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**?

The bank has an influx of new loan applications and decisions need to be made about which loan applicants should be approved or declined. To make this decision a predictive model is required and therefore decisions need to be made about which predictor variables to include to ensure the most accurate model is used to evaluate each loan applicant.

- **What data is needed to inform those decisions**?

To evaluate the loan applicants, personal financial data is needed such as current loans and assets, income, savings and past credit information.

To make the predictive model data is needed on current and past loan customers with all relevant financial information which could be a relevant in predicting credit worthiness such as savings, credit amounts, lengths and payments of loans, types of loans, age.

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

A binary classification model is required as we have a binary decision to be made – should the applicant be approved for a loan 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.***

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

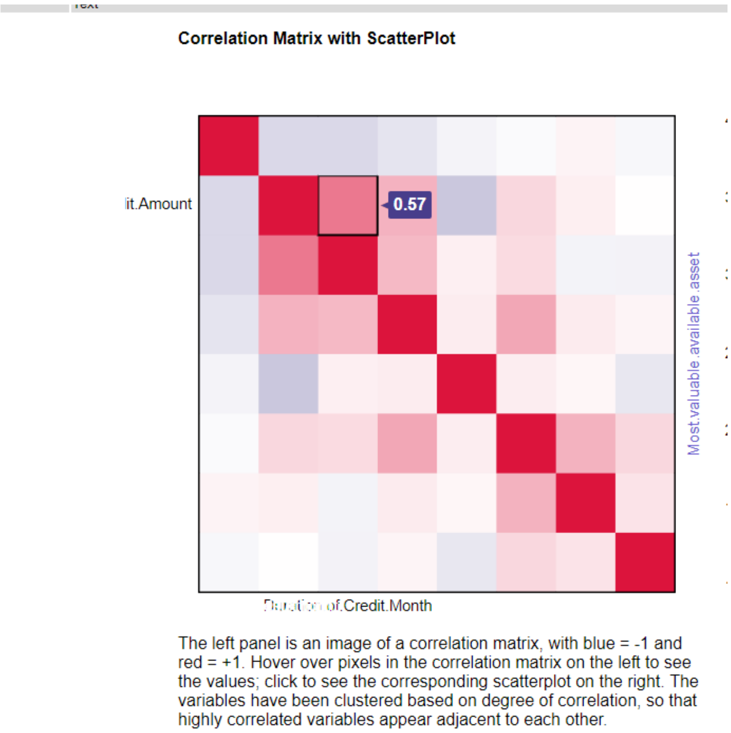
Foreign-Worker	Double

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**.

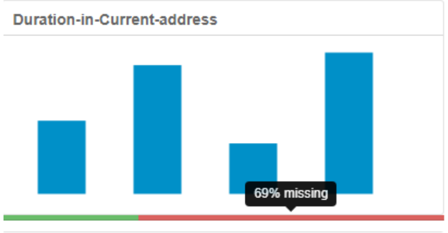
A correlation scatter plot was run and shown below. There were no highly correlated numerical data fields, with the highest correlation being 0.57 as shown. Therefore, no fields were removed based on correlation.


Correlation Matrix with ScatterPlot
it.Amount
0.57
Most.valuable.available.asset
Duration.of.Credit.Month

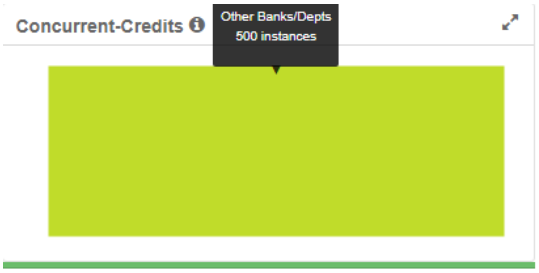
The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

The Field Summary tool showed the output below and resulting in the following actions -

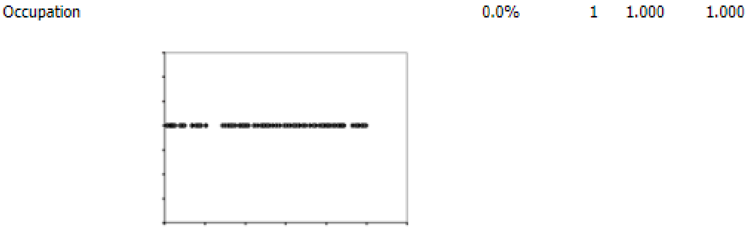
- Found Age has 12 Null values and imputed the Median age
- Removed Telephone as it is not a logical variable
- Removed Duration In Current Address as 344 of the 500 were Null values and therefore too many to impute



