Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

- What decisions needs to be made?
  Provide a list of how many of 500 new customers would be qualify for a loan.

- What data is needed to inform those decisions?
  1. Account Balance
  2. Duration of Credit Month
  3. Purpose
  4. Payment Status of Previous Credit
  5. Credit Amount
  6. Instalment Per Cent
  7. Age Years
  8. Type of Apartment
  9. Length of Current Employment
  10. No of Credits at This Bank
  11. Value Savings Stocks
  12. Most Valuable Available Asset

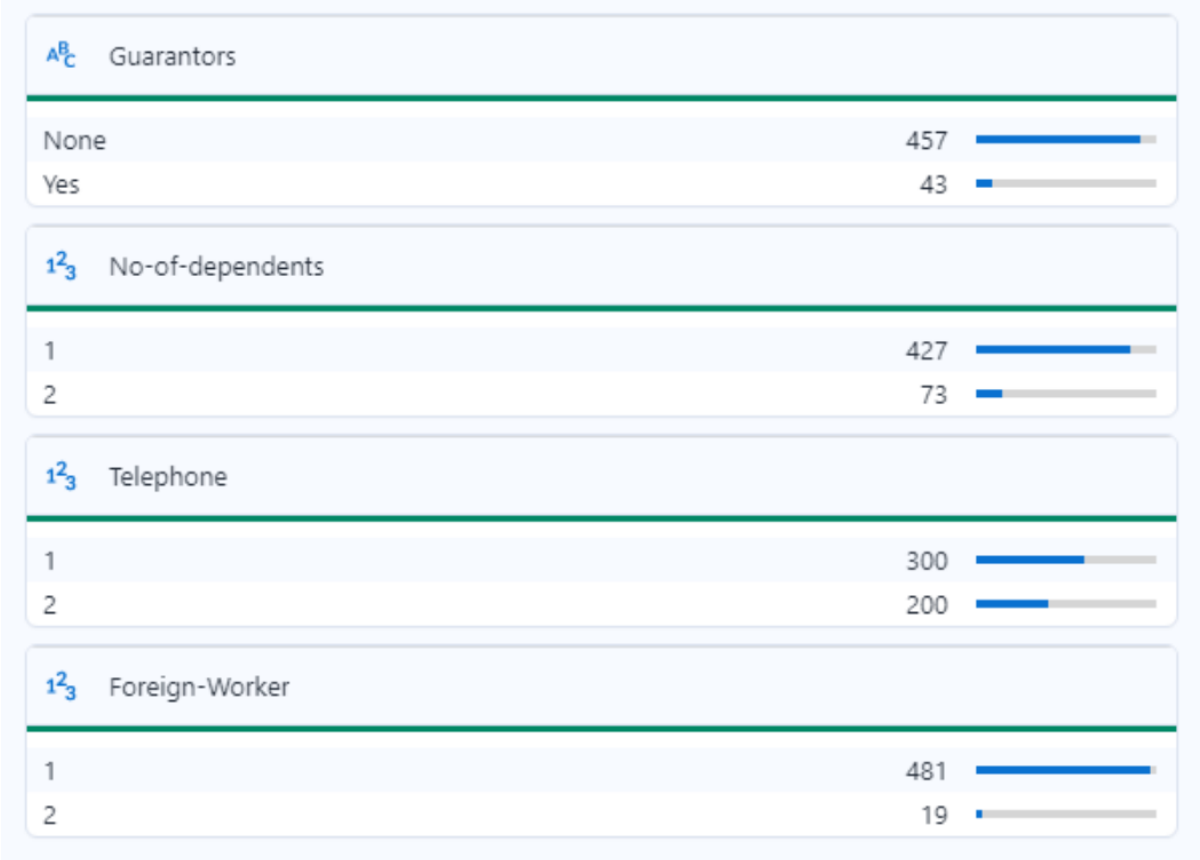
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  The model is Binary, because the target value contains only two value (creditworthy or (non-creditworthy).

Step 2: Building the Training Set

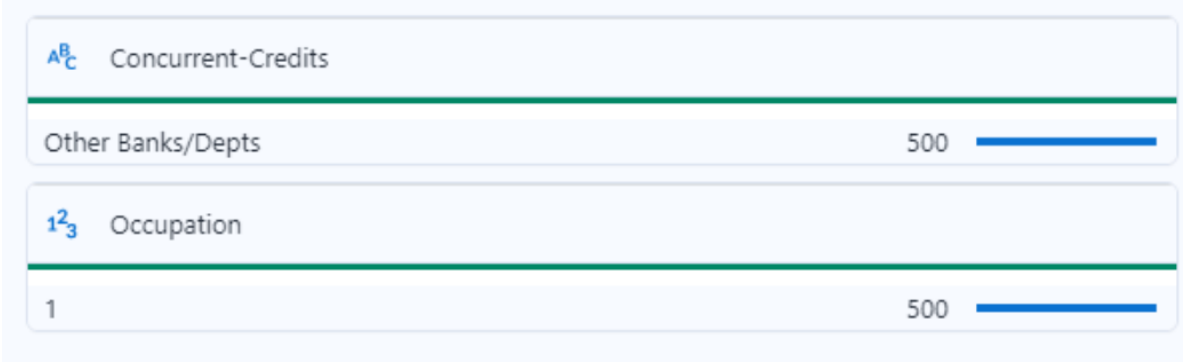
Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

