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 key decision point is to identify whether customers who applied for loan are creditworthy or not. Basis this the required loan can be extended to them.

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

Parameters like Credit Amount, Age_years and Duration of Credit Month are required to help predict.

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

We would be using Binary classification models to analyze and determine credit worthiness. These include:

1. Logistics regression
2. Decision tree
3. Forest model
4. Boosted tree

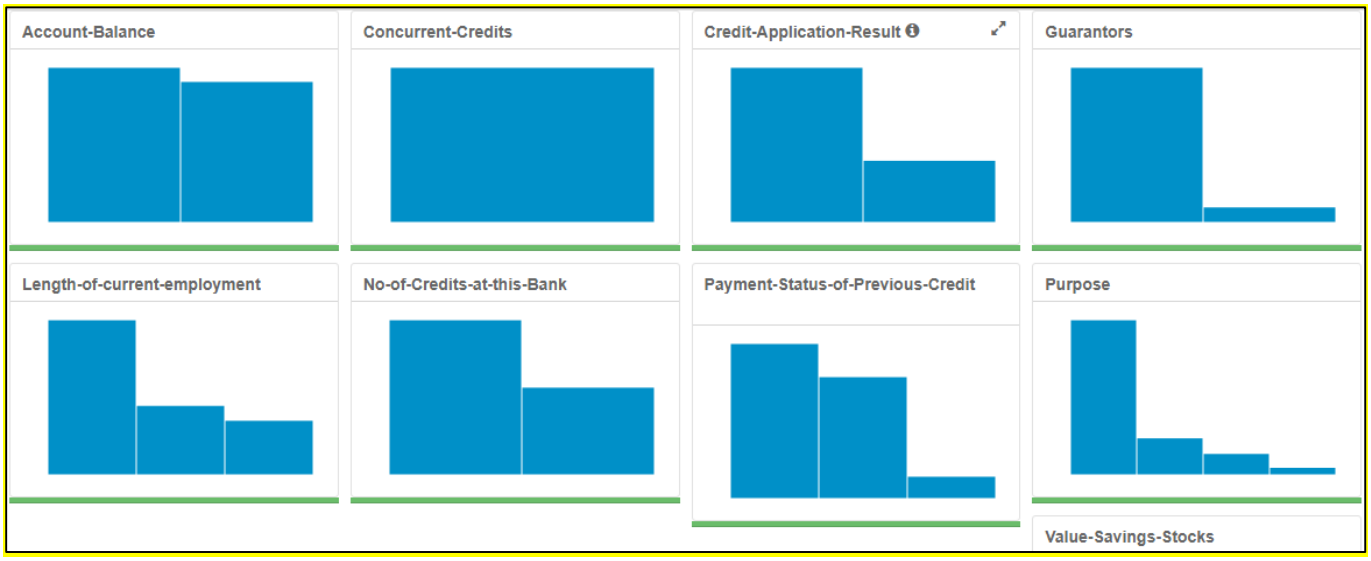
Step 2: Building the Training Set

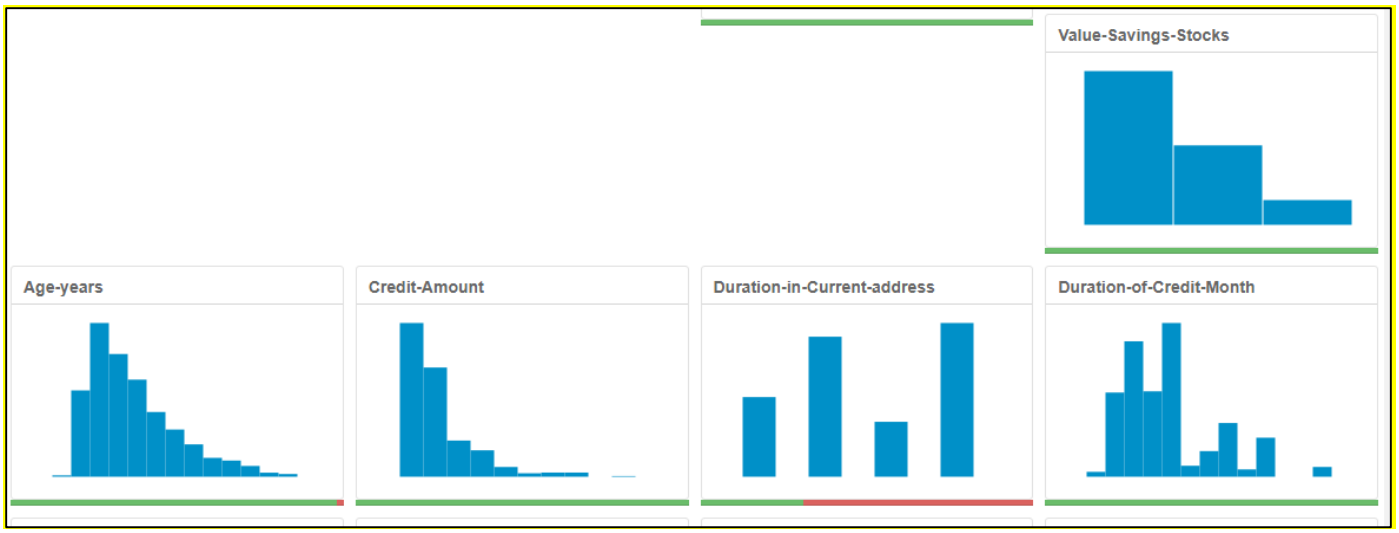
Answer this question:

