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?

We need to develop and test out an efficient model to classify new customers based on whether they can be approved for a loan or not.

- What data is needed to inform those decisions?

We will be using historical data from pervious customers (credit-data-training) to create a model that will be applied against new applicants (customers-to-score) to qualify them for a loan.

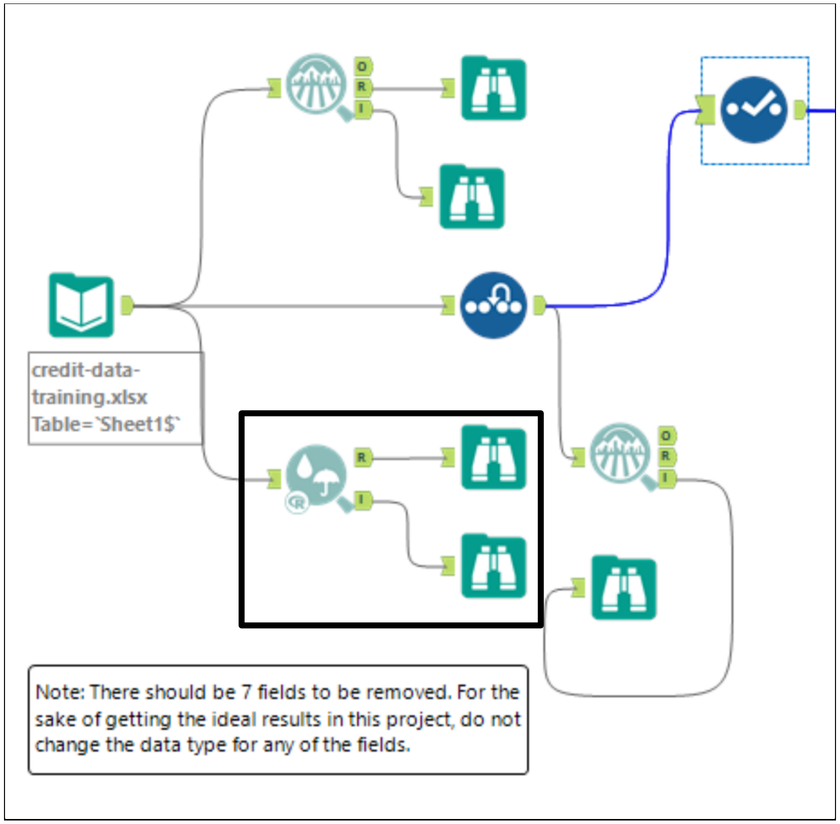
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

We will be using the following Classification models (Logistic regression, Decision tree, Random Forest and Boosted model) to determine if a customer is credit worthy or not. The outcome is a binary value of either yes or no.

Step 2: Building the Training Set

- 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”.

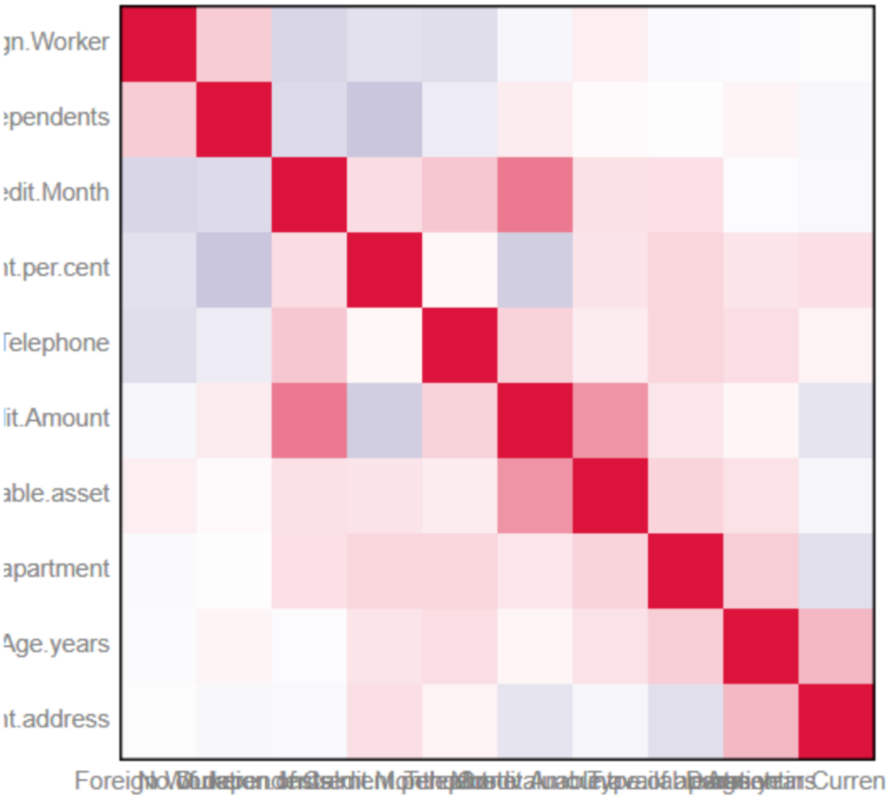
We utilize the Association analysis tool to check for the correlation between the various input variables. From the correlation matrix we see that there are no variables that exceed 0.7. We do not have highly-correlated variables.



credit-data-training.xlsx
Table=`Sheet1$`

Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.

