

## 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>

### 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*

1. What decisions needs to be made?

The decision is whether new customers qualify for loan approvals. Once this decision is made, the bank can then extend loans to these qualified customers. The bank is looking to come up with an better and automatic way to manage the huge increase in credit applications. The short term decision is dealing with the 500 applications. The long term decision is which model to use to automate the process and deal with increased business on a consistent basis.

2. What data is needed to inform those decisions?

The data needed for this decision comprises non-financial information about the applicant such as age, occupation, and dependents, and financial information about the applicant such as assets and stocks. Additional relevant information on all past applications will also be used to build the model.

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

A customer can either be approved or not approved. This is a classification problem in which a customer can fall into one of two classes. Therefore, we need to use a binary classification model to make this decision.

### 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 average 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

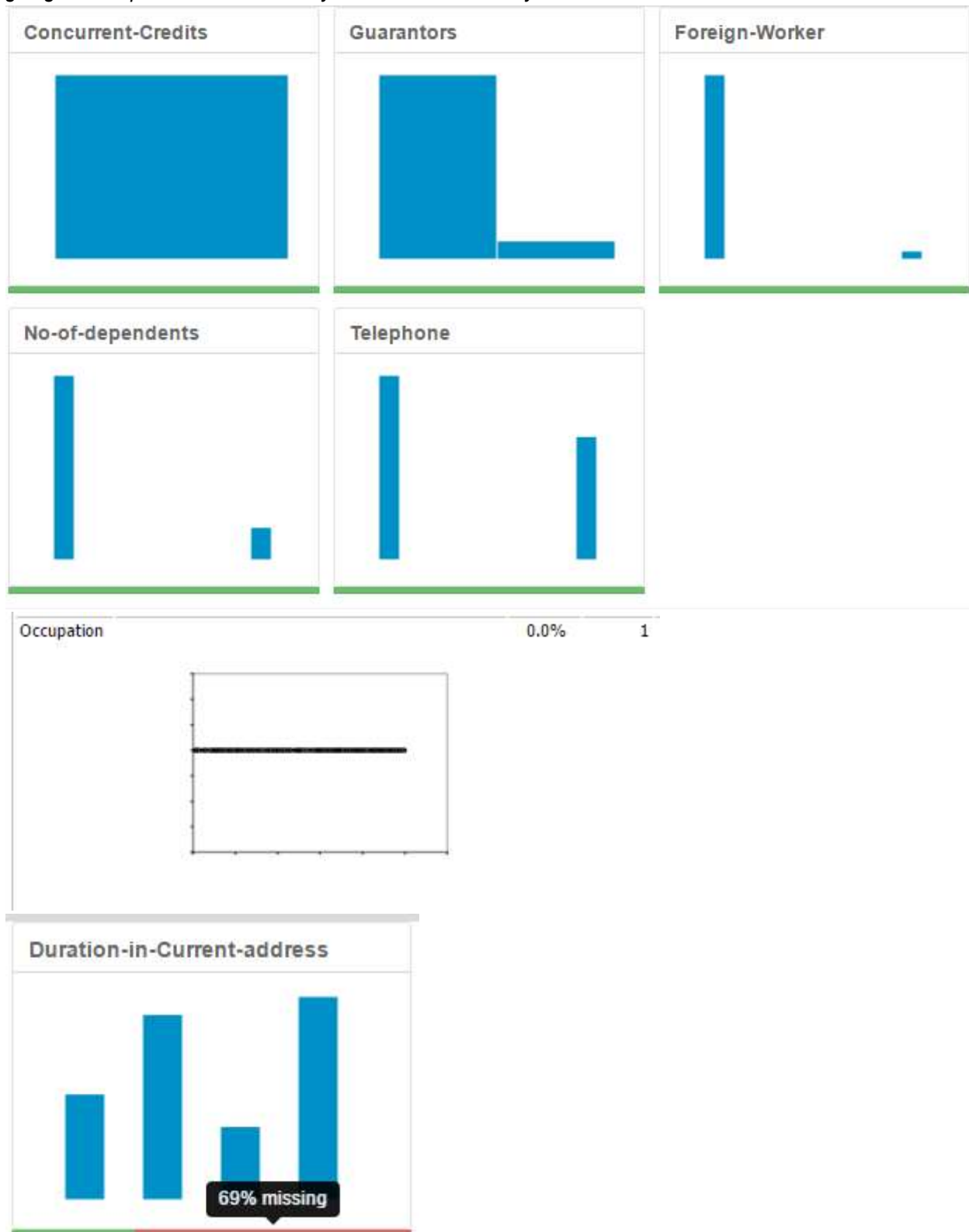
*To achieve consistent results reviewers expect.*

