Project: Creditworthiness

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?

  A small bank has an influx of new people applying for loans and management needs to decide if these applicants are creditworthy.

- What data is needed to inform those decisions?

  In order to inform this decision, we need to know their bank account balance, prior credit payment status, employment status, savings and other assets, debt, current loan amount request and income.

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

  We need to decide whether an applicant is creditworthy, hence we need a Binary 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".

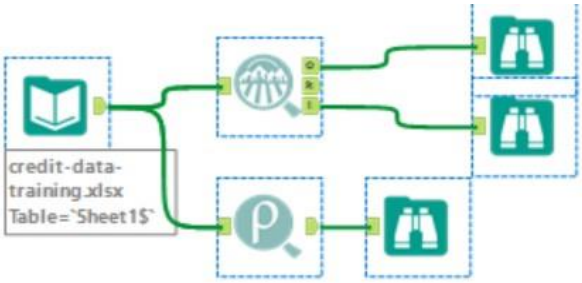
- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed

- 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.

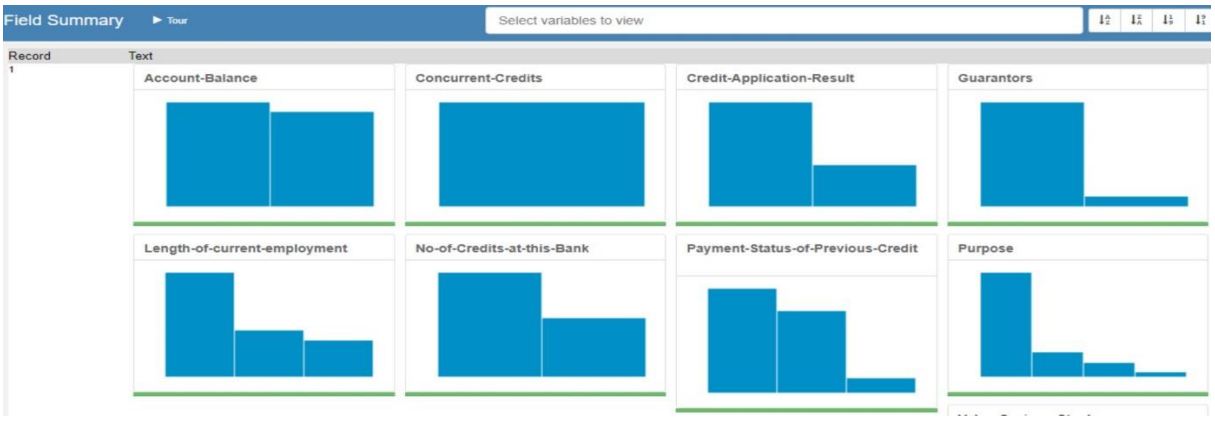
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

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

In order to analyze the data for any missing fields and correlation, I created the workflow below.

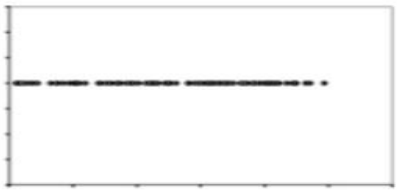


First, I looked at the field summary report. From this report, I was able to identify that the Duration-in-Current-address field was missing 69% of its data and the Age-years field was missing 2% of its data. I decided to remove Duration-in-Current-address field because of the large amount of missing data. Additionally, I imputed the Age-years field with median age of 33. Moreover, I decided to remove the Occupation, Foreign Worker, Number of dependents, Guarantors and Concurrent – Credits fields because of Low variability. See report below and next page.



Field Summary for Occupation (Low Variability)

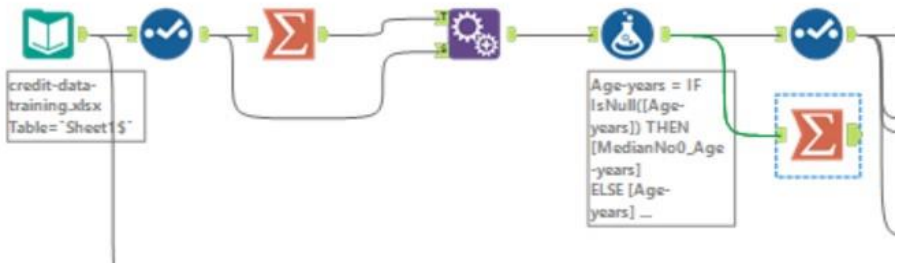
Name	Field Category	Min	Max	Median	Std. Dev.	Percent Missing	Unique Values	Mean	Remarks
Occupation	Numeric	1	1	1	0	0	1	1	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".



