

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 need to be made?

To decide on if a customer is creditworthy to give a loan to.

Then we need to come out with a list of customer who are creditworthy and who are not.

- What data is needed to inform those decisions?

i) Data on all past loan applications.

ii) The list of customers that need to be processed in the next few days.

iii) Data aka predictor variables like Credit Amount, Amount Balance etc.

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

Binary Classification Models.

They are Logistic Regression, Forest Model, Decision Tree and Boosted Model.

We will decide and choose the fittest model.

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

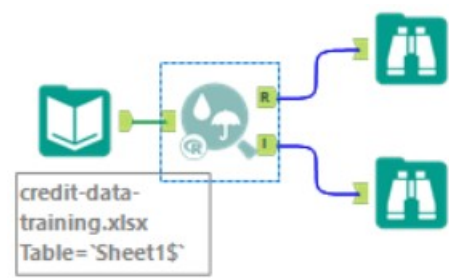
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

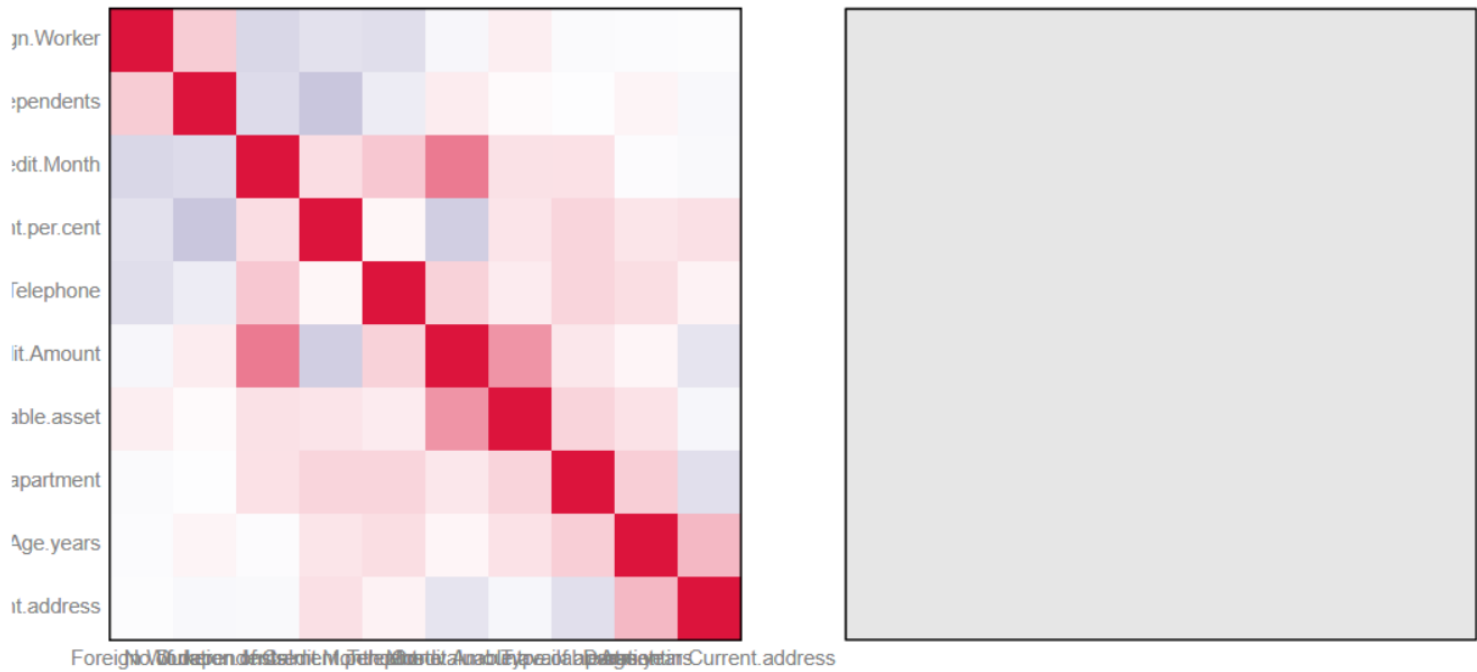
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

Referring to Correlation Matrix, there is no field with correlation value at least 0.70.



