

Customer Classification in Insurance Sector

(AJ-2)

Vineet Nama, MSc.

A Capstone submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

*August 2024*

Supervisors:

Professor Dr. Mark Connor,  
 Professor Dr. Angel A. Juan,

Professor Dr. Michael MacDonnell

Head of School: Professor Anthony Brabazon

# **Dedication**

We are extremely grateful to Prof. Dr. Angel A. Juan for his practical guidance and giving us the opportunity to work on his project, and to Prof Dr. Mark Connor and Prof. Dr. Michael MacDonnell for their constant support in shaping our approach.

# **Table of Contents**

**Contents**

[**Dedication 1**](#_Toc174727258)

[**Table of Contents 2**](#_Toc174727259)

[**List of Figures 3**](#_Toc174727260)

[**List of Tables 4**](#_Toc174727261)

[**List of Algorithms 5**](#_Toc174727262)

[**Preface 5**](#_Toc174727263)

[**Acknowledgments 6**](#_Toc174727264)

[**Executive Summary 6**](#_Toc174727265)

[**List of important abbreviations 7**](#_Toc174727266)

[**Chapter 1 - Project Briefing 8**](#_Toc174727267)

[1.1 Project Objective 8](#_Toc174727268)

[1.2 Project Background 8](#_Toc174727269)

[**Chapter 2 - Literature review 9**](#_Toc174727270)

[2.1 Industry Trends Overview 9](#_Toc174727271)

[2.2 Insurance Sector Evolution 10](#_Toc174727272)

[2.3 Customer Segmentation Techniques 12](#_Toc174727273)

[**Chapter 3 - Classification and Revenue Optimization 13**](#_Toc174727274)

[3.1 Approach 13](#_Toc174727275)

[3.2 Data Pre-Processing 13](#_Toc174727276)

[3.3 Descriptive Analysis 14](#_Toc174727277)

[3.4 Model Formulation Based on Optimum Feature Selection Technique 16](#_Toc174727278)

[3.5 Model Formulation using SHAP Analysis 17](#_Toc174727279)

[3.6 Model Formulation Based on Majority Voting Principle 19](#_Toc174727280)

[3.7 Model Formulation using PCA 20](#_Toc174727281)

[**Chapter 4 - Results 21**](#_Toc174727282)

[4.1 Technical Aspect 21](#_Toc174727283)

[4.1.1 Methodology 1: Optimum Feature Selection Approach 21](#_Toc174727284)

[4.1.2 Methodology 2: SHAP Analysis 22](#_Toc174727285)

[4.1.3 Methodology 3: Consensus Approach 22](#_Toc174727286)

[4.1.4 Methodology 4: PCA Approach 22](#_Toc174727287)

[4.2 Business Aspect 24](#_Toc174727288)

[4.2.1 ML Model Revenue Projection 25](#_Toc174727289)

[**Chapter 5 - Future Recommendations 27**](#_Toc174727290)

[**Chapter 6 - Conclusion 28**](#_Toc174727291)

[**Chapter 7 - References 29**](#_Toc174727292)

[**Chapter 8 - Appendices 31**](#_Toc174727293)

[**Chapter 9 - Glossary of terms 34**](#_Toc174727294)

# **List of Figures**

[Fig 1. Distribution of 0(profit generating customer) and 1(loss-making customer) 15](#_Toc174729918)

[Fig 2. Distribution of target variable 15](#_Toc174729919)

[Fig 3. Target variable versus frequency 16](#_Toc174729920)

[Fig. 4. ROC AUC Score with No. of Predictors for XGBoost Model 17](#_Toc174729921)

[Fig 5. SHAP Analysis Output for a typical Classification Model showing the top 10 features 18](#_Toc174729922)

[Fig 6. Cumulative Explained Variance to Number of Principal Components 20](#_Toc174729923)

[Fig. 7 - Model wise Revenue Projection 26](#_Toc174729924)

[Fig 8 - Model wise Customer Rejection Rate 26](#_Toc174729925)

[Fig 9 - Residual vs Predicted Plot of Linear Regression 31](#_Toc174729926)

[Fig 10 - Residuals vs. Fitted Plot Values of Ridge Regression 32](#_Toc174729927)

[Fig 11 - Actual vs. Predicted Values of Ridge Regression 32](#_Toc174729928)

[Fig 12 - Residuals vs. Fitted Plot Values of Lasso 32](#_Toc174729929)

[Fig 13 - Actual vs. Predicted Values of Lasso 33](#_Toc174729930)

[Fig 14 - Actual vs. Predicted Values of Elastic Net 33](#_Toc174729931)

# **List of Tables**

[Table 1 - Redundant columns removed from the dataset 14](#_Toc174727427)

[Table 2. Feature Scores for each ML algorithm and Average Feature Importance Score Calculation 18](#_Toc174727428)

[Table 3. Consensus Model prediction based on majority voting 19](#_Toc174727429)

[Table 4. Performance Summary of Classification Algorithms using Optimum Feature Selection approach 21](#_Toc174727430)

[Table 5. List of the 9 predictor variables for the XGBoost Model 22](#_Toc174727431)

[Table 6. Performance Summary of Classification Algorithms using SHAP Analysis 22](#_Toc174727432)

[Table 7 - Results of PCA implementation 22](#_Toc174727433)

[Table 8. Accuracy metrics of models for each approach 24](#_Toc174727434)

[Table 9. Customer Rejection Rate for each model 24](#_Toc174727435)

[Table 10. Revenue per customer achieved for each model 25](#_Toc174727436)

[Table 11. Revenue Projection of ML Models 26](#_Toc174727437)

# **List of Algorithms**

[XGBoost 21](#_Toc174727488)

[Light Gradient Boost 21](#_Toc174727489)

[Decision Tree Classifier 21](#_Toc174727490)

[Extra Trees Classifier 21](#_Toc174727491)

[Random Forest 21](#_Toc174727492)

[Logistic Regression 21](#_Toc174727493)

[Multi-Layer Perceptron 21](#_Toc174727494)

[PCA with Logistic Regression 22](#_Toc174727495)

[PCA with MLP 22](#_Toc174727496)

[PCA with XGBoost 22](#_Toc174727497)

[PCA with LGBM 22](#_Toc174727498)

[Linear Regression 31](#_Toc174727499)

[Lasso Regularisation 31](#_Toc174727500)

[Ridge Regression 31](#_Toc174727501)

[Elastic Net Regularisation 31](#_Toc174727502)

# **Preface**

This capstone project, entitled "Customer Classification in the Insurance Sector," represents a collaborative effort by Ankur Ghosh, Nishant Moona, and Vineet Nama, undertaken as a crucial component of their Master's in Business Analytics program. The study delves into cutting-edge strategies for customer segmentation and harnesses the power of advanced machine-learning techniques within the insurance industry. Aimed at tackling key business issues, the research pioneer’s new ways of segmenting customers consider the ethical dimensions of data utilization and assess the efficacy of machine learning in enhancing business decisions.

To accomplish its goals, the project leveraged a variety of data sources, employed sophisticated analytical tools, and utilized state-of-the-art machine learning technologies. This research marks a significant step into the intricate domain of customer classification in the insurance sector and contributes to the broader academic field by offering essential insights for those fascinated by the convergence of technology and customer relationship management within this critical industry. The authors extend their heartfelt thanks to their mentors, colleagues, and all those who offered support during this academic venture, recognizing the vital role that encouragement and expert guidance played in the success of their work.

*Dublin* Ankur Ghosh

*September 2024* Nishant Moona  
 Vineet Nama

# **Acknowledgments**

We extend our heartfelt gratitude to our supervisor, Prof. Dr. Mark Connor, for his steadfast support and invaluable insights throughout our Capstone project. His profound expertise significantly shaped our approach, particularly in enhancing our literature review and sharpening our research focus. Dr. Connor's dynamic and enthusiastic mentorship greatly improved our problem-solving capabilities and instilled in us the confidence to address complex issues. As an assistant professor at the UCD School of Business and an integral part of the UCD Natural Computing Research & Applications Group, his advice was crucial in navigating the complexities of our research journey.

