

Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: <https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project>

Done By: Manar Alharbi

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

Q1) What decisions needs to be made?

Decide which of the new customers are creditworthy to be given a loan

Q2) What data is needed to inform those decisions?

In order to make this decision, we need to know if a customer is creditworthy or not. To determine this, there are a variety of factors which could be considered. Some of the things which could influence our decision are their current length of employment, income, credit score, if the customer carries a credit balance from month to month, age, and their current savings.

Q3) What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

- The **Binary Classification models** will be used.
- The output is either the customer is creditworthy or the customer is not 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**.*

Here are some guidelines to help guide your data cleanup:

- For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Note: For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)

Note: For students using software other than Alteryx, please format each variable as:

Variable	Data Type
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

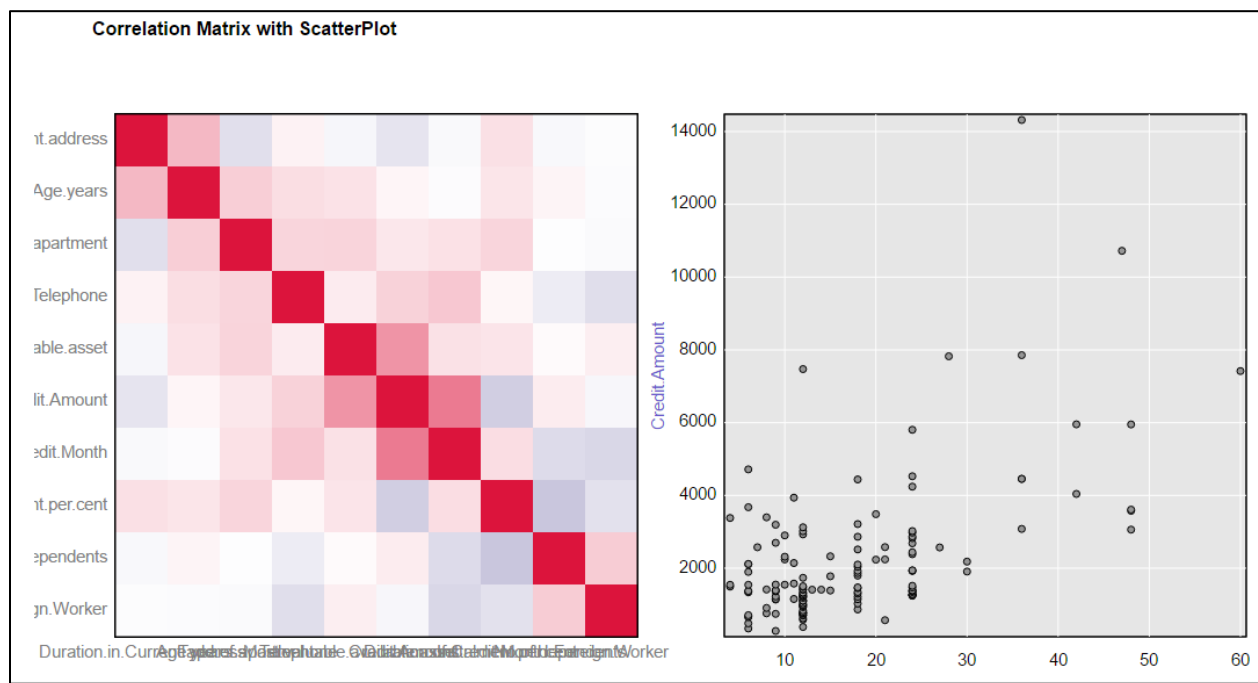
Answer this question:

Q1) In your cleanup process, which fields did you remove or impute?
Please justify why you removed or imputed these fields. Visualizations are encouraged.