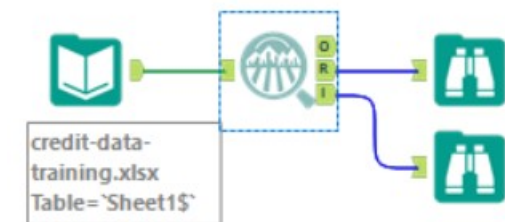
Picture 2.1 Association Analysis Tool.

Correlation Matrix with ScatterPlot

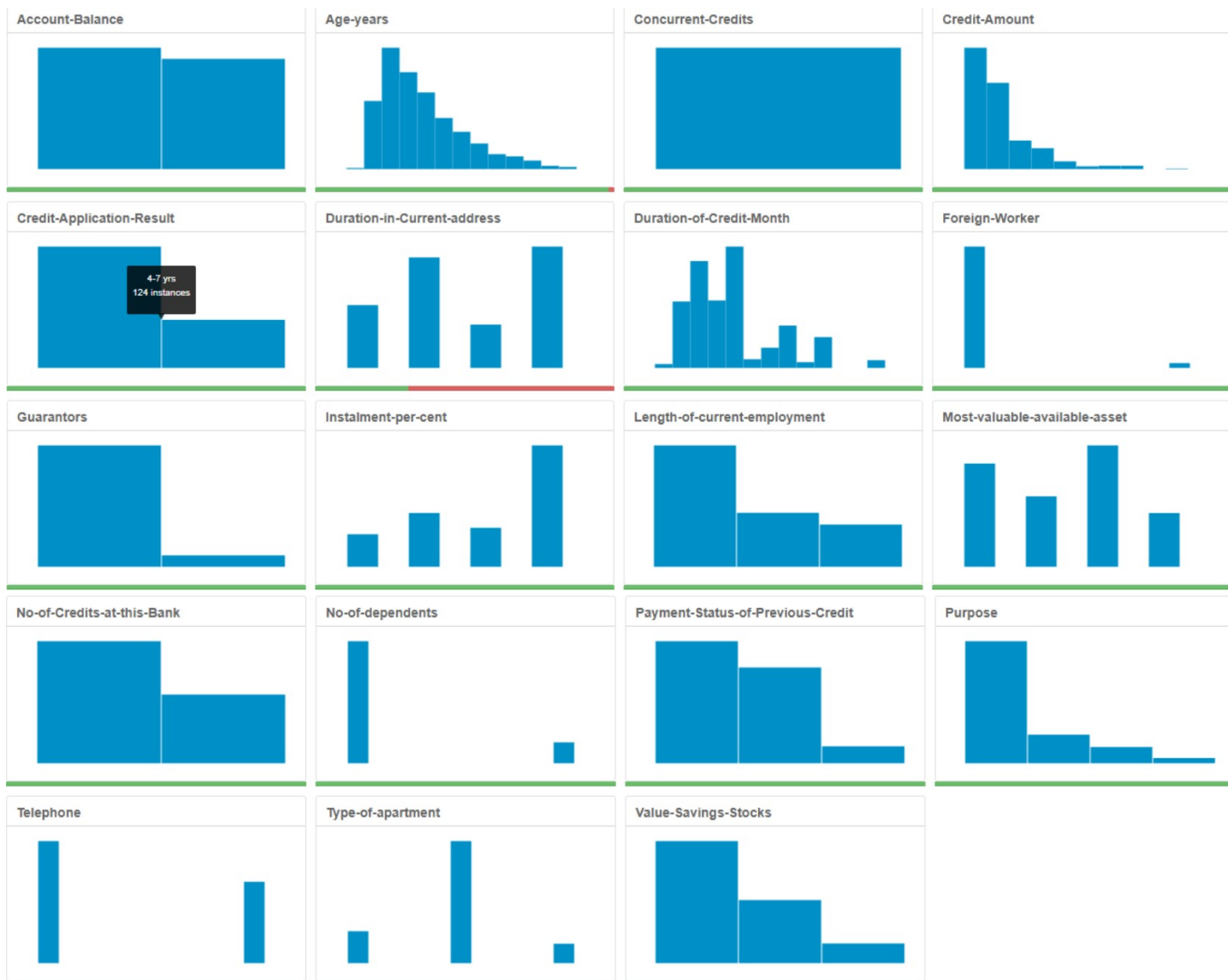


The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated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

Picture 2.2 Correlation Matrix




Picture 2.3 Field Summary tool.



