Machine Learning and Credit Default

Tom O'Connell[1], Thomas Birkner[2], and Theyab Alrashdi[3]

[1]Arizona State University, Tempe, AZ 85281, USA

Abstract

This paper applies machine learning (ML) techniques to predict credit defaults, leveraging a dataset from Home Credit, which includes over 300,000 instances and 122 features per instance. Our analysis focuses on the impact of borrowers' attributes, such as age, education, and other demographic factors, on their likelihood of defaulting on loans.

Exploratory Data Analysis (EDA) revealed that younger borrowers and those with lower educational attainment are more likely to default. Various ML models, including logistic regression, XGBoost, neural networks, and Random Forest, were evaluated to improve prediction accuracy over traditional statistical methods. The models' performance was assessed based on accuracy, precision, recall, F1 score, and AUC score.

Our findings indicate that ML can significantly enhance the predictive power of credit default models, offering potential benefits to lenders in terms of risk reduction and borrowers through better-managed credit terms. The paper illustrates the potential of ML in transforming credit risk assessment and provides a foundation for further research into its application across different datasets.

Contents

1 Introduction

The provision of credit is a cornerstone of modern financial systems, vital for both personal and business endeavors. As the demand for credit has expanded, so has the complexity and risk associated with lending. This paper seeks to evaluate the performance of ML techniques to enhance the prediction of credit defaults, leveraging a dataset provided by Home Credit, a multinational non-bank financial institution specializing in loans for consumers with little or no credit history.

1.1 Background

Between 1987 and 2005, the United States consumer credit rose by 182%, and credit card debt rose by 416% [8]. The consistent rise in credit card balances, except for 2009 and 2010 during the Great Recession and the COVID-19 pandemic, emphasizes the cyclical vulnerability of consumers to economic downturns. The record-high balance of over a trillion dollars for three consecutive quarters beginning in 2023 is a testament to this trend. Accurate credit evaluation is essential for protecting the financial health of lending institutions and ensuring consumers are not overextended in their credit obligations. The historical data demonstrates the propensity of consumers to accumulate large balances rapidly, driven by factors such as record interest rates and persistent inflation [9] [13]. Historically, banks have relied on various credit scoring systems to assess lending risk to individuals and businesses. These systems evaluate the ability of borrowers to repay debts based on financial histories and other related metrics [2]. However, the advantages of ML offer a transformative approach by utilizing large datasets and complex algorithms to predict outcomes more accurately.

1.2 Previous Research

Currently, ML techniques are being widely studied and evaluated to predict loan defaults, with popular methods including Logistic Regression, Decision Trees, Random Forests, and XGBoost.

According to research by Zhang et al., machine learning-based credit risk assessment models achieved an accuracy level of over 80%, compared to around 70% accuracy for traditional statistical models [21].

Logistic Regression is favored for its simplicity, robust performance, and ease of use [17]. It excels over Linear Regression in default prediction by providing probabilities ranging from 0 to 1, indicating the likelihood of a loan default [5]. Research conducted by Han et al. demonstrated the efficacy of Logistic Regression combined with the Cox proportional hazard model in predicting student loan defaults, pinpointing key factors such as age, income, repayment amounts, and college degree. The model achieved an AUC of 0.697 in test scenarios, highlighting its accuracy and reliability [11].

Decision Trees segment data into branches to form a tree-like structure, where each node represents a decision point, and each leaf represents a potential outcome [5]. In previous studies, Wang et al. explored the efficacy of Decision Trees in conjunction with Logistic Regression for predicting delinquencies in business service accounts. Their findings indicated that Decision Trees performed better than Logistic Regression, particularly when managing a small number of attributes within a large dataset [20].

Random Forest improves upon Decision Trees by constructing multiple trees during training and selecting the most common outcome among them as the final prediction [14]. Malekipirbazari and Aksakalli's study employed a Random Forest model to identify viable peer-to-peer borrowers, outperforming traditional FICO credit scores [16].

XGBoost, an enhanced version of the Gradient Boosting algorithm, leverages decision trees and has achieved top-tier results across various ML challenges. Li et al. research involving an XGBoost-based model for predicting peer-to-peer loan defaults showed a remarkable accuracy of 97.705%, fitting the results better than Logistic Regression and Decision Trees.

1.3 Project Plan

The primary objective of this paper is to evaluate the efficiency of various ML models in predicting credit defaults. This involves a detailed comparison of Logistic Regression, XGBoost, Random Forest Classifier, and Neural Networks. The paper aims to determine which model performs best in accuracy, precision, recall, F1-score, and AUC score for predicting a client defaulting on their loan. The paper will utilize a comprehensive Home Credit dataset, including detailed information on loan applicants' demographics, loan history, and repayment records. This data will be used to train and test the ML models.

1.4 Other Analysis

This study will also explore correlations between default and borrowers' education levels, as education is often a critical yet overlooked factor influencing loan default risks [15]. Questions regarding how default rates vary by education level and the impact of borrowers' age on default probabilities will also be addressed, providing a broader understanding of the factors that influence credit risk [15] [18].

1.5 Benefits

By integrating ML into credit risk assessment, this paper aims to improve the accuracy of default predictions and explore how these technologies can be applied in real-world financial settings to enhance decision-making processes and risk management strategies. The outcomes of this paper are expected to provide valuable insights into the strengths and limitations of various ML models in the context of credit risk and to lay the groundwork for further research and development in this critical area of finance.

