AUDIT RISK MODELLING TO PREDICT FRAUDULENT FIRMS

INDEX

Data Analysis of Audit Risk Dataset

Plots

Data Analysis of Audit Risk Dataset

Problem Statement: To help the auditors by building a classification model that can predict the fraudulent firm on the basis the present and historical risk factors.

Descriptive Analysis: The dataset has 776 records with 27 variables out of which ‘Risk’ is the target variable i.e. a binary variable where ‘0’ represents Non-Fraudulent firm and ‘1’ represented Fraudulent firm. It is a combination of continuous variables like 'Sector_score', 'PARA_A', 'Score_A', 'Risk_A', 'PARA_B', 'Score_B', 'Risk_B', 'TOTAL', 'numbers', 'Score_B.1', 'Risk_C', 'Money_Value', 'Score_MV', 'Risk_D', 'PROB', 'RiSk_E', 'Prob', 'Risk_F', 'Score', 'Inherent_Risk', 'CONTROL_RISK', 'Detection_Risk', 'Audit_Risk', 'Risk' & discrete variables like 'LOCATION_ID', 'District_Loss', 'History'.

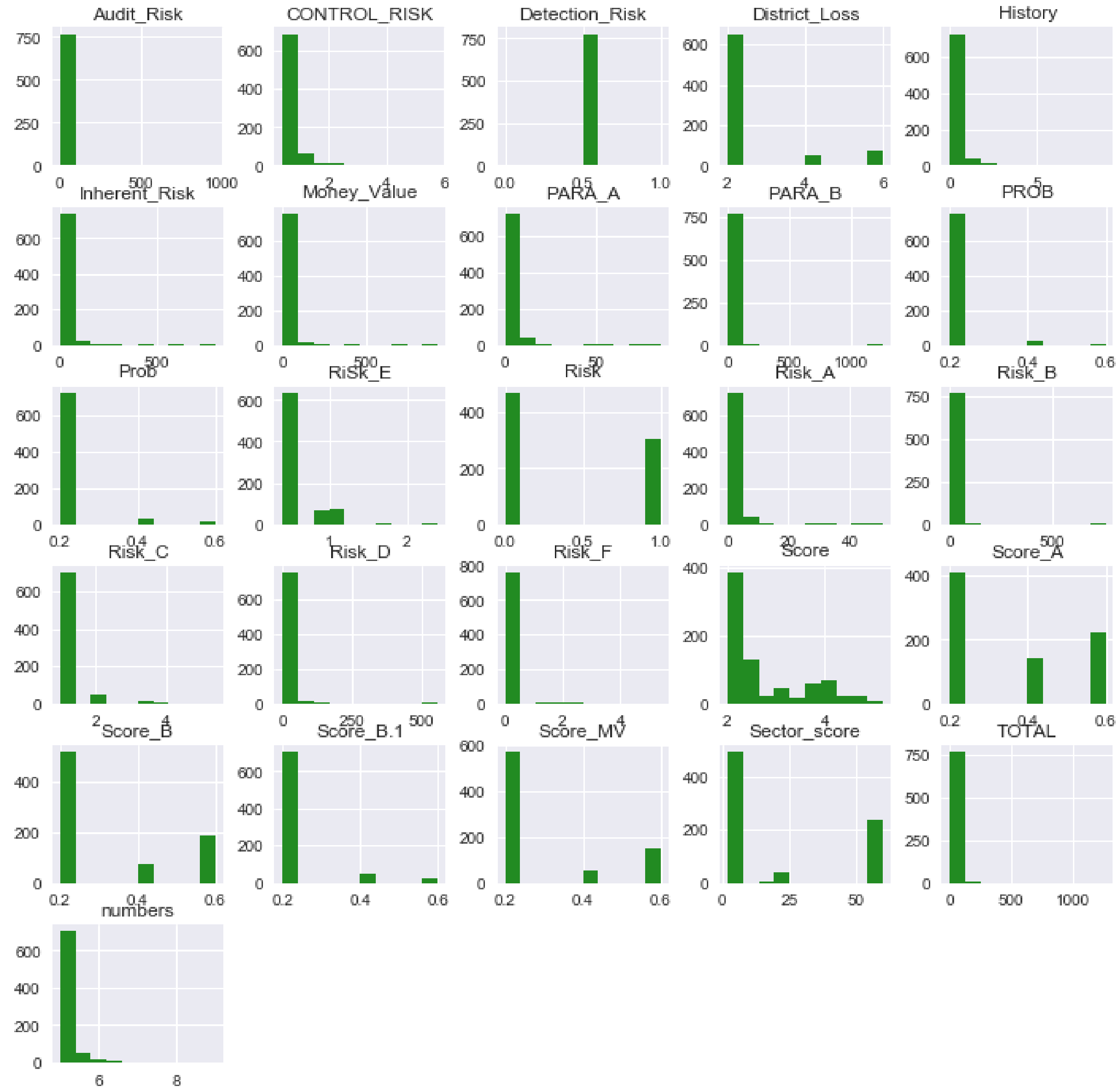
Missing/Outliers & Distribution Analysis: There was one missing value and the row containing the missing value was dropped. The histogram suggests that there not many significant outliers which needs immediate attention. Variables like 'LOCATION_ID','TOTAL', 'Detection_Risk' were dropped as Location_ID had many unique values which are non-essential for our model analysis. Also, Total variable is the summation of PARA_A and PARA_B. Also, Detection_Risk had only one unique value for all the 775 cases. It is prudent to mention that 40% of the total 775 cases were found to be having Risk = 1. Hence, it is not considered as an imbalanced dataset and none of the under sampling/oversampling techniques were applied. It is observed with the help of a stacked bar graph that Money_Value & Sector_score variables have high and low values at Risk levels 0 & 1. It is to be noted that the target variable is affected by Inherent_Risk as it is directly proportional to Risk. The mean value of Inhenrent_Risk is almost 1 when Risk=0 whereas it is 41 when Risk=1.

