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?
  Based on the previous applications for credit approval, determine whether each incoming application is Creditworthy or Non-Creditworthy.

- What data is needed to inform those decisions?
  Information on customers who were Creditworthy and information on customers who were Non-Creditworthy is needed to evaluate their differences. What is provided is datasets complete with features such as the purpose of the loan, the credit amount, the age of the applicant, whether the applicant has a guarantor, living condition (apartment, house, etc.), collateral, and other predictors the bankers have used to assess whether someone is Creditworthy or Non-Creditworthy.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  A binary model is needed, since the outcome is Creditworthy or Non-Creditworthy (only 2 opposite possibilities).

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)*

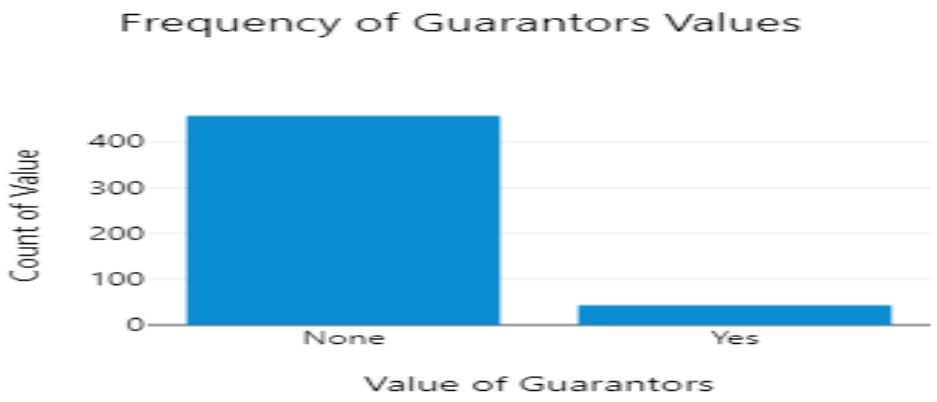
Note: *For students using software other than Alteryx, please format each variable as:*

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:

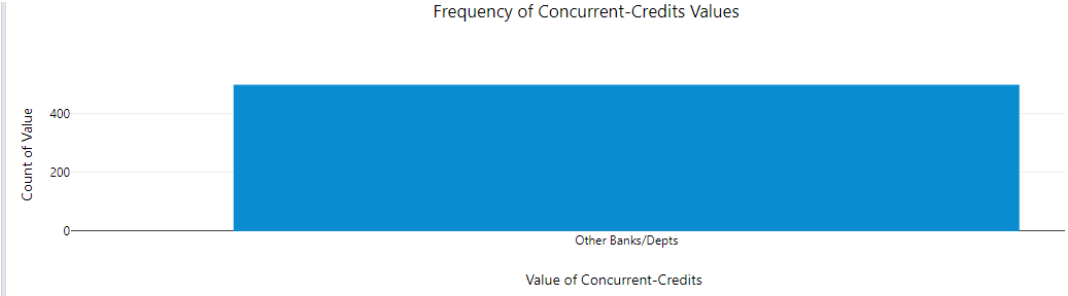
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
  I will go through each column and explain my decision to keep, remove, or impute data. If not mentioned, there were not any null values. For Credit-Application-Result, this is the target variable so I kept. For Account Balance, according to the frequency table, there is not low variability, so I kept it. For Duration-of-Credit-Month, there aren't any null values, so I kept it. For Payment-Status-of-Previous-Credit, there was only 7.2% of the values as 'Some Problems' but this is a good value to discern from so I kept the variable intact. For Purpose, there was not low variability, so I kept it. For Credit-Amount, this is a crucial variable so I kept it. For Value-Savings-Stocks, there was not low variability, so I kept it. For Length-of-current-employment, there was not low variability, so I kept it. For Installment-per-cent, there was not low variability, so I kept it. For Guarantors, there is low variability, so I removed it.