- I removed 6 fields: Duration-in-Current-address, Concurrent-Credits, Occupation, Guarantors, Foreign-Worker, No-of-dependents and Telephone.
- I imputed 1 field: Age-years.

a) Highly related variables

- The highest correlation found was 0.57 (which's not considered "high".)
- Hence, There's no highly-correlate variables, as shown in below correlation matrix



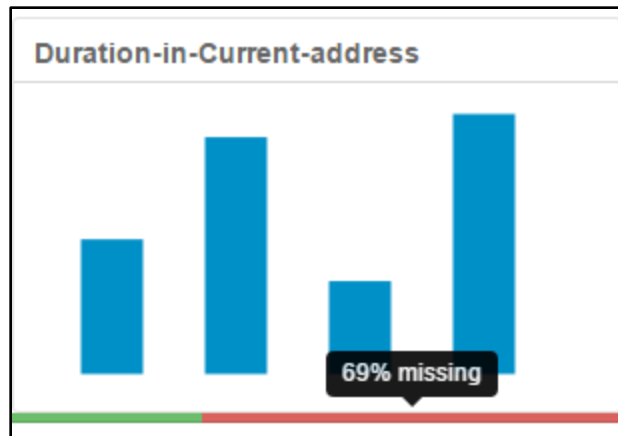
b) Variables with Missing data



- The fields with null values were **Duration-in-Current-address** and **Age-years**

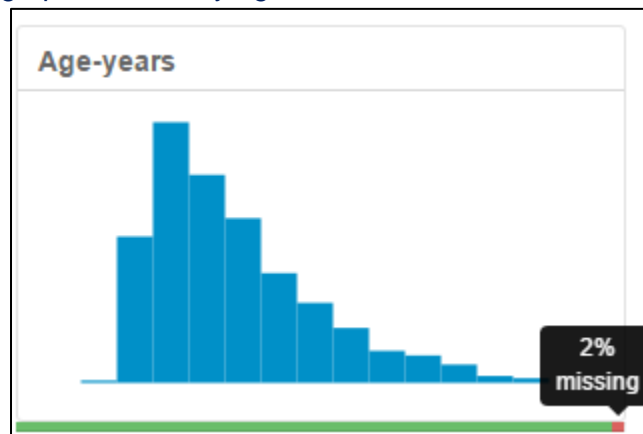
1. Duration-in-Current-address

- The **Duration-in-Current-address** field was removed, since it has over %69 of missing data.



2. Age-years

- The **Age-years** field has only %2 of missing data., so it was imputed by the median of the entire data field (33). I chose to impute the missing data of age-years field using the median, because the graph is a heavily right-skewed distribution.

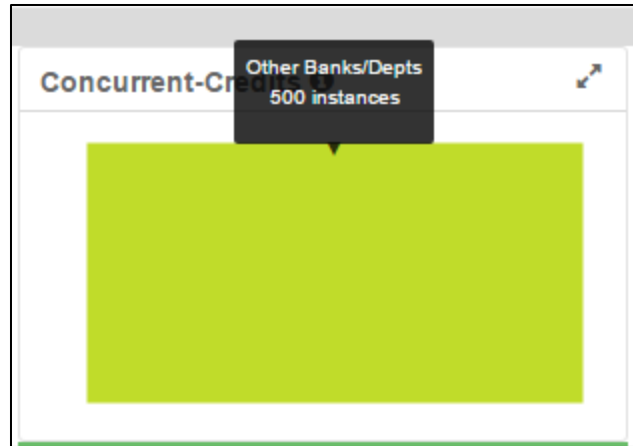


c) Variables with low variability

- The fields with low variability were **Concurrent-Credits, Occupation, Guarantors, Foreign-Worker, and No-of-dependents.**

