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 to be made is to process all the 500 loan applicants, whether they are creditworthy or not, according the applicants data.

   We can predict the creditworthiness of the applicant data by creating a model using all our past applications data

2. **What data is needed to inform those decisions?**

   We would need data related with clients that can :

1. **Show that client is capable to pay back the loan:**
   Example : Account Balance, Payment status of previous credit, length of current employment, most available assets, age , guarantors, value savings stocks, income etc.

2. **The risk client may have that can cause client may unable to pay the loan:**
   For Example: Installment percentage, Duration of credit month, number of dependents, other current credit, job risk level etc.

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

   We need a binary model with end result of predicting whether the applicant is creditworthy or not.

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.***

<u>*Here are some guidelines to help guide your data cleanup:*</u>

- *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".*

  According to Pearson Correlation Analysis, there is no highly correlation (all less than 0.70)with the Credit result variable and with each other. So we can use all the numerical variable to take it to the next test (missing data & variability test)

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value
Most.valuable.available.asset	-0.232248	0.0050930**
Duration.of.Credit.Month	-0.215149	0.0096065**
Instalment.per.cent	-0.130496	0.1190020
Age.years	0.123088	0.1416213
Credit.Amount	-0.092205	0.2717004
Foreign.Worker	0.072525	0.3876717
Duration.in.Current.address	0.067284	0.4229716
Type.of.apartment	-0.039360	0.6395134
No.of.dependents	0.038037	0.6508161
Telephone	0.030838	0.7136766

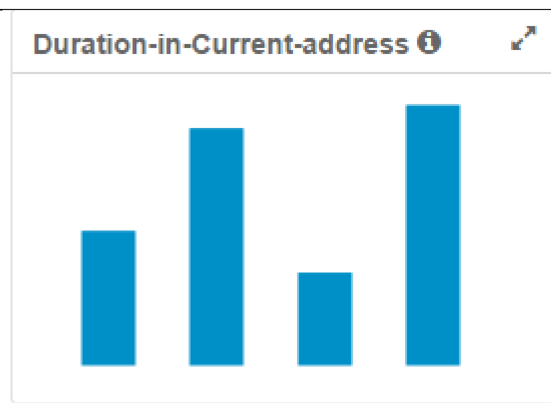
Full Correlation Matrix

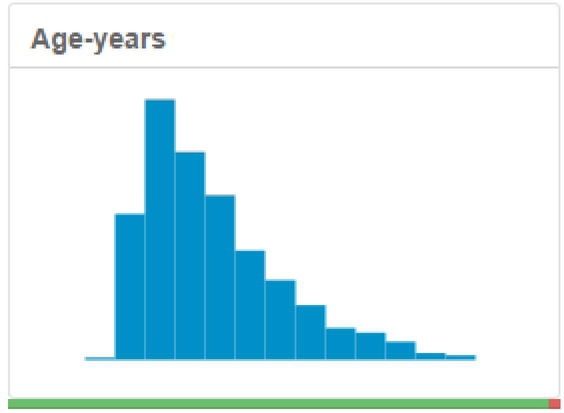
	Credit.Applicat	Duration.of.	Credit.A	Instalmen	Duration.in.C	Most.valuable.
Credit.Applicat	1.000000	-0.215149	-0.092205	-0.130496	0.067284	-0.232248
Duration.of.Cr	-0.215149	1.000000	0.565054	0.145637	-0.032494	0.128814
Credit.Amount	-0.092205	0.565054	1.000000	-0.253286	-0.136621	0.457147
Instalment.per	-0.130496	0.145637	-0.253286	1.000000	0.131231	0.115114
Duration.in.Cu	0.067284	-0.032494	-0.136621	0.131231	1.000000	-0.047386
Most.valuable.	-0.232248	0.128814	0.457147	0.115114	-0.047386	1.000000
Age.years	0.123088	-0.018171	0.040486	0.111456	0.301966	0.123579
Type.of.apart	-0.039360	0.126967	0.100413	0.178926	-0.163386	0.182744
No.of.depende	0.038037	-0.185180	0.082721	-0.293380	-0.036814	0.019435
Telephone	0.030838	0.238437	0.192532	0.038515	0.055112	0.083395
Foreign.Worke	0.072525	-0.207298	-0.045994	-0.155458	-0.015787	0.071932

