UNIVERSITY OF MICHIGAN

Fraudulent Firm Classification

-- STATS 503 Final Project Group 22

Motivation

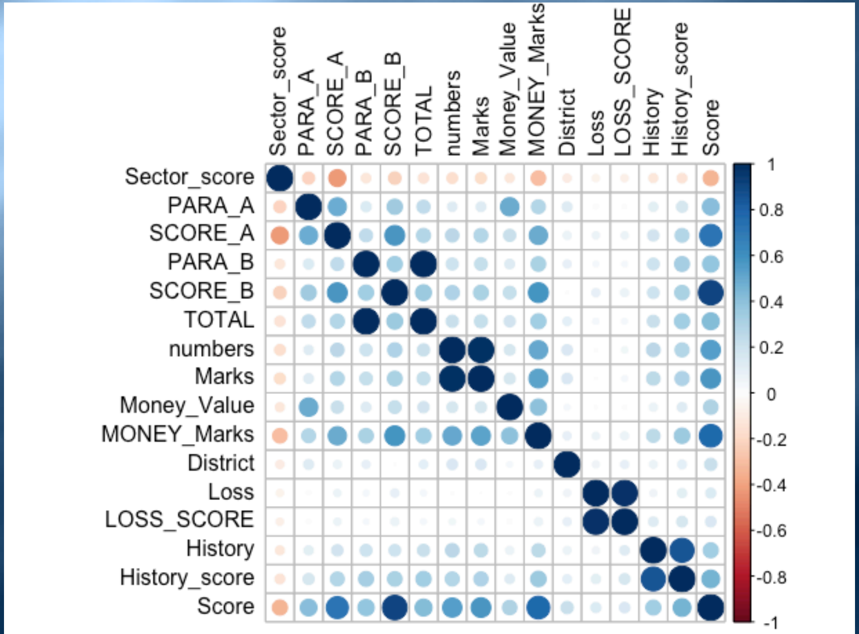
- Fraud is a critical issue worldwide.
- Auditing practices are responsible for fraud detection
- To explore and test the applicability of classification models in the prediction of a ***Risk*** class

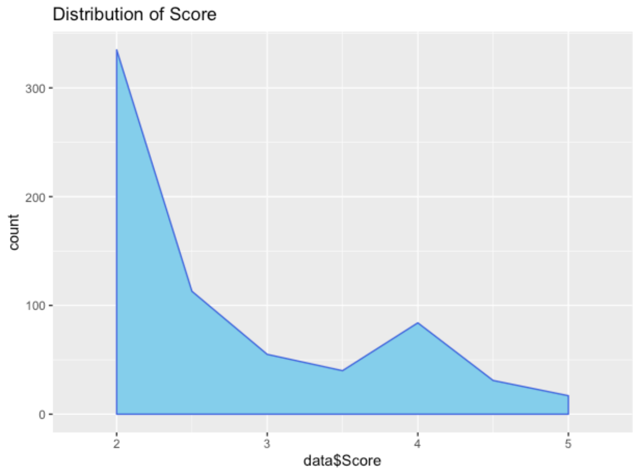
Dataset

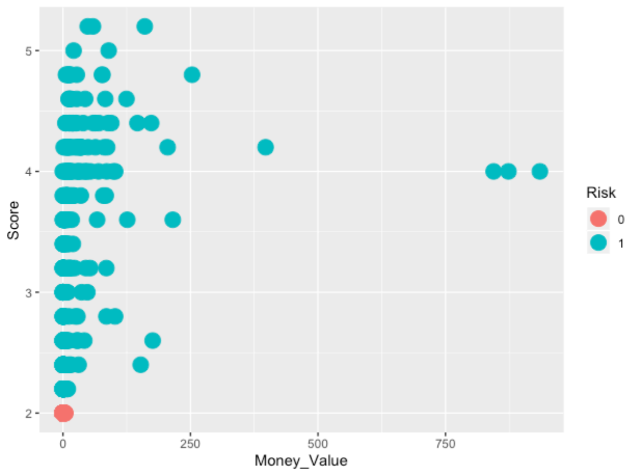
- UCI Machine Learning Repository
- Data Cleaning: Missing Value
- The number of observations: 776
- The number of Variables: 18
- Response/Target : ***Risk*** Class