2 Methods

2.1 Introduction to Dataset

The dataset employed in this study is sourced from the Home Credit Default Risk competition on Kaggle. Home Credit, an international consumer finance provider primarily operating in Europe, aims to offer loan opportunities to consumers with limited access to traditional financial services. The primary datasets utilized in this research are labeled as application_train and application_test, both of which have been pre-split for training and testing purposes. The application_train dataset includes the target variable indicating whether a loan was defaulted on.

2.2 Data Processing

Upon loading the data into our environment using Python libraries such as Pandas, NumPy, and Seaborn, we first assessed the structure and completeness of the dataset. The application_train dataset contained 307,511 instances and 122 features, encompassing various applicant details from demographic information to financial history.

A critical step in our data processing pipeline involved handling missing values, a common issue in real-world data sets. Missing data was pervasive across several features, from simple demographic information to more complex financial metrics. We addressed these gaps by implementing a dual-strategy approach:

- Numerical Features: Missing values in numerical features were filled using the median of each respective feature, providing a robust central tendency measure less sensitive to outliers.
- Categorical Features: For categorical data, we applied the mode to fill missing values, ensuring that the most frequently occurring category within each feature was used as a placeholder for missing data.

Here are examples of the types of features that experienced issues with missing values and how they were handled:

- **AMT_ANNUITY**: The loan's monthly payment amount had missing values in some records. These were filled using the median of the existing values to maintain consistency in loan payment estimations.
- **NAME_HOUSING_TYPE**: This feature describes the type of housing the applicant lives in; options include House/apartment, With parents, and Municipal apartment, among others. For entries where this information was missing, the mode was used for imputation, aligning the missing entries with the most prevalent type of housing among the applicants.

Finally, we utilized Standardization. This was a crucial step to ensure that each feature contributed equally to our predictive models. We used the StandardScaler from scikit-learn, a powerful preprocessing tool that transforms each feature into having zero mean and unit variance. This normalization is essential, mainly when dealing with features that vary widely in magnitudes, units, and range. Scaling the features onto a standard scale prevented any feature from dominating the model's predictions due to its scale, leading to more stable and consistent performance across all inputs. Standardization was essential for the following features:

- **AMT_INCOME_TOTAL**: This feature represents an applicant's total income ranging from a few thousand to several million. The range can be quite vast, necessitating scaling to prevent it from outweighing other features in the model.
- **DAYS_BIRTH**: Recorded as the number of days since birth (negative because it's counted backward from the current application date), this feature's scale is significantly different as it typically ranges in the tens of thousands.
- **AMT_CREDIT**: The total credit amount of the loan, which, similar to income, can vary drastically from small sums to large values, depending on the applicant's borrowing needs.
- **REGION_POPULATION_RELATIVE**: This feature is a normalized score that quantifies the population density of the applicant's region. It is typically a small decimal, far less in magnitude compared to income or credit amounts.

2.3 Exploratory Data Analysis

We conducted thorough EDA and utilized sophisticated visualization techniques to understand the underlying patterns and relationships within the Home Credit dataset. This approach was pivotal in establishing clear correlations between various features and the probability of default.

2.3.1 Data Distribution Analysis

- **Histograms**: These were used to visualize the distribution of various numerical features such as DAYS_BIRTH, AMT_CREDIT, AMT_INCOME_TOTAL, and AMT_ANNUITY. These histograms helped assess the spread and central tendencies of these variables.
- **Bar Charts**: Bar charts were generated for categorical data to display the frequency of different categories within features like NAME_CONTRACT_TYPE and CODE_GENDER. This visualization helped identify dominant categories and potential imbalances in the data.

2.3.2 Relationship Analysis

- **Correlation Heatmaps**: A correlation matrix was generated and visualized through heatmaps to examine the relationships between numerical features and the target variable (TARGET). This method is crucial for identifying features with either a strong positive or negative association with the likelihood of a credit default.

2.3.3 Comparative Analysis

- **Grouped Bar Charts**: These charts were used to compare data points across different subgroups, such as comparing default rates among different genders and age groups. This type of visualization is particularly effective in highlighting differences or similarities across categorical variables.
- **Kernel Density Estimates (KDE)**: KDE plots were employed to analyze the density of loan amounts (AMT_CREDIT) for loans that were repaid versus those that were not. This helps in understanding how loan amount distributions differ between the two groups.

2.3.4 Advanced Visual Techniques

- **Box Plots**: These were useful for visualizing the distribution of credit amounts or age in terms of quartiles and for spotting outliers.
- **Scatter Plots**: These helped visualize the relationship between continuous variables and identify potential clusters or anomalies. For example, plotting AMT_CREDIT against DAYS_EMPLOYED helped us observe how employment duration relates to the credit amounts granted.

These EDA techniques are instrumental in gaining a deeper insight into the data, which aids in better hypothesis building and the subsequent development of more accurate predictive models. Each method uncovers different aspects of the data, from general distributions and central tendencies to more complex relationships and group dynamics within the dataset.

2.4 Introduction to Machine Learning Models

As mentioned, ML techniques are currently being widely studied and evaluated to predict loan defaults. We the following models due to the extensive research conducted into their performance, specifically in predicting borrower defaults [7] [3] [11].

