Warm-up of Statistical Machine Learning

Wenjing Xue 21028457
Hongxi Chen 21025223
Guobin Ding 21018385
Guan Qiao 21017331

CONTRIBUTION OF WORK
Wenjing Xue & Hongxi Chen: Data Analysis and Processing
Guobin Ding & Guan Qiao: Model Evaluation and Enhancement

1 *INTRODUCTION*

1.1 *Background*

With the development of the economy and financial market, people's ideas and ways of spending have changed, and credit cards have become an indispensable tool in most people's lives with this change. In addition to this, the increase in the type and intensity of loans issued has allowed more people to take out loans to solve a range of problems caused by shortages of funds through short periods of time. However, it also provides an opportunity for many unscrupulous individuals who want to profit from the situation to take advantage of other people to obtain loans and often cause problems with many people's credit histories, making it difficult to obtain loans. Home Credit hopes to give more people a good credit experience and provide positive access to loans for those with insufficient credit history. To visually analyze whether a customer applying for a loan has the ability to repay the loan would take a great deal of time and manpower, and it is unlikely that very accurate results would be obtained. Therefore, Home Credit is using a variety of data to analyze and design optimal quantitative models to calculate the probability of repayment and to identify and predict the repayment ability of customers, so as to obtain a more accurate judgment on whether to grant them a loan, and to ensure that customers who have the ability to repay will not be rejected to apply for a loan.

1.2 *Restatement of Problems*

According to the background introduction, in order to refine this situation, we decide to solve following questions so that we may get some ideas:

1.What is the profiling of the customer base?
2.How those attributes influence the probability of default?
3.How to use models to accurately predict whether the customer would defualt or not?

2 *RELATED WORKS AND EXISTING TECHNIQUES*

In recent years, various algorithms of machine learning have been widely used in various fields, including medical and financial fields. Vattsal.S (2022) used five different machine learning algorithms: SVM, Random Forest, Logistic Regression, Decision Tree and KNN model to predict and diagnose breast cancer and compared these five models by confusion matrix and accuracy, found that SVM has the highest accuracy. Bartosz.S et al (2012) classify the company financial data by Logistic Regression and SVM and classify them into five categories to evaluate the financial status of the company. Machine learning is also widely used in the business world. Sulim.K and Heeseok.L (2022) applied a Decision Tree model to predict customer churn in the influence commerce and obtained a high model accuracy, demonstrating that the model can be used well in this field.

Identifying and predicting an applicant's ability to repay a loan is a very important financial issue that can help a mortgage lender make a more accurate judgment about whether or not to grant an applicant a loan. There are many scholars who have used different machine learning methods for modeling and analysis on this problem. As early as 2010, Tsai and Chen found that the 'Classification + Classification' hybrid model based on the combination of logistic regression and neural networks can provide the highest prediction accuracy and maximize the profit based on four different types of hybrid models are compared by four different types of techniques.

Addo et al. (2018) built binary classifiers based on machine and deep learning models on real data in predicting loan default probability and work with machine and deep learning models for credit risk analysis. They observe that the tree-based models are more stable than the models based on multilayer artificial neural networks. Deng (2019) used SASEM tool to build three algorithms on predicting credit card customer churn based on Decision Tree, Logistic Regression and Neural Network models, found that Neural Networks and Logistic Regression outperformed decision trees as a way to optimize the bank's service to credit card customers.

3 *METHODOLOGY*

A variety of machine learning algorithms can be used to build models when dealing with binary classification problems. The following is a short introduction to theoretical and mathematical principles to Logistic Regression, SVM, KNN, Gaussian GB, Random Forest, and GBDT.

