Machine Learning Report

Qiaotong Huang
College of Engineering
Northeastern University
Toronto, ON
huang.qiaot@northeastern.edu

Abstract

In this report, I explored the performance of supervised learning algorithms on two interesting datasets. The four learning algorithms include Decision Tree, Neural Network, AdaBoost, Logistic Regression. The two datasets are bank marketing and credit score classification. The learning curve, model complexity and training time of each algorithm on both datasets have been explored and analyzed.

1 Datasets

I used two train datasets which are dividually saved in two CSV files.

Table 1: The basic feature of both datasets.

	Data Set Characteristics	Attribute Characteristics	Associated Tasks	Number of Instances	Number of Attributes
Bank Marketing Prediction	Multivariate	Real	Classification	5000	20
Credit Score Classification	Multivariate	Real	Classification	5000	27

1.1 Data characteristics

The histograms of classes from both datasets are shown in Figure 1. Class unbalance is evident in both datasets and must be accounted for in calculating the accuracy scoring function. Weighted scoring function is thus used in my analysis.

Figure 1: The class frequency of two datasets.

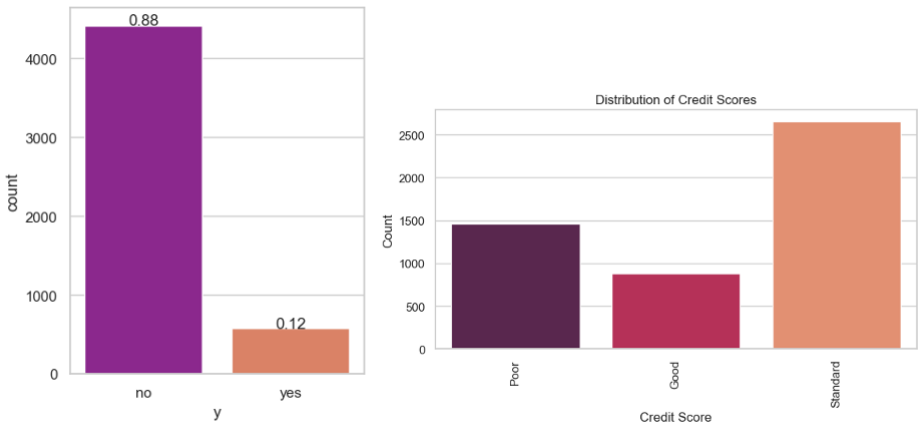


Figure 2: Attribute Characteristics of Bank Marketing prediction

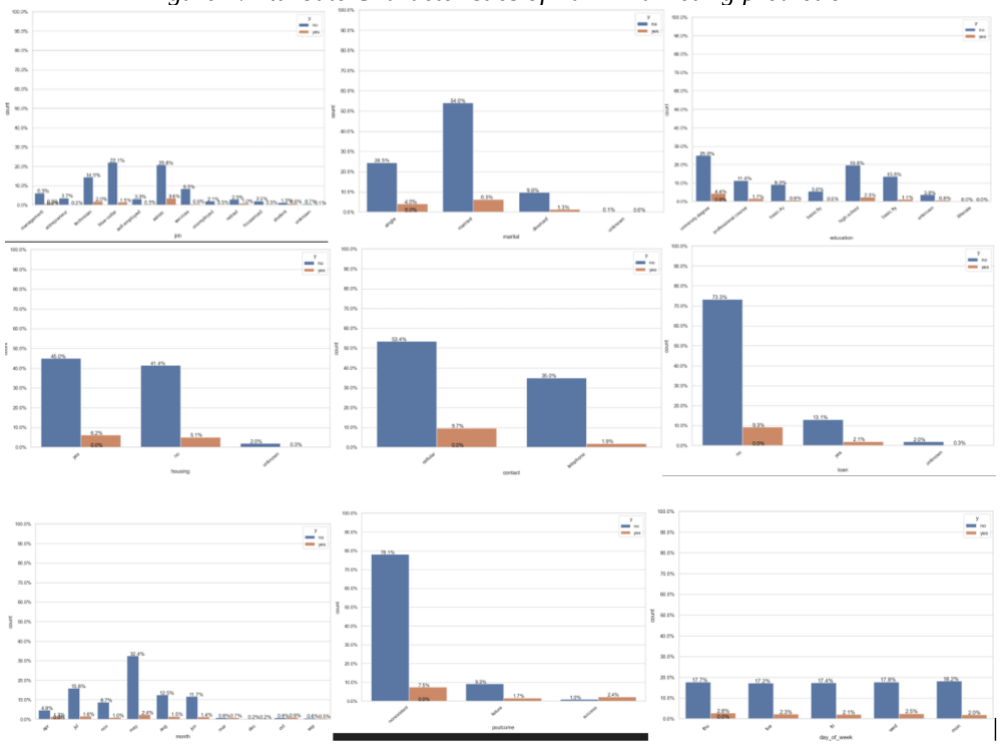


Figure 3: Attribute Characteristics of Credit Score Classification

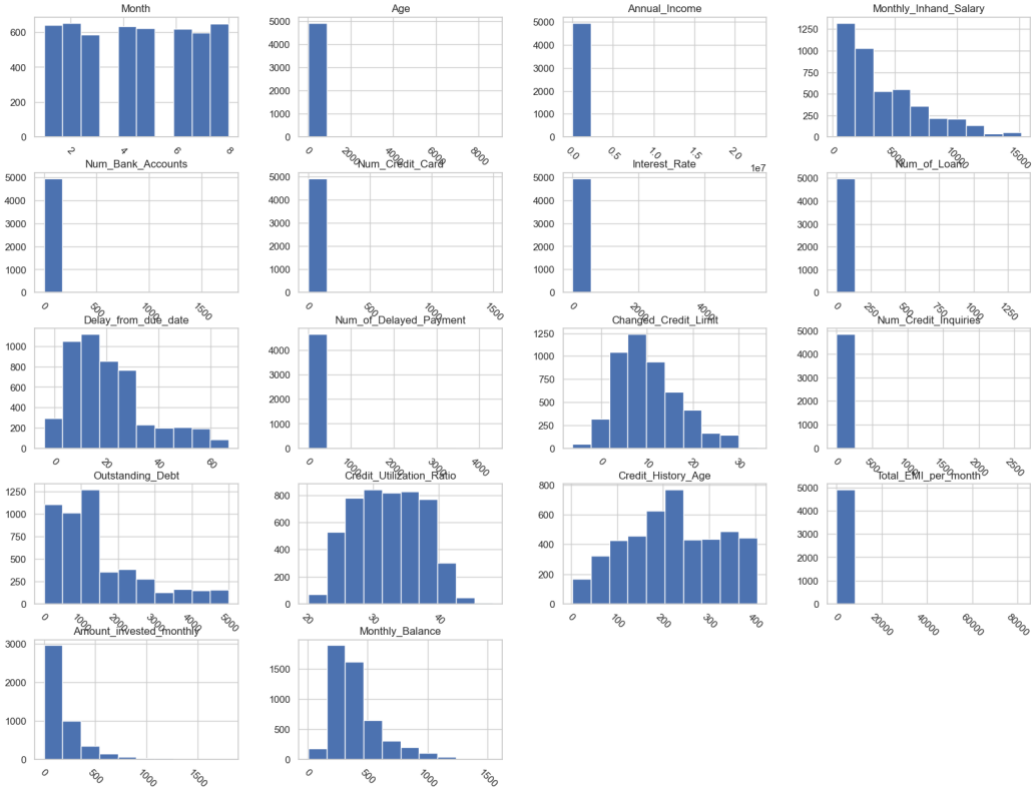


Figure 4: Distribution comparison chart of numerical features of Bank Marketing prediction(Left: Before Process, Right: After Process)

