Project: Creditworthiness

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?
- What data is needed to inform those decisions?
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

We have been asked to make a classification model to determine whether the new loan applicant is creditworthy of the loan or not. To make this prediction, we have the data of the previous applicants and their credit worthy result we found out 12 significant variables namely Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset, Age-years, Type-of-apartment, No-of-Credits-at-this-Bank.

We need to use a binary classification model in order to classify if the applicant is creditworthy or non-creditworthy.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

<u>Here are some guidelines to help guide your data cleanup:</u>

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and

you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

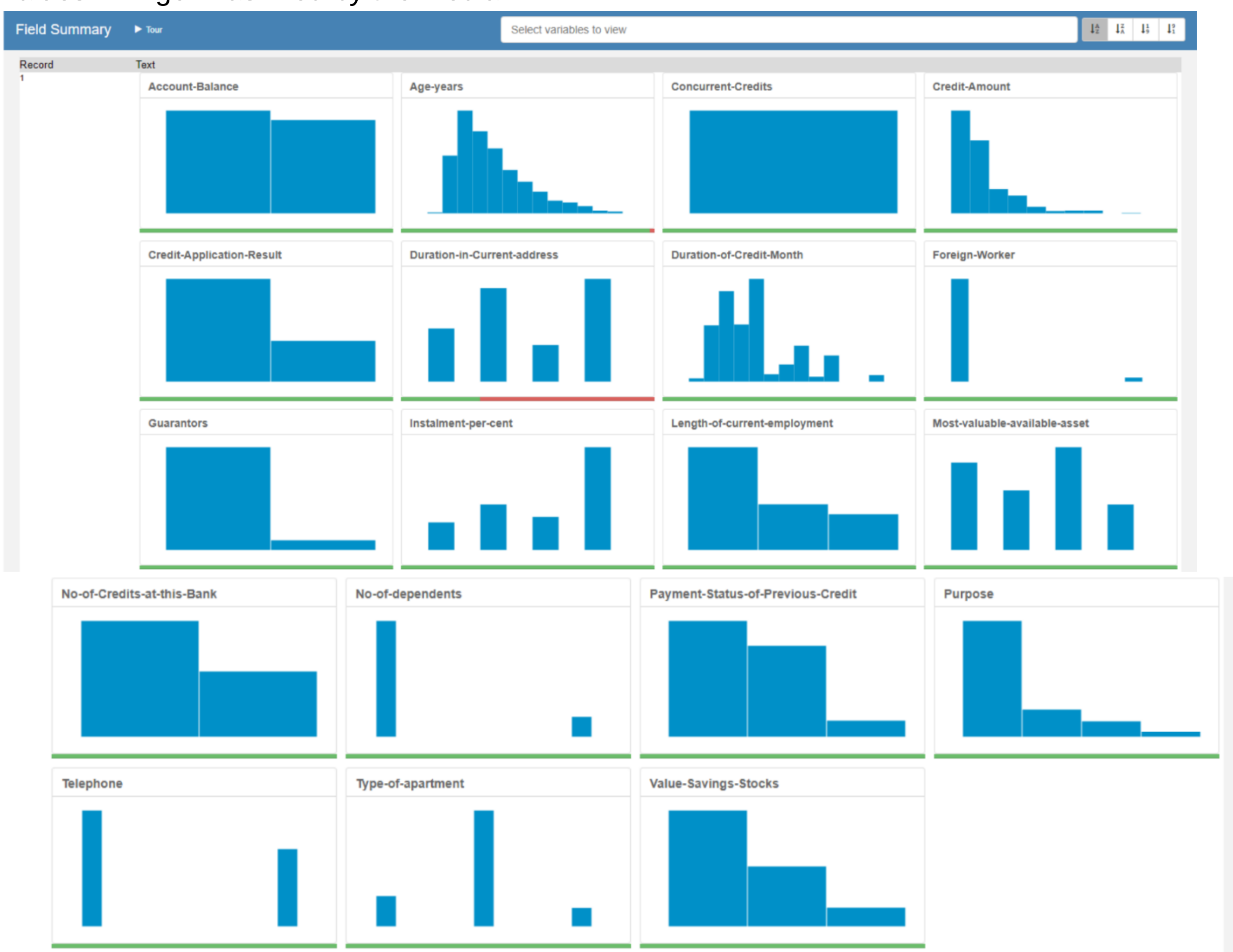
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Answer this question:

- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

As a clean-up process, many variables were filtered out. “Duration in current address” was removed as it consists of large missing values. ”Concurrent credits” and “occupation” were removed for having zero variance. ”Foreign Workers”, ”No of dependents”, “Guarantors” had low variability, “telephone” was an irrelevant attribute and hence these were removed. The missing values in “Age” was filled by the median.



There were no variables with correlation greater than 0.7.

Pearson Correlation Analysis

Full Correlation Matrix

	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Duration.of.Credit.Month	1.0000000	0.5704408	0.0795146	0.3047342	-0.0663189	0.1531405
Credit.Amount	0.5704408	1.0000000	-0.2856309	0.3277621	0.0686430	0.1686831
Instalment.per.cent	0.0795146	-0.2856309	1.0000000	0.0781104	0.0405397	0.0829360
Most.valuable.available.asset	0.3047342	0.3277621	0.0781104	1.0000000	0.0854367	0.3796504
Age.years	-0.0663189	0.0686430	0.0405397	0.0854367	1.0000000	0.3330748
Type.of.apartment	0.1531405	0.1686831	0.0829360	0.3796504	0.3330748	1.0000000
No.of.dependents	-0.0604413	0.0055003	-0.1164661	0.0507817	0.1177351	0.1707221
Telephone	0.1475443	0.2920589	0.0255102	0.1909078	0.1764790	0.0953716
Foreign.Worker	-0.1064163	0.0318954	-0.1182555	-0.1405878	-0.0032847	-0.0968173

	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	-0.0604413	0.1475443	-0.1064163
Credit.Amount	0.0055003	0.2920589	0.0318954
Instalment.per.cent	-0.1164661	0.0255102	-0.1182555
Most.valuable.available.asset	0.0507817	0.1909078	-0.1405878
Age.years	0.1177351	0.1764790	-0.0032847
Type.of.apartment	0.1707221	0.0953716	-0.0968173
No.of.dependents	1.0000000	-0.0461802	0.0412103
Telephone	-0.0461802	1.0000000	-0.0494452
Foreign.Worker	0.0412103	-0.0494452	1.0000000

Matrix of Corresponding p-values

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

*Answer these questions for **each model** you created:*

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

You should have four sets of questions answered. (500 word limit)

Logistic Regression

All twelve selected features were used for the logistic regression model. It is seen that eight features were found to be significant excluding the intercept which is shown in the below figure. The significant features have asterisk at the end of the row.

Record	Report
	Report for Logistic Regression Model LogReg
	Basic Summary
	Call: glm(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years, family = binomial("logit"), data = the.data)
	Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 ***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

