DS 504 - Capstone Project

Capstone Project Report

MBA-INTELLIGENT DATA SCIENCE (BATCH 2018-20)

UNDER THE GUIDANCE OF
Mr. Shounak Pal

Submitted by: Group-10

Nikhil Goyal	MB18GID258
Ritika Yadav	MB18GID292
Bansi Lal Pathak	MB18GID267
Anuradha Mohapatra	MB18GID278
Namrata Yadav	MB18GID284

Table of Figures:

Table of Contents

ABSTRACT

Credit and lending business still runs under a lot of uncertainty. With no proper automation method available for assessment and analysis of credibility of individuals, creditors/banks always carry the risk of their clients not meeting their debt obligations – often called as risk of non-payment. Another concern for them is the exact amount which should be given based on the estimated risk.

As a scope of our work, we have tried to cover different factors which should be considered into the process of flagging whether a customer is a good risk or bad risk and predict the loan amount which can be granted to them. Our study involved working with different machine learning models and feature engineering to get a model which can predict the 'risk' and 'amount'. Our findings showed that existing checking account balance had the primary role in customer flagging and deciding the limit of loan followed by duration of the loan and a customer's existing. Also, on the contrary to popular belief, gender had no role to play in deciding the limit or risk. Our work further motivates other researcher to find various other factors on separate data, affecting the decision making which might vary based on geographies and demographics. It also helps managers and institutions take important decisions of knowing their customer better, hence reducing risk and increase profit.

1 INTRODUCTION

It's been a while that banking sector has tried to explore the full potential of technology but with changing technology, there is always something new to implement. The role of technology started by simply storing customer data to centralizing the banking system and has now reached to implementation of machine learning to predict customer behavior, acquire new user, up-sell new products and use them to understand and retain the existing users.

Loans have been one of the most profitable businesses for banking institutions, but it always involved risk of non-payment or the loan going bad. Every loan given by the bank is a risk which can be broadly described into two major categories: Good Risk and Bad Risk. Good Risks are the ones which has better chances of repayment compared to others whereas on the other hand bad risks have more chances of defaulting. In both the cases, the results are not certain.

Risk classification helps bank to evaluate if a loan applicant can be a defaulter at a later stage if loan is granted, and what should be the amount in either case. As they cannot deny the loan grant, knowing the risk minimizes the risk and increase profit.

A lot of research has already gone using different machine learning models of regression and classification to find out the probability of default (Stephen Zamore, Kwame Ohene Djan, Ilan Alon, Bersant Hobdari, 2018). Whereas many people have only considered only on one aspect, either finding the limit or the risk, we have tried to cover both in our research.

One of the important aspects of the research is to find the factors which are actually responsible for decision and by how much. (Changjun Zheng, Niluthpaul Sarker, Shamsun Nahar, 2018) **Factors affecting bank credit risk: An empirical insight** focused on characteristics of bank and showcased how profitability, capital and bank size are inversely associated with bank credit risk whereas net interest margin and inefficiency have positive effect. (Garr, 2013) in his paper **Determinants of Credit Risk in the Banking Industry of Ghana** examined bank-specific, industry-specific and macroeconomic factors that influence credit risk in developing

In this research, we have compared different machine learning classification models and picked up the one with maximum accuracy preceded by feature selection and data cleaning aspects.

Some of the most previously done most relatable works have been mentioned in “Literature Work” section considering the techniques used and results/findings. The “Methodology” section covers different steps and algorithms used to get the maximum accuracy through different models. The “Results” and “Discussions” sections explore the output and what sense they make in business scenarios.

It is worth mentioning that the metrics selected and models implemented are done to get the maximum accuracy and efficiency based on data available.

2 LITERATURE REVIEW

For the research work we have gone through many research papers and shortlisted the ones which as best suited for scope of our project. We have reviewed those papers and highlighted some key features like: features used, models implemented and research questions need to be proved. The papers key features are as follows:

Author's name, year of publication and articles name	Dependent Variable	Independent Variable/s	Technique/ Model Used	Results/Findings
(Li Gan, 2008) An empirical study of the credit market with unobserved consumer types	default status	age of the debt, demographic characteristics of each individual, some economic variables and the credit worthiness.	Authors have implemented logistic regression to model probability of default.	Two alternatives compared with the results of the logit model: i. A "naïve" approach which uses the unconditional probability of default. ii. A "risk averse" approach based on the distribution of the estimated probability *for* each individual to minimize the no. of "bad" clients classified by the model as "good". The performance of the "risk averse" approach is lower than both the logit model and the "naïve" approach.
(Khashman, 2010) Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes	Credit Risk	Mix of 20 categorical and numerical attributes had been used (Status of existing checking account, Credit history, Purpose, Credit account, Personal status and sex etc.)	Neural Network	The credit risk evaluation neural network model performs best when using the LS4 learning scheme. Accuracy rates of 99.25% and 73.17% were obtained using the training and validation data sets

(Zala Herga, 2016) Modeling probability of default and credit limits	Probability of default & Credit Limit	Sum of trades, outstanding debts, disputed claims and delayed payments.	Author have implemented 1. Logistic Regression to model probability of default. 2. a linear programming based approach for credit limit.	It is observed by the author that this method provides an optimal portfolio as monthly credit limit for each company takes value between the provided credit limit bounds. The default values used for these graphs are alpha = 0.95, Lower_b = 0, upper_b = max trading volume & margin = 0:01. Most companies get either zero or maximum credit approved
(N., Kristovska I., & M., 2016) Credit risk management in commercial banks	Problems with loan repayment	Loan Terms, Loan Amount, Client's age, Number of children, sex, Average monthly income	Clustering Methods	In overseeing credit hazard, one needs to make an arrangement of interconnected and associated techniques for purposeful activity went for limiting danger and vulnerability in crediting-related exercises. Utilizing the proposed model of credit chance evaluation makes it conceivable to adopt a separated strategy to credit chance administration.
(HON & BELLOTTI) Models and forecasts of credit card balance	Credit card balance	Application variables: age of applicant, employment status, tenure, months at current address and application channel and Behavioral variables: statement number, total outstanding balance, credit limit, cash balance, account status (i.e. open, closed, charged-off or fraud), delinquency, payment amount and number of payments.	Statistical Models used: 1. Ordinary Least Squares 2. Two-stage regression model 3. Mixture regression 4. Random effects panel model	When comparing all the models random effect panel model is the best model based on using the MAE measure of performance. There were few factors which came out to be important variables such as age where as Employment status is not that significant factor for the credit limit

(Flood, July 2017) Early identification of high-risk credit card customers based on behavioural data	Credit Risk	month-ending balance, credit limit, payment amount, account activity, delinquency, borrower income etc.	1. Linear Regression 2. Logistic regression 3. Decision trees 4. Random Forests 5. Ada Boost 6. Gradient Boosting SMOTE was used to deal with imbalanced data	The results were compared using different combinations of dependent variables, forecast horizon lengths and training window lengths.
(Liu, 2018) Machine Learning Approaches to Predict Default of Credit Card Clients	Credit Risk	The amount of credit, gender, education, marital status, age, history of delayed payment, amount of bill statement and amount paid	This paper compares below traditional machine learning models: 1.Support Vector Machine 2.k-Nearest Neighbors 3.Decision Tree and Random Forest 4.Feedforward Neural Network	neural networks achieve higher accuracies than traditional models. This paper also tries to figure out whether dropout can improve accuracy of neural networks.
(Motwani, Chaurasiya, & Bajaj, 2018) Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud	Credit Risk	Details of payments, demographic factors, credit data, history of payment, and bill statements of credit card clients	Author have used Azure Technology for applying Model on the Dataset Models Applied: 1. Bayes Point Machine 2. Logistic Regression 3. Decision Tree 4. Neural Network (NN) (Proposed)	All the Models are compared on the basis of their Accuracy, Prediction Rate, Recall calculated. Neural Network results were on top in case of Accuracy (82.20), Prediction Rate (0.110), Recall (0.411)

Table 1: Literature Review

3 DATA

Data Source: UCI Machine Learning Repository
No. of Features: 21
No. of Records: 1000
No. of Classes (Categorical Variable): 56

4 MODEL AND HYPOTHESIS

On the basis of literature review below mentioned hypothesis are tested as part of this research:

H_1: Customer Credit Risk effects the credit Limit offered by Banks.

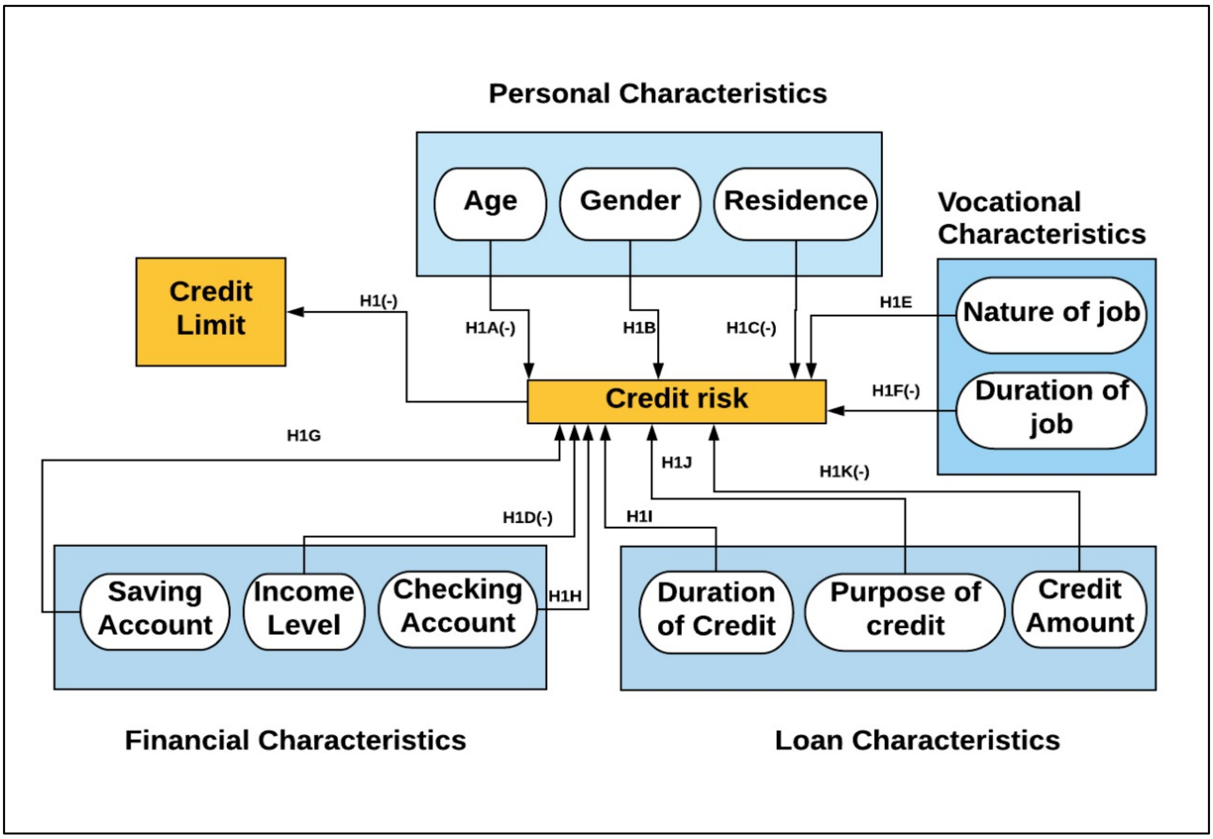


Figure 1: Model and Hypothesis

Features/characteristics which directly affects credit risk and hence credit limit can be broadly classifying in 4 characteristics as below:

4.1 Personal Characteristics of Borrowers

Personal Characteristics of a borrower as age, sex and duration of residence or residence type can highly effect of being a customer a Bad or Good Credit risk:

- **Age:** The index of bad Borrower is high for the customer with less age.
- **Sex:** Credit risk seems to be affected also by the sex of borrowers, but its relation to marital status is questionable.
- **Residence:** Residence type and duration of residence to present address can affect the credit worthiness of a customer.

