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?
  Using a model we need to identify the creditworthy customer who are eligible for a loan.

- What data is needed to inform those decisions?
  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

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

  Binary model is used to make the decision, as we have to predict either to give the loan or refuse the loan to the customer.

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.

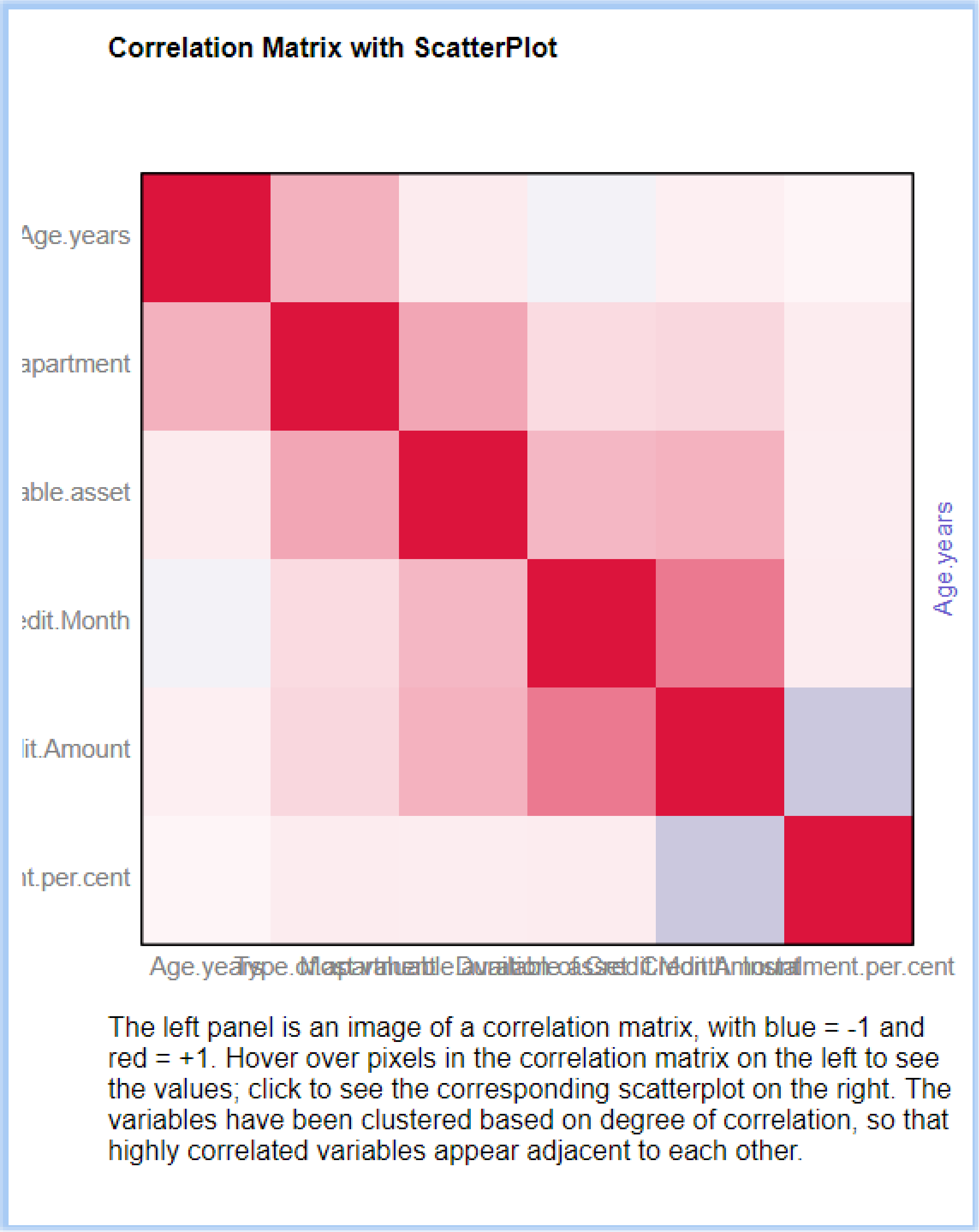
Answer this question:

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

- **“Duration in Current Address” has 68.8% missing data and hence is removed.**
- **The telephone number is also removed as per the question guidelines.**
- **Guarantors, Concurrent Credits, Occupation, No of Dependents, Foreign worker are also removed as some of these values don’t add any significance and have low variability also. Hence they are dropped from the model.**

- **I imputed AGE-years field. It has 2.4 per cent values missing, it was imputed using the median rather than mean and mode as it was less affected by outliers. I did not remove the missing data fields as it may have affected the quality of the model.**

- **Correlation matrix does not show any association of 0.70 or above:**



Correlation Matrix with ScatterPlot

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.

