Credit Default Classification and Customer Segmentation

Student Researcher: Yuxuan Li, ‘24 B&F Data Science, NYU Shanghai
Faculty Mentor: Junpei Komiyama, Assistant Professor of Operations and Statistics, NYU Stern School of Business

Introduction

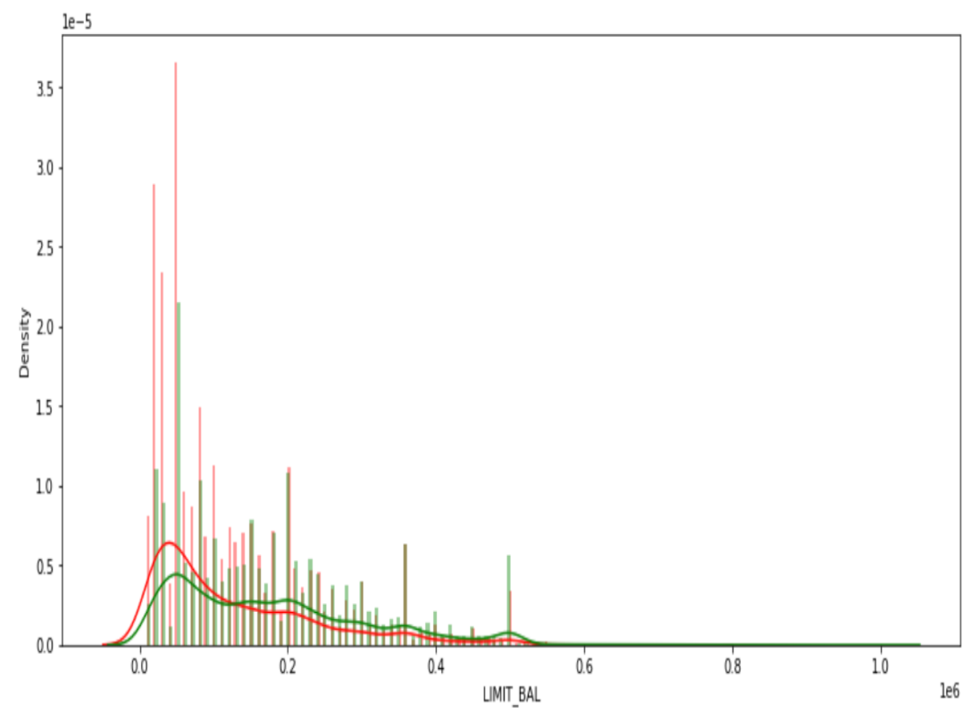
- This project is separated into three parts to get a full understanding of the credit card default dataset. Credit card default happens when you have become severely delinquent on your credit card payment. In order to increase market share, card-issuing banks in Taiwan over-issued credit cards to unqualified applicants. At the same time, most cardholders, irrespective of their repayment ability, overused credit card for consumption and accumulated heavy credit bills.
- The clustering is to build a model to generally categorize the risk of the repayment ability of the credit card holders considering various metrics including credit history and other available personal information. This applies an unsupervised learning process that customers are classified based on their risk level of default.
- The goal for machine learning algorithms constructed is to build an automated model for both identifying the key factors and predicting the credit card default based on the information about the client and historical transactions. In particular, Random Forest, Gradient Boosting Algorithm, and Ensembled models have been used.
- Time series model (Long Short-Term Memory Networks) analysis is applied to leverage time-related credit default dataset. This model is more robust than traditional non-time dependency models for complicated real-world data and applications.

Exploratory Data Analysis

The dataset consists of 30000 observations that represent distinct credit card clients. Each observation has 24 attributes that contain information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

The main aim of the data is to discriminate clients that are predicted to credit card default the next month, which is set to “0” for non-defaulters and “1” for defaulters. A number of 6636 out of 30000 (22.1%) of clients will default next month.