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

I have used the Field Summary reports to analysis each column. Basis the graphs added below for reference:

1. Fields to Remove:
    a. Having Only 1 Value: Concurrent Credits and Occupation
    b. Large amount of missing data: Duration in Current Address has 69% missing data
    c. Low Variability: Guarantors, Foreign Worker and No of Dependents- These have more than 80% of the data having a single value
    d. Telephone Field- Being not relevant to the credit- worthy decision
2. Fields to impute:
    a. Age_Years has 2% missing data. Here I have imputed the missing values by the median age. Mean has not been used as the data is skewed to the left as per the graphs below.




Value-Savings-Stocks
Age-years
Credit-Amount
Duration-in-Current-address
Duration-of-Credit-Month


Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset
No-of-dependents
Telephone
Type-of-apartment

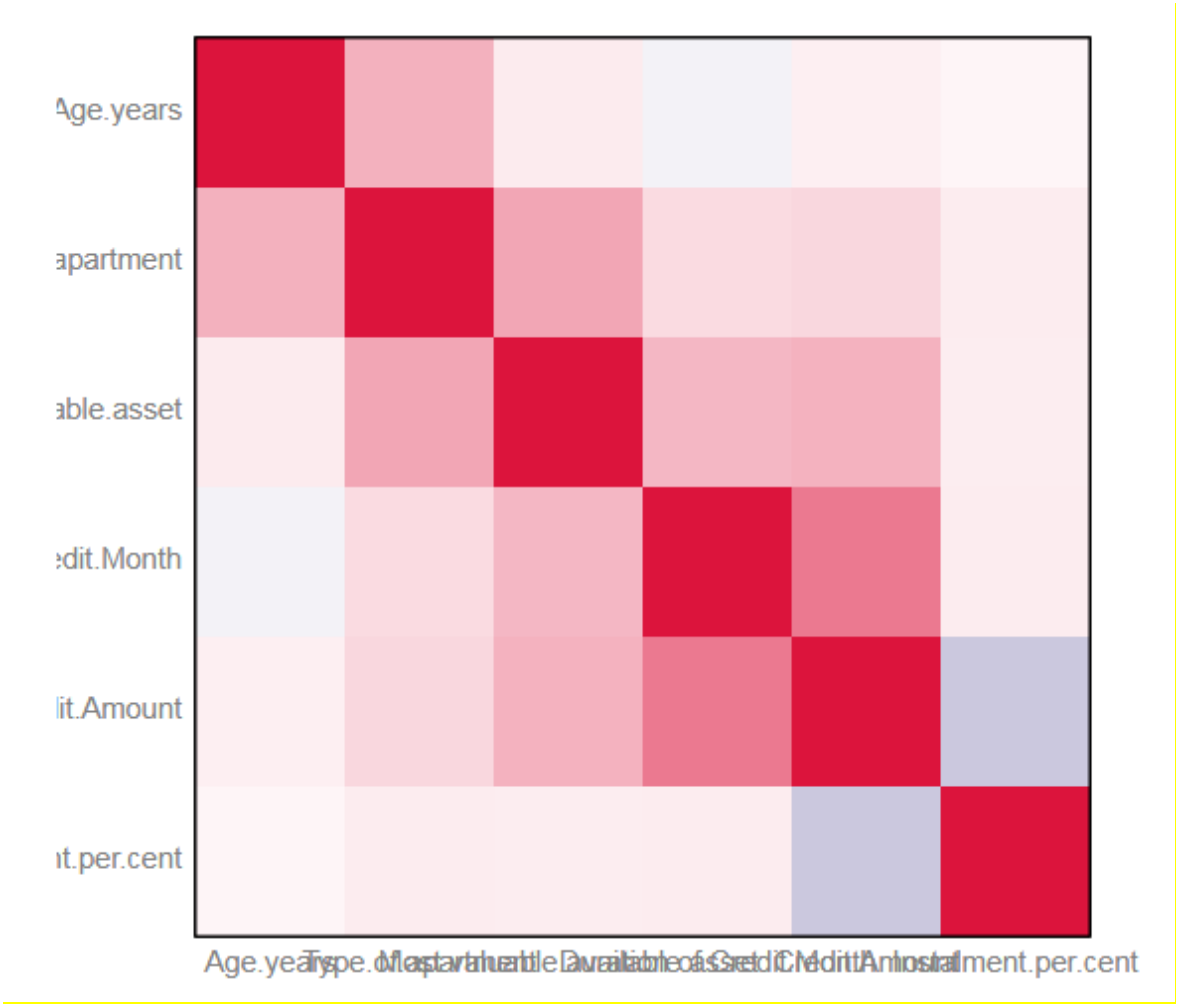
Step 3: Train your Classification Models