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

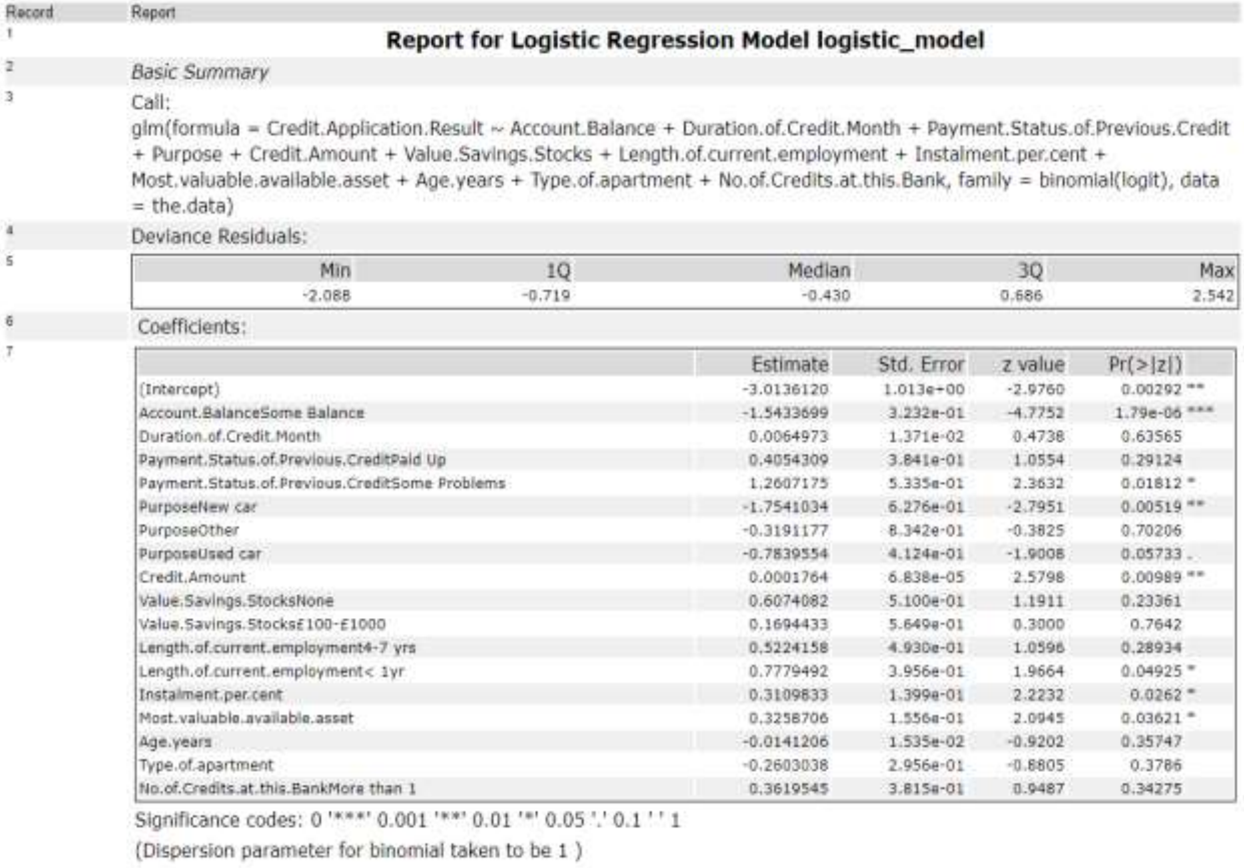
Answer these questions for ***each model*** *you created:*

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

- 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



Record Report

1 **Report for Logistic Regression Model logistic_model**

2 *Basic Summary*

3 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 + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank, family = binomial(logit), data
= the.data)
```

4 Deviance Residuals:

5

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

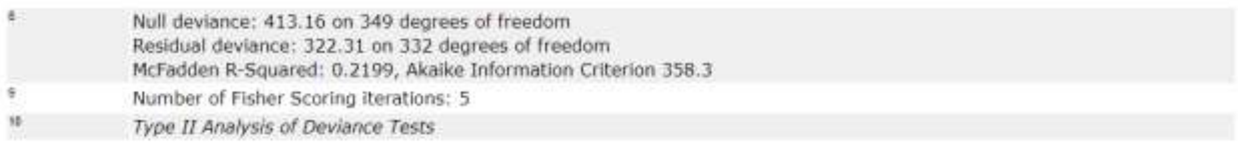
6 Coefficients:

7

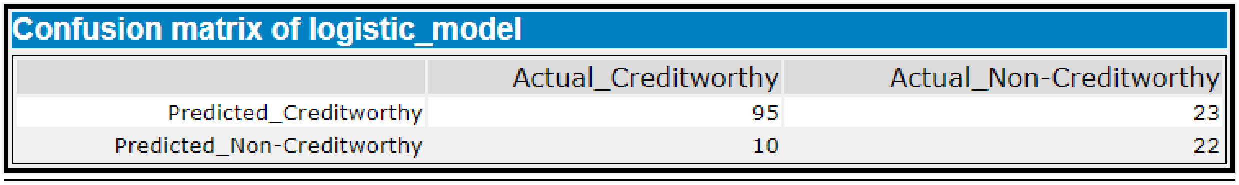
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome 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	*
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747	
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	

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

(Dispersion parameter for binomial taken to be 1)

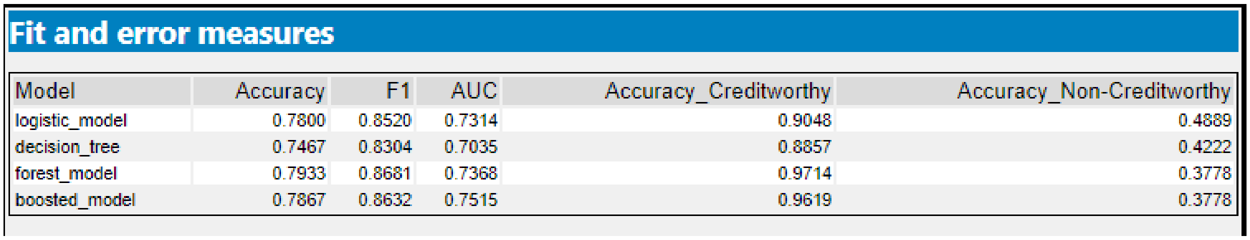