For primary investigations, most of defaulters are for credit limits 0-100,000 and density for this interval is larger for defaults than for non-defaults (Figure 1). Besides, correlation among features can also affect the model performances. As from the correlation matrix (Figure 2), some features show high correlations with each other (e.g. series of BILL_AMT features). This observation leads to dimensionality reduction techniques to remove correlated features.



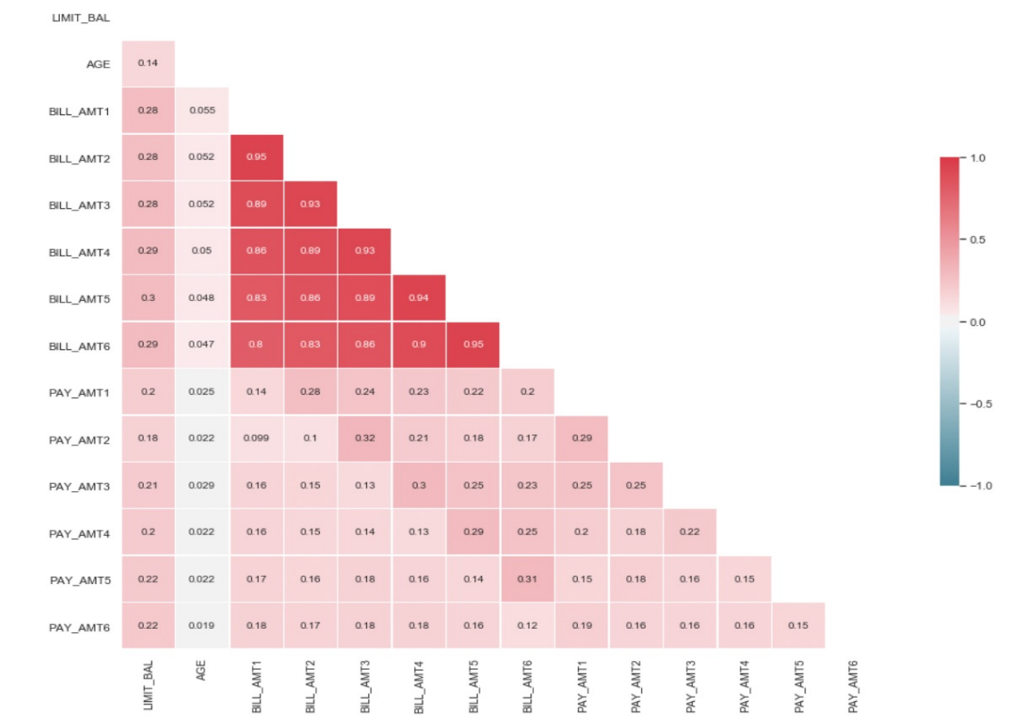
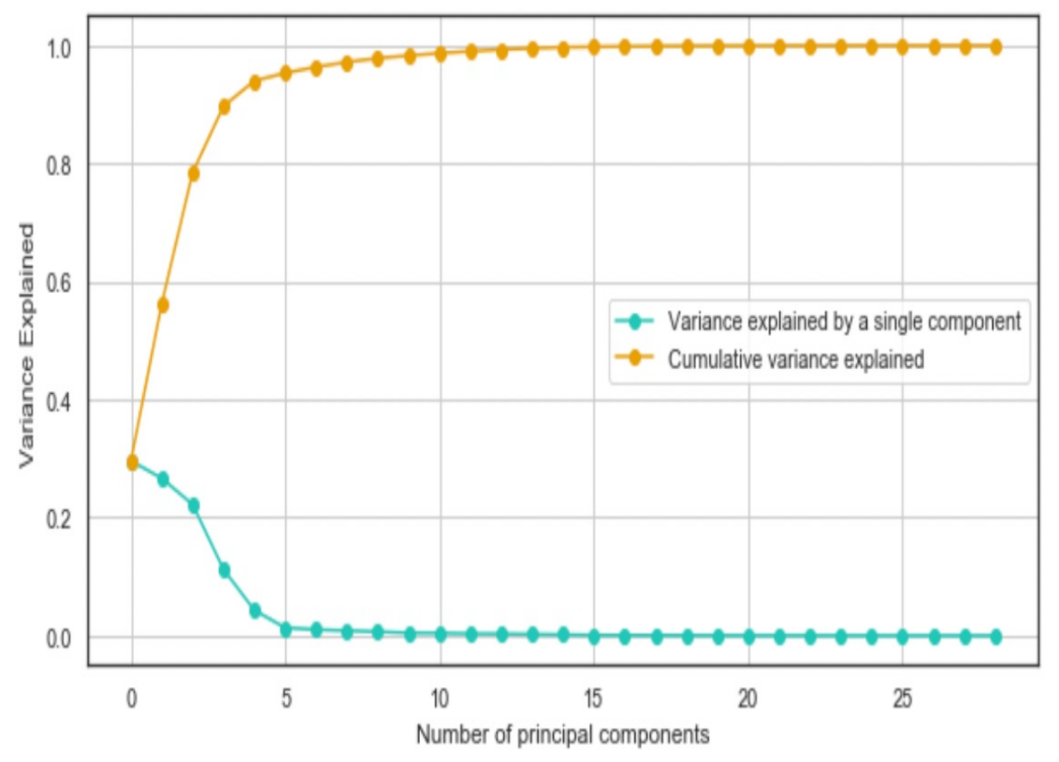