*Answer this question:*

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

*After I reviewed the variability of each field using the field summary tool, I decided to remove the Concurrent Credits, Foreign worker, No of dependents, Guarantors, and Occupation fields due to their low*

variability, as depicted below. Although the telephone field also displays low variability, it is probably not going to be a predictive factor in my model and this is why I will remove it.



I decided to remove Duration in Current address because this field is missing 68.8% of its values.

*Finally, I decided to impute Age years because it is only missing 2.4% of its values. I imputed using the median age of 33.*

: Awesome: All the variables, that you have removed are the right ones. Bonus points for imputing the missing records for the age field with the median of the age field.

## 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:*

LOG

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

Account balance is the most statistically significant variable of the LOG model.

### Coefficients:

	Estimate	Std. Error	z	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	.

Significance codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

2. 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?

The LOG model is 76% accurate. There is a bias towards Creditworthiness.

### Confusion matrix of LOG

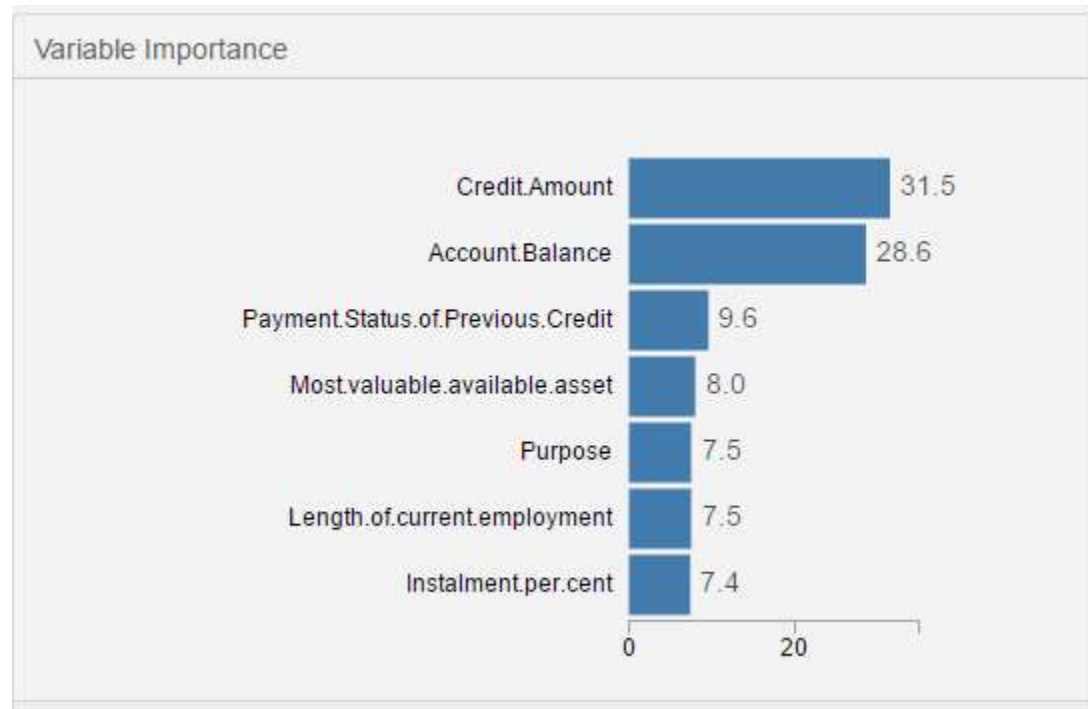
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

: Awesome: The confusion matrix for Log regression looks good.

### TREE

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

Credit amount and Account balance are the most important variables of this model.



: Required: The variable importance plot for Tree model is not correct.

The confusion matrix for Tree model is also off.

- 1) Please ensure that the all the variables are used in variable importance plot.
- 2) Use the most important variables for building the model.

