Credit Score Classification

Ting-Hsuan, Chen
Institute of Statistics and Data Science
National Taiwan University
r13250005@ntu.edu.tw

Chu-Yun, Hsueh
Institute of Statistics and Data Science
National Taiwan University
r13250003@ntu.edu.tw

Yu-Wen, Chen
Institute of Industrial Engineering
National Taiwan University
r13546016@ntu.edu.tw

Hung-Ta, Wang
Institute of Industrial Engineering
National Taiwan University
r13546017@ntu.edu.tw

1 Introduction

1.1 Research Background

Credit scores are crucial for bank issues. We aim to classify credit scores using tabular data with mixed variables via machine learning models such as Logistic Regression, XGBoost, DNN, and TabNet. By comparing these models, we evaluate their accuracy and identify the best scenario for each model.

1.2 Dataset Introduction

Our project is to classify credit scores through personal information to support lending decisions. This dataset includes 100,000 samples divided into 80,000 for training and 20,000 for testing with 27 features (10 categorical, 17 continuous) and a target variable categorized into three classes: Good, Standard, and Poor.

2 Data Preprocessing

In the Data Preprocessing stage, we categorized variables into four types based on their significance and the extent of missing values, applying corresponding treatments for each type.

First, we removed variables that were either insignificant to the target variable or had excessive missing values like ID, Name, and SSN or type of Loan was excluded as more than 50 percent of its values were missing. The second type includes variables with no missing values to process like Customer ID, Month. The third type consists of variables that remain constant over the eight-month period. Observations revealed that certain variables did not change over time for each individual. Missing values for these variables were filled using their corresponding Customer ID values.

Lastly, the remaining five variables were handled individually as follows:

(a) Age : Missing values were filled using Customer ID, considering age change from time.

(b) Credit History Age: Text format was converted into a numerical format.

(c) Amount invested monthly : Missing values or invalid entries were replaced with average value.

(d) Monthly Balance: Outliers were identified using the range of "Average ś 3 times the standard deviation." Both outliers and missing values were replaced with the individual average.

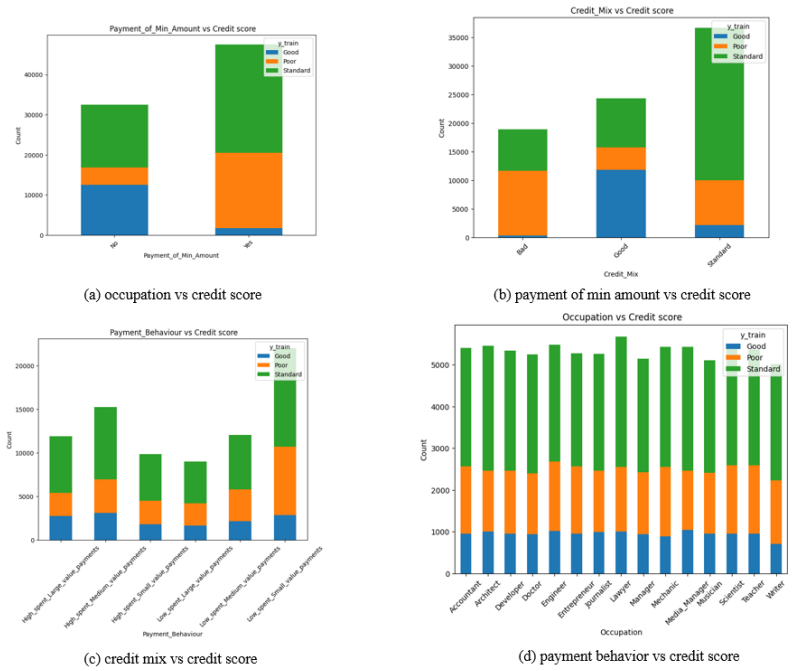
(e) Num of Delayed Payment : Values with a frequency of less than 10 would be replaced with the previous valid value.

3 Feature Selection

We handle categorical and continuous variables separately.