Figure 1: Density Plot of amount of given credit (LIMIT_BAL)
Figure 2: Correlation matrix by means of the Pearson’s coefficient for all feature pairs

Input variables may have different units so different scales and magnitude. So, a MinMaxScalor() is applied in order to scale the features between a range (0,1). This transformation is applied on numerical features only as the categorical variables has been already transformed into one-hot vectors. The dataset is divided in training set and test set, with the proportion 4:1. Given the imbalance fashion of the original dataset a stratified sampling strategy has been applied during the split.
To obtain a lower dimensional dataset, I applied a Principal Component Analysis (PCA) to the data to deal with the multicollinearity problem and reduce the number of dimensions. Figure 3 shows how the variance has been redistributed on the new features extracted. Regarding the number of components to keep, Table 1 shows that the first 12 PCs are able to capture almost the whole variance (99%) of the data points.



Number of PCs	Cumulative Variance Explained
4	89.7 %
6	95.3 %
8	97.1 %
12	99.0 %

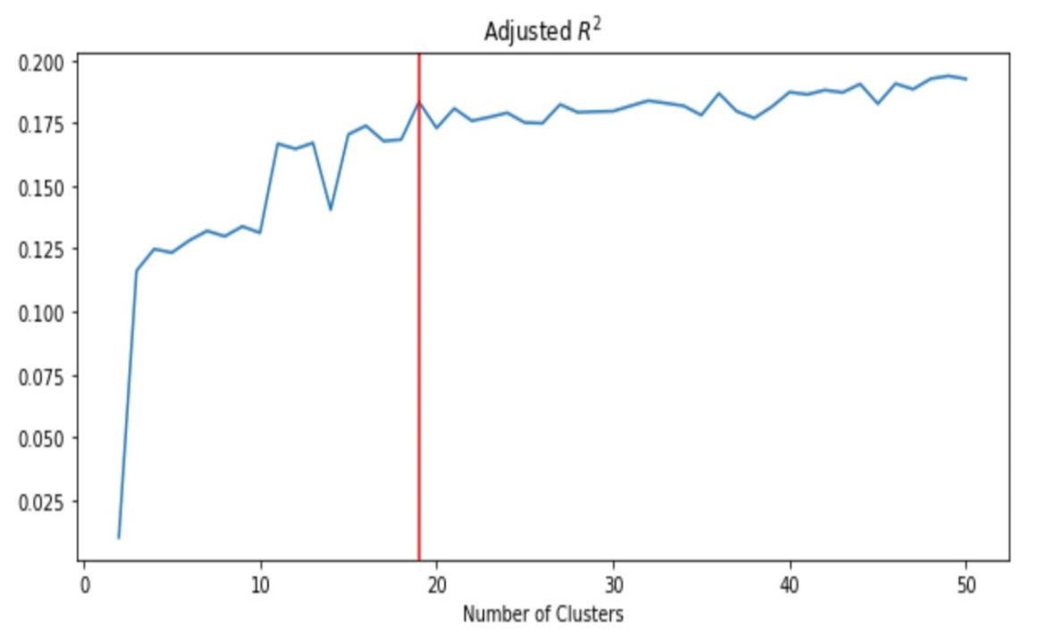
Figure 3: Explained variance ratio and cumulative variance explained of each principal component
Table 1: Variance captured of different number of PCs

