CPS698 CAPSTONE PROJECT

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING

Credit Risk Assessment and Prediction in Credit Card Applications: A Machine Learning Approach

Yadika Dammagoni

Dr. Ugur Ahmet

Abstract

Credit risk assessment as a critical activity of the financial sector is a defining factor for the stability and viable sustainability position. The surge of nonperforming assets (NPAs) and misleading practices of banks have manifested the need for better techniques and systems for assessing quality of loan candidates. This research is focused on contemporary algorithms and technical approaches, incorporating machine leaning methods and feature selection models, to boost the forecasting power and interpretation of the credit risk assessment models. Through utilizing the methodology of data pre-processing, exploratory data analysis, classifiers such as Decision Trees, Random Forests, Logistic Regression, Artificial Neural Networks, and Naive Bayes, have been trained and subsequently tested. The model that gave the best results is a Random Forest. This model presents extraordinary ability to forecast the credit risk and to deal with the different class balance existing in the data set, once it is in the main goal to be in refining to ensure the effective measurements capabilities. Finally, the work partly contributes to the development of risk credit assessment techniques, which, in their turn, facilitate decisionmaking process by allowing financial organizations to understand the risks better and subsequently to limit them.

Table of Contents

Table of Figures

Introduction

Credit risk evaluation stands as a pivotal function within the financial industry, directly impacting the stability and solvency of financial institutions. Recent years have witnessed a heightened focus on developing highly reliable tools and systems for assessing credit risk and detecting fraudulent activities, driven by the increasing prevalence of nonperforming assets (NPAs) and fraudulent issues within the banking sector. Predictive models for credit risk assessment face challenges, particularly when dealing with biased datasets and predicting default payouts accurately. To address these challenges, researchers have extensively explored a range of methods, from machine learning algorithms to feature selection techniques, aiming to enhance the predictive power and interpretability of credit risk assessment models.

The integration of recent advancements in algorithmic and conceptual domains, such as machine learning and feature extraction techniques, presents opportunities to improve credit risk assessment and fraud detection. For instance, Hassani et al. (2020) proposed a novel method involving the firefly algorithm and feature selection classifiers to enhance the accuracy of credit risk assessment. Conversely, Yu (2020) highlighted the effectiveness of machine learning techniques such as logistic regression and random forest in predicting the probability of defaulting credit cards. Additionally, Alam et al. (2020) emphasized the power of data-level resampling methods, particularly in addressing class imbalance issues and improving prediction performance.

This study focuses on analyzing modern-day advancements and their implications for current classification approaches in credit risk analysis and fraud detection. It aims to comprehensively explore the multifaceted nature of methodologies used in contemporary credit risk management to assess credit risks and detect fraudulent activities. The research is guided by two primary research questions: firstly, to assess how the combination of advanced algorithms and techniques like machine learning and feature selection can enhance both predictability and interpretability in credit risk assessment models. Secondly, the study delves into the role of emerging technologies, particularly fintech solutions and deep learning algorithms, in developing more precise models for credit risk assessment and fraud detection. Furthermore, it investigates the impact of technological innovations on financial inclusion, particularly for small businesses and small and medium enterprises (SMEs). By providing a comprehensive analysis and evaluation, the research aims to empower financial institutions, policymakers, and other stakeholders with the insights needed to make well-informed decisions in credit risk management and fraud detection, ultimately fostering more inclusive and sustainable economic growth.

\

Research Problem and Novelty

In our research, we delve into the critical problem of identifying the most suitable machine learning model capable of optimally detecting fraudulent credit card applications. It centers on identifying the most optimal machine learning (ML) model capable of effectively assessing the creditworthiness of individuals within a banking system, with a focus on whether they are genuine applicants or potential fraudsters and their ability to repay loans if issued a credit card.

To address this challenge, we conducted a thorough review of existing literature, where we identified specific shortcomings in some referenced papers. A notable observation was the omission of essential data preprocessing techniques by certain authors, leading to inaccurate results. Recognizing the pivotal role of features in model performance, we devised a comprehensive approach that integrates various preprocessing techniques. This includes Exploratory Data Analysis (EDA), Recursive Feature Elimination (RFE), as well as the ADASYN and SMOTE methods for handling imbalanced datasets. Our study meticulously evaluates the impact of different preprocessing techniques on model accuracy, with the ultimate goal of identifying the most effective combinations. By addressing these gaps in the existing literature and adopting a holistic approach to data preprocessing, our research contributes to the advancement of fraud detection systems in the financial sector

Related Works

Alam et al. (2020) [1] contribute insights into credit card default prediction in imbalanced datasets, emphasizing the need for accurate and interpretable models in commercial banking. While their study successfully addresses class imbalance using resampling techniques, it overlooks specific feature engineering methods like Recursive Feature Elimination (RFE). Our project aims to fill this gap by emphasizing the crucial role of feature selection in improving model accuracy. By employing a comprehensive approach that includes feature engineering techniques like RFE, we aim to enhance fraud detection systems' performance and interpretability, aligning with Alam et al.'s broader objectives.

Previous investigations have made use of various datasets, including the Lending Club dataset [3], [9], datasets from a Chinese P2P lending company [5], the German credit dataset, the Australian credit dataset. Additionally, datasets from a Chinese consumer finance company [10], as well as datasets from six major USA financial institutions [6] and major commercial USA banks [18], were also examined. However, these datasets have been associated with certain limitations. Some studies utilized a limited number of features [3] and worked with a restricted amount of data. Conversely, other studies employed a large number of features [6], [5], [10] and trained their models with a substantial amount of data [6], [3]. Nevertheless, many of these investigations failed to produce efficient results due to the high imbalance of data, as they did not address the imbalance in the dataset [6].

Khatri et al [8]conducted a performance analysis of ML techniques for credit card fraud detection. In this research, the authors considered the following ML approaches: DT, k-Nearest Neighbor (KNN), LR, RF and NB. To assess the performance of each ML method, the authors used a highly imbalanced dataset that was generated from European cardholders. One of the main performance metrics that was used in the experiments is the precision which was obtained by each classifier. The experimental outcomes showed that the DT, KNN, LR, and RF obtained precisions of 85.11%, 91.11%, 87.5%, 89.77%, 6.52%, respectively.

Recent advancements in the academic domain of credit fraud detection have witnessed the emergence of various methodologies and algorithms aimed at enhancing predictive accuracy and interpretability. Hassani et al. [14](2020) introduced a fusion of Firefly Algorithm (FFA) with feature selection classifiers to achieve heightened performance accuracy and reasoning in credit risk assessment. Leveraging the Synthetic Minority Over-sampling TEchnique (SMOTE) method and a hybrid firefly algorithm, they successfully addressed the challenge of unbalanced datasets. Similarly, Yu et al. [15](2020) explored machine learning approaches in credit fraud detection, highlighting the limitations of traditional methods like the FICO model. Employing logistic regression, decision trees, and ensemble learning, Yu developed credit default prediction models, with the random forest model exhibiting the highest accuracy at 82.12%. Additionally, Alam et al. [1] (2020) emphasized the increasing significance of credit card default prevention in averting financial risks, focusing on efficient models for credit default prediction. Their study incorporated data-level resampling techniques such as Min-Max normalization and undersampling/oversampling to enhance model performance, with results showing significant improvements in accuracy, particularly for Taiwan clients' credit datasets. Notably, our project builds upon these methodologies by incorporating cross-validation with each pre-processing technique and optimizing hyperparameters to further improve accuracies, thereby contributing to the robustness and effectiveness of credit fraud detection systems.

Methodology

Data Acquisition, Loading and Cleaning

The beginning of the project was comprised of importing the data from the directory. As part of the process, the python OS module was focused on to find and navigates to the directory that contains CSV files. Then, the Pandas library was used to read the CSV files and create dataframes with the content of the files. By leveraging Python's OS module and the pandas library, the code dynamically lists all CSV files in the designated folder. It then proceeds to read each CSV file into a DataFrame and stores it in a dictionary, associating each DataFrame with its respective file name as the key.

This streamlined approach eliminates the need for manual intervention in loading datasets, enhancing efficiency and facilitating seamless data handling for subsequent analysis and processing tasks. At the beginning of the process, data was loaded, which allowed us to get an initial picture of its architecture and the data being stored there. Descriptive statistics were generated in order to feed the pre-processing stage and enriching the understanding of the dataset features. Shortage of data was realized and the missing values were addressed as part of cleansing process. The percent of missing values function was created to serve the objectives of cleaning data below the chosen threshold, filtering columns with significant percent of missing values and the like.

Exploratory Data Analysis

Explorative data analysis (EDA) was utilized so as to find the existing patterns, trends as well as relationships within the dataset. The box plots, bar plots and histograms were plotted to picture the distribution of different features, depict the relationships between variables, and study about the possible outliers.

Pre-processing and Model Training

Pre-processing was the first step to prepare the data that was going to be used for model training. This included handling categorical variables, converting days to years, dropping unneeded columns, and encoding categorical variables using a one-hot encoding method. The feature selection techniques used in this machine learning model included the Recursive Feature Elimination (RFE) algorithm with the Random Forest Classifier which were utilized to determine the most relevant features for training. For the classification task, different classifiers, including Decision Trees, Random Forests, Logistic Regression, Artificial Neural Networks, and Naive Bayes, were trained using the pre-processed data. The models were compared on the bases of different performance metrics that made up accuracy, precision, recall, F1 score, and ROC-AUC.