Picture 2.4 Field Summary.

Numeric Fields

Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
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".

Picture 2.5 Field Summary Report (Occupation).

Fields **Age-years** and **Duration-in-Current-address** have 2% and 69% missing data respectively. We will impute field **Age-years** missing data with the Median age because the graph is left-skewed. We will remove field **Duration-in-Current-address** because more than half of the data is missing.

Fields **Concurrent-Credits** and **Occupation** will be removed because they have 1 unique value and do not have significant effect on the model result.

Fields **Guarantors**, **Foreign-Worker** and **No-of-dependents** show low variability because they have more than 80% of the data skewed towards one data.

These can skew our result thus will be removed.

Field **Telephone** have 2 unique values and they show irrelevancy to the result.

This field will be removed.

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)

1. Logistic Regression Model

Fields Account-Balance, Purpose and Credit-Amount are the 3 most significant variables with their p-values less than 0.05.

The overall Logistic Regression Model accuracy is 76.0%.

The accuracy for creditworthy is 80% whereas for non-creditworthy it is 62.8% respectively.

The difference between accuracies is 17.2% which is greater than 10%.

Hence, the model is biased towards predicting customers as Creditworthy.

Report for Logistic Regression Model Stepwise

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)

Picture 2.6 Logistic Regression Stepwise Alteryx-generated result.

Null deviance: 413.16 on 349 degrees of freedom
 Residual deviance: 328.55 on 338 degrees of freedom
 McFadden R-Squared: 0.2048, AIC: 352.5
 Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Picture 2.7 Logistic Regression Stepwise Alteryx-generated result part 2.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predicted to be Class [class name]
 AUC: area under the ROC curve, only available for two-class classification.
 F1: F1 score, precision * recall / (precision + recall)

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

Picture 2.8 Logistic Regression Stepwise Alteryx-generated Model Comparison Report result.

2. Decision Tree

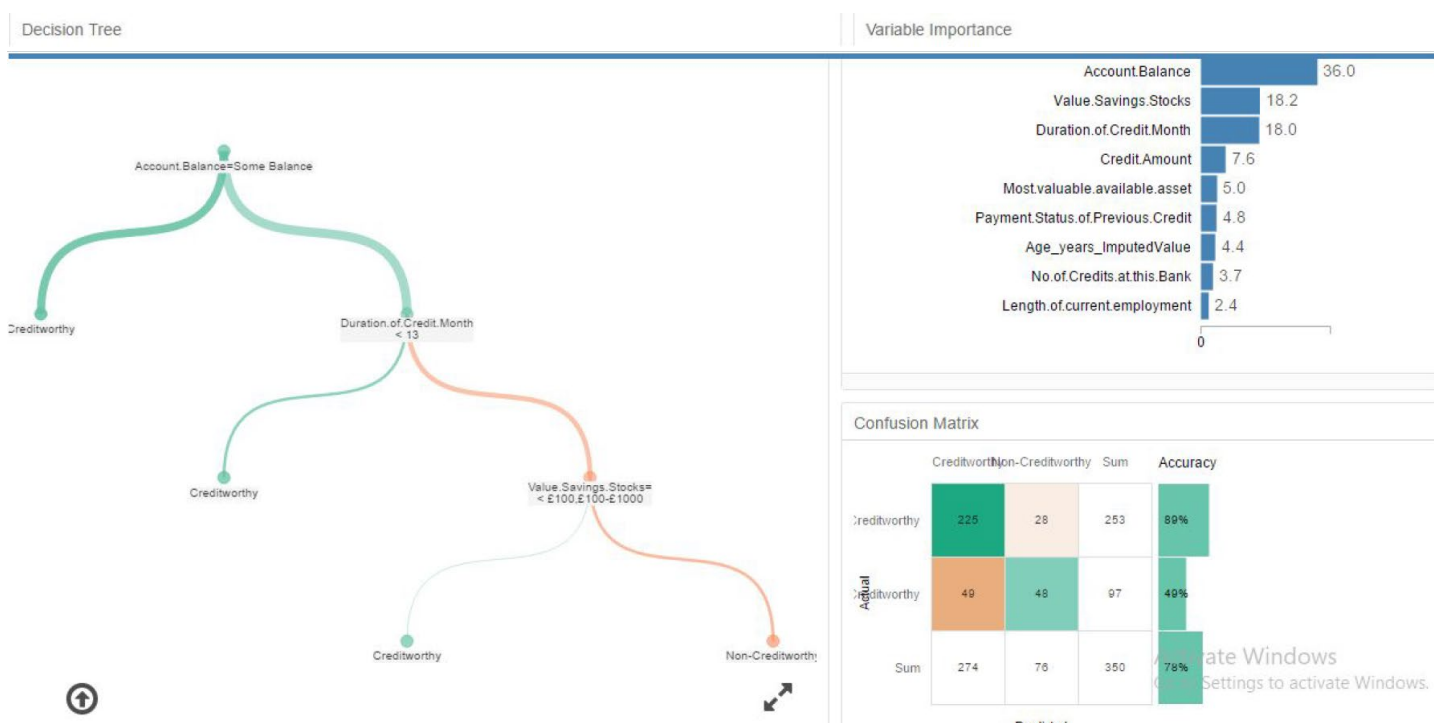
Fields Account-Balance, Value-Savings-Stocks and Duration-of-Credit-Month are the 3 most important variables.

