ISFS 620: Data Science in Financial Services

Individual Assignment

1. Introduction

My project is Credit Scorecard. My goal is to train a good model based on the training data to predict the probability of default, and then use the model to predict the probability of default (PD) for the first 300 values in the test set.

2. Hypothesis

I hope that my model can predict the probability of user defaults as accurately as possible, so as to provide decision-making basis for financial institutions to issue loans. I will divide my training dataset into training set validation set(the ratio will be 4:1). And I will train model on training set and see the performance on the validation set, according to which I will adjust the parameters and find the best model. If my model work well in the validation dataset, like having high accuracy, f1 score and ROC value, we have good reason to believe that it should also perform well on the real test set.

3. Data description

The training dataset contains 150,000 rows and 11 columns of data (i.e., 10 features and 1 label). The label is **SeriousDlqin2yrs**, which means whether the person experienced 90 days past due delinquency or worse. And the features include age, monthly income and etc., all of them are continuous variables.

4. Data pre-processing

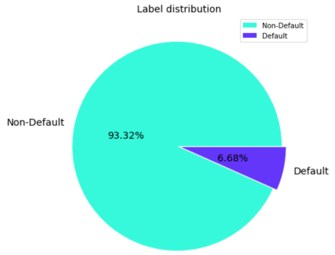
(1) **Fill the null values**: I find that there are two features containing null values: monthly income and number of dependents. And I use mean value of the variable to fill them.

(2) **Standardize the features**: Because I will use linear model to predict so the feature need standardization.

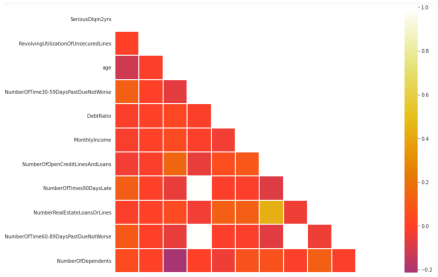
(3) SMOTE sampling: The labels' distribution is imbalanced; I use SMOTE sampling to improve the overall performance of the data.

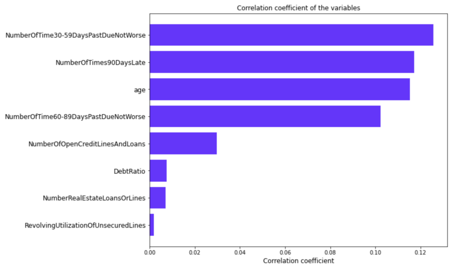
5. Data analysis

(1) I first observe the distribution of the labels and find it's an imbalanced data.



(2) I have analyzed the correlation of the data and Then I found that there is strong collinearity among the three features, but considering the limited number of features I decide to keep them to avoid losing info. And 4 features are strongly correlated with label.





(3) Draw the box plots to observe the distribution of variables. It seems all the variables are not evenly distributed but the variances of them are also not small enough for me remove them.

6. Solution and Results

(1) Training three models based on the training set: Random Forest, XGBoost tree and logistic regression. And the results are as follows (on the sampled validation set):

Model Name	f1 score	ROC AUC score	Accuracy
Random Forest	0.9708	0.9699	0.9696
XGBoost tree	0.9604	0.9599	0.9592
logistic regression	0.7966	0.8084	0.8047

(2) Test the model without SOMTE Sampling. I also use the data set without smote to train the XGBoost with same parameter, and the former one ,to both predict the validation set, the results are as follows:

	f1 score	ROC AUC score	Accuracy
SMOTE	0.6481	0.9021	0.9365
Without SMOTE	0.3035	0.5993	0.9339

(3) Try training the XGBoost model without standardization.

	f1 score	ROC AUC score	Accuracy
Standardization	0.3035	0.5993	0.9339
Without Standardization	0.2821	0.5897	0.9354

7. Conclusion

(1) Random Forest model is the best so far.

(2) SMOTE and Standardization truly improve the performance of the model.

So in the end I decide to use the random forest model with the best parameter that I find, training it based in the whole training dataset (with smote sampling). And then use the model to predict the PD of the test dataset. In the classification model, probability >0.5 means it will be classified as 1(will default). Higher probability means higher risk that issuing a loan.

As for the future plan, I believe trying stacking method may be useful in this case, because I find overfitting trend of the model. Therefore, if we can combine models together, maybe we can lower the variance. What surprises is that Standardization actually improved the model’s performance since we know that tree models do not need it. I guess this is because Standardization helps the regularization.