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?
  The main objective is to inform the manager of the bank if the new 200 clients deserve the loan or not.

- What data is needed to inform those decisions?
  Basically, two data sets; historical data about the old clients and their loan status if it was approved or not and their ages, their account balance, the purpose of the loan and some information which increase the model accuracy to get accurate prediction. Moreover, the same data for the new 200 clients is needed to use it in predict the probability of the creditworthy or not

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

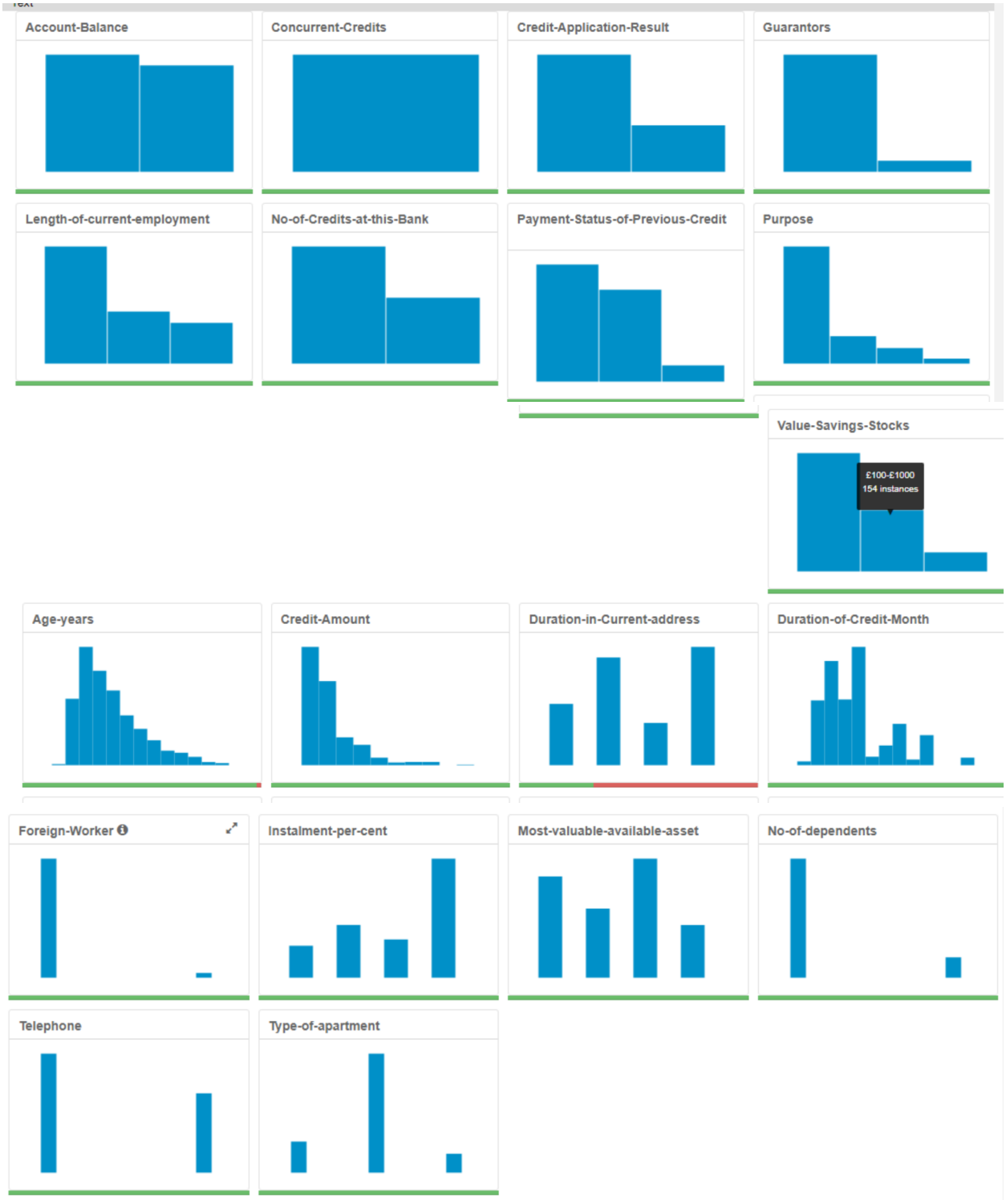
Binary analysis is the right model because the decision can take two options only which are give the client the loan or not.

