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?

  As a loan officer, I need to process 500 credit loans within one week based on applications' creditworthiness. So to solve the problem, I would need to build a classification model based on past loan data, and make predictions for new loan applications by implementing this prediction mode.

- What data is needed to inform those decisions?

  Data on all past applications:
  This dataset contains lots of background information and performance of past applicants, such as their age, credit amount and credit history, etc.
  I would generate my classification model based on this dataset.

  The list of customers that need to be processed in the next few days
  This dataset contains new applications' background information, combined this dataset and prediction model, I can generate predictions about each applicant's creditworthiness.

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

  Binary model, because we only concern about whether or not the applicant is creditworthy.

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.***

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)*

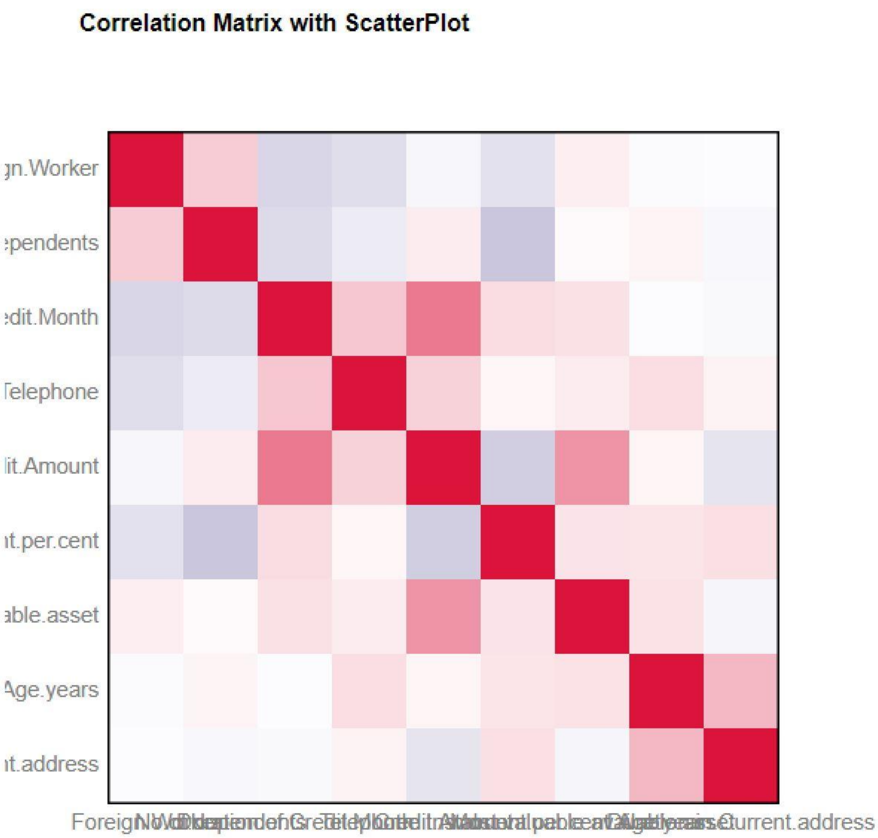
To achieve consistent results reviewers expect.

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.

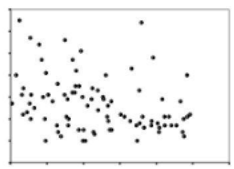
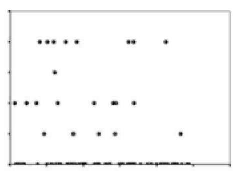
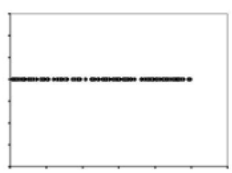
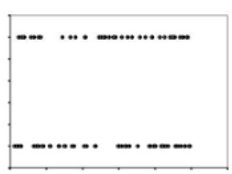
Data Cleanup

Below is the EDA for checking the health of the dataset.
Correlations:



Field Summary:


Value-Savings-Stocks
Type-of-apartment
Telephone
Purpose
Payment-Status-of-Previous-Credit
No-of-dependents
No-of-Credits-at-this-Bank
Most-valuable-available-asset
Length-of-current-employment
Instalment-per-cent
Guarantors
Foreign-Worker
Duration-of-Credit-Month
Duration-in-Current-address
Credit-Application-Result
Credit-Amount
Concurrent-Credits
Age-years
Account-Balance

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	
Duration-in-Current-address		68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Telephone		0.0%	2	1.000	1.400	1.000	2.000	0.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

<u>I removed 7 columns of the original dataset following the steps:</u>

Columns Removed:

Duration in Current Address, Concurrent-Credits, Occupation, Guarantors, Telephone, No of Dependents, Foreign Workers.