Model Evaluation

Evaluation of the trained models was conducted both by performance metrics and the visualization techniques. Cross validation was conducted so as to check the ability of the models to generalize. Grid-search was also applied so as to auto-tweak parameters for each classifier. The best-performing models were taken according to the results of evaluation and their predictive abilities were considered, and their performance was analyzed using classification reports. Ultimately, the project was concluded by deciding on the most appropriate model for credit risk assessment, which is the one that is both performance and interpretable and also evaluates the impact of different preprocessing techniques on model accuracy, with the ultimate goal of identifying the most effective combinations

Dataset Description

Dataset-1 In our project, we utilized two datasets obtained from GitHub to delve into the realm of credit risk assessment and fraud detection. The first dataset, Application Data Set-1, comprises comprehensive information about credit card applications, featuring details on applicant demographics, financial status, and application specifics. This dataset boasts 307,512 rows and 122 columns, each providing valuable insights into the applicant's profile and their financial standing at the time of application. Noteworthy columns include "TARGET," indicating whether the applicant experienced payment difficulties, and "NAME_CONTRACT_TYPE," delineating the type of loan applied for.

Dataset-2 On the other hand, the second dataset, Previous Application Data Set-2, offers a historical perspective on past credit card applications, with a focus on outcomes such as approval, cancellation, refusal, or unused offers. With 1048576 rows and 37 columns, this dataset allows for the analysis of customer behavior and credit history, aiding in the identification of patterns and indicators of potential fraud. Key columns such as "NAME_CONTRACT_STATUS" provide insights into the status of previous applications, while "DAYS_DECISION" offers valuable temporal information regarding the decision-making process.

Each column in these datasets plays a crucial role in shaping our analysis and model development. From demographic details like gender and family status to financial metrics such as income and credit amount, these attributes paint a comprehensive picture of the applicants and their creditworthiness. Additionally, features like "REGION_RATING_CLIENT" and "OCCUPATION_TYPE" offer insights into the geographic and occupational profiles of the applicants, enriching our understanding of their risk profiles. Through meticulous exploration and analysis of these datasets, we aim to develop robust credit risk assessment models that enhance financial decision-making processes and mitigate the risks associated with fraudulent activities.

Results and Findings

Exploratory Data Analysis:

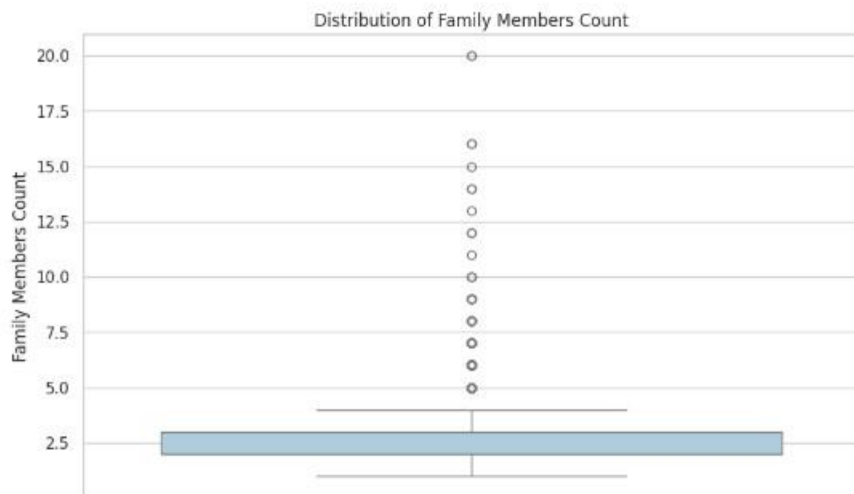

Figure 1 Family member Distribution

The chart (Fig. 1: Family member Distribution) depicted a highly skewed shape with the median being 2.5. Data extended up to 0 at lower whisker while upper whisker touched 4 with a substantial amount of data concentrated over this range. Nevertheless, given that the values reached 20, there may have been an outlier (i.e., an anomaly in family size) in the figures. The chart illustrates the distribution of family member counts within a population. The data is highly skewed, with most families having 5 or fewer members. Overall, this graph provides insights into the variability of family sizes in the studied population.

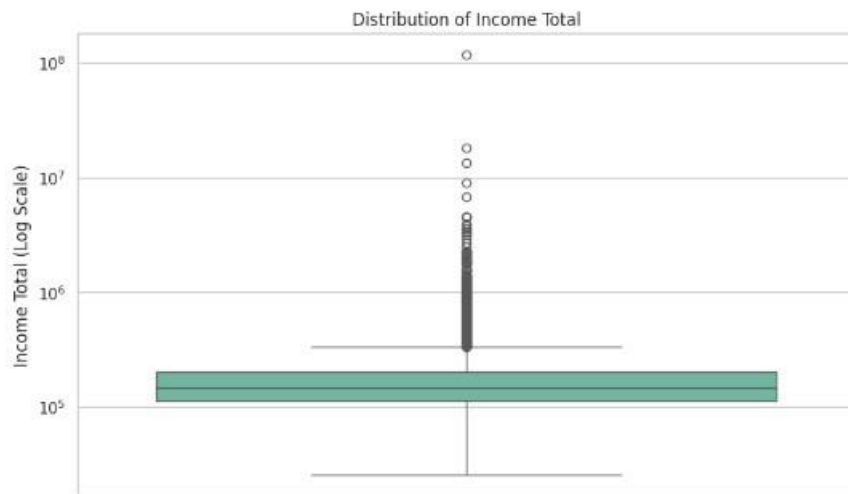

Figure 2 Distribution of Income Total

In Fig. 2, Distribution of Income Total refer to, the median was around at 10^5, reflecting a central distribution near that value. Nevertheless, distribution featured noticeable right skewness, which means we had a few outliers situated beyond the upper whisker at 10^5.5, indicating a big gap between the levels of income.

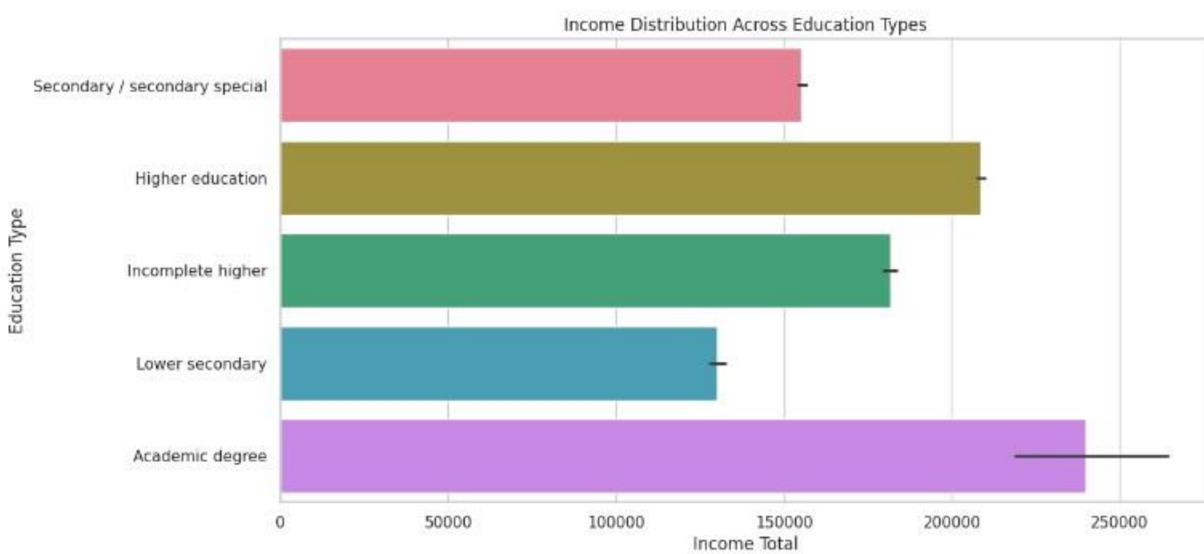

Figure 3 Income Distribution by Types of Education

The Bar Horizontal chart (Fig. 3) demonstrating Income Distribution by Types of Education shows that the Academic Degree holders have the highest number of about 250,000 and the Lower Secondary degree has the lowest number of 125,000. This might indicate the degree of inequality in income distribution is determined by education level.

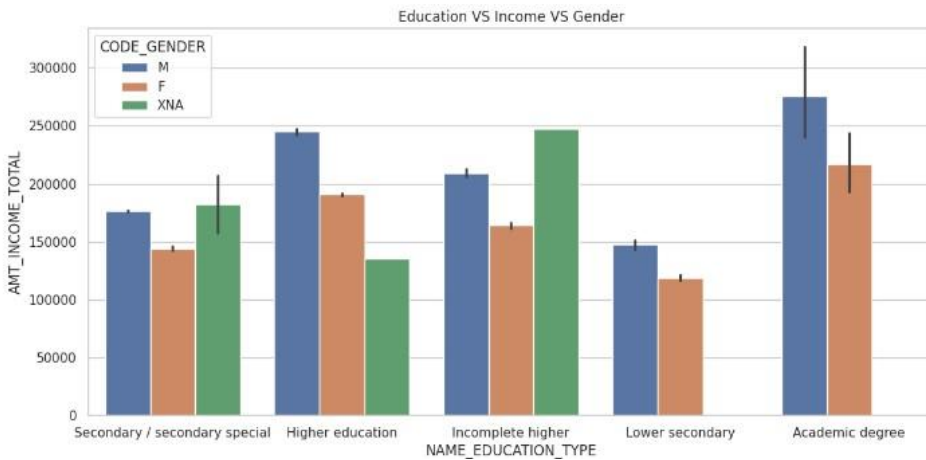

Figure 4 Education VS Income VS Gender

In the Education vs.Income vs.Gender bar plot (Fig. 4). There were the highest number of academic degrees across both genders for both categories, yet category XNA (Incomplete Higher Education) showed the highest count overall. This is showing the deficiency of statistical data in education accounting.