On the basis of personal characteristics of borrowers, the below mention hypothesis can be tested over our dataset:

H_{1A}: Age of the Customer effect the Credit Risk to Bank.

H_{1B}: Gender of the Customer effect the Credit Risk to Bank.

H_{1C}: Residence Type of the Customer effect the Credit Risk to Bank.

H_{1D}: Duration of stay at present address of the Customer effect the Credit Risk to Bank.

4.2 Vocational Characteristics of Borrowers

As Vocational characteristics dependent to personal attributes up to a certain degree so it is essential to analyse the same for better understanding of credit risk. Vocational Characteristics includes the features like: nature of the borrower's work or occupation and tenure of employment.

- **Nature of Borrower's Job:** The nature of Borrower job can be an important feature for classifying the credit risk. The risk can be different for a person doing skilled, Semiskilled or unskilled job.
- **Tenure of Employment:** Tenure of employment effect the credit worthiness for a customer. Tenure of employment is directly proportional to credit worthiness of a customer.

As per the Vocational characteristic of borrowers in our dataset we can tested below Hypothesis:

H_{1E}: Nature of Job of the Customer effects the Credit Risk to Bank.

H_{1F}: Duration of Job of the Customer effects the Credit Risk to Bank.

4.3 Financial Characteristics of Borrowers

Financial Characteristics of a customer is the most important feature to determine the Credit risk. Financial factors tell the credit worthiness and the obligation of customer towards his/her debtors and creditors.

- **Annual Income:** People with higher income per year tends to default less than people with low income
- **Assets (saving account):** Assets like saving account tells the details of assets a person had and can give the significant information about a person to default.
- **Liabilities (Checking account):** No. of transaction, cheque deposit in checking account can be a relevant feature to classifying good or bad risk.

Features given at our dataset related to financial characteristic can be used to test below Hypothesis:

H_{1G}: Saving Account of the Customer effects the Credit Risk to Bank.
H_{1H}: Checking Account of the Customer effects the Credit Risk to Bank.

4.4 Characteristics of the Loan

In an examination of factors contributing to credit risk, consideration must be given not only to characteristics of the borrower but also to certain features of the loan transaction itself.

- **Credit Amount:** Higher amount of credit asked by customer tends to default less.
- **Purpose for Credit:** Purpose of credit can be an important feature. It tells how a person will use this credit like Education, Shopping or assets.
- **Duration of the repayment period:** The credit asked for less than 12 months default more but this can be dependent to other factors as well.

How different loan characteristics given at dataset can affect the Bad and Good risk assessed by testing below mentioned Hypothesis:

H_{1I}: Duration of the credit effects the Credit Risk to Bank.
H_{1J}: Purpose for which the credit is given effects the Credit Risk to Bank.
H_{1G}: Credit Amount of the Customer effects the Credit Risk to Bank.

5 METHODOLOGY

5.1 Proposed technique

Our research tries to overcome the shortcomings of previous research works in many ways by using:

- Latest machine learning and prediction models and comparing the results
- Automation of feature selection process
- Automation of model selection

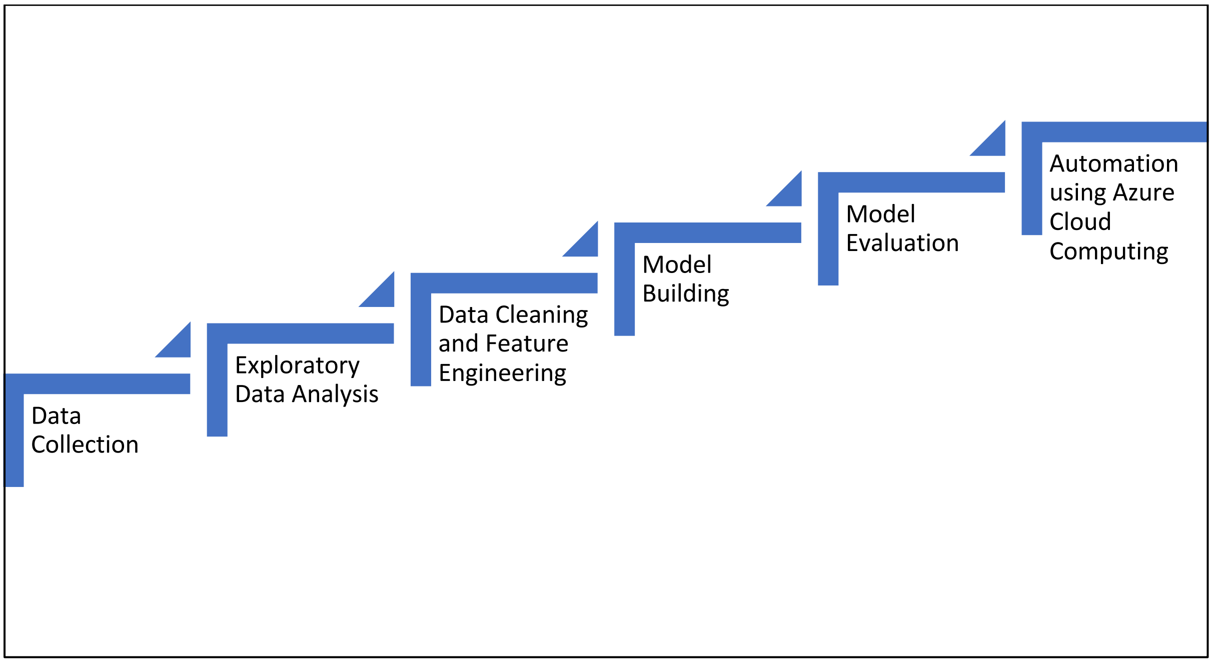

Data Collection
Exploratory Data Analysis
Data Cleaning and Feature Engineering
Model Building
Model Evaluation
Automation using Azure Cloud Computing

Figure 2: Proposed Flow for Modelling

5.2 Exploratory Data Analysis

To understand the data better, the data is visualized in different ways to know the relationships, correlation and inter-dependencies. Some of the key findings are:

- Countplot of our target variable(default) and found that data is imbalance as in the ration of 7:3. Synthetic Minority Over-Sampling Technique (SMOTE) (Nitesh V. Chawla, 2002) has been used to balance the data.

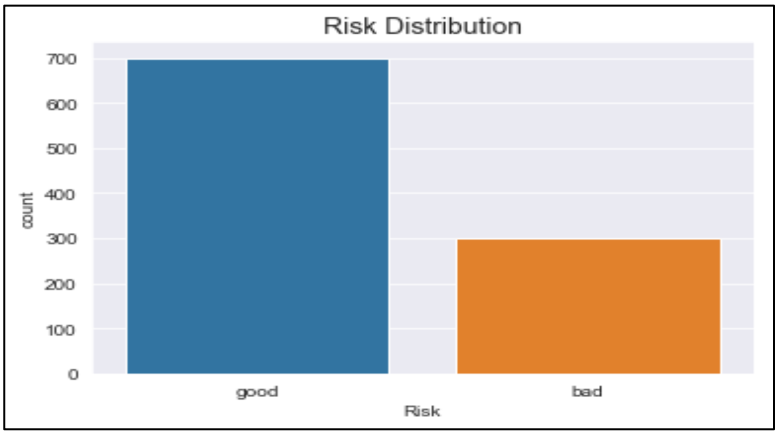


Figure 3: Counterplot of target variable(default)

- Both both the Risks (Good and Bad) are positively skewed

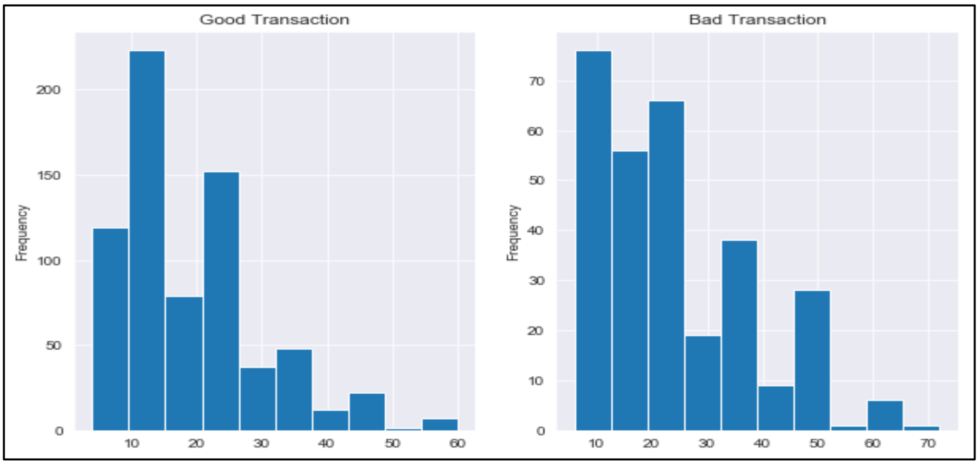
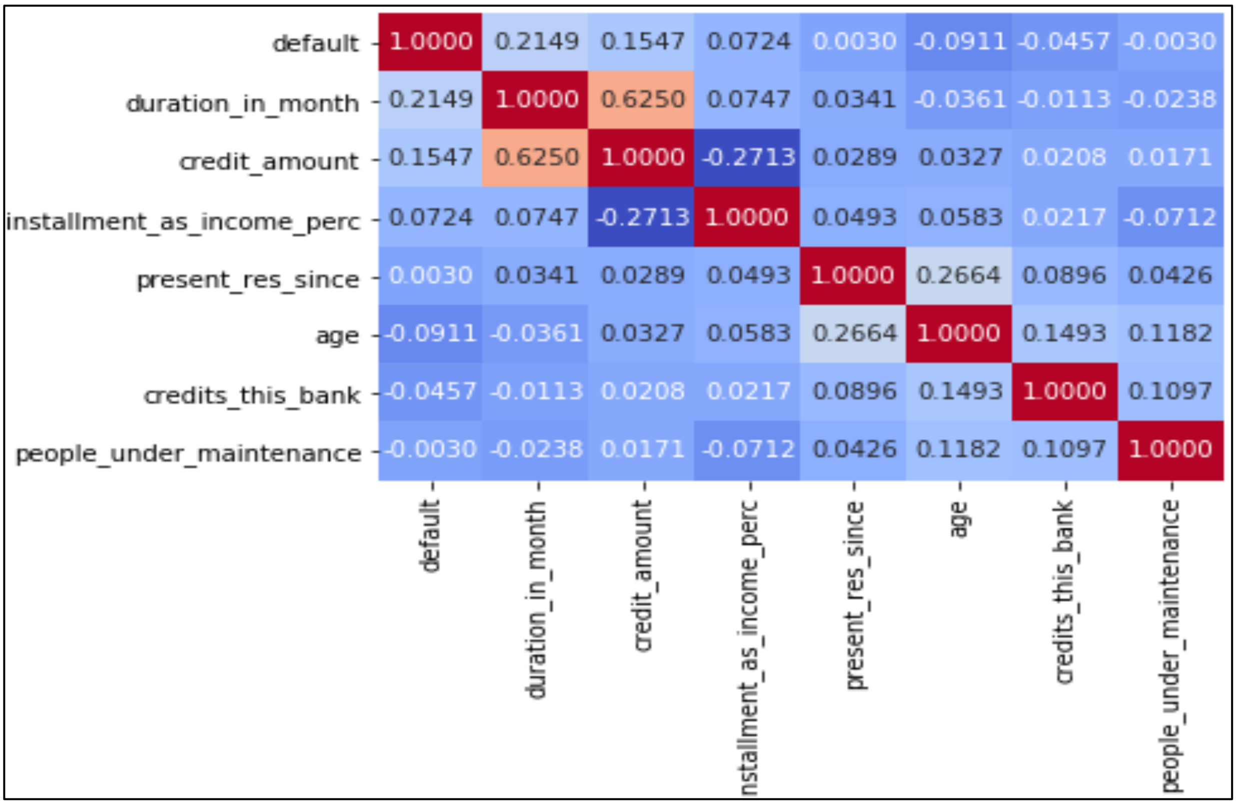