3.1 Categorical Variables

We use the chi-square test ($\alpha = 0.05$) to individually examine the significance of each categorical variable. From the results of the chi-square tests, we found that all four categorical variables are significant. Thus, we keep these four categorical variables. The results of the chi-square tests are as figure 1 :



(a) occupation vs credit score

(b) payment of min amount vs credit score

(c) credit mix vs credit score

(d) payment behavior vs credit score

Figure 1: Comparison of chi-square tests

3.2 Continuous Variables

First, we applied the PCA method for variable selection, and the results are as figure 2:

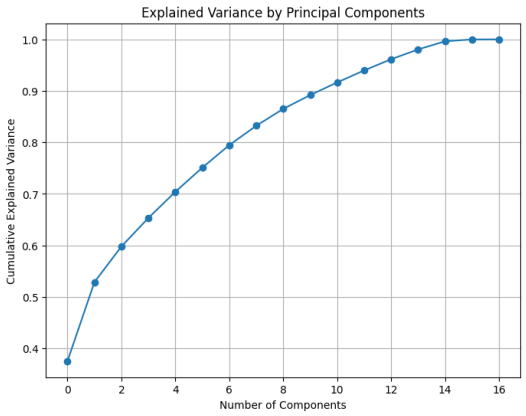


Figure 2: PCA result

We found that after performing PCA, the first 13 PCs were retained. However, we considered that PCA represents a recombination of the original variables, and the plot shows no significant difference in the proportion of explained variance. Additionally, we feel that reducing the number of variables from 17 to 13 does not bring substantial benefits. Therefore, we decided to retain all 17 original variables and explore other methods for selecting continuous variables.

Second, we used Random Forest[1] to evaluate the importance of each variable to the model. The results are as figure 3:

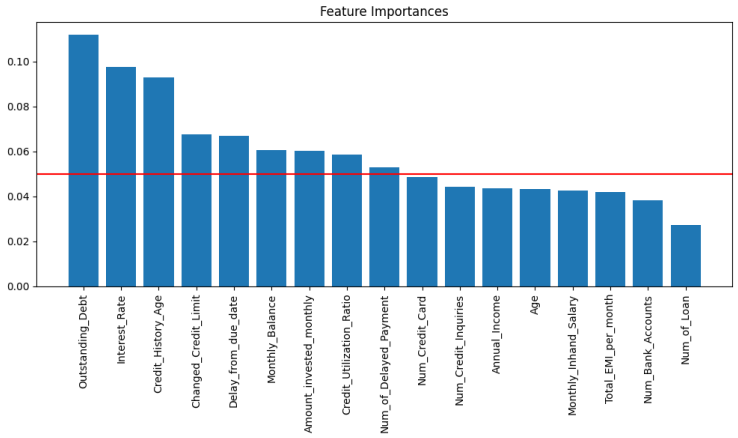


Figure 3: random forest result: importance of each variable

From the results, we observed that the importance of each variable does not show significant differences. Therefore, it is difficult to establish a clear boundary for selecting specific variables.

Third, we used LightGBM to verify whether the variables identified as important in Random Forest exhibit consistent results under the LightGBM algorithm. LightGBM is an algorithm capable of handling both categorical and numerical variables, allowing us to evaluate the importance of both types of variables simultaneously. The results of LightGBM are as Table 1:

Table 1: Feature Importance Table

Feature	Description	Importance
17	Credit_Mix	99954.025783
9	Outstanding_Debt	82863.438280
5	Interest_Rate	45585.298458
4	Num_Credit_Card	19087.964773
16	Payment_of_Min_Amount	17954.439403
7	Delay_from_due_date	15618.453998
11	Credit_History_Age	10864.401712
12	Total_EMI_per_month	10136.986126
15	Occupation	9370.076226
1	Annual_Income	8963.548150
2	Monthly_Inhand_Salary	8883.359397
3	Num_Bank_Accounts	7855.303784
8	Num_Credit_Inquiries	6826.400855
0	Age	6082.533182
13	Amount_invested_monthly	4136.549455
6	Num_of_Loan	3286.195930
14	Monthly_Balance	1892.656028
10	Credit_Utilization_Ratio	962.478561

From the results, we observed that the outcomes of LightGBM are consistent with those of Random Forest and the chi-square test. For example, the variable Credit Mix, which has the highest

importance in LightGBM, also demonstrated significant results in the chi-square test. Similarly, Outstanding Debt and Interest Rate, which rank second and third in importance in LightGBM, are also among the most important variables in Random Forest.