	Age.years	Type.of.apa	No.of.de	Telephon	Foreign.Work
Credit.Applicat	0.123088	-0.039360	0.038037	0.030838	0.072525
Duration.of.Cr	-0.018171	0.126967	-0.185180	0.238437	-0.207298
Credit.Amount	0.040486	0.100413	0.082721	0.192532	-0.045994
Instalment.per	0.111456	0.178926	-0.293380	0.038515	-0.155458
Duration.in.Cu	0.301966	-0.163386	-0.036814	0.055112	-0.015787
Most.valuable.	0.123579	0.182744	0.019435	0.083395	0.071932
Age.years	1.000000	0.208552	0.046996	0.141103	-0.020939
Type.of.apart	0.208552	1.000000	-0.010189	0.179688	-0.026742
No.of.depende	0.046996	-0.010189	1.000000	-0.097632	0.218454
Telephone	0.141103	0.179688	-0.097632	1.000000	-0.168472
Foreign.Worke	-0.020939	-0.026742	0.218454	-0.168472	1.000000

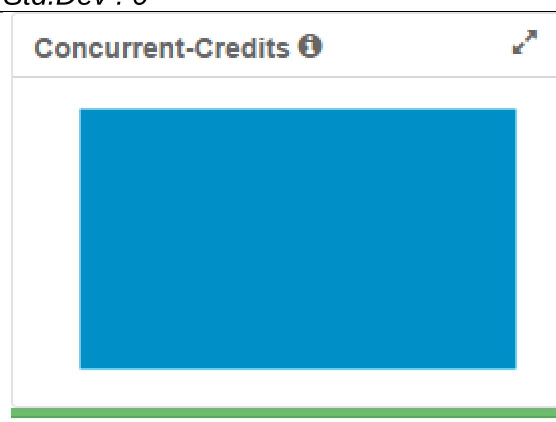
- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*

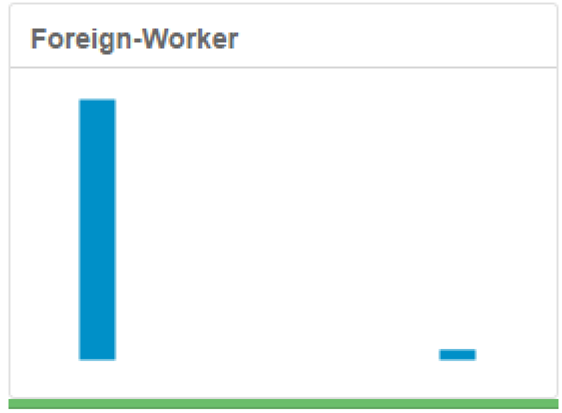
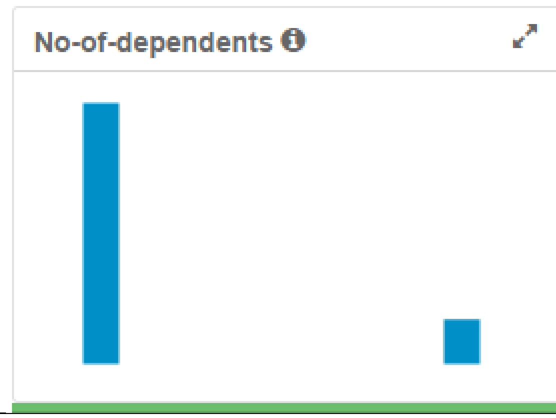
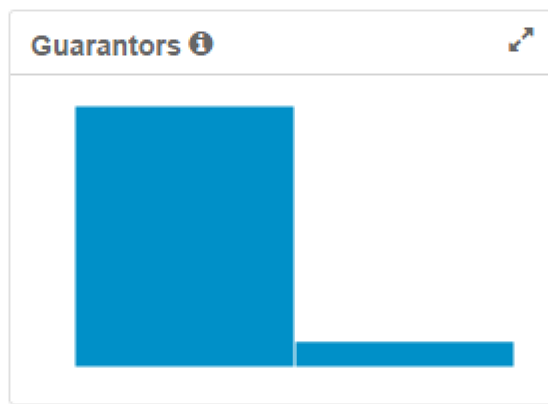
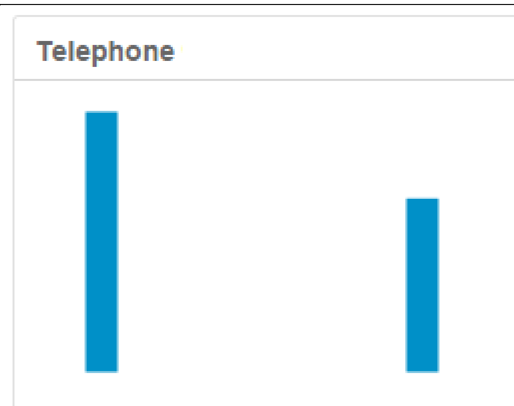
Duration in current address variable has 68,8% missing data out of 500 records, I decided to exclude this variable.