Correlation Matrix with ScatterPlot

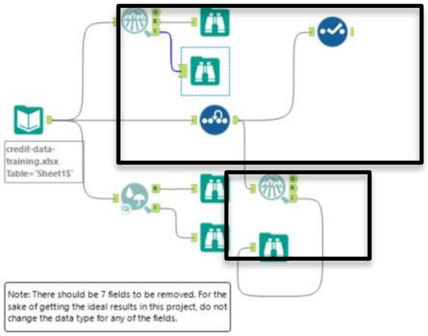


Pearson Correlation Analysis

Full Correlation Matrix

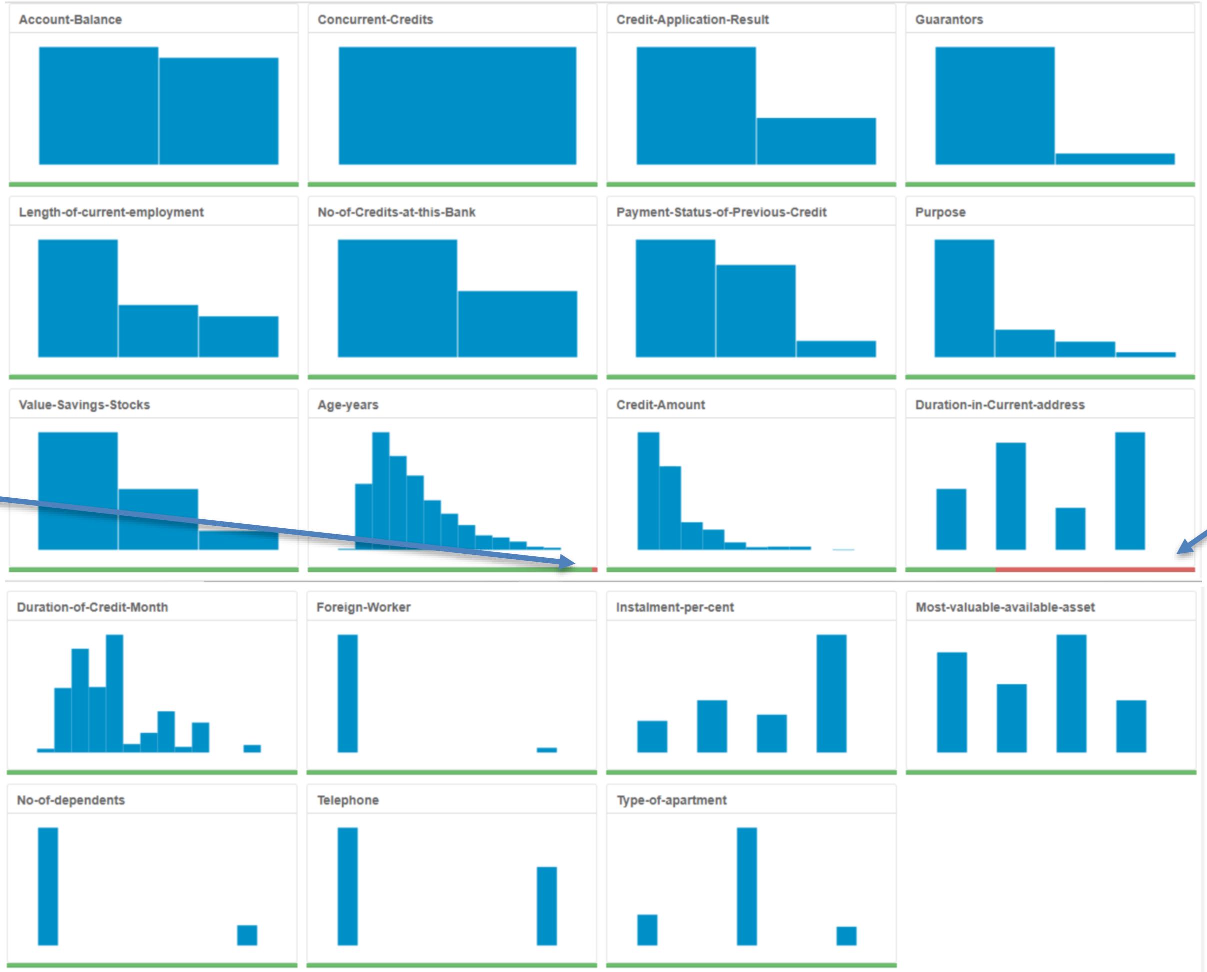
	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Duration.in.Current.address	Most.valuable.available.asset	Age.years	Type.of.apartment	No.of.dependents	Telephone	Foreign.Worker
Duration.of.Credit.Month	1.000000	0.565054	0.145637	-0.032494	0.128814	-0.018171	0.126967	-0.185180	0.238437	-0.207298
Credit.Amount	0.565054	1.000000	-0.253286	-0.136621	0.457147	0.040486	0.100413	0.082721	0.192532	-0.045994
Instalment.per.cent	0.145637	-0.253286	1.000000	0.131231	0.115114	0.111456	0.178926	-0.293380	0.038515	-0.155458
Duration.in.Current.address	-0.032494	-0.136621	0.131231	1.000000	-0.047386	0.301966	-0.163386	-0.036814	0.055112	-0.015787
Most.valuable.available.asset	0.128814	0.457147	0.115114	-0.047386	1.000000	0.123579	0.182744	0.019435	0.083395	0.071932
Age.years	-0.018171	0.040486	0.111456	0.301966	0.123579	1.000000	0.208552	0.046996	0.141103	-0.020939
Type.of.apartment	0.126967	0.100413	0.178926	-0.163386	0.182744	0.208552	1.000000	-0.010189	0.179688	-0.026742
No.of.dependents	-0.185180	0.082721	-0.293380	-0.036814	0.019435	0.046996	-0.010189	1.000000	-0.097632	0.218454
Telephone	0.238437	0.192532	0.038515	0.055112	0.083395	0.141103	0.179688	-0.097632	1.000000	-0.168472
Foreign.Worker	-0.207298	-0.045994	-0.155458	-0.015787	0.071932	-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



Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.

The Field Summary tool is used to analyzes data and create a summary report. We can see from the summary report that we have a huge amount of data that is missing from the field ‘Duration-in-current-address. Since we will not be imputing these values we will go ahead and omit them. The Field Age-years will be imputed using the Imputation tool with the Median value for age since it does not have a huge amount of missing data.



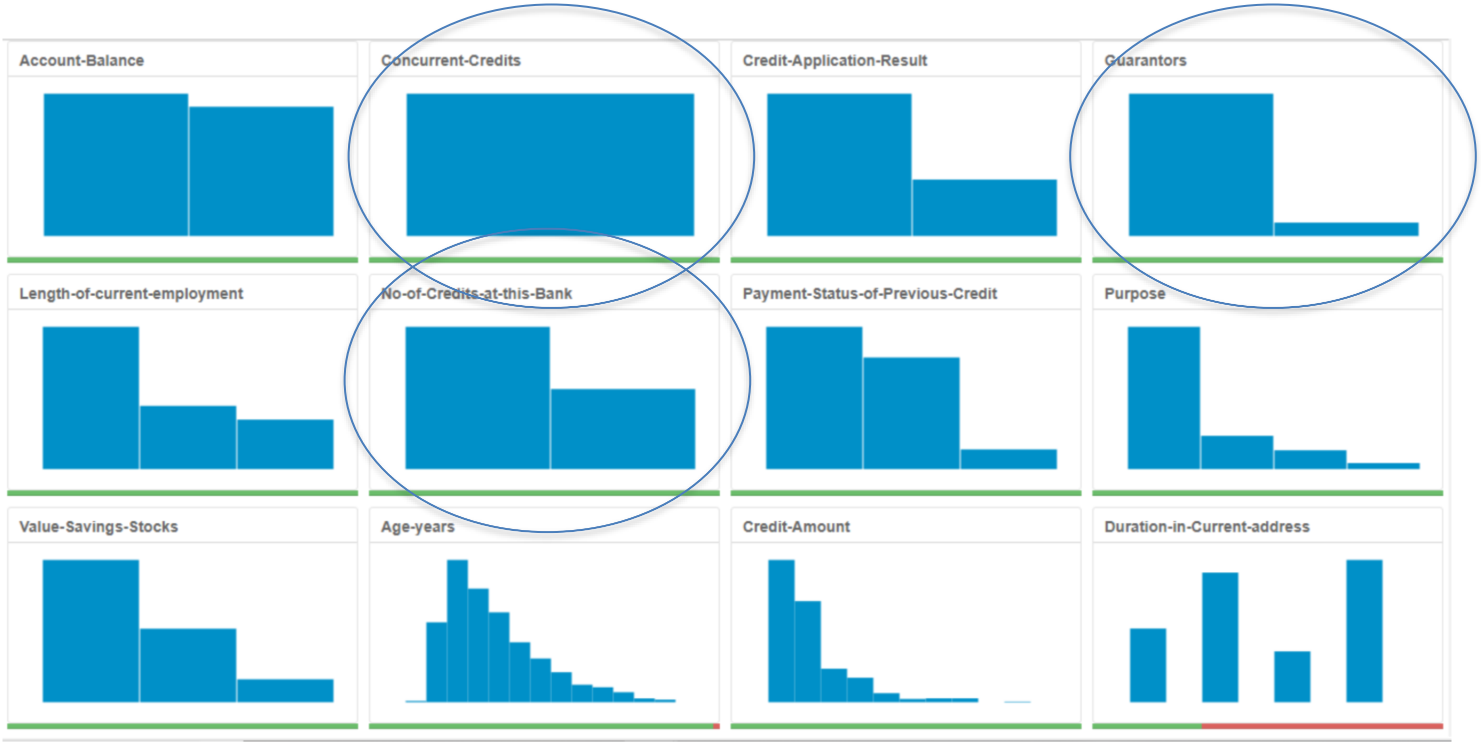
2% data missing from the Age-years. The missing data will be imputed using the Median value.

