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

- What decisions needs to be made?
- What data is needed to inform those decisions?
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

We need to build a model based on previous data of what type of applicants were given loans. We may try four different models and compare them to choose the best fitting one. And then, we apply the new data to that model and predict each applicant of whether whom shall be given a loan.

We need to use binary model to help make the decisions, since we only need to answer creditworthy or non-creditworthy to the application result. And we need to find out what are significant predictor variables and the best fitting model.

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

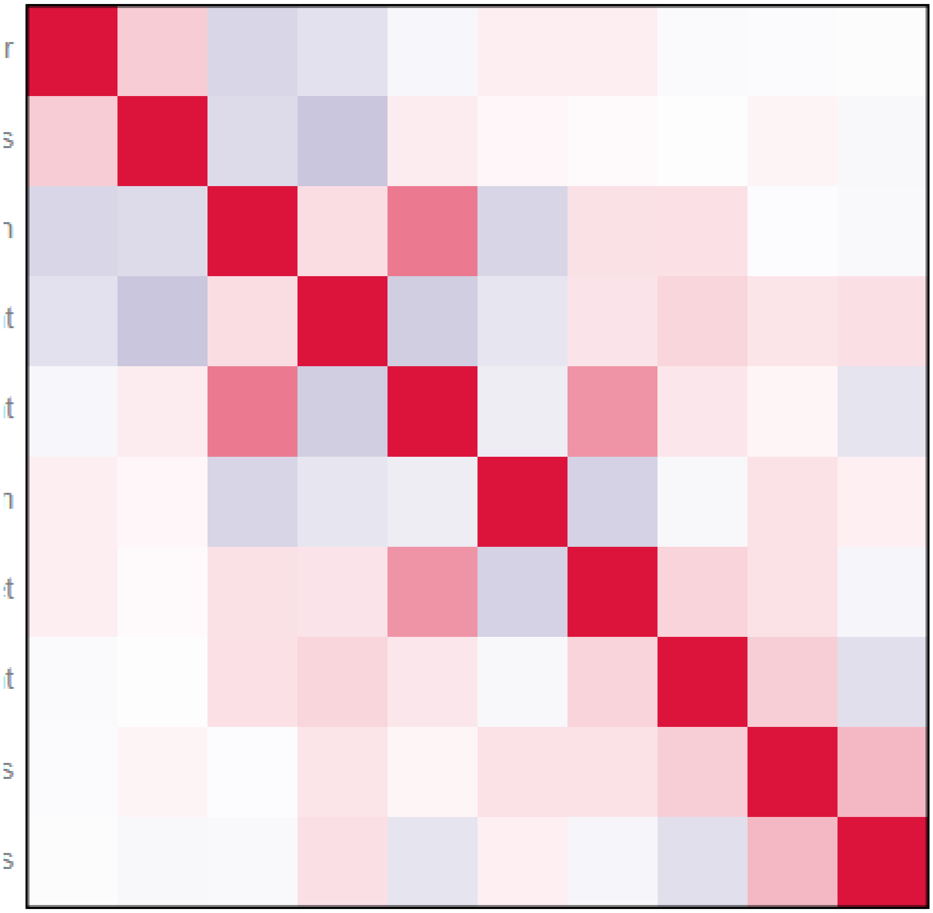
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

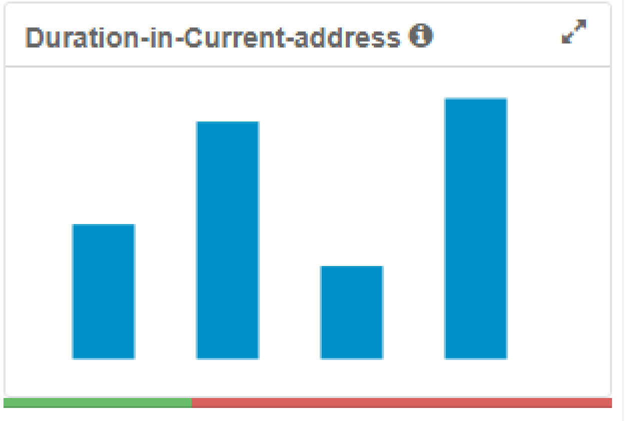
Answer this question:

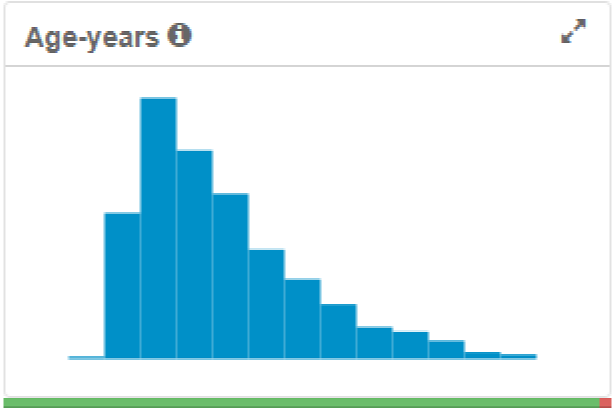
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

There is no field that highly-correlates with each other for numerical data fields, since there is no as high as .70 correlation showed up in the report of Association Analysis.
However, when using the Association Analysis, I deselected the Occupation field, since this field has only one data for every entry. There would be no meaning analyze this field anymore.
Further, I purposely deselected the Telephone field, since this one has no logic meaning related to the creditability of applicant.

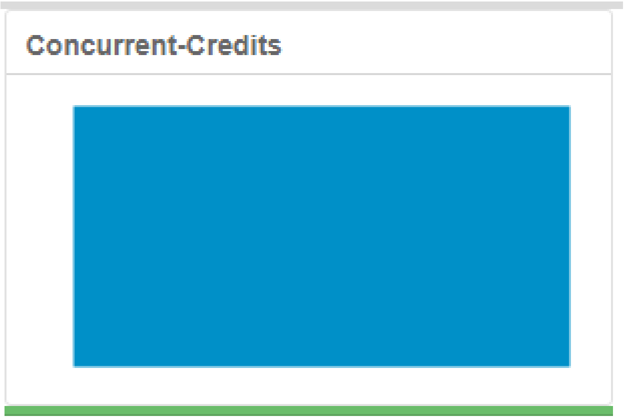


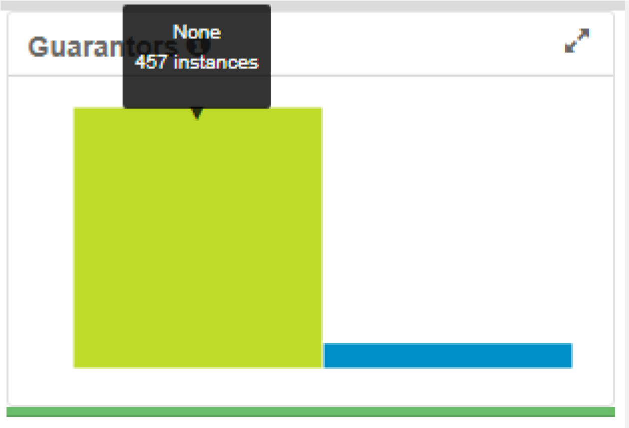
The Duration-in-Current-address field has 69% missing data showed in the report of Field Summary. This one shall be deleted.
In the meantime, the Age-yeas field has a 3% missing data, which shall be saved for further cleaning.

