Lisbon School of Economics & Management
Universidade de Lisboa

MASTERS IN MANAGEMENT (MIM)

MASTERS FINAL WORK

PROJECT

PREDICTING CREDIT INSURANCE SUBSCRIPTION: A COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR CLIENT RANKING

SARA ISABEL RITA GUTIERREZ

SUPERVISOR: PROF. JOÃO AFONSO BASTOS

JURY:
PRESIDENT: PROF. MARIA EDUARDA SOARES
RAPPORTEUR: PROF. LUÍS SILVEIRA SANTOS
SUPERVISOR: PROF. JOÃO AFONSO BASTOS

FEBRUARY - 2025

GLOSSARY

AUC Area Under the ROC Curve. ii–v, 1, 5–7, 24–27, 29–31, 33, 35

EDA Exploratory Data Analysis. ii, 9

EFB Exclusive Feature Bundling. ii, 5, 22, 23

GOSS Gradient-based One-Side Sampling. ii, 5, 22

JEL Journal of Economic Literature. ii–iv

MDA Multiple Discriminant Analysis. ii, 3

ROC Receiver Operating Characteristic. ii, 6, 25, 26

SHAP SHapley Additive exPlanations. ii–v, vii, 2, 7, 24, 26, 31–35

RESUMO, PALAVRAS-CHAVE E CÓDIGOS JEL

Esta dissertação investiga o desenvolvimento de um modelo preditivo de ordenação com o objetivo de aumentar a eficiência da equipa de Inside Sales de uma instituição financeira, através da previsão da subscrição de seguros de proteção ao crédito. Recorrendo a registos de crédito ao consumo de 2024, o estudo passa por uma preparação de dados abrangente, análise exploratória e feature engineering para tratar um conjunto de dados de elevada dimensionalidade e com uma representação desigual entre classes. Foram implementados e comparados vários modelos preditivos, nomeadamente Regressão Logística, Random Forest, LightGBM e CatBoost, com afinação de hiperparâmetros orientada por validação cruzada e avaliação baseada em múltiplas métricas e análise por decis.

Os resultados revelam que, embora a Regressão Logística seja valorizada pela sua interpretabilidade, o seu desempenho preditivo é inferior ao de métodos mais avançados baseados em ensemble e boosting. O modelo Random Forest demonstra forte capacidade discriminativa global, evidenciada pelos seus valores superiores de AUC e coeficiente de Gini no conjunto de teste, mas apresenta sinais de sobreajuste na análise por decis. O LightGBM revela um desempenho competitivo, destacando-se especialmente pelo valor do F1 score na classe positiva. Contudo, é o CatBoost que se destaca como o modelo mais equilibrado, com desempenho consistente nas métricas validadas por validação cruzada, nas avaliações no conjunto de teste e na análise por decis. Adicionalmente, os valores SHAP oferecem uma análise pormenorizada da importância das variáveis, identificando vários atributos-chave como determinantes na previsão da subscrição.

Este estudo representa um contributo relevante para a literatura emergente sobre a subscrição de seguros de proteção ao crédito, um tema ainda pouco explorado tanto em contextos académicos como empresariais. O modelo de ordenação desenvolvido constitui um avanço significativo face à abordagem tradicional baseada em Regressão Logística, oferecendo ganhos em precisão preditiva e interpretabilidade, permitindo decisões mais informadas e maior eficiência operacional. Investigações futuras deverão incidir sobre o aperfeiçoamento dos parâmetros do modelo, a exploração de técnicas nativas de tratamento de variáveis categóricas e a implementação de estratégias de agregação de modelos para otimizar o desempenho e a transparência na previsão da subscrição. Adicionalmente, recomenda-se a inclusão de uma dimensão económica relacionada com o potencial de comissão dos clientes, com vista a aumentar a relevância prática do modelo.

PALAVRAS-CHAVE: CatBoost; LightGBM; Modelação Preditiva; Random Forest; Regressão Logística; Subscrição de Seguro de Proteção ao Crédito.

CÓDIGOS JEL: C45; C53; C55; G21; G22; G32; G33.

ABSTRACT, KEYWORDS, AND JEL CODES

This thesis investigates the development of a predictive ranking model to enhance the efficiency of an inside sales team at a financial institution by accurately forecasting credit insurance subscription. Using consumer credit records collected in 2024, the study employs comprehensive data preprocessing, exploratory data analysis, and feature engineering to prepare high dimensional, imbalanced data for modelling. Predictive models such as Logistic Regression, Random Forest, LightGBM, and CatBoost were implemented and compared, with hyperparameter tuning guided by cross-validation and evaluation via multiple metrics and decile analysis.

The findings reveal that while Logistic Regression is considered within the field to have superior interpretability, its overall predictive performance is inferior to that of more advanced ensemble and boosting methods. Random Forest shows high global discrimination, as evidenced by its superior test set AUC and Gini coefficients, yet it exhibits signs of overfitting in the decile analysis. LightGBM achieves competitive performance, particularly in its F1 score for the positive class, but CatBoost emerges as the most balanced model, with consistent performance across cross-validated metrics, test set evaluations, and decile analysis. Additionally, SHAP values provide granular insights into feature importance, identifying several key variables as decisive drivers of subscription predictions.

This research contributes significantly to the emerging literature on credit insurance subscription, a topic that remains underexplored in both academic and business contexts. The developed ranking model represents a substantial advancement beyond the traditional baseline of Logistic Regression, offering enhanced predictive accuracy and interpretability that enable more informed decision-making and improved operational efficiency. Future research should focus on further refining model parameters, exploring native categorical processing, and investigating ensemble strategies to optimise performance and transparency in predicting subscription. Additionally, incorporating an economic dimension to capture clients' commission potential is recommended to enhance the model's practical relevance.

KEYWORDS: CatBoost; Credit Insurance Subscription; LightGBM; Logistic Regression; Predictive Modelling; Random Forest.

JEL CODES: C45; C53; C55; G21; G22; G32; G33.

TABLE OF CONTENTS

LIST OF FIGURES

List of Tables

ACKNOWLEDGEMENTS

Firstly, I would like to thank Professor João Bastos for captivating my interest during his classes and for agreeing to guide me through this important stage of my academic journey. His expertise and thoughtful advice have paved the way for this work, which is undoubtedly enriched by his mentorship.

My sincere gratitude also goes to the financial institution that provided the dataset forming the basis of this research study, thereby presenting an interesting challenge with practical real-world implications. I am especially grateful for the opportunity to explore this dataset and potentially make a meaningful impact with this project. I would like to acknowledge the invaluable support and insights provided by Jorge Mendes, whose modelling expertise and business acumen have been the cornerstone of this research. His patience and dedication have greatly contributed to my learning and development throughout this project.

I must also express my heartfelt thanks to my outstanding team of colleagues, who were there for me every day, providing both motivation and support during challenging times. Their encouragement and the light-hearted moments we shared have been instrumental in keeping my spirits high, and I am profoundly grateful for their unwavering support. I extend my gratitude to all my friends from work, university, and beyond for their steadfast support and for the thought-provoking discussions that enriched my experience throughout this process.

Lastly, I would like to thank my loved ones for continuously inspiring me to perform to the best of my abilities while forging my own path in whichever field I am passionate about. My deepest appreciation goes to my mother, a pillar of resilience and the strongest woman I know; to my father, for always offering a shoulder to lean on during moments of fatigue; to my brother, whose uplifting spirit has consistently bolstered my confidence; to my grandparents, whose lifelong support has helped shape the person I am today; and to my wonderful boyfriend, for being my steadfast rock, for fully believing in my abilities, and for encouraging me to shine wherever I venture.

1 INTRODUCTION

The ability to accurately predict consumer behavior has long been a critical focus in the financial industry, particularly in areas where customer decision-making directly impacts business profitability. One such area is credit insurance, a financial product designed to protect lenders and borrowers from potential defaults due to unforeseen circumstances. Despite its benefits, not all clients choose to purchase credit insurance, making it challenging for financial institutions to identify the most promising leads. This study seeks to develop a machine learning-based propensity model for a financial institution to classify clients based on their likelihood of subscribing to credit insurance after obtaining credit. The primary objective is to provide the company's inside sales team with a data-driven approach to prioritize potential customers for insurance subscription, improve efficiency, and increase the conversion rate of insurance sales. Using machine learning, this research aims to enhance targeted sales efforts, streamline marketing strategies, and ultimately optimise customer engagement in financial services.

The literature reveals a significant evolution in credit scoring and risk assessment methods. Early studies by Altman (1968) established the foundation for quantitative evaluation using techniques such as discriminant analysis, while subsequent research by Wiginton (1980) proposed maximum likelihood estimation of the logit model as a superior alternative for scoring consumer credit behaviour. Later, Thomas (2000) further advanced the field, and methods such as logistic regression became prized for their simplicity and interpretability (Hand & Henley, 1997). However, as data complexity and availability increased, more sophisticated machine learning techniques emerged. Research by Khandani et al. (2010) demonstrated that these advanced methods capture nonlinear relationships and complex interactions that traditional models might overlook. Furthermore, the work of Thomas et al. (2002) and benchmarking studies by Baesens et al. (2003) and Lessmann et al. (2015) highlight that while traditional models remain robust in certain contexts, ensemble methods—particularly Random Forest, LightGBM, and CatBoost—offer enhanced discriminatory power and calibration. These developments underscore the necessity of balancing predictive accuracy with interpretability in financial applications.

To address this research question, a comprehensive methodology was developed that encompasses extensive data preprocessing, exploratory data analysis, and feature engineering on a high-dimensional, imbalanced dataset of consumer credit records collected in 2024. Multiple predictive models, including Logistic Regression, Random Forest, LightGBM, and CatBoost, were implemented with rigorous hyperparameter tuning via cross-validation. Evaluation metrics—such as AUC, Gini coefficient, F1 score, and decile analysis—were employed to assess and compare the models' abilities to rank clients ef-

fectively by their likelihood of subscribing to credit insurance. In addition, SHAP values were utilised to provide both global and local interpretability of the best performing model, elucidating the contribution of individual features to the final predictions.

This study's primary contribution is the development of an overall robust, data-driven ranking model that significantly advances the industry baseline of logistic regression. By integrating state-of-the-art ensemble and boosting methods with comprehensive evaluation techniques, the research offers new insights into feature importance and model interpretability through the application of SHAP values. These findings not only enhance predictive accuracy and provide practical guidelines for improving operational efficiency in credit insurance sales, but also demonstrate that it is possible to employ more complex methods in credit-related areas while still retaining explainability and actionable real-world insights.

Due to the focus of the master's in sustainability and the SDGs, it is important to note that this work directly aligns with SDG 9 (Industry, Innovation and Infrastructure) by integrating advanced machine learning techniques into credit insurance subscription modelling. More specifically, by developing a comprehensive, data-driven propensity model, it contributes to Target 9.5, which focuses on fostering innovation and technological advancement across all industries, particularly in developed countries. This work not only enhances predictive performance through state-of-the-art methodologies, but also supports the digital transformation of the financial industry by improving risk assessment and operational efficiency, ultimately contributing to a more resilient and innovative infrastructure.

The subsequent chapters are organized as follows: Chapter 2 presents a detailed literature review on quantitative models, the evolution of credit-related modelling, and the specific modelling techniques used in this study. Chapter 3 outlines the methodology, including data overview and preprocessing, Exploratory Data Analysis, model implementation, and optimization strategies. Chapter 4 details the results, including a comparative analysis of the models based on various evaluation metrics and decile analysis. Finally, Chapter 5 discusses the findings in the context of existing literature, outlines implications for practice, and suggests directions for future research.

2 LITERATURE REVIEW

2.1 Credit Insurance and Risk Mitigation in Consumer Lending

Credit insurance has played a pivotal role in mitigating the risks associated with consumer lending for over a century. Since its introduction in 1919, credit insurance products have been designed to protect both borrowers and lenders from the financial fallout of adverse events such as death, disability, or involuntary unemployment. By either extinguishing a consumer's debt or suspending periodic payments when such events occur, these products distribute risk across a broad portfolio, thus reducing the likelihood of catastrophic financial loss for any single party.

This risk-spreading mechanism not only benefits individual consumers by shielding them from unexpected hardships but also contributes to the stability of credit markets by minimizing defaults. As discussed by Durkin and Elliehausen (2018), the evolution of credit insurance has transformed the landscape of consumer lending. The widespread adoption of these products demonstrates the financial industry's commitment to managing risk through innovative solutions. This historical and practical background provides an essential context for current research efforts aimed at predicting client subscription to credit insurance.

Due to the limited availability of literature specifically addressing machine learning models for credit insurance, this review also draws on research from related fields such as credit scoring and credit risk assessment to provide a comprehensive methodological framework.

2.2 Quantitative Models in Financial Contexts

Quantitative models have long been the backbone of financial decision-making, offering systematic approaches to risk assessment and investment strategy. Altman (1968) was among the first to demonstrate that financial ratio analysis, combined with Multiple Discriminant Analysis (MDA), could effectively predict corporate bankruptcy. Subsequently, Wiginton (1980) proposed maximum likelihood estimation of the logit model as an alternative to the linear discriminant model commonly used in credit scoring. Their findings indicated that the logit model yields parameter estimates producing a higher proportion of correct classifications, offering a more accurate method for scoring consumer credit behaviour.

Building on this foundation, later studies—such as those by Thomas et al. (2002)—have extended the use of quantitative models to various aspects of credit evaluation and

risk management. These models enable financial institutions to assess the likelihood of default, estimate potential losses, and make informed decisions regarding credit allocation. By leveraging historical financial data and sophisticated statistical techniques, quantitative models facilitate a more objective evaluation of risk, thereby reducing reliance on subjective judgment and contributing to more stable financial markets.

2.3 Evolution of Credit Scoring and Risk Assessment

The field of credit scoring has undergone a significant transformation over the past few decades. In its early stages, credit scoring relied on methods such as discriminant analysis and Logistic Regression to classify applicants into "good" and "bad" risk categories (Hand and Henley, 1997; Thomas, 2000). These early techniques were prized for their simplicity and transparency, as they provided interpretable outputs that allowed lenders to understand which variables influenced credit decisions.

As data availability increased and computational power advanced, the evolution of credit scoring embraced more sophisticated machine-learning approaches. Khandani et al. (2010) demonstrated that by using machine-learning algorithms, it is possible to capture nonlinear relationships and intricate patterns within consumer data that traditional methods might miss. Moreover, Thomas et al. (2002) observed that while Logistic Regression remains the most commonly used technique for developing scorecards, a range of alternative methods, including mathematical programming, neural networks, and classification trees, have been explored to capture more complex borrower behaviour. These alternatives often lack the transparency of Logistic Regression, underscoring the enduring challenge of balancing predictive accuracy with interpretability in credit risk assessment.

2.4 Specific Models for Credit Risk and Scoring

Modern credit risk modelling employs a variety of techniques, each with its distinct advantages and limitations. Logistic Regression has traditionally served as the industry standard due to its simplicity and clear interpretability. As highlighted by Hand and Henley (1997) and Thomas (2000), Logistic Regression provides probabilistic predictions with coefficients that directly reflect the influence of each predictor. This transparency allows for rigorous statistical testing and straightforward communication of credit decisions, a crucial aspect in regulated environments.

Building on this foundational method, ensemble approaches have been introduced to capture complex, nonlinear relationships that linear models may overlook. Random Forest (Hastie et al., 2009) is an extension of bagging that build a large ensemble of de-correlated

decision trees by randomly selecting a subset of predictors at each split, thereby reducing the overall variance without substantially increasing bias. The final prediction is then obtained by averaging the outputs of individual trees by majority voting for classification, effectively capturing complex interactions and nonlinear relationships.