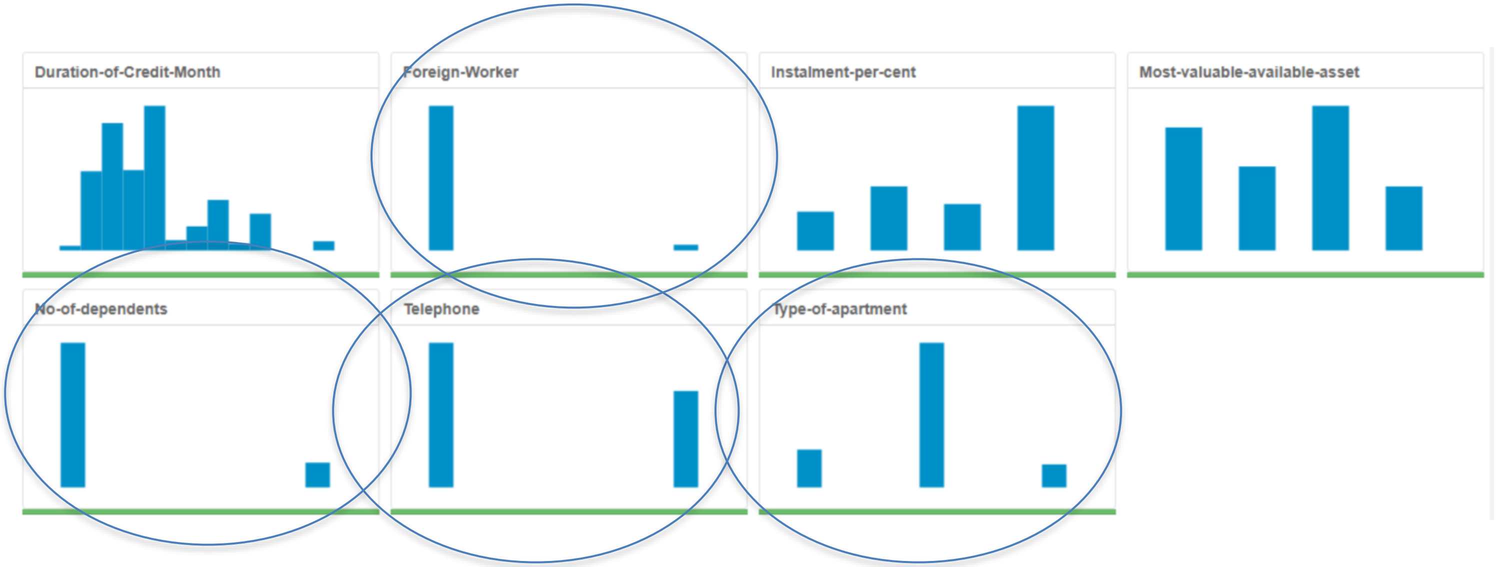
69% data missing from the Duration-in-current-address. This field will be omitted from the data set.

- 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.

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

The Following fields have been identified and removed from the data set since they have either very low variability or there is only a single data point. The telephone data was removed from the data set since it does not have any logical reason for including the variable. We can see from the charts that concurrent – credits and Occupation has only 1 value, Guarantors, Foreign-Workers, No-of-Dependents have 2 values with a majority skew towards a single instance these will also be removed.

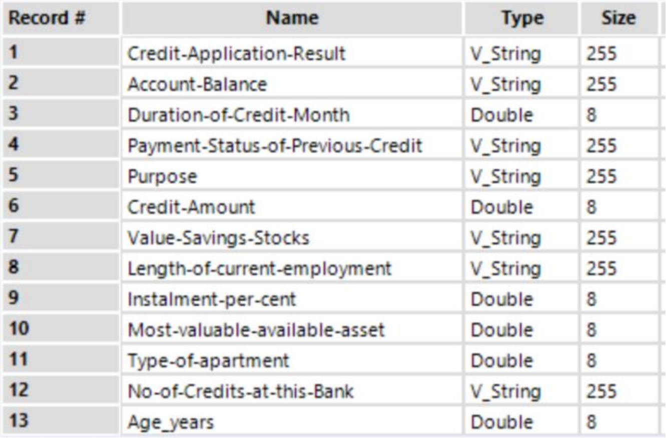




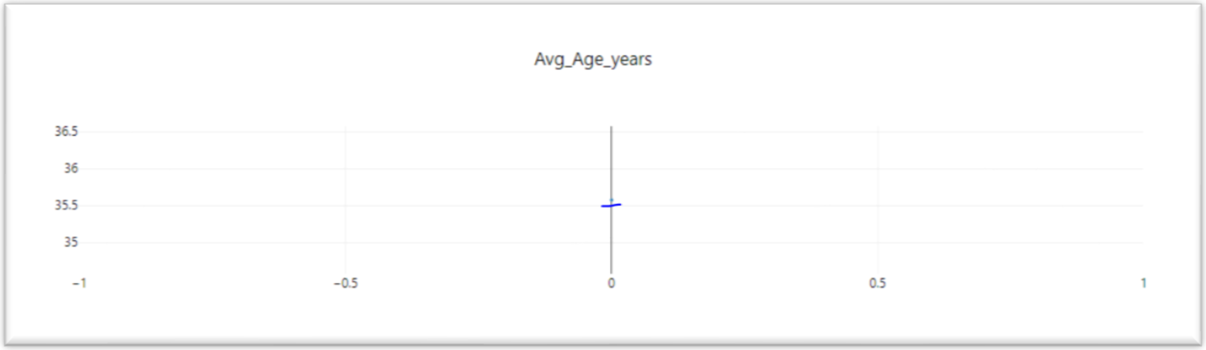
- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

***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)*

The Following 13 fields are used for the analysis the average age rounded up to 36.



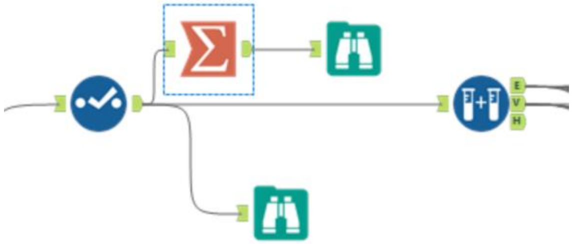
Record #	Name	Type	Size
1	Credit-Application-Result	V_String	255
2	Account-Balance	V_String	255
3	Duration-of-Credit-Month	Double	8
4	Payment-Status-of-Previous-Credit	V_String	255
5	Purpose	V_String	255
6	Credit-Amount	Double	8
7	Value-Savings-Stocks	V_String	255
8	Length-of-current-employment	V_String	255
9	Instalment-per-cent	Double	8
10	Most-valuable-available-asset	Double	8
11	Type-of-apartment	Double	8
12	No-of-Credits-at-this-Bank	V_String	255
13	Age_years	Double	8



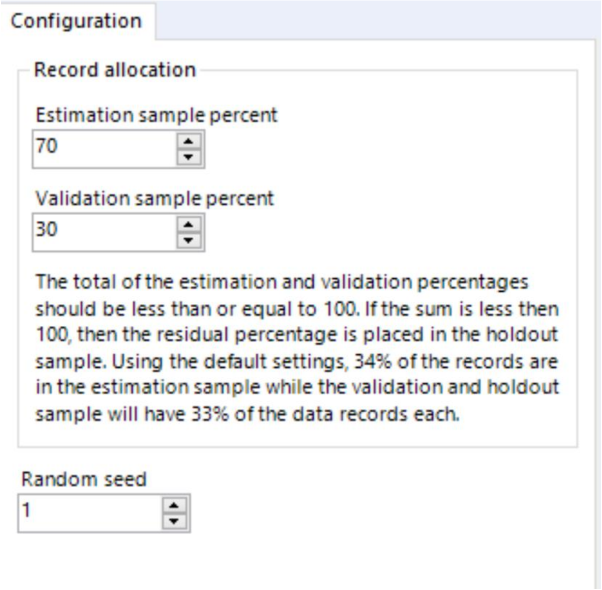
Answer this question:

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.



The following configuration has been used for the data split with 70% as Estimation and 30% for Validation. Random seed value has been set to 1.



Configuration

Record allocation

Estimation sample percent
70

Validation sample percent
30

The total of the estimation and validation percentages should be less than or equal to 100. If the sum is less then 100, then the residual percentage is placed in the holdout sample. Using the default settings, 34% of the records are in the estimation sample while the validation and holdout sample will have 33% of the data records each.