Feature Engineering: A correlation matrix between continuous variables using Pearson method suggest that many variables like PARA_A & Risk_A, PARA_B & Risk_B etc. are highly correlated. Hence, variables with correlation index greater than 0.9 have been removed to address multi-collinearity issues. Creation of dummies for the discrete variables like label encoding & one hot encoding of different levels was not required as most of the variables are continuous and in clean state.

Training/Evaluation: The dataset is split into 70% of training & 30% of test & was standardized in view of retaining variance in the data. Several models were used to fit the data like Logistics Regression, Decision Tree and Random Forest as the target variable is categorical & has <100K variables. It has been observed that Decision Tree & Random Forest gave the highest accuracy (100%) with highest scores of F1 Score, Recall, Precision, ROC confirming that the data is correctly fit. An ensemble technique was used called Bagging with Decision Tree and was found to be fitting the data well with Accuracy 100%. Also, confusion matrix is made for all the models and observed that all of the Positive & Negative cases are correctly predicted. Hence, to keep it simple, a Decision Tree model is perfectly fitting the data & can be used to predict the fraudulent firm on the basis of the present & historical factors. Audit_Risk is the most potential variable having high feature importance in determining the target variable with the least important variables being Prob, History, PROB, numbers etc.

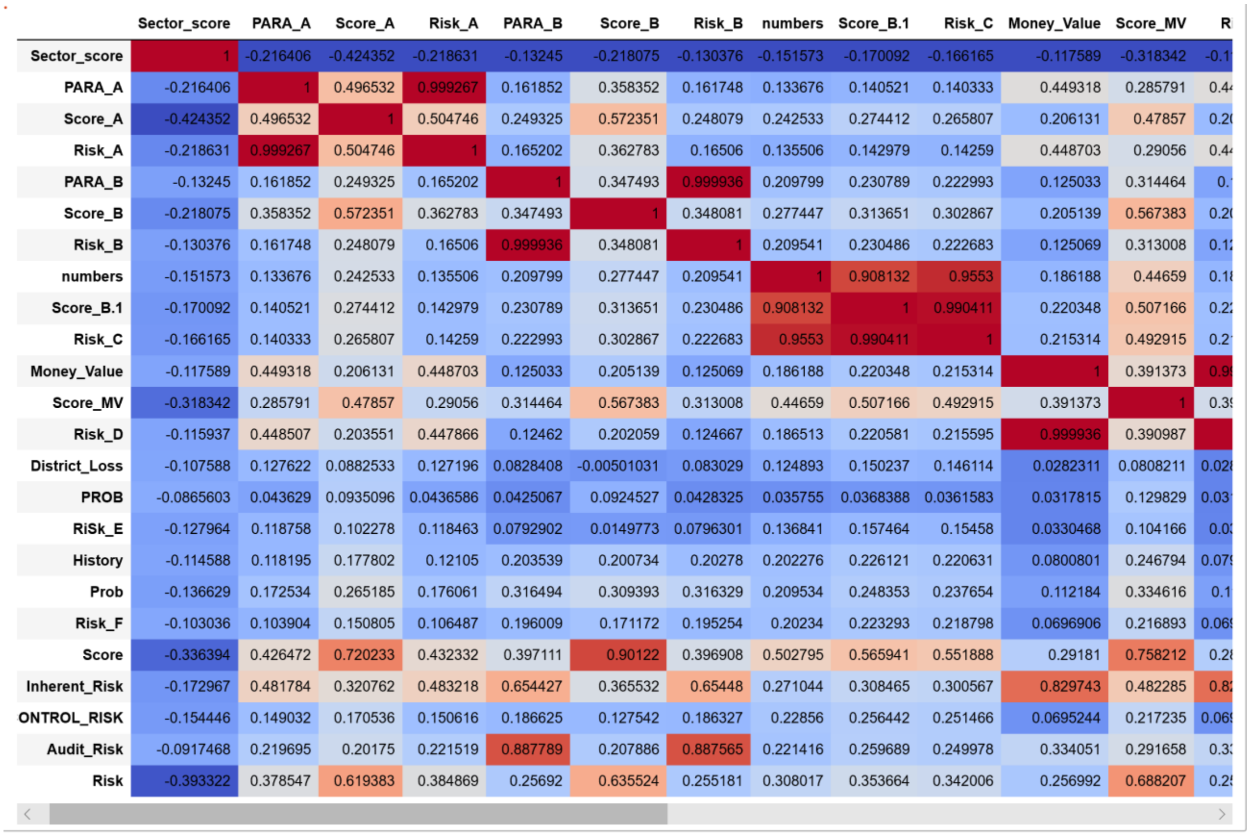
Plots

Histogram



There are many variables which are positively or negatively skewed as per the histogram