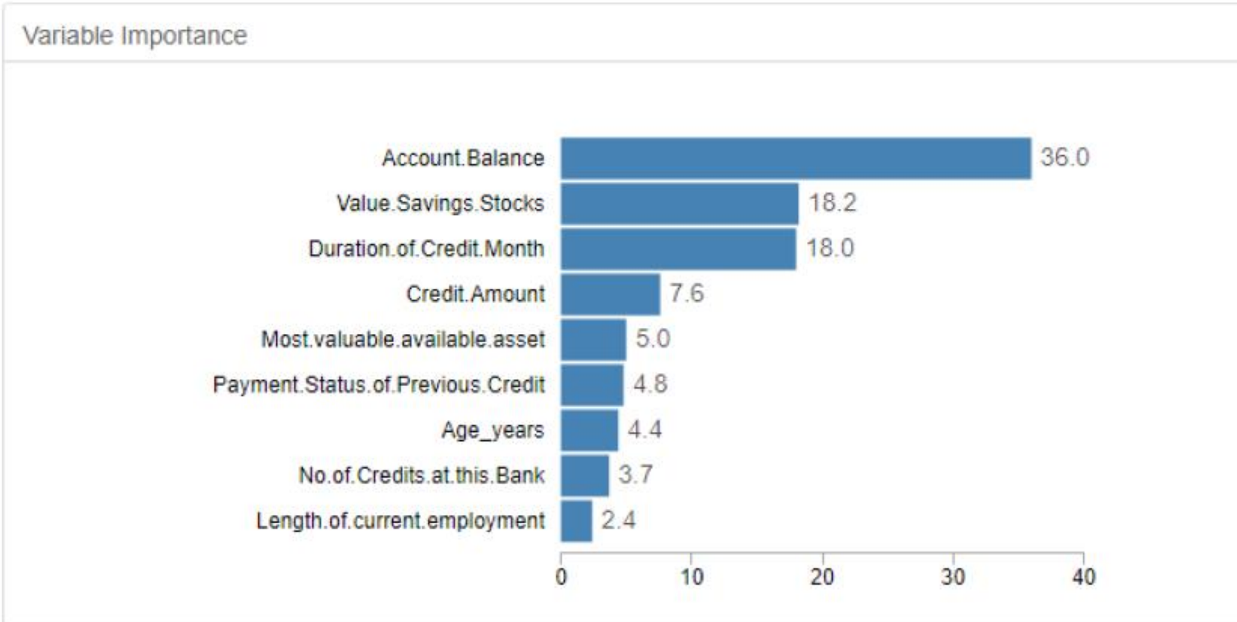
I achieved the accuracy of 78% on the validation dataset. Looking at the confusion matrix, we can see that the model is a biased towards the creditworthy applicants as prediction accuracy for non-credit worthy applicants is less.

Confusion matrix of LogReg

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Decision Tree

Nine features were present in the decision tree. The significant values found using Gini impurity values were Account Balance, Duration of credits and value.savings.stock.



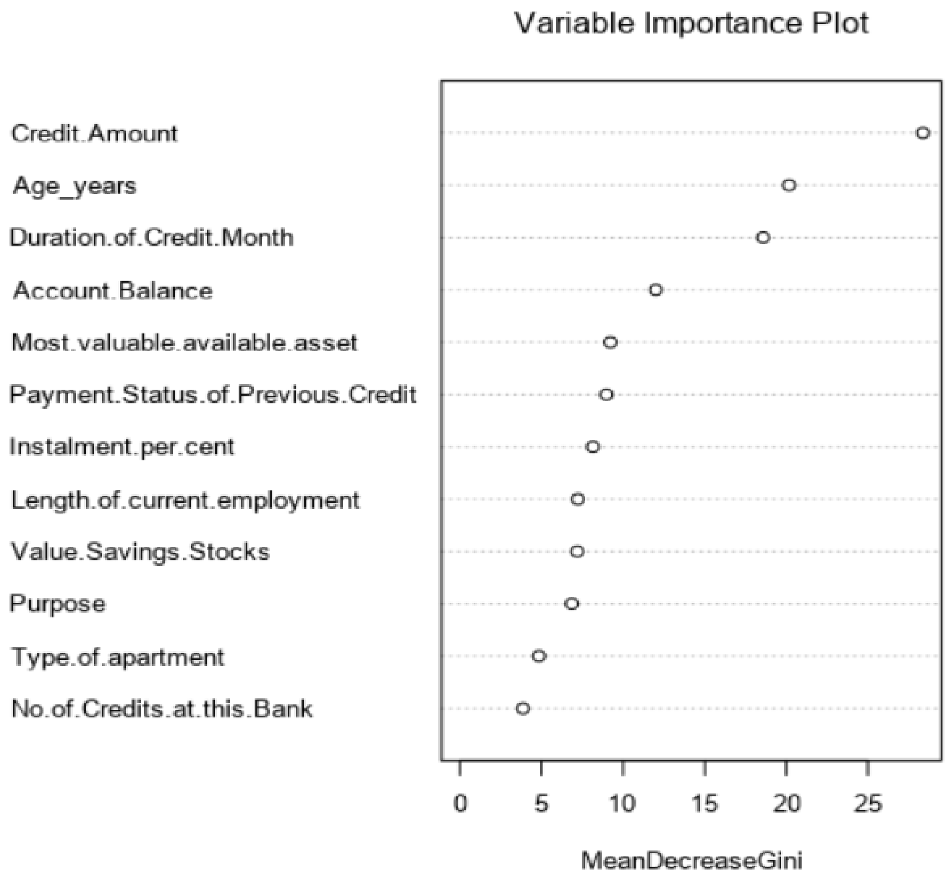
For the validation dataset , the accuracy of the model was 74.67%. Looking at the confusion matrix, it is seen that this model also doesnot perform very good for not creditworthy data. It is biased towards creditworthy applicant.