More recent advancements in gradient boosting have led to the development of highly efficient frameworks that address the challenges of large-scale credit scoring. LightGBM, as presented by Ke et al. (2017), employs innovative techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to significantly reduce computational overhead while maintaining high accuracy. Its design is particularly well-suited for environments characterized by vast feature sets and imbalanced data distributions. Following LightGBM, CatBoost (Prokhorenkova et al., 2018) further refines the gradient boosting paradigm by introducing ordered boosting and specialized algorithms for processing categorical variables without extensive preprocessing. This targeted approach not only mitigates prediction shift but also enhances model robustness on datasets with mixed type variables.

2.5 Benchmarking and Comparative Studies

Benchmarking studies are crucial for validating the effectiveness of various credit scoring methods and for guiding model selection in practice. Baesens et al. (2003) evaluated various classification algorithms on eight real-life credit scoring datasets, comparing traditional methods—such as Logistic Regression, discriminant analysis, and k-nearest neighbors—with more advanced techniques including neural networks. Their findings revealed that although sophisticated models often achieved high accuracy and AUC, simpler, more interpretable models also performed robustly regarding evaluation metrics, suggesting that many credit scoring datasets are only weakly nonlinear.

Later on, Lessmann et al. (2015) extended this evaluation by benchmarking novel classification algorithms and heterogeneous ensemble methods. Their study found that while some advanced classifiers yielded statistically significant improvements over traditional models, the gains were often marginal from a managerial perspective. Notably, Random Forest consistently emerged as a strong benchmark, reinforcing the idea that any incremental improvement must be balanced against considerations of interpretability and operational complexity in practical credit scoring applications.

2.6 Tuning Techniques

Modern credit scoring applications often involve high-dimensional datasets characterized by a multitude of features and significant class imbalance. These challenges require meth-

ods that preserve critical information while mitigating the adverse effects of noisy data and skewed class distributions. For instance, tree-based methods, such as those implemented in LightGBM and CatBoost, inherently perform feature selection through their splitting criteria and regularization strategies, thereby reducing dimensionality without discarding potentially informative predictors.

Imbalanced datasets, where the number of positive target cases is much lower than negative, can severely compromise the performance of standard learning algorithms. He and Garcia (2009) emphasize that imbalanced learning presents unique challenges, as most algorithms assume balanced class distributions or equal misclassification costs. To address this, techniques such as stratified sampling, cost-sensitive learning, and the use of comprehensive evaluation metrics are essential for accurately assessing model performance under these conditions.

In parallel, model optimization is achieved through regularization and hyperparameter tuning. Techniques such as L1 (lasso) and L2 (ridge) regularization constrain coefficient magnitudes to prevent overfitting—especially in high-dimensional settings where multicollinearity poses a challenge (Friedman et al., 2010; Hastie et al., 2009). Additionally, the elastic net, which combines both L1 and L2 penalties, offers a balanced approach for stabilizing models when predictors are highly correlated. Furthermore, hyperparameter tuning via cross-validation is critical for identifying the optimal balance between bias and variance (Stone, 1974). As described by Kohavi (1995), ten-fold stratified cross-validation provides a robust framework for model selection, ensuring that performance estimates are unbiased and that parameter settings are well-calibrated to generalize to unseen data.

2.7 Evaluation Metrics: F1-Score, AUC, and Gini Coefficient

In the evaluation of credit scoring models, it is essential to employ metrics that capture the nuances of imbalanced datasets. The F1 score, which combines precision and recall, is particularly useful when the costs of false positives and false negatives differ (Powers, 2011). Complementary metrics, such as the Area Under the ROC Curve (AUC) and the derived Gini coefficient—calculated as $\text{Gini} = 2 \times \text{AUC} - 1$—provide a robust, scalar measure of a model's overall discriminatory power, offering a more holistic assessment that is less sensitive to class imbalance (Fawcett, 2006).

In addition to these global performance metrics, decile analysis is employed to assess model calibration and ranking capability. By dividing the population into ten groups based on predicted risk scores, decile analysis compares observed rates with predicted probabilities across segments. This method provides granular insight into the model's

ability to rank individuals accurately, complementing the AUC and Gini measures by demonstrating how well the model discriminates between different levels.

2.8 Model Interpretability: Coefficients and SHAP Values

In regulated financial environments, model interpretability is as crucial as predictive accuracy. Traditional models like Logistic Regression offer straightforward interpretability through their coefficients, which directly quantify the relationship between predictors and the likelihood of a specific outcome. This transparency is vital for regulatory compliance and for providing clear explanations to stakeholders.

However, as more complex models such as Random Forest, LightGBM, and CatBoost are increasingly used in credit-related problems, their decision-making processes become less transparent. To address this challenge, Lundberg et al. (2020) introduced TreeExplainer, which leverages SHAP (SHapley Additive exPlanations) values to decompose individual predictions into additive contributions from each feature. This approach not only guarantees local accuracy but also provides a global understanding of feature impacts. By offering detailed insights into how each variable influences the model's predictions, SHAP values enhance the interpretability of complex models, ensuring that they remain accessible and trustworthy to both developers and regulators.

3 METHODOLOGY

In this section, the dataset is described, along with the exploratory and preprocessing methods applied, the models tested, and the evaluation metrics employed.

3.1 Data Overview and Preprocessing

This study utilised data from a consumer finance company that offers a diverse array of credit solutions—including unsecured personal loans, revolving credit, and complementary services such as debt consolidation and credit insurance—to meet various financial needs. The dataset, collected in 2024, pertains to clients who have acquired personal credit from the institution. To protect client confidentiality and safeguard company assets, all features were anonymized; consequently, this impacts the interpretation of features and the analysis of feature importance in the modelling process. Nonetheless, it is known that the features represent client characteristics, particularly financial conditions, with some of the information being supplied by the Bank of Portugal.

The dataset comprises 20164 observations across 38 features. A detailed characterization of these features, including their data types, is presented in Table I:

TABLE I: OVERVIEW OF FEATURES.

Feature	Data Type	Unique Values	Missing Values (% of Total)
ID	int64	20164	0
target	int64	2	0
VAR_1	object	14	0
VAR_2	int64	56	0
VAR_3	object	5	0
VAR_4	object	5	9 (0.04)
VAR_5	object	2	0
VAR_6	int64	2	0
VAR_7	int64	2673	0
VAR_8	object	7	0
VAR_9	int64	2	0
VAR_10	int64	2	0
VAR_11	int64	2	0
VAR_12	int64	564	0
VAR_13	int64	2	0
VAR_14	int64	58	0
VAR_15	float64	3321	11822 (58.63)
VAR_16	int64	2	0
VAR_17	int64	15	0
VAR_18	float64	11363	2177 (10.80)
VAR_19	int64	12178	0
VAR_20	int64	5399	0
VAR_21	int64	4253	0
VAR_22	int64	6924	0
VAR_23	int64	5320	0
VAR_24	int64	5313	0
VAR_25	int64	4770	0
VAR_26	float64	9391	0
VAR_27	int64	3	0
VAR_28	int64	3	0
VAR_29	int64	3	0
VAR_30	int64	3	0
VAR_31	float64	15268	28 (0.14)
VAR_32	float64	12365	28 (0.14)
VAR_33	float64	3881	0
VAR_34	int64	73	0
VAR_35	object	9	0
VAR_36	float64	9526	0

The *ID* variable serves as a unique identifier for ordering the observations, while the *target* variable is the binary objective feature that partitions the observations into two distinct classes.

$$target = \begin{cases} 1, & \text{client subscribed to credit insurance} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The remaining variables are classified into three types: *VAR_6*, *VAR_9*, *VAR_10*, *VAR_11*, *VAR_13*, and *VAR_16* are binary variables; *VAR_1*, *VAR_3*, *VAR_4*, *VAR_5*, *VAR_8*, *VAR_27*, *VAR_28*, *VAR_29*, *VAR_30*, and *VAR_35* are categorical; and all other features are continuous.

The dataset was preprocessed by the company—before being made available—to handle features with special placeholder values that denote missing or unavailable information. In particular, financial data derived from the Bank of Portugal included specific codes: a value of -9999 was used to indicate that no information was available for the client in question, while -999 was assigned when a particular value was not available. These standardized codes were integrated into the dataset to ensure consistency and to enable subsequent analyses to correctly interpret these missing data markers.

Manual inspection of the variable behaviour reveals an evident correlation among several features. Specifically, when *VAR_16* equals 1, *VAR_17* consistently assumes a value of 0, and for the variables ranging from *VAR_18* through *VAR_30*, the only observed value is -9999. This pattern suggests that *VAR_16* may exert an influential role on its subsequent variables.

As shown in Table I, several variables contain missing values. For *VAR_4*, *VAR_31*, *VAR_32*, the number of missing entries is minimal and these variables do not exhibit the standardized codes described previously; consequently, the affected rows were removed from the dataset. In the case of *VAR_15*, which exhibited the highest number of missing values, the standardized code logic was applied—specifically, a value of -999 was used to indicate that no information is available for that client regarding the particular characteristic. Furthermore, because *VAR_18* is functionally related to *VAR_16*, missing values in *VAR_18* were imputed based on the value of *VAR_16*. Specifically, if *VAR_16* equals 1, then *VAR_18* is set to -9999; otherwise, *VAR_18* is set to -999.

3.2 Exploratory Data Analysis

The EDA aims to characterize the distribution and relationships within the dataset. Initially, the distribution of the *target* variable was visualized using a bar plot that displayed

both relative frequencies and observation counts, thereby providing a comprehensive overview. This visualization is critical for identifying class imbalance—a common challenge in credit-related modelling.

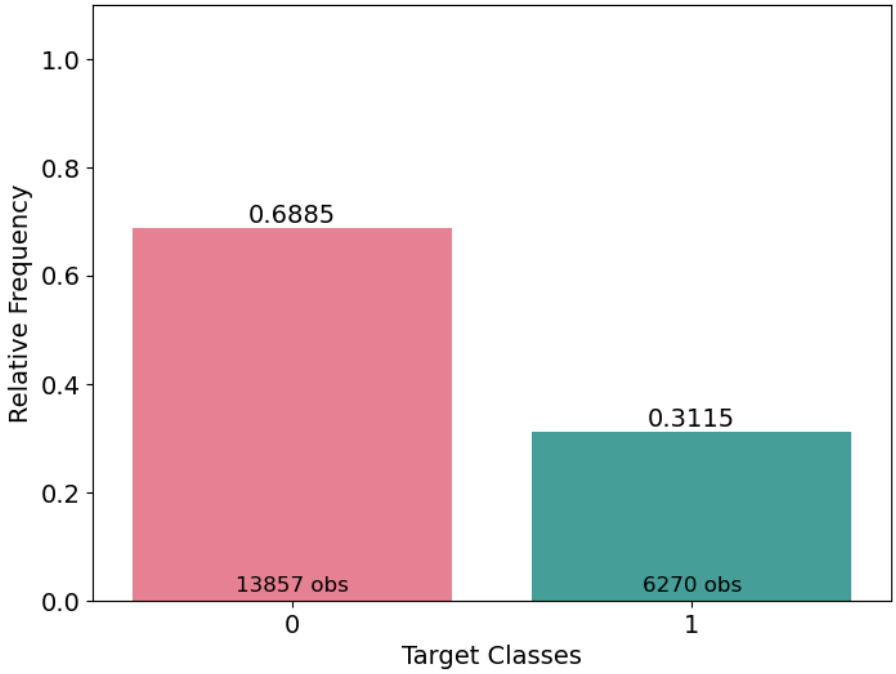


FIGURE 1: Class Distribution of Target Variable.

The dataset exhibits a pronounced class imbalance, as evidenced by Figure 1. Specifically, one class (clients subscribing to credit insurance) is underrepresented compared to the other, which can introduce bias in model training. This imbalance may lead to standard algorithms favoring the majority class, potentially degrading the predictive performance on the minority class. Consequently, it is essential to ensure robust model performance despite the skewed class distribution.

Furthermore, the prevalence of the target variable was visualised across binary and categorical features. These plots display both the mean target values for each feature class—meaning the probability of credit insurance subscription. The corresponding observation counts for each feature class are also plotted, thereby indicating whether a certain class is statistically significant for the model. This dual presentation not only reveals intrinsic relationships between the target and the features but also highlights additional class imbalances at the feature level, thereby aiding in the identification of potential predictors for credit insurance subscription. The insights drawn from these visualizations will later be compared with the results obtained in this study.

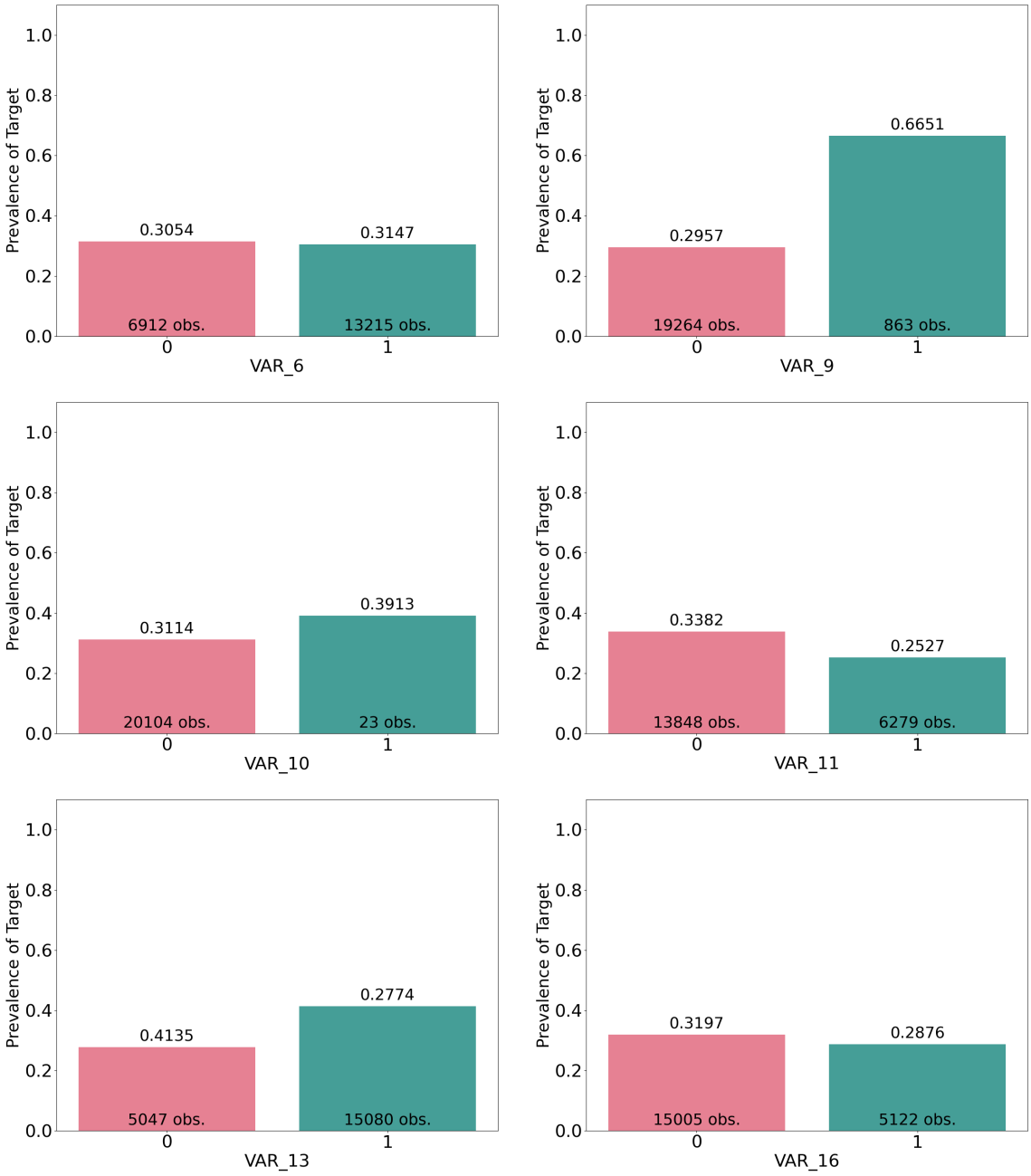
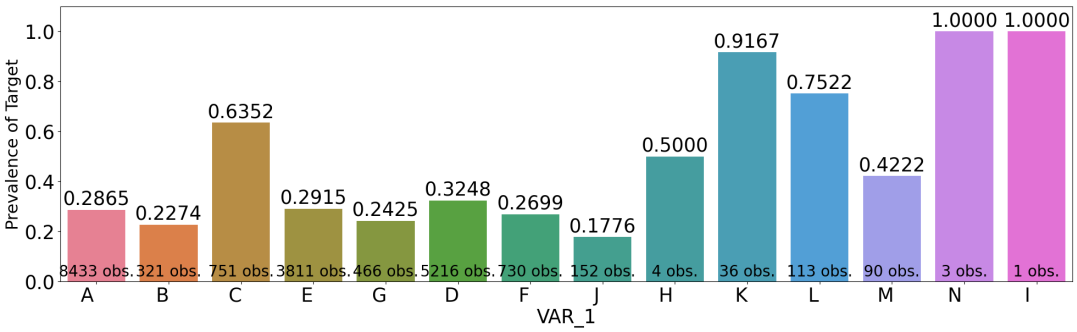
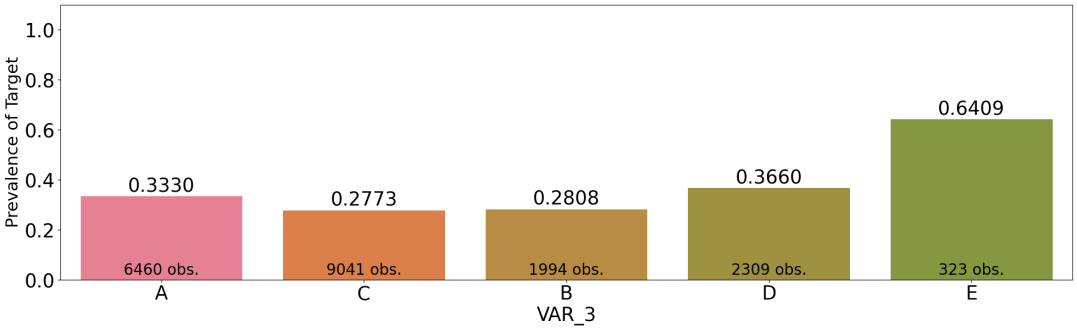
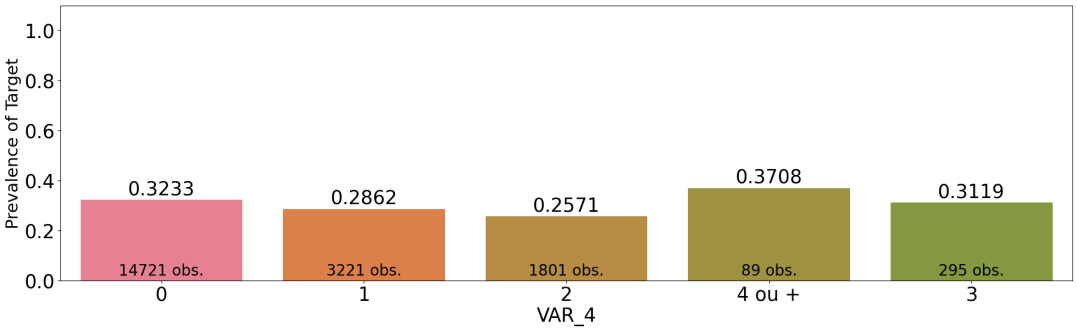


