**Capstone Project – Report**

**Prediction of the Fraudulent Firm**

**Nneamaka Okafor**

1. **INTRODUCTION**
   1. **BACKGROUND**

This is based on the case study of visiting an externl audit company and knowing how Machine Learning would help in improving the audit quality.

### **1.2 PROBLEM**

Fraud is one of the main issues in the whole world. People find different ways to overcome frauds. This problem shows how the machine learning helps in overcoming the fraud.  
Fraud consists of dishonest and illegal activities perpetrated by the companies in order an advantageous financial outcome to the people in the firm or the firm.

Audit is an official inspection of an individual’s or organization’s accounts, typically by an independent body. 763 firms have been targeted in this project.

This project would help me gain knowledge on how which factors would effect the risk of the fraud, what are the laws that the firm has to follow and which would help me improve my knowledge in analysing the data and help me understand how Machine Learning can be used in finding the fraudulent firm.

1. **DATA ACQUISITION AND ANALYSIS**
   1. **Data Sources**

I found the data from the UCI Machine Learning Repository. I also took some help from my professor and have done some research for the missing values.

* 1. **Data set information:**

This dataset is taken from a research explained here.

The goal of the research is to help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors. The information about the sectors and the counts of firms are listed respectively as Irrigation (114), Public Health (77), Buildings and Roads (82), Forest (70), Corporate (47), Animal Husbandry (95), Communication (1), Electrical (4), Land (5), Science and Technology (3), Tourism (1), Fisheries (41), Industries (37), Agriculture (200).

There are two csv files to present data. Merged these two datasets into one dataframe.

Considered ``Audit\_Risk`` as target columns for regression tasks, and ``Risk`` as the target column for classification tasks.

There were two data files for which I had to create the dataframes and merge them both.

* 1. **Description of Data:**

There are two files for this data. So I merged both of them and the total size came upto:

The number of observations = 763

The number of features = 33

The data features are as follows:

1. SECTOR\_SCORE

2. LOCATION\_ID

3. PARA\_A

4. SCORE\_A

5. RISK\_A

6. PARA\_B

7. SCORE\_B

8. RISK\_B

9. TOTAL

10. NUMBERS

11. SCORE\_B.1

12. RISK\_C

13. MONEY\_VALUE

14. SCORE\_MV

15. RISK\_D

16. DISTRICT\_LOSS

17. PROB\_1

18. RISK\_E

19. HISTORY

20. PROB

21. RISK\_F

22. SCORE

23. INHERENT\_RISK

24. CONTROL\_RISK

25. AUDIT\_RISK

26. RISK\_x

27. MARKS

28. MONEY\_MARKS

29. DISTRICT

30. LOSS

31. LOSS\_SCORE

32. HISTORY\_SCORE

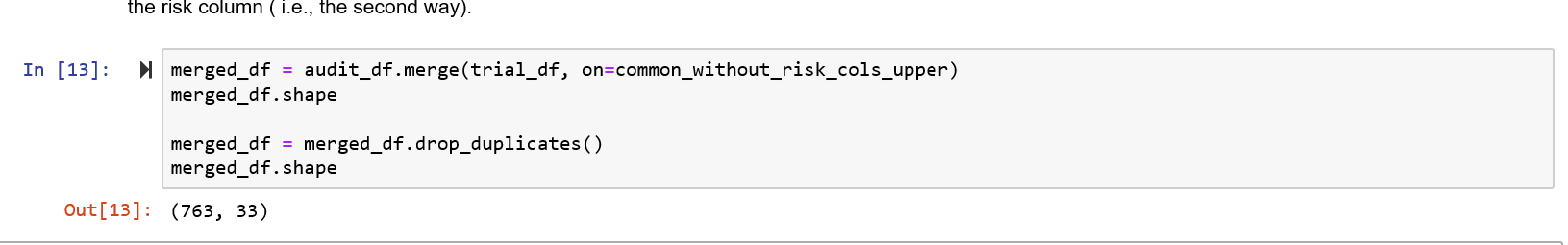
33. RISK\_y'

1. **Methodology:**

While analysing the data I observed that there we some inter relations between the columns which are as follows:

1. Score\_A feature and Score\_B feature in the trial\_df (one of the data frame) are equal to 10 \* Score\_A and 10 \* Score\_B of the audit\_risk\_df (another dataframe).
2. Also I observed that the Detection\_risk column had zero variance and hence dropped that feature.
3. It was observed that on merging the two dataframes there would be two columns formed for the risk as the risk columns and values are different for both the dataframes. Hence by performing the OR operation the RISK column was built.
   1. **Merging of the two dataframes**

I merged the two dataframes using the inner merge on the common columns between the two dataframes while not considering the “risk” column.