Confusion matrix of Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Forest Model

This model was generated using 500 trees. The out of bag error rate is found to be 24%. To find the most important features, we use mean decrease Gini values. Credit_Amount Age_years and Duration_of_credit_month were found to be most important.



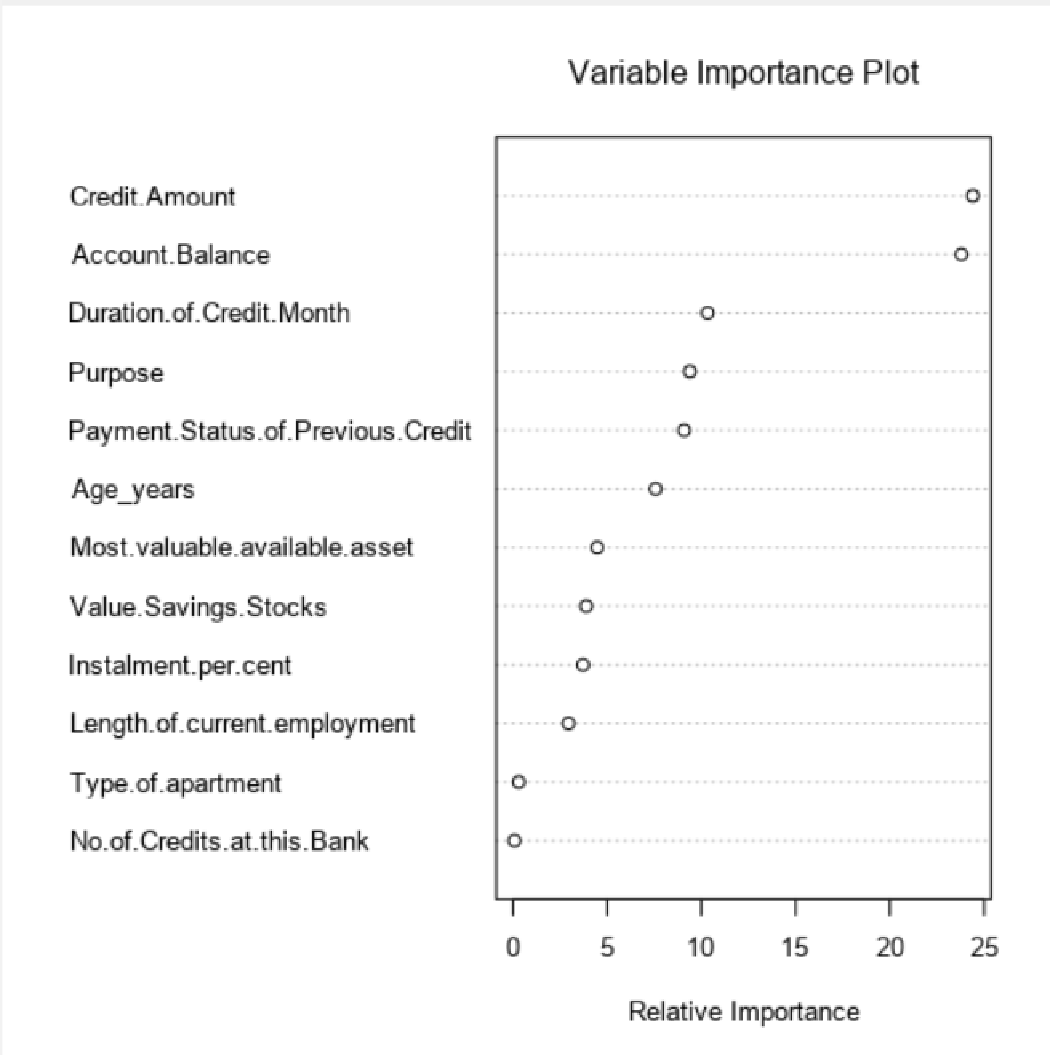
We achieved the accuracy of 80% with this model in the validation dataset. From the confusion matrix, we can illustrate that there in improvement in correctly classifying creditworthy (96.2%) and only 42% for non-credit worthy applicant. The model is biased for the credit worthy applicants.