The Pearson Correlation tool showed that none of the numeric data fields have an association greater than 0.70. Hence, they are not highly correlated. See below.

FieldName	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Duration-in-Current-address	Most-valuable-available-asset	Age-years	Type-of-apartment	Occupation	No-of-dependents	Telephone	Foreign-Worker
Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269	0.143176	-0.115916
Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986	0.286338	0.025493
Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894	0.029354	-0.133411
Duration-in-Current-address	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]
Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454	0.203509	-0.146005
Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738	0.101443	-0.089848
Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1	-0.048559	0.065943
Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559	1	-0.055516
Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943	-0.055516	1

I made the changes mentioned on page 2 through the workflow below and confirmed that the clean data set had 13 columns and the average of Age Years is 36. See below and next page.



Average of Age Years = 36.

Sample Data showing 13 columns.

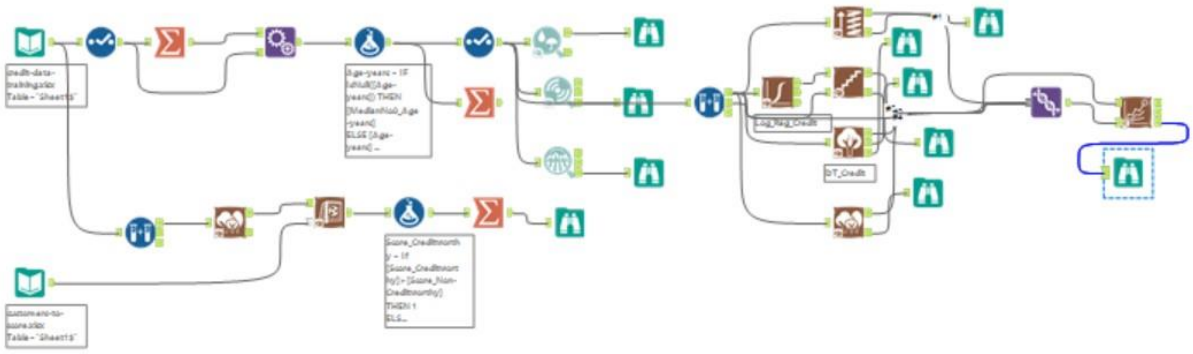
Credit-Application-Result	Account-Balance	Duration-of-Credit-Month	Payment-Status-of-Previous-Credit	Purpose	Credit-Amount	Value-Savings-Stocks	Length-of-current-employment	Instalment-per-cent	Most-valuable-available-asset	Age-years	Type-of-apartment	No-of-Credits-at-this-Bank
Creditworthy	Some Balance	4	Paid Up	Other	1494	£100-£1000	< 1yr	1	1	33	2	1
Creditworthy	Some Balance	4	Paid Up	Home Related	1494	£100-£1000	< 1yr	1	1	29	2	1
Creditworthy	Some Balance	4	No Problems (in this bank)	Home Related	1544	None	1-4 yrs	2	1	42	2	More than 1
Creditworthy	Some Balance	4	No Problems (in this bank)	Home Related	3380	None	1-4 yrs	1	1	37	2	1
Creditworthy	No Account	6	Paid Up	Home Related	343	None	< 1yr	4	1	27	2	1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Related	362	< £100	< 1yr	4	3	52	2	More than 1
Non-Creditworthy	No Account	6	Some Problems	Home Related	433	£100-£1000	< 1yr	4	2	24	1	1
Creditworthy	No Account	6	Paid Up	Home Related	454	None	< 1yr	3	2	22	2	1
Creditworthy	No Account	6	Paid Up	Home Related	484	None	1-4 yrs	3	1	28	2	1
Creditworthy	Some Balance	6	Paid Up	Home Related	660	£100-£1000	1-4 yrs	2	1	23	1	1
Creditworthy	No Account	6	No Problems (in this bank)	Home Related	666	£100-£1000	1-4 yrs	3	1	39	2	More than 1
Creditworthy	Some Balance	6	No Problems (in this bank)	Home Related	700	£100-£1000	4-7 yrs	4	4	36	3	More than 1
Creditworthy	Some Balance	6	Paid Up	Home Related	709	£100-£1000	< 1yr	2	1	27	2	1

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

See Workflow below.



