Predicting Credit Risk and Building a Credit Risk Score to Detect Fraud

H.C.D Hapuarachchi
University of Sri Jayawardenapura
hchamikadilshan@gmail.com

P.R.Annassiwatte
University of Sri Jayawardenapura
annasiwatte55@gmail.com

Introduction

Lending is the backbone of the financial industry, but it comes with risks, especially fraudulent activity. Fraud occurs when individuals exploit lending systems to obtain credit without intent to repay, leading to significant financial losses for institutions. A key indicator of this risk is the "charge-off," where loans are written off as uncollectible, often signaling intentional fraud.

This poster presents a predictive model designed to identify high-risk customers and detect potential fraud by assigning a credit risk score. The model helps banks reduce losses, improve decision-making, and ensure regulatory compliance, while maintaining transparency and explainability in its predictions

Data Preprocessing

The dataset consisted of records or 7000 customers, described by 23 variables , each reflecting financial and behavioral data related to their credit history. Dateset was preprocessed to handle missing and out-of-range values before analysis

Unusual Submission Pattern (910 missing values)

Imputed using a **XGBoostClassifier** model, by feature '***multiple_applications_short_time_period***', '***applications_submitted_during_odd_hours***', and '***payment_methods_high_risk***'

Number of Delinquent Accounts (700 missing values)

Imputed using a **XGBoostRegressor** model, by feature '***number_of_defaulted_accounts***', and '***delinquency_status***'

FICO Score (210 missing values)

missing FICO scores were **removed**, as these were limited in number and provided by an external organization

Average Balance Last 12 Months (350 missing values)

Imputed using the ***median***, as the distribution of this feature was slightly skewed, and the median was a robust choice to avoid distortion

FICO Score (42 out of range values)

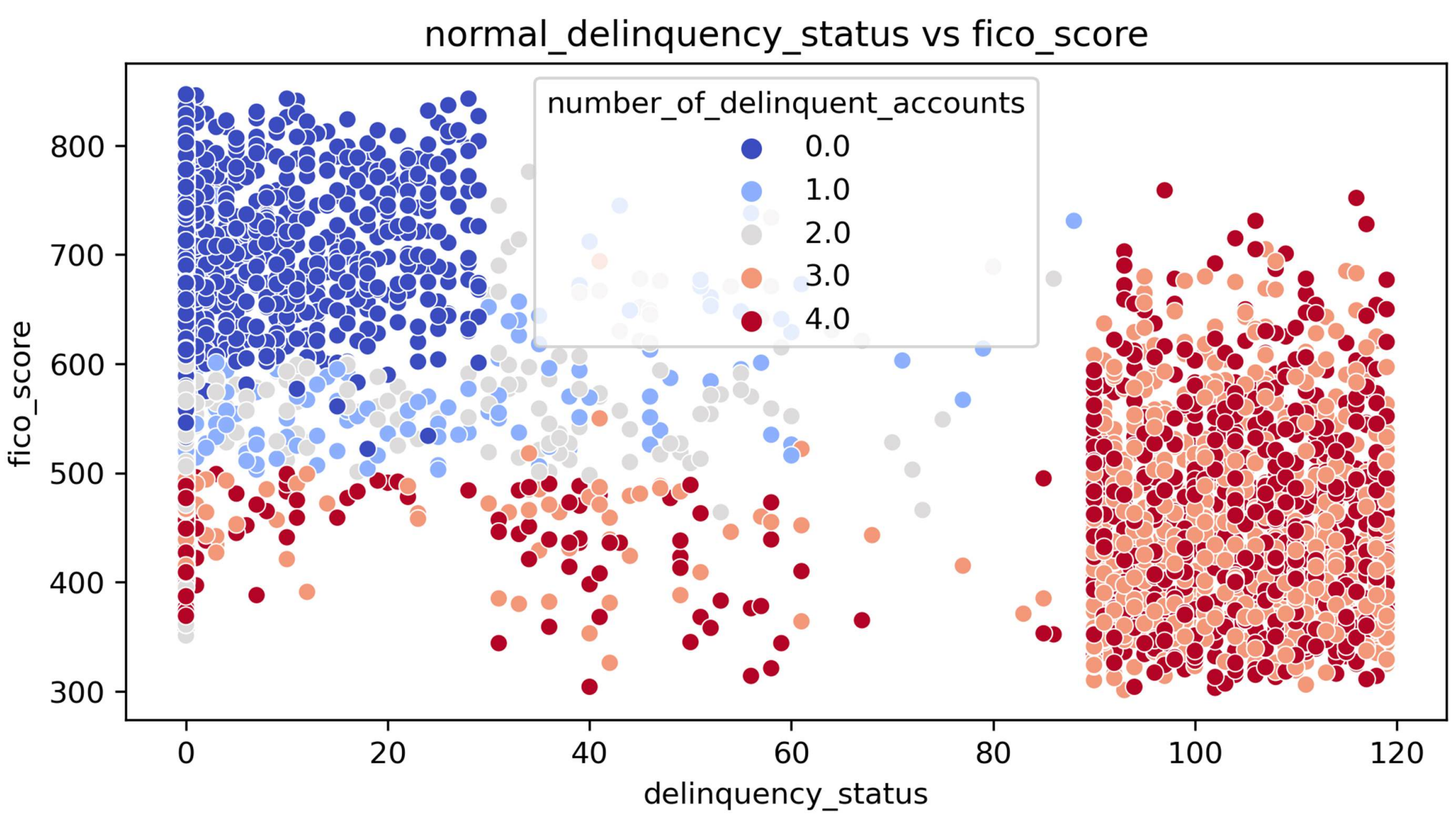
These values were out of the range 300 - 850.
These values were removed

