**MSc Project - Reflective Essay**

|  |  |
| --- | --- |
| **Project Title:** | A Comparison Of Binary Classification Methods for Loan Eligibility Prediction |
| **Student Name:** | Zhuli Wang |
| **Student Number:** | 210244908 |
| **Supervisor Name:** | Aisha Abuelmaatti |
| **Programme of Study:** | MSc Computing and Information Systems |

This reflective essay is going to review the thorough experience of the dissertation research, analyze the success and failure of the research and propose a future advance for this project.

The project mainly focuses on comparing the performance of 4 different types of classification models, Logistic Regression, Random Forest, Support Vector Machine and Extreme gradient Booster on the specific data set. Analyzing the results of each classification model and discussing the results and concluding the best performance classification model for the specific data set.

At the beginning of project selection, I wanted to try machine learning work in the area of finance or other business because my previous career was related to commercial advertisement data analysis, and I wanted to broaden my experience in other business areas. Then I went to Kaggle.com trying to find a complete data set for research and found a loan approval prediction data set from SBA. After checking the amount and structure of the data set, I found this data set was quite suitable for me a new learner in machine learning to start with. In the literature review, I generally read the relating thesis and understand the current research direction is mainly focused on finding out which machine learning algorithms are best performing under the data structure of specific loan applications. From these researches, I found four types of classification models were of high recommendation and extensive research, then I decided to focus on these four algorithms.

1. Discussion of Method – success and failure

Loan eligibility prediction is attributed to the binary classification problem which is one of the classification problems in machine learning. The rest of the other classification problems are multiclass, multilabel and imbalance classification which is not concerned in this project so far. Before I investigate this topic, I think there should be a mature algorithm being developed and utilized in the financial industry but there is not. Therefore, this area remains lots of key points to be researched.

The first thing I started with the project after the literature review is to investigate the data set. Because I am using Python to code the models, the packages Pandas and NumPy are the tools I am mainly utilizing. I successfully cleaned the null values, reformatted and unified the data types, extracted features from original variables and created several high correlative features to help improve the performance of models. However, there was something more I can do with the preprocessing part. According to the description of the final processed data set, there are some more features that could be created to support the learning.

The first thing is a flag can be set to identify whether the actual disbursed amount is greater than the actual loan approval amount, as we know the extra disbursed part should be the penalty for deferring repaying the loan. This flag feature potentially reveals the loan default risk for each industry, and this also relates to the Revolving line of credit for urgent but bad currency circulation will make the disbursement worse after each loan period.

Second thing is to discover whether the loan is secured by the real estate of the loan applicants. Because with the real estate as collateral, the risk of loan default will be greatly reduced for the value of the estate. However, we can learn from the data that the application period of the data set is from 1984 to 2014 a big event occurred during this period, the great recession from 2007-2009. This means the value of real estate experienced a great decrement which highly increased the risk of loan default.

I think creating these two features to the data set would make the learning model more efficiently and prediction more precise. Because it takes more dimensions of loan approval decisions as features to help train the learning models.

There is a severe failure after fitting the model and predicting. The first one is the imbalance of the data set did not affect the prediction model as I proposed. When analyzing the data set, I assume the imbalance of the target variable might cause a negative impact on the prediction because the distribution bias rate is close to 1:3.7 which is quite severe. And because the sample amount is big enough, the minority class still has 98,382 cases of the sample, I choose undersampling as the method to handle the imbalanced data. I assume that the imbalanced data would lead the model to lean on the majority class which would increase the false negatives and false positives, therefore decreasing the accuracy and precision making the model not worth trusting. However, the prediction results of each model based on undersampling data made me confused when the anticipation of performance metric increase did not happen. Then I tried to tune the hyperparameter to investigate if the default parameters are suitable for the undersampling data or not. But the results still did not meet my proposal, the best score for hyperparameter was default parameters. This means two possibilities, one is the features of the processed data set are completely suitable for these models and the other is the data set was not completely raw, it was being designed already. Also, there is a failure in the hyperparameter tuning. Due to the hardware limitation of my device, I did not simulate parameters as many as possible to find out the optimal solution for each algorithm.

