

https://dspace.org/

The impact of time-based factors on outbound life insurance sales in call centres

Schneider, David

2021
https://hdl.handle.net/10539/33362

Downloaded from a DSpace repository. https://dspace.org/



The impact of time-based factors on outbound life insurance sales in call centres

David Schneider

(Student number: 0709524F)

School of Mechanical, Industrial and Aeronautical Engineering

University of the Witwatersrand

Johannesburg, South Africa.

Supervisors:

Lorrainne Doherty

A Research Project **proposal** submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in fulfilment of the requirements for the degree of Masters in Engineering.

4 October 2021

Johannesburg 2020

DECLARATION WITH TASK SUBMITTED FOR ASSESSMENT

I, the undersigned, am registered for the course **MECN7018 Research Project**, in the year 2020. I herewith submit the following **Research Project Report** in fulfilment of the requirements of the above course.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong;
- I confirm that the work submitted herewith for assessment in the above course is my own unaided work except where I have explicitly stated otherwise;
- This task has not been submitted before, either individually or jointly, for any course requirement, examination or degree at this or any other tertiary institution;
- I have followed the required conventions in referencing the thoughts and ideas of others;
- I understand that the University of the Witwatersrand may take disciplinary action against me if it can be shown that this task is not our own unaided work or that I have failed to acknowledge the sources of the ideas or words in my writing in this task.

Student Name	Student Number	Signature
D. Schneider	0709524F	

ABSTRACT

The end-to-end value chain for outbound life insurance sales in call centres is one that has in-built time-based factors that can affect its overall efficiency, impacting cost efficiency for the insurer.

This research considers two such time-based relationships. The first is the difference in days between an outbound insurance sales call and a customer's salary date and its relationship to the likelihood of the customer purchasing the offered insurance product. The second is the difference in days between the date of the outbound insurance sales call and the date of the first attempt to collect the policy's monthly premium, and its relationship to the likelihood of successfully collecting the premium.

Historical leads of a life insurer's funeral plan book are analysed, and two clear patterns are observed. With regards to likelihood to take up an insurance offer versus days between salary and date of call, there is an initial low average rate of conversion of leads to sales for the first few days of a month. This is followed by a steep spike in conversions with a gradual decline as the number of days between date of call and date of salary increases. In the case of likelihood to successfully collect the first premium, a more linear relationship is seen with average first debit collection rates at their highest when the difference in days between date of call and date of debit orders are at a minimum, with the average rate of collection declining as this difference increases. In both relationships, higher income customers appear to be marginally less sensitive to these time-bound factors.

Recommendations are made to maximise the average monthly sales-to-lead conversions throughout a given month and modifications are proposed for the sales process using the choice architecture concepts to improve the average first debit collection rates.

KEY WORDS: Life insurance, call centres, time-based factors, consumer behaviour

DEDICATION

To my late father, Brian Schneider,
who instilled in me the value of education

ACKNOWLEDGEMENTS

Throughout this research, it has become increasingly clear that abundant gratitude and recognition is owed to those who have made this endeavour possible.

Firstly, and always, to G-d for all that He does for me and my family and for the wisdom and insight that He bestows.

To my wife, Danielle, and son, Elad Chanan (who was born during the course of this research), for all the unceasing encouragement, deserved tough love and unconditional love without which this undertaking could not have been achieved. I would also like to thank my mother, Beverley, for her unwavering belief in my ability and for giving me the confidence that I could achieve this goal. Thank you for all that you do and sacrifice for us all every day. I love you all.

My employer for supporting me throughout my journey of further education, in guidance, expertise, and resources (both in time and financial support). This research would have unlikely been possible without even one of these supports, for which I am extremely grateful. I thank you for this opportunity and the many others with which you continue to provide me.

Finally, but most certainly not least, I would like to thank my supervisor, Mrs. Lorraine Doherty, whose enthusiasm, passion, and encouragement for this research was only surpassed by her immense patience, support and kindness throughout. I cannot adequately express how much you have contributed to both this work and me personally.

TABLE OF CONTENTS

DECLARATION ..... **i**
ABSTRACT ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENTS ..... **iv**
TABLE OF CONTENTS ..... **v**
LIST OF FIGURES ..... **viii**
LIST OF TABLES ..... **x**
NOMENCLATURE ..... **xi**
1. CHAPTER 1: INTRODUCTION **1**
1.1 Background **1**
1.1.1 Demand for life insurance 1
1.1.2 The direct insurer value chain 1
1.1.3 The insurer 3
1.2 Problem statement/Motivation **5**
1.2.1 The cost of inefficiency 5
1.2.2 Timing is everything 6
1.2.3 The problem of collections 6
1.2.4 Opportunity 7
1.3 Research Questions **8**
1.4 Research Objectives **8**
1.5 Summary of Research Method **9**
1.6 Scope/Delimitations **10**
1.7 Outline of Chapters **11**
2. CHAPTER 2: LITERATURE REVIEW **13**
2.1 Introduction 13
2.2 Time-based factors affecting consumer behaviour 13
2.3 Factor improving likelihood of sales in retail environments 14
2.4 Consumer behaviour in the service industry 16
2.5 Consumer behaviour in Insurance 18
2.6 Call centres: management and strategies 20
2.7 Importance of the broader value chain and first debit rates 21
2.8 Techniques for predicting consumer behaviour 21
2.9 Light Gradient Boosting Machine 23
2.10 Conceptual framework 25
3. CHAPTER 3: RESEARCH METHODOLOGY **27**

3.1 Introduction **27**
3.2 Research Design and Methodology **27**
3.3 Research instruments **28**
3.4 Managing confounding factors **28**
3.5 Sampling 29
3.5.1 Datasets 29
3.5.2 Interviews 32
3.5.3 Data Collection, analysis procedures and instrumentation 32
3.6 Developing a predictive model 34
3.6.1 Pre-processing 35
3.6.2 Training, validation and test data 37
3.6.3 Correlation of variables 38
3.6.4 Fitting and evaluating the models 39
3.7 Validity and reliability 40
3.7.1 Predictive models 40
3.7.2 Survey 40
3.8 Ethical considerations 41
3.8.1 Ethical clearance 41
3.8.2 Permission for use of data 41
3.8.2 Confidentiality of personal data 41
4. CHAPTER 4: ANALYSIS AND INTERPRETATION OF RESULTS 42
4.1 Introduction **42**
4.2 Sample Data **42**
4.2.1 Salary deposits and debit order data 42
4.2.2 Sampling of sales records 45
4.2.3 Survey results 47
4.2.4 Variables considered in predictive model analysis 51
4.2.5 Analysis of relationships 54
4.3 Implementation of predictive models **70**
4.3.1 Correlation of variables 70
4.4 Model Results **73**
4.4.1 Leads conversion model results 73
4.4.2 Feature ranking of leads conversion model 73
4.4.3 First debit collection model results 75
4.4.4 Feature ranking of first debit collection model 76
5. CHAPTER 5: DISCUSSION **78**
5.1 Introduction 78

5.2 Importance of number of days between salary receipt and date of call 78
5.3 The impact of agent behaviour 79
5.4 Importance of customer income on first debit collection 80
5.5 Importance of monthly premium 80
5.6 Importance of time-based factors in first debit collection 81
6. CHAPTER 6: CONCLUSIONS 82
6.1 Introduction 82
6.2 The impact of time-based factors in the outbound insurance value chain 82
6.3 The impact of days difference between salary receipt and date of sales call 83
6.4 Validation of first debit rate as a metric for collections analysts 84
6.5 The impact of days difference between date of sale and date of first debit collection 85
6.6 The use of LightGBM-based predictive models to validate the impact of time-based factors 85
6.7 The impact of monthly income 86
6.8 Next steps arising for improving the efficiency of the outbound sales process for life insurance sales 87
7. CHAPTER 7: RECOMMENDATIONS 89
7.1 Introduction 89
7.2 Recommendations for management of call centre and agents 89
7.3 Recommendations for the supply of monthly leads 92
7.4 Recommendations for enhancements to collections processes 94
7.5 Combined improvements to the value chain 95
7.6 Recommendations for future work 95
8. REFERENCES 98
APPENDIX A: SURVEY FOR COLLECTION ANALYSTS 102
APPENDIX B: MODEL CODE 105

LIST OF FIGURES

Figure 1: Value chain of the direct insurer analysed in this study....1
Figure 2: Simplified view of time-based variability in customer interactions for the direct insurer studied....6
Figure 3: Conceptual framework of literature considered....25
Figure 4: Process followed for implementation of models and feature ranking....35
Figure 5: Distribution of salary dates for the insurer's funeral plan book....43
Figure 6: Distribution of debit days for the funeral plan book....43
Figure 7: Tenure of analysts surveyed....47
Figure 8: Number of years of experience within sample with life insurance analytics....48
Figure 9: Position held within Analytics community by surveyed candidates....48
Figure 10: Ranking of efficacy of variables in predicting off-movements in new outbound life insurance sales by interview candidates....50
Figure 11: Overall conversion rate versus day since salary deposit (overall base)....54
Figure 12: Relative performance of different elapsed times since salary receipt....56
Figure 13: Conversion rate versus day since salary deposit for R0 - R5 000 monthly income....57
Figure 14: Conversion rate versus day since salary deposit for R5000 – R10 000 monthly income....58
Figure 15: Conversion rate versus day since salary deposit for R10 000 – R15 000 monthly income....58
Figure 16: Conversion rate versus day since salary deposit for R15 000 – R20 000 monthly income....59
Figure 17: Conversion rate versus day since salary deposit for R20 000 – R25 000 monthly income....60
Figure 18: Conversion rate versus day since salary deposit for R25 000 – R30 000 monthly income....60
Figure 19: Conversion rate versus day since salary deposit for R30 000+ monthly income....61
Figure 20: Distribution of income bands within the set of call centre leads analysed....62
Figure 21: First debit rate versus days between sale and date of first debit for overall base....63
Figure 22: First debit rate versus days between sale and date of first debit for monthly income between R0 and R5 000....64
Figure 23: First debit rate versus days between sale and date of first debit for monthly income between R5 000 and R10 000....65
Figure 24: First debit rate versus days between sale and date of first debit for monthly income between R10 000 and R15 000....66
Figure 25: First debit rate versus days between sale and date of first debit for monthly income between R15 000 and R20 000....66
Figure 26: First debit rate versus days between sale and date of first debit for monthly income between R20 000 and R25 000....67
Figure 27: First debit rate versus days between sale and date of first debit for monthly income between R25 000 and R30 000....67
Figure 28: First debit rate versus days between sale and date of first debit for monthly income greater than R30 000....68
Figure 29: correlation matrix of features considered in the leads conversion model....70

Figure 30: correlation matrix of variables for debit order analysis model 72
Figure 31: Ranking of feature importance for lead conversion model generated by the model code 74
Figure 32: Ranking of feature importance for first debit model generated by the model code 77

LIST OF TABLES

Table 1: Sampling factors to be normalised 30
Table 2: Required features for datasets.....31
Table 3: Selection criteria of interviewees.....32
Table 4: Relationships to be determined.....33
Table 5: Variables considered.....51

NOMENCLATURE

- **EFB**: Exclusive Feature Bundling
- **GBDT**: Gradient Boosting Decision Tree
- **GOSS**: Gradient-Biased One-Sided Sampling
- **FSP**: Financial Services Provider
- **LightGBM**: Light Gradient Boosting Machine

1. CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 Demand for life insurance

Life insurance is a critical part of risk management in financial planning (Parrish, 2018), yet has a very low rate of penetration in markets across the globe, especially in Africa. Despite this critical role in personal finance, the penetration of life insurance in South Africa is estimated to be less than 12% (van der Merwe & Gerin, 2018). In order to improve financial inclusion (and to capitalise on this unsaturated market), some insurers are moving towards the use of direct distribution models (PWC, 2016) These models often include the use of outbound call centres (Binder & Mußhoff, 2017) to make telephonic sales of life insurance products.

1.1.2 The direct insurer value chain

This research considers the value chain (illustrated in Figure 1) of a particular South African direct insurer operated by an integrated financial services provider (FSP).

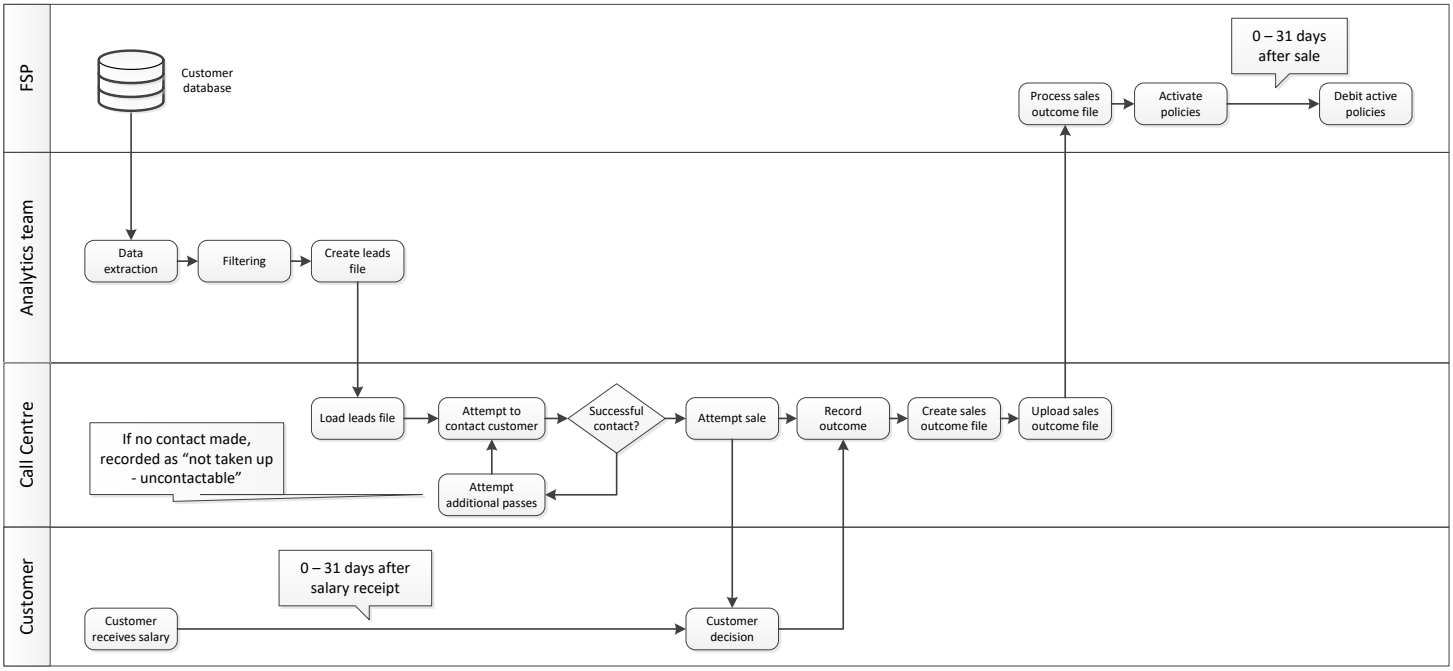

Figure 1: Value chain of the direct insurer analysed in this study

The call centre sits at the heart of the distribution but relies on other supporting nodes to ensure continuous production and ultimate sales quality.

The call centre is supplied by leads by the insurer. These leads are details of potential customers that have been identified as having the particular propensity to purchase insurance products. This propensity that qualifies the leads is typically identified by the insurer from the analysis of internal company data sources, or simply through interest expressed by the customer (for example, a customer may request for the insurer to call them back for a quotation). Whichever way the leads are identified, only customers who have provided consent to be contacted for marketing purposes may be contacted by a service provider, in accordance with legislation and the internal policies of the FSP.

The leads provided to the call centres are then worked by the call centre agents. In this context "work" means the agent draws from the supplied leads and attempts to contact one of these supplied leads. Once contact has been made, the agent then pitches the offering and attempts to make a sale to the customer. If the sale is successfully made, the lead is considered to be a "successful lead conversion". If the lead does not convert, it could be due either to inability to successfully contact the lead or lack of interest from the customer.

Once successful sales from the call centre have been activated by the insurer, they are loaded on to a policy administration system and then require monthly billing to remain active.

As such, the success of the value chain can be measured using two key metrics: sales conversion rate and first debit collection rate. These rates have the same basis in customer acquisition – that is, for a particular interaction, the number of successful interactions divided by total number of interactions (Kumar & Reinartz, 2006).

In a similar vein to the basis provided by Kumar and Reinartz (2006), the insurer defines these two metrics as follows:

- sales conversion rate – defined for a particular campaign as number of successful sales over the total number of leads supplied for a particular campaign, or:

$$Conversion\ (\%) = \frac{sales\ from\ supplied\ leads}{Total\ leads\ supplied} \quad (1)$$

- First debit collection rate – defined as number of sales from a campaign that have their full premium successfully collected the first month the customer is debited, or:

$$First\ debit\ rate\ (\%) = \frac{Number\ of\ polices\ with\ first\ premium\ successfully\ collected}{Total\ number\ of\ monthly\ policies\ debited\ for\ the\ first\ time} \quad (2)$$

Thus, it can be seen that a critical driver of the success for the overall value chain (in terms of maximising revenue while minimising acquisition costs) is the underlying efficiency of the supplied campaigns (containing leads). The efficiency of the campaigns (sales to leads) can then be managed by reducing wasted leads, where waste refers to:

- Leads that do not convert to sales, due to either:
  - Unsuccessful sale
  - Uncontactable (out of the scope of this investigation)
- Sales that do not successfully debit – also lead to multiple forms of waste, including:
  - Increased expenditure on collections processes (particularly if multiple attempts need to be made to collect the required premium)
  - If policies lapse, used acquisition costs can be considered waste

1.1.3 The insurer

The particular insurer considered by this study is a life insurance division within a South African bank, primarily making use of direct distribution strategies. This direct

distribution includes both outbound sales (the insurer initiates contact with a customer in order to offer an insurance product) and inbound sales (a customer initiates contact with the bank – or directly with the insurer – in order to purchase an insurance product).

Outbound sales of stand-alone products include both non-underwritten insurance plans (for example, funeral plans) and underwritten life insurance policies (for example, death, disability and critical illness cover). In this context underwriting refers to the characterisation of a client's risk profile in order to determine eligibility for insurance, and, if eligible, personalised pricing for their individual risk.

The insurer generates sets of leads by extracting and filtering profiles of customers who have consented to marketing from existing datasets within the bank. These internal sources of data are refreshed regularly and are considered accurate. As the bank has many functional areas, the combined data sets are rich in the total number of features available for selecting leads (examples of these features include information regarding income, age and gender, among many others).

The sets of leads that are ultimately extracted are sent to the call centre and loaded as bulk files at the beginning of each month. The volume of leads contained in each set sent is calculated to be sufficient for the entirety of the upcoming calendar month. These leads are then worked in no particular order by the call centre agents and multiple attempts to contact the customer (referred to as passes) are usually allowed. Three passes are typically allowed for a lead by the insurer (the allowed number may vary from campaign to campaign).

A file containing successful sales is returned to the insurer on a daily basis to be uploaded and activated by the insurer so that billing can take place on the debit date (specified by the customer).

1.2 Problem statement/Motivation

In order to reduce operational complexity, the insurance company compiles a single bulk leads file to be worked by the outbound call centre. This file is generated and submitted on a monthly basis. The call centre then draws down from this file, selecting leads in no particular order and supplies the leads to available agents. This method of loading and drawing leads can result in various inefficiencies:

1.2.1 The cost of inefficiency

Leads need to be worked as efficiently as possible for two reasons:

- the source of these leads is limited as only existing bank customers are allowed to be contacted by the call centres for insurance sales. Once one of these customers have been contacted by an area in the bank, there is usually a "cooling-off" period (30 days) before they can be contacted again by a different area.
- there is an acquisition cost associated with working the leads

Any leads that do not result in a sale can be considered "waste" leads and as a result, they increase the insurer's cost base without improving the ratio of cost to income – a key metric for the business.

Another potential source of process inefficiency is the variability (Bicheno & Holweg, 2016) inherent in the process. The two sources of variation specifically considered are:

- the variation in the time between calling a customer and the date of receipt of salary ("payday")
- the variation in time between a successful sale and first debit (the debit order date is chosen by the customer, and as a result the first debit may only occur up to a full month after the sale is actually made)

These sources of variation are illustrated in Figure 1 (above) and Figure 2 (below).



Figure 2: Simplified view of time-based variability in customer interactions for the direct insurer studied

On paper (or rather, on process map), the impact of these sources of variability is not obvious. They may, however, have significant impact on the overall efficiency of the end-to-end process.

1.2.2 Timing is everything

It has been suggested by some of the insurer's sales managers that the best sales days are those on, or shortly following the most popular pay days (usually, the 25th or end of the month). They further suggest this trend exists due to a psychological aspect of consumer behaviour: when customers are feeling short on money (usually further away from pay day), they are less likely to purchase new products due to the anxiety surrounding their cash flow (or lack thereof). Conversely, when greater funds are in their account (usually on pay day, or shortly thereafter), they relax and are more open to these offers, as the bulk of their discretionary spending has not been done yet. The idea that consumers have a higher affinity to promotional offers shortly after payday is supported by research (Mishra, Mishra, & Nayakankuppam, 2010; Stephens, 2002), although the studies do have significant gaps that are relevant to the context of outbound insurance sales (this is discussed in detail in Chapter 2).

1.2.3 The problem of collections

Even if a supplied lead results in a sale, the lead can still become a wasted lead through unsuccessful premium collections. If the premiums of a newly sold policy are not successfully collected, the policy lapses, resulting in a wasted lead (as well as the associated costs of acquisition and attempted collection which drive up the cost base without bringing in any income)

To this end, the insurer's collections and analytics teams track a variety of metrics in order to proactively identify at-risk policies. These metrics are monitored as key indicators of whether a policy will experience an early lapse (and thus become loss-making). One such metric tracked is the aforementioned first debit collection rate (the measure of how many policies had their first premium successfully collected).

Even though first debit collection rate can intuitively be seen to be a key piece of the outbound insurance value chain, its relevance does require verification.

1.2.4 Opportunity

An opportunity therefore presents itself to improve the efficiency of the sales process (and by extension, improve the insurer's cost-to-income ratio) by modifying the supply of leads and the timing of when they are worked, such that:

- sales activities are better aligned to customers' paydays in an attempt to drive up the sales conversion ratio
- Sales processes are geared towards improving first debit rate

1.3 Research Questions

The critical research question that this research will attempt to answer is:
"What is the impact of time-based factors on outbound life insurance sales made through call centres?"

1.4 Research Objectives

The overall objective of this study is to assess the impact of particular time-based factors on outbound insurance sales.

This will be done by:

- Determine the relationship between the proximity of an outbound insurance sales call to a customer's payday and the likelihood of a customer to take up an insurance offer from the call centre
- Establish the extent of the relationship between the time from the insurance sale through the call centre to debit date and the likelihood of a policy to successfully collect on the first attempted debit
- Propose recommendations for making cost improvements to the end-end insurance call centre sales process through the use of these relationships.

1.5 Summary of Research Method

In order to assess what the impact of time-based factors is on outbound life insurance sales made through call centres, the following steps were followed:

1. The area of research was identified and a problem statement, set of objectives and the scope of the research were defined.
2. A review of the literature pertinent to the topic and the related methodologies required for the study was conducted. This literature was used to create a framework with which to base the analysis and to inform best practice for the implementation of the methodologies.
3. A research proposal, including submissions for ethics approval and permission for use of data from the FSP, was submitted and presented in order to obtain approval to proceed with the study.
4. Amendments were made to the research proposal per the feedback from the examiners.
5. Once final approval for the proposal was obtained, a survey was conducted in order to verify some of the underlying assumptions of the problem statement, in line with the relevant research objectives.
6. The results of this survey were aggregated and analysed to verify the assumptions of the relevant research objectives
7. Further literature was considered, reviewed and incorporated into the body of the research paper.
8. Data was extracted from the FSPs databases of sales records for outbound funeral sales made through the FSP's call centres
9. The books of insurance products under analysis were characterised and an initial analysis was done on the relationship between time-based factors and outbound insurance sales.
10. In order to determine which variables or factors have the largest impact on outbound insurance sales, predictive models were developed in order to output a ranking of the most impactful model features (that is, factors)
11. This output was then analysed and discussed.

12. Conclusions with regards to the objectives of this study were made on the basis of this analysis
13. Finally, recommendations were proposed on the basis of these conclusions

It is noted that, in particular, steps 7-13 were iterative – as they were completed and new opportunities or challenges arose, further research and refinement of modelling was conducted until acceptable levels of reliability were achieved.