Age year variable has 2.4% missing data. We will include this variable and conduct imputation for the missing record, using average of age values in the data.	

- *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.*

Occupation has uniform value, which mean it has no variability and it will have no impact on prediction analysis. So we will Exclude it	*Occupation:* *Min : 1* *Max: 1* *Med : 1* *Std.Dev : 0*
Concurrent Credit has uniform value, which mean it has no variability and it will have no impact on prediction analysis. So we will Exclude it	

Foreign Worker variable has low variability and very skewed, so we will exclude it	
No od Dependents variable has low variability and very skewed, so we will exclude it	
Guarantors variable has low variability and very skewed, so we will exclude it	
While the data is not skewed, Telephone variable still has low variability, and I don't think number of telephone owned will relate much with the creditworthiness, so we will exclude the variable .	

- *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)*

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.

The variable that I decided to exclude are:

1. Duration in Current Address - Too many missing data at 68%
2. Concurrent Credit - Low variability, the data is entirely uniform
3. Occupation - Low Variability, the data is entirely uniform
4. Guarantors - Low variability
5. Foreign Worker - Low variability
6. No-of-Dependents - Low variability
7. Telephone - Low variability

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:*

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.

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?

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

	Logistic Regression	Decision Tree	Forest Model	Boosted Model
Important Variable	Account Balance***	Account Balance***	Credit Amount***	Account Balance***
	Purpose **	Duration of Credit Month***	Age Year**	Credit Amount**
	Credit Amount **	Value Savings Stock***	Duration of Credit Month**	Duration of Credit*
	Instalment percent*		Account Balance*	Payment Status of Previous Credit*
	Payment Status of previous Credit.			Purpose*
	Most valuable available asset			
Accuracy	76%	74.67%	80%	78.67%

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistic_Regression_Model	0.7600	0.8364	0.7306	0.8000	0.6286
DecisionTree	0.7467	0.8273	0.7054	0.7913	0.6000
Random_Forest_Model	0.8000	0.8718	0.7426	0.7907	0.8571
Boosted_Model	0.7867	0.8621	0.7526	0.7874	0.7826

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DecisionTree

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

Confusion matrix of Logistic_Regression_Model

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

Confusion matrix of Random_Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Based on validation against the current models, the highest accuracy goes to Random forest model. Therefore, we are using Random Forest Model to test the 500 new loan applicants

LOGISTIC REGRESSION

Type II Analysis of Deviance Tests

Response: Credit.Application.Result

	LR Chi-Sq	DF	Pr(>Chi-Sq)
Account.Balance	31.129	1	2.41e-08***
Payment.Status.of.Previous.Credit	5.687	2	0.05823.
Purpose	12.225	3	0.00665**
Credit.Amount	9.882	1	0.00167**
Length.of.current.employment	5.522	2	0.06324.
Instalment.per.cent	5.198	1	0.02261*
Most.valuable.available.asset	3.509	1	0.06104.