3.1 *Logistic Regression*

Logistic Regression is a linear model commonly used to solve classification problems. It maps input features to probabilities for output by using a sigmoid function. The goal of logistic regression is to predict the probability of a binary classification based on a linear combination of the input features. The core idea of logistic regression is to use a parametric linear equation to weight and sum the input features with weights and map the result to a probability value between 0 and 1 by a sigmoid function.

$$P(y=1|x) = \frac{1}{1+e^{-\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n}} \tag{1}$$

This activation is capable of mapping any real number to a range between 0 and 1 so it outputs the probability that a sample belongs to a certain category. By maximizing the log-likelihood function, we can find the parameters that make the model fit the sample data best.

Logistic regression is highly interpretable because it estimates weights that can explain the influence of features on results. Combined with simple implement and computationally efficiency, logistic regression become the preferred algorithm for classification problems.

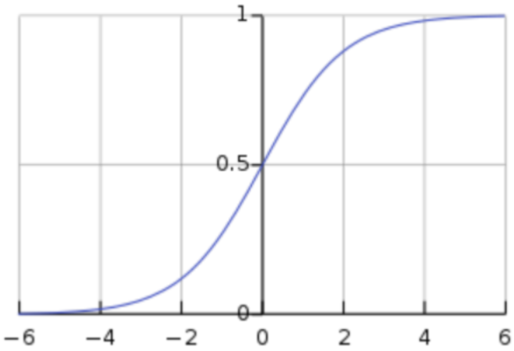


Figure 1: Logistic Function

3.2 *SVM*

SVM is a widely used supervised learning algorithm for model training based on the principle of structural risk minimization. As shown in the Figure, SVM works by finding a hyperplane with a maximum margin which effectively separates sample points. These closest sample points to the hyperplane are support vectors because they play a key role in defining the maximum interval hyperplane.

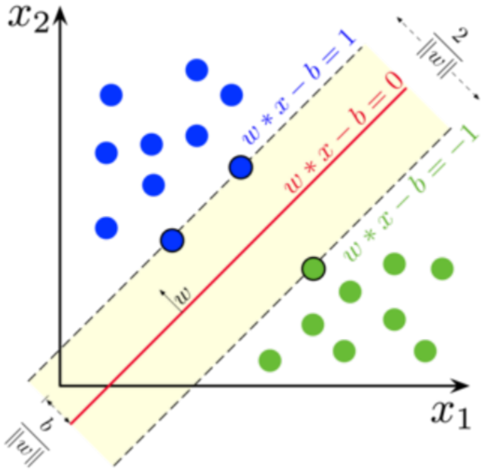


Figure 2: SVM

$$\begin{aligned} \min_{w,b,\xi} \quad & \tfrac{1}{2} w^t w + C \textstyle\sum_{i=1}^{N} \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T(x_i+b)) + \xi_i - 1 \geq 0 \\ & \xi_i \geq 0 \end{aligned} \tag{2}$$

where w is the normal vector of the hyperplane, b the bias term, ξ_i the relaxation variable, and C the regularization parameter balancing the interval and the penalty. And the decision function is like the following.

$$f(x) = sign(\mathbf{w}^T x + b) \tag{3}$$

SVM performs well when dealing with high-dimensional data and smaller sample sets and has better generalization ability for smaller sample sizes.

3.3 *KNN*

KNN is an instance-based classification method that makes predictions based on the labels of nearest neighbor samples. The algorithm compares the similarity of a new sample with the samples in the training set and determines its classification based on the votes of the nearest neighbor samples. KNN has no explicit training process, but requires the calculation of the distance between samples.

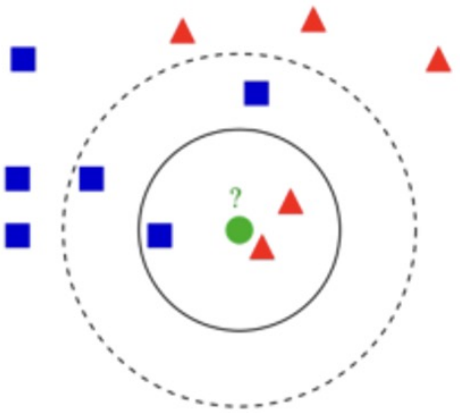


Figure 3: KNN

3.4 *Gaussian NB*

Gaussian NB is a form of Naive Bayes classifier. It assumes that the probability distribution of each feature is Gaussian and uses Bayes' theorem to compute the posterior probabilities. The algorithm is particularly useful when dealing with continuous features and can perform well for high-dimensional datasets.