Logistic Regression evaluates the relationship between a dependent variable and one or more independent variables by estimating probabilities categorized as 0 (default) or 1 (non-default). It effectively measures and provides a direction of association (either positive or negative) between predictors and the outcome [19] [1].

XGBoost stands for eXtreme Gradient Boosting. It enhances model performance and speed through parallel and distributed computing and an iterative ensemble of decision trees (weak learners), focusing progressively on correcting previous errors [6] [4].

Random Forest utilizes the technique of 'bagging' to create a less correlated ensemble of decision trees by selecting random subsets of features for each tree. This method significantly reduces the risk of overfitting, common in individual decision trees, by averaging multiple uncorrelated trees, which decreases variance and prediction errors [12].

Neural Networks are designed to simulate the behavior of biological neural networks. A neural network consists of layers of interconnected nodes (neurons), where each connection (synapse) can transmit a signal from one neuron to another [10].

2.5 Models Implementation

- Logistic Regression: The logistic regression model was configured to maximize iteration for optimal learning, with the max_iter parameter set to 1000. This model helped establish a baseline for performance metrics such as accuracy, precision, recall, and F1 score.
- Random Forest Classifier: This model used an ensemble of decision trees to improve prediction accuracy and control overfitting, leveraging the robustness of multiple learning estimators. The Random Forest model was particularly tuned for its depth and number of estimators, ensuring a comprehensive learning process. A grid search was used which systematically explores various combinations of hyperparameters

to identify the optimal settings. The results of the grid search led to the optimal settings being: 'maxdepth': 20, 'min samples leaf': 4, 'min samples split': 2, 'n estimators': 300.

- XGBoost: As a highly efficient and scalable implementation of gradient boosting, XGBoost was applied due to its ability to handle large datasets quickly and accurately. It was used to refine the prediction accuracy further and handle the feature-laden dataset effectively.
- Neural Networks: The first neural network architecture, consists of three densely connected layers: an input layer with 242 neurons representing input features, followed by two hidden layers with 128 and 64 neurons, respectively, all utilizing the rectified linear unit (ReLU) activation function for introducing non-linearity. Dropout layers with a dropout rate of 0.3 are applied after each hidden layer to prevent overfitting, and the output layer comprises a single neuron with a sigmoid activation function for binary classification tasks, such as loan default prediction
- Neural Networks: The second neural network featured the same architecture as the first, however it added the parameter of class weights. Class weights help handle imbalanced datasets, in this dataset the amount of defaulted loans is significantly smaller than non-defaulted ones. By assigning higher weights to defaulted loans and lower weights to non-defaulted ones, the model focuses more on learning from the minority class instances, which in this case is defaults.
- Neural Networks: The third neural network that was tested included an input layer of 25 neurons representing 25 features after employing the feature selection technique of SelectKBest. It featured two densely connected hidden layers: the first with 64 neurons and the second with 16 neurons, both employing the rectified linear unit (ReLU) activation function. A dropout layer with a dropout rate of 0.2 is applied after the first hidden layer to prevent overfitting. The output layer comprises a single neuron with a sigmoid activation function.

Each model was trained on the standardized and encoded training data. Predictive performance was then measured on the training set to understand the model's learning capability and applicability using standard metrics like accuracy, precision, recall, and the F1 score. Moreover, ROC curves were plotted to evaluate the models' actual positive rate against the false positive rate, providing insight into their effectiveness at various threshold settings.

3 Results

3.1 Visualization and Exploratory Data Analysis Results

The EDA conducted as part of our study on the Home Credit dataset provided significant insights into the factors influencing loan defaults. The EDA revealed numerous connections and relationships regarding the likelihood of borrower default. The key findings are as follows:

3.1.1 Age Analysis

The EDA revealed that younger applicants have higher default rates than older borrowers. This pattern suggests that age and implicit factors like financial maturity and stability significantly affect loan repayment behaviors. Histograms illustrating the distribution of defaults across different age groups clearly showed that default rates are higher among the younger demographics.

3.1.2 Education and Loan Defaults

Education level proved to be a strong indicator of default likelihood. Applicants with 'Lower Secondary' education had the highest default rates, while those holding an 'Academic Degree' exhibited the lowest. This finding underscores the importance of educational attainment in predicting financial behavior and

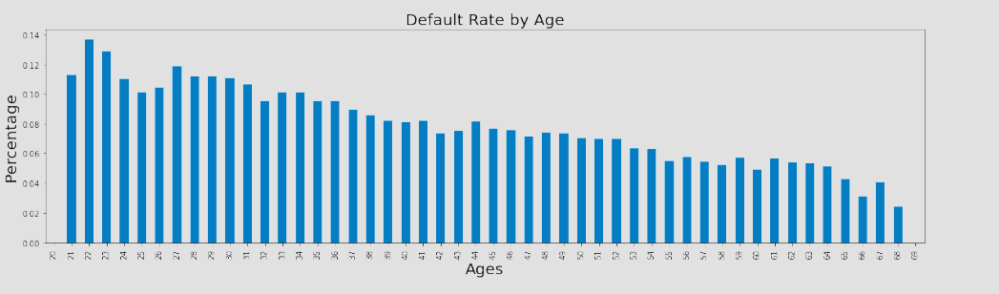

Figure 1: Age Analysis

creditworthiness. Bar charts displaying education-level default rates highlighted these disparities, suggesting that higher educational levels correlate with lower default risks.