1.6 Scope/Delimitations

The scope of this research was limited to a single funeral book maintained by the insurer. Therefore, this research only considers:

- non-underwritten funeral plans (as opposed to underwritten life insurance products or other non-underwritten life insurance products such as accidental death cover)
- The study focused on the relationship between time-based factors that may have an impact on the overall value chain for outbound insurance sales through call centres. (other influences such as the macroeconomic environment, demographics and personal income are not the focus as the insurer has little to no control over these factors).
- The behaviour observed is only for a single call centre conducting outbound sales (inbound sales are not considered).

1.7 Outline of Chapters

This research consists of seven chapters, references and appendices. A brief overview of the chapters is as follows:

- **Chapter one** provides background regarding the value chain of outbound life insurance sales through call centres and the insurer. The problem statement, research question and research objectives are provided, as well as a summary of the research method and scope of the research.

- **Chapter two** is a review of the key literature considered regarding consumer behaviour with regards to services, insurance and the analytical methods and tools used to determine relationships between consumption and particular time factors.

- **Chapter three** provides the detail of the research design and the methodology followed, including the required research instruments, treatment of possible confounding factors, approach to data sampling and how the data was collected and analysed. An overview of the predictive modelling process is also provided. Finally, considerations regarding reliability and validation of the research are provided, as well as ethical considerations.

- **Chapter four** presents the results from the research conducted, including a characterisation of the book of insurance policies analysed and detail pertaining to the sample data of the sales records that was obtained for the study. It also presents the results of the survey conducted, the initial analysis of the impact of the studied time-based factors and, finally, the details of the practical implementation of the predictive models and their performance.

- **Chapter five** discusses the results obtained from the analyses conducted with regards to the objectives of the research

- **Chapter six** presents the conclusions drawn from the study with regards to the overall objectives of the research

- **Chapter seven** details the recommendations are made for the FSP and for further possible research

- **References** to existing literature used throughout this research

- **Appendices:**
    - **Appendix A:** Survey questionnaire
    - **Appendix B:** Full code listing of model implementation

2. CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This section presents a review of literature pertaining to understanding and predicting the factors that drive consumer behaviour related to purchasing services, with focus on life insurance as a financial services product. In particular, this review focuses on:

- understanding the previous research of time-based factors
- improving likelihood of sales in sales environments
- consumer behaviour in the service industry
- Consumer behaviour in the life insurance industry
- call centre management theory
- the importance of evaluating an end-end value chain in the context of call centre-based sales channels
- finally, literature on predicting customer purchase decisions and appropriate techniques and tools

2.2 Time-based factors affecting consumer behaviour

While much research has been published showing empirical evidence against the Lifecycle/Permanent income hypothesis (LCPIH), of note are the various works that describe how the spend behaviour of consumers is sensitive to time-based factors and triggers. Shapiro and Slemrod (1993) provide evidence that in practice consumers do respond to the timing of income, although their findings are inconclusive with regards to the role played by liquidity (perceived or otherwise).

Zeldes (1989) had previously demonstrated that consumption is impacted by the consumer's ability to borrow against future income, thus providing a link between liquidity and consumer behaviour. This concept is developed further by Stephens (2002) in an attempt to determine the role of the monthly receipt of salary. Stephens (2002) finds the spend behaviour of consumers is sensitive to this monthly inflow of salary. Around the same time that Stephens (2002) was developing this work, Huffman

and Barenstein (2005) put forward an analysis based on the same data set. While their work similarly concluded that consumer spending decreases between paydays, their approach focused more on aspects of behavioural economics, specifically that of "mental accounting" as developed by Thaler (2008). The difference in this approach suggests that it is not just the actual liquidity that consumers experience (that is, how much money they actually have on hand in an account at a particular point in time) that influence decisions around making purchases, but also their perception of their personal liquidity.

2.3 Factor improving likelihood of sales in retail environments

Mishra et al. (2010) builds on this body of research with the goal of showing how consumer preferences are influenced by the amount of time that has elapsed since their last salary was received. Through their work, they show that consumer's preference to promotional messages changes cyclically within the month as a function of proximity to salary receipt. Where they differ from Huffman, Barenstein (2005) and Zeldes (1989), is they assert the concept of mental accounting and liquidity, perceived or otherwise, (termed "liquidity-alone accounts") cannot account for the patterns of purchasing behaviour observed in their research. Instead they claim to demonstrate a clear improvement in the uptake of promotions of new products closer to receipt of salary, which they attribute to regulatory focus theory - that is, that decision-making and the assignment of value can be affected by the framing of a scenario or given choice due to the way it resonates with the goals pursued by the person (Avnet & Higgins, 2003).

However, several limitations are apparent in this study, some of which may prove challenging in applying the findings to outbound sales processes and financial services:

- *Direction of sale:* The study revolves around scenarios whereby customers choose to go out and purchase an item, thus assuming an "inbound" (that is customer-initiated) scenario. In "outbound" (that is, business-initiated) scenarios where a particular offering is promoted, different behavioural aspects may come in to play.

- *Transactional Frequency:* This study only considers single-purchase items (that is, in order to purchase an item, payment has to be made once, upfront), such as toothpaste, whereas products with recurring monthly fees, such as financial services, are not considered. There are two implications to this: firstly, a different set of biases may be at play for these sorts of purchases. Additionally, it doesn't consider the ongoing persistence of using the product (for example, maintaining an insurance policy)
- *Absolute pricing and value:* The items used in the study for potential purchases have relatively low prices and seem to ignore possible conditions where the purchase could be high. Additionally, there is little consequence to the decision (purchasing the wrong kind of toothpaste in contrast to purchasing the wrong – or no – life insurance)
- *Product type:* the study uses fairly arbitrary goods for the participants to purchase. Different behaviours and decision-making factors may come in to play when making more purposeful purchases such as insurance products. Even within the insurance context there are different needs at play (for example, potential customers in different contexts may have differing preferences for various insurance benefits such as life cover or disability).
- *Introduction of priming effects:* As part of the methodology, it appears that researchers inadvertently introduce an element of priming to initial study (despite considering and factoring in priming in later studies) which is likely to have significant impact on initial and ongoing decision-making throughout the study (Dolan et al., 2012)
- *Period of study*: The study only considers a single month. This relatively short period of time allows for the influence of seasonal consumer patterns and other external sources of noise to skew results in the study.
- *Known observation:* Participants were aware their purchasing decisions were to be monitored, which may influence their purchasing decisions

Finally, the arguments presented against the sole influence of liquidity-alone account theory assume that consumers are constantly aware of their precise liquidity (including the occurrence of bill payments that may be automatically debited). However, this may not be the case with regards to perceived liquidity (the initial position of the authors)

– this being the broader awareness of the consumer with respect to relative levels of liquidity ("funds are low" versus "there's money in the account"). Given this, salary receipt may still be a prominent predictor.

2.4 Consumer behaviour in the service industry

Some of the above limitations are highlighted by McKechnie (1992) in an attempt to understand the then available state of research relating to consumer behaviour in the hopes of creating a framework that catered for generic services marketing. Even though more recent research has been conducted since (Chaudhary & Kaur, 2016), McKechnie's framework provides a comprehensive and foundational basis for understanding consumer behaviour as it pertains to the service industry. This framework would then be used to better understand consumer behaviour as it specifically pertained to personal and corporate financial services.

McKechnie analyses the development of the literature on service marketing (Gummesson, 1979; Shostack, 1977) and concludes that development of this line of research is based on three key assumptions:

- Factors exist which distinguish services from goods
- These factors impose unique challenges for service marketers not faced by goods marketers.
- Service marketing requires a specific approach to marketing, distinct from that of marketing physical goods.

McKechnie (1992) details that services are distinguished from typical consumer goods on several fronts, including:

- intangibility (the inability to evaluate something physical)
- inseparability (the service is simultaneously rendered by supplier and consumed by the purchaser)
- perishability (services cannot be stored to be consumed for use later)

- heterogeneity (potential for higher variability due to the often-personal nature of the interactions).

While these aspects do not create perfect inverses between services and goods, they do result in challenges in the marketing of services which are unique to that of physical goods. These challenges will then have an impact on consumer behaviour for services in general, as well as financial services.

McKechnie (1992) further highlights that an import factor influencing the consumption of services is the actual interaction between the buyer and the supplier. In the case of financial services, this would be an agent or representative of the financial services provider. A particularly illustrative finding considered is the observation that after actual pricing of financial services, the interface with customers (that is, the actual contact with sales channel or agent) is the next most important aspect of marketing and the area with second highest potential for improving overall performance of the FSP.

While McKechnie's research has been discussed up to this point as it provides fundamentals of understanding consumer behaviour in the service industry, at least with regards to consumption of services in the general sense, similar findings are supported by more modern studies (Chaudhary & Kaur, 2016) show that these intangible customer experience-related factors are pertinent to the purchase of insurance products.. It is important to highlight again that when purchasing insurance products (and to varying extents, other financial services), a customer receives no physical or immediately tangible benefit (and in fact, it may never be felt in their lifetime – which is usually not the case with other classes of financial products).

Physical goods can be immediately utilised or consumed, providing immediate satisfaction of needs (at least, theoretically). Besides the obvious benefit of this immediate consumption, consumers are also able to evaluate the quality of a given product before purchase (that is, by touch, feel, previous consumption, etc.) when making a decision to purchase. When considering financial products - particularly insurance, where the utilisation of the product may only occur in the future (if ever, taking in to account the nature of the particular benefit. In the case of benefits covering mortality – as in the case of this research which considers funeral insurance -, the

policyholder will never see direct enjoyment of the product. Additionally, there is the possibility of a claim being declined, the policy lapsing, or the policy being cancelled), the consumer has few opportunities to evaluate the quality of the product during the decision to purchase aside from the interface with the sales agent or broker.

Assuming the fundamental aspects of a financial services offering, such as product features and pricing, cannot be changed (or at the very least, there is insufficient business appetite to do so), an area of interest for improving the efficiency of the sales value chain would be improving the actual point of interaction between a prospective customer and the FSP (Chaudhary & Kaur, 2016; Mckechnie, 1992) (for example, an agent in a call centre and the associated experience).

2.5 Consumer behaviour in Insurance

It has been observed that an oft-neglected behavioural aspect relating to consumer decision-making, particularly in the context of insurance purchases, is that of a concept referred to as *preference formation*. This concept relates to a person being influenced by non-financial attributes when assessing the overall value of a purchase (or subscribing to a particular service such as in the case of choosing to purchase insurance) which may include emotional aspects such as personal concerns and worries that extrinsic to the actual nature or quality of the item or service in question (Kunreuther & Pauly, 2005).

Kunreuther and Pauly (2005) further point out there is considerable evidence that consumers make purchase decisions by evaluating how the purchase decision will impact their ultimate financial position relative to their current financial position. It follows logically that as consumers' available income, and hence, financial position, tends to fluctuate in a largely cyclical fashion every month (at least in the case of regularly employed individuals – although this still holds true in the case of commission-based or freelance workers, albeit not necessarily in a regular or cyclical fashion), there should follow some form of regular, cyclical influence to consumer behaviour and decision making (Kahneman & Tversky, 1979).

Kunreuther and Pauly (2005) further assert that given the many and varied decisions and concerns that a person has to contend with on an ongoing basis as part of their daily lives, it is reasonable to assume that various heuristics, or mental shortcuts, are used by consumers when making purchase decisions. These mental shortcuts may extend to consumers making negative decisions specifically in the case of insurance purchases. These shortcuts may be triggered by feeling constrained by budgetary considerations such as personal income and current liquidity. This assertation accords with previously noted findings of Stephens (2002), as well as those of Huffman and Barenstein (2005)

Finally, Kunreuther and Pauly also cite the findings of Thaler (1999) that individuals may not purchase insurance due to prior allocation of expenditure in to other mental accounts. They further note that in the instance of insurance, consumers may mentally categorise various form of insurance and services under a single mental account (that is, simply as "insurance" or "protection services") or category and thus may be hesitant to increase expenditure on an existing category when feeling constrained by monthly cashflow considerations. As an illustrative example, individuals may already have short-term insurance such as car insurance and household insurance, but view funeral insurance (a long-term, life insurance product) as increasing their overall "protection" spend - even though the insured risk is fundamentally different and unrelated to their existing cover. It is also noted that this constraint may also be prevalent in higher income individuals. Further, lower income families are found to be more sensitive to cyclical budget constraints with shorter time frames and thus more sensitive to specific weekly and monthly expenditure (Thaler, 1999). It is reasonable to assume that this sensitivity also extends to weekly and monthly inflows due to wages and salaries.

Finally, It has also been demonstrated that individuals with sufficient income and assets to purchase insurance may choose not to due to these same mental budgeting heuristics and sensitivities(Bundorf & Pauly, 2002).

These various finding suggest that the same general consumer biases and decision-making processes that exist for consumption of generic goods and services hold true in the case of insurance consumption, and importantly, that these biases hold true across consumers with various levels of income (and not just in lower-income markets).

2.6 Call centres: management and strategies

While much research has been undertaken regarding call centres, the bulk of the research surveyed tends to focus on two broad yet distinct domains. The first of these domains is that of the management of the human capital and elements of industrial relations within call centres (broadly referred to here as the Human Resources approach). This approach is particularly prevalent in research within the context of South African call centres (Benner, Lewis, & Omar, 2007). Russel (2008) provides an extensive view of this domain and considers various theoretical viewpoints including high-performance work systems, processes relating to labour, perspectives relating to human resources, studies on gender and the impact of globalisation. The main thrust of this area of research tends to focus more practical aspects relating to staffing and working conditions (Kinnie, Hutchinson, & Purcell, 2000; van den Broek, Barnes, & Townsend, 2008) or more performance-oriented considerations such as high performance work systems (Batt, 2002). While the Human Resources approach is critical in optimising sustainability and reliability (in the form of employee retention and skills development), these aspects are largely outside the scope of this research.

The second domain is that of analytically understanding the performance of call centres (broadly referred to here as the Quantitative approach). The Quantitative approach addresses a broad range of topics such as the application of queueing theory (Brown & Gans, 2005) and quantitatively modelling customer satisfaction (Feinberg, Kim, & Hokama, 2015; Joia, Oliveira, & Oliveira, 2010). The research in this domain tends to heavily focus on the context of inbound service centres, with less focus both on the sales context and outbound call centre context. Given the bulk of call volumes are inbound in nature (78% in the South African context (Benner et al., 2007)), this can be expected but does create a gap in the available body of knowledge

Finally, and of greater note, it is argued the majority of this research fails to consider the call centre as part of a larger value chain or system. While these analyses are often rigorous, this narrow view of modelling and understanding of the greater tele-sales value chain can create blind spots with regards to inefficiencies both upstream

(management of leads) and downstream (management of the customer interaction after the phone has been put down).

2.7 Importance of the broader value chain and first debit rates

The use of a production facility as a model for applying lean practices to service industries, in particular FSPs, has been previously suggested by Swank (Swank, 2003). This approach can often be useful, particularly for multi-stage application processes (like insurance which often requires underwriting and other evaluations between initial application and the start of the contract), but often fails to make a critical distinction. In the context of production of physical goods, once a unit has been produced and sold there is generally little to no ongoing relationship between the supplier and the consumer – the consumer, however, remains satisfied with the ultimate flow of value. Within financial services, there is often a monthly fee collected by the FSP, which is a requirement for ongoing service (or premiums and coverage in the context of insurance). If these premiums are not successfully collected, the service terminates, resulting in the consumer not having their original needs met. It therefore appears that a critical element of the value chain is ignored by not including collections processes. Ignoring this element of the value chain (and resulting metrics) can also result in lower quality sales (that is, sales that do not successfully collect, and thus do not activate) and lower longer term profitability (Hauser & Katz, 1998)

2.8 Techniques for predicting consumer behaviour

In order to predict the likelihood of a particular customer behaviour (such as the purchase/not purchase or debit successfully/unsuccessfully of life insurance products sold on an outbound basis), several techniques can be applied. These techniques are not specific to the case of outbound life insurance sales, but rather, are techniques that can be used to predict behaviours in general using software-based supervised learning. Various supervised learning techniques such as support vector machines and artificial neural networks may be used, but in the scenario of modelling multi-stage leads-based sales processes, logistic regression is considered one appropriate approach. Reasons for

this include ease of implementation and the ability to directly produce probabilities from a variety of inputs (Monat, 2011) for predicted lead conversions (as opposed to a more abstract score or measure) that are more easily understood by non-technical members of the organisation (Yan, Gong, Sun, Huang, & Chu, 2015). In cases where well-structured commercial data (as opposed to noisier, unstructured data gathered from online scraping or other common sources), the use of either decision trees (and by extension, random forest models and other tree-based models) or logistic regression are appropriate, with both showing similar predictive performance (D'Haen, Van den Poel, & Thorleuchter, 2012).

In the context of machine learning, decision trees are a type of model that uses a branching tree structure to model possible paths to an end outcome (that is, identify a particular class or value) using series of if/then-style decisions based on the model's features. Using the available training data, decision tree algorithms recursively add nodes to the tree by partitioning the data using "tests" (that is, the if/then-style decisions) until each partition (or "leaf") contains only a single class or value. When new data is then provided to the tree for classification (or regression), the outcome is determined by going through the various nodes of the trees and using the tests at each node junction to determine which branch to then follow until a final classification (or value in the case of regression) is reached (Muller & Guido, 2017).

Although, despite being robust in terms of the data they can handle and the ease in which they can be understood and interpreted, decision trees do suffer from some drawbacks. The most notorious of these drawbacks is that standalone trees are prone to being overfitted to their training data and thus may struggle to accurately predict outcomes for previously unseen data (Grus, 2015). Various techniques are therefore used to avoid overfitting and provide high predictive accuracy. Examples of these techniques include the use of random forests and gradient boosted trees.

Random forests models are a type of ensemble models that make use of multiple, relatively simple decision tree models to generate a prediction. Essentially, random forest models aggregate several weak leaner models (hence, "ensemble" models) in order to produce an overall strong model. As a result of this, random forest models have

an advantage of traditional regression techniques and decision trees in that they are less prone to overfitting of data (Grus, 2015).

Each decision tree model within the random forest is trained using different samples from the same training data (referred to as "bootstrapping" or "bagging"). Bootstrapping thus results in a variety of decision tree models. It is important to note that each of the generated trees are independent from each other. These independent models each predict an outcome (referred to as a "vote"). The random forest model tallies (that is, aggregates and averages) the votes in order to arrive at an ultimate prediction (Foreman, 2014).

2.9 Light Gradient Boosting Machine

In contrast to the bagging/bootstrapping techniques used by traditional random forest methods, gradient boosting creates sequential (as opposed to independent or parallel) models, with each model created to incrementally improve on the results of the previous model (Muller & Guido, 2017). As in bootstrapping methods, gradient boosting ensemble models using aggregation of the outputs of all the models generated to arrive at a final output. This aggregation considers the value of the residuals (that is, the difference between the predicted output and the actual value) and attempts to minimize the overall error of the final model. When gradient boosting is used with decision trees, the models are often composed of shallow trees, allowing for more efficient use of available memory and a faster time to prediction.

The gradient boosting decision tree (GBDT) has therefore become a popular technique in machine learning as a result of this combination of accuracy, efficiency and interpretability. Additionally, this technique provides the ability to provide increased accuracy with new data while remaining computationally efficient (especially when compared with rerunning the entire analysis on the new data) (Friedman, 2001). GBDT has also been shown to obtain class-leading performance in a variety of machine learning problems, including tasks such as learning to rank data with real-life applications (Burges, 2010).

As larger and more complex datasets emerge (both in terms of the number of available features and the number of instances), GBDT now faces new obstacles, especially in the trade-off between accuracy and efficiency – in response to this, novel and innovative implementations of GBDT have been developed to overcome these challenges. One such example of these more recent implementations is Light Gradient Boosting Machine (LightGBM). LightGBM is a newer implementation of GBDT that makes use of two novel techniques known as Gradient-biased One-side sampling (GOSS) and Exclusive Feature Bundling (EFB) (Ke et al., 2017).

Essentially, GOSS considers the relative gradients of the available data for analysis. Simply speaking, a gradient provides a measure of how significant a change in input to a model is to the ultimate output. The higher the gradient, the more rapidly the model is able to learn. GOSS attempts to trim instances of data with smaller gradients and only look at the remaining instances to estimate the information gain. Ke et al. (2017) demonstrate that by focusing on instances with larger gradients, this technique can accurately estimate information gain while making use of a significantly smaller set of data (additionally providing some computational efficiency).

EFB attempts to reduce the number of features the model has to be fitted with by combining or "bundling" mutually exclusive variables. In this context, mutual exclusivity implies that the features do not often have non-zero values simultaneously.

Ke et al. (2017) further prove that in using various public datasets LightGBM can increase the speed of the training process of more standard implementations of GBDT models (including other versions such as the popular XGBOOST) by over 20 times while maintaining comparable levels of accuracy.

In order to avoid challenges of overfitting in this research, the LightGBM implementation of GBDT is selected due to the balance it provides in terms of computational efficiency of training and predictive accuracy.

2.10 Conceptual framework

The insights from the above literature may be organised into a conceptual framework. This framework groups the insights obtained in to five categories (Swaen, 2021):

- Target outcome: the behaviour (likelihood to purchase insurance in this case) that is studied
- Independent variable: the primary predictive factor studied that impacts the target outcome
- Mediating factor/s: variables that further connect the independent variable to the target outcome by providing additional insight
- Moderating factor/s: alters the impact of the independent variables
- Extrinsic factors to control: these variables may play roles in mediating or moderating the independent variable but need to be controlled for as, practically, little (if anything at all) can be done in select or optimise these variables in practice. Therefore, these variables need to be held constant or isolated when studying the other factors.

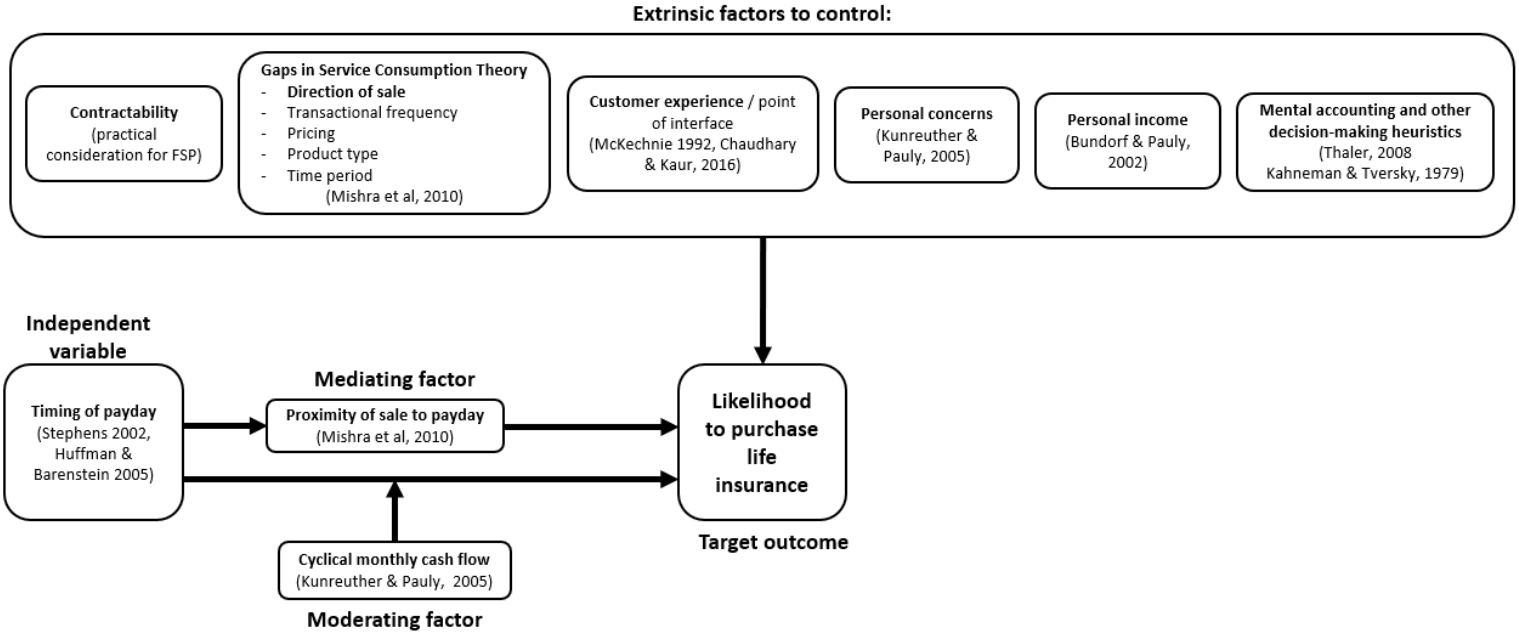

Figure 3: Conceptual framework of literature considered

The framework suggests that, in isolation of a host of extrinsic factors that the FSP cannot practically control, the timing of a consumer's payday will impact their likelihood to purchase life insurance. Specifically, the closer a sale is attempted to the consumer's payday the more likely they should be to purchase life insurance. However, this effect may be further altered by the time in the month that the sale is attempted in accordance with the consumer's monthly cash flow.