3.5 *Random Forest*

Random forest is an integrated learning method, based on decision tree construction (as in figure 4). It constructs multiple decision trees by randomly selecting a subset of features and samples with replacement. The final prediction result is based on the collective vote of these decision trees. Random forest has good generalization ability and resistance to overfitting.

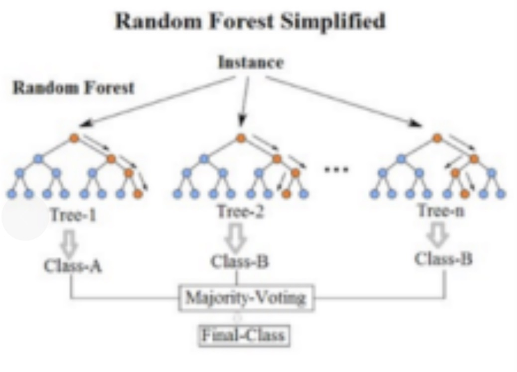


Figure 4: Random Forest

3.6 *GBDT*

Gradient boosting decision tree is an integrated learning method based on decision trees, which improves prediction performance by training multiple weak learners in succession. Each new decision tree is fitted on top of the residuals of the previous trees. With the gradient descent optimization algorithm, each tree tries to reduce the residuals of the previous tree's predictions.

The use of decision trees as weak classifiers makes the GBDT model have better interpretation and robustness, and can automatically discover higher-order relationships between features.

4 *EXPLORATORY DATA ANALYSIS*

In order to predict clients' repayment ability, we select all datasets from Kaggle about applicants' information, including their previous telco and trading behaviors from Credit Bureau and Home Credit. There are totally 307511 pieces of clients' information and 221 features.

4.1 *Numerical Variables*

For the main dataset, $application\_train$, there are 28 numerical variables. And about the loan they have taken: The average value of total income of customers is $168,798, ranging from $25,650 to $117,000,000; The credit those customers hold is $[402,491, 4,050,000]; The annuity is $[1,615.50, 258,025.50].

According to their distributions, we can find that, most of them are close to normal distribution and all data are on the rational range, like the following.

But some of them still have outliers, hence, we treat those irrational outliers by adding some dummy variables to sift by checking the significance of these outliers.

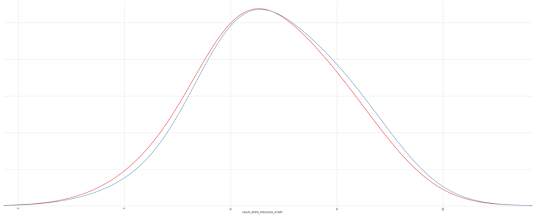


Figure 5: Normal Distribution

4.2 *Categorical Variables*

First, this paper focuses on visualizing and analyzing the dependent variable.

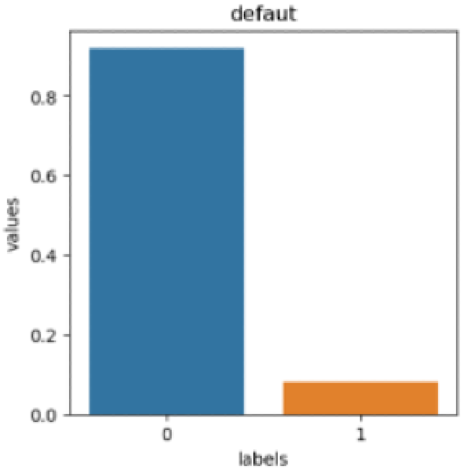


Figure 6: Target Column

From the distribution plot of the TARGET column, we notice that our data suffers from a category imbalance, which may affect the model and lead to predictions that are biased towards non-default categories. Due to the data imbalance, we must use techniques to address this issue.