Confusion matrix of Tree_Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Boosted Model

A gradient boosted model with ensemble of 4000 trees and Bernoulli loss function was generated. The below graph suggest that credit amount and account balance are the most important variable to this model.



The boosted model gave an accuracy of 78.8% for the validation dataset. The evaluation of confusion matrix shows that the boosted method performed worst for non-credit worthy applicants (37.3%) and good for creditworthy (96.2%). Even this model is biased for creditworthy applicants.

Confusion matrix of Tree_Boost

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
  - Overall Accuracy against your Validation set
  - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
  - ROC graph
  - Bias in the Confusion Matrices

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

ANSWER

We had to create a model for correctly classifying the creditworthiness of new loan applicants. I created 4 standard models i.e. Logistic Regression, Decision Tree, Forest Model and Boosted Model. Then I used a model comparison tool to compare the models.

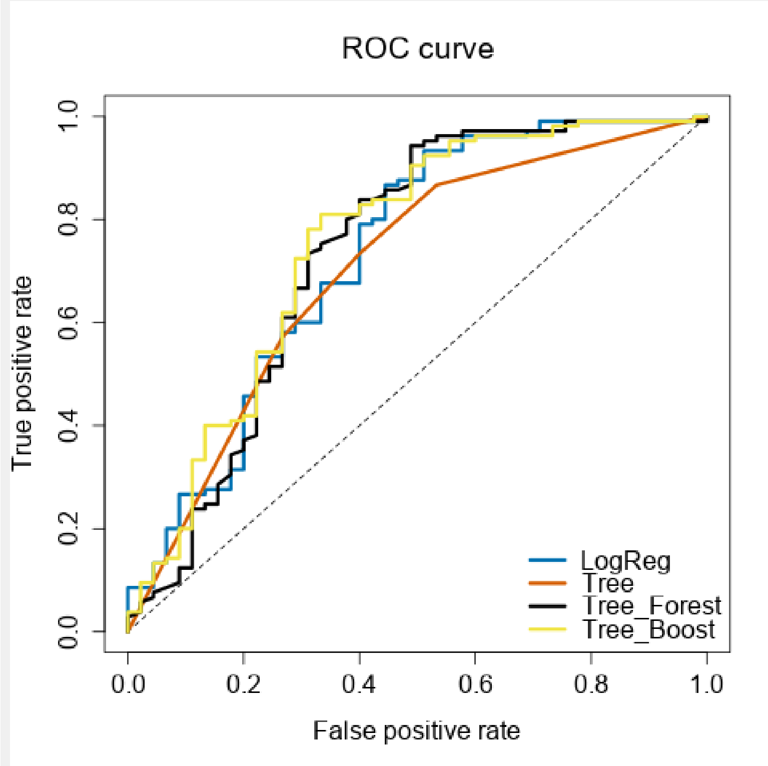
Model Comparison Report