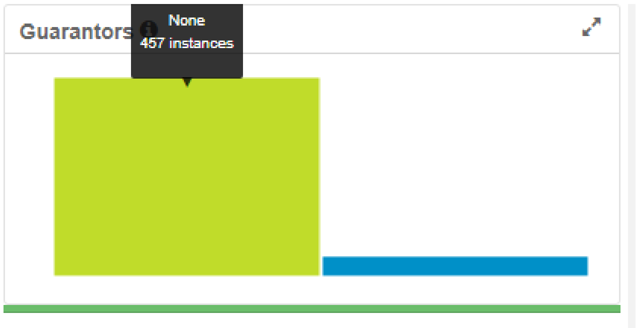
- Removed Concurrent-Credits as it only held one value



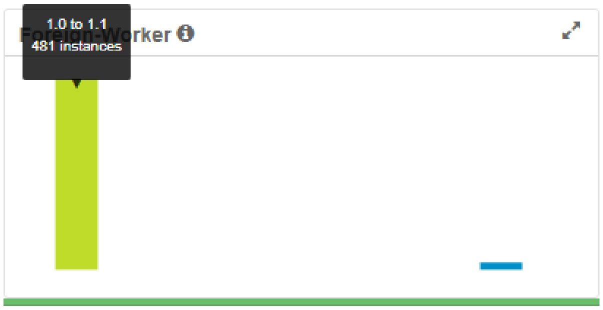
- Removed Occupation as it only held one value


Occupation 0.0% 1 1.000 1.000

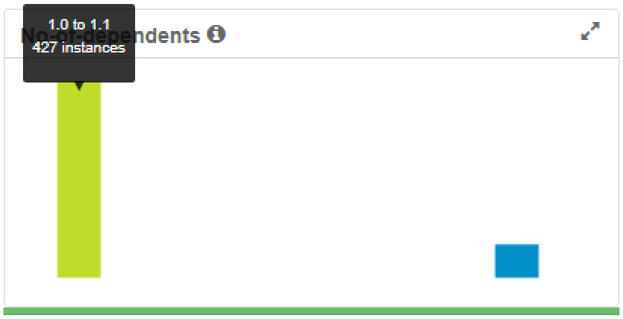
- Removed Guarantors due to low variability of the data



- Removed Foreign Worker due to low variability of the data



- Removed Dependants due to low variability of the data



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

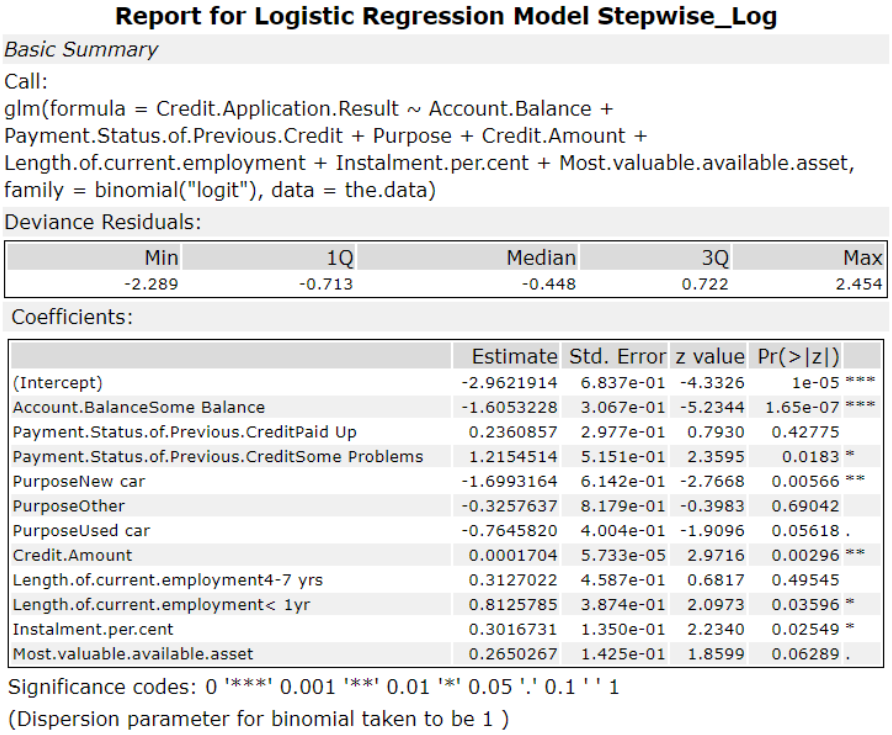
*Answer these questions for **each model** you created:*