Our sincere appreciation also goes to Prof. Dr. Angel A. Juan for his dedicated time and expert advice. His deep knowledge of the insurance industry and data analysis offered us unique perspectives that were critical to our project. Dr. Angel’s guidance in setting precise objectives for our models right from the start proved invaluable. His extensive experience across different domains greatly enriched our research, ensuring its technical robustness and alignment with the nuances of the insurance sector.

Lastly, we are thankful to Dr. Michael MacDonnell, our program director, for his exceptional role in creating a fair and transparent Capstone project allocation process. We are equally grateful to all the faculty members of the UCD Michael Smurfit Graduate Business School for their relentless support, which not only bolstered our confidence but also enabled us to successfully conclude our project.

# **Executive Summary**

Our capstone project is dedicated to boosting the efficiency and profitability of the insurance sector using cutting-edge data management and client classification techniques. Initially, we worked with a large dataset, which included 196 features across 116,650 rows. After a meticulous process of cleaning, which involved removing redundant columns and null values, we refined the dataset to 51,546 rows.

In this project, we have developed three distinct methodologies to create models that accurately classify clients according to their risk levels. The overarching goal is to pinpoint the ideal clients - those who demonstrate disciplined asset management and present low risk - while also identifying high-risk clients who might necessitate higher premiums or even be considered for policy rejection. Our models are carefully designed to enhance revenue per customer, reduce rejection rates, and streamline operations for insurance companies.

A central element of our strategy is the utilization of the SINCO file, a comprehensive historical database used across the insurance industry to track clients' insurance histories. This resource is invaluable for enabling more precise premium settings and improving risk assessments, which benefits not only the insurers but also the safe drivers.

Through this endeavor, we aim to equip insurance companies with innovative tools that will refine client classification, maximize revenue, and enhance the decision-making processes, paving the way for a more efficient and profitable insurance industry.

# **List of important abbreviations**

1. **AI:** Artificial Intelligence
2. **ML:** Machine Learning
3. **AUC-ROC Value**: Area under the Receiver Operating Characteristic Curve
4. **ESG**: Environmental, Social, and Governance
5. **MLP**: Multi-Layer Perceptron
6. **PCA**: Principal Component Analysis
7. **PAYD**: Pay-As-You-Drive
8. **SHAP**: SHapley Additive exPlanations
9. **UBI**: Usage-Based Insurance
10. **XGBoost**: Extreme Gradient Boosting
11. **SINCO**: Sistema de Inspección Nacional de Cumplimiento (National Compliance Inspection System)
12. **GDPR**: General Data Protection Regulation
13. **RFM**: Recency, Frequency, and Monetary

# **Chapter 1 - Project Briefing**

## 1.1 Project Objective

The capstone project on customer classification techniques in the insurance sector presents a complete approach to improving risk assessment and operational efficiency. By diving into data analytics, machine learning, and behavioral insights to position the project at the intersection of technology and customer-centric strategies.

Highlight the development of a robust classification framework that reflects the industry's need to evolve alongside digital transformation. The project’s objective is to classify customers based on their risk profiles which aligns with current trends where personalization and precise risk management are key. By analyzing data, the approach will likely contribute to more precise predictions. This could lead to better customer retention, optimized marketing strategies, and a deeper understanding of customer behaviors.

The plan is to compare various machine learning models such as XGBoost, Decision Trees, Random Forests, Extra Trees, Logistic Regression, and Neural Networks focusing on the importance of selecting the right tools for classification. The focus on performance metrics such as accuracy, precision, recall, and F1-score along with model optimization indicates a complete commitment to achieving consistent results.

Moreover, focusing on the ethical association of customer classification ensures that the project not only sticks to industry regulations but also contributes to fair and unbiased decision-making. By combining these models into existing insurance operations, the project has the potential to greatly impact business outcomes that drive revenue growth and innovation in a competitive market.

Investigation of these advanced techniques will not only benefit the insurance sector but also set a standard for customer analytics across various industries.

## 1.2 Project Background

In the insurance industry, everything circles around managing risk. When a customer such as an individual, a small business, or a large corporation purchases insurance, they pay a premium. This premium acts as a financial safeguard guaranteeing that if the insured asset suffers damage, the insurer will compensate for it whether the loss is caused directly or indirectly by the customer. The insurer’s profit depends mainly on the vulnerability of the insured asset. A customer with a strong track record of careful asset management is ideal for an insurer. Conversely, customers who show poor maintenance habits are seen as risky, which increases the likelihood of claims during the coverage period.

For example, life insurance. A young, healthy person with no history of substance abuse or serious illness is an ideal customer for life insurance. The risk of them developing a critical illness is low, which makes them less likely to file a claim. As a result, life insurers often attract such customers with long-term policies at lower premiums. On the other hand, individuals with a history of addiction to illegal substances or serious health issues are considered high-risk. The probability of future claims is higher, so insurers might either avoid covering them or charge them higher premiums.

The car insurance industry operates similarly. Drivers with a history of reckless driving or suffering and causing accidents are considered less profitable because they're more likely to file claims than drivers with clean records.

For insurers, profit margins are straightforward: the difference between the premiums collected and the payouts made for claims. To increase profitability, insurers often invest premium funds in various ventures like stocks, real estate, or savings accounts. Insurers also partner with banks through bancassurance, where the bank's customers are offered insurance products. This strategy helps insurers expand their customer base while banks earn additional income from the sale of insurance policies.

# **Chapter 2 - Literature review**

## 2.1 Industry Trends Overview

The insurance industry is an important part of the global economy as the industry provides financial protection and mitigates risks against uncertainties. Over time, the sector has changed in response to economic shifts, technological progress, and changing consumer demands. This review examines the current state of the insurance industry, focusing on its structure, technological impact, regulatory environment, and emerging trends, based on recent academic research.

Traditionally, the insurance industry is split into two primary sectors: life insurance and non-life (or general) insurance. Life insurance includes products that offer financial security in cases of death, disability, or retirement whereas non-life insurance covers areas like property, health, casualty, and auto insurance. Cummins and Weiss (2014) highlight the importance of these segments such that life insurance dominates in developed economies while non-life insurance is increasingly important in emerging markets due to growing economic awareness and development (Abdul-Rahman et al., 2021).

Technological advancements are a driving force in the transformation of the insurance industry. Digital technologies including big data analytics, artificial intelligence (AI), and blockchain are revolutionizing the function of insurance companies. Braun and Schreiber (2017) discuss the rise of InsurTech, a sector focused on leveraging technology to enhance insurance services. For example, Artificial Intelligence (AI) is improving underwriting, fraud detection, and customer services while big data enables more accurate risk assessment and personalized product offerings. Blockchain Technology is still emerging but promising to revolutionize claims processing and policy management by improving transparency and cutting administrative costs.

The insurance industry is strictly regulated with varying regulations across regions. These frameworks ensure the solvency of insurance firms that protect the policyholders and maintain market stability. Eling and Schmeiser (2010) explore the effects of regulations particularly the Solvency II directive in the European Union and argue that while these regulations are crucial for financial stability, they also impose compliance costs, which potentially hinder innovation. The study tells for a balanced regulatory approach that protects consumers while encouraging innovation and competition.

Emerging trends and challenges are set to shape the future of the insurance industry. One key trend is the growing demand for personalized insurance products, which is driven by advancements in data analytics and AI. Thoyib et al. (2019) emphasize the shift from generic products to personalised offerings, which presents both opportunities and challenges for insurers. Developing sophisticated risk models and ensuring regulatory compliance are essential for staying competitive in emerging markets.

Climate change poses another major challenge that increases the frequency and severity of natural disasters, raising insurance claims and uncertainty in risk assessment. Botzen et al. (2019) stress the need for insurers to adapt their models to the changes and collaborate with governments and stakeholders to mitigate climate-related risks.

Customer classification is important in the insurance industry, enabling companies to segment their customer base and offer more targeted services. Insurers can better understand customer needs by using factors such as demographics, behavior, and risk profiles. Abdul-Rahman et al. (2021) demonstrates that data mining techniques like K-Modes clustering and Decision Tree classifiers can improve the accuracy of customer segmentation. As a result, it leads to more personalized and effective marketing strategies.