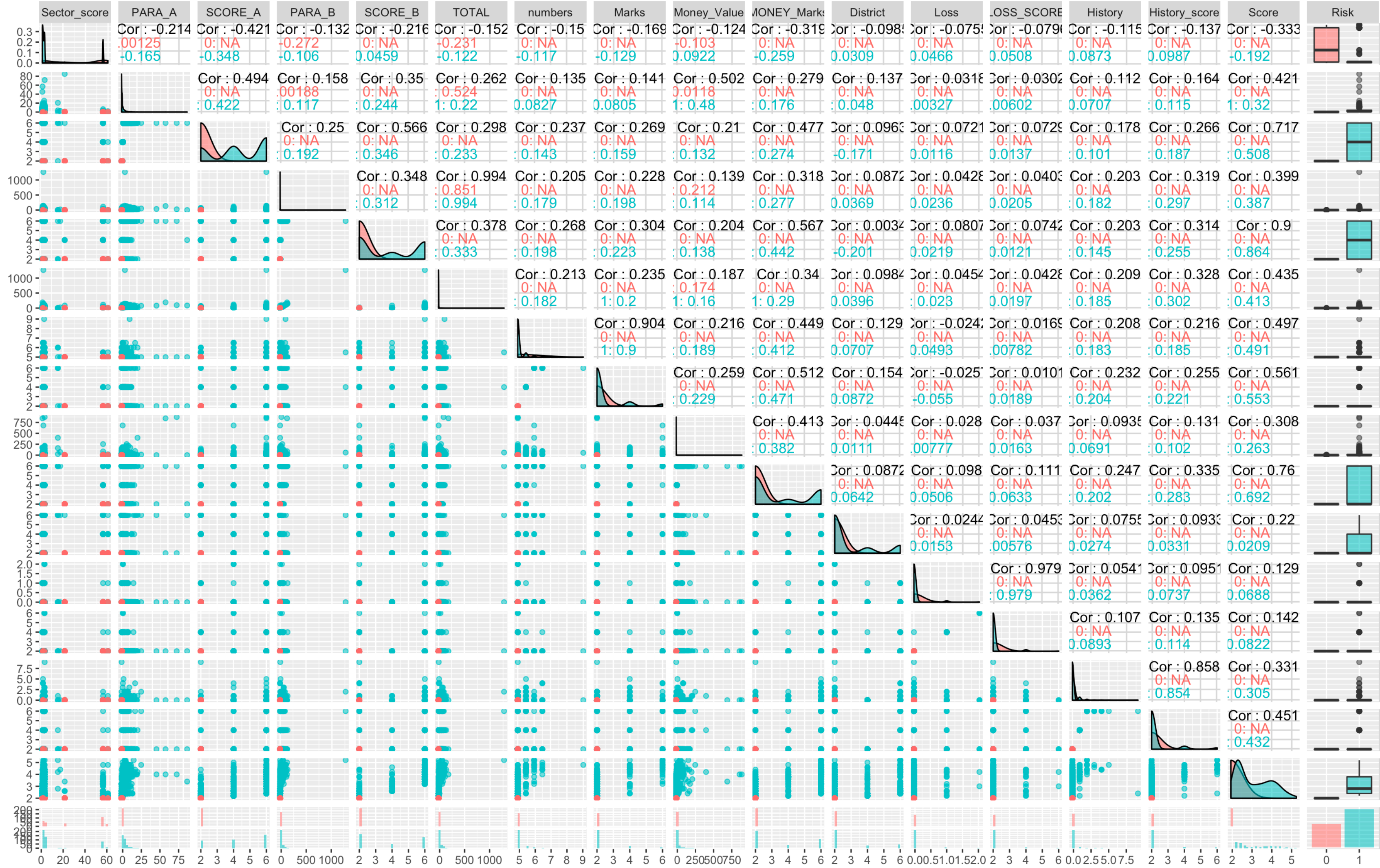
Predictor	Information
Para A value (in Rs in crore)	Discrepancy found in the planned-expenditure of inspection and report A
Para B value (in Rs in crore)	Discrepancy found in the unplanned-expenditure of inspection and report B
Total	Total amount of discrepancy in other reports
Number	Historical discrepancy score
Money value	Amount of money involved in misstatements in the past audits
Sector score	Historical risk score of the target unit
Loss	Amount of loss suffered by firm last year.
History	Average historical loss in last 10 years
District score	Historical risk score of a district in last 10 years
Sector ID	Unique ID of the target sector
ARS	Total risk score using analytical procedure
Location ID	Unique ID of the city/province
Audit ID	Unique ID assigned to an audit case