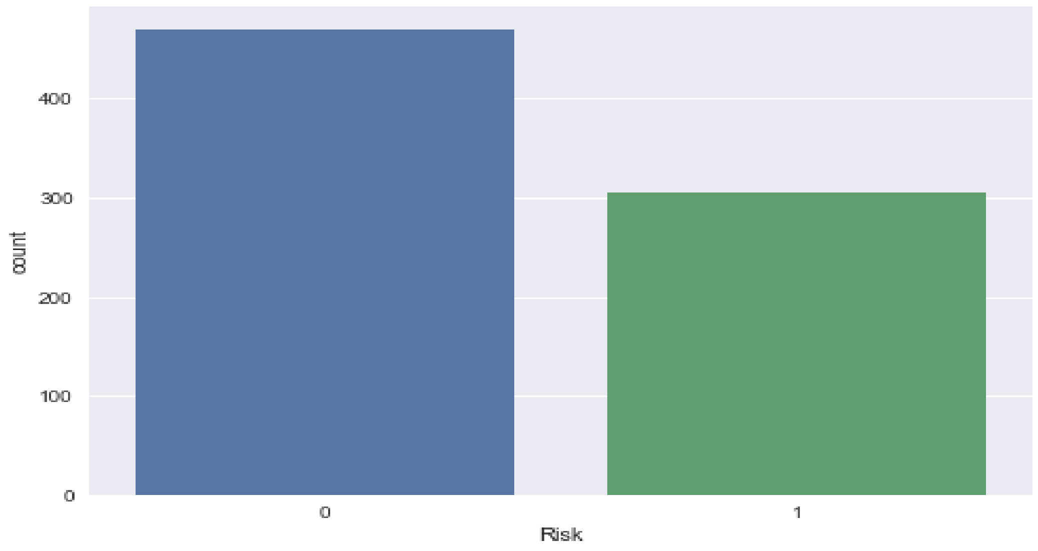
Correlation Matrix



	Sector_score	PARA_A	Score_A	Risk_A	PARA_B	Score_B	Risk_B	numbers	Score_B.1	Risk_C	Money_Value	Score_MV	R
Sector_score	1	-0.216406	-0.424352	-0.218631	-0.13245	-0.218075	-0.130376	-0.151573	-0.170092	-0.166165	-0.117589	-0.318342	-0.1
PARA_A	-0.216406	1	0.496532	0.999267	0.161852	0.358352	0.161748	0.133676	0.140521	0.140333	0.449318	0.285791	0.44
Score_A	-0.424352	0.496532	1	0.504746	0.249325	0.572351	0.248079	0.242533	0.274412	0.265807	0.206131	0.47857	0.2
Risk_A	-0.218631	0.999267	0.504746	1	0.165202	0.362783	0.16506	0.135506	0.142979	0.14259	0.448703	0.29056	0.44
PARA_B	-0.13245	0.161852	0.249325	0.165202	1	0.347493	0.999936	0.209799	0.230789	0.222993	0.125033	0.314464	0.
Score_B	-0.218075	0.358352	0.572351	0.362783	0.347493	1	0.348081	0.277447	0.313651	0.302867	0.205139	0.567383	0.2
Risk_B	-0.130376	0.161748	0.248079	0.16506	0.999936	0.348081	1	0.209541	0.230486	0.222683	0.125069	0.313008	0.1
numbers	-0.151573	0.133676	0.242533	0.135506	0.209799	0.277447	0.209541	1	0.908132	0.9553	0.186188	0.44659	0.1
Score_B.1	-0.170092	0.140521	0.274412	0.142979	0.230789	0.313651	0.230486	0.908132	1	0.990411	0.220348	0.507166	0.2
Risk_C	-0.166165	0.140333	0.265807	0.14259	0.222993	0.302867	0.222683	0.9553	0.990411	1	0.215314	0.492915	0.2
Money_Value	-0.117589	0.449318	0.206131	0.448703	0.125033	0.205139	0.125069	0.186188	0.220348	0.215314	1	0.391373	0.9
Score_MV	-0.318342	0.285791	0.47857	0.29056	0.314464	0.567383	0.313008	0.44659	0.507166	0.492915	0.391373	1	0.3