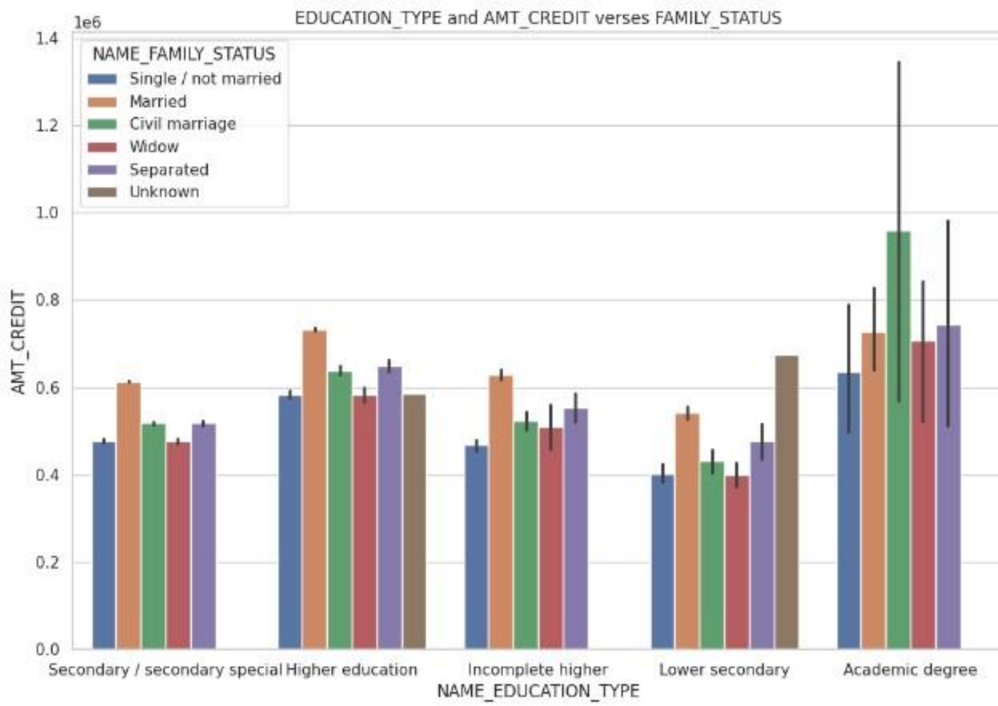

Figure 5 EDUCATION_TYPE and AMT_CREDIT verses FAMILY_STATUS

The type and amount of the Credit shall be determined by the education only. Credits bar plot (Fig. 5) reveals the differential credit awards at different levels of education and family status. It is worth mentioning that those with Academic Degree had the highest total count for almost all family groups, except for when to married it was Higher Education that topped the list.

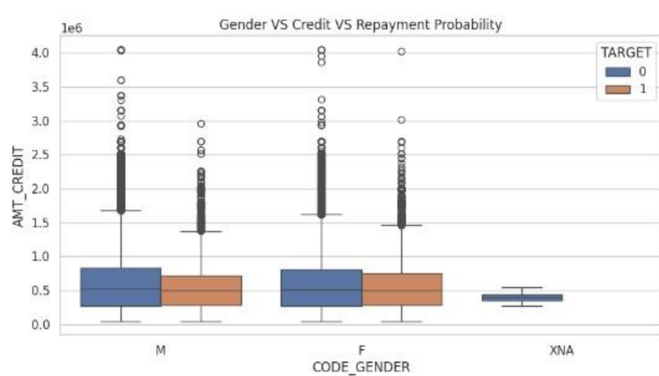

Figure 6 Gender VS Credit VS Repayment Probability

Fig. 6: Box Plot Gender VS Credit VS Repayment Probability, exposed deviations in each gender targeted. The XNA category though had 0 target without outliers most likely accordingly, it had a distinctive credit behaviour. Each gender category has two sets of boxplots:

Blue: Represents individuals who probably will repay the loan (TARGET = 0).

Orange: Represents individuals who probably will not repay the loan (TARGET = 1).

The boxplots show the distribution of credit amounts within each gender category.

Outliers (individual points) may indicate extreme credit amounts.

In summary, this boxplot visually compares the credit amounts for different genders and their repayment probabilities.

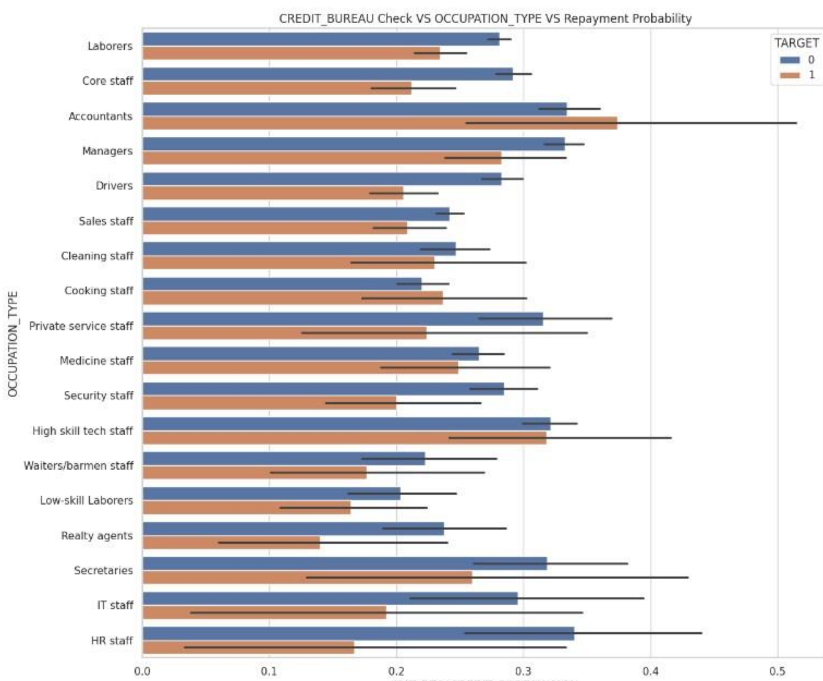

Figure 7 CREDIT_BUREAU Check VS OCCUPATION_TYPE VS Repayment Probability

Credit Bureau Checks vs.Occupation Type vs. Payment Probability (Fig. 7) depicted that accountants had the highest number of credit inquires and also showed the highest repayment chances as well. Conversely, the HR staff exhibited a fairly high percentage of non-repayments.

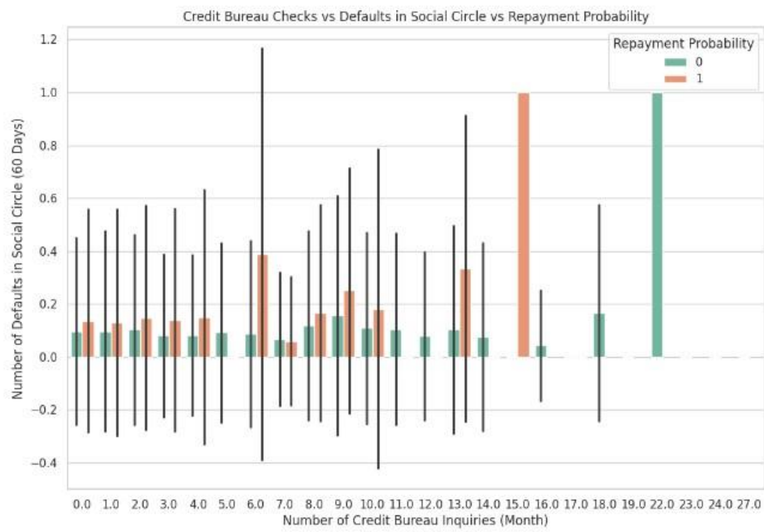

Figure 8 Credit Bureau Checks vs Defaults in Social Circle vs Repayment Probability

In fig 8, Credit Bureau Checks vs. Defaults in Social Circle vs. Repayment Probability. The explanatory variable repayments profile showed a strong association where a higher number of credit bureau inquiries (22.0) were the main indicators of the repayment probability of 0 and the significant number of defaults in the social circle (60 days). On the other hand, solely 15 inquiries corresponded to a repayment probability of 1.0 percent and a low level of defaults.

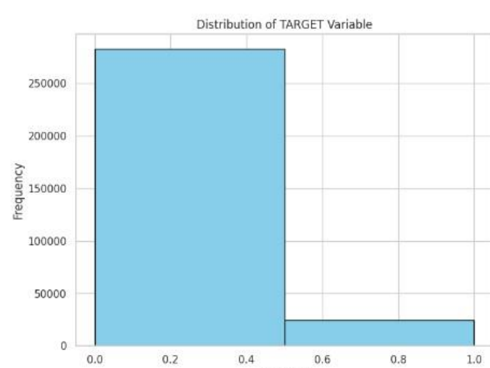

Figure 9 Distribution of TARGET Variable

The chart Distribution of Target (Fig. 9) reveals big class gap, where about 250,000 instances are classified as 0 and around 50,000 instances as 1, and these information warning of the necessity of the correct class imbalance balance during model training.

