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?
    - Whether a new customer is creditworthy. A list of creditworthy customers is to be provided based on which it will be decided whether the new customer is creditworthy or not.

- What data is needed to inform those decisions?
    - Data required to build the model will be all the data on past application or customers. Data for new customers which is to be predicted.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    - Binary classification model will help us to decide.

Step 2: Building the Training Set

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

The fields that were removed or imputed are:

- *Duration in Current address.*
- *Guarantors*
- *Concurrent credits*
- *Occupation*
- *No. of dependents*
- *Foreign Worker*
- *Age*

Duration in Current Address:

In this variable there are around 69 % missing values. So, whenever there are missing values more than 50%, we avoid imputation and drop the variable.

Type	Records	Data Type Size
Double	**500**	**8**

Ok	156	31.20%
Null	344	68.80%

Concurrent Credits and Occupation:

In this variable all 100 % of the values are of same class. So, we have nothing to predict and this variable will not be helpful in model building so we drop it.

Concurrent-Credits

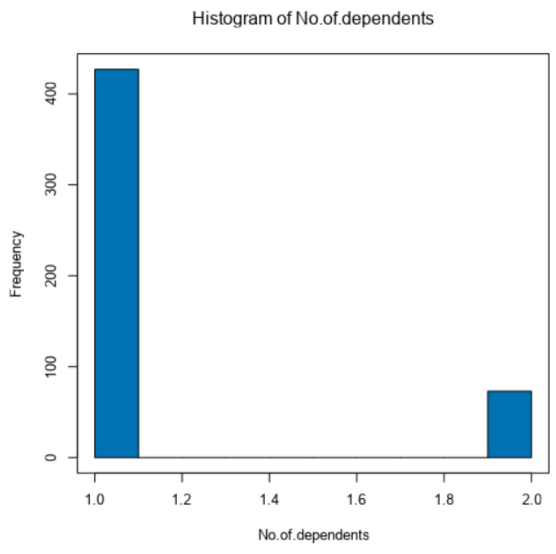
Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Other Banks/Depts	500	100.00	500	100.00

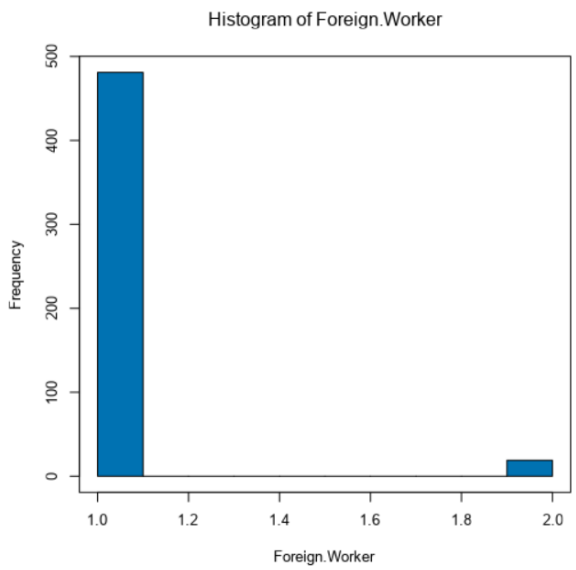
Guarantors

Value	Frequency	Percent	Cumulative Frequency	Cumulative Percent
None	457	91.40	457	91.40
Yes	43	8.60	500	100.00

Guarantors, No of Dependents and Foreign Worker:

In these variables there is low variability between classes. As we can see more than 90 % of the values are of class 1. So, it will be difficult to predict the other class due to a smaller number of observations. We drop the variable.





Age:

In this variable there are around 3 % of missing values. So, we will impute missing values in this case. We have done imputation using median for the column.

Step 3: Train your Classification Models

*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 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?

Logistic Regression:

	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	.

- The significant variables are account balance, purpose of loan, credit amount.
- The overall accuracy percentage is 76 %.