*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.

Logistic Regression

Significant predictor variables for the Logistic Regression Model are Account Balance, Payment Status, Purpose, Credit Amount, Length of Current Employment and Installment Percentage. See below.

Report for Logistic Regression Model Step_Log_Credit

Basic Summary

Call:

glm(formula = Credit.Application.Result ~ Account.Balance + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset, family = binomial("logit"), data = the.data)

Deviance Residuals:

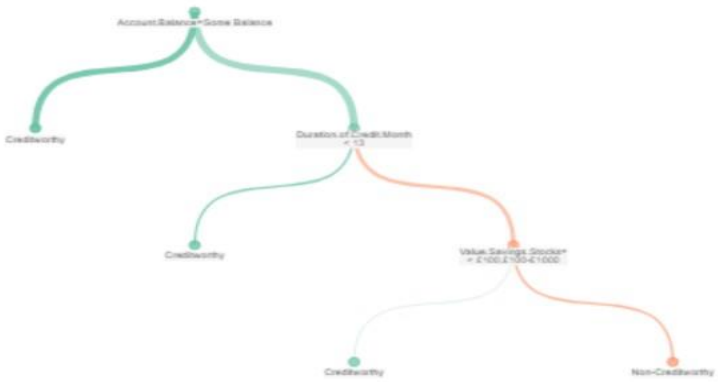
Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

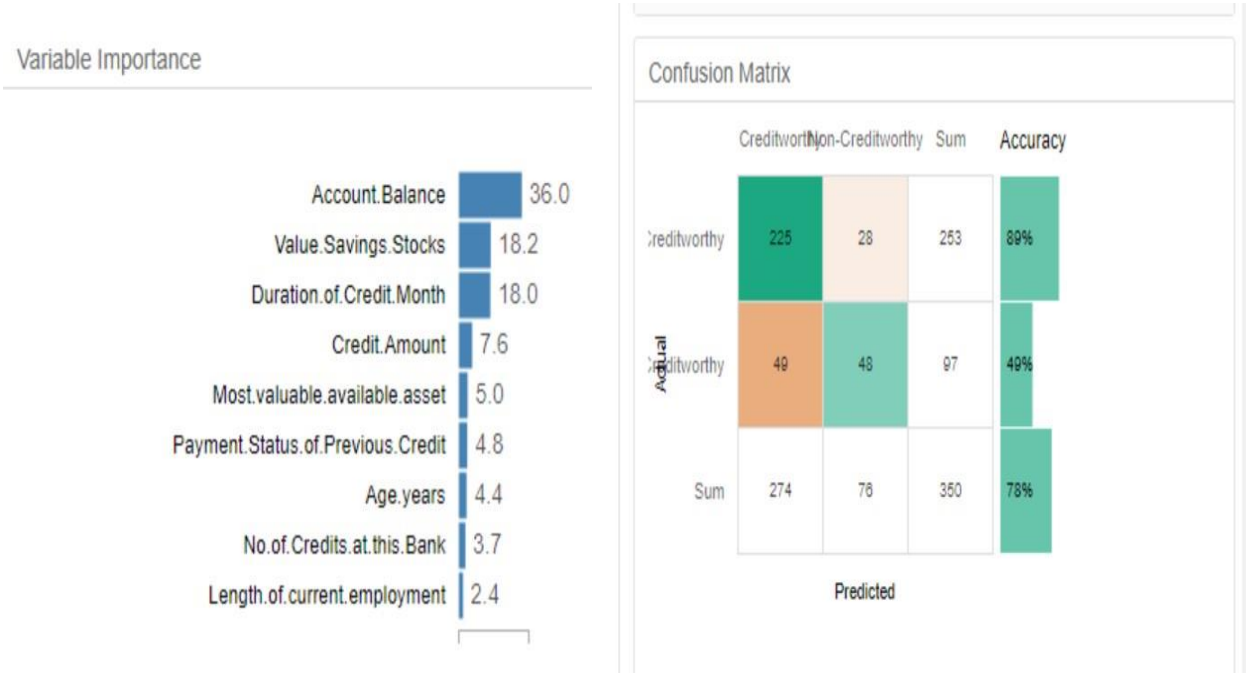
Coefficients:

	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	.