Random seed
1

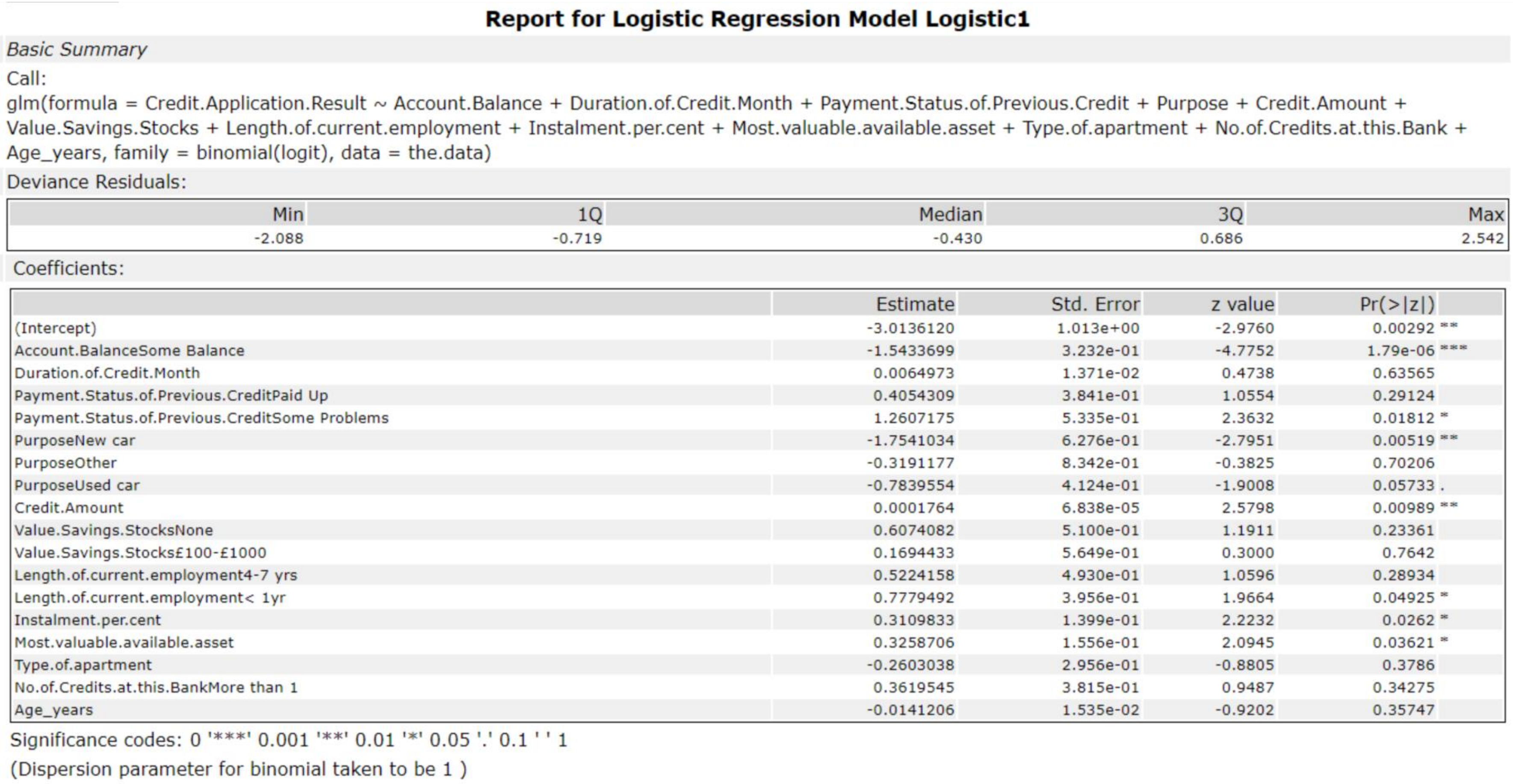
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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

Model Logistic Regression:

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

The predictor variable chart indicates that we have **p-values** (typically ≤ 0.05) with high significance (Account.Balance , Purpose, Credit-Amount ,Length-of-current.employment and Instalment-per-cent.



Report for Logistic Regression Model Logistic1

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:

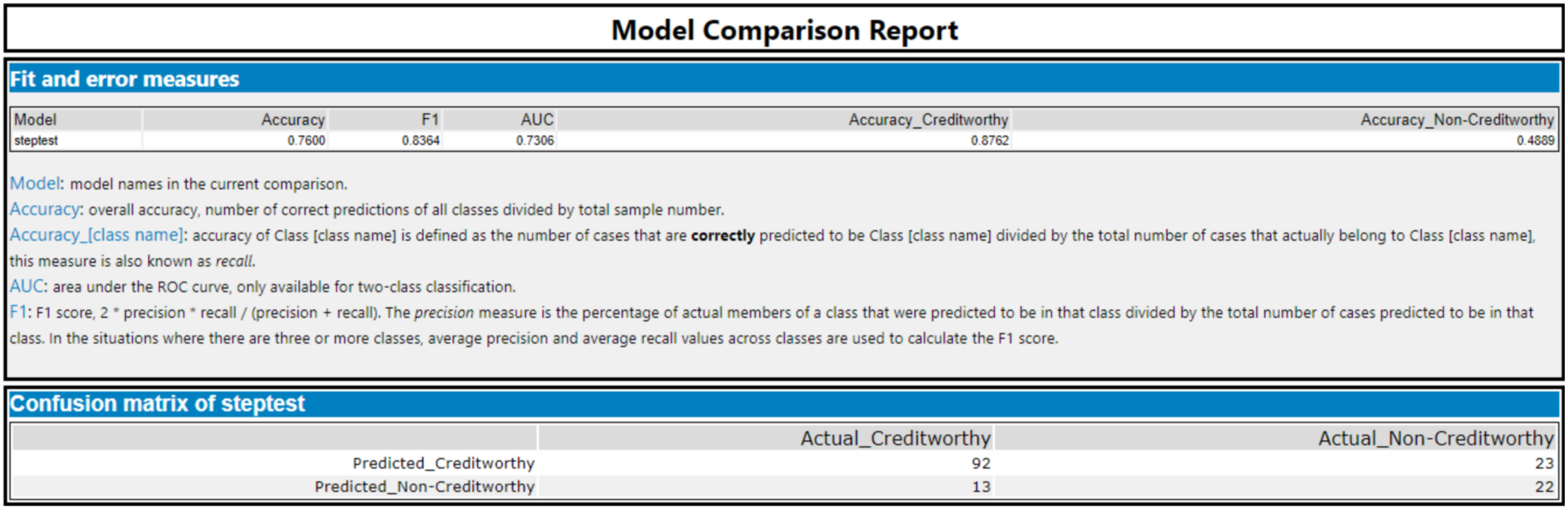
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome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

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 percentage is at 76%. We see from the confusion matrix that the model tends to predict a higher number of users creditworthy as compared to non-creditworthy here we see 23 predicted as creditworthy falsely while only 13 creditworthy were predicted as non-creditworthy. We see the model has a low accuracy of predicting Non-creditworthy of 49% which makes the model biased towards predicting customers as creditworthy falsely which makes a bias towards creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
steptest	0.7600	0.8364	0.7306	0.8762	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