```
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.31 on 332 degrees of freedom
McFadden R-Squared: 0.2199, Akaike Information Criterion 358.3
Number of Fisher Scoring iterations: 5
Type II Analysis of Deviance Tests
```



Confusion matrix of logistic_model

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



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_model	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7515	0.9619	0.3778

Predictor Variables:

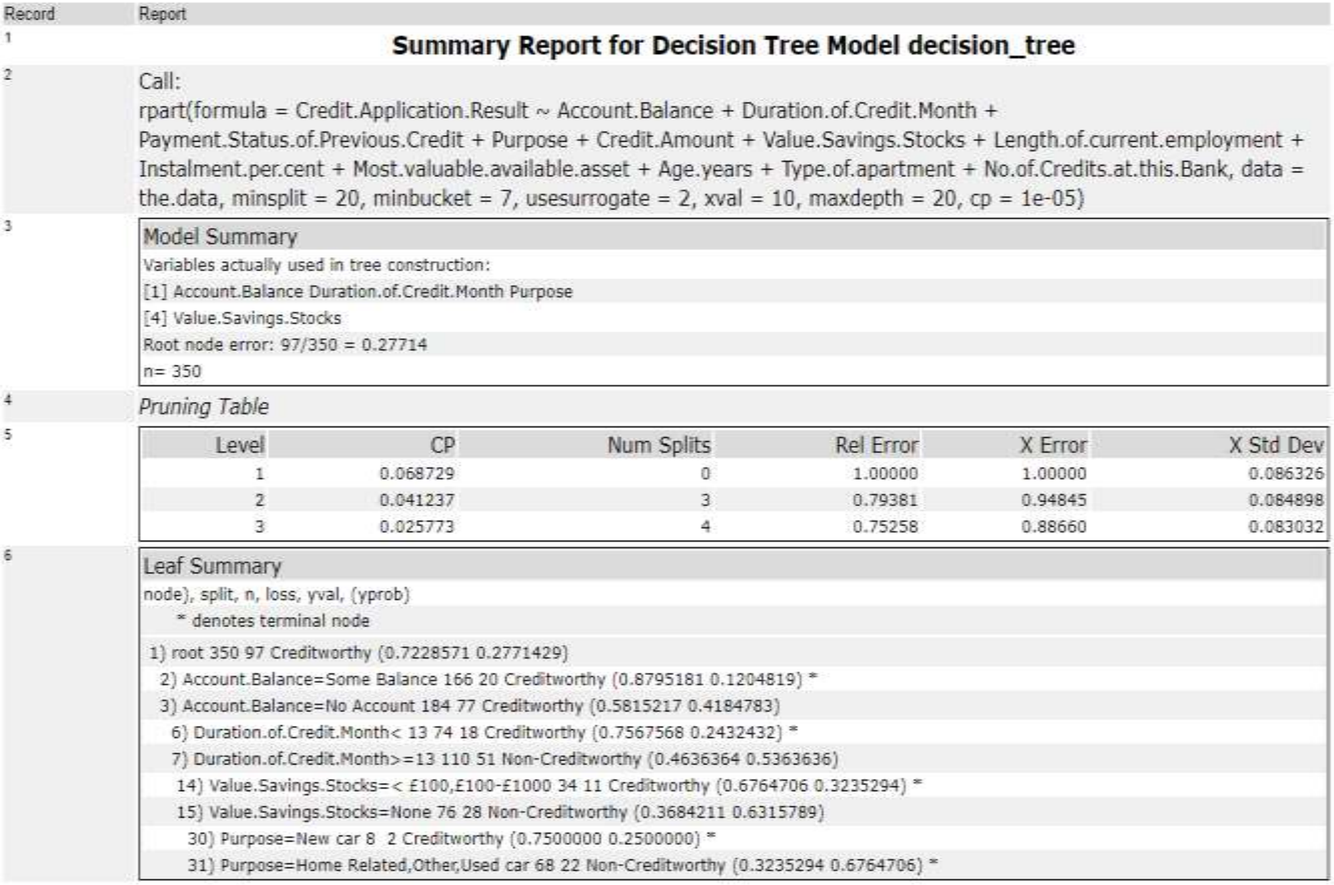
Account.Balance ,Duration.of.Credit.Month , Payment.Status.of.Previous.Credit ,Purpose ,Credit.Amount , Value.Savings.Stocks , Length.of.current.employment ,Instalment.per.cent , Most.valuable.available.asset , Age.years ,Type.of.apartment ,No.of.Credits.at.this.Bank

Important Predictor Variables: **Account.Balance, Credit.Amount, Purpose, Length.of.current.employment, Instalment%**

Overall Model accuracy: 78%

Model is creditworthy biased

Decision Tree



Record | Report

1

Summary Report for Decision Tree Model decision_tree

2

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)
```

