Project: Creditworthiness

Step 1: Business and Data Understanding

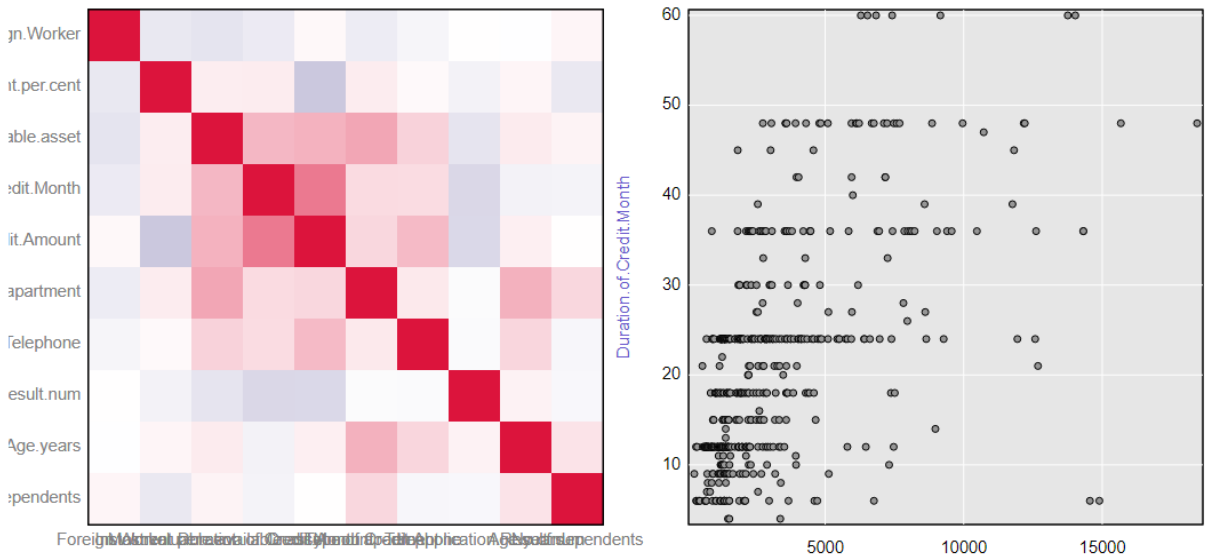
Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

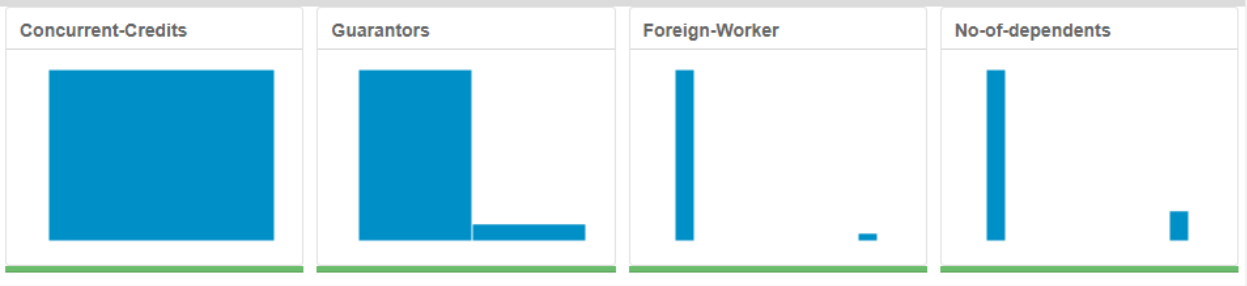
- We have an influx of new clients applying for a loan and we therefore, need to identify how many clients qualify for a loan (creditworthy) and how many clients do not qualify for loan (non-creditworthy).
- Data on all past applications is needed to inform those decisions on the list of customers that need to be processed in the next few days
- We need to use a Binary model since the expected outcome is whether the clients are creditworthy or non-creditworthy.