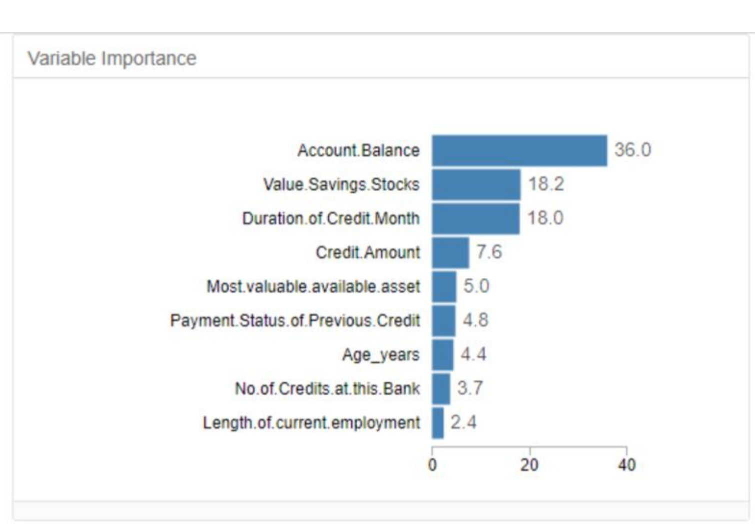
Confusion matrix of steptest

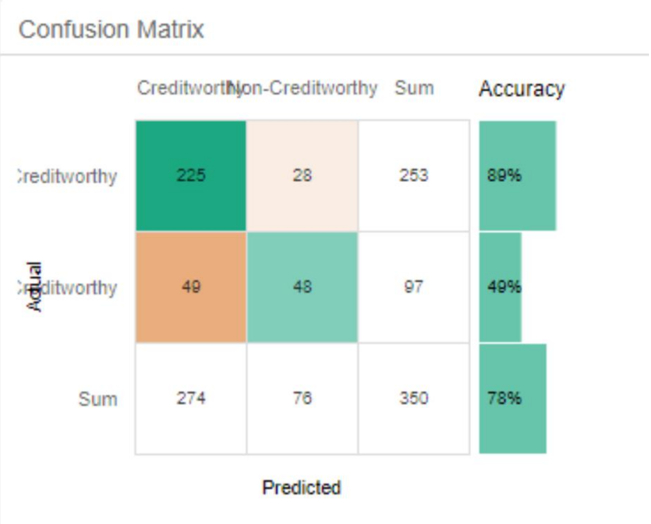
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Model Decision Tree:

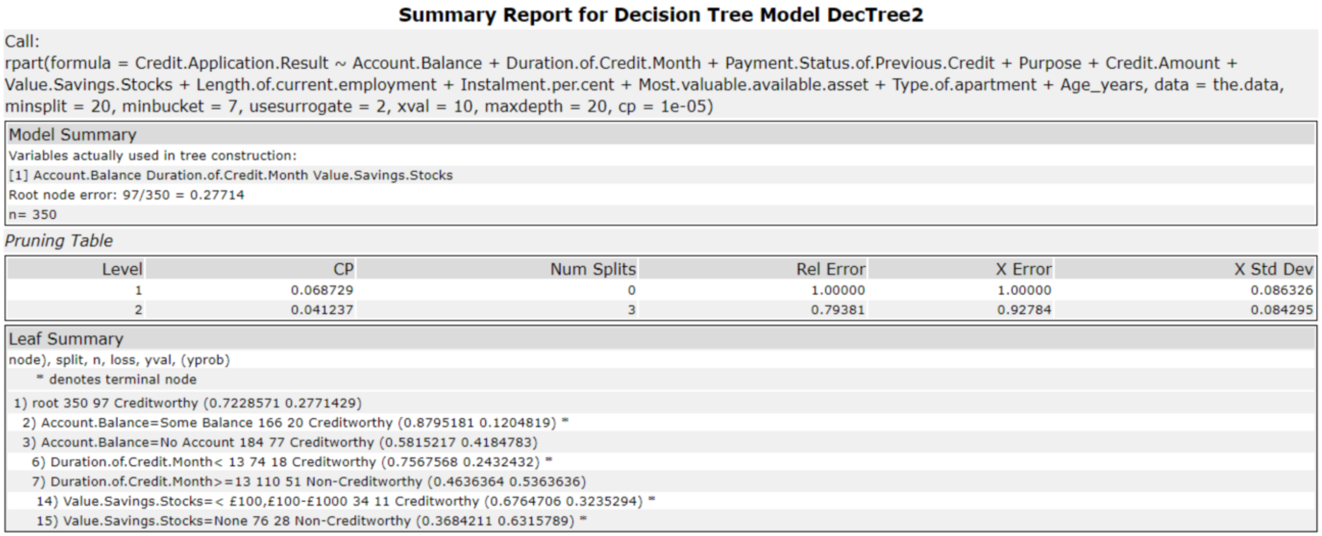
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all your predictor variables.

The Variable importance graph shows Account.Balance ,Value.savings.Stocks and Duration.of.credit.Month having the highest importance.




Confusion Matrix

Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
Creditworthy	225	28	253	89%
Non-Creditworthy	49	48	97	49%
Sum	274	76	350	78%


Summary Report for Decision Tree Model DecTree2

```
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 + Type.of.apartment + Age_years, 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

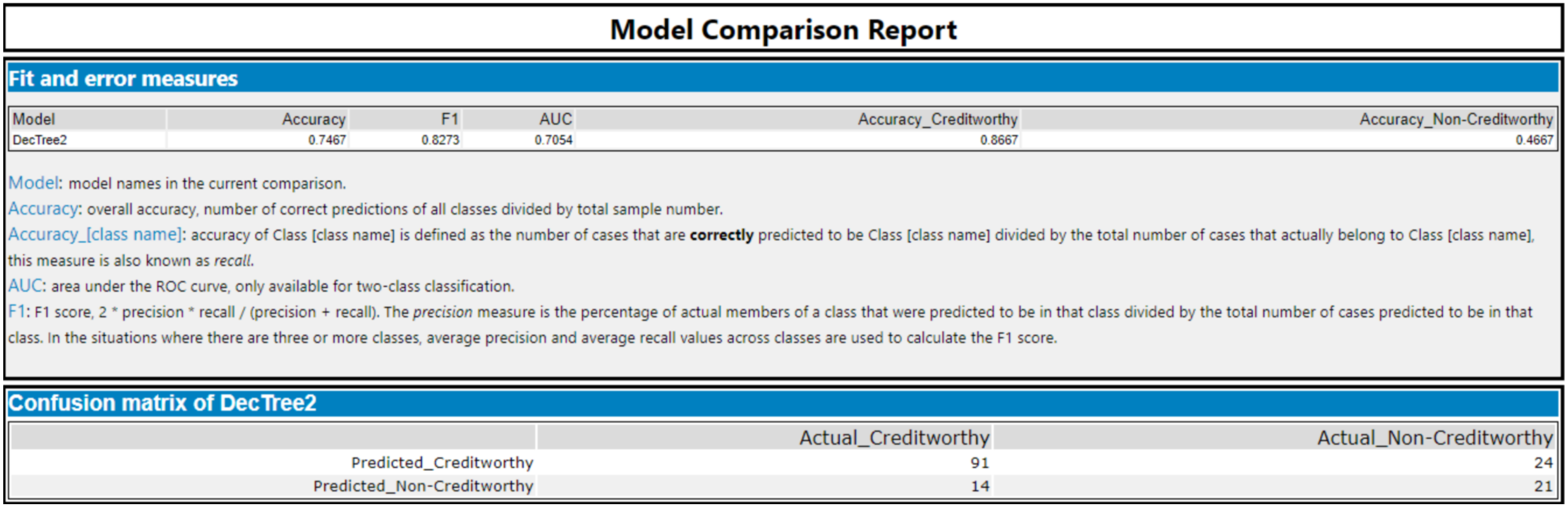
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

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) *
```

- 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 for our model is at 75%. Looking at the confusion matrix we see that the accuracy of predicting non-creditworthy is at 47% which is lower compared to the accuracy of creditworthy. Hence, we can say that the model has a higher bias towards predicting customer as creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecTree2	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