The research methodology that follows (Chapter 3), will elaborate on how the interplay between these factors was analysed as well as how the extrinsic factors to control were isolated throughout the study to ensure that these confounding factors were appropriately managed.

3. CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter provides the detail of the research design and the methodology followed, including the required research instruments, treatment of possible confounding factors, approach to data sampling and how the data was collected and analysed.

A detailed overview of the approach used to develop the predictive models is also provided, describing the key steps of pre-processing, training and validation, analysis of the correlated variables and evaluation of the final model output.

Finally, considerations regarding overall reliability and validation of the study are provided, as well as ethical considerations taken into account for the study.

3.2 Research Design and Methodology

As the objective of this study was to determine the nature of relationships between time-based factors and process efficiency, this research was a quantitative study, requiring analysis of available data using decision-tree based method such as the LightGBM framework as recommended in Section 1.5.7. Although this study was primarily quantitative in nature, it did entail a minor qualitative component whereby interviews were conducted. The key outcome of these interviews was to determine which metrics the Collections analysts felt were most important for predicting off-movement of newly sold policies.

The design of the study required four major components:

1. Structured surveys with Collections analysts within the insurer and broader bank
2. Collection and analysis of available data for non-underwritten life insurance sales over the course of a calendar year.

3. Collection and analysis of available first debit rate data for non-underwritten life insurance products over the course of a calendar year.
4. Making recommendations for improving the efficiency of the outbound sales process for life insurance sales.

3.3 Research instruments

The first component that took place in the initial phase of the study comprised of structured surveys with the Collections analysts. The purpose of these interviews was to validate the significance of first debit rates within the direct insurance value chain.

For the second and third components, anonymised data was collected from the insurer's data logs of previously worked leads. For the first component, the information recording the final status of the lead (*sale, not taken up,* or *uncontactable*) and the anonymized customer's salary date (along with the period of time between salary date and final status in the case of leads with sale or *not taken* up statuses) were extracted. These data were used to determine what relationship exists between the likelihood of sale and proximity to salary date.

For the third component, a similar process was carried out. Data was extracted for leads that resulted in a sale, with the features of the data containing the status of the first debit (*successful or unsuccessful)* and the period of time between date of sale and date of first debit (a numeric range from 1 to 31). These data were used to determine the nature of the relationship between first debit rate and proximity of debit to sales date.

3.4 Managing confounding factors

In a sense, the intention of these investigations was to determine how a lead's cashflow impacts their likelihood to accept an outbound insurance offer (in line with section 2 in Chapter 2). As such, the most critical variable that needed to be controlled that is available in the data was that of the lead's income, as this is a major influence on their monthly solvency. In order to normalise for income, the investigations stratified the data by splitting the leads under investigation into income bands. The correlations

between the time-based factor and relevant quality measure (that is, sale to leads conversion or first debit rate) were then be established per income band to mitigate the influence of personal income on the study.

A related consideration is that of the macro-economic environment. In order to control for the influence of the performance of the economy (which may have an influence on a consumer's propensity to make purchases), the analyses conducted were done over a relatively small period (not greater than three years).

The data are warehoused in various databases available to the insurer and the required features could be extracted and anonymised as required.

3.5 Sampling

3.5.1 Datasets

Various factors are considered when the Insurer evaluates the performance of outbound telesales campaigns. The performance of a sales campaign could be impacted by any of the following considerations (among others, but these have been historically noted by the Insurer):

- Seasonality (particular months/larger periods in the year have clear patterns with regards to call centre productivity and customer response)
- Medical underwriting (the process of determining a customer's individual risk profile – which may result in a decline by the insurer in cases of high risk – which in turn plays a major influence in the final value of the customer's premium)
- Maturity of campaigns (this influences the experience and/or skill of the sales agent for a particular product or campaign)
- Customer bias to marketing
- Affordability
- The essentially random time of the month when the lead is contacted (the focus of this study)

In order to control factors which may result in non-take up but not due to customer agency (such as the factors above), the population of the sample analysed will contain the characteristics listed in Table 1. These characteristics described are those that were practically available from the insurer's datasets that relate to the above factors and directly control for these factors that allow for the analysis of the impact of the time of the month the call is conducted.

Table 1: Sampling factors to be normalised

Characteristic	*Value*	*Factor/s to normalize*	*Description*
Time period	Sample size	Seasonality of consumer behaviour	The sample of sales records are analysed over the course of a full year with average rates taken to smooth out seasonal effects
Product set	Funeral plans	Medical underwriting and pricing	The funeral plans offered by the insurer do not make use of customer-level underwriting – rather a blended, flat underwriting profile is used to create a structure with flat pricing
		Sample size	Funeral plan campaigns make use of large leads file (several thousand per month)
Sales campaign	Mature campaigns	Campaign maturity	Steady state campaigns with similar profiles will be sampled, as agents will have developed similar proficiency over time. These campaigns also make use of somewhat standardized age ranges, incomes and other demographic information.
	Outbound campaigns	Customer bias to marketing	Inbound sales campaigns (where customer contact the provider) have a

			different sales conversion profile to outbound sales. Additionally, outbound sales are typically structured around a particular profile of customer resulting in a more level degree of suitableness to the customers contacted
Customer income	Banded affordability	affordability	Customers can be stratified/banded by income within campaigns to create subpopulations

Similarly, Table 2 lists the features that are practically available from the insurer's datasets that could be used to analyse the status (that is, successful or unsuccessful) of sales data and first debit rate:

Table 2: Available features for status datasets

Feature	***Description***	***Possible states***
Unique key	Anonymised, unique reference of a particular sales lead	Alphanumeric string
Contact status	Indicator of whether the lead was successfully contacted	Contacted; not contacted; uncontactable
Product uptake	Outcome of sales conversation where lead was contacted	Taken up; not taken up; uncontactable
Salary date	Date on which a lead's salary is paid into their bank account	Date
Date of contracting	Date on which sale occurred	Date
Date of first debit	Date on which first debit was scheduled	Date

3.5.2 Interviews

Table 3 lists the criteria to be used when selecting potential participants for the structured interviews investigating first debit rates.

Table 3: Selection criteria of interviewees

Criteria	*Explanation*
Functional role	Role entails functional appreciation and implementation of the quantitative methods associated with the operations of collections and the interpretation thereof
Tenure in collections	Experience in collections and appreciation of associated seasonal variances and trends
Insurance experience	Possesses a specific understanding of, and experience in collections in the context of life insurance

At the time the survey was conducted the total possible sample consisted of 17 analysts of various levels of seniority (ranking from quantitative analyst to team lead/analytics head) and experience.

In order to verify the importance of first debit rates as a metric to be tracked by the Collections team analysts (see section 1.2.3), a sample of these analysts within the Insurance company were interviewed via a digital questionnaire (Appendix A) with regards to the efficacy of various metrics in terms of predicting premature off-movements (that is, lapses and cancellations) of newly sold policies. It is important to note this survey is not a key component of this research but is conducted in order to verify the choice of variable analysed.

3.5.3 Data Collection, analysis procedures and instrumentation

3.5.3.1 Sales and debit order data

Collection of data for sales performance in the call centres and first debit success rates was handled using the enterprise software package, SAS. The data was extracted,

cleaned, filtered and categorised as necessary to normalise the data (in addition to the considerations in Table 1).

Aggregation, analysis, predictive modelling and visualization of the data was also performed using software by using a combination of the Python programming language and Microsoft Excel. The data was analysed to determine relationships as listed in table 4.

Table 4: Relationships to be determined

Analysis	***Independent variable***	***Dependent variable***
Influence of salary date	Proximity to salary date	Sales conversion (sales in sample divided by total leads in sample)
Influence of time to first debit	Proximity of sales date to first debit date	First debit rate

3.5.3.2 Survey data collection

Surveys were captured electronically on digital forms. These surveys provide structured questions with a given range of answers for participants to complete. The results of these surveys were aggregated, tabulated and analysed using Microsoft Excel. The final structure of these surveys is presented in Appendix A. The data was analysed and ranked.

Where ranking of results was required, the procedure conducted was as follows: the candidates were asked to rank the metrics (or indicate whether they were unsure) from 1 (being most important) to 3 (being least important). Once recorded, the rank given was mapped to a number of points. Being ranked first earned three points, second earned two points and third earned a single point. If an option for “Unsure” was

available as an answer, it was awarded three points. The total number of points were then tallied, and the results were expressed as a percentage of the total available points

3.6 Developing a predictive model

In order to quantify the predictive power of the time-based factors, a gradient boosted decision tree classification model was developed using the Python programming language (full code listings appear in Appendix B). Once it had been established that the classification models were valid and producing acceptably accurate predictions, the models could then be reviewed using various analyses. One such analysis was determining feature importance. This analysis (also implemented and visualised using Python) essentially looks at each of the model features (that is, the variables considered when making a prediction for a particular data point – in this case whether an outbound sale is successful or not) and quantified on a relative basis as to how much a particular feature contributed to the overall predicted output.

In this ranking, the higher the score, the more predictive the feature was. Clearly, understanding which variables had the greatest influence on the final outcome has enormous value in terms of understanding how to optimise the selection of leads to be worked at any given time.

In order to carry out an analysis using feature ranking, the following steps were followed:

- Formatting and transformation of the available data (also referred to as "pre-processing")
- Separating the data into separate sets for training, testing and validating a predictive model
- Identifying correlated variables
- Implementation and fitting of the model
- Assessment of the accuracy of the model
- Once accuracy has been deemed acceptable, carry out a feature importance ranking analysis on the model (discussed in Section 4 of Chapter 4)

These steps are depicted in Figure 4 and will now be explained in more detail.

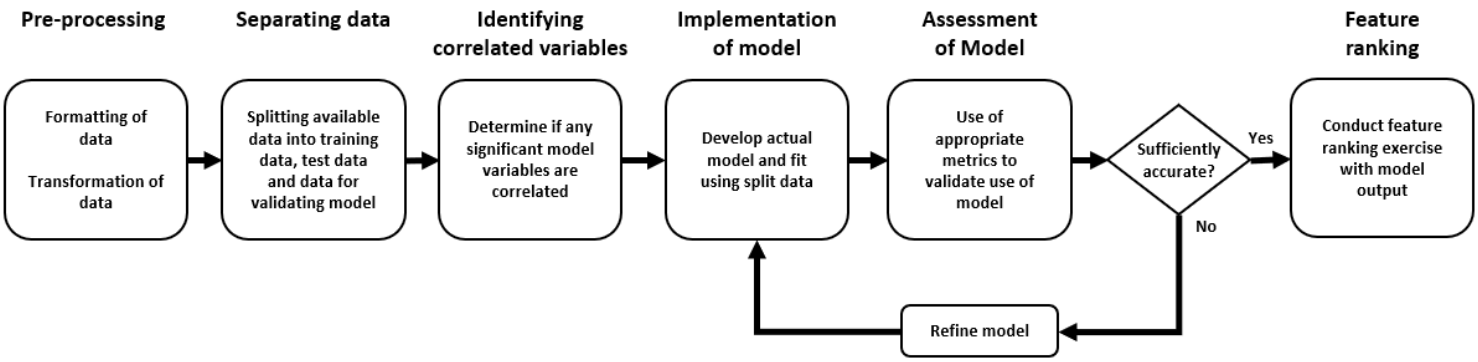

Figure 4: Process followed for implementation of models and feature ranking

3.6.1 Pre-processing

Before the data could be used to train a predictive model, several preparatory steps – referred to as pre-processing - were necessary in order to ensure that the data was in fact suitable for this purpose.

The first step of the pre-processing procedure was to review all the fields in the available dataset and to select only those that would be patently useful to the exercise and to exclude and remove those which would clearly play no role in predicting the outcome that the model generates. These fields typically include information that is endemic to the particular dataset and are specific to the fundamental use of the dataset (in this case, managing leads in a call centre). In this scenario, these are fields that served more administrative purposes (for example, campaign name, lead identifying number used by the call centre, and the like) or fields with the same value for the entire set (for example “product type” may be “Funeral Plan” for the entire set). This is used by both the insurer and call centre for analytical exercises but obviously had no impact on predicting the outcome of a lead. In these cases, the entire field was ignored by code when setting up the predictive model.

Additionally, leads with certain properties (that is, a particular value of specific field) would also be ignored by the code. A particularly salient example is that contactability (see section 1.1.2). If a lead was classified as being uncontactable (that is, the agent was unable to actually engage in conversation with the lead and begin a sales pitch), it was excluded from the analysis. Since these kind of leads provide no insight in to whether a customer would be more receptive to a sales offer based on proximity to their salary deposit date, they should be excluded (not only will these data records not only not provide insight, but they may actually introduce noise in to the model and thereby reduce its predictive ability).

There are two reasons why this pre-processing step is so essential. The first is a more pragmatic one – by reducing the features used for training a predicting model, significantly less computational effort is required and reduces the time required to train and validate models. The second reason is that by eliminating unnecessary features earlier on, fewer sources of noise are introduced in to the model training (Muller & Guido, 2017)

The next step in the pre-processing procedure was the encoding of categorical variables. At their essence, predictive models are mathematical functions that use a collection of input variables to produce a result. As such, these inputs need to be understandable by a mathematical model. Categorical variables are those that are qualitative in nature. Typical examples would include features such as race or level of education (as an illustrative example, a mathematical model will most likely be unable to interpret and consume an input such as “postgraduate degree”). If these variables are to be used, they need to be translated into a format that can be consumed and understood by the predictive model. This process is referred to as “encoding”.

In this analysis, a technique referred to as “label encoding” was used. In concept, it is a simple technique. Essentially, label encoding looks at a particular feature and considers all the possible values found in the dataset for that feature. Each value is then assigned a corresponding numeric value (as an illustrative example, if a feature called “day of the week” was being encoded using this technique, the categorical values “Monday”, “Tuesday”, “Wednesday”, “Thursday”, “Friday”, “Saturday” and “Sunday” might be assigned the numeric values of 1,2,3,4,5,6 and 7). This technique was selected (as

opposed to other methods such one-hot encoding), due its ease of implementation and its superior performance in tree-based models such as the one used in this analysis. In this analysis, the categorical variables that required encoding are segment, gender and campaign month.

As the final conceptual step in pre-processing, a new dataset was created using the newly encoded features and combined with (or "joined to") the remaining non-categorical features that were initially chosen as part of the feature selection step. This new, smaller dataset was used for constructing and verifying a predictive model.

3.6.2 Training, validation and test data

Now that the data could be consumed and understood by a predictive model, the data had to be separated in to three separate groups – namely, training data, testing data and validation data.

Training data, as the name suggests was used, to train (also referred to as "fitting") the predictive model, using sales interactions with known outcomes (that is, a successful sale or not). In other words, this is the dataset that was used to actually create the model.

Validation data is a smaller set of data used after the initial training to fine tune some of the parameters of the model and improve performance.

Finally, test data (also known as "holdout data") was used to evaluate the actual performance of the model that had been created and tuned. By using a final set of data that the model had never encountered during the "learning" phases with training and validation data, it could be determined how effectively the model could discriminated between predicted outcomes (Muller & Guido, 2017).

The bulk of the data was used for training to expose the model to a wider range of possible cases, but a balance had to be struck to prevent overfitting. When a model suffers from overfitting, it is extremely adept at providing outcomes from the given

training data set but is liable to exhibit poorer performance at generalising when exposed to new, previously unseen data (Grus, 2015).

For large datasets, such as the one used in this analysis, a training/validation/testing split of 60/20/20 is typically used (von Neumann, 2016) – that is:

- 60% of the data is separated as a training set
- 20% is separated as a validation set; and
- 20% is finally reserved as a test set for evaluation of the model

This split was used for both the analysis of the lead conversion versus time between salary receipt and call, as well as the analysis of collections versus time between call and first debit.

3.6.3 Correlation of variables

Before training a model, the various features were analysed for correlation. If any of the variables showed significant correlation then including both features in the model could make it difficult to interpret the results of the ultimate analysis due to the confounding nature of the correlated variables (Grus, 2015). These correlations are computed using the Python programming language and visualised as heatmaps for analysis.

The heatmaps are structured as a matrix that compares each feature against every other feature (and against itself for completeness). The correlation between each set of features is given where they intersect in the matrix. These values are unitless and range between -1 and 1. A value of -1 implies that the two features are perfectly anti-correlated, while a value of 1 implies perfect correlation. (Grus, 2015). A correlation of zero implies that no correlation exists at all between the variables. The heatmap then assigns a colour to each point of intersection with red assigned to negative correlations and blue assigned to positive correlations. The stronger the correlation, the more intense the colour assigned. As an illustration, a strong negative correlation will have a deep red colour assigned.

In contrast, when each variable is compared to itself it is assigned a correlation of 1 (since a variable is obviously perfectly correlated to itself) and thus is the deepest shade of blue. The heatmap helped to quickly identify significant correlations by simple inspection.

If two variables had weaker correlation (that is, between -0.5 and 0.5), they could be used together in the model.

It was noted that the features were labelled on the heatmap using the feature names they were assigned in the dataset and the code used to create the predictive model.

3.6.4 Fitting and evaluating the models

The actual predictive model was constructed using the Python programming language, specifically using the LightGBM framework to fit the random forest model. The Pandas library (an additional Python library designed for data manipulation, aggregation and analysis) was used in order to facilitate some of the key pre-processing steps, including reading of the data, selection of appropriate features and the joining of the newly encoded features to form a new baseline dataset to train, test and validate the predictive model.

The models were evaluated using the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC). The ROC metric is essentially a measure of how well a model correctly identifies true positives (also referred to as “recall” or “sensitivity”) versus false positive outcomes. ROC plots the sensitivity of the model against 1 – specificity of the model (where specificity is the rate of true negatives).

In the context of ROC, AUC provides a summary of the overall predictive accuracy of the model. AUC ranges from zero to one, with a value of zero corresponding to being perfectly inaccurate and one being perfectly accurate. A ROC AUC of 0.5 would indicate that the model has an even chance of correctly the outcome (or phrased differently, the model provides no better predictions that random guessing would

provide). Hence, any model with an AUC higher than 0.5 can provide (to varying degrees) some predictive insight. A ROC AUC of 0.7 and above is considered to indicate a model with good ability to discriminate between correct and incorrect predictions, with scores of 0.8 and above indicating superlative models. (Mandrekar, 2010).

In the context of this first analysis, the ROC AUC was used in assessing the model's ability to correctly predict a successfully converting lead (that is, a sale) versus an unsuccessful sales pitch.

3.7 Validity and reliability

3.7.1 Predictive models

In order to ensure that the research could be replicated and validated by other researchers, standardised Python libraries were used. Further, the full methodology of developing the models was provided (including the approaches to model validation and evaluation), as well as the full code listing for the models (Appendix B) including the various parameter settings. This allows for other researchers to use the same model with the dataset – or other related datasets – in order to replicate the outputs or further the study.

3.7.2 Survey

The survey used in this research can be found in Appendix A and a full explanation of how the applicable results were ranked is described in subsection 3.4.3 of this chapter. While the exact results may not be able to be exactly replicated in other insurers (as they will likely have a different mix of analysts in terms of tenure and experience), the survey may be used in different organisations to further this study using a standardised format.

3.8 Ethical considerations

3.8.1 Ethical clearance

Ethical clearance was obtained through the School Ethics Committee. The clearance number provided for this research was MIAEC 241/19.

3.8.2 Permission for use of data

Permission was requested and provided from the Insurer to allow the data required for this study to be extracted and analysed.

Individual participants for the survey participated voluntarily and were not paid nor forced to participate in the study.

3.8.2 Confidentiality of personal data

All records of data, including those provided by the participants in the survey, were anonymised for the purposes of this study.

4. CHAPTER 4: ANALYSIS AND INTERPRETATION OF RESULTS

4.1 Introduction

This chapter outlines the results obtained from the methodology described in the previous chapter. The results include a characterisation of the book of insurance policies analysed and detail pertaining to the sample data of the sales records that was obtained for the study.

Following this, the results of the survey to the collections analysts is presented, as well as the initial analysis of the impact of the studied time-based factors.

Finally, the details of the practical implementation of the predictive models are described and the results obtained are presented.

4.2 Sample Data

4.2.1 Salary deposits and debit order data

In order to characterise the distribution of salary dates, an analysis was performed on a particular book of funeral plans underwritten by the insurer. The book under analysis contained 714 531 records (that is, customers with active funeral plans) and the distribution of the salary days (recorded as the day in the month on which a salary is determined to be regularly deposited into the customer's bank account) is shown in Figure 5. Similarly, Figure 6 shows the distribution of debit order days for the funeral plans on the book (that is, on which day of the month collection is attempted for funeral plan debit order). It was noted there is a total number of 991 113 debits recorded for this book. This discrepancy (that is, more debits than policies on book) was due to the fact that a unique customer may hold multiple funeral plans, with each having their own separately recorded debit order.

The bulk of salary receipt occurred between the 24th of a month up to the first day of the following month (totally 72.3% of all salary receipt dates), with a notable spike on the 15th of the month (representing 7.8% of salary receipt dates). This mid-month spike is attributable to wage-earners who do not receive a single salary deposit every month, but rather a deposit made every two weeks.

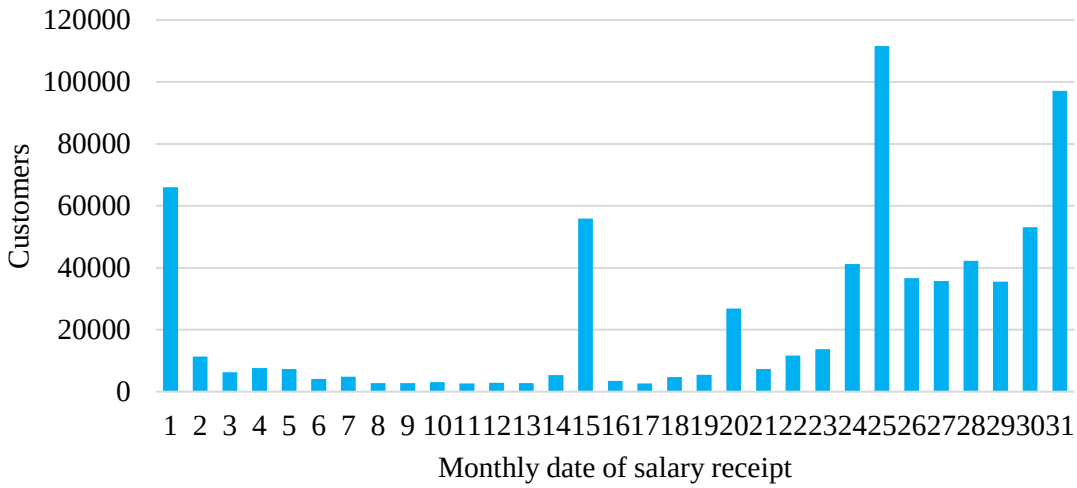

Figure 5: Distribution of salary dates for the insurer's funeral plan book

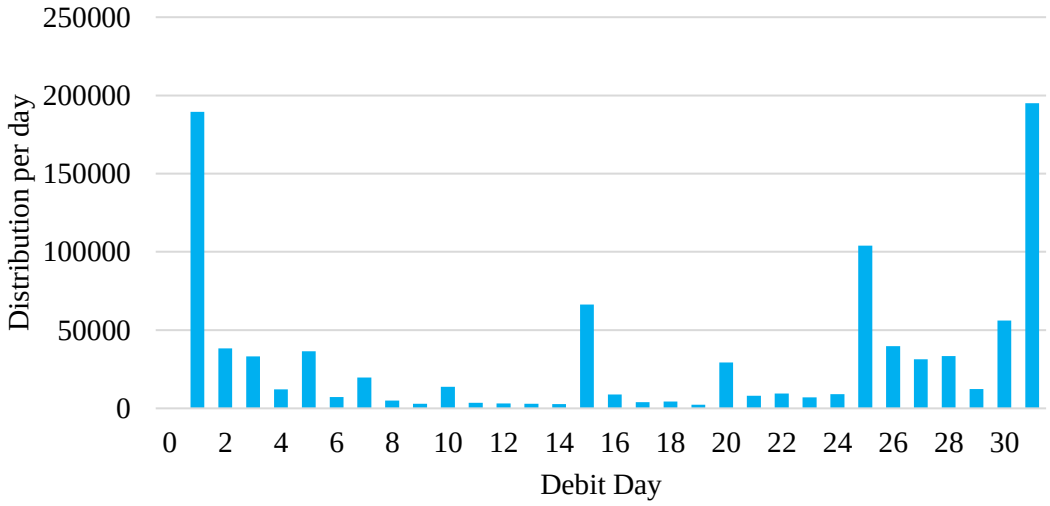

Figure 6: Distribution of debit days for the funeral plan book

While the distribution of debit order days showed a similar spread (that is, ranging from the 25th of one month to the first of the next – with a mid-month spike on the 15th of the month, it did not show a complete overlap with that of salary receipt (only 67.7% of all debits occur between the 24th of the month and the first of the next month). Instead, it could be seen there is a further concentration of debit days on the 31st (or last day of the month – debit orders for the 30th and 31st comprise 25.3% of debit order days) and the first day of the month (19.1% of debit order days), constituting 44.4% of all debits order dates, despite only representing 30.3% of all salary dates.

