CONTINUOUS RISK EVALUATION OF LOANS

Shiqi BAI (20219140)

CONTENT

EXECUTIVE SUMMARY

In this report, we applied various machine learning methods to differentiate the 'good' and 'bad' loans, so that help the company monitor the risks before and after a loan issued. With the rigid comparison of different models, we finally chosen the Random Forest Classifier (Recall=0.78) as our best model to deal with the classification problem.

Regarding loan products, it is recommended to issue further small amount and short-term loan products. In addition, whether the current loan term is 36 or 60 periods, it is relatively long for customers and companies. The possibility of overdue customers and the possibility of bad debts of the company will increase. It is recommended that loan products less than 36 periods can be developed. In addition, for customers who can accept high interest rates, they need to pay extra attention to it, they can control the loan amount and loan term, or supplement financial proof and manually intervene in credit.

With regard to the purpose of customer loans, most of the customers are debt consolidation. Although they cannot be denied in one fell swoop, they can be specifically aimed at customers of debt consolidation and understand their debt scale and composition. Approval rules, or provide manual approvers to control quotas, interest rates, and deadlines, in order to prevent possible risks in advance.

Regarding customers, the current credit rating division can better distinguish high-quality customers, so many high-quality customers are marketed in the front-end marketing process. Secondly, we can offer preferential treatment for high-quality customers in terms of quota, term and interest rate, and improve the approval efficiency of these customers, improve other service quality, and enhance customer experience.

The rest of the report is organized as the following: we explored the data and visualized some important variables in the second section. In the third section, we introduced the methods of data processing, model constructions and hyperparameters tuning applied in this survey, and then demonstrated the results and evaluation of all models. Finally, we illustrated both business and technical insights to help the company develop the strategies.

EXPLORATORY DATA ANALYSIS

Dataset exploration

The original datasets consist of historical payment and loans/borrowers' information, the latter as our main predictor source has 2601 observations and 145 variables in total. With the consideration that the task is to help the company predict loan default and evaluate the potential risks, we constructed an effective model to monitor a loan after the before it issued according to both loan/borrower information and the past payment files provided by the Lending Club.

At the first stage, we explored the data and undertaken a beginning view on the dataset to find some characters of it:

- Of all 2601 observations, there are 2161 loans with the state of fully paid (Negative label) and the rest of 440 loans has been charged off (Positive label). It is an unbalanced dataset for a prediction task.
- This is a semi-structured dataset with both numerical and categorized features, and many columns have various missing values. As a result, we dealt with the data problems in the data preparation section.

variable exploration

Through the analysis and visualization of this dataset including the analysis of the overall loan situation, the comparative analysis against the target variable ('Fully paid'/'Charged off') among various variables, we can have an unambiguous understanding of the given loan situation and use it to guide the improvement of subsequent loan policies and the development of loan business, so that the company can avoid risks in a better way and increase more profits.

- 'loan_amnt': The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value. [1]

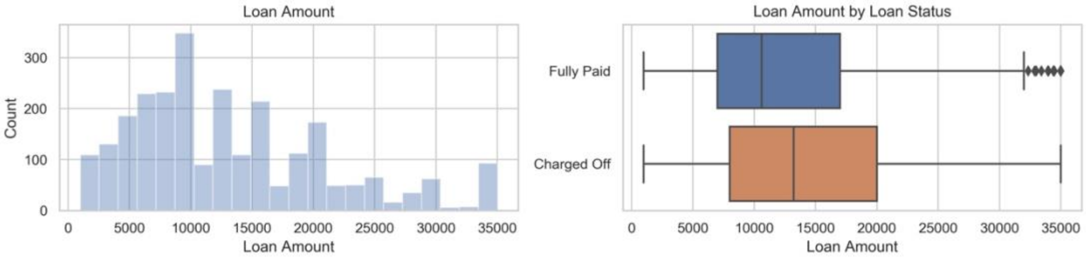


Figure 1:

Loan amounts range from 1000 to 35,000 with the median of 11,200 and Charged-off loans tend to have higher loan amounts. The distribution of loan amount shown in Figure 1 demonstrates that all loan borrowers prefer small amounts of loans, and the demand for large amounts of loans is not very high.