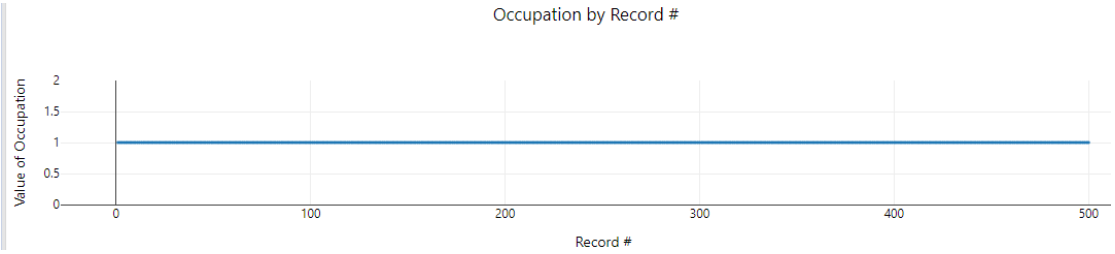
For Duration-in-Current-address, there was more than 50% missing, so I removed it.

Duration-in-Current-address	
Data Type	Double
Size	8
Non-Nulls	156
Nulls	344

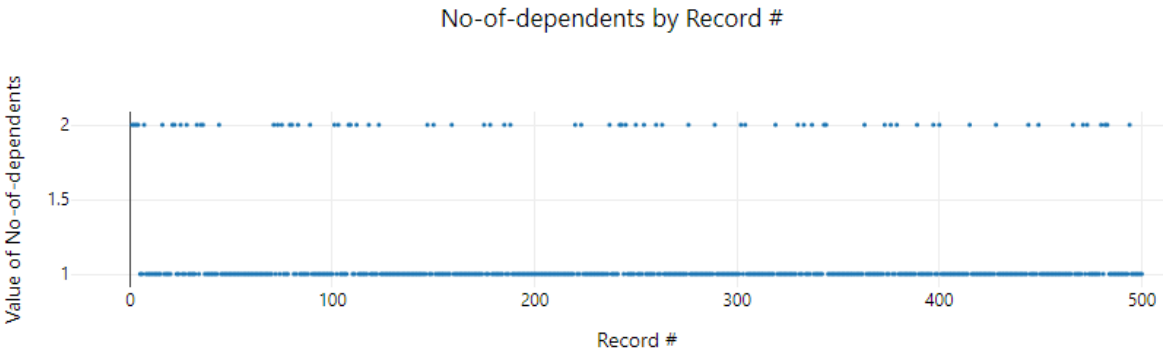
For Most-valuable-available-asset, this could be important for collateral, so I kept it. For Age-years, there were 12 missing values so I imputed them with the median and kept it. For Concurrent-Credits, there was only one value, so I removed the column.



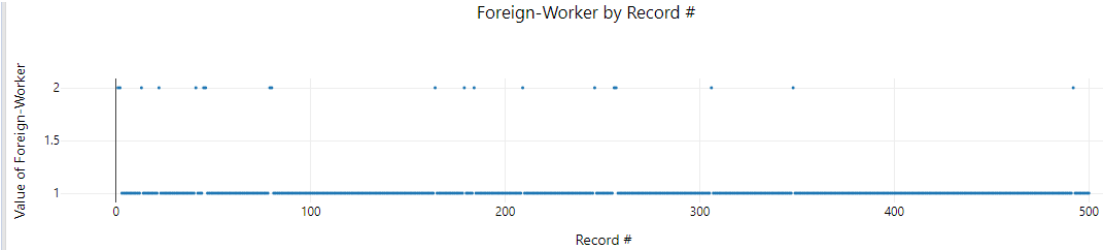
For Type-of-apartment, there was not low variability, so I kept it. For No-of-Credits-at-this-Bank, there was not low variability, so I kept it. For Occupation, there was only one values, so I removed it.



For No-of-dependents, there was low variability, so I removed it.



For Telephone, it does not make any sense to keep, so I removed it. For Foreign-Worker, there was low variability, so I removed it.



Upon doing a Pearson correlation against all the continuous variables, there were no correlations above 70%. Therefore, no additional variables were removed.

FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Age-years	No-of-dependents
Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	-0.047155	-0.065269
Credit-Amount	0.57398	1	-0.288852	0.325545	0.074901	0.003986
Instalment-per-cent	0.068106	-0.288852	1	0.081493	0.026562	-0.125894
Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.103999	0.046454
Age-years	-0.047155	0.074901	0.026562	0.103999	1	0.117349
No-of-dependents	-0.065269	0.003986	-0.125894	0.046454	0.117349	1

I came up with 13 total variables. Once I have trained the classifiers, I can see which variables produce no impact and make adjustments.

Variables removed: Guarantors, Duration-in-Current-address, Concurrent-credits, Occupation, Telephone, Foreign-Worker, No-of-dependents
Variables imputed: 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.