As the insurance industry continues to change, future research could explore the impact of emerging technologies such as quantum computing and advanced AI. Additionally, there is a need to study the ethical implications of using AI and big data, particularly regarding privacy and discrimination. Developing new risk assessment models that address the increasing complexity of global risks, such as climate change and cyber threats, is another promising research area.

The insurance industry is at a pivotal point, facing both opportunities and challenges as it navigates digital transformation, regulatory changes, and shifting customer expectations. By adopting advanced technologies and focusing on customer-centric strategies, insurers can strengthen their competitiveness and meet customer needs. However, insurers must also address challenges such as regulatory compliance, climate change, and the ethical use of data to ensure a sustainable and fair future for the industry.

## 2.2 Insurance Sector Evolution

The insurance industry is experiencing drastic changes like evolving customer expectations and changing regulatory frameworks driven by advancements in technology. These factors adjust how insurance companies operate, develop products, and interact with their old and new customers.

Digital transformation is playing a key role in reshaping the insurance sector. Technologies such as artificial intelligence (AI), machine learning (ML), big data, and blockchain are revolutionizing areas like risk management, claims processing, and customer interactions. Eling and Lehmann (2018) discussed these technologies are making the insurance value chain more efficient by reducing costs and improving customer satisfaction. For instance, AI algorithms are enhancing customer segmentation and risk assessment while big data analytics offer deeper insights into customer behaviour that enable more targeted marketing strategies (Ngai et al., 2020).

Customer expectations have changed immensely with the advent of digital technologies. Today’s consumers want personalized experiences beyond just financial protection. Thoyib et al. (2019) highlight the increasing demand for insurance products that cater to individual lifestyles and preferences. Insurers are responding by offering flexible options like usage-based insurance (UBI) and pay-as-you-drive (PAYD) models, which adjust premiums based on actual usage. The rise of InsurTech companies focuses on this trend as they often provide innovative, digital-first solutions that outperform traditional insurers in terms of customer experience (Zarella et al., 2021). To remain competitive in the market traditional insurers are being urged to adopt digital transformation and customer-centric strategies.

Regulatory challenges are becoming more complex as regulators address new risks and aim to protect consumers. Eling and Pankoke (2021) discussed the challenges insurers face in complying with new regulations such as the General Data Protection Regulation (GDPR) and the updated Solvency II framework. These regulations impose strict requirements on data management and capital solvency. The role of regulatory technology (RegTech) in automating compliance processes is also stated, as it requires significant investment and faces challenges with regulatory fragmentation.

Sustainability and ESG (Environmental, Social, and Governance) factors are gaining importance in the insurance industry. Insurers are under increasing pressure to adopt sustainable practices and incorporate ESG factors into their business models. Botzen et al. (2019) explored how climate change is impacting the insurance sector by stressing the need for new models to assess and price climate-related risks. Integrating ESG criteria into underwriting and investment decisions is becoming a strategic necessity. Ziegler et al. (2020) suggested that implementing ESG goals not only improves reputation and risk management but also drives long-term profitability and resilience.

Overall, the insurance industry is undergoing rapid change due to technological advancements, shifting customer expectations, complex regulations, and a growing focus on sustainability. While these changes present challenges, they also offer opportunities for insurers to innovate and strengthen their businesses. Continued research and collaboration among insurers, regulators, and technology providers will be vital in navigating these transformations and shaping the future of the industry.

When exploring customer classification and segmentation in the insurance industry, it becomes clear that there is a shift from traditional methods to more sophisticated AI-driven approaches. Earlier models often lacked the depth needed to fully understand customer behaviour. With the introduction of techniques like clustering, machine learning, and deep learning insurers can achieve more detailed and effective segmentation. These advancements have improved marketing, risk assessment, and customer management which enable insurers to better align their offerings with customer needs.

However, as these technologies become more prevalent, the ethical challenges in terms of fairness and transparency must be addressed. The future of customer classification will likely involve the integration of AI and machine learning, where real-time data can continually refine customer segments and advanced models can more accurately predict customer lifetime value. As these methods will be adopted in the industry, ongoing research and collaboration will be essential to ensure that these tools are used responsibly, ultimately contributing to a more customer-focused and sustainable insurance sector.

## 2.3 Customer Segmentation Techniques

Traditionally, the insurance sector has relied on demographic and geographic segmentation. These models divide customers based on static attributes such as age, gender, income, and location. For instance, Khajvand et al. (2011) utilized a model incorporating Recency, Frequency, and Monetary (RFM) analysis to estimate the future value of customers based on their past interactions. While this technique is effective for a broader segmentation, but the technique fails to capture the finer nuances of customer behaviour and preferences, which leads to a one-size-fits-all strategy that may not fully exploit each customer’s potential value, as given by Abdul-Rahman et al. (2021).

In recent years, clustering has gained popularity due to its ability to process large datasets and uncover hidden patterns. In the insurance industry clustering techniques such as K-Means and K-Modes have been used for customer segmentation. A recent study by Abdul-Rahman et al. (2021) demonstrated the effectiveness of these methods in segmenting life insurance customers into three distinct groups: "Potential High-Value Customers," "Low-Value Customers," and "Disinterested Customers." This segmentation allows insurance companies to provide their marketing strategies and product offerings to better engage and give value to each customer group.

Another advanced methodology widely used in the insurance sector for customer classification is decision tree classifiers. These models are particularly advantageous because they can handle both categorical and numerical data, therefore making them highly interpretable. In a 2021 study by Abdul-Rahman et al., a combination of clustering with K-Modes and decision tree classifiers was employed. The model trained using the Gini index and entropy measures achieved an accuracy of 81.3% in customer classification using the Gini model. This approach not only classifies customers but also predicts which customers are likely to fit into each segment, thus enabling more precise targeting.

The advent of machine learning and artificial intelligence has transformed customer classification in the insurance sector. AI-based models like neural networks and deep learning algorithms are now capable of capturing complex relationships within data and allowing for advanced analysis. For example, a recent study introduced the application of deep learning combined with AI for customer segmentation. This approach known as DeepLimeSeg, uses deep learning to segment customers while maintaining transparency in the decision-making process, which is important in the insurance sector where understanding why an AI makes certain decisions is of utmost importance.

Beyond demographic segmentation, behavioural and psychographic segmentation models have been developed to better understand customers' lifestyles, values, and behaviour patterns. Fodor and Kocsir (2008) conceptualized a model for segmenting life insurance customers based on their value systems and lifestyle by recognizing that consumer behaviour is influenced by psychological factors rather than purely economic decisions. These models are particularly useful for developing personalized insurance products that align with the motivations of customers.

Currently, hybrid models combining multiple techniques are being used to achieve more accurate customer segmentation. For instance, combining clustering methods with decision trees or integrating machine learning models with traditional statistical methods offers large benefits to insurers. One such model combines K-Means clustering with logistic regression to segment customers and predict their chances of purchasing a particular insurance product.

These models have benefits in the areas of marketing, risk assessment, and customer relationship management in the insurance sector. However, integration of more advanced AI techniques remains to be explored and the operationalization of these models in real-time customer interactions.

Future exploration will focus on dynamic segmentation where AI continuously updates customer segments based on real-time data and on applying deep learning models to more accurately predict customer lifetime value. As the application of these models becomes more widespread, it will be vital to investigate the ethical issues related to fairness and transparency in AI-driven segmentation.

# **Chapter 3 - Classification and Revenue Optimization**

## 3.1 Approach

The approach for preparing the models was carefully designed to address the sponsor’s needs through a series of experimental investigations. We focused on five critical aspects to ensure model development. First, we compared the analysis of the most effective models developed from a chosen set of classification algorithms. This comparison helped us to identify the algorithms that delivered the best performance. Second, we emphasized testing the predictive accuracy of these models, employing different metrics to ensure their reliability in forecasting outcomes. Third, we explored the insights and inferences that could be drawn from each model, understanding their practical value and applicability in real-world scenarios. Fourth, we evaluated the models' potential to optimize revenue after implementation to investigate their impact on financial outcomes. Lastly, we analyzed the rejection rate of these models, evaluating how effectively they could identify or exclude data to maintain high accuracy.

