledge generates competitive advantage by providing organizations different and often new ways to overtake their competitors
- knowledge can spawn an entire new business, often from within the organisation's current operation.

Big data specifically is “a term describing the storage and analysis of large and or complex data sets using a series of techniques” (Ward & Barker, 2013); the explosive growth of data globally resulted in the term *big-data* being used to describe and designate huge datasets, often including huge masses of data in an unstructured format that required additional analysis, often in real-time (M. Chen et al., 2014). Big data, and its high variety information requires new methods of processing to allow superior decision-making, the discovery of new insights, and the optimization of processes – there are furthermore up to 10 factors which impact big data: volume, variety, velocity, veracity, validity, value, variability, venue, vocabulary, and vagueness (Moorthy et al., 2015). *Variety* is a special factor and is what makes big-data really big (Sagiroglu & Sinanc, 2013); it is specifically relevant to this proposed research considering data from various databases were integrated to create the research instrument. This data came from an insurer’s investments, long-term insurance, wellness and health insurance databases to create a single view of entities in the research population.

With a cellphone in almost every pocket, computers being used by billions around the world, and big information technology platforms used in many back offices globally – the fruit of the information age is evident (Mayer-Schönberger & Cukier, 2013). The massive strides in technology and various connected technologies has also resulted in an acceleration of data gatherers and consumers. The *Internet of Things* (IoT) and huge growth in cloud computing has further stimulated the growth of data. An example is the over 500 hours of video content being uploaded to Youtube every minute (Tankovska, 2021), or the sensors positioned across the world which not only collect but also transmit data which is then stored and thereafter processed within a cloud.

Presently, data has evolved in importance as a factor of production that is compared with human capital and material assets. IoT, social media and multimedia are developing, and organisations will continue to collect this additional information resulting in exponential data volume growth. This data

and big-data overall will continue to show massive and increasing potential in value creation for customers and business alike (M. Chen et al., 2014).

2.3.2 Data and personalisation

There is a resurgence of B2C, and personalised, directable and measurable communication (Kitchen, 2010). There is, however, a concern with many companies being unwilling, or perhaps unable to move beyond existing measurement tools and devices to any form of behavioural segmentation and associated measurement techniques. Behavioural measurement is dependent on understanding market dynamics, impossible without ongoing data and knowledge inputs. Through the years though, research into personalisation has resulted in various techniques which enable the creation of personalised services specific to customers' needs, constraints and interests; these are expressed by a user model (UM) which is an integral input for any personalisation technique (Berkovsky et al., 2007).

There is a real need for smart systems to advise and pre-empt while acknowledging personal interests and needs of customers, and thus to provide tailored offers in a manner which is most valuable and appropriate to the customer (Ricci, Rokach, & Shapira, 2011). Data and associated knowledge provides a foundation on which to personalize marketing communication and / or product offerings; personalization is the new marketing responsibility, and illustrates the transition of marketing from mass marketing to market segmentation, to niche marketing, to micro-marketing, to mass-customization and finally to personalization; furthermore the implications of personalization are segmentation and positioning on a more granular level, and personalization should therefore become a standard part of the marketing mix, forming the "8th P (Goldsmith, 1999).

There are specific applications for big data, the first two being essential for marketing application (H. Chen, Chiang, & Storey, 2012):

- Recommender systems
- Social media monitoring and analysis
- Crowd-sourcing systems
- Social and virtual games

Besides using big data for recommender systems (essential in cross-selling campaigns) big data allows a marketer to tailor personalised marketing campaigns, such as coupon offers, and enhance the customer experience (Goller & Hoffmann, 2013). Recommendation systems generally operate on three variations of data models – these can be generally classed as a ratings data model, a unary data model and a binary data model (Bodapati, 2008). Also vital to consider is the aspect of context in recommendation systems (especially multidimensional recommendation models); context relating to preferences of the customer (also linked to consumer behaviour) is vital and can often provide a better recommendation and thus better results (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005).

In a ratings data model (the most common) (Herlocker, 2000), a user records their vote in a scale system - if the user isn't willing to place their vote, or is unaware of the item, the element is recorded as N/A; the binary data system is a shortened version of the ratings model where a user indicates either a positive result (recorded as 1) or a negative result (recorded as 0) should they not have an intention to purchase the product. If the intention of the user is missing, it is recorded as N/A. The unary data model is an amended version of the binary data model where only positive results are detected, and negative results are noted as N/A – therefore the N/A results reflect the user having a negative result or not being aware of the item (Bodapati, 2008).

Cellular devices also offer a platform on which a marketer can secure detailed profiling and data mining based on the user's web history - browser cache and keyword extracts from emails and social media can be used to send customized advertisements to potential consumers via cellular devices (Haddadi, Hui, & Brown, 2010).

Based on big data collected from loyalty cards, retailers (for example Tesco), and many e-commerce retailers use this data to personalise their offerings and their pricing – sellers can thereby compete for the customer, instead of an "average" customer; it is suggested that personalisation is good for the consumer as more consumers will purchase the product (Shaw & Vulkan, 2012).

In financial services, specifically banking, there is an emphasis being placed on customer relationship management through the accelerated use of information technology (Bremner, 2001); furthermore there is the view that connections to consumers via the internet can be utilized to construct relationships with customers, also allowing banks to tailor and personalize offerings (Ibbotson & Moran, 2003). This is reinforced by the view that the internet (and associated data) offers firms unparalleled opportunities to secure detailed customer insights and for customizing (personalizing) value propositions to satisfy their preferences (Reichheld & Schefter, 2000). Furthermore up to 92% of advertising professionals believe a marketing message must be personalised to be successful (Kim & Jun, 2008).

Important to add, with the current focus on integrated marketing communication (IMC), big-data driven personalisation can empower the transition from mass-market advertisements to more tailored and targeted messaging and business development activities (D. E. Schultz, 1999); this approach is a basic first step in the evolution from product-driven, outbound approaches to the modern consumer-and-behaviour-orientated, interactive marketing approaches seen today (Kitchen & Schultz, 1999). Consequently, IMC is often recommended as the solution for realising advertising planning and execution synergy, with downstream results including improved productivity, efficiency and overall performance (Peltier, Schibrowsky, & Schultz, 2003).

It must be noted though that previous purchasing behaviour data stored within customer databases does have limitations when used to target new services or products, as the new service or product is likely not to be identical or even

similar to existing products and services; there is thus the need to generalise customer purchase patterns from their existing product responses in relation to the new products on offer (Kamakura, Kossar, & Wedel, 2004). Furthermore, customers are often aware of deals and prices in a market, and when examining purchase data, one can witness households acquiring more value from their purchases when compared to an average buyer (Krishna, Currim, & Shoemaker, 1991).

Also important to note is that big data-empowered cross-selling can possibly be detrimental to the organisation's relationship with their client, as numerous attempts to upsell or cross-sell may result in the customer becoming non-responsive, perhaps even encouraging the customer to switch to a competitor (Kamakura, Wedel, De Rosa, & Mazzon, 2003); personalisation and focused business development offers some protection against this as the organisation can potentially offer a product or service to the right client, at the right time.

2.3.3 Big data and predictive analytics

Various studies involving database marketing have occurred: in 1991, cross-selling opportunities were first studied, where a latest trait analysis positioned financial services providers along a common continuum along with investors (Kamakura, Ramaswami, & Srivastava, 1991). In 2004 a split hazard rate model predicted a physician time to adopt new drugs linking in prior adoptions of various drugs (Kamakura et al., 2004). In 2002, a NPTB model was presented looking at logistical regression, multinomial logit, neural nets and discriminant analysis, which predicted what product a customer was potentially likely to acquire next (Knott, Hayes, & Neslin, 2002). In 2003 a broad approach to recognise multivariate binomial probit model restrictions was proposed (Edwards & Allenby, 2003), and in the same year a mixed data analysis was conducted which extended factor analysis to include various other data types allowing a marketer to tailor their customer approach (Kamakura et al., 2003). In 2005, a structural multivariate model evaluated how customers demand various

products over time, and the sequential patterns of acquisition, specifically in the banking sector (Li, Sun, & Wilcox, 2005).

Refer to Appendix B for the map of where database mining and analytics fits together. Business analytics is essentially view from three perspectives, being descriptive, prescriptive, and lastly predictive (Evans & Lindner, 2012). Descriptive analytics is most commonly used and summarises data into charts and reports (for example about budgets, or revenue). Prescriptive analytics utilises optimisation to select the best alternatives to maximise or minimise a given objective (for example when selecting the best possible pricing strategy to increase revenue). Predictive analytics analyses past data and thereafter identifies patterns, thereafter extrapolating those relationships or patterns forward in time (an example is a marketing specialist wanting to predict responses from different consumer segments to an advertisement).

The above literature is all important as it highlights the foundation and environment the intended propensity model and predictive analytics are linked to. Predictive analytics includes statistical models and various other empirical methods which aim to create empirical predictions (versus theoretical predictions only), in addition to evaluating the prediction quality (prediction power) in practice; it also includes the building and assessment of a quantitative empirical model which often includes data mining algorithms intended for predicting future or new scenarios (Shmueli & Koppius, 2011). It is understood that big data and data mining through data science provides an opportunity to modify business model designs as well as daily decision-making models though emerging data analysis and predictive analytics (Waller & Fawcett, 2013).

In marketing specifically, predictive analytics can play a crucial role where models can assess historic behaviour to evaluate the propensity of a client to specific behaviours in the future. Descriptive models define data relationships and are utilised to classify prospects or events into groups. Decision models define the relationships between all elements of a decision, including the known factors of the data relationships, the actual decision, as well as the predicted

decision outcome (Hair, 2007). See the Appendix C for the data mining and predictive analytics process chain.

Predictive analytics modelling has been successfully rolled out in wireless telecommunications, fraud in healthcare and insurance, as well as predicting consumer losses in financial services (Rubel, 2006). In marketing specifically, Google has made vast inroads into adaptive marketing which utilizes insights from customer interactions. Customer interactions against marketing campaigns are stored in databases where results are available to be used again to retrain and refine predictive models, as well as to add additional insights into how customers respond to marketing campaign-delivered offers (Reed, Ansusinha, & Hariharan, 2010).

Research has also been conducted into the application of propensity score methodology to assess marketing interventions, such as promotions and advertising. It is understood that companies when engaging on outbound sales campaigns should work from a list of potential customers, starting with the most likely to create further revenue to the least likely in doing so. The reasoning for this is that these efforts require an investment, and so basic return on investment principles apply when a company should target segments of potential customers based on importance (Rust, Lemon, & Zeithaml, 2004), i.e. those with the highest sales conversion potential.

Propensity scoring methods are applicable where there is a requirement to evaluate the effect of an intervention; they are also extremely relevant where the intervention is not applied via a randomised basis and where background variables influence the result; an example of such is an e-commerce example, where an online shop can test the propensity between customers receiving newsletters and their purchase patterns and revenue contributions (Rubin & Waterman, 2006). Tiny correlations in vast datasets can translate into meaningful revenue streams, and one can increase profitability when even small but meaningful relationships are identified, being predictors that increase ROMI.

2.3.3.1 Research question 1:

How can big data be used to expose customer-related features from non-short term insurance products which have the biggest impact on a client's short-term insurance purchasing propensity?

2.3.4 Tight margins and a soft market

South African short-term insurance providers specifically are seeing huge pressure with multiple players offering short-term products, a consumer base with disposable income pressure and client perceptions of insurance being commoditised. Disposable income pressure is expected to become worse following the economic factors in South Africa over the last few years which intensifies the struggle insurers have over new business. Following various finance minister reshuffles and the Rand's at times above-30% depreciation against the US Dollar (Old-Mutual, 2016), high levels of unemployment exceeding 27% (Trading-Economics, 2017), the downgrades by various ratings agencies (Fitch-Ratings, 2017), the pre-Covid potential for even higher inflation which can again result in further interest rate increases (Personal-Finance, 2017). The South African consumer is expected to experience continued financial pressure: this exacerbates an already soft market. Covid-19 has added complexity where the pandemic has placed additional pressure on the South African economy with increased debt-to-GDP ratio and loss of additional jobs (de Villiers, Cerbone, & Van Zijl, 2020) – these factors placed even more strain on household affordability and contributing to soft markets.

As the market becomes more saturated and more competitive, these new consumers are often switchers, who are often likely to switch in the future to another attractive offer they may encounter (Kamakura et al., 2003). This churn can be combatted by an organisation strengthening their relationships with their clients. Customer relationship management is imperative, with one key tool for strengthening customer relationships being cross-selling (Kamakura et al., 1991). A significant benefit of cross-selling not always considered is that the

firm learns more about their client preferences (due to higher share of wallet) and can potentially better satisfy their client needs versus competitors; cross-selling and upselling also reduce a customer churn rate due to switching costs which increase as additional products or services are acquired by the customer (Kamakura et al., 2003).

Cross-selling also empowers an organisation to learn more about their customers' wants and needs, whereupon the organisation is able to foster increasing client loyalty which defends them from competitor attacks (Kamakura et al., 2003); simultaneously, an increase in loyalty results in increased customer lifetime value, which increases customer equity and resultant profitability (Kahreh & Kahreh, 2012).

With the constant search for higher revenue with embedded efficiencies, cross-selling effectiveness can be further enhanced by a NPTB model, which based on customer knowledge specifies which product to target which customer with (Knott et al., 2002). Research has discovered that response rates and revenue per customer is far higher with NPTB campaigns. The NPTB model:

- commences with the compiling of customer-specific data through a period of time, in addition to observing what the customer purchased in this period
- requires a statistical model to be selected, which can include multinomial regression, logistical regression, neural nets and discriminant analysis
- necessitates one to estimate and evaluate the model, which is similar irrespective of the technique used: the user observes the customer predictor variables together with the product type the customer chooses, and the software evaluates the coefficients required to predict the results
- finally scores and targets customers: all the above-mentioned techniques produce probabilities that score each customer for a specific product, with the issue then being how far down the database-generated list to target (Knott et al., 2002).

In addition, five themes that are emerging and most noticeable (Day & Montgomery, 1999), and especially relevant in the short-term insurance's tightly margined market today are:

- An economy of connected knowledge
- Industries consolidating and converging
- Frictionless markets
- Empowered customers who are also very demanding
- Organisations which need to adapt.

Big data-driven marketing takes connected knowledge from various data sources, for a company which is willing to adapt (which may have various financial services offerings which have converged) to provide tailored and market relevant offers to a demanding and empowered (and soft) market, with potentially higher returns.

This is especially important where in the short-term insurance industry, underwriting performances are not expected to be stellar. Even in years when there are not significant losses (claims costs), insurers continue to battle to obtain the desired premium rates considering the marketing remains incredibly competitive and soft (KPMG, 2016). Data-driven sales initiatives can reduce the cost of marketing and associated business acquisition costs, lessening overall expenses and allowing for a greater underwriting and profit margins.

2.3.5 Return on marketing investment (ROMI)

Return on investment is a metric which is extensively used and recognised in practice and fulfils the business need to understand profit / or benefit from an investment; in a simple form, return on investment is defined as the net benefit from an investment (or undertaking) over the actual cost of the investment (Kaske, Kugler, & Smolnik, 2012). Marketing return on investment is therefore the calculation of the proceeds or net returns from a marketing investment, in relation to the cost of the marketing activity.

Returns can also be measured in terms of economic value add (EVA), as well as intangible assets like brand equity, relational equity and customer equity (Seggie, Cavusgil, & Phelan, 2007). EVA measures look at the financial performance of residual equity which is calculated by subtracting the cost of capital from the firm's operating profit after tax (Stewart, 1993).

Brand equity return on the other hand, is an intangible asset return which can be measured through perceived quality measures, awareness measures, differentiation measures, and loyalty measures which may include measures of switching costs, satisfaction, and commitment (Aaker, 1996). Relational equity draws on relationship marketing and is the potential to create wealth which lies in an organisation's stakeholder relationships (Sawhney & Zabin, 2002), and the relational equity scorecard (RES) offers overall measures which concentrate on relationship quality (Seggie et al., 2007).

Customer equity focuses on the customer (unlike brand equity which focuses on the brand) and is calculated by summing customer lifetime value (CLV); this equity measure concentrates on three drivers: what the customer values, how the customer assesses the brand (above perceived value), and loyalty to the brand (Lemon, Rust, & Zeithaml, 2001). Academic studies have shown a relationship between marketing investments and increases in CLV, which as a consequence also increases customer equity (Rust et al., 2004), effectively making customer equity a ROMI measure.

Using these alternative measures of marketing return is important as it is difficult to assess the returns offered by marketing communication and business development efforts when sales are the main basis for the evaluation, simply because there are many other factors which contribute to sales (Koekemoer, 2014). For the purpose of this study, ROMI is measured via the quote conversion rate, where a higher conversion rate means a higher sales count for the same number of quotes and associated quote costs.

2.3.6 ROMI Constraints

With slowed economic growth resulting in tighter budgets, many organisations are learning how to do more, with less. Marketing is no exception, and the use of data for smarter, tailored marketing communication and business development is relevant in today's environment of tighter margins, where advertising increasingly finds itself in a climate of cost cutting, and so smarter and more efficient options must be explored (Danaher & Rust, 1996). It is no secret that market budgets remain under close scrutiny, especially when in many industries the general management and operational costs decline, and marketing officers are facing increasing pressure to become more productive and efficient (Weber, 2002). Gatekeepers of corporate budgets are often concerned with these cost differences and it has been an ongoing concern.

In one study, manufacturing went from 50% to 30% of total costs, management costs decreased from 30% to 20%, yet marketing went from 20% to 50% of total costs (Sheth & Sisodia, 1995). Marketing teams witness a gradual decrease in their resources if they don't show real value and often times must again do more, with less. Work has been done that shows marketing results can increase while spending less with the objective being not the increase in the marketing budget, but the reduction of marketing costs (DE Schultz, 1996).

Furthermore, in another study where 500 marketing programmes were analysed, most sales promotions were found to be unprofitable, and advertising ROI was under 4% (Stone & Clancy, 2005). With direct mail or telemarketing, it is vital that the people the organisation communicates with are those people who have the highest probability of response; this is most relevant when the individualised information (available through company databases and big data manipulation) about the targeted population is available (Bult & Wansbeek, 1995).

There has been success with personalisation producing higher revenue and greater client retention in certain sectors such as hospitality (Arora et al., 2008).

There has also been success with customizing (personalizing) pricing to consumers based on their market segments and associated preferred value propositions, where market share was increased by 17% and profits by 45% (Simon & Butscher, 2001).

Database marketing is argued to besides developing the customer-company relationship and creating sustainable competitive advantage, to also increase marketing productivity (R. C. Blattberg, Kim, & Neslin, 2008). In addition, organisations which are customer-orientated are privy to a massive amount of information about their clients, which is becoming very valuable to marketers (Goodman, 1992). Marketers can use these databases to engage in database marketing which effectively is building, arranging, supplementing and thereafter mining client transactional databases to improve the accuracy and efficiency of marketing efforts by identifying the best prospects for marketing initiatives (Labe, 1994). Additionally, further enhancements through NPTB models can see marketing returns as high as 530% (Knott et al., 2002).

South Africa overall has shown slowed economic growth over the last few years with 2015 showing 1.3% growth (Statistics-South-Africa, 2016a), and a continued decline in GDP per capita between 2016 and 2020 (Statistics-South-Africa, 2020), refer to Appendix D; the private sector has also become less profitable due to higher costs (Statistics-South-Africa, 2015). Looking at the last 15 years overall, the South African business sector does generate higher revenue amounts every year, however, costs have increased at an even faster rate. Accordingly, profit margins have declined in relative terms looking at the percentage of revenue (Statistics-South-Africa, 2016b). When looking at the short-term insurance industry specifically, expense ratios (which impacts margin) continue to increase with 2015 showing a 1.2% expense ratio increase for the industry overall (KPMG, 2016). When expenses and expense ratios increase, it’s a matter of time before budgets are scrutinised in more detail and often times reduced to curb the upward expenditure trend.

The following research questions are therefore applicable, and are also relevant due to the tighter budgets linked to lower growth and / or lower profitability witnessed within the SA business sector today:

2.3.5.2 Research question 2:

What is the ROMI impact of applying these big-data driven insights into customised business development activities?

2.3.7 Hypothesis

The key variables in this study are as follows:

- Independent variable 1 (IV_1): short-term insurance conversion rate from non-customized business development
- Independent variable 2 (IV_2): short-term insurance conversion rate from big data-driven customized business development
- Dependent variable (DV): return on marketing investment (ROMI)

There are 2 hypotheses:

- Hypothesis 1 (H_1): short-term insurance sales per 100 quotes resulting from non-customized business development attains lower ROMI
- Hypothesis 2 (H_2): short-term insurance sales per 100 quotes resulting from big data-driven customized business development offers increased revenue and associated higher ROMI.