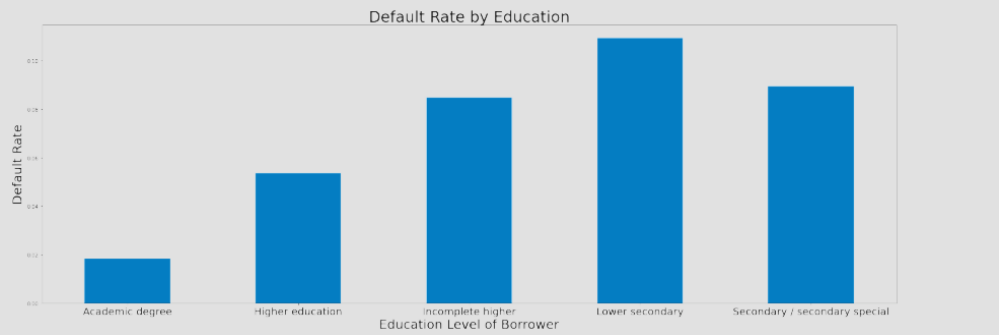

Figure 2: Education Analysis

3.1.3 Impact of Gender on Defaults

Our analysis also explored gender differences in loan defaults. Despite a higher number of female applicants, the data showed no significant variation in default rates between genders. This observation was supported by histograms and bar charts comparing the default rates across male and female applicants, indicating that gender alone does not significantly influence default probabilities.

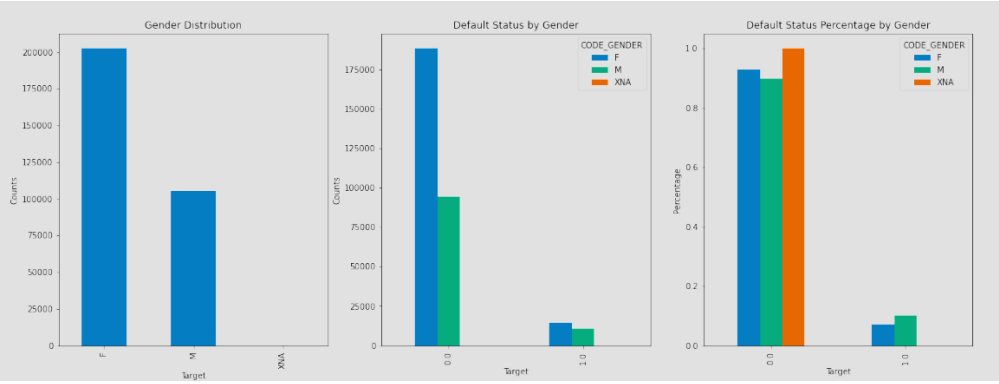

Figure 3: Gender Analysis

3.1.4 Home Ownership and Loan Defaults

Analysis of the real estate ownership showed that applicants owning real property ('Y' category) are less likely to default compared to those who do not own real estate ('N' category). This trend was depicted through bar charts illustrating the default rates for each category, suggesting that property ownership might provide financial stability that mitigates default risk.

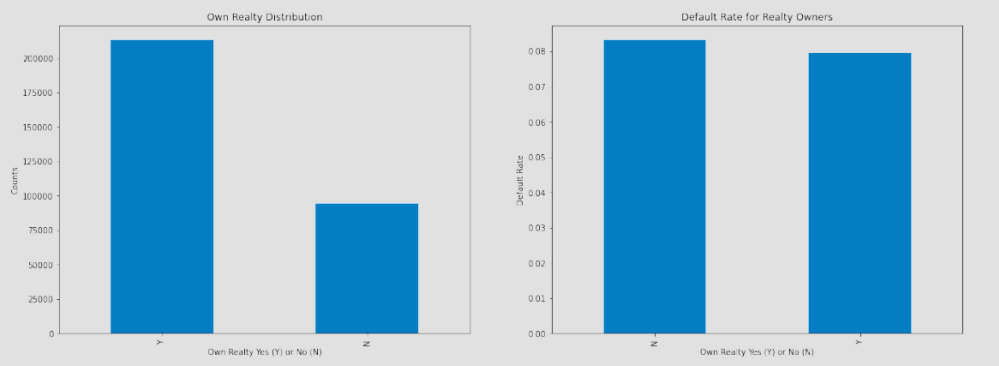

Figure 4: Real Estate Analysis

3.1.5 Correlation Analysis

A heatmap of the top 20 features with the highest correlation to the target variable ('TARGET') was generated. This analysis helped identify key variables most influential in predicting defaults, such as external credit ratings (EXT_SOURCE variables) and DAYS_BIRTH (age). The heatmap provided a visual representation of how various features interrelate with the probability of default.

3.1.6 Loan Amount Distribution

KDE plots of the loan amount (AMT_CREDIT) showed distinct distributions for repaid and defaulted loans, indicating that higher loan amounts have a slightly increased risk of default. This visualization was important because it supported the hypothesis that loan size impacts default risk.