This tighter concentration of debit orders may be a result of various behaviours employed by customers in managing their personal finance. An example of these behaviours is that customers want their salary deposit to clearly reflect in their accounts before debit orders take place, thus ensuring these debit orders will be successfully collected for their various financial commitments. A time buffer can thus be established by setting debit orders to occur a few days after salary has been deposited to allow for deposited funds to clear. In the context of funeral plan debit order dates, customers are careful to ensure that funds are collected, as unsuccessful collections result in a funeral plan from going into arrears (thus impacting the next month's cash flow) or lapsing (resulting in losing out on the actual cover in the event of death and "wasting" all of the money previously spent on premiums up until this point).

Another possible reason is that customers have their own sets of priorities for paying off their various sources of debt. Accordingly, they may attempt to triage their debit orders, giving precedence to larger and/or more tangible debt (that is, debt for assets or loans that they currently make use of, like home loans or personal loans) and leave their insurance debit orders (which represents a possible payoff at some unknown point in the future) for later in the month in the event of an unexpected shortage of funds. This system of personal money management accords with the notion of temporal discounting (Critchfield & Kollins, 2001) - sometimes referred to as "intertemporal discounting" (Soman et al., 2005) - in the fields of behavioural economics, psychology and neuroscience (Ersner-hershfield, Wimmer, & Knutson, 2009).

While these statistics do not necessarily provide an understanding of customers behaviour around the time of sales calls, it does provide some context and illustrate the

point that customers may in fact be sensitive to time-based factors, such as the timing between receipt of salary and an optimal date for collections on their recurring debit orders for financial services.

With this in mind, the data collected pertaining to behaviour during sales will be now be discussed.

4.2.2 Sampling of sales records

In order to carry out the analysis set out in this study, a sample of funeral plan sales records was required. Each record provided information about a lead supplied for the outbound Funeral plan sales made through a call centre, as well as the outcome of the interaction. It is important to note the information that could be used to personally identify customers was completely ignored by the analysis – only basic demographic features were used.

A total of 289 102 such records were extracted from sales files from 2019. While these records represented the sum total of all attempted sales by the call centre over the period sampled, not all of the records were appropriate to consider in the analysis. As specified in Table 3, sales leads are ultimately assigned one of three broad statuses (referred to as product uptake). These statuses are:

- "taken up" (that is, the customer was successfully contacted and opted to purchase the product – in short, a sale)
- "not taken up" (that is, the customer was successfully contacted but opted not to purchase the product offered); or
- "uncontactable" (this means that the call centre was unable to successfully make contact with the lead during any of the passes made). Leads that have a product uptake status of "uncontactable" are not considered in the analysis as they provide no information regarding the potential response of the lead to an insurance offer.

Once the uncontactable leads have been filtered from the total dataset, 172 382 records remained that had conclusive product uptake statuses (that is “taken up” or “not taken up”). These conclusive records represented 59.6% of the records from the total available data.

The period sampled was an 11-month sample, beginning in February 2019 and continuing until the end of December 2019. January’s conspicuous omission is due to concerns over the validity of the data recorded for the month. In January 2019 technical issues between the bank and the call centre were experienced resulting in misclassified data. Only one calendar year was considered in order to isolate larger macroeconomic factors that may have changed in other years.

It is noted that January (in all years, even when accurately recorded) can represent an erratic sales month. This is possibly due to two reasons. Firstly, call centres typically start sales activity later than usual in the month due to agents returning from festive leave and often require some time to regain momentum.

Secondly, employers often pay salaries earlier in the month in December (ostensibly to enable their employees to do festive shopping), resulting in a significantly longer period until the next pay day (at the very least, longer than the typical possible maximum of 30 days difference during the rest of the year. The extent of this increased payday gap is variable and difficult to analytically determine. This introduces noise into the data when trying to analyse the relationship between date of last salary receipt and lead-to-sales conversion for January.

Considering these concerns (namely, veracity of the originally received data for January 2019, reduced call centre activities and sources of noise in the data), it was prudent to exclude January 2019 from the period sampled.

Within this sample, there were 16 829 instances where a lead successfully converted (that is, a lead was successfully contacted and took up a policy on the back of the offer – simply put, a sale was made). This yield of sales resulted in an overall sales conversion of 5.82% for original batch of 289 102 sales records. This conversion improved to 9.76% when only the contactable leads are considered.

Once the fundamental sample of policies had been characterised as above, the results of the survey could then be considered.

4.2.3 Survey results

Nine surveys were returned and reflected a range of experience with the analytics teams. The responding survey thus represents 53% of the available candidates. The overall experience of the responding candidates was summarised in Figures 7 and 8.

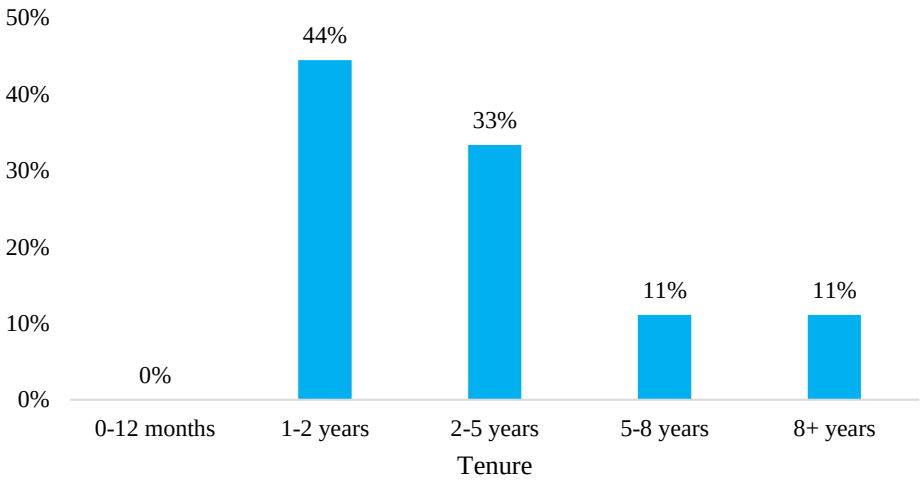

Figure 7: Tenure of analysts surveyed

All but one of the analysts surveyed had experience in collections analytics but had experience in insurance analytics (in this case, specifically sales analytics). Similarly, one candidate did not have experience specifically in insurance collections, but did have experience in broader collections analytics and acts as a business solutions head for other collections projects in the wider financial services provider that the insurer belongs to. The overall tenure within life insurance analytics is shown in Figure 8:

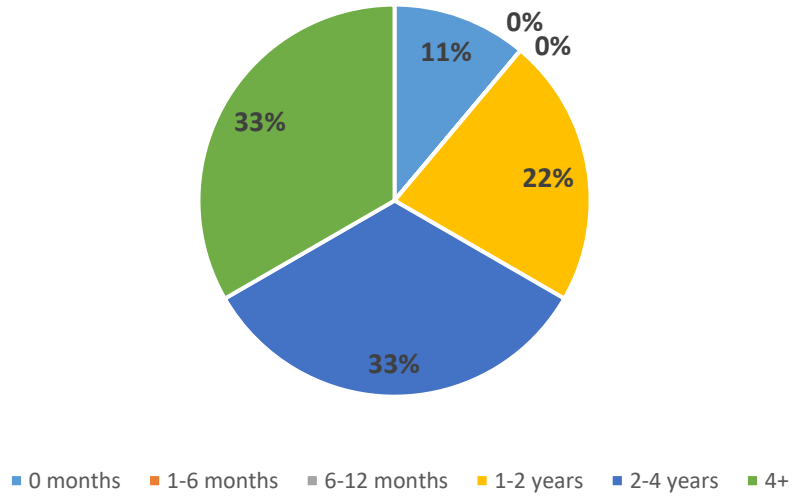

Figure 8: Number of years of experience within sample with life insurance analytics

The sample also contained a range of seniority amongst the analysts, as shown in figure 9:

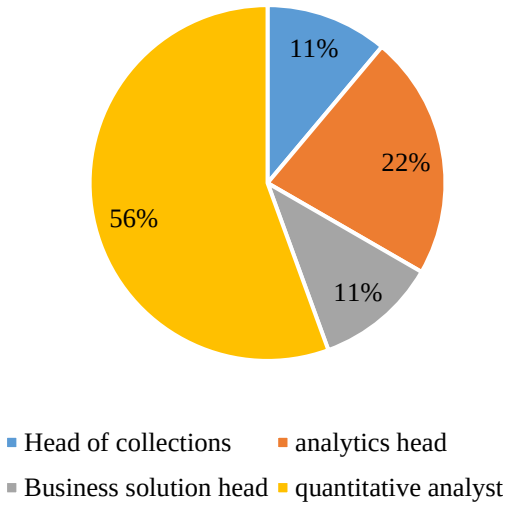

Figure 9: Position held within Analytics community by surveyed candidates

It should be noted that even though the available survey was quantitatively on the smaller side, the depth of knowledge and experience represented was significant – approximately half of the responding candidates held senior positions (44%) within

their analytics teams. These positions generally required the candidate to have first worked within the team (or related teams within the broader financial services provider), thus accumulating significant hands-on experience with the required variety of data and associated analytics. Further, a third of responding candidates indicated having four or more years of experience specifically in life insurance analytics. It was thus reasonable to consider this sample as sufficient to provide the required broad context around collections-centric analytics within life insurance.

As part of the survey, candidates were asked to rank the following metrics in terms of their perceived effectiveness in predicting early off-movements:

- First debit collection rate
- Arrears rate
- First collection rate
- Other

First debit collections rate has been explained previously, but arrears rate and first collection rate require some explanation before discussing the results of the survey.

Arrears rate refers to the percentage of policies that have not had their previous month's premium collected (usually due to insufficient funds in the policyholder's bank account or incorrect billing details being provided).

First collection rate refers to the percentage of policies that collected successfully on the first attempt within that month (several attempts are made to collect the policy if unsuccessful on the first attempt). The distinction between this metric and the first debit collection rate is a subtle one. Essentially, first debit collection rate asks: "was the very first premium collected on a new policy without it going into arrears?", whereas first collection rate asks; "was this month's premium (whether it was the first month or not) collected on the first attempt"?

Other simply provided an opportunity to note that another metric not asked about is important in predicting early off-movements. An option was also provided to allow

candidates to indicate if they were not certain how to rank the metrics (labelled as "unsure"). If a candidate marked "unsure", no other rankings were provided.

The candidates were asked to rank the metrics (or indicate whether they were unsure) from 1 (being most important) to 3 (being least important) in predicting early off-movements in new outbound life insurance sales. Once recorded, the rank given was mapped to a number of points. Being ranked first earned three points, second earned two points and third earned a single point. "Unsure" answers earned three points. The total number of points were then tallied, and the results are expressed as a percentage of the available points. The results of this poll thus provide the following tallies, show in Figure 10:

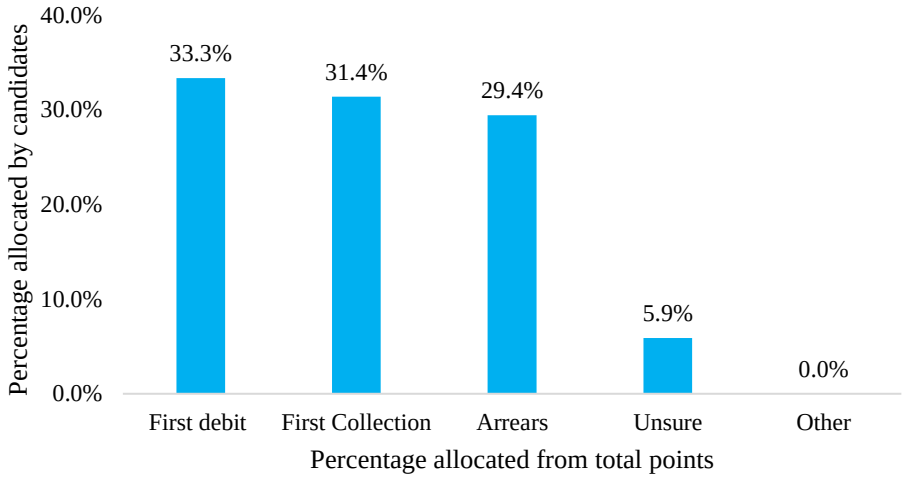

Figure 10: Ranking of efficacy of variables in predicting off-movements in new outbound life insurance sales by interview candidates

It could be seen that first debit collection rate was ranked as the most effective predictor of early off-movements in outbound sales of life insurance policies (first collection rate is a close second, but it could be argued that in practice this metric forms a broader set of which first debit collection rate is a part of) and thus was verified as a key variable related to collections of premiums that was analysed as part of this research.

Once the results of the survey had been aggregated and analysed (with the key assumption regarding the importance of first debit rate being validated), the development of the predictive models could then be considered.

4.2.4 Variables considered in predictive model analysis

Table 2 listed some of the basic features used for analyses. However, other variables may provide further insight and, as such, were considered in this study. These variables were considered as they were available in the campaign datasets from the insurer. The additional considered variables were only those that were non-administrative features.

This more expansive range of variables used in the predictive model analyses are now listed and explained in Table 5.

Table 5: Variables considered for implementation of predictive model

Variable	***Description***	***Comments***
Customer segment	A category assigned to customers by the bank related to the cheque account held.	This category is fundamentally related to income, but not always the case.
Gender	Refers to the customer's sex. In this dataset, the range of values only contains male and female.	
Campaign month	Refers to the month in which a particular lead was submitted to a call centre for dialling.	
Average monthly income	The average amount deposited into a lead's bank account every month	An average amount is considered as salaries may vary throughout the year (for example, in the case of commission-based earners or people with variable incomes) or change due to annual

		increases or change in occupation.
Date of salary receipt	Day of the month (ranging from values from 1 to 31) that a customer's monthly salary is typically deposited into their account.	
Debit order day	Day of the month (ranging from values from 1 to 31) that a customer's premium for their insurance product is scheduled for collection by the Insurer.	
Number of attempted contacts	The number of attempts made by the call centre to contact a dialled customer within a campaign month in order to get hold of them	Also referred to as *passes*
Date of contact in the month	Day (ranging from 1 to 31) on which a customer was successfully contacted by the call centre.	
Difference in number of days between salary receipt and date of call	The number of days between date of salary receipt and date of contact in the month (ranging between 0 and 30)	A value of zero implies that the lead was called on the day that they received their salary. The maximum possible value would occur if a salary was received on the first of the month and the call was made on the 31st day of that month (or vice versa)
Outcome	The result assigned to a lead after the conclusion of a sales	

	call has been completed. The outcome can be assigned as "taken up" (that is, a sale) or "not taken up" (that is, sale not made)	

The *Customer segment* variable is a nuanced one that requires some further detail. As stated in Table 5, it broadly speaks to the Bank cheque account held by the customer. This is of relevance as in order to be eligible to purchase an insurance product from the Insurer, the customer must hold a qualifying cheque account from the Bank.

There are five varieties of cheque accounts, and the account held is usually a function of the customer's income. However, if the only consideration was income then this variable would be redundant. At times, customers may be upgraded to different products by the Bank for a variety of reasons (examples include tenure with the bank, number of products held and credit score).

On the other hand, customers who qualify for higher level accounts (that is account requiring a higher annual income to qualify) may opt for a lower level account (that is one that requires a lower annual income to qualify). This voluntary "downgrading" may be due to customers wishing to pay lower monthly fees (higher level accounts attract higher monthly fees and offer more benefits) or to optimise their performance in various rewards and loyalty programs offered by the Bank. As such, this variable may be viewed as more complex than simply a measure of income, and instead is more a compound of broader customer financial behaviours.

4.2.5 Analysis of relationships

4.2.5.1 Relationship between date of salary receipt and date of sales call

In order to determine if any significant relationship exists between conversion rates of sales to leads and the time since receipt of salary, an initial analysis was performed by plotting the average sales to leads conversion rates for the particular differentials between salary receipt and time to call.

To illustrate how this analysis was carried out, consider the following example: for every instance where a customer was presented with an insurance offer one day after receiving their salary over the entire sample period, the total number of leads and successful sales resulting from those leads were considered to calculate an average conversion for a days difference of "one". This exercise is repeated for every value of days difference ranging from zero – that is, the offer is made on the date of salary receipt – up to a difference of 30 days.

Figure 11 shows the actual plot of the overall conversion for rate for the various values of days difference between salary receipt and date of call. The overall rate covers the entire sampling period for all included months and all income bands aggregated into a single view.

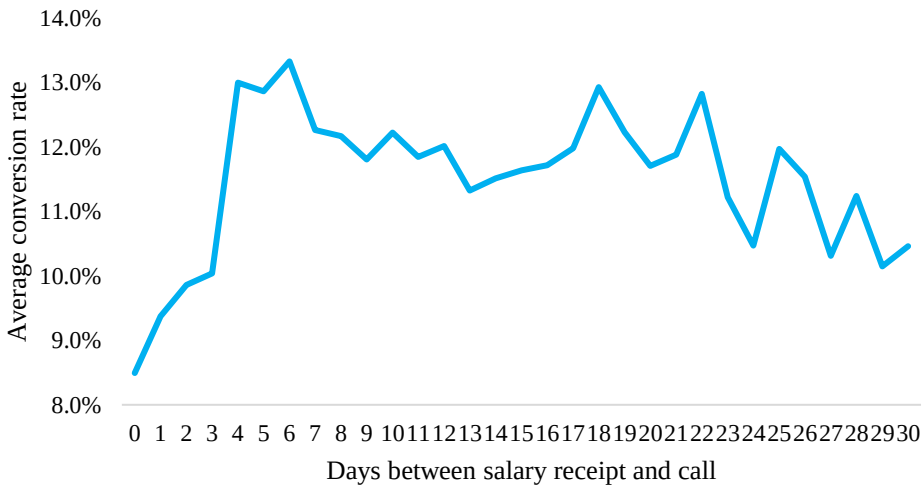

Figure 11: Overall conversion rate versus day since salary deposit (overall base)

According to the heuristics used by the call centre manager (see Sub-section 2.2 in Chapter 1), it would have been intuitively expected that peak conversion occurs on the day of salary receipt (that is, on the zero value) and monotonically decreases as the value of difference increases and tends to a value of 30.

When considering Figure 11, there are two key characteristics that stand out. Firstly, and most notably, the lowest average conversion rates occur from the values of zero to three days difference between salary receipt and date of call. The average conversion sharply increased as the differential grows.

Secondly, beyond that point, the plot appears to follow a pattern more in line with the expected behaviour described above (that is, a clear trend downwards following an initial peak in conversion - albeit without uniform movement downwards as the differential increases). As shown in Figure 12, it was noted there was in fact a significant difference between this peak and the subsequent trough towards the tail end of differential values. At a peak value of 13.3% at a differential of six days and 10.1% at a differential of 29 days (there was a minor increase to 10.5% at a differential of 30 days), there was effectively a 31.7% better chance that a given lead will convert into a sale if it is made six days after the lead has received their salary as opposed to 29 days after receiving their salary (the implications of this will be discussed in Chapter 5).

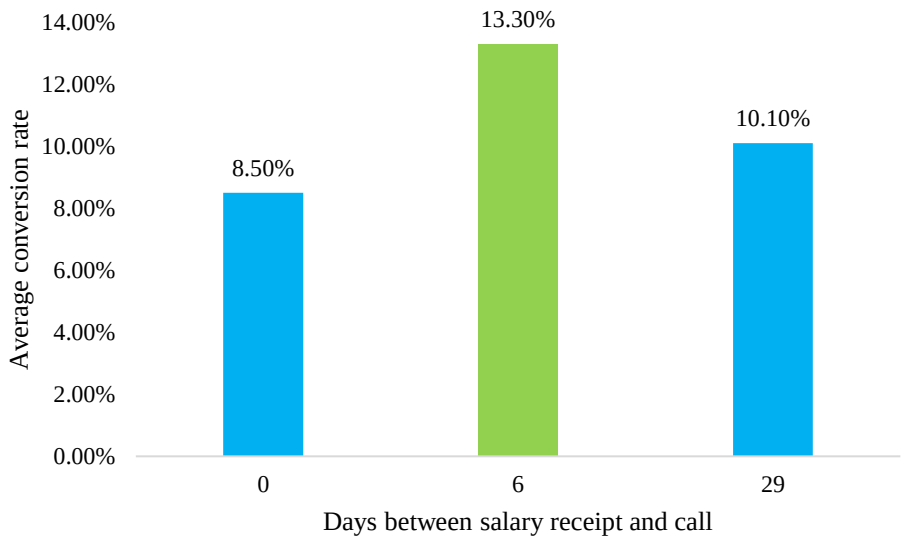

Figure 12: Relative performance of different elapsed times since salary receipt

As stark as this contrast in conversion rates was, it was even more noteworthy when considering the average conversion when calling a lead on the day of salary receipt (that is, on day "zero" with a conversion rate of 8.5%) versus the best possible waiting period after six days (figure 12). In this comparison, calling a customer earlier can result in a lead being up to 36.1% *less likely* to convert to a successful sale. Comparing it again to the lowest conversion after a differential of 29 days, the analysis suggested that a lead is up to 15.8% less likely to convert to a sale if called on the date of salary receipt than if they were called nearly a full calendar month after receiving their salary.

The leads supplied to the call centre covered a variety of incomes ranging from less than five thousand rand per month to more than thirty thousand rand per month. As such, it was critical to determine if the average conversion rates were influenced purely by income. Intuitively, it would have been expected that at higher incomes bands average conversion would flatten out, eliminating the trend in Figure 11 - as income increases, concerns about available monthly cashflow should be alleviated resulting in a relatively stable, flat conversion rate across the month.

Therefore, in order to determine if any meaningful general relationship exists between conversion and the differential between salary date and date of contact, this trend had to be established as a general characteristic independent of income. Figures 13 – 19

show a stratified view by plotting the average conversion rates versus differential between salary date and date of contact for various income bands from the data.

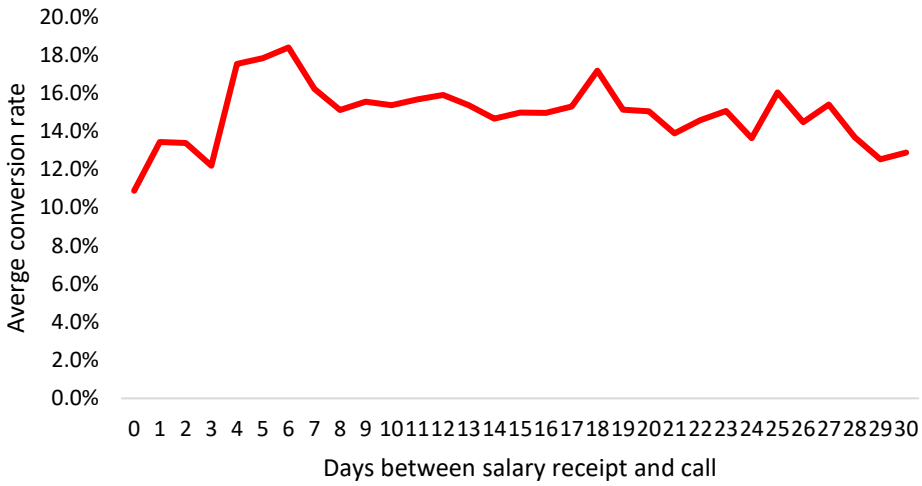

Figure 13: Conversion rate versus day since salary deposit for R0 - R5 000 monthly income

Figure 13 shows the average conversion rates versus differential between salary date and date of contact for leads with a monthly income of up to R5000. It can be seen that the same general pattern observed in Figure 11 can be seen, albeit with higher absolute conversion rate at both the peak and minimum average rates. This may indicate that customers in this income band may have a higher affinity to these products as compared to the total lead studied.