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

Answer these questions for each model you created:

I have performed an association analysis on the numerical variable. Basis the below graph, we can conclude that there are no variables which are highly correlated with each other. High co-relation here being co-relation more than 0.7.

Co-relation Matrix with Scatterplot



a) Logistic Regression:

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

Using *Credit Application Result* as the target variable, the 3 most significant variables area as below:

- *Account Balance*
- *Purpose*
- *Credit Amount*

These variables have p-value of less than 0.05.

Basic Summary

Call:

glm(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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age_years, family = binomial(logit), data = the.data)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.088	-0.719	-0.430	0.686	2.542

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812	*
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519	**
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206	
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733	.
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989	**
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361	
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642	
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934	
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925	*
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262	*
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621	*
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786	
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275	
Age_years	-0.0141206	1.535e-02	-0.9202	0.35747	

- 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?

- Overall accuracy is around 78.0%
- Accuracy for creditworthy is higher than non-creditworthy at 90.0% and 49.0% respectively
- The model is biased towards predicting customers as non-creditworthy

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_CW	0.7800	0.8520	0.7314	0.9048	0.4889

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of LR_CW

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

b) 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 Important Variables are as below:

- Account Balance
- Value Savings Stocks
- Duration of Credit Month

- 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 accuracy is as below:

- Overall Accuracy- 75%
- Creditworthy- 87%
- Non- creditworthy- 47%
- The model is biased towards predicting customers as non-creditworthy