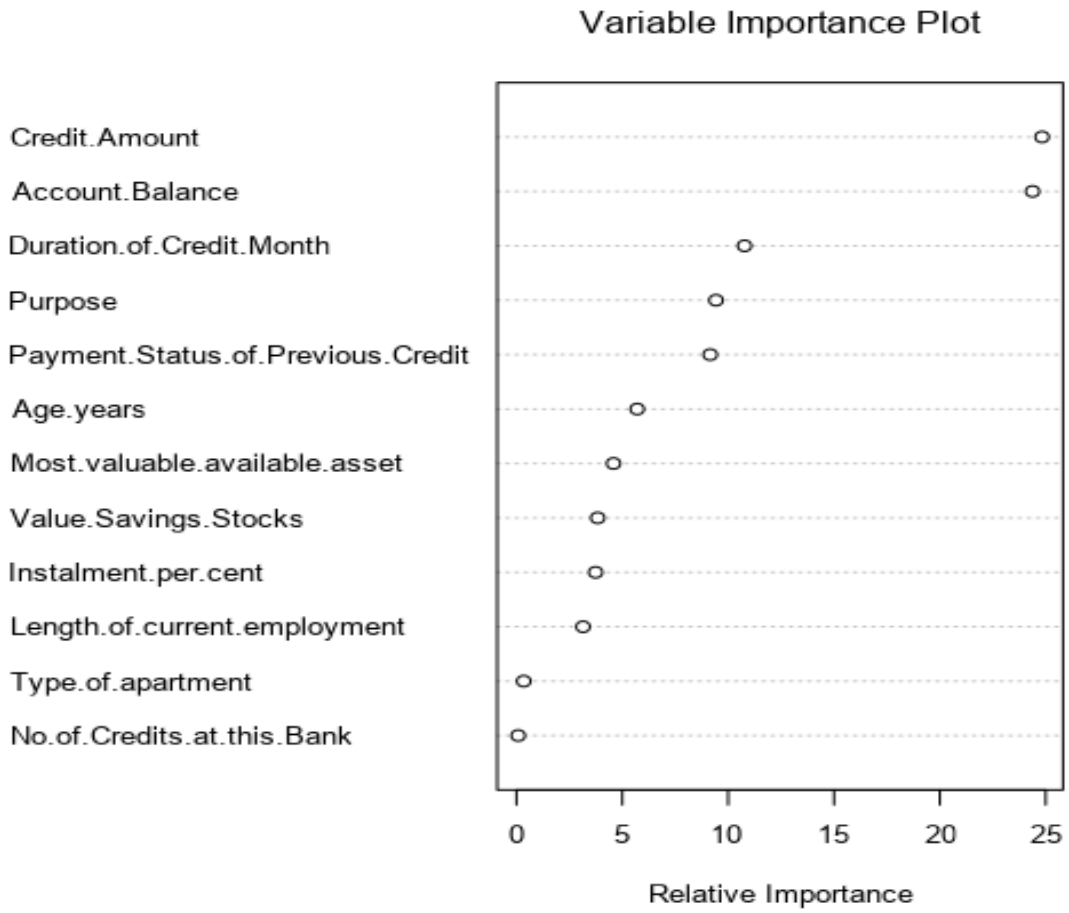
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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