Figure 4: Bar plot of good risk and bad risk

- The correlation matrix between variables and the numerical variables in data set shows found duration_in_month and credit_amount have the highest correlation with target variable. While variables like age, credits_this_bank and people_under_maintenace, are negatively correlated with target variable. Correlation between the independent variable is not high so can be used directly for model building.

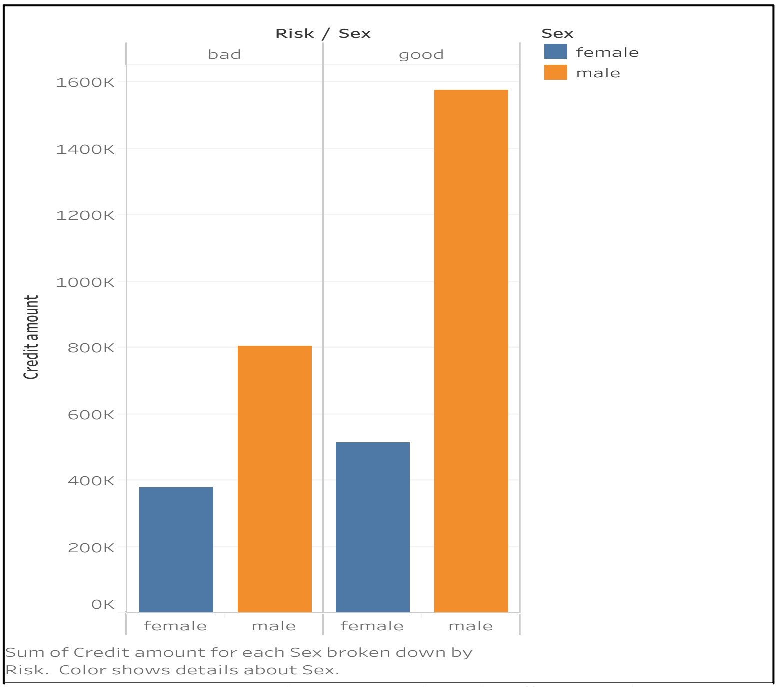


	default	duration_in_month	credit_amount	installment_as_income_perc	present_res_since	age	credits_this_bank	people_under_maintenance
default	1.0000	0.2149	0.1547	0.0724	0.0030	-0.0911	-0.0457	-0.0030
duration_in_month	0.2149	1.0000	0.6250	0.0747	0.0341	-0.0361	-0.0113	-0.0238
credit_amount	0.1547	0.6250	1.0000	-0.2713	0.0289	0.0327	0.0208	0.0171
installment_as_income_perc	0.0724	0.0747	-0.2713	1.0000	0.0493	0.0583	0.0217	-0.0712
present_res_since	0.0030	0.0341	0.0289	0.0493	1.0000	0.2664	0.0896	0.0426
age	-0.0911	-0.0361	0.0327	0.0583	0.2664	1.0000	0.1493	0.1182
credits_this_bank	-0.0457	-0.0113	0.0208	0.0217	0.0896	0.1493	1.0000	0.1097
people_under_maintenance	-0.0030	-0.0238	0.0171	-0.0712	0.0426	0.1182	0.1097	1.0000

Figure 5: Correlation matrix between variables and the numerical variables in data set

Now considering multi variables effect on the Risk keeping 0 for Bad Risk and 1 for Good Risk.

- **Credit Amount and Gender effect on Risk:**



Sum of Credit amount for each Sex broken down by Risk. Color shows details about Sex.

Figure 6: Credit Amount and Gender effect on Risk

We can clearly see that the difference in Good Risk is more between Males and Female whereas it is quite less when we compare for Bad Risk.

- **Duration and Age effect of Risk:**

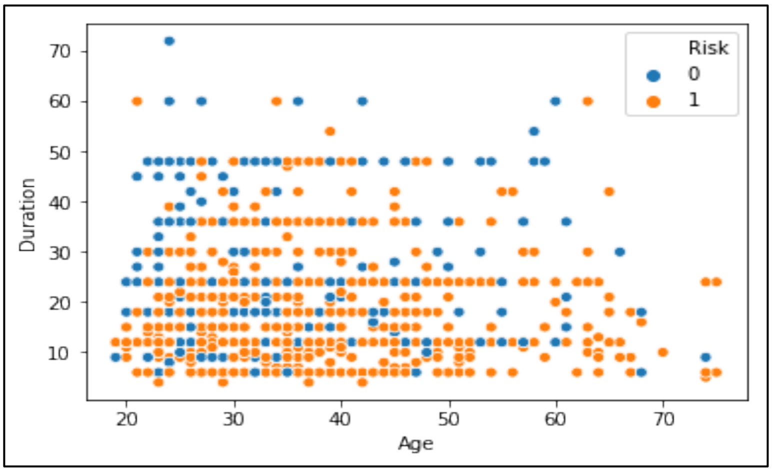


Figure 7: Duration and Age effect of Risk

It is observed from the scatter plot that as the duration is going high the Bad Risk is increasing with the increase in the Age group as well. However age group between 20-30 have the more Bad Risk as seen the figure

- **Credit Amount and Age effect on Risk:**

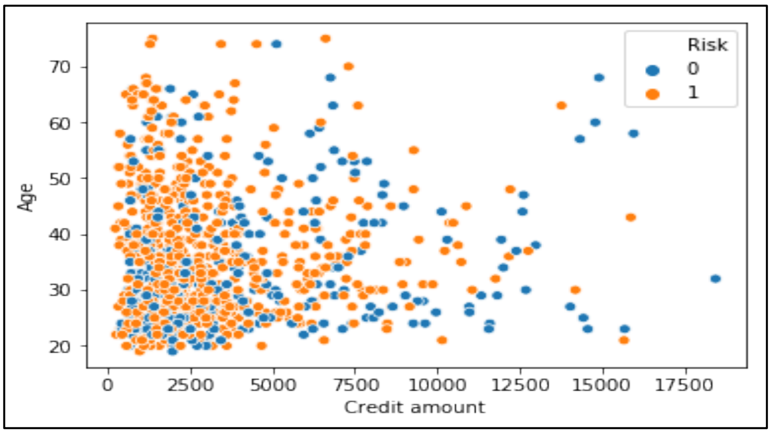


Figure 8: Credit Amount and Age effect on Risk

The above figure of scatter plot gave us the insight that the Credit amount involving lower credit amount are likely to be at lower risks (Good).

- **Credit Amount on basis of Job and Purpose:**

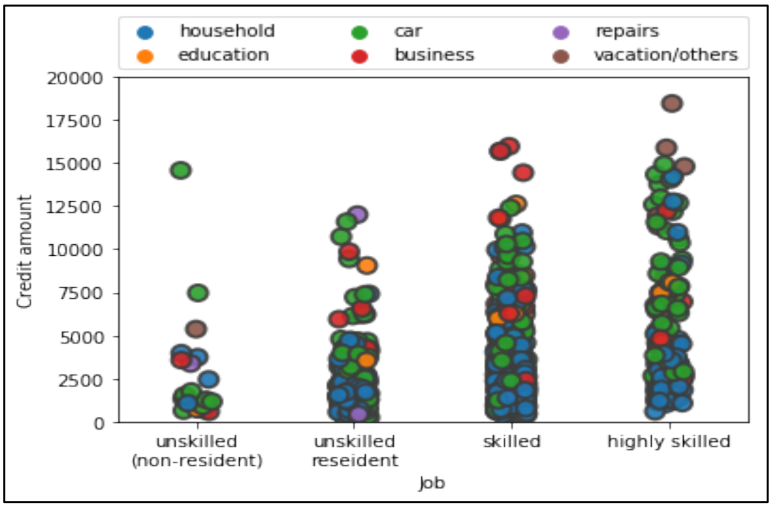
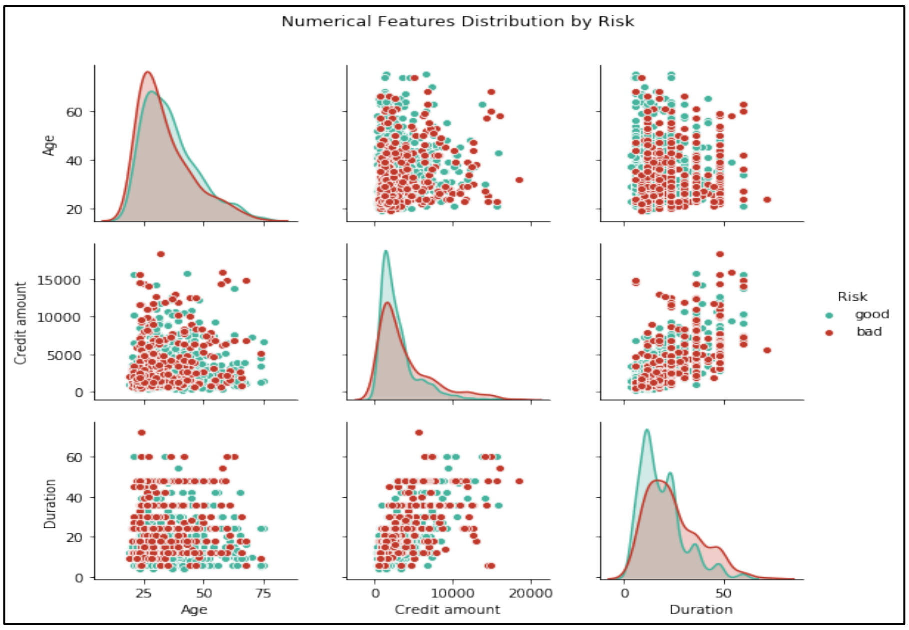


Figure 9: Credit Amount on basis of Job and Purpose

Skills, job and purpose are highly correlated. Highly skilled people tend to take more loans of higher amount than other. Car remains a common 'purpose' among all categories.

- **Numerical Features Distribution by Risk**



We could say from the pair plot that duration of credit is directly proportional to the credit amount and with low duration & high credit amount it represents a bad risk.