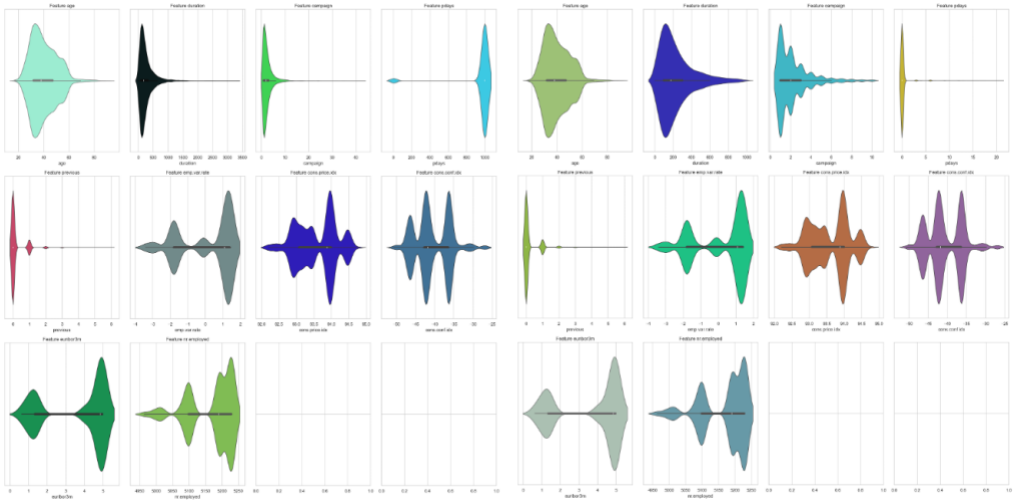
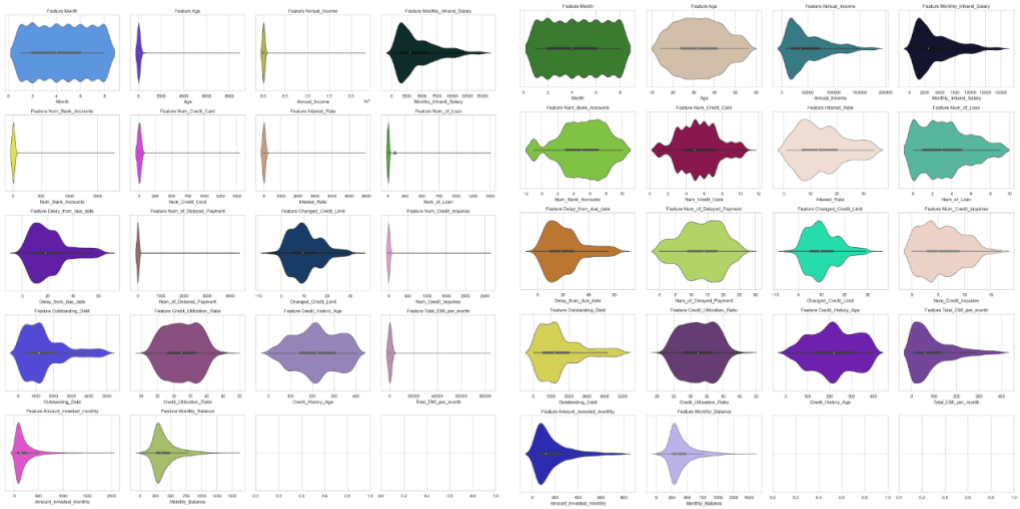


Figure 5: Distribution comparison chart of numerical features of Credit Score Classification(Left: Before Process, Right: After Process)



1.2 Why are these interesting datasets?

Both datasets are interesting for their practical applications. In finance, credit is at the core. A customer's credit rating is the most important basis for deciding whether to grant a loan or approve a credit card. How banks classify customers' credit is a crucial issue, because you don't want to shut out any high-quality customers, nor do you want any potential default risk to affect your own safety. So, we need a prediction model with the highest possible accuracy, which is crucial for banks.

How to judge whether a bank's recommendation of a deposit product to a customer through telemarketing will be successful is a more challenging question. We can see the promotion of financial products in many scenarios. Nowadays, it is certainly not limited to telemarketing, but the basic information of customers is similar. If we can accurately obtain potential users based on user portraits, then the sales of financial products promoted by banks will undoubtedly reach a higher level.

The two datasets have some similarities, as well as differences. They are all based on commercial real data sets, have the same size, some attributes are the same, and their class frequencies are all unbalanced. I want to analyze whether different machine learning

algorithms will perform similarly on these two data sets, and what are the similarities and differences in their learning curves, prediction accuracy, and time consumption. That's what this report can be about.

2 Decision Tree with pre-pruning

All the best classifiers are based on scoring = 'balanced_accuracy'

Figure 4:default decision trees for two datasets(left: Bank Marketing Prediction, Right: Credit Score Classification)