<u>Logistic Regression</u>

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

I ran the logistic regression and then added the Stepwise tool. The results are below and show the most significant variables by low p values and stars for statistical significance. These are

- Account balance – most significant with ***
- Purpose of loan - **
- Credit amount - **
- Payment status of previous credit
- Length of current loan
- Instalment per cent
- Most valuable available asset



Report for Logistic Regression Model Stepwise_Log

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

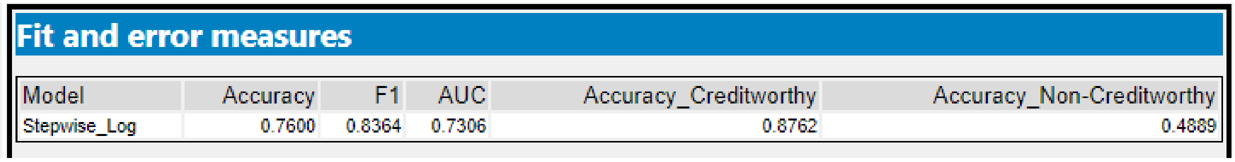
Coefficients:

	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	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

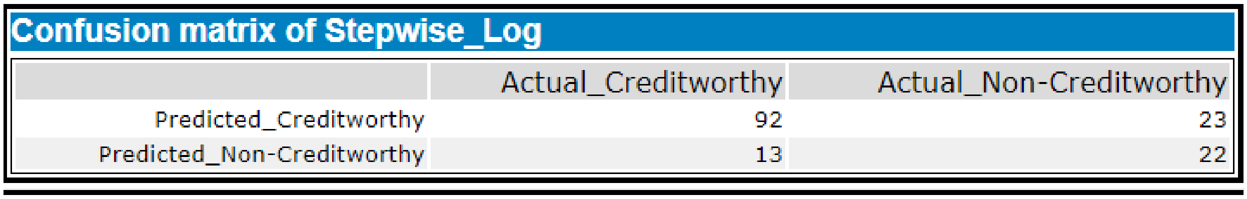
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

The overall accuracy of the model is 76%. The accuracy of predicting Creditworthy correctly was 87.6% and predicting Non-Creditworthy is 49%. This shows a bias toward categorising 1 out of 2 people as Creditworthy instead of Non-Creditworthy. Confusion matrix supports this – 23 out of 45 predicted incorrectly.



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Log	0.7600	0.8364	0.7306	0.8762	0.4889