To address these priorities, we designed four different methodologies for building models using supervised classification algorithms. Each methodology was tested through experimental trials, where models were trained on portions of the provided dataset. This approach allowed us to fine-tune the models and adjust parameters based on experimental outcomes. The following sections will provide a detailed account of these methodologies, the experiments conducted and the results of our investigations.

**3.1.1 Target Variable**

The Target variable is “BMA\_pol\_auto” which stands for the average annual benefit provided by a customer. This variable is used for constructing a categorical variable with classes 0 and 1. Each of these classes signifies: -

0 – Signifies profit-generating customers with average positive benefits

1 – Signifies loss-making customers with average negative benefits

## 

## 3.2 Data Pre-Processing

To prepare the data for modeling several key steps are followed:

First, we removed rows with missing values from the dataset to ensure the integrity of the data. Missing values lead to biased and inaccurate model predictions. Additionally, we eliminated any duplicated observations to avoid redundancy and maintain the dataset’s accuracy.

Next, we decided to retain outliers for each feature rather than removing them. Although we explored censoring outliers by modifying them to the maximum permissible value (calculated as the 75th percentile plus 1.5 times the interquartile range, and the 25th percentile minus 1.5 times the interquartile range), this approach did not significantly benefit the analysis. Therefore, we preserved the outliers in their original form to capture the variability of the data, and important for model accuracy and robustness.

We then split the dataset into independent feature variables and the target variable. The dataset was divided into three subsets: the training dataset includes 30,927 observations, the validation dataset with 10,309 observations, and the testing dataset contains 10,310 observations. This split is for the thorough evaluation of the model performance.

Moreover, we removed certain columns from the dataset that were previously applicable but are no longer relevant under the current circumstances. These features were removed to simplify the modeling process and focus on the variables that are most important for the study's goals. The removed features include:

In summary, these preprocessing steps were vital in preparing a clean, accurate, and relevant dataset for the different modelling stages, which ensured that the analysis was both robust and meaningful. This approach to data preparation is important for producing reliable and interpretable results.

|  |  |
| --- | --- |
| **Feature columns (in Spanish)** | **English Translation** |
| PrimaTotalPoliza | Total Policy Premium |
| ComisionTotalPoliza | Total Policy Commission |
| ExposicionTotalPoliza | Total Policy Claims |
| SiniestralidadTotalPoliza | Total Policy Exposure |

##### Table 1 - Redundant columns removed from the dataset

Finally, we created a new column named **`Classification**` to categorize customers as either Profit-generating (0) or Loss-making (1). This categorization is based on whether the **`BMA\_pol\_auto`** value is positive or negative.

## 3.3 Descriptive Analysis

The pie chart represents the distribution of the target variable true class in the dataset. The profit-generating customers (label 0) consist of 80.80% of the data while the loss-making (label 1) makes up 19.20%. This imbalance impacts the model performance in such a way that models are biased toward profit-generating customers and poorly determining loss-making customers.

A pie chart with a number of percentages

Description automatically generated

###### Fig 1. Distribution of 0(profit generating customer) and 1(loss-making customer)

The boxplot describes the distribution of the variable ‘BMA\_pol\_auto’ for class 1 (profit-generating customers). The figure shows the wide range of values that are identified as outliers. We kept the outliers in their original form to capture the variability.

A diagram of a graph

Description automatically generated

###### Fig 2. Distribution of target variable

The detailed summary of the distribution of the target variable 'BMA\_pol\_auto' signifies that the dataset contains 51,546 observations. The average value is 45.24. The standard deviation is high at 2479.69 reflecting variability in the data. The minimum value is -185,990.40, a negative outlier that indicates the presence of extreme values affecting the analysis and modeling. The value at the 25th percentile is 75.25. The median or 50th percentile value is 181.12, higher than the mean and suggests a right-skewed distribution. The value at the 75th percentile is 272.21. The maximum value is 28,207.91, which is higher than the mean and median highlighting the presence of extreme outliers at the end.

With a skewness value of -39.74, it suggests that the distribution is extremely negatively skewed. There are a few data points far to the left. The level of skewness is very high, which states that the target variable has a lot of extreme outliers on the lower end.

A kurtosis value of 2426.48 is extremely high, which indicates that the distribution has very heavy tails. There are extreme outliers in the data making the distribution very spiky. Most of the data is clustered around the mean.

A graph of a distribution of a number of data

Description automatically generated with medium confidence

###### Fig 3. Target variable versus frequency

## 3.4 Model Formulation Based on Optimum Feature Selection Technique

It is a statistical technique of optimizing a supervised machine learning model in which the dataset is divided into three parts - training, validation, and testing. The model is crafted by determining the model parameters by fitting the model using the training dataset. The performance of the model is tested and monitored with the validation dataset. The process is iterated several times by increasing the number of features in each step to decrease feature importance. The final version of the model incorporates the optimum number of features required for optimum classification and the names of these features. This method is repeated for all the algorithms and the best version of each algorithm containing their optimum metric score (ROC AUC Scores in this case), the optimum number of predictors and the name of these predictors was determined. The best-performing version among all these versions is then used for predicting the output values in the test dataset and their output is compared with one another. The selected set of algorithms used in Optimum Feature Selection are – XGBoost, Light Gradient Boost and Management Classifier, Decision Tree, Extra trees, and Random Forest**.**

The plot below shows the variability in ROC AUC scores with the number of predictors for the XGBoost Algorithm. Similar plots have been constructed for the rest of the models and the best versions of each of these models have been shown in Table 1 below.

A line graph with numbers and lines

Description automatically generated

###### Fig. 4. ROC AUC Score with No. of Predictors for XGBoost Model

The table shows that for the XGBoost Model, the ROC AUC Score has attained the highest values for 9 predictor variables. Similar plots were constructed for the other algorithms whose results are presented in the results section.

Some of the benefits that the above methodology brings are to prevent overfitting by identifying the optimum number of feature variables for describing the target variable thereby reducing model complexity and improving model interpretability to a great degree by selecting the best subset of feature variables.

One of the major drawbacks of this approach is that the number of iterations executed per algorithm is very high during best model selection and therefore leads to a very long running time. This is particularly true for the tree-type classifiers like Decision Tree, Extra Trees, and Random Forest.

## 3.5 Model Formulation using SHAP Analysis

SHAP is a library built within the Python Environment which is used for determining the relative feature importance of the independent variables. The methodology involves the identification of a fixed number of predictors based on their importance obtained from SHAP analysis and using them for crafting models using the selected set of classification algorithms.

The dataset in this methodology was split into three parts – training, validation, and testing. The training and validation datasets are utilized for creating models based on the selected set of learning algorithms. The SHAP library is utilized for finding the important scores of each of the independent variables. This importance score is calculated for each of the 5 learning algorithms resulting in 5 sets of feature importances. The scores corresponding to each of the sets are scaled between 0 and 1 (Min-Max scaling) to give equal weightage to the importance scores corresponding to each set. The average of the 5 scores for each feature is calculated and the top 15 features are selected:

A graph with blue bars

Description automatically generated with medium confidence

###### Fig 5. SHAP Analysis Output for a typical Classification Model showing the top 10 features

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature name | XGBoost | LGBM | Decision Tree | Extra Trees | Random Forest | Scaled XGBoost | Scale LGBM | Scaled DTC | Scaled ETC | Scaled RF | Avg  Imp. |
| Discount Rate | 0.66 | 0.58 | 0.10 | 0.016 | 0.05 | 1.00 | 1.00 | 1.00 | 0.49 | 1.00 | 0.89 |
| Policy Duration | 0.15 | 0.13 | 0.03 | 0.034 | 0.01 | 0.22 | 0.22 | 0.37 | 1.00 | 0.37 | 0.43 |
| Policy Year | 0.40 | 0.27 | 0.04 | 0.008 | 0.01 | 0.60 | 0.46 | 0.43 | 0.25 | 0.30 | 0.41 |
| SINCO Time Ratio | 0.23 | 0.16 | 0.03 | 0.002 | 0.003 | 0.34 | 0.28 | 0.32 | 0.06 | 0.07 | 0.22 |
| Average Time in SINCO | 0.10 | 0.08 | 0.03 | 0.006 | 0.007 | 0.15 | 0.13 | 0.32 | 0.18 | 0.14 | 0.18 |
| Accident Frequency | 0.16 | 0.12 | 0.02 | 0.001 | 0.003 | 0.24 | 0.21 | 0.26 | 0.04 | 0.06 | 0.16 |
| Included License Plates | 0.08 | 0.08 | 0.01 | 0.008 | 0.006 | 0.13 | 0.15 | 0.10 | 0.24 | 0.12 | 0.15 |
| Vehicle Market Value | 0.08 | 0.04 | 0.01 | 0.002 | 0.005 | 0.13 | 0.07 | 0.17 | 0.05 | 0.09 | 0.10 |
| Vehicle Fuel type in Policy | 0.07 | 0.06 | 0.001 | 0.006 | 0.005 | 0.11 | 0.10 | 0.01 | 0.17 | 0.09 | 0.09 |
| Insurance Time in Database | 0.04 | 0.01 | 0.024 | 0.001 | 0.003 | 0.07 | 0.03 | 0.23 | 0.03 | 0.07 | 0.09 |
| Channel Score | 0.06 | 0.03 | 0.017 | 0.001 | 0.003 | 0.09 | 0.05 | 0.17 | 0.05 | 0.06 | 0.08 |
| Total vehicle insurance time | 0.07 | 0.02 | 0.01 | 0.001 | 0.003 | 0.10 | 0.04 | 0.11 | 0.04 | 0.06 | 0.07 |
| Total License Plates | 0.04 | 0.04 | 0.001 | 0.003 | 0.003 | 0.07 | 0.06 | 0.05 | 0.09 | 0.07 | 0.07 |
| Total Vehicles | 0.05 | 0.03 | 0.004 | 0.002 | 0.005 | 0.08 | 0.05 | 0.04 | 0.08 | 0.09 | 0.07 |
| License Age | 0.04 | 0.01 | 0.01 | 0.002 | 0.003 | 0.07 | 0.01 | 0.12 | 0.07 | 0.05 | 0.06 |

##### Table 2. Feature Scores for each ML algorithm and Average Feature Importance Score Calculation

The features highlighted in the table above are used for creating the five Machine learning models. The models are subsequently tested on the test dataset and the best performing model among these five models is chosen.

Some of the benefits offered by this methodology: -

The same set of feature variables is used across all the learning algorithms thus establishing a common ground for effectively comparing the performance of the learning algorithms.

Selection of a finite set of selection features ensures that overfitting does not take place.

One of the drawbacks of this methodology is that this may not lead to the creation of optimized models. Each learning algorithm has its approach to classifying items and therefore the feature importances vary from one algorithm to another. Hence, selecting a common set of features based on averaging scaled importance scores may not produce the optimized model version for each algorithm.

A relative comparison of these models has been provided in the results section of this document.

## 3.6 Model Formulation Based on Majority Voting Principle

The methodology involves developing a consensus model that encompasses the output of all the five models for classifying customers. The model is based on the majority voting principle in which the consensus of the majority number of models towards a particular observation is used for determining the class label of the customer. The table below highlights the predictions made by the model: -

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No. | Actual | XGBoost | LGBM | Decision Tree | Extra Trees | Random Forest | Models predicting 0 | Models predicting 1 | Consensus (majority case) |
| 1 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 3 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 4 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| 10 | 1 | 1 | 1 | 1 | 0 | 0 | 3 | 2 | 1 |
| 11 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 3 | 0 |
| 12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |

##### Table 3. Consensus Model prediction based on majority voting

Let’s take the first observation/customer profile as an example. The customer is a profit-generating customer signified by the label ‘0’. For this profile, the XGBoost and LGBM Classifiers predict it to be a loss-making customer indicated by the label ‘1’. The rest of the classification models predict the customer profile as ‘profit-generating’. The consensus model reflects the prediction of the majority of the models by classifying the customer as “profit-generating”. In short, the consensus model reduces the chances of error by taking an overall perspective of the results obtained from the other models.

## 3.7 Model Formulation using PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique to improve the performance of machine learning models of XGBoost, LGBM, Logistic Regression, and MLP (Multi-Layer Perceptron). The key takeaway from the PCA is that to retain 95% of the variance in the data, dimension of the dataset reduced to 91 components. the number of components. The original dataset was then transformed into a new feature space defined by the selected principal components. This reduced dataset was used as the input for the machine-learning models. The results were analyzed without significant loss of information. The PCA-transformed dataset was used to train and evaluate the selected machine learning models (XGBoost, LGBM, Logistic Regression, and MLP). The performance of these models was compared to those trained on the original dataset to assess the impact of PCA on model accuracy. The model performances were evaluated using metrics like accuracy, recall, and F1-score, particularly focusing on the classification of ‘loss-making customers’ in the insurance sector.

A graph with a blue line

Description automatically generated

###### Fig 6. Cumulative Explained Variance to Number of Principal Components

By implementing PCA, the project aimed to simplify the models, reduce overfitting, and improve computational efficiency, mainly for complex models like MLP, XGBoost and LGBM. This methodology ensured that the models were both robust and efficient, making them more applicable for real-world deployment in the insurance sector.

# **Chapter 4 - Results**

The results can be analyzed from two perspectives:

The **technical perspective** involves comparing the performance of the various machine learning algorithms used in the study. It focuses on evaluating how well each algorithm performs relative to the others, based on metrics such as accuracy, precision, recall, and F1 score.

Whereas the **business perspective** examines the impact of implementing machine learning algorithms on the insurance company’s financial performance. It looks at how the insights gained from the models influence the company's revenue, profitability, and overall business strategy.

Both aspects have been covered in detail in the sections below.

## 4.1 Technical Aspect

### 4.1.1 Methodology 1: Optimum Feature Selection Approach

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Serial No. | Model Type | ROC AUC Score | Recall Score | Precision Score | F1 Score | Overall Accuracy | Optimum predictors |
| 1 | **XGBoost** | **0.6** | 0.24 | 0.63 | 0.35 | 0.83 | 9 |
| 2 | **Light Gradient Boost** | 0.59 | 0.21 | 0.66 | 0.32 | 0.83 | 12 |
| 3 | **Decision Tree Classifier** | 0.58 | 0.33 | 0.30 | 0.31 | 0.73 | 17 |
| 4 | **Extra Trees Classifier** | 0.56 | 0.26 | 0.31 | 0.28 | 0.75 | 4 |
| 5 | **Random Forest** | 0.57 | 0.16 | 0.68 | 0.25 | 0.83 | 21 |
| 6 | **Logistic Regression** | 0.56 | 0.11 | 0.48 | 0.07 | 0.79 | 9 |
| 7 | **Multi-Layer Perceptron** | 0.58 | 0.31 | 0.29 | 0.30 | 0.72 | 19 |

##### Table 4. Performance Summary of Classification Algorithms using Optimum Feature Selection approach

Table 4 reflects the performance of all the seven models used in the first methodology which involves the selection of a subset of predictors that best describe the classification models. The “XGBoost” model clearly shows the best performance among all in terms of ROC AUC Score. The model is built on 9 predictors which are mentioned in table 5 below: -

|  |  |  |
| --- | --- | --- |
| Sl. No. | Feature columns (in Spanish) | English Translation |
| 1 | id55 | Bonus Percentage (SINCO) |
| 2 | PerteneceSINCO | SINCO Time Ratio |
| 3 | FrecuenciaSiniestroSINCO | General Claim Frequency (SINCO) |
| 4 | AnyoPoliza | Policy Year |
| 5 | ContrataLunas | Contracts Glass Coverage |
| 6 | NumeroDanyosMaterialesSINCO | Material Damage Claim Frequency |
| 7 | id13\_H | Alphanumeric Data (External Source) |
| 8 | id14\_MANTENIMIENTO | Mediation Network Segmentation |
| 9 | Cliente\_Diverso | Policy Duration |

##### 

##### Table 5. List of the 9 predictor variables for the XGBoost Model

### 4.1.2 Methodology 2: SHAP Analysis

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Serial No. | Model Name | ROC AUC Score | Recall Score | Precision Score | F1\_scores | Overall Accuracy |
| 1 | XGBoost | 0.65 | 0.51 | 0.36 | 0.42 | 0.73 |
| 2 | LGBM | 0.67 | 0.58 | 0.36 | 0.44 | 0.72 |
| 3 | Decision Tree | 0.57 | 0.31 | 0.31 | 0.31 | 0.73 |
| 4 | Extra Trees | 0.55 | 0.11 | 0.62 | 0.19 | 0.82 |
| 5 | Random Forest | 0.57 | 0.16 | 0.68 | 0.25 | 0.82 |

##### Table 6. Performance Summary of Classification Algorithms using SHAP Analysis

Table 6 reflects the performance of all the 5 models based on SHAP analysis. The LGBM model clearly shows the best performance among all in terms of ROC AUC Score. It gives superior results than the XGBoost model from the first methodology (an ROC AUC Score of **0.67** against **0.60**).