4. 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?

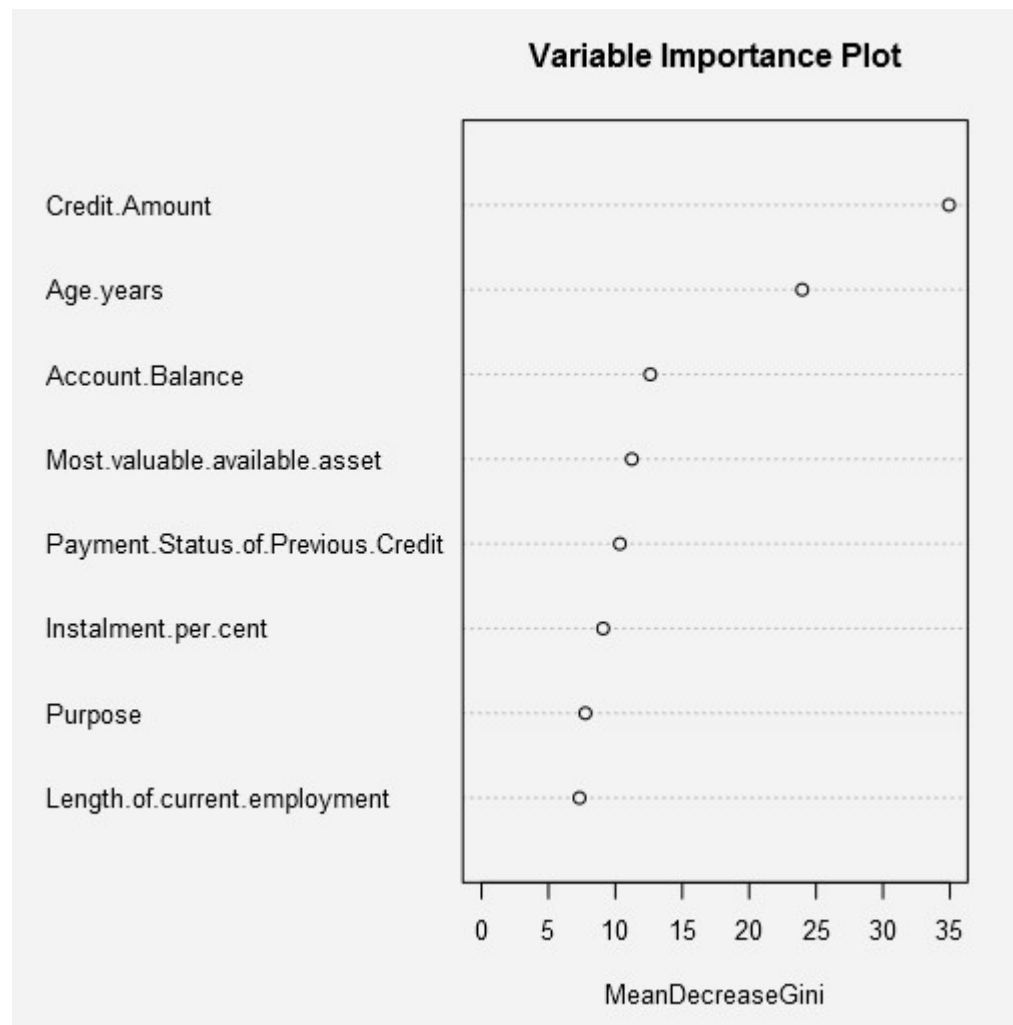
The TREE model is 74% accurate. There is a bias towards Creditworthiness.

Confusion matrix of TREE		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	89	23
Predicted_Non-Creditworthy	16	22

FOREST

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

Credit amount is by far the most important variable of this model.



: Required: The variable importance plot is correct up until Payment.Status.of.Previous.Credit.

However, some variables are missing from the variable importance plot.

Please use all the variables in the variable importance plot, other than the seven variables we decided to exclude.

The confusion matrix is also not correct.

6. 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?

The FOREST model is 78.67% accurate. There is a bias towards Creditworthiness.