Fit and error measures

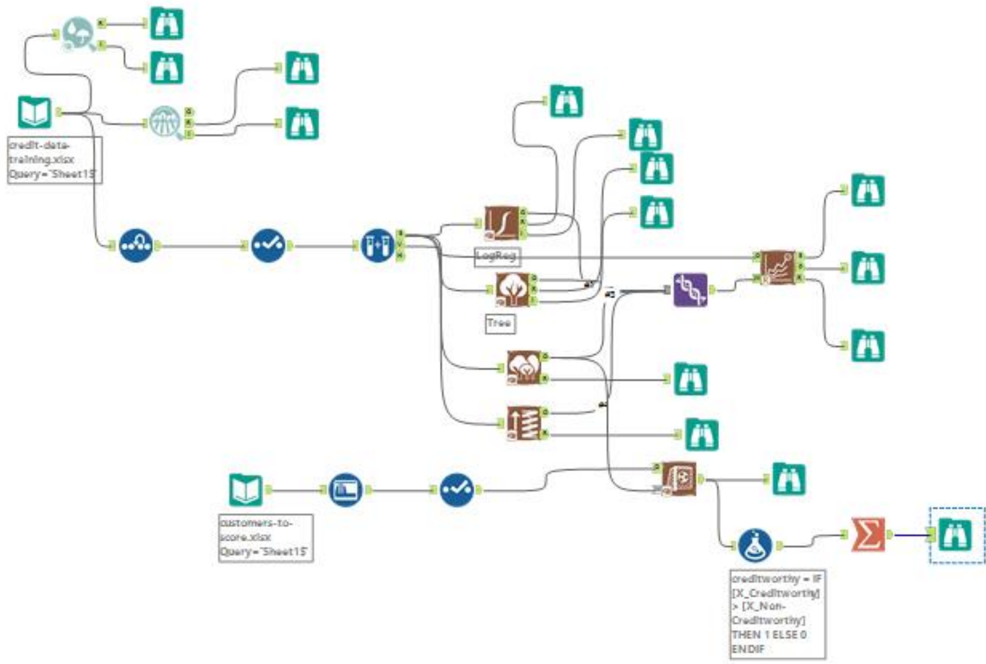
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogReg	0.7800	0.8520	0.7314	0.9048	0.4889
Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Tree_Forest	0.8000	0.8707	0.7361	0.9619	0.4222
Tree_Boost	0.7867	0.8632	0.7524	0.9619	0.3778

All the models had high accuracy (74.67%-80%). Also, all the models are biased towards the prediction of creditworthy applicants. The accuracy for predicting the non-creditworthy applicants was very bad (37.78%-48.89%) while for creditworthy applicants, it was good (86.67%-96.19%). The reason for this is the highly imbalanced dataset.

The ROC curve suggest that the boosted method performs best and then the forest method as shown in the below plot.



Since, the boss only cares about the prediction accuracy, forest model has the highest accuracy of 80%. Also, when compared to boosted model, it has high specificity (42.2%), we prefer Forest model. This model classified 406 new applicants as creditworthy.


credit-data-training.xlsx Query=`Sheet1$`
LogReg
Tree
customers-to-score.xlsx Query=`Sheet1$`
creditworthy = IF [X_Creditworthy] > [X_Non-Creditworthy] THEN 1 ELSE 0 ENDIF