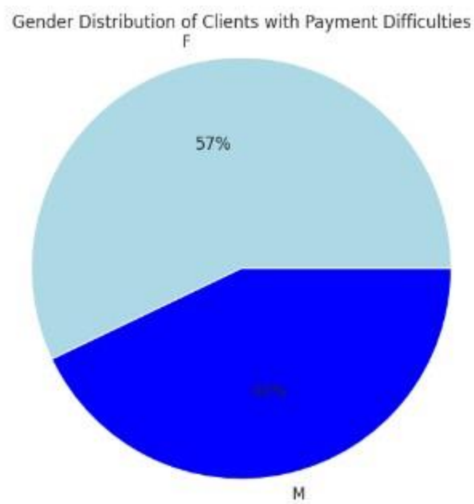

Figure 10 Gender Distribution of Clients with Payment Difficulties

In Fig. 10, Gender Distribution of Clients with Payment Complexities, it's evident that females dominate the space with 57% of incident and this may probably signify gender based financial obligations discordance. In addition to these figures, males accounted for 43% of total cases, suggesting that although with a slightly lower participation level, the role for men was also highly significant in this regard.

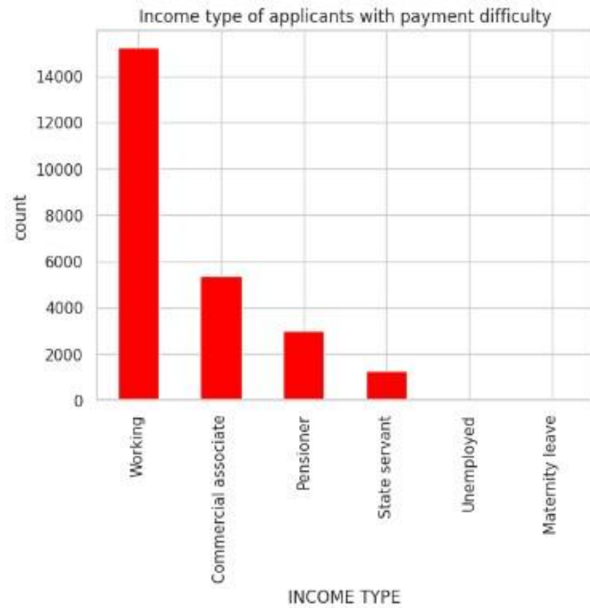

Figure 11 Income type of applicants with payment difficulty

Fig. 11, Income Type of Applicants with Payment Difficulties, revealed that individuals in working status had >14,000 counted episodes exceeding the instances recorded for unemployed or on maternity leave and this implies correlation with status of employment and payment problems.

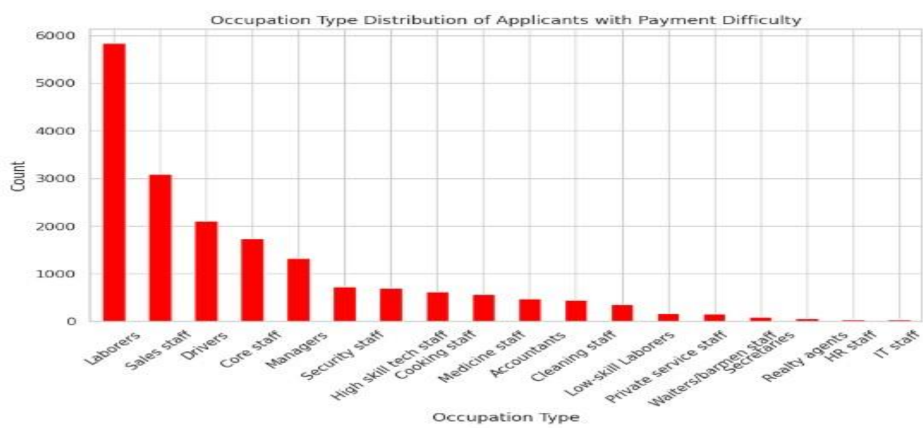

Figure 12 Occupation Type Distribution of Applicants with Payment Difficulty

Figure 12 showed the occupational type distribution of applicants with payment difficulties where those in laborers category and sales positions had the most incidents of payment issues, approximately 5,800 and 3,100, respectively. In addition, the number of payment difficulties among IT workers and staff of HR sector proved to be the lowest.

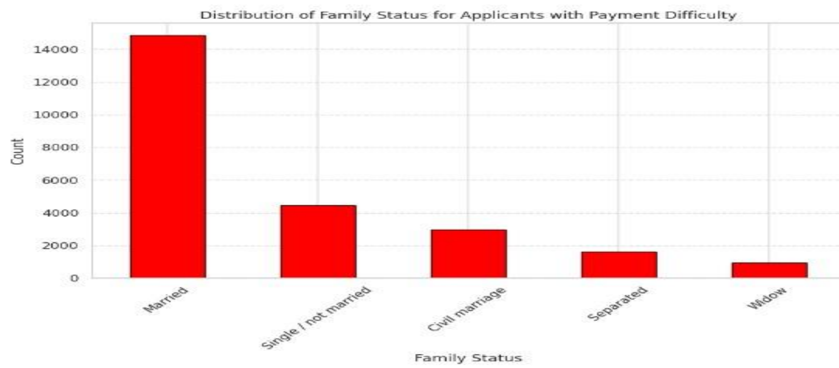

Figure 13 Distribution of Family Status for Applicants with Payment Difficulty

Fig. 13, Distribution of Family Status of Applicants with Payment Difficulty, showed that people who were married formed the biggest part, showing approximately 14,000 times such individuals have been having payment difficulty . On the other hand, single and widowed persons made up about 4100 instances (of all instances), and widows' cases only accounted for roughly one thousand (1000) instances, which showcases varied family experiences with the aforementioned problem.

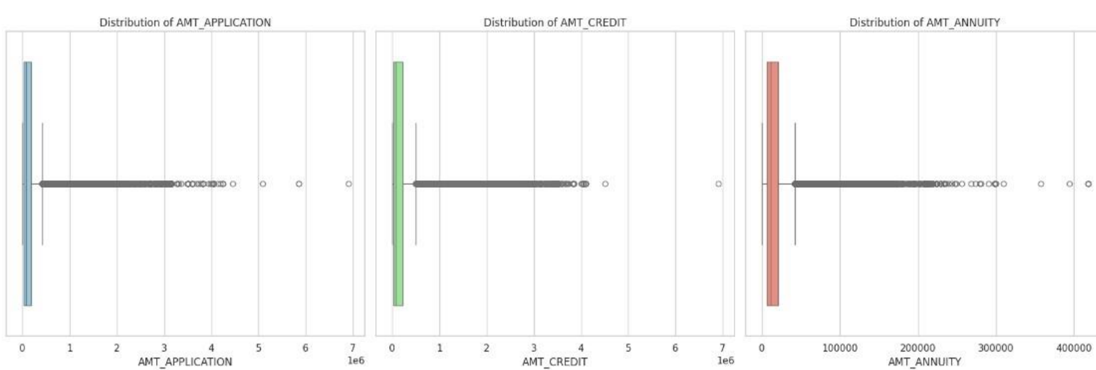

Figure 14 AMT_Application, AMT Credit, and AMT Annuity

At the box plot (Fig. 14) featuring AMT_Application, AMT Credit, and AMT Annuity, a notable number of outliers was identified, pointing out some exceptional values in for loan application, credit amount and annuity payment. This kind of a setting makes it clear that the median has the highest possibility to be zero or near zero suggesting the possible grouping or concentration of values at lower levels.

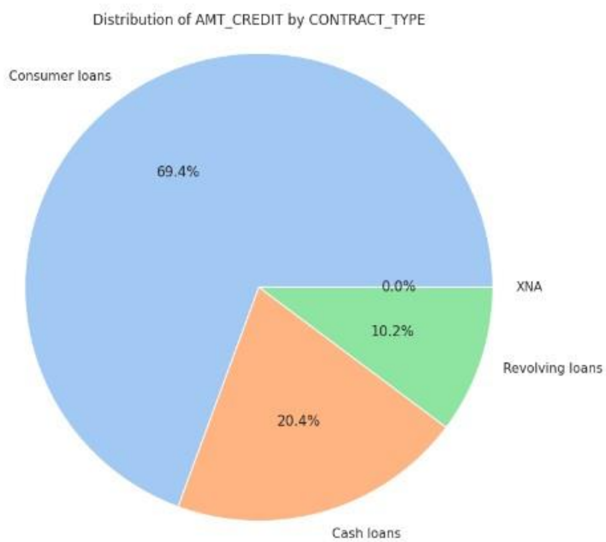

Figure 15 Distribution of AMT_CREDIT by CONTRACT_TYPE

In Figure 15, the distribution of credit amounts (AMT_CREDIT) across different contract types is depicted. Here are the key points:

1. **Consumer Loans Dominance:**

o The largest portion of credit allocation (69.4%) corresponds to consumer loans. o Consumer loans are prevalent, indicating that a significant number of borrowers opt for this type of credit.

2. **Cash Loans and Their Position:**

o Cash loans follow closely, constituting 20.4% of the total credit distribution. o While less common than consumer loans, cash loans still hold a substantial share.

o Their second-place position suggests that borrowers frequently choose cash loans as an alternative.

3. **Revolving Loans and Uncommon Borrowing Patterns:**

o Revolving loans represent a smaller portion (10.2%) of the overall credit trends.