- 'int_rate': Interest Rate on the loan. [1]

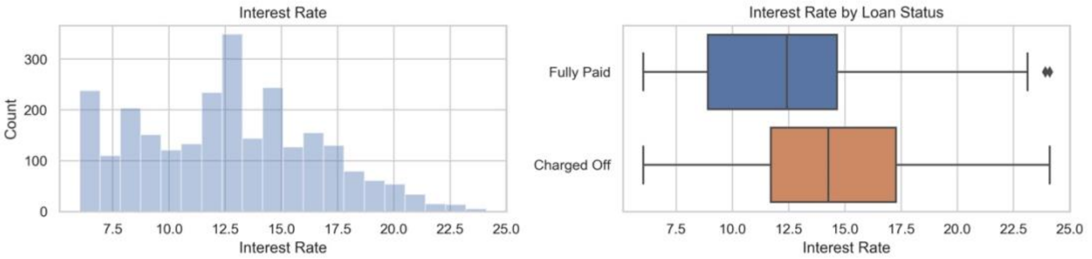


Figure 2:

The loan interest rate with the highest proportion of normal repayment customers is about 9% - 12%, while the loan interest rate with the highest proportion of overdue customers is about 15% - 17%. In addition, the distribution of interest rate for the overdue customers has shifted to the right obviously compared with that of normal customers. In general, loan customers who prefer low interest rate are more likely to repay normally, while those who accept high interest rate are more likely to delay or default their loans.

METHODOLOGY

Data preparation and Feature enginerring

MISSING VALUES TREATMENTS: As we discussed in our initial dataset exploration, we deleted the columns with data missing over 80% in order to limit the feature space. For other columns with missing data, we manipulated data imputation to fill the missing values.

ONLY KEEP FEATURES KNOWN TO POTENTIAL INVESTORS: We must also deal with a common problem in data science, namely 'leakage from the future', to ensure that we do not use information that cannot be known in advance when deciding whether to issue loans to customers. This information about the future can lead to overfitting problem and reduce the capacity of our models to make accurate predictions.

MODIFY FEATURES TO ENHANCE THE DATASET: 1. Calculate the sum of interest fees received of each loan provided from the payment dataset into a new variable 'INT_PAID_TOTAL'; 2. Calculate the average value of some related variables (e.g. 'FICO score') by checking the correlation matrix into new variables; 3. Convert the original variables into new types (e.g. 'purpose', 'earliest_cr_line').

CONVERT THE CATEGORIZED FEATURES TO NUMERICAL TYPE: When we studied at the dataset, we found that different data types exist in this semi-structure dataset. We must convert object data into numerical, because it cannot be processed by many machine learning algorithms directly.

DROP FEATURES ARE HIGHLY RELEVANT TO REDUCE MULTICOLLINEARITY: we dropped features to ensure that all correlation coefficients are under 0.8.

DEAL WITH THE UNBALANCED DATASET PROBLEM: Obviously, the data amount is unbalanced between 'fully paid' and 'charged off' label (Target variable: encode the Charged-off label into '1' and another is '0'), which means that the model would predict the loan results based on the majority data if we leave the problem alone. Our solution is over-sampling the data by applying SMOTE algorithm since that the volume of dataset is not big for the model training and minority label (i.e. '1') is more important for our analysis.

STANDARDIZATION: Standardization is useful when our data has different scales, and the algorithms you use do assume that your data has a gaussian distribution.

Modeling and Tuning

The goal of this report is to implement various predictive pipelines to distinguish between 'good' and 'bad' loans (possible default) and whether a borrower would repay in the next month. After comparing different models by using proper classification measures in the evaluation section, we finally picked Random Forest Classifier as our winner model.