In this case, we only get about 22.1% of the data are labelled as defaulters (y = 1), oversampling technique SMOTE (Synthetic Minority Over-Sampling Technique) is used to artificially augment the minority class.

Methods

- Clustering on Credit Card Defaulters

The model is completed using the KMean Cluster of machine learning combined with the Linear Probability Model. KMean Cluster is firstly used to classify users according to their credit card payment status, and then the classified groups are used to calculate the default probability of the group using the Linear Probability Model. In this process, the number of clusters finally selected is determined based on the adjusted R^2 (Figure 4) and the number of insignificant coefficients (Figure 5). After the number of clusters is 19, the rate of increase in adjusted R2 slows down, and the number of insignificant coefficients begins to rise rapidly.



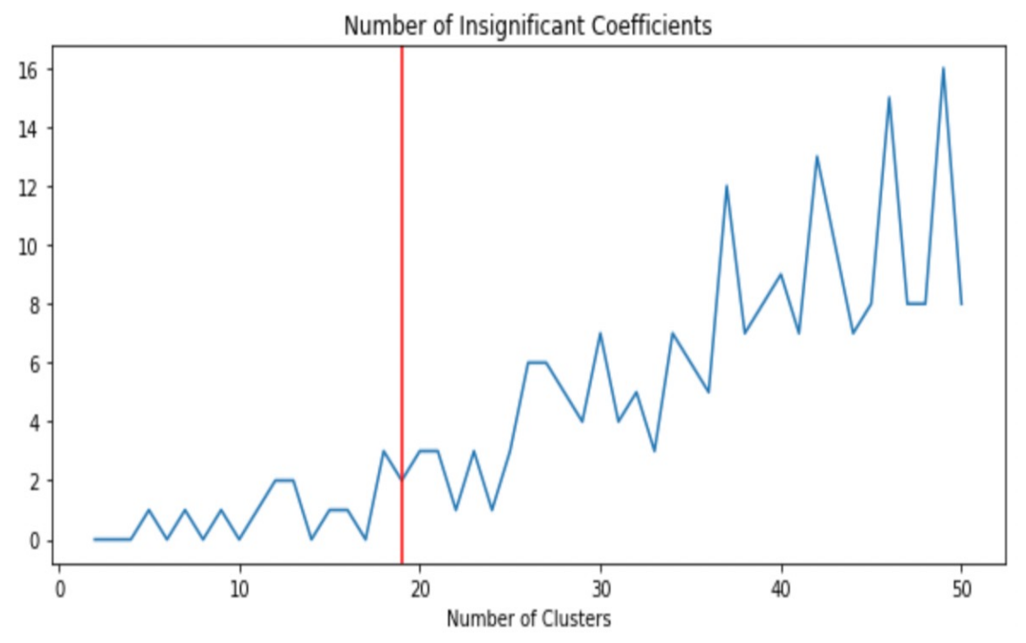


Figure 4: Adjusted R2 vs. the number of clusters
Figure 5: The number of insignificant coefficients vs. the number of clusters

Assume a given default probability threshold, if the user’s default probability after grouping exceeds the threshold, it’s assumed that the user will default, and the overall accuracy is calculated based on this rule.