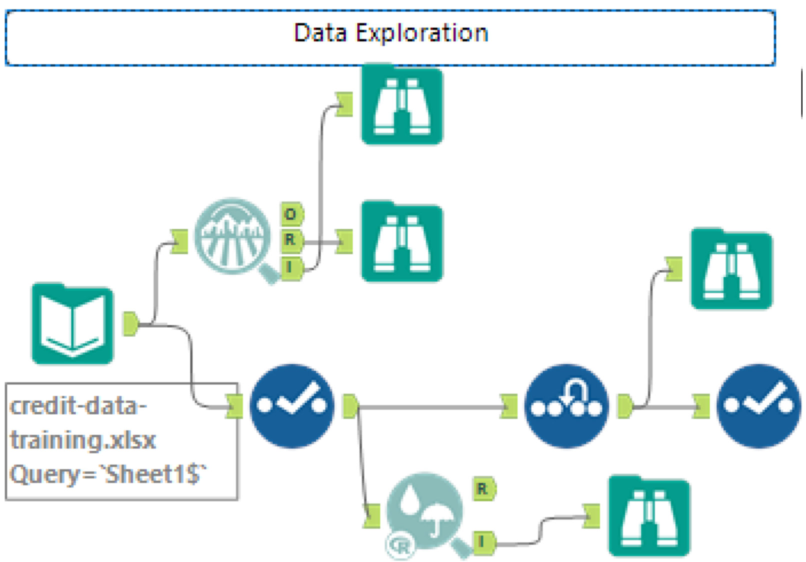
Columns Imputation:

Age of Years (Replace NULL value with Average Age of Years)

Steps:	**Results:**

1. Numerical data fields, are there any fields that highly-correlate with each other?	The Heatmap showed there were no variables highly correlated with other variables. (all correlations were < 0.5)
2. Missing Data	Remove 'Duration in Current Address', with 68.8% missing data. (Conduct imputation for 'Years of Age', replace 2.4% missing data.)
3. Low- Variability Data	'Concurrent-Credits': One value 'Occupation': One value 'Guarantors','Telephone', 'No of Dependents', 'Foreign Workers': High skewed toward one side.

Alteryx Workflow for reference:



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?

Logistic Regression:

Report

Report for Logistic Regression Model Step_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 .

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: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5
Number of Fisher Scoring iterations: 5

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Step_credit	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 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 Step_credit

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

From the logistic regression report, we can see there are following significant Predictor Variables, (with p-value <0.05)

- Account-Balance
- Payment-Status-of-Previous-Credit
- Purpose
- Credit-Amount

- Length-of-current-employment
- Instalment-per-cent

Accuracy: The overall accuracy for the model is 0.76. The confusion matrix exhibited that, for Accuracy of creditworthy is 0.80, and Accuracy of non-creditworthy is 0.63. We can infer that the prediction model is biased, toward predicting Creditworthy.

Decision Tree:

Report

Summary Report for Decision Tree Model Decision_Tree_Credit

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)