3.3 Result of Feature Selection

Finally, we decided to select the top 10 continuous variables based on the importance in Random Forest model, along with all four categorical variables.

The final selected variables are listed below, totaling 14 variables, including 10 continuous variables and 4 categorical variables.

- categorical variables : Occupation, Payment of Min Amount, Credit Mix, Payment Behaviour
- continuous variables : Num Credit Card, Interest Rate, Delay from due date, Num of Delayed Payment, Changed Credit Limit, Outstanding Debt, Credit Utilization Ratio, Credit History Age, Amount invested monthly, Monthly Balance

4 Model Construction

We used various machine learning and deep learning models for analysis, including the traditional Logistic Regression model, the widely used XGBoost, Deep Neural Network (DNN), and TabNet. Table 2 provides a comparative overview of these four models, highlighting their applicability under different circumstances.

Since our dataset consists of tabular data containing both categorical and continuous variables, with distinct preprocessing requirements for each type, model selection was tailored to the characteristics of the data. Initially, we trained the models using raw data without feature selection to assess whether feature selection would improve model performance. After conducting feature selection and identifying the most relevant variables, we proceeded to train new models using the selected features.

Finally, we evaluated the performance of each model to determine which was most suitable for our dataset. Under different models, after obtaining the predicted results using the test dataset, we construct a confusion matrix for each model to evaluate its performance. The classification accuracy of each model is calculated to compare the results obtained from the original dataset with those from the dataset after feature selection. Additionally, we compare the accuracy across different models to assess their relative performance. This systematic approach ensures that the chosen model aligns with the specific characteristics and requirements of the data.

Table 2: Comparison of Different Models

Aspect	**Logistic Regression**	**XGBoost**	**DNN**	**TabNet**
Interpretability	High, easy to interpret	Low, difficult to explain internal process	Low, difficult to explain internal process	High, focuses on feature selection
Suitable Data Scale	Small-scale data	Medium-scale data	Large-scale data	Medium-scale data
Feature Engineering	Relies on preprocessing	Highly adaptive	Relies on preprocessing	Automatically selects features

4.1 Logistic Regression Model

Logistic Regression is a supervised machine learning algorithm used for binary and multi-class classification tasks. It predicts the probability of a given input belonging to a specific category, making

it particularly effective in scenarios where outcomes are categorical.In this project, our response variable consists of three categories; therefore, we employ multinomial logistic regression.

4.2 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful and scalable machine learning algorithm specifically designed for supervised learning tasks such as classification and regression. It is an implementation of the gradient boosting framework that focuses on speed, performance, and efficiency, making it particularly well-suited for handling large and complex datasets.

Here, we used n_estimators=100, learning_rate=0.1, max_depth=6, objective='multi:softmax' to train the model.

4.3 Deep Neural Network

DNN (Deep Neural Network)[3] is a type of artificial neural network (ANN) that has multiple layers, allowing it to process and learn from large, complex data patterns. DNNs are capable of learning high-level abstractions in data, which makes them powerful for tasks such as image recognition, speech processing, and natural language processing.

We use three hidden layers, with 'tanh' as the activation function in each. We also tried other activation functions to achieve better results, but the final accuracy is still quite similar.

Here, we used epochs=10, batch_size=32, validation_split=0.2 to train the model.

Figure 4 is our model architecture of DNN.