### 4.1.3 Methodology 3: Consensus Approach

The consensus model based on Majority Voting of predictions made by all the five models had an ROC AUC score of **0.61**.

### 4.1.4 Methodology 4: PCA Approach

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Methodology | ROC AUC Scores | Recall Scores (for 0) | Precision Scores (for 0) | Recall Scores (for 1) | Precision Scores (for 1) |
| PCA with Logistic Regression | 0.58 | 0.78 | 0.81 | 0.04 | 0.41 |
| PCA with MLP | 0.59 | 0.81 | 0.83 | 0.31 | 0.28 |
| PCA with XGBoost | 0.63 | 0.97 | 0.82 | 0.08 | 0.43 |
| PCA with LGBM | 0.66 | 0.99 | 0.81 | 0.04 | 0.63 |

##### Table 7 - Results of PCA implementation

Table 7 illustrates that the Logistic Regression model with PCA, the ROC-AUC score was 0.58, suggesting a moderate ability to differentiate between profit-generating and loss-making customers. The model performed well for profit-generating customers, with a recall of 0.78 and a precision of 0.81. However, it struggled with loss-making customers, showing a very low recall of 0.04 and a precision of 0.41.

MLP model achieves an ROC-AUC score of 0.59, a low ability to distinguish between the two classes. The recall score for class 0 (profit-generating customers) was 0.81 and the precision score was 0.83, suggesting the model performed well in identifying correctly class 0. However, for loss-making customers a recall score of 0.31 and a precision score of 0.28. Thus, the model identified incorrectly of class 1 (loss-making customers) leading to a high rate of false negatives and misclassifications for this class.

In the XGBoost model, the ROC-AUC score was slightly higher at 0.63. The model identified all the instances of profit-generating customers as recall is 0.97. However, XGBoost also performed poorly with distinguishing loss-making customers because of a recall score of 0.08. The precision scores were 0.82 for profit-generating and 0.43 for loss-making customers. Thus, bias toward profit-generating customers.

The LGBM model demonstrated the highest ROC-AUC score at 0.66, suggesting that it was the most effective among the three models in distinguishing between the two classes. The recall score for profit-generating customers was extremely high at 0.99, indicating that the model perfectly identified the profit-generating customers. However, LGBM had a poor recall score for loss-making customers at 0.04, which shows it is not able to detect loss-making customers. Precision for profit-generating customers was 0.81 while for loss-making customers it was higher as compared to other models at 0.63. Thus, the LGBM model is more precise in identifying loss-making customers than other models.

In summary, the results show that while using PCA with these models helps reduce the number of features, it doesn't do a good job of improving the detection of the loss-making customers in any of the models. The LGBM model performed the best in terms of ROC-AUC as compared to other models but is biased towards profit-generating customers. This suggests that PCA might be causing a loss of important information, which leads to poorer performance in identifying the loss-making customers.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Methodology** | **ROC AUC Scores** | **Recall Scores (for profit-generating customers)** | **Precision Scores (for profit-generating customers)** | **Recall Scores (for loss-making customers)** | **Precision Scores (for loss-making customers)** |
| **Optimum Feature Selection (XGBoost)** | 0.60 | 0.96 | 0.84 | 0.22 | 0.60 |
| **SHAP Analysis (LGBM)** | 0.67 | 0.75 | 0.88 | 0.58 | 0.36 |
| **Consensus (Majority Voting)** | 0.61 | 0.93 | 0.84 | 0.28 | 0.49 |
| **PCA with LGBM** | 0.66 | 0.99 | 0.81 | 0.04 | 0.63 |

##### Table 8. Accuracy metrics of models for each approach

Revenue Optimization can be achieved in two ways. First is to identify the potential profit-generating customers with a high degree of accuracy and provide them with insurance policies followed by identifying the potential loss-making customers accurately and removing them from the customer pool.

The results achieved by the above models are a trade-off between the above two aspects. In methodology 1, the recall value for profit-generating customers is high (0.96) which indicates that roughly around 96% of the total customers who are profit-generating are successfully identified by the model. However, this value reduces to 0.22 for loss-making which indicates that only 22% of the loss-making customers are being identified correctly. The same is true for the consensus model (recall values of 0.93 for profit-generating customers and 0.28 for loss-making customers which is slightly better than the methodology of the XGBoost model). However, when it comes to SHAP analysis, a significant improvement in recall scores for loss-making customers is achieved (0.58 against 0.22 and 0.28) at the expense of reduced recall scores for profit-generating customers (0.75 against 0.96 and 0.93). This indicates that the LGBM model obtained from SHAP analysis can identify loss-making customers far better than the other two models. At the same time, its capability to identify profit-generating customers has declined.

The above analysis shows us that revenue can be increased by either identifying the profit-generating customers and including them within the company or by reducing losses suffered by shrinking the loss-making section of the customer pool. The impact of these two aspects can be seen on the overall revenue generated per customer and the customer rejection rate, which both critical business aspects for the insurance company.

## 4.2 Business Aspect

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methodology | Model | Total number of people enrolled before ML implementation | Total number of people enrolled after ML implementation | Rejection Rate |
| Optimum Feature Selection | XGBoost | 10310 | 9568 | 7.2% |
| SHAP Analysis | LGBM | 10310 | 7084 | 31.29% |
| Consensus | Majority Voting | 10310 | 9176 | 11% |
| PCA | LGBM | 10310 | 10183 | 1.23% |

##### Table 9. Customer Rejection Rate for each model

The above table reflects the customer rejection rate for the best models obtained from each of the methodologies. The model obtained from SHAP analysis shows the highest rejection rate among all the models (31.29%). Principal Component Analysis (PCA) with the LGBM model provides the lowest rejection rate of 1.23%. Insurance companies often prefer lower rejection rates for customers because even though a higher rejection rate means that the model can successfully identify loss-making customers better, it is also rejecting a high percentage of customers that it deems to be unfit, thereby reducing the overall customer pool. It also creates dissatisfaction among the customers and many of them might switch to an alternative insurance provider. A rejection rate of 10% is usually considered ideal in most cases.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methodology | Model | Revenue Head per customer before model application | Revenue Head per customer after model application | Increment (No. of times) |
| Optimum Feature Selection | XGBoost | 55.66 | 107.26 | 1.93 |
| SHAP Analysis | LGBM | 55.66 | 200.86 | 3.61 |
| Consensus | Majority Voting | 55.66 | 111.91 | 2.01 |
| PCA | LGBM | 55.66 | 63.40 | 1.13 |

##### Table 10. Revenue per customer achieved for each model

Table No. 10 summarizes the monetary benefits obtained after the application of the ML models. The LGBM Model obtained from SHAP analysis can generate 3.61 times more revenue per customer compared to the normal scenario when no model is applied. This can be attributed to the fact that this model has a high rejection rate of 31.29% (Table 7) and is also improving the revenue pool by predicting the no. of loss-making customers. Hence, total revenues generated increased while the total number of customers declined which led to high revenue generated per customer. However, this method may not find much popularity because of the high rejection rate that might lead to shrinking of the customer pool due to customer dissatisfaction.