* 1. **Central Imputation**

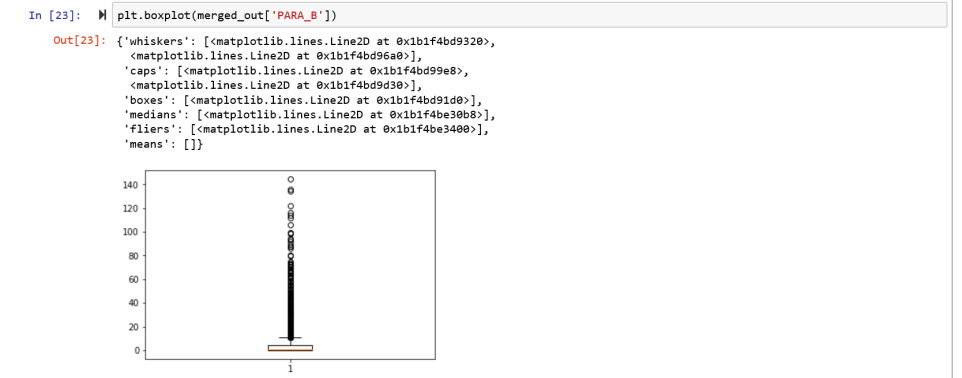
There was only one null value present in the dataframe in the feature MONEY\_VALUE. This null value was replaced by the mean of MONEY\_VALUE.

* 1. **Type Conversion**

Some of the observations for the feature LOCATION\_ID were in strings. I have assigned and converted it into the categorical variable.

* 1. **Dealing with the Outliers:**

I have used describe() function to observe the statistics of the dataframe. The I observed the outliers by plotting the boxplot and removed some of the outliers.



There were outliers in PARA\_B, INHERENT\_RISK, TOTAL, MONEY\_VALUE, RISK\_D

* 1. **Dealing with the data leakage:**

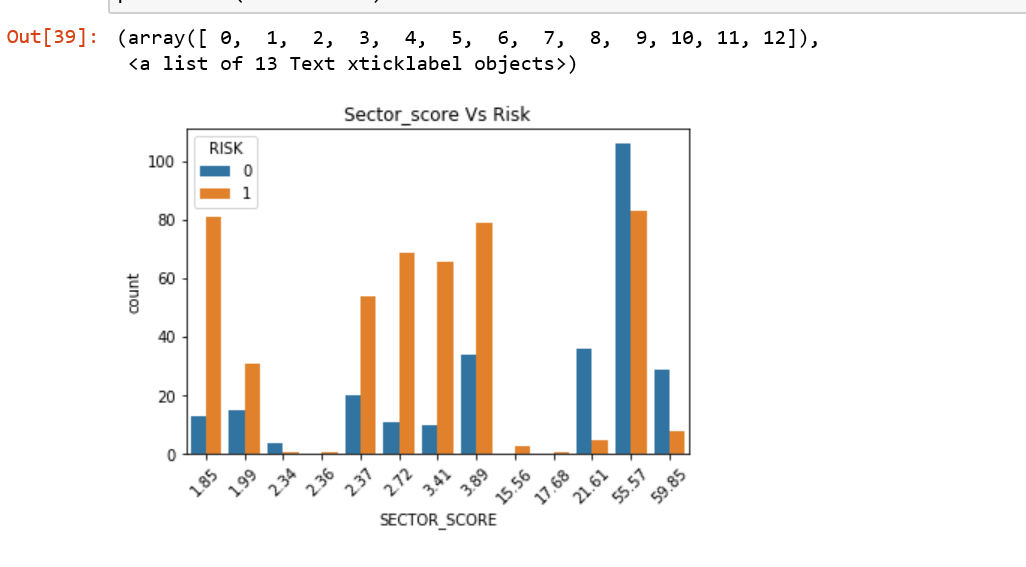
It is observed that District loss and district have the same values and same effect on the target so dropping one of the two attributes (which is DISTRICT) is done.

It is also observed that MONEY\_Marks and Score\_MV differ by a constant factor which is the multiplication of 10 (i.e., SCORE\_MV\*10 is \MONEY\_MARKS), so dropping the MONEY\_MARKS attribute.