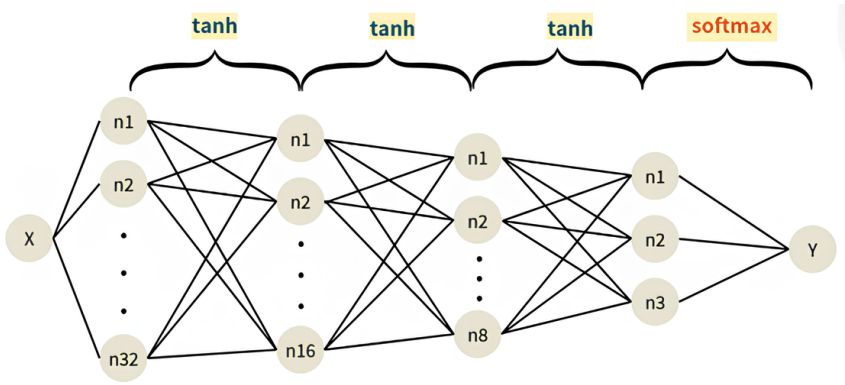


Figure 4: Model architecture

4.4 TabNet

TabNet is an unsupervised learning method specifically designed for tabular data. Unlike traditional machine learning methods that strongly depend on data preprocessing and feature selection, TabNet utilizes attention mechanisms[2] to focus on the most crucial features automatically. It can generate a report of feature importance and interactions, making it highly effective for tasks involving structured data.

We want to compare the results of unsupervised learning and other traditional ways and expect to clarify when we shall use each method.

	Feature	Importance
8	Interest_Rate	0.284992
13	Num_Credit_Inquiries	0.257347
19	Total_EMI_per_month	0.109376
12	Changed_Credit_Limit	0.088531
14	Credit_Mix	0.072899
20	Amount_invested_monthly	0.066508
7	Num_Credit_Card	0.055440
18	Payment_of_Min_Amount	0.054975
4	Annual_Income	0.009933
0	Customer_ID	0.000000
21	Payment_Behaviour	0.000000
17	Credit_History_Age	0.000000
16	Credit_Utilization_Ratio	0.000000
15	Outstanding_Debt	0.000000
11	Num_of_Delayed_Payment	0.000000
1	Month	0.000000
10	Delay_from_due_date	0.000000
9	Num_of_Loan	0.000000
6	Num_Bank_Accounts	0.000000
5	Monthly_Inhand_Salary	0.000000
3	Occupation	0.000000
2	Age	0.000000
22	Monthly_Balance	0.000000

Figure 5: TabNet Report of Feature Importance

4.5 Results

Table 3 is the classification accuracy obtained after the model training.It can be observed that the classification accuracy of the models using the data after feature selection is higher than that of the original data. Moreover, DNN has the highest accuracy among these models.

Table 3: Classfication Accuracy of Different Models

	Accuracy	
	Without Feature Selection	With Feature Selection
Logistic Regression	0.5749	0.6611
DNN	0.5300	0.7196
XGBoost	0.7400	0.7300
TabNet	0.7111	0.7074

5 Conclusion

Besides comprehensive data preprocessing and feature selection, our project classifies credit scores by Logistic Regression, XGBoost, DNN, and TabNet. Among the models, XGBoost and TabNet performed the highest accuracy with minimal preprocessing whereas other methods' accuracy decreased dramatically while analyzing raw data. Logistic Regression can be an initial attempt, and DNN may be more suitable in large and more structured data.

These findings highlight the importance of data preprocessing and model selection for tabular data. Future work may explore advanced methods to improve robustness and accuracy.

References

[1] JA Alexander and MC Mozer. Template-based algorithms for connectionist rule extraction. *Advances in Neural Information Processing Systems*, 7:609–616, 1995.

[2] JM Bower and D Beeman. *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* TELOS/Springer–Verlag, 1995.

[3] ME Hasselmo, E Schnell, and E Barkai. Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region ca3. *Journal of Neuroscience*, 15(7):5249–5262, 1995.