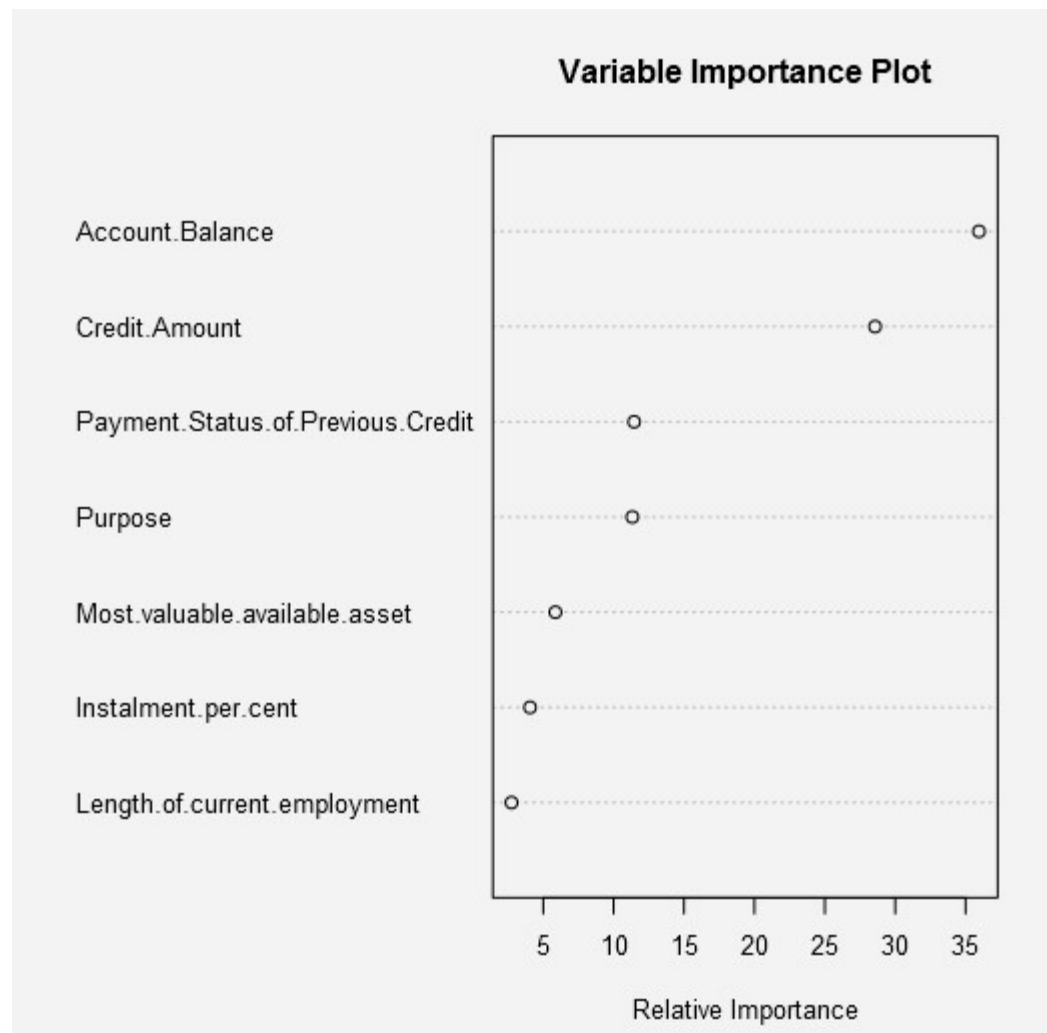
Confusion matrix of FOREST		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	98	25
Predicted_Non-Creditworthy	7	20

BOOST

7. 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 Account balance and Credit amount variables seem to be the most important.





: Required: The variable importance plot and confusion matrix for Boosted model are not correct.

Please make sure you use all the variables for making variable importance plot.

8. 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?

The BOOST model is 76.67% accurate. There is a bias towards Creditworthiness.