Decision Tree

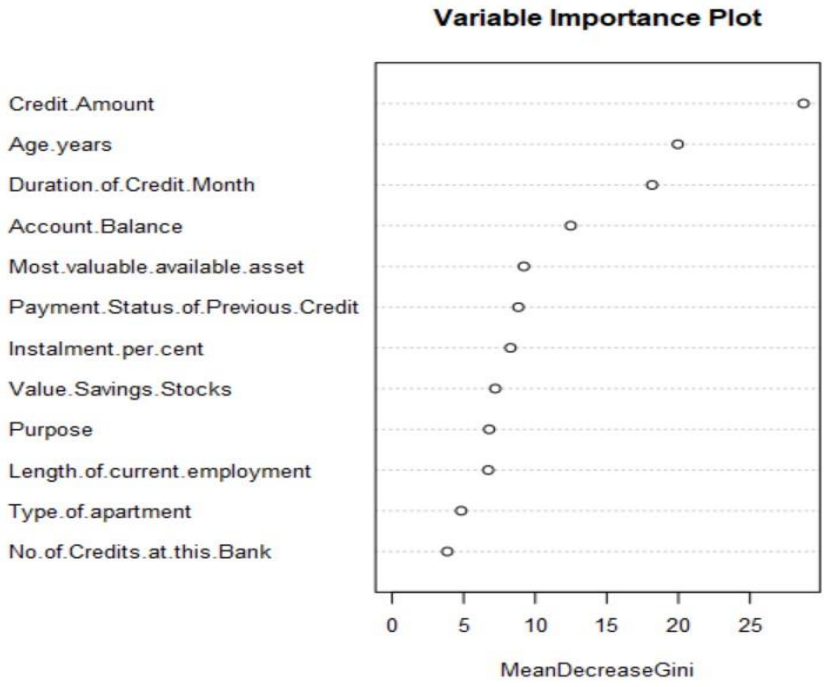
Significant predictor variables for the Decision Tree Model are Account Balance, Duration of Credit in months and Value of Savings Stocks. See below and next page.





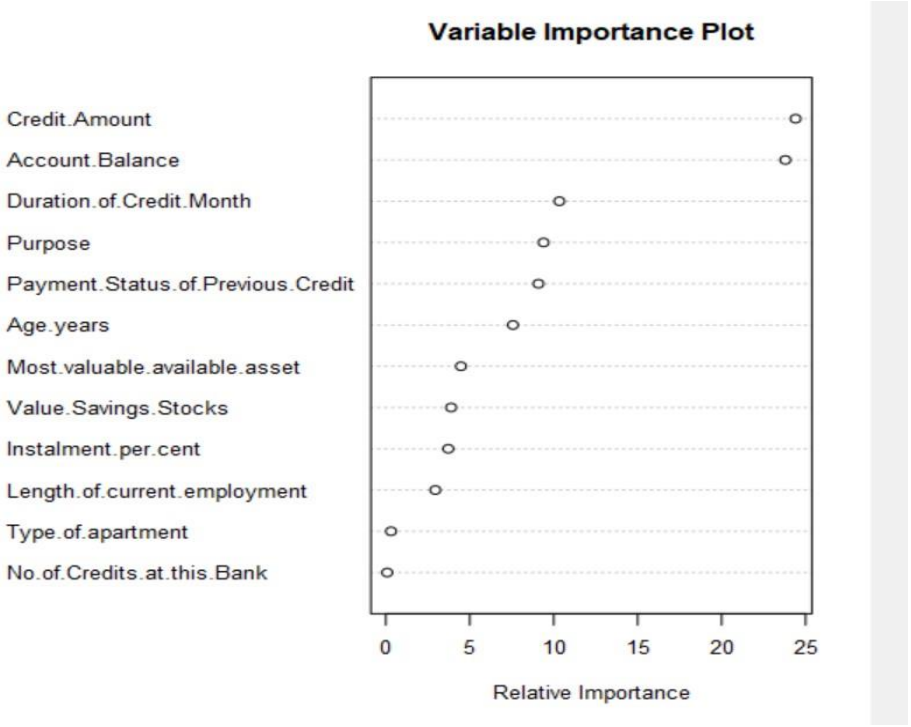
Forest Model

Significant predictor variables for the Forest Model are Credit Amount, Age in years, Duration of Credit in months and Account Balance. See below.



