Credit Card Approval Modeling Report

1. Executive Summary

This report delves into the development and evaluation of models for credit card approval using historical data from regional bank XYZ. Logistic regression and random forest and XGBoosting algorithms were implemented, and their performance was assessed to aid in the decision-making process for credit approval.

2. Baseline Model

2.1. Simple data pre-processing

The training and validation datasets, comprising 20,000 and 3,000 accounts respectively, were loaded and divided into features and target variables. Since we had a huge amount of data to work with, all of the rows that contained NaN values were removed.

2.2. Baseline Logistic Regression Model

The logistic regression model achieved an accuracy of 93% on the validation set, indicating a reliable performance in classifying credit card applications. When it came to a AuROC of 0.58 we observe the main weakness of this model. It has highly incorrect predictions for the minority class of our data.

3. Building Comparison Models

3.1. Further data pre-processing for Comparison Models

Used various plots and Chi-square test to arrive at the conclusion of excluding the "States" column due to lack of useful information. For the numerical columns, removed highly correlated columns (corr() value greater than 0.9) to remove highly proportional columns to reduce spam data. The data is also standardized for easy

3.2. Why this data is bad?

After the baseline model statistics it is easy to observe that there is plenty of data for the "no default" part of the predictions while there is significantly lesser data for "default". This is the real-world scenario and hence we need further steps to mitigate the data imbalance. One of the common methods adopted for this is the SMOTE technique.

3.3. SMOTE:

SMOTE, or Synthetic Minority Over-sampling Technique, is a resampling technique commonly used in machine learning to address class imbalance in datasets. SMOTE works by creating synthetic examples of the minority class, thereby balancing the class distribution. It does this by generating synthetic instances along the line segments joining existing minority class instances. We have used smote to repopulate data for the Logistic regression model for comparison.

3.4. Detailed Comparison between Models used:

Logistic Regression: Logistic Regression is a statistical model suitable for binary classification tasks.
Random Forest: Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.
XGBoost: XGBoost (Extreme Gradient Boosting) is a scalable and efficient gradient boosting framework that utilizes a combination of additive models and regularization to enhance the predictive performance of decision trees, making it a powerful algorithm for classification and regression tasks.

Comparison between the XGBoost and LogisticRegression:

Accuracy: XGBoost outperforms Logistic Regression in accuracy, suggesting it makes correct predictions for a higher proportion of instances.
AUROC: XGBoost also achieves a higher AUROC, indicating better overall discrimination between positive and negative instances.
Precision and Recall: While XGBoost shows higher precision for Class 1 (47% vs. 27%), indicating a better ability to correctly identify positive instances, recall remains the same at 57%.

3.5. Comparison Summary

Model Name	Model Accuracy	Model AuROC	Speed of Testing and Prediction
Baseline Model	0.9332794177112819	0.584374925430119	Both are very fast
Linear Regression	0.8552365547917509	0.810788173813444	Both are very fast
Random Forest	0.9413667610190053	0.83105447777221	Very slow training and slow testing
XGBoost	0.9276182773958754	0.8669840360798913	Slow training and fast prediction

4. Performance Metrics

4.1. AuROC

Area Under the Receiver Operating Characteristic curve is a performance metric for binary classification models, quantifying their ability to discriminate between classes across various classification thresholds. AUROC considers the performance of a classifier across various classification thresholds, providing a comprehensive assessment of the model's ability to distinguish between positive and negative instances without being tied to a specific decision threshold and hence was chosen as the primary metric.

4.2. Accuracy

Accuracy was chosen as the primary metric for model evaluation. It represents the ratio of correctly predicted instances to the total instances. It is a non-negotiable metric in evaluating any prediction model.

4.3. Precision, Recall, and F1-score

Precision, recall, and F1-score provide a more nuanced evaluation, especially when dealing with imbalanced datasets and hence were chosen as the tertiary deep dive metrics for the above models.

5. Final Suggestion

Finally, we arrive at the conclusion that XGBoost gives us the best result in terms of speed and accuracy among all the models developed XGBoost provides the most accurate predictions on unseen data and hence might be the most appropriate model for our usecase.

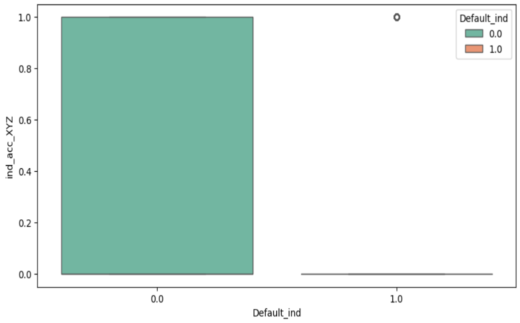
6. Answering Unanswered Questions:

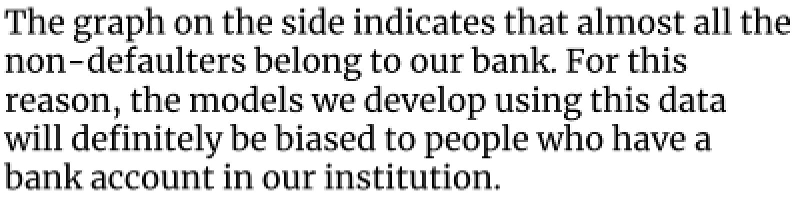
- Describe how you would use it to make decisions on future credit card applications.

Answer: The XGBoost model suggests a probability of default that is further used to make our prediction for a default. The model can be tuned to be more forgiving about the prediction of a default but it is currently configured to be unbiased. Further based on default prediction, we make a decision on credit card application.

- Do customers who already have an account with the financial institution receive any favorable treatment in your model? Support your answer with appropriate analysis.

Answer:




The graph on the side indicates that almost all the non-defaulters belong to our bank. For this reason, the models we develop using this data will definitely be biased to people who have a bank account in our institution.