Figure 10: Numerical Features Distribution by Risk

- **Credit Risk by Credit Amount and Saving/Checking Accounts**

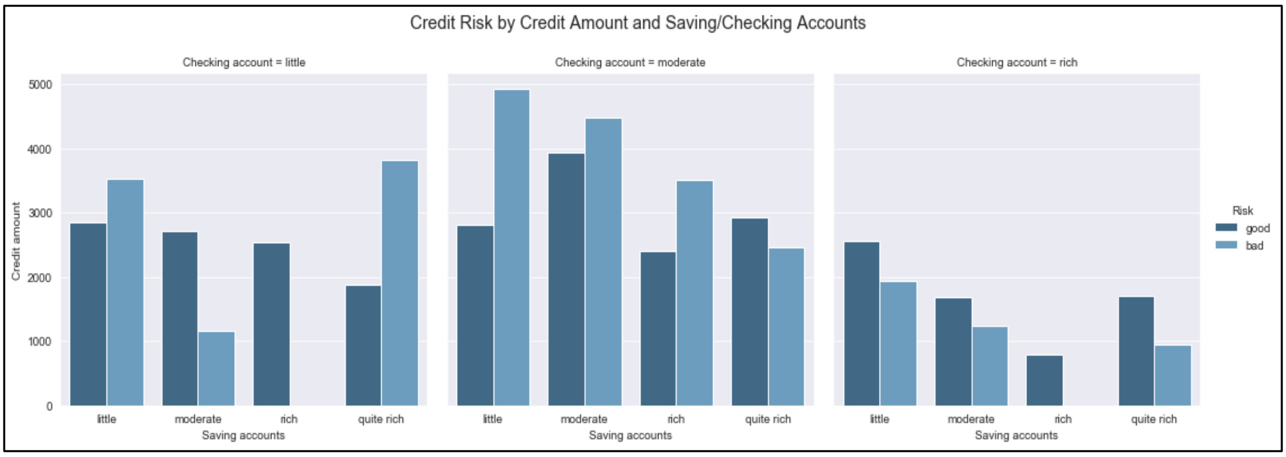
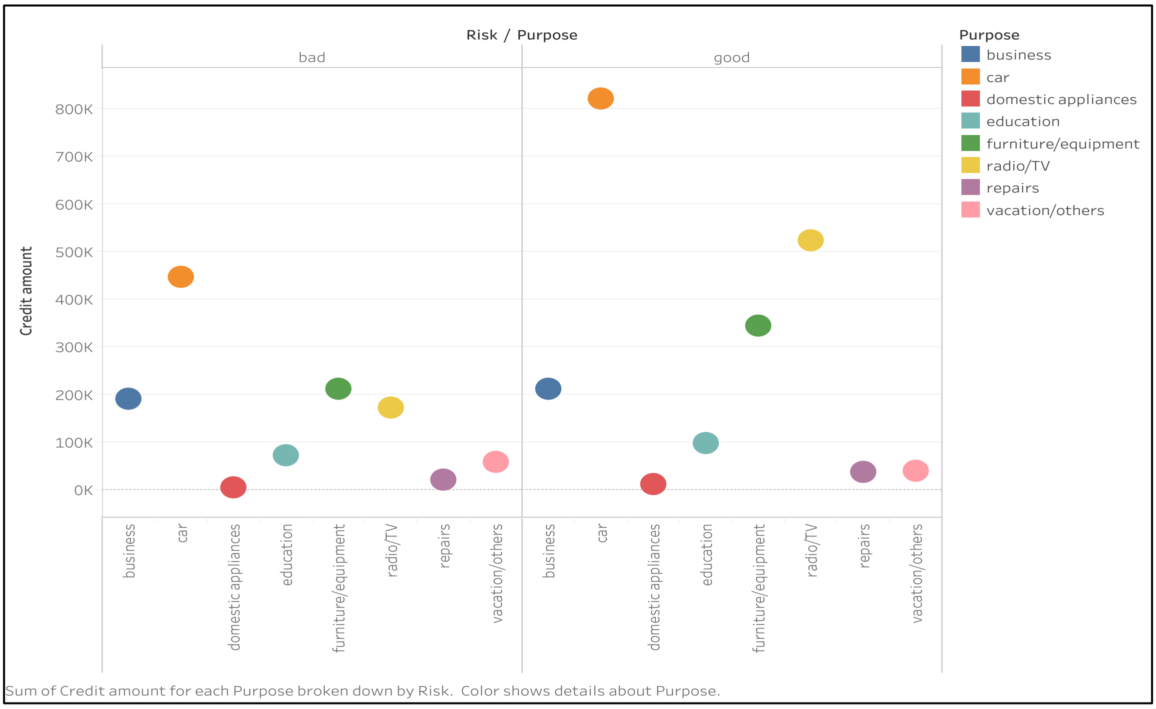


Figure 11: Credit Risk by Credit Amount and Saving/Checking Accounts

We could say that the borrowers with very good checking accounts (rich) have the lowest mean credit requests and there is no bad risk for clients with good saving accounts (rich).

- **Credit Amount and Purpose effect on Risk:**



Sum of Credit amount for each Purpose broken down by Risk. Color shows details about Purpose.

Figure 12: Credit Amount and Purpose effect on Risk

The visualization displayed above helps us to understand the Risk effect on the basis of Purpose and amount. Car comes out to be the most prominent purpose in Good and Bad risk when Credit amount is considered as one of the factor where as domestic appliances is the least affecting factor. Business is equal risk on the same credit amount whereas Radio/TV has bad risk with lower amount.

- **Credit Distribution by purpose**

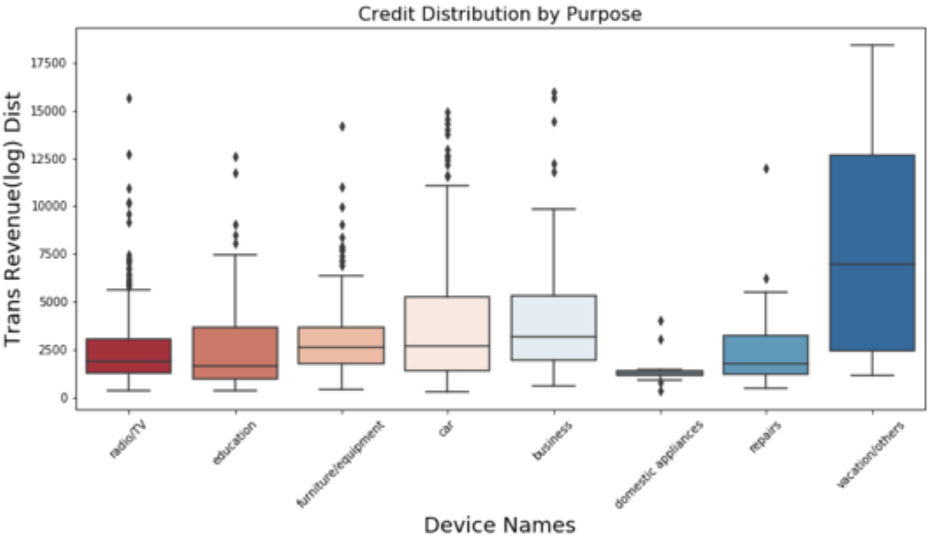


Figure 13: Credit Distribution by purpose

- **Application Reasons by Gender**

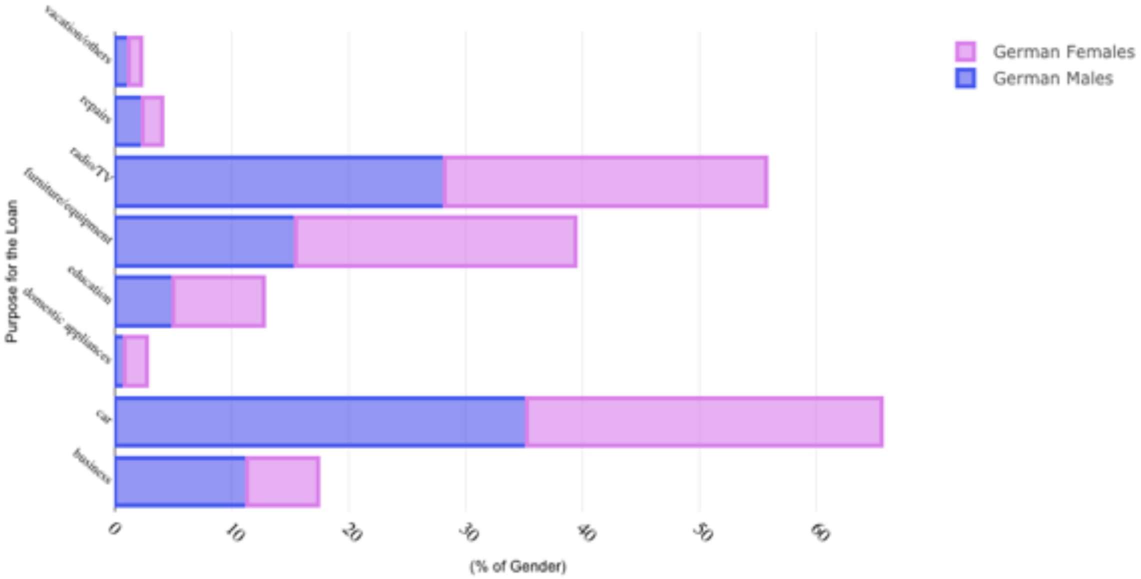


Figure 14: Application Reasons by Gender

- **Types of Loan by Age Group Segment**
  - **Young:** Clients age ranges from (19 - 29).
  - **Young Adults:** Clients age ranges from (30-40)
  - **Senior:** Clients age ranges from (41-55)
  - **Elder:** Clients age is more than 55 years old

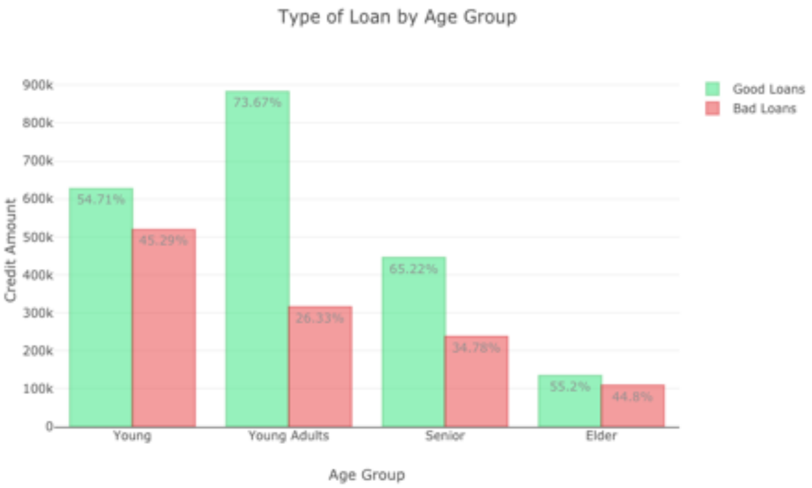


Figure 15: Types of Loan by Age Group Segment

5.3 Data Manipulation

Following data manipulation techniques were applied to implement the models mentioned in next step:

- We have changed the target variable output to 1 and 0 in place of 1 & 2 in original dataset.
- Numerical variables are standardized using StandardScalar.
- The categorical variables are treated by using dummy variables.
- The standard numerical variables are merged and categorical variables are treated.
- The number of columns in dataset increases from 21 to 62 as there is more than 50 categories in dataset (dummy variables included).
- The dataset is split into train and test in to 8:2 ratio.