1. Concurrent-Credits

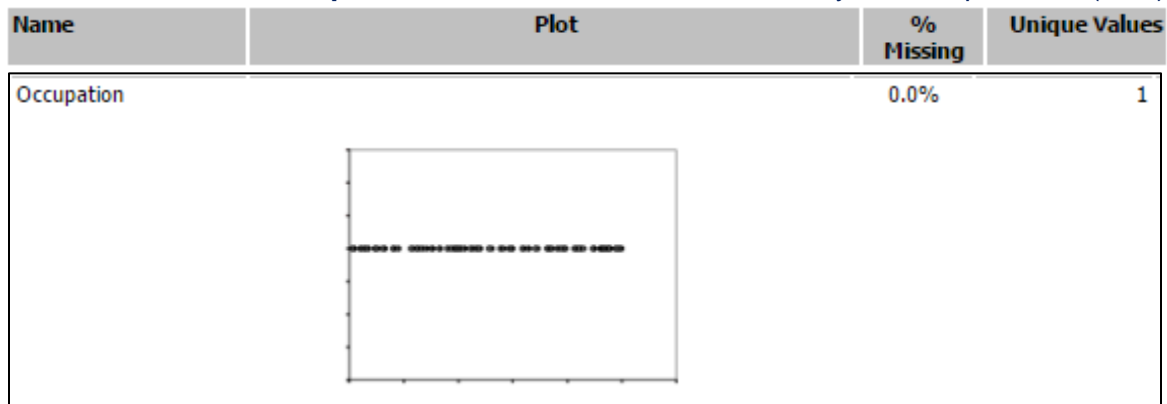
- The **Concurrent-Credits** field was removed, as it has only one unique value (Other Banks/Depts)



Name	% Missing	Unique Values	Shortest Value	Longest Value
Account-Balance	0.0%	2	No Account	Some Balance
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts
Credit-Application-Result	0.0%	2	Creditworthy	Non-Creditworthy

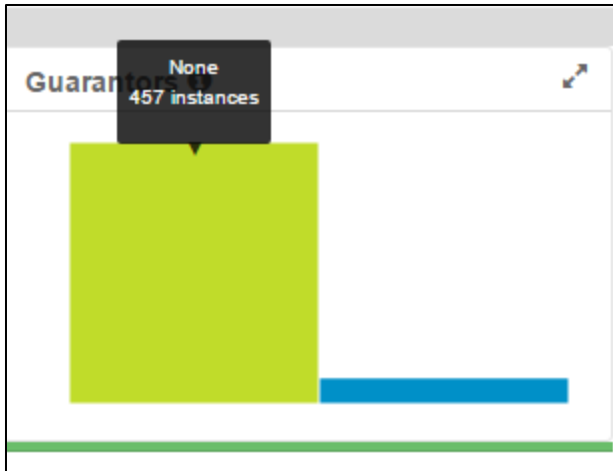
2. Occupation

- The **Occupation** field was removed, as it has only one unique value (1.00)



3. Guarantors

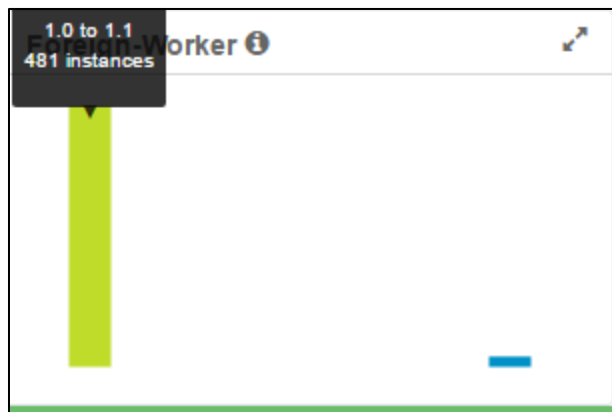
- The **Guarantors** field was removed, as it has two unique value and it heavily skews towards "None".