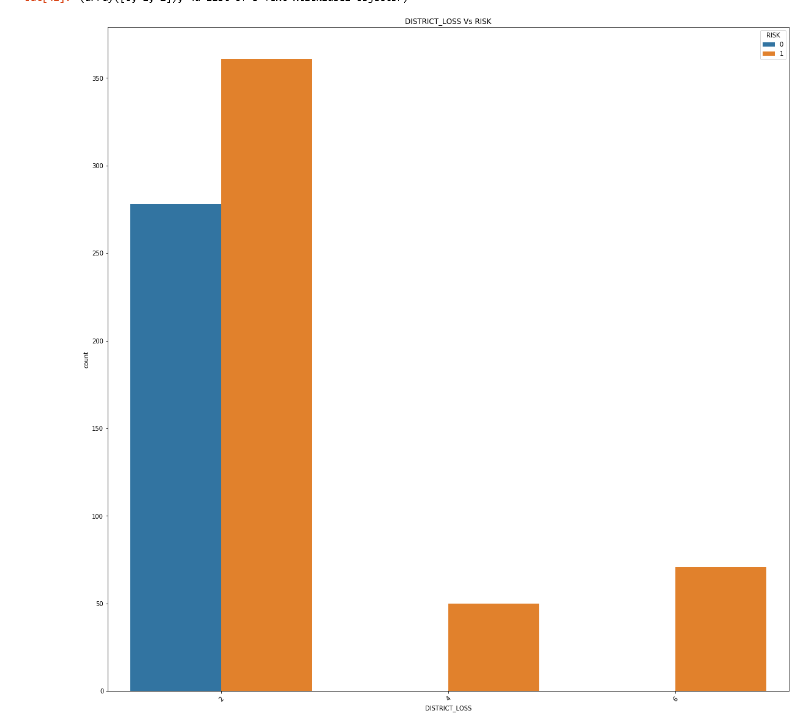
1. **Visualizations**

With the help of visualisations I could observe some of the following:

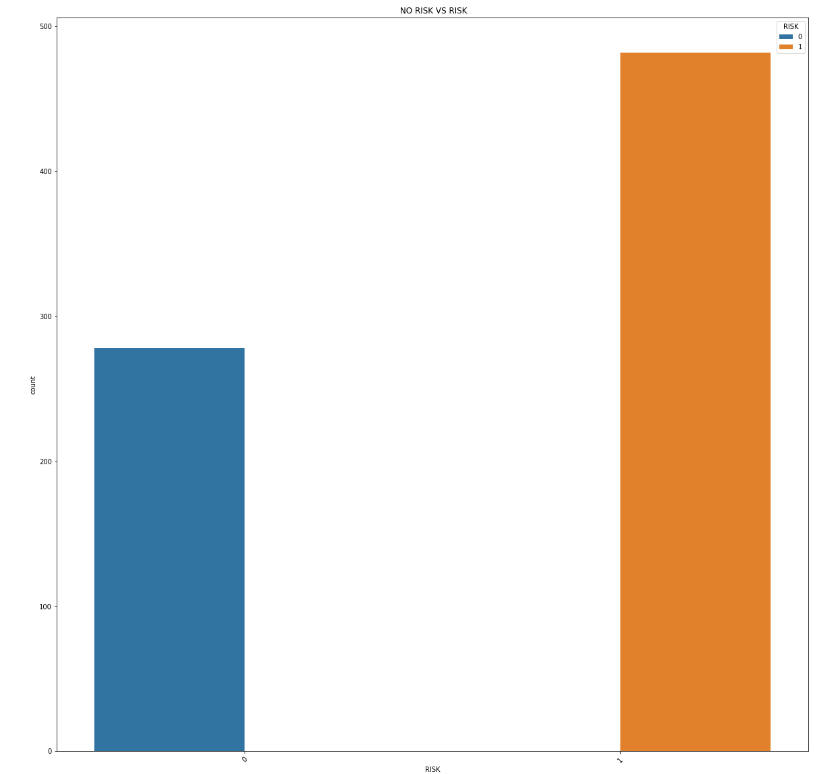
1. There was risk that the firm is a fraudulent firm for sector\_score between 2.72 and 3.89.



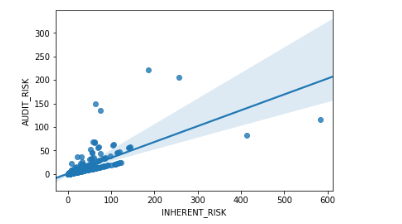
1. It was observed that district loss = 2 had less fraudulent firms.