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

Basic Diagnostic Plots

DECISION TREE

Summary Report for Decision Tree Model X

Call:

rpart(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 1e-05)

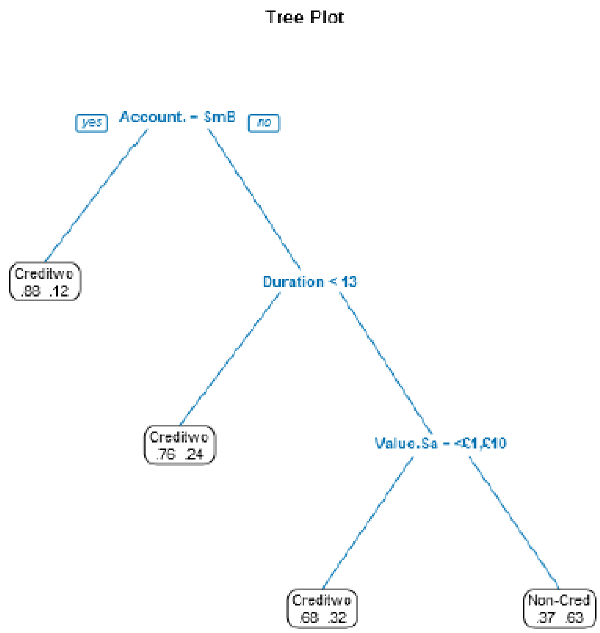
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

Pruning Table

Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

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) *

Plots


Tree Plot
yes
Account. = $mB
no
Creditwo
.88 .12
Duration < 13
Creditwo
.76 .24
Value.Sa = <€1,€10
Creditwo
.68 .32
Non-Cred
.37 .63

FOREST MODEL

Basic Summary

Call:

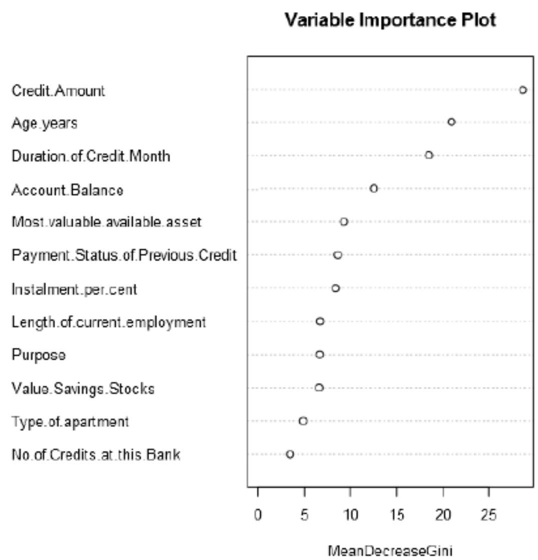
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 500)

Type of forest: classification

Number of trees: 500

Number of variables tried at each split: 3

OOB estimate of the error rate: 36.6%



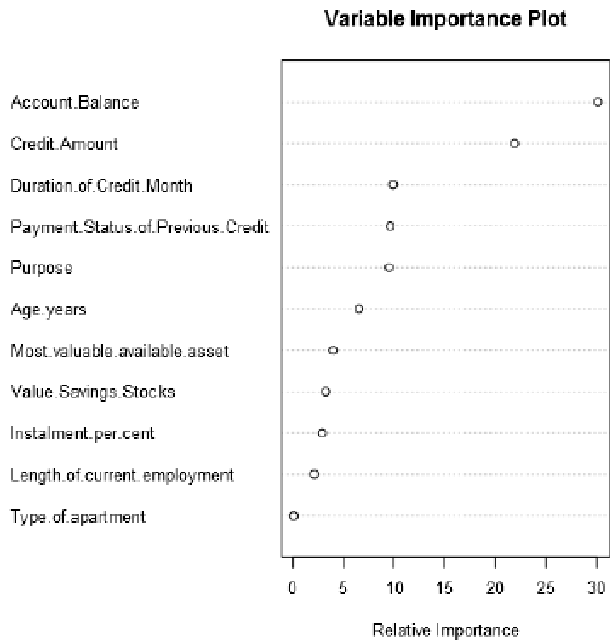
BOOSTED MODEL

Report for Boosted Model X

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2377

Plots:



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

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

2. How many individuals are creditworthy?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Random_Forest_Model	0.8000	0.8718	0.7426	0.7907	0.8571

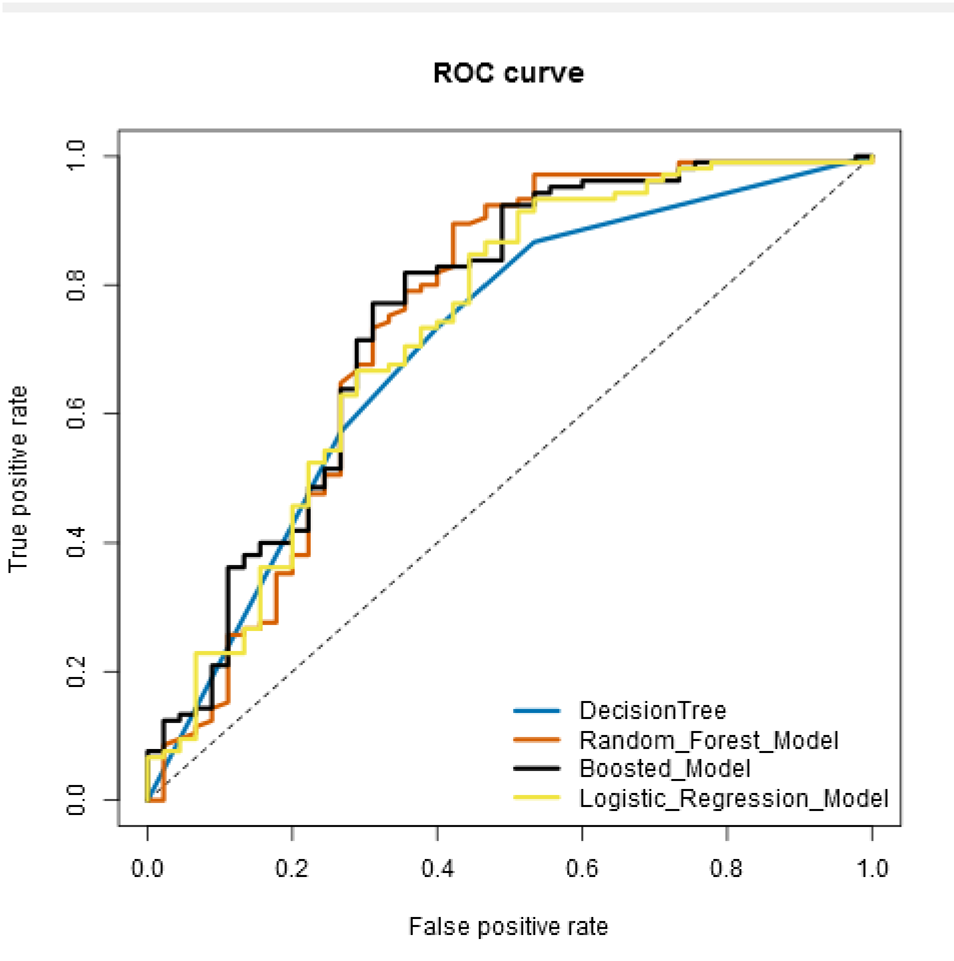
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Random_Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Using random forest model, the accuracy is 80% against validation data set
With 79.07% Accuracy within “Creditworthy”
And 85.71% Accuracy within “Non-Creditworthy”

The ROC graph is as follow:



And using formula , if possibility of creditworthy > possibility non-creditworthy , then it is considered credit worthy, we will get **413** approved clients

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.