Concordia University

Proposal

SOEN6111

Credit Card Score Forecaster with Supervised Learning

Author:	*Author ID:*
Somayeh Ghahary	40106359
Mehrnoosh Amjadi	40091264

Submitted to:
Dr. Tristan Glatard

February 19, 2021

1 Abstract

Credit card is one of the most crucial services that banks offer to their customers. However, if customers cannot pay their statements on time, it might have negative consequences for the financial industry. To tackle this challenge, in this project, we will propose a novel system that can predict the score of the credit cards of the customers in prior to issuing the card to see if clients would be able to pay their credit cards on time or not. We perform supervised learning methods on customers datasets and operate our analysis based on some powerful Machine Learning algorithms including Neural Networks, Logistic Regression, Random Forest, Decision Tree, etc. Also, we might take advantage of more models or other datasets due to the progress of our project.

2 Introduction

With the unprecedented growth of the data in financial industry, banks are interested to have a automated mechanism to decide about issuing credit cards to their costumers from a lot of applications that they have received[1]. It is vital for these financial institutes to make sure that customers will pay statements of their credit cards on-time. Although many literature and competitions have valuable results in this field[3][1], we want to propose a novel idea for forecasting the credit card scores by feature engineering.

Machine learning models facilitate the process of predicting the score of credit cards in compare to traditional methods. The primary goal of this project is to apply Machine Learning methods on the chosen datasets to implement a supervised learning for predicting the score of credit card of customers according to their historical and financial records. Based on the result of the prediction, banks and financial institutes would be able to decide to accept or reject the credit card application of a customer by categorizing that customer to a ”good” or ”bad” client. This project comprises different phases:

- Data preprocessing: including handling missing values, removing unnecessary features, adding new features, visualizing the data to better understand them, splitting the data into train, validation and test sets, and labeling targets.
- Training the models: including selecting the appropriate algorithms (since this project’s objective is to classify the customers based on the historical data, the Machine Learning algorithms which are suitable for classification problem must be chosen), using Hyper parameter Tuning technique to get good performance about our models.
- Evaluating the models based on the test set: predicting the test set based on models trained on previous phase. Calculating the accuracy of the models and comparing them to understand which models perform better. We can also use Confusion matrix to visualize the performance of the models.

In the following, the details of the progress of the project are described.

3 Materials and Methods

3.1 Dataset

We aim to utilize two datasets provided by Kaggle (Application Record and Bank Churners) with having about 440,000 and 10,000 records respectively. This datasets contain heterogeneous data

Dataset Name and Link	*Source*	*Record Number*	*Columns*
Bank Churners	Kaggle	Over 10,000	Customer Age, Gender, Dependent, Card Category, Credit Limit, Income Category, Marital Stats, etc
Application Record	Kaggle	About 440,000	Income Type, Family Status, Occupation Type, Housing Type, Education Type, etc

Table 1: Data-sets Table

(tabular, textual, categorical, numeric, etc), which might have missing values. Data preprocessing is the first step of our project to improve the data quality, which can directly affect our model's ability to learn. Handling the missing values and normalizing the features are some of the tasks in this phase. In the next step, the dataset is sliced randomly into train, validation and test set with the ratio of 60/20/20 normally. The train dataset is used for training phase, while the test set is used later for the prediction purpose.

3.2 Technologies

In the progress of the project, we use the following tools and technologies for the project implementation. However, we might update this list during the progress of the project:

- *Programming Language and IDEs:* Python3, Jupyter Notebook, Anaconda
- *Numerical Analysis:* Numpy, Pandas, Pytorch and Tensor Flow
- *Machine Learning Packages:* Scikit-learn library [2]
- *Visualization and Plotting:* Matplotlib and Seaborn

3.3 Algorithms

Beside Logistic Regression classifier that is a popular methods for score predicting, we utilize Neural Networks, Gradient Boost Regressor and its variant, Decision Tree, Random Forest, Support Vector Machines models, etc. Finally we compare the performance of applied models.

To achieve a successful score prediction we follow the following steps. First in the preprocessoing, we extract new features, merge attributes, handle null values and remove irrelevant features. We might normalize some attributes and scale raw data to make them suitable for training. Then in feature engineering phase, we add new features, label targets and split data to train, validate and test datasets. Next steps are training different models and predicting. In the final step, we analyze and evaluate the results for instance by false positive or true negative techniques. We visualize data during progress of the project using bar charts, box plots, line charts, confusion matrix and other methods to get a better sense of understanding.

References

[1] Omareltouny. Credit card approval, Aug 2020.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Rikdifos. Credit card approval prediction using ml, Nov 2020.