Model Comparison Report

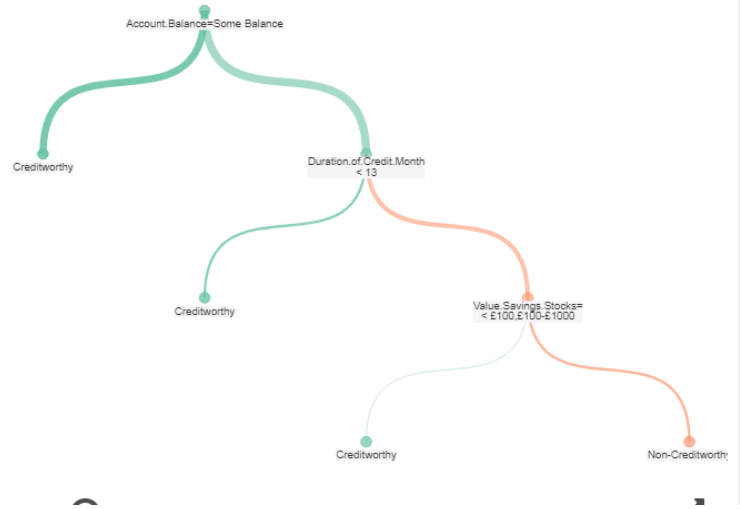
Fit and error measures

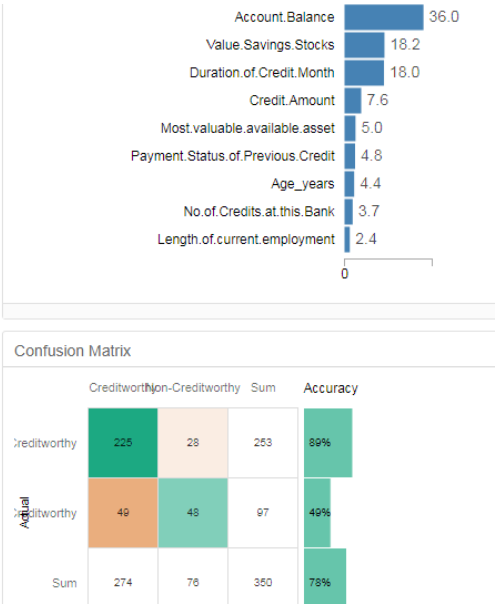
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecisionT_CW	0.7467	0.8273	0.7054	0.8667	0.4667

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of DecisionT_CW

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





c) Forest Model- Overall accuracy is 80%

- 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 as below:

- Credit Amount
- Age_years
- Duration of Credit Month

- 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 accuracy is as below:

- Overall Accuracy- 80%
- Creditworthy- 96%
- Non- creditworthy- 42%
- The model is biased towards predicting customers as non-creditworthy.

Model Comparison Report

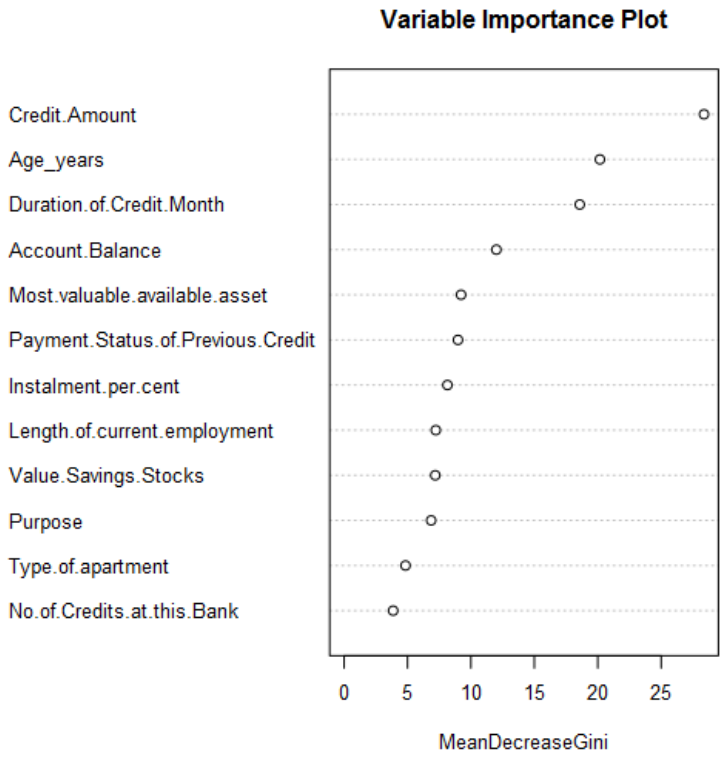
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
ForestM_CW	0.8000	0.8707	0.7361	0.9619	0.4222

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of ForestM_CW

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19



d) Boosted Model- Overall accuracy is 78%

- 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 as below:

- Account Balance
- Credit Amount
- Payment status of previous credit
- Duration of Credit Month
- Purpose

- 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 accuracy is as below:

- Overall Accuracy- 80%
- Creditworthy- 96%
- Non- creditworthy- 42%
- The model is biased towards predicting customers as non-creditworthy.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BoostedModel_CW	0.7867	0.8632	0.7524	0.9619	0.3778

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BoostedModel_CW

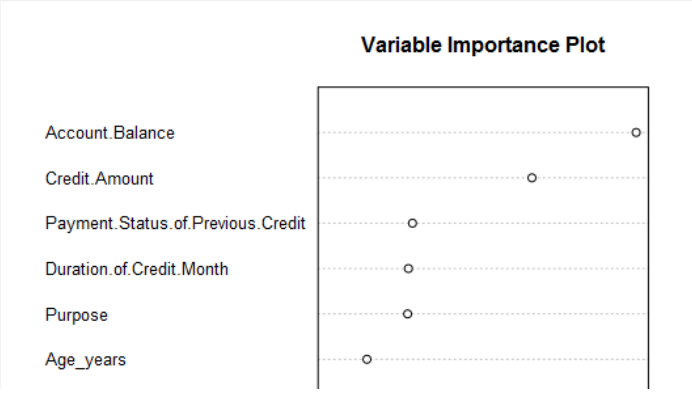
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Report for Boosted Model BoostedModel_CW

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:



Step 4: Writeup

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

Below is model comparison:

1. Overall Accuracy is highest for Forest Model
2. In Accuracy for Creditworthy- highest is for Forest and Boosted
3. In Accuracy for Non-Creditworthy- highest is for Linear Regression

Basis above, the **forest model** has been chosen as it offers the highest accuracy of 80%.