Step 2: Building the Training Set

- There are no numerical fields that highly-correlate with each other since the correlation for all these fields is less .70 as shown in the correlation matrix below



- The field *Duration-in-Current-address* contains 69% missing data and therefore this field is removed from the dataset. The field *Age-years* contains 2% missing data, and since removing this records will reduce our dataset significantly, we have decided to impute the missing values with the median.
- The fields *Concurrent-Credits, Occupation, Foreign-Worker, No-of-dependents, Telephone, and Guarantors* are removed from the dataset due to low variability. Shown below



Step 3: Train your Classification Models

1 Logistic Regression

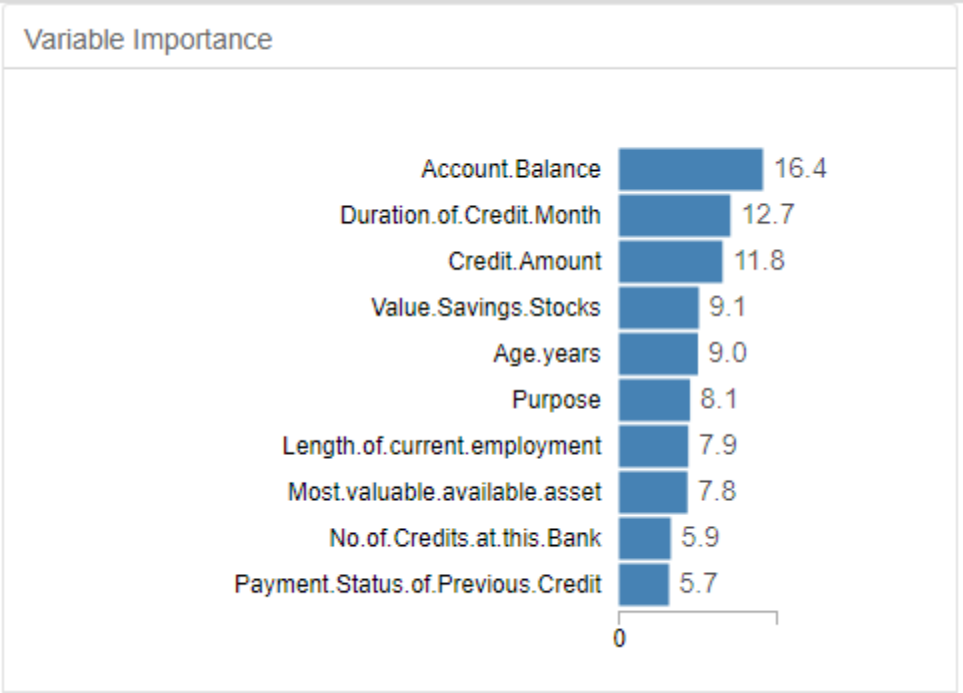
The following predictor variables are significant and their p-values are shown below:

➢ Account-Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Length-of-current-employment, and Instalment-per-cent

	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	.

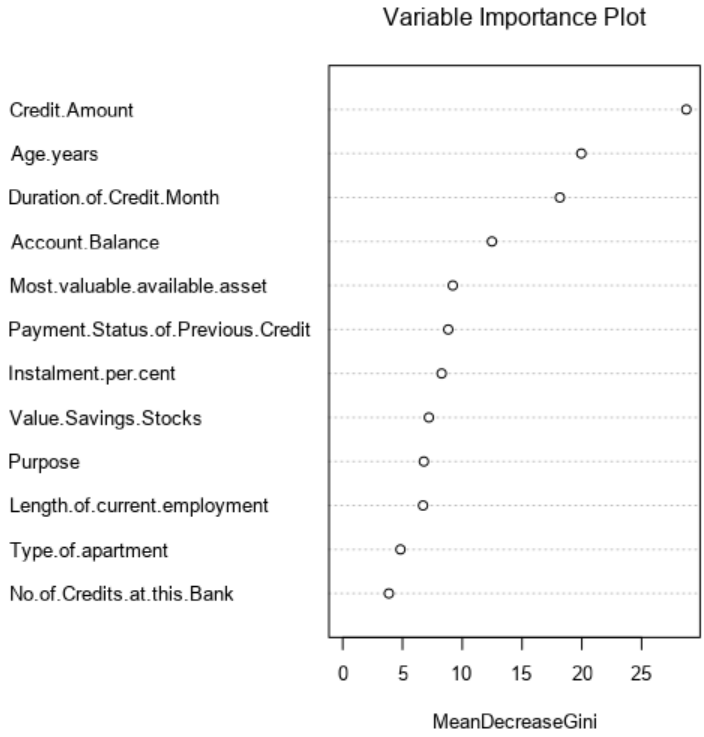
2 Decision Tree

The variable importance charts shown next depicts that the most important predictor variables for the Decision Tree model are *Account-Balance, Duration-of-Credit-Month and Credit-Amount*