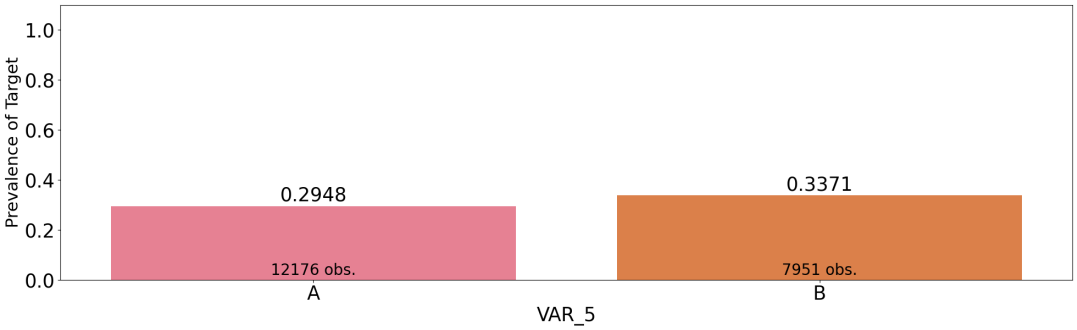
FIGURE 2: Prevalence of Target for Binary Variables.

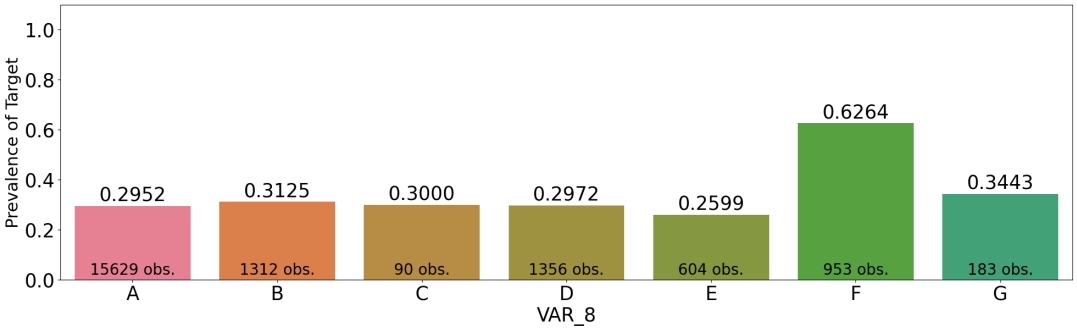
To draw reliable conclusions from Figure 2, both the target prevalence and the observation count at each feature level must be considered. For example, the bar plot for *VAR_9* shows a marked difference in height between the classes, suggesting distinct behaviour; however, the low observation count in the 1 class renders this conclusion tentative. In contrast, *VAR_11* and *VAR_13* appear more robust predictors, as even a slight difference in mean target values is supported by a minority class observation count that exceeds 10% of the majority class. This suggests that these features may be more reliable predictors within the binary variables.


Prevalence of Target
1.0
0.8
0.6
0.4
0.2
0.0
0.2865
0.2274
0.6352
0.2915
0.2425
0.3248
0.2699
0.1776
0.5000
0.9167
0.7522
0.4222
1.0000
1.0000
8433 obs.
321 obs.
751 obs.
3811 obs.
466 obs.
5216 obs.
730 obs.
152 obs.
4 obs.
36 obs.
113 obs.
90 obs.
3 obs.
1 obs.
A
B
C
E
G
D
F
J
H
K
L
M
N
I
VAR_1


Prevalence of Target
1.0
0.8
0.6
0.4
0.2
0.0
0.3330
0.2773
0.2808
0.3660
0.6409
6460 obs.
9041 obs.
1994 obs.
2309 obs.
323 obs.
A
C
B
D
E
VAR_3


Prevalence of Target
1.0
0.8
0.6
0.4
0.2
0.0
0.3233
0.2862
0.2571
0.3708
0.3119
14721 obs.
3221 obs.
1801 obs.
89 obs.
295 obs.
0
1
2
4 ou +
3
VAR_4


Prevalence of Target
1.0
0.8
0.6
0.4
0.2
0.0
0.2948
0.3371
12176 obs.
7951 obs.
A
B
VAR_5


Prevalence of Target
1.0
0.8
0.6
0.4
0.2
0.0
0.2952
0.3125
0.3000
0.2972
0.2599
0.6264
0.3443
15629 obs.
1312 obs.
90 obs.
1356 obs.
604 obs.
953 obs.
183 obs.
A
B
C
D
E
F
G
VAR_8

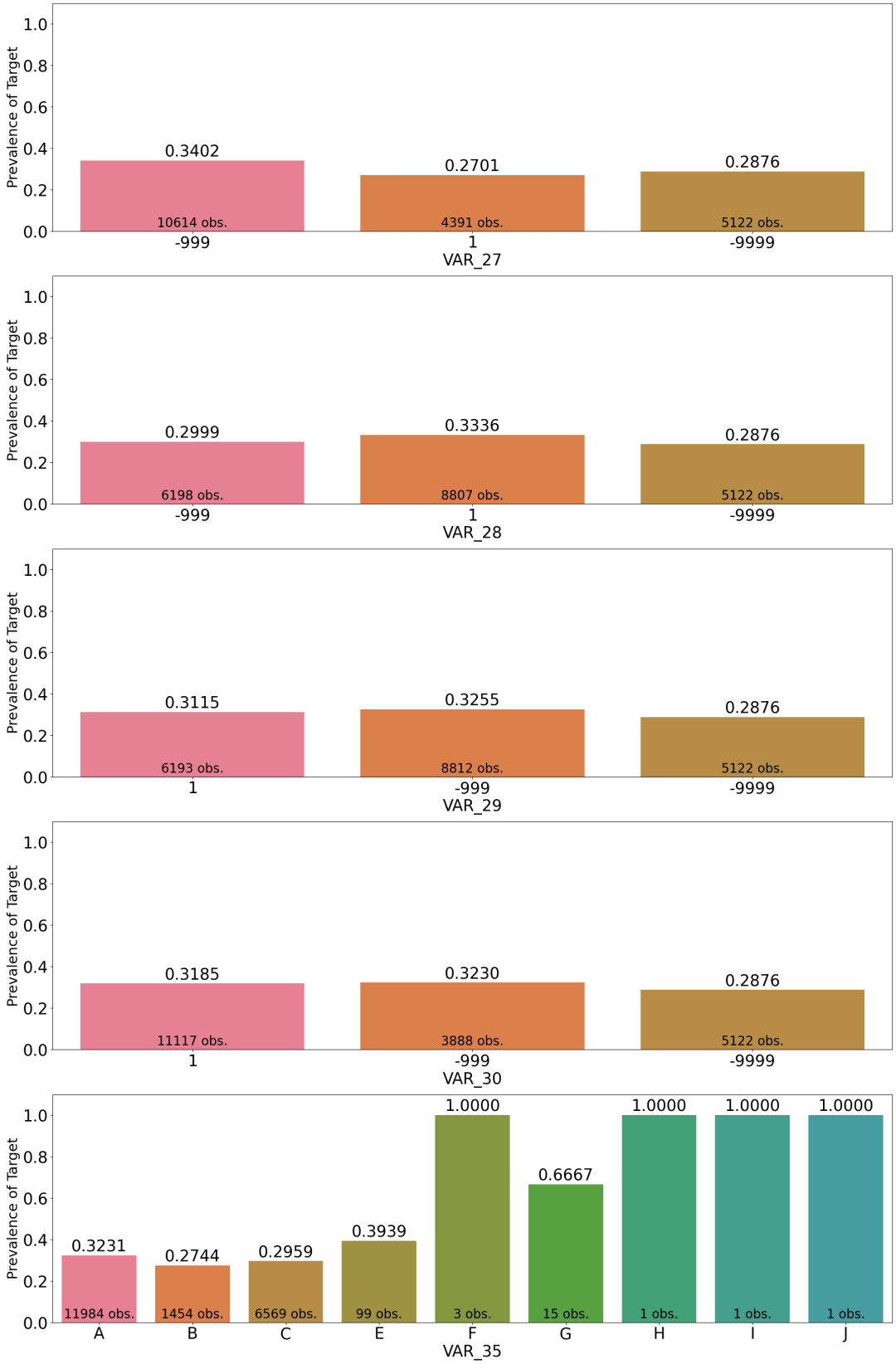
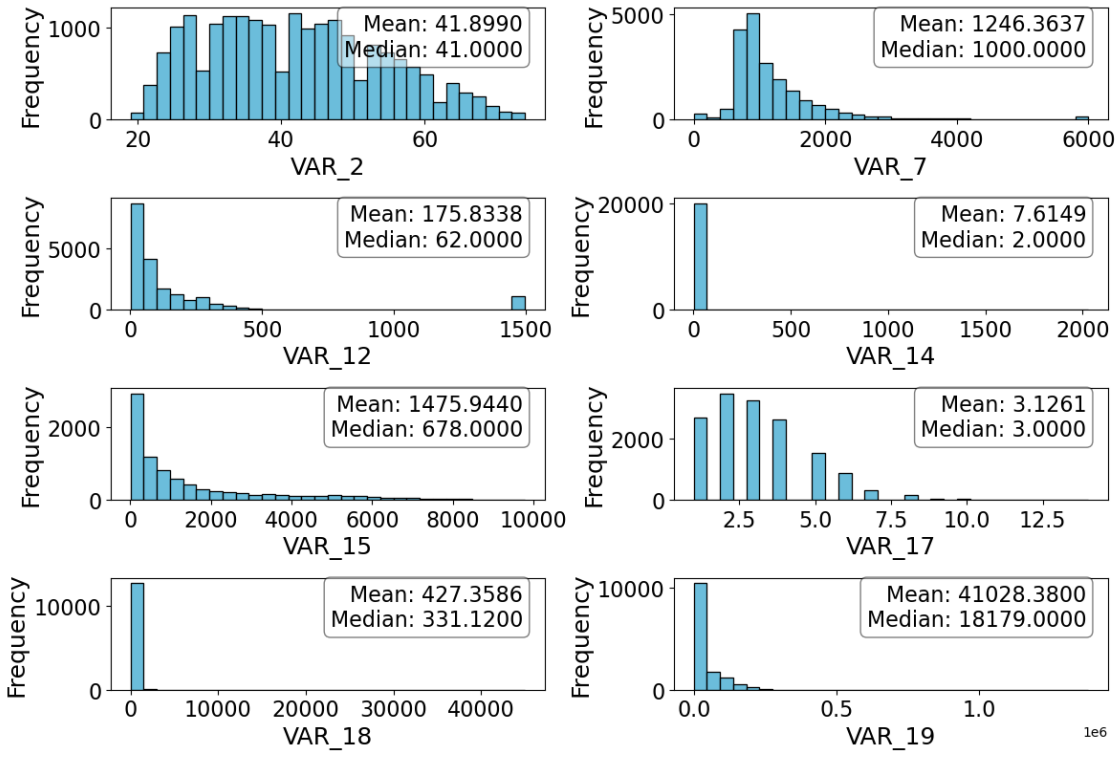


FIGURE 3: Prevalence of Target for Categorical Variables.

Focusing on the target prevalence plots for the categorical variables and applying the same analytical approach as used for the binary variables, several noteworthy patterns emerge. In the case of *VAR_1*, the C and D levels appear to be significant. For *VAR_3*, level D exhibits the highest mean target prevalence with an acceptable observation count, whereas level C shows the lowest mean despite having the highest observation count. Moreover, it is important to identify predictors of the negative class—those features strongly associated with a lower likelihood of credit insurance subscription. In this regard, *VAR_4* indicates that level 2 is linked to the lowest mean target value, and similarly, *VAR_35* demonstrates that level B corresponds to the lowest mean. In contrast, the remaining variables display minimal differences in target prevalence across their levels or suffer from insufficient observation counts, suggesting that they may have limited predictive value.

Shifting focus to the continuous variables, it is essential to understand their distributional characteristics to inform subsequent modelling decisions. Histograms are employed as a fundamental visual tool to reveal the central tendency, dispersion, skewness, and potential outliers within the data. In addition to these visualisations, each plot is annotated with the mean and median values of the variable, providing clear numerical insights into the impact of outliers on the overall distribution. This dual approach helps highlighting areas where additional data transformation or robust modelling techniques might be required.



