Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?
  Predict creditworthiness of the new customers by using best performing model

- What data is needed to inform those decisions?
  Historic customer data having independent variables such as 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 and having data with categorisation of creditworthiness or not creditworthiness. So that model can be trained to predict the creditworthiness.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  The model is Binary, and we need to predict only whether the customer is creditworthy or not creditworthy.

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.

  In the clean-up process,
  - Duration in Current Address was removed because it has 69% missing data
  - While Age has only 2% missing data, median age is used for imputation. Median age is used instead of mean as the data is skewed to the left as shown below.
  - Concurrent Credits and Occupation is removed because it has only one value
  - Guarantors, Foreign Worker and No of Dependents are removed because of low variability.
  - Telephone field is removed due to its irrelevancy to the customer creditworthy.

    Please refer the below field summary for reference

Step 3: Train your Classification Models

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

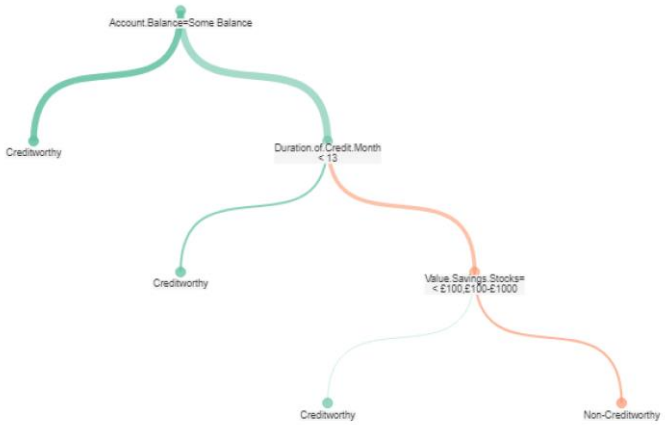
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

  After careful consideration,

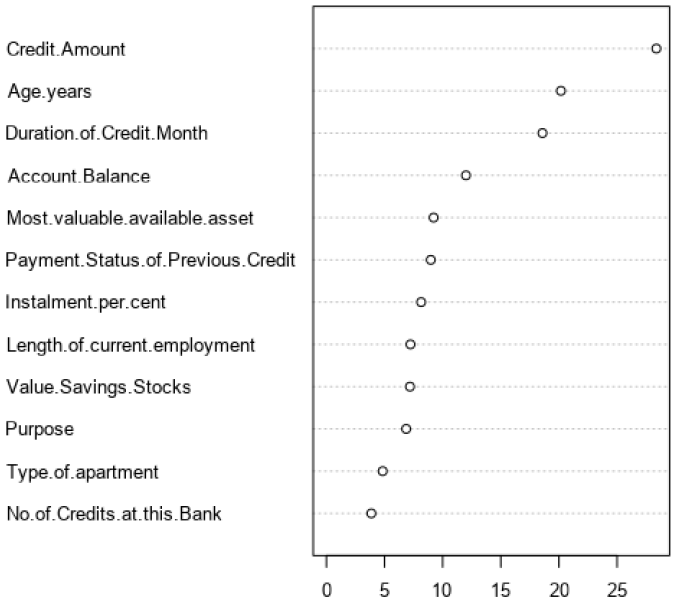
  - **Logistic Regression (Stepwise):** Account balance, payment-status of previous credit, purpose, credit amount, length of current employment, instalment-per-cent and most valuable available asset are the important variables considered for logistics stepwise model. Refer below for the report

	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	.