Then we batch visualize and analyze the 13 categorical features which are in the main dataset, we can get the distribution of each feature of the dependent variable in different cases, and we can see that there are single-category discrete data and multi-category discrete data in our categorical features, which we will deal with separately by choosing different coding methods in the subsequent processing.

About the customer base, we find that: Most customers borrowing from Home Credit are at the maternity leave or unemployed. Losing the source of money is reasonable to explain why they are hard to borrow money from Credit Bureau; Most customers only have lower secondary educational level, but only small part own the academic degree; Most are low-skill laborers.

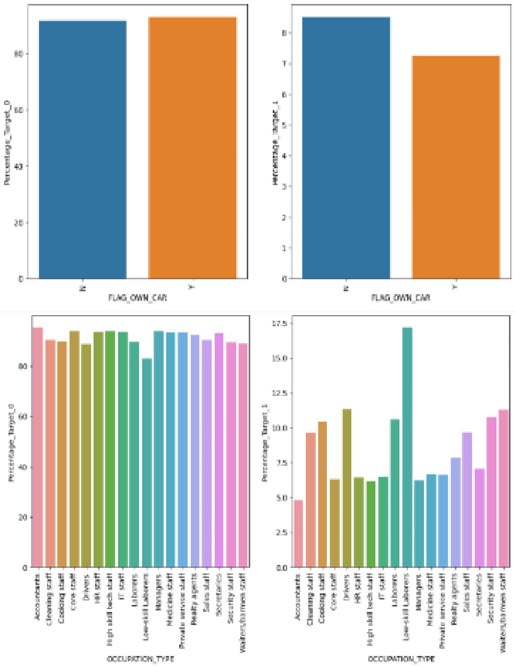


Figure 7: Categorical Features

5 *DATA PROCESSING*

5.1 *One Hot Encoder & Label Encoder*

13 features in the dataset are non-numeric data, therefore such data need to be recoded. Furthermore, the categorical data are divided into two-category and multi-category data, and the data with only two cases of a feature are coded by Label Encoder, but those with multiple cases of a feature need to be coded by One-hot. The classifier often defaults to continuous and ordered data, but for multi-category discrete data, the data is not ordered but randomly assigned. One-Hot coding is a method that uses N-bit status registers to encode N states, each with its own independent register bits, and only one of them is valid at any given time. And if the extra degrees of freedom generated by using one hot encoding will lead to overfitting during model building, the article may choose to deal with it by using regularization.

5.2 *Feature Construction*

In order to make our model with more economical meanings and rational, along with reducing dimensions of the data, we decide to construct some variables. For example, for bureau and bureau_balance, there are new linear-combined variables, like CREDIT_DUARATION, indicating the total length of time for borrowing from Credit Bureau, and also rational-combined variables, like BUREAU_CREDIT_DEBT_RATIO, indicating current debt borrowing as a percentage of total borrowing.

5.3 *Missing Value Padding*

Because many people struggle to get loans due to insufficient or non-existent credit histories, there are lots of unknown information. Thus, it is essential to treat those to make sure model construction is successful.

For missing values occurred at the categorical variables, we apply one-hot method. We face unknown one as one category: if it is unknown, after applying one hot, there not only would treat multi-categorical variables into several dummy variables, but also would have one dummy specified to unknown (1=nan, otherwise, 0).

For that at numerical variables, we apply dummy variable adjustment. We add dummy so that variables with missing values become interactions involving dummy variables (0=nan, otherwise, 1).

$$x_{withmissingvalue} \rightarrow d * x_{withmissingvalue}$$

Both adjustments successfully help us divide unknown values from the known values, but also transform it into a useful indicator.