- Machine Learning Algorithms for Predicting Defaulters

The goal is to build an automated model for both identifying the key factors and predicting a credit card default based on the information about clients and historical transactions. Random Forest and Gradient Boost techniques have been applied. The best configuration is selected by comparing different metrics, principally based on the AUC score.
For the Random Forest algorithm, hyperparameter tuning on the number of trees and pruning factors for each tree has been performed. Tree-based model can also be used as an alternative method for feature selection by providing how much the gini index has been affected on various splits with the given feature. Random Search has been applied for model hyperparameter tuning and class imbalance problem has also been solved.

- Credit Default via Deep Learning

Traditionally, banks use non-time dependency models to deal with credit scoring problems. However, the model would not be robust for more complicated real-world data and applications. Thus, I leveraged time-related deep learning models to provide better insights.

The features are divided into ‘static’ group and ‘dynamic’ group to understand that some of the features are time-correlated. Use LSTM to extract the latent from dynamic features. Then, I concatenated static features such as amount of given credit, gender, education, marriage and age with the latent together. In the end, I passed them into a fully-connected DNN for the binary classification. Time-steps in this process is set to be 4 months and the working process is shown in Figure 6.

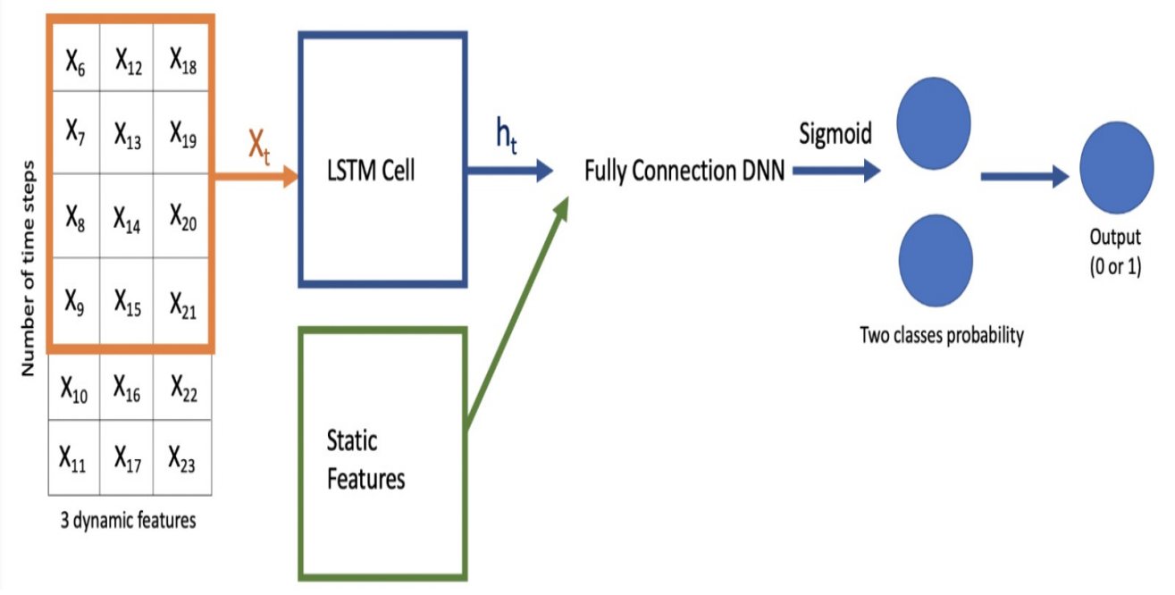


Figure 6: Deep Learning Model Architecture

nn.BCEWithLogitsLoss from Pytorch was used as the loss function. This loss combines a Sigmoid later and BCELoss into one single class.