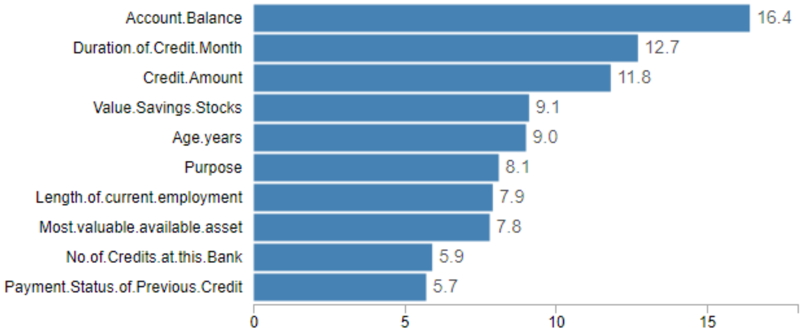
Confusion matrix of Stepwise_Log

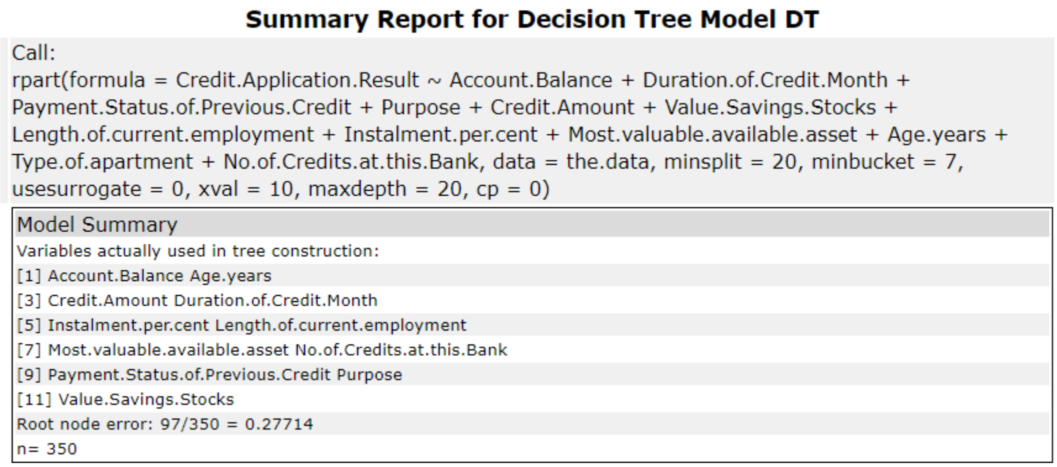
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Decision Tree

- 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 most significant predictor variables were Account Balance, Duration of Credit and Credit Amount. Shown below.





