Project 4: Creditworthiness

The Business Problem

You work for a small bank and are responsible for determining if customers are creditworthy to give a loan to. Your team typically gets 200 loan applications per week and approves them by hand.

Due to a financial scandal that hit a competitive bank last week, you suddenly have an influx of new people applying for loans for your bank instead of the other bank in your city. All of a sudden you have nearly 500 loan applications to process this week!

Your manager sees this new influx as a great opportunity and wants you to figure out how to process all of these loan applications within one week.

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

The decision needs to be made whether an applicant is creditworthy of loan approval. How many applicants are creditworthy?

- What data is needed to inform those decisions?

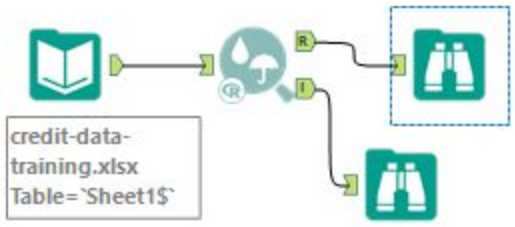
Data on past applications such as Account Balance and Credit Amount and list of applicants to be processed are required to make decisions.

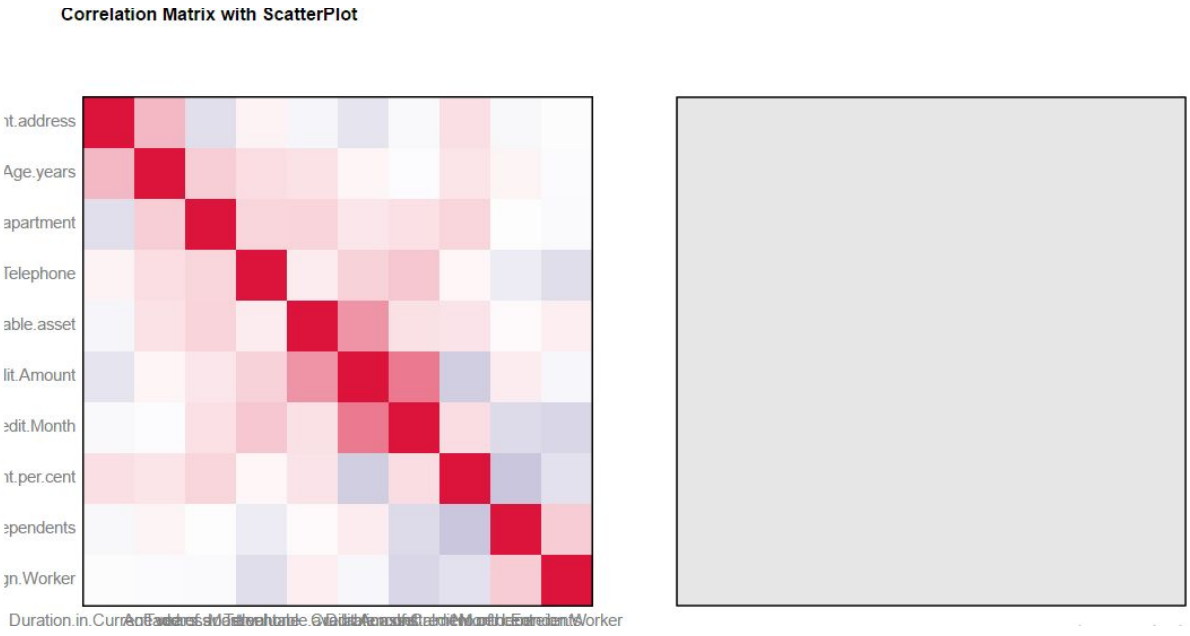
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Binary Classification Models (Logistic Regression, Forest Model, Decision Tree and Boosted Tree Model) are needed for analysis.

Step 2: Building the Training Set

An association analysis is performed in Alteryx and there are no variables which are highly correlated with each other.



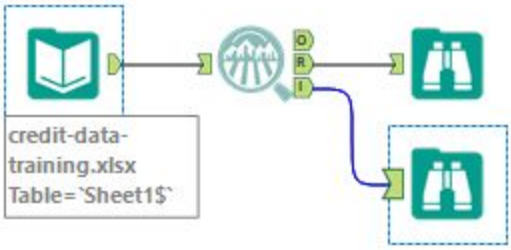


Field summary tool Results:

Duration in Current Address has 69% missing data and should be removed. While Age Years has 2% missing data, it is appropriate to impute the missing data with the median age. Median age is used instead of mean as the data is skewed to the left.

Concurrent Credits and Occupation have 1 value while Guarantors, Foreign Worker and No of Dependents show low variability where more than 80% of the data skewed towards one data. These data should be removed in order not to skew our analysis results.

Telephone field should also be removed due to its irrelevance to the decision to be made.