These EDA findings were crucial for understanding the dynamics within the Home Credit dataset and guided the subsequent machine learning model development by highlighting which features should be considered more closely due to their impact on loan default outcomes.

3.2 Model Evaluation Methods

The evaluation methods were carefully chosen to assess each model's accuracy, predictive power, and generalization ability to unseen data. The key model evaluation techniques used are as follows:

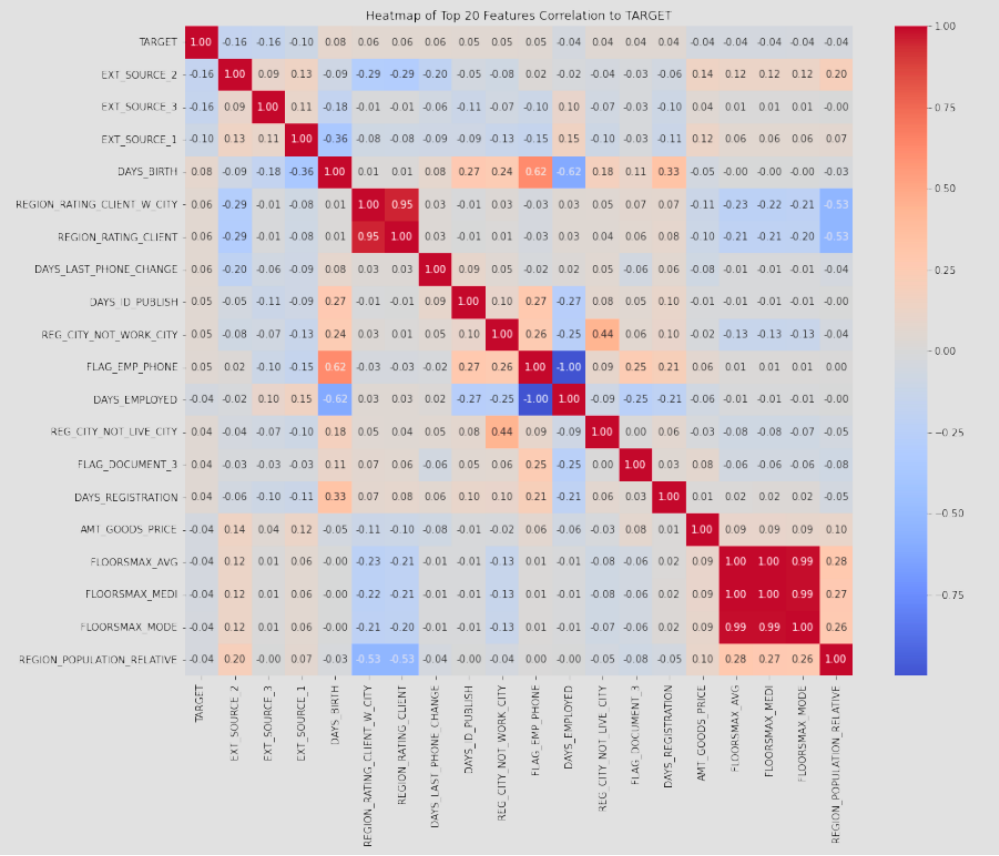

Figure 5: Heatmap of most prevalent features

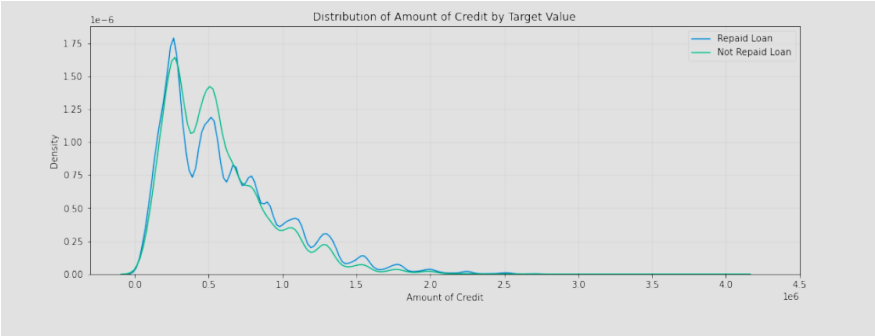

Figure 6: Loan Analysis

3.2.1 Accuracy, Precision, Recall, and F1 Score

- Accuracy measures the overall correctness of the model and is a useful metric when the class distribution is similar. It is calculated as the ratio of correctly predicted observations to the total observations.
- Precision assesses the model's accuracy in predicting positive labels and is particularly important in scenarios where the cost of a false positive is high.
- Recall (or Sensitivity) evaluates how well the model can identify all relevant instances within a dataset.

This metric is crucial when missing a positive (false negative) instance, which carries a greater risk.

- F1 Score is the harmonic mean of precision and recall, providing a balance between the two. It is especially useful when dealing with imbalanced datasets, as it takes both false positives and false negatives into account.

3.2.2 ROC Curve and AUC (Area Under the Curve)

- The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The Area Under the ROC Curve (AUC) measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). It provides an aggregate measure of performance across all possible classification thresholds. An AUC of 0.5 suggests no discrimination (i.e., random chance), while an AUC of 1.0 indicates perfect discrimination.