5.4 *Synthetic Minority Oversampling Technique [SMOTE]*

According to the previous analysis, we can find that the target is unbalanced (like the following), which means the model constructed based on it directly cannot fully depict figures of default ones. We apply SMOTE to solve this problem.

SMOTE “was proposed for enlarging the region of minority class by generating synthetic instances in feature space” (Zhu et. al.,2017). After SMOTE, we can find the effect is obvious through following figures that it is much more balanced.

5.5 *Feature Selection Engineering*

5.5.1 *Data Leakage Processing*

In order to prevent the data leakage, which may lead to great performance in the training dataset but poor in the testing dataset. Therefore, we drop those features with high causality based on experts’ research.

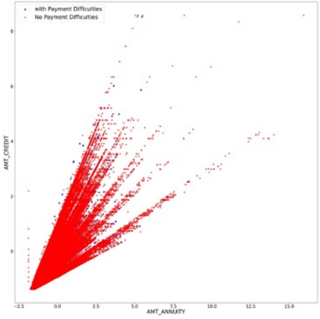


Figure 8: Before SMOTE

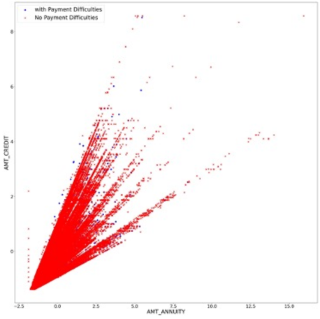


Figure 9: After SMOTE

5.5.2 *Principal Components Analysis [PCA]*

For reducing dimensions of the dataset and making sure final dataset could fully depict characteristics of original one, for categorical variables, we apply PCA.

The goal of PCA is to map the high-dimensional data into the low-dimensional space through some linear projection, and expect to maximize the information content of data in the projected dimension. And finally, we reduced the dimension of categorical variables to 10. We can observe the explained variance ratio is almost 50%.

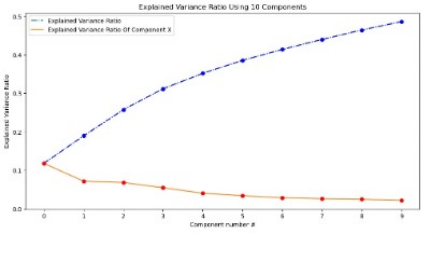


Figure 10: PCA

Even though it is not very high score, as considering correlations, they are not highly correlated (Figure 10). Thus, we think those principal components are rational to select for this research.

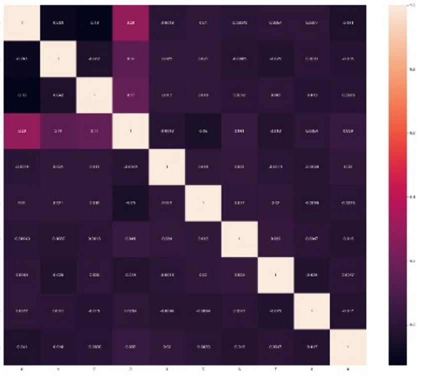


Figure 11: Correlation

5.5.3 *Random Forest*

For numeric features, this paper uses the Random Forest algorithm to rank the importance of the feature independent variables before building a classifier model. This allows the most appropriate number of feature variables and feature variable results to be obtained, making it a more accurate and concise classifier model.

Table below shows some of the numeric features that have an importance higher than the set threshold of 0.01, the rest of the numeric features are not considered in the modeling.

Features	Importance
EXT_SOURCE_2	0.122906
EXT_SOURCE_3	0.112986
INSTAL_PAYMENT_PERC_MEAN	0.060744
BURO_AMT_CREDIT_SUM_DEBT_MEAN	0.047532
BURO_DAYS_CREDIT_MEAN	0.046792
...	...