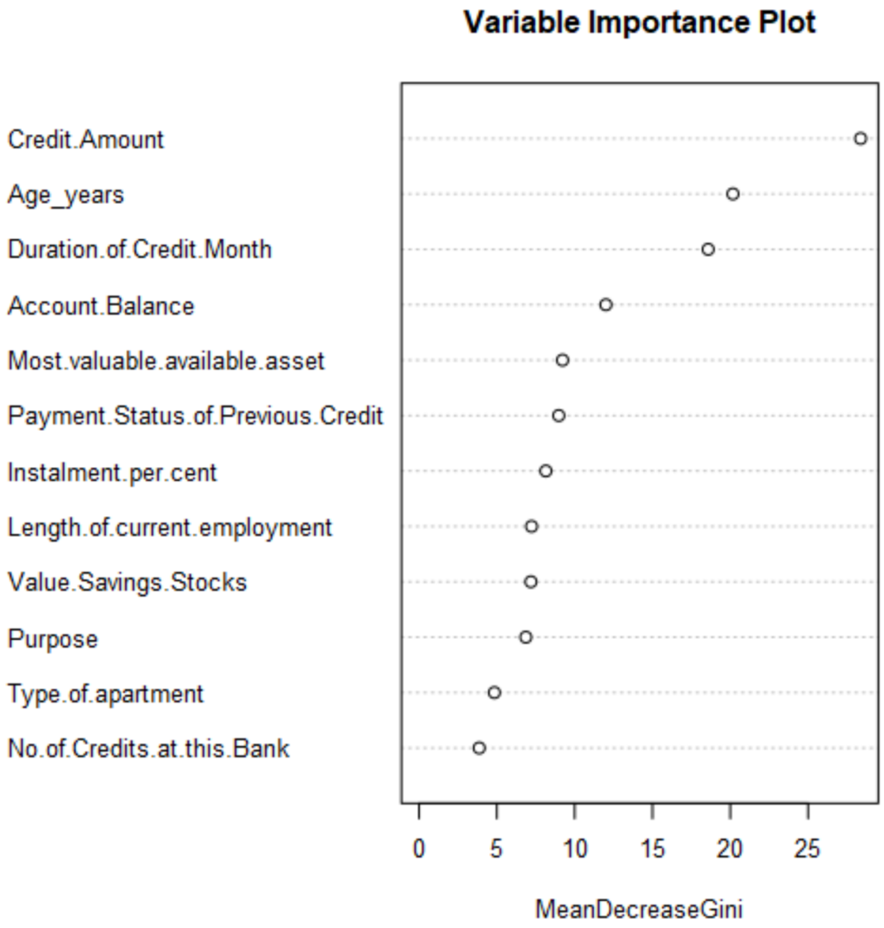
Confusion matrix of DecTree2

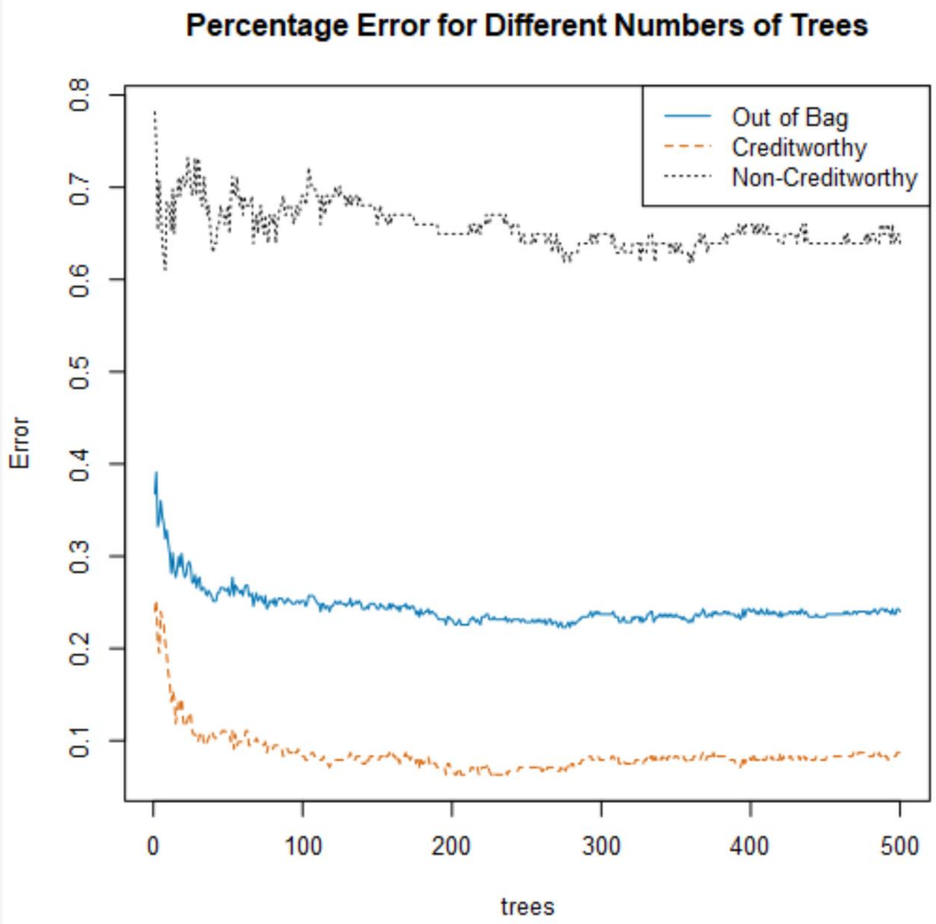
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Model Random Forest:

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

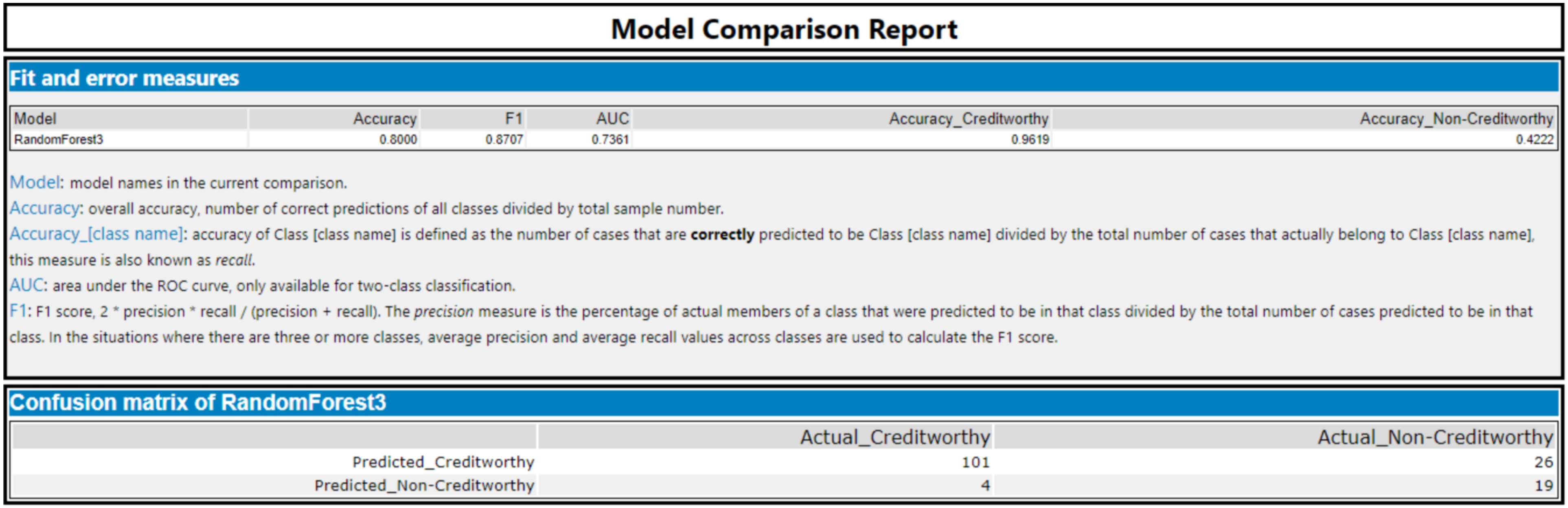
The Variable importance charts indicate that the credit-Amount, Age_years and Duration-of-Credit-Month have the highest Variable Importance. The Percentage Error for different numbers of trees shows that the error value flattens out after 200 trees.