Results

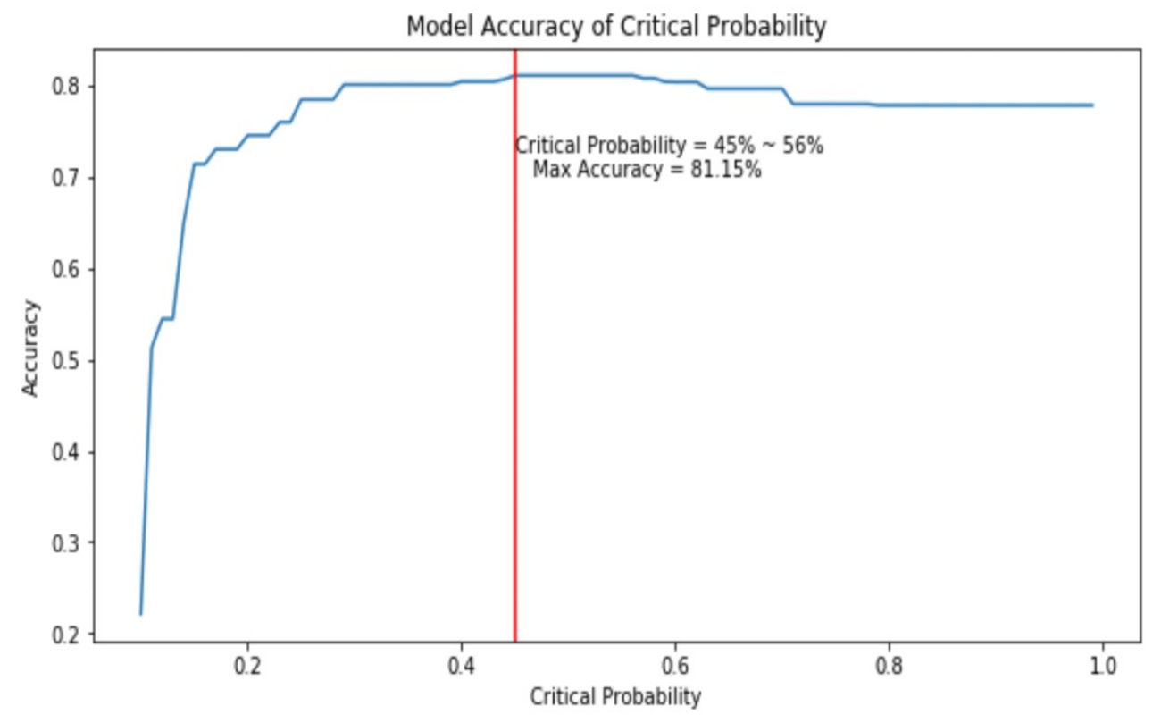
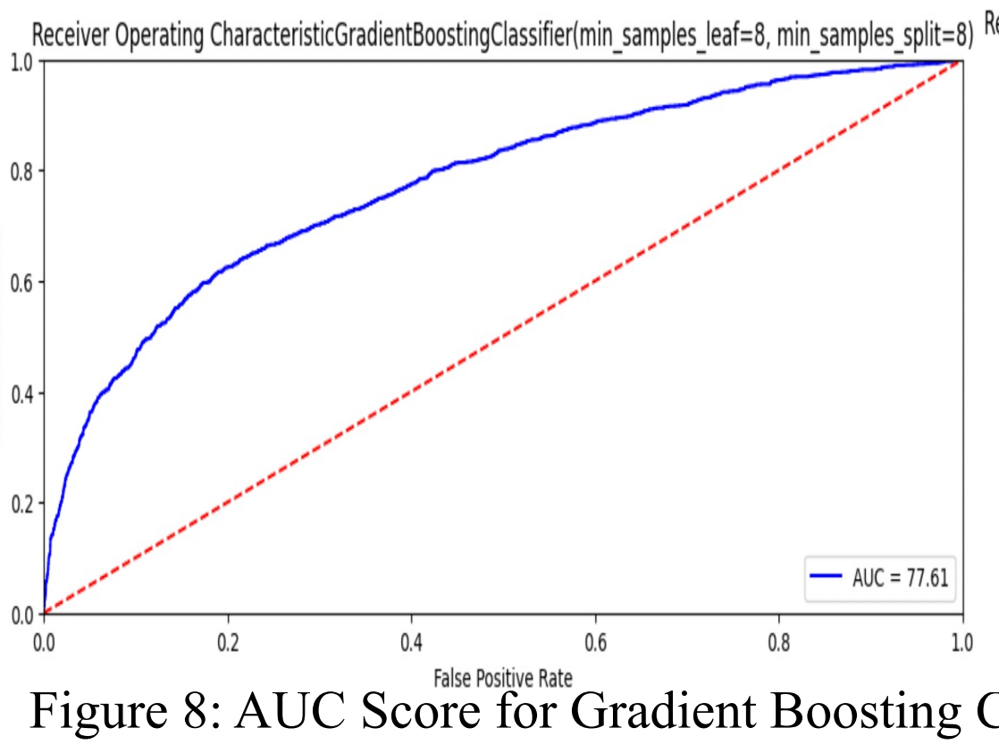


