Author: Fejiro Obor
BrainStation Capstone Report
Data Science Diploma Program Jan - Sept 2023
oobor@hotmail.com
www.linkedin.com/in/fejiro-obor/
github.com/FejiroObor/

Loan Default Prediction

Executive Summary

This report presents the summary of my capstone project which was aimed at using machine learning techniques to predict loan default.

Problem Statement

In the financial industry, accurate loan default prediction is a paramount concern for lenders and financial institutions. According to StatsCan, Credit card debt in Canada rose by 21.9% from $75 billion in January 2021 to $91.5 billion in December 2022[1]. By providing financing opportunities to clients, lenders are always at risk of losing their principal in addition to expected profits (interests). To maximize profits, such businesses must minimize their losses by identifying borrowers who are at a higher risk of defaulting on their loan payments. The objective of my capstone project is to develop predictive models that can accurately differentiate between borrowers likely to repay their loans versus those more prone to default. This predictive capability will empower lenders to make more informed decisions regarding loan approvals and improve their risk management strategies.

Dataset

Loan default prediction is usually a complex task that involves analyzing various factors such as borrower's income, current debt, credit history, past delinquencies, etc. The dataset used for modeling in this project was obtained from Kaggle[2]. The publicly available dataset was provided by Lending Club, which is a US-based peer-to-peer lending platform. It ceased to exist as of October 2020 but prior to that, it operated as an online marketplace that connected borrowers seeking personal loans with investors looking to lend some money for potential returns.

Data Preprocessing

The initial dataset contained 396,030 entries and 27 features which are explained in the attached data dictionary file. Each entry corresponds to an individual loan application completed on the Lending Club platform and the target column was labeled (charged off vs. fully repaid loans). The dataset was moderately imbalanced considering it contained 20% default vs. 80% repaid loans. Although the dataset contained no duplicate entries, 6 columns initially had null entries which accounted for 20.6% of the total observations (81,589 rows). As part of the data

preprocessing steps, null values were filled in and outliers were removed. New features were created to enhance the model's performance and the relevant predictive features were narrowed down using statistics and hypothesis testing. Lastly, features with high multicollinearity were excluded to avoid creating unstable models.

EDA and Feature Selection

Figure 1 below presents some insights into the data exploration conducted. As seen below, the dataset contained only 20% of loan defaults. However, the amount of clients who defaulted on their loans increased as grade type changed from A to G.