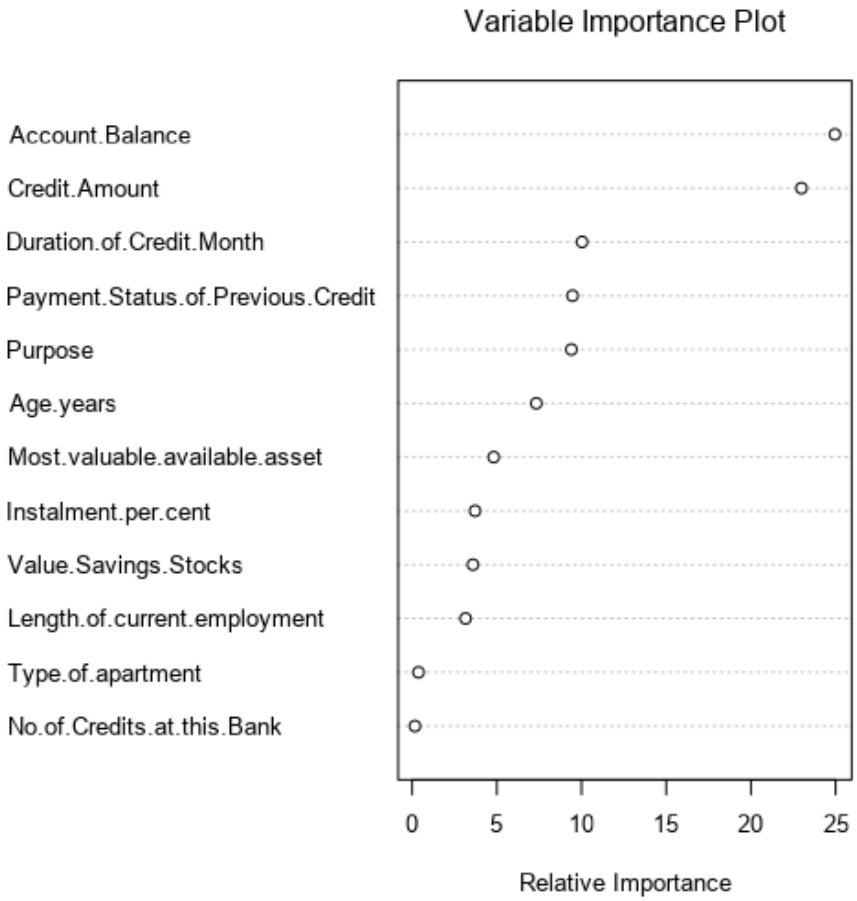
3 Forest Model

The variable importance plot shown below depicts that the predictor variables *Credit-Amount, Age-years, Duration-of-Credit-Month and Account-Balance,* are the most important predictor variables for the Forest Model



4 Boosted Model

The variable importance plot shown below depicts that the predictor variables *Account-Balance* and *Credit-Amount* are the most important predictor variables for the Boosted Model



After validating the models against the validation set, the overall model's accuracy for each model are as follows:

Model	**Overall Accuracy**
Logistic Regression	0.7600
Decision Tree	0.6733
Forest Model	0.7933
Boosted Model	0.7867

The confusion matrices are shown next. There is bias in the prediction of the Decision Tree and Linear Regression Models as it can be seen from the Overall accuracy as compared to the predicted accuracy for the Creditworthy individuals.