Figure 7: Model Accuracy of Critical Probability

Figure 6 shows the result of the defaulter clustering and prediction. In the case of 19 groups, when the default probability threshold is set at 45% to 56%, the accuracy of this model is the highest, reaching 81.15%.

The best performing configurations of hyperparameters on the training set have been chosen for rebuilding all the machine learning algorithms for classification and evaluating them on the test set.



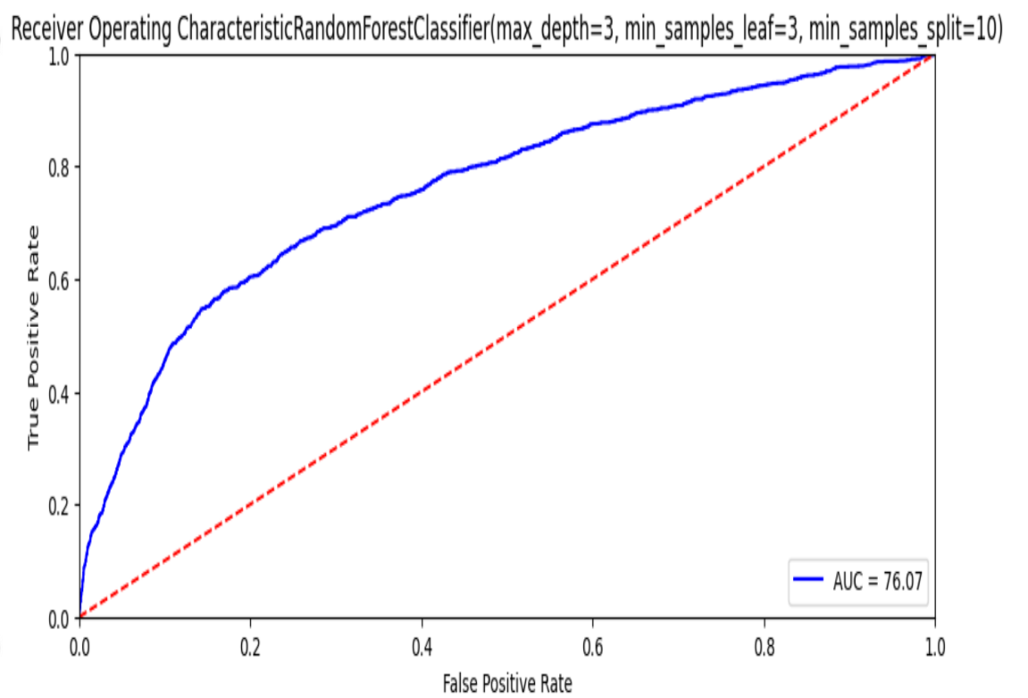


Figure 8: AUC Score for Gradient Boosting Classifier
Figure 9: AUC Score for Random Forest Classifier