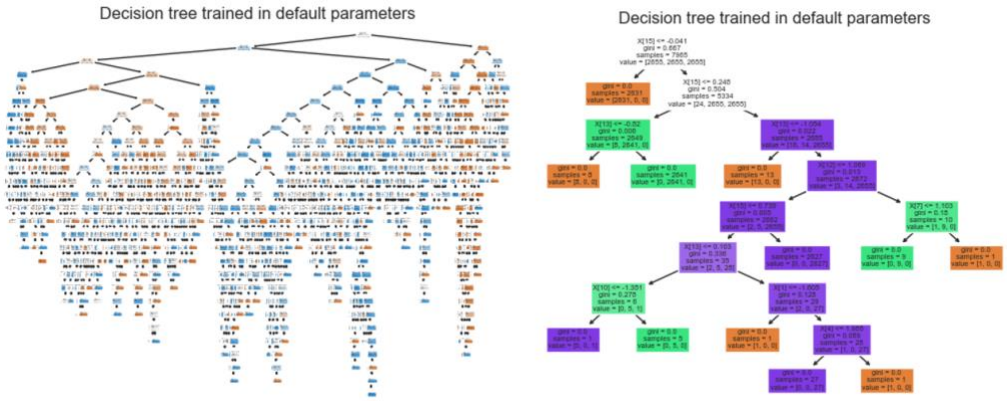


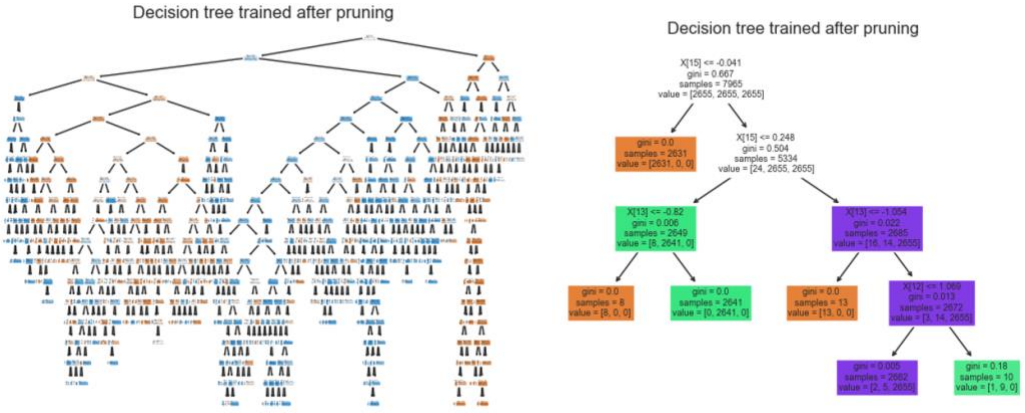
Table 2: grid search best estimator for Decision Tree

Grid Search	Bank Marketing Best Estimator	Credit Score Best Estimator
Max depth	18	5
Min samples leaf	9	6

I chose pre-pruning by controlling the maximum depth of the tree, and the minimum samples per leaf.

The erformance of the decision tree in these two datasets result in a balanced accuracy score of 79.44% for Bank Marketing Prediction and 99.86% for Credit Score Classification.

Figure 5: *Decision trees for two datasets after pre-pruning (Left:Bank Marketing Prediction, Right:Credit Score Classification)*



3 Neural Network(Multi-Layer Perceptron MLP)

Tuning hyperparameters by trying different activation functions, solvers and hidden layer sizes using grid search.

Table 3: grid search best estimator *for Neural Network*

Grid Search	Bank Marketing Best Estimator	Credit Score Best Estimator
activate	logistic	relu
solver	adam	adam
Hidden layer sizes	(10, 10)	(50,)

The performance of the MLP model in these two datasets result in a balanced accuracy score of 81.31% for Bank Marketing Prediction and 99.91% for Credit Score Classification.

In the test of the neural network, I tried several different sets of parameters, and the results obtained by different parameters were obviously different. Although the parameters obtained by greedy search have good results, they are still not guaranteed to be the optimal solution.

4 Boost(adaBoost)

Tuning hyperparameters by trying different n_estimators using grid search.

Table *4*: *grid search best* estimator *for Boost*

Grid Search	Bank Marketing Best Estimator	Credit Score Best Estimator
N estimators	91	11

The performance of the AdaBoost model in these two datasets result in a balanced accuracy score of 81.83% for Bank Marketing Prediction and 99.66% for Credit Score Classification.

In this model, we found the best classifiers for two datasets of the same size, with significant differences in accuracy. However, compared with other algorithms on the same data set, the Boost model is undoubtedly highly accurate.

5 Logistic Regression

The performance of the Logistic Regression model in these two datasets result in a balanced accuracy score of 78.04% for Bank Marketing Prediction and 99.8% for Credit Score Classification