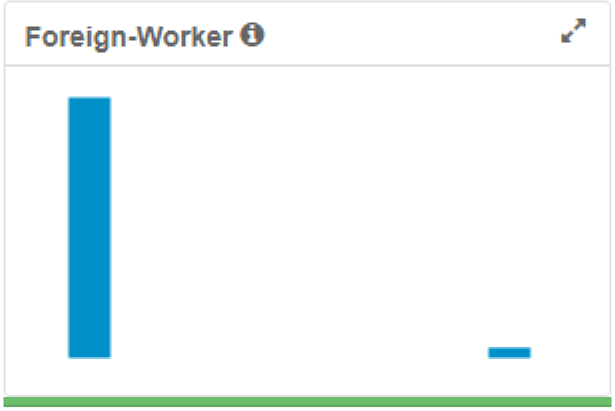


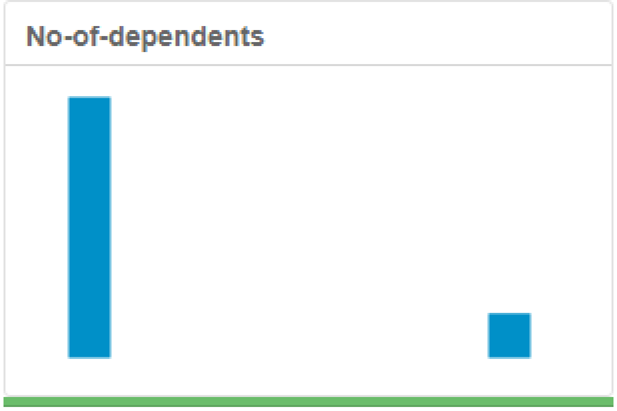


There are some fields with low variability. They are Concurrent-Credits, Guarantors, No-of-dependents, and Foreign-Worker. They shall be deselected.