5.4 Model Building

5.4.1 5.4.1 Base-Line model

To build our base line model all the features were taken into consideration and applied different models after treating the categorical features by using categorical variable and standardization of all the features.

5.4.2 Improvement Over Base-Line model

We have three improvements over our base line model which are sometimes used independently and in some case, one over another.

5.4.2.1 *Dummy Trap Improvement*

As we had replaced our categories with dummy variables for model building, so there must be a possibility of one dummy variable is highly correlated with other dummy variables. Using all dummy variables lead to dummy variable trap. So, we had designed our models excluding one dummy variable for each categorical feature.

5.4.2.2 *Unbalanced data-set Improvement*

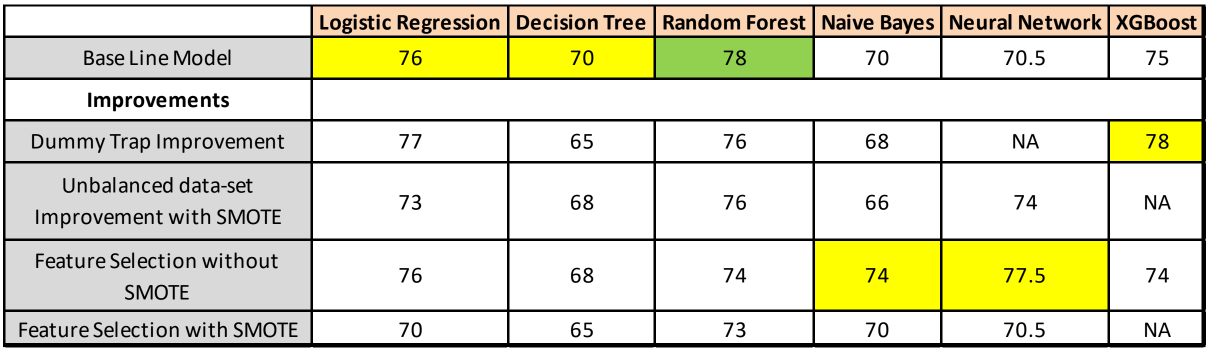
We had used oversampling technique SMOTE to take care of unbalance dataset as our dataset has ratio of 7:3 between good and bad risk. We had applied SMOTE over dataset used in base line model as well as Dummy trap improvement dataset.

5.4.2.3 *Important Feature model*

Random Forest classifier to get Rank of 15 most important features which are as follows in descending order: Credit Amount, Duration, Age, Existing Checking, Instalment rate, existing checking, credit history, savings, existing credits, purpose, other instalment plans, housing, existing checking.

Then we had used these features to build our model and compare the accuracy to find a better model. We had also applied SMOTE for the above important feature dataset and build model separately to compare.

We had built all the previous model after above improvements and compare the results. To find a better model than our base line model with higher accuracy.



	Logistic Regression	Decision Tree	Random Forest	Naive Bayes	Neural Network	XGBoost
Base Line Model	76	70	78	70	70.5	75
Improvements						
Dummy Trap Improvement	77	65	76	68	NA	78
Unbalanced data-set Improvement with SMOTE	73	68	76	66	74	NA
Feature Selection without SMOTE	76	68	74	74	77.5	74
Feature Selection with SMOTE	70	65	73	70	70.5	NA

Figure 16: Comparison of implemented model against base and improved datasets

5.5 Comparison of Models AUC - ROC curve

ROC bends normally highlight genuine positive rate on the Y axis, and false positive rate on the X axis. This implies the upper left corner of the plot is the "perfect" point - a FP rate of zero, and a TP rate of one. This isn't practical, however it means that a bigger territory under the bend (AUC) is typically better. The "steepness" of ROC bends is additionally significant, since it is perfect to amplify the genuine positive rate while limiting the bogus positive rate.

AUC-ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.

We can easily see that AUC of Logistic Regression is the best following with XG Boost and Random Forest.

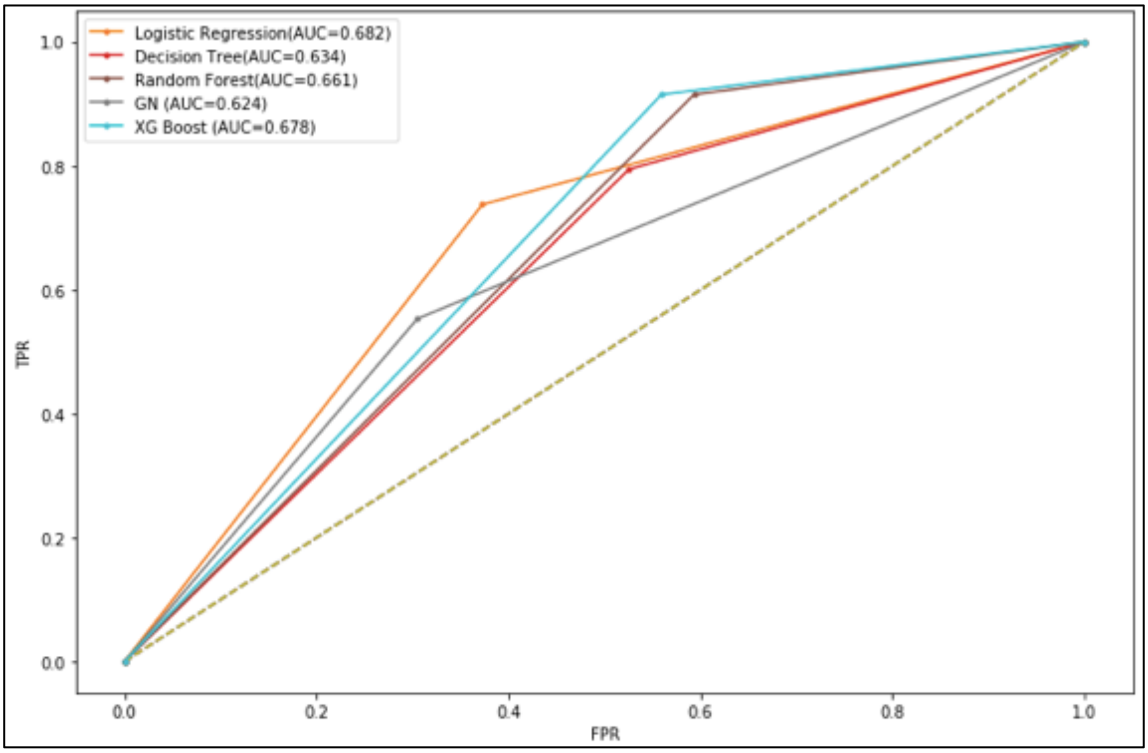
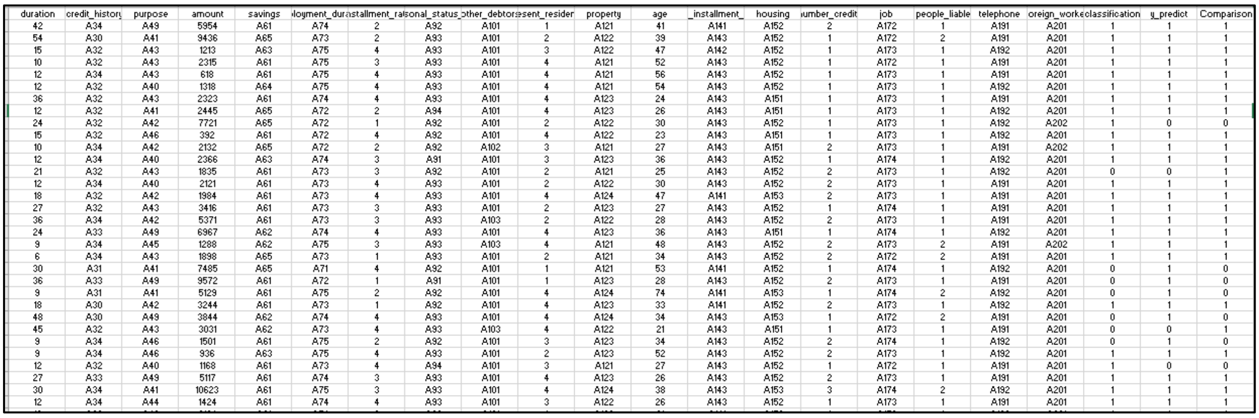


Figure 17: Comparison of Models AUC - ROC curve

6 RESULT

As we find the maximum accuracy in Random forest so we had used the result of that model and compared the predicted value of our dependent variable with its original value to find possible anomalies in our dataset.

For doing so we had combined our training data x variable unscaled value with corresponding rows of original y variable and its predicted value through our model. We had done it through python and Microsoft Access by joining our dataset tables through access queries.