Confusion matrix of BOOST		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	30
Predicted_Non-Creditworthy	5	15

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

## 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:

- 1. Which model did you choose to use? Please justify your decision using only the following techniques:
  - a. Overall Accuracy against your Validation set
  - b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
  - c. ROC graph
  - d. Bias in the Confusion Matrices

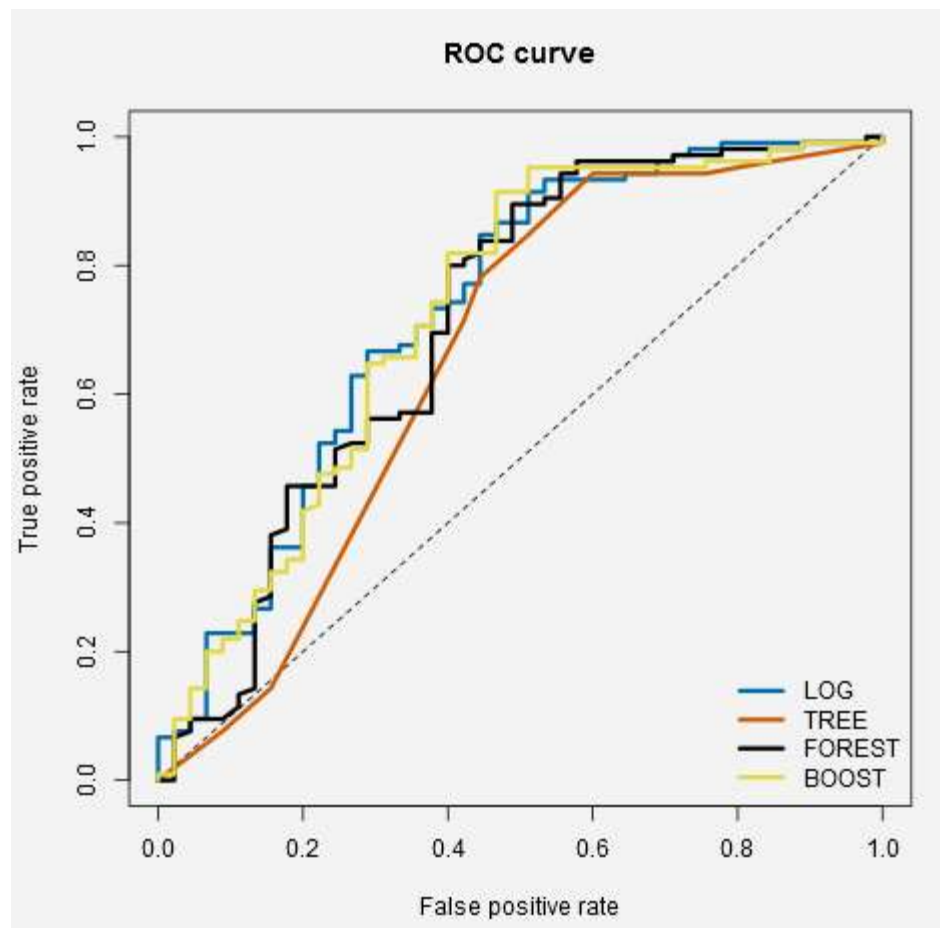
I chose to use the Random Forest Model, FOREST, because it slightly outperforms the other models.

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LOG	0.7600	0.8364	0.7306	0.8000	0.6286
TREE	0.7400	0.8203	0.6585	0.7946	0.5789
FOREST	0.7867	0.8596	0.7117	0.7967	0.7407
BOOST	0.7867	0.8632	0.7303	0.7829	0.8095

The chart above demonstrates that the FOREST model has a slightly higher accuracy rate than the other models except the BOOST model. However, the Creditworthy accuracy of the FOREST model is higher than the BOOST model. The FOREST model has a 78.67% accuracy rate and an accuracy of Creditworthiness of 79.67%. This means that the predictive power of the model is strong.

: Awesome: The accuracy and F1 score for Log regression and Boosted model are correct.

Required: The accuracy and F1 score for Tree and Forest model are not correct.



: Required: Please correct for the ROC curves after building the correct model.

According to the ROC graph above, the FOREST model performs well in general. It performs well in comparison to the other models.


Confusion matrix of BOOST		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	30
Predicted_Non-Creditworthy	5	15

Confusion matrix of FOREST		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	98	25
Predicted_Non-Creditworthy	7	20

Per the confusion matrix above, the FOREST model displays a bias towards Creditworthiness that is less pronounced, and therefore better, than the BOOST model..

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

2. How many individuals are creditworthy?  
405 individuals are creditworthy out of the 500 new applicants.



: Suggestion: The number of creditworthy individuals is an acceptable solution, however, differs from ideal solution.

### 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.