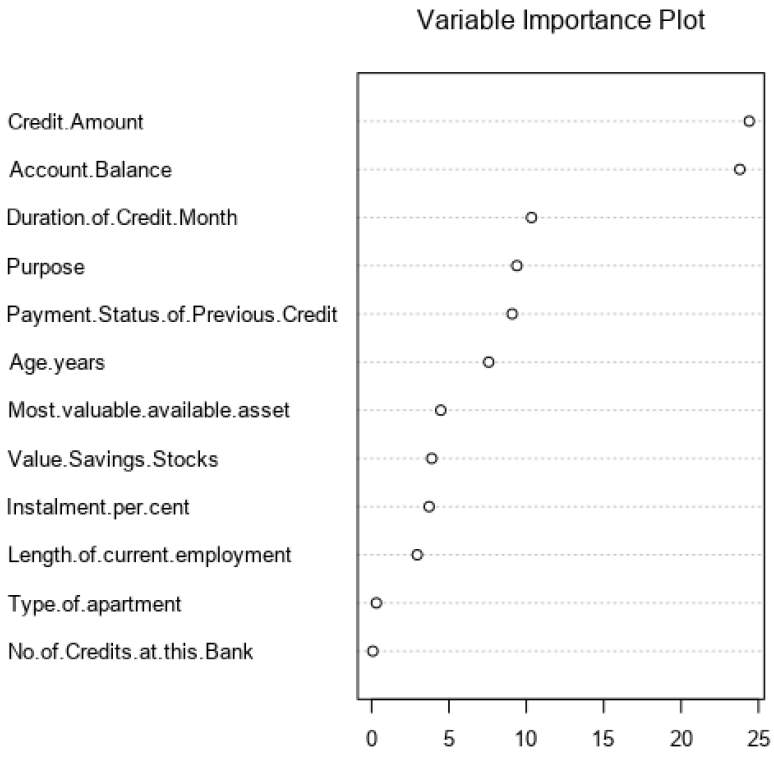
- **Decision Tree:** Account balance, duration-of-credit-month and value saving stocks are important variables for decision tree. Refer the below report



- **Random forest:** Credit amount, age-years, duration of credit month are top 3 important variables for random forest. Refer the below report



- **Boosted Model:** Credit amount, account balance are top 2 important variables for random forest. Refer the below report



- 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 Stepwise Model:** Logistics stepwise model has the 3rd highest accuracy and there seems to be a high biasness towards creditworthiness.
    - **Decision Tree Model:** Decision tree has the lowest accuracy with slight biasness towards creditworthiness.
    - **Forest Model:** Has the highest accuracy and no prominent bias between creditworthiness and not creditworthiness.
    - **Boosted Model:** Has the 2nd highest accuracy and no prominent bias between creditworthiness and not creditworthiness.

    After careful consideration Forest model has been selected because the highest overall accuracy and highest accuracy between classification no prominent bias between creditworthiness and not creditworthiness. Hence selecting the forest model.
    Refer the below report

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7361	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095
Logistic_Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

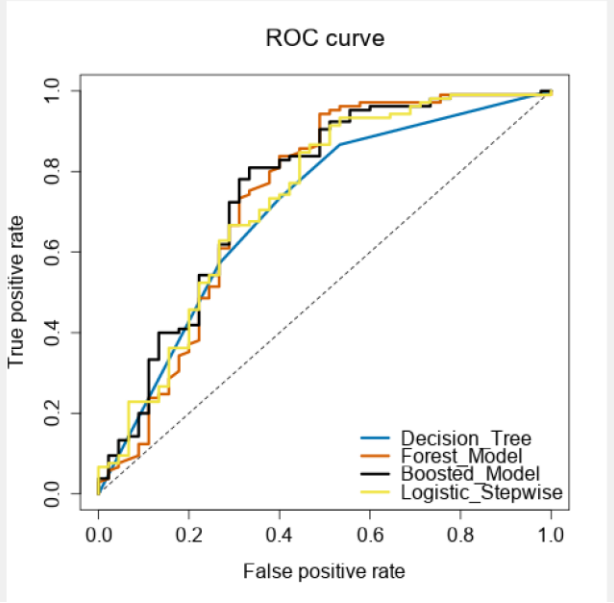
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Forest_Model

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

Confusion matrix of Logistic_Stepwise

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



Step 4: Writeup

- 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:

  After careful consideration forest model has the highest overall accuracy against the validation set, highest accuracies within “Creditworthy” and “Non-Creditworthy” segments and a higher ROC graph and less bias in the confusion matrices.
  Hence choosing forest model.

- How many individuals are creditworthy?
  406 new customers are creditworthy as per the model.