The overall Decision Tree model accuracy is 74.6%.

The accuracy for creditworthy is 79.1% whereas for non-creditworthy is 60.0%.

The difference between accuracies is 19.1% which is greater than 10%.

Hence, the model seems to be biased towards predicting customers as Creditworthy.



Picture 2.9 Decision Tree Report that shows Decision Tree, Variable Importance and Confusion Matrix.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
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], number of samples that are correctly predicted to be Class [class name] divided by number of samples predicted to be Class [class name] AUC: area under the ROC curve, only available for two-class classification. F1: F1 score, precision * recall / (precision + recall)					
Confusion matrix of Decision_tree					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	91		24		
Predicted_Non-Creditworthy	14		21		

Picture 2.10 Decision Tree Alteryx-generated Model Comparison Report

3. Forest Model

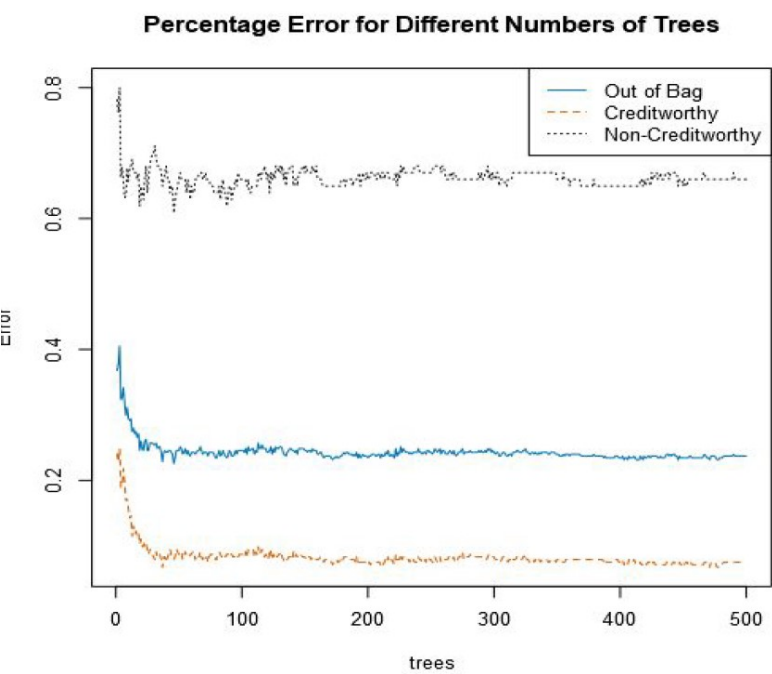
Fields Credit-Amount, Age-Years and Duration-of-Credit-Month are the 3 most important variables.

The overall Forest Model accuracy is 80.0%.