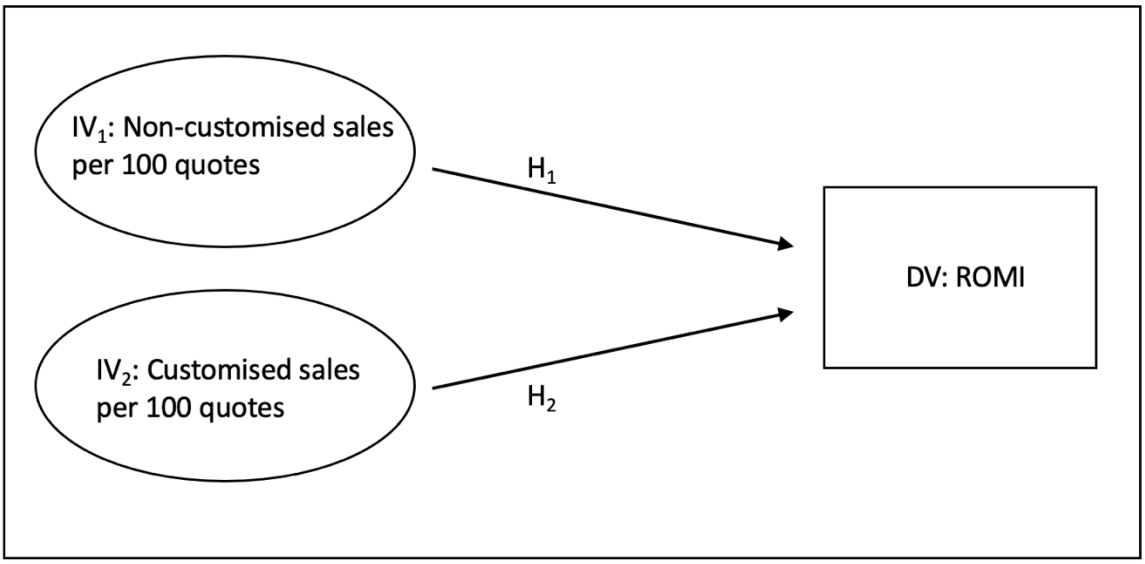

Figure 1: Hypothesis

2.4 Conclusion of literature review

In summary, the literature discusses the differences between data, information and knowledge - the information revolution's importance is also revealed. The importance of personalisation and understanding personal attributes and associated customer factors (and its link to big data) is evident, and is an emerging marketing responsibility. There has been a resurgence of personalised communications and business development but there is still concern that many companies are not undertaking any behavioural segmentation or associated measurement techniques, neither are they understanding market dynamics, which is impossible without ongoing data and knowledge inputs.

Big data can be used in marketing in areas such as social media analysis and sales recommender systems, which both have personalisation linkages. With the current prevalence of cellular and other digital devices, a vast platform is available off which marketers can secure personalised information. Personalised communication's importance is evident with many advertising professionals stating communication success is dependent on personalisation.

Use of databases for marketing empowers organisations to efficiently leverage their customer knowledge - this improves the sales effort yield whilst reducing the risk of irritating customers with unexciting or irrelevant offers; the vast amounts of demographic and behavioural customer data firms have allows them to tailor their marketing needs and strengthen their ties with their customers (Kamakura et al., 2003). Data mining and associated business analytics, more specifically predictive analytics, is a cost-effective tool to run smart, personalised cross-sell offers, which should render higher sales conversation performance, increase an insurer's share of their customer's wallet, strengthen the customer relationship, and reduce churn.

Promotional and marketing efficiency is also important in today's business environment with higher costs, lower margins, tighter budgets and economic pressure causing a softer short-term insurance market. This is coupled with traditional advertising often realising extremely low returns (ROMI), whilst increased revenue and better client retention is seen with personalization in some sectors.

The next chapter (chapter three), presents the research methodology.

3 RESEARCH METHODOLOGY

3.1 Research philosophy

This chapter classifies and describes the research methodology which was used in this research. On a high level, it has three intentions; to name and define the research strategy (in section 3.2), research design (in section 3.3), in addition to the research procedure and methods (in section 3.4). This chapter thereafter discusses the target population and sampling (section 3.5), and then expands on the ethical considerations when collecting data (section 3.6). Section 3.7 unpacks data collection and storage, with the following section discussing the processing and analysis of the data (section 3.8). The description of the respondents is discussed (section 3.9) and reliability and the validity measures are thereafter expanded on (in section 3.10) to ensure credibility. Lastly, administrative and technical limitations relating to the research procedure and methods are listed in section 3.11.

3.2 Research strategy

Research strategy is the research approach, which depends on the questions being researched, and the nature of information being sought (Kolb, 2010). This strategy can also be defined as the path adopted to discover the answers to research questions, either being the structured, or unstructured approach (Kumar, 2011), or as the broad orientation to the manner of research (Bryman, 2012). There are three types of research strategies: quantitative, qualitative and mixed methods, of which the latter is a combination of quantitative and qualitative methods. The strategy for this research is strictly quantitative.

Quantitative research is a research approach which is grounded on scientific-based principles utilized when proof of fact is required, or when the questions being researched deals with descriptive facts, for example “how many” or “who”

(Kolb, 2010). This research strategy is also a way to assess objective theories by analysing the relationships among variables – instruments are able to measure these variables, in order for the numbered data to be analysed by utilizing statistical procedures (Creswell, 2014). Quantitative research furthermore stresses quantification in both the collection and analysis of the data, entailing a deductive approach when analysing the relationships between research and theory - the focus is placed on testing of the theories (Bryman, 2012).

The research committed to a quantitative strategy considering a big-data dataset was analysed to determine the relationships between multiple variables in business development initiatives which impact sales conversion in a South African short-term insurer. This research paper took note of previous quantitative studies, some of which involved the use of big data.

In a 2008 study evaluating the passive mobile positioning (big) data specifically for tourism surveys, quantitative research was applied (Ahas, Aasa, Roose, Mark, & Silm, 2008). The objective of this research was to evaluate and introduce the applicability of mobile positioning data in studying tourism, and considering precise measurement was required for analyzing the mobile positioning data, quantitative research was applied. This research report, being also focused on big data is also purely numerical with no room for information being sought on feelings, beliefs or attitudes.

In another big-data-related study, (Datar & Sturm, 2004), a national sample of 9,751 kindergarten children in the United States were reported on for two years in a research exercise which investigated the relationship between physical education and body mass index (BMI). A quantitative research approach was applied as it was an empirical study with a large dataset.

A further research study (Ang & Buttle, 2006), investigating the linkages between management processes and customer retention outcomes, (including client retention planning, budgeting, and the existence of a formal complaint-handling process), utilized a quantitative strategy; this was necessary in order to

study the relationship between the multiple variables harvested from a questionnaire.

A quantitative approach is beneficial to this research considering this study assesses the empirical and numerical relationships between multiple factors from a big-data dataset.

3.3 Research design

The research design details how the research is conducted, inclusive of by whom, when and where (Kolb, 2010). The research design can also be defined as an arrangement of decisions relating to what topic is being studied, in what population, with which research methods, and finally for what purpose (E. R. Babbie, 2014). The research design is the framework covering the collection and analysis of data, which indicates decisions relating to the priority provided to a range of dimensions within the research process (Bryman, 2012). Bryman furthermore notes five broad research designs, being cross-sectional, longitudinal, case study, comparative, and experimental.

This proposed research uses two research design elements, being longitudinal, and cross-sectional. Cross-sectional analysis is the the grouping of data on numerous cases, at a single point in time, in order to collect quantifiable / quantitative data with linkage to at least two (often more) variables, which are examined to identify patterns of association (Bryman, 2012). Cross-sectional designs can further be explained as involving observations of a cross-section of a population or phenomenon, which are made at a single point in time (E. R. Babbie, 2014). An advantage of cross-sectional design that this design is well suited to studies attempting to identify the prevalence of a phenomenon (Kumar, 2011). The disadvantage however, is that this analysis cannot be utilized over a period of time.

The longitudinal element of this research is to mitigate the disadvantage of the cross-sectional design being unable to collect data over a period of time. This is

possible because the data being studied has been collected on the data subjects repeatedly, over an extended period of time. Longitudinal research is a design where data is collected from a sample on a minimum of two occasions (Bryman, 2012), effectively data collected over time (Creswell, 2014). This research study also examined prior studies that utilized a longitudinal and cross-section designs.

In a 2004 research study examining the relationship between levels of blood vitamin D and prostate cancer, a longitudinal quantitative research design was applied (Tuohimaa et al., 2004). This research benefited from such a research design as the research required the multiple samples to be harvested from the research subjects at various points in time, similar to how the current research proposed analyses data points collected from consumers also at various points in time.

In another study where researchers sought to clarify the relationship between personality and social networking sites examining the social and informational use of Facebook and Twitter, a cross-sectional design was utilized (Hughes, Rowe, Batey, & Lee, 2012). The benefits these researchers enjoyed was that multiple variables at a single point in time within the data snapshot were able to be analyzed. This research also contains multiple variables secured at the time the big-data dataset was compiled.

3.4 Research procedure and methods

3.4.1 Data collection instrument

A data collection instrument, is a precise tool utilized for collecting data from a sample, in order for the research questions to be answered (Bryman, 2012). It can also be described as the research tool, being a specific mechanism that the researcher utilizes to gather, manipulate, or interpret the data (Leedy & Ormrod, 2013). Anything used as a means of gathering information for research is termed a research tool or a research instrument, and the formation of the

research instrument is the initial and practical stage in conducting research (Kumar, 2011). There are various forms of a data collection instrument: unstructured and structured interviews, observations and questionnaires. This research was conducted on a big-data dataset which has been customized to the research requirements. As noted prior, this study investigated previous research papers where big-data was utilized.

In a 2008 study where passive mobile positioning data was investigated as a tool for studying tourism (Ahas et al., 2008), a big-data dataset was used. This instrument was convenient, as the research team was based in the same region the cellular data was derived from, and furthermore the research team enjoyed access to the data, provided by a local cellular provider. The same benefits apply to this research as it is beneficial to perform research on customers within the South African market, using data already being stored in the vast data warehouses within a financial services provider.

In another study, longitudinal data fusion was conducted to provide evidence of the influence of social media on the expansion of retail brands within the American marketplace (Don Schultz & Block, 2014). This big-data instrument was again chosen due to availability of data from a commercial database holding the multiple responses to online questionnaires, which were based on a sample that was nationally projectable of the American population. Also, in the already-mentioned study (Datar & Sturm, 2004), where 9,751 kindergarten-age children were researched to identify the relationship between physical education and body mass index, the big-data dataset was available as the National Center for Educational Statistics was already in possession of the data.

This relates in a similar fashion to this research utilizing data extracted from multiple internal databases and curated to provide the big-data dataset, representative of the South African population. The data being used was from the various staff members within the financial services provider inputting information into the customer relationship manager (CRM) platform – this commences at lead generation stage, to quoting for new business, to acceptance of new products, to amending, upselling and cross-selling of

existing products, to finally cancellation of products. A specific data extract collected this information on 566,758 customer profiles, inclusive of details of financial products they enjoyed, which is pertinent to the study.

There is no instrument such as a questionnaire, as the research was conducted on a quantitative dataset (proprietary data the researcher was privy to) which was tailored to the research requirements (hence no research instrument in the appendix).

The data used in this study excludes personal client information which cannot be included due to any privacy restrictions and / or POPI requirements. The dataset includes specific details and transactions per customer (Kamakura et al., 2003), for example (but not limited to):

- age of client
- number of persons in client's household
- age of policy
- number of various financial products owned at points in time
- types of various financial products owned at points in time
- engagement and utilisation of product(s) at points in time
- internally developed financial credit scores at points in time
- historic purchasing behaviour at a point in time.

The necessary permissions have been obtained from the financial services provider in accordance with ethical requirements, see Appendix E.

3.5 Target population and sampling

The target population is the people or records or events that contain the sought-after information which can therefore answer the measurement questions (Cooper & Schindler, 2011). It can also be described as the universe of components where the sample is selected from; the researcher could sample from a universe of cities, regions, nations, etc (Bryman, 2012). It can further be

termed as the sample frame where the ultimate participants will be chosen from (Kolb, 2010). The research conducted here considered previous studies in this regard.

In the 2014 case study where big data was used to research the relationship between social media and brand preference, the target population was a nationally-projectable sample of the U.S. population (Don Schultz & Block, 2014). The big-data dataset consisted of responses from the US population, and since the researchers were based in the US, this data was convenient and accessible.

In another 2008 study, researchers in Estonia analyzed the applicability of mobile-positioning data when studying tourism (Ahas et al., 2008). The target population was all visitors to Estonia, again geographically convenient to the research team who were based in Estonia, especially since the data was readily available from a local cellular provider.

In 2004, Nordic researchers based their research focusing on blood vitamin D with prostate cancer on a target population of Nordic males from Sweden, Finland and Norway (Tuohimaa et al., 2004) – the team of researchers had access to three Nordic bio-banks with appropriate medical samples, and so this again was convenient and accessible.

This research focuses on a target population of a South African FSP's customer base. The population dataset has over three-million individual consumer profiles, representative of all religions, races, cultures, occupations and ages (above eighteen years old) – this target population is representative and the data from the sample is both convenient and accessible.

Sampling is the selection of the segments of the population for investigation, where a sample is a subset within the population (Bryman, 2012). Researchers should first have access to everyone in the target population, thereafter a sample within the population will be selected to participate (Kolb, 2010). An important aspect of research is deciding what to observe, and the procedure of

selecting observations is termed sampling (E. R. Babbie, 2014). There are various types of sampling techniques in quantitative research, which include probability, random, simple-random, systematic stratified random, and multi-stage cluster sampling. This research utilized a stratified sampling technique, where the sample frame for this research study were the customers of the FSP who did not have a short-term insurance product who were quoted between 01 January 2018 and 15 March 2021 for a short-term insurance quote.

Stratified sampling is when the sample is controlled to include factors from each of the segments (Cooper & Schindler, 2011), and requires the population being researched to be divided by criteria into groups, termed strata; furthermore random samples can then be taken from the required strata (Bryman, 2012). This sampling technique is used when the research calls for compelling results between particular groups within a population (Kolb, 2010). This research however used the complete sample.

In the afore-mentioned case study, the Nordic researchers investigating the linkage of blood vitamin D with prostate cancer divided their population between those with and those without cancer, being placed in strata's accordingly (Tuohimaa et al., 2004). The Estonian case study analyzing the applicability of mobile-positioning data when studying tourism placed the population in tourist / non-tourist strata's (Ahas et al., 2008). The benefit of the proposed research using this sampling technique allows the population to be divided into various strata's, with each strata representing customers within a particular product-house within the insurer holding company.

The population was segmented into various strata's, with each strata representing clients of a specific product house within the FSP holding company. The various strata's can be found in Appendix F.

The number of subjects from where data is collected from is termed the sample size (Kumar, 2005). The overall sample of this study is 566,758 customers - this large sample size is appropriate considering big data is a research point. Secondly, this large sample reduces bias risk.

There are a vast number of data points, considering each customer could also be a client of the insurer's other product houses, for example health insurance, wellness and other risk products such as investments. Each product house stores relevant data on these individuals, which was integrated and cross referenced in the research.

3.6 Ethical considerations when collecting data

When one gathers data from respondents or involves study subjects in the experiment, one should carefully examine if harm from their involvement is probable (Kumar, 2011). It is imperative that research is designed to ensure a participant does not endure physical harm, pain, discomfort, embarrassment, or the loss of privacy (Cooper & Schindler, 2011). Furthermore, researchers must not expose research subjects to unnecessary psychological or physical harm; the common rule of thumb is that the risk associated to participating in a research study must not be greater than the regular risk of daily living (Leedy & Ormrod, 2013).

The research conducted utilized existing big data, lying within an insurer's CRM database. The ethical risk of this data is therefore negligible, and because there is no further data collection from respondents required, there is no possibility of deceiving, harming or stressing any respondents. In line with Protection of Personal Information Act (SA-Government, 2014), sensitive information, such as identity numbers, names, addresses, and contact details are not revealed, or even included in the dataset. The necessary permissions have been obtained from the insurer in accordance with ethical requirements, see Appendix E for details.

3.7 Data collection rigour and storage

Data collection is the process of collecting data from a sample so the relevant research questions can be answered and outcomes analysed (Bryman, 2012).

Data collection is obviously undertaken after selecting the research approach, and methods include focus groups, surveys, observation, interviews (Kolb, 2010), or even the reanalysis of already-created statistics (E. R. Babbie, 2014). This research involved the analysis of existing statistics from existing databases, extracted from an insurer's system, and is therefore cost-effective.
Numerous historic studies have made use of existing data, including the afore-mentioned studies where big data was used to research the relationship between social media and brand (Don Schultz & Block, 2014), as well as the applicability of mobile-positioning data when studying tourism (Ahas et al., 2008), in addition to when Nordic researchers analysed the relationship between on blood vitamin D and prostate cancer (Tuohimaa et al., 2004). The data in such studies was accessible, readily available and therefore convenient to use, in addition to being cost-effective. These same benefits apply to this research which analyses the data which already exists. A further advantage of this proposed data collection is the generally detailed granularity of the data held by an insurer within their datasets.

This research utilized data from various databases within the insurer's long-term insurance, health insurance, wellness, credit card and investment business areas. The targeted insurer uses a single unique identifier for each client throughout their different business areas, improving the quality of the data. Pricing, risk management factors and client interaction data was included in the dataset requirement specification, from which a SQL script was written to extract the data from the various databases. A statistician was responsible to extract the data into CSV format, ensuring consistency, quality and completeness of data.

The data, in its final and usable form was stored within the insurer's own server, ensuring the researcher worked within the insurer's environment to safeguard the security of the data. Post research, the insurer can further enhance this dataset for further internal analysis and business improvement where the insurer remains in possession of the data; the insurer was also notified of the option to discard the dataset to prevent the possibility of the data landing into the possession of others who may misappropriate it (Creswell, 2014).

3.8 Data processing and analysis

Data processing includes all actions undertaken from where a set of data is gathered until it is ready for analysis, either by a computer, or manually (Kumar, 2011). This can also refer to the organization of data which includes the gathering and transformation of collected information (Kolb, 2010). Furthermore, quantitative researchers select methods which allow them to measure the variables of interest objectively; this also allows the researcher to remain disconnected from the participants and phenomena they are noticing - unbiased conclusions are therefore possible (Leedy & Ormrod, 2013).

Once the data had been extracted, the subsequent steps are the loading of data, pre-processing and finally model building and evaluation of results. Coding is the process where data is broken down into various component parts, with these parts then given labels (Bryman, 2012). It involves assigning symbols or numbers to answers, so the responses can be segmented into a certain number of categories (Cooper & Schindler, 2011). It is important to further note that this process is coupled with a retrieval system of sorts (E. R. Babbie, 2014).

Data entry transforms information secured from primary or secondary methods into a computer for viewing and manipulation (Cooper & Schindler, 2011). This process is usually automated, so researchers do not necessarily have to manually enter the data via a spreadsheet – errors in data entry are mostly then avoided (Bryman, 2012). Considering the research utilized existing data, the data was already be in an electronic format, similar to the afore-mentioned study evaluating mobile positioning data for tourism surveys (Ahas et al., 2008) and the already-mentioned big-data file fusion research (Don Schultz & Block, 2014). An electronic spreadsheet is an important tool, which empowers researchers to input and thereafter manipulate data; software can therefore easily and quickly arrange the data, along with making calculations (Leedy & Ormrod, 2013).

After data entry, the data was cleaned or edited, which ensures it is free from incompleteness and inconsistencies (Kumar, 2011). Many software programmes include the functionality to, for example, produce simple counts / frequencies; they can also compile cross-tabulations, for example concerning the occurrence of a coded theme (Bryman, 2012). As per the research conducted on passive mobile positioning, (Ahas et al., 2008), where the dataset used in the study was very carefully compiled by the Estonian cellular provider (EMT), this proposed research also made use of a very carefully prepared big-data dataset.

After the carefully-prepared data was cleaned, the data was analyzed. This analysis involved presenting the results in figures or tables and interpreting the results via a statistical test; the interpretation in quantitative research infers the researcher makes conclusions from the findings of the research question(s), hypotheses, and the bigger meaning of the results (Creswell, 2014). More simply, data analysis is the route where information (data) is managed, inspected and interpreted so useful conclusions can be drawn (Bryman, 2012). There are multiple quantitative methods for analyzing data, including discriminant analysis, structural equation modeling (including confirmatory factor analysis), inferential analysis and regression analysis.