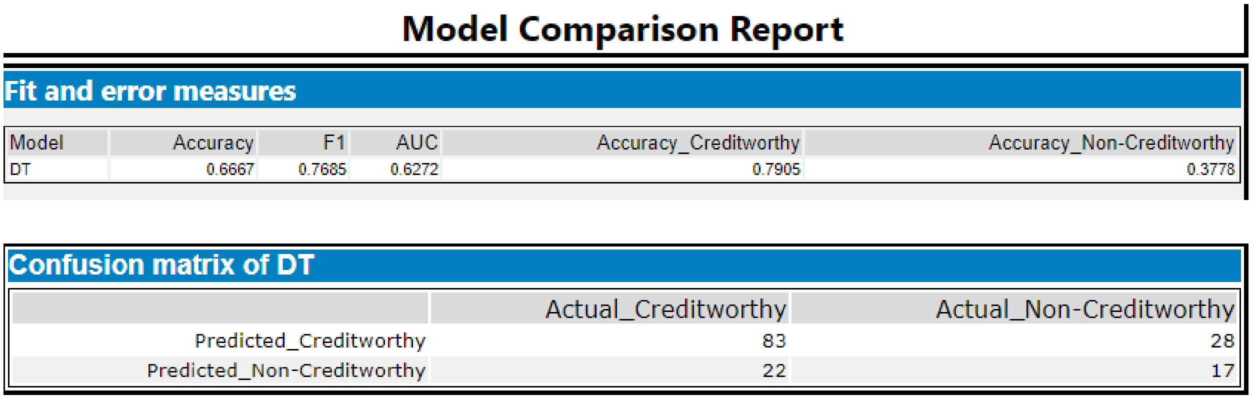
Summary Report for Decision Tree Model DT

```
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 = 0, xval = 10, maxdepth = 20, cp = 0)
```

```
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age.years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

- 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?

Using the model comparison tool, the overall accuracy of the model was 66.7%. Accuracy to predict creditworthiness was 79% and accuracy to predict non-creditworthiness was low – 37.8% - even lower than the decision tree confusion matrix. This is again showing a strong bias for the model to incorrectly predict non-creditworthy as creditworthy. Confusion matrix supports this – 28 out of 45 predicted incorrectly.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.6667	0.7685	0.6272	0.7905	0.3778

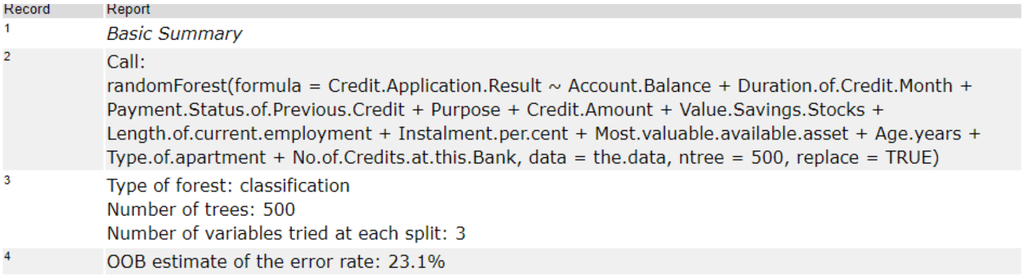
Confusion matrix of DT

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

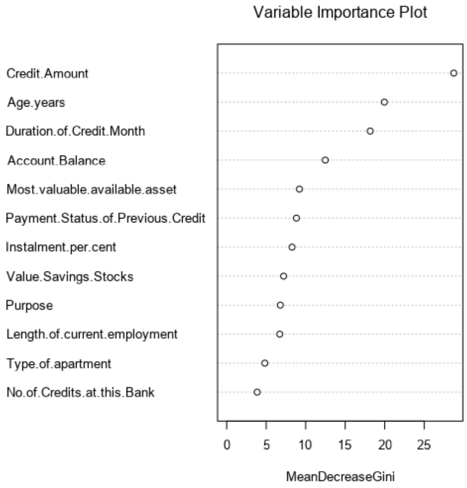
Forest Model

- 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 most important variables are Credit amount, age and duration of credit. These are shown below.

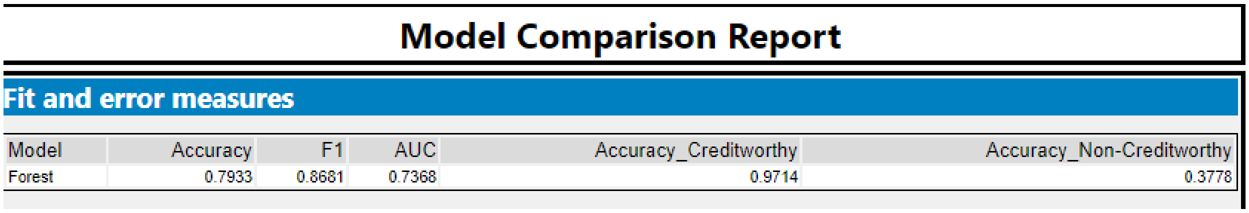


Record	Report
1	*Basic Summary*
2	Call: randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + Length.of.current.employment + Instalment.per.cent + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, data = the.data, ntree = 500, replace = TRUE)
3	Type of forest: classification Number of trees: 500 Number of variables tried at each split: 3
4	OOB estimate of the error rate: 23.1%



- 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?

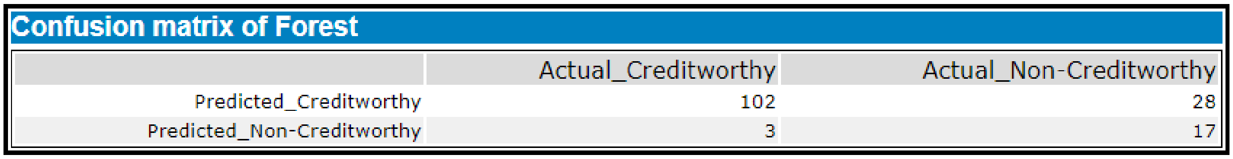
Using the model comparison tool, the overall accuracy of the model is 79%.