Step 2: Building the Training Set

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.
  The fields that have been removed are:
    - Guarantors (Low variation)
    - Occupation (Low variation)
    - No of dependent (Low variation)
    - Telephone (No logical reason)
    - Foreign Worker (Low variation)
    - Concurrent Credit (Low variation)
    - Duration in Current (Too many missing data 69%)

On the other hand, the field that has been impute is age years because only 2% are missing so, it's better to impute it rather than delete it.



Step 3: Train your Classification Models

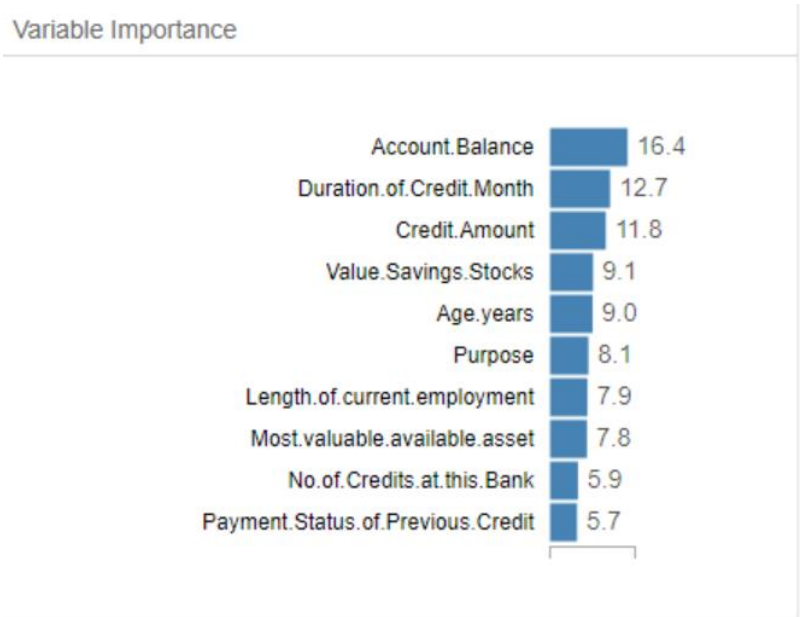
*Answer these questions for **each model** you created:*

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  The significant variables are different from each model. Therefore, I will show each one separately.

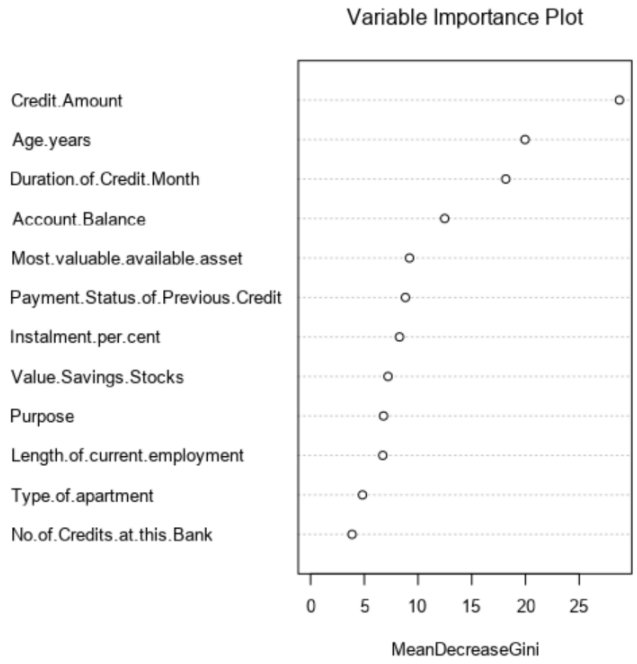
    - Logistic regression:
      Based on p-value the most significant variables are Account balance and Credit amount.

	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	.

