Project: Creditworthiness

Business and Data Understanding

- What decisions needs to be made?
  Because of financial problem , there is sudden influx of 500 loan applications. The aim is to find out whether the applicants are enough creditworthy or not.
  Also we need to find how many applicants are creditworthy for the loan approval.

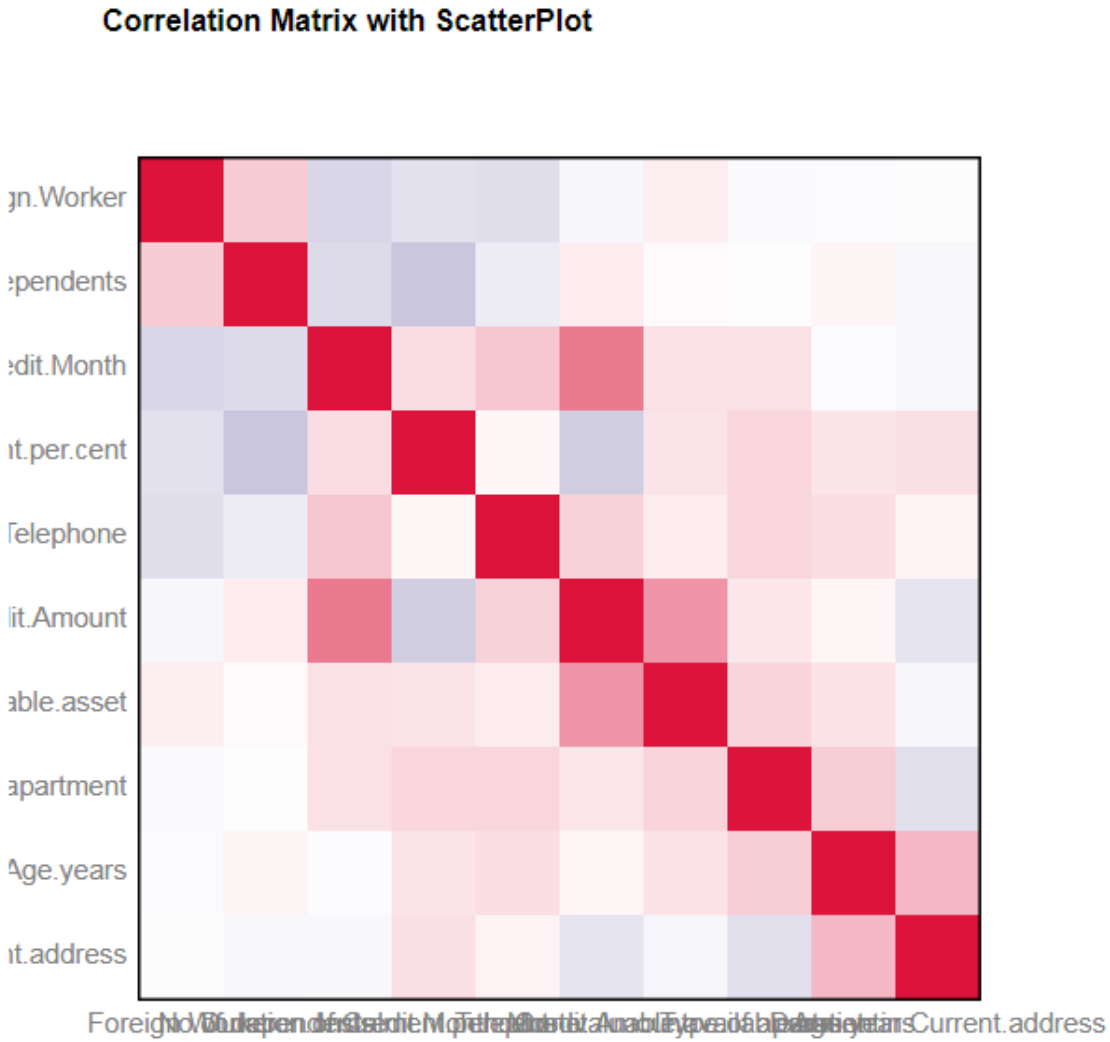
- What data is needed to inform those decisions?
  There is lot of data available with us but the data which is needed or we can use for finding creditworthy applicants is -
  Data of applications which were applied previously,
  Account Balance of the applicants,
  Credit amount and the list of applicants.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  As we have to make choice between two decisions whether the applicant is creditworthy or not, we will use binary model to make the decision.