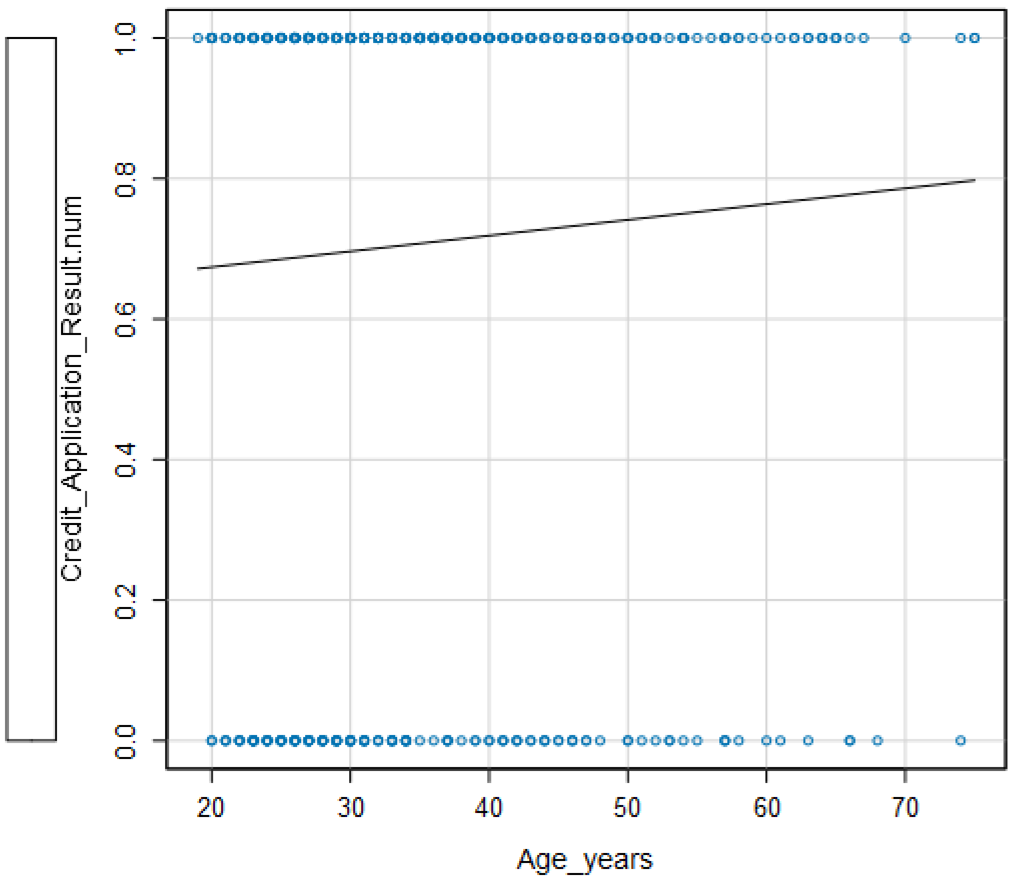


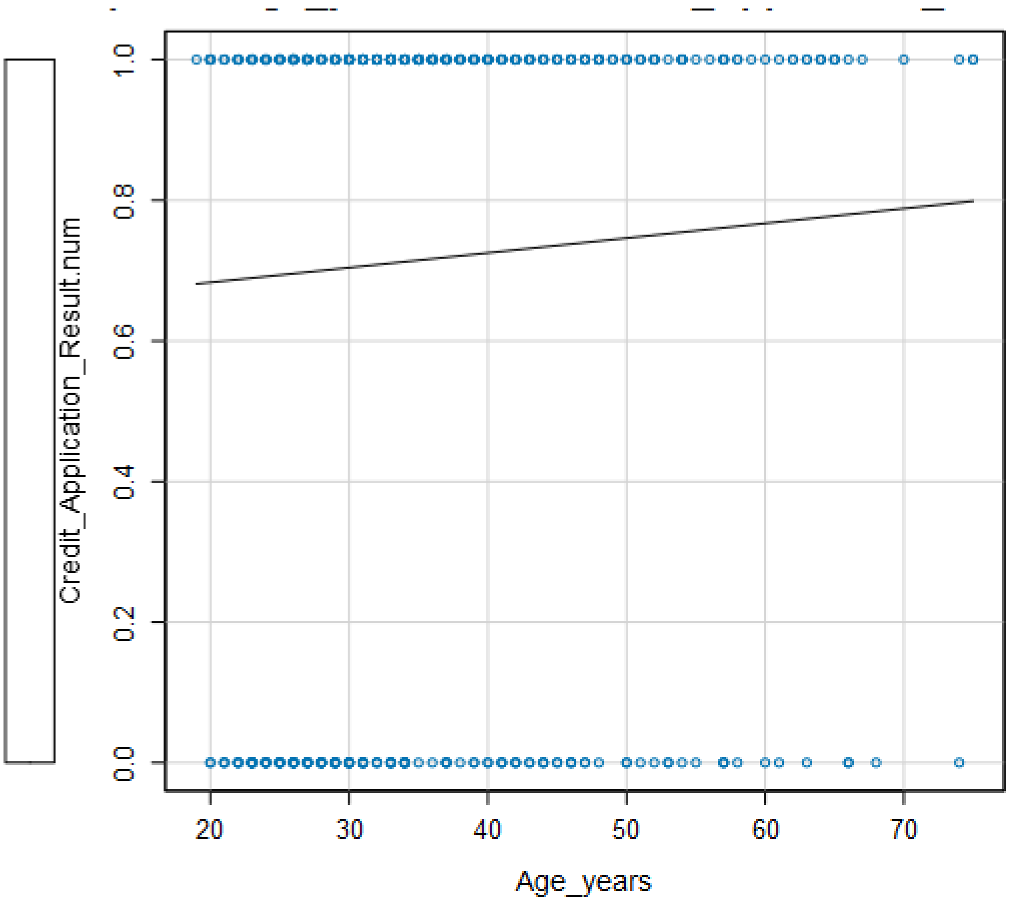






For the Age-year field, I impute the data using the median of the entire data field. By comparing the scatter plots before and after the imputing, the trends of the data are similar. Therefore the imputing is good.


Credit_Application_Result.num
1.0
0.8
0.6
0.4
0.2
0.0
20
30
40
50
60
70
Age_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.

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

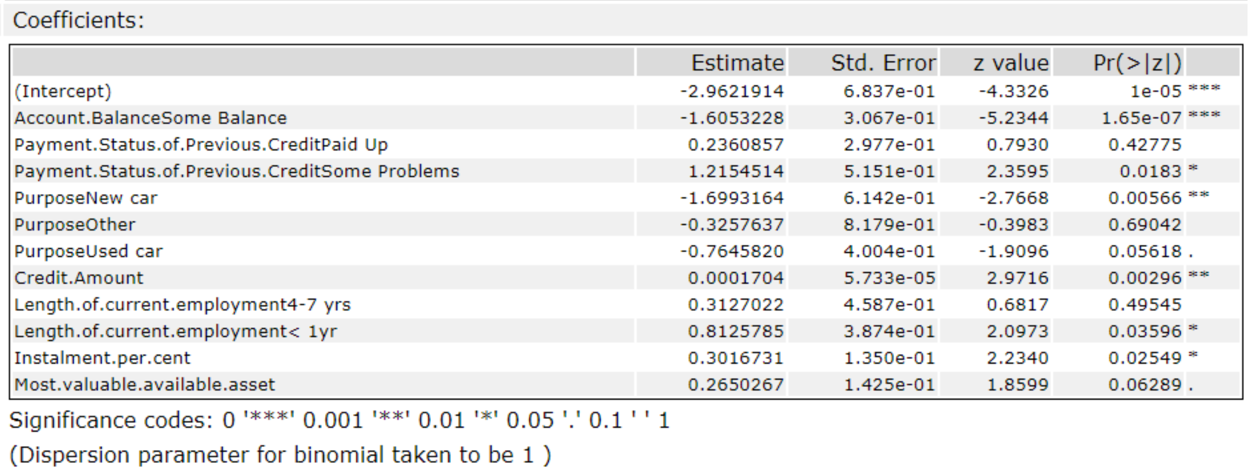
*Answer these questions for **each model** you created:*

- 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)

For the Logistic Regression plus Stepwise model, among all the predictor variables, the Account-Balance, Purpose and Credit-Amount are significant. Here goes the chart.



