Credit Score Prediction of Loan Applicants

Ben Liu

SharpestMinds

Contents

- Introduction
- Methodology
- Data Preparation
- EDA
- Data Cleansing
- Modeling
- Conclusion
- Future

Introduction

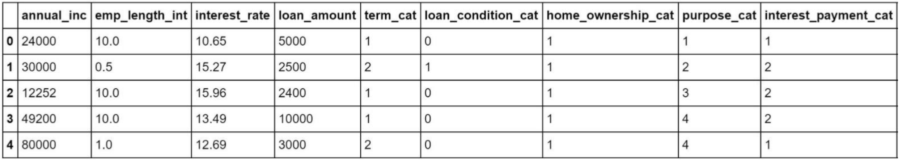
- Background: What major factors does the management team of a bank consider when they decide to grant a loan? This question is preoccupying the minds of loan applicants around the world. Understanding this, I discovered several factors which might influence how a bank calculated the credit score.
- Data: This project is based on a dataset that contains information about 1 million potential borrowers of a bank.
- Objective: The goal of this report is to predict the credit score of loan applicants with high accuracy.

Methodology

- Since this bank utilized various categories ("A","B",...) to evaluate client's credit score, I implemented several supervised classification machine learning models, such as k-NN, Decision Tree, Random Forest, and LightGBM.
- The primary metric would be the prediction accuracy of the test set.

Data Preparation

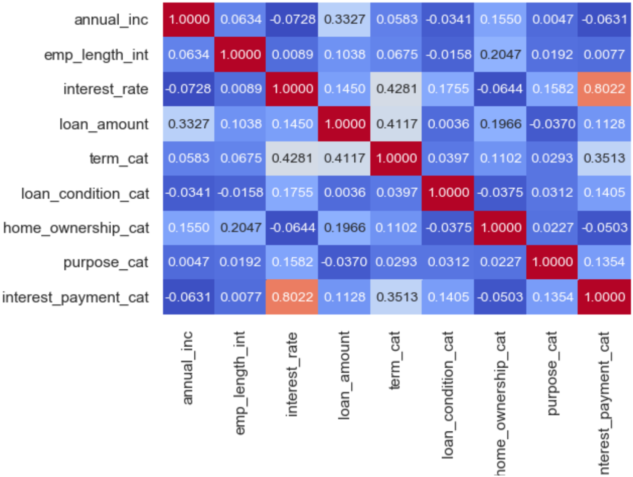
Based on my domain knowledge in finance, I selected 9 major features as explanatory variables, including 4 numeric variables and 5 categorical variables.



	annual_inc	emp_length_int	interest_rate	loan_amount	term_cat	loan_condition_cat	home_ownership_cat	purpose_cat	interest_payment_cat
0	24000	10.0	10.65	5000	1	0	1	1	1
1	30000	0.5	15.27	2500	2	1	1	2	2
2	12252	10.0	15.96	2400	1	0	1	3	2
3	49200	10.0	13.49	10000	1	0	1	4	2
4	80000	1.0	12.69	3000	2	0	1	4	1

EDA

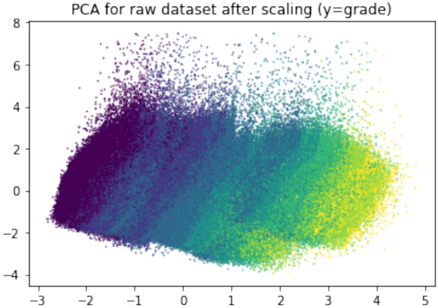
I conducted descriptive analysis about 9 features respectively, and I also explored the correlation between the features.