Building the Training Set

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

Association analysis was done on the data provided and the result with the correlation matrix shows that there is no correlation between the variables i.e no variables are correlated with each other (>0.7)



After association analysis, field summary on variables is done which shows the following visuals.

From the visuals above we can see that Duration in Current address has about 69% of missing values therefore it is removed.

Since Age-years has only 2 % of missing data and age plays vital role in determining whether the applicant should be granted loan or not, so age is imputed.

Guarantors, Foreign Workers , No. of Dependents have 80% of the values towards one data which shows that the data is skewed, so these are removed in order to make our analysis unbiased.

Also Concurrent Credits & Occupation has just one unique value so they are removed.

Telephone is removed as it has no relevance for identifying whether the applicant is creditworthy or not.

Train your Classification Models

a. Logistic Regression (Stepwise)

As we want to check the creditworthiness of applicant Credit Application Result is considered as target variable.

Report for Logistic Regression Model Default_Risk

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	2.3595	0.0183 *
PurposeNew car	-1.6993164	6.142e-01	-2.7668	0.00566 **
PurposeOther	-0.3257637	8.179e-01	-0.3983	0.69042
PurposeUsed car	-0.7645820	4.004e-01	-1.9096	0.05618 .
Credit.Amount	0.0001704	5.733e-05	2.9716	0.00296 **
Length.of.current.employment4-7 yrs	0.3127022	4.587e-01	0.6817	0.49545
Length.of.current.employment< 1yr	0.8125785	3.874e-01	2.0973	0.03596 *
Instalment.per.cent	0.3016731	1.350e-01	2.2340	0.02549 *
Most.valuable.available.asset	0.2650267	1.425e-01	1.8599	0.06289 .

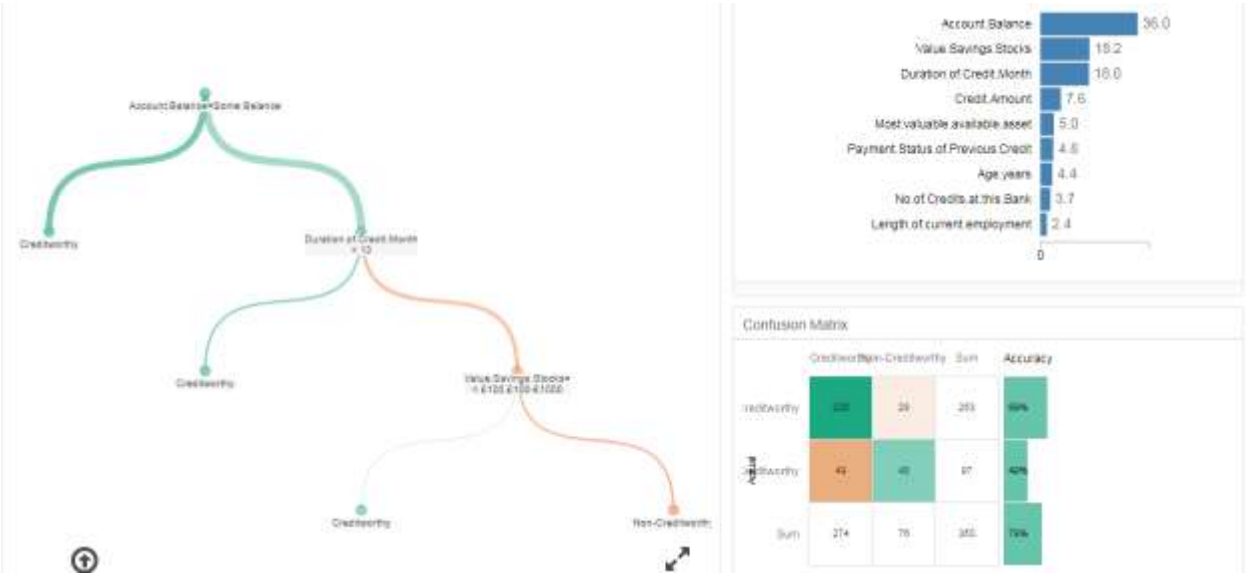
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

After running logistic regression model we can see from the above report that -
Account Balance, Purpose and Credit Amount are significant variables.