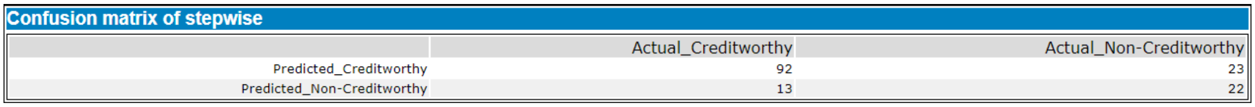
Coefficients:

	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	.

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

(Dispersion parameter for binomial taken to be 1)

The overall percent of accuracy is 0.76. The confusion Matrix goes here.

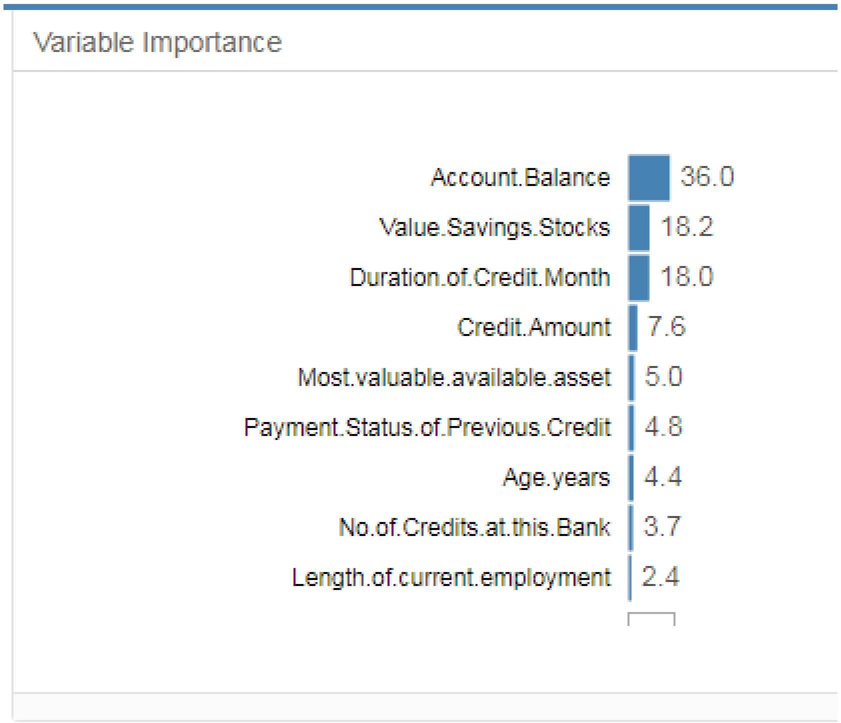


Confusion matrix of stepwise

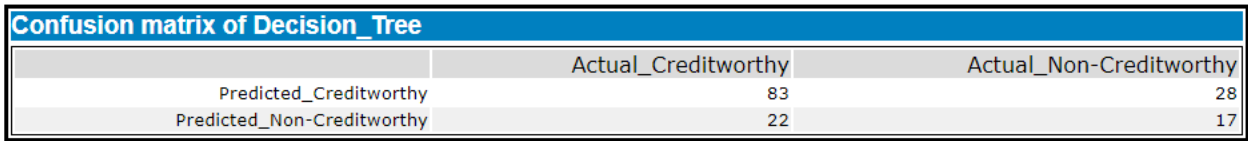
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

There is a bias in the model's prediction. The Non-creditworthy is much harder to predict for this model. The accuracy is as low as 0.4889. The Confusion Matrix also show much predicted non-creditworthy are actual creditworthy.

