Machine Learning Engineer Nanodegree
Capstone Proposal

Credit Scoring

by Deviyanti Aryani Mariam
November 2021

Domain Background

As technology is evolving so fast that it is getting to the point where it is able to create a new digital revolution in the finance sector by providing access to more people, lending companies face a challenge in identifying customers' creditworthiness, the ability of the customers to pay back. This project are meant to help company "X" become more efficient and effective in making decision and providing financial services to its customers. The credit scoring model will be used as a tool for accepting or rejecting a loan application and it will be also for determining the cutoff of the credit score as the company tries to satisfy their expected approval rate (the number of loans will be approved out of the number of application) and default rate (the number of non performing loans out of the number of approved loans). Machine learning would be the approach used for the credit scoring. The model defines customer scoring based on past borrowers' characteristics.

Problem Statement

The upper management of company "X" wants the overall default rate of their portfolio to be below 3% while the approval rate is not less than 70%. Below is the main objective of the project:

- Build a model that predicts whether a loan would be default or not, the result would be the predicted probability of default which indicates a customer is unlikely to pay
- Provide recommendations on the optimal credit score cutoff.

Datasets and Inputs

Dataset can be accessed at
https://github.com/deviyantiam/credit_scoring/blob/main/data.csv
It is artificially created based on the real data to replace privacy sensitive information.

Feature	Description
x	user_id
number_of_cards	number of cards owned by customer
outstanding	total outstanding amount of credit card usage
credit_limit	credit limit amount that can be used

bill	last month customer bill amount
total_cash_usage	last month total cash usage of customer
total_retail_usage	last month total retail usage of customer
remaining_bill	remaining bill that has not been paid in the last month
branch_code	branch code
payment_ratio	payment per bill ratio in the last month
overlimit_percentage	overlimit percentage
payment_ratio_3month	payment per bill ratio in the last 3 month
payment_ratio_6month	payment per bill ratio in the last 6 month
delinquency_score	delinquency score
years_since_card_issuing	total year since first card issued
total_usage	total usage
remaining_bill_per_number_of_cards	ratio remaining bill per number of cards
remaining_bill_per_limit	ratio remaining bill per credit limit
total_usage_per_limit	ratio total usage per limit
total_3mo_usage_per_limit	ratio total 3 months usage per limit
total_6mo_usage_per_limit	ratio total 3 months usage per limit
utilization_3month	Credit card utilization for past 3 months
utilization_6month	Credit card utilization for past 6 months
default_flag	Credit default flag (1: default; 0: non_default)

Solution Statement

The proposed solution is to build machine learning classifiers. We would do visualizations and feature engineering. As it takes a lot of iterative work, we also try to select multiple feature combination to be fed into the algorithms and do hyperparemeter tuning and choose the model with best evaluation metric value and the model should satisfy the requirements from management. We also will provide the credit scorecard that contains information about score bins, its approval rate and default rate. The scorecard enables the management to accept a certain number of loans based on their willingness to take on risks.

Benchmark Model

We will build a baseline model to which the actual model's performance will be compared.
We will use Logistic Regression.

Evaluation Metrics

The model performance will be graded based on the area-under-the-ROC-curve score (AUC) or GINI (2*AUC-1). The Receiving Operating Characteristic (ROC) curve is a graphical plot of the True Positive Rate (percentage of bads rejected) versus the False Positive Rate (percentage of goods rejected) for every threshold or cut-off (see **Figure 1**). The Area under the Curve (AUC) measures the percentage of the area that is under this curve. The higher the percentage, the better the model performance. Gini is a normalised or linear transformation of AUC, a random classifier scores 0, and a perfect classifier scores 1.

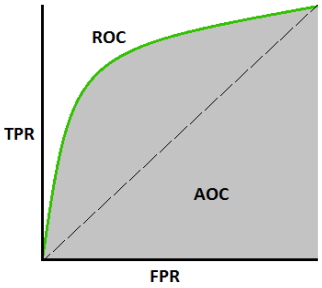

Figure 1 ROC-AUC

Project Design

- Data Analysis. It will provide a better understanding of dataset variables and the relationships between them.
- Data Prepocessing. We will perform some of data processing as we see fit, that could be import all the crucial libraries and dataset, identifying and handling the missing values, encoding the categorical data, splitting the dataset, feature scaling
- Modeling and Evaluation. We will train classification models and calculate the AUC. After choosing one model, we calculate the approval rate and default rate. We also perform hyperparameter tuning if necessary or even go back to the feature engineering step until we can get the model that meets the requirements.
- Credit scorecard creation. Once the final model is chosen, we also create scorecard to help non-technical team understand the model usage.

Reference

- https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03
- https://medium.com/henry-jia/how-to-score-your-credit-1c08dd73e2ed
- https://towardsdatascience.com/machine-learning-gridsearchcv-randomizedsearchcv-d36b89231b10