	annual_inc	emp_length_int	interest_rate	loan_amount	term_cat	loan_condition_cat	home_ownership_cat	purpose_cat	nterest_payment_cat
annual_inc	1.0000	0.0634	-0.0728	0.3327	0.0583	-0.0341	0.1550	0.0047	-0.0631
emp_length_int	0.0634	1.0000	0.0089	0.1038	0.0675	-0.0158	0.2047	0.0192	0.0077
interest_rate	-0.0728	0.0089	1.0000	0.1450	0.4281	0.1755	-0.0644	0.1582	0.8022
loan_amount	0.3327	0.1038	0.1450	1.0000	0.4117	0.0036	0.1966	-0.0370	0.1128
term_cat	0.0583	0.0675	0.4281	0.4117	1.0000	0.0397	0.1102	0.0293	0.3513
loan_condition_cat	-0.0341	-0.0158	0.1755	0.0036	0.0397	1.0000	-0.0375	0.0312	0.1405
home_ownership_cat	0.1550	0.2047	-0.0644	0.1966	0.1102	-0.0375	1.0000	0.0227	-0.0503
purpose_cat	0.0047	0.0192	0.1582	-0.0370	0.0293	0.0312	0.0227	1.0000	0.1354
interest_payment_cat	-0.0631	0.0077	0.8022	0.1128	0.3513	0.1405	-0.0503	0.1354	1.0000

PCA

To analyze the relationship between "grade" and other 9 explanatory features, principal component analysis (PCA) was utilized to summarize 9 features with 2 new features.



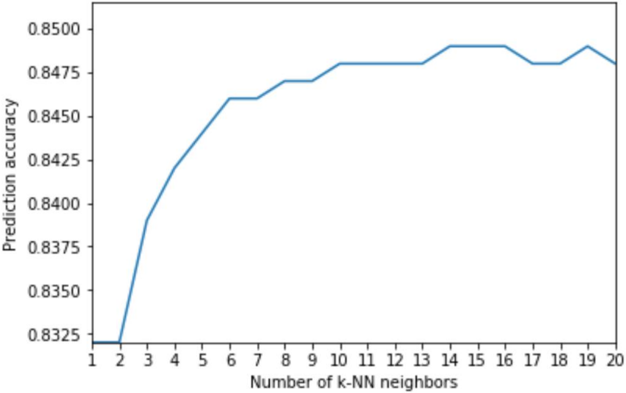
Although there were overlaps, it was noteworthy that the color was becoming lighter from the left to the right in the plot. Therefore, it was feasible to classify the dataset with typical classification models.

Data Cleansing

- Impute missing values.
- Deal with outliers.

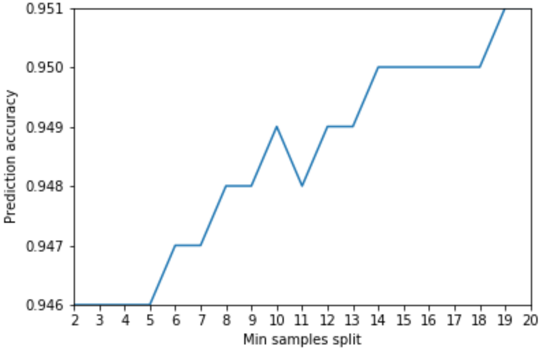
Model1 - k-NN

After data scaling, train-test split, model fitting, I tuned the parameters of the k-NN model, including the number of neighbors as follows. Through multiple similar experiments, the prediction accuracy of the optimal k-NN model was 86.7%.



Model2 - Decision Tree

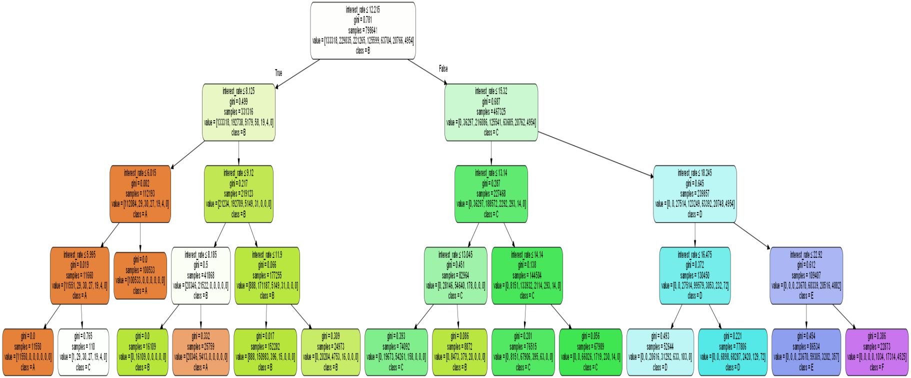
Similarly, I tuned the parameters of the Decision Tree model to increase the accuracy, take the “Min samples split” for example:



After dozens of experiments, the prediction accuracy of the optimal Decision Tree model was 95.8%.

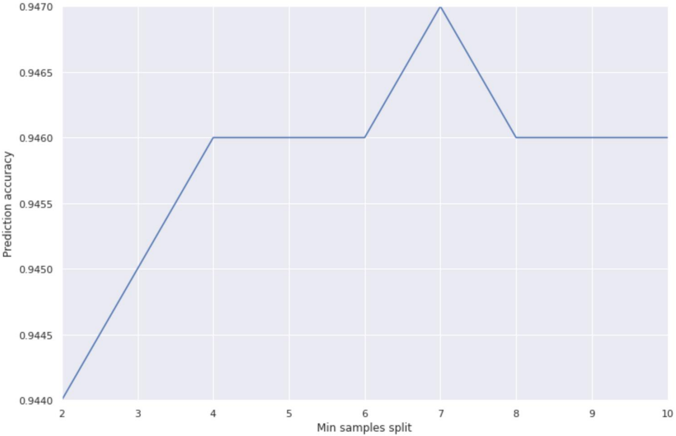
Model2 - Decision Tree

To demonstrate how each node was split by a feature, I visualized the process of a Decision Tree model when the "Max depth" was 4.



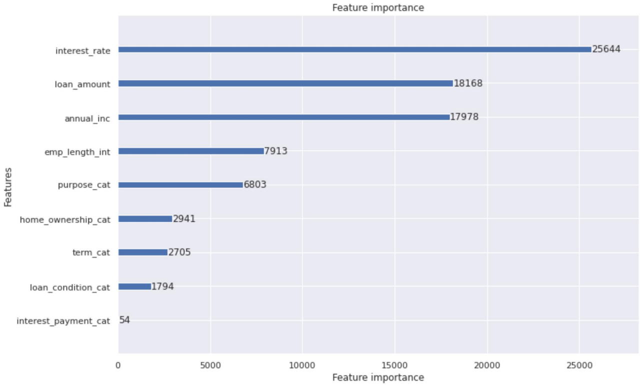
Model3 - Random Forest

Since Random Forest was a collection of various decision trees, I also applied it to the dataset. After parameter tuning experiments, the prediction accuracy of the optimal model was 94.7%.



Model4 - LightGBM

For the LightGBM model, I applied Grid Search to conduct the hyperparameter tuning (learning rate, number of estimators). The prediction accuracy of the optimal model was 94.9%. From the below graph, it could concluded that the "interest rate" was the most significant feature.



Result

According to the results of the previous machine learning models, the Decision Tree model achieved the highest prediction accuracy as 95.8%, increasing the base value by 8%.

	Prediction Accuracy
k-NN	86.7%
Decision Tree	95.8%
Random Forest	94.7%
LightGBM	94.9%

Conclusion

- Accuracy: The prediction accuracy of the k-NN model was only 85.8%, while that of the Tree-based models was around 95%.
- Efficiency: The Decision Tree model outperformed other models dramatically, in terms of computational time.
- Explainability: It was easier to interpret and visualize the relation between the response variable and explanatory variables, utilizing the Decision Tree model.
- Feature importance: The variable "interest rate" was the most significant factor when evaluating the credit score in this case.

Future

- Feature Engineering.
- More machine learning algorithms.
- Various types of hyperparameter tuning.