Risk_D	-0.115937	0.448507	0.203551	0.447866	0.12462	0.202059	0.124667	0.186513	0.220581	0.215595	0.999936	0.390987	
District_Loss	-0.107588	0.127622	0.0882533	0.127196	0.0828408	-0.00501031	0.083029	0.124893	0.150237	0.146114	0.0282311	0.0808211	0.02
PROB	-0.0865603	0.043629	0.0935096	0.0436586	0.0425067	0.0924527	0.0428325	0.035755	0.0368388	0.0361583	0.0317815	0.129829	0.03
RiSk_E	-0.127964	0.118758	0.102278	0.118463	0.0792902	0.0149773	0.0796301	0.136841	0.157464	0.15458	0.0330468	0.104166	0.0
History	-0.114588	0.118195	0.177802	0.12105	0.203539	0.200734	0.20278	0.202276	0.226121	0.220631	0.0800801	0.246794	0.07
Prob	-0.136629	0.172534	0.265185	0.176061	0.316494	0.309393	0.316329	0.209534	0.248353	0.237654	0.112184	0.334616	0.1
Risk_F	-0.103036	0.103904	0.150805	0.106487	0.196009	0.171172	0.195254	0.20234	0.223293	0.218798	0.0696906	0.216893	0.06
Score	-0.336394	0.426472	0.720233	0.432332	0.397111	0.90122	0.396908	0.502795	0.565941	0.551888	0.29181	0.758212	0.2
Inherent_Risk	-0.172967	0.481784	0.320762	0.483218	0.654427	0.365532	0.65448	0.271044	0.308465	0.300567	0.829743	0.482285	0.8
ONTROL_RISK	-0.154446	0.149032	0.170536	0.150616	0.186625	0.127542	0.186327	0.22856	0.256442	0.251466	0.0695244	0.217235	0.06
Audit_Risk	-0.0917468	0.219695	0.20175	0.221519	0.887789	0.207886	0.887565	0.221416	0.259689	0.249978	0.334051	0.291658	0.3
Risk	-0.393322	0.378547	0.619383	0.384869	0.25692	0.635524	0.255181	0.308017	0.353664	0.342006	0.256992	0.688207	0.2