The Python programming language was also used considering its ecosystem of data and scientific analytics libraries which are available, as well as Python's ease to integrate with most other technologies (Hilpisch, 2014). Furthermore, for data analysis, exploratory computing, interactive data visualization, the platform is often preferred due to improved library support and is therefore a common alternative when undertaking data manipulation tasks; furthermore, when considering Python's strength for outcomes such as general purpose programming, the platform remains a brilliant choice in term of a single language used for constructing data-centric applications (McKinney, 2012).

Python has also been used successfully in analyzing various complex data studies. Some of these include forecasting stock market index returns, forecasting of wind speed, inflation forecasting, fraud detection of credit cards,

credit scoring applications – along with the integration of machine learning (Subasi, 2020). Python has also been used extensively in not only data analysis but data visualization to, which is essential in business intelligence. Python's ability to illustrate data in graphical formats also enables its users to detect correlations and highlight questions of importance; considering the massive volumes of data being collected from the operations of say a multi-market insurer, Python metamorphizes busines intelligence into reality by identifying faster and better ways to spot trends, patterns and identify correlations in datasets which may otherwise remain undetected (Embarak, 2018).

Python programming language is also very useful when the user analyzing data commences from a simple and basic concept, and migrates towards vast quantities of concepts; with such complex data networks, the data the user is most interested in is often in an unformatted or "raw" form, which can result in obstacles to the research – this applies to numerous fields such as ecology, biology, economic, finance and medicine (Caldarelli & Chessa, 2016).

Python programming language is also very successful in designing recommender systems, purchase predictions, demand forecasting and rankings, which is conducted via exploratory data analysis, also known as EDA; such analysis is fundamental where knowledge, review and understanding of the customer is a pivotal role in a company (Sahoo, Samal, Pramanik, & Pani, 2019). The software plays an integral role in predictive analytics and modelling techniques, specifically in promotion and advertising, market basket analysis, analysis of preference and choice and economic data analysis; Python proves useful in analysis involving market segmentation, positioning of product machine learning and importantly – recommender systems (Miller, 2014).

Some previous studies investigated included Python programming language where neural activity patterns were decoded for alongside brain imaging, as well as other multi-variate analysis approaches which included a classifier-based analysis which was initially used to assess neural representations of objects and faces in the ventral temporal cortex, which indicated the

representations of various and different object categories were spatially-distributed as well as overlapping (Hanke et al., 2009).

Data can be described both descriptively and inferentially (Laerd-Statistics, 2013). Descriptive analysis was conducted via:

- measures of central tendency for the variables: this illustrates the central position of a frequency's distribution for a dataset – this was conducted by measures such as a mean, median and mode (Hays, 1981) – refer to descriptive statistics expanded on later in this research paper
- measures of spread (Nick, 2007): this summarizes a dataset by defining how spread out the various scores or values are – this was conducted by measures such as quartiles, variance measures and standard deviation measures, also summarised in the descriptive statistics noted later in the research paper.

Inferential analysis is where generalizations regarding the populations are drawn, simply because there is a limited amount of data available (for example the data from the targeted FSP and not all the FSPs within RSA). As a result, the estimation of parameters is useful (Hasselblad, 1966), as is testing of statistical hypotheses (Holický, 2013).

Regression analysis is often used in big data studies and is a technique organizations can going forward see huge value from (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011). Essentially, the regression analysis describes how the dependent variable varies as the independent variables differ. Similarly in this study, there is a need to understand the manipulation of IV_1 and IV_2 to ascertain their effects and influence over the DV. Since the target variable in this research is binary, standard regression analysis would have been inappropriate, and instead, a binary classification model is used.

Logistic regression, which is in fact a classification technique, is known to be a linear model as the outcome always depends on a linear combination of the parameters and input. Logistic regression is utilised to present a hyperplane to

separate observations which belong to a certain class versus all other observations which don't belong to that specific class; the decision boundary is therefore linear in hyperspace. This is a main drawback of linear models. XGBoost (a popular gradient boosting machine learning algorithm) was chosen with a binary logistic objective function as the classification technique. XGBoost was favoured over a traditional logistic regression as it is able to capture non-linearity in the data (Tian, Cai, Wen, Li, & Li, 2019).

3.9 Description of the respondents

The 566,758 subjects being studied are the present and previous customers of a South African FSP, more specifically an insurer. These research subjects display the following characteristics:

- demographic:
    - age eighteen years and older
    - males / female
- key attributes:
    - may have / have had a life insurance product
    - may also have / have had a health insurance product
    - may also have / have had a short-term insurance product
    - may also have / have had a credit card
    - may also have / have had an investment portfolio
    - may also have / have had a wellness product
- duration as customer: between 1 day and 6 years
- permanently-located and reside within South Africa.

The dataset is summarised below with the 566,768 customer profiles analysed – the missing cells would be due to that specific client not always having all the variables being observed, for example not every customer will have a banking product. Overall the missing cells are not statistically material.

Table 1: Dataset description

Dataset statistics	
Number of variables	41
Number of observations	566758
Missing cells	1998864
Missing cells (%)	8.6%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	90.3 MiB
Average record size in memory	167.0 B

3.10 Validity and reliability

Although the words reliability and validity almost appear synonymous, they do have rather dissimilar meanings, where reliability speaks to the stability and consistency of results, and validity speaks to the integrity of the research findings (Bryman, 2012). Furthermore, where reliability (and measurement validity (Bryman, 2012)) measures relate to how reliable and consistent are the results, validity is the degree to which the test measures what the research actually wants to measure (Cooper & Schindler, 2011).

The comparison between reliability and validity is illustrated and summarised in the tabulation below:

Table 2: Reliability and validity summary

	Reliability	**Validity**
What it tells you	This measures the extent to which results from the research can be replicated when the said research is (under the same conditions) repeated	This measures the extent in which the results actually do measure what the research is supposed to be measuring
How it assesses	It checks the result's consistency over a period of time, across differing observers, as well as across sections of the research test itself	It checks how well the research results correspond to the established theory, as well as other measures linked to the same concept
How they relate	A reliable measure does not necessarily mean it us valid - the result may be repeatable, but they may not necessarily be correct	A valid measurement often is generally reliable - also if a research test delivers accurate results, the results should be repeatable and reproducible

3.10.1 Validity

Validity is crucial in determining the quality and value of the research (Babbie & Mouton, 2001) with the aim and purples of validity being to ensure factors to be measured are actually measured.

Referring back to the afore-mentioned delimitations, research involving vast datasets is often impacted by the power issue and validity findings can be impacted due to the sheer size of the sample (Kaplan, Chambers, & Glasgow, 2014).

Ecological validity, which looks at the extent to which the conclusions of the research study can be generalized to real life scenarios, is applicable in this study. There are other examples where big data studies have ecological validity (Mehl & Gill, 2010).

It is optimum to maximise internal as well as external validity, however at times it is not always possible due to logistical and practical constraints of a research paradigm (McDermott, 2011)

3.10.1.1 External validity

External validity ascertains if a causal relationship can be confirmed from a sample or not and often determines if the study's results will hold for other settings, times, places or persons (Calder, Phillips, & Tybout, 1983). External validity therefore unpacks to how valid the sample selection is and how any sample selection should be representative of the size of the sample. The size of the dataset involving 566,758 customer profiles reinforces external validity as representative of the greater market of an FSP who offer similar financial products.

3.10.1.2 Internal validity

Internal validity determines if relationships linked to the dependent variable are influenced by solely the independent variable, and not by additional variables (Slack & Draugalis, 2001). By means of an example, if various and different researchers were to investigate a specific and single research point and thereafter also identify the same conclusion, that initial research would be termed internally valid.

This research study noted the decision tree as detailed in the Appendix G where a dependent metric variable (return on marketing investment) being analysed requires multiple regression in terms of the correct statistical technique (Black, Babin, & Anderson, 2010). However in this case, since the target variable is binary and used to predict the propensity to convert a quote, a non-linear implementation of a logistic classification model was used (see Appendix H).

3 strategies were employed to ensure validity of the developed model:

1. **An upfront test-train and validation split** of the original dataset (see Appendix I)

- If this is not done, one can overfit (model fits to the noise and any artifacts in the training set which essentially means it memorises the training data). The effect of this means the model will not be able to generalise to unseen data.
- Without this performance on new data, the performance of the model would be degraded
- Important to note that the test train validation split was stratified by target variable to ensure a consistent distribution across the 3 splits, which are noted in the graph below.
- In the graphical illustration below, the test curve would have dropped sharply if there was over-fitting. Early-stopping (for over-fitting) criteria was applied to avoid this and maintain model validity.

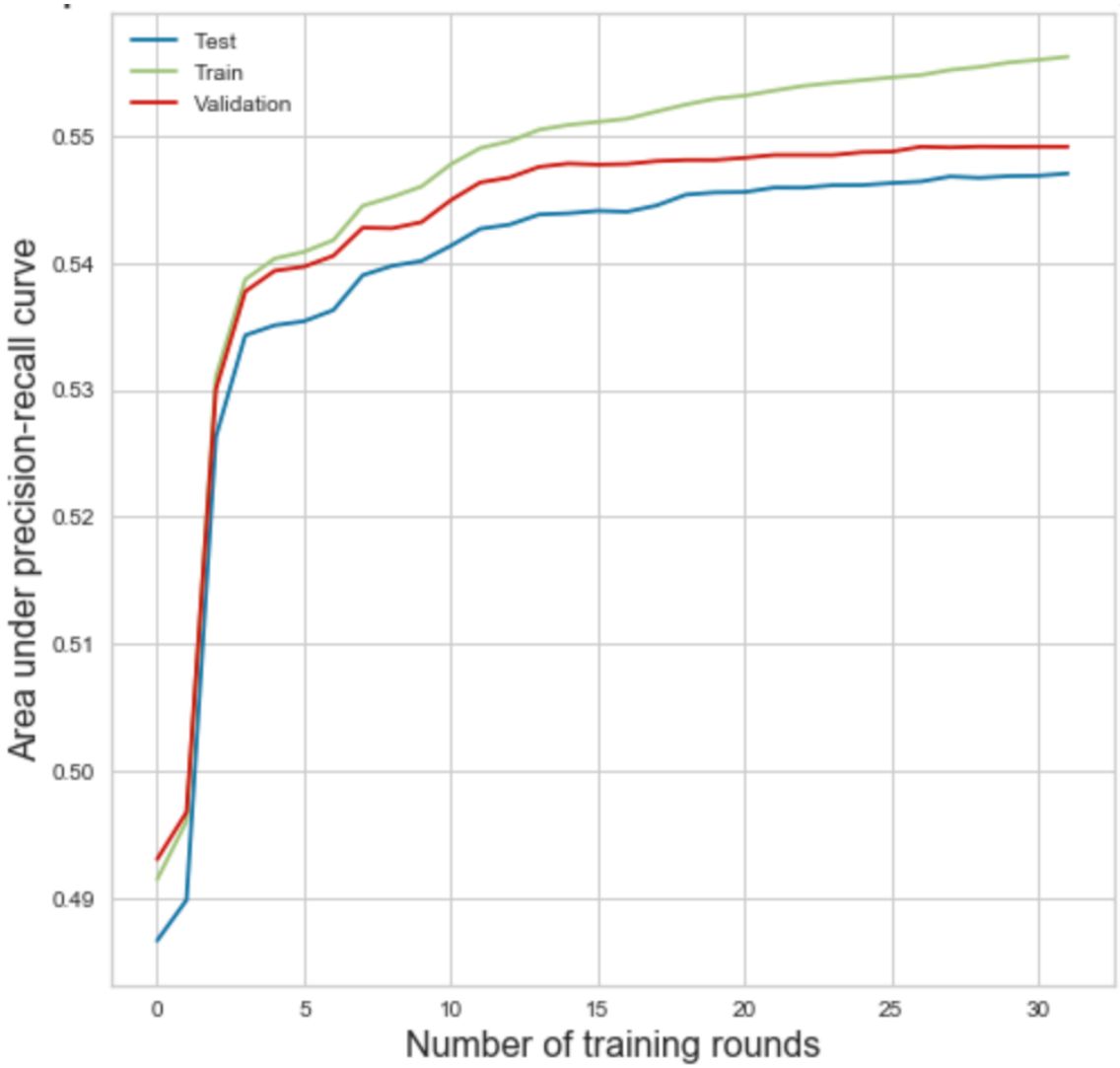

Figure 2: Model performance on test and train datasets vs. training rounds

2. **Bayesian hyperparameter tuning** (Martinez-Cantin, 2014) on the validation set (T. Chen & Guestrin, 2016), refer to Appendix J.
    - This is a technique to ensure overfitting doesn't happen
    - The tabulation in Appendix K illustrates the iterations of the Bayesian optimisation algorithm, where the final line indicates the optimum hyperparameter set

3. The third strategy was the post-hoc **probability calibration check** which effectively tests whether the output of the model as a probability reflects reality – reality in such a case being probability of future incidence (refer to the Appendix L) of which the graphical results are published below

(the agreement of the green calibration line with the 45-degree line indicates that the model probabilities are effectively calibrated). This probability calibration curve illustrates if the model output can be interpreted as probabilities. The result indicates that a segment of short-term insurance product quotes with an average prediction value of 20% would convert 20% of the time. This is why the graph is reflected as a 45 degree line.

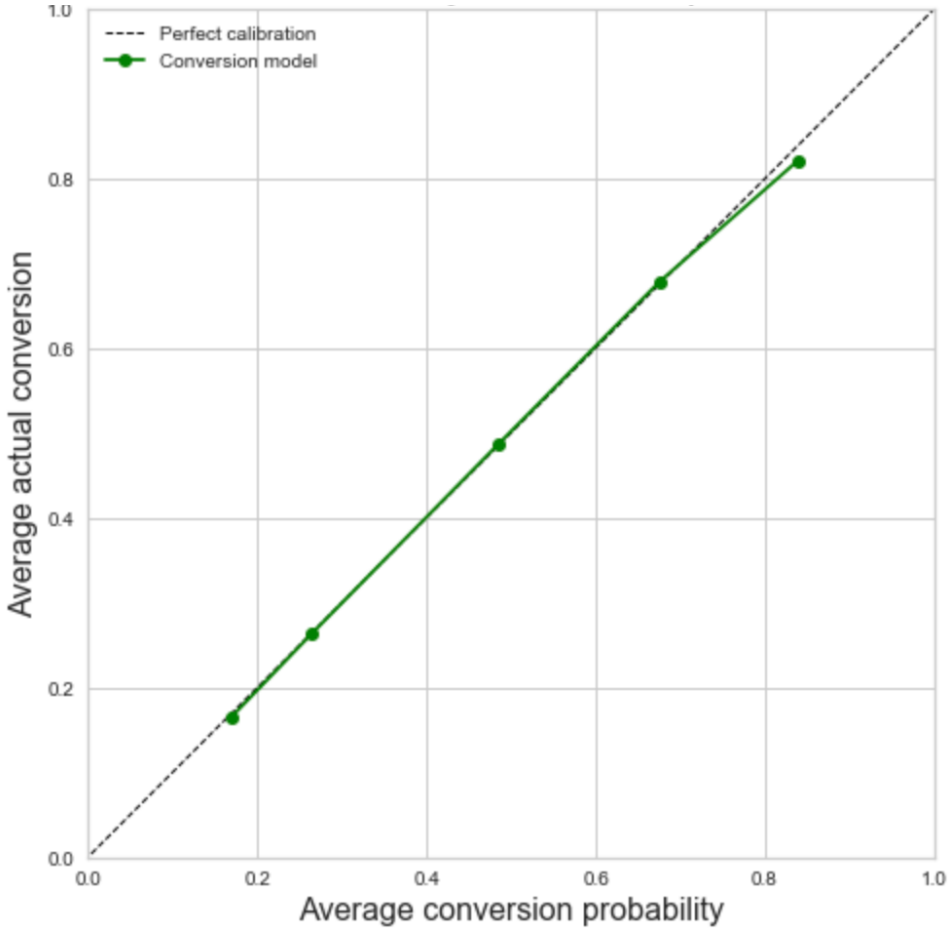

Figure 3: Probability calibration plot

3.10.2 Reliability

Research is seen as reliable when it could be utilized by various researchers and still render consistent results, which don't vary (under stable conditions); in addition, reliability can be interpreted as the extent to which the test is free from errors in measurement as naturally the more errors in the measurement then the less reliable the results of that test, and vice versa (Fraenkel, Wallen, &

Hyun, 1993). Reliability can therefore have a definition as the scale which consistently ascertains and measures what this scale set out to measure (Connaway & Powell, 2010).

Interestingly, Messick in 1989 metamorphized established definitions of validity and reliability and acknowledged a combined concept of validity that included reliability as a type of validity and therefore reliability contributing to overall construct validity (Messick, 1989).

As reliability measures how accurate one's research instrument is, for example: if the researcher was conducting a survey would that researcher receive the same answers this week as next week? This is not relevant to this study as the hard metrics were used from an dataset and the research was not conducted on attitudinal measures demonstrated via say a survey or questionnaire.

As part of assessing the reliability of model performance, the below graphical illustration notes confidence intervals around metrics of interest, namely precision and recall, to assess their robustness given random subsets of the underlying data. For thresholds greater than 0.8, it is clearly seen that the confidence interval for precision includes 0; for this reason the researcher does not recommend a decision (discrimination) threshold greater than 0.8 as the researcher would risk severely-degraded precision.

The below reinforces that precision is sensitive to random changes in underlying datasets, reinforming reliability of the model. Since the research opted to use threshold of 0.3, the research ensures robustness of the precision and recall metric of the model, (refer to Appendix M).

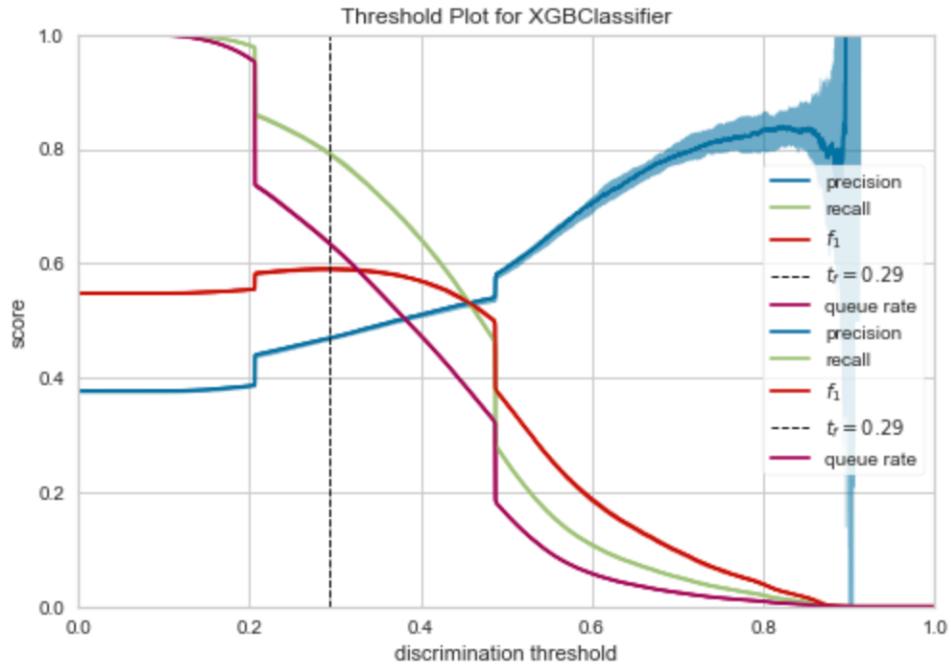

Figure 4: Discrimination threshold sensitivity plot

4 PRESENTATION OF RESULTS

4.1 Introduction

The propensity model tested the success of quoting a customer profile based on 41 different variables, where these variable were attributed to the customer profile outside the short-term insurer's environment.

These variable tests include:

- Quote numbers
- Quote identifier (2 channels)
- Quote activated indicator
- Policy activated indicator
- Health insurance client indicator
- Health insurance product type indicator (5 different health insurance products)
- Health insurance duration indicator
- Hospital gap cover indicator
- Wellness product indicator
- Wellness product duration indicator
- Wellness product status level indicator (5 engagement statuses)
- Life insurance client indictor
- Credit card indicator
- Credit card health indicator
- Credit card duration indicator
- Credit card type indicator (5 credit card types)
- Investment product indicator
- Age of client indicator
- Duration of client indicator
- Social economic status indicator

- Family size indicator
- Number of products added in last 6 months indicator
- Quote platform indicator

The next session highlights the findings and describes the key statistics and correlations identified.

4.2 Descriptive statistics and correlations

The various analyses conducted below indicate the statistical findings for each variable conducted, which were the factors from which the main propensity model was constructed.