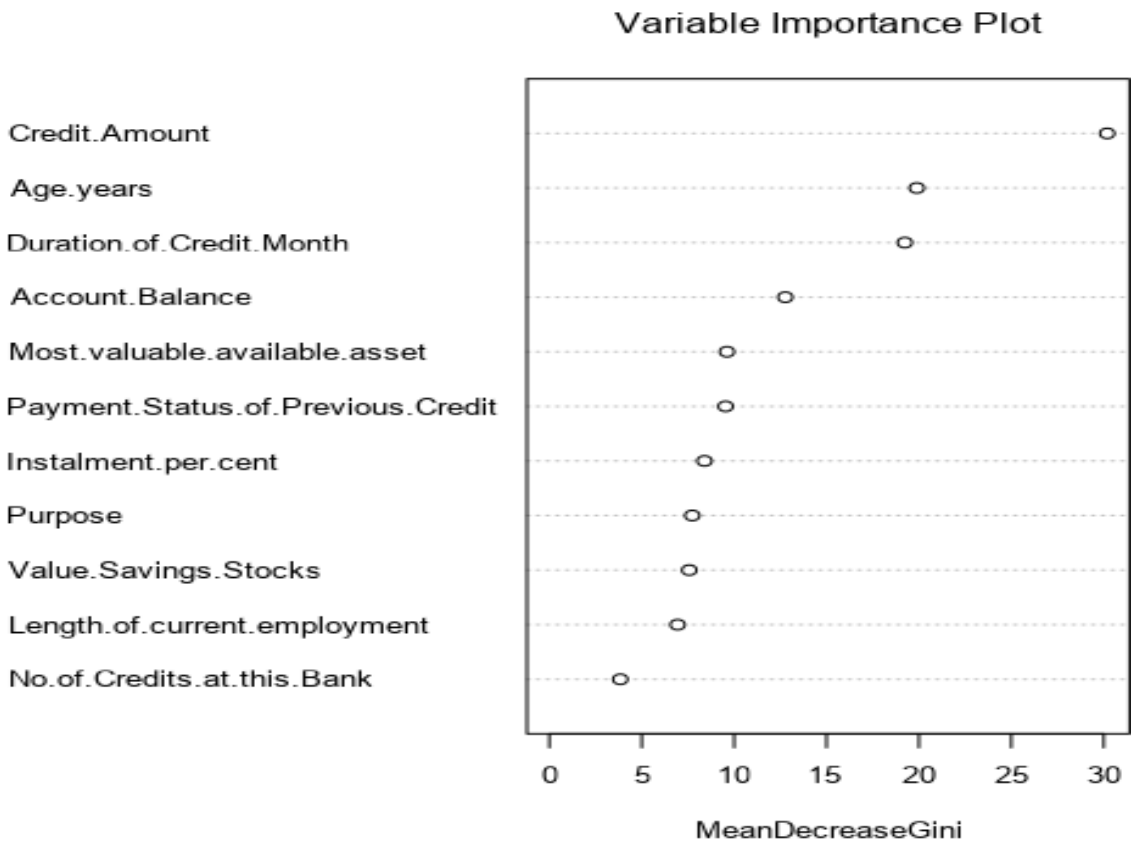
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
  For DecisionTree, the most important predictor variables are Account Balance, Duration of Credit Months, and Value Savings Stocks.

```
Leaf Summary
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
   2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
   3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
     6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
     7) Duration.of.Credit.Month>=13 110 51 Non-
Creditworthy (0.4636364 0.5363636)
      14) Value.Savings.Stocks=< £100,£100-
£1000 34 11 Creditworthy (0.6764706 0.3235294) *
      15) Value.Savings.Stocks=None 76 28 Non-
Creditworthy (0.3684211 0.6315789) *
```

For BoostedModel, the most important predictor variables are Credit Amount, Account Balance, Duration of Credit Month, Purpose, and Payment Status of Previous Credit.



For RandomForest, the most important predictor variables are Credit Amount, Age years, Duration of Credit Month, Account Balance, Most Valuable Available Asset, and Payment Status of Previous Credit.



For StepwiseLogisticReg, the most important predictor variables are the Account Balance, Credit Amount, Purpose, Payment Status of Previous Credit, Installment Per Cent, and Length of Current Employment.

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	.

- 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 Model Comparison is found in the ModelComparisonReport pdf. Creditworthy was provided as the true value for the Model Comparison. All listed metrics are provided as Decision Tree, Forest Model, Boosted Model, and Stepwise Logistic Regression Model, respectively.

Looking at the Fit and Error Measures table, the overall accuracies are 74.67%, 80.67%, 78.67%, and 76%. It is apparent Random Forest is the best with an overall accuracy of 80.67%. The F1 scores are 0.8273, 0.8745, 0.8632, and 0.8364. Random Forest is best in this statistic also.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionTree	0.7467	0.8273	0.7054	0.8667	0.4667
ForestModel	0.8067	0.8745	0.7468	0.9619	0.4444
BoostedModel	0.7867	0.8632	0.7513	0.9619	0.3778
StepwiseLogisticReg	0.7600	0.8364	0.7306	0.8762	0.4889

Looking at the Confusion Matrices below the Fit and Error Measures table illustrates other factors. The Boosted model performs the same as the Random Forest when predicting the actual Creditworthy applicants. Other models perform better predicting Non-Creditworthy customers, making the True value (Creditworthy in this situation) important in evaluating this model.