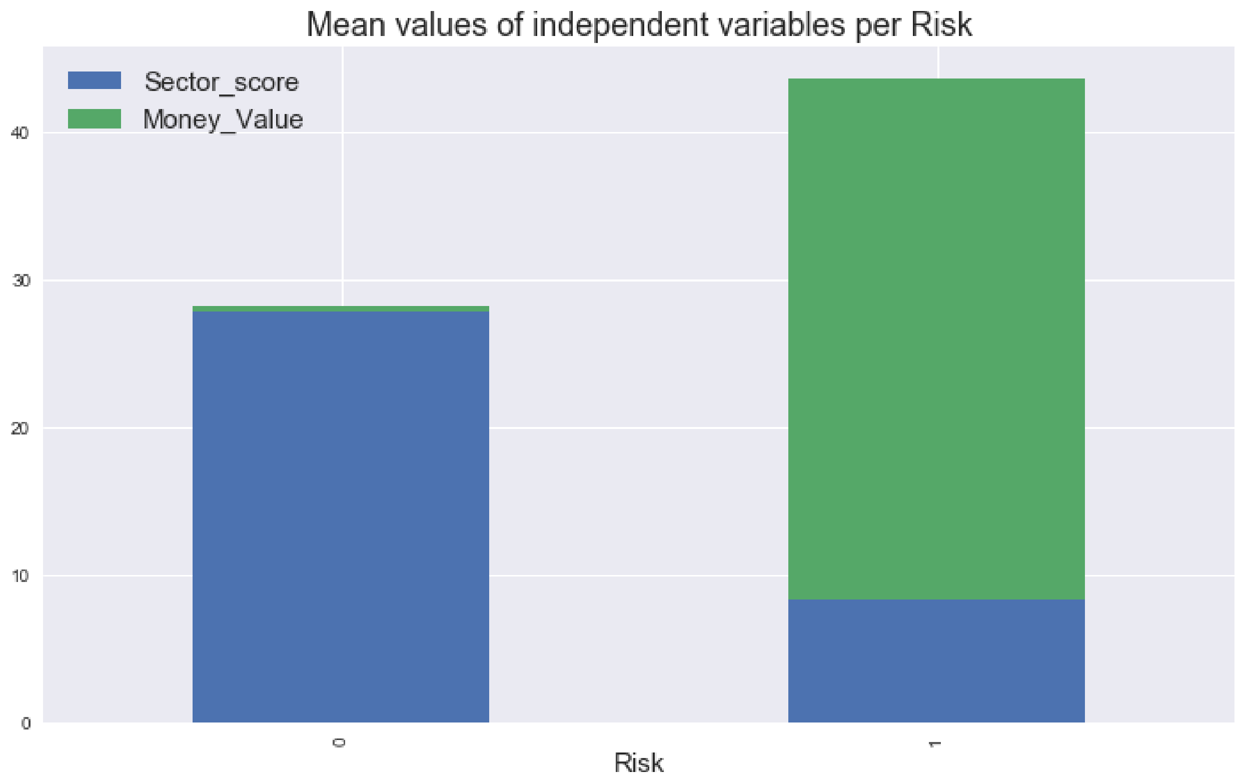
It is observed that "PARA_A" & "Risk_A" are highly correlated. Similarly, PARA_B & Risk_B, numbers, Score_B.1 & Risk_C etc. are highly correlated Hence, it is proposed to consider any one of the variable to check its influence on the dependent variable.