5.5.4 *Pearson Correlation Coefficient*

The correlation test between two of numeric features after the previous partial feature filtering and visualizing them by thermodynamic diagram can be obtained where the correlation between $BURO\_DAYS\_CREDIT\_MIN$ and $BURO\_DAYS\_CREDIT\_MEAN$, $REGION\_RATING\_CLIENT$ and $REGION\_RATING\_CLIENT\_W\_CITY$ are higher than 0.85. Therefore, only one of the features is selected to represent the information contained in these two features into the classifier model to reduce the correlation of the features in the model.

After normalization and initialization, we start to construct models.

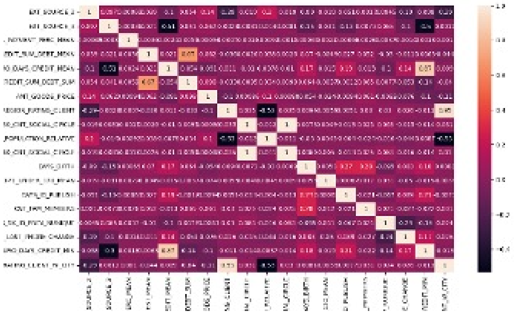


Figure 12: Correlation

6 *MODEL CONSTRUCTION*

After getting the expected data, six different models are trained by the data. The models are trained by conducting cross validation. After the training, different characteristics of the model, including confusion matrix, accuracy, recall, F1-measure and precision etc., are calculated to evaluate the performance of models. What's more, ROC curve and AUC are also verified to have a more accurate and comprehensive view of the performance of different models.

6.1 *Logistic Regression Model*

Logistic regression model is commonly used in classification. Instead of using a linear model, a logistic function is used in the logistic model to control the output between 0 and 1, which means, by defining an appropriate threshold, the output can be classified easily.

After applying 5 folds cross validation, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.6411220904181308
Precision	0.6283850843612647
Recall	0.6283850843612647
F1	0.6360276970544971

What's more, the ROC curve is drawn and AUC = 0.70.

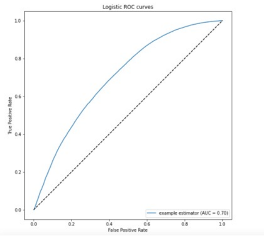


Figure 13: Logistic ROC Curve

6.2 *Support Vector Machine*

The second model is support vector machine, which maps the data to points in space in order to maximize the width of the distance between two categories. 5 folds cross validation is used to train the model. Besides, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.6829469612999081
Precision	0.7172597948863367
Recall	0.7172597948863367
F1	0.6930383359590839

What's more, the ROC curve is drawn and AUC = 0.74.

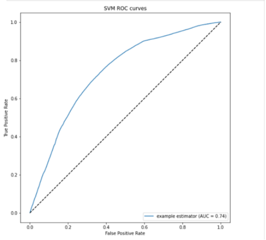


Figure 14: SVM ROC Curve

6.3 *K-Nearest Neighbors*

The third model is k-Nearest Neighbors, which is a non-parametric algorithm that can be used for regression or classification. This algorithm will assign the input to its k-th closest neighbor training data. Here, k is set to be 1 in order to find its closest neighbor. 5 folds cross validation is used to train the model. Besides, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.649329056906351
Precision	0.6562455692613073
Recall	0.6562455692613073
F1	0.6512822316811482

What's more, the ROC curve is drawn and AUC = 0.70.

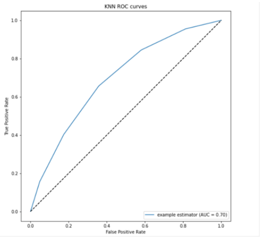


Figure 15: K-Nearest Neighbors ROC Curve

6.4 *Gaussian Naive Bayes*

Gaussian Naive Bayes algorithm is always utilized to conduct the classification. It assumes that each variables are independent that follows Gaussian distribution. Besides, Bayes' theorem is used to calculate the probability of each class. 5 folds cross validation is used to train the model. Besides, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.6622172959460415
Precision	0.6484710997684201
Recall	0.6484710997684201
F1	0.6570573446665868

What's more, the ROC curve is drawn and AUC = 0.72.