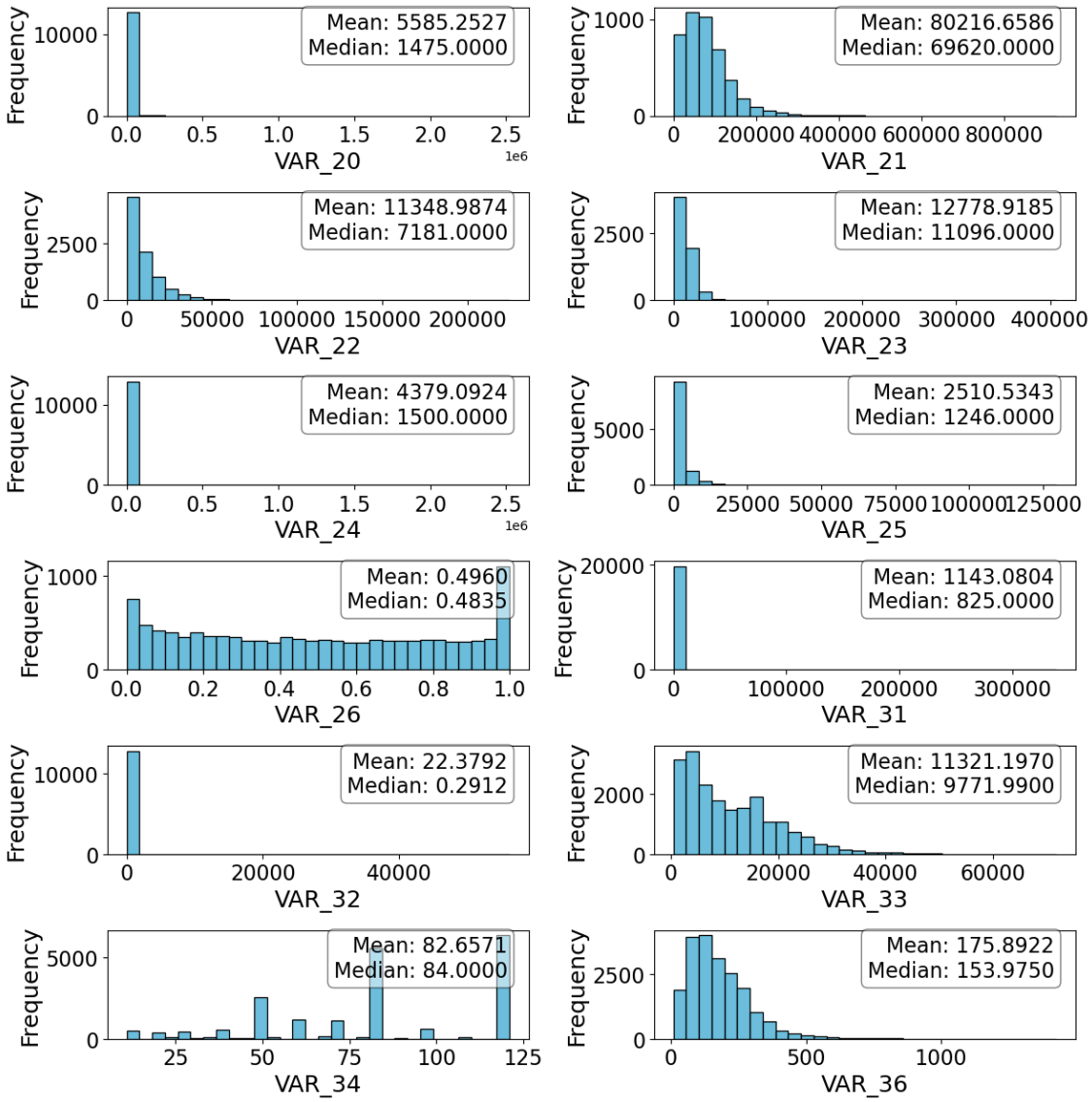


FIGURE 4: Histograms of Continuous Variables.

In Figure 4, the histograms reveal that most continuous variables are heavily right-skewed, with a significant concentration of observations near zero and a long tail extending towards high values. This skewness may present challenges for models that assume normally distributed features. Notably, variables such as *VAR_14*, *VAR_18*, *VAR_20*, *VAR_24*, *VAR_31*, and *VAR_32* exhibit exceptionally high maximum values relative to the bulk of their distributions, suggesting the presence of pronounced outliers that could overshadow the majority of smaller observations. This can impede interpretability and comparability across variables. Nonetheless, recognizing outlier presence is essential, which serves as motivation for a subsequent outlier-focused plot. By contrast, *VAR_2* displays a much narrower range—capped at around 70—implying that it may measure a different scale or type of characteristic compared to the more expansive variables.

To further understand the relationship between continuous variables and the target, the mean values for each continuous variable were computed across the target categories. This analysis provides a critical reference for understanding the central tendency within each target group and aids in identifying variables with strong discriminative power. The following table displays the average value of each feature across the target classes, with the values -9999 and -999 excluded to ensure they do not distort the resulting means.

Target	VAR_2	VAR_7	VAR_12	VAR_14	VAR_15
0	40.2568	1257.3868	142.5877	7.6597	1444.2835
1	45.5284	1222.0048	249.2331	7.5158	1537.0169
Target	VAR_17	VAR_18	VAR_19	VAR_20	VAR_21
0	3.1418	433.5716	45088.3422	5468.0662	86517.1997
1	3.0928	414.0395	32369.4990	5841.4578	63190.3229
Target	VAR_22	VAR_23	VAR_24	VAR_25	VAR_26
0	11573.5703	12880.3752	4248.4932	2512.6507	0.4900
1	10900.3567	12554.6511	4663.9951	2506.0062	0.5088
Target	VAR_31	VAR_32	VAR_33	VAR_34	VAR_36
0	1171.0334	12.0322	11882.6578	83.4896	183.1183
1	1081.1525	44.5609	10079.9270	80.8165	159.9220

TABLE II: Mean values of Continuous Variables across Target Classes.

To properly evaluate the impact of the results presented in Table II, it is essential to consider the histograms generated earlier. Several continuous features exhibit distinct differences in their average values across the target classes. To assess the significance of these differences, both the mean and median of each feature were analysed, along with an examination of their overall distributions. The analysis indicates that variables such as *VAR_2*, *VAR_12*, *VAR_24*, and *VAR_32* tend to have higher mean values for the positive target class, suggesting a direct correlation with credit insurance subscription. Conversely, variables like *VAR_19*, *VAR_21*, *VAR_33*, and *VAR_36* exhibit lower mean values for the positive class, which may also provide meaningful insights for predictive modelling.

To further explore potential patterns between continuous variables and the target, these features were discretised into ten equally populated bins (deciles). This theoretical approach transforms complex, continuous data into more interpretable categorical segments, facilitating easier comparison across target classes. By dividing the data into deciles, it becomes possible to detect non-linear relationships and identify critical thresholds that may be obscured on a continuous scale. Moreover, by examining the average target value within each bin, this method mitigates the influence of extreme values and provides a

clearer assessment of each variable's predictive power.

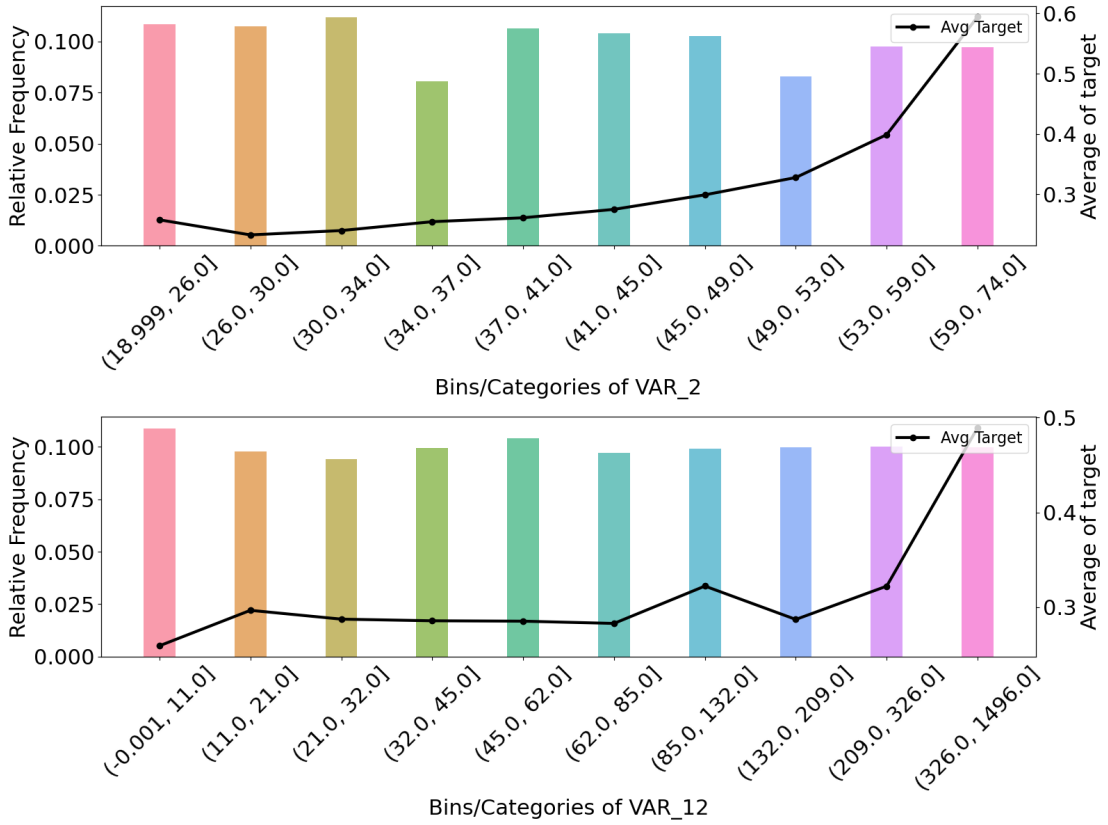


FIGURE 5: Relative Frequency of Continuous Variables with Average Target.

Figure 5 plots the relative frequency of the binned continuous variables with the average target. To avoid unnecessary plots, the ones represented are the only ones that showed relevant findings. The rest can be found in Appendix A. As it can be seen by the plots, the average target value has an increasing exponential behaviour in both cases. This indicates that higher values of *VAR_2* and *VAR_12* are more related to the positive target class.

The histogram analysis revealed that several continuous variables exhibit high skewness, suggesting the presence of significant outliers. It is imperative to identify these extreme values and carefully consider whether to replace them or retain them in the analysis, especially given the goal of preserving the original data as much as possible. To further investigate the extent and impact of outliers, the following figure presents a boxplot of the continuous variables after they have been standardized. This transformation, achieved via scaling to a mean of 0 and a standard deviation of 1, ensures that variables with different scales are directly comparable. The boxplot, utilising the Tukey method to determine the whiskers, provides clearer insights into each variable's dispersion and distributional characteristics, thereby facilitating a more accurate assessment of outlier effects.

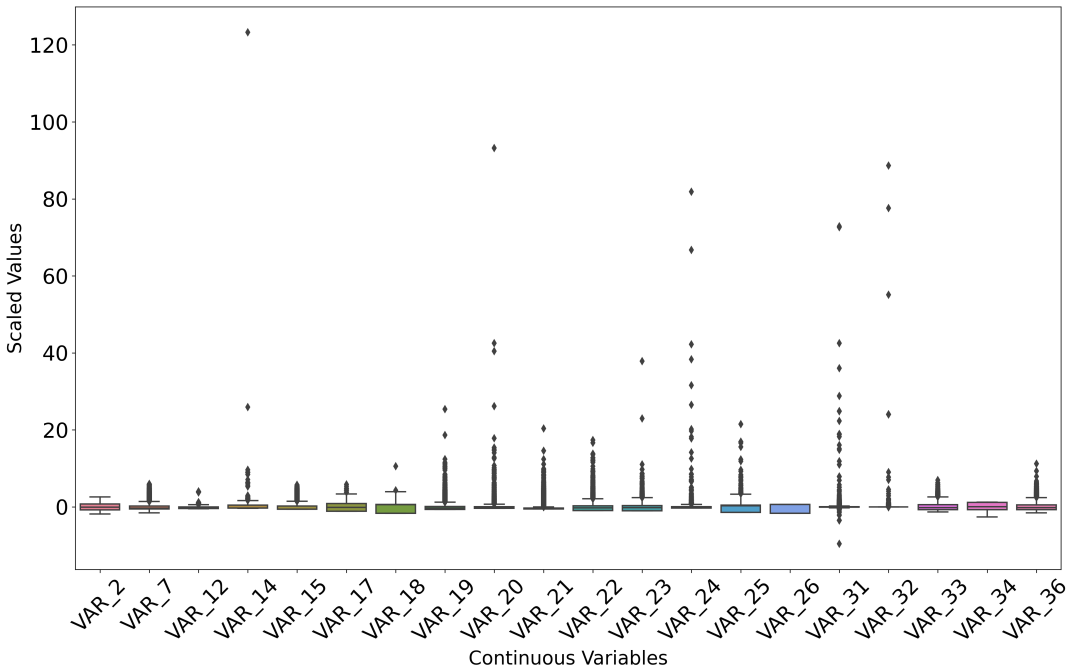


FIGURE 6: Boxplot of Continuous Variables.

Figure 6 demonstrates that many of the continuous variables contain extreme values, as evidenced by points lying well beyond the upper whiskers. Notably, variables such as *VAR_14*, *VAR_20*, *VAR_24*, *VAR_31* and *VAR_32* exhibit particularly large outliers, while other features remain more tightly clustered. This pattern reflects considerable variability within the data, with some features being especially prone to extreme values. Although such outliers have the potential to skew parameter estimates and adversely affect model performance, their relative scarcity—as well as the possibility that they contain valuable information about rare but significant events—justifies their retention in the dataset. Preserving these extreme observations ensures that the model is exposed to the full range of data variability, which is crucial for capturing underlying patterns that may be predictive.

Another critical aspect of the analysis involves examining the correlations among continuous variables. The correlation matrix, visualised through a heatmap, provides insights into the degree of multicollinearity present in the dataset. High correlations between predictors can lead to challenges for certain models, such as logistic regression, where multicollinearity may compromise interpretability and inflate the variance of coefficient estimates. However, as mentioned, in the context of this study, the objective is to retain as much of the original information as possible. Therefore, instead of preemptively discarding highly correlated features, the analysis incorporates these relationships which can be fruitful in future modelling efforts.

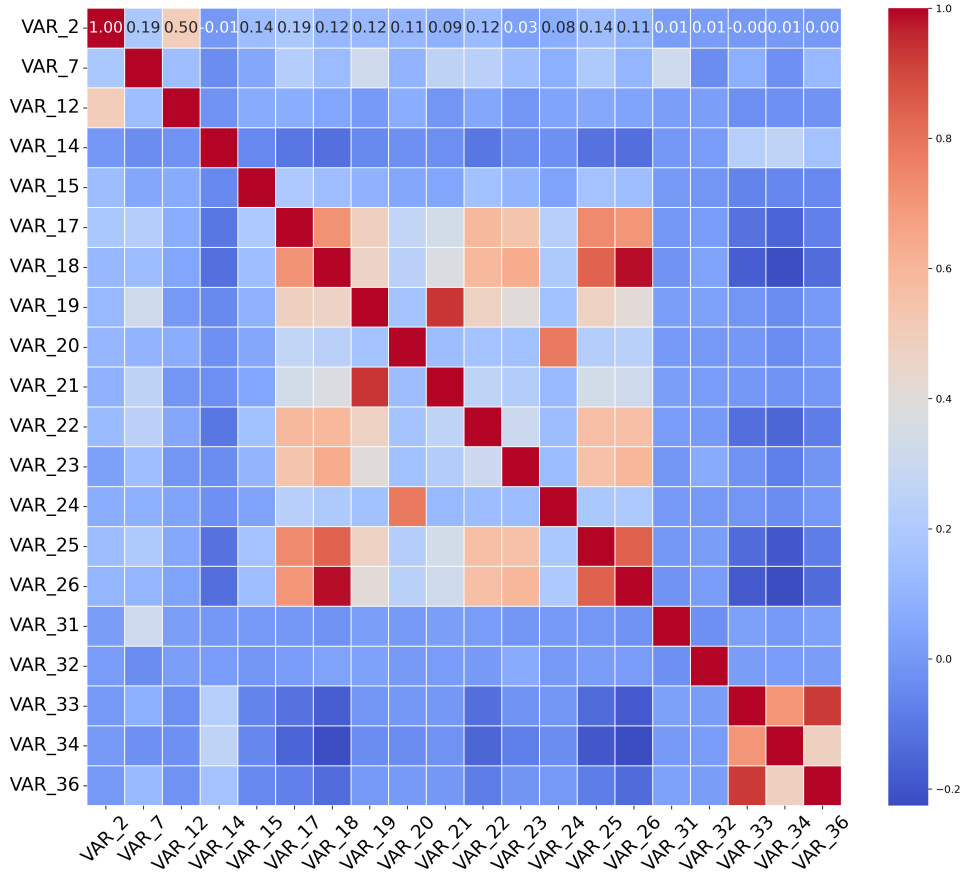


FIGURE 7: Correlation Heatmap for Continuous Variables.