1. Future works

Besides this, comparing the results of LR, RF, SVM and XGBoost, it is easily found that algorithms of ensemble methods have a better performance than the others. The ensemble methods mean there is a general classifier which consists of lots of small classifiers which can significantly help reduce residual loss and help improve general performance. In the loan eligibility prediction problem, we are dedicated to reducing the false prediction even if highly potentially rejecting an implicitly available application in order to avoid default risk, especially when manual audit could intervene. When random forest adopts the bagging technique to average the results from different overfitted classifiers and XGBoost adopt boosting technique to fit several classifiers one by one to iterate the result. These two algorithms might highly suitably meet this principle and therefore can have a better performance.

To improve the performance of the learning model, I think model stacking can significantly help improve the performance based on random forest and xgboost. From the previous discussion, we noticed that ensemble methods will have better performance than logistic regression and support vector machine. However, logistic regression and support vector machine can efficiently separate the binary label though not too precise if the noise distribution is existed, at the meantime the random forest and xgboost can modify mild noise but predict consuming lots of time. Model stacking is a multi-layer ensembled method in which predictions are made by a second layer algorithm. The second layer algorithms take several algorithms models’ predictions as input data to make the optimal predictions. It can build several experimental sets to compare different stacking combinations. I assume that random forest and xgboost to be the first layer algorithms and support vector machine to be the second layer algorithm might potentially achieve the best performance.

To better help the finance institution find their target customer and improve the automation of their business, a recommendation system can be built to combine the prediction model. The requirements of loan approval are varied for different levels of loan and different operation capital amounts of financial institutions. A user-based collaborative filtering technique could be applied to the recommendation system. The assumption of this work is to build a comprehensive loan platform that gathers the loan applications in advance and then labels the features of each application. Based on the participating financial institution’s loan requirement and previous successful cases, the system will recommend a similar application and predict the result. Push the predicted approval cases to the participated institution to promote the loan transaction process.

1. The relationship between theory and practical works

In the data pre-processing part of the method, my practical works followed the general guideline recommendation and were based on specific data situations. In statistics, the missing value should be handled properly because it might cause analysis bias. For machine learning, imputation is a good technique if the data analysis does not sensitive to data bias and is familiar with the missing values. However, I found that apart from the ChgOffDate, the number of other columns’ missing values is relatively small compared to the whole data set, therefore I decided to drop all the null values beside ChgOffDate. The date of loan default has a mild relationship with loan eligibility because we focus on the status of repayment, and it is not necessary to drop ChgOffDate. The reason for not adopting the imputation technique is there are over tens of thousands of samples are left after dropping NA. Also, some missing values occur in important features such as RevLineCr, LowDoc, and MIS\_Status which the wrong imputation might affect the accuracy of prediction.

In machine learning, feature extraction is dedicated to improving the performance of a feature by extracting the raw relation and meaning to identify its useful information. I found many variables’ value is chaotic and contains lots of irrelevant values according to the meaning table. For example, variables RevLineCr and LowDoc contain useless values which can be learned by humans but are hard for a machine. I easily extract the relation of it, which identifies whether an application is a revolving line of credit or a LowDoc loan program or not and convert these two variables into flag fields of value 0/1 to represent its status. We also want to study the relationship between loan approval and business category but found too many unique numbers from variable NAICS which are hard to understand. Therefore, based on the reference NAICS number of each industry, we try to extract the first two digits of NAICS and use them to identify the industry category.

From the results of the models, we can learn that the random forest and xgboost are of best performance among logistic regression and support vector machine. As the key theory of logistic regression and support vector machine demonstrate, both two classification methods are linear models while applying to binary classification, but have different loss functions, logistic loss for logistic regression and hinge loss for support vector machine. Therefore, we can learn that logistic regression and support vector machine has different reason to perform badly. For logistic regression, it will perform badly when the feature space is too large, and the data set contains too many categorical variables. However, to prevent this from happening, we adopt get\_dummy to separate the categorical variables into binary flag fields. Then trimming features with low importance and correlation might be another way to improve the performance of logistic regression. For support vector machine, it will perform badly when there are too many samples and too much noise. This is because when trying to find out the maximum margin between the hyperplanes respectively built by negative vectors and positive vectors, with undersampling data set the best hyperplane might intersect with the best hyperplane of the raw data set which means the prediction has a bias from the undersampling data set. Reviewing the data set being processed, it is not optimal for these two algorithms. Therefore, to improve the performance of linear models, another method of preprocessing should be applied.

1. Conclusion

This dissertation remains some future works to be completed due to a lack of sufficient knowledge, ability, and time. I will continuously be learning and practicing in the machine learning area to advance my professional knowledge and skill which could help me fulfil this topic of research. Also, I need to enhance my ability in literature review which can make my research more efficient and productive.