There are also appropriate findings noted under each heading, specifically indicating the number of clients who had these below-noted characteristics, and the relationship with these said characteristics on the performance of short-term insurance product conversion when these customers are quoted.

The line in each of the graph indicates the conversion rates for the descriptive statistics based on the raw, non-processed features that are used in the propensity model.

Business channel description

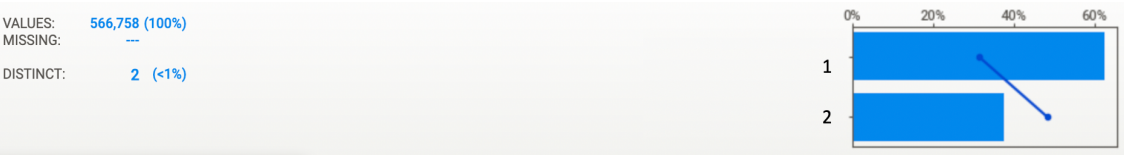

Figure 5: Business channel from where quotes are received

This looks at the conversion rates between 2 differing quote platforms within the insurer, channel 1 and channel 2:

- There were less quotes via channel 2.
- Channel 2 showcased a higher conversation performance at 48.3%.

Health insurance product indicator

Figure 6: Health insurance product indicator

This investigate the difference of potential customers who enjoyed health insurance and their associated conversion rates whether they do or do not have a health insurance product:

- More customers had a health insurance product than not.
- Whilst more customers had a health product, there was a marginally higher conversation performance (38.3%) versus than customers with no health product (conversion of 37.0%).

Health plan type indicator

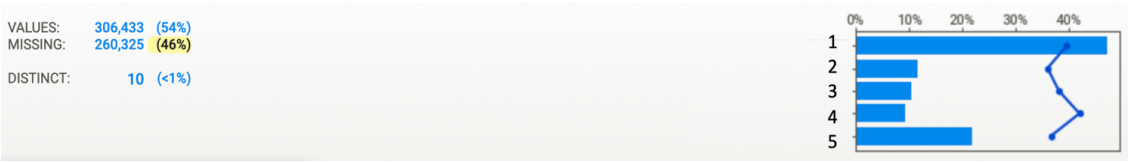

Figure 7: Health plan type indicator

- The insurer offers 5 different health insurance products, with majority of the clients on the product 1 category.
- Clients with health product 4 illustrated the highest conversion rates (42.0%).

Health product duration indicator

Figure 8: Health plan type indicator

- This calculation indicated that of that average client with a health insurance product had enjoyed that product for under 60 months.

- Clients enjoying the health insurance product for 150 months had the highest conversion performance (just over 40%).

Hospital gap cover indicator

Figure 9: Hospital gap cover indicator

- Very few customers had a hospital gap cover product (1.4%)
- Customers that did have the hospital gap cover product show better conversion performance at 41.6%.

Wellness product indicator

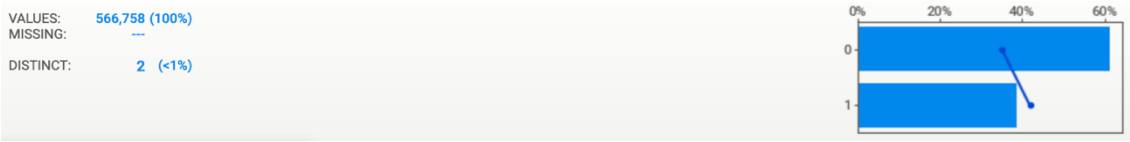

Figure 10: Wellness product indicator

- 38.7% of customers had a wellness product.
- Even though lesser customers enjoyed a wellness product, there was better conversion rate on this segment of customer (41.8%).

Wellness product engagement indicator

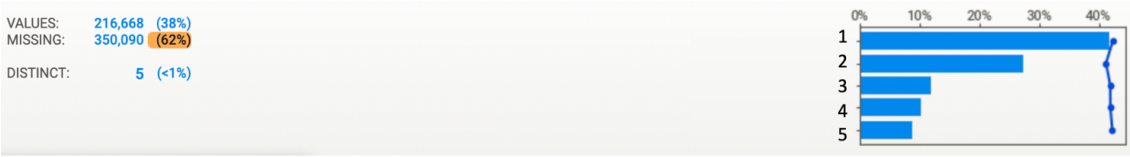

Figure 11: Wellness product engagement indicator

- Majority of customers with a wellness product were on level 1.
- Slight decrease in conversion performance for wellness product level 2, however very little difference in conversion rates between 40 and 43% across all levels.

Wellness product duration

Figure 12: Wellness product duration

- Majority of clients with a wellness product had enjoyed the product for 60 or less months.
- Better conversion performances were experienced in customers who had enjoyed a wellness product for less than 100 months.

Life product indicator

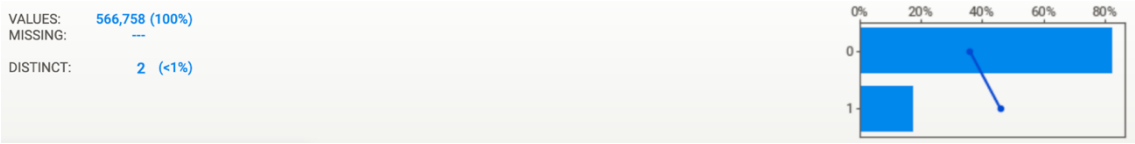

Figure 13: Life product indicator

- More customers don't have a life insurance product than do have – 17.5% have a life insurance product.
- Customers with a life insurance product also see an optimised conversion rate of 46.0% versus 35.9% for non-life insurance product customers.

Credit card product indicator

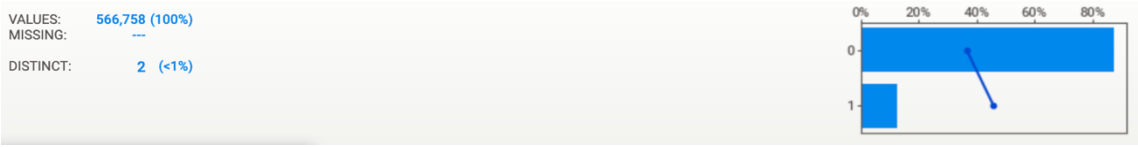

Figure 14: Credit card product indicator

- 13% of customers had a credit card product
- Of the customers that did have a credit card product, a superior conversion rate was seen versus customers who did not have a credit card product.

Credit card product type indicator

Figure 15: Credit card product type indicator

- Credit card type 3 contributed only 15.9% of the credit card customer population, versus credit card types 1 and 2 both owning 40.2% and 40.0% respectively
- Credit card type 3, however, witnessed a far superior conversion performance at 52.2%.

Credit card product health indicator

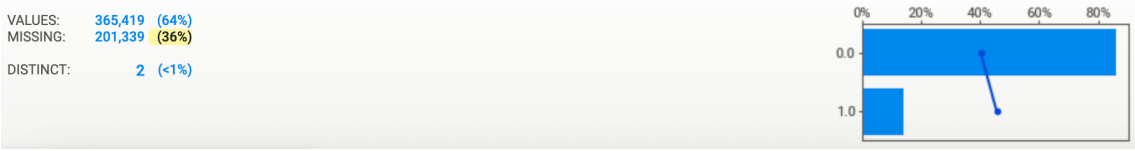

Figure 16: Credit card product health indicator

- 13.9% of the population sample managed their credit card product optimally in term of both exposure and debt repayments.
- This same segment of the credit card customer base also provided a superior conversion performance result of 45.6%.

Credit card duration indicator

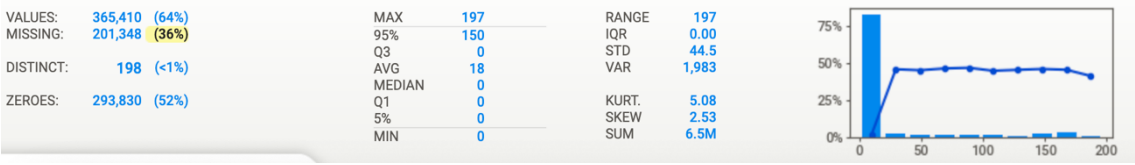

Figure 17: Credit card duration indicator

- This measure indicated that for credit card customers who had a credit card product for more than 25 months, the conversion performance was consistently between 40% and 50%.

Investment product indicator

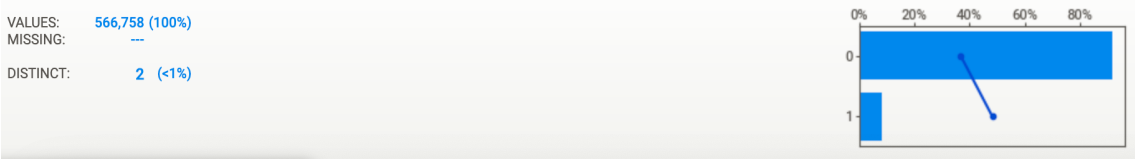

Figure 18: Investment product indicator

- The analysis reflected that 8.3% of the sample had an investment product.
- The customer base who possessed an investment product illustrated a 48.2% conversion result, better than the segment with no investment product.

Age of customer

Figure 19: Investment product indicator

- The sample population returned the largest age segment being in the 30-40 years old category (just over 30%).
- The age bracket which enjoyed the best conversion result however, was the 40-50 years old category with a conversion performance slightly above 50%).

Social economic status (SES) indicator

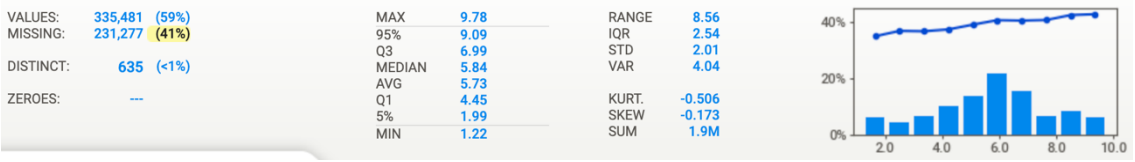

Figure 20: SES Indicator

- The inherent targeted customer base of the insurer is towards a higher LSM population.

- The conversion performance shows a directly proportion relationship to economic status, with higher SES scores returning improved conversion results, directly related to more disposable income.
- An SES score of 6 or higher enjoyed conversion performances exceeding 40%.

Family size indicator

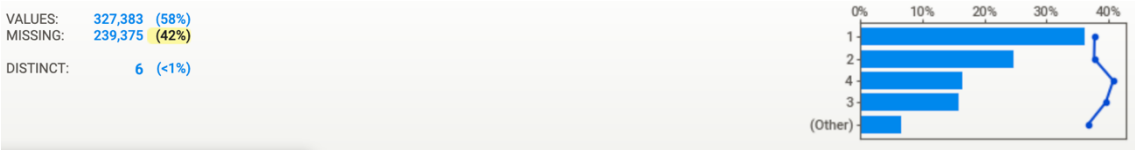

Figure 21: Family size indicator

- The analysis indicated that the majority of the sample set was a single-person household (36.2%), but there would also be scenarios where the customer was cohabiting or married but was only noting themselves within the household.
- A slightly dampened conversion result was seen when the household grew from 1 to 2, potentially as a new couple was in a growth stage financially.
- A household of 3 returned a better conversion performance than households of 1 and 2, potentially as the household improved their financial positions
- A far greater conversion result at 40.7% was returned for households consisting of 4 individuals – due to a potential level of financial stability as the household matured.
- Households of 5 and more saw conversion rates decrease to worse than households consisting of 1 individual.

Number of products added in last 6 months indicator

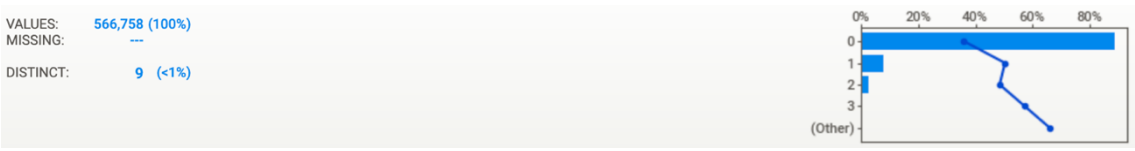

Figure 22: Number of products added in last 6 months indicator

- This measure indicated that the majority of customers (88.7%) had added no additional products in the last 6 months.
- Except for the addition of 2 additional products in last 6 months with a conversion performance of 48.5%, the general trend was the more products added in the last 6 months, the greater the propensity of the customer accepting a short-term insurance product, with:
    - 1 product added in last 6 months showing a conversion of 50.4%.
    - 3 additional products added in last 6 months reflecting a conversion of 57.3%.
    - 4 or more additional products added in last 6 months returning a conversion performance of 66.1%.

4.3 Model architecture

XGBoost (Extreme Gradient Boosting Package, also known as an ensemble technique) framework was used considering it is a scalable and efficient implementation of the gradient boosting framework (Friedman, Hastie, & Tibshirani, 2000) and also because it not only includes a tree learning algorithm and linear model solver, but also ranking, classification and regression functions (T. Chen et al., 2015). *XGboost* has extra regularisation, which is a built-in to mitigate overfitting and thus improves performance for out of sample (test dataset) data.

The underlying architecture relied on decision trees so that non-linear interactions could be modelled – imperative for this study. Furthermore, decision trees also possess the following advantages:

- They are simple to understand and interpret, even non-experts due to their graphical display (James, Witten, Hastie, & Tibshirani, 2013)
- The decision tree showcases a hierarchy of attributes which reflect the associated importance of these attributes; therefore features positioning on the top are thus the most informative (Piryonesi & El-Diraby, 2020)

- They often require less data preparation as other techniques mostly require data to be normalized – trees can accommodate qualitative predictors and therefore there isn't a need to generate dummy variables (James et al., 2013).

4.4 Model evaluation

The performance metric, namely area under the precision recall curve (Buckland & Gey, 1994), was used to evaluate performance of the model being sensitive to the imbalanced nature of the dataset (more quotes that didn't convert versus quotes that did convert).

- The performance was evaluated on all splits of the data (test train validation sets), however final performance was exclusively evaluated on the test set.
- The overall model indicated that customer profiles were identified as preferable where they were reflecting above 0.3 probability, thus indicating a higher chance of conversion success when being quoted. This threshold was selected due to it being the F1 optimal score which maximises the harmonic mean between precision and recall.
- The model uses personalised variables as detailed above to identify segments of customer profile who when offered marketing communication or business development engagements, show a far higher conversion rate due to such personalised variables being targeted: this is effectively the managerial problem being solved.
- The final model performs at (0.5316-0.3768)/0.3768 where:
    - 0.5316 is the area under the precision recall curve on the test set used to balance the trade-off between precision and recall to maximise the overall model's performance
    - 0.3768 is the baseline curve in a PR curve plot (the overall quote conversion experience in the train dataset) and is a horizontal line with height equal to the number of positive examples over the total

number of training data i.e. the proportion of positive examples in the train dataset.

This equates to 41.1% better performance than the baseline model, indicating that the user can score all quotes with a propensity to convert and can therefore rank the priority of quotes to process the highest propensity quotes first, specifically before too much time has elapsed and the quote and data goes “cold”.

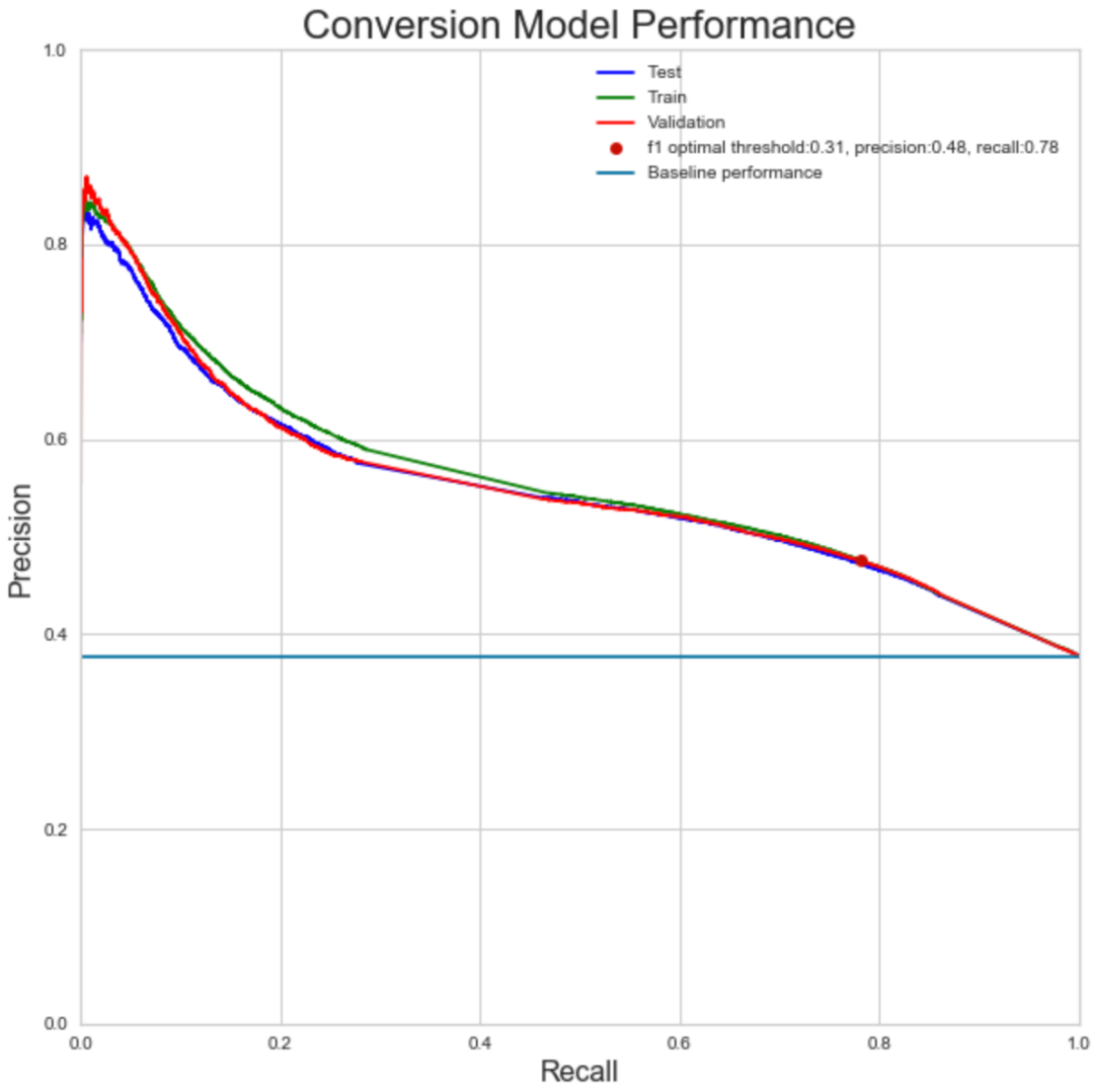

Figure 23: Conversion model performance

Furthermore, a confusion matrix below (Figure 24) was constructed (Luque, Carrasco, Martín, & de las Heras, 2019), which effectively is a table utilized to describe a classification model's performance on a test dataset where the true values are known. The true label collectively is 42,714 (quotes), and the model correctly predicted 33,134 quotes which converted which equates to a recall of 77.6% (see red dot above in Figure 23).

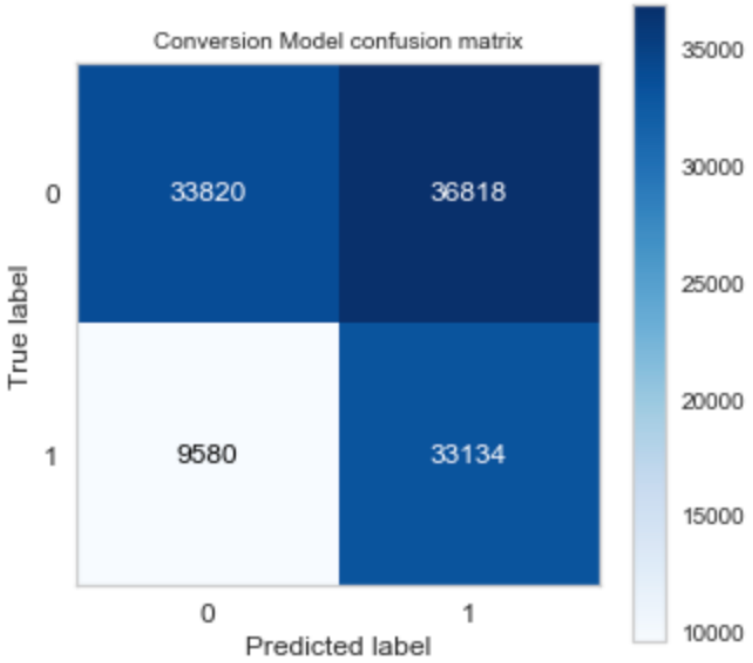

Figure 24: Conversion model confusion matrix

The above confirms a satisfactory model was employed for the research in question.