Confusion matrix of X

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

Decision Tree:

Model Summary

Variables actually used in tree construction:

[1] Account.Balance Duration.of.Credit.Month Purpose Value.Savings.Stocks

Root node error: 97/350 = 0.27714

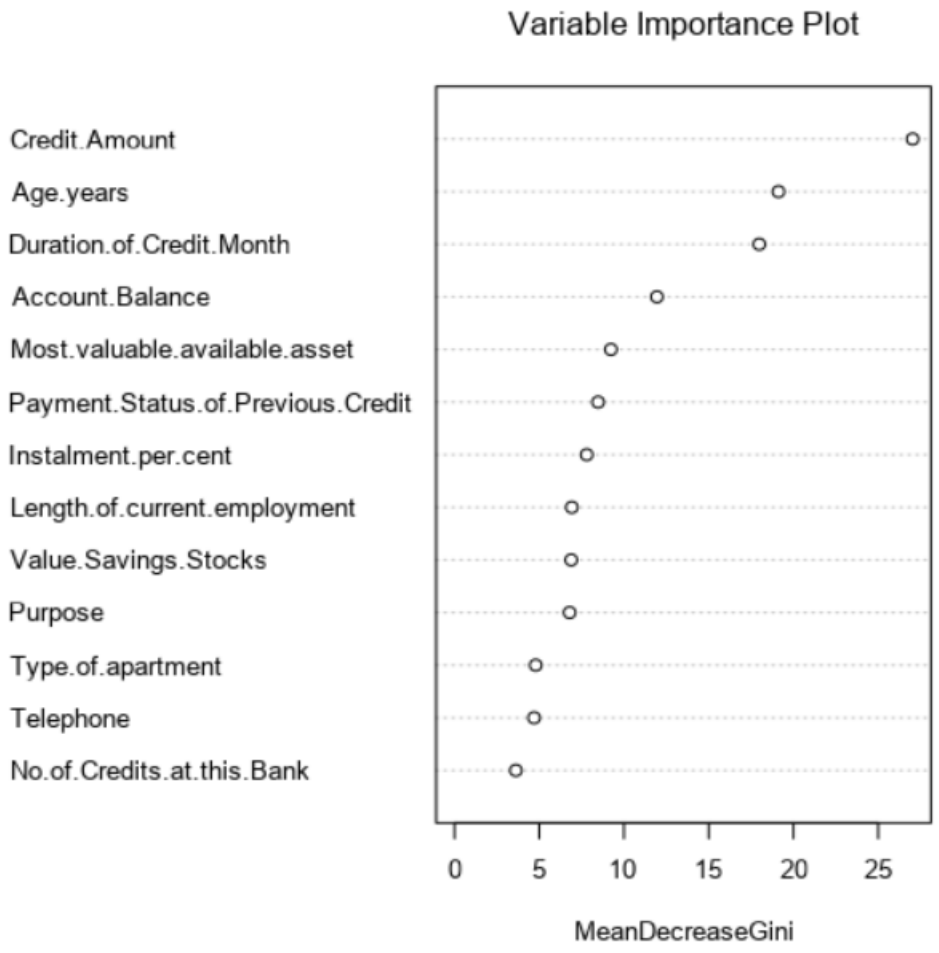
n= 350

- The significant variables are account balance, duration of credit, purpose.
- The overall accuracy percentage is 74.67 %.

Confusion matrix of Decision_Tree_19		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Random Forest:

- In the plot below we can clearly see that credit amount is the most important variable.