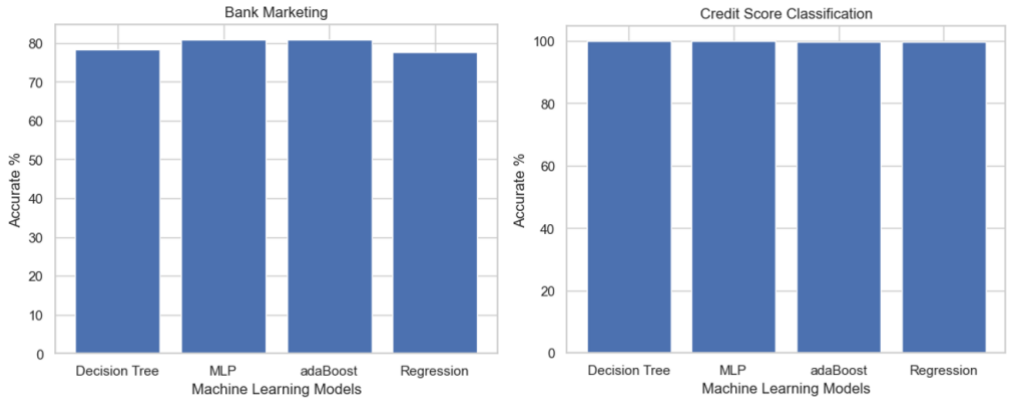
Obviously, the linear accuracy is lower, but for credit score classification, the logistic regression algorithm still achieves high results, which has a lot to do with the quality of the training and testing data sets.

6 Conclusions

In this report, I have analyzed the performance of 4 supervised learning algorithms on two interesting datasets. The balanced accuracy bar graph is shown in fig 6.

First of all, I noticed that there is a clear difference in the accuracy of the two data sets, because the accuracy of credit score classification is close to full score, while the accuracy of marketing success prediction is concentrated around 80%. This is because credit classification problems are easier to predict than marketing problems. This is difficult to change by different machine learning models.

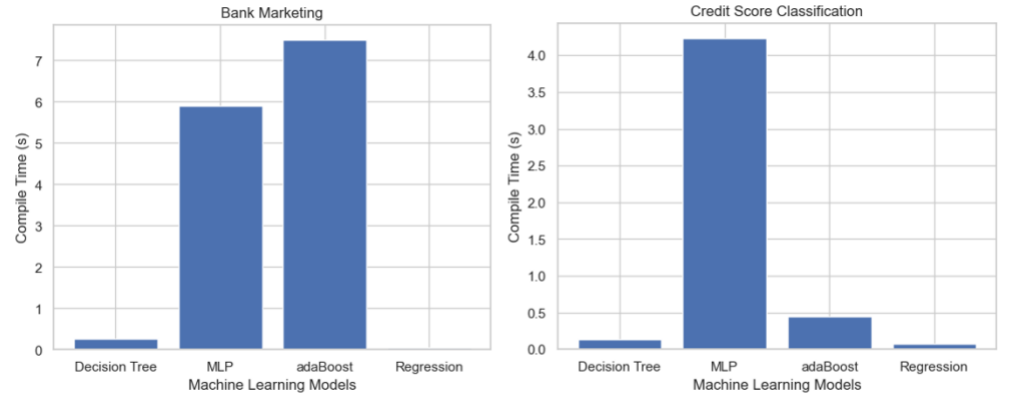
Figure 6: Balanced accuracy of 5 models(Left:Bank Marketing Prediction, Right:Credit Score Classification)



After greedy search and repeated parameter adjustment, these algorithms have achieved a good balance of accuracy. The accuracy of these four algorithms is within an acceptable range, and there is not much difference. This shows that supervised learning algorithms can almost always play a good role when the parameter selection is accurate. Of course, this is also largely due to the fact that the missing data was filled in during data preprocessing and the data was standardized.

Use SMOTE to handle imbalanced classification problems by generating synthetic samples to balance the dataset, ensuring there are enough samples for each category to train the model. This helps improve model performance, especially when dealing with imbalanced data. And analyzing VIF can help determine which independent variables have multicollinearity, so that steps can be taken to improve the stability and interpretability of the multiple linear regression model.

Figure 7: fitting and prediction compile time(Left:Bank Marketing Prediction, Right:Credit Score Classification)



Since the sizes of our data sets are the same, from the comparison of various models, the regression algorithm takes the least time, and the neural network takes a lot of time. However, from the comparison of different datasets, the difference in time consumption of adaBoost is very big. It may be due to the large difference in the selected estimators. At the same time, decision trees take very little time, and this is when the decision tree has a high accuracy.

Acknowledgments

The learning code is adapted from: https://github.com/scikit-learn/scikit-learn and references therein.

References

[1] Sukanta Roy (2019) Machine Learning Case Study: A data-driven dpproach to predict the success of bank telemarketing. https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c

[2] Aya Nada, Kaggle.com "Credit Score Classification" https://www.kaggle.com/code/ayanada/credit-score-classification

[3] Wikipedia "Loan" https://en.wikipedia.org/wiki/Loan