Model Summary

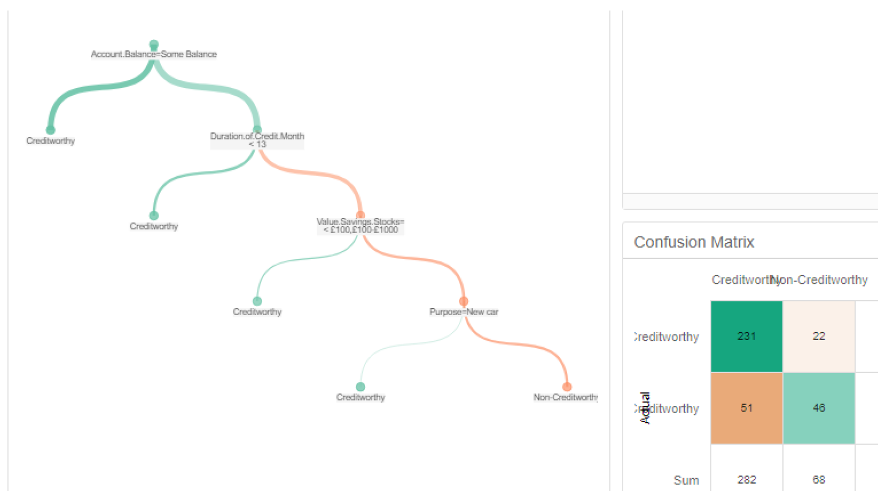
```
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose
[4] Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

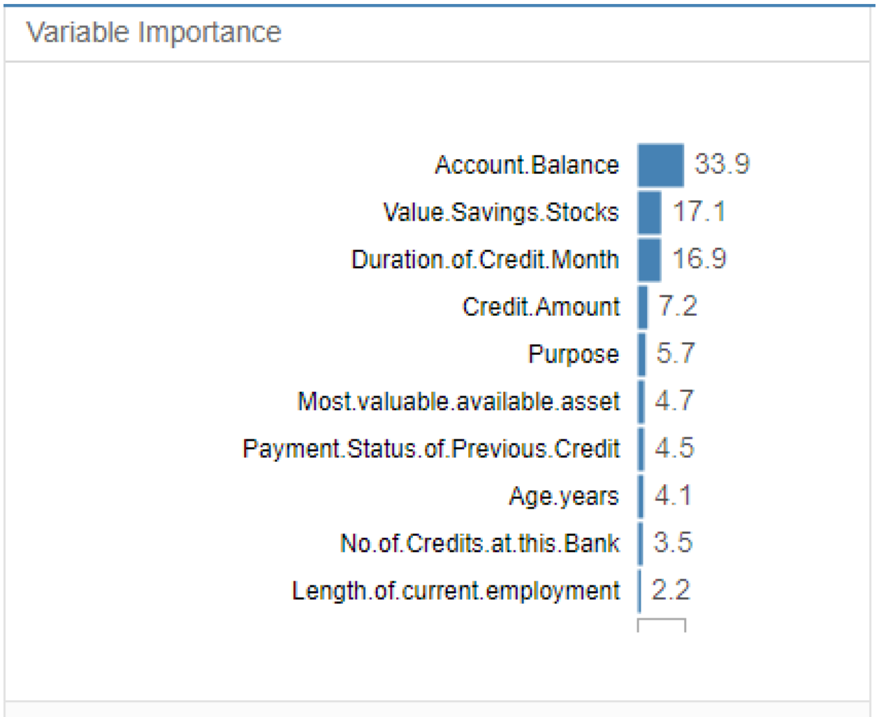
Pruning Table

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	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032

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) *
```





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Credit	0.7467	0.8304	0.7035	0.8857	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 Decision_Tree_Credit

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

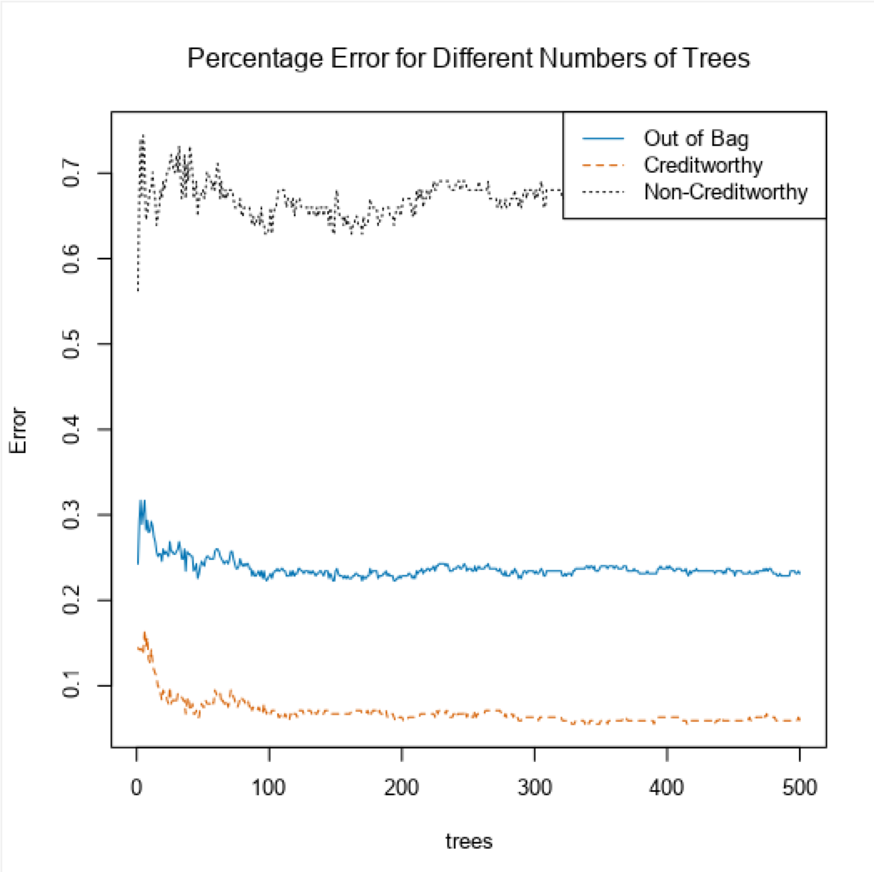
From the Decision Tree report, we can see the three most important variables are:

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

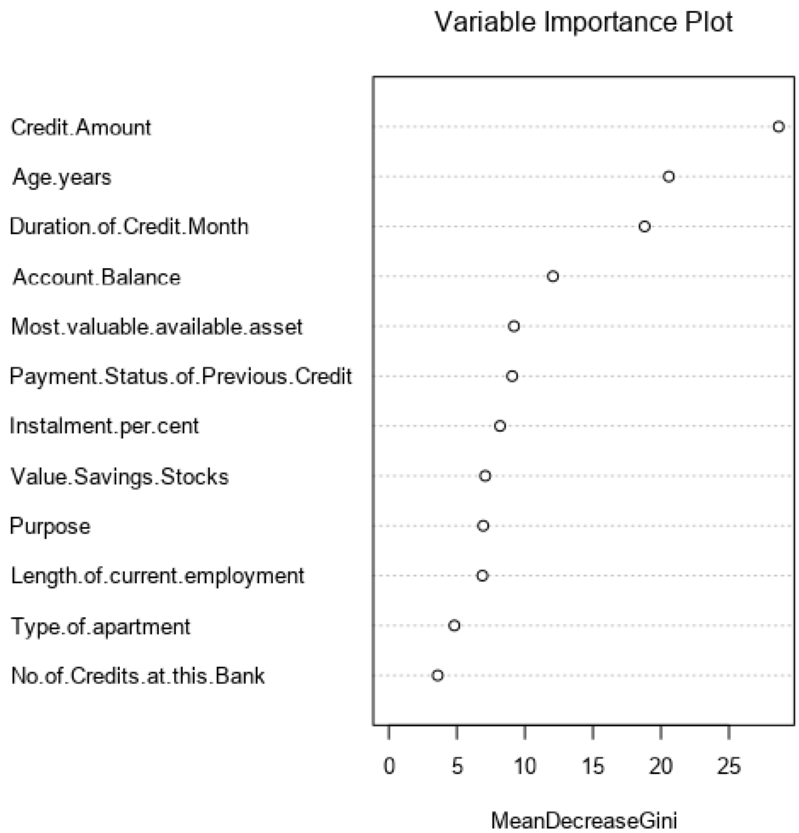
Accuracy: The overall accuracy for the model is 0.75. The confusion matrix exhibited that, for Accuracy of creditworthy is 0.78, and Accuracy of non-creditworthy is 0.61. We can infer that the prediction model is biased, toward predicting Creditworthy.

Forest Model:

Plots



Percentage Error for Different Numbers of Trees
Out of Bag
Creditworthy
Non-Creditworthy
Error
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500
trees



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.7867	0.8644	0.7389	0.9714	0.3556

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 X

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	29
Predicted_Non-Creditworthy	3	16

From the Forest Model report, we can see the three most important variables are:

- Credit-Amount
- Age-years
- Duration-of-Credit-Month

Accuracy: The overall accuracy for the model is 0.79. The confusion matrix exhibited that, for Accuracy of creditworthy is 0.78, and Accuracy of non-creditworthy is 0.84. We can infer that the prediction model is not biased, since the accuracy for predicting creditworthy and non-creditworthy were pretty close.