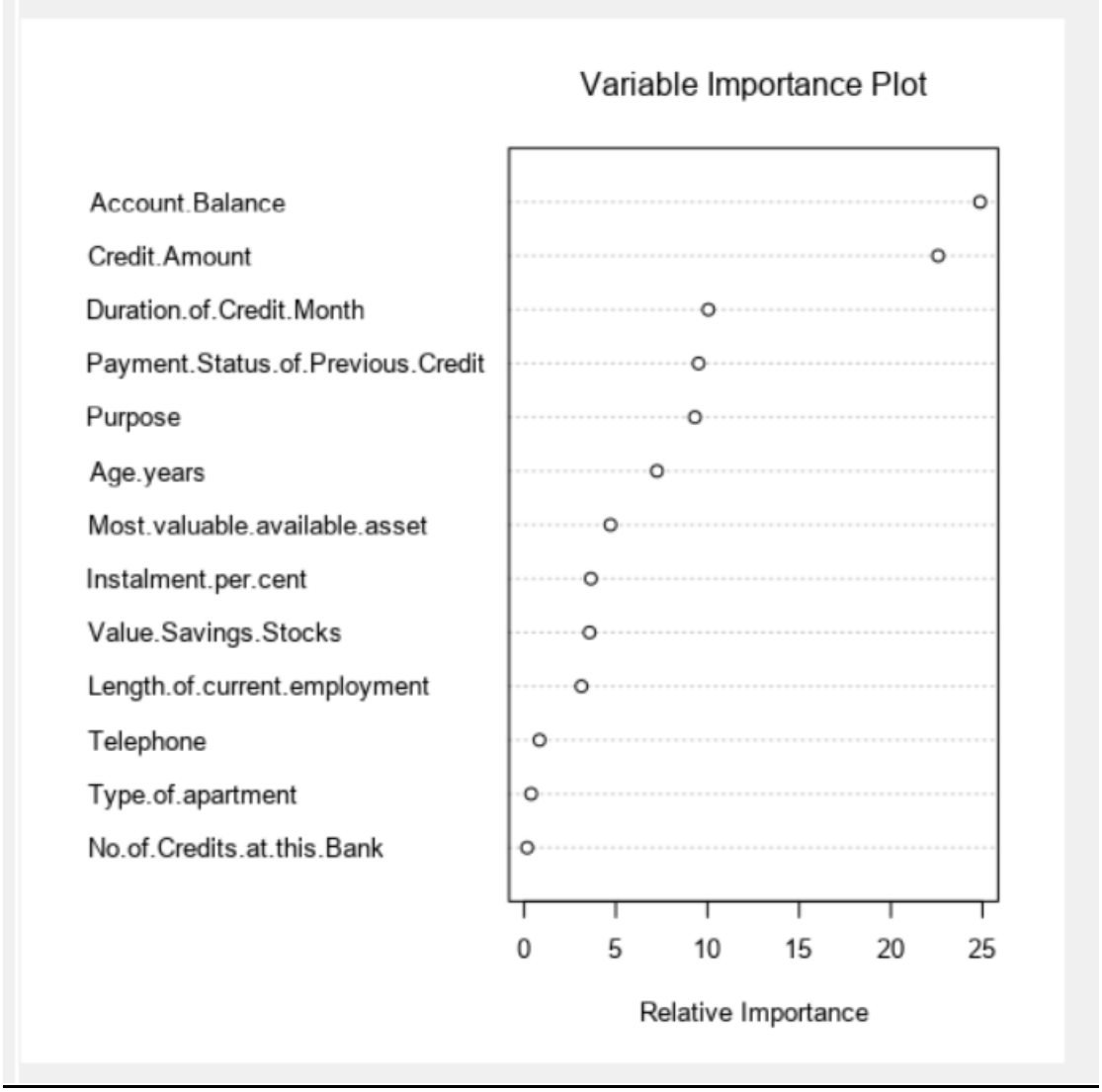


- The overall accuracy percentage is 81 %.

Confusion matrix of Forest_model_1		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	103	27
Predicted_Non-Creditworthy	2	18

Boosted Model:

- According to the plot the account balance is the most important variable.



- The overall accuracy percentage is 78 %.

Confusion matrix of bosted_model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	29
Predicted_Non-Creditworthy	4	16

Step 4: Writeup

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** 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

  The model chosen by me to predict is Random Forest model as it has the highest accuracy on validation data among the 4 models.
  It is having highest rate for accuracy credit worthy which is 98.10 % and accuracy non credit worthy of 40 %. So, comparing it with other models it is a good bargain.
  The area under the ROC graph in decision tree case is 0.735 which is a good score considering the model and it is better than all other models.

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

- How many individuals are creditworthy?
    - 410 individuals are creditworthy out of 500 for random forest.