We applied a basic Logistic Regression as our baseline model due to its comparative interpretability and performance to the traditional methods. In this section, we implemented the following pipelines and tuned the hyper-parameters to improve the performance with the training data:

Models	Hyper-parameters tuned
Baseline: Logistic Regression	
1 - Logistic Regression with SGD training	'alpha': 0.01, 'penalty': 'l2'
2 - K-Nearest Neighbors	'lda__n_components': 3, 'n_neighbors': 125
3 - Decision Tree	'max_depth': 20, 'min_samples_leaf': 10, 'min_samples_split': 2

4 - Random Forest	'n_estimators': 50, 'max_depth': 3; 'max_features': log2
5 - Extra Tree Classifier	'n_estimators': 200, 'max_depth': 3
6 - Neural Network with MLP Classifier	'hidden_layer_sizes': (5, 2)
7 - Naive Bayes Classifier	
8 - Linear Discriminant Analysis	'solver': 'svd'
9 - Gradient Boosting	'n_estimators': 200, 'max_depth': 3
10 - Bagging Classifier	'n_estimators': 20
11 - Multi Layer ANN	3 layers, epoch: 10
12 - Single Layer ANN	epoch: 20

Table 1: List of different models and hyper-parameters tuned

Results and Evaluation

A lot of time has been spent in data preparation section, which is enough to demonstrate how important data preparation in machine learning. Only with 'good' data can motivate substantial classification results be predicted. For binary classification problems, logistic regression is generally preferred. Start by defining measures for evaluating the models' effects. According to the actual circumstance of the lending industry, we assume that an investor lends money to a person who has no ability to repay, which means that the investor can't obtain the interest but also the principle. However, if an investor lend money to a solvent person, he can make profits from the investment. The ROC curves are shown in Figure 3, which help us get an overview of all models.

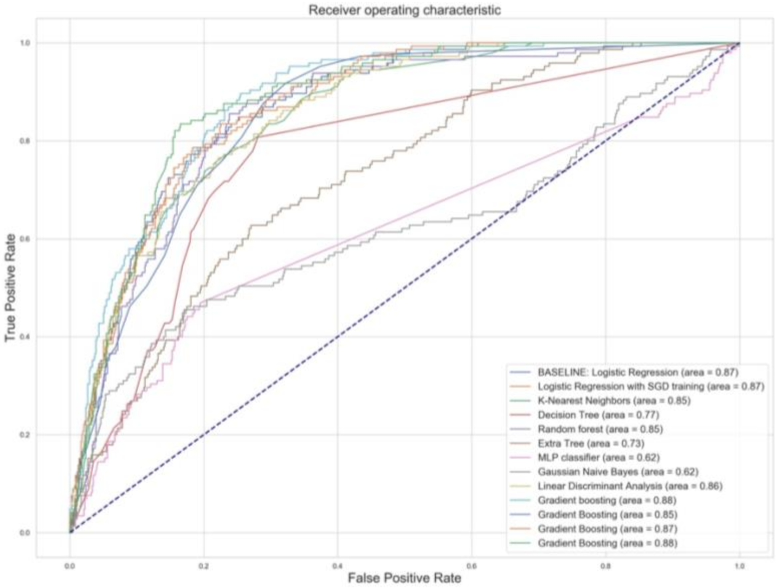


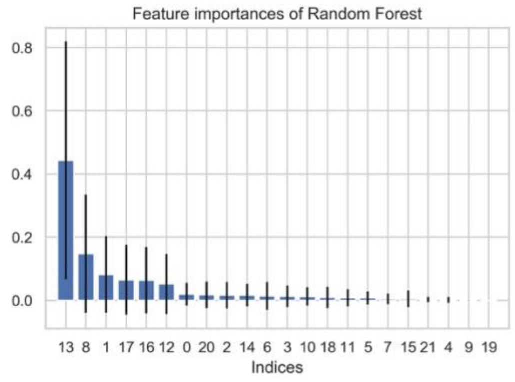
Figure 3: ROC curves of different models

For both the prediction and business perspectives, the loss is considerable if we predict a person can repay the loan improperly. So, the precision no longer applies to this business context, in order to realize the profit maximization, not only requires models have higher recall scores, as well as fall-out ratio. Therefore, True Positive Rate and False Positive Rate are adopted here.