For Decision Tree Model, the predictor variables of Account-Balance, Value-Savings-Stocks are significant. The Variable Importance Chart goes as:



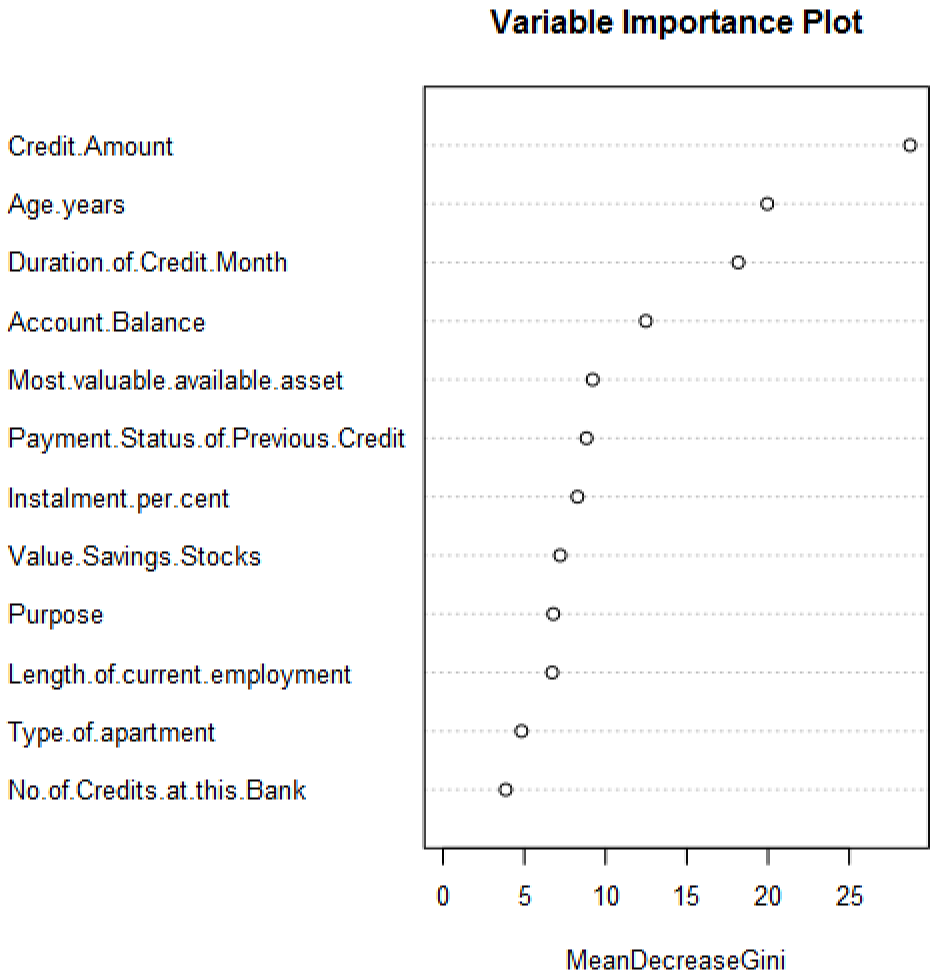
The overall percent of accuracy is 0.7467. There is a bias in the model's prediction. The Non-creditworthy is much harder to predict for this model. The accuracy is as low as 0.4667. . The Confusion Matrix also show much predicted non-creditworthy are actual creditworthy.



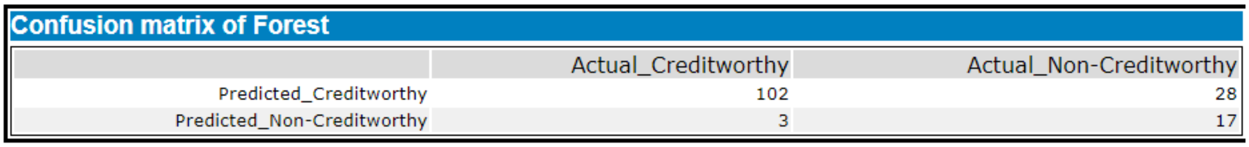
Confusion matrix of Decision_Tree

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

For the Forest Model, among all the predictor variables, the Credit Amount is most important, and Age-Years and Duration of Credit Month are significant. Here goes the chart.