Overall Accuracy – 76%

b. Decision Tree



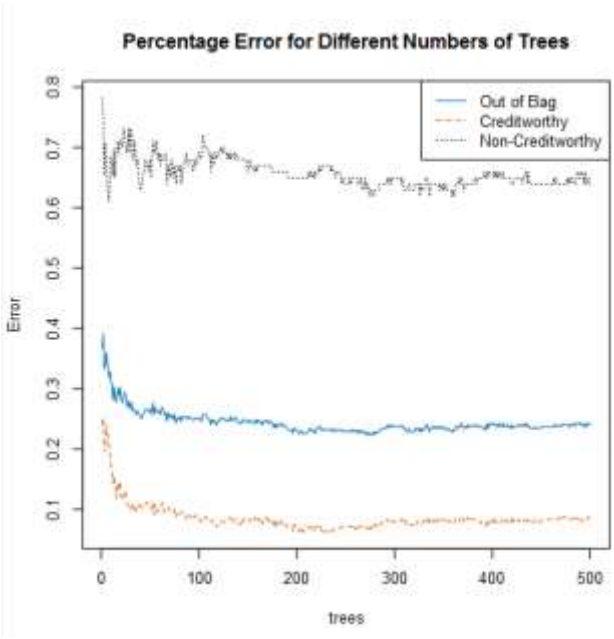
After logistic regression we run decision tree model to make our decision

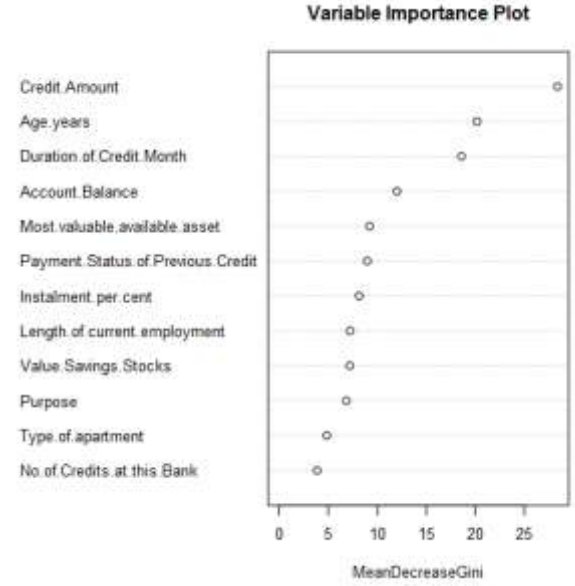
The results we can see is that -

Account Balance ,Values saving stocks and Duration for Credit month are variables with importance.

Overall Accuracy – 74%

c. Forest Model

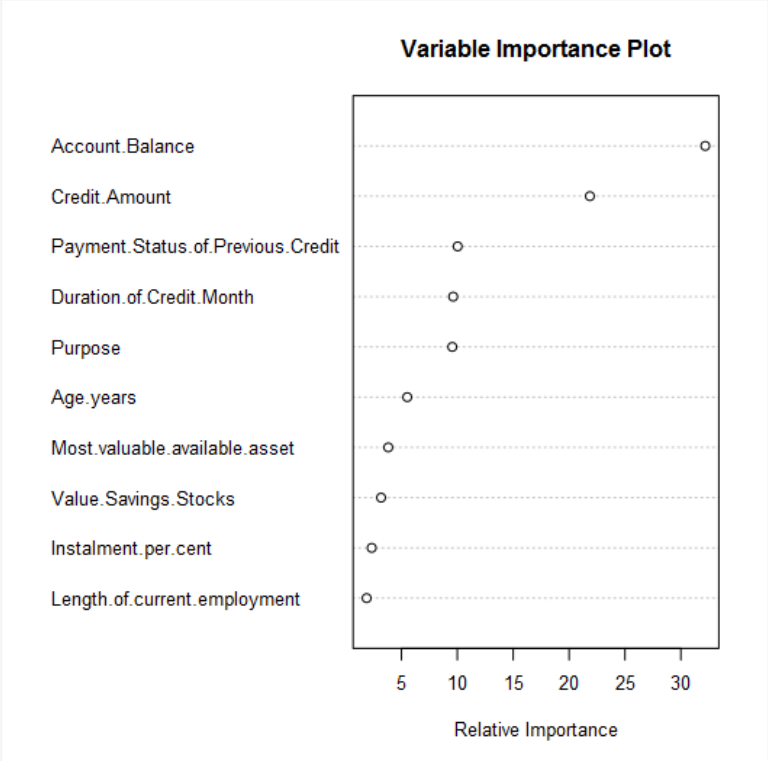




Executing the forest model on our data we see that
Credit Amount, Age Years & Duration of Credit Month are important variables.