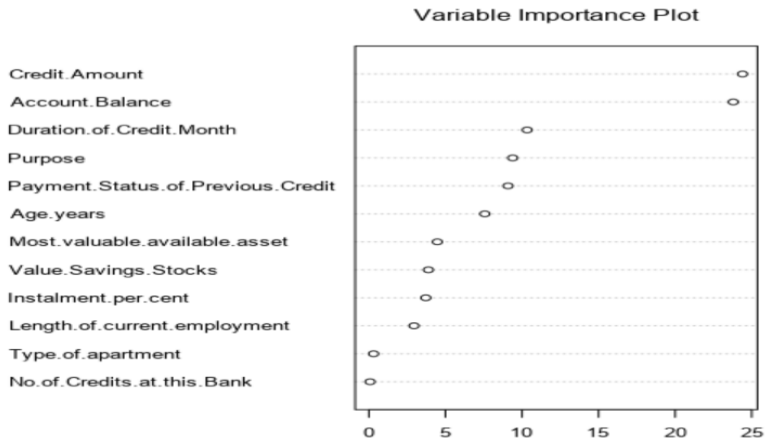
    - Decision tree:
      Based on variable importance chart Account balance and duration of credit month are the most significant variables.



- Forest Model:
  Based on variable importance chart Credit Amount and Age years are the most significant variables.



- Boosted Model:
  Based on variable importance chart Credit Amount and Account Balance are the most significant 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:

Confusion matrix of X

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

PPV= true positives \ (true positives + false positives) = 92/ (92+23) =.8
NPV= true negatives \ (true negatives + false negatives) = 22/ (22+13) = .63
There is a bias for creditworthy.

- Decision tree:

Confusion matrix of Decision_Tree_17

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

PPV= true positives \ (true positives + false positives) = 83/ (83+28) =.0.75
NPV= true negatives \ (true negatives + false negatives) = 22/ (22+17) = .56
There is a bias for creditworthy.

- Forest Model:

Confusion matrix of FM

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

PPV= true positives \ (true positives + false positives) = 102/ (102+28) =.0.78
NPV= true negatives \ (true negatives + false negatives) = 17/ (3+17) = 0.85
There is no bias in the model.

- Boosted Model:

Confusion matrix of BM

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

PPV= true positives \ (true positives + false positives) = 101/ (101+28) =.0.78
NPV= true negatives \ (true negatives + false negatives) = 17/ (4+17) = 0.81
There is no bias in the model.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM	0.7933	0.8681	0.7368	0.9714	0.3778
Decision_Tree_17	0.6667	0.7685	0.6272	0.7905	0.3778
X	0.7600	0.8364	0.7306	0.8762	0.4889
BM	0.7867	0.8632	0.7524	0.9619	0.3778

The forest model is the highest accuracy.

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"

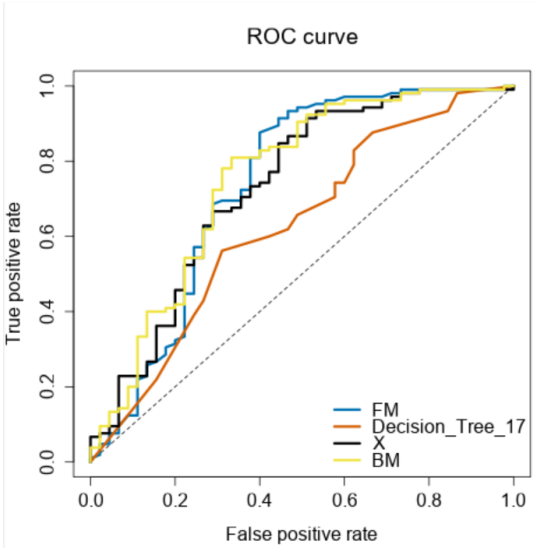
Answer these questions:

- 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

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM	0.7933	0.8681	0.7368	0.9714	0.3778
Decision_Tree_17	0.6667	0.7685	0.6272	0.7905	0.3778
X	0.7600	0.8364	0.7306	0.8762	0.4889
BM	0.7867	0.8632	0.7524	0.9619	0.3778

Because forest model is the highest accuracy and the manager concerned in the accuracy, I will choose it. Moreover, the individual accuracy for "Creditworthy" is the highest in Forest model and "Non-Creditworthy" is the highest but it equals Decision tree and Boosted model. Although, the overall accuracy superiors Forest Model.

    - ROC graph



According to ROC curve Forest Model has the highest curve, which is good indicator to use it.

- Bias in the Confusion Matrices

Confusion matrix of BM

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

Confusion matrix of Decision_Tree_17

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Confusion matrix of FM

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

Confusion matrix of X

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

As mentioned above there is no bias in the chosen model which is Forest model.

- How many individuals are creditworthy?

408 clients are creditworthy.