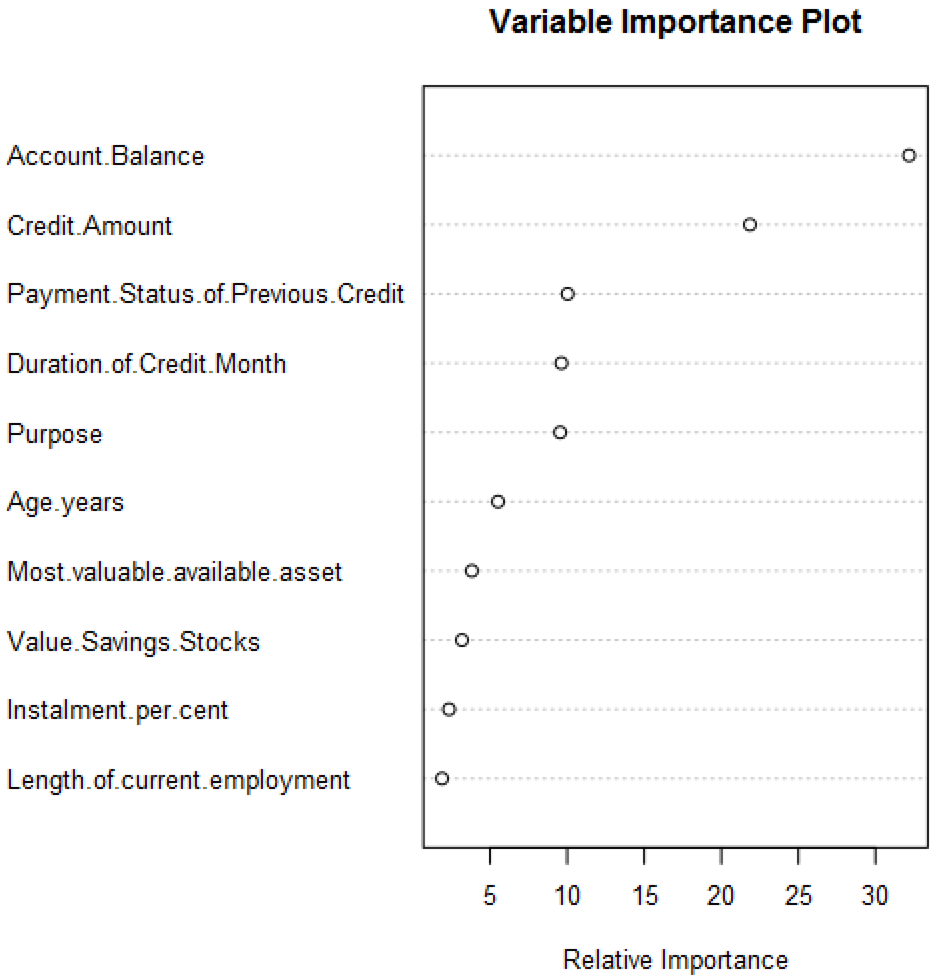
The overall percent of accuracy is 0.7933. There is a bias in the model's prediction. The Non-creditworthy is much harder to predict for this model. The accuracy is as low as 0.3778. . The Confusion Matrix goes as:



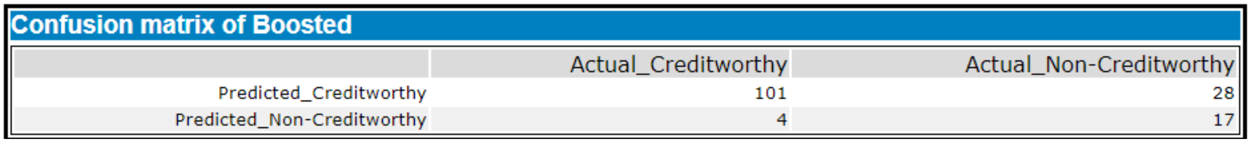
Confusion matrix of Forest

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

For the Boosted Tree Model, among all the predictor variables, the Account Balance is most important, and Credit-Amount is significant. Here goes the chart:



The overall percent of accuracy is 0.7867. There is a bias in the model's prediction. The Non-creditworthy is much harder to predict for this model. The accuracy is as low as 0.3778. . The Confusion Matrix goes as:



Confusion matrix of Boosted

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

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)

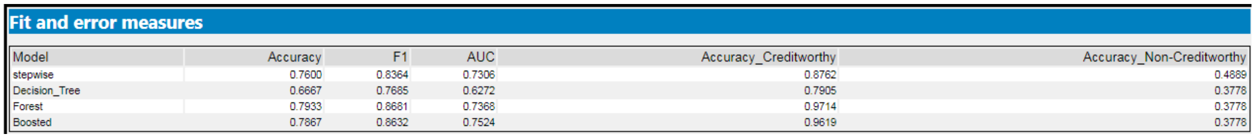
Answer these questions:

- 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
    - ○ ROC graph
    - ○ Bias in the Confusion Matrices

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

- How many individuals are creditworthy?