5 DISCUSSION OF RESULTS

5.1 Introduction

Chapter 4 prior noted the presentation, description and the analysis of the statistical data results collected. The main focus and resultant structure of chapter 5 discusses and explains the results of the research in relation to literature reviewed as well as the specific focus and reference to the afore-mentioned hypotheses.

This research paper intended to contribute to knowledge in the financial services environment, where the objective of the research was aimed at showcasing the impact of knowledge of customer profiles derived from big data, and how such knowledge can increase sales performances at a reduced cost per sale and thus a higher return on marketing investment (due to more sales from the same number of quotes). Against the hypothesis outlined in this research, the intended outcome was confirmed. The results of this research are illustrated in a classification model and the conversion propensity of the customised customer segment is illustrated in Figure 27, with the hypotheses being resultantly proven and summarised in Figure 28.

5.2 Description of results

The target market of a business is at the heart and very centre of distribution channels, supply chain, procurement and the marketing mix selection (Aghdaie & Alimardani, 2015). Target marketing involves taking various steps which include assessing the attractiveness of each various market segment before selecting one (Armstrong, Adam, Denize, & Kotler, 2014). In consequence, the selection of the target market is an output of the market segment evaluation as well as the selection decision; market segment evaluation should therefore be

utilised as the natural approach by organisations for target market selection (Aghdaie & Alimardani, 2015).

Using big data to refine market segmentation and target market selection is also important for sectors where there are various financial services providers who have the same or similar offerings and value propositions. Such big data insights enable the company (such as a financial services provider) to understand what matters most in increasing customer satisfaction and predict customer behaviour (Spiess, T'Joens, Dragnea, Spencer, & Philippart, 2014). In addition, an organisation will not be able to always satisfy all of their customers or potential customers, all the time – it is difficult to satisfy the exact and specific requirements for each individual and unique customer (Camilleri, 2018), which is the reason why many companies and organisations undertake the target marketing strategy.

The research study identifies a segment of customers within this specific financial services provider existing customer base that reflects the optimum segment from which the insurer can prioritise business development and quote generation for a cross-sell campaign. Customer profiles who demonstrate the following attributes have the highest propensity to accept a short-term insurance quote and therefore can be prioritised as the target market for a specific short-term insurance product campaign where they render the highest return on marketing investment (ROMI).

This initial preferred segment made visible through the analysis of big data is where the customer:

- is between the ages of 40 to 50 years old
- demonstrates a social economic status score of between 5,6 and 7 with 6 indicating the highest conversion propensity
- lives in a household of 4 and 5 members, with a 4 indicating the best score and a score of 5 marginally lower, but still demonstrating high conversion performance

- received the quote via channel 2
- possesses a health insurance product, which was further optimised when the customer:
    - possesses health insurance product 4
    - has enjoyed the health insurance product between 80 and 160 months
- enjoys a hospital gap cover product
- possesses wellness product, where optimised conversion is experienced where the customer
    - is on level 1,3,4 and 5 of the wellness programme
    - has enjoyed the wellness product for less than 100 months
- possesses a life insurance product
- has a credit card product; conversion rates are increased where the customer enjoys:
    - credit card product 3
    - a healthy credit card status
    - the credit card product for longer than 25 months
- enjoys an investment product
- has accepted 1 or more additional products from the financial services provider; the optimum segment is the group of customers who have accepted 4 or more products from the insurer.

The above segment of customer reflects zero quotes based on too many criteria being simultaneously active.

Reducing some of the above features from the targeted segment increases the number of prospective customers to be targeted, as an example:

1. a 68% conversion propensity with customer segment who demonstrate below features (only 28 potential customers):
    - Investment product
    - wellness product
    - 4 or more financial products added in last 6 months
    - wellness product duration under 100 months

    - health insurance
    - credit card product with best performing credit card types
2. a 58% conversion propensity with a customer segment who portray the following features (333 prospective customers):
    - Investment product
    - wellness product
    - 4 or more financial products added in last 6 months
    - wellness product duration under 100 months
    - health insurance

This presents the challenge to balance increasing conversion propensity with decreasing number of quotes and more importantly vice-versa: balancing reduced conversion propensity as the number of quotes and targeted customer segment is increased.

The above results can still be beneficial to a large multi-product financial services provider. Some FSPs who offer a broad suite of products including life insurance, health insurance, banking and short-term insurance enjoy extremely large customer bases, such as Ping-An, who service over 200 million retail clients (Ping-An, 2021). A similar proportion, where 333 quotes out of a sample population of 566,758 customers applied to a customer base of 200 million clients renders 117,510 quotes, which at a 58% conversion propensity indicates propensity to grow the specific product customer base by over 68,000 new customers in such an example. Such a campaign would be beneficial to such a large FSP who can focus sales and business development resources into such a higher-return segment.

For a smaller FSP, additional segmentation is required to render a larger target segment whilst maintaining above-average conversion propensity. Accordingly, sub-cohorts need to be defined based on a subset of “ideal client” features; to enable this, model explainability tools such as feature-importance plots (which include SHAP) and partial dependences plots were used. To clarify, the feature-importance plots inform the research which features are most important,

whereas the partial-dependence plots inform the research at what values the features will render higher propensities for the quote to convert.

Enhanced model interpretation and explanation

SHAP-feature importance plots (Lubo-Robles et al., 2020) and partial dependence plots were analysed. In the graph below, the 16 most influential features are plotted in order of importance (see Appendix N):

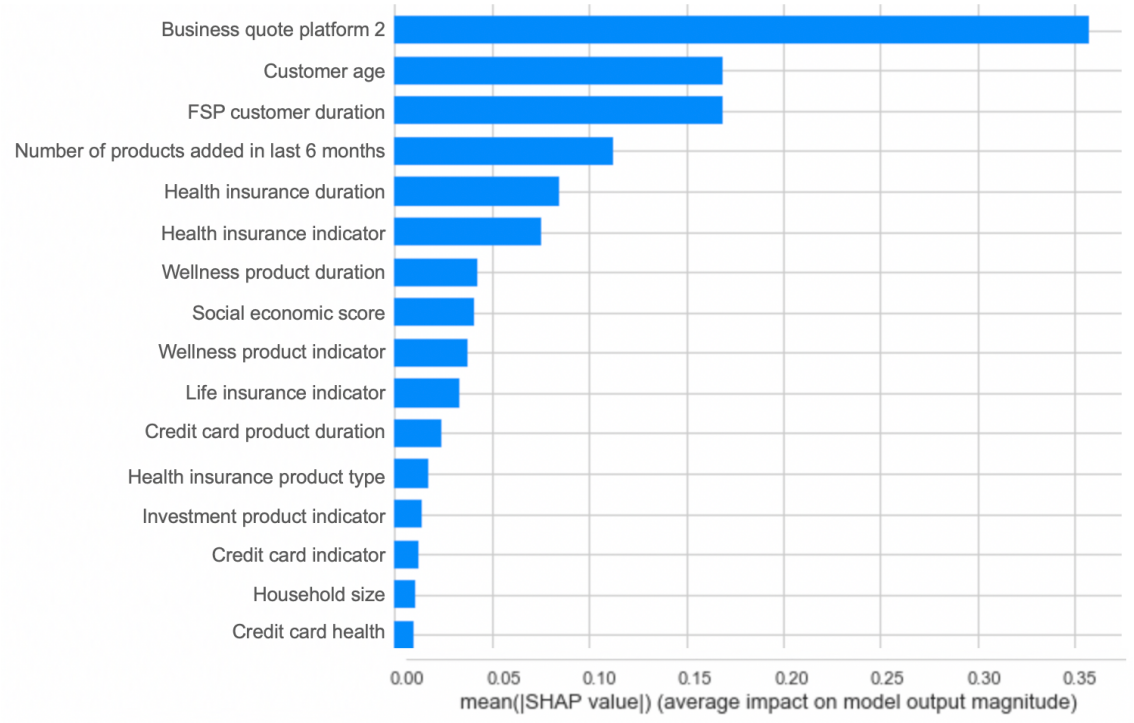

Figure 25: SHAP Analysis

Feature importance plots predict which features are most influential in the model, however this does not provide enough details to construct an ideal customer with. Therefore a partial dependence plot was used to explain the magnitude and direction of the effect of a feature on propensity to convert when quoted.

Each of the features (as per figure 23) were analysed in order of their influence. The optimum range was identified via the SHAP partial-dependence plots. As this is proprietary information and not all features can be published, one example being the age of the customer was analysed and illustrated, where the

desired range is clearly between the ages of 40 and 60 years old which is the interval where the graph peaks indicating the highest predicted propensity to convert. A special note is the grey section in the graph which indicates the exposure across the feature. Results where lower exposure exists should be interpreted with caution; graphical illustration below (see Appendix O):

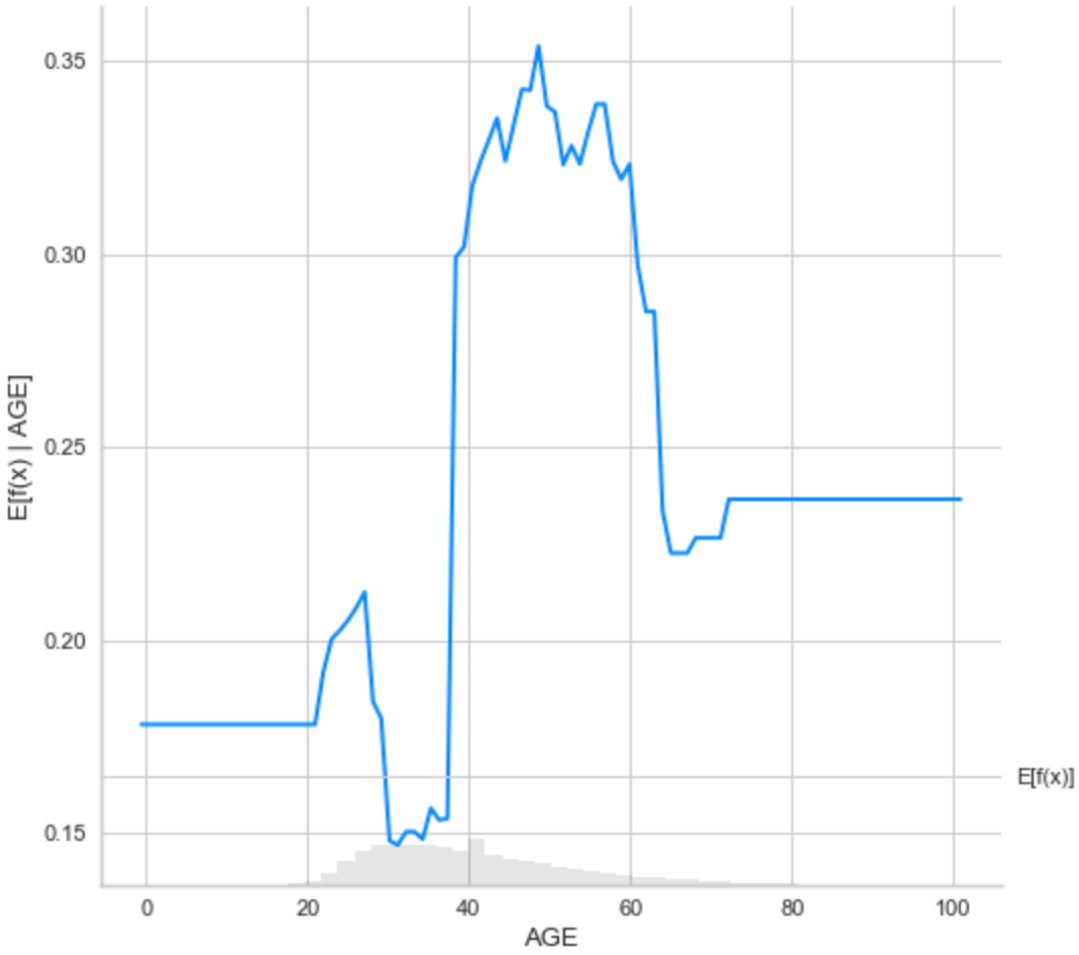

Figure 26: SHAP partial-dependence plot on age of customer

The cumulative result of these SHAP partial dependence plots were used to construct various customer segments to target, one of which is published below:

1. 2,889 customer profiles demonstrating a conversion propensity of 57.98% versus baseline standard conversion of 37.68% (see Appendix P):
    a. Age of customer between 40 and 60 years
    b. More than 1 product added in last 6 months

c. Wellness product duration under 50 months
d. A client of the FSP less than 60 months
e. Quote conducted via quote channel 2

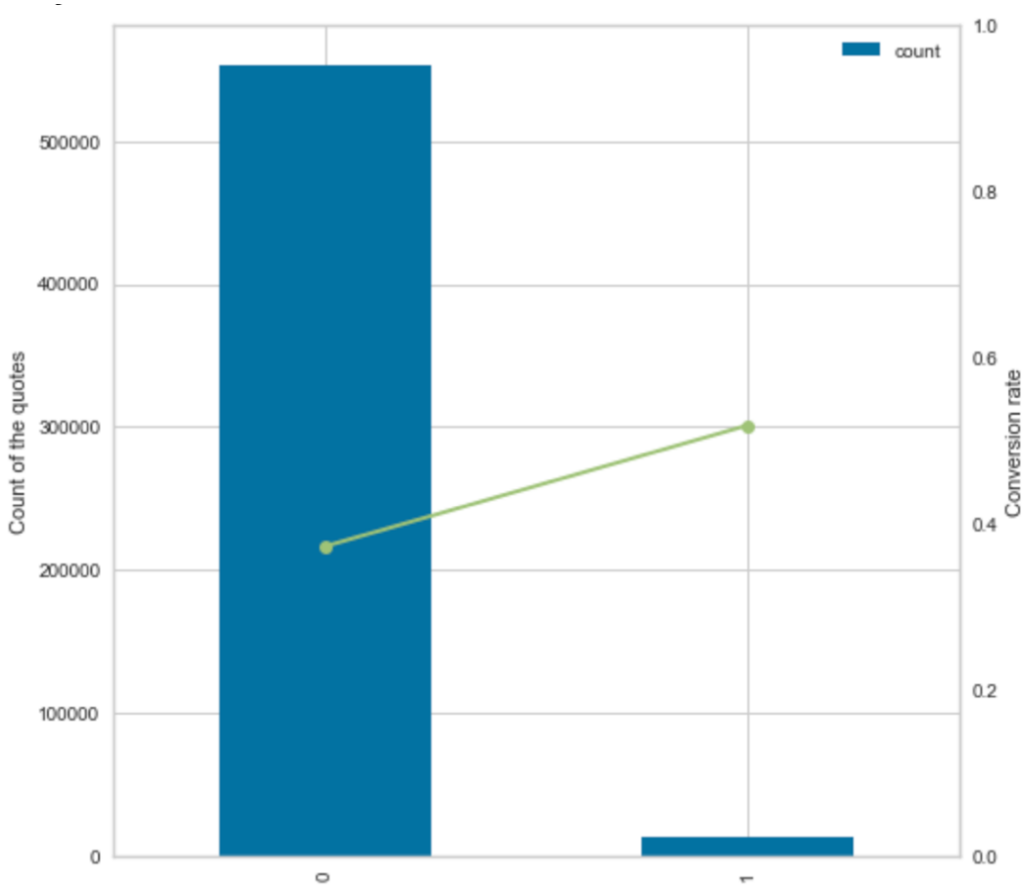

Figure 27: Analysis of conversion rates

Application of this customised customer segment

This customised segment for short-term insurance sales made visible through the analysis of big data when run through the classification model performs at a conversion rate of 57.98%, which is significantly higher versus overall quote conversion (37.68%).

Reverting back to the two hypotheses:

- Hypothesis 1 (H_1): short-term insurance sales per 100 quotes resulting from non-customized business development attains lower ROMI – this is proven and correct.
- Hypothesis 2 (H_2): short-term insurance sales per 100 quotes resulting from big data-driven customized business development offers increased revenue and associated higher ROMI – this too is proven and correct.

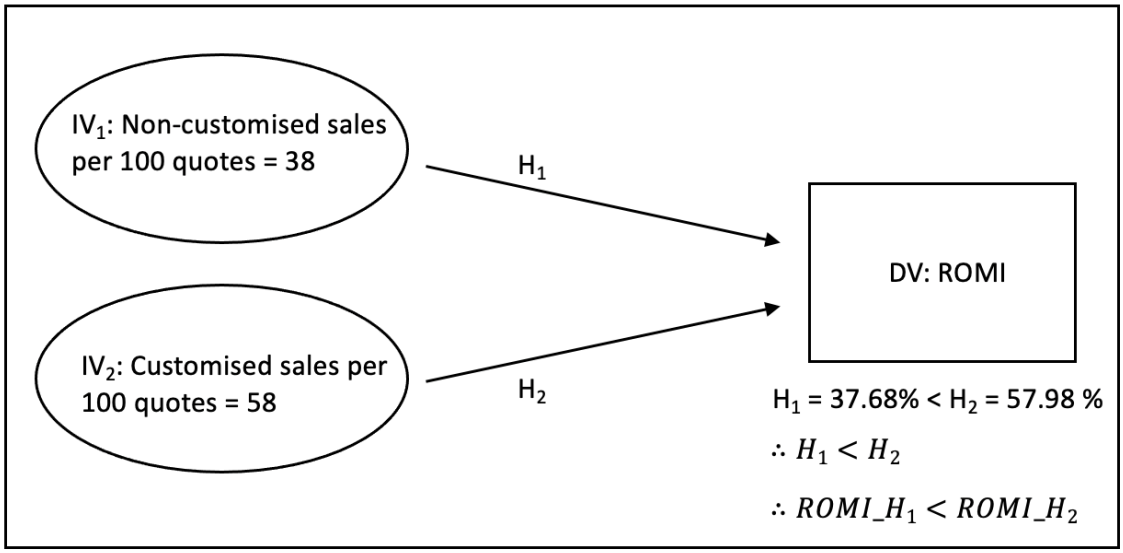

Figure 28: Hypothesis proven

Companies, CEOs and boards are more demanding on actual empirical evidence which indicate proposed marketing and business development expenditures will actually yield greater profitability (Cook & Talluri, 2004). Using this identified segment results in a greater return on marketing investment as resources and infrastructure are deployed to generate short-term insurance quotes which witness a superior conversion performance. Furthermore, the digital revolution we currently find ourselves within enables organisations who are able to analyse customer behaviour with a competitive advantage (Ertemel, 2015); the massive volumes of business data many financial services providers possess (such as insurance companies) truly enables marketers to use such insights into unique customer insights, accelerating cross-sell opportunities within high-conversion market segments and rendering undoubtable competitive advantage and a definite return on marketing investment. This is especially important as organisations strive to improve the effectiveness and efficiency of

their salesforce, especially those with a strong dependences of new business (Jones, Brown, Zoltners, & Weitz, 2005). In addition, marketers and salesperson who can learn how to integrate such technology into their teams should see a positive impact as the salesforce is able to establish longer-term and profitable relationships (Ahearne, M., Jelinek, R., & Rapp, A, 2005). The Insurance sector's decreasing investment returns from the 1990s has resulted in a more competitive market with increased focus on correct pricing and premium income; this has encouraged the search for more granular big-data analytics-based models to optimise pricing and profit through more granular and detailed risk segmentation, improved risk selection, and pricing (EIOPA, 2019).

5.3 Conclusion

In relation to this research paper's research objective to identify if big data can be used to expose customer features which have a material impact on a client's short-term insurance purchasing propensity – the above findings indeed prove that big data can be used to achieve such a result. Secondly, through use the use of these identified customer features to segment a customized quote target market, the higher conversion propensity at the same number of quotes answers the second research objective of this research being used to increase return on marketing investment.

This research study's findings contribute to knowledge for marketers and business development executives in that it showcases the relationships between seemingly disconnected customer features in relation to their insurance sales propensity and awakens cross-sell sales campaign potential across different financial product markets. Finally, this research has presented evidence which supports the hypotheses being tested.

6 CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

Chapter 5 concentrated on the details of the results from the statistical data collected. This chapter focuses and details the conclusion of this research paper and thereafter will delve into managerial implications and associated recommendations for marketing, sales and business development skillsets within financial services providers, particularly from short-term insurance providers.

As this chapter concludes, a discussion on topics of potential future research is noted, which can leapfrog off the chassis this study provides. A list of limitations of this research study is also noted to give context into possible constraints which could potentially impact the results.