The accuracy of predicting creditworthy cases was high – 97%. However, the accuracy of predicting non-creditworthy was low – 38%. This is showing the model has a bias for incorrectly predicting non-creditworthy cases as creditworthy. Confusion matrix supports this – 28 out of 45 predicted incorrectly.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	0.7933	0.8681	0.7368	0.9714	0.3778



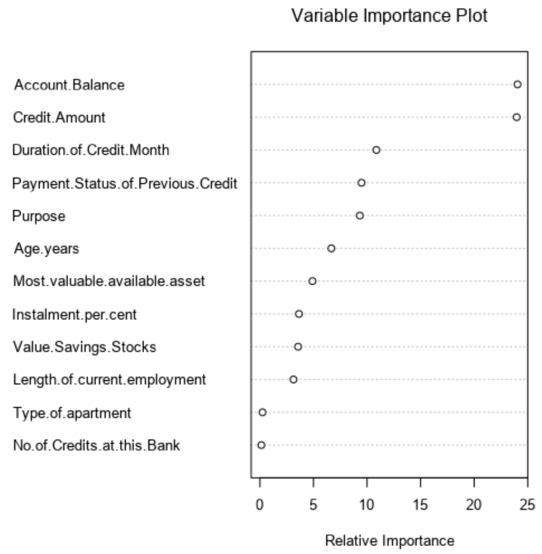
Confusion matrix of Forest

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

Boosted Model

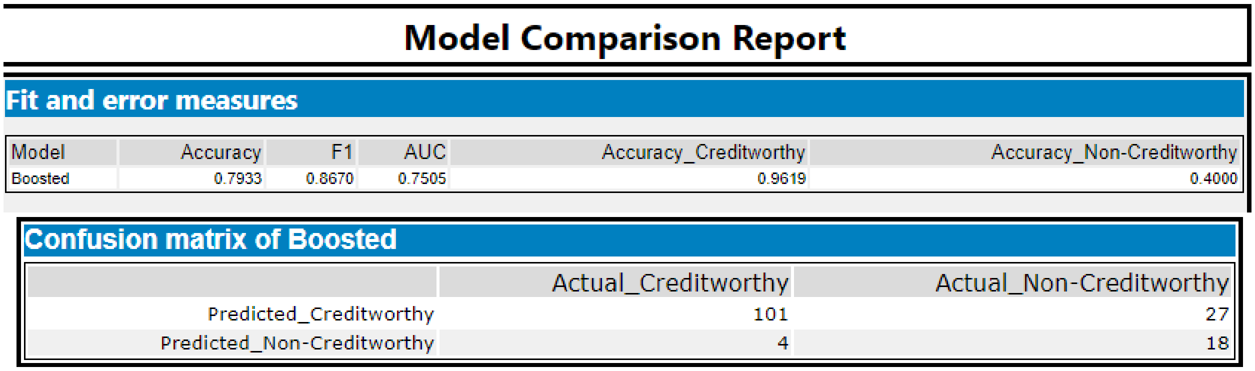
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

The most significant variables are account balance and credit amount, shown 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?

Using the model comparison tool, the overall accuracy of the model is 79%.
The accuracy of predicting creditworthy cases is high – 96%. The accuracy of predicting non-creditworthy cases is 40%. This shows a bias to predict a non-creditworthy case as creditworthy incorrectly 60% of the time using the sample. Confusion matrix supports this – 27 out of 45 predicted incorrectly.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted	0.7933	0.8670	0.7505	0.9619	0.4000

Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

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"

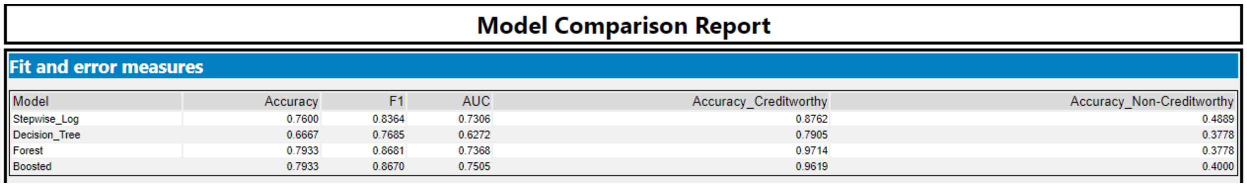
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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**

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

Running the 4 models into the model comparison tool produces the following results. All models have an overall high accuracy rate but a low accuracy rate for predicting non-creditworthy cases – all under 50% accuracy.

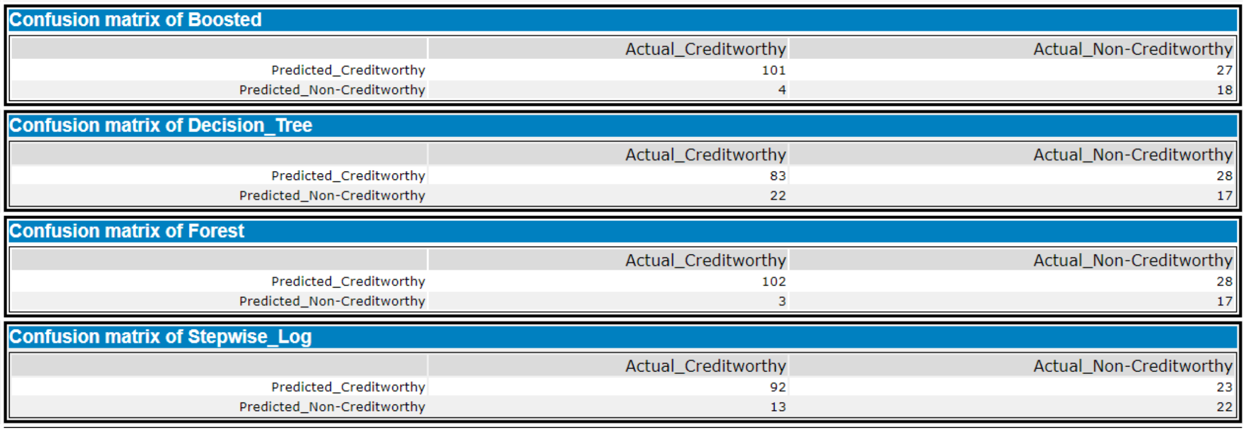


Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Log	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	0.7933	0.8670	0.7505	0.9619	0.4000