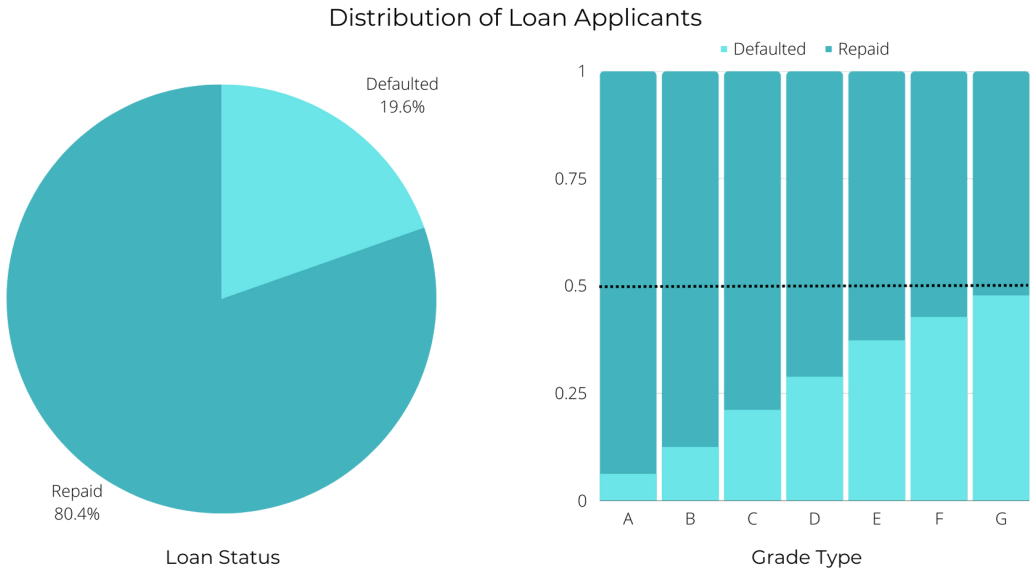

Figure 1: Loan default distribution

Figure 2 below shows the correlations of the numeric features with the loan outcome. We can see that clients who defaulted on their loans had higher average loan amounts, interest rates, number of installments, debt-to-income ratio, open accounts, revolving balance, revolving utilization, public bankruptcies, and public derogatory records overall. On the other hand, they had lower average annual income, mortgage accounts, total number of accounts overall, and credit history length.

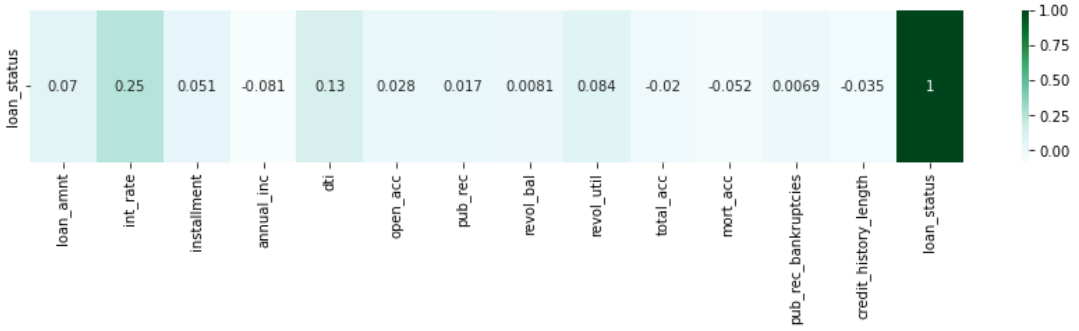

Figure 2: Correlations of dependent variables with the target (loan_status)

Model Selection and Evaluation

To predict customers most likely to default on their loans, I employed four machine learning algorithms including Logistic Regression, Decision Tree, Random Forest, and XGBOOST. Due to the moderate class imbalance in the target variable, I focused on being able to strike an ideal balance between identifying true defaulters without misclassifying too many credit-worthy applicants. I compared the precision, recall and F1 scores while optimizing for AUC. Figure 3 below presents the baseline model performance while Figure 4 compares with the optimized model performance:

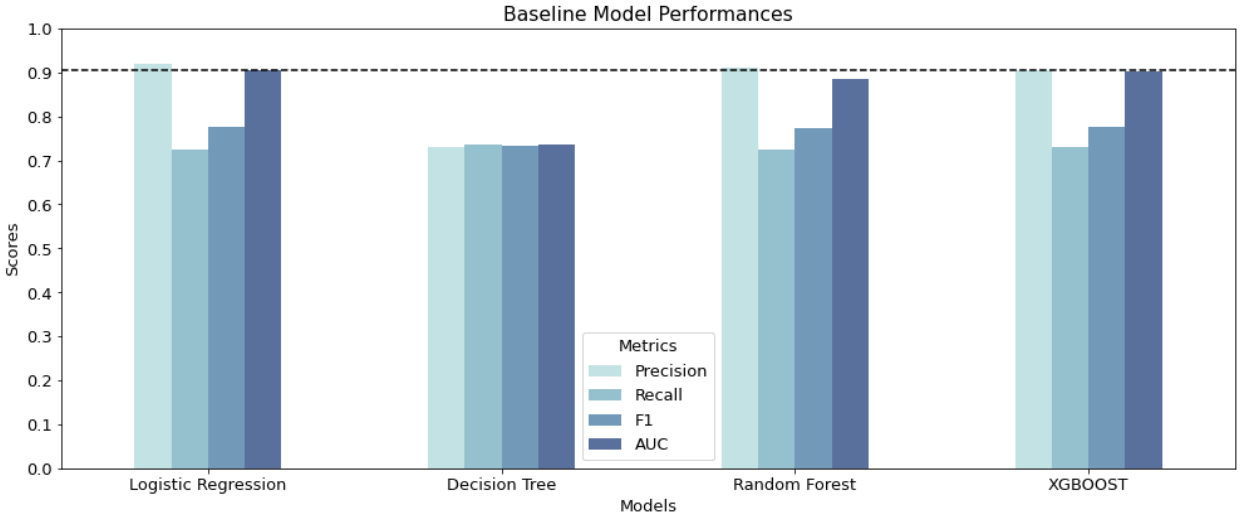

Figure 3: Baseline model performance

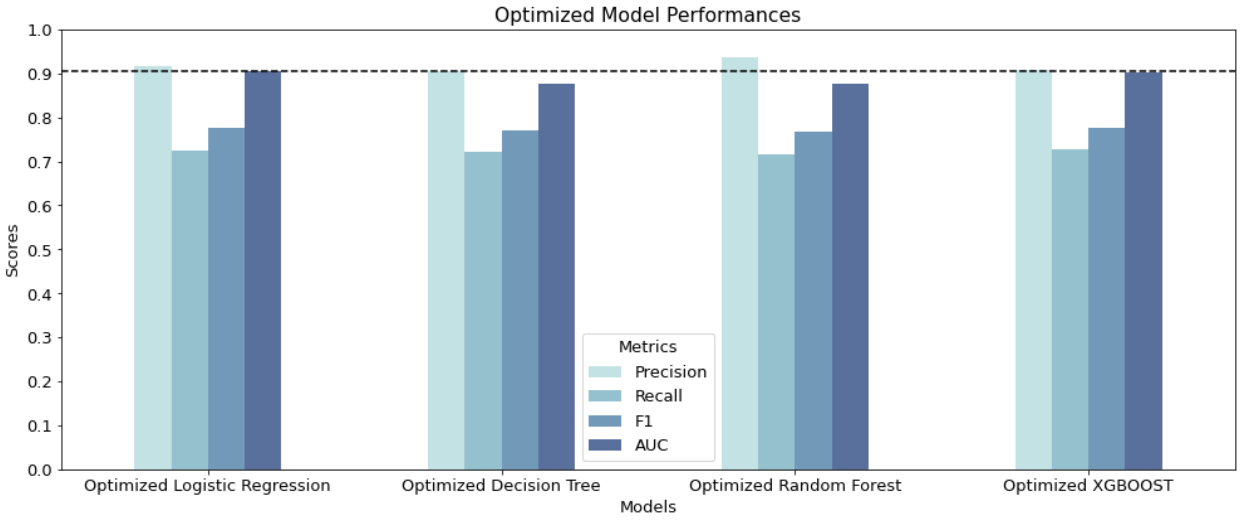

Figure 4: Optimized model performance

The best model after optimization was the Logistic Regression model which outperformed the others with the highest AUC of 0.9043. The model is doing very well in segregating clients into class default vs. repaid with 92% average precision. With such high precision, we can be confident that individuals will default or not based on the model's prediction.

Conclusion

In conclusion, four machine learning models were applied to the Lending Club dataset with the goal of accurately predicting which clients will most likely to default on their loans. These models include Logistic Regression, Decision Tree, Random Forest, and XGBOOST. The Logistic regression model provided the best performance with a ROC_AUC score of 0.9043. It identifies the true loan defaulters with 96% precision while having 88% precision on the loan repaid class.

I believe this model can be very useful for a wide range of financial institutions including banks, peer-to-peer lenders, credit rating agencies, insurance companies, etc. These companies can use this tool to improve their credit risk management practices while maximizing their profits. As part of the next steps of this project, I plan to deploy the Logistic Regression model using Streamlit.

References

1. https://www.statcan.gc.ca/o1/en/plus/3222-traditional-credit-card-debt-hangover-following-holidays-or-something-more-ominous
2. https://www.kaggle.com/datasets/epsilon22/lending-club-loan-two