Overall Accuracy – 80%

d. Boosted Model



In this last model i.e. boosted model the results we can interpret that only two variables are significant Account Balance & Credit Amount.

Overall Accuracy – 78%

Writeup

After executing all the four models, these models are compared with each other in order to select the best model which gives accurate results

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
decision_tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest	0.8000	0.8707	0.7361	0.9619	0.4222
Boost	0.7867	0.8632	0.7524	0.9619	0.3778
Default_Risk	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boost

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

Confusion matrix of Default_Risk

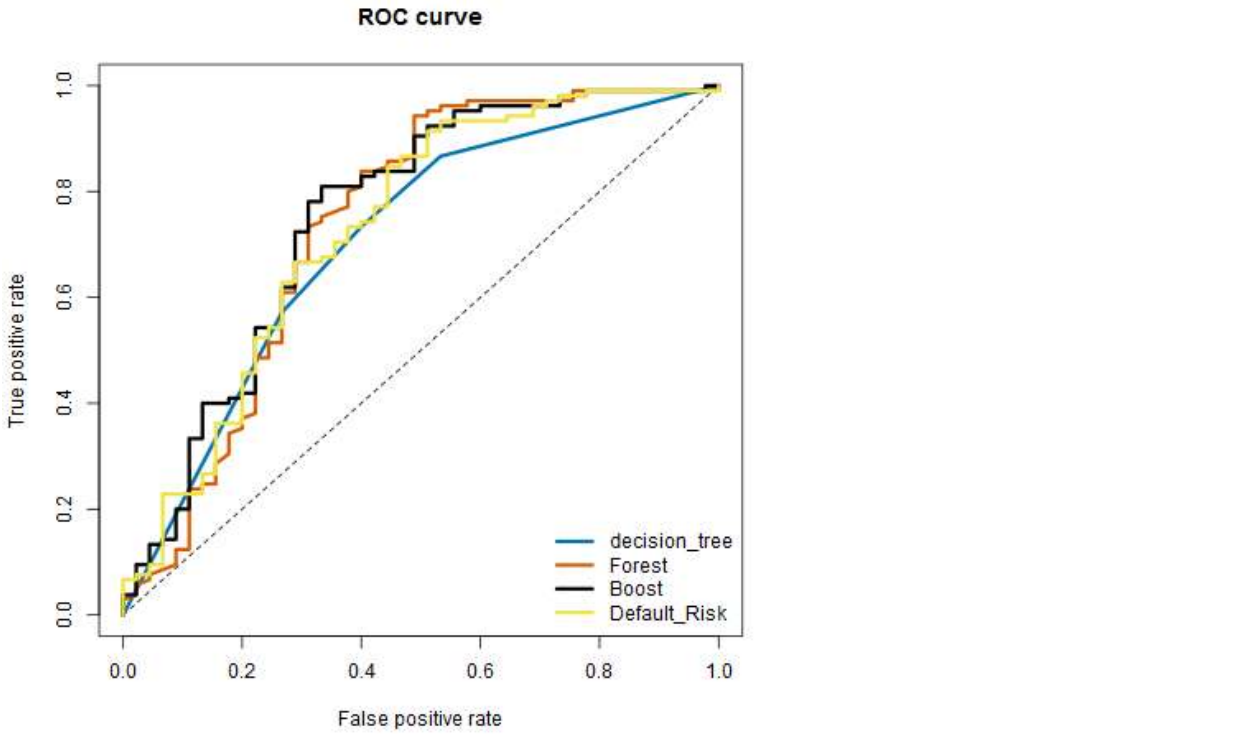
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Confusion matrix of Forest

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

Confusion matrix of decision_tree

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



After comparing the models with each other, from the report we can conclude that accuracy of Forest Model is highest among all of them.
Its accuracy for credit worthy and non creditworthy is also highest among all.

The Roc curve shows that the forest model has the best overall true positive rate. Also the difference between the creditworthy and non creditworthy is second least but the overall accuracy is highest so this model is preferred.

So we use forest model in order to make our decision and calculate creditworthy applicants.

After executing forest model on our data, we conclude that

There are **406** creditworthy applicants which can be approved for loan.