STATS 503 Final Project: Fraudulent Firm Classification

Enhao Li Benhan Liu Lei Zhang Wenjing Zhou

April 28, 2020

1 Introduction and Motivation

Fraud is a critical issue worldwide. Firms which resort to the unfair practices have a grievous consequence on the economy and individuals in the society, while auditing practices are responsible for fraud detection. Hence, the goal of our project is to explore the usefulness of machine learning algorithms for improving the quality of an audit work, and to check the performance of the classification models for the fraud prediction.

We are using the dataset, Hooda, Bawa, and Rana (2018) from UCI machine learning repository. It consists of 777 observations of non-confidential data from 14 different sectors. We plan to predict the fraudulent firm based on the 27 present and historical risk factors. The response variable is "Risk" with two classes, 1 indicating a company having a risk of being fraudulent, and 0 indicating not having a fraudulent risk. The predictors include sector factor, location factor, corporate value, and several risks, such as control risk, detection risk, audit risk, and inherent risk. All predictors are continuous variables.Specifically, we deselect the 'LOCATION ID' variable since it is a categorical predictor with 45 levels, not correlated to our classification target.

Based on the features of the data, We have implemented four classification methods, including logistic regression, support vector machine, classification tree and random forest. We aim to find the one that has the best prediction performance and want to decide the significant predictors for classifying the fraud companies.

2 Exploratory Data Analysis

2.1 Standard Deviation Analysis

We have calculated the standard deviations of predictors in both "Risk = 0" class and "Risk = 1" class. The result is shown in the table 1 below.

	Risk = 0	Risk = 1		Risk = 0	Risk = 1
Sector_score	25.23	21.09	Money_Value	0.71	75.85
PARA_A	0.31	6.53	MONEY_Marks	0.00	1.82
SCORE_A	0.00	1.63	District	0.00	1.46
PARA_B	0.50	63.63	Loss	0.00	0.23
SCORE_B	0.00	1.84	LOSS_SCORE	0.00	0.46
TOTAL	0.59	64.56	History	0.00	0.65
numbers	0.00	0.32	History_score	0.00	0.82
Marks	0.00	0.99	Score	0.00	0.84

Table 1: Standard Deviations of Predictors in Each Class

From the table above, it is notable that the standard deviations for some variables in 'Risk = 0' class are exactly 0, which means observations in this class take a constant value on these variables. This can lead to several findings.

1 The assumption for Linear Discriminant Analysis (LDA) does not hold as the two classes have entirely different covariances.

2 These variables can be useful in building a classification tree. For example, in the class "Risk = 0", the predictor "numbers" takes a constant value on 5, so classifying the cases with "numbers" unequal to 5 as 'Risk = 1' will generate a pure node.

3 If predictors in 'Risk = 0' class are assumed to follow a multi-normal distribution, the covariance matrix may be singular.