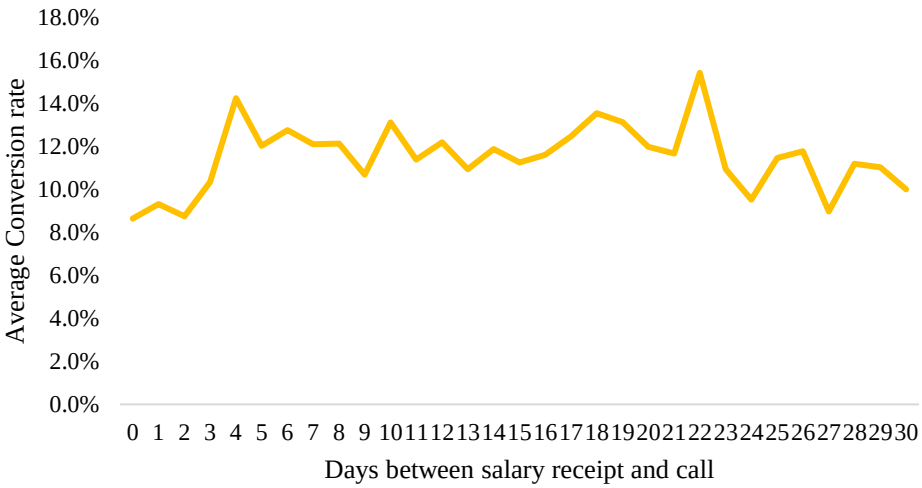

Figure 14: Conversion rate versus day since salary deposit for R5000 – R10 000 monthly income

Figure 14 shows the average conversion rates versus differential between salary date and date of contact for leads with a monthly income between R5000 and R10 000. It can be seen that the same general pattern observed in Figure 11 can be seen. It is noted that the peak average conversion for this income band is higher than that of the averaged total set of leads, with the minimum average conversion rate at a similar level.

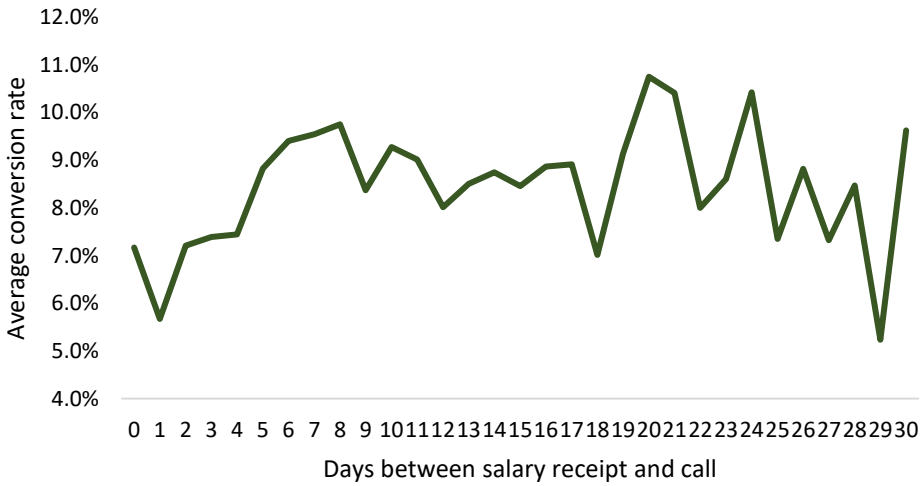

Figure 15: Conversion rate versus day since salary deposit for R10 000 – R15 000 monthly income

Figure 15 shows the average conversion rates versus differential between salary date and date of contact for leads with a monthly income between R10 000 and R15 000. Although the trend is more erratic than that of the total book (Figure 11), the same general behaviour is observed with an initial lower conversion, followed by a relatively sharp increase in conversion and then gradual decline as the number of days between salary receipt and date of call increases.

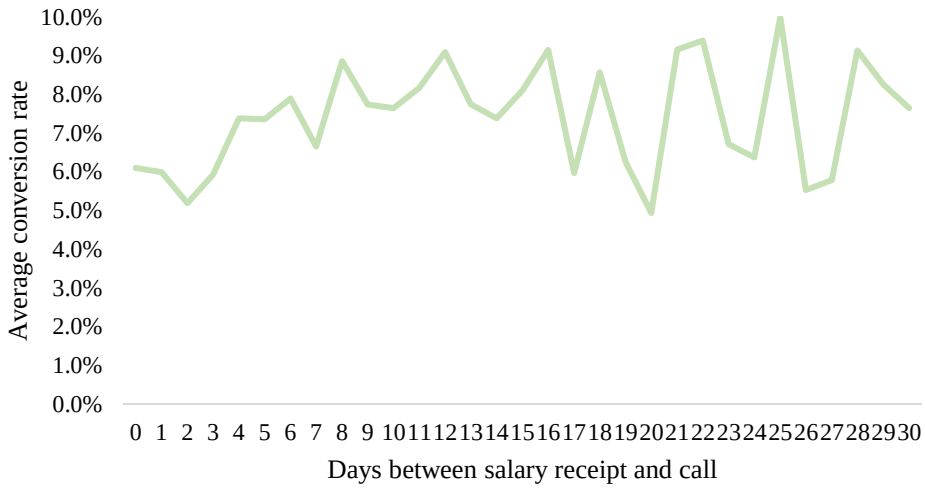

Figure 16: Conversion rate versus day since salary deposit for R15 000 – R20 000 monthly income

Figure 16 shows the average conversion rates versus differential between salary date and date of contact for leads with a monthly income between R15 000 and R20 000. Again, the same general trend observed in Figure 11 is present here although the absolute conversions are lower and the decline as number of days increases is flatter.

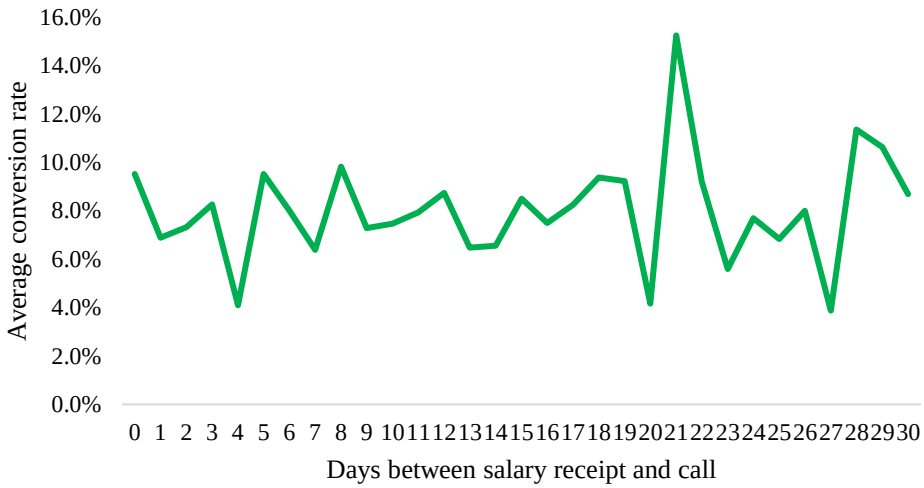

Figure 17: Conversion rate versus day since salary deposit for R20 000 – R25 000 monthly income

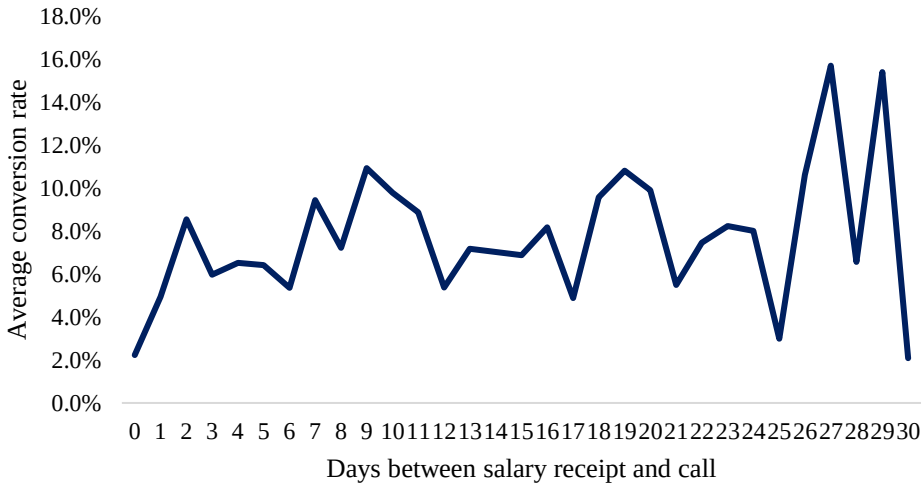

Figure 18: Conversion rate versus day since salary deposit for R25 000 – R30 000 monthly income

Figures 17 and 18 show the average conversion rates versus differential between salary date and date of contact for leads with a monthly income between R20 000 and R25 000, and R25 000 and R30 000, respectively. These trends still show an initial low average conversion with an increase in average conversion rate, but the eventual decline is more gradual. Additionally, the overall movement appears to be more erratic and less defined. This is to be expected as these income bands comprised the lowest number of leads in the dataset analysed.

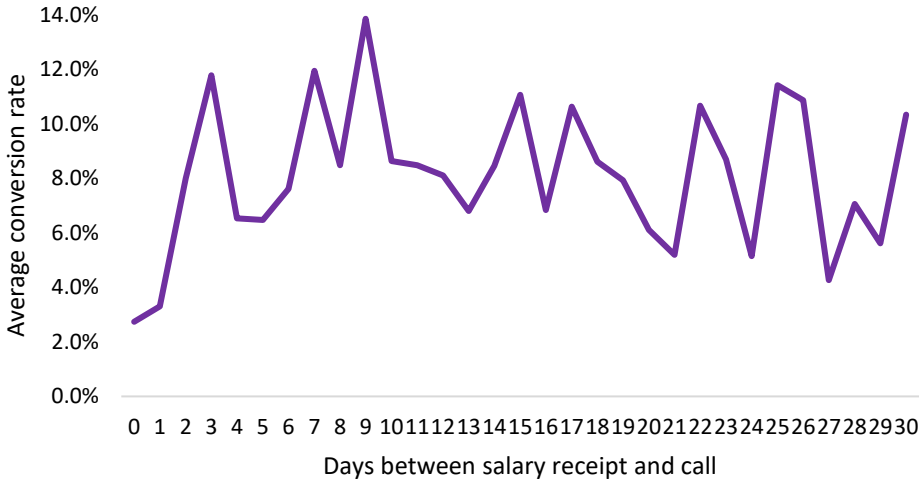

Figure 19: Conversion rate versus day since salary deposit for R30 000+ monthly income

Finally, Figure 19 shows the average conversion rates versus differential between salary date and date of contact for leads with a monthly income of more than R30 000. Again, the general trend seen in Figure 11 is indeed present although it is noted the subsequent decline in average conversion rate as the number of days since salary receipt increases. As in the previous two figures, this is to be expected as there were relatively few leads in this income band.

In summary, it could be seen in these figures that the same general trend held true across income bands for the dataset – that is, for each income band an initial trough in average conversion rates was observed for the first few days following salary receipt, followed by sharp increase in average conversion rate and then followed by a gradual decline in conversion as the differential between salary date and date of contact increases. This signature pattern is observed to some degree in each of the above figures (11-19) and indicated a general trend.

It was noteworthy that in the higher income bands (R20 000 and above), this trend became more erratic as a result of the more spurious data due to fewer leads in the higher income bands. Figure 20 shows the distribution of income bands within the aggregated leads considered in the analysis.

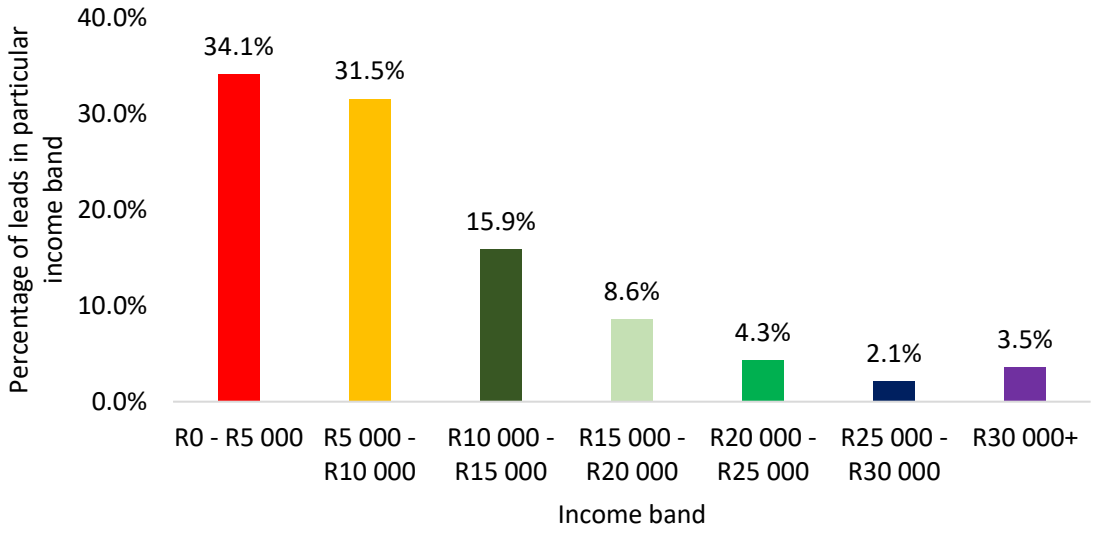

Figure 20: Distribution of income bands within the set of call centre leads analysed

It was clearly seen there were significantly fewer leads from incomes of R20 000 and above, constituting only 9.9% of the total leads, thus explaining the above-mentioned erratic data. It was therefore seen that the general trend observed in Figure 11 held true across the range of different incomes. It is important to clarify that this does not imply that income has no influence on conversion, nor that is not an important factor in predicting conversion of a lead into a successful sale – only that the observed trend generally held true.

The profile of average conversion rates (Figure 11) suggested the intuition of the call centre managers (namely, that leads are more likely to convert shortly after pay day) is generally reasonable in terms of overall trend, although the early sharp troughs in conversion indicated a need for further analysis. In order to determine if there was in fact a meaningful relationship between lead-to-sales conversion rates and difference in time between receiving salary and receiving an outbound sales call for insurance that existed, a more sophisticated analysis was required to determine whether this differential has any predictive ability.

4.2.5.2 Relationship between date of sale and date of first debit

When analysing the relationship between the first debit rate and the number of days that have elapsed between date of sale and date of first debit, a more linear relationship was observed.

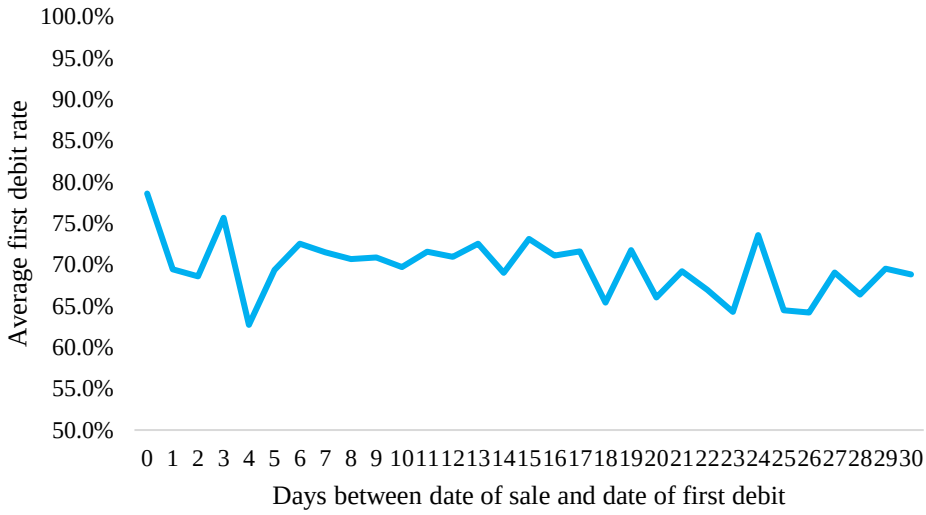

Figure 21: First debit rate versus days between sale and date of first debit for overall base

Figure 21 shows this relationship, and a more monotonic trend can be seen. As the number of days between date of call and date of first debit increased, the likelihood that the first debit would successfully collect decreased. This likelihood of collection was at its peak of 78.6% when the selected debit order day was the for the same day of the month as the date of call (represented by day "zero"), and gradually decreased with the last five days not having a likelihood to collect higher than 68.8%. This implies that successful collections can be as much as 12.5% less likely towards the end of the month than the beginning of the month.

Stated differently, the sooner the debit is scheduled to occur after the date of sale, the more likely the debit order will collect and result in an activated policy.

As in the analysis of leads conversions, the impact of customer income needed to be investigated. Figures 22-28 show this stratified view of the relationship between days from sale to first debit and first debit collection rate for different income bands.

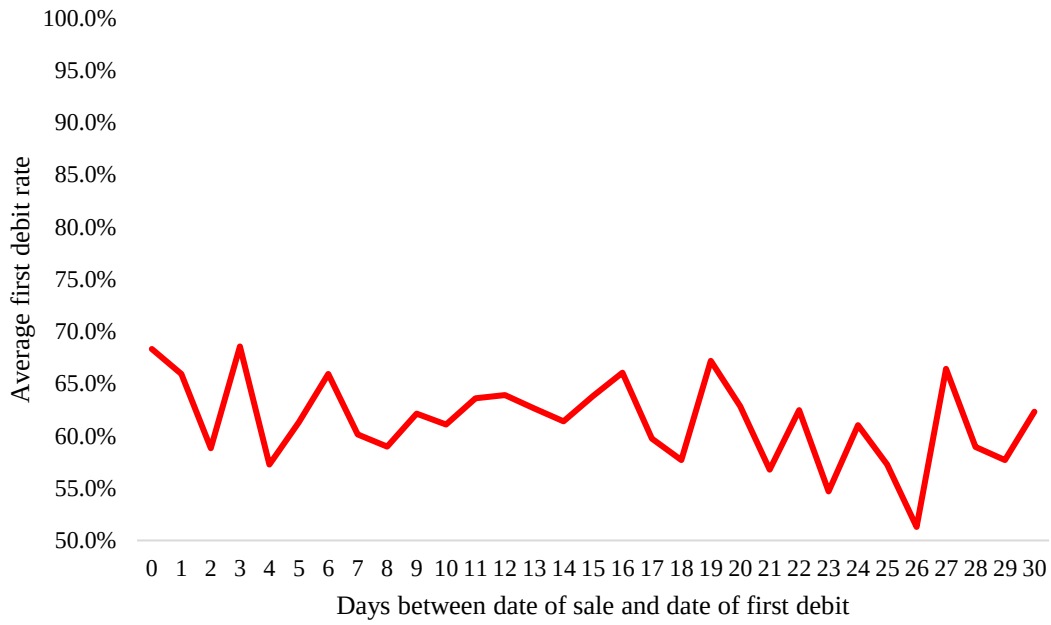

Figure 22: First debit rate versus days between sale and date of first debit for monthly income between R0 and R5 000

Figure 22 shows the average first debit rate versus number of days that have elapsed since the date of sale for leads with monthly income of R5 000 or less. As may have been expected, the customers with the least income showed a lower overall rate of average collection than the overall base, although the same linear trend downward observed in Figure 21 is present.

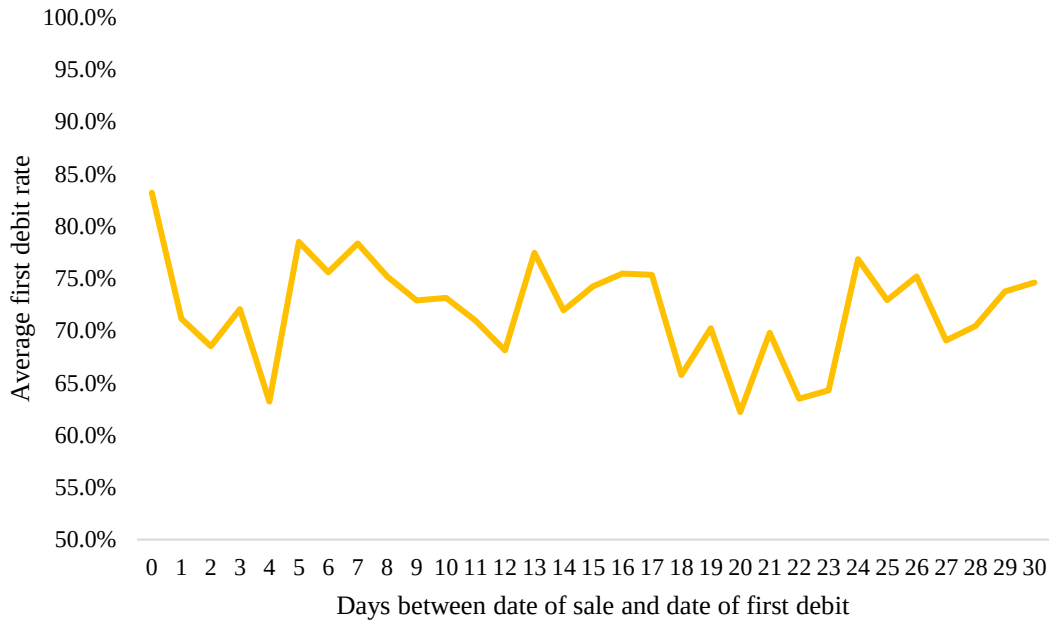

Figure 23: First debit rate versus days between sale and date of first debit for monthly income between R5 000 and R10 000

Figure 23 shows the average first debit rate versus number of days that have elapsed since the date of sale for leads with monthly incomes between R5000 and R1000. It can be seen that the pattern is closer to the trend for the overall base in Figure 21 and shows significantly higher average collection rates than for the lowest income leads.

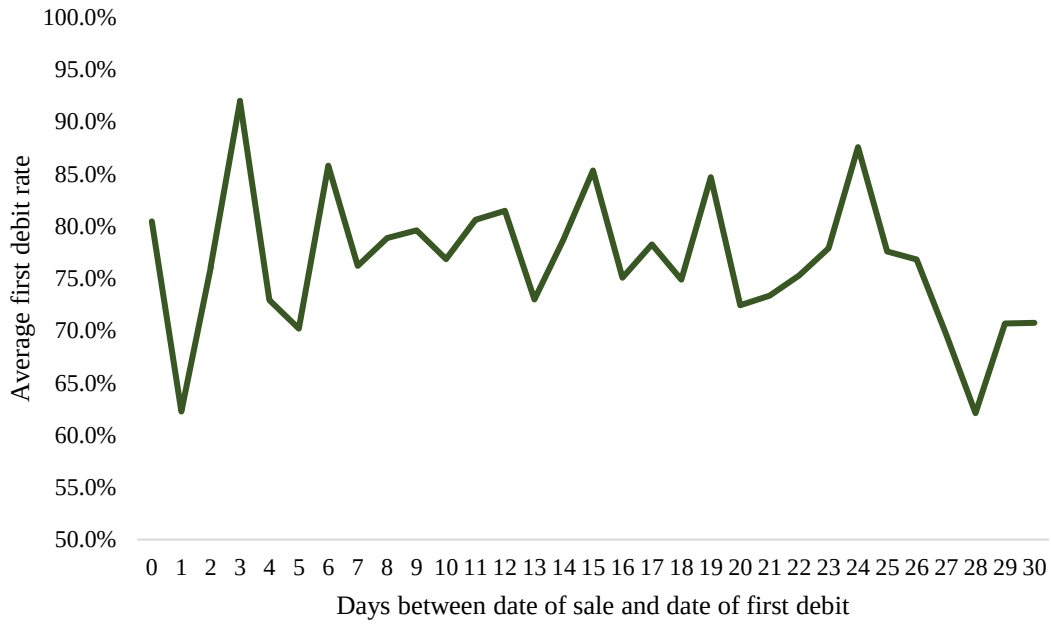

Figure 24: First debit rate versus days between sale and date of first debit for monthly income between R10 000 and R15 000

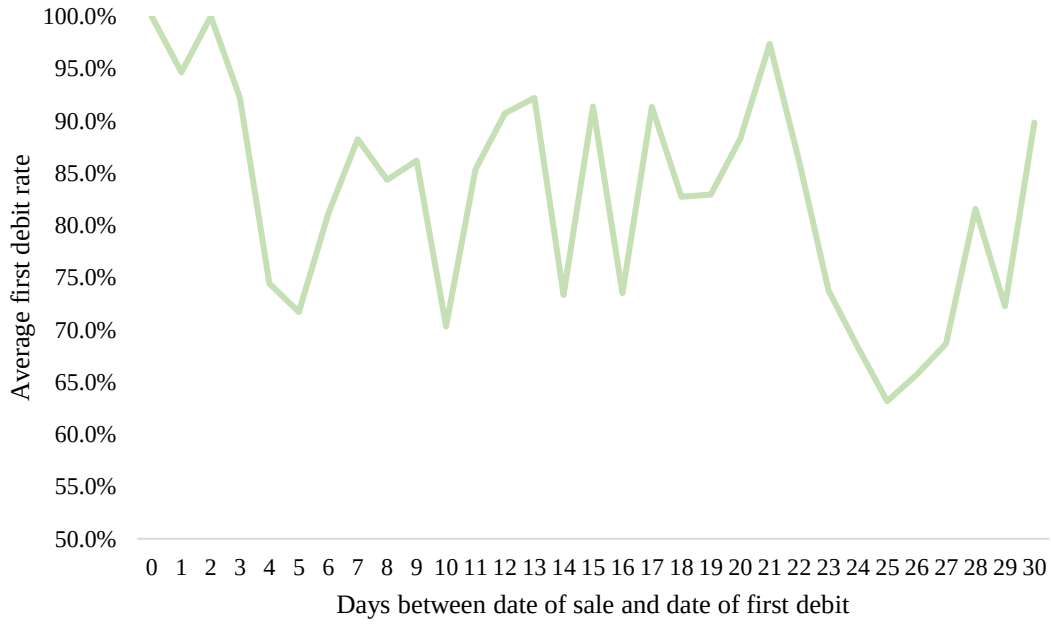

Figure 25: First debit rate versus days between sale and date of first debit for monthly income between R15 000 and R20 000

Figures 24 and 25 show the average first debit rate versus number of days that have elapsed since the date of sale for leads with monthly incomes between R10 000 and

R15 000, and R15 000 and R20 000 respectively. It can be seen that a similar overall pattern of an initial peak in average collection and relatively linear decrease as the number of days since the date of date sale increases. It is noted that in these higher income bands (relative to Figures 22 and 23) average collection rates are overall higher.