For all models, it can be seen predicting Creditworthy customers as creditworthy has a much higher accuracy than predicting Non-Creditworthy customers as Non-Creditworthy. If you observe the columns on the left, it can be seen in all models the Creditworthy applicants predicted as Non-Creditworthy is much smaller than the Creditworthy applicants correctly identified. However, when looking at the right columns, it can be observed that the Non-Creditworthy applicants predicted as Creditworthy and Non-Creditworthy are very in the number range. Therefore, the bias is in all models' prediction to Creditworthy applicants.

Confusion matrix of BoostedModel

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

Confusion matrix of DecisionTree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

Confusion matrix of StepwiseLogisticReg

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

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"

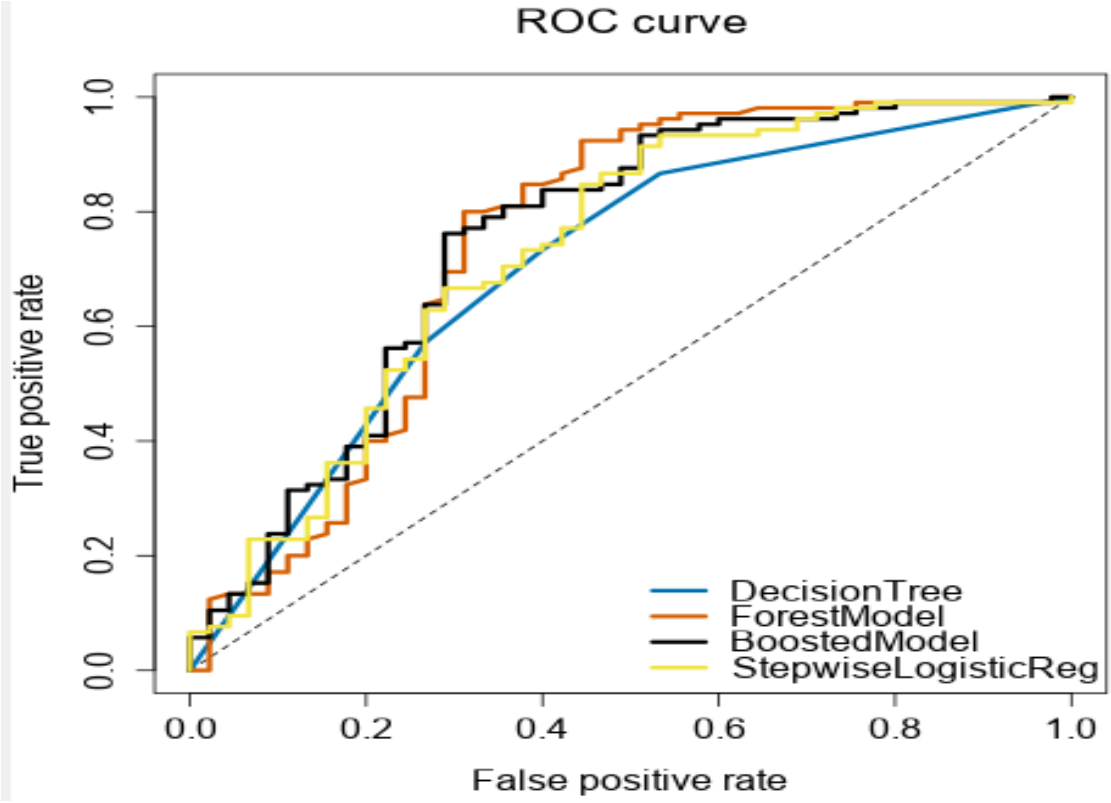
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- 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

  Looking at the Fit and Error Measures table, it is apparent Random Forest is the best choice with an overall accuracy of 80.67%. It outperforms the other three models in

Creditworthy accuracy (96.19%) and the F1 score (0.8632). Since we deemed Creditworthy as the true value for our Model Comparison, these two accuracies matter most. Looking at the Confusion Matrices below the Fit and Error Measures table illustrates other factors. The Boosted model performs similar to the Random Forest. Other models perform better predicting Non-Creditworthy customers, making the decision of the True value important in evaluating this model. Therefore, the bias is in all models' prediction to Creditworthy applicants. The ROC curve in the Model Comparison pdf shows the Boosted Model and the Forest Model above the other two models in most occasions. The Forest Model reaches the top the fastest.



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

- How many individuals are creditworthy?
  Using the Random Forest, there are 406 out of the 500 new applicants are predicted as Creditworthy.

Record	Sum_Creditworthy
1	406

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.