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


Account-Balance
Concurrent-Credits
Credit-Application-Result
Guarantors
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit
Purpose
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

1. Logistic Stepwise Regression

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

Overall accuracy is 76.0% while accuracy for creditworthy is higher than non-creditworthy at 80.0% and 62.8% respectively. The difference between accuracies 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)

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

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 predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

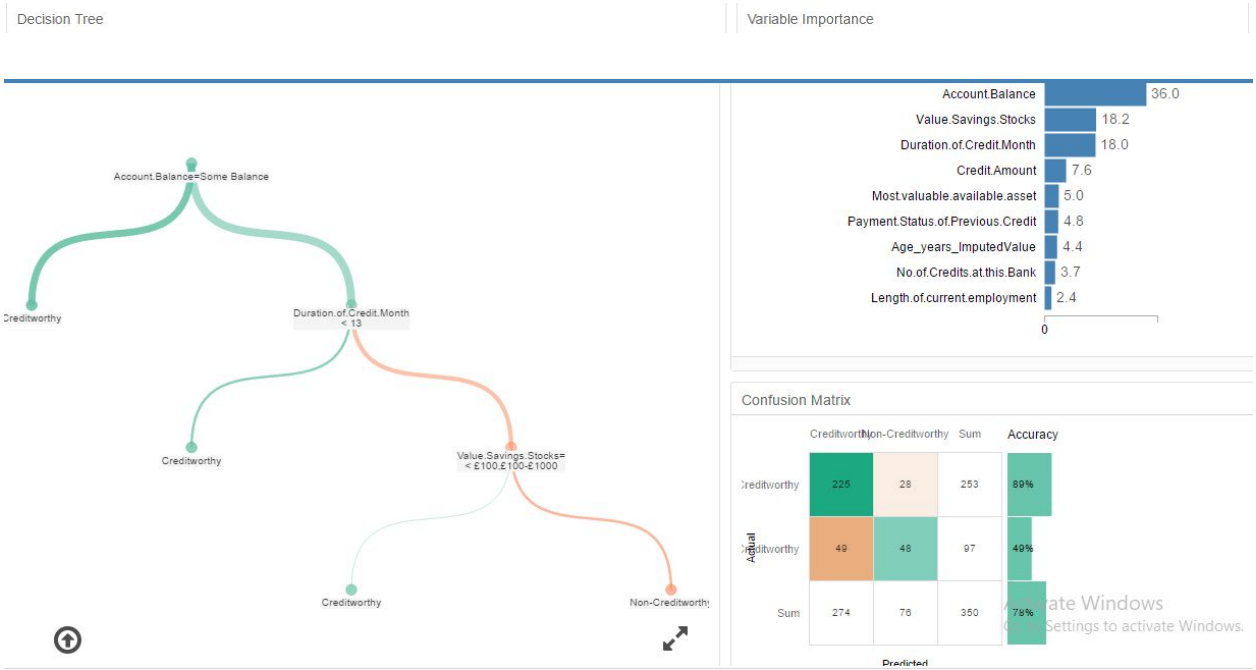
Confusion matrix of Stepwise

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

2. Decision Tree

Account Balance, Value Savings Stocks and Duration of Credit Month are the 3 most important variables. The overall accuracy is 74.6%.

Accuracy for creditworthy is 79.1% while accuracy for non-creditworthy is 60.0%. The difference between accuracies is greater than 10%. The model seems to be biased towards predicting customers as Creditworthy.



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 predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