3.2.3 Cross-Validation

- Cross-validation was used extensively to ensure the models were robust and not overfitting the training data. Typically, 10-fold cross-validation was applied, where the data set was split into ten smaller sets; the model was trained on nine of these and tested on the remaining set. This process was repeated ten times, and each of the subsets was used exactly once as the test data.
- Metrics such as ROC and AUC were used to score each fold, providing insights into how the model's performance might generalize to an independent data set.

3.3 Model Performance

The performance of the ML models in predicting credit defaults was rigorously evaluated using the aforementioned metrics and visualizations.

3.3.1 Logistic Regression

The logistic regression model, serving as the baseline, showed high accuracy due to the imbalanced nature of the dataset. However, the low recall and F1 score indicated its limited capability in correctly identifying default cases, which is critical in practical applications.

Accuracy	Precision	Recall	F1 Score	ROC Score
91.9%	49.6%	1.2%	2.4%	0.7491

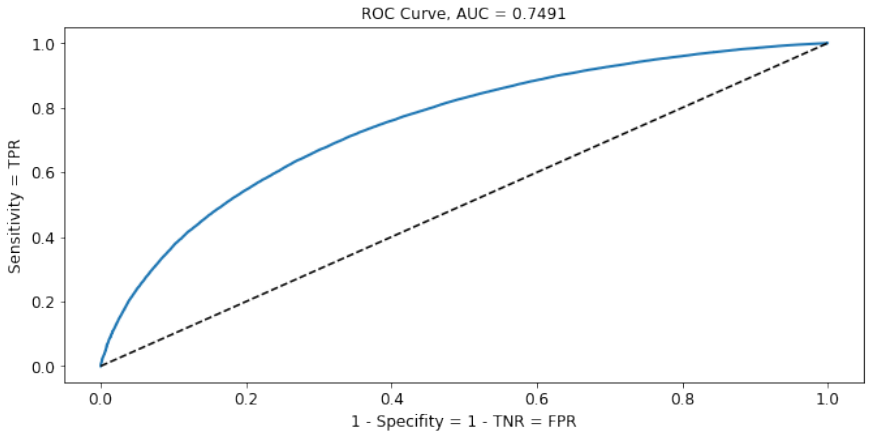

Figure 7: ROC & AUC Curve for Logistic Regression Model

3.3.2 XG Boost

XGBoost performed better than the logistic regression in terms of precision and ability to distinguish between the classes, as evidenced by the ROC and AUC. However, the recall remained low, highlighting challenges in identifying a substantial portion of true defaults.

Accuracy	Precision	Recall	F1 Score	ROC Score
92.2%	78.0%	6.4%	11.8%	0.8412

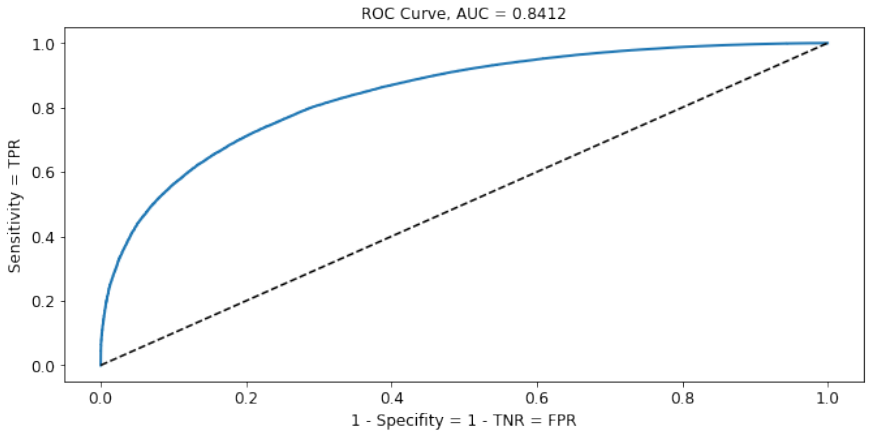

Figure 8: ROC & AUC Curve for XG Boost Model

3.3.3 Random Forest

Random Forest model had high overall accuracy but poor performance in correctly identifying positive instances, as evidenced by low precision and recall scores. Despite a moderately high ROC score, which measures discriminative ability, the model's inability to effectively capture positive instances raises concerns about its reliability, especially when identifying defaults.

Accuracy	Precision	Recall	F1 Score	ROC Score
91.2%	44.4%	2.4%	1.2%	.7317

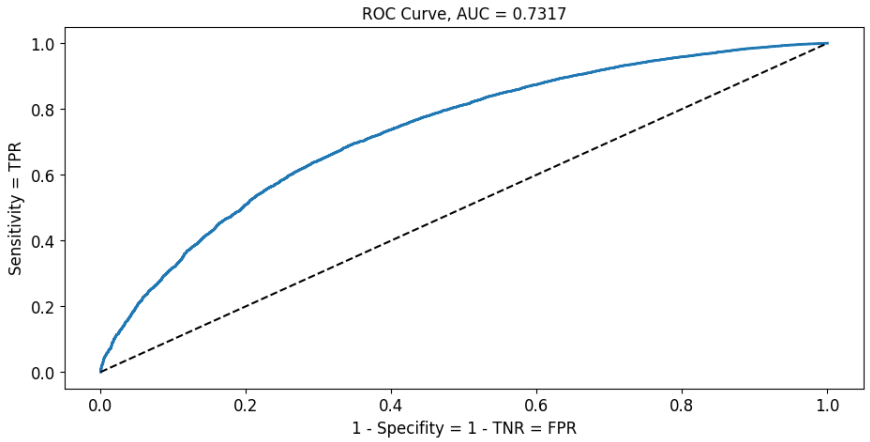

Figure 9: ROC & AUC Curve for Random Forest Model

3.3.4 Neural Networks

Despite a high accuracy, the first neural network struggled with a very low recall rate, similar to logistic regression. Its ability to predict defaults accurately was limited, indicating the need for further optimization of its architecture and training process.