3

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

4

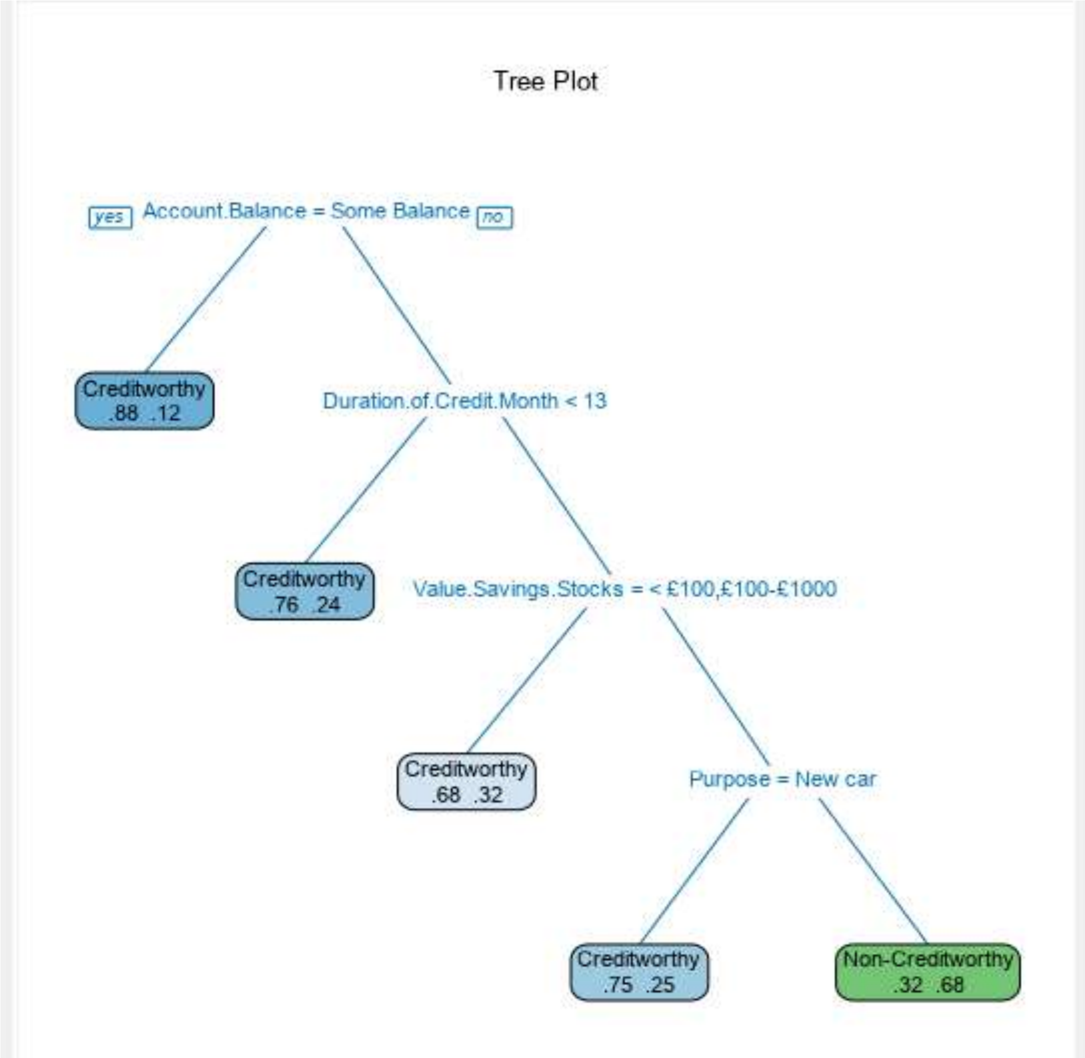
Pruning Table

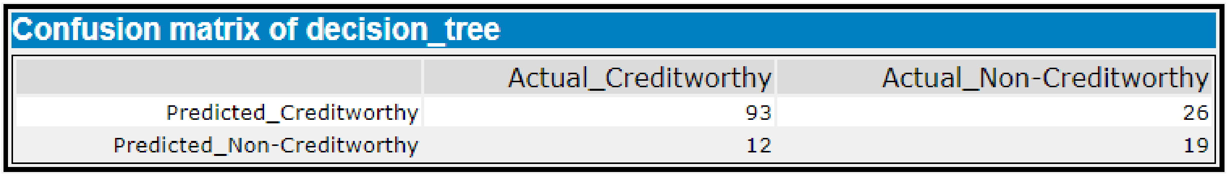
5

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	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032

6

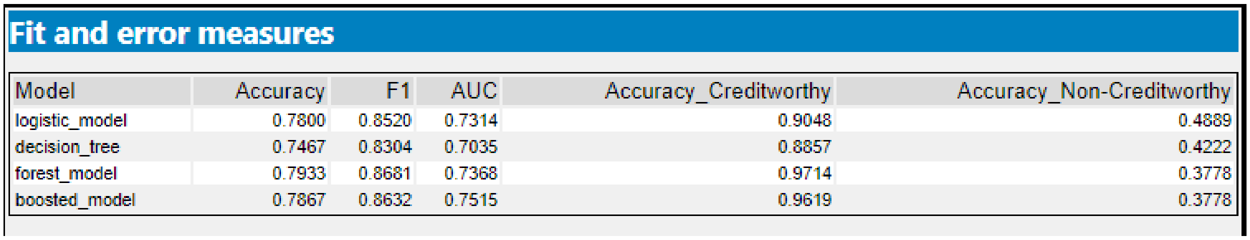
Leaf Summary
```
node), split, n, loss, yval, (yprob)
     * denotes terminal node
 1) root 350 97 Creditworthy (0.7228571 0.2771429)
   2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
   3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
     6) Duration.of.Credit.Month< 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)
        30) Purpose=New car 8  2 Creditworthy (0.7500000 0.2500000) *
        31) Purpose=Home Related,Other,Used car 68 22 Non-Creditworthy (0.3235294 0.6764706) *
```