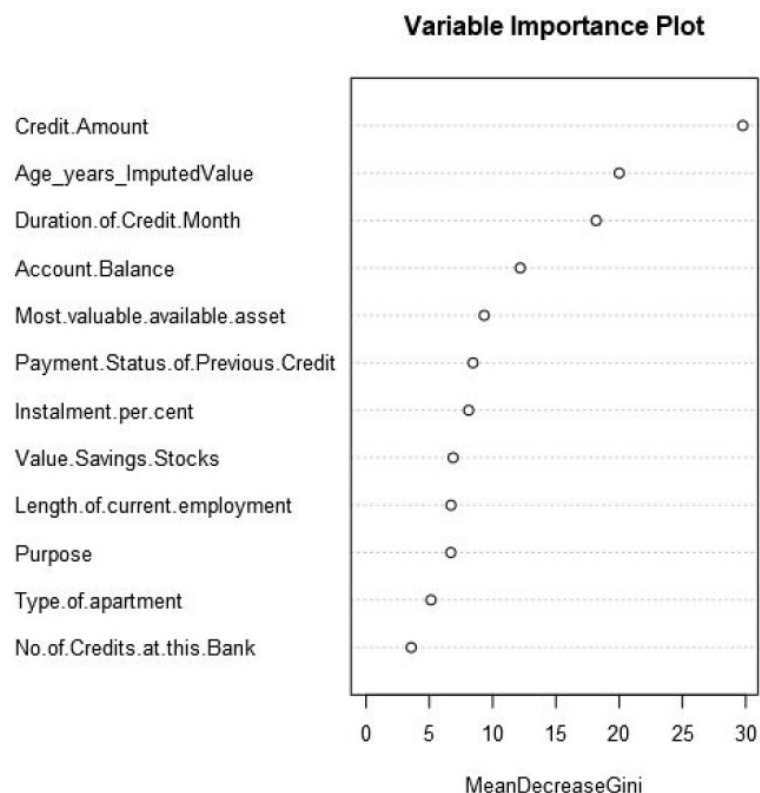
The accuracy for creditworthy is 79.5% and for non-creditworthy it is 82.6%.

The difference of the accuracies is 3.1% which is lower than 10%.

Hence, the model is not biased to predicting customer as creditworthy or non-creditworthy.



Picture 2.11 Forest Model Percentage Error Report



Picture 2.12 Forest Model Variable Importance Plot Report

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
<p>Model: model names in the current comparison.</p> <p>Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.</p> <p>Accuracy_[class name]: accuracy of Class [class name], number of samples that are correctly predicted to be Class [class name] divided by number of samples predicted to be Class [class name]</p> <p>AUC: area under the ROC curve, only available for two-class classification.</p> <p>F1: F1 score, precision * recall / (precision + recall)</p>					
Confusion matrix of Forest_Model					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	101		26		
Predicted_Non-Creditworthy	4		19		

Picture 2.13 Forest Model Model Comparison Report

4. Boosted Model

Fields Account-Balance and Credit-Amount are the most significant variables.

The overall Boosted Model accuracy is 78.6%.

The accuracies for creditworthy is 78.2% and for non-creditworthy it is 80.9%.

The difference between the accuracies is 2.7% which is less than 10%.

Hence, the model is not biased to predicting customer as creditworthy or non-creditworthy.

Report for Boosted Model Boosted_Model

Basic Summary:

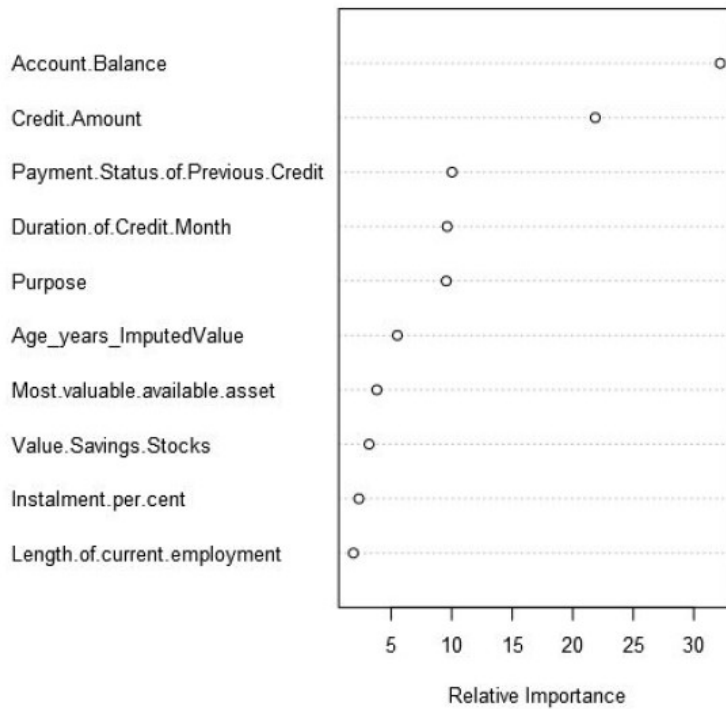
Loss function distribution: Bernoulli

Total number of trees used: 4000

Best number of trees based on 5-fold cross validation: 2036

Picture 2.14 Boosted Model Summary Report

Variable Importance Plot



The Variable Importance Plot provides information about the relative importance of each predictor field. The measures are normalized to sum to 100, and the value for each field gives the relative percentage importance of that field to the overall model.

Picture 2.15 Boosted Model Alteryx-generated Variable Importance Plot Report.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predicted to be Class [class name]

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, precision * recall / (precision + recall)

Confusion matrix of Boosted_Model

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

Picture 2.16 Boosted Model Alteryx-generated Model Comparison Report.

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

I choose Forest Model.

It has the highest overall accuracy of 80.0% against the validation set.

It has the highest accuracy in Creditworthy (79.5%) and Non-creditworthy (82.6%).

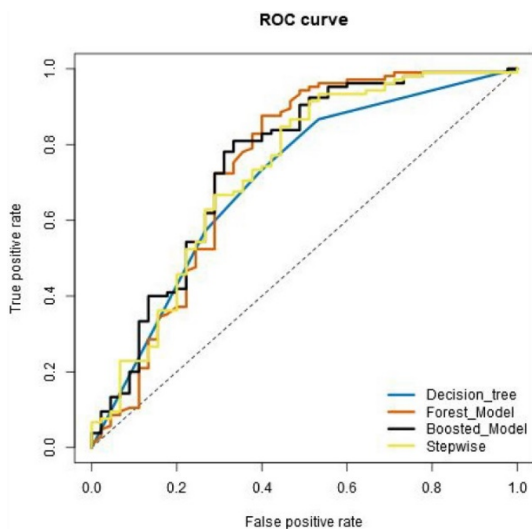
It reaches the true positive rate at the fastest rate.

The accuracy difference between creditworthy and non-creditworthy is small thus makes it least bias towards predicting any decisions.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

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], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predicted to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