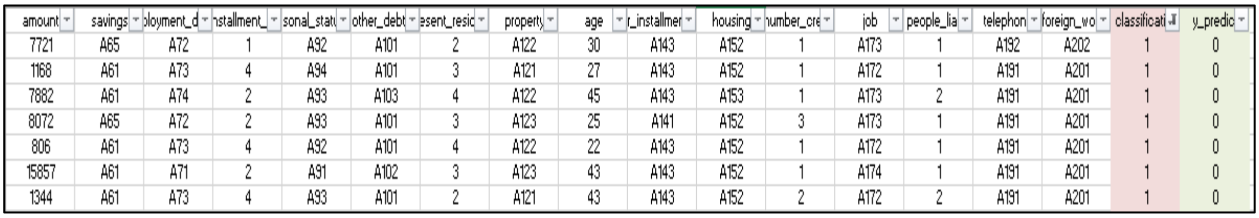


duration	credit_history	purpose	amount	savings	loyment_dur	nstallment_ra	sonal_status_	other_debtor	esent_resider	property	age	_installment_	housing	umber_credit	job	people_liable	telephone	oreign_worke	classification	y_predict	Comparison
42	A34	A49	5954	A61	A74	2	A92	A101	1	A121	41	A141	A152	2	A172	1	A191	A201	1	1	1
54	A30	A41	9436	A65	A73	2	A93	A101	2	A122	39	A143	A152	1	A172	2	A191	A201	1	1	1
15	A32	A43	1213	A63	A75	4	A93	A101	3	A122	47	A142	A152	1	A173	1	A192	A201	1	1	1
10	A32	A43	2315	A61	A75	3	A93	A101	4	A121	52	A143	A152	1	A172	1	A191	A201	1	1	1
12	A34	A43	618	A61	A75	4	A93	A101	4	A121	56	A143	A152	1	A173	1	A191	A201	1	1	1
12	A32	A40	1318	A64	A75	4	A93	A101	4	A121	54	A143	A152	1	A173	1	A192	A201	1	1	1
36	A32	A43	2323	A61	A74	4	A93	A101	4	A123	24	A143	A151	1	A173	1	A191	A201	1	1	1
12	A32	A41	2445	A65	A72	2	A94	A101	4	A123	26	A143	A151	1	A173	1	A192	A201	1	1	1
24	A32	A42	7721	A65	A72	1	A92	A101	2	A122	30	A143	A152	1	A173	1	A192	A202	1	0	0
15	A32	A46	392	A61	A72	4	A92	A101	4	A122	23	A143	A151	1	A173	1	A192	A201	1	1	1
10	A34	A42	2132	A65	A72	2	A92	A102	3	A121	27	A143	A151	2	A173	1	A191	A202	1	1	1
12	A34	A40	2366	A63	A74	3	A91	A101	3	A123	36	A143	A152	1	A174	1	A192	A201	1	1	1
21	A32	A43	1835	A61	A73	3	A92	A101	2	A121	25	A143	A152	2	A173	1	A192	A201	0	0	1
12	A34	A40	2121	A61	A73	4	A93	A101	2	A122	30	A143	A152	2	A173	1	A191	A201	1	1	1
18	A32	A42	1984	A61	A73	4	A93	A101	4	A124	47	A141	A153	2	A173	1	A191	A201	1	1	1
27	A32	A43	3416	A61	A73	3	A93	A101	2	A123	27	A143	A152	1	A174	1	A191	A201	1	1	1
36	A34	A42	5371	A61	A73	3	A93	A103	2	A122	28	A143	A152	2	A173	1	A191	A201	1	1	1
24	A33	A49	6967	A62	A74	4	A93	A101	4	A123	36	A143	A151	1	A174	1	A192	A201	1	1	1
9	A34	A45	1288	A62	A75	3	A93	A103	4	A121	48	A143	A152	2	A173	2	A191	A202	1	1	1
6	A34	A43	1898	A65	A73	1	A93	A101	2	A121	34	A143	A152	2	A172	2	A191	A201	1	1	1
30	A31	A41	7485	A65	A71	4	A92	A101	1	A121	53	A141	A152	1	A174	1	A192	A201	0	1	0
36	A33	A49	9572	A61	A72	1	A91	A101	1	A123	28	A143	A152	2	A173	1	A191	A201	0	1	0
9	A31	A41	5129	A61	A75	2	A92	A101	4	A124	74	A141	A153	1	A174	2	A192	A201	0	1	0
18	A30	A42	3244	A61	A73	1	A92	A101	4	A123	33	A141	A152	2	A173	1	A192	A201	1	1	1
48	A30	A49	3844	A62	A74	4	A93	A101	4	A124	34	A143	A153	1	A172	2	A191	A201	0	1	0
45	A32	A43	3031	A62	A73	4	A93	A103	4	A122	21	A143	A151	1	A173	1	A191	A201	0	0	1
9	A34	A46	1501	A61	A75	2	A92	A101	3	A123	34	A143	A152	2	A174	1	A192	A201	0	1	0
9	A34	A46	936	A63	A75	4	A93	A101	2	A123	52	A143	A152	2	A173	1	A192	A201	1	1	1
12	A32	A40	1168	A61	A73	4	A94	A101	3	A121	27	A143	A152	1	A172	1	A191	A201	1	0	0
27	A33	A49	5117	A61	A74	3	A93	A101	4	A123	26	A143	A152	2	A173	1	A191	A201	1	1	1
30	A34	A41	10623	A61	A75	3	A93	A101	4	A124	38	A143	A153	3	A174	2	A192	A201	1	1	1
12	A34	A44	1424	A61	A74	4	A93	A101	3	A122	26	A143	A152	1	A173	1	A191	A201	1	1	1

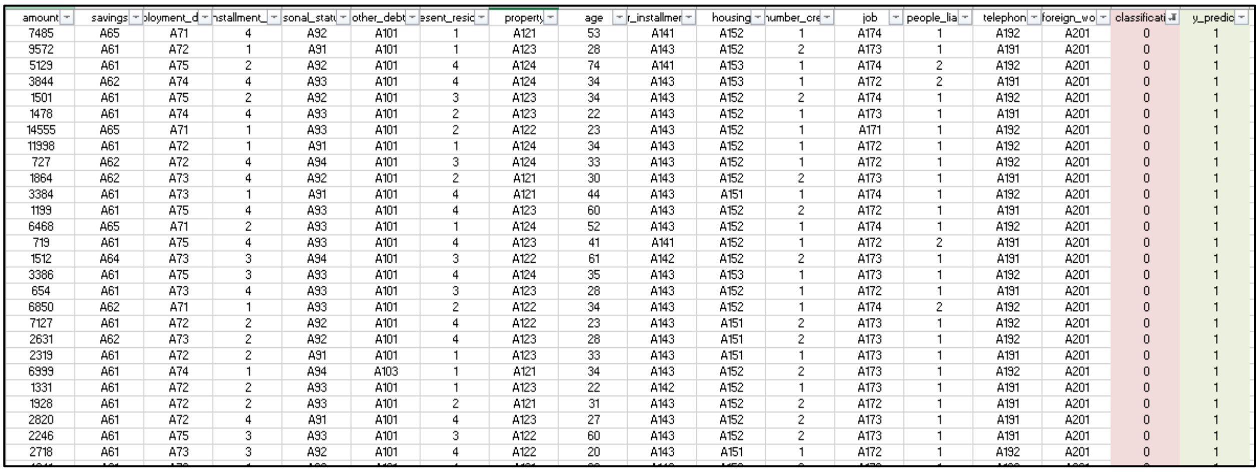
Figure 18: Excel sheet for results comparison

As we compare the prediction result with original values we find following anomalies in data for wrong prediction.

- We found that out of 200 observations our model predicts 46 observations falsely and 154 observations truly. As we drill down more to our result 39 observation predicted falsely were Bad risk or 0 to which model predict as a Good Risk or 1. Rest 7 observation were Good Risk or 1 to which model predict as Bad risk or 0.



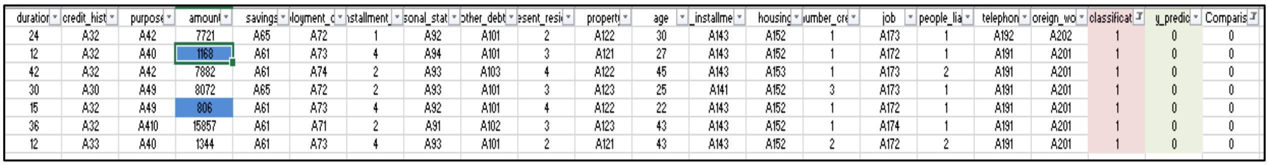
amount	savings	ployment_d	nstallment_	sonal_stat	other_debt	esent_resic	property	age	r_installmer	housing	number_cre	job	people_lia	telephon	foreign_wo	classificati	y_predic
7721	A65	A72	1	A92	A101	2	A122	30	A143	A152	1	A173	1	A192	A202	1	0
1168	A61	A73	4	A94	A101	3	A121	27	A143	A152	1	A172	1	A191	A201	1	0
7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1	0
8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1	0
806	A61	A73	4	A92	A101	4	A122	22	A143	A152	1	A172	1	A191	A201	1	0
15857	A61	A71	2	A91	A102	3	A123	43	A143	A152	1	A174	1	A191	A201	1	0
1344	A61	A73	4	A93	A101	2	A121	43	A143	A152	2	A172	2	A191	A201	1	0



amount	savings	ployment_d	nstallment_	sonal_statu	other_debt	esent_resid	property	age	r_installmer	housing	number_cre	job	people_lia	telephon	foreign_wo	classificati	y_predic
7485	A65	A71	4	A92	A101	1	A121	53	A141	A152	1	A174	1	A192	A201	0	1
9572	A61	A72	1	A91	A101	1	A123	28	A143	A152	2	A173	1	A191	A201	0	1
5129	A61	A75	2	A92	A101	4	A124	74	A141	A153	1	A174	2	A192	A201	0	1
3844	A62	A74	4	A93	A101	4	A124	34	A143	A153	1	A172	2	A191	A201	0	1
1501	A61	A75	2	A92	A101	3	A123	34	A143	A152	2	A174	1	A192	A201	0	1
1478	A61	A74	4	A93	A101	2	A123	22	A143	A152	1	A173	1	A191	A201	0	1
14555	A65	A71	1	A93	A101	2	A122	23	A143	A152	1	A171	1	A192	A201	0	1
11998	A61	A72	1	A91	A101	1	A124	34	A143	A152	1	A172	1	A192	A201	0	1
727	A62	A72	4	A94	A101	3	A124	33	A143	A152	1	A172	1	A192	A201	0	1
1864	A62	A73	4	A92	A101	2	A121	30	A143	A152	2	A173	1	A191	A201	0	1
3384	A61	A73	1	A91	A101	4	A121	44	A143	A151	1	A174	1	A192	A201	0	1
1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	0	1
6468	A65	A71	2	A93	A101	1	A124	52	A143	A152	1	A174	1	A192	A201	0	1
719	A61	A75	4	A93	A101	4	A123	41	A141	A152	1	A172	2	A191	A201	0	1
1512	A64	A73	3	A94	A101	3	A122	61	A142	A152	2	A173	1	A191	A201	0	1
3386	A61	A75	3	A93	A101	4	A124	35	A143	A153	1	A173	1	A192	A201	0	1
654	A61	A73	4	A93	A101	3	A123	28	A143	A152	1	A172	1	A191	A201	0	1
6850	A62	A71	1	A93	A101	2	A122	34	A143	A152	1	A174	2	A192	A201	0	1
7127	A61	A72	2	A92	A101	4	A122	23	A143	A151	2	A173	1	A192	A201	0	1
2631	A62	A73	2	A92	A101	4	A123	28	A143	A151	2	A173	1	A192	A201	0	1
2319	A61	A72	2	A91	A101	1	A123	33	A143	A151	1	A173	1	A191	A201	0	1
6999	A61	A74	1	A94	A103	1	A121	34	A143	A152	2	A173	1	A192	A201	0	1
1331	A61	A72	2	A93	A101	1	A123	22	A142	A152	1	A173	1	A191	A201	0	1
1928	A61	A72	2	A93	A101	2	A121	31	A143	A152	2	A172	1	A191	A201	0	1
2820	A61	A72	4	A91	A101	4	A123	27	A143	A152	2	A173	1	A191	A201	0	1
2246	A61	A75	3	A93	A101	3	A122	60	A143	A152	2	A173	1	A191	A201	0	1
2718	A61	A73	3	A92	A101	4	A122	20	A143	A151	1	A172	1	A192	A201	0	1

Figure 19: Comparison of false prediction

- To check the anomalies in data we first checked the 7 observation or False negative observation.