Data preprocessing makes algorithms perform slightly better than when trained with original data. The oversampling has been combined with PCA to assess the dataset imbalance problem. For models I constructed, Gradient Boosting Classifier with appropriate hyperparameter tuning received the highest AUC score of 0.7761 (Figure 7). Random Forest Model implemented achieved comparable results in terms of AUC scores as well (0.7607, Figure 8).

The traditional prediction can be improved by using time-correlated features extraction model of LSTM+DNN. The testing AUC score can achieve 0.79 (Table 2).

Model	Accuracy	Precision	Recall	F1-Score	AUC
LSTM+DNN (Training)	0.84	0.86	0.87	0.86	0.78
LSTM+DNN (Testing)	0.79	0.84	0.72	0.77	0.79

Table 2: Training and Testing Results for the Deep Learning Model

Note that the scores obtained on the test should never be treated as additional information to change how the training is performed, but only as a final evaluation of the model.

Conclusions

The comprehensive analysis conducted in this project aimed to address the critical issue of credit card default, a challenge stemming from over-issuance of cards and unsustainable credit usage. By employing clustering techniques and machine learning algorithms, a structure was developed to categorize and predict the risk of credit card default based on the available credit history and individual information.

Moreover, the application of time series modeling, particularly Long Short-Term Memory (LSTM) networks combined with DNN, demonstrated enhanced efficiency in handling temporal dependencies within the credit default dataset. The model exhibited its effectiveness in analyzing time-related patterns and dynamics, providing a robust approach for complex real-world datasets and their practical applications.

This study not only offers a significant step towards understanding and predicting credit card defaults but also showcases the efficacy of leveraging machine learning and time series modeling techniques in addressing the complexities associated with such financial phenomena. The insights garnered from this research could potentially aid financial institutions in implementing proactive measures to mitigate credit defaults and better assess client risk, thereby contributing to more prudent lending practices.

References

Yeh,I-Cheng. (2016). default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.

Understanding Machine Learning: From Theory to Algorithms, S. Shalev-Shwartz, S. Ben-David, 2014

B. Hassani D. Guegan, P. Addo. Credit Risk Analysis using machine and deep learning models. Risks, 6(2): 38, 2018.

T. Liu C. M. Kuan. Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks. Journal of Applied Econometrics, 10: 347-364, 1995.

Thomas Fischer and Christopher Krauss. Deep Learning with Long Short-Term Memory Networks for Financial Market Predictions. European Journal of Operational Research, 270, 2017.

A. L. Perrone and G. Basti. A New Criterion of NN Structure Selection for Financial Forecasting. 6: 3898-3903 vol.6, 1999.

Gavira-Durón N, Gutierrez-Vargas O, Cruz-Aké S. Markov Chain K-Means Cluster Models and Their Use for Companies’ Credit Quality and Default Probability Estimation. Mathematics. 2021; 9(8):879.

T. M. Alam et al., "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," in IEEE Access, vol. 8, pp. 201173-201198, 2020.

Acknoledgement and Contact

I would like to express my sincere thanks of gratitude to my mentor Prof. Junpei Komiyama who gave me invaluable guidance throughout this project for devoting his precious time and efforts. I would also like to thank the Dean’s Undergraduate Research Fund (DURF) from New York University Shanghai for financially supporting me with the golden opportunity to do further research on this topic.

Contact Information: yl8095@nyu.edu Yuxuan Li, ‘24 B&F and Data Science, NYU Shanghai.

Github page for this project: https://github.com/kohathyli/Credeit-Default