Picture 2.17 All Models Comparison Report

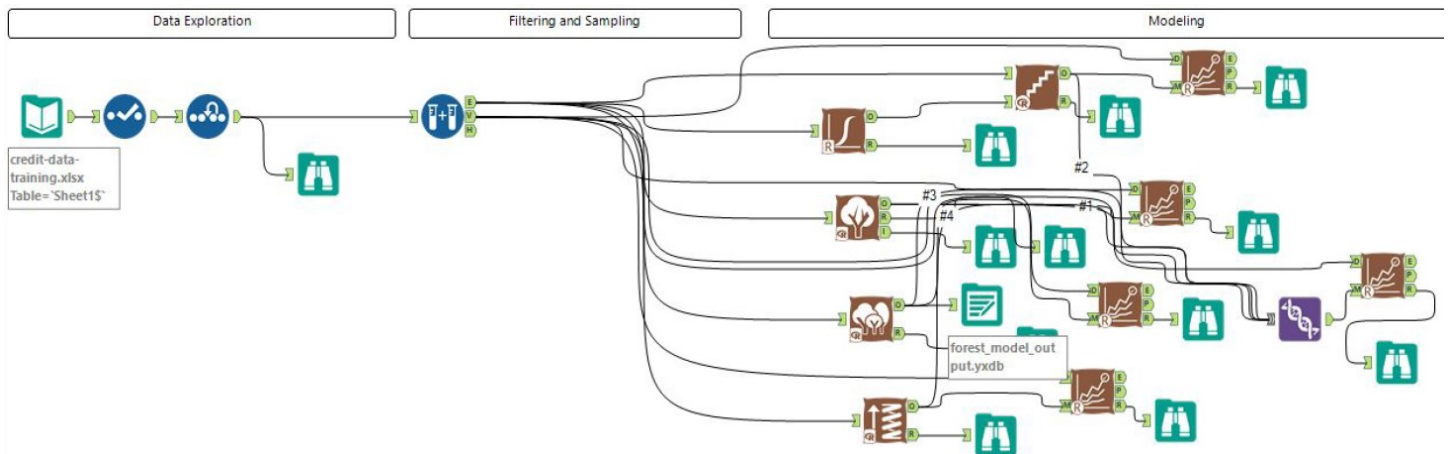


Picture 2.18 All models ROC graph

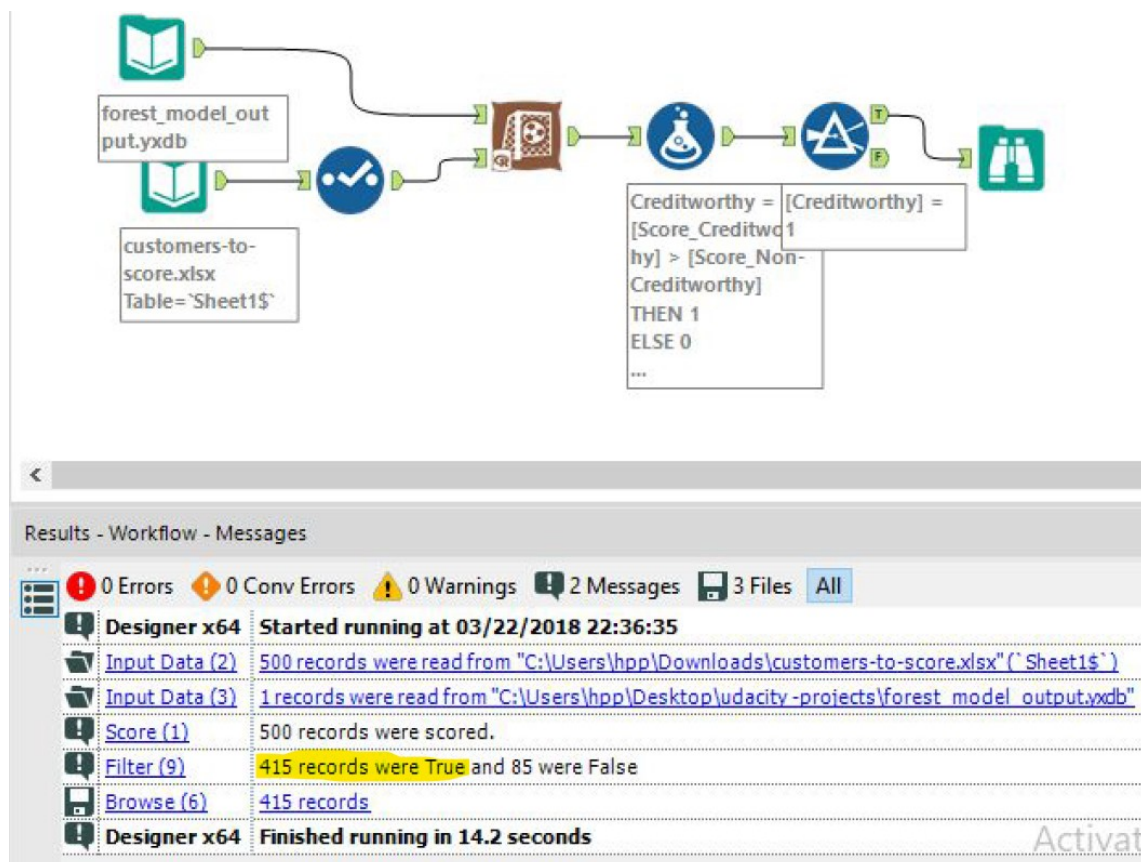
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

Using Forest Model to predict creditworthiness of the new customers, there are 415 creditworthy customers.



Picture 2.19 Alteryx workflow for all models.



Picture 2.20 Alteryx workflow to predict how many customers are 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.