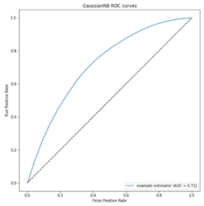


Figure 16: Gaussian Naive Bayes ROC Curve

6.5 *Random Forest*

Random forest is also utilized to have a better prediction of the data. This model is a tree-based algorithm, which means multiple decision trees are trained on random subsets of the dataset and results are combined to get the final prediction. 5 folds cross validation is used to train the model. Besides, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.9184565745858897
Precision	0.8848860227957285
Recall	0.9184565745858897
F1	0.8797834015318446

What's more, the ROC curve is drawn and AUC 0.71.

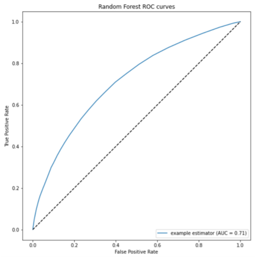


Figure 17: Random Forest ROC Curve

6.6 *Gradient Boosting Decision Tree*

The final model is Gradient Boosting Decision Tree, which builds trees in a sequential order. In this algorithm, each new tree attempts to correct the error of the previous trees, which can have a better prediction of the dataset. 5 folds cross validation is used to train the model. Besides, the accuracy, precision, recall and F1-measure are calculated.

Accuracy	0.9185131731868564
Precision	0.886942645518914
Recall	0.9185131731868564
F1	0.880895869232118

What's more, the ROC curve is drawn and AUC = 0.75.

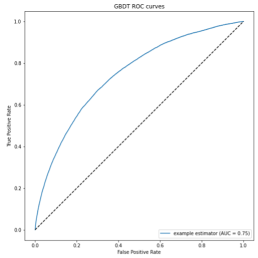


Figure 18: Gradient Boosting Decision Tree ROC Curve

7 *MODEL COMPARISON*

After the construction of the above seven models, its necessary to compare their performance based on their accuracy, precision, recall, F-1 score and AUC. It's obvious that the Decision Tree, Random Forest and Gradient Boosting Decision Tree have the highest accuracy, recall, precision and F-1 score among all the models that are trained to predict the test data. However, by observing the ROC curve of the three models, clearly, the ROC curve of the decision tree is the break line, which achieve the smallest AUC (0.54) while the Random Forest and Gradient Boosting Decision Tree achieve AUC 0.71 and 0.75, respectively. The reason for the result is mainly because the output of the decision tree is discrete instead of continuous so that it can only have a limited number of unique output values. Therefore, based on the overall performance on the training dataset, we decide to use Random Forest and Gradient Boosting Decision Tree to predict the testing dataset.

8 *MODEL PREDICT*

In our previous rounds of training and testing, we observed that Random Forest and Gradient Boosting Decision Trees [GBDT] performed the best in terms of AUC and accuracy. Building upon this insight, we proceeded to fine-tune the parameters and code for application to the actual test dataset, and then uploaded the results to Kaggle to obtain scores.

For Random Forest, we began with feature engineering, selecting the optimal 8 features, followed by further hyper-parameter tuning. Specifically, we set max_depth to 11, $min\_samples\_split$ to 10, and $n\_estimators$ to 1000. With these adjustments, the Kaggle score for Random Forest reached 0.7189.

Subsequently, we conducted parameter tuning and application for GBDT, setting $n\_estimators$ to 100 and $learning\_rate$ to 0.1. With these parameter adjustments, the Kaggle score for GBDT improved to 0.731.