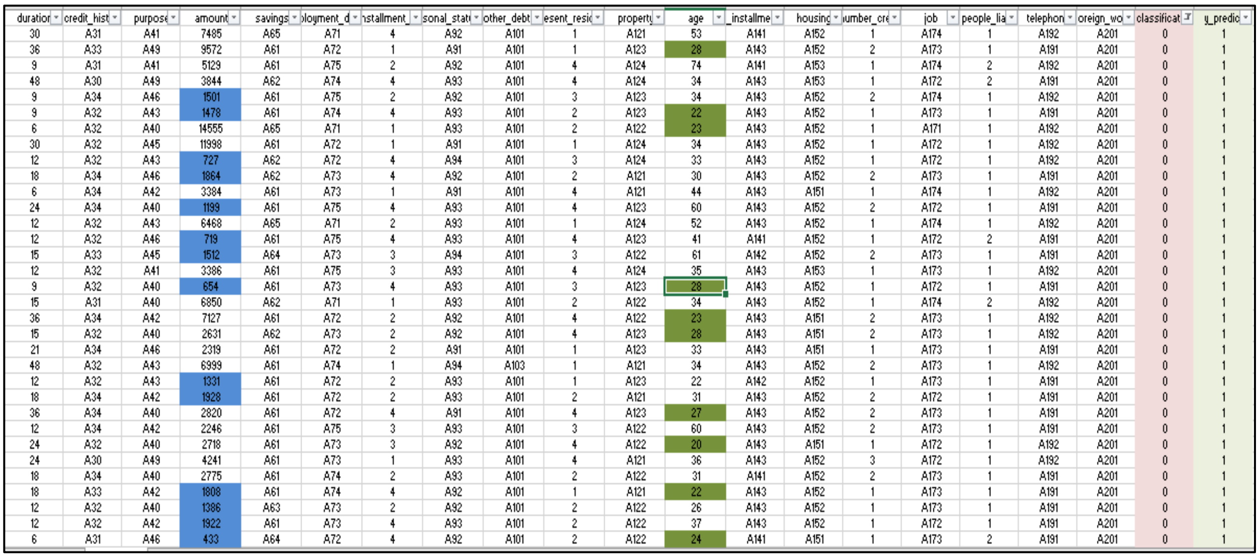
duration	credit_hist	purpose	amount	savings	loyment_d	nstallment_	sonal_stat	other_debt	esent_resi	property	age	_installme	housing	number_cre	job	people_lia	telephon	oreign_wo	classificat	y_predic	Comparis
24	A32	A42	7721	A65	A72	1	A92	A101	2	A122	30	A143	A152	1	A173	1	A192	A202	1	0	0
12	A32	A40	1168	A61	A73	4	A94	A101	3	A121	27	A143	A152	1	A172	1	A191	A201	1	0	0
42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1	0	0
30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1	0	0
15	A32	A49	806	A61	A73	4	A92	A101	4	A122	22	A143	A152	1	A172	1	A191	A201	1	0	0
36	A32	A410	15857	A61	A71	2	A91	A102	3	A123	43	A143	A152	1	A174	1	A191	A201	1	0	0
12	A33	A40	1344	A61	A73	4	A93	A101	2	A121	43	A143	A152	2	A172	2	A191	A201	1	0	0

Figure 20: Data anomalies in false negative observations

As we know from our previous phases the credit amount is the most important feature of our prediction. And as per our data visualization result the customer with Bad risk asked for higher credit amount on an average. And it is clear from the result that all the customer falsely predicts as bad risk though they are not being actually looking for higher credit amount.

However, two customers looking for less credit amount are falsely predict as the age of both the customers is less than 30. Age being the second most feature of dataset and our EDA results shows age group between 20-30 have the higher Bad Risk.

- Next we had checked the rest falsely predict 39 observation for False positive observation.



duratior	credit_hist	purpose	amount	savings	ployment_d	nstallment_	sonal_stat	other_debt	esent_resid	property	age	_installme	housing	umber_cre	job	people_lia	telephon	oreign_wo	classificat	y_predic
30	A31	A41	7485	A65	A71	4	A92	A101	1	A121	53	A141	A152	1	A174	1	A192	A201	0	1
36	A33	A49	9572	A61	A72	1	A91	A101	1	A123	28	A143	A152	2	A173	1	A191	A201	0	1
9	A31	A41	5129	A61	A75	2	A92	A101	4	A124	74	A141	A153	1	A174	2	A192	A201	0	1
48	A30	A49	3844	A62	A74	4	A93	A101	4	A124	34	A143	A153	1	A172	2	A191	A201	0	1
9	A34	A46	1501	A61	A75	2	A92	A101	3	A123	34	A143	A152	2	A174	1	A192	A201	0	1
9	A32	A43	1478	A61	A74	4	A93	A101	2	A123	22	A143	A152	1	A173	1	A191	A201	0	1
6	A32	A40	14555	A65	A71	1	A93	A101	2	A122	23	A143	A152	1	A171	1	A192	A201	0	1
30	A32	A45	11998	A61	A72	1	A91	A101	1	A124	34	A143	A152	1	A172	1	A192	A201	0	1
12	A32	A43	727	A62	A72	4	A94	A101	3	A124	33	A143	A152	1	A172	1	A192	A201	0	1
18	A34	A46	1864	A62	A73	4	A92	A101	2	A121	30	A143	A152	2	A173	1	A191	A201	0	1
6	A34	A42	3384	A61	A73	1	A91	A101	4	A121	44	A143	A151	1	A174	1	A192	A201	0	1
24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	0	1
12	A32	A43	6468	A65	A71	2	A93	A101	1	A124	52	A143	A152	1	A174	1	A192	A201	0	1
12	A32	A46	719	A61	A75	4	A93	A101	4	A123	41	A141	A152	1	A172	2	A191	A201	0	1
15	A33	A45	1512	A64	A73	3	A94	A101	3	A122	61	A142	A152	2	A173	1	A191	A201	0	1
12	A32	A41	3386	A61	A75	3	A93	A101	4	A124	35	A143	A153	1	A173	1	A192	A201	0	1
9	A32	A40	654	A61	A73	4	A93	A101	3	A123	28	A143	A152	1	A172	1	A191	A201	0	1
15	A31	A40	6850	A62	A71	1	A93	A101	2	A122	34	A143	A152	1	A174	2	A192	A201	0	1
36	A34	A42	7127	A61	A72	2	A92	A101	4	A122	23	A143	A151	2	A173	1	A192	A201	0	1
15	A32	A40	2631	A62	A73	2	A92	A101	4	A123	28	A143	A151	2	A173	1	A192	A201	0	1
21	A34	A46	2319	A61	A72	2	A91	A101	1	A123	33	A143	A151	1	A173	1	A191	A201	0	1
48	A32	A43	6999	A61	A74	1	A94	A103	1	A121	34	A143	A152	2	A173	1	A192	A201	0	1
12	A32	A43	1331	A61	A72	2	A93	A101	1	A123	22	A142	A152	1	A173	1	A191	A201	0	1
18	A34	A42	1928	A61	A72	2	A93	A101	2	A121	31	A143	A152	2	A172	1	A191	A201	0	1
36	A34	A40	2820	A61	A72	4	A91	A101	4	A123	27	A143	A152	2	A173	1	A191	A201	0	1
12	A34	A42	2246	A61	A75	3	A93	A101	3	A122	60	A143	A152	2	A173	1	A191	A201	0	1
24	A32	A40	2718	A61	A73	3	A92	A101	4	A122	20	A143	A151	1	A172	1	A192	A201	0	1
24	A30	A49	4241	A61	A73	1	A93	A101	4	A121	36	A143	A152	3	A172	1	A192	A201	0	1
18	A34	A40	2775	A61	A74	2	A93	A101	2	A122	31	A141	A152	2	A173	1	A191	A201	0	1
18	A33	A42	1808	A61	A74	4	A92	A101	1	A121	22	A143	A152	1	A173	1	A191	A201	0	1
12	A32	A40	1386	A63	A73	2	A92	A101	2	A122	26	A143	A152	1	A173	1	A191	A201	0	1
12	A32	A42	1922	A61	A73	4	A93	A101	2	A122	37	A143	A152	1	A172	1	A191	A201	0	1
6	A31	A46	433	A64	A72	4	A92	A101	2	A122	24	A141	A151	1	A173	2	A191	A201	0	1

Figure 21: Predicted false positive observation

In case of false positive observations, we can see maximum of the customers asking for low amount are predicted as a Good Risk or 1 though they are 0. While rest predicted falsely either have age between 20 to 30 or due to high credit duration.

7 DISCUSSION

7.1 EDA

- Most females that applied for a credit loan were **less than 30** while most of the males that applied for a loan ranged from their **20s-40s**
- Females were more likely to apply for a credit loan to buy **furniture and equipment**. (10% more than males)
- Males applied 2x more than females for a credit loan to invest in a **business**.
- The **younger age group** tended to ask slightly for higher loans compared to the older age groups.
- The young and elderly groups had the **highest ratio** of high risk loans which can be explained by their unemployment or part-time jobs
    - **45.29%** are in Bad Risk of all the clients that belong to the young age group
    - **44.28%** are in Bad Risk of the total amount of people considered in the elderly group.

- The loan are more prone to be classified as Bad Risk if the credit amount is high or the duration is high
- Loans required to buy **Cars**, **Radio/TV** and **Furniture and Equipment** are more at risk of being Bad, i.e., 50% of total bad risk.

7.2 Decision Tree

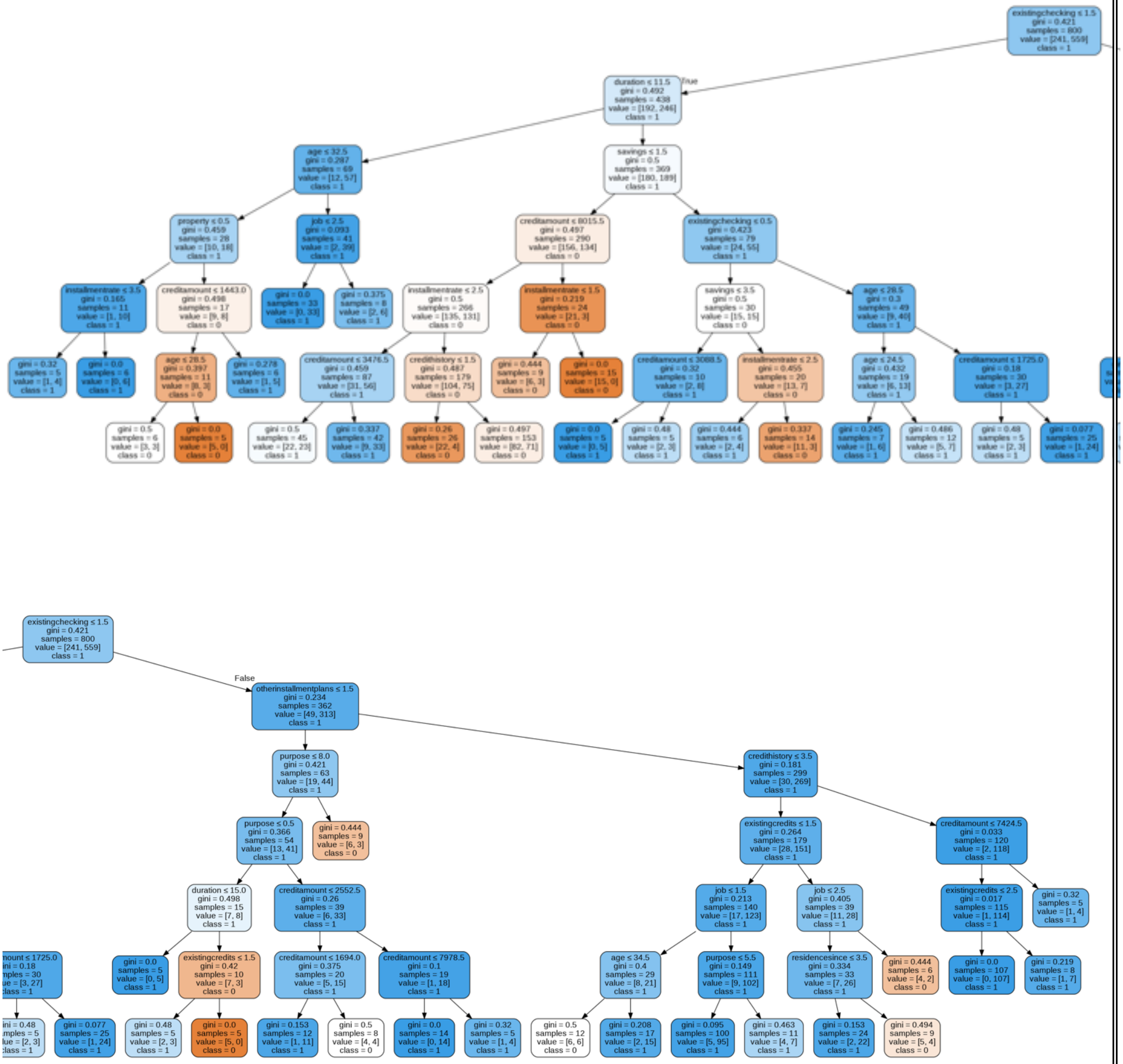


Figure 22: Decision Tree

Note: Please refer to appendix at the end to get details about categorical attribute references used below