Confusion matrix of decision_tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_model	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7515	0.9619	0.3778

Predictor Variables:

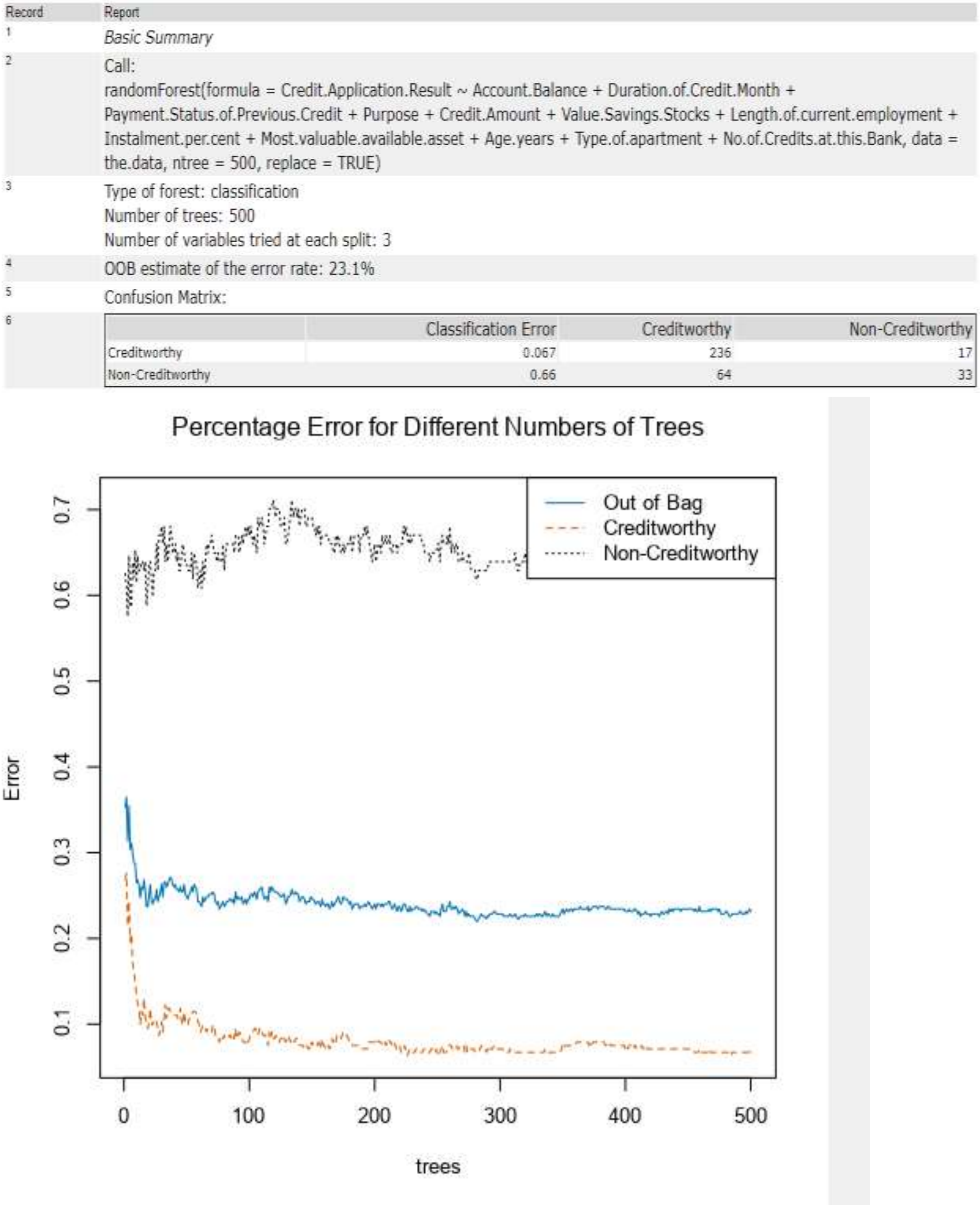
Account.Balance ,Duration.of.Credit.Month , Payment.Status.of.Previous.Credit ,Purpose ,Credit.Amount , Value.Savings.Stocks , Length.of.current.employment , Instalment.per.cent , Most.valuable.available.asset , Age.years ,Type.of.apartment ,No.of.Credits.at.this.Bank

Important Predictor Variables: **Account.Balance , Duration.of.Credit.Month , Value.Savings.Stocks**