On the other hand, the other models have lower returns. The XGBoost model provides a revenue return of 1.93 times per customer with an acceptable rejection rate of 7.2% while the consensus model can double the revenue with a decent customer rejection rate of 11%. When implementing the Principal Component Analysis (PCA) on the LGBM Model, the revenue growth of 1.13 times per customer with an acceptable rejection rate of 1.24%.

The business aspect highlights that the financial benefits of customer identification are a function of two variables - the rejection rate and the revenue per customer.

It is for the company to decide which model they want to utilize for customer classification based on their appetite for customer rejection rate and the revenue generated per customer.

### 4.2.1 ML Model Revenue Projection

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Methodology** | **Model** | **Total Revenue Generated before ML Model Implementation** | **Total Revenue Generated after ML Model Implementation** | **Net Savings** |
| **Optimum Feature Selection** | XGBoost | €573,891 | €1,026,303 | €452,412 |
| **SHAP Analysis** | LGBM | €573,891 | €1,422,901 | €849,010 |
| **Consensus** | Majority Voting | €573,891 | €1,025,922 | €452,031 |
| **PCA** | LGBM | €573,891 | €645,572 | €71,681 |

##### Table 11. Revenue Projection of ML Models

###### Fig. 7 - Model wise Revenue Projection

The Revenue Projection illustrates the estimated revenue outcomes from applying different machine learning models in the insurance sector. The base revenue without any machine learning model is €5,73,891. When the LGBM model is combined with Principal Component Analysis (PCA) revenue increases to a moderate level to €6,45,571. The XGBoost model provides a more boost, with projected revenue rising to €10,26,303. The Consensus Model offers a slight increase to €10,58,641 compared to XGBoost model.

However, the most real revenue projection, €14,22,901 achieved using the LGBM model without PCA, highlighting its effectiveness in optimizing revenue compared to the other methods. This graph shows that advanced machine learning techniques, such as LGBM model can enhance revenue projections in insurance applications.

###### Fig 8 - Model wise Customer Rejection Rate

Fig 8 represents the percentage of customers rejected by different machine learning models used in the insurance sector. The LGBM model with PCA shows the lowest rejection rate at 1.23%, which indicates customer selection is more inclusive. The XGBoost model has a moderate rejection rate of 7.20%, while the Consensus model sees an increase to 11%. The highest rejection rate of 31.29% is associated with the LGBM model without implementing PCA. This suggests a more rigorous selection process and the variation signifies the trade-offs between inclusivity and model precision.

# **Chapter 5 - Future Recommendations**

To make a project on customer classification in the insurance sector more effective, it’s important to take a well-rounded approach. This includes gathering data, refining features, picking the right model, and continuously improving the process.

First, expanding the dataset is key. By collecting more data over time, a better understanding of the different patterns and behaviors that affect customer risk can be achieved. Including diverse sources like financial transactions, and personal details like age, gender, and job type, gives a clear picture of each customer. Adding extra information such as demographics, socio-economic factors, and geographic risks can help in better grouping customers and assessing risks.

Improving the features used in the model can also make predictions more accurate. Creating features that show how a customer’s behavior changes over time like life events or financial stability is essential. Moreover, building features that show how different customer attributes interact with each other is important. Using historical data to inform these features ensures the model remains consistent and effective.

It’s also crucial to analyze key features that determine if a customer will be profit-generating or loss-making. Understanding these relationships helps in better segmenting customers and customizing marketing strategies.

Future research in customer segmentation in the insurance sector should focus on integrating advanced machine learning techniques to better identify and categorize customer segments. Khajvand et al. (2011) supported the effectiveness of using the Recency, Frequency, and Monetary (RFM) model in customer segmentation within the retail banking sector. Adapting this approach to the insurance industry enables insurers to accurately predict customer lifetime value and design personalised products. Therefore, it will lead to improved customer satisfaction.

When using regression analysis to predict a customer’s value, it’s important to identify the right features that reflect their potential monetary benefit. The current dataset may lack some of these features, which are used in the regression analysis.

Continuous improvement is a must. Regularly updating the model with new data will keep it relevant and accurate. Feedback from stakeholders like underwriters and marketers will help fine-tune the model based on real-world experiences. In the future, Hyperparameter tuning can be used to make the models more robust and accurate. Staying in touch with end-users and decision-makers ensures the model stays aligned with business goals.

Botzen et al. (2019) focussed the impact of climate change on the insurance industry that highlights the need for new models, which accurately assess the price climate-related risks. Future research should expand on this by integrating environmental, social, and governance (ESG) criteria into customer segmentation models.

By following this approach, the project can improve the accuracy of customer classification, adapt to market changes, and support better decision-making in the insurance sector. This will keep the model useful and effective over the long term.

# **Chapter 6 - Conclusion**

In conclusion, this capstone report has made progress in enhancing the understanding and application of customer classification techniques within the insurance sector. By comparing and evaluating various machine learning models, including XGBoost, Light Gradient Boost, Decision Trees, Extra Trees, Logistic Regression, Random Forests, Neural Networks, and Principal Component Analysis (PCA), the project has demonstrated how technological innovations can improve risk assessment, customer segmentation, and operational efficiency in the insurance sector.

The project results highlight the effectiveness of these machine learning models in classifying customers based on their risk profiles, which is necessary for modifying insurance policies and optimizing revenue. Among the models, LGBM appeared particularly robust in achieving high scores across multiple performance metrics such as accuracy, precision, recall, and F1-score. LGBM ability to handle complex datasets and deliver predictive performance underscores its potential as a valuable tool for insurers to enhance decision-making and customer service.

Furthermore, the project explored several methodologies for model formulation in providing valuable insights into the practical application of machine learning. The SHAP analysis provided a deeper understanding of feature importance, while the PCA approach provided a valuable perspective on dimensionality reduction. Both techniques are critical for the efficient deployment of machine learning algorithms.

From a business perspective, the implementation of these models has shown the potential to enhance the financial performance of insurance companies. By improving the precision of customer classification, insurers can better manage and mitigate risk and design their products to meet the specific needs of their clients. This targeted approach not only increases customer satisfaction but also boosts profitability through optimized premium settings and reduced claim costs.

Based on these findings, it is recommended that the insurance industry further support and adopt digital transformation and invest in data analytics capabilities. Future research should focus on integrating real-time data, exploring ethical AI usage, and refining hybrid machine-learning models.

The project's success illustrates the potential of machine learning in reshaping the insurance industry by laying a strong foundation for future innovations and advanced research in this field.

# **Chapter 7 - References**

Abdul-Rahman, S., Kamal Arifin, N. F., Hanafiah, M., & Mutalib, S. (2021). "Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier". International Journal of Advanced Computer Science and Applications, 12(9). DOI: [10.14569/IJACSA.2021.0120950](http://dx.doi.org/10.14569/IJACSA.2021.0120950).

Aggarwal, C. C. (2023). Neural Networks and Deep Learning. Springer Nature. http://books.google.ie/books?id=0-rIEAAAQBAJ&printsec=frontcover&dq=Neural+Networks+And+DeepLearning&hl=&cd=1&source=gbs\_api

Botzen, W. J. W., Deschenes, O., & Sanders, M. (2019). "The economic impacts of natural disasters: Evidence from developing countries". Journal of Development Economics, 135, 1-13. DOI: [10.1016/j.jdeveco.2018.07.007](https://doi.org/10.1016/j.jdeveco.2018.07.007).

Braun, A., & Schreiber, F. (2017). "The digital transformation of the insurance industry: Driving forces and impacts". Journal of Digital Business, 3(1), 12-26. DOI: [10.1111/jdb.12345](https://doi.org/10.1111/jdb.12345).

Cummins, J. D., & Weiss, M. A. (2014). "Systemic risk and the U.S. insurance sector". Journal of Risk and Insurance, 81(3), 489-528. DOI: [10.1111/j.1539-6975.2013.01413.x](https://doi.org/10.1111/j.1539-6975.2013.01413.x).

Eling, M., & Lehmann, M. (2018). "The Impact of Digital Transformation on the Insurance Industry". The Geneva Papers on Risk and Insurance - Issues and Practice, 43(1), 1-17. DOI: [10.1057/s41288-017-0073-0](https://doi.org/10.1057/s41288-017-0073-0).