2.2 Pairs Plot

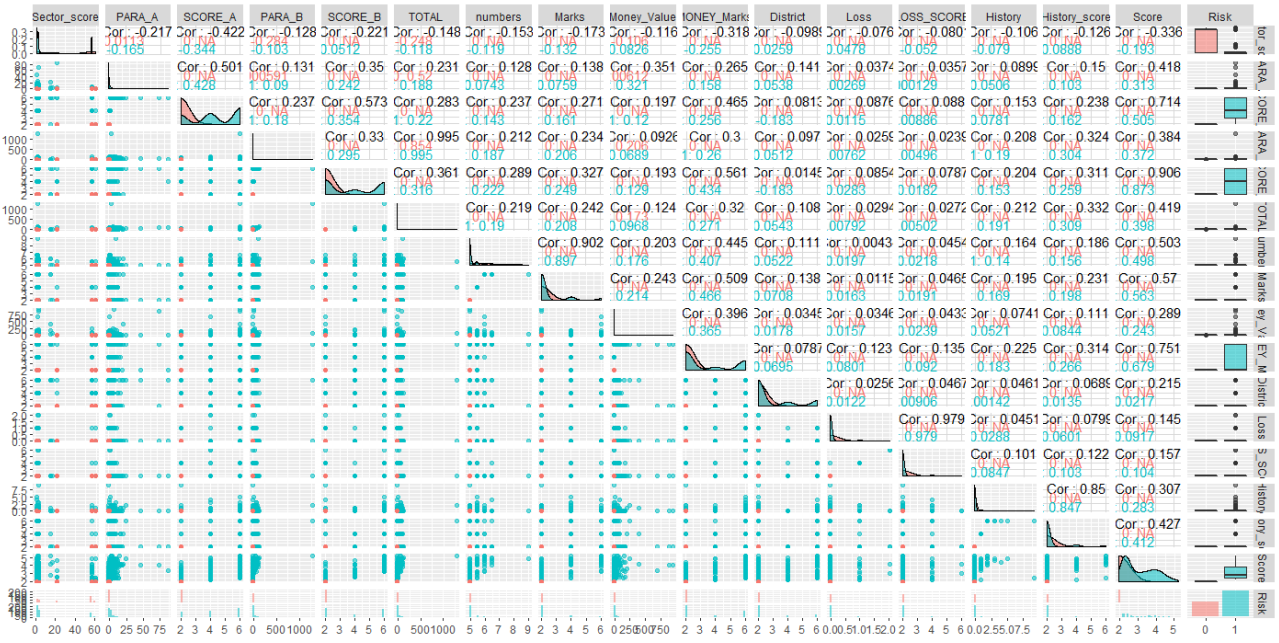


Figure 1: Pairs plot

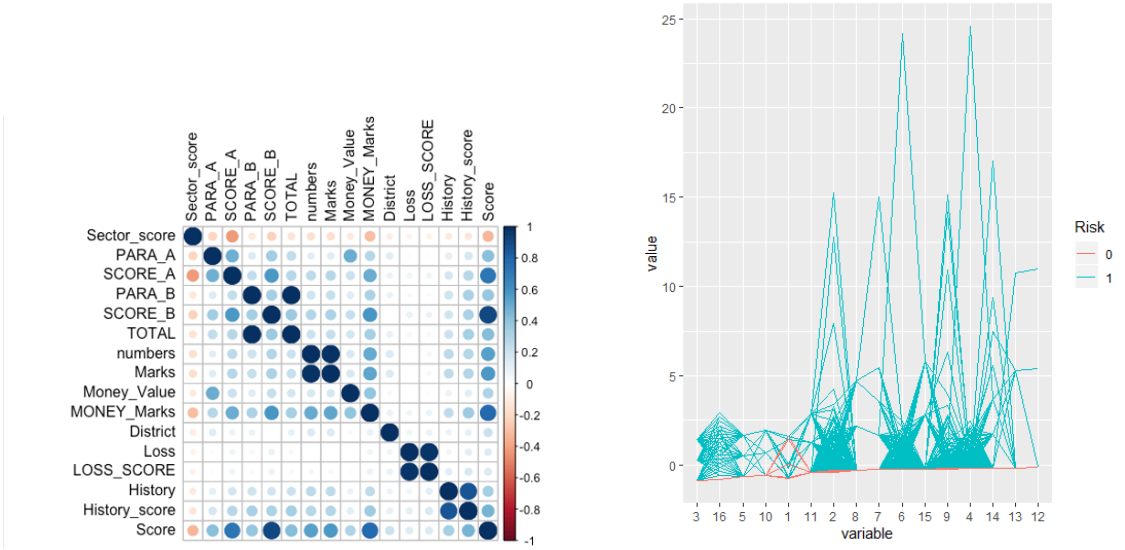
The figure 1 contains a lot of useful information for deciding classification models.

1 As shown in the diagonal plots, the distributions of predictors for different classes are skewed and not alike normal distribution.

2 The rightmost column is the box plots by two classes. The plots indicate that some of the predictors have significantly different means between the two classes.

3 The upper right plots show the pairwise correlations overall and in each class. NA means at least one of the predictors remains constant in the class "Risk = 0". And some of the predictors are highly correlated, such as 'Total' and 'PARA B', indicating a problem of collinearity. Figure 2(a) shows it as well.

4 From the scatter plots in the bottom left, there seems to exist a hyperplane in predictor space, which can divide the two classes.



(a) Correlation Plot

(b) Parallel Coordinate Plot

2.3 Parallel Coordinate Plot

A parallel coordinate plot shows how the cases change among all the predictors. It dose not have a natural order to arrange the variables. In the figure 2(b), the variables are arranged in the order that F-statistics from one way ANOVA decreasing from left to right. We may conclude that the distributions of two classes among these predictors are quite different.

3 Classification Models

Based on the previous data analysis results, we find that discriminant analysis models may not be suitable for our dataset. Aiming to improve the quality of audit work, we have built

four classification models, including classification tree, random forest, logistic regression, and support vector machine.

In order to compare their prediction performance, we first randomly split the dataset into training and testing parts at a ratio about 9:1. Classification models are built on training set, and testing set is used to evaluate their test error. In addition, we have detected there is one observation containing missing value, so we remove this observation before constructing the classification models.

3.1 Classification Tree (CART) and Random Forest (RF)

We firstly perform classification tree on the training data. As shown in the figure 2, which describes the importance of each feature, the predictor "Score" has strong predicting power.