Looking at the confusion matrix output below, the low levels of accuracy in predicting non-creditworthy cases are demonstrated in the incorrect numbers and therefore bias for predicting non-creditworthy cases as creditworthy. The Confusion matrix supports this with the corresponding numbers of cases being incorrectly predicted Creditworthy when they were Non-Creditworthy.



Confusion matrix of Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of Decision_Tree

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

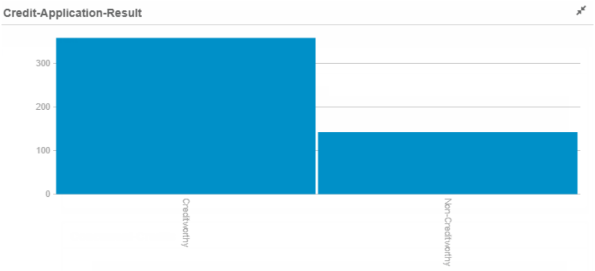
Confusion matrix of Forest

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

Confusion matrix of Stepwise_Log

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

It is noted the low levels of non-creditworthy cases in the sample could be leading to the low levels of accuracy in predicting the Non-Creditworthy cases and the outlined bias in the predictability of the model.



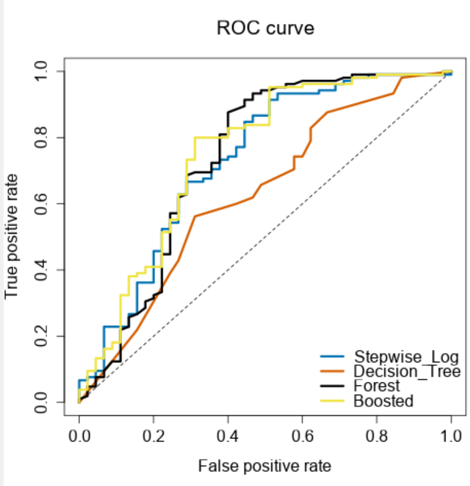
The Forest model and the Boosted model are the most accurate overall – both 79% overall accuracy.

The Forest model predicts Creditworthy accurately 97% and Non-Creditworthy accurately 38%.

The Boosted model predicts Creditworthy accurately 96% and Non-Creditworthy accurately 40%.

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest	79%	87%	74%	97%	38%
Boosted	79%	87%	75%	96%	40%

Looking at the ROC graph below, both the Forest and Boosted model have similar lines. However, the Boosted model Area Under Curve is 75% and Forest is 74%.



Based on the above information I have decided to use the Boosted model due to the overall accuracy, high accuracy for Creditworthy, slightly higher AUC and the slightly higher accuracy for non-Creditworthy than the Forest Model.

- How many individuals are creditworthy?

Using the Boosted model on the customers to predict 441 (88%) of the loan applications would be deemed Creditworthy.

Creditworthy	Non-Creditworthy	Total
441	59	500
88%	12%	

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.