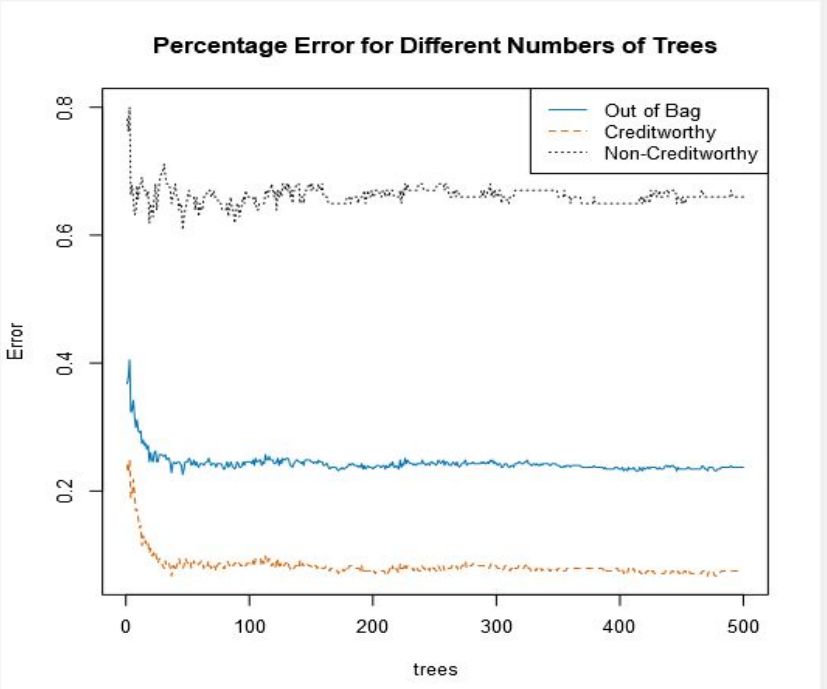
Confusion matrix of Decision_tree

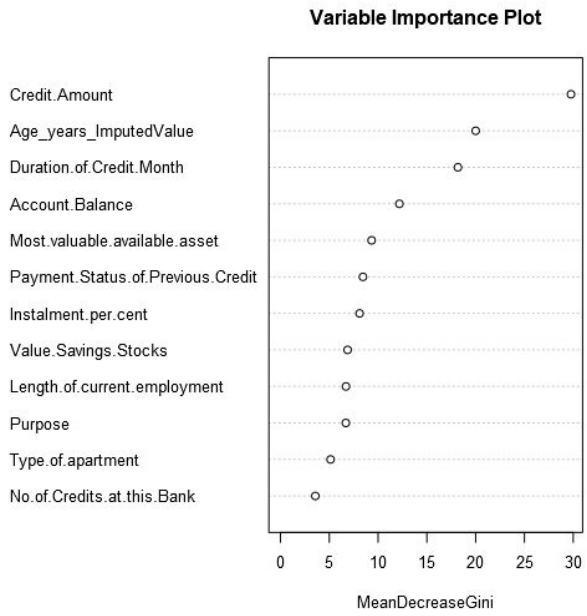
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

3. Forest Model

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

Overall accuracy is 80.0%. The model is not biased as the accuracies for creditworthy and non-creditworthy are 79.5% and 82.6% respectively, which are comparable.





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

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 predited 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 Forest_Model

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

Performance Diagnostic Plots

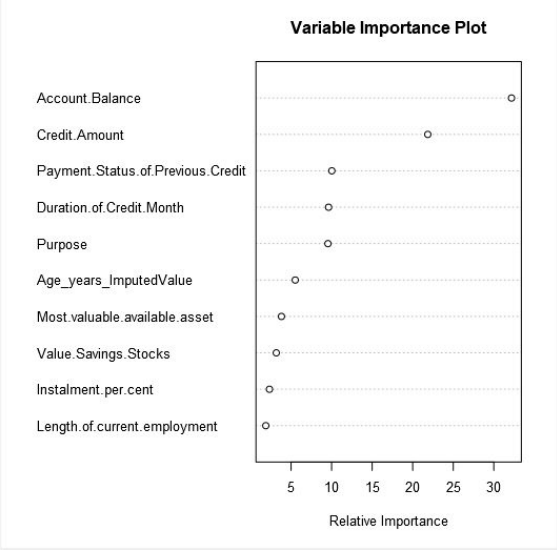
4. Boosted Model

Account Balance and Credit Amount are the most significant variables. Overall accuracy is 78.6%. Accuracies for creditworthy and non-creditworthy are 78.2% and 80.9% respectively which shows a lack of bias in predicting whether customers are creditworthy or not.

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



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.

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 predited 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

Step 4: Writeup

Forest model is the best choice.

It gives the highest accuracy at 80% against validation set.

Its accuracies for creditworthy and non-creditworthy are among the highest of all other models.

Forest model reaches the true positive rate at the fastest rate.

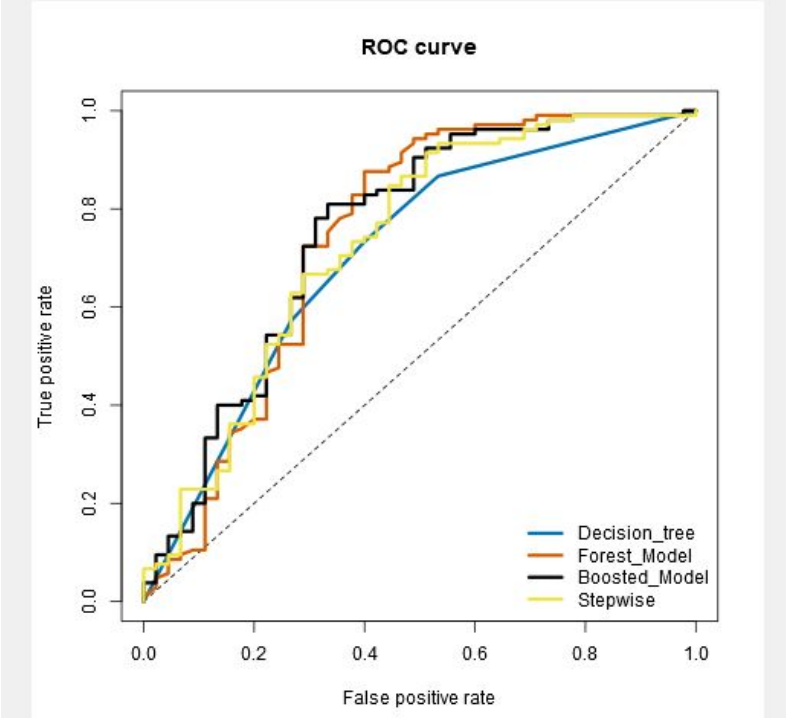
The accuracy difference between creditworthy and non-creditworthy is small which makes it least bias towards any decisions.

Model Comparison Report

