**Should a loan be approved or denied?**

Prepared by Yuying (Jean) Dong

12/10/2020

Abstract: To predict the small business loan should be approved or denied. I had 9 models to predict the loan status. The best model I found is Radom Forest. The top two most important predictors are Charged-Off Amount and Term.

**Key Takeaways**

This research focus on predicting small business loans that will be charged off. There is some information may be important to know about the dataset as below.

* The Paid in full is much more than Loan status charged off. 21.5% small business loans are charged off.
* Retail trade and Manufacturing obviously get more loan than the other industries.
* Most loan borrowers and banks are not in the same state.

**Analysis Summary**

*Analysis Matric*

When I look into the prediction results, the most important matric I was looking for is recall rate. My emphasis is to see the charged-off loans that has not been predicted rightly, which is the smaller the better. This could be calculated by the formular (1- recall rate). In another word, I was looking for the biggest recall rate.

The recall rate tells us the proportion of loans that we predicted correctly that actually charged off. The higher the recall, the more accurate we could predict on loans that may charge off. The recall rate of the prediction models I selected is below (Table 1.).

Table 1. Recall Rates of Models

|  |  |  |
| --- | --- | --- |
| Rank | Models | Recall Rate |
| 1 | Random Forest | 1.00 |
| 2 | Bagging Classifier | 1.00 |
| 3 | Extra Trees | 0.97 |
| 4 | Multilayer Perceptron Model | 0.97 |
| 5 | Gradient Boosting | 0.97 |
| 6 | Decision Tree | 0.97 |
| 7 | Stochastic Gradient Descent | 0.97 |
| 8 | Adaboost | 0.97 |
| 9 | KNN | 0.88 |

In case you interested, I list out the accuracy rate either. The accuracy rate is the number of correct predictions divided by the total number of predictions. The accuracy rates of models are below (Table 2.). Random Forest have the highest accuracy rate.

Table 2. Accuracy Rates of Models

|  |  |  |
| --- | --- | --- |
| Rank | Models | Accuracy Rate |
| 1 | Random Forest | 99.39% |
| 2 | Gradient Boosting | 99.39% |
| 3 | Bagging Classifier | 99.39% |
| 4 | Multilayer Perceptron Model | 98.90% |
| 5 | Extra Trees | 98.83% |
| 6 | Stochastic Gradient Descent | 98.80% |
| 7 | Decision Tree | 98.75% |
| 8 | Adaboost | 98.69% |
| 9 | KNN | 96.08% |

*Cross Validation*

I also checked if the models are overfitting. As the recall rate, accuracy rate is very high, this may result from overfitting. I used cross validation to test the overfitting. Cross validation could give an insight on how well the prediction model perform with unknow dataset (Wikipedia). The average accuracy rates of cross validation are almost the same with the prediction accuracy rate.

Table 3. Difference between Cross Validation and Accuracy Rates of Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | Models | Accuracy Rate | Cross Validation | Difference |
| 1 | Random Forest | 99.4% | 99.3% | 0.1% |
| 2 | Gradient Boosting | 99.4% | 99.4% | 0 |
| 3 | Bagging Classifier | 99.4% | 99.3% | 0.1% |
| 4 | Multilayer Perceptron Model | 98.9% | 98.9% | 0 |
| 5 | Decision Tree | 98.8% | 98.7% | 0.1% |
| 6 | Adaboost | 98.8% | 98.7% | 0.1% |
| 7 | Stochastic Gradient Descent | 98.8% | 98.7% | 0.1% |
| 8 | Extra Trees | 98.8% | 98.8% | 0 |
| 9 | KNN | 96.1% | 96.0% | 0.1% |

In the difference column of Table 3, we can see the difference between models’ accuracy rate and cross validation. The differences are less than 0.1%, which are every small. So that the prediction results are not overfitting.

*Feature Importance*

The important feature (Table 4) shows how important the features when predict the default status.

Table 4. Feature Importance (Random Forest)

|  |  |  |
| --- | --- | --- |
| Rank | Feature | Feature Importance |
| 1 | ChgOffPrinCr | 0.62 |
| 2 | Term | 0.20 |
| 3 | UrbanRural\_0 | 0.03 |
| 4 | SBA\_Appv | 0.02 |
| 5 | RetainedJob | 0.01 |
| 6 | InSameState | 0.01 |
| 7 | GrAppv | 0.01 |
| 8 | DisbursementGross | 0.01 |
| 9 | UrbanRural\_1 | 0.01 |
| 10 | RealEstate | 0.01 |

The most important feature is ChgOffPrinCr (charged-off amount), the feature importance of which is 0.62. The second important feature is Term (loan term in months) , the feature importance of which is 0.2. The importance of the rest features is less than 0.03. As we emphasis on the prediction of charged-off loans, it is reasonable to have charged-off amount as the most important feature.

**Commentary**

Basing on current analysis, I have several recommendations to share here.

* Looking for small businesses who may need SBA loan all over the nation. Over half of the small businesses borrowed SBA loan from banks in different states. Small businesses may prefer borrow loans from a different state. Instead of focus on the local small businesses, the banks may attract small businesses all over the nation.
* Pay attention to manufacturing businesses. The amount of loan borrowed to manufactory is almost the same as the amount borrowed to retail trade businesses. However, the number of manufactory businesses is nearly half of the number of retail trade businesses. This means the average loan borrowed by manufacturing almost twice the average loan borrowed by each retail-trade business.
* Define urban/rural in data input process. Urban or Rural column contains nearly one sixth data shows “Undefined”, which is the third important feature influencing the prediction of default status (Table 5).

**Reference**

Machine Learning Crash Course, retrieved from [https://developers.google.com/machine-learning/crash-course/classification/accuracy)](https://developers.google.com/machine-learning/crash-course/classification/accuracy)T)

Wikipedia, retrieved from <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>