Fit and error measures

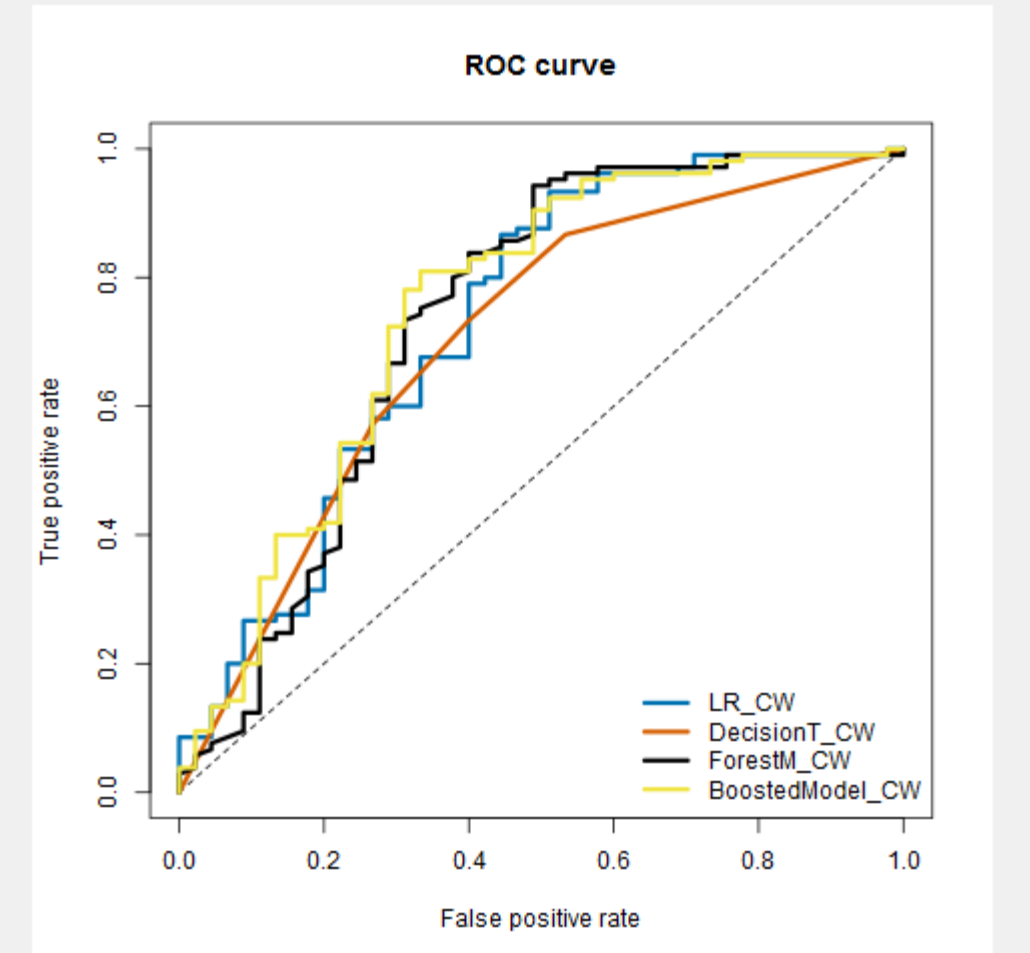
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_CW	0.7800	0.8520	0.7314	0.9048	0.4889
DecisionT_CW	0.7467	0.8273	0.7054	0.8667	0.4667
ForestM_CW	0.8000	0.8707	0.7361	0.9619	0.4222
BoostedModel_CW	0.7867	0.8632	0.7524	0.9619	0.3778

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] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of BoostedModel_CW

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

The ROC curve comparison is as below: Boosted model appears to have the highest area under the graph.



- How many individuals are creditworthy?

Basis the scoring of the forest model, 406 customers are creditworthy.
.