Models	Evaluate on testing data						
	Accuracy	F1-score	AUC	Recall	Precision	Total: 145 observations	
						TP	FP
Baseline Model	0.84	0.59	0.87	0.68	0.52	99	46
LR with SGD training	0.84	0.57	0.87	0.66	0.51	96	49
K-Nearest Neighbors	0.83	0.55	0.85	0.63	0.49	92	53
Decision Tree	0.80	0.50	0.76	0.59	0.43	85	60
Random Forest	0.80	0.56	0.85	**0.78**	0.44	113	32
Extra Tree Classifier	0.77	0.39	0.85	0.45	0.35	65	80
Neural Network with MLP	0.78	0.32	0.62	0.30	0.33	44	101
Naive Bayes Classifier	0.70	0.36	0.62	0.50	0.28	73	72
LDA	0.82	0.52	0.86	0.57	0.48	83	62
Gradient Boosting	0.85	0.56	0.88	0.58	0.54	84	61
Bagging Classifier	0.82	0.49	0.85	0.50	0.47	73	72
Multi-Layer ANN	0.85	0.60	0.85	0.67	0.54	97	48
Single-Layer ANN	0.87	0.61	0.88	0.76	0.52	110	35

Table 2: Various performance measures of different models

According to the results shown on the above, we picked the Random Forest model (with 'n_estimators': 50, 'max_depth': 3; 'max_features': log2) as our final prediction machine even if Single-Layer ANN might has a better performance if we can collect more data. The reason we made the decision is that we hope the model has considerable interpretability, especially for this kind of risk evaluation problem. As shown in Figure xx, we can find that which features have more power to drive the model, in other words, which factors are more important when the company decide a loan whether to issue.



13	last_fico_avf
8	total_rec_late_fee
1	term
17	verification_status_Source Verified
16	home_ownership_RENT
12	fico_avg

Figure 4: Feature importance ranking of Random Forest Classifier

From the level of supervision, first of all, we must prevent the systemic risks that the out-of-control of some industry credit businesses may bring to the whole society, especially for mainstream credit products, supervision will pay more attention. Supervision requires considerable details about the credit business that industries engage in, including specific measures to prevent fraud risks and credit risks, to have a penetrating understanding and control. Therefore, from a macro perspective, in a whole category, regulation has an interpretable demand for the data risk control model used by industries in credit business.

INSIGHT REPORT

This report demonstrates the use of machine learning methods to train a predictive model that can identify default loans and other good loans with an accuracy rate of up to 80%. This is achieved through information obtained before the loan is approved; this includes but is not limited to the borrower 's credit history, unemployment rate, reference short-term interest rate, and so on. However, its accuracy is a little bit less than the baseline accuracy (up to 84%). Of course, we have reason to ask: If blind guessing is better than the prediction model trained by machine learning, what is the point? First, some of the models shown in this article explain well why a particular borrower may default. His or her loan.

More importantly, there can be clear reasons for inferring future loans. From the logistic regression model, we can infer that if the borrower has a higher debt-to-income ratio and plans to borrow a high-value loan, then the loan is likely to default. Although reasoning is very intuitive, it has specific numerical support, which is something that a benchmark model and random guessing cannot provide. Secondly, this is an unbalanced way of unreasonably expecting loan data in the real world. Fully paid loans are always the minority of the majority and default loans-if this is true, most people, including authors and readers, can invest billion A millionaire always buys most stocks.

In fact, these data were provided by Lending Club, which had previously processed many loans that did not meet the requirements of the underwriters and rejected them. The unfiltered loan request initially seen by Lending Club is likely to be more balanced because the baseline model is no longer useful. Therefore, the model presented in this article may be very useful if exposed to unfiltered loan requests. Finally, as can be seen from Table 1, all models can maximize investment returns, which is superior to the strategy currently used by Lending Club. Of course, this is at the expense of refusing nearly half of the existing loans. If some investors require a low-risk investment plan, this is still an effective strategy to reduce losses.

Obviously, although the multi layers neural network model achieves the highest accuracy, the performance improvement is very small compared to other machine learning methods. This raises two possibilities. First of all, this is the data limitation of machine learning algorithms. Only when more data is available can we achieve higher accuracy-collecting or designing data from scratch. Second, this is because the neural network is not optimized. The second point is of course correct, because the accuracy of the model reaches 87%. In fact, it was trained at a certain point, but it was lost due to computer problems. As long as there is enough time, it is entirely possible to optimize the neural network to the extent that there is a huge gap with other traditional machine learning methods. Nonetheless, the Random Forest Classifier has achieved a great balance between good performance and interpretability if we do not choose a neural network. The neural network-based model can of course be used as a selection model if we do not consider of the interpretability.