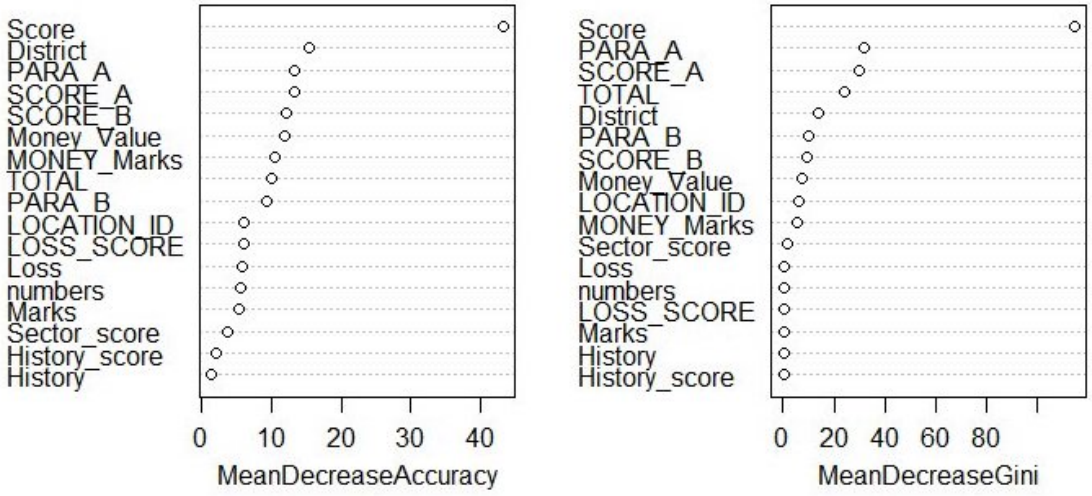


Figure 2: The Importance of the Features

When directly fitting the classification tree, there is only one split by the variable "score", which classify the cases perfectly. All cases in class 0 is 2 in 'score' while all cases in class 1 is strictly greater than 2. By removing the feature predictor with extreme significance, we fit the tree model again ending with 6 splits using 6 different feature variables. The testing error rate for classification tree is 2.56% , which can be considered as a satisfactory result. To further decorrelate the trees, we implement random forest. When building the trees, each time a split in tree is considered,we choose a random sample of predictors. In building a random forest, we don't even consider a majority of the available predictors. It is clever to do so since we have a very strong predictors that can make our trees similar to each other and therefore being highly correlated. The worst case is what we encounter. We only obtain a single tree splitted by one predictor. It is necessary to implement random forest to further decoorealte the trees. The predicting performance is similar to the case in which we eliminate the "score" variable. The testing error rate for random forest is 1.28%, which is better than classification tree as expected.

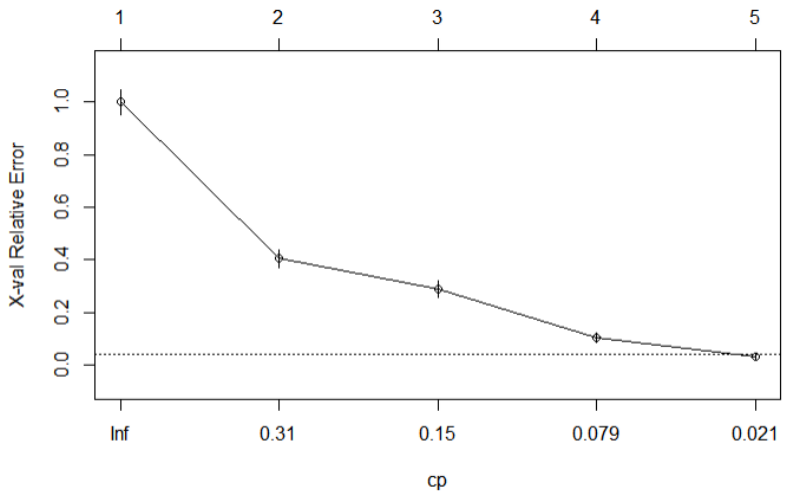


Figure 3: Plot of CP

During process of performing classification tree (CART), we should control tree size to avoid overfitting, and we use the cost-complexity pruning parameter to decide the size of tree. We can perform this by looking at figure 3. Although Cp = 0.079 has a simpler model than Cp = 0.021, its performance is much worse. For Cp = 0.021, the cross validation error is close to the dashed line enough, and accordingly, we choose the tree with number of terminal nodes equal to five.