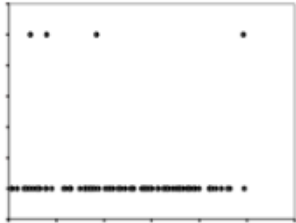


Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count
Account-Balance	0.0%	2	No Account	Some Balance	238	262
Concurrent-Credits	0.0%	1	Other Banks/Depts	Other Banks/Depts	500	500
Credit-Application-Result	0.0%	2	Creditworthy	Non-Creditworthy	142	358
Guarantors	0.0%	2	Yes	None	43	457

4. Foreign-Worker

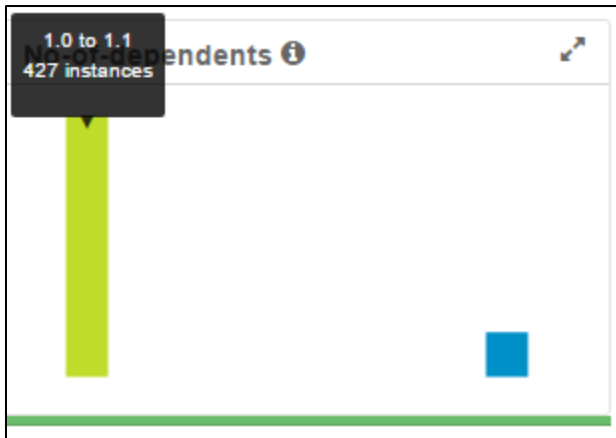
- The **Foreign-Worker** field was removed, as it has two unique value and it heavily skews towards "1".



Name	Plot	% Missing	Unique Values
Foreign-Worker		0.0%	2

5. No-of-dependents

- The **No-of-dependents** field was removed, as it has two unique value and it heavily skews towards "1".



d) Variables not related to prediction output

Telephone

The **Telephone** field was removed, since it has no logical connection to the outcome we want to predict.

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.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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.

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?

You should have four sets of questions answered. (500 word limit)

Q1) Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

1. Logistic Regression

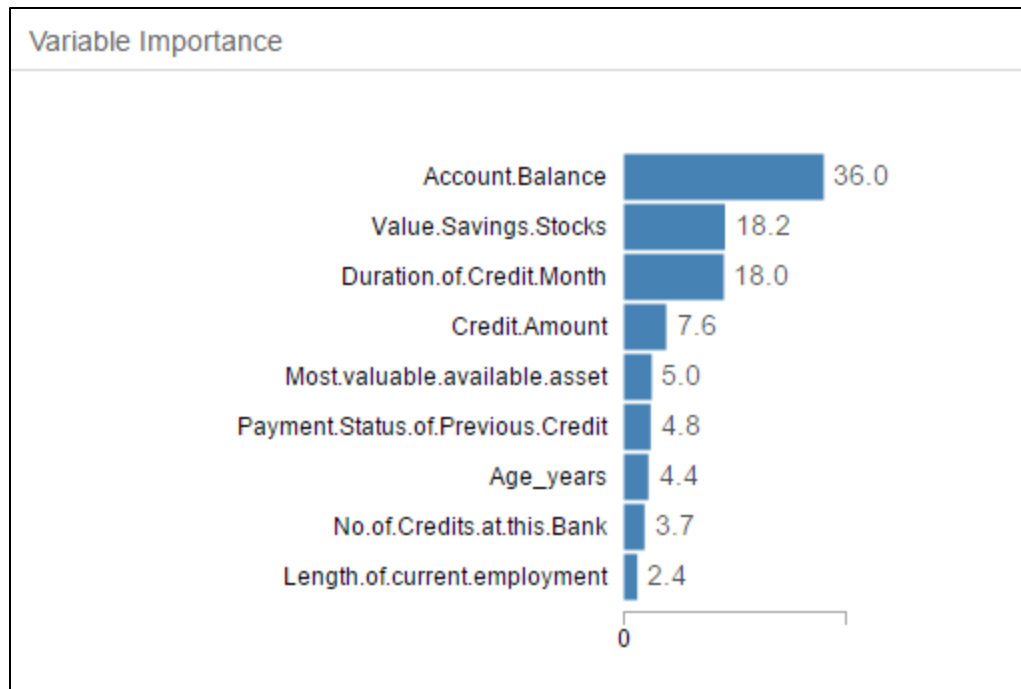
- **After running the Stepwise Logistic Regression, The chosen predictor variables were:**

- Account.BalanceSome Balance
- Payment.Status.of.Previous.Credit
- Purpose
- Credit.Amount
- Length.of.current.employment
- Instalment.per.cent
- Most.valuable.available.asset

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	.