Methodology

Exploratary Data Analysis

We conducted a comprehensive exploratory data analysis (EDA) to examine the relationships between various predictor variables and the target variable, ***Charge-Off Status***. The following steps were undertaken

- We began by visualizing the distribution of each individual variable using histograms, bar charts, and boxplots. This helped in understanding the central tendency, spread, and presence of extreme values in the data
- Next, we compared pairs of variables using stacked bar charts, clustered bar graphs, and scatter plots. This allowed us to identify potential interactions and correlations between the predictors
- To understand the relationship between each predictor variable and the target, Charge-Off Status, we visualized the data using techniques such as bar charts and scatter plots, focusing on the impact of each predictor on the outcome
- Key Features identified in the EDA
  - FICO Score
  - Number of Delinquent Accounts
  - Delinquency_status
  - Debt-to-Income Ratio



Feature Engineering

- Using the date type features("*Account_open_date*", *"Earliest_credit_account"*, ""*Recent_trade_activity*") , new features were derived by counting the number of days from the specific date to current date
- Categorical features ("location" and "Occupation") were encoded using One-Hot encoding

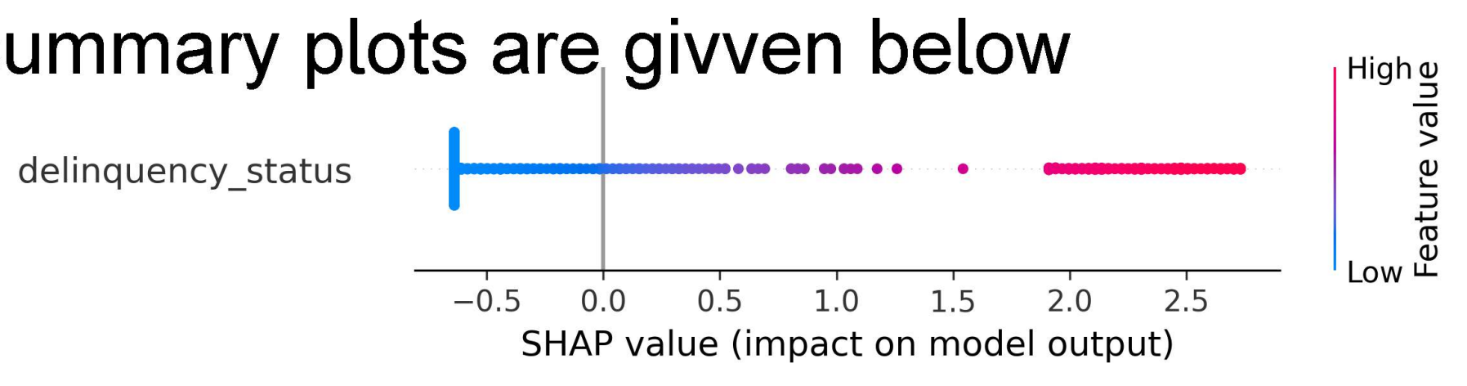
Model Building & Selection

Two models were fitted (XGBoost and Logistic Regression). For the XGBoost model GridSearchCV hyperparamter tuning was performed

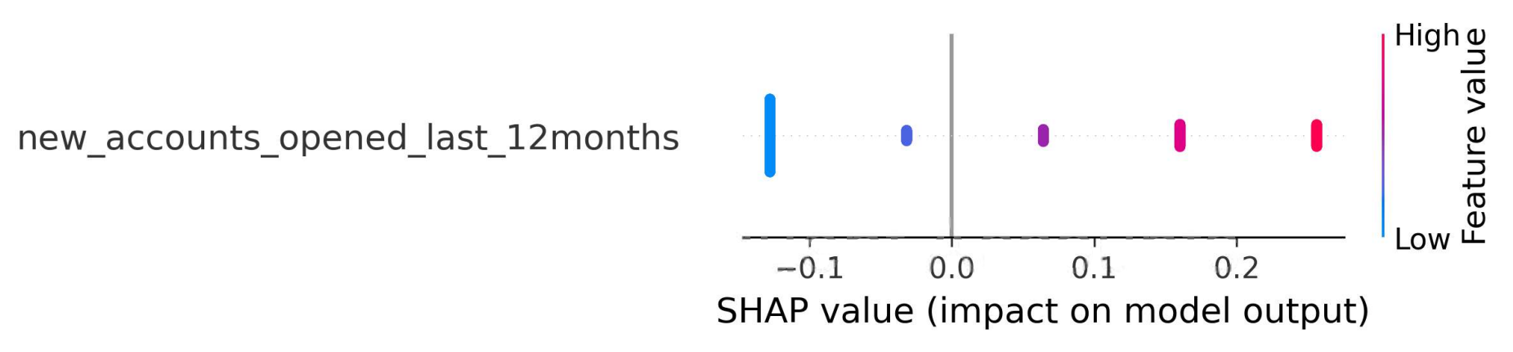
Metric	**Logistic Regression**	**XGBoost**
Accuracy	0.8674	0.8644
ROC-AUC	0.8261	0.8335
Precision (Charged-off)	0.76	0.75
Recall (Charged-off)	0.68	0.69
F1-Score (Charged-off)	0.72	0.72
Weighted Avg Recall	0.87	0.86
Weighted Avg F1-Score	0.86	0.86