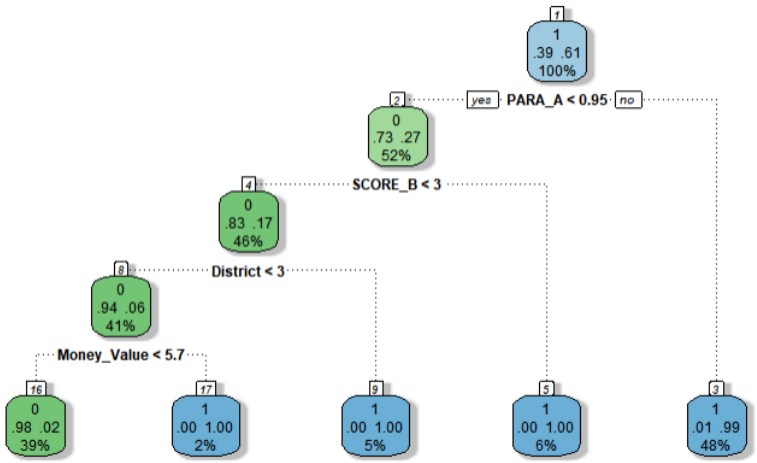


Figure 4: Classification Tree

Figure 4 display the results of fitting and prunning of a classification tree, using four of the features. We can find that these features are also top important features in Figur 2, and using these variables can make a good classification on testing data set.

3.2 Logistic Regression(LR)

We then implement logistic regression(LR) and support vector machine (SVM) classifiers. Given that the variable 'score' gives a perfect prediction result, we deselect this variable

and then construct the models.

For logistic regression, we are only interested in prediction, so we do not drop predictors other than 'score' and 'LOCATION ID', although there exists collinearity. The result is organized in table 2. The test error is 2.56%.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1722.3411	1861077.6427	0.00	0.9993
Sector_score	-0.0061	177.4953	-0.00	1.0000
PARA_A	-8.7869	19405.6584	-0.00	0.9996
SCORE_A	24.1664	7634.6999	0.00	0.9975
PARA_B	-8.0995	15656.3616	-0.00	0.9996
SCORE_B	23.6281	6410.4747	0.00	0.9971
TOTAL	7.9194	15640.0042	0.00	0.9996
numbers	-427.4261	413759.1237	-0.00	0.9992
Marks	129.5412	107539.5302	0.00	0.9990
Money_Value	-0.1948	580.6217	-0.00	0.9997
MONEY_Marks	23.3775	6685.0282	0.00	0.9972
District	23.7279	8764.0501	0.00	0.9978
Loss	11.0272	388471.6058	0.00	1.0000
LOSS_SCORE	18.1057	192922.8794	0.00	0.9999
History	26.5537	76507.7076	0.00	0.9997
History_score	-46.8708	42579.5559	-0.00	0.9991

Table 2: Result of LR

3.3 Support Vector Machine(SVM)

We also try the Support Vector Machine (SVM) classifiers with radial kernel and linear kernel. The optimal parameters are chosen by minimizing the validation errors. Here we use validation set instead of cross validation under consideration of the computational cost. For support vector machine(SVM), we first scale the predictors and similarly use all the predictors except 'score' and 'LOCATION ID'. For the SVM with radial kernel, cost is 1 and the number of support vectors is 80, and we get training error and testing error are both zero. For the SVM with linear kernel, cost is 10 and number of support vectors is 26, and we get training error is 0.01282051 while testing error is 0, which may slightly indicate overfitting .

4 Conclusion and Discussion

To compare the model performance, we arrange the test errors in the table 3.

Classification Model	Test error
Classification Tree	2.56%
Random Forest	1.28%
Logistic Regression	2.58%
SVM (linear and radial)	0%

Table 3: Comparison Among Models

From CART, we can find that the para_A and score_B are the most important predictors in fraud company classification, as they contain the information of discrepancy found in the planned-expenditure of inspection and report.

It is easily understood that random forest outperforms classification tree. In this problem, SVM predicts better than random forest probably because in this dataset there exists a hyperplane that can separate the cases perfectly. And the decision tree is not suitable for smooth boundary.

Although both logistic regression and SVM give linear boundary, it seems that SVM outperforms logistic regression. One of the reason might be that SVM aims to maximize the margin distances and it can easily handle the outliers. Hence, in this problem, SVM is a better approach.

For future works, we are targeting to improve the performance of the classifiers by the ensemble machine learning approach, using a hybrid of the best performing classifiers.

References

Hooda, N., Bawa, S., & Rana, P. S. (2018). Fraudulent firm classification: a case study of an external audit. *Applied Artificial Intelligence*, *32*(1), 48–64.