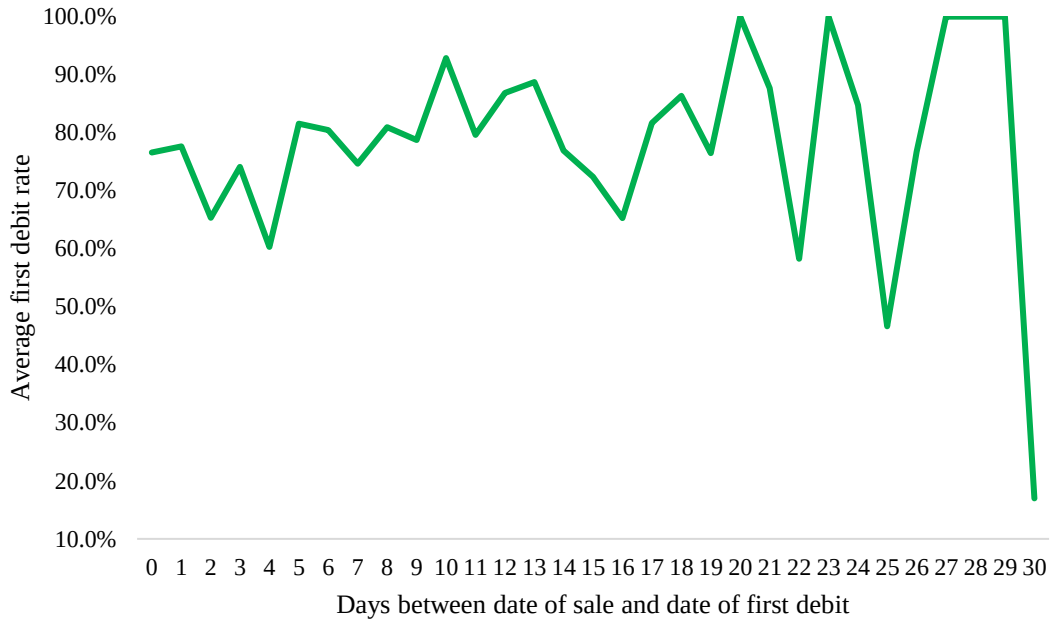

Figure 26: First debit rate versus days between sale and date of first debit for monthly income between R20 000 and R25 000

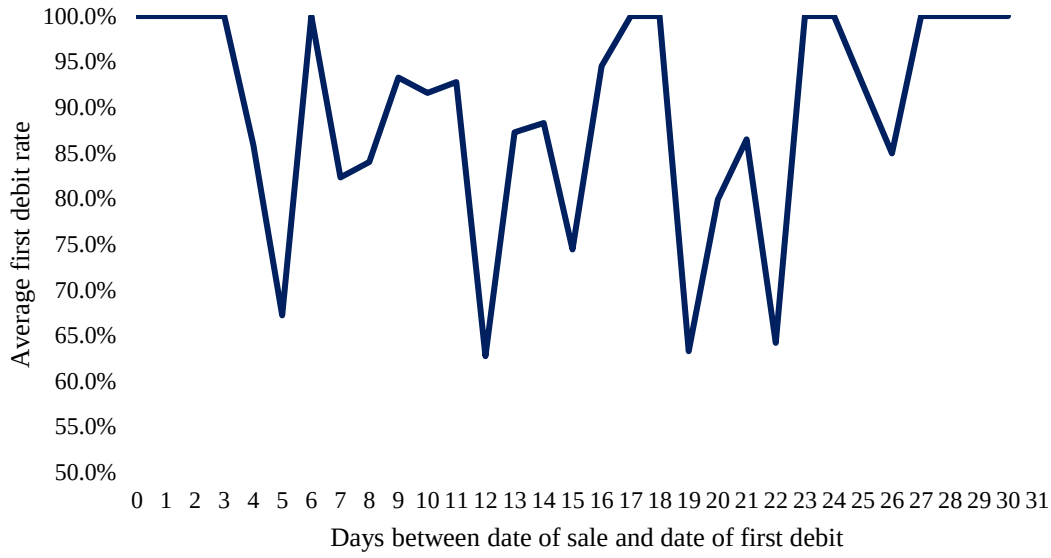

Figure 27: First debit rate versus days between sale and date of first debit for monthly income between R25 000 and R30 000

Figures 26 and 27 show the average first debit rate versus number of days that have elapsed since the date of sale for leads with monthly incomes between R20 000 and R25 000, and R25 000 and R30 000 respectively. As in the previous analysis of das between salary receipt and date of call, the relatively sparse data for these income bands results in erratic trends (including instances where first debit rates are at 100%, implying that all first debits successfully collected. This is likely due to the small sample of policies that debited their first premium at those time periods from date of sale). Coupled with the higher income of these leads, the trend observed in Figure 21 is less pronounced.

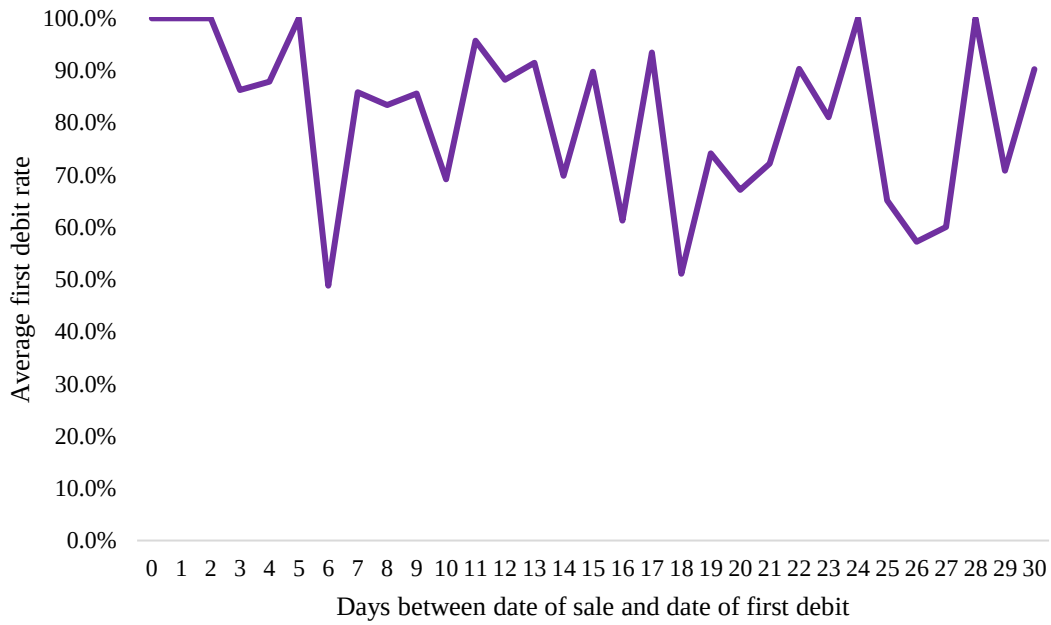

Figure 28: First debit rate versus days between sale and date of first debit for monthly income greater than R30 000

Finally, Figure 28 shows the average first debit rate versus number of days that have elapsed since the date of sale for leads with monthly incomes greater than R30 000. While less pronounced, a similar pattern is observed to that of Figure 21 with an initial high in collection rates when the number of days since date of sale is at a minimum and a gradual decrease in collection rate as the number of days since the sale increases. As has been observed in the preceding Figures, as monthly income increases for the bands

of leads, the decline in collection rate as number of days between sale and first debit increases has attenuated to become more gradual.

In summary, in these Figures showing the first debit collection rate trends for different incomes, the same general pattern can be observed – that is, a gradual decline in collection efficiency as the number of days between sale and first debit collection increases.

Of note, it was also seen the lowest overall collection rates are seen in the two lowest income groupings (as may have been expected). The lowest income grouping (Figure 22) had a maximum first debit collection rate of 68.6% with the next lowest grouping (Figure 23) having a maximum of 83.2%. A second observation of note was that for incomes of R20 000 and greater (Figures 26-28), the decline of first debit collection rates was noticeably less pronounced. These two observations are both clear illustrations of the impact of a customer's cashflow (and accord with intuitive thinking) – customers with lower monthly incomes are more difficult to collect premiums from (especially with debit dates further away from the date of sale), while customers with higher incomes are easier to collect from with the impact of a later debit order date being less pronounced (but still visible). While this observation may seem obvious, it is not clear whether the lower income customers exhibit lower average collection rates due to their particular circumstances (Barsotti, Milhaud, & Salhi, 2016) or due to greater sensitivity to macroeconomic shocks (Fier & Liebenberg, 2012).

As with the case of the lead conversion trends (Figures 13-19), the data for incomes greater than R20 000 became more erratic owing to the smaller sample sizes in higher income groups (Figure 20), resulting in more spurious data. Importantly, even though the data in the higher income groupings was more erratic, the observations noted above were still visible.

4.3 Implementation of predictive models

The method described in section 3.6 was implemented and the following results were obtained for the correlation analysis.

4.3.1 Correlation of variables

Figure 29 shows a heatmap of the correlation between each variable considered in the model for predicting lead conversion.

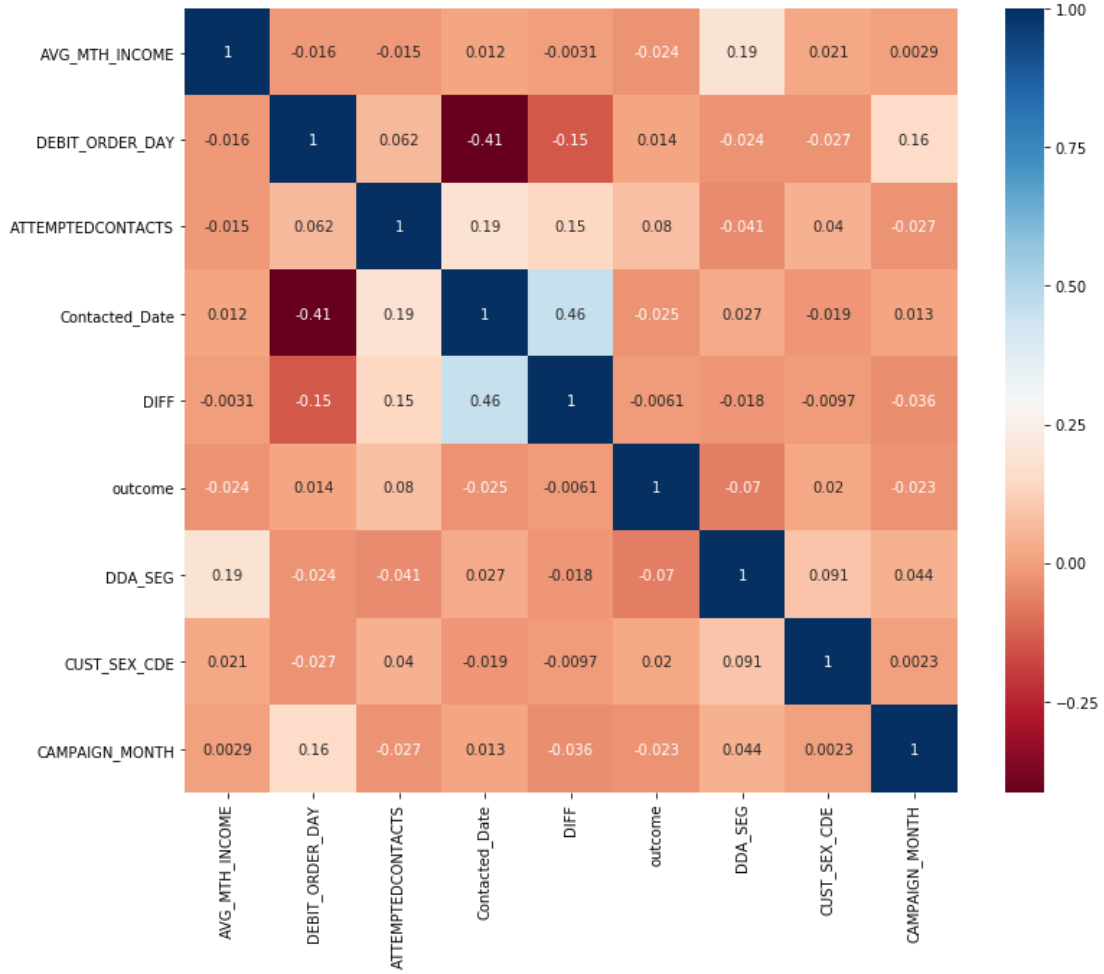

Figure 29: correlation matrix of features considered in the leads conversion model

It is noted again that in the heatmap generated, colour is assigned to each point of intersection with red assigned to negative correlations (tending to a correlation of -1) and blue assigned to positive correlations (tending to a correlation of 1). The stronger the correlation, the more intense the colour assigned.

From Figure 29 it could be seen that no significant correlations existed between the variables. The strongest variables appeared to be between date of contact (shown as "Contacted_Date", as in the dataset and model code) and the difference in days between date of salary receipt and date of contact by the call centre (shown as "DIFF", also as per the label in data and model code) with a positive correlation of 0.46. The next most significant correlation was between date of contact (again, "Contacted_Date") and the day a salary is deposited into a lead's account (shown as "DEBIT_ORDER_DAY", as in the dataset and model code) with a negative correlation of -0.41.

While these values are not considered to be particularly strong correlations, it may have been expected given the tight clustering of salary dates seen in the data (see Figure 5).

For the model predicting successful collection of first debit, the heatmap of correlations for the model is shown in Figure 30.

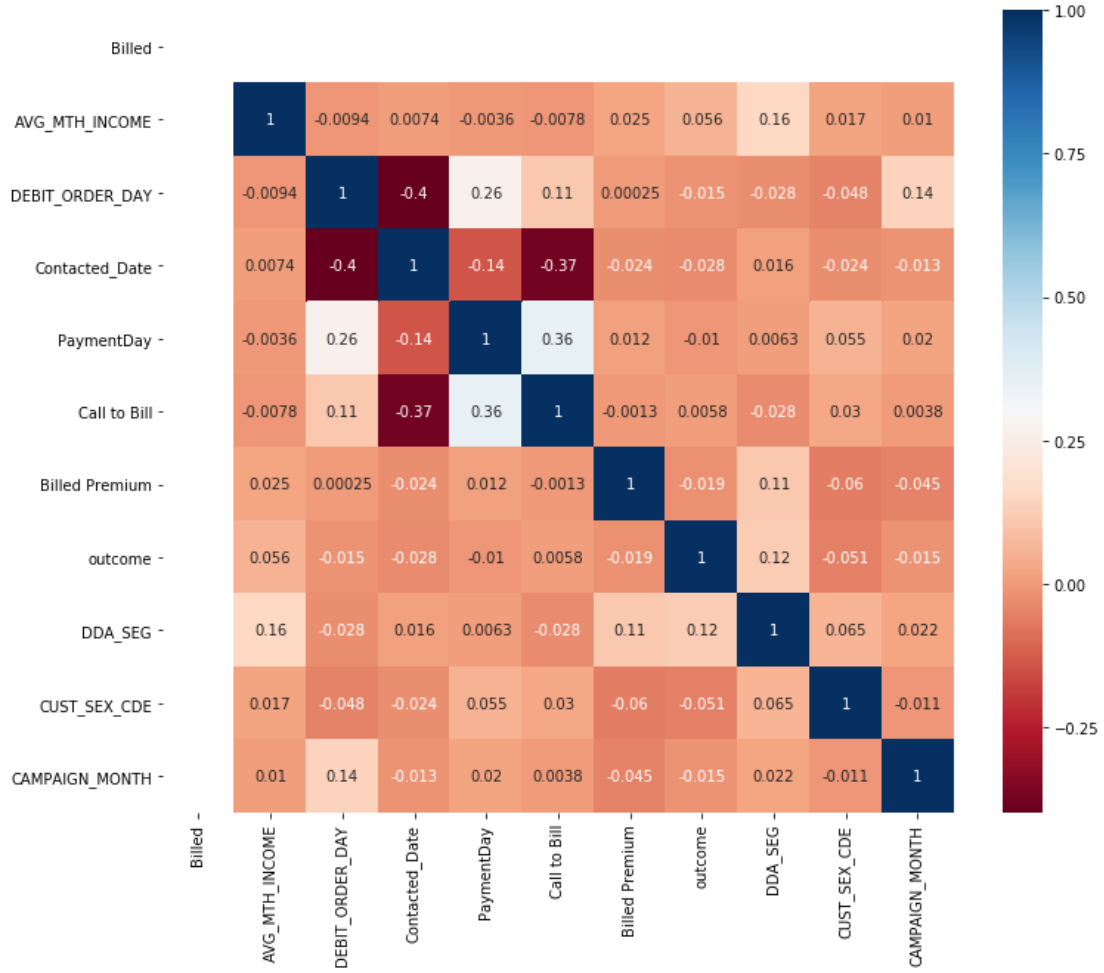

Figure 30: correlation matrix of variables for debit order analysis model

It was seen that, again, no significant correlations existed between any of the variables. The strongest correlation existed between the date that a lead's salary is deposited in their account (labelled as "DEBIT_ORDER_DAY" in the code and model – not be confused with the day of the month that the debit order for the insurance product will occur) and the date of the call ("Contacted_Date" in the code and the model) with a correlation of -0.4.

The next biggest correlation of -0.37 existed between the date of the call and the number of days between the call and the date of the first debit (labelled as "Call to Bill" in the code and the model). This was closely followed by the correlation between "Call to Bill" and the date in the month that the debit order for the insurance product is scheduled

to go off (labelled as “PaymentDay” in the code and the model) with a correlation of 0.36.

Again, these correlations were not considered significant enough to exclude any of the features from the model to be developed.

4.4 Model Results

4.4.1 Leads conversion model results

The model created in the analysis of leads conversion obtained a ROC AUC score of 0.7360 and was therefore considered to be more than sufficiently accurate at predicting the outcome of a sales pitch based on the lead profile.

In an effort to improve the accuracy of the model a technique known as feature engineering was employed. Using built-in methods, the model code could generate new hybrid variables to be used in fitting the model. These new hybrid features created combinations of the existing features in the baseline dataset and attempted to explore the impact of interactions between particular features. Using this approach, the model attempted to create hybrid features using the categorical variables that represent customer segment (labelled as “DDA_SEG” in the model code), gender (labelled as “CUST_SEX_CDE” in the model code), and campaign month (labelled “CAMPAIGN_MONTH” in the model code).

Using these hybrid features to retrain the model, the model obtained a ROC AUC score of 0.7328. This score is similar to (albeit slightly worse than) the initial model that was fitted without the hybrid features.

4.4.2 Feature ranking of leads conversion model

Now that the model was demonstrated to be sufficiently reliable, it could be used to derive relative feature importance and thus, the predictive power of individual

variables. A visualisation of the ranking of importance of the various features (including the hybrid features generated in the second version of the model) is shown in Figure 31, below (again, the labels assigned to the features in the figure follow the labels used in the actual model code).

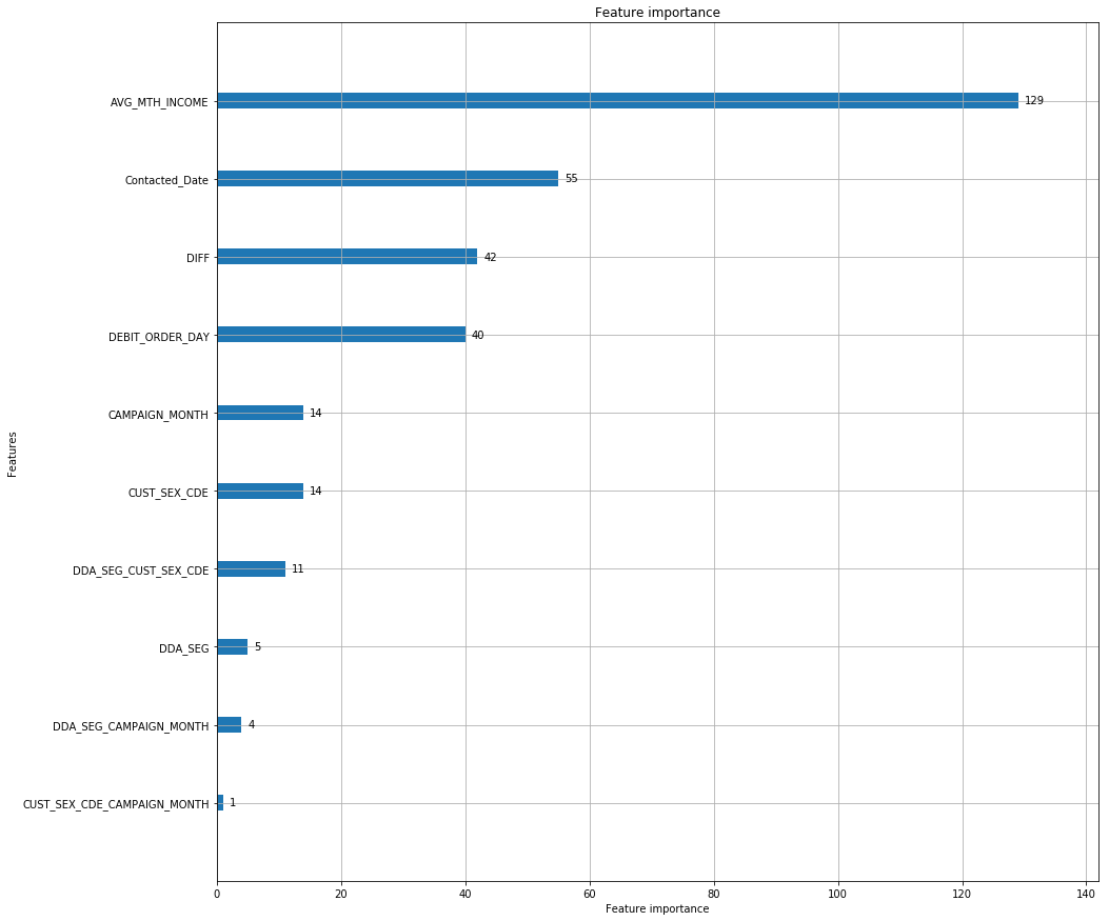

Figure 31: Ranking of feature importance for lead conversion model generated by the model code

The ranking of the importance of the various features was conducted using the "split" parameter for metrics in the LightGBM library. This method of ranking considers the number of times that each feature was used throughout the running of the model and generation of trees. The higher the count of instances of the use of a particular feature, the more important it is considered (or stated differently, how much does that particular feature contribute to a prediction for a particular record).

According to the ranking generated, the most prominent feature in the model was a customer's average monthly income (labelled as "AVG_MTH_INCOME" in the model code and the figure) with a usage count of 129.

This finding supports the position of Stephens (2002) that consumers are acutely sensitive to the inflow of their monthly salary when making purchase decisions (in addition to according to the intuitive notion that customers that earn more money are more likely to purchase something than customers with lower incomes). The practical significance of this will be discussed in Section 5.