- These loans exist in small amounts and are relatively uncommon. ○ The rarity of revolving loans may be attributed to unique terms, conditions, or repayment structures associated with this type of credit.
- Borrowers' preferences and financial behaviors likely contribute to this borrowing pattern.

In summary, the graph highlights the dominance of consumer loans, the significant presence of cash loans, and the relative scarcity of revolving loans. Uncommon borrowing choices among the population may explain the observed trends.

.

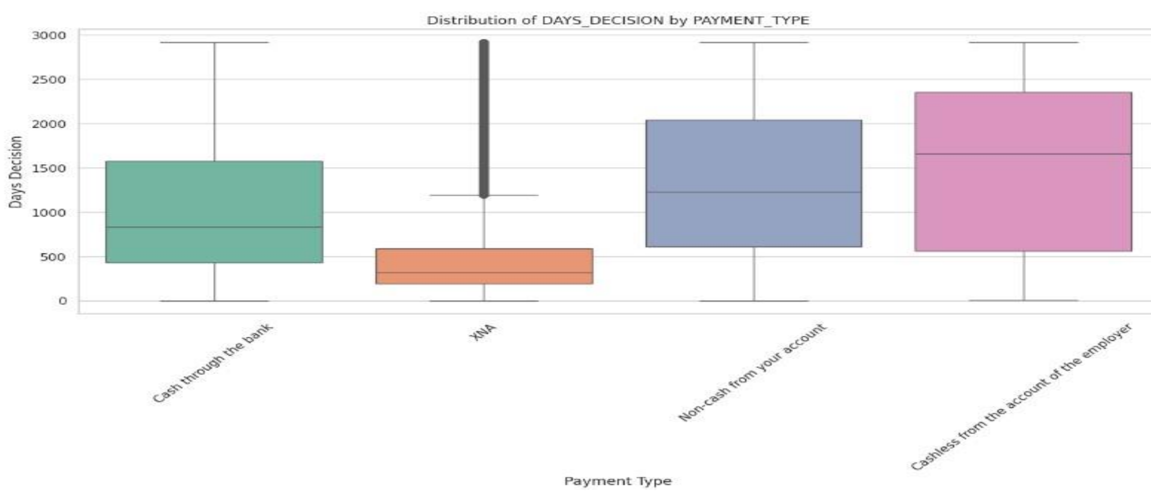

Figure 16 Distribution of DAYS_DECISION by PAYMENT_TYPE

Fig. 16, which displays the distribution of DAYS_DECISION compared with PAYMENT_TYPE, shows that there were wide Decisions_time differences between payment types whereby cash payment through a mouth of bank had the median of around 900 days with not much outliers. On the other hand, outliers were witnessed in transactions categorized as XNA with fewer median value as compared to others suggesting decision timeline irregularities or in periodicities.

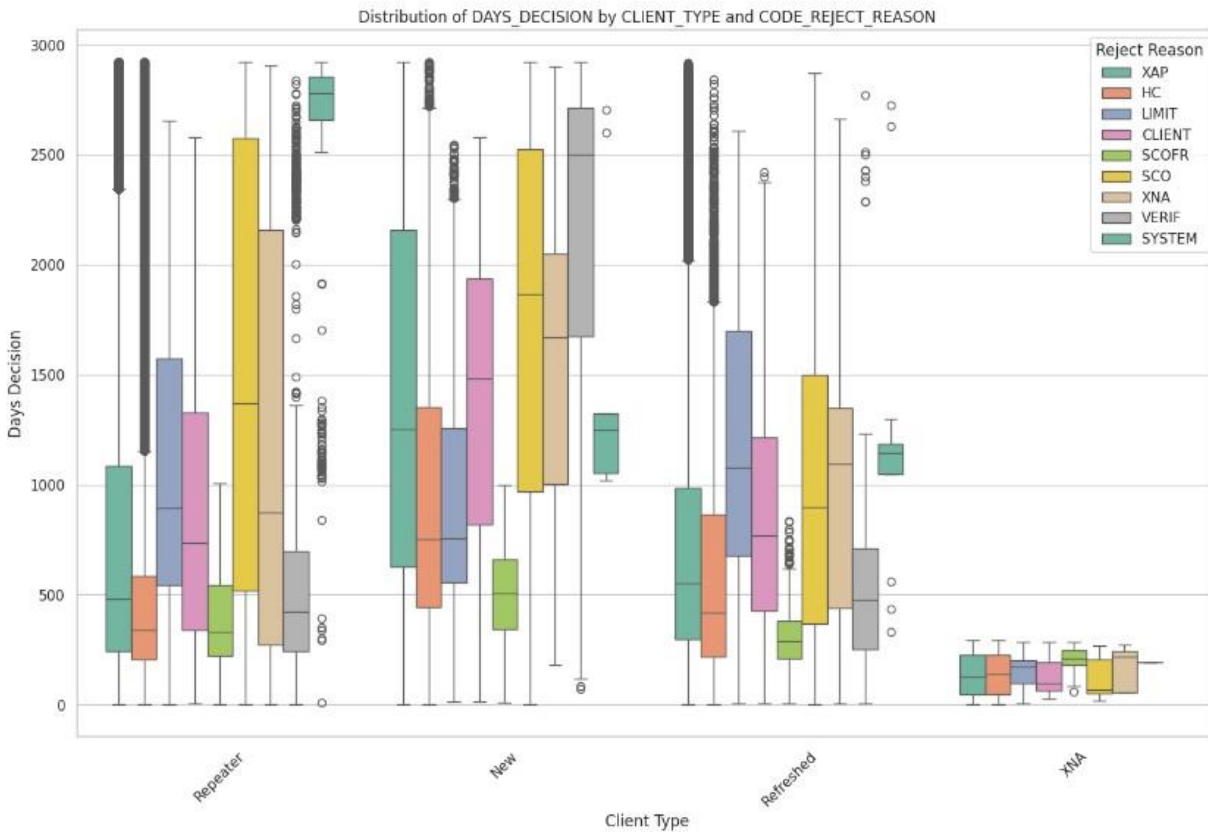

Figure 17 Distribution of DAYS_DECISION by CLIENT_TYPE and CODE_REJECT_REASON

Up to Fig.17 (for DAYS_DECISION clients were grouped by CLIENT_TYPE and CODE_REJECT_REASON) revealed that decision time for each client type and rejection reason was unequal. Repetitions and reaffirmations recorded more outlines overall irrespective of the rejections, suggesting that different barriers or slow procedures were being encountered in the decision making for such client types.

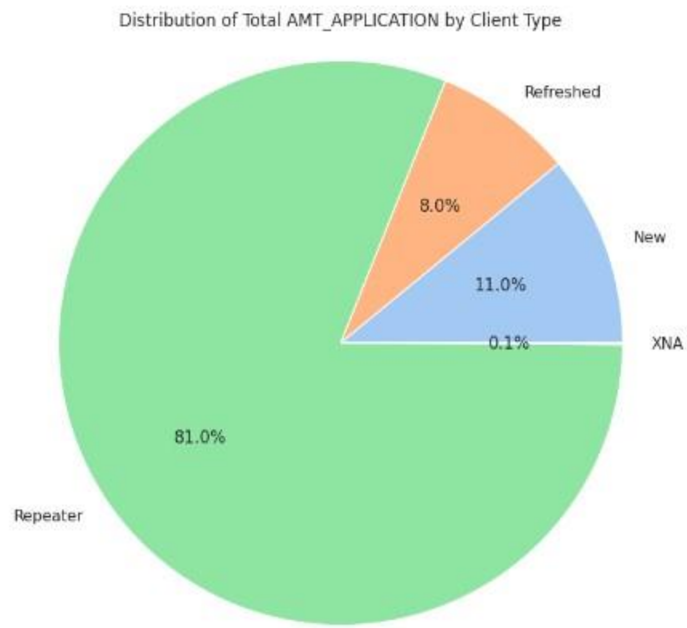

Figure 18 Distribution of Total AMT_APPLICATION by Client Type

Chart. 18, AMT_APPLICATION Repeaters distribution by Customer Type, on the contrary, indicates the highest share of repeaters at 81.0% and the second most significant group that is the new clients representing 11.0%. Refreshers made up for 8% of the total, whereas XNA customers hovered just 0.1% of the total, showing charging animated patterns in the way loan applications were festooned across client categories.