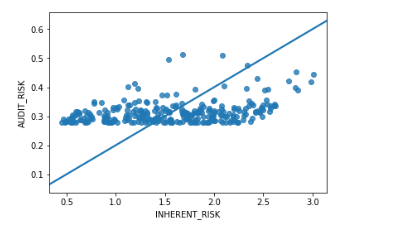
1. Number of Fraudulent (i.e., risk) and Number of non\_fraudulent firms (i.e., no risk)



1. With the help of scatter plot it is also observed that there is a good linear relationship between INHERENT\_RISK and AUDIT\_RISK when the RISK =1 .



1. With the help of scatter plot it is also observed that there is the linear relationship between INHERENT\_RISK and AUDIT\_RISK when the RISK =0 is not good.



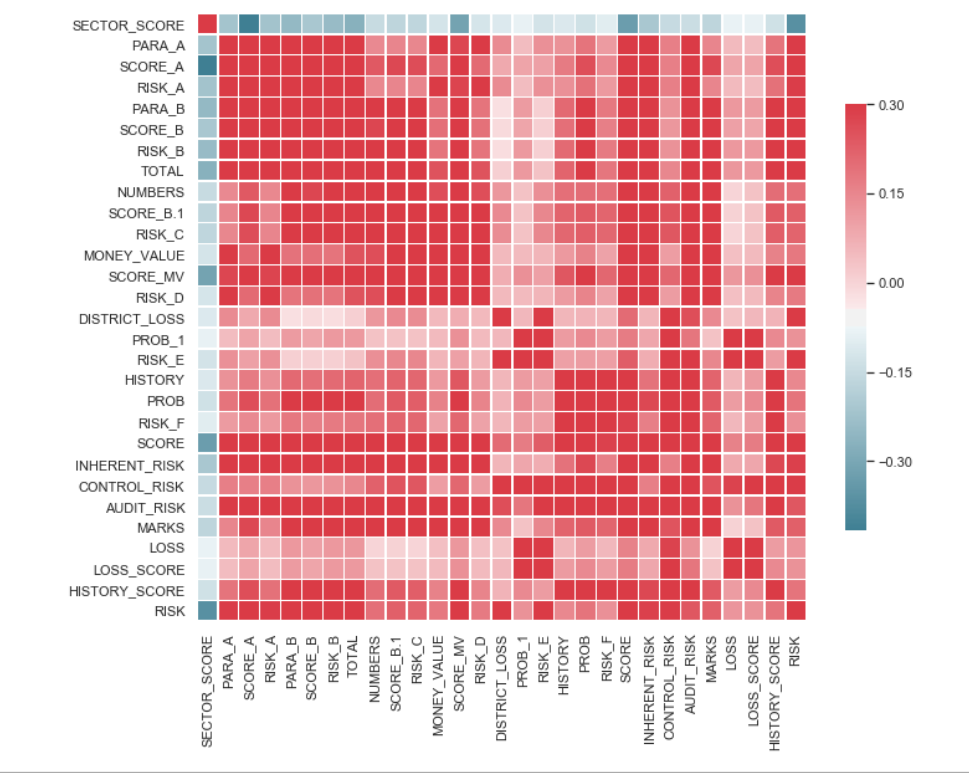
1. **Feature Scaling**

The data is divided into to\_scale\_x\_df and y\_final\_reg which is the target.

Then the data is scaled using both min\_max\_scaler and standard\_scaler.

1. **Correlation Matrix**

I also analysed the data and the correlations using the correlation matrix.



1. **Regression Models**

- Applied all the regression models. Used Grid Search to find the best scaling parameter. Used plots and graphs to help get a better glimpse of the results.

- Used cross validation to find average training and testing score.

- Found the best regressor for this dataset and train your model on the entire dataset using the best parameters and predict buzz for the test\_set.

- Performed evaluation for the regression models.

1. **Classification Models**

- Found the best parameters for following classification models: KNN classifcation, Logistic Regression, Linear Support Vector Machine, Kerenilzed Support Vector Machine, Decision Tree.

- Performed evaluation for the classification models.

1. **RESULTS:**

After performing the regression and classification models , the polynomial regression gave the best results with the minimal error of 1.15. The classification models also predicted well with an accuracy of 100%.

1. **Conclusion:**

After performing all the regression models it is observed that polynomial is giving the best results as the mean square error is less and the scores are optimal.

It shows the best results against the other models

And according to the classification, all the models are giving an accuracy of 100% with very-less or no miss-classification.

According to the Feature Importance found by the Random Forest, the SCORE and INHERENT\_RISK are the most important features which effect the audit risk.