Following average monthly salary, the next three most important features ranked by the model are (in order):

- Date of contact (labelled as "Contacted_Date" in the code and figure) with a usage count of 55
- Days elapsed between salary receipt and date of contact (labelled as "DIFF" in the code and figure) with a usage count of 42; and
- Date of salary receipt (labelled as "DEBIT_ORDER_DAY" in the code and model) with a usage count of 40.

Following these features there is a sharp drop in usage of the other features, none of which are used more than 14 times (and as little as once). Included in these lower ranking features are the generated hybrid features. This lower usage was to be expected given that adding the new hybrid features actually marginally reduced the predictive ability of the model.

4.4.3 First debit collection model results

In the second analysis, the ROC AUC was used in assessing the model's ability to correctly predict which leads would result in a successful collection of first debit for the insurance product (the alternative obviously being an unsuccessful collection).

The model created in this second analysis obtained a ROC AUC score of 0.6283. While this value was lower than that obtained by the lead conversion model, it was still

considerably higher than 0.5 (which would imply predictive capabilities close to random guessing).

In a similar attempt to the first analysis, feature engineering methods were used to generate new hybrid features that were used to fit a new model. The features selected for creating new hybrid variables are the number of days between date of call and first debit (labelled as "Call to Bill" in the code and the model) as well as customer segment, gender and campaign month (respectively labelled ("DDA_SEG", "CUST_SEX_CDE" and "CAMPAIGN_MONTH" in the code and the model), as in the previous model.

Unlike in the case of the lead conversion model, by using hybrid features to retrain the model, the model obtained an improved ROC AUC score of 0.6389. This improvement in score implies that the newly created hybrid features possess predictive value.

The purpose of this exercise was to create a model that is able to provide insight as to which features are most predictive of a successful first debit and not necessarily a model optimised for production use. Benchmark models designed to predict churn (that is, the likelihood that customers will terminate their relationship with a service provider) for other service-based (as opposed to physical goods-based) industries, such as telecoms, have similar AUC scores, depending on the implementation of the model (models using bagging methods can have AUC scores as low as 0.6168 with the higher scores being 0.6273 and 0.6307 for random forest and logistic regression models, respectively) (De Bock & Van den Poel, 2012). Given this, the model created for this analysis can be considered to be sufficiently accurate to use for a feature ranking analysis.

4.4.4 Feature ranking of first debit collection model

A visualisation of the ranking of importance of the various features (including the hybrid features generated in the second version of the model) is shown in Figure 32, below (as before, the labels assigned to the features in the figure follow the labels used in the actual model code).

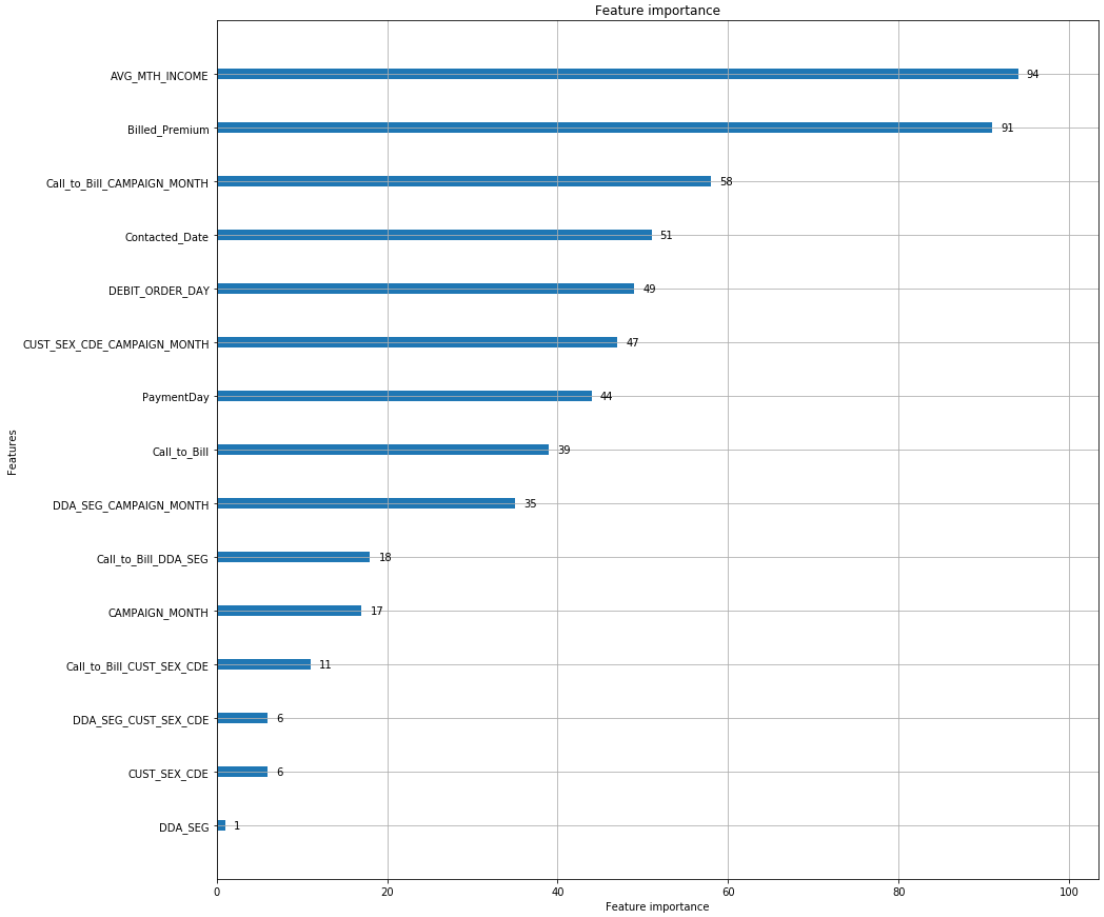

Figure 32: Ranking of feature importance for first debit model generated by the model code

Again, this ranking of features was conducted using the "split" parameter for metrics in the LightGBM library.

In the ranking generated, the most prominent features identified were the customer's average monthly income and the actual monthly premium of the insurance product (labelled as AVG_MTH_INCOME and Billed_Premium, respectively, in the model code and figure) with usage counts of 94 and 91, respectively. The next highest ranking was a feature that is a combination of the amount of days between calling and first billing, and the campaign month (this gestalt variable is labelled as Call_to_Bill_CAMPAIGN_MONTH in the model code and the figure) with a usage count of 58.

Again, the implications of these rankings will be discussed in Section 5.

5. CHAPTER 5: DISCUSSION

5.1 Introduction

The findings of this research, as detailed in the previous chapters, and their practical significance are discussed presented in this chapter.

5.2 Importance of number of days between salary receipt and date of call

The feature ranking exercise showed that the most predictive feature used by the model (and more importantly, as a feature of each of the leads) was average monthly income. It is important to note that there will generally be a variety of average monthly salaries for the various leads within a monthly batch and there is little control over selecting customers with narrow ranges of incomes as there is a finite amount of leads, particularly those with higher incomes. As such as it becomes unfeasible (or at least, short-sighted) to create campaigns only using higher income leads that can be sustained every month. Given this practical limitation it is not pragmatic to propose changes to any ongoing sales and marketing processes or targeting of leads around narrower bands of income than those already used.

Further, there are possible market conduct considerations to be factored in. Lower income customers who could still benefit from the insurance products being offered may have less opportunities to access the offerings. The implications of excluding these lower income customers from such offers in the name of higher efficiency (that is, more cost effective) campaigns must be considered carefully.

Given the rankings assigned to the model features it appears that difference between salary date and call date was *not* the most important factor and ranked behind actual date of call. The reason behind this could possibly be due to the fact that there is such a dense clustering of salary receipt dates (see Figure 5). Even though no strong correlation exists in reality between date of contact and the difference between salary receipt and date of contact, they still possessed the most significant correlation (0.46), thus implying that if one features ranks as important in the model then the other feature may also rank highly.

5.3 The impact of agent behaviour

This dense clustering of salary date and weak correlation may be coupled with the reality that other behavioural factors may come in to play. The agents working within the Insurer's call centre are typically remunerated with variable commission-based payments at the beginning of each month. These commission-based payments are linked to each's agent's individual performance. The agents' performance can be measured across a range of metrics but will generally be primarily driven by number of sales made in the month. Naji (2014) demonstrates there is significant connection between remuneration structures, particularly those that include an indirect compensation portion, such as commission-based payments determined by individual performance, and job satisfaction. It is shown that these forms of remuneration, especially without other forms of job satisfaction (such as the nature of the working environment and the agents' relationship to their supervisors) can lead to lower overall engagement with their work activities.

South African call centres have been noted as sources of anxiety and stress for their employees due to the nature of the inherent environment and degree of supervision (Benner et al., 2007). Given this, it is feasible that the closer to the immediate receipt of commission-based payments a decrease in job engagement and motivation could be experienced by the call centre agents, resulting in reduced levels of performance. This reduced motivation and performance may manifest as a lower conversion rate of sales to leads. As time passes from this feeling of disengagement (as well as the practical need to manage performance levels to maintain or improve variable pay), conversion of leads could then return to more optimal levels and thus be more driven by consumer behaviour, as opposed to any feeling of despondency on the part of the sales agent.

Given that agents may regularly be demotivated in the first few days of each month, a mitigation strategy is required in order to reduce the impact of this on leads conversion in order to optimise the cost efficiency of the call centre's handling of the leads.

There are two possible immediate options for tackling this issue, namely:

- New incentive structures
- Alternative allocations of call centre agent capacity

These options are discussed in further detail in Chapter 7.

5.4 Importance of customer income on first debit collection

The picture painted by the predictive model for successful first debit collection is a much more intuitive one. From the ranking of model features (Figure 32), the two most predictive features of the model correspond to monthly income and the actual monetary value of the monthly debit order - neither of which should be surprising. It is intuitively expected that the higher a person's monthly income is, the more likely they will be to reliably meet their various financial obligations and successfully pay for new purchases and services. It is similarly intuitive, that the lower the monthly premium required for a financial product or service, the more likely a person will be able to meet their obligation (irrespective of their monthly income) – this intuition is confirmed by the billed premium being the second most predictive feature identified by the predictive model (Figure 32). Most importantly though, it is noted that little control of over either of these features is possible.

As noted in Section 1 in chapter 4 (4.1) in the case of leads conversion, there will generally be a range of monthly incomes in a given lead set and it is often unfeasible to restrict this feature to a narrower range of values (that is, it may prove unsustainable to only use lead sets with leads from higher incomes).

5.5 Importance of monthly premium

With regards to monthly premium, there is also limited control available. In the non-underwritten funeral plan space, the primary driver of the monthly premium is the plan cover amount (that is, the actual sum assured for the insured person) which is selected by the customer during the course of the sales call. A secondary driver of the premium may be the age of additional lives the customer would like to add to their plan (such as

parents, children or other relatives) – the monthly premiums for these additional lives are a function of the cover amount, as well as the age of the particular insured live. It is clear that it is not practically possible to know in advance what cover amounts a given lead will select, or which (if any) additional lives they will elect to add to the plan (or the ages of these additional lives if added). As such no material changes can be made to influence this feature short of repricing all the different funeral plan cover amounts which would entail far-reaching commercial and strategic considerations, as well as significant actuarial modelling exercises and accompanying risk analyses – all of which are beyond the scope of this research.

5.6 Importance of time-based factors in first debit collection

The next most predictive feature, however, is one that is able to be influenced directly by the call centre processes and thus open to modification. This third-most predictive feature is labelled in the model as "Call_to_Bill_CAMPAIGN_MONTH" and is one of the features that was generated by the model as part of the feature engineering process. This feature is a hybrid between "Call_to_Bill" (the feature in the model that describes the number of days between the date of call and first debit) and "CAMPAIGN MONTH" (the actual month of the year in which the customer was called). This suggests there is a significant impact to the amount of days that elapse between the sales call and first debit, and that this is a key driver of the behaviour observed in relationship between this difference and the rate of first debit collections (see Figure 21). The additional dimension of sales month comes in to play as the different number of days in each month, as well as the presence of the many public holidays spread throughout the South African calendar, may also contribute to the ease or difficulty of collecting a premium. As illustrative examples, debit orders are often changed – both by the insurer and many other of the customer's debtors – in December in an effort to collect before holiday spending. Similarly, in earlier months of the year, customers are often catching up on the impact of significant holiday spending and preparing for the new year (which may include new expenses such as annual increases on previous debit orders and preparing for the new school year), which may put customers under significant financial strain.

6. CHAPTER 6: CONCLUSIONS

6.1 Introduction

This chapter also serves to highlight how the original research objectives of assessing the impact of particular time-based factors on outbound life insurance sales were met. Specific focus was provided on the following objectives:

- The determination of the relationship between the proximity of an outbound insurance sales call to a customer's payday and the likelihood of a customer to take up an insurance offer from a call centre
- The establishment of the extent of the relationship between the time from the insurance sale through the call centre to debit date and the likelihood of a policy to successfully collect on the first attempted debit

The final objective of proposing recommendations for improvements to the end-end insurance call centre sales process through the use of these relationships is discussed in detail in Chapter 7.

6.2 The impact of time-based factors in the outbound insurance value chain

The end-to-end value chain for outbound life insurance sales in call centres is one that has in-built time-based factors that can affect its efficiency as a distribution channel. This research demonstrated that two instances of these time-based factors had significant impact on the overall of efficiency of this value chain with the potential for improving the cost efficiency of the operations of the life insurer and the broader financial services provider of which it is a part. These two time-based factors are:

- The number of days between the date a lead receives their salary and the date that they are called to be presented an insurance sales offer
- The number of days between the date of a successful insurance sale and the date of first debit collection

This research analysed the impact of these two above-mentioned time-based factors. The significance of this research lies in the fact that there appeared to be a relative gap in the available literature on consumer behaviour with regards to service-based offerings (as opposed to tangible products), even more so in the case of financial products and even more pronounced in the specific domain of life insurance products where the customer will likely never see the benefit of their purchase in their lifetime.

Therefore, for this research, these two key relationships were investigated for the historical leads of a life insurer's non-underwritten funeral plan portfolio over the course of a year in order to identify opportunities for improving efficiency within the end-to-end life insurance outbound call centre sales value chain. All the objectives of this research were completed, and a clear pattern was observed for each relationship. Based on the analyses conducted, several conclusions and recommendations emerged.

6.3 The impact of days difference between salary receipt and date of sales call

One of the main time-based factors considered by call centre managers is that of the proximity to month end (a crude proxy for when customers receive their monthly salaries). This factor is used to inform a heuristic that predicts whether the performance in call centres, as measured using sales-to-lead conversions, will be higher at the beginning of a month, closer to the notional payday of the average customer, or lower at the end of the month, further away from payday and when disposable income is at its lowest.

This research demonstrated that while there is some merit to the broad generalisation that sales conversion rates tend to be higher at the beginning of the month and lower at the end, there is an important caveat that needs to be taken in to account. An initial nadir in conversion rates occurred for the first days of a month (a period between the first four and six days of a month), which was both counterintuitive and diametrically opposed to the rule of thumb used by the call centre managers – not only was it not the best time in the month for sales conversions rates, but it was actually the worst (often as bad, if not worse than the end of the month).

For the general trend analysed, independent of more granular income band segmentation, the difference in sales-to-lead conversion increased from a minimum of 8.8% at the start of the month to a maximum of 13.3% six days later, representing an improvement of 51.1% (conversion rates decreased to a low of 10.1% just before the end of the month). The impact of this significant improvement can be appreciated when considering that the insurer typically submits lead batches in the order of thousands of leads per batch per month. By just optimising the timing of contacting particular customers, a significant reduction in the "wasting" of leads (that is, leads that do not successfully convert into a sale).

Consequently, call centre managers and sales analysts can improve the efficiency of their sales processes and operations by taking this finding in to account instead of relying on legacy heuristics that may actually the compromise the success of their operations.

6.4 Validation of first debit rate as a metric for collections analysts

Another example of time-based factors presents itself in the space of operations for collecting monthly premiums. Various metrics are considered by the insurer's collections teams in order to predict which policyholders will remain on book and how to reduce the likelihood of customers missing their premiums payments which ultimately result in costly collections and retentions exercises.

As part of this research, a survey was circulated amongst the collections analysts to determine which of the metrics they track is most pertinent in managing off-book movements. While this survey was not a key component of the overall research, it was vital to establish whether first debit rate was the correct metric to analyse in subsequent critical components.

The results of the survey indicated the most pertinent of the metrics tracked by the analysts is that of first debit collection rate, thus validating the premise of the collections-related portion of the research. The question then emerged of when the

optimal time is to attempt this collection of the first debit for a new policy on the insurer's book.

6.5 The impact of days difference between date of sale and date of first debit collection

This research demonstrates a trend for successful first debit collections related to the amount of time that has elapsed since the date of sale. Generally speaking, as the number of days increase from the date of the sale, the likelihood of a successful first debit collection decreases, with the best time to collect being on the date of sale (that is, "zero" days since the date of sale)

In the case of likelihood to successfully collect the first premium, the analysis yielded a more roughly monotonic, linear relationship compared to the leads conversion analysis. In the general income-independent trend, the average first debit collection rates dropped from a high of 78.6% at day zero, to a significantly lower value of 68.8% at day 30 (the furthest possible date from date of sale). This represents a decrease of 14.2% in likelihood to successfully the first debit.

While this difference in rate (in absolute terms) at either end of the spectrum isn't as startling as in the case of leads conversion, in the context of the hundreds or thousands of new sales made by the insurer each month, this drop in collections rate can result in an extremely significant financial impact due to the wasted acquisition costs, lost revenue and second order costs associated with further collections and retentions activities.

6.6 The use of LightGBM-based predictive models to validate the impact of time-based factors

In order to determine if these time-bound factors are in fact driving these patterns of customer behaviour for both sales conversions and first debit rates, predictive models were created using the LightGBM framework in the Python programming language to

develop a ranking of the various variables (or factors) in terms of their contribution to predicting a successful outcome for either lead conversion or successful debit order collections. As part of the upfront preparatory steps, correlation analyses were conducted to ensure that there was no meaningful entanglement of variables. The analyses showed that no significant correlations were found between the model variables in either case.

In both cases the predictive models were found to be sufficiently accurate in making predictions for each analysis, demonstrating that the use of the LightGBM framework is appropriate for the analyses conducted.

Both predictive models ranked the time-bound factors highly relative to the other model features, showing that they are strongly predictive of the model target outcomes (that is, whether a lead converts into a sale in the first analysis, and whether a first debit will successfully collect in the second). This implies they significantly contribute to the observed behavioural patterns.

While the time-bound factors were not the single most predictive feature in either case (this top spot is occupied by average monthly income in both models), the time-based factors ranked as the third most predictive features for each model, with the importance of the lower ranking features decreasing at varying rates beyond this. As a result, it can be concluded from this research that time-based factors do have significant impact and potential commercial application to the life insurance sales value chain in the context of outbound call centres.

6.7 The impact of monthly income

It was noted that for both relationships analysed in this research, average monthly income was ranked as the most significant factor in predicting both the likelihood of a successful sales conversion and the likelihood of a successful first debit collection.

Customers with higher monthly incomes (approximately R20 000 monthly income or higher) appear to be marginally less sensitive to the impact of these time-bound factors. This reduced sensitivity may be attributed to their better financial resilience and larger disposable income for purchasing and maintaining new services such as life insurance.

Since the monthly income of leads is difficult to control (or at least inadvisable due to both practical operational and market conduct considerations), the time-based factors become the most viable aspects of the end-to-end value to chain to optimise through modifications to their respective operational processes.

6.8 Next steps arising for improving the efficiency of the outbound sales process for life insurance sales

Given the findings of this research, several modifications could be made to the supply and delivery of the monthly lead files by the insurer's sales analytics teams, as well as the sales process for the call centre agents. These modifications, as well as suggested avenues for furthering this line of research, are discussed in detail in Chapter 7.

A first modification that can be concluded from the research entails sending multiple smaller batches of leads throughout the month instead of one larger file delivered at the start of the month. This modification is also aligned with principles of lean manufacturing whereby reducing batch sizes to be worked in production processes typically yields higher quality throughput for the process (Bicheno & Holweg, 2016).

An additional modification that can be concluded from the research entails the call centre optimally selecting the timing and processing of leads to increase conversions later in the batch cycles.

Further, it could be concluded that a modification cold be made to the actual sales process using the choice architecture concepts of defaults to improve the average first debit collection rates whereby agents suggest the default collection date of the first debit

order to be made on the date of the sale (as opposed to the current process whereby the customer is prompted to suggest a date).

Finally, it was also noted all these modifications could be combined to achieve greater synergy. If implemented, these modifications have the potential to yield improved efficiencies to the end-to-end insurance value chain for the Insurer's outbound sales processes and ultimately reduce the cost-to-income ratio for the Insurer and FSP as a whole.

7. CHAPTER 7: RECOMMENDATIONS

7.1 Introduction

Following the findings, resulting discussions and conclusions from the previous Chapters, recommendations are made for the insurer (both for individual areas as well as the overall outbound call centre value chain) and for further possible research.

7.2 Recommendations for management of call centre and agents

Two possible avenues for addressing performance in call centres were highlighted in section 5.2. These pertained to new incentive structures and exploring alternative capacity planning schemes.

7.2.1 Revised incentive structures

The first option for tackling the call centre agents' possible dissatisfaction (and resulting lack of motivation) is to counteract the feeling of not being appropriately remunerated with additional, alternative remuneration policies. These policies should not be directly linked to ongoing performance and be more of a "spot prize" nature for rapid initial displays of engagement. Illustrative examples of this approach could be to offer a top up incentive for the first sale made by a call centre agent in a particular team or to offer increased commission for the first few hours of the first few days of every sales month. Optimally, the additional incentives should be a non-monetary incentive (for example, a gift or at least a gift voucher) in order to avoid creating an obvious show of appeasement for any perceived inadequacy in monthly remuneration.

While this approach does have its obvious merits in directly addressing and counteracting one of the root causes of reduced early-month engagement, it does have a dark side. By "fighting fire with fire" (as it were), call centre agents may become conditioned to only respond to incentives (monetary or otherwise) as a stimulus for modifying or improving sales performance, potentially leading to a vicious circle of

increasing compensation. Another potential pitfall with this approach is that incentives can lose their efficacy in stimulating ongoing engagement when they become too predictable (Eyal, 2014), requiring ongoing monitoring, evaluation and reformulation – all of which require additional man-hours which can result in an increase in staff costs, or at the very least, a decrease in efficient use of existing staff's capacity.

Such a recommendation may likely result in increased overhead costs associated with running the call centre and maintaining stable performance. This potential spiral of increasing costs could ultimately undo the good the additional incentives would be designed to alleviate and runs the risk of actually increasing the cost base of the outbound call centre value chain (or at least, return only marginal gain for relatively significant effort).

7.2.2 Alternative capacity planning

The second strategy noted is that of taking a different approach to allocating the activities and capacity of the call centre's agents. Essentially, this strategy takes the approach that instead of continuing sales activities in the first few days of the months when conversions are known to be at their lowest points, to rather not conduct sales at all in these days and reallocate those leads to later in the month when conversions are likely to be significantly higher. The question that remains then is how to best utilise the agents' capacity during the first few days of the months.

The immediately apparent option is to simply not have the call centre options work on these days drastically reducing overheads during that time. However, as simple as this approach seems, especially given the changing nature of ways of working and the rise of the on-demand economy, it does lend itself to introducing new human resources issues. Firstly, by relying on part-time or occasional staff to facilitate this arrangement, it may be difficult to maintain a stable staff contingent in the call centre with a predictable number of agents at any given time. Having a relatively stable quantum for the call centre's agents is a key criteria in making sure that leads do not go wasted (that is, too many leads supplied for a month with too few agents to effectively work them all) and to prevent idle agents attracting unnecessary seat costs (that is, too many agents

without enough leads to keep them occupied while they are in the call centre). Both of these scenarios inevitably lead to a reduction in cost efficiency by losing out on the potential value of unworked leads in the under-capacity scenario and paying excess staff overheads in the over-capacity scenario.

Further, rolling forces of agents require training to ensure they are suitably capable to conduct the sales calls. Besides the costs (both financial and time) associated with these ongoing training activities, poorly trained or unnecessarily inexperienced agents would be prone to lower conversions on their leads worked due to their lack of consistent engagement with the product and campaigns. Again, this could result in lower cost efficiencies associated with the campaigns.

7.2.3 Dynamically assigning call centre agents to different activities

Given these considerations, another approach may be to then reassign the existing full-time staff's (or at least portions of the full team's) time during the first few days of the month to other value adding activities such as customer service and complaint resolution. This approach allows the agents to move past their initial performance trough of dissatisfaction while still being kept busy with productive activities.