In Figure 7, the correlation analysis reinforces an earlier observation: *VAR_16* directly influences the values of variables *VAR_17* through *VAR_30*. Although the heatmap displays correlations only among continuous variables, it is evident that those within this range are moderately correlated with at least one other feature. Notably, *VAR_33*, *VAR_34*, and *VAR_36* exhibit moderate to high correlations, highlighting a significant interrelationship among these predictors.

3.3 Modelling Overview

This section delineates the selection, fine-tuning, and development of predictive algorithms for credit insurance subscription prediction, drawing on both classical and contemporary methodologies. Traditional techniques, such as Logistic Regression, have long been employed due to their interpretability and ease of implementation (Hand and Henley, 1997; Thomas, 2000). However, as demonstrated by Khandani et al. (2010), the integration of advanced machine-learning algorithms—capable of capturing nonlinear relationships and subtle interactions—can substantially enhance predictive accuracy in large-scale consumer credit data.

A diverse suite of modelling approaches is incorporated, ranging from Logistic Regression and ensemble methods, such as Random Forest (Baesens et al., 2003; Lessmann et al., 2015), to state-of-the-art gradient boosting frameworks like LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018). These models were selected for their proven ability to handle high-dimensional and imbalanced datasets, as well as for their capacity to provide insights into the underlying determinants of credit insurance subscription. Robust regularization and hyperparameter tuning procedures, as highlighted by Friedman et al. (2010) and Hastie et al. (2009), ensure that the models generalise well to unseen data. Collectively, these techniques form a comprehensive modelling framework that leverages both traditional statistical foundations and modern machine-learning advancements to address the complex problem of predicting credit insurance subscription.

3.3.1 Logistic Regression

Logistic Regression is a widely used method for binary classification and has been a staple in credit scoring literature (Hand and Henley, 1997; Thomas, 2000). In the present study, Logistic Regression is employed to model the probability that a client subscribes to credit insurance as a function of a set of predictor variables. Let Y denote the binary target variable, with $Y = 1$ indicating that the client subscribes to credit insurance, and let $X = (x_1, x_2, ..., x_k)$ represent the predictor variables. The model is then defined by the equation 2

$$P(Y = 1 \mid \mathbf{X}) = \frac{1}{1 + \exp(-z)}, \tag{2}$$

with $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Taking the logit transformation yields a linear relationship

$$\text{logit}\left[P(Y = 1 \mid \mathbf{X})\right] = \ln\left(\frac{P(Y = 1 \mid \mathbf{X})}{1 - P(Y = 1 \mid \mathbf{X})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \tag{3}$$

Each coefficient β_i represents the change in the log-odds of a client subscribing to credit insurance for a one-unit increase in the predictor x_i. These coefficients will be interpreted later in the analysis to elucidate the impact of each predictor on the likelihood of a client subscribing to credit insurance. Note that maximum likelihood estimation is used to determine the parameter values that maximize the likelihood of observing the given data.

For classification purposes, the conventional decision rule is applied

$$\hat{Y} = \begin{cases} 1, & \text{if } P(Y = 1 \mid \mathbf{X}) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The default threshold of 0.5 was adopted as it represents the natural midpoint in probability space, ensuring a balanced trade-off between sensitivity and specificity.

Logistic Regression is renowned for its interpretability and ease of implementation. Its straightforward linear relationship between predictor variables and the log-odds of an outcome facilitates both statistical inference and transparent communication of credit decisions. This interpretability is particularly valuable in regulated financial environments, where the ability to explain decisions to stakeholders is imperative.

However, several studies have also highlighted the inherent limitations of Logistic Regression. For instance, Khandani et al. (2010) observe that while Logistic Regression provides a solid baseline, it is less adept at capturing nonlinear relationships and complex interactions among predictors. This limitation can result in suboptimal performance when the underlying data structure exhibits curvature or other nonlinearity that a linear model cannot adequately model. Furthermore, Logistic Regression is sensitive to multicollinearity; when predictors are highly correlated, the variance of coefficient estimates increases, making it challenging to discern the individual impact of each predictor on the target outcome (Thomas, 2000). Such issues are particularly pertinent in financial datasets, where correlated features are common.

While Logistic Regression remains a valuable tool due to its clarity and efficiency, its limitations in capturing nonlinear relationships and handling correlated predictors underscore the necessity for more sophisticated methods in certain contexts. Moreover, advanced models possess the ability to capture intrinsic relationships between features, effectively incorporating interaction terms within their algorithmic framework. This capability is particularly crucial in a study where the goal is to utilise the data with minimal manipulation. Nevertheless, owing to its simplicity and reliable interpretability, Logistic Regression is employed as a benchmark model, serving as a baseline against which the performance of more advanced machine-learning techniques can be compared.

3.3.2 Random Forest

Random Forest is an ensemble learning method designed to improve predictive performance by reducing the high variance typically associated with single decision trees. This modelling technique represents a significant advancement over linear models such as Logistic Regression, particularly in settings where relationships between predictors and the target variable are complex and nonlinear. As noted by Lessmann et al. (2015), Random Forest often yields superior predictive performance in credit scoring applications. As described by Hastie et al. (2009), the method builds on the idea of bagging—bootstrap aggregation—where multiple decision trees are trained on different bootstrap samples of

the training data. In the Random Forest algorithm, each tree is grown by recursively splitting the data at each node. However, unlike traditional bagging, Random Forest introduces additional randomness by selecting a random subset of predictors at each split. This strategy reduces the correlation between individual trees, which is crucial because the variance of an average of B identically distributed random variables with pairwise correlation ρ is given by

$$\text{Var}(\bar{T}) = \rho\,\sigma^2 + \frac{1-\rho}{B}\,\sigma^2, \tag{5}$$

where σ^2 is the variance of a single tree's prediction. As B increases, the second term diminishes, yet the first term—proportional to ρ—remains; thus, reducing ρ through random feature selection is essential to maximize the variance reduction benefit.

For classification tasks, such as predicting whether a client subscribes to credit insurance (with $Y = 1$ indicating subscription), each tree in the forest casts a vote for the predicted class. The final prediction is then determined by majority vote across all trees. This ensemble approach not only captures complex, nonlinear relationships and interactions among predictors—capabilities that linear models like Logistic Regression lack—but also tends to outperform simpler models, especially in high-dimensional, heterogeneous datasets common in credit risk and credit insurance contexts.

By leveraging these de-correlation techniques, Random Forest achieves a favorable balance between bias and variance. The relative ease of training and tuning, coupled with robust performance across a variety of problems, makes it an attractive option for modelling credit insurance subscription. This method offers a more flexible alternative to traditional linear models, providing improved predictive accuracy without compromising on interpretability when combined with modern interpretability tools.

3.3.3 LightGBM

Gradient boosting is an overall robust machine-learning approach that delivers state-of-the-art performance across a range of applications. It has long been favored for tackling problems characterized by diverse features, noisy data, and complex interactions, as it builds an ensemble predictor through gradient descent in a functional space.

LightGBM represents a substantial advancement regarding gradient boosting technology, specifically designed to address the challenges associated with large-scale, high-dimensional datasets. As described by Ke et al. (2017), LightGBM employs innovative techniques such as Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) to significantly reduce computational complexity while maintaining high predictive accuracy. GOSS selectively retains instances with large gradients—those

that contribute most to the learning process—thus ensuring that the algorithm focuses on the most informative data points. Meanwhile, EFB reduces the number of features by bundling mutually exclusive features together, thereby mitigating the issue of dimensionality without sacrificing crucial information.

In contrast to Random Forest—which builds numerous de-correlated trees and aggregates their predictions via majority voting—LightGBM employs a sequential boosting approach that iteratively constructs trees to correct the errors of prior iterations. This methodology not only is capable of capturing subtle nonlinear relationships and complex interactions among predictors more effectively but also leads to faster training times and lower memory consumption, making LightGBM highly scalable in environments where computational efficiency is essential. In the context of credit-related modelling and predicting credit insurance subscription, these innovations are particularly valuable. Financial datasets in this domain are typically large and complex and often suffer from imbalanced class distributions. LightGBM's ability to be robust to class imbalance, coupled with its competitive performance achieved with relatively little tuning, renders it a highly suitable choice for modelling in this present study.

3.3.4 CatBoost

CatBoost has been proved to be a cutting-edge gradient boosting algorithm that builds upon the strengths of traditional boosting while addressing some of its key limitations. According to Prokhorenkova et al. (2018), two critical innovations in CatBoost are its implementation of ordered boosting—a permutation-driven alternative designed to combat prediction shift caused by target leakage—and its specialized algorithm for processing categorical features. These advancements allow CatBoost to construct an ensemble predictor using binary, oblivious decision trees that are inherently balanced and less prone to overfitting, which in turn speeds up execution during testing.

It is important to note that in the context of this study, maximum comparability across models is achieved by pre-encoding all categorical variables. Therefore, while CatBoost's native handling of categorical features represents a significant advantage in many applications, this benefit is not leveraged here. Instead, CatBoost is employed on an equal footing with LightGBM and Random Forest, allowing for a fair assessment of predictive accuracy and computational efficiency in modelling credit insurance subscription behaviour.

3.3.5 Tuning Methods

Before applying the models, interaction terms composed of pairs of continuous and binary/categorical variables were tested. Although this process was not central to the pri-

mary research objective—namely, preserving the integrity of the original dataset—it provided valuable insights. A LightGBM model, combined with SHAP values (discussed in detail later), was used to extract the most relevant interaction terms. To assess the impact of these terms, correlated individual variables were removed before the augmented dataset was constructed. A comparison of the evaluation metrics for models trained on the original dataset versus those trained on the dataset including interactions showed that the original dataset yielded superior performance for all models except CatBoost, which exhibited only minimal improvements in AUC and Gini scores. Consequently, the interaction features were not included in the final modelling dataset.

In terms of changes to the original data, aside from the initial imputation or removal of missing values, two key modifications were applied. The first modification involved one-hot encoding the categorical variables. As noted earlier, not all models can effectively handle categorical variables in their native form; one-hot encoding ensures that all models treat these features consistently, thereby reinforcing comparability across different modelling approaches. The second modification was the scaling of data for Logistic Regression, which proves to be essential because it ensures that all predictor variables contribute equally to the model. When features are measured on different scales, those with larger numerical ranges can disproportionately influence parameter estimates, potentially skewing the model's performance. Standardising features to have zero mean and unit variance not only promotes numerical stability but also facilitates faster convergence of optimisation algorithms, such as gradient descent.

One tuning mechanism associated with scaling in the Logistic Regression model is the use of shrinkage terms. As explained by Hastie et al. (2009), while subset selection yields interpretable models by retaining only a limited number of predictors, its discrete nature can result in high variance. In contrast, shrinkage methods apply continuous penalties that reduce variability and enhance model stability. In high-dimensional settings where predictors are often highly correlated, the elastic net offers additional benefits by combining the L1 (lasso) and L2 (ridge) penalties. Friedman et al. (2010) note that the lasso promotes sparsity by driving some coefficients to zero, whereas ridge regression shrinks coefficients of correlated predictors towards one another, allowing them to "borrow strength." The elastic net leverages these complementary properties to enable simultaneous variable selection and coefficient shrinkage, thereby mitigating overfitting and enhancing model stability. Since the effectiveness of these penalty terms depends on the features being on a comparable scale, proper scaling is critical to impose dual regularization effectively, resulting in a model that is both robust and interpretable.

Due to the imbalanced nature of the dataset, all models incorporated a class weighting

parameter to adjust for unequal distributions in the target variable. In addition, stratified k-fold cross-validation—employing both 5-fold and 10-fold splits with shuffling enabled—was utilised to ensure that each fold accurately reflected the overall class distribution, thereby avoiding any systematic ordering effects. This rigorous cross-validation framework not only provided robust and realistic performance estimates but also played a pivotal role in hyperparameter tuning, ensuring that the model configurations generalised effectively to unseen data.

Hyperparameter tuning is essential for optimising model performance, as it involves systematically adjusting parameters to achieve the best trade-off between bias and variance. This process is particularly critical in imbalanced classification tasks, where models must be finely tuned to enhance overall predictive accuracy and discriminatory power. In this study, the (AUC)—detailed later—was employed as the primary metric during hyperparameter tuning, emphasising its importance in effectively ranking clients by their likelihood of subscribing to credit insurance. (AUC) is especially appropriate for this purpose, as it quantifies a model's ability to distinguish between classes across all thresholds, thereby ensuring that the model reliably prioritises the minority class. This approach is supported by prior literature, which underscores the utility of (AUC) in scenarios with skewed class distributions (Fawcett, 2006).

Furthermore, for computationally intensive models such as LightGBM and CatBoost, early stopping rounds were implemented. This strategy not only prevents excessive computation times but also serves as a safeguard against overfitting, ensuring that the model maintains its ability to generalize to new data.

3.4 Evaluation Metrics and Interpretability

Evaluation metrics are essential for assessing the performance of predictive models, particularly in imbalanced classification tasks typical in credit-related modelling. In practical applications, financial institutions require models to achieve a minimum level of performance to ensure that associated risks are effectively managed. A fundamental component in this evaluation procedure is the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various classification thresholds. The Area Under the ROC Curve (AUC) is derived directly from the ROC curve and represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance (Fawcett, 2006). In financial risk contexts, AUC provides an aggregate measure of the model's discriminatory power, independent of any particular threshold, while the Gini coefficient, calculated as

$$\text{Gini} = 2 \times \text{AUC} - 1, \tag{6}$$

quantifies the inequality in the distribution of predicted probabilities (Hand, 2009).

In addition to these global performance measures, the F1 measure—particularly the F1 score for the positive class—remains a vital evaluation metric in this study. Although AUC is used as the primary metric for hyperparameter tuning and model selection, the F1-positive score is crucial for gauging the balance between precision and recall, thereby capturing the model's ability to accurately detect the minority class, namely, clients who subscribe to credit insurance. This metric provides additional insight into the model's performance in imbalanced settings, ensuring that improvements in sensitivity and precision for the positive class are adequately monitored (Powers, 2011).

Together, these metrics—AUC derived from the ROC curve, the Gini coefficient, and the F1-positive score—form a comprehensive evaluation framework that not only quantifies overall discriminatory ability but also specifically addresses performance on the critical minority class. In addition to these performance measures, model interpretability remains a crucial factor in financial applications. Financial institutions are often hesitant to adopt advanced models that operate as "black boxes," due to the difficulty in elucidating their internal mechanisms and translating their outputs into actionable, real-world insights. Such interpretability is essential to satisfy regulatory requirements and build trust among stakeholders, emphasizing the need for evaluation frameworks that balance predictive power with transparency.