Exploratory Data Analysis



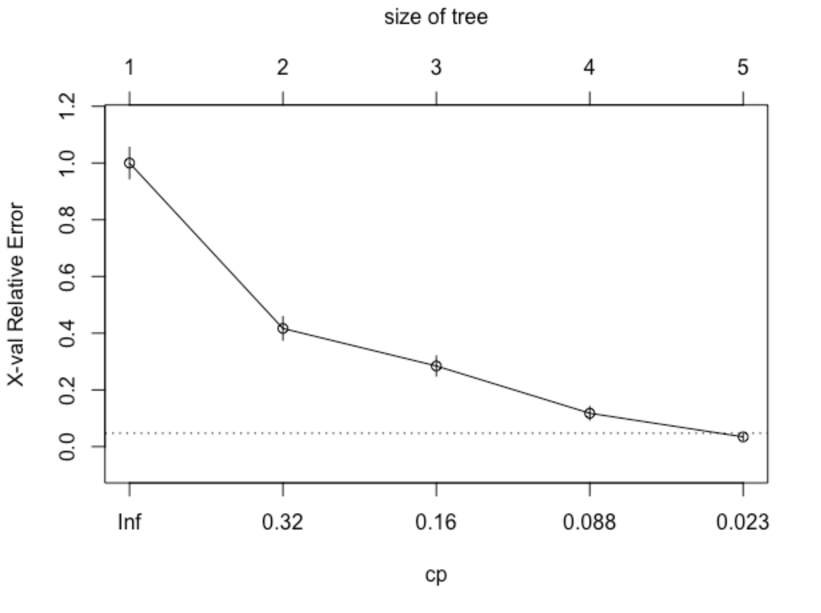

Distribution of Score



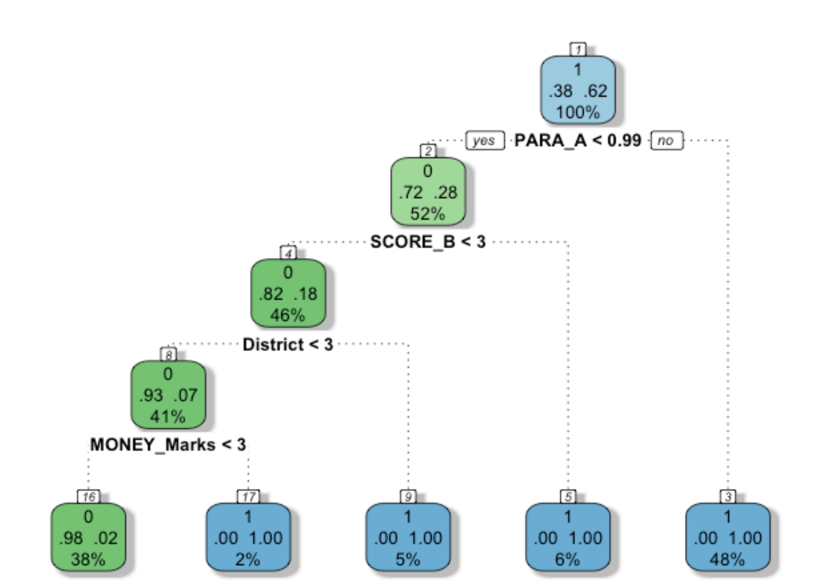


Explore Data

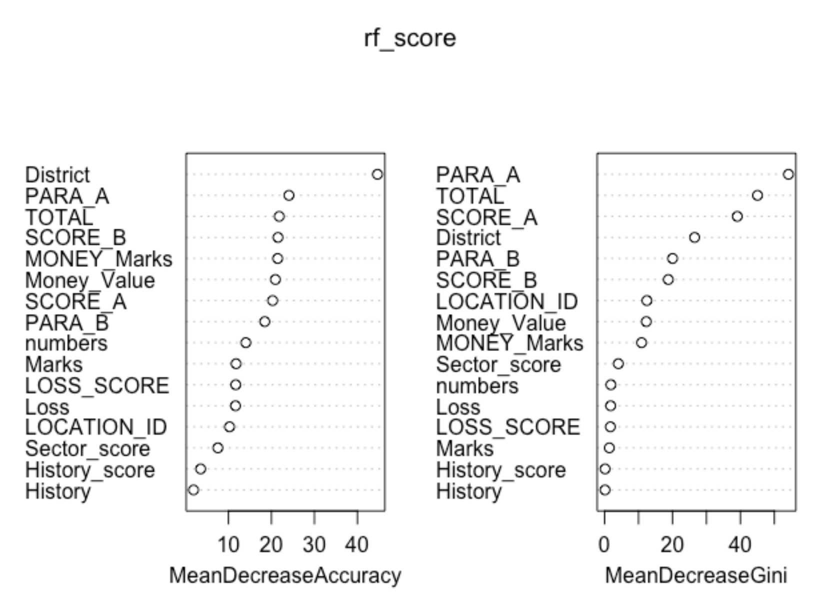
Classification Tree



Classification Tree



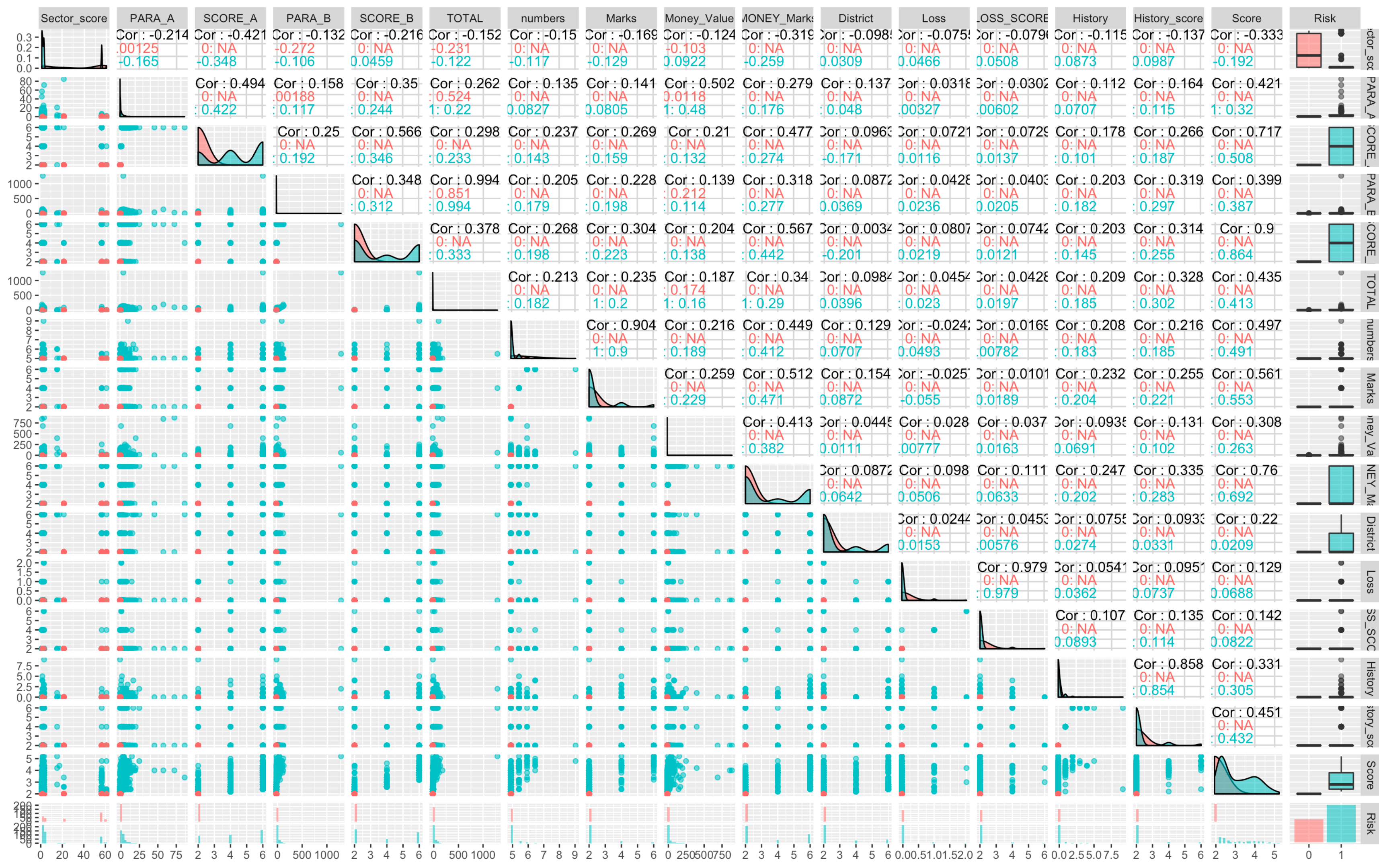
Random Forest



Model Comparison

Model	Training Error	Testing Error
LDA	17.517 %	22 %
QDA	12.643 %	16 %
Classification Tree	0.368 %	1.716 %
Random Forest	0	1.288 %

- The data size really matters.
- The distribution may ***not*** satisfy the prerequisites of LDA and QDA.
- Plan to try ***Naive Bayes*** and ***Logistic Regression***



Explore Data

Model Comparison

Model	Training Error	Testing Error	
LDA	17.517 %	22 %	✖
QDA	12.643 %	16 %	
Classification Tree	0.368 %	1.716 %	
Random Forest	0 ?	1.288 %	★

- The data size really matters.
- The distribution may ***not*** satisfy the prerequisites of LDA and QDA.
- Plan to try ***Naive Bayes*** and ***Logistic Regression***

Thank you!