Results

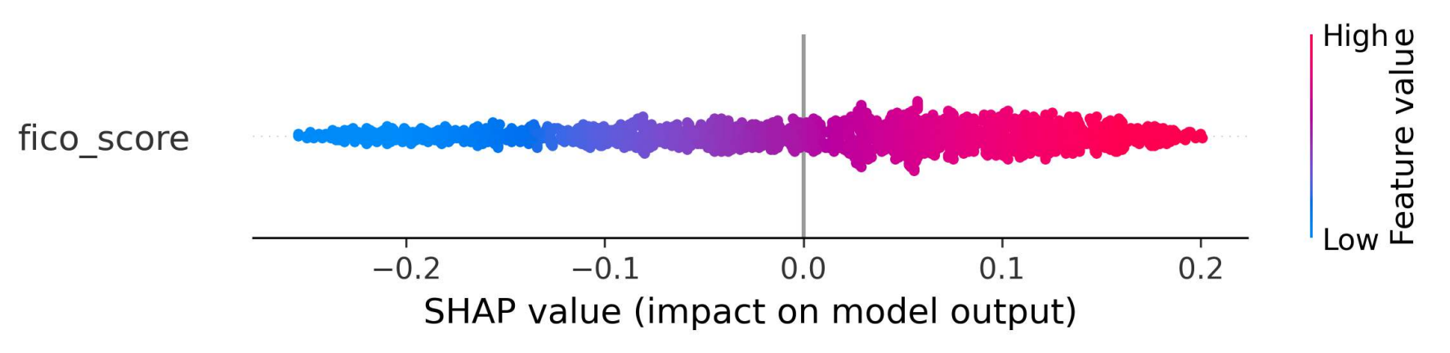
Since the both models' performances are similar. We considerd model explanability when choosing the model. As a result **Logistic Regression Model** was selected. In this model following 4 features were identified as cruicial in predicting. SHAP summary plots are givven below



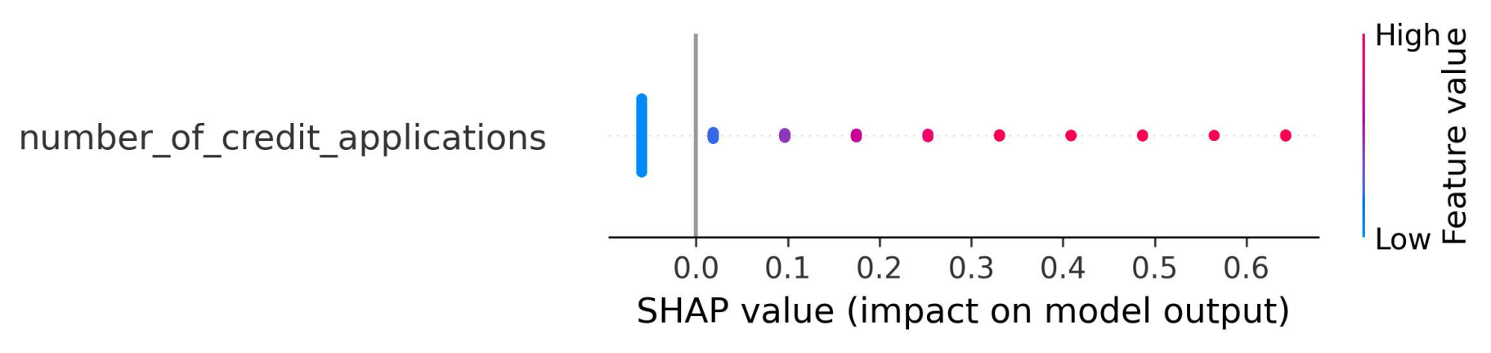
1 - Delinquency_status



2 - new_accounts_opened_last_12months



3 - fico_score



4 - number_of_credit_applications

Discussion

We encountered challenges with missing values and class imbalance. Missing values in critical variables required careful imputation or removal to ensure data integrity. The class imbalance between charge-off and non-charge-off customers affected model accuracy, and adjusting class weights helped mitigate this issue. Future work may explore more advanced techniques for imputation and balancing to improve model performance.

Conclusion

In conclusion, this study developed a predictive model for credit risk and fraud detection, identifying key features like delinquency status, FICO score, new accounts opened, and credit application frequency. Despite challenges with missing data and class imbalance, the model effectively predicted charge-off status, helping financial institutions make informed decisions. Future work could focus on refining data imputation and balancing techniques to improve model performance.