In Logistic Regression, the model coefficients (β) provide a direct and interpretable measure of the impact that each predictor has on the log-odds of the target outcome. As explained previously, a coefficient (β_i) indicates the change in the log-odds of a client subscribing to credit insurance for a one-unit increase in the corresponding predictor (x_i). This direct interpretability facilitates clear communication of model insights, allowing stakeholders to understand which factors most significantly influence credit risk decisions.

In contrast, more complex models such as Random Forest, LightGBM, and CatBoost, while often yielding higher predictive accuracy, do not offer straightforward coefficient-based interpretations. To address this, SHAP (SHapley Additive exPlanations) values are employed to explain model predictions. TreeExplainer, as introduced by Lundberg et al. (2020), is a specialized method for tree-based models that computes SHAP values efficiently in polynomial time. This approach decomposes each prediction into additive contributions from individual features, ensuring that the sum of these contributions equals the difference between the model's output and its expected output. By doing so, SHAP values provide both local explanations for individual predictions and a global perspective

on feature importance. Thus, it becomes possible to translate the complex, nonlinear interactions captured by advanced machine-learning models into interpretable and human-intuitive insights, maintaining the required transparency in this area.

4 RESULTS

In this chapter, the outcomes of the modelling implementation—using Logistic Regression, Random Forest, LightGBM, and CatBoost—are presented and analysed. The optimal hyperparameters for each model are detailed, and evaluation metrics are compared to identify the best-performing approach. The chosen model is subsequently interpreted to elucidate its predictive behaviour.

4.1 Best Models

In this subsection, the optimal hyperparameters for each model—Logistic Regression, Random Forest, LightGBM, and CatBoost—are presented. These configurations were determined through a rigorous hyperparameter tuning process using stratified cross-validation, with the AUC as the primary evaluation metric. The selected settings were chosen for their superior ability to discriminate between classes, thereby enhancing overall predictive stability. The best parameters for each model are detailed below, followed by a comprehensive discussion of their impact on model performance.

Hyperparameter	**Description**	**Value**
penalty	Specifies the norm used in the penalization. The 'elasticnet' option combines L1 and L2 regularization.	elasticnet
C	Inverse of regularization strength.	0.5
solver	Algorithm used for optimization. The 'saga' solver supports elastic net regularization.	saga
l1_ratio	The mixing parameter for elastic net, balancing L1 and L2 regularization.	0.8
class_weight	Adjusts weights inversely proportional to class frequencies to address class imbalance.	balanced

TABLE III: Selected Hyperparameters for Logistic Regression.

As shown in Table III and the subsequent tables, all models were configured with a balanced class weight parameter to address the inherent class imbalance in the dataset. For Logistic Regression specifically, the elastic net penalty and the *saga* solver were manually selected to effectively manage the skewed class distribution and correlated features typical of high-dimensional data.

For Logistic Regression, a low C value (0.05) enforces strong regularization, promoting a parsimonious model that mitigates overfitting in high-dimensional settings. An $l1\_ratio$ of 0.8 indicates a heavy emphasis on L1 regularization, encouraging sparsity in the model.

Hyperparameter	Description	Value
n_estimators	Number of trees in the forest.	500
max_depth	Maximum depth of each tree.	15
min_samples_split	Minimum number of samples required to split an internal node.	7
min_samples_leaf	Minimum number of samples required to be at a leaf node.	2
class_weight	Adjusts weights inversely proportional to class frequencies.	balanced

TABLE IV: Selected Hyperparameters for Random Forest.

For Random Forest, Table IV indicates a configuration aimed at balancing complexity and generalization. The use of 500 trees provides sufficient ensemble diversity to reduce variance and capture robust patterns, while a maximum depth of 15 allows the model to capture complex interactions without overfitting. Moreover, setting a minimum of 7 samples for splitting and 2 samples per leaf helps to avoid overly granular splits.

Hyperparameter	Description	Value
n_estimators	Number of boosting rounds (trees).	700
learning_rate	Step size at each iteration.	0.01
max_depth	Maximum tree depth.	7
num_leaves	Maximum number of leaves per tree.	20
class_weight	Adjusts weights inversely proportional to class frequencies to handle imbalance.	balanced

TABLE V: Selected Hyperparameters for LightGBM.

For LightGBM, the results in Table V indicate that the model builds a sufficiently large ensemble to capture complex patterns in high-dimensional data while maintaining stability and generalization. The low learning rate ensures gradual convergence during the boosting process, reducing the risk of overshooting the optimal solution. Furthermore, by limiting the maximum depth and the number of leaves, the model effectively controls overfitting, striking a balance between complexity and robustness.

For CatBoost, the values in Table VI reflect a configuration designed to capture complex, nonlinear relationships while maintaining stability and avoiding overfitting. A moderate depth ensures that the model can learn intricate interactions among features without

Hyperparameter	Description	Value
depth	Maximum depth of the trees.	8
iterations	Number of boosting rounds.	1000
learning_rate	Step size for updating the model.	0.01
auto_class_weights	Automatically sets class weights to handle imbalanced datasets.	Balanced

TABLE VI: Selected Hyperparameters for CatBoost.

becoming overly complex, and the iterations allow the boosting process sufficient capacity to refine its predictions gradually. The low learning rate further promotes gradual convergence, contributing to robust overall performance.

4.2 Model Evaluation

The models are evaluated and compared based on their cross-validated AUC scores as well as their performance on the test set, measured by the F1 score, AUC, and Gini coefficient. Additionally, a complementary decile analysis is conducted to assess each model's ability to correctly rank clients by likelihood of subscribing to credit insurance.

The table below presents the evaluation metrics for each of the models described in the previous subsection.

Model	CV AUC Score	Test F1 Score	Test AUC	Test Gini
Logistic Regression	0.672	0.509	0.679	0.358
Random Forest	0.688	0.472	0.708	0.416
LightGBM	0.687	0.517	0.698	0.397
CatBoost	0.693	0.515	0.704	0.408

TABLE VII: Comparison of performance metrics for the four models.

In terms of cross-validated AUC, CatBoost achieved the highest score—slightly exceeding those of Random Forest and LightGBM—while Logistic Regression lagged behind, indicating that CatBoost exhibits superior and robust discrimination.

For the F1 score, particularly for the positive class, LightGBM recorded the highest value, followed closely by CatBoost and Logistic Regression, with Random Forest performing noticeably lower. This lower F1 score for Random Forest suggests that, despite strong overall discrimination, the model may struggle more than others to accurately identify the minority class—clients who subscribe to credit insurance—resulting in a higher rate of false negatives and/or false positives.

Regarding AUC and Gini value, Random Forest demonstrated the highest values, underscoring its ability to correctly rank clients by their likelihood of subscribing to credit insurance. In contrast, Logistic Regression exhibited the lowest test AUC, indicative of comparatively weaker overall discrimination.

Taken together, the ensemble models do not exhibit significantly different values across most metrics, suggesting that their practical performance is comparable. Still, while Random Forest is the best in overall discrimination and LightGBM shows the most strength in identifying the positive class, the results suggest that CatBoost provides the most balanced performance overall. This equilibrium across cross-validated AUC and test set metrics indicates that CatBoost is the most effective model for predicting and ranking clients by their likelihood of subscribing to credit insurance.

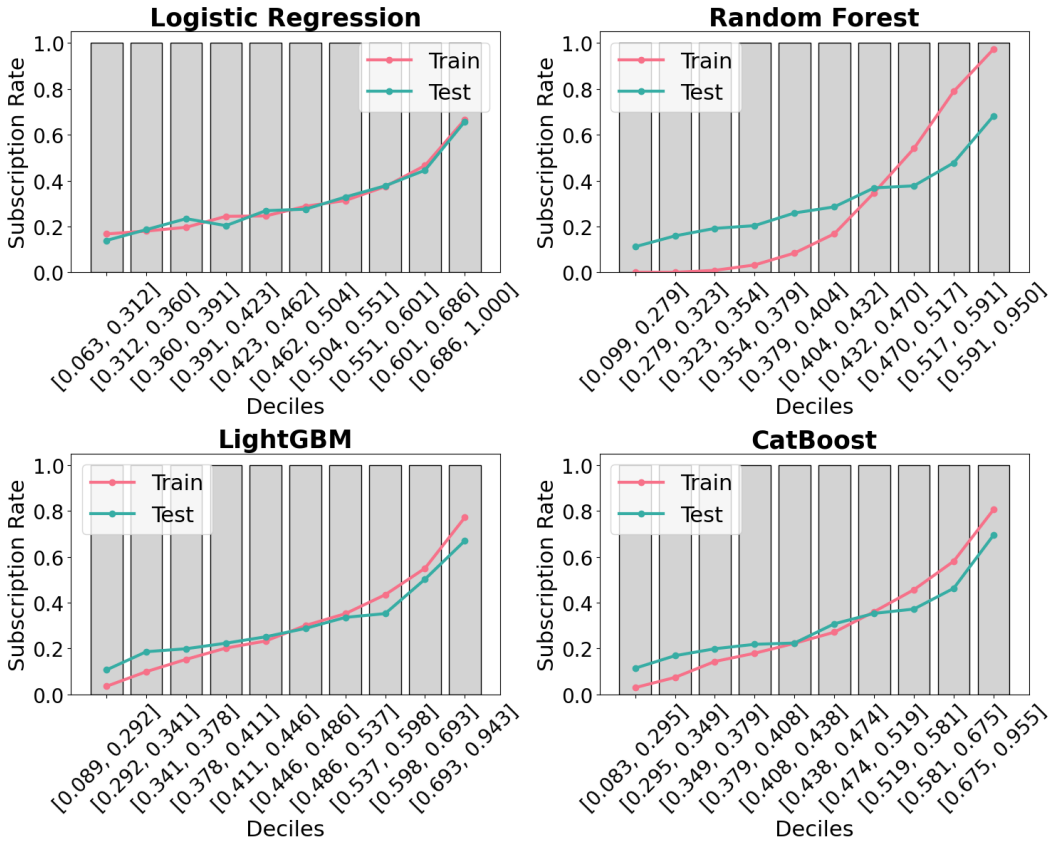


FIGURE 8: Decile Analysis of Subscription Rates.

A review of the decile analysis presented in Figure 8 indicates that all four models exhibit an upward trend in subscription rates as the predicted probability increases, confirming that each method effectively ranks clients to a certain degree. Notably, Logistic Regression achieves the highest subscription rate in the top decile, suggesting that it excels at isolating the most likely subscribers in the highest risk bracket. However, this observation must be interpreted with caution, as Logistic Regression, while interpretable, underper-

forms on global evaluation metrics. In contrast, Random Forest displays strong separation on the training set, yet a pronounced gap between the training and test decile curves suggests overfitting, thereby limiting the generalization in ranking clients. Both LightGBM and CatBoost demonstrate relatively stable train–test performance across deciles, providing consistent stratification of clients, though neither achieves the top-decile subscription rate observed for Logistic Regression.

When considering the evaluation metrics in tandem with the insights provided by the decile analysis, CatBoost emerges as the most balanced model overall. Although Logistic Regression excels in the top decile, its overall predictive performance is the lowest among the models. Random Forest achieves the highest AUC and Gini coefficients, reflecting strong global discriminatory power, yet its instability in decile ranking raises concerns about overfitting. LightGBM performs well, particularly in identifying positive instances, but its results remain slightly below those of CatBoost. Collectively, these findings suggest that while each model exhibits particular strengths, CatBoost offers the most reliable and consistent overall performance, making it the most effective model for predicting and ranking clients by their likelihood of subscribing to credit insurance.

4.3 Model Interpretability

In this subsection, as CatBoost emerged as the best model, its interpretability is examined through the use of SHAP values, which quantify the contribution of each feature to the final prediction.

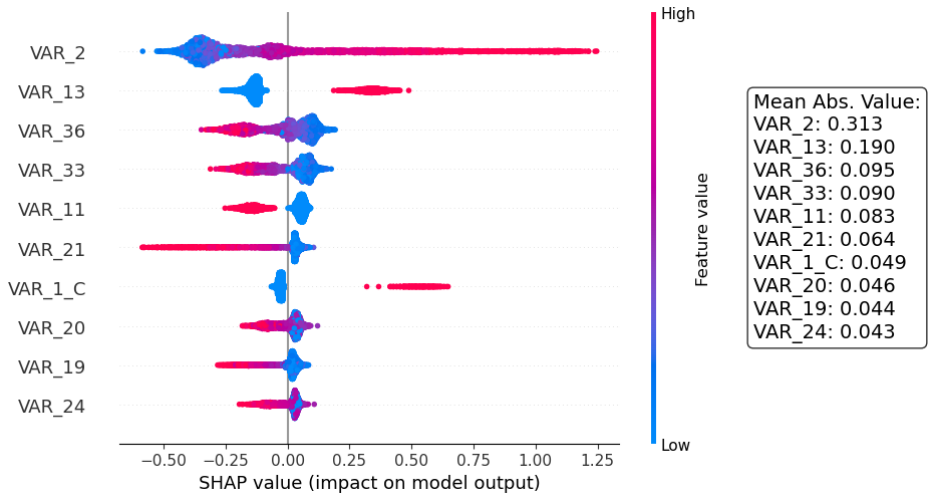


FIGURE 9: Top 10 Features by SHAP Value Contribution for CatBoost.

By examining the SHAP summary plot in Figure 9, it becomes clear that *VAR_2* exerts the largest average influence on the model's predictions, with the widest range of SHAP

values and the highest mean absolute contribution. This indicates that *VAR_2* is particularly pivotal in distinguishing whether an individual will subscribe, as even moderate variations in its value can lead to marked shifts in the predicted outcome. The remaining features, such as *VAR_13*, *VAR_36*, and *VAR_33*, also demonstrate meaningful impact but on a comparatively smaller scale. The color gradient in the summary plot, transitioning from blue to red, further illustrates how higher or lower values of each feature can push the prediction in a positive or negative direction. Notably, features with a broad spread of SHAP values have a more variable effect across different individuals, whereas those with a narrower spread influence fewer observations in a consistently moderate manner.

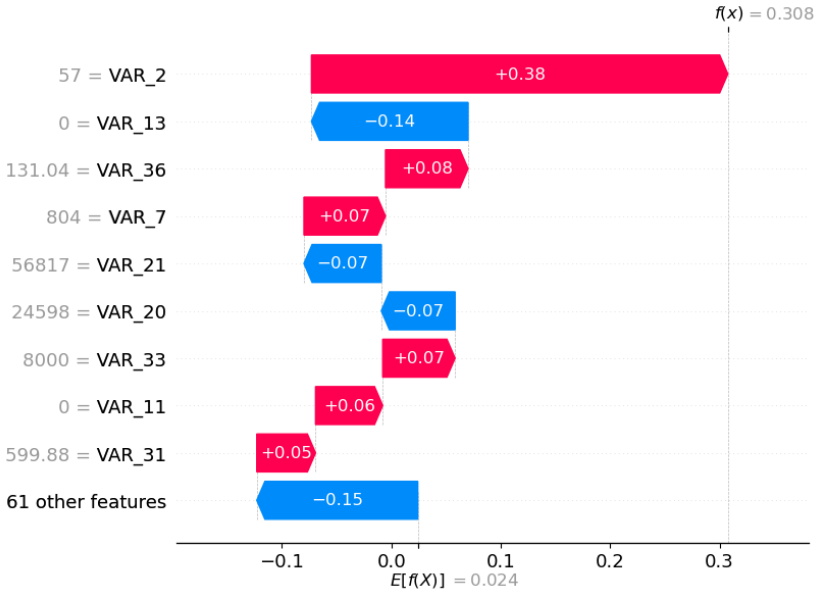


FIGURE 10: Waterfall Chart for Individual Contribution for CatBoost.

Figure 10 consists of the waterfall plot, which focuses on a single random instance, where the baseline prediction (the model's expected value) of 0.024 is incrementally adjusted by each feature's contribution until arriving at the final prediction. Here, *VAR_2* again stands out with a substantial positive effect, raising the predicted value considerably above the baseline. By contrast, other features (e.g., *VAR_13* and *VAR_21*) offset some of that increase through negative contributions, indicating that, for this particular instance, their values lower the likelihood of subscription. Ultimately, these contrasting upward and downward pushes settle on a final prediction around 0.308. Such an individualized breakdown underscores how the interplay of multiple variables can decisively shape the model's conclusion for any given client.

Overall, these SHAP plots demonstrate not only which features hold the most sway in a broad sense—particularly *VAR_2*—but also how each of them can vary substantially in effect from one individual to another. By combining a global perspective on the model's

primary drivers with a granular view of how those features converge on a single prediction, it becomes possible to both prioritize the most impactful variables for further investigation and understand precisely why a given client is deemed more or less likely to subscribe.

5 DISCUSSION

5.1 Key Findings

The primary goal of this study was to identify the best predictive model for ranking clients by their likelihood of subscribing to credit insurance. To achieve this, several modelling approaches were implemented, including Logistic Regression, Random Forest, LightGBM, and CatBoost. Logistic Regression, widely regarded for its simplicity and interpretability, is considered as the baseline model in the industry. However, despite its high degree of transparency, Logistic Regression exhibited lower overall predictive performance compared to advanced ensemble and boosting methods. Interestingly, Logistic Regression achieved the best results in decile analysis, indicating that it excels in ranking clients in the highest risk bracket, even if its global evaluation metrics are inferior.

Random Forest and LightGBM, both of which employ ensemble methods to capture complex nonlinear interactions, demonstrated superior performance on global discrimination metrics and in identifying the positive class, respectively. Random Forest, in particular, achieved excellent overall ranking ability; yet, the lower F1 score for the positive class suggested that it may produce a higher rate of misclassification for the minority class, reducing its effectiveness in accurately identifying clients likely to subscribe. LightGBM, with its low learning rate and controlled tree complexity, achieved a superior F1 score, indicating an enhanced ability to detect positive instances. However, its performance in overall discrimination was slightly lower than that of Random Forest.

CatBoost emerged as the most balanced model overall, attaining competitive values across cross-validated AUC, test AUC, Gini, and F1 scores. The decile analysis further underscored its strength by demonstrating consistent ranking performance across all risk segments. Notably, the SHAP summary plot reveals that certain features—ranked in order of decreasing importance as *VAR_2*, *VAR_13*, *VAR_36*, *VAR_33*, *VAR_11*, *VAR_21*, *VAR_1_C*, *VAR_20*, *VAR_19*, and *VAR_24*—play a decisive role in the model's predictions. For example, *VAR_2*, which exhibits a narrower value range in the histograms, consistently shows the largest average influence, aligning with earlier exploratory analyses where *VAR_2* demonstrated distinct distributional properties compared to other continuous variables. Similarly, features such as *VAR_11* and *VAR_13*, highlighted in the

target prevalence plots for binary variables due to robust observation counts, emerge as significant predictors. *VAR_1_C*, *VAR_33*, *VAR_36* were also initially referenced. In contrast, variables like *VAR_9*, although initially suggestive of distinct behaviour, proved less reliable owing to low observation counts. This congruence between the initial assumptions and the SHAP findings validates the relevance of these features in driving the final ranking. Moreover, SHAP values provide both global summaries and individualized explanations, thereby enhancing understanding of which features are most impactful in the decision-making process. This dual perspective is crucial for refining the model, as it enables the prioritization of the most influential predictors and offers insights into why certain clients are ranked as highly likely to subscribe to credit insurance.

5.2 Future Research

The promising findings of this study open several avenues for research in the future. While this study pre-encoded categorical variables to ensure comparability across models, future research could explore the benefits of native categorical processing in algorithms like CatBoost, potentially yielding performance improvements and deeper insights into feature interactions. In addition, Logistic Regression could benefit from a tailored threshold, while Random Forest—despite exhibiting strong classification on the training set—should be tuned more effectively to mitigate overfitting. Further fine-tuning of hyperparameters through advanced optimisation techniques and refined cross-validation strategies—coupled with an expanded evaluation framework—could provide a more nuanced understanding of model performance under imbalanced conditions. Moreover, exploring model stacking or ensemble methods that combine the robust discriminatory power of Random Forest with the balanced performance of CatBoost may reveal synergistic effects, ultimately enhancing prediction accuracy and reliability. Finally, integrating advanced interpretability techniques alongside SHAP values will further improve transparency and foster greater stakeholder trust in model-based decision-making.

Finally, on a business level, it would be beneficial to integrate an economic feature into the model. Although this study focused on ranking clients based solely on their propensity to subscribe, it would be even more advantageous for the company to also rank clients according to the commission they generate upon subscription. Typically, clients with lower loan amounts pay lower premiums, and consequently, the company earns less commission, even though these clients can be among the most likely to subscribe. Thus, identifying those clients who are both inclined to subscribe and generate substantial commission remains a challenging yet promising area for future research.

6 CONCLUSION

This thesis project set out to identify the best predictive model for ranking clients by their likelihood of subscribing to credit insurance, with the objective of advancing beyond the traditional industry baseline of Logistic Regression. The study implemented and compared multiple models—including Logistic Regression, Random Forest, LightGBM, and CatBoost—using a comprehensive evaluation framework that combined cross-validated AUC, decile analysis, and interpretability measures through SHAP values.

The results confirm that while Logistic Regression remains attractive in this context due to its interpretability, it falls short in capturing the complex, nonlinear interactions present in high-dimensional, imbalanced data. Although Random Forest demonstrated strong global discrimination, as evidenced by its high test set AUC and Gini coefficients, it also exhibited signs of overfitting and lower sensitivity to the minority class, as indicated by its reduced F1 score. LightGBM provided competitive performance, particularly in identifying positive cases; however, CatBoost emerged as the most balanced model overall, delivering consistent performance across evaluation metrics and risk segments.

Moreover, the application of SHAP values not only validated the global discriminatory power of the models but also offered granular insights into feature importance. In particular, features such as *VAR_2*, *VAR_13*, and *VAR_36* were identified as key drivers of predictions, reinforcing initial exploratory findings and enhancing the transparency of the complex models.

Overall, the findings demonstrate that advanced ensemble methods, particularly CatBoost, provide a significant step forward from the conventional Logistic Regression baseline, offering both improved predictive accuracy and enhanced interpretability. Future research should explore further model refinement, including native categorical processing and advanced ensemble strategies such as model stacking, to continue enhancing predictive performance in credit insurance subscription prediction. Additionally, integrating an economic dimension to account for clients' commission potential could further optimise the model's practical relevance.

From a business perspective, these methodologies and results contribute to greater efficiency in client outreach by highlighting the key features that characterise clients likely to opt for credit insurance, thereby enabling the prioritisation of customers. This targeted approach not only facilitates cost optimisation for the insurer but also increases the rate of insurance subscriptions.

REFERENCES

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*, 589–609. https://www.jstor.org/stable/2978933

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*, 627–635. https://doi.org/10.1057/palgrave.jors.2601545

Durkin, T., & Elliehausen, G. (2018). New evidence on an old unanswered question: The decision to purchase credit insurance and other debt protection products. *Journal of Insurance Regulation*, *37*. https://doi.org/10.52227/26007.2018

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*. https://doi.org/10.18637/jss.v033.i01

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*, 523–541. https://doi.org/10.1111/j.1467-985x.1997.00078.x

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the roc curve. *Machine Learning*, *77*, 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning* (2nd ed.). Springer.

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284. https://doi.org/10.1109/tkde.2008.239

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Neural Information Processing Systems*.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*, 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*, 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*, 37–63.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). Catboost: Unbiased boosting with categorical features. *Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1706.09516

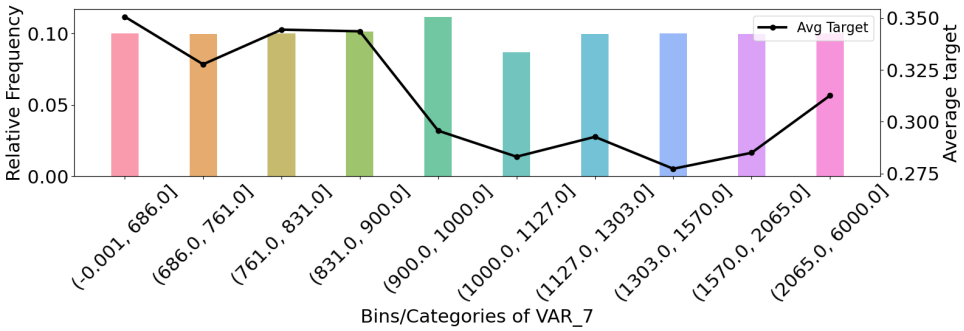
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*, 111–147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

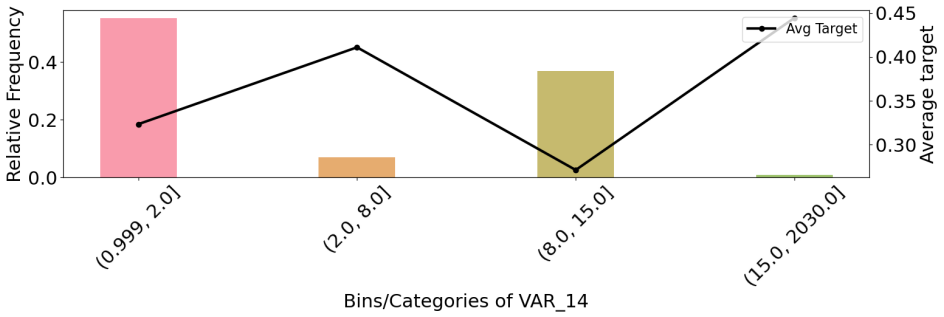
Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*, 149–172. https://doi.org/10.1016/s0169-2070(00)00034-0

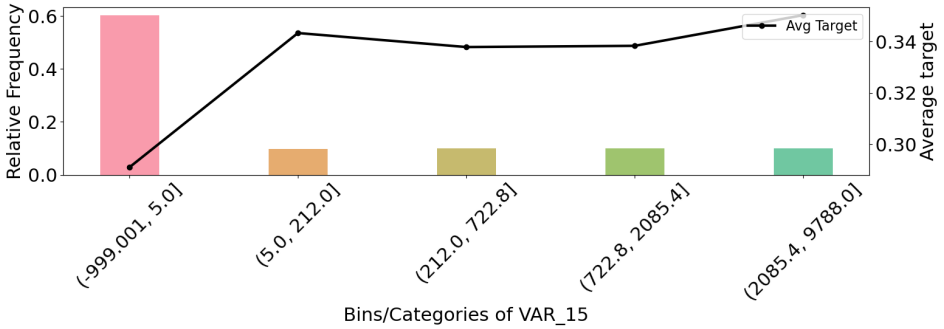
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. Society For Industrial; Applied Mathematics.

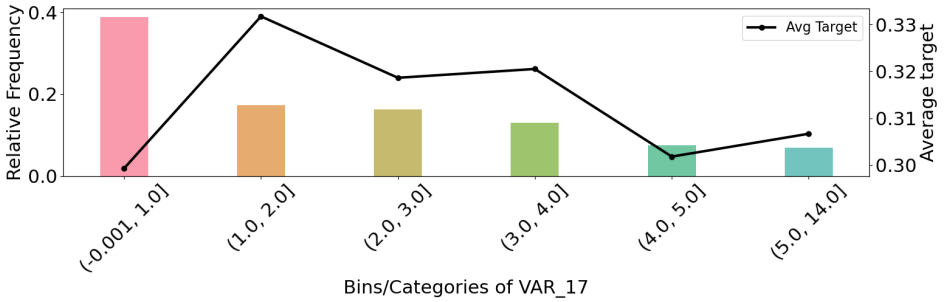
Wiginton, J. C. (1980). A note on the comparison of logit and discriminant models of consumer credit behavior. *The Journal of Financial and Quantitative Analysis*, *15*, 757. https://doi.org/10.2307/2330408

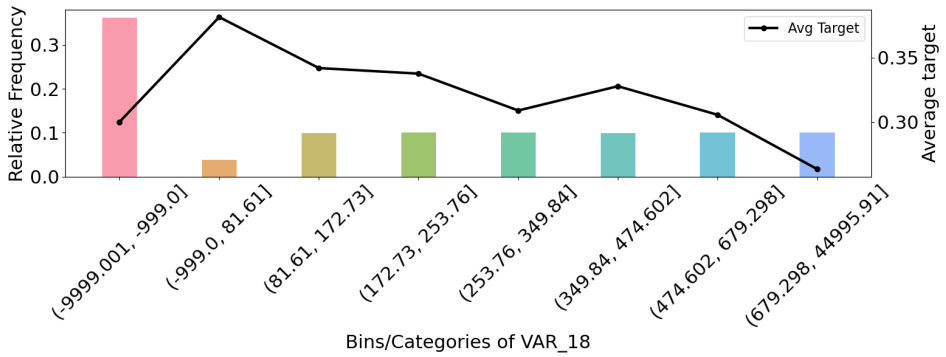
A APPENDICES

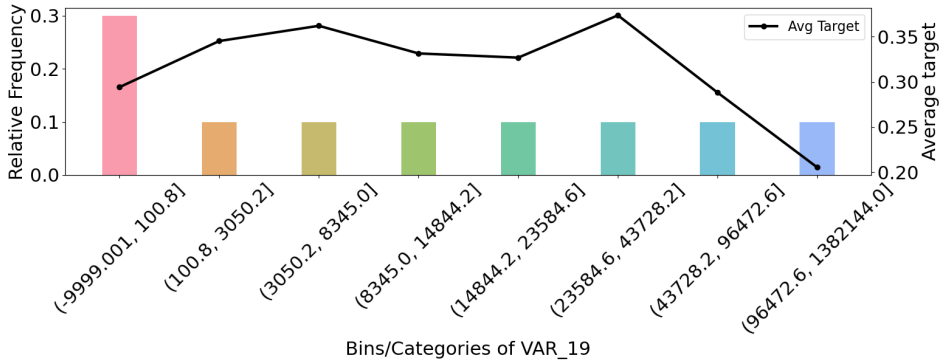


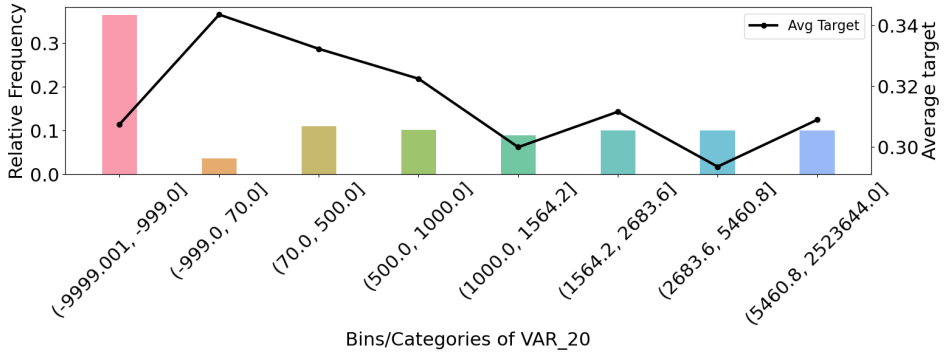


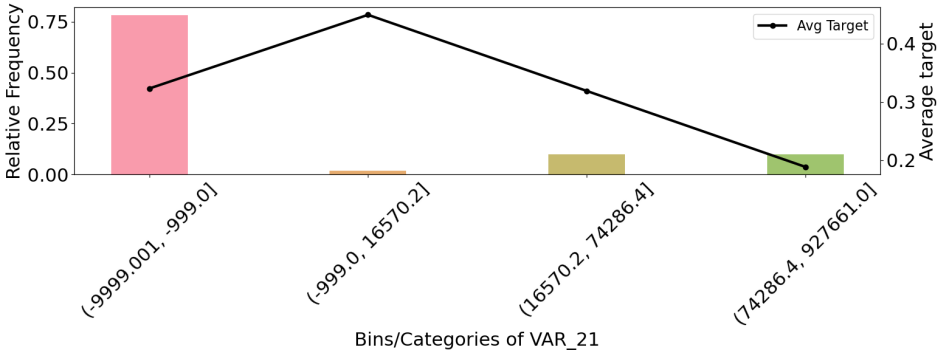


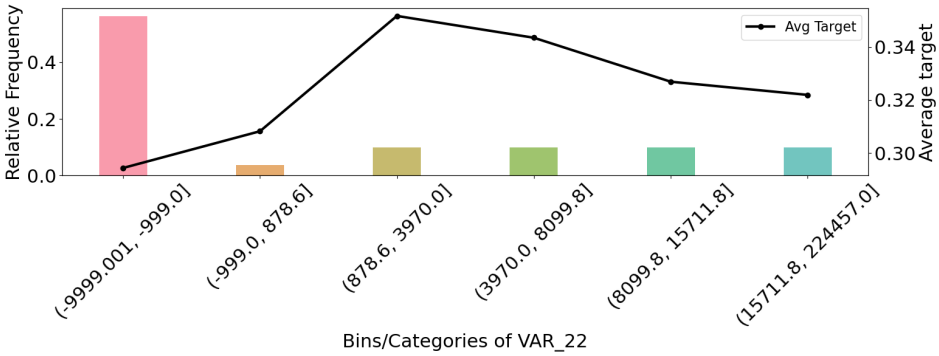


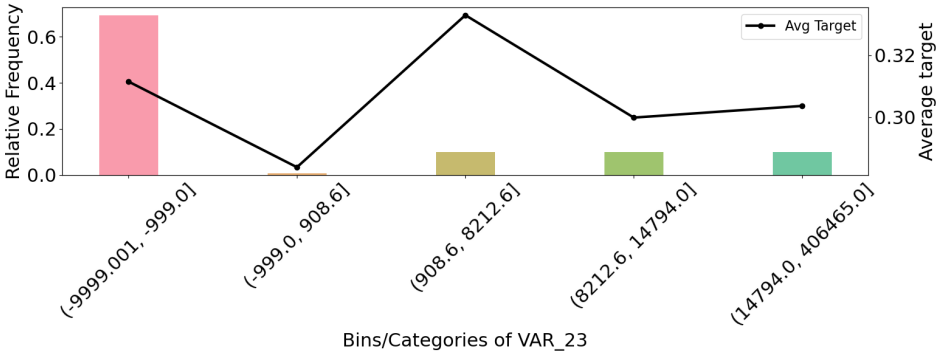

Relative Frequency
Average target
Avg Target
0.0
0.1
0.2
0.3
0.30
0.35
(-9999.001, -999.0]
(-999.0, 81.61]
(81.61, 172.73]
(172.73, 253.76]
(253.76, 349.84]
(349.84, 474.602]
(474.602, 679.298]
(679.298, 44995.91]
Bins/Categories of VAR_18

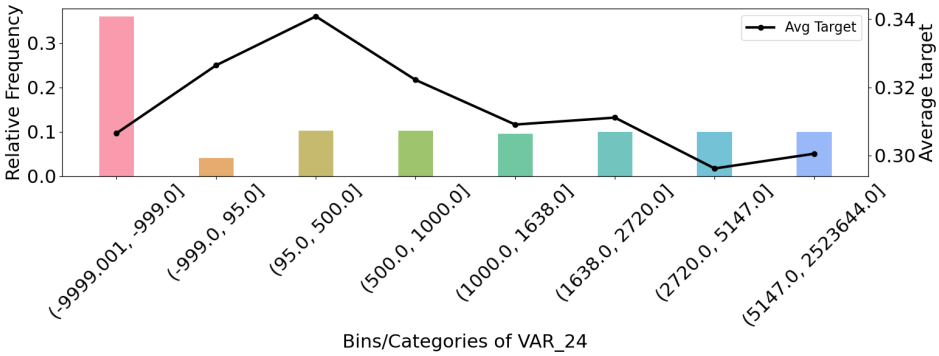

Relative Frequency
Average target
Avg Target
0.0
0.1
0.2
0.3
0.20
0.25
0.30
0.35
(-9999.001, 100.8]
(100.8, 3050.2]
(3050.2, 8345.0]
(8345.0, 14844.2]
(14844.2, 23584.6]
(23584.6, 43728.2]
(43728.2, 96472.6]
(96472.6, 1382144.0]
Bins/Categories of VAR_19

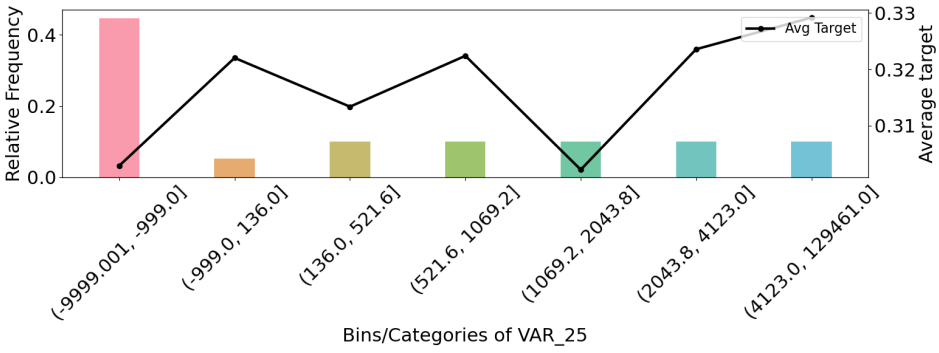

Relative Frequency
Average target
Avg Target
0.0
0.1
0.2
0.3
0.30
0.32
0.34
(-9999.001, -999.0]
(-999.0, 70.0]
(70.0, 500.0]
(500.0, 1000.0]
(1000.0, 1564.2]
(1564.2, 2683.6]
(2683.6, 5460.8]
(5460.8, 2523644.0]
Bins/Categories of VAR_20

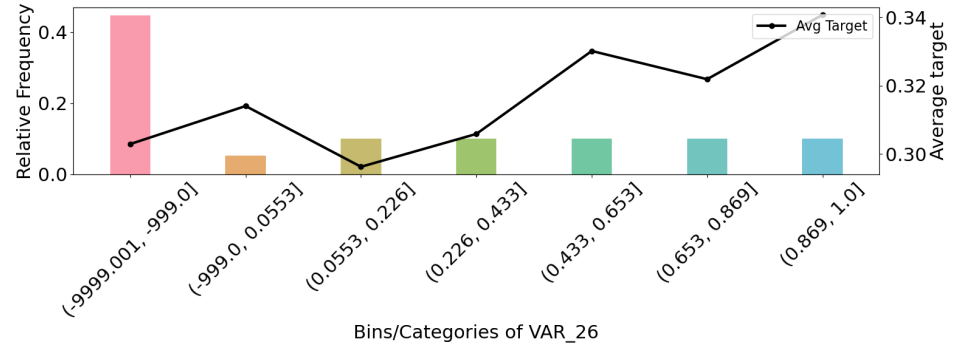

Relative Frequency
Average target
Avg Target
0.00
0.25
0.50
0.75
0.2
0.3
0.4
(-9999.001, -999.0]
(-999.0, 16570.2]
(16570.2, 74286.4]
(74286.4, 927661.0]
Bins/Categories of VAR_21

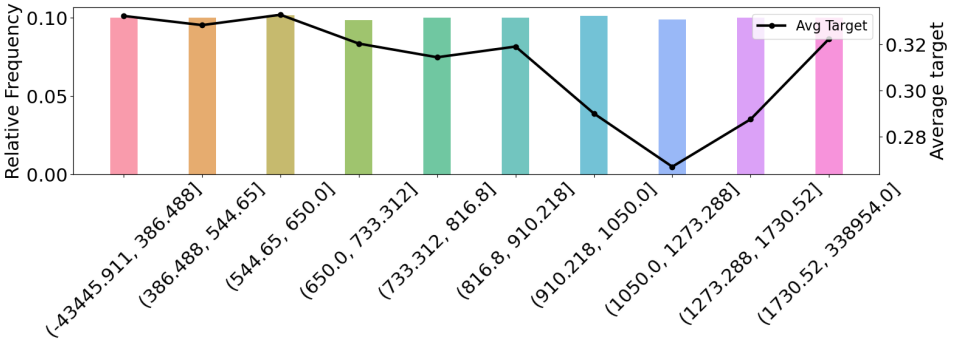

Relative Frequency
Average target
Avg Target
0.0
0.2
0.4
0.30
0.32
0.34
(-9999.001, -999.0]
(-999.0, 878.6]
(878.6, 3970.0]
(3970.0, 8099.8]
(8099.8, 15711.8]
(15711.8, 224457.0]
Bins/Categories of VAR_22

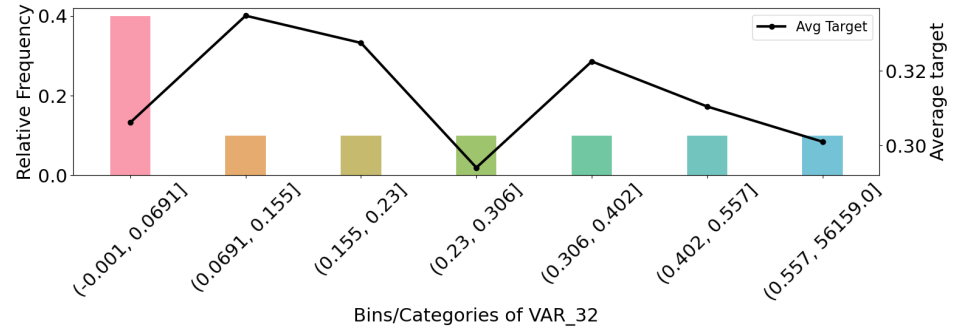

Relative Frequency
Average target
Avg Target
0.0
0.2
0.4
0.6
0.30
0.32
(-9999.001, -999.0]
(-999.0, 908.6]
(908.6, 8212.6]
(8212.6, 14794.0]
(14794.0, 406465.0]
Bins/Categories of VAR_23

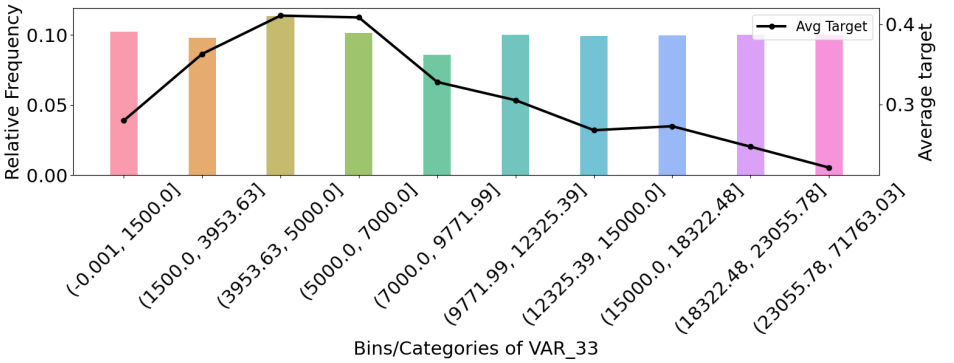

Relative Frequency
Average target
Avg Target
0.0
0.1
0.2
0.3
0.30
0.32
0.34
(-9999.001, -999.0]
(-999.0, 95.0]
(95.0, 500.0]
(500.0, 1000.0]
(1000.0, 1638.0]
(1638.0, 2720.0]
(2720.0, 5147.0]
(5147.0, 2523644.0]
Bins/Categories of VAR_24


Relative Frequency
Average target
Avg Target
0.0
0.2
0.4
0.31
0.32
0.33
(-9999.001, -999.0]
(-999.0, 136.0]
(136.0, 521.6]
(521.6, 1069.2]
(1069.2, 2043.8]
(2043.8, 4123.0]
(4123.0, 129461.0]
Bins/Categories of VAR_25


Relative Frequency
Average target
Avg Target
0.0
0.2
0.4
0.30
0.32
0.34
(-9999.001, -999.0]
(-999.0, 0.0553]
(0.0553, 0.226]
(0.226, 0.433]
(0.433, 0.653]
(0.653, 0.869]
(0.869, 1.0]
Bins/Categories of VAR_26


Relative Frequency
Average target
Avg Target
0.00
0.05
0.10
0.28
0.30
0.32
(-43445.911, 386.488]
(386.488, 544.65]
(544.65, 650.0]
(650.0, 733.312]
(733.312, 816.8]
(816.8, 910.218]
(910.218, 1050.0]
(1050.0, 1273.288]
(1273.288, 1730.52]
(1730.52, 338954.0]
Bins/Categories of VAR_31


Relative Frequency
Average target
Avg Target
0.0
0.2
0.4
0.30
0.32
(-0.001, 0.0691]
(0.0691, 0.155]
(0.155, 0.23]
(0.23, 0.306]
(0.306, 0.402]
(0.402, 0.557]
(0.557, 56159.0]
Bins/Categories of VAR_32


Relative Frequency
Average target
Avg Target
0.00
0.05
0.10
0.3
0.4
(-0.001, 1500.0]
(1500.0, 3953.63]
(3953.63, 5000.0]
(5000.0, 7000.0]
(7000.0, 9771.99]
(9771.99, 12325.39]
(12325.39, 15000.0]
(15000.0, 18322.48]
(18322.48, 23055.78]
(23055.78, 71763.03]
Bins/Categories of VAR_33

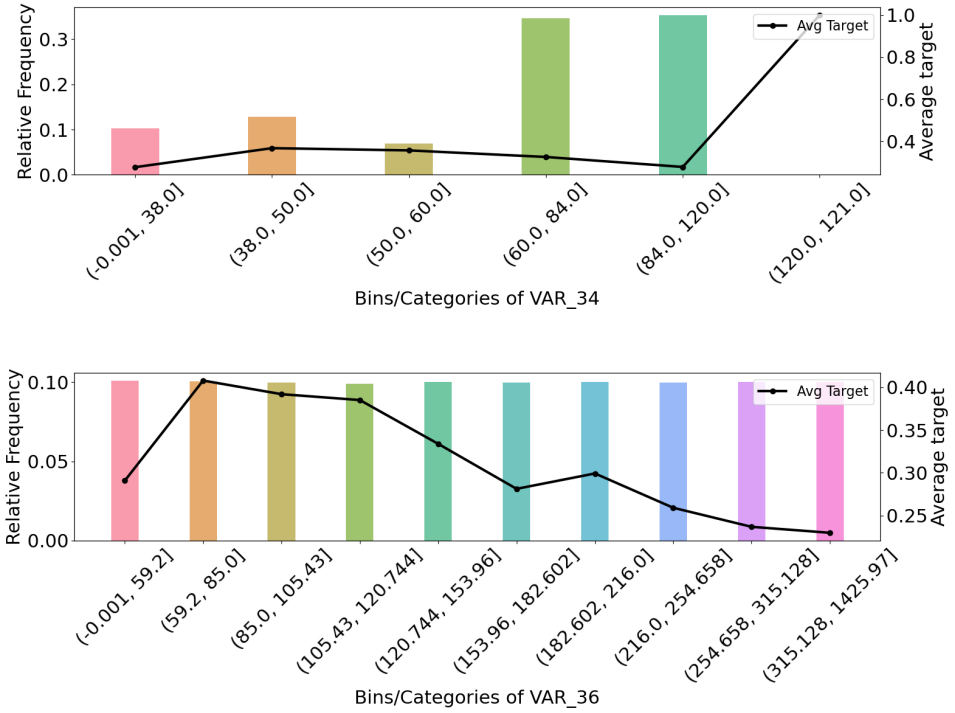


FIGURE 11: Relative Frequency of Continuous Variables with Average Target (Continuation).