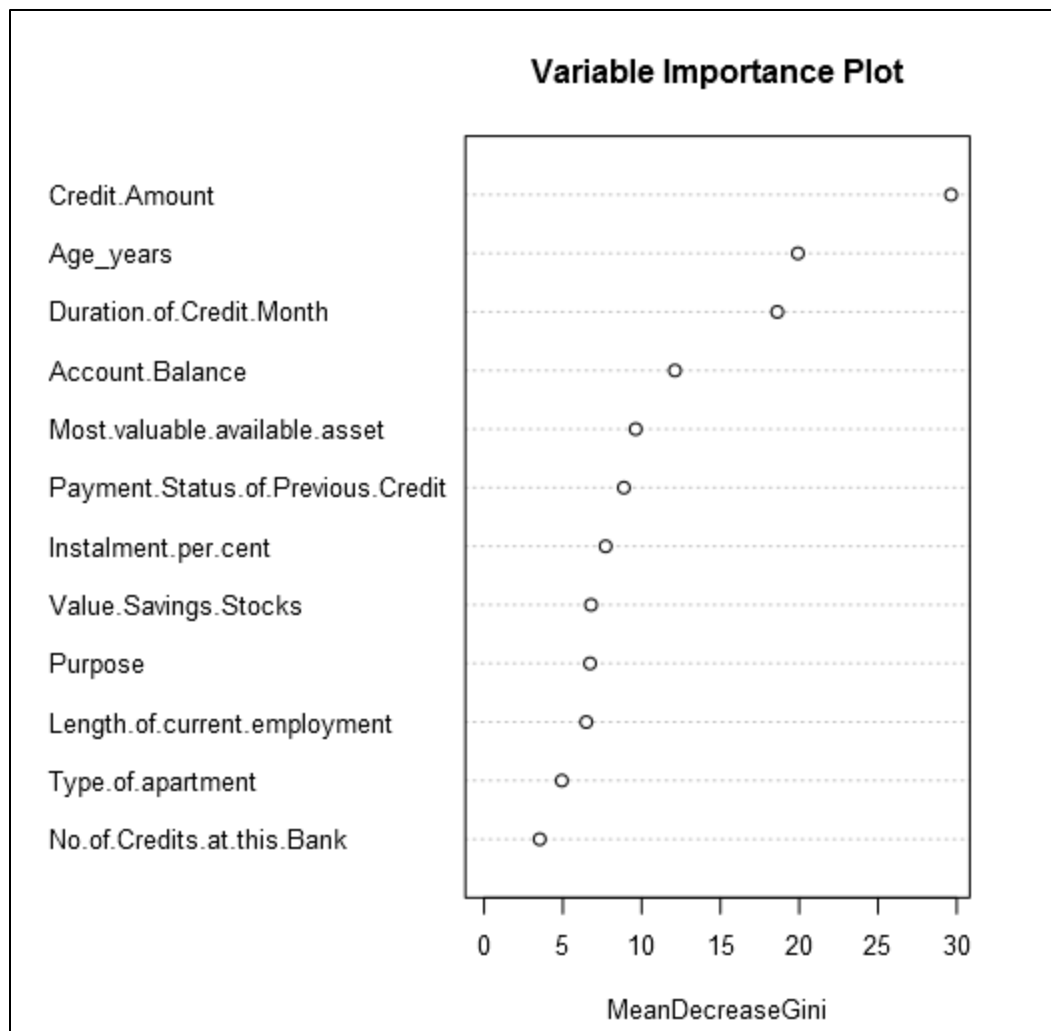
2. Decision Tree

- **Variable importance chart:**



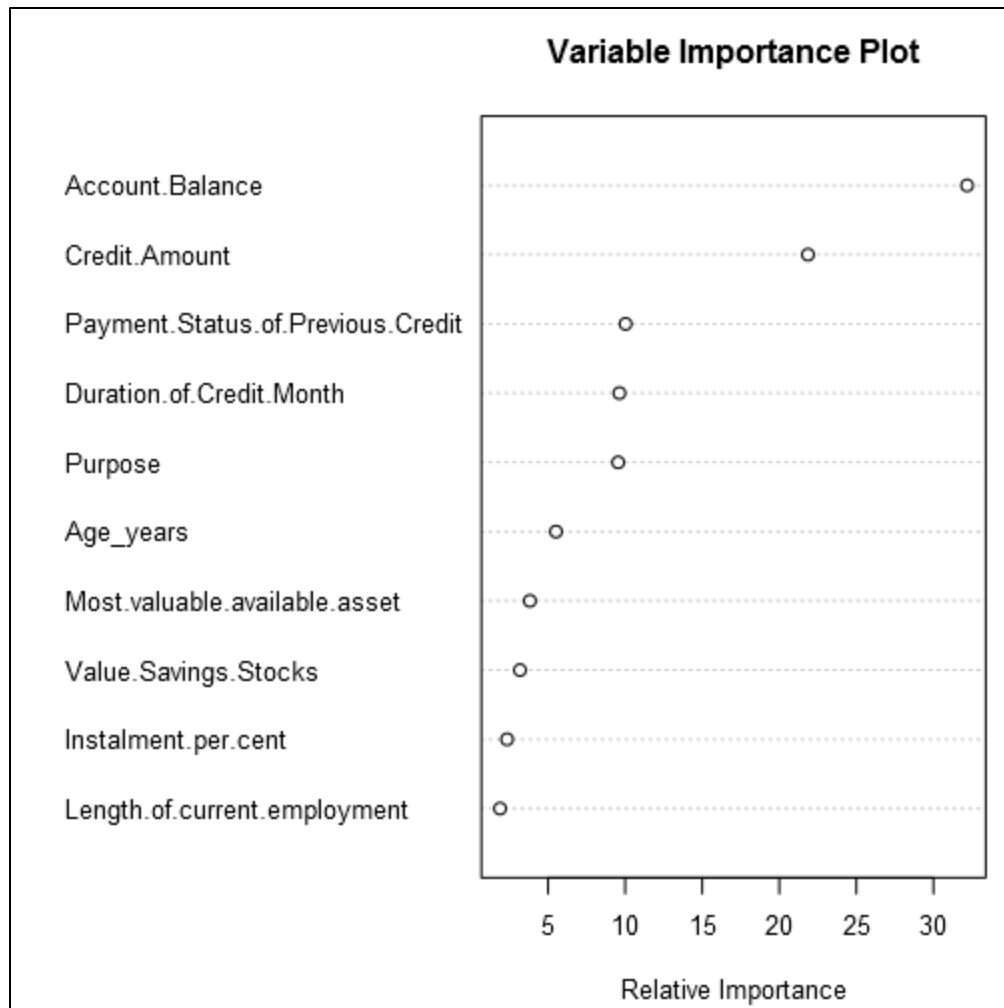
3. Forest Model

- **The variable Importance Plot:**



4. Boosted Model

- **The variable Importance Plot:**



Q2) 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?

a. Overall percent accuracy

- **The overall percent accuracy of the four models:**

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model_Creditworthy	0.8067	0.8755	0.7392	0.7969	0.8636
Decision_Tree_Creditworthy	0.7467	0.8273	0.7054	0.7913	0.6000
Boosted_Model_Creditworthy	0.7867	0.8632	0.7524	0.7829	0.8095
Stepwise_Logistic_Regression_Creditworthy	0.7600	0.8364	0.7306	0.8000	0.6286

b. confusion matrix

- The confusion matrix of the four models:

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

Confusion matrix of Decision_Tree_Creditworthy		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of Forest_Model_Creditworthy		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Confusion matrix of Stepwise_Logistic_Regression_Creditworthy		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

c. Bias

- The decision tree and logistic regression models have a bias towards correctly predicting creditworthy.
- While, the forest and boosted model have little bias towards correctly predicting Non-creditworthy.