Confusion matrix of BoostedModel

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

Confusion matrix of DecisionTree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of stepwise_logistic

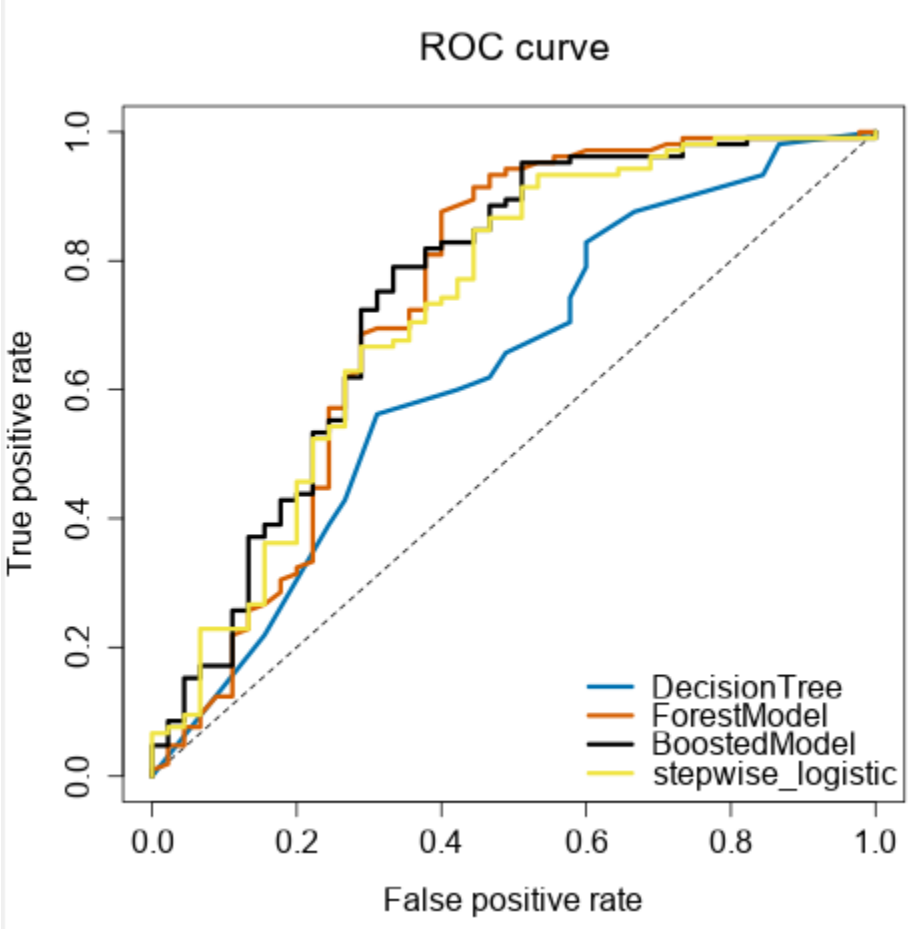
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Step 4: Writeup

I have decided to use the **Forest Model** since it has the highest overall accuracy of 79.33% outperforming the stepwise Logistic Regression (76%), Decision Tree (67.33%) and Boosted Models (78.67%). The Forest Model also has a better prediction accuracy (97.14%) in terms of predicting creditworthy segment, which is the highest compared to the Decision Tree, Logistic Regression and Boosted Models. However, Logistic Regression, with an accuracy of 48.89%, performs best in terms of predicting the Non-Creditworthy segments. The results are shown below:

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree	0.6733	0.7721	0.6296	0.7905	0.4000
ForestModel	0.7933	0.8681	0.7368	0.9714	0.3778
BoostedModel	0.7867	0.8632	0.7515	0.9619	0.3778
stepwise_logistic	0.7600	0.8364	0.7306	0.8762	0.4889

The decision to select the Forest Model is also supported by the ROC Curve shown next, which depicts that the Forest Model is slightly above the other models, suggesting that the Forest Model has a higher True Positive rate in comparison to that of the Decision Tree, Logistic Regression and Boosted Models.



Lastly, the Confusion Matrices shown next depicts that the Forest Model has the most actual creditworthy clients predicted as creditworthy. Based on the confusion matrix and the determination of the bias of the Forest model, it is depicted that the Forest Model is an unbiased model with regards to predicting the Creditworthy clients.

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

- **408 individuals are creditworthy.**

The workflow for the project is given below