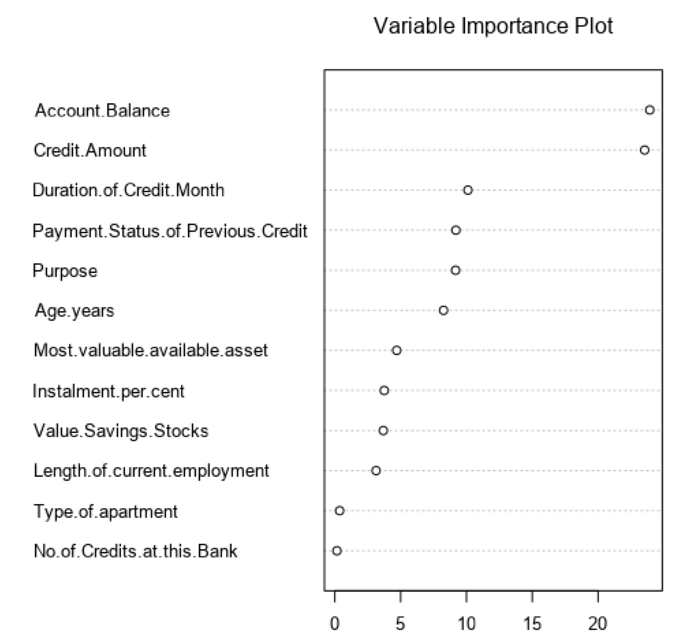
Boosted Mode:

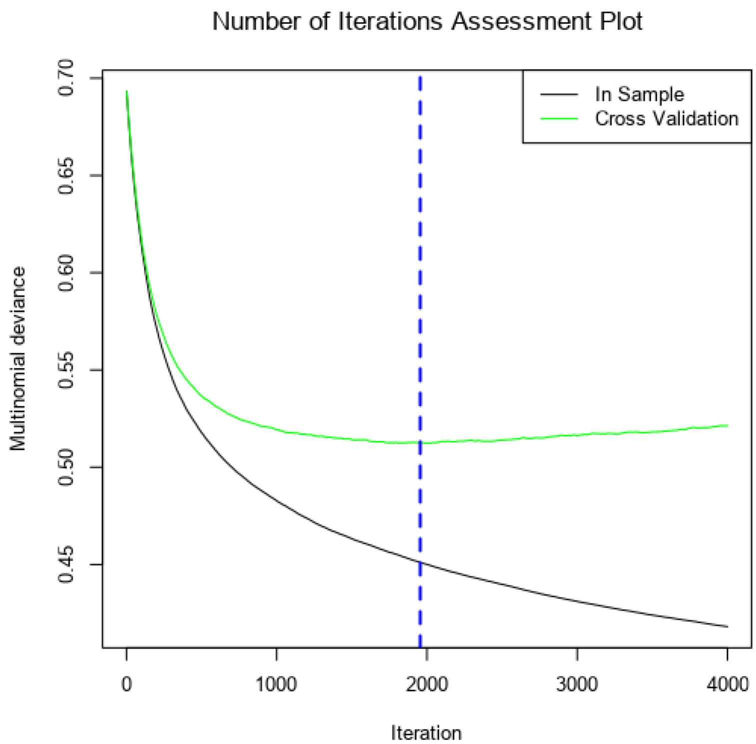
Report for Boosted Model Test_Model

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1955

Plots:





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Test_Model	0.7933	0.8670	0.7539	0.9619	0.4000

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 Test_Model

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

From the Boosted Model report, we can see the three most important variables are:

- Account-Balance
- Credit-Amount

Accuracy: The overall accuracy for the model is 0.79. The confusion matrix exhibited that, for Accuracy of creditworthy is 0.79, and Accuracy of non-creditworthy is 0.82. We can infer that the prediction model is not biased, since the accuracy for predicting creditworthy and non-creditworthy were pretty close.

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)

<u>Which model did you choose to use?</u>

For this question, I would choose the Forest Model to perform my prediction. Because it has overall best Accuracy with 0.79 (slightly less than Boosted Model). Forest Model also has the highest Accuracy within Creditworthy.
Based on our ROC curve, the Forest Model reached the highest True Positive first and has the highest True Positive rate.
From the Confusion Matrices, we can calculate that the accuracy for Creditworthy is 0.78 and accuracy for Non-Creditworthy is 0.84. This results shows that the prediction model is not biased.
Please see the following graphs for reference.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Credit	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Credit	0.7867	0.8644	0.7389	0.9714	0.3556
Boosted_Credit	0.7933	0.8670	0.7539	0.9619	0.4000
Step_credit	0.7600	0.8364	0.7306	0.8762	0.4889