Target Variable – Count Plot



It is observed that Risk is 1 or positive for over 40% out of 775 cases

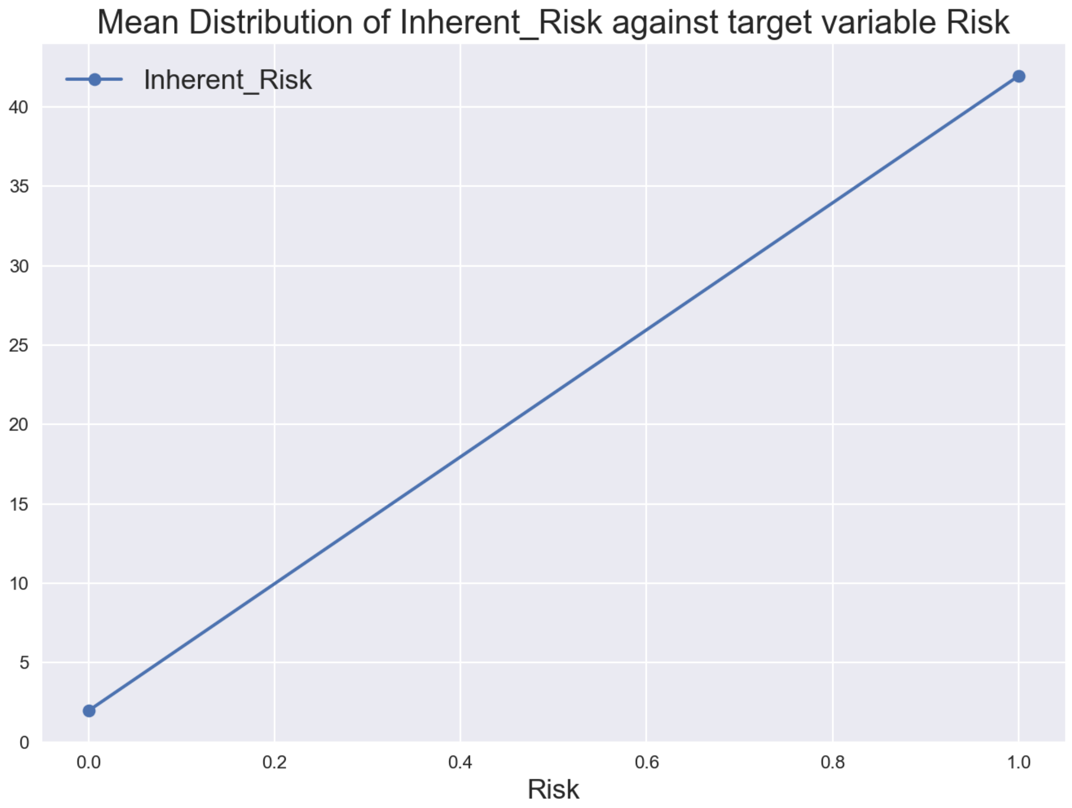
Independent Variables Distribution



It is observed that Money_Value & Sector_score variables are quite different and has unrelatable values at Risk 0 & 1 i.e. Sector_Score is high when Risk=0 but Money_Value is low and vice versa.

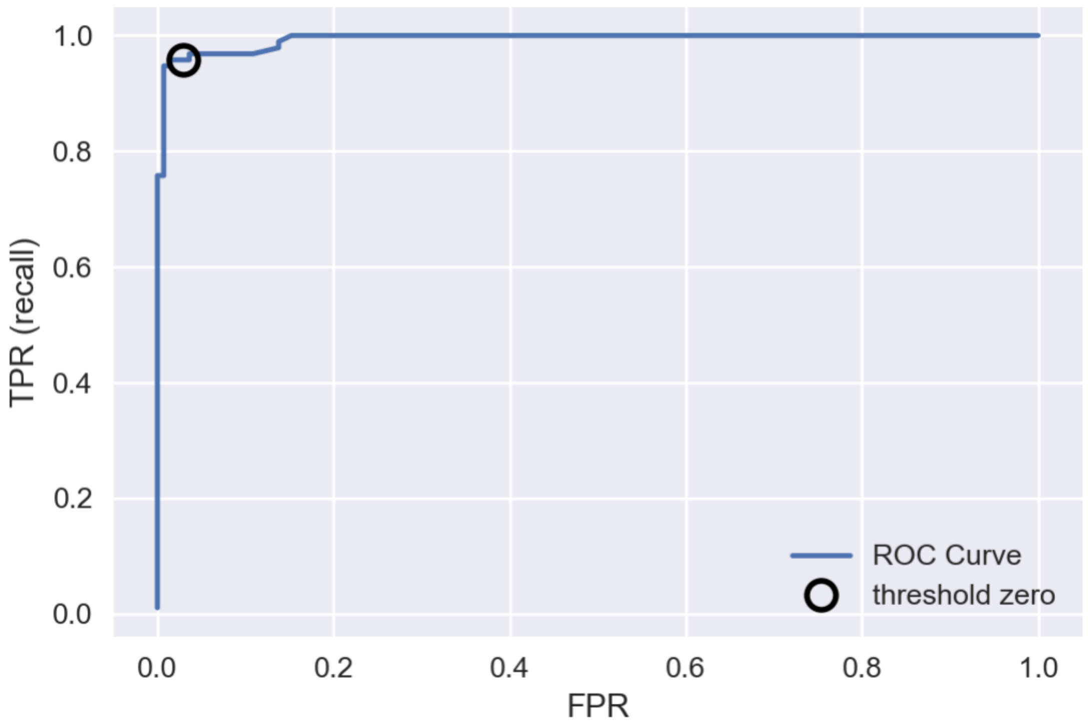
Inherent_Risk vs Risk

The above mentioned graph tells that the target variable is affected by Inherent_Risk as it is directly proportional to Risk. The mean value of Inhenrent_Risk is almost 1 when Risk=0 whereas it is 41 when Risk=1