6.2 Conclusion of the study

This research study endeavoured to provide meaningful value and knowledge to marketers in an insurer, particularly where the insurer offered products in multiple markets and thus possessed data on their customers which enables such propensity modelling and where the resultant higher returns on marketing investment can be realised. Based on the hypotheses which were outlined, the sought-after outcome (being a higher conversion rate) was realised.

From the data which was collected and the findings which were analysed, it is clear that there are definite customer attributes which make them more likely to accept a financial services quote when offered. The research was able to present seemingly unrelated customer factors and characteristics derived from completely different financial services products and utilise such factors to model quote conversion behaviour for short-term insurance.

6.3 Recommendations and managerial implications

The following section is intending on highlighting certain potential and possible adoption strategies for the financial services sector, particularly in the South African short-term insurance market, which are based on previously highlighted proposed outcomes what were noted against a base of results of the set hypothesis.

Simplistically summarised, insurance is good for society. It hedges and mitigates risk, bridges financial losses and therefore enables sustainability - specifically when the customer is faced with a sudden and unforeseen event which results in such a financial loss; beside the insurance market's positive impact due to its risk-sharing mechanisms, the insurance sector has a positive and direct impact on an economy's development as insurers search for new, additional markets, for optimised strategies and business models, with an objective to secure and grow customer relationships and elevate operational infrastructure: all these factors result in a positive knockon effect on the economy (Liedtke, 2007).

For many insurers, their data is well understood for underwriting, risk-selection, and loss ratio management. These insights are mostly utilised in premium calculation which is imperative for the insurer to charge the correct premium against a targeted loss ratio. The value of the premium is most important for the insurer to attract preferred customers (by offering a competitive premium to attract the desired customers). The offered premium is also important to push away undesirable risks, where the premiums charged and the terms and conditions are onerous to a level of either the unpreferred customer can be profitably managed at such a premium value, or the unpreferred customers simply walks away from the deal – both of which are preferred outcomes for the undesirable segment of customers.

This research paper indicates and highlights the possibilities for a financial services provider such a short-term insurance company to harness their databases and expand their data capabilities to include more evolved and advanced marketing perspectives, and therefore to metamorphosize their raw data to information, then to knowledge, and lastly to wisdom (Bellinger, Castro, & Mills, 2004).

If the financial services provider, in particular a short-term insurance company, possesses an understanding of factors that influence acceptance of a short-term insurance policy as well as understanding of consumer behaviour influences, the said financial services provider can enable their financial advisers, independent brokers, and marketers to capitalize on this data-led competitive advantage, and demonstrate a market-leading position in terms of channelling sales resources and architecture to an optimised sales efficiency level. This not only maximises intermediary satisfaction (due to higher quote conversion and resultant sales performances), but also has the benefit of increasing rate of revenue growth and reducing lapse rates due to a more engaged customer. Clients who enjoys additional products with the said brand, and thus experiences additional “lock-in”, therefore improve embedded value and return on shareholder equity for the said insurer.

A recommendation which is suggested for financial services providers who also offer short-term insurance is for such companies is to harness such insights and offer, or at least market the short-term insurance product when customers within the targeted threshold (as detailed in the “Model evaluation” section of this research paper) contact the company. This has the benefit of marketing and offering the insurance product to the customer at a time they are able to communicate with the company. Although the customer may need to be transferred to a licenced or accredited division of the company that can render the appropriate advice (Millard, 2019), this allows an immediate stage to engage targeted customers who statistically demonstrate high-propensity to covert convert when quoted.

Another recommendation for the financial services provider is to engage on a personalised and specific correspondence campaign. The most cost-effective route would be an email campaign that is published to the customer segment as detailed in the “Model evaluation” section of this research paper. Such emails could test the customer’s appetite for a quote against their short-term insurance portfolio with an attractive description of benefits and differentiators the financial services providers offers. Positive responses would result in a call back to the customer to run through the quote underwriting process and close the sale. For customers in this selected and preferred segment that either did not respond, or declined to quote, the financial services provider could thereafter offer guaranteed discount based on their holistic understanding of the customer’s overall portfolio, where it may make financial sense to offer a lower premium based on the additional customer cashflow to the insurer and therefore an increase in overall customer equity (Hansotia, 2004).

A third recommendation which may cost more and potentially take longer is a dedicated call-based campaign where targeted customers from the desired segment (as illustrated in the “Model evaluation” section) are called by financial adviser accredited to render the advice for such an insurance product sale (FAIS, 2002). This has the benefit of using skilled financial advisers who can massage the quote engagement to defend against potential customer push-back, and maximise the sale opportunity with a personalised interaction guided by the data associated to this ring-fenced preferred customer segment. This enables a nimble and flexible quote experience which isn’t always possible with email campaign and uses the data to provide tailored insurance risk solutions in a manner which is most valuable and appropriate to the customer (Ricci et al., 2011).

The recent developments in the technological landscape enable marketers and business development executives to render some of the above recommendations with close to immediate implementation. A further recommendation for such a call-based campaign would be to utilize call-centre dialers (Alberto & da Silva, 2014). A dailer is an application that provides automation for outbound telephone calls from a call-centre and is configurable

and in most instances integrated with the insurers CRM system – it therefore optimises the sales productivity if used within a sales call-centre.

A further recommendation is for financial services business development executives, marketers and sales divisions to accelerate their adoption and their use of such targeted cross-sell techniques to guarantee their upsell and cross-sell communication plans are laser-beam focused on their customers which demonstrate a high propensity to convert when quoted, versus sales initiatives which are often executed in a broad fashion with a shotgun approach.

A final recommendation is to embed discount on profiles that demonstrate lower propensity to accept the insurance policy when quoted. Instead of an insurer reducing their insurance rates (premiums) broadly, the insurer can rather dynamically discount quotes for customer profiles less likely to accept the insurance product.

In addition, this research paper's findings can aid marketers in other segments of financial services (not just in short-term insurance) in investigating additional propensity modelling of their existing customer base to identify similar factors which indicate a higher propensity for the acceptance of additional and different financial services products. These improved marketing efforts will unlock the expansion of their market shares and accelerate product penetration outcomes for their existing customer bases.

6.4 Recommendations for further research

During this process and based on the findings and results from this research paper, the study offers and encourages possible paths of future research which can be conducted. New research opportunities and newly discovered data structures should not be ignored and many interesting and relevant research studies are excellently-suited to big data and associated analysis (Mahrt & Scharkow, 2013).

There are many insights and characteristics of customers which numerous financial services providers and insurers collect in their CRM and databases, some of which were not included in this research study. Future research could identify customer perception based on satisfaction indexes to be added as weightings in business development propensity models. Customer perception scores and satisfaction rates and their link on loyalty in a cross-sell campaign was not included or covered in this research paper, however there are previous studies which have touched on the link between customer satisfaction and sales performance in both tangible products (Gomez, McLaughlin, & Wittink, 2004), and financial services such as insurance (Maroofi, Ardalan, & Tabarzadi, 2017). Such previous studies can also incorporate the finding of this research paper to ventilate the weighting and impact of customer satisfaction on their propensity to accept further quotes from the same financial services provider.

Further research could also consider the roles of sales employees in attaining and delivering customer satisfaction (Evanschitzky, Sharma, & Prykop, 2012) and how such sales satisfaction has an impact of recommender models and cross-sell propensity models. In addition, some research papers have provided context on the impact and influence of acquisition channels in relation to cross-selling and loyalty (Verhoef & Donkers, 2005), however such findings and research can be amplified in the context of insurance and financial services providers.

There are existing research papers that have focused on customer retention in financial services and how the perception of service quality as well as satisfaction scores are linked to retention (Kassim & Souiden, 2007); it is known that customer satisfaction may not be the single determinant of retention and further research may be conducted in how the impact of a customer adding a short-term insurance product has on overall retention for the financial services provider retaining that said customer.

This research which was conducted focused on propensity modelling and recommendation of which customers to offer a short-term insurance quote to.

Further research can consider using similar customer profiles and characteristics to yield a higher return in marketing investment on other financial services products, such a life insurance, health insurance, commercial insurance, and investment products.

The research conducted in this study used quote conversion (specifically client acceptance when quoted) as the metric to measure return on marketing investment (ROMI). Further studies using existing knowledge on embedded value (calculated as the present value of the future profits in addition to the net asset value of the business) specifically in insurance (Bennett, Campbell, Kaster, & Segal, 2002) can further knowledge in ascertaining how short-term insurance product acquisition for an existing financial services customer base can increase the embedded value of the organization and therefore also contribute positively to the return on marketing investment.

Finally, variation of future research could look into using customer insights and characteristics on propensity models for the up-selling of an existing short-term insurance customer base with additional short-term insurance benefits. Existing research concluded on up-sell insights (Kannan & Balakrishnan, 2009) as well as research that looks at both up-sell and cross-sell (Salazar, Harrison, & Ansell, 2007) can be expanded on for a South African financial services provider that offers short-term insurance products.

Further to this, the study was limited to customer profiles who reside permanently in South Africa. Further research can therefore be undertaken where customer profiles from customers located around the globe can be utilised, even if on a lesser global scale and rather concentrated with continental parameters, or even geographically ring-fenced trade-bloc regions such as the Southern African Development Community (SADC).

6.5 Limitations of the study

The limitations to this research are where the insurer may not necessarily possess additional data which may supplement the research. There may also be further limitations, and not only limited to:

- errors in the insurer's data generation
- administration / user input errors into the CRM which may translate into inaccuracies in the data
- details relating to customers which cannot be revealed due to POPI requirements, and / or privacy limitations (for example details relating to race, medical conditions, religion)
- situations where customer data is stored independently at each product house within the insurer, resulting in data integration challenges.

An additional limitation includes the population and sampling-related constraints, where data was only harvested from a customer-base within a single insurer within South Africa. Furthermore, 566,758 customer profiles were analysed in this research and there may be larger customer bases to use to conduct a similar study on.

Further research could be undertaken in other states within the African continent, as well as other continents; a more macro view of a similar study could reduce national limitations and expand the study to include customer profiles from around the world.

Finally, this research study included data from existing customers only and focused on personal lines short-term insurance quote conversion and not on other financial services products' quote conversions.

6.6 Conclusion

There has been a considerable upsurge in competition amongst short-term insurance companies and the market is becoming more commoditised with often the only differentiator about price (also known as premium). South African short-term insurance companies are constantly having to increase their efficiencies to reduce their cost base, or minimise their loss ratios to keep their underwriting costs minimal thus enabling the best and most affordable premium to the market.

This is at times difficult when the insurance market overall often makes a very slender margin of profit, exacerbated by unpredictable weather patterns, and more recently the Covid-19 pandemic which has brought about a massive increase in unemployment (International-Labour-Organisation, 2021) and general economic downturn (Jackson, Weiss, Schwarzenberg, & Nelson, 2020). Shareholders still, however, insist on and expect acceptable returns in terms of their investments and insurers are expected to maintain adequate growth levels, where often the rate of new business growth has a direct association with their share price and where introducing innovations impact market share and the very survival of the business (Banbury & Mitchell, 1995).

With competition levels high and the market saturated with a high number of short-term insurance companies (SAIA, 2016) marketers and sales executive need to implement effective and smart ways to grow their customer bases. Insurance companies that have the benefit of products in different and various financial services markets possess a competitive advantage in that they possess unique insights into their customers at a customer level, and can thus use an enhanced version of personalisation to target very specific segments in their customer base for short-term insurance products. This enables success for the insurer when they customize and personalize pricing to customers based on their linked market segments as well as associated and preferred value propositions (Simon & Butscher, 2001).

Furthermore, for companies like insurers who often possess hundreds of thousands, and often millions of customers, smart targeted business development is key to ensure an efficient sales chassis. If an insurer has, for example, 3.5 million customers overall across financial services markets which include credit card / banking, health insurance, life insurance, asset management / investments and short-term insurance, then this insurer would need to implement smart business development and sales strategies if targeting its existing customer base for additional financial services products – it would not make sense for their sales channel to work down a list of 3.5 million customers!

Propensity modelling and recommendation systems therefore enable the insurer to laser-beam focus on specific segments of their customer population, where sales resources can approach these targeted customers who statistically demonstrate a higher propensity to accept a quote. Higher conversion rates in such business development activities therefore reduce the number of sales resources required, and furthermore reduce the number of quotes required to maximise new business revenue – a clear outcome of improved returns on marketing investment.

From the findings, it is clear and evident that personalisation, i.e. producing or designing the quote to cater for a customer's unique product appetite is a cost-effective approach to maximising new busines revenue from an existing customer base, and thus improve the return on marketing investment.

It is a recommendation from this research study that financial services providers adopt such data science and propensity modelling techniques to exploit the additional product conversion rate of existing customers, thereby not only increasing new business revenues, but also reducing their lapse rates. A customer is more likely to remain with a company the more products (value) they receive from that said company, as they become more "locked in" and develop increased dependency on the company (Harrison & Ansell, 2002).

7 BIBLIOGRAPHY

Aaker, D. A. (1996). Measuring brand equity across products and markets. *California management review, 38*(3), 102-120.

Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis, 16*(1), 3-9.

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS), 23*(1), 103-145.

Aghdaie, M. H., & Alimardani, M. (2015). Target market selection based on market segment evaluation: a multiple attribute decision making approach. *International Journal of Operational Research, 24*(3), 262-278.

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management, 29*(3), 469-486.

Ahearne, M., Jelinek, R., & Rapp, A. (2005). Moving beyond the direct effect of SFA adoption on salesperson performance: training and support as key moderating factors. Industrial Marketing Management, 34(4), 379−388.

Alberto, T., & da Silva, P. M. (2014). Simplified Customer Segmentation Applied to an Outbound Contact Center Dialer. *The Seventh International Conference on Advances in Computer-Human Interactions [AHCI]*, 23-27.

Ang, L., & Buttle, F. (2006). Customer retention management processes: A quantitative study. *European Journal of marketing, 40*(1/2), 83-99.

Armstrong, G., Adam, S., Denize, S., & Kotler, P. (2014). *Principles of marketing*: Pearson Australia.

Arora, N., Dreze, X., Ghose, A., Hess, J. D., Iyengar, R., Jing, B., . . . Neslin, S. (2008). Putting one-to-one marketing to work: Personalization, customization, and choice. *Marketing Letters, 19*(3-4), 305-321.

Babbie, & Mouton, J. (2001). The practice of social research: South African edition. *Cape Town: Oxford University Press Southern Africa*.

Babbie, E. R. (2014). *The Basics of Social Research*: Cengage Learning.

Banbury, C. M., & Mitchell, W. (1995). The effect of introducing important incremental innovations on market share and business survival. *Strategic management journal, 16*(S1), 161-182.

Beamish, N. G., & Armistead, C. G. (2001). Selected debate from the arena of knowledge management: new endorsements for established organizational practices. *International Journal of Management Reviews, 3*(2), 101-111.

Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom. In.

Bennett, N. E., Campbell, B. C., Kaster, M. L., & Segal, S. (2002). Embedded Value of Life Insurance Product Lines. *Record of the Society of Actuaries, 1*.

Berkovsky, S., Aroyo, L., Heckmann, D., Houben, G.-J., Kröner, A., Kuflik, T., & Ricci, F. (2007). Providing context-aware personalization through cross-context reasoning of user modeling data. *UBIDEUM 2007*, 2.

Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective*: Pearson.

Blattberg, R., Buesing, T., Peacock, P., & Sen, S. (1978). Identifying the deal prone segment. *Journal of Marketing Research*, 369-377.

Blattberg, R., Getz, G., & Thomas, J. (2001). *Customer equity: Building and managing relationships as valuable assets*: Harvard Business Press.

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). *Why Database Marketing?* : Springer.

Bodapati, A. V. (2008). Recommendation systems with purchase data. *Journal of Marketing Research, 45*(1), 77-93.

Bremner, A. (2001). CRM Comes of Age. *The Banker*(March).

Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of ‘big data’. *McKinsey Quarterly, 4*(1), 24-35.

Bryman, A. (2012). *Social Research Methods* (4th ed.): Oxford university press.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science, 45*(1), 12-19.

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science, 14*(4), 378-394.

Caldarelli, G., & Chessa, A. (2016). *Data science and complex networks: real case studies with Python*: Oxford University Press.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1983). Beyond external validity. *Journal of Consumer Research, 10*(1), 112-114.

Camilleri, M. A. (2018). Market segmentation, targeting and positioning. In *Travel marketing, tourism economics and the airline product* (pp. 69-83): Springer.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly, 36*(4), 1165-1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: a survey. *Springer Science+Business Media*.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system.* Paper presented at the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2, 1*(4).

Connaway, L. S., & Powell, R. R. (2010). *Basic research methods for librarians*: ABC-CLIO.

Conti, N., Jennett, C., Maestre, J., & Sasse, M. A. (2012). *When did my mobile turn into a'sellphone'?: a study of consumer responses to tailored smartphone ads.* Paper presented at the Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers.

Cook, W. A., & Talluri, V. S. (2004). How the pursuit of ROMI is changing marketing management. *Journal of Advertising Research, 44*(3), 244-254.

Cooper, D., & Schindler, P. (2011). *Business Research Methods* (11th ed.).

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Method Approaches*: Sage Publications.

Danaher, P. J., & Rust, R. T. (1996). Determining the optimal return on investment for an advertising campaign. *European Journal of Operational Research, 95*(3), 511-521.

Datar, A., & Sturm, R. (2004). Physical education in elementary school and body mass index: evidence from the early childhood longitudinal study. *American journal of public health, 94*(9), 1501-1506.

Davenport, T. H., & Dyché, J. (2013). Big data in big companies. *International Institute for Analytics*, 3.

Day, G. S., & Montgomery, D. B. (1999). Charting new directions for marketing. *The Journal of Marketing*, 3-13.

de Villiers, C., Cerbone, D., & Van Zijl, W. (2020). The South African government's response to COVID-19. *Journal of Public Budgeting, Accounting & Financial Management*.

Discovery. (2017). About Us. Retrieved from https://www.discovery.co.za/portal/individual/about-us

Eckerson, W. W. (2007). Predictive analytics. *Extending the Value of Your Data Warehousing Investment ". TDWI Best Practices Report, 1*, 1-36.

Edwards, Y. D., & Allenby, G. M. (2003). Multivariate analysis of multiple response data. *Journal of Marketing Research, 40*(3), 321-334.

EIOPA (European Insurance and Occupational Pensions Authority) (2019) Big Data analytics in motor and health insurance: A thematic review. Available at: https://register.eiopa.europa.eu/Publications/EIOPA_ BigDataAnalytics_ThematicReview_April2019.pdf (accessed 05 September 2021).

Embarak, D. O. (2018). *Data Analysis and Visualization Using Python*: Apress.

Ertemel, A. V. (2015). Consumer insight as competitive advantage using big data and analytics. *International Journal of Commerce and Finance, 1*(1), 45-51.

Evans, J. R., & Lindner, C. H. (2012). Business analytics: the next frontier for decision sciences. *Decision Line, 43*(2), 4-6.

Evanschitzky, H., Sharma, A., & Prykop, C. (2012). The role of the sales employee in securing customer satisfaction. *European Journal of Marketing*.

Financial Advisory and Intermediary Services Act 37 of 2002, (2002).

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review, 1*(2), 293-314.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fitch-Ratings. (2017). *Fitch Downgrades South Africa to 'BB*+. Retrieved from https://www.fitchratings.com/site/pr/1021851

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (1993). *How to design and evaluate research in education* (Vol. 7): McGraw-hill New York.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics, 28*(2), 337-407.

Goldsmith, R. E. (1999). The personalised marketplace: beyond the 4Ps. *Marketing Intelligence & Planning, 17*(4), 178-185.

Goller, B., & Hoffmann, S. (2013). Leveraging "Big Data" for Precision In-Store Marketing. *Retail Property Insights, 20*(1), 30-36. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=87956920&site=ehost-live

Gomez, M. I., McLaughlin, E. W., & Wittink, D. R. (2004). Customer satisfaction and retail sales performance: an empirical investigation. *Journal of retailing, 80*(4), 265-278.

Goodman, J. (1992). Leveraging the customer database to your competitive advantage. *DIRECT MARKETING-GARDEN CITY-, 55*, 26-26.

Haddadi, H., Hui, P., & Brown, I. (2010). *MobiAd: private and scalable mobile advertising.* Paper presented at the Proceedings of the fifth ACM international workshop on Mobility in the evolving internet architecture.

Hair, J. (2007). Knowledge creation in marketing: the role of predictive analytics. *European Business Review, 19*(4), 303-315.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics, 7*(1), 37-53.

Hansotia, B. (2004). Company activities for managing customer equity. *Journal of Database Marketing & Customer Strategy Management, 11*(4), 319-332.