Boosted Model

Significant predictor variables for the Boosted Model are Credit Amount and Account Balance. See below.



- 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?

  I used the Model Comparison tool for accuracy comparison and validation. See results below.
  The Regression and Decision Tree Models have an overall accuracy of 76% and 74.7%. These two models are better at predicting non-creditworthy applicants (48.9%, 46.7%) accuracy vs Forest and Boosted Models (37.8%) accuracy. In comparison the Forest and Boosted Models have an overall accuracy 79.3% and 78.7%. These two models are better at predicting creditworthy clients (97.1%, 96.1%) vs Regression and Decision Tree Models (87.6%, 86.7%).
  I also noticed that all the models are better at predicting creditworthy vs non-creditworthy clients. This is because our sample data has more creditworthy clients vs non-creditworthy clients (72% vs 28%).

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Step_Log_Credit	0.7600	0.8364	0.7306	0.8762	0.4889
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Credit	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Credit	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix of Boosted_Credit

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

Confusion matrix of DT_Credit

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

Confusion matrix of Forest_Credit

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

Confusion matrix of Step_Log_Credit

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

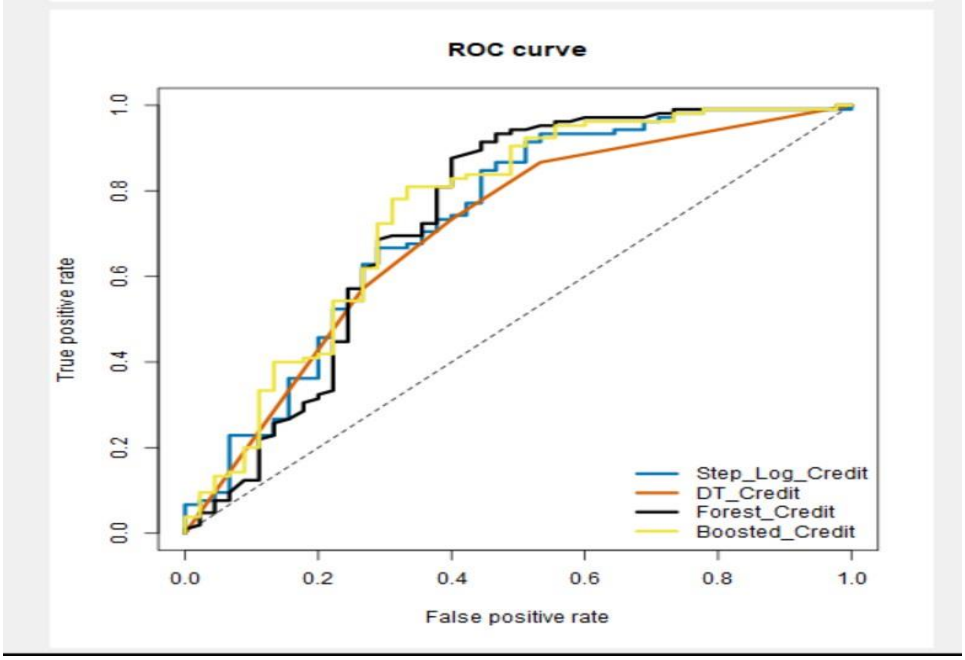
Step 4: Writeup

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

  Consider the ROC graph below.



From this chart, we can see that the Forest and Boosted Model perform better than the Decision Tree and Regression Models. Additionally, the AUC for the Boosted Model is 0.75.
The Confusion Matrix (prior page) shows that the Boosted and Forest Models classify more Non-Creditworthy clients as Creditworthy while the Regression and Decision Tree Models classify more Creditworthy clients as Non-Creditworthy.
Since our boss is only concerned with model accuracy, we should use the Forest Model which has an overall accuracy of 79.3%.

- How many individuals are creditworthy?

  I used the workflow below to score the model and according to this model the number of creditworthy applicants is 408.