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"

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:

Q1) 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:

The Forest Model was chosen.

a. Overall Accuracy against your Validation set

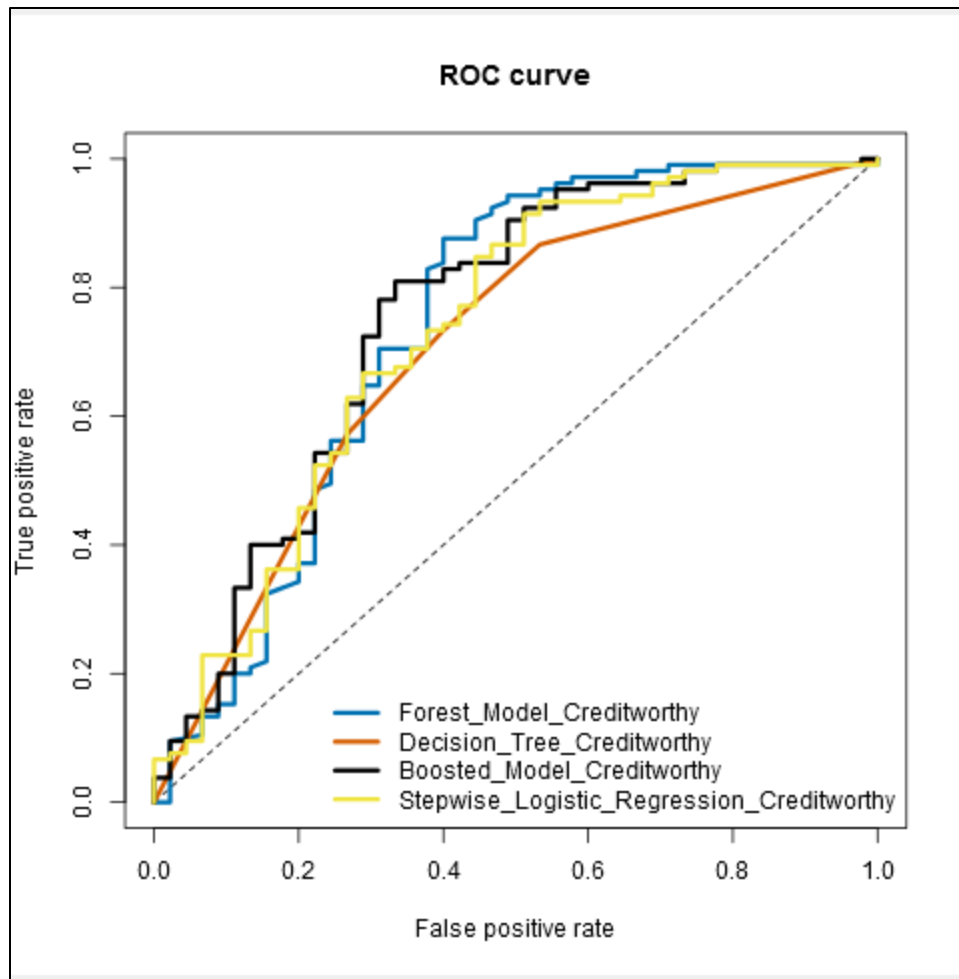
The Forest Model has the highest accuracy with 0.8067.

b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments

- The Accuracy_Creditworthy is 0.7969
- The Accuracy_Non-Creditworthy is 0.8636

c. ROC graph

The AUC of forest model was 0.7392, which represents how well a model can distinguish between the two customers segments (Creditworthy/ Non-Creditworthy).



d. Bias in the Confusion Matrices

The forest model have little bias towards correctly predicting Non-creditworthy.

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

Q2) How many individuals are creditworthy?

There are 408 of customers are creditworthy

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](#) here. Reviewers will use this rubric to grade your project.