Harrison, T., & Ansell, J. (2002). Customer retention in the insurance industry: using survival analysis to predict cross-selling opportunities. *Journal of Financial Services Marketing, 6*(3), 229-239.

Hasselblad, V. (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics, 8*(3), 431-444.

Hays, W. L. (1981). *Statistics* (3rd ed.): Harcourt Brace College Publishers.

Herlocker, J. L. (2000). *Understanding and improving automated collaborative filtering systems.* Citeseer,

Hilpisch, Y. (2014). *Python for Finance: Analyze big financial data*: " O'Reilly Media, Inc.".

Holický, M. (2013). Testing of Statistical Hypotheses. In *Introduction to Probability and Statistics for Engineers* (pp. 125-138): Springer.

Hughes, D. J., Rowe, M., Batey, M., & Lee, A. (2012). A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior, 28*(2), 561-569.

Ibbotson, P., & Moran, L. (2003). E-banking and the SME/bank relationship in Northern Ireland. *International Journal of Bank Marketing, 21*(2), 94-103.

International-Labour-Organisation. (2021). *COVID-19 and the world of work*. Retrieved from https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms_767028.pdf

Jackson, J. K., Weiss, M. A., Schwarzenberg, A. B., & Nelson, R. M. (2020). Global economic effects of COVID-19.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jones, E., Brown, S. P., Zoltners, A. A., & Weitz, B. A. (2005). The changing environment of selling and sales management. Journal of Personal Selling and Sales Management, XXV(2), 105−111.

Kahreh, M. S., & Kahreh, Z. S. (2012). An empirical analysis to design enhanced customer lifetime value based on customer loyalty: evidences from Iranian banking sector. *Iranian Journal of Management Studies, 5*(2), 145.

Kamakura, W. A., Kossar, B. S., & Wedel, M. (2004). Identifying innovators for the cross-selling of new products. *Management Science, 50*(8), 1120-1133.

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in marketing, 8*(4), 329-349.

Kamakura, W. A., Wedel, M., De Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in marketing, 20*(1), 45-65.

Kannan, S., & Balakrishnan, P. (2009). Ensemble Approach to Predict Churn, Appetency and Up-Sell.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and translational science, 7*(4), 342-346.

Kaske, F., Kugler, M., & Smolnik, S. (2012). *Return on investment in social media--Does the hype pay off? Towards an assessment of the profitability of social media in*

organizations. Paper presented at the System Science (HICSS), 2012 45th Hawaii International Conference on.

Kassim, N. M., & Souiden, N. (2007). Customer retention measurement in the UAE banking sector. *Journal of Financial Services Marketing, 11*(3), 217-228.

Kim, M. J., & Jun, J. W. (2008). A case study of mobile advertising in South Korea: Personalisation and digital multimedia broadcasting (DMB). *Journal of Targeting, Measurement and Analysis for Marketing, 16*(2), 129-138.

KingPrice. (2012). Once upon a time. Retrieved from https://www.kingprice.co.za/about/Index

Kitchen, P. J. (2010). *Integrated brand marketing and measuring returns*: Springer.

Kitchen, P. J., & Schultz, D. E. (1999). A multi-country comparison of the drive for IMC. *Journal of Advertising Research, 39*(1), 21-21.

Knott, A., Hayes, A., & Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of Interactive Marketing, 16*(3), 59-75.

Koekemoer, L. (2014). *Marketing Communication - An Integrated Approach*: Juta & Co, Cape Town: South Africa.

Kolb, B. (2010). *Marketing research: a practical approach*: Sage.

KPMG. (2014). *The South African Insurance Industry Survey 2014*. Retrieved from https://www.kpmg.com/ZA/en/IssuesAndInsights/ArticlesPublications/Financial-Services/Documents/2014%20KPMG%20Insurance%20Survey.pdf

KPMG. (2016). *The South African Insurance Industry Survey 2016*. Retrieved from https://assets.kpmg.com/content/dam/kpmg/pdf/2016/07/2016-Insurance-Survey-v2.pdf

Krishna, A., Currim, I. S., & Shoemaker, R. (1991). Consumer perceptions of promotional activity.

Kumar, R. (2005). *Research Methodology: a Step-By-Step Guide for Beginners* (2nd ed.). London: Sage.

Kumar, R. (2011). *Research Methodology: A step by step guide for beginners* (3rd Edition ed.).

Labe, R. P. (1994). Database marketing increases prospecting effectiveness at Merrill Lynch. *Interfaces, 24*(5), 1-12.

Laerd-Statistics. (2013). Descriptive and Inferential Statistics. Retrieved from https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT sloan management review, 52*(2), 21.

Leedy, P., & Ormrod, J. (2013). *Practical Research: Planning and Design* (10th ed.): Macmillan.

Lemon, K. N., Rust, R. T., & Zeithaml, V. A. (2001). What drives customer equity. *Marketing management, 10*(1), 20.

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research, 42*(2), 233-239.

Liedtke, P. M. (2007). What's insurance to a modern economy? *The Geneva Papers on Risk and Insurance-Issues and Practice, 32*(2), 211-221.

Lubo-Robles, D., Devegowda, D., Jayaram, V., Bedle, H., Marfurt, K. J., & Pranter, M. J. (2020). Machine Learning model interpretability using SHAP values: application to a seismic facies classification task. In *SEG Technical Program Expanded Abstracts 2020* (pp. 1460-1464): Society of Exploration Geophysicists.

Luque, A., Carrasco, A., Martín, A., & de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91*, 216-231.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20-33.

Maroofi, F., Ardalan, A. G., & Tabarzadi, J. (2017). The effect of sales strategies in the financial performance of insurance companies. *International Journal of Asian Social Science, 7*(2), 150-160.

Martinez-Cantin, R. (2014). BayesOpt: a Bayesian optimization library for nonlinear optimization, experimental design and bandits. *J. Mach. Learn. Res., 15*(1), 3735-3739.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt.

McDermott, R. (2011). Internal and external validity. *Cambridge handbook of experimental political science*, 27-40.

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*: " O'Reilly Media, Inc.".

Mehl, M. R., & Gill, A. J. (2010). Automatic text analysis.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*.

Millard, D. (2019). A comparison between the COFI Bill and the FAIS Act in light of the TCF requirements.

Miller, T. W. (2014). *Modeling techniques in predictive analytics with Python and R: A guide to data science*: FT Press.

Moneyweb. (2017). Liberty, Standard Bank developing short-term insurance offering [Press release]. Retrieved from https://www.moneyweb.co.za/news/companies-and-deals/liberty-standard-bank-developing-short-term-insurance-offering/

Moorthy, J., Lahiri, R., Biswas, N., Sanyal, D., Ranjan, J., Nanath, K., & Ghosh, P. (2015). Big data: prospects and challenges. *J. Decis. Makers, 40*, 74-96.

Myers, J. L., Well, A., & Lorch, R. F. (2010). *Research design and statistical analysis*: Routledge.

Nick, T. G. (2007). Descriptive statistics. *Topics in biostatistics*, 33-52.

Old-Mutual. (2016). *Fundamentals*. Retrieved from https://www.oldmutual.co.za/docs/default-source/unit-trust/annual-reports/fundamentals/fundamentals-may-2016.pdf?sfvrsn=0

OUTsurance. (1998). About us. Retrieved from https://www.outsurance.co.za/about-outsurance/

Peltier, J. W., Schibrowsky, J. A., & Schultz, D. E. (2003). Interactive integrated marketing communication: combining the power of IMC, the new media and database marketing. *International Journal of Advertising, 22*(1), 93-115.

Perrey, J., Spillecke, D., & Umblijs, A. (2013). Smart analytics: How marketing drives short-term and long-term growth. *McKinsey Quarterly*, 00425-00423.

Personal-Finance. (2017). How the downgrades affect you, the consumer [Press release]. Retrieved from http://www.iol.co.za/personal-finance/how-the-downgrades-affect-you-the-consumer-8547250

Ping-An. (2021). Who We Are. Retrieved from https://group.pingan.com/about_us/who_we_are.html

Piryonesi, S. M., & El-Diraby, T. E. (2020). Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements, 146*(2), 04020022.

Porter, M. E., & Millar, V. E. (1985). How information gives you competitive advantage. In: Harvard Business Review, Reprint Service.

PWC. (2016). Challenges of insurance marketing. Retrieved from http://smallbusiness.chron.com/challenges-insurance-marketing-71593.html

Reed, K. L., Ansusinha, T., & Hariharan, H. S. (2010). Adaptive marketing using insight driven customer interaction. In: Google Patents.

Reichheld, F. F., & Schefter, P. (2000). E-loyalty. *Harvard business review, 78*(4), 105-113.

Ricci, F., Rokach, L., & Shapira, B. (2011). *Introduction to recommender systems handbook*: Springer.

Rubel, T. (2006). *The next wave in customer relationship analytics*. Retrieved from www.dmreview.com/article_sub.cfm?articleId = 1005301

Rubin, D. B., & Waterman, R. P. (2006). Estimating the causal effects of marketing interventions using propensity score methadology. *Institute of Mathematical Statistics*. Retrieved from http://projecteuclid.org/download/pdfview_1/euclid.ss/1154979822

Rumbauskas Jr, F. J. (2010). *Never Cold Call Again: Achieve Sales Greatness Without Cold Calling*: John Wiley & Sons.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. *Journal of Marketing, 68*(1), 109-127.

SA-Government. (2014). Protection of Personal Information Act. Retrieved from http://www.gov.za/sites/www.gov.za/files/37067_26-11_Act4of2013ProtectionOfPersonalInfor_correct.pdf

Sagiroglu, S., & Sinanc, D. (2013). *Big data: A review.* Paper presented at the Collaboration Technologies and Systems (CTS), 2013 International Conference on.

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(12), 2019.

SAIA. (2016, 2016). South African Insurance Association. Retrieved from http://www.saia.co.za

Salazar, M. T., Harrison, T., & Ansell, J. (2007). An approach for the identification of cross-sell and up-sell opportunities using a financial services customer database. *Journal of Financial Services Marketing, 12*(2), 115-131.

Sawhney, M., & Zabin, J. (2002). Managing and measuring relational equity in the network economy. *Journal of the Academy of Marketing Science, 30*(4), 313-332.

Schultz, D. (1996). Can you improve marketing results by spending less? *American Marketing Association, 30*(8), 4-5.

Schultz, D., & Block, M. (2014). Using big data file fusion to determine the effects of social media on retail brand preference. *Applied Marketing Analytics, 1*(1), 81-102.

Schultz, D. E. (1999). Integrated marketing communications and how it relates to traditional media advertising. *The advertising business: Operations, creativity, media planning, integrated communications*, 325-338.

Seggie, S. H., Cavusgil, E., & Phelan, S. E. (2007). Measurement of return on marketing investment: A conceptual framework and the future of marketing metrics. *Industrial Marketing Management, 36*(6), 834-841.

Shan, L. H., Chin, T. A., Sulaiman, Z., & Muharam, F. M. (2016). EFFECTIVE MOBILE ADVERTISING ON MOBILE DEVICES.

Shaw, I., & Vulkan, N. (2012). Competitive personalized pricing: An experimental investigation. *Review of Economic Studies (Submitted)*.

Sheth, J. N., & Sisodia, R. S. (1995). Feeling the heat: More than ever before, marketing is under fire to account for what it spends. Retrieved from http://www.jagsheth.net/publications_subject.html

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *Mis Quarterly*, 553-572.

Sicular, S. (2013). Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. *Forbes Tech*. Retrieved from http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#12f621af3bf6

Simon, H., & Butscher, S. A. (2001). Individualised pricing:: boosting profitability with the higher art of power pricing. *European Management Journal, 19*(2), 109-114.

Slack, M. K., & Draugalis, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy, 58*(22), 2173-2181.

Spiess, J., T'Joens, Y., Dragnea, R., Spencer, P., & Philippart, L. (2014). Using big data to improve customer experience and business performance. *Bell labs technical journal, 18*(4), 3-17.

Statistics-South-Africa. (2015). *Annual Financial Statistics 2014*. Retrieved from http://www.statssa.gov.za/publications/P0021/P00212014.pdf

Statistics-South-Africa. (2016a). *Annual Report 2015/2016*. Retrieved from http://www.statssa.gov.za/wp-content/uploads/2016/10/Annual_Report_2016_Book_1_.pdf

Statistics-South-Africa. (2016b). *The last 15 years: business income, spending and profit*. Retrieved from http://www.statssa.gov.za/?p=9227

Statistics-South-Africa. (2020). *GDP: Quantifying SA's economic performance in 2020*. Retrieved from http://www.statssa.gov.za/wp-content/uploads/2021/03/gdp2.jpg

Stats-SA. (2015). *Annual Financial Statistics 2014*. Retrieved from http://www.statssa.gov.za/publications/P0021/P00212014.pdf

Stewart, S. (1993). The Stern Stewart Performance 1000 Database Package: Introduction and documentation. *New York: Stern Stewart Management Services*.

Stone, R. L., & Clancy, K. J. (2005). Don't blame the metrics. *Harvard Business Review, 83*(6), 26-27.

Subasi, A. (2020). *Practical Machine Learning for Data Analysis Using Python*: Academic Press.

Tankovska, H. (2021). *Hours of video uploaded to YouTube every minute 2007-2019*. Retrieved from https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

Tian, H., Cai, H., Wen, J., Li, S., & Li, Y. (2019). *A Music Recommendation System Based on logistic regression and eXtreme Gradient Boosting.* Paper presented at the 2019 International Joint Conference on Neural Networks (IJCNN).

Trading-Economics. (2017). *South African Unemployment Rate*. Retrieved from http://www.tradingeconomics.com/south-africa/unemployment-rate

Tuohimaa, P., Tenkanen, L., Ahonen, M., Lumme, S., Jellum, E., Hallmans, G., . . . Luostarinen, T. (2004). Both high and low levels of blood vitamin D are associated with a higher prostate cancer risk: A longitudinal, nested case-control study in the Nordic countries. *International Journal of Cancer, 108*(1), 104-108.

Verhoef, P. C., & Donkers, B. (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing, 19*(2), 31-43.

Waller, M., & Fawcett, S. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*.

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.

Weber, J. A. (2002). Managing the marketing budget in a cost-constrained environment. *Industrial Marketing Management, 31*(8), 705-717.

Ziady, H. (2016). *Short-term insurers fighting for the same customers*. Retrieved from https://www.moneyweb.co.za/news/industry/short-term-insurers-fighting-customers/

8 APPENDIX

8.1 Appendix A

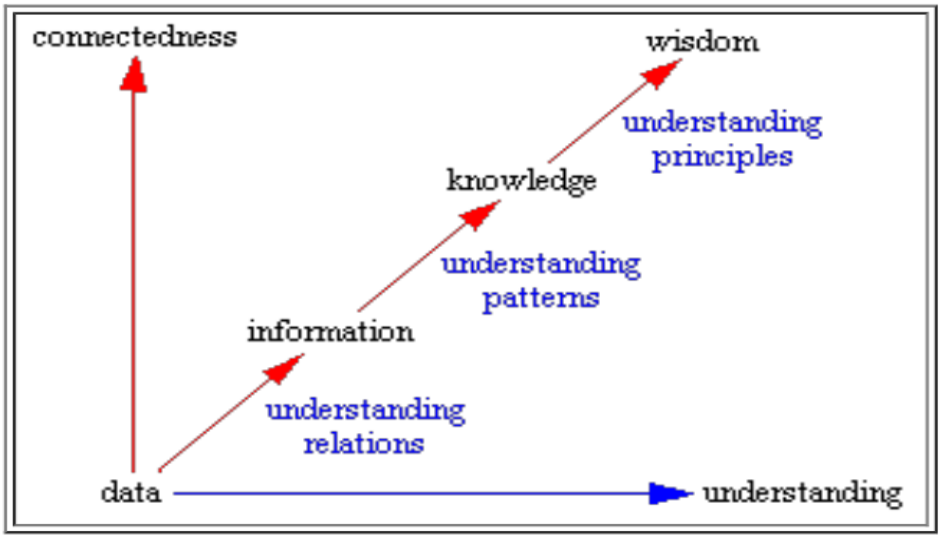

Figure 29: Data to wisdom relationship (Bellinger et al., 2004)

8.2 Appendix B

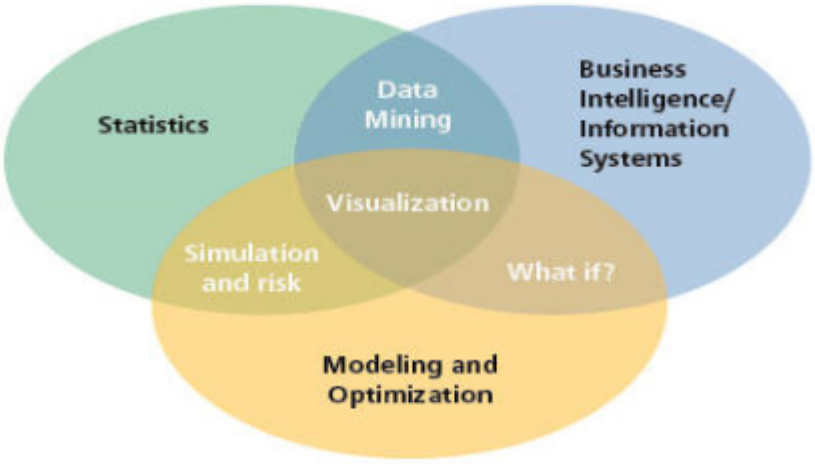

Figure 30: Modelling and analytics (Evans & Lindner, 2012)

8.3 Appendix C

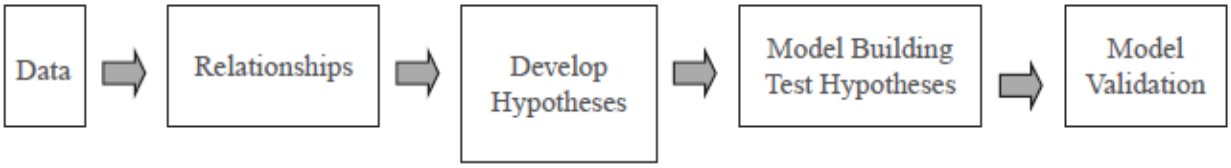

Figure 31: Data mining and predictive analytics process chain (Hair, 2007)

8.4 Appendix D

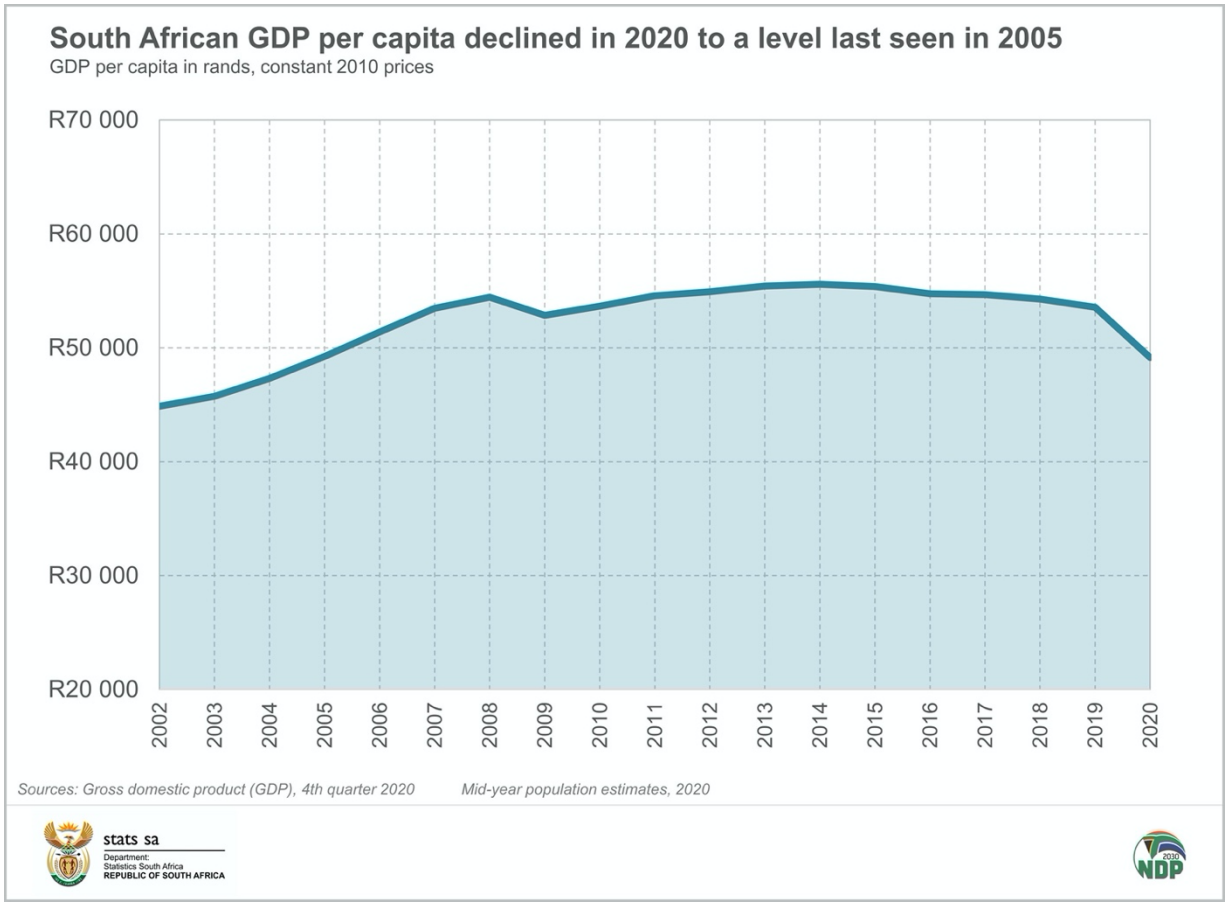

Figure 32: (Statistics-South-Africa, 2020)

8.5 Appendix E

Dissertation Ethical Clearance
Master of Management: Strategic Marketing (Wits Business School)
Darryl Grater

To whom it may concern,

RE: Dissertation: *Using Big Data to increase ROMI in a South African Insurer*

Discovery Ltd. hereby confirms that Darryl Grater (RSA ID 8707025190085 / Wits Student Number 1539545) has approached the company for permission to utilize Discovery's corporate data as his dissertation's research instrument for the above mentioned Masters degree.