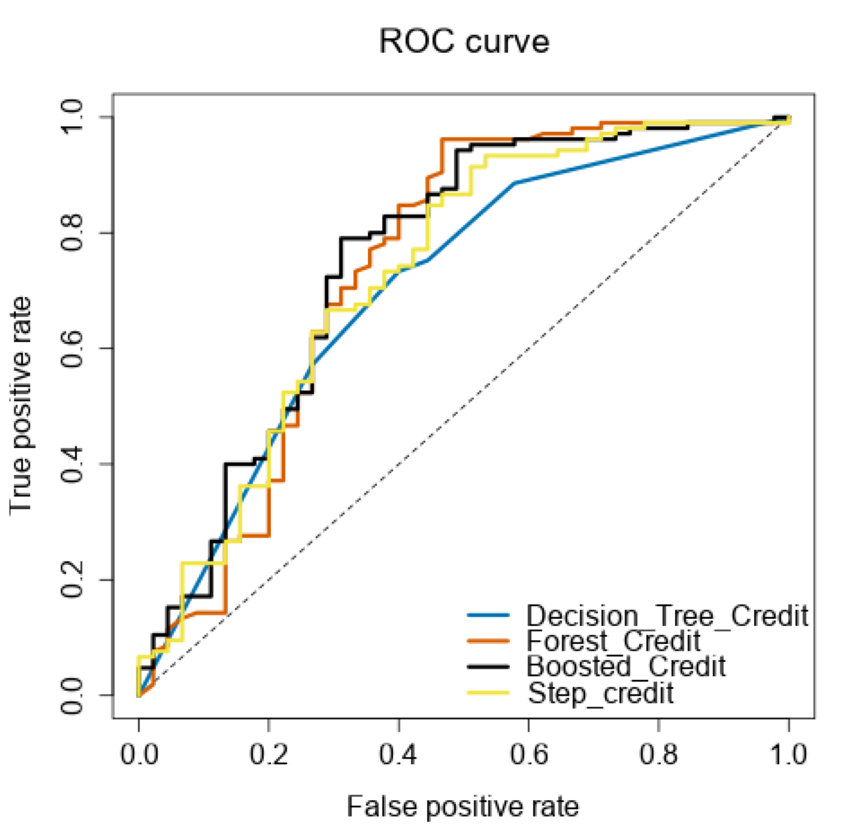
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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 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 Boosted_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of Decision_Tree_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of Forest_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	29
Predicted_Non-Creditworthy	3	16

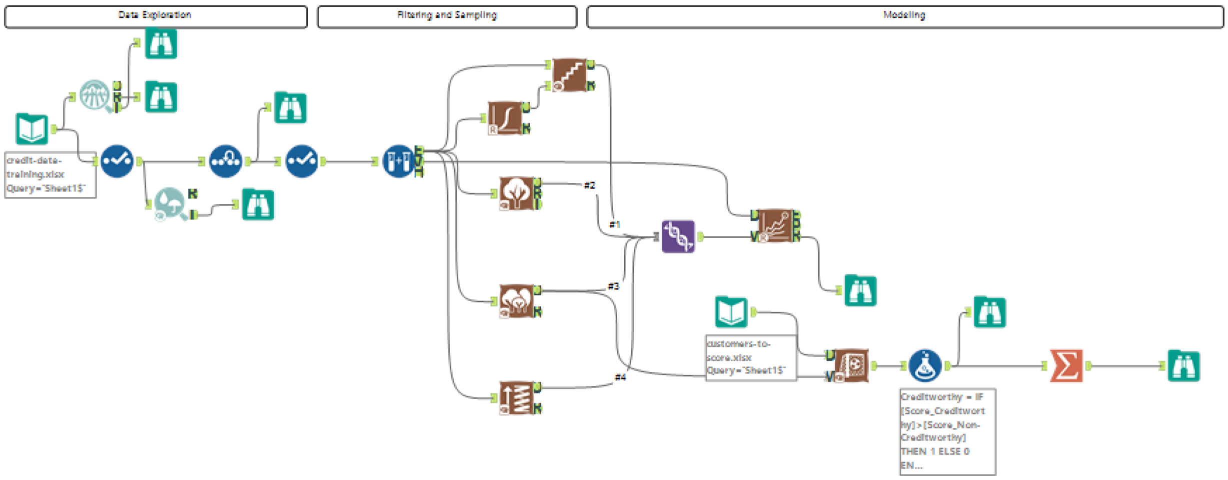
Confusion matrix of Step_credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22



<u>How many individuals are creditworthy?</u>

Based on the Forest Model we built, there would total 412 individuals predicted creditworthy.

Alteryx Workflow for Score:


Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx Query=`Sheet1$`
#1
#2
#3
#4
customers-to-score.xlsx Query=`Sheet1$`
Creditworthy = IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN 1 ELSE 0 EN...