- If installment rate is less than 3.5, property is less than 0.5, age is less 32.5, duration is less than 11.5, installment rate is less than 1.5, the customer is a good risk. Same is the case if job is less than 2.5 keeping age, duration and installment rate is less than 1.5.
- On the other hand If credit amount is less than 1443.0, property is more than 0.5, age is less than 32.5, duration is less than 11.5, installment rate is less than 1.5 – customer is a good risk.
- Users are more likely to default is instalment rate is less than 2.5 and credit amount is less than 8015.5, considering savings is less than 2.5, duration is less than 11.5 and installment rate is less than 1.5. But if the expected credit amount is 3476.5 user is a good risk, considering other factors remain same.
- Data also shows that if the installment rate is less than 1.5 for expected credit is more than 8015.5, the user is very certain to default for all loan amount and types regardless of other factors. People with savings of more than 3.5 and instalment rate more than 2.5 are more prone to default (are bad risk). But if credit amount is less than 3088.5, it is still a risk worth taking (Good Risk).
- Users with characteristics of existing checking of more than 0.5 and savings of more than 1.5 are good risk regardless of age and credit amount.
- Customers falling in the category of duration of less than 15 months, purpose of 8.0 or less and other instalment plans of less than 1.5 are a good risk but if existing credits is more than 1.5, the loan can default.
- But if the existing checking is more than 1.5, other instalment plans is less than 1.5 and purpose is more than 8, he/she at more risk of defaulting whereas if the purpose is more than 0.5 and expected credit amount is between 2552.5 and 7978.5 or less than 1694.0, it is a good risk.
- It is a good risk for a bank to give loan to customers with expected loan amount of less than 7424.5, credit history of more than 3.5 and other instalments of more than 1.5 if their existing checking account balance is more than 1.5.
- If the job is more than 1.5 and the age is more than 34.5, regardless of the purpose the customer will turn out to be a good risk, considering his existing credit is less than 1.5, credit history is less than 3.5, other instalment plans is more than 1.5 and existing checking is more than 1.5. But the risk is higher is the age is less than 34.5. But if the job is more than 2.5, the risk increases. In case the customer is resident for less than and the job is less 2.5, he is a good risk, else the bank can deny the loan.

- The customer is a good risk in all cases of existing credits, credit history or less than 3.5, other installment plans of more than 1.5 and existing checking of more 1.5. Except in either of cases where:
    - Age is less than 34.5, or
    - Job is greater than 2.5
    - Customer is a resident with more than 3.5

8 CONCLUSION

The study was conducted by using a dataset from UCI Machine Learning Repository with an objective to build a machine learning classification based model to segregate between a good risk and a bad risk. We have built six classifiers- Logistic Regression, Naive Baye's, Decision Tree, Random Forest, Neural Network and XGBoost. Results showed that Random forest classifier has the best accuracy (78%) amongst all, given the data set and methods applied. To improve the accuracy further, we have done few other improvements to our based line model which was our worst case scenario with all the features taken into consideration, to balance the dataset include Dummy Trap Improvement, Unbalanced dataset using SMOTE, Feature Selection with SMOTE and Feature Selection without SMOTE. We got the same accuracy of 78% in Dummy trap improvement using XGboost.

Even though Random Forest had the maximum accuracy among all applied models, logistics regression has the maximum AUC – 68.2, followed by XGBoost and Logistics regression with AUC values of 67.8 and 66.1 respectively. The reason can be found in the False Positives and False Negatives values shown in the results. Almost 23% of the values are False Negatives which is a result of biasness created in test data using available un-balanced data.

Also using decision tree some of the key factors considered for classification based on GINI scores are: existing checking account balance, duration of EMI, other instalment plans, age, savings, purpose and credit history. Other factors included, property, job, credit amount and residence since. One thing worth noticing is, characteristics like gender, personal status, people liable and telephone had no role to play in credit risk decision.

Once a customer is classified based on risk assessment, the credit limit value can be calculated – a credit amount which is based on different factors to minimize the risk and maximize the profit.

9 LIMITATION

As our research was based on limited German credit data which was available online, the research was limited by factors and data was unbalanced. Risk assessment prediction is limited by data availability and factors/features captured. Richer availability of data allows better prediction and higher accuracy.

The results cannot be generalized for all geographies and demographics as the factors taken into this research are specific to Germany and that too for the time period when the data was collected. As the credit patterns are highly dependent on economic status of individuals and hence of the country, it cannot be generalized. Considering the similar aspects, the models cannot not be used on archived data and should only be used on latest data to give better estimation of credit risk and credit amount.

Similarly, in case credit amount calculation, the training data has to have previous credit amount given and its dependency on credit risk to give better insights and ease of calculation.

The predicted outcomes are based on previous data using different machine learning models and can be should as a reference only. The results are not to be considered as 100% certain.

Also, most of the previous literature as well as this research incudes only one side – either the customer side or bank's side factors.

The most obvious limitation is that all models and predictions are based on available data which is as reliable as how it is obtained. There are always chances of data entry error, implementation error, time inconsistencies and economic uncertainties. Hence a constant monitoring is always needed.

10 FUTURE SCOPE

The research can be extended to be a model of generalized application. A model which can take into consideration different geographic and demographic independence. For example, models which can give same results in India as well in US considering different factors.

As future scope of this research, a model can be developed to include both external as well internal factors – factors of customers as well as of banks'. Along with that, the current model only uses data with loan purposes of smaller values, a more accurate model can be developed for business loans, consumer loans, real estate loans and non-financial risks.

Models currently being developed uses mostly the archived data; a real-time analysis would make better impact on decision making, giving more weights to latest data than previous data. Also, the models being developed uses a few factors whereas it should have as many factors possible which needs models like neural network. Such models will allow better customer profiling using latest technologies of Big Data and Cloud Architecture.

An advancement to the model can be an interface which can be used by both the sides – customer as well as bank. A portal where either of the parties can add respective details to know the risk involved and credit eligibility.

11 BUSINESS IMPLICATIONS

One of the major concerns of banking industry is credit default by the customer. While giving credit to a customer bank faces two scenarios, firstly it is likely that the customer will repay the credit amount & get designated with a good risk for the bank and secondly, customer could not able to pay & became a bad risk to the bank. Bad risk is a gap area and impact a lot to its bottom line.

In addition to loan, credit risk assessments can be applied to other financial instruments like foreign exchange transactions, swaps, bonds, equities, options acceptances, trade financing, and interbank transactions etc.

By finding the credit worthiness of banks can rule out the possibility of default, also minimising the risk will maximize the profit of the bank. In order to decide the risk status of a customer, the business manager needs a decision rule for approval of the loan application which this binary classification based model will provide.

The models used in the research allows a business to identify factors and their weightage in decision making of risk assessment.

12 BIBLIOGRAPHY

Li Gan, &. R. (2008). AN EMPIRICAL STUDY OF THE CREDIT MARKET WITH UNOBSERVED CONSUMER. *NATIONAL BUREAU OF ECONOMIC RESEARCH*.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural. *Neural networks for credit risk evaluation: Investigation of different neural*.

Zala Herga, &. J. (2016). Modeling Probability of Default and Credit Limits. *researchgate*.

N., K., Kristovska I., & M., K. (2016). CREDIT RISK MANAGEMENT IN COMMERCIAL BANKS . *CREDIT RISK MANAGEMENT IN COMMERCIAL BANKS* .

HON, P. S., & BELLOTTI, T. (n.d.). Models and forecasts of credit card balance. *Models and forecasts of credit card balance*.

Flood, M. H. (July 2017). *Early identification of high-risk credit card customers based on behavioral data.* Norwegian University of Science and Technology.

Liu, R. (2018). Machine Learning Approaches to Predict Default of Credit Card Clients. *Machine Learning Approaches to Predict Default of Credit Card Clients*.

Motwani, A., Chaurasiya, P., & Bajaj, G. (2018). Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud. *Predicting Credit Worthiness of Bank Customer with Machine Learning Over Cloud*.

Stephen Zamore, Kwame Ohene Djan, Ilan Alon, Bersant Hobdari. (2018, March). *Credit Risk Research: Review and Agenda.* Retrieved from Research Gate: https://www.researchgate.net/publication/323430569_Credit_Risk_Research_Review_and_Agenda

Changjun Zheng, Niluthpaul Sarker, Shamsun Nahar. (2018, January). *Factors affecting bank credit risk: An empirical insight.* Retrieved from Research Gate: https://www.researchgate.net/publication/325115644_Factors_affecting_bank_credit_risk_An_empirical_insight

Garr, D. K. (2013). *Determinants of Credit Risk in the Banking Industry of Ghana.* Retrieved from Semantic Scholar: https://pdfs.semanticscholar.org/6bd7/43874fae07f997f0af07a8bcc20580e8f164.pdf

Nitesh V. Chawla, K. W. (2002, June). *SMOTE: Synthetic Minority Over-sampling Technique.* Retrieved from arxiv.org: https://arxiv.org/pdf/1106.1813.pdf

13 APPENDIX

Please refer to the below categorical variables (numeric codes for attribute description) for discussion and outcome section above:

Existing Checking:

0 : ... < 0 DM
1 : 0 <= ... < 200 DM
2 :>= 200 DM / salary assignments for at least 1 year
3 : no checking account

Duration: In Months (in numerical)
Age: In Years (in numerical)
Property:

0: real estate
1 : if not A121: building society savings agreement/ life insurance
2 : if not A121/A122: car or other, not in attribute 6
3 : unknown / no property

Instalment rate (in numerical)

Credit Amount : In Numbers (in numerical)

Job:

0 : unemployed/ unskilled - non-resident
1 : unskilled - resident
2 : skilled employee / official
3 : management/ self-employed/ highly qualified employee/ officer

Savings:

0 :< 100 DM
1 : 100 <= ... < 500 DM
2 : 500 <= ... < 1000 DM
3 : ... >= 1000 DM
4 : unknown/ no savings account

Credit History:

0 : no credits taken/ all credits paid back duly
1 : all credits at this bank paid back duly
2 : existing credits paid back duly till now
3 : delay in paying off in the past
4 : critical account/ other credits existing (not at this bank)

Other Instalment Plans:

0 : none
1 : co-applicant
2 : guarantor

Purpose:

0 : car (new)

1 : car (used)
2 : furniture/equipment
3 : radio/television
4 : domestic appliances
5 : repairs
6 : education
7 :(vacation - does not exist?)
8 : retraining
9 : business
10 : others

Existing Credits:

0 : bank
1: stores
2 : none

Residence Since: (in numerical)