Accuracy	Precision	Recall	F1 Score	ROC Score
91.9%	57.9%	0.9%	1.7%	0.7524

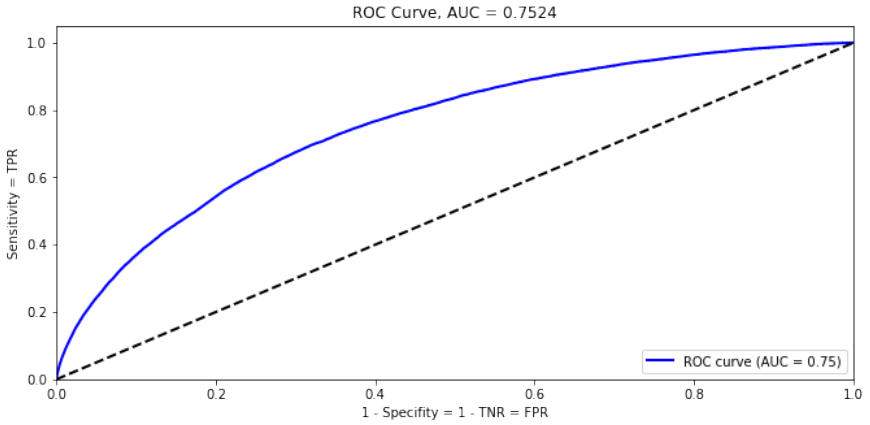

Figure 10: ROC & AUC Curve for Neural Network Model 1

The second neural net which featured the same architecture as the first but added class weights was able to focus more on learning from the default class instances, enhancing its ability to accurately identify potential defaults amidst the majority of non-defaults.

Accuracy	Precision	Recall	F1 Score	ROC Score
72%	16.4%	60.9%	25.9%	0.7328

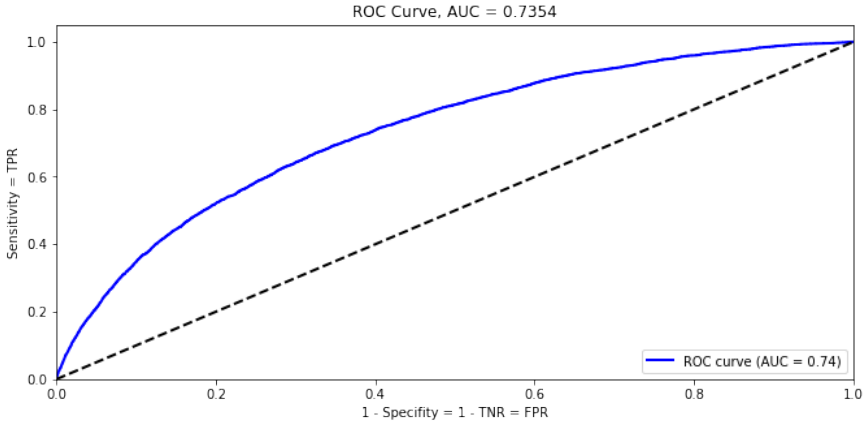

Figure 11: ROC & AUC Curve for Neural Network Model 2

The third neural net results were similar to the second as they employed the class weights but only took 25 features as input. The ability to predict defaults was slightly higher.

Accuracy	Precision	Recall	F1 Score	ROC Score
66.7%	15.1%	67.9%	24.6%	0.7384

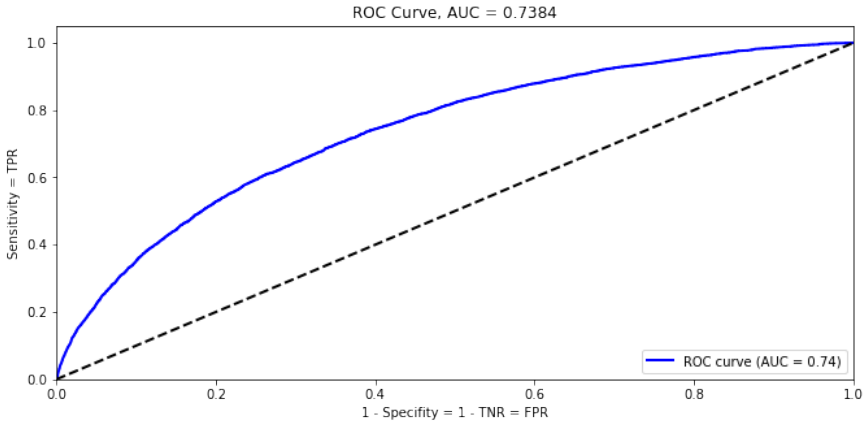

Figure 12: ROC & AUC Curve for Neural Network Model 3

4 Discussion

This paper showcases successfully applying various ML techniques to predict credit defaults, leveraging a comprehensive dataset provided by Home Credit. Our EDA revealed significant trends, such as the higher propensity of younger borrowers and those with lower educational levels to default on loans. These insights are crucial for financial institutions aiming to refine risk assessment methods and improve lending practices.