This may require significant training to upskill the agents to perform servicing activities and effectively deal with customer queries, but the costs associated with this then become an upfront cost that can yield significantly more value and engagement over time as opposed to adding incentive structures which increase costs on an ongoing basis with diminishing returns over time. It is, however, important to note this specific approach to the capacity planning is not always feasible and susceptible to the less predictable volumes of customer queries that arrive early in the month (that is, without being able to know in advance how many queries will come in – as opposed to knowing exactly how many leads will be supplied – it is difficult to consistently estimate how many agents need to be assigned to servicing functions). Finally, if not all the agents are able to engage in other value-adding activities then the problem of agents working fewer leads during a time of lower expected conversions remains (albeit at a reduced scale).

7.3 Recommendations for the supply of monthly leads

A final approach that can be considered for reallocating the time of the call centre agents is to more thoroughly regulate the leads worked at different times during the month. While the analysis showed that in order to be most effectively converted into sales the time between payday and date of call must be minimized, an important secondary finding was noted. For higher income customers, the impact of time-bound factors on outbound insurance sales was present, however, the effect was relatively less pronounced.

Combining these two findings, an effective strategy would then consist of two modifications to the monthly supply of leads to the call centre agents.

7.3.1 Recommendations for sales analytics teams

Monthly batches of leads can be divided into smaller batches in an attempt to maximise the overall conversion per batch. The batches would primarily need to be split such that the maximum difference in days between date of call and salary date per lead in each batch is minimised.

In the general trend (Figure 11) and the income-specific drill downs (Figures 13-19), the peak conversion after the initial trough occurred approximately at calls made on the sixth day after salary receipt with the average conversion starting to drop for calls made after approximately 18 days from salary receipt. Practically speaking this would mean optimising delivery of leads batches so they only contain leads with a days difference of at least 6 and not more than 18 for days from possible date of call.

In order to achieve this first modification, there are two practical considerations to take into account:

- In order to meet monthly leads volumes, analysts perform significant work to ensure that both monthly lead volumes and the overall quality of the leads are maintained. This newly proposed delivery scheme would therefore require the analysts to perform this exercise an additional time for each lead batch required.
- Secondly, the dates of salary dates are not evenly distributed throughout any given month – the bulk of salaries are received between the 24th day of the month until the first day of the next month (See Figure 5). This would mean the batches would probably then have to be split asymmetrically in order to keep the days difference within the desired range.

Given these practical concerns, a pragmatic recommendation for the Insurer's sales analytics team would then be to split the monthly leads batch in to just two smaller batches. By keeping the number of splits low, the amount of additional work placed on the analysts (as well as the logistics required to be performed by the call centre loading and allocating the leads batches) is kept to a minimum. Secondly, if the batches are supplied to the call centre at the beginning of the month and the middle of the month it can be ensured that at least no lead will have a difference of more than 15 days from any possible date of call.

7.3.2 Recommendations for leads consumption by agents

The result of this newly proposed split in batches of leads is that more leads may be allocated to the first batch and fewer to the second batch. This asymmetry can however be compensated for by limiting the number of allowed call attempts ("passes") to a lower number (as an illustrative example, no more than three passes for the first leads batch) and allowing a higher number of passes for the second batch. Alternatively, where practical, agents can be upskilled and assigned to also assist with burden of servicing calls during the second half of the month or be assigned to other outbound campaigns.

A second key modification to the leads supply would consist of keeping higher income leads in each batch for later in the batch cycle (as opposed to randomly sampling every day). The higher income leads still exhibited early minimum average conversion rates with days difference between salary receipt and date of call between zero and four days but showed more stable average conversions as the difference grew larger (see Figures 17-19). Therefore, by keeping the higher income leads (leads with approximately R20 000 monthly income or higher) for later in the batch when this difference is likely to be at a maximum, a hedging effect can be created to keep conversions at a stable, higher rate.

7.4 Recommendations for enhancements to collections processes

The relationship between first debit collection rates and time between sales call and first debit can be summarised as follows: the fewer days between a sale and the first debit, the more likely the first debit will successfully collect (and as a result, the more likely that the policy will be maintained over the long term). This relationship was clearly seen in the analysis on first debit collection rate, with the optimum time for collecting the first policy premium being on the date of sale (or as close to as possible). The rates of collection then deteriorate as the difference between date of call and first debit day increases (see Figures 21-28). A change to the sales process can therefore be proposed that is practically simpler than that of splitting the leads batches, but perhaps no less impactful to the overall sales value chain for the insurer.

7.4.1 Defaulting of debit order dates

Instead of the status quo process of asking the customer what day of the month to set the regular debit (including the first debit), a default date can instead be presented to customer with the option to change it if they desire. The default date first offered would be for the first debit to be debited the same day as the sale. This change is a classic example of choice architecture (popularly referred to as “nudges” in the behavioural economics context) utilising the significant power of defaults in helping people and organisations achieve better mutual outcomes, (a successful first collection ensures the customer will have the best chance of remaining covered in the long term, while the

insurer benefits by expending less efforts on retaining non-collecting policyholders), as in other well-publicised examples such as defaulting rates for pension savings schemes (R. Thaler & Sunstein, 2008). The implementation of this change would only require minor changes to the call centre agents' scripts and training to agents already accustomed to the current system.

7.5 Combined improvements to the value chain

The advantage of the modifications proposed for improving leads conversion and first debit collection rate is that they are somewhat independent of each other in the broader value chain (see Figure 1). Therefore, it is possible to implement both sets of modifications simultaneously to the value chain – monthly lead sets can be split in to smaller, more workable batches and optimised for conversion (including basing the timing consumption of leads on monthly income) and sales processes can be adapted to default the first collection of the monthly debit order to be on the day of the sale.

While it is expected that combining these two sets of modifications will result in a linear net improvement to the efficiency of the overall value chain, more work is required to test and validate the individual contributions of the proposed modifications in the real world, as the well analysis to determine if any unexpected interplay exists between these changes to determine if combing the changes result in performance that is more (or less) than the sum of their parts.

7.6 Recommendations for future work

While the key objectives of this investigation were met, resulting in several practical recommendations for improving the end-to-end value chain of the Insurer's outbound call centre, several avenues are available for furthering this line of research. These avenues were beyond the scope of this initial research but may yield additional efficiencies and scope, leading to a greater overall reduction in cost-to-income ratio for the broader FSP.

Recommendations for further work includes:

- *Validation of recommendations made:* Based on the findings of this research, several modifications are proposed to the operations of the insurer's analytics team's process and to the call centre's operations. These modifications need to be validated by implementing them for a pilot test group (along with a well-calibrated control group) of call centre agents. The results of this pilot group would then need to be compared with the control group for the same period and profile of leads in order to determine if the proposed validations have merit in the practice. This validation was not practical in the timeframe of this research due the operational complexities involved, which were exacerbated by the global pandemic which required many companies, including financial services providers (and insurers in particular) to focus on maintaining operational stability and the ability to keep pace with rising tides of various claims. Additionally, many organisations, including both FSPs and call centres, have had to adjust to radically different ways of working with decreased appetite for speculative new changes to process that could compromise operational stability. Finally, it is also possible that consumer attitudes and behaviour to various financial services and products, particularly life insurance, may have drastically changed during the pandemic (this final consideration may constitute a research topic in and of itself).

- *Replicate with other data:* This research could be further validated by employing the same methodology but with different data samples (for example, other time periods, either before or after the period considered by this study,

- *Expansion of study to other financial products:* Similar investigations to the ones performed in this research could be conducted for other insurance products, including other non-underwritten products (such as credit life insurance), underwritten products (traditional life insurance for a range of benefits that include individual assessments of policyholders' medical particular demographics and medical wellbeing), short-term insurance products (for example car insurance and homeowners insurance), other financial products (for

example, transactional cheque accounts or investment vehicles) and other regular service offering (for example, subscriptions for monthly cell phone data)

- *Remuneration policies for outbound call centres:* the timing and structure of remuneration of call centre agents was suggested as a possible cause for the early-month trough in leads conversions. The required investigation of this is a significant undertaking and beyond the scope of this research but is likely to be of great commercial interest to managers of agent-based distribution channels in a variety of contexts. Additionally, alternative remuneration schemes (for example, just-in-time payments on completion of successful activities – like making a sale) could also be explored as another avenue to create more stable performance in call centres throughout any given month.

8. REFERENCES

Avnet, T., & Higgins, E. T. (2003). Locomotion , assessment , and regulatory fit : Value transfer from ‘“ how ”’ to ‘“ what .”’ *Journal of Experimental Social Psychology*, *39*, 525–530.

Barsotti, F., Milhaud, X., & Salhi, Y. (2016). Lapse risk in life insurance : correlation and contagion effects among policyholders ’ behaviors. *Insurance: Mathematics and Economics*, (71), 317–331.

Batt, R. (2002). Managing customer services: Human resource practices, quit rates, and sales growth. *Academy of Management Journal*, *45*(3), 587–597.

Benner, C., Lewis, C., & Omar, R. (2007). *The South African Call Centre Industry :* (June). Retrieved from https://www.wits.ac.za/media/migration/files/South Africa GCC Report.pdf

Bicheno, J., & Holweg, M. (2016). *The Lean Toolbox: a handbook for lean transformation* (5th edition). Buckingham: PICSIE Books.

Binder, S., & Mußhoff, J. (2017). *McKinsey Global Insurance Pools-seventh edition: 2017 Global Insurance Industry Insights An in-depth perspective*. Retrieved from https://www.mckinsey.com/~/media/mckinsey/industries/financial services/our insights/global insurance insights an overview of recent trends in life p and c and health/global-insurance-industry-insights-an-in-depth-perspecti

Brown, L. D., & Gans, N. (2005). *Statistical Analysis of a Telephone Call Center : A Queueing-Science Perspective Statistical Analysis of a Telephone Call Center : A Queueing-Science*. *100*, 36–50.

Bundorf, M. K., & Pauly, M. V. (2002). *Is Health Insurance Affordable for the Uninsured* (No. 9281).

Burges, C. J. C. (2010). From RankNet to LambdaRank to LambdaMART : An Overview. *Microsoft Research Technical Report*, *MSR-TR-201*.

Chaudhary, S., & Kaur, J. (2016). Consumer Perception Regarding Life Insurance Policies: a Factor Analytical Approach. *International Journal of Information Movement*, *1*(III), 30–40.

Critchfield, T. S., & Kollins, S. H. (2001). Temporal discounting: basic research and the analysis of socially important behavior. *Journal of Applied Behaviour Analysis*, *34*(1), 101–122.

D’Haen, J., Van den Poel, D., & Thorleuchter, D. (2012). *Predicting Customer*

Profitability During Acquisition : Finding the Optimal Combination of Data Source and Data. Ghent University.

De Bock, K. W., & Van den Poel, D. (2012). Reconciling Performance and Interpretability in Customer Churn Prediction using Ensemble Learning based on Generalized Additive Models. *Expert Systems with Applications*, *39*(8).

Dolan, P., Hallsworth, M., Halpern, D., King, D., Metcalfe, R., & Vlaev, I. (2012). Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, *33*(1), 264–277. https://doi.org/10.1016/j.joep.2011.10.009

Ersner-hershfield, H., Wimmer, G. E., & Knutson, B. (2009). Saving for the future self : Neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience*, *4*(1), 85–92.

Eyal, N. (2014). *Hooked: How to Build Habit-Forming Products* (1st ed.). Portfolio Penguin.

Feinberg, R. A., Kim, I., & Hokama, L. (2015). *Operational determinants of caller satisfaction in the call center*. (January).

Fier, S. G., & Liebenberg, A. P. (2012). *Life Insurance Lapse Behaviour*.

Foreman, J. W. (2014). *Data Smart: Using Data Science to Transform Information into Insight* (First). Indianoplis, Indiana: John Wiley & Sons.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Grus, J. (2015). *Data Science from Scratch: First Principles with Python* (First). Sebastopol, CA: O'reilly Media, Inc.

Gummesson, E. (1979). The Marketing of Professional Services — An Organisational Dilemma. *European Journal of Marketing*, *13*(5). https://doi.org/10.1108/EUM0000000004951

Hauser, J. R., & Katz, G. M. (1998). *Metrics : You Are What You Measure !*

Huffman, D., & Barenstein, M. (2005). A Monthly Struggle for Self-Control ? Hyperbolic Discounting , Mental Accounting , and the Fall in Consumption Between Paydays. *Institute for the Study of Labor (IZA) Discussion Paper*, *1430*.

Joia, L. A., Oliveira, A. F., & Oliveira, A. F. (2010). *Call Center Key Performance Indicators and Customer Satisfaction*.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, *47*(2), 263–292.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T. (2017).

LightGBM : A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems*, (Nips), 1–9. Long beach, CA, USA.

Kinnie, N., Hutchinson, S., & Purcell, J. (2000). "Fun and surveillance": The paradox of high commitment management in call centres. *International Journal of Human Resource Management*, *11*(5), 967–985. https://doi.org/10.1080/095851900422375

Kumar, V., & Reinartz, W. J. (2006). *Customer Relationship Management: A Databased Approach*. Hoboken, N.J.: John Wiley & Sons.

Kunreuther, H., & Pauly, M. V. (2005). Insurance Decision-Making and Market Behavior. *Foundations and Trends in Microeconomics*, *1*(2), 63–127.

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Mckechnie, S. (1992). Consumer Buying Behaviour in Financial Services : *International Journal of Bank Marketing*, *10*(5), 4–12.

Mishra, H., Mishra, A., & Nayakankuppam, D. (2010). How Salary Receipt Affects Consumers ' Regulatory Motivations and Product Preferences. *Journal of Marketing*, *74*(September), 93–103.

Monat, J. P. (2011). A Good Lead is Hard to Find. *Marketing Research*, *23*(2), 18–23. Retrieved from http://0-search.ebscohost.com.innopac.wits.ac.za/login.aspx?direct=true&db=bsu&AN=67267156&site=ehost-live&scope=site

Muller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python: a Guide for Data Scientists* (First edit). Sebastopol, CA: O'reilly Media, Inc.

Naji, A. (2014). Components of Remuneration and Employee Satisfaction: The Impact of Effort Rewards and Career. *International Journal of Arts and Sciences*, *07*(02), 425–436.

Parrish, S. (2018). The Life Insurance Agent as Financial Planner. *Journal of Financial Services Professionals*, (March), 39–43.

PWC. (2016). *Insurance through challenging times: Insurance industry analysis*. Retrieved from www.pwc.co.za/insurance

Russell, B. (2008). *Call centres : A decade of research*. *10*(3), 195–219. https://doi.org/10.1111/j.1468-2370.2008.00241.x

Shapiro, M. D., & Slemrod, J. (1993). Consumer response to the timing of income: Evidence from a change in withholding tax. *National Bureau of Economic Research*.

Shostack, G. L. (1977). Breaking Free from Product Marketing. *Journal of Marketing*, *41*(2), 73–80.

Soman, D., Ainslie, G., Frederick, S., Li, X., Lynch, J. G., Moreau, P., … Zauberman, G. (2005). The Psychology of Intertemporal Discounting : Why are Distant Events Valued Differently from Proximal Ones ? The Psychology of Intertemporal Discounting : Why are Distant Events Valued Differently from Proximal Ones ? *Marketing Letters*, *16*(3/4), 347–360.

Stephens, M. J. (2002). Paycheck receipt and the timing of consumption. *National Bureau of Economic Research*, (November).

Swaen, B. (2021). Constructing a conceptual framework. *Scribbr* website: scribbr.com/methodology/conceptual-framework/ (accessed: 30 September 2021)

Swank, C. K. The Lean Service Machine. , Harvard Business Review (2003).

Thaler, R. (2008). Mental Accounting and Consumer Choice. *Marketing Science*, *27*(1), 15–25.

Thaler, R. H. (1999). Mental Accounting Matters. *Journal of Behavioral Decision Making*, *12*, 1983–206.

Thaler, R., & Sunstein, C. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. London: Penguin Books.

van den Broek, D., Barnes, A., & Townsend, K. (2008). "Teaming up": Teams and team sharing in call centres. *Journal of Industrial Relations*, *50*(2), 257–269.

van der Merwe, J., & Gerin, J. (2018). *Emerging markets: Growing insurance & challenges with a focus on Africa*. Cape Town.

von Neumann, J. (2016). Model selection and overfitting. *Nature Methods*, *13*(9), 703–704.

Yan, J., Gong, M., Sun, C., Huang, J., & Chu, S. M. (2015). *Sales pipeline win propensity prediction : a regression approach*.

Zeldes, S. P. (1989). Consumption and Liquidity Constraints: An Empirical Investigation. *Journal of Political Economy*, *97*(2), 305–346.

APPENDIX A: SURVEY FOR COLLECTION ANALYSTS

Survey for research entitled: The impact of time-based factors on outbound life insurance sales in call centres

(self-completed)

Administrative (Don't fill in):

Candidate number:

Date of interview:

Candidate experience:

1. What is your position within your organisation?

2. How long have you worked in this position?

0-12 months		**1-2 years**		**2-5 years**		**5-8 years**		**8+ years**	

3. Do you have any experience in analytics related to collections and/or customer retentions?

Select with "**X**":

YES		**NO**	

4. If **NO**, please describe your experience (previous and current) in analytics, if any:

5. Do you have any experience in analytics (collections, retentions or otherwise) in the field of life insurance?

Select with "**X**":

YES		**NO**	

6. If **YES,** please specify how many years/months of experience you have in analytics in life insurance:

Select with "**X**":

0–6 months		**6-12 months**		**1-2 years**		**2-4 years**		**4+ years**	

Metrics collected and analysed

7. As an analyst (current or former) do you or your team monitor and/or analyse any of the following metrics:

Select with "**X**":

First debit rate		**Lapse rate**		**Cancellation rate**		**Averaged debit rate**	
Arrears rate		**First collection rate**		**Other**		**None of these metrics**	

8. If any of these are not monitored or analysed, what is the reason for not monitoring or analysing them?

9. Of the metrics mentioned that are monitored and/or analysed, which do you believe are the strongest predictors of whether of off-movements (that is lapses and cancellations) for life insurance (on a total book level)?

Indicate **1/2/3** next to best predictor (1), second choice (2) and third choice (3) :

Unsure/not my area of expertise (only indicate 1 for this and leave the other options blank)		**First debit rate**		**Arrears rate**		**First Collection rate**		**Other**	

10. Are there any other metrics that are monitored and/or analysed that you believe have greater ability in predicting overall off-movements for a life insurance book?

Select with "**X**":

YES		**NO**	

11. If **YES** to question 10, what are these metrics and why do you believe they provide greater predictive power for predicting off-movements for life insurance?

APPENDIX B: MODEL CODE

B.1 Python code for leads conversion model

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import lightgbm as lgb
from sklearn import metrics
from sklearn.preprocessing import LabelEncoder

ks = pd.read_csv("Copy_of_S…a.csv")

#Drop uncontactables
ks = ks.query('CONTACTEDINCAMPAIGN != "N"')
#Convert target values to 1/0
ks = ks.assign(outcome=(ks['CONTACTOUTCOME']=='POLICY SOLD').astype(int))

#Label encoding
cat_features = ['DDA_SEG','CUST_SEX_CDE','CAMPAIGN_MONTH']
encoder = LabelEncoder()
encoded = ks[cat_features].apply(encoder.fit_transform)

#Joining to required fields

#DEBIT_ORDER DAY -> moved here
#maybe drop attempted contacts as you do not know that before calling the
customer
data_cols = ['AVG_MTH_INCOME','DEBIT_ORDER_DAY','Contacted_Date', 'DIFF',
'outcome']
baseline_data = ks[data_cols].join(encoded)

#Correlation matrix plotting
plt.figure(figsize=(12,10))
sns.heatmap(baseline_data.corr(), annot=True, cmap='RdBu')
plt.show()

#Functions to help test encoding

def get_data_splits(dataframe, valid_fraction=0.1):
    valid_fraction=0.1
    valid_size = int(len(dataframe)*valid_fraction)

    train = dataframe[:-valid_size*2]
    valid = dataframe[-valid_size*2:-valid_size]
    test = dataframe[-valid_size:]

    return train, valid, test

def train_model (train, valid):
    feature_cols = train.columns.drop('outcome')

    dtrain = lgb.Dataset(train[feature_cols], label = train['outcome'])
    dvalid = lgb.Dataset(valid[feature_cols], label = valid['outcome'])
```

```
    param = {'num_leaves': 64, 'objective': 'binary', 'metric': 'auc',
'seed': 7}
    #Delete this next line
    print("Training model!")
    bst = lgb.train(param, dtrain,
num_boost_round=1000,valid_sets=[dvalid], early_stopping_rounds=10,
verbose_eval=False)

    valid_pred = bst.predict(valid[feature_cols])
    valid_score = metrics.roc_auc_score(valid['outcome'], valid_pred)
    print(f"Validation score: {valid_score:.4f}")

    return bst

#training model on baseline data:
train, valid, _ = get_data_splits(baseline_data)
bst = train_model(train,valid)

#Feature generation
import itertools

cat_features = ['DDA_SEG','CUST_SEX_CDE','CAMPAIGN_MONTH']
interactions = pd.DataFrame(index=ks.index)

for col1,col2 in itertools.combinations(cat_features, 2):
    new_col_name = '_'.join([col1,col2])
    #convert to strings and combine
    new_values = ks[col1].map(str) + "_" + ks[col2].map(str)
    label_enc = LabelEncoder()
    interactions[new_col_name] = label_enc.fit_transform(new_values)

baseline_data = baseline_data.join(interactions)
baseline_data.head(20)

train, valid, _ = get_data_splits(baseline_data)
bst = train_model(train,valid)

ax = lgb.plot_importance(bst, importance_type="split", max_num_features=59,
figsize=(15,15))
```

B.2 Python code for first debit collection model

```
import pandas as pd
import lightgbm as lgb
from sklearn import metrics
from sklearn.preprocessing import LabelEncoder

import matplotlib.pyplot as plt
import seaborn as sns

ks = pd.read_csv("Debit Or…S_5.csv")

#Drop uncontactables, only keep successful sales
ks = ks.query('CONTACTEDINCAMPAIGN != "N"')
ks = ks.query('CONTACTOUTCOME != "POLICY NOT SOLD"')
ks = ks.query('Billed != 0')

ks = ks.assign(outcome=(ks['SuccessfulCollection']))

#Label encoding
cat_features = ['DDA_SEG','CUST_SEX_CDE','CAMPAIGN_MONTH']
encoder = LabelEncoder()
encoded = ks[cat_features].apply(encoder.fit_transform)

#Joining to required fields
data_cols = ['Billed','AVG_MTH_INCOME','DEBIT_ORDER_DAY','Contacted_Date',
'PaymentDay', 'Call to Bill', 'Billed Premium', 'outcome']
baseline_data = ks[data_cols].join(encoded)

#Correlation matrix plotting
plt.figure(figsize=(12,10))
sns.heatmap(baseline_data.corr(), annot=True, cmap='RdBu')
plt.show()

#Functions to help test encoding

def get_data_splits(dataframe, valid_fraction=0.1):
    valid_fraction=0.1
    valid_size = int(len(dataframe)*valid_fraction)

    train = dataframe[:-valid_size*2]
    valid = dataframe[-valid_size*2:-valid_size]
    test = dataframe[-valid_size:]

    return train, valid, test

def train_model (train, valid):
    feature_cols = train.columns.drop('outcome')

    dtrain = lgb.Dataset(train[feature_cols], label = train['outcome'])
    dvalid = lgb.Dataset(valid[feature_cols], label = valid['outcome'])

    param = {'num_leaves': 64, 'objective': 'binary', 'metric': 'auc',
'seed': 7}
    #Delete this next line
    print("Training model!")
```

```
    bst = lgb.train(param, dtrain,
num_boost_round=1000,valid_sets=[dvalid], early_stopping_rounds=10,
verbose_eval=False)

    valid_pred = bst.predict(valid[feature_cols])
    valid_score = metrics.roc_auc_score(valid['outcome'], valid_pred)
    print(f"Validation score: {valid_score:.4f}")

    return bst

#training model on baseline data:
train, valid, _ = get_data_splits(baseline_data)
bst = train_model(train,valid)

#Feature generation
import itertools

cat_features = ['Call to Bill', 'DDA_SEG','CUST_SEX_CDE','CAMPAIGN_MONTH']
interactions = pd.DataFrame(index=ks.index)

for col1,col2 in itertools.combinations(cat_features, 2):
    new_col_name = '_'.join([col1,col2])
    #convert to strings and combine
    new_values = ks[col1].map(str) + "_" + ks[col2].map(str)
    label_enc = LabelEncoder()
    interactions[new_col_name] = label_enc.fit_transform(new_values)

baseline_data = baseline_data.join(interactions)
baseline_data.head(20)

train, valid, _ = get_data_splits(baseline_data)
bst = train_model(train,valid)

ax = lgb.plot_importance(bst, importance_type="split", max_num_features=59,
figsize=(15,15))
```