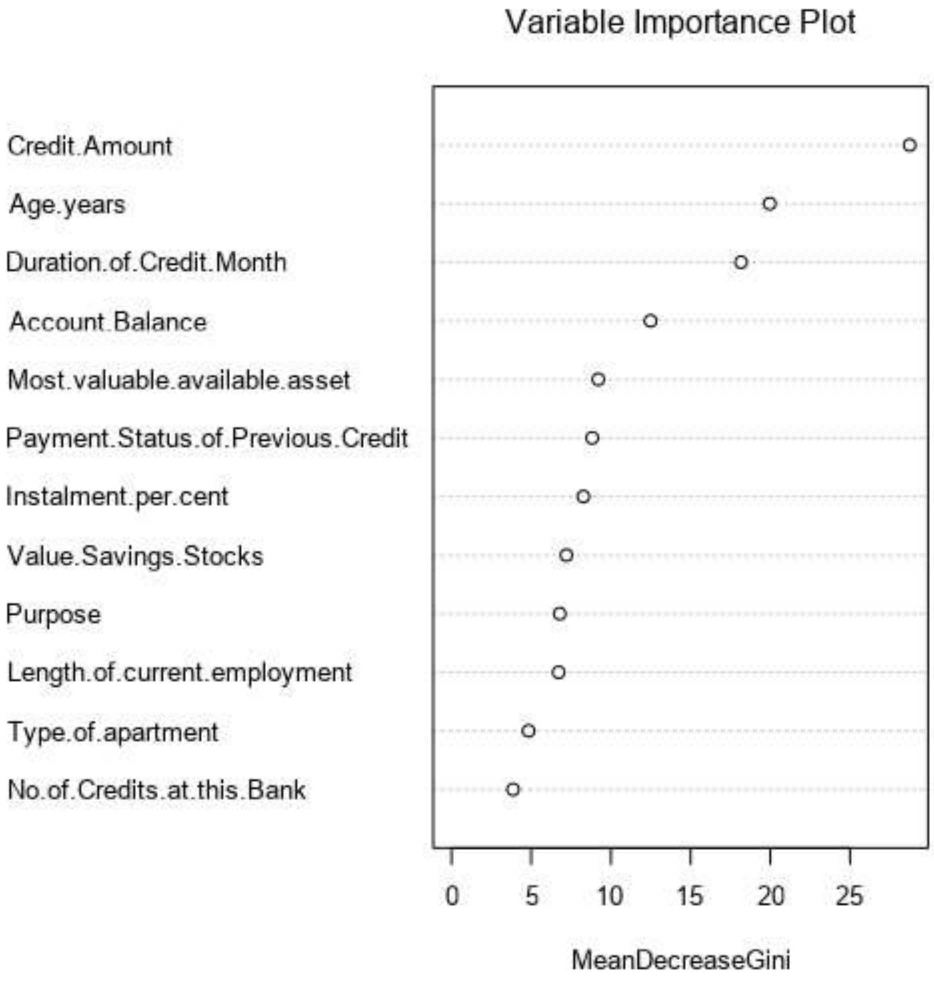
Overall Model accuracy: 75%

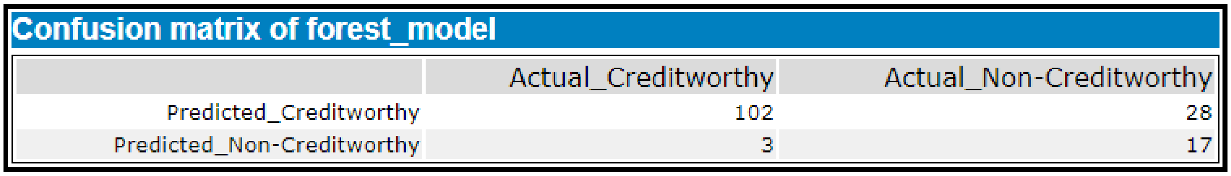
Model is creditworthy biased.

Forest Model



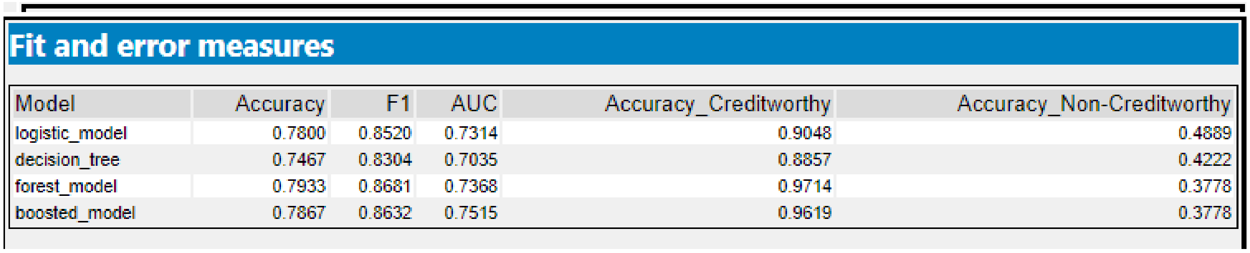
<table>
<tr><th>Record</th><th>Report</th></tr>
<tr><td>1</td><td><i>Basic Summary</i></td></tr>
<tr><td>2</td><td>Call:
randomForest(formula = Credit.Application.Result ~ Account.Balance + Duration.of.Credit.Month + Payment.Status.of.Previous.Credit + Purpose + Credit.Amount + Value.Savings.Stocks + 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)</td></tr>
<tr><td>3</td><td>Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3</td></tr>
<tr><td>4</td><td>OOB estimate of the error rate: 23.1%</td></tr>
<tr><td>5</td><td>Confusion Matrix:</td></tr>
<tr><td>6</td><td>
<table>
<tr><th></th><th>Classification Error</th><th>Creditworthy</th><th>Non-Creditworthy</th></tr>
<tr><td>Creditworthy</td><td>0.067</td><td>236</td><td>17</td></tr>
<tr><td>Non-Creditworthy</td><td>0.66</td><td>64</td><td>33</td></tr>
</table>
</td></tr>
</table>





Confusion matrix of forest_model

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



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_model	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7515	0.9619	0.3778

Predictor Variables:

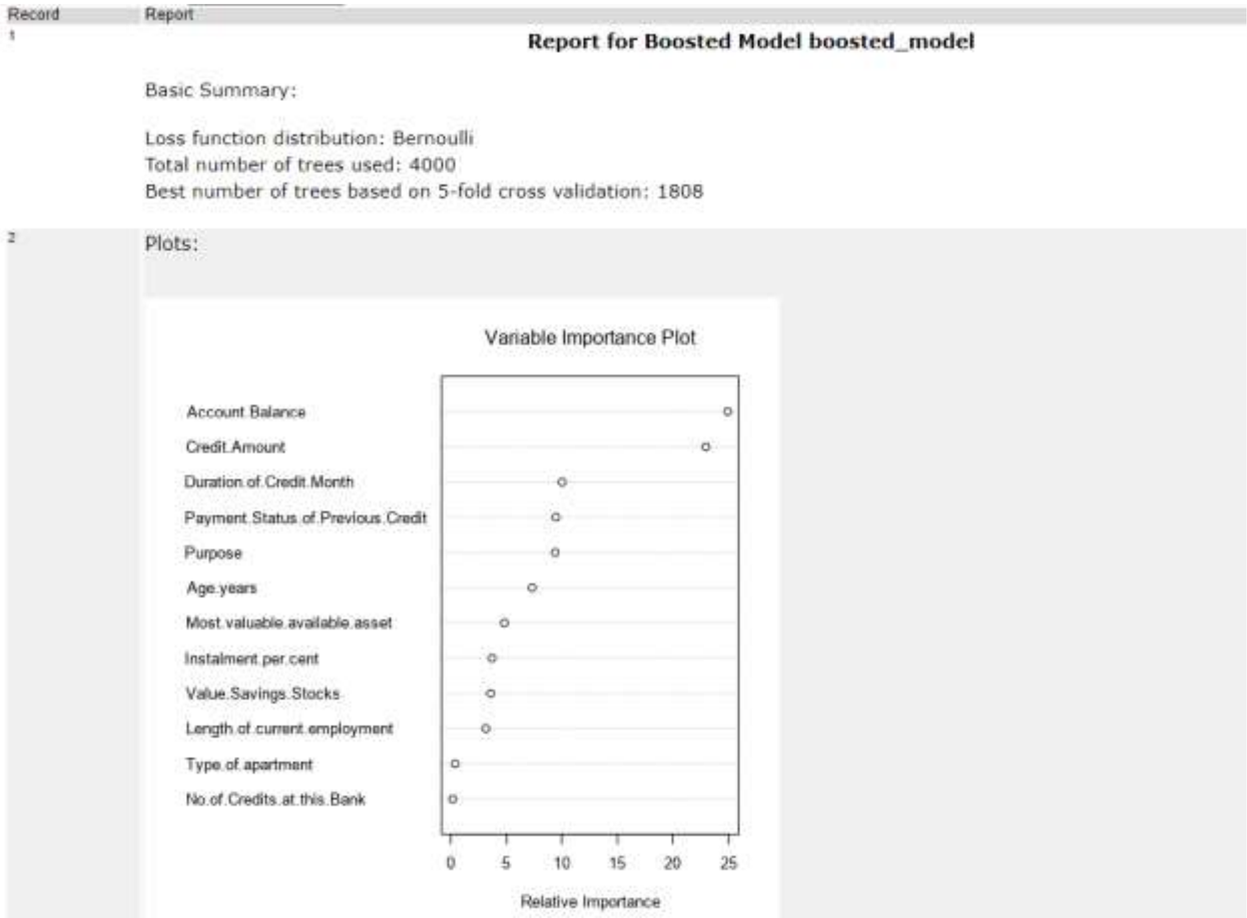
Account.Balance ,Duration.of.Credit.Month , Payment.Status.of.Previous.Credit ,Purpose ,Credit.Amount , Value.Savings.Stocks , Length.of.current.employment , Instalment.per.cent , Most.valuable.available.asset , Age.years ,Type.of.apartment ,No.of.Credits.at.this.Bank

Important Predictor Variables: Credit.Amount, Age.years, Duration.of.Credit.Month, Account.Balance

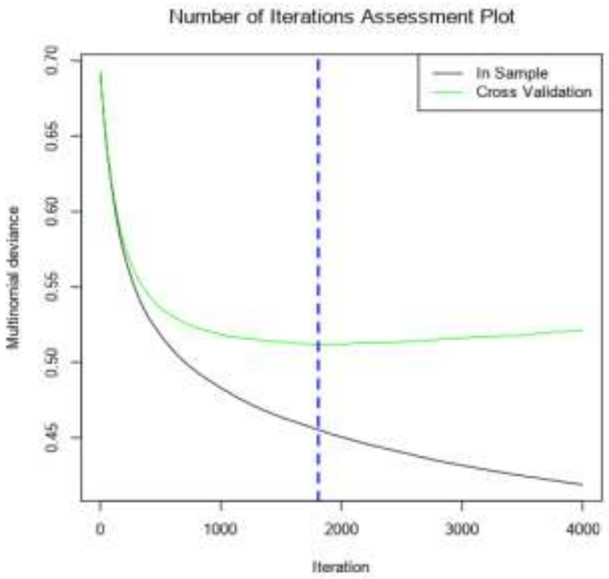
Overall percent accuracy: 79%

Model is creditworthy biased.

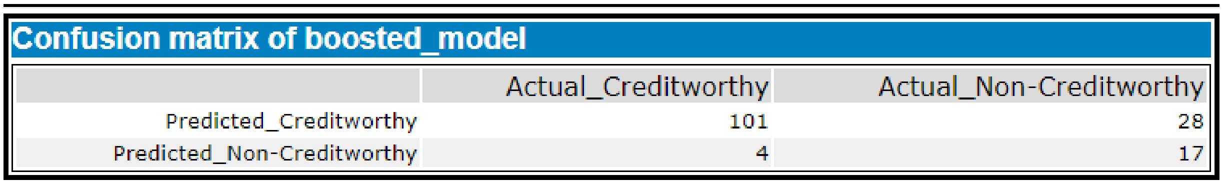
Boosted Model



Record	Report
1	**Report for Boosted Model boosted_model** Basic Summary: Loss function distribution: Bernoulli Total number of trees used: 4000 Best number of trees based on 5-fold cross validation: 1808
2	Plots:

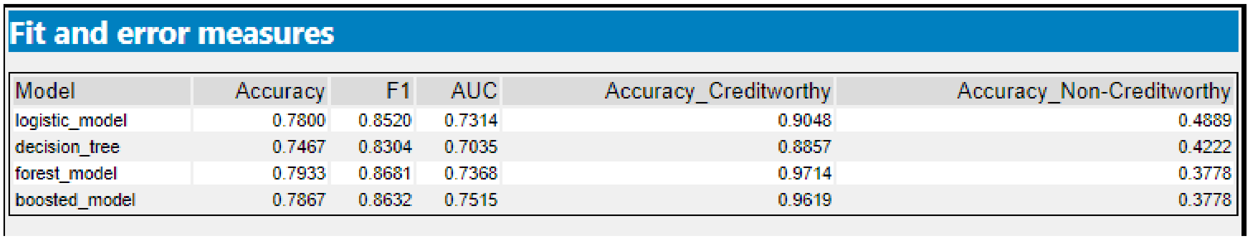


The Number of Iterations Assessment Plot illustrates how the deviance (loss) changes with the number of trees included in the model. The vertical blue dashed line indicates where the minimum deviance occurs using the specfied assessment criteria (cross validation, the use of a test sample, or out-of-bag prediction).



Confusion matrix of boosted_model

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



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_model	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7515	0.9619	0.3778

Predictor Variables:

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

Important Predictor Variables: Credit.Amount, Age.years, Payment.Status.of.Previous.Credit Duration.of.Credit.Month, Account.Balance

Overall percent accuracy :78%

Model is creditworthy biased.

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"

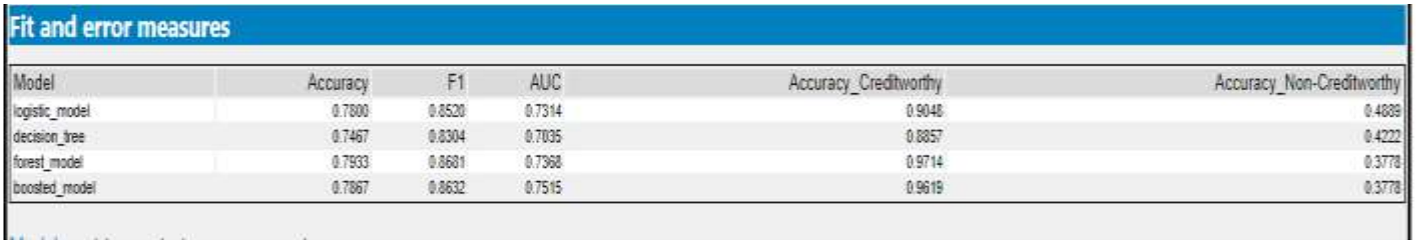
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

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.

Based on my analyses I will use the Logistic model. While it does not have the highest accuracy it is still high for the Creditworthy segments at 90%, it does have the highest accuracy for Non-Creditworthy segments (49%). The ROC curve is also within the parameters of the other models.



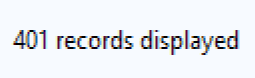
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
logistic_model	0.7800	0.8520	0.7314	0.9048	0.4889
decision_tree	0.7467	0.8304	0.7035	0.8857	0.4222
forest_model	0.7933	0.8681	0.7368	0.9714	0.3778
boosted_model	0.7867	0.8632	0.7515	0.9619	0.3778

.

- How many individuals are creditworthy?

401 customer are creditworthy.



401 records displayed

<u>Before you Submit</u>

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.