ML models, including logistic regression, XGBoost, random forest, and neural networks, demonstrated varying degrees of effectiveness in predicting defaults. Our paper highlighted the strengths of each model and their specific applicability to different aspects of the credit default prediction problem. For example, while logistic regression provided a good baseline, advanced models like random forest offered exceptionally high accuracy, albeit with potential overfitting issues.

The implications of these findings are significant. They suggest that incorporating machine learning can greatly enhance the predictive power of credit default models, leading to better risk management and more tailored credit offerings. Furthermore, the nuanced understanding of how demographic factors influence credit risk can help lenders develop more equitable lending practices, potentially reducing the financial vulnerabilities of certain borrower groups.

5 Conclusions

Our paper conclusively demonstrates that machine learning models have substantial potential to enhance credit default predictions, outperforming traditional statistical methods. Each model brought unique benefits to the table, with ensemble methods like random forest showing particularly strong performance metrics, albeit with cautions regarding overfitting.

Moving forward, it is clear that further optimization and validation of these models are required. This involves not only technical improvements, such as adjusting model parameters and exploring new features but also ensuring that the use of machine learning in credit scoring complies with ethical standards and legal regulations. As machine learning continues to evolve, its integration into financial practices promises significant advancements in assessing and managing credit risk, offering both reduced risks for lenders and fairer, more accessible credit for borrowers.

Overall, this research provides a strong foundation for future studies and practical applications, highlighting both the capabilities and the challenges of using machine learning in the financial sector.

6 References

References

[1] Advantages and disadvantages of logistic regression. https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/, 2023.

[2] Five cs of credit - what lenders look for. `https://www.wellsfargo.com/financial-education/credit-management/five-c/`, 2023. Accessed: 2024-02-08.

[3] Andrés Alonso and Jose Manuel Carbo. Understanding the performance of machine learning models to predict credit default: a novel approach for supervisory evaluation. 2021.

[4] Analytics Vidhya. An end-to-end guide to understand the math behind xgboost. `https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/`, 2018. Accessed: 2024-02-09.

[5] Bart Baesens, Daniel Roesch, and Harald Scheule. *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley & Sons, Hoboken, 2016.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[7] Wei Chen, Yang Li, Weifeng Xue, Himan Shahabi, Shaojun Li, Haoyuan Hong, Xiaojing Wang, Huiyuan Bian, Shuai Zhang, Biswajeet Pradhan, et al. Modeling flood susceptibility using data-driven approaches of naïve bayes tree, alternating decision tree, and random forest methods. *Science of The Total Environment*, 701:134979, 2020.

[8] Jonathan N Crook, David B Edelman, and Lyn C Thomas. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465, 2007.

[9] Federal Reserve Bank of New York. Household debt and credit report, 2023. Accessed: [4/23/2002].

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[11] J.T. Han, J.S. Choi, M.J. Kim, et al. Developing a risk group predictive model for korean students falling into bad debt. *Asian Economic Journal*, 32:3–14, 2018.

[12] IBM. Random forest. `https://www.ibm.com/topics/random-forest`, 2023. Accessed: 2024-02-09.

[13] LendingTree. Credit card debt statistics, 2023. Accessed: [2/23/2002].

[14] X. H. Li. Using "random forest" for classification and regression. *Chinese Journal of Applied Entomology*, 50:1190–1197, 2013.

[15] Mehul Madaan, Aniket Kumar, Chirag Keshri, Rachna Jain, and Preeti Nagrath. Loan default prediction using decision trees and random forest: A comparative study. In *IOP Conference Series: Materials Science and Engineering*, volume 1022, page 012042. IOP Publishing, 2021.

[16] M. Malekipirbazari and V. Aksakalli. Risk assessment in social lending via random forests. *Expert Systems with Applications*, 42:4621–4631, 2015.

[17] J. Nalić and A. Švraka. Using data mining approaches to build credit scoring model: Case study—implementation of credit scoring model in microfinance institution. In *2018 17th International Symposium Infoteh-Jahorina (INFOTEH)*, pages 1–5, East Sarajevo, 2018. IEEE.

[18] Y. Rachmansyah, A. Dorkas R.A, Harijono Harijono, and R. Prabowo. The determinants of home mortgage default probability: The effect of loan and borrower's characteristics. EAI, 2 2021.

[19] Spiceworks. What is logistic regression? https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/, 2024. Accessed: 2024-02-09.

[20] Y. Wang and J.L. Priestley. Binary classification on past due of service accounts using logistic regression and decision tree. Grey Literature from PhD Candidates, 2017.

[21] Weiguo Zhang, Chao Wang, Yue Zhang, and Junbo Wang. Credit risk evaluation model with textual features from loan descriptions for p2p lending. *Electronic Commerce Research and Applications*, 42:100989, 2020.

7 Code Appendix

Code for the paper is available at our GitHub repository: https://github.com/tomoc99/DAT490_Capstone/tree/main