To compare the four models, I start with overall accuracy. Within them, the Forest Model is the best as 0.7933. The following is the Boosted Tree as 0.7867, the Logistic Stepwise 0.7600, and the Decision Tree as 0.6667. So I would rule out the Decision Tree first.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.6667	0.7685	0.6272	0.7905	0.3778
Forest	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted	0.7867	0.8632	0.7524	0.9619	0.3778

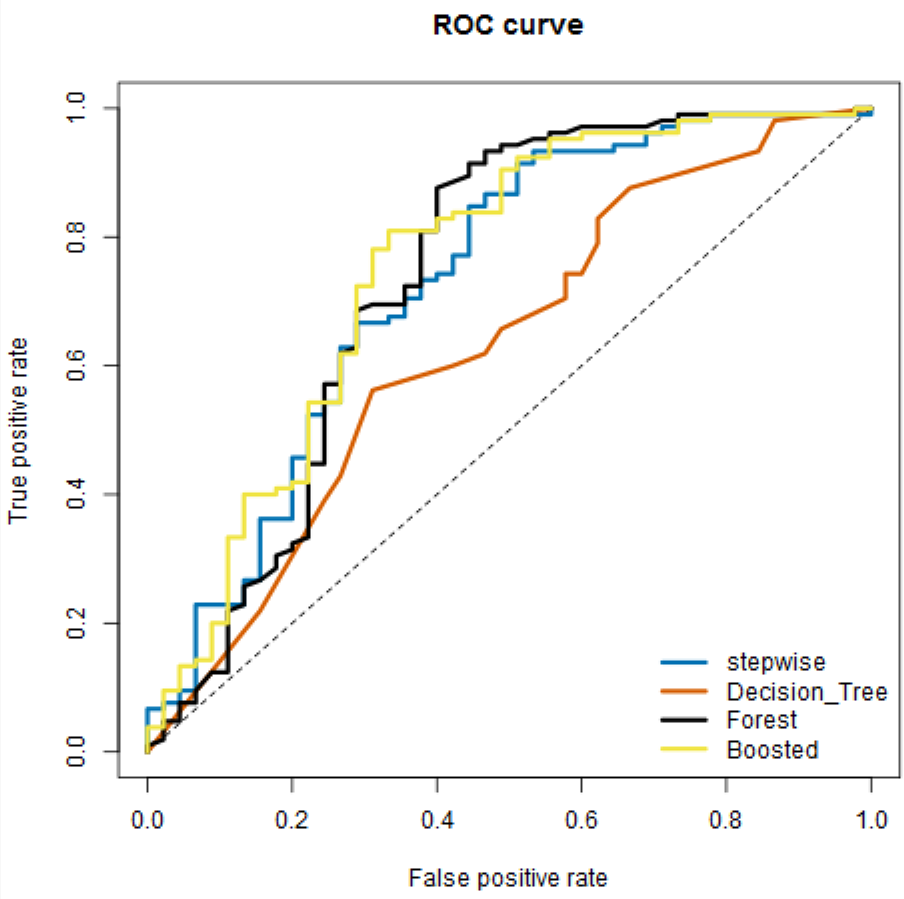
Then, I will see the accuracies within “Creditworthy”. Among the left three, the Forest is still the best as 0.9714. The following is the Boosted as 0.9619, and the logistic stepwise as 0.8762. For the accuracies within “non-Creditworthy”, the Forest is as good as the Boosted, as 0.3778. And the Logistic Stepwise is better as 0.4889.

Consider the three accuracy data, the Forest looks to have the best comprehensive performance.

All the four models have similar issue with predicting non-credit worthy. The bias exists, but does not change the previous choose of the Forest.

For ROC graph,

Both the Boosted and the Forest do good job separating the classes, since they have ROC curves that hug the upper left corner of the plot.



Therefore, I choose the Forrest to score.

408 individuals 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.