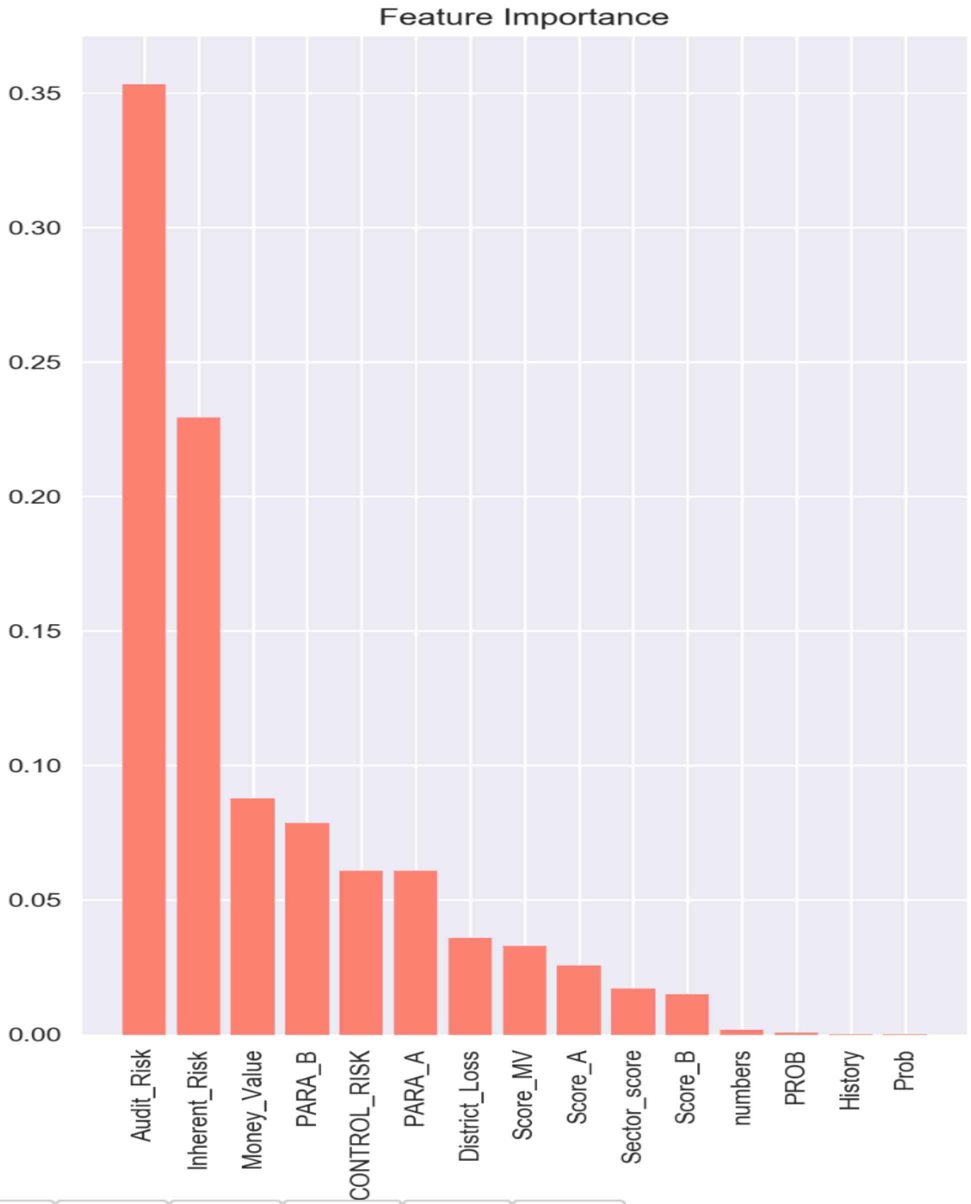


The above-mentioned graph tells that the target variable is affected by Inherent_Risk as it is directly proportional to Risk. The mean value of Inherent_Risk is almost 1 when Risk=0 whereas it is 41 when Risk=1

ROC Curve of Logistic Regression



Feature Importance Graph



This tells that Audit_Risk is the most potential variable which has high importance in determining the target variable with the least important variables being Prob, History, PROB, numbers etc.