517 Group Project

Group Members:

Chenyi Mao Netid:chenyim2

Youshi Wang Netid:

MLF\_GP1\_CreditScore

1.Introduction/Exploratory Data Analysis

1.1 Introduction

When we first get the data, checking the shape of the data gives us an idea of the number of observations and features in the dataset, and checking for missing values to prevent potential influence on quality of our analysis and model building. Getting summary statistics of the dataset can help us identify outliers, skewness, and distribution of the data.

图形用户界面

低可信度描述已自动生成

1.2 Exploratory Data Analysis

Then, we start to do Exploratory Data Analysis to obtain some insights into the underlying structure and relationships in the data.

correlation matrix heat map:

Chart, scatter chart

Description automatically generated

Distribution of variables:

Diagram, schematic

Description automatically generated

Scatter matrix:

Chart, calendar

Description automatically generated

Pair plot:

Calendar

Description automatically generated

A picture containing shape

Description automatically generatedBox plot:

Histogram:

Chart, histogram

Description automatically generated

2. Feature Extraction, and Feature Selection

2.1.1 Feature Extraction (Binary)

PCA (Principal Component Analysis) is a dimensionality reduction technique that is used to reduce the number of variables in a dataset while retaining as much of the variability in the data as possible. The technique works by identifying the principal components of the data, which are the directions in the data that have the most variance.

When we set the number of PCA components to 2, we are specifying that we want to reduce the number of variables in our data to two dimensions. This is useful for visualization purposes, as we can plot the data in two dimensions and visualize the relationship between the variables. By selecting the two dimensions that capture the most variance in the data, we can ensure that the plotted points are spread out as much as possible and that we can see any patterns that may exist in the data. We get the data with shape (1700,2).

2.1.2 Feature Extraction (Multiclass)

To differentiate from the binary classification mode, the selectkbest method was employed in this case.

SelectKBest is a feature selection technique in which k number of features are selected based on their scores on a given metric, such as chi-square, mutual information, or f-score. The metric is calculated for each feature and the top k features with the highest scores are selected for the model.

Setting k=5 in SelectKBest means that the algorithm will select the top 5 features with the highest scores on the chosen metric. These top 5 features are then used as input variables in the model, and the remaining features are discarded. SelectKBest is useful when dealing with datasets that have a large number of features, and where not all of the features are equally important for the model. By selecting only the top k features, the model can be simplified and the performance can potentially be improved by removing noisy or irrelevant features. We get the data with shape (1700,5).

2.2.1 Feature Selection (Binary)

In this case, we use random forst to do feature selection. The process involves building multiple decision trees using randomly sampled subsets of the original dataset. For each tree, the importance of each feature is calculated based on the decrease in impurity that results from using that feature for splitting.

This importance score is then aggregated over all trees to give an overall measure of the importance of each feature. The importance scores can be used to rank the features in order of importance and select the top features for use in a final model. A graph was created with the feature importance scores on the y-axis and the feature indices on the x-axis, with the features ranked in descending order of importance. This graph can help to visualize which features are most important and which can be dropped without significantly impacting the performance of the model.

Chart, bar chart

Description automatically generatedA picture containing text, plaque

Description automatically generated

2.2.2 Feature Selection (Multiclass)

Use similar technique we used in binary one, but the result is based on our SelectKBest method. Therefore, there are five variables in the selection and measurement of importance.

文本

中度可信度描述已自动生成Chart, bar chart

Description automatically generated

3.Model Fitting and Evaluation

3.1 Fitting and Evaluation (Binary)

For Binary one, we use three different machine learning models. And we calculate the out of sample accuracy for each of them.

Logisticregression: 

Decision tree: 

For random forest method, we use five n\_estimators to get 10 different values(Both train dataset and test dataset). And we calculate the mean of them to get the accuracy score. Text

Description automatically generated

3.2 Fitting and Evaluation (Multiclass)

In this case, we only use random forest method to calculate the accuracy and precision. 

4.Hyperparameter Tuning

4.1 Fitting and Evaluation (Binary)

For binary one, we use both gird method and random choose method. The result is best score and best parameter.

Gird:

For randomized search, it is based on cross-validation. In this case, we have tried different CVs, but with CV over2, it will took over thirty mintues to get the result. So we have no choice but to set CV = 2 and n\_iter =5.

4.2 Fitting and Evaluation (Multiclass)

We choose to use Grid in this case:

5.Ensembling

5.1 Ensembling (Binary)

We used bagging, Ada boost, and voting in binary case.

Bagging: 

Ada boost:

Text

Description automatically generatedVoting:

5.2 Ensembling (Multiclass)

Voting: 

6.Conclusion

Based on ourresearch, it appears that random forest classification had the highest accuracy for the binary target classification, followed by logistic regression and then decision tree.

For the multiclass target classification, the accuracy and precision were both relatively low. The low accuracy and precision for the multiclass target may be due to the complexity and diversity of the data. In multiclass classification, it can be difficult to accurately predict each class, especially if there are many classes and they are not well-separated. Additionally, the features may not be informative enough to distinguish between the different classes. It may be necessary to perform additional feature selection or feature engineering to improve the performance of the model. Furthermore, different algorithms or ensembling techniques may need to be applied to better handle the complexity of the data.

One possible method to address this problem class imbalance correction. If the dataset is imbalanced, where some classes have significantly fewer samples than others, we can use techniques such as oversampling or undersampling to balance the dataset.

After we using hyperparameter tunning, based on the results of hyperparameter tuning using grid search and randomized search, it can be concluded that optimizing the hyperparameters of the machine learning models can significantly improve the performance of the models. The binary classification accuracy was able to improve from 0.758 to 0.8157 using the grid search method, and to 0.76 using the randomized search method. These results highlight the importance of hyperparameter tuning in the machine learning workflow, and the potential to achieve better results with more sophisticated tuning methods or more computational resources.

For multiclass target model, the test drop from 0.46 to 0.29. It is a significant decrease in performance. This may indicate that the hyperparameters selected in the tuning process were not optimal for the multiclass problem, or that the model itself may not be well-suited for this type of classification task. It is also possible that there is an issue with the dataset, such as class imbalance or lack of representative samples.

The ensembling techniques used for the binary target case were able to significantly improve the accuracy compared to individual models. Bagging achieved the highest accuracy of 0.894, followed by voting method with an accuracy of 0.82 and ada boost with an accuracy of 0.771. This indicates that combining multiple models can lead to better predictions by reducing overfitting and improving the overall generalization of the model.

The accuracy of 41.18% using voting for the multiclass target indicates that the model is not performing well. It may be due to the complexity of the problem and the difficulty in identifying all target variables at once. It is possible that focusing on two variables at a time may improve the accuracy. Additionally, other techniques such as feature engineering, selecting better features, and trying different algorithms may also improve the accuracy of the model.