This document hereby confirms that Discovery Ltd. Is aware of the context of the dissertation and has provided the necessary permission to proceed with the research.

If you have any enquiry in this regard, please contact the office of Discovery group Chief Marketing Officer, on 011 529 3835.

Sincerely

Dinesh Govender
Chief Marketing Officer: Discovery

Figure 33: Ethical Clearance (Discovery Ltd)

8.6 Appendix F

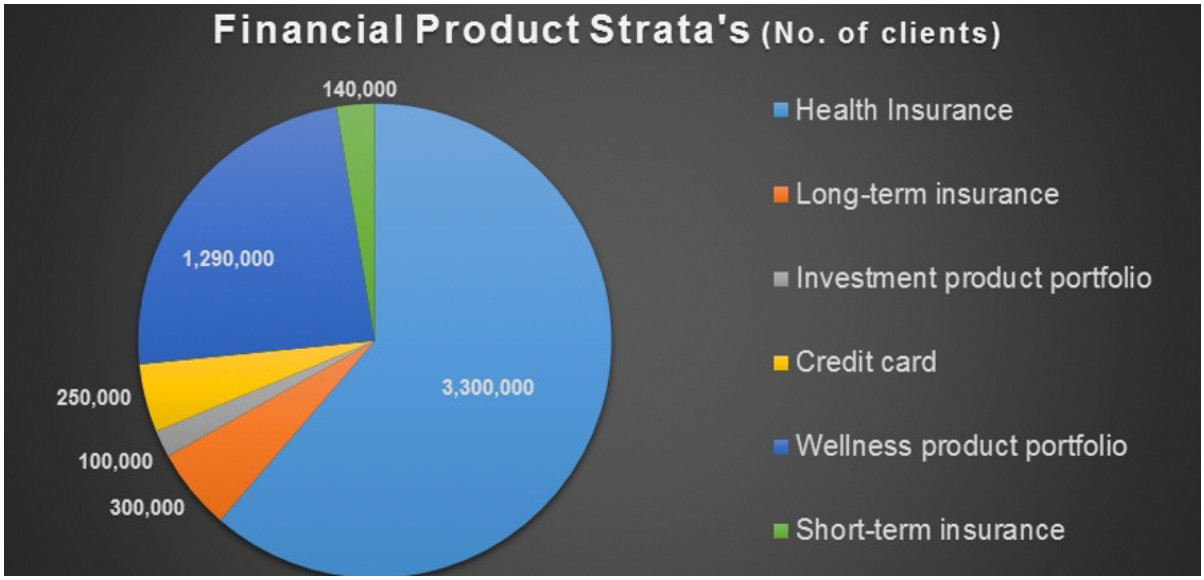

Figure 34: Research population split (Author)

8.7 Appendix G

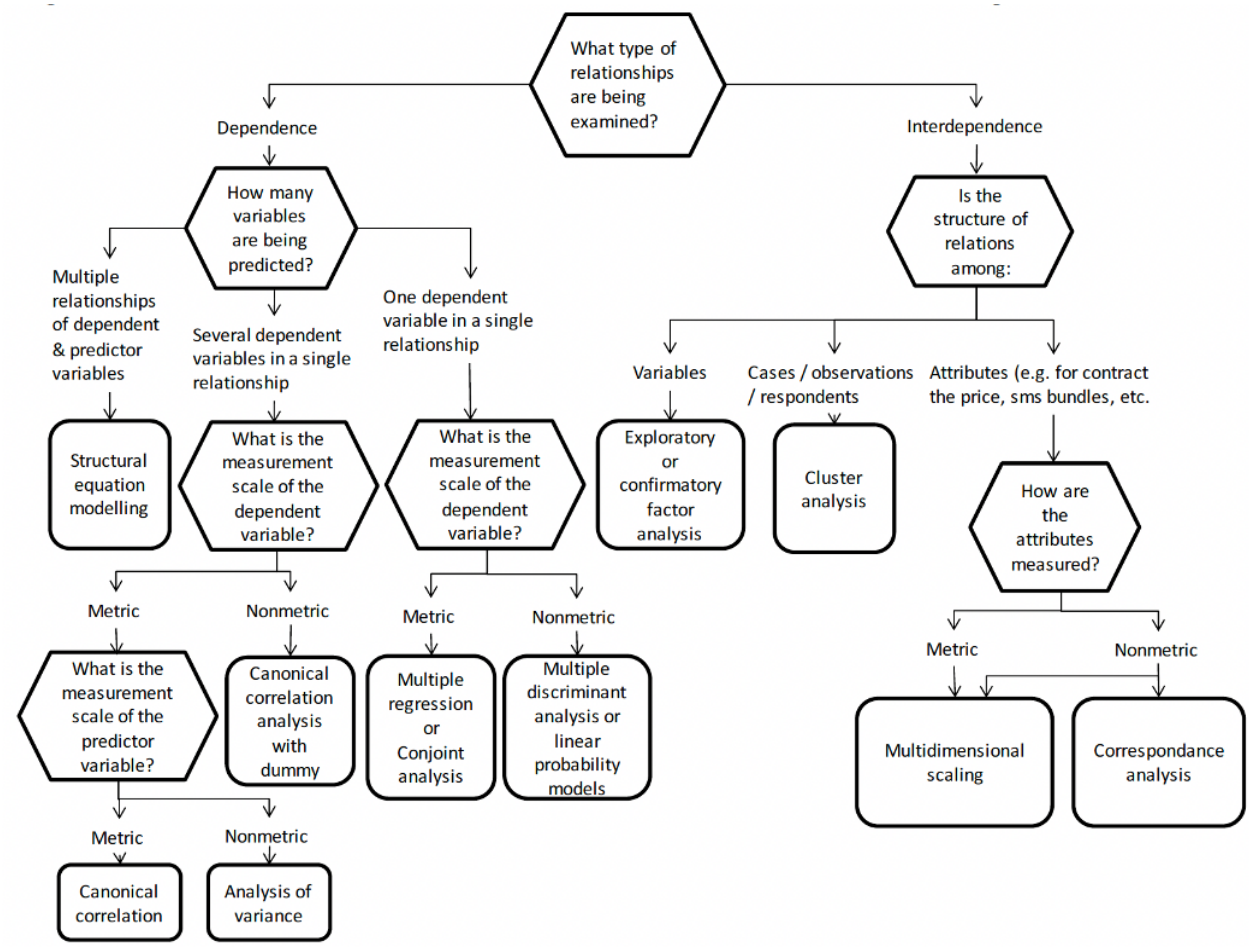

Figure 35: Decision tree for choosing a correct statistical technique (Black et al., 2010)

8.8 Appendix H

```
def xgbc_cv(max_depth,learning_rate,n_estimators,reg_alpha, subsample, colsample_bytree, colsample_bylevel,min_child_weight):


    estimator_function = xgb.XGBClassifier(max_depth=int(max_depth),
                                           learning_rate= learning_rate,
                                           n_estimators= int(n_estimators),
                                           reg_alpha = reg_alpha,
                                           nthread = -1,
                                           objective='binary:logistic',
                                           seed = seed,
                                           eval_metric = 'aucpr',
                                           subsample = subsample,
                                           colsample_bytree = colsample_bytree,
                                           colsample_bylevel = colsample_bylevel,
                                           min_child_weight = min_child_weight
                                          )
     # Fit the estimator
    estimator_function.fit(X_train,y_train)

    # calculate out-of-the-box roc_score using validation set 1
    probs = estimator_function.predict_proba(X_val)
    probs = probs[:,1]
    val1_roc = average_precision_score(y_val,probs)


    # return the mean validation score to be maximized
    return val1_roc
```

Figure 36: Construct function to maximise validation performance

8.9 Appendix I

```
rcParams['figure.figsize'] = 8, 8
test = pd.DataFrame(progress['test']['aucpr'])
train = pd.DataFrame(progress['train']['aucpr'])
val = pd.DataFrame(progress['val']['aucpr'])
plt.plot(test, label = "Test")
plt.plot(train, label = "Train")
plt.plot(val, label = "Validation")
plt.legend()
plt.xlabel('Number of training rounds', fontsize= 16)
plt.ylabel('Area under precision-recall curve' , fontsize= 16)
plt.legend(loc="upper left")
plt.title('Model performance on test and train datasets vs training rounds' , fontsize= 22)
plt.savefig('Model performance on test, train and validation datasets.png', dpi=1200)
```

FIGURE 37: Evaluation on model on test, train and validation sets

8.10 Appendix J

```
import warnings
warnings.filterwarnings('ignore')

from bayes_opt import BayesianOptimization

# alpha is a parameter for the gaussian process
# Note that this is itself a hyperparemter that can be optimized.
gp_params = {"alpha": 1e-10}

# We create the BayesianOptimization objects using the functions that utilize
# the respective classifiers and return cross-validated scores to be optimized.

seed = 112 # Random seed

# We create the bayes_opt object and pass the function to be maximized
# together with the parameters names and their bounds.
# Note the syntax of bayes_opt package: bounds of hyperparameters are passed as two-tuples

hyperparameter_space = {
    'max_depth': (3, 8),
    'learning_rate': (0, 1),
    'n_estimators' : (10,50),
    'reg_alpha': (0,1),
    'subsample' : (0.6, 0.9),
    'colsample_bytree': (0.6, 0.9),
    'colsample_bylevel': (0.6, 0.9),
    'min_child_weight': (500, 2000)
}

xgbcBO = BayesianOptimization(f = xgbc_cv,
                              pbounds =  hyperparameter_space,
                              random_state = seed,
                              verbose = 10)

# Finally we call .maximize method of the optimizer with the appropriate arguments
# kappa is a measure of 'aggressiveness' of the bayesian optimization process
# The algorithm will randomly choose 3 points to establish a 'prior', then will perform
# 10 interations to maximize the value of estimator function
xgbcBO.maximize(init_points=20,n_iter=100,acq='ucb', kappa= 3, **gp_params)
```

Figure 38: Bayesian hyperparameter tuning

8.11 Appendix K

Table 3: Iterations of the Bayesian optimisation algorithm

iter	target	colsam...	colsam...	learni...	max_depth	min_ch...	n_esti...	reg_alpha	subsample
1	0.5345	0.7125	0.7921	0.95	3.378	1.665e+0	43.31	0.05481	0.8453
2	0.5237	0.8656	0.8167	0.002556	7.906	1.015e+0	13.79	0.3946	0.6015
3	0.5392	0.821	0.8867	0.8206	4.725	1.068e+0	41.39	0.08623	0.7638
4	0.5255	0.6487	0.687	0.04501	4.661	1.605e+0	43.57	0.7094	0.6565
5	0.5417	0.8957	0.7164	0.5032	6.818	1.526e+0	38.88	0.2142	0.8289
6	0.5354	0.6226	0.7833	0.4256	4.172	1.324e+0	33.08	0.548	0.6465
7	0.535	0.7898	0.6327	0.1147	5.481	1.066e+0	49.16	0.7133	0.6184
8	0.5344	0.8701	0.8842	0.8292	3.77	1.48e+03	42.26	0.1011	0.6303
9	0.5269	0.6699	0.7476	0.3623	3.162	801.0	17.66	0.3966	0.6638
10	0.5324	0.6608	0.8761	0.4856	4.102	666.6	10.81	0.228	0.6046
11	0.5326	0.6223	0.7497	0.2999	3.165	569.8	34.27	0.7251	0.7369
12	0.5263	0.6182	0.694	0.01406	7.068	729.3	13.28	0.2656	0.6736
13	0.5304	0.8805	0.6318	0.2162	4.008	1.512e+0	16.69	0.4781	0.8043
14	0.5344	0.6639	0.7185	0.204	4.006	1.016e+0	37.31	0.8783	0.7352
15	0.5337	0.8461	0.8056	0.8032	4.306	1.36e+03	16.19	0.951	0.656
16	0.5403	0.6153	0.6785	0.1633	7.069	709.5	38.72	0.4271	0.6557
17	0.5347	0.6442	0.7545	0.9266	3.009	713.4	24.46	0.9582	0.6109
18	0.5407	0.7865	0.7011	0.5926	7.993	1.138e+0	48.25	0.157	0.6651
19	0.5322	0.6138	0.618	0.5135	3.573	1.805e+0	44.26	0.4445	0.8877
20	0.5357	0.8839	0.6043	0.8254	6.047	1.668e+0	27.85	0.7123	0.6335
21	0.5285	0.7034	0.6212	0.6695	5.535	1.999e+0	11.94	0.1155	0.883
22	0.5406	0.7309	0.614	0.7149	7.264	1.999e+0	49.21	0.08838	0.864
23	0.5395	0.8485	0.892	0.7729	6.602	501.6	10.53	0.4954	0.6548
24	0.5206	0.8212	0.8556	0.6233	3.796	1.204e+0	10.23	0.2849	0.8187
25	0.5381	0.7425	0.651	0.1997	4.579	504.7	49.23	0.09148	0.6935
26	0.5397	0.7986	0.6818	0.8477	7.782	1.915e+0	46.36	0.1697	0.6474
27	0.5424	0.7262	0.8638	0.8021	6.087	871.9	49.54	0.06814	0.8186
28	0.5393	0.8528	0.7709	0.6458	7.968	1.711e+0	48.8	0.05312	0.6498
29	0.5423	0.6837	0.8691	0.4838	4.645	868.0	47.44	0.8005	0.8534
30	0.5275	0.6936	0.8367	0.2373	7.683	1.854e+0	11.33	0.03775	0.6508
31	0.5396	0.872	0.7954	0.8631	7.886	1.376e+0	49.45	0.3553	0.8067
32	0.5431	0.8784	0.6923	0.6667	4.537	794.9	49.42	0.2569	0.7237
33	0.5415	0.737	0.773	0.7559	7.832	644.6	48.56	0.4347	0.7107
34	0.5322	0.6859	0.7839	0.4315	3.183	1.969e+0	49.44	0.2355	0.756
35	0.5373	0.7278	0.6964	0.9449	3.006	928.8	49.24	0.1066	0.8568
36	0.5404	0.8585	0.6575	0.9837	4.186	1.226e+0	50.0	0.3669	0.8642
37	0.5342	0.8874	0.799	0.6996	3.052	1.132e+0	48.07	0.3093	0.6388
38	0.537	0.6078	0.7739	0.845	7.859	1.275e+0	39.51	0.6592	0.6146
39	0.5308	0.8919	0.6707	0.9628	5.856	1.738e+0	10.46	0.191	0.7568
40	0.5301	0.7503	0.8056	0.1269	3.127	667.9	49.58	0.1175	0.806
41	0.5419	0.6386	0.7484	0.6565	7.536	798.1	49.46	0.08957	0.6734
42	0.5402	0.663	0.7331	0.9852	7.915	530.1	45.95	0.6321	0.7146
43	0.5389	0.6411	0.7479	0.9456	7.901	605.7	13.48	0.3316	0.888
44	0.5381	0.8202	0.7678	0.8315	7.074	1.87e+03	48.85	0.197	0.7009
45	0.5436	0.8067	0.6748	0.4889	7.829	1.191e+0	48.65	0.7033	0.6794
46	0.54	0.7507	0.788	0.8664	7.295	500.3	30.03	0.7423	0.797
47	0.5368	0.6011	0.7511	0.9474	7.378	1.12e+03	13.47	0.9031	0.8317
48	0.5399	0.7377	0.7757	0.8926	7.449	754.3	49.0	0.4261	0.7202
49	0.536	0.6824	0.8553	0.7324	7.891	1.421e+0	12.22	0.0536	0.7771
50	0.5459	0.883	0.6095	0.467	7.872	585.8	48.9	0.3189	0.8332

8.12 Appendix L

```
calibrate = pd.DataFrame(list(zip(y_test, y_pred_test)), columns=['label','prediction'])

calibrate[['Group']] = pd.cut(calibrate.prediction,bins=[0,0.2,0.4,0.6,0.8,1],
       labels=['0-20','20-40','40-60','60-80', '80-100'], )

calibrate_grouped = calibrate.groupby(['Group']).mean()

fig = plt.figure()
ax = fig.add_subplot(1,1,1)
ax.plot([-1,1],[-1,1], 'black', linewidth=1, linestyle = '--', label = "Perfect calibration")
ax.plot(calibrate_grouped.prediction, calibrate_grouped.label, color = 'green',
         linestyle = 'solid', marker = 'o',
         markerfacecolor = 'green',  label = "Conversion model")

plt.axes().set_aspect('equal')
ax.set_xlim(0,1)
ax.set_ylim(0,1)
plt.legend()

ax.set_xlabel('Average conversion probability',  fontsize= 16)
ax.set_ylabel('Average actual conversion', fontsize= 16)
ax.set_title('Probability calibration plot', fontsize= 22)
plt
plt.savefig('Probability calibration plot.png', dpi=1200)
```

Figure 39: Probability calibration plot

8.13 Appendix M

```
from yellowbrick.classifier import DiscriminationThreshold
# Instantiate the classification model and visualizer
model = xgb.XGBClassifier(max_depth=int(max_depth),
                                          learning_rate= learning_rate,
                                          n_estimators= int(n_estimators),
                                          reg_alpha = reg_alpha,
                                          nthread = -1,
                                          objective='binary:logistic',
                                          seed = seed,
                                          eval_metric = 'aucpr',
                                          subsample = subsample,
                                          colsample_bytree = colsample_bytree,
                                          colsample_bylevel = colsample_bylevel,
                                          min_child_weight = min_child_weight
                          )

visualizer = DiscriminationThreshold(model)

visualizer.fit(X_train, y_train)        # Fit the data to the visualizer
# visualizer.score(X_test,y_test)
visualizer.draw()
visualizer.poof(outpath="Discrimination threshold analysis.png" )
```

Figure 40: Threshold plot for the XGBClassifier

8.14 Appendix N

```
import shap
explainer = shap.TreeExplainer(xgboost_model)
shap_values = explainer.shap_values(X_test)
shap.summary_plot(shap_values,X_train, plot_type="bar")
```

Figure 41: SHAP global feature importance code

8.15 Appendix O

```
shap.plots.partial_dependence(
    "AGE", model.predict, X_test, ice=False,
    model_expected_value=True, feature_expected_value=True,
)
```

Figure 42: SHAP partial-dependence plot of "age" feature

8.16 Appendix P

```
df = pd.read_csv("query_20210317.csv", sep = "|")
df.loc[
         ((df['AGE'] > 40) & (df['AGE'] < 60) ) &
        (df['PRODUCTS_ADDED_IP6M'] >= 1) &
#         (df['VIT_ENTITY_DURATION'] <= 50) &
        (df['MONTHS_ON_DSY'] <= 60) &
        (df['BUSINESS_TYPE_DESCR'] == 'New Business - Smart Advice')
    ,'comparison2'

      ] = '1'
df['comparison2'] = df['comparison2'].apply(lambda x: x if x == '1' else 0)
df.groupby('comparison2')['ACTIVATED_IND'].agg(['count', 'mean'])
summary_2 = df.groupby('comparison2')['ACTIVATED_IND'].agg(['count', 'mean']).reset_index()
```

Figure 43: Code of customised customer segment for 57.98% conversion propensity