1. Remove fields:

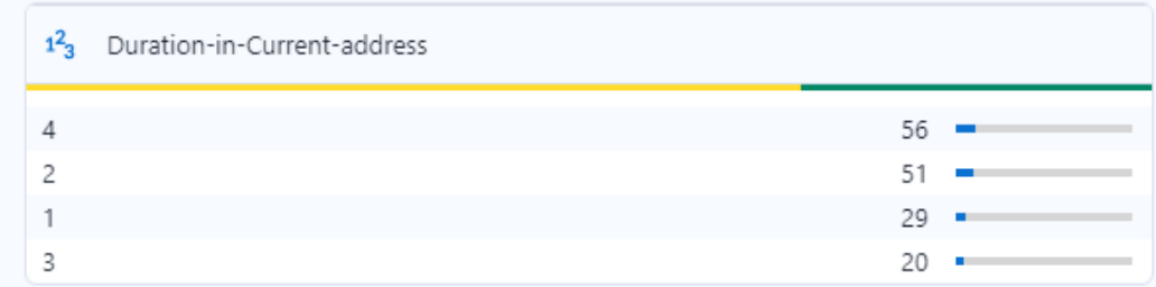
1. **Foreign Worker**
   It's low variability, showing 481 records are "1"and 19 records are "2" and only two unique variables.
2. **No-of-dependents**
   Also, It's low variability, showing 427 records are "1"and 73 records are "2" and only unique two variables.
3. **Guarantors**
   As those two above, It's low variability, showing 457 records are "none" and 53 records are "yes" and only two unique variables.
4. **Telephone**
   It has only two unique value, and dose not useful for the model.



5. **Concurrent Credits**
   Has only one value for 500 records.
6. **Occupation**
   Has only one value for 500 records.



7. **Duration-in-Current-address**
   343 of 500 records are messing (68.6%).



2. Imputed

1. **Age years**
   has 13 records empties (2.4%) since the variable is numeric and it's not that much affect on the dataset if we replace every missing value with the median "33".
   The average age before imputed is = 35.63 years.
   The average age after imputed is = 35.57 years.

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

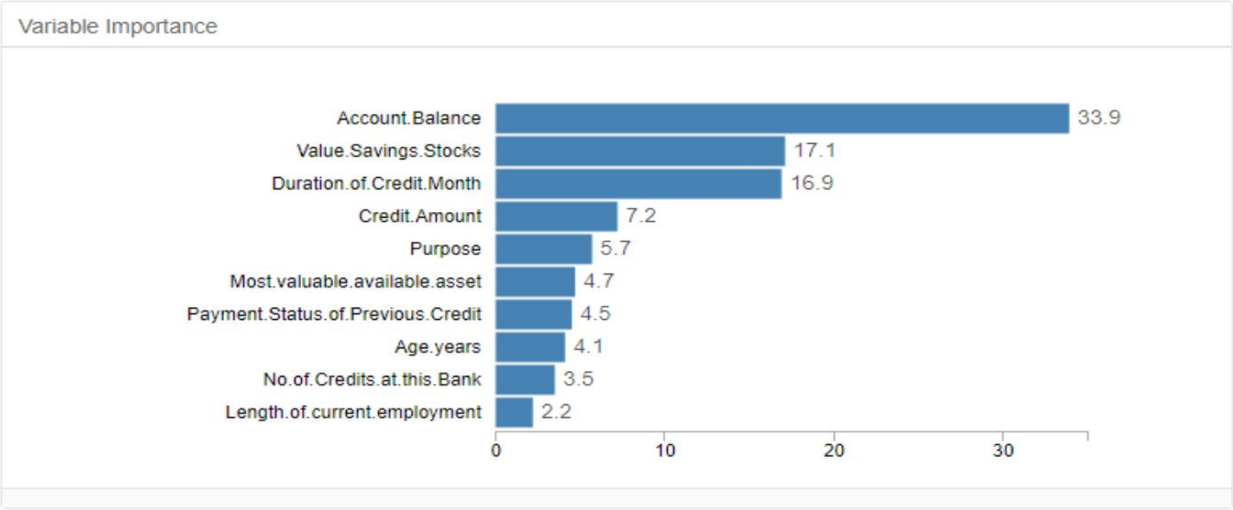
*Answer these questions for **each model** you created:*

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.
    1. **Logistic regression + stepwise**
        - Account Balance (Some Balance), P-value = 1.65e-07
        - Purpose (New car), P-value = 5.66e-03
        - Credit Amount, P-value = 2.96e-03