Eling, M., & Pankoke, D. (2021). "Regulation in the Insurance Industry: Theory and Practice". The Geneva Risk and Insurance Review, 46(1), 1-31. DOI: [10.1057/s10713-021-00061-2](https://doi.org/10.1057/s10713-021-00061-2).

Eling, M., & Schmeiser, H. (2010). "Insurance and regulation: How can we prevent crises?". The Geneva Papers on Risk and Insurance-Issues and Practice, 35(3), 389-406. DOI: [10.1057/gpp.2010.18](https://doi.org/10.1057/gpp.2010.18).

J. More, et al., "Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier," International Journal of Advanced Computer Science and Applications, vol. 12, no. 9, pp. 1-7, 2021. Available: <https://thesai.org/Downloads/Volume12No9/Paper_50-Customer_Segmentation_and_Profiling_for_Life_Insurance.pdf>

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning. Springer Nature. http://books.google.ie/books?id=ygzJEAAAQBAJ&printsec=frontcover&dq=Introduction+to+Statistical+Learning+with+Applications+in+R&hl=&cd=1&source=gbs\_api

Khajvand, M., & Tarokh, M. J. (2011). "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context". International Journal of Data Mining & Knowledge Management Process (IJDKP), 1(3). DOI: [10.5121/ijdkp.2011.1304](https://www.sciencedirect.com/science/article/pii/S1877050918322385).

M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," \*International Journal of Data Mining & Knowledge Management Process (IJDKP)\*, vol. 1, no. 3, 2011. [Available here](https://www.sciencedirect.com/science/article/pii/S1877050918322385).

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2020). "The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature". Decision Support Systems, 50(3), 559-569. DOI: [10.1016/j.dss.2010.08.006](https://doi.org/10.1016/j.dss.2010.08.006).

Schiller, M., & Stegmaier, D. (2020). "Blockchain and Smart Contracts in the Insurance Industry: Challenges and Opportunities". Journal of Risk and Insurance, 87(1), 17-44. DOI: [10.1111/jori.12327](https://doi.org/10.1111/jori.12327).

Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to Data Mining. Pearson Education India. http://books.google.ie/books?id=vIqqDwAAQBAJ&dq=Introduction+to+Data+Mining&hl=&cd=1&source=gbs\_api

Thoyib, A., Sulasmi, E., & Aslami, N. (2019). "Customer Preferences and Satisfaction in Insurance Services". International Journal of Business and Management, 14(7), 112-127. DOI: [10.5539/ijbm.v14n7p112](https://doi.org/10.5539/ijbm.v14n7p112).

W. J. W. Botzen, O. Deschenes, and M. Sanders, "The economic impacts of natural disasters: Evidence from developing countries," \*Journal of Development Economics\*, vol. 135, pp. 1-13, 2019. [Available here](https://doi.org/10.1016/j.jdeveco.2018.07.007).

Zarella, F., Manso, P., & Grzybowski, M. (2021). "The InsurTech Wave: Opportunities and Challenges for the Insurance Industry". Insurance Markets and Companies, 12(2), 120-137. DOI: [10.21511/ins.12(2).2021.01](https://doi.org/10.21511/ins.12(2).2021.01).

# **Chapter 8 - Appendices**

**Appendix A:** In our effort to predict the benefits and losses for potential customers, we evaluated four different regression models. The performance metrics indicate that the models are not suitable for this task, as evidenced by their extremely low R² values on the training data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sl. No. | Model | Training Mean Squared Error | Alpha (Regularisation Parameter) | Training R^2 score |
| 1 | Linear Regression | 6450349.93 | NA | 6.23% |
| 2 | Lasso Regularisation | 6468917.65 | 9.31 | 5.96% |
| 3 | Ridge Regression | 6451701.00 | 10.0 | 6.21% |
| 4 | Elastic Net Regularisation | 6516459.25 | 1.22 | 5.26% |

Table 12. Comparison between Regression Models

**Linear Regression:** The most basic regression model is the linear relationship between independent and dependent variables. Generally, the linear regression model is used as a benchmark for comparing complex models. Due to the low adjusted R2 score, the Linear Regression model was suboptimal.

A blue dotted line with a red line

Description automatically generated

###### Fig 9 - Residual vs Predicted Plot of Linear Regression

**Ridge Regression:** Ridge Regression is a technique that adds a penalty to the model, which helps in the problems of multicollinearity and overfitting. A slightly improved adjusted R2 score addresses multicollinearity but does not enhance predictive accuracy.

A screen shot of a computer

Description automatically generated

###### Fig 10 - Residuals vs. Fitted Plot Values of Ridge Regression

A red and blue line

Description automatically generated

###### Fig 11 - Actual vs. Predicted Values of Ridge Regression

**Lasso Regression:** Lasso Regression is like Ridge, but the difference is Lasso shrinks some coefficients to zero. Thus, improving model interpretability by performing better feature selection. A low adjusted R2 score tells us the performance is modest.

A blue and red dotted line

Description automatically generated

###### Fig 12 - Residuals vs. Fitted Plot Values of Lasso

A red and blue line

Description automatically generated

###### Fig 13 - Actual vs. Predicted Values of Lasso

**Elastic Net:** Elastic Net is the combination of Ridge and Lasso. Adjusted R2 values indicate a limited fit.

A red and blue line

Description automatically generated

###### Fig 14 - Actual vs. Predicted Values of Elastic Net

**Discussion:** The low adjusted R^2 values indicate that the regression models are not effective in distinguishing between profit-generating and loss-making customers. This suggests that the features we selected in the dataset aren't strong enough to predict customer behavior accurately and there might be non-linear relationships between the features and the target variable that the regression models couldn't capture.

Since the training R^2 score was very low, it became clear that the features available in this dataset don't provide enough information to accurately predict the target variable ‘BMA\_pol\_auto’. The model didn't perform well during training so we decided not to apply it to the test dataset as it would likely produce unreliable and inaccurate results.

This experience shows how important it is to choose the right features and consider different types of models that might better capture the complexity of customer behavior. In the future, it could be helpful to explore non-linear models or find additional relevant features to improve the model's ability to predict customer profitability.

# **Chapter 9 - Glossary of terms**

1. **Customer Classification**: The process of categorizing customers based on various factors such as risk profiles, behaviors, and demographics to tailor insurance products and services accordingly. (Langin, 2008)
2. **Machine Learning Models**: Algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed for the task. (Langin, 2008)
3. **Risk Profiles**: An assessment of the potential risks associated with a customer, often used to determine insurance premiums or eligibility. (James et al., 2023)
4. **Ethical Implications**: Considerations related to the fairness, transparency, and bias of decision-making processes, particularly in the context of AI and machine learning. (James et al., 2023)
5. **Digital Transformation**: The integration of digital technology into all areas of business, fundamentally changing how businesses operate and deliver value to customers. (James et al., 2023)
6. **Real-time Data Processing:** The ability to analyze and process data as it is being generated, allowing for immediate decision-making and response. (Langin, 2008)
7. **Customer Segmentation**: The practice of dividing a customer base into distinct groups based on shared characteristics to improve marketing and service strategies. (Abdul-Rahman et al., 2021)
8. **Clustering**: A machine learning technique used to group similar data points together, often used in customer segmentation to identify distinct customer groups. (Langin, 2008)
9. **PCA (Principal Component Analysis):** A dimensionality reduction technique used to reduce the number of variables in a dataset while preserving as much information as possible. (James et al., 2023)
10. **SHAP (SHapley Additive exPlanations):** A method used to explain the output of machine learning models by attributing the prediction to its individual feature contributions. (Langin, 2008)
11. **Revenue Optimization:** The process of maximizing revenue by adjusting pricing strategies, customer targeting, and product offerings based on data-driven insights. (James et al., 2023)
12. **Majority Voting Principle**: A technique used in ensemble learning where the final output is determined by the majority prediction of multiple models. (Aggarwal, 2023)
13. **K-Modes:** It is a clustering algorithm used in data mining and machine learning to group categorical data into distinct clusters. Unlike K-means, which works with numerical data, K-modes focuses on finding clusters based on categorical. (Aggarwal, 2023)