Now, let's analyze why GBDT's performance slightly outperformed Random Forest:

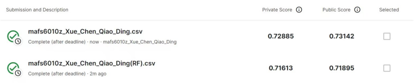
1. Ensemble Method Dynamics: GBDT is an ensemble method that builds decision trees sequentially, each focusing on correcting the errors made by the previous ones. This sequential approach often leads to stronger predictive power, as each tree tries to compensate for the weaknesses of its predecessors. In contrast, Random Forest builds multiple decision trees independently and averages their predictions. While Random Forest reduces overfitting, GBDT's sequential nature can adapt more effectively to complex patterns in the data.

2. Gradient Boosting: GBDT uses gradient descent to minimize the loss function, which allows it to learn from mistakes in a targeted way. This can result in a more efficient and accurate model compared to Random Forest's approach of aggregating multiple trees with equal weight.

3. Parameter Tuning: The parameter tuning process may have favored GBDT in this specific scenario. Setting an appropriate learning rate and the number of estimators (trees) can significantly impact GBDT's performance. By finding the right combination of hyper-parameters, we achieved a slightly higher Kaggle score for GBDT.

9 *CONCLUSION*

In summary, GBDT's performance edge over Random Forest in this context can be attributed to its sequential learning strategy, gradient boosting mechanism, and the fine-tuning of key hyper-parameters. These factors allowed GBDT to capture complex relationships within the data more effectively, resulting in a slightly better Kaggle score.



Submission and Description	Private Score ⓘ	Public Score ⓘ	Selected
mafs6010z_Xue_Chen_Qiao_Ding.csv Complete (after deadline) · now · mafs6010z_Xue_Chen_Qiao_Ding	0.72885	0.73142	□
mafs6010z_Xue_Chen_Qiao_Ding(RF).csv Complete (after deadline) · 2m ago	0.71613	0.71895	□

Figure 19: Kaggle Score

Thus, we can find that, for those customers who have no and insufficient credit information, we could use GBDT model to get the accurate predictions. The most important thing is that, relied on this prediction results, those people do not need to suffer high risk to borrow money from illegal intuitions, which highly improve the security of the society and wellness of people. Besides, according to those features that are highly significant sifted by our data processing stage, those customers know how to improve their credit grade to pursue lower interest rate and more credit amount. And we make sure that, if this model is used and promoted by those creditable companies like Home Credit, the society could be much more stable and the loan system could be much more efficient and valuable.

REFERENCE

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. Risks (Basel), 6(2), 1–20. https://doi.org/10.3390/risks6020038

Bowles, M. (2015). Machine learning in python: essential techniques for predictive analysis. Wiley. Liang, W., Luo, S., Zhao, G., Wu, H. (2020). Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. Mathematics (Basel), 8(5), 765–. https://doi.org/10.3390/math8050765

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications. Journal of Marketing Research, 56(6), 960–980. https://doi.org/10.1177/0022243719852959

Shen, J. (2020). Secure Training of Random Forest Classifiers over Continuous Data. Thesis (Master's)–University of Washington, 2020.

Singhal, V., Chaudhary, Y., Verma, S., Agarwal, U., & Sharma, M. P. (2022). Breast Cancer Prediction using KNN, SVM, Logistic Regression and Decision Tree. International Journal for Research in Applied Science and Engineering Technology, 10(5), 1877–1881. https://doi.org/10.22214/ijraset.2022.42688

Swiderski, B., Kurek, J., & Osowski, S. (2012). Multistage classification by using logistic regression and neural networks for assessment of financial condition of company. Decision Support Systems, 52(2), 539–547. https://doi.org/10.1016/j.dss.2011.10.018

Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. Applied Soft Computing, 10(2), 374–380. https://doi.org/10.1016/j.asoc.2009.08.003

Zhou, Z.-H. (2021). Machine learning. Springer.

Zhu, T., Lin, Y., & Liu, Y. (2017). Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition, 72, 327-340. https://doi.org/10.1016/j.patcog.2017.07.024