- 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 percentage Accuracy is about 80%. The model might have a slight bias towards predicting as credit worthy with 26 predicted falsely as creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
RandomForest3	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 the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

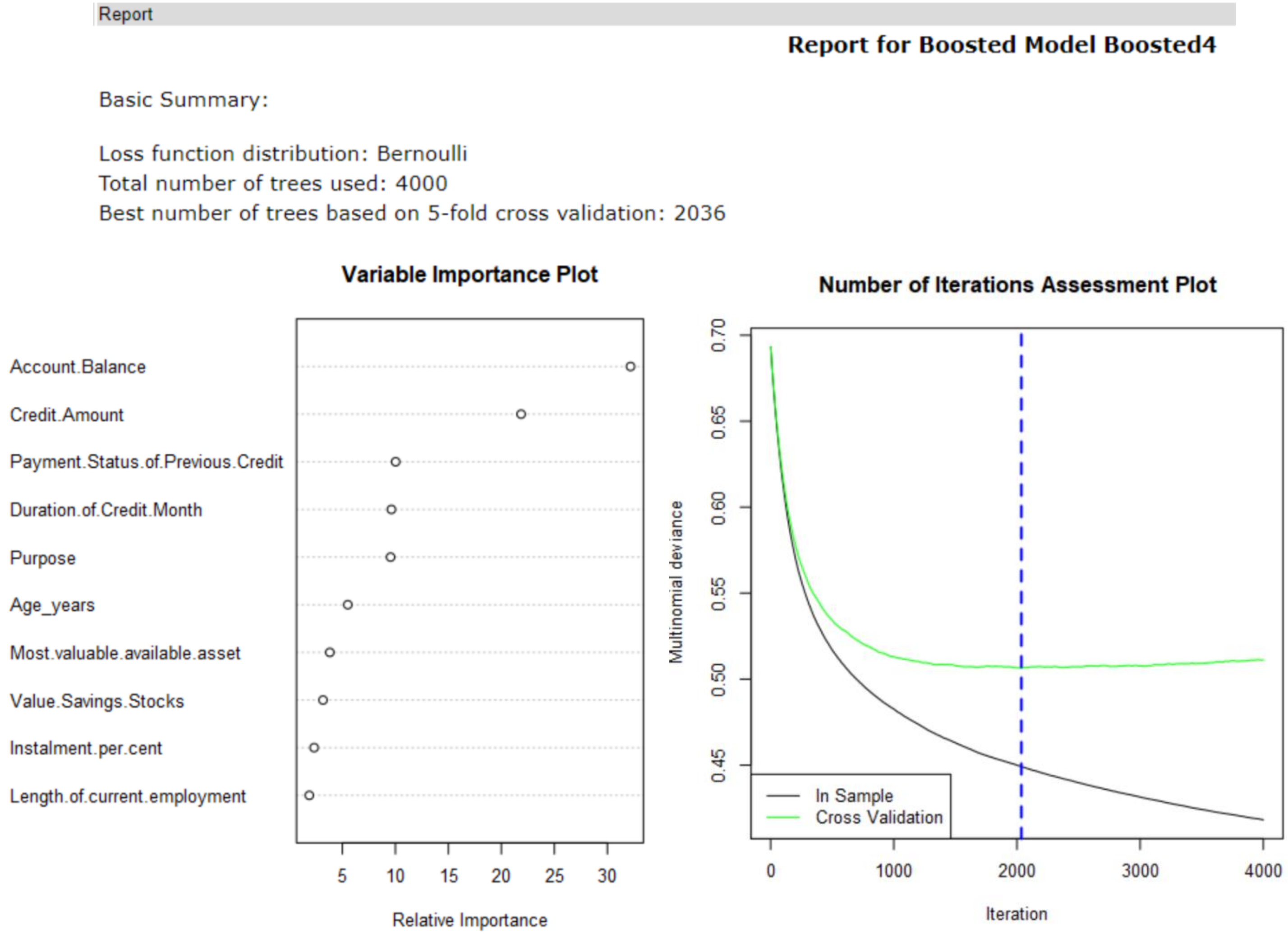
Confusion matrix of RandomForest3

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

Model Boosted:

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

From the report we see that the two most important variables are Account.Balance and Credit.Amount. We also see from the Number of iterations the best umber of tress are 2036.



Report

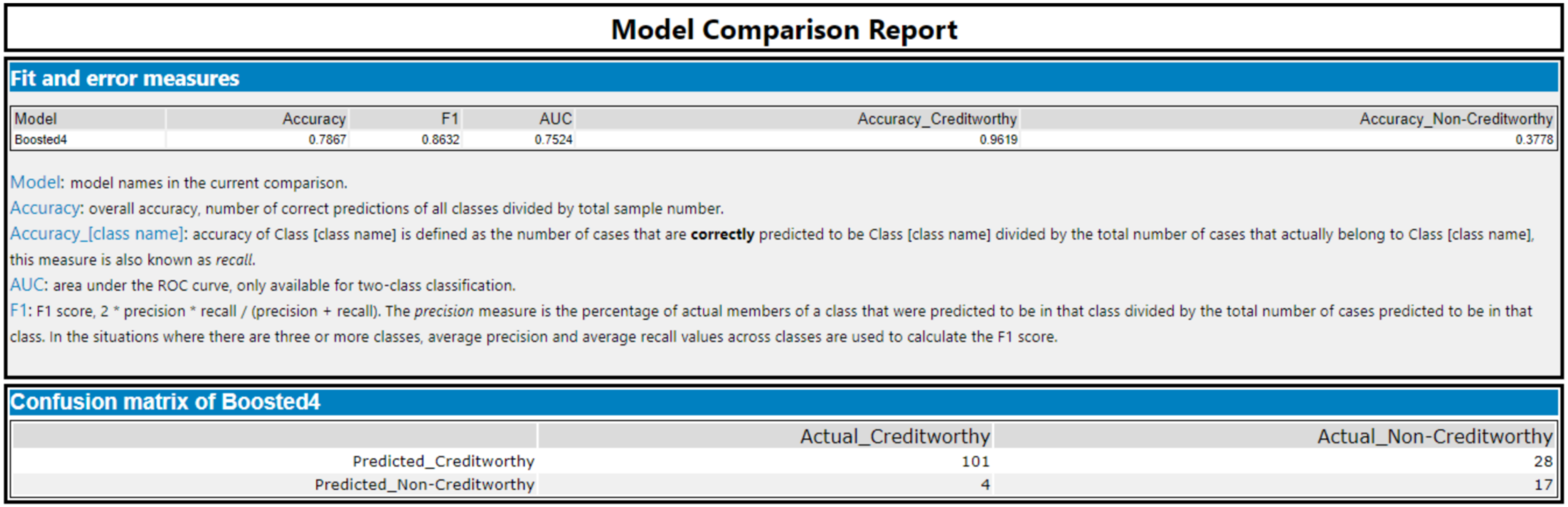
Report for Boosted Model Boosted4

Basic Summary:

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

- 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 this model is at 79%. We see from the confusion matrix that the model is highly biased towards predicting customers as credit worthy. Since it has an accuracy of only 38% for predicting non-creditworthy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted4	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 the 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 Boosted4