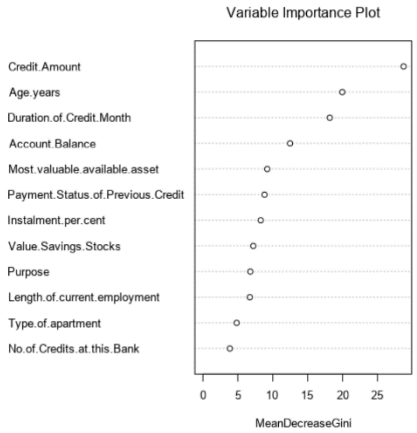
Coefficients:

	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	.

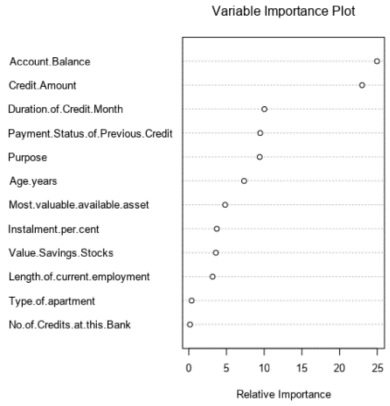
    2. **Decision Tree**
        - Account Balance
        - Value Savings Stocks
        - Duration of Credit Month



3. **Forest Model**
    - Credit Amount
    - Age years
    - Duration of Credit Month



4. **Boosted Model**
    - Account Balance
    - Credit Amount



- 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?

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model	0.7867	0.8632	0.7515	0.9619	0.3778
Logistic_regression___stepwise_tool	0.7600	0.8364	0.7306	0.8762	0.4889

1. **Logistic regression + stepwise**
   Accuracy = 76%, Accuracy Creditworthy = 87.6%,
   Accuracy Non-Creditworthy = 48.8%

Confusion matrix of Logistic_regression___stepwise_tool

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

2. **Decision Tree**
   Accuracy = 74.7%, Accuracy Creditworthy = 88.6%,
   Accuracy Non-Creditworthy = 43.2%

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

3. **Forest Model**
   Accuracy = 79.3%, Accuracy Creditworthy = 97.2%,
   Accuracy Non-Creditworthy = 37.8%

Confusion matrix of Forest_Model

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

4. **Boosted Model**
   Accuracy = 78.7%, Accuracy Creditworthy = 96.2%,
   Accuracy Non-Creditworthy = 37.8%

Confusion matrix of Boosted_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

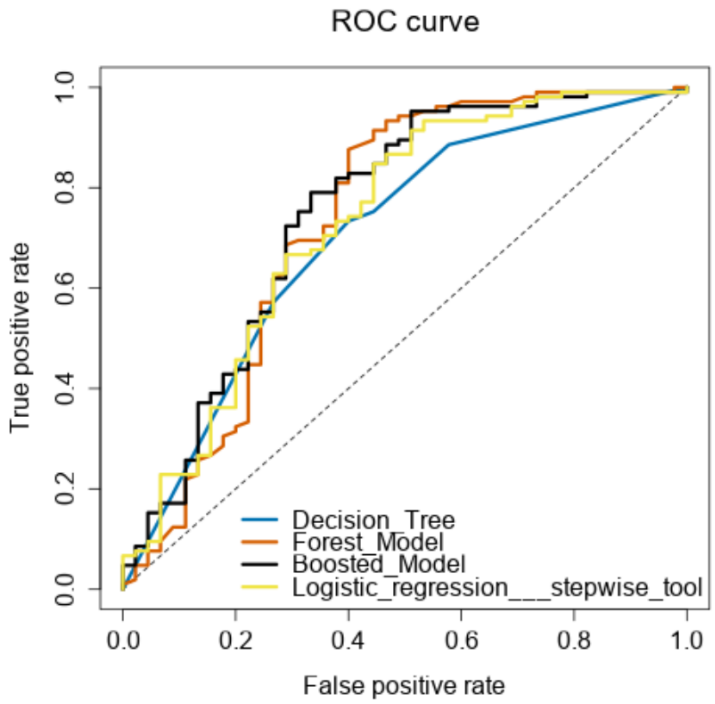
Step 4: Writeup

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:
  1. Overall Accuracy against your Validation set
     The Forest Model is high accurate against another model by 0.6% more than Boosted Model.

  2. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
     The Forest Model is a highest accurate in Creditworthy 0.9714, but not a highest in Non-Creditworthy it takes a 3rd place.

  3. ROC graph



4. Bias in the Confusion Matrices

There are some issues Non-Creditworthy, the model predicted 28 records as creditworthy that ware actually Non-Creditworthy.

The problem is the model may decide the customer is qualified for the loan when he is not.

- How many individuals are creditworthy?

  408 requests are qualified for loan.

- **Alteryx workflow:**