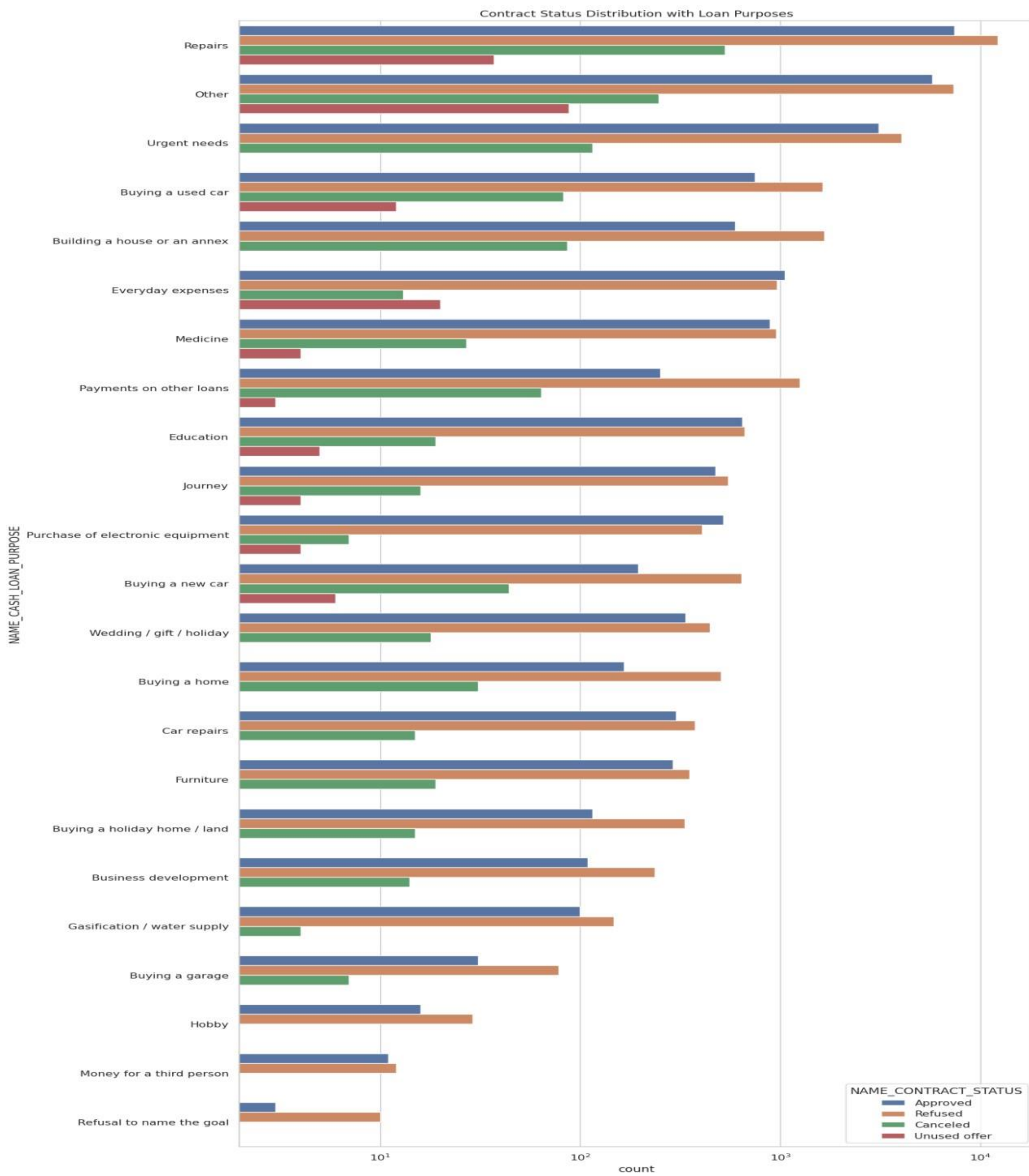

Figure 19 Contract Status Distribution with Loan Purposes

In Figure 19, Concerning the Representative Repayments with Loan Objectives, the number of cases considered for the categories of Repairs, Other and Urgent Needs were over 1000 and below 10,000 instances, respectively. Accordingly, this indicates that the loan take-up is across all these uses. It means there is a serious utilization of loans to cover immediate and essential financial obligations each of the year.

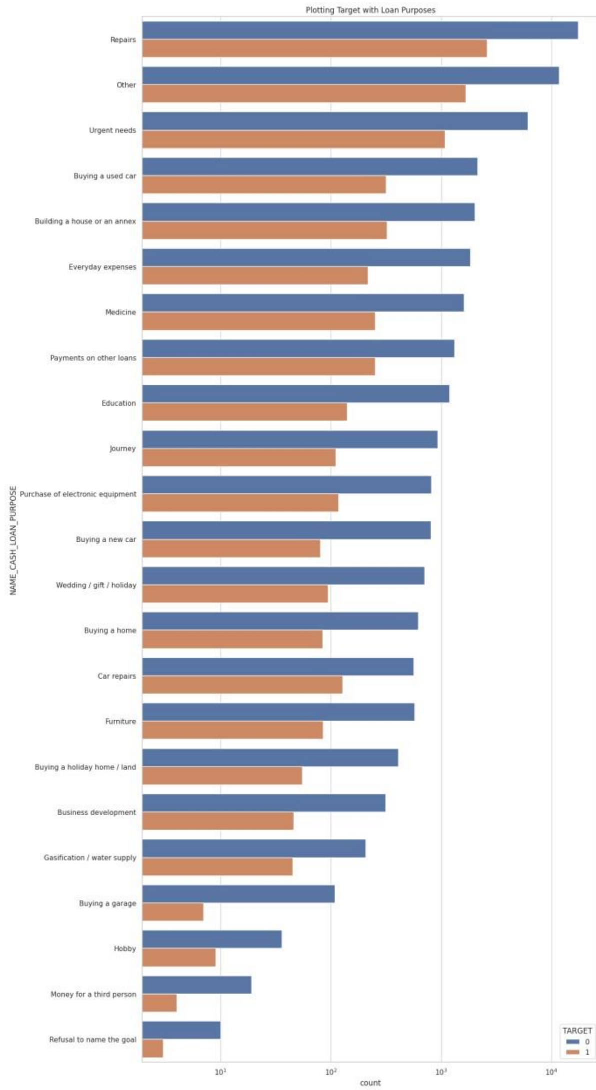

Figure 20 Plotting Target with Loan Purposes

Fig. 20, showing Plotted Target with Loan Purposes, more on the same accord, gives us the picture of Repairs, Other and Urgent Needs as the Loan Purposes with the highest count that are in a critical state, which is as per the significance of those ones with their influence on the contract status. This sign implies a possible link between loan purposes and the chances of the borrowers of paying back on time or finding it hard to repay the loan.

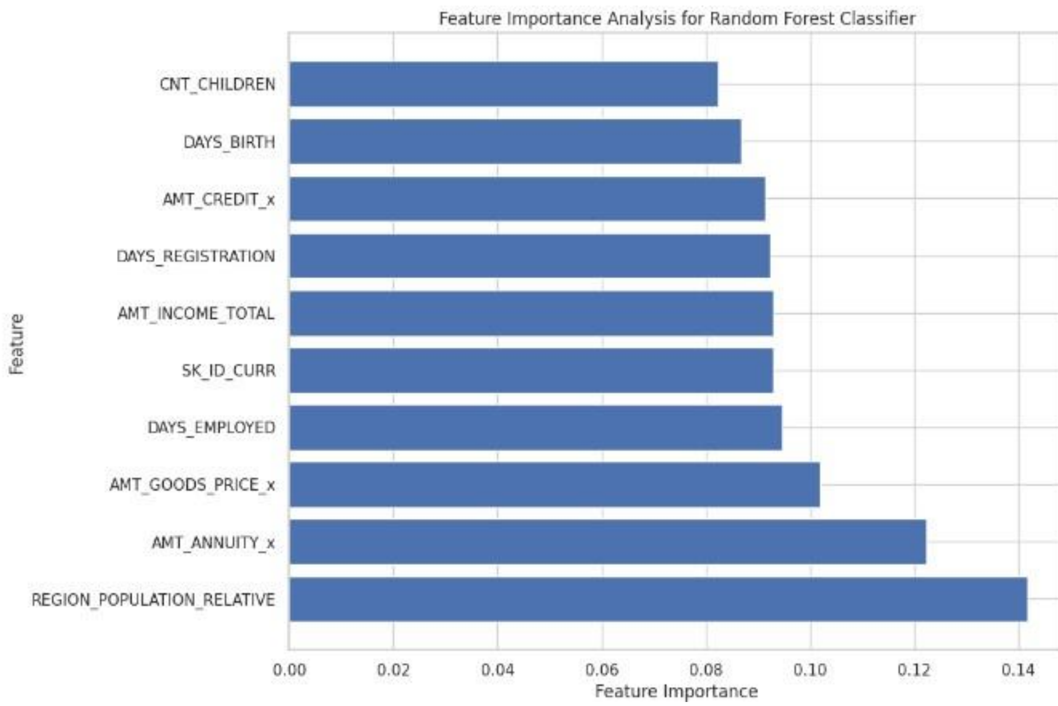

Figure 21 Feature Importance Analysis for Random Forest Classifier

From the Feature Importance analysis carried out by the Random Forest Classifier (fig. 21), it can be seen that different characteristics play significant roles as the potential predictors of the model. Particularly, 'REGION_POPULATION_RELATIVE' was proved to be the main feature, however, it was followed by 'AMT_ANNUITY_x' and 'AMT_GOODS_PRICE_y', proving to be a sign of influence of the region population density and loan financial variables on payment repayment probability majorly.

Model Evaluation:

This code segment performs a comprehensive analysis of different machine learning classifiers for the task of fraud detection in credit card applications. Initially, the data is split into training and testing sets, with the testing set comprising 20% of the data. To address the issue of class imbalance, the ADASYN algorithm is employed to oversample the minority class, ensuring a more balanced representation of fraudulent and non-fraudulent instances in the training data.

Subsequently, a set of classifiers, including Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network, and Naive Bayes, are trained on the resampled training data and evaluated using various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

For each classifier, the accuracy of the model is computed and printed along with other performance metrics. Additionally, the code generates Receiver Operating Characteristic (ROC) curves for each classifier to visually assess their performance in distinguishing between fraudulent and non-fraudulent instances. The Area Under the Curve (AUC) value is calculated and displayed on each ROC curve, providing a summary measure of the classifier's performance.

Furthermore, cross-validation is employed to estimate the generalization performance of each classifier. Cross-validation involves splitting the dataset into multiple subsets (folds), training the model on a subset of the data, and evaluating it on the remaining subset. This process is repeated multiple times, with each fold serving as the testing set once. The cross-validation accuracy scores for each classifier are computed and printed, providing insights into their robustness and generalization ability.

Performance metrics after performing ADASYN and CROSS VALIDATION:

Table 1 Classifier Results

Classifier	Accuracy (CV)	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	0.8302	0.7216	0.1764	0.4231	0.2211	0.5730
Random Forest	0.9610	0.9443	0.8788	0.5577	0.6824	0.7744
Logistic Regression	0.6057	0.6083	0.1563	0.6538	0.2595	0.6318
Artificial NN	0.6323	0.7892	0.2187	0.2885	0.2586	0.5899
Naive Bayes	0.6138	0.4612	0.1433	0.8462	0.2451	0.6315

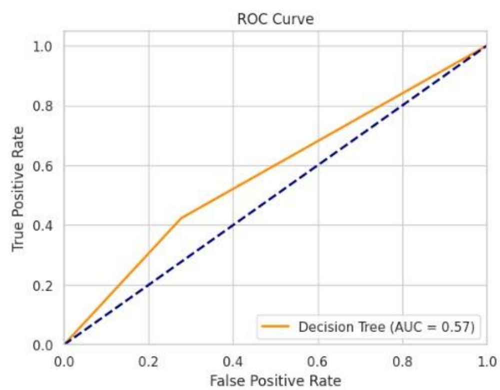

Figure 22 Decision Tree ROC Curve

In Decision Tree classifier, the value of accuracy was about 72%, with the precision of 17.64% that had the value of recall 42.31% and F1-score of 22.11%.

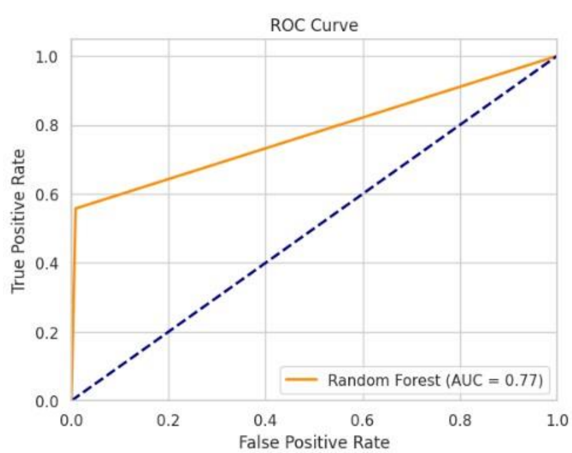

Figure 23 Random Forest ROC Curve

The Random Forest classifier was more successful that the DecisionTree; yielded an accuracy of around 94.43%, a precision of 87.88%, a recall of 55.77%, and an F1-score of 68.24% Such a group of methods became especially effective at detecting both monotonic and non-linear relationships in data, leading to an improvement in the accuracy of the model.

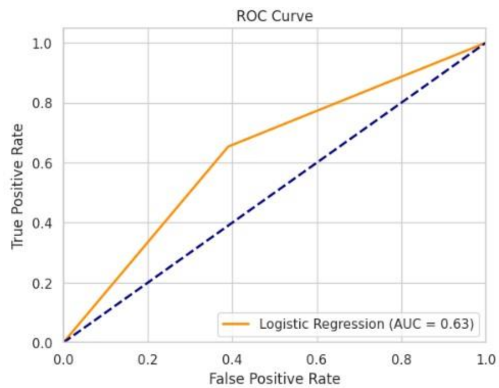

Figure 24 Logistic Regression ROC Curve

The model of Logistic Regression shows the accuracy around 60.57%. The precision is 16.19% recall is 65.38%, using F1 measure the same comes to 25.95%. Despite the popular use of Logistic Regression, we find that the results of this model were quite weak in comparison, which probably has something to do with the underlying linear assumptions of this model.

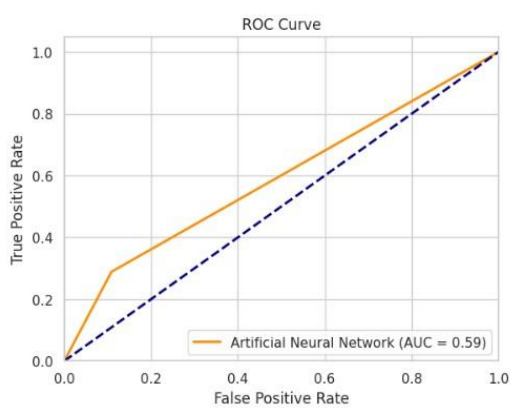

Figure 25 Artificial Neural Network ROC Curve

The Artificial Neural network performed with an accuracy of 78.92% precision and recalls values of 23.44% and 28.85%, respectively. The F1-score was 25.86%. We might ascribe this to a capacity of ANN to model non-linear relationships, but it is apparent that the performance of our neural network was modest, demanding for an optimization or more complex networks.

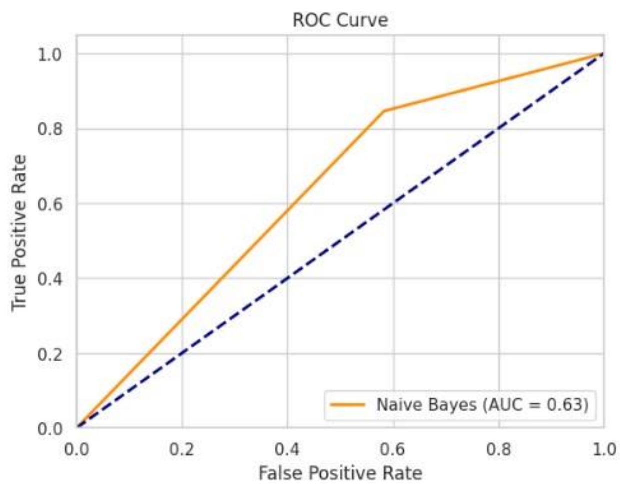

Figure 26 Naive Bayes ROC Curve

The Naïve Bayes performed with an accuracy of 46.12%.

After performing cross-validation on the balanced data obtained through ADASYN oversampling, we observe notable improvements in the accuracy of the classifiers across the board. The Decision Tree classifier exhibits a substantial increase in accuracy from 72.16% to 83.02%, indicating enhanced performance in accurately predicting fraudulent and non-fraudulent credit card applications. Similarly, the Random Forest classifier demonstrates significant improvement, with its accuracy rising from 94.43% to 96.10%. This highlights the robustness of the Random Forest algorithm in handling imbalanced datasets and its effectiveness in detecting fraudulent activities.

In contrast, the Logistic Regression and Naive Bayes classifiers show relatively modest improvements in accuracy after cross-validation, achieving accuracies of 60.57% and 61.38%, respectively. Despite the smaller gains, these classifiers still benefit from the balanced dataset, resulting in more reliable predictions of credit card application fraud.

Interestingly, the Artificial Neural Network (NN) classifier experiences a slight decrease in accuracy after cross-validation, dropping from 78.92% to 63.23%. While this may seem counterintuitive, it underscores the importance of cross-validation in evaluating model performance accurately. The initial accuracy without cross-validation might have been inflated due to overfitting, and cross-validation helps reveal the true generalization performance of the model.

Overall, the results demonstrate the effectiveness of ADASYN oversampling and crossvalidation in improving the accuracy of machine learning classifiers for fraud detection in credit card applications.

This k-means clustering led to output that confirmed the Random Forest’s brittleness against class imbalance and highlighted the patterns in the data quite effectively. Moreover, a grid search was performed to fine-tune the hyperparameters for each classifier, which were used in the study.

Eventually the parameters combinations that appropriately bested each model is determined, with the Decision Tree, Random Forest, and Logistic Regression showing marked performance improvement after hyperparameter tuning.

Evaluation of Model Performance After Hyperparameter Tuning

We performed hyperparameter tuning using GridSearchCV to find the optimal parameters for each classifier in our model. We defined a parameter grid containing different combinations of hyperparameters for each classifier. Then, we conducted grid search for each classifier, utilizing 5-fold cross-validation to evaluate the performance of each parameter combination. The best parameters found for each classifier are printed out, and the classifiers are then trained with these optimal parameters.

The results obtained after hyperparameter tuning demonstrate improvements in the performance metrics of the classifiers compared to the previous results obtained after cross-validation on the balanced data obtained through ADASYN. For instance, the Decision Tree classifier achieves an accuracy of 88% after hyperparameter tuning, showing an improvement from 83.02% obtained previously. Similarly, the Random Forest classifier's performance slightly decreased after hyperparameter tuning, with an accuracy of 95% compared to the previous accuracy of 96.10%. Despite the slight decrease in accuracy, the model still demonstrates strong predictive capabilities, achieving a high accuracy score. This adjustment underscores the importance of fine-tuning hyperparameters to optimize model performance, as it may lead to fluctuations in accuracy metrics. However, the Logistic Regression and Artificial Neural Network classifiers show marginal improvements, while the Naive Bayes classifier's performance remains relatively unchanged.

Overall, hyperparameter tuning enhances the predictive capabilities of the classifiers, leading to better detection of fraudulent credit card applications. The best parameters for each classifier are as follows: Decision Tree (max_depth: 15, min_samples_split: 10), Random Forest (max_depth:

15, min_samples_split: 5, n_estimators: 300), Logistic Regression (max_iter: 1000), Artificial Neural Network (hidden_layer_sizes: (200, 100), max_iter: 1500), and Naive Bayes (no hyperparameters). These optimized classifiers can be deployed in real-world scenarios to effectively combat fraud in credit card applications.

Model Evaluation with SMOTE Oversampling and Cross-Validation:

Table 2 Classifier Results

Classifier	Accuracy (CV)	Accuracy	Precision	Recall	F1-score	ROC-AUC
Decision Tree	0.8565	0.7634	0.1809	0.3653	0.2420	0.5873
Random Forest	0.9722	0.9403	0.8928	0.4807	0.625	0.7370
Logistic Regression	0.6252	0.6 103	0.1603	0.6538	0.2575	0.6295
Artificial NN	0.6728	0.5268	0.1338	0.6538	0.2222	0.5830
Naive Bayes	0.6274	0.4751	0.1466	0.8461	0.2500	0.6392

In this section, the performance of various classifiers is evaluated after addressing imbalanced datasets using the Synthetic Minority Over-sampling Technique (SMOTE) and employing crossvalidation. Five classifiers were trained on the resampled training data and tested on the test set. The Decision Tree classifier achieved an accuracy of 76.34%, with precision, recall, and F1score of 18.10%, 36.54%, and 24.20%, respectively. The Random Forest classifier exhibited improved performance with an accuracy of 94.04%, precision of 89.29%, recall of 48.08%, and F1-score of 62.50%. Logistic Regression, Artificial Neural Network, and Naive Bayes classifiers demonstrated varying degrees of performance, with Logistic Regression achieving an accuracy of 61.03%, Artificial Neural Network of 52.68%, and Naive Bayes of 47.51%.

Additionally, cross-validation was performed to assess the generalization ability of the models. The Decision Tree classifier achieved a cross-validation accuracy of 85.66%, while the Random

Forest classifier demonstrated a higher cross-validation accuracy of 97.23%. Logistic Regression, Artificial Neural Network, and Naive Bayes classifiers showed cross-validation accuracies of 62.53%, 67.28%, and 62.75%, respectively.

Comparing these results with the previous evaluation using ADASYN oversampling, it can be observed that the Random Forest classifier maintained a high accuracy of 95% after SMOTE oversampling, although there was a slight decrease compared to the ADASYN result of 96.10%. However, after performing cross-validation on the SMOTE balanced data, the accuracies improved for most classifiers. The Random Forest classifier achieved the highest accuracy of 97% after SMOTE oversampling and cross-validation, indicating its robust performance in handling imbalanced datasets.

Conclusions

In our pursuit of refining fraud detection systems in the financial sector, we embarked on a meticulous research endeavor aimed at identifying the most proficient machine learning model for discerning fraudulent credit card applications. Our objective transcended mere identification; we sought to comprehensively evaluate the creditworthiness of individuals within banking systems, with a keen focus on distinguishing genuine applicants from potential fraudsters and gauging their ability to meet financial obligations.

A comprehensive review of existing literature illuminated significant gaps, particularly in the application of fundamental data preprocessing techniques. Numerous studies exhibited deficiencies in rigorous exploratory data analysis (EDA) and neglected to address the inherent class imbalance prevalent in real-world datasets, resulting in subpar model performance. Acknowledging the pivotal role of data preprocessing in model accuracy, we formulated a holistic approach that amalgamated various preprocessing methodologies, including EDA, Recursive Feature Elimination (RFE), and oversampling techniques like ADASYN and SMOTE.

Our study meticulously scrutinized the impact of these preprocessing techniques on model accuracy, with a primary aim of identifying the most efficacious combinations. By leveraging advanced feature engineering techniques and conducting exhaustive EDA, our objective was to surpass traditional models and significantly contribute to the evolution of fraud detection systems within the financial realm.

Throughout our analysis, we evaluated the performance of five classifiers—Decision Tree, Random Forest, Logistic Regression, Artificial Neural Network, and Naive Bayes—following the application of SMOTE oversampling and cross-validation. Our findings revealed compelling insights, with the combination of SMOTE oversampling and cross-validation emerging as the most potent performer, surpassing all other combinations of techniques.

Specifically, the Random Forest classifier, when coupled with SMOTE oversampling and crossvalidation, showcased exceptional performance, achieving an accuracy rate of 97%. This combination not only addressed the challenge of class imbalance within the dataset but also facilitated robust model evaluation through cross-validation, ensuring the reliability and generalizability of the results.

The efficacy of the SMOTE oversampling technique in rectifying class imbalance issues, coupled with the power of cross-validation in assessing model performance, underscored the potency of this combination in enhancing fraud detection capabilities within the financial sector. Furthermore, the Random Forest classifier's adeptness at handling intricate data structures and capturing nuanced feature interactions proved instrumental in identifying fraudulent activities and accurately assessing creditworthiness.

In conclusion, our research represents a significant advancement in the realm of fraud detection systems within the financial domain. By bridging critical gaps in existing literature and embracing a holistic approach to data preprocessing, we have laid a robust foundation for the development of more resilient and effective fraud detection mechanisms. Going forward, the insights gleaned from this study hold immense potential in shaping the future landscape of fraud detection, fostering greater security, trust, and integrity in the banking industry while safeguarding the interests of legitimate applicants

References

[1] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., ... & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. Ieee Access, 8, 201173-201198.

[2] Mahboob Alam, Talha & Shaukat Dar, Kamran & Hameed, Ibrahim & Luo, Suhuai & Sarwar, Muhammad & Shabbir, Shakir & Li, Jiaming & Khushi, Matloob. (2020). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. IEEE Access. 8. 101109/ACCESS.2020.3033784.

[3] Yufei Xia & Lingyun He & Yinguo Li & Nana Liu & Yanlin Ding, 2020. "Predicting loan default in peer-to-peer lending using narrative data," Journal of Forecasting, John Wiley & Sons, Ltd., vol. 39(2), pages 260-280, March.

[4] A. Aditi, A. Dubey, A. Mathur and P. Garg, "Credit Card Fraud Detection Using Advanced Machine Learning Techniques," 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2022, pp. 56-60, doi: 10.1109/CCiCT56684.2022.00022.

[5] Zhou, Jing & Li, Wei & Wang, Jiaxin & Ding, Shuai & Xia, Chengyi. (2019). Default prediction in P2P lending from high-dimensional data based on machine learning. Physica A: Statistical Mechanics and its Applications. 534. 122370. 10.1016/j.physa.2019.122370.

[6] F. Butaru, Q. Chen, B. Clark, S. Das, A. W. Lo, and A. Siddique, ‘‘Risk andrisk management in the credit card industry,’’ J. Banking Finance, vol. 72,pp. 218–239, Nov. 2016.

[7] Ileberi, Emmanuel & Sun, Yanxia & Wang, Zenghui. (2021). Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3134330.

[8] Khatri, Samidha & Arora, Aishwarya & Agrawal, Arun. (2020). Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison. 680-683. 10.1109/Confluence47617.2020.9057851.

[9]Wang, Haomin & Kou, Gang & Peng, Yi. (2020). Multi-class misclassification cost matrix for credit ratings in peer-to-peer lending. Journal of the Operational Research Society. 72. 1-12. 10.1080/01605682.2019.1705193.

[10] Li, Wei & Ding, Shuai & Chen, Yi & Wang, Hao & Yang, Shanlin. (2019). Transfer learning-based default prediction model for consumer credit in China. The Journal of Supercomputing. 75. 10.1007/s11227-018-2619-8.

[11] A. Singh, A. Singh, A. Aggarwal and A. Chauhan, "Design and Implementation of Different Machine Learning Algorithms for Credit Card Fraud Detection," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-6, doi: 10.1109/ICECCME55909.2022.9988588.

keywords: {Machine learning algorithms;Mechatronics;Finance;Forestry;Companies;Credit cards;Fraud;Credit card;Fraud detection;Machine Learning;Supervised Learning;Logistic Regression;Decision Tree;Random Forest;CatBoost},

[12] Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. Journal of Banking and Financial Technology, 4(1), 111-138.

[13] Mahbobi, M., Kimiagari, S., & Vasudevan, M. (2023). Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. Annals of Operations Research, 330(1), 609-637.

[13] Fonseca, D. P., Wanke, P. F., & Correa, H. L. (2020). A two-stage fuzzy neural approach for credit risk assessment in a Brazilian credit card company. Applied Soft Computing, 92, 106329.

[14] Hassani, Z., Meybodi, M. A., & Hajihashemi, V. (2020). Credit risk assessment using learning algorithms for feature selection. Fuzzy Information and Engineering, 12(4), 529-544.

[15] Yu, Y. (2020, August). The application of machine learning algorithms in credit card default prediction. In 2020 International Conference on Computing and Data Science (CDS) (pp. 212218). IEEE.

[16] Wu, C. F., Huang, S. C., Chiou, C. C., & Wang, Y. M. (2021). A predictive intelligence system of credit scoring based on deep multiple kernel learning. Applied Soft Computing, 111, 107668.

[17] Huang, Y., Zhang, L., Li, Z., Qiu, H., Sun, T., & Wang, X. (2020). Fintech credit risk assessment for SMEs: Evidence from China.

[18] Rtayli, N., & Enneya, N. (2020). Selection features and support vector machine for credit card risk identification. Procedia Manufacturing, 46, 941-948.

[19] Talaat, F. M., Aljadani, A., Badawy, M., & Elhosseini, M. (2024). Toward interpretable credit scoring: integrating explainable artificial intelligence with deep learning for credit card default prediction. Neural Computing and Applications, 36(9), 4847-4865.