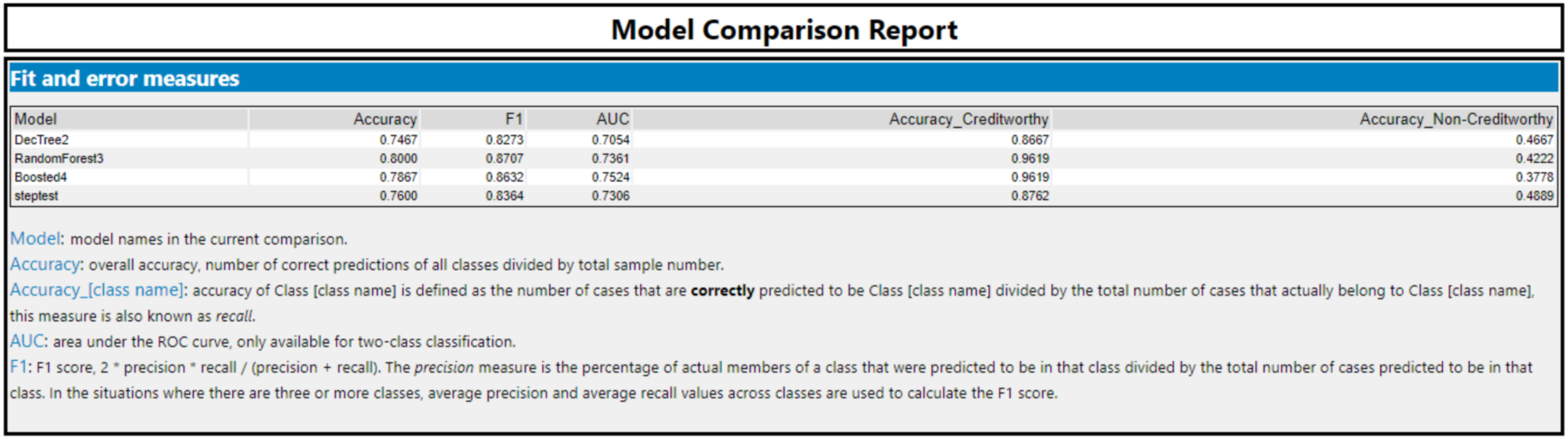
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

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”

- 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
    - ○

From the model comparison tool we see that the over all highest accuracy comes from the Random Forest Prediction model with 80% accuracy.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DecTree2	0.7467	0.8273	0.7054	0.8667	0.4667
RandomForest3	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted4	0.7867	0.8632	0.7524	0.9619	0.3778
steptest	0.7600	0.8364	0.7306	0.8762	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.

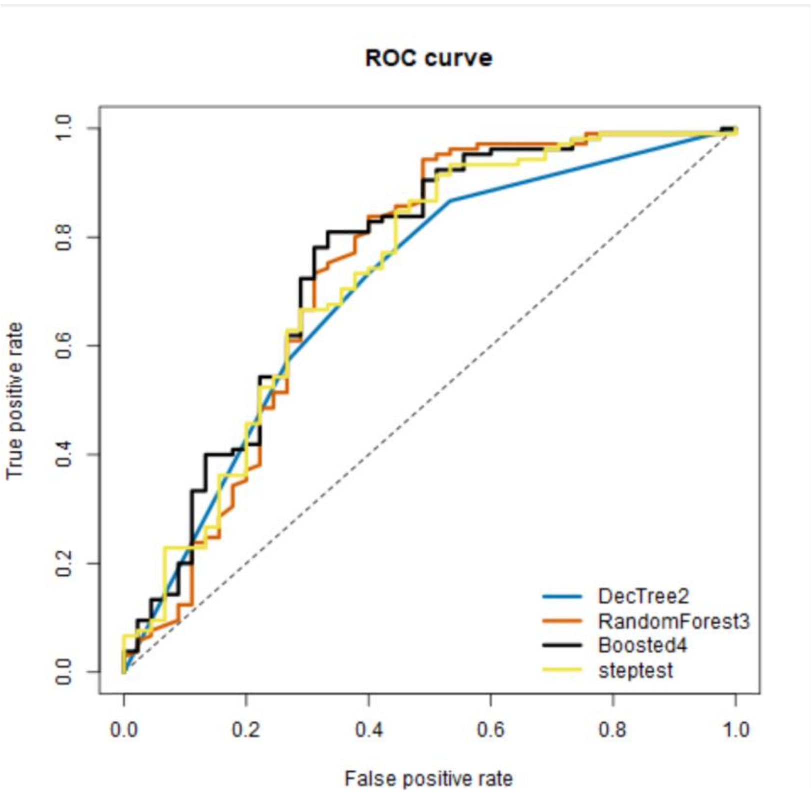
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In the situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

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)

- ○ Accuracies within “Creditworthy” and “Non-Creditworthy” segments
- ○ Bias in the Confusion Matrices

We see that the Random Forest Model Has an accuracy of 96% for predicting Creditworthy and 42% accuracy predicting non-Creditworthy. We will use the random Forest Model since it has a high degree of accuracy with a low bias tendency. We Also notice from the ROC curve we see a comparable Rate on the True Positive values compared to other models.

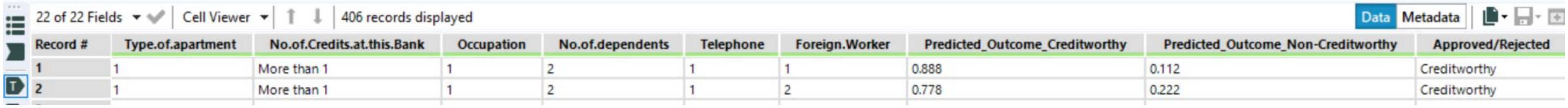
- ROC graph



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

- How many individuals are creditworthy?

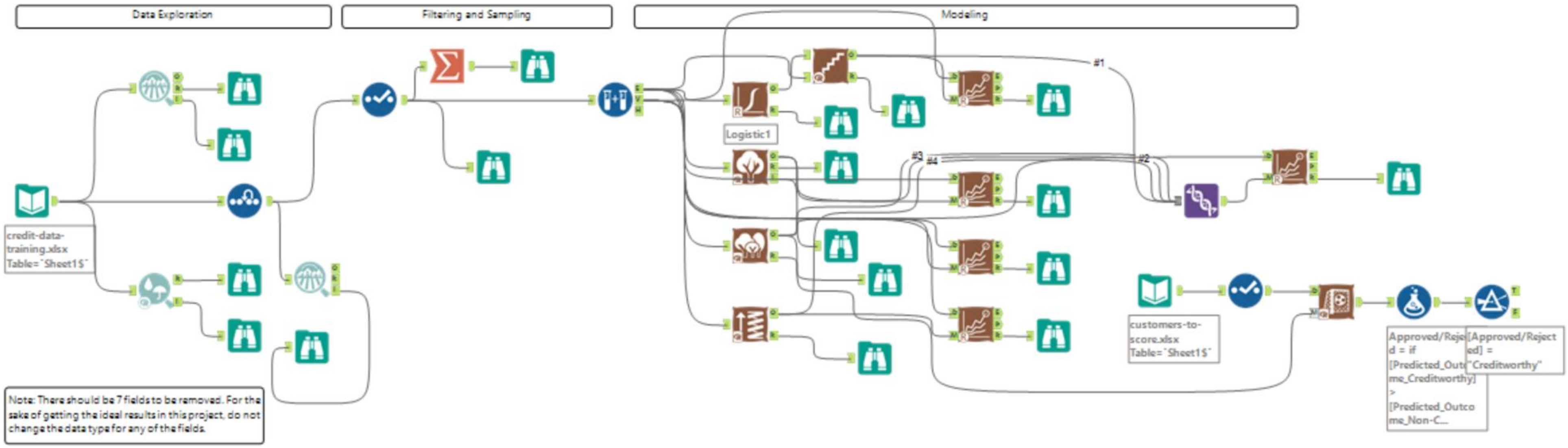
406 customers have been deemed creditworthy.



Record #	Type.of.apartment	No.of.Credits.at.this.Bank	Occupation	No.of.dependents	Telephone	Foreign.Worker	Predicted_Outcome_Creditworthy	Predicted_Outcome_Non-Creditworthy	Approved/Rejected
1	1	More than 1	1	2	1	1	0.888	0.112	Creditworthy
2	1	More than 1	1	2	1	2	0.778	0.222	Creditworthy

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.


Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx Table=`Sheet1$`
Logistic1
#1
#2
#3
#4
customers-to-score.xlsx Table=`Sheet1$`
Approved/Rejected = if [Predicted_Outcome_Creditworthy] > [Predicted_Outcome_Non-C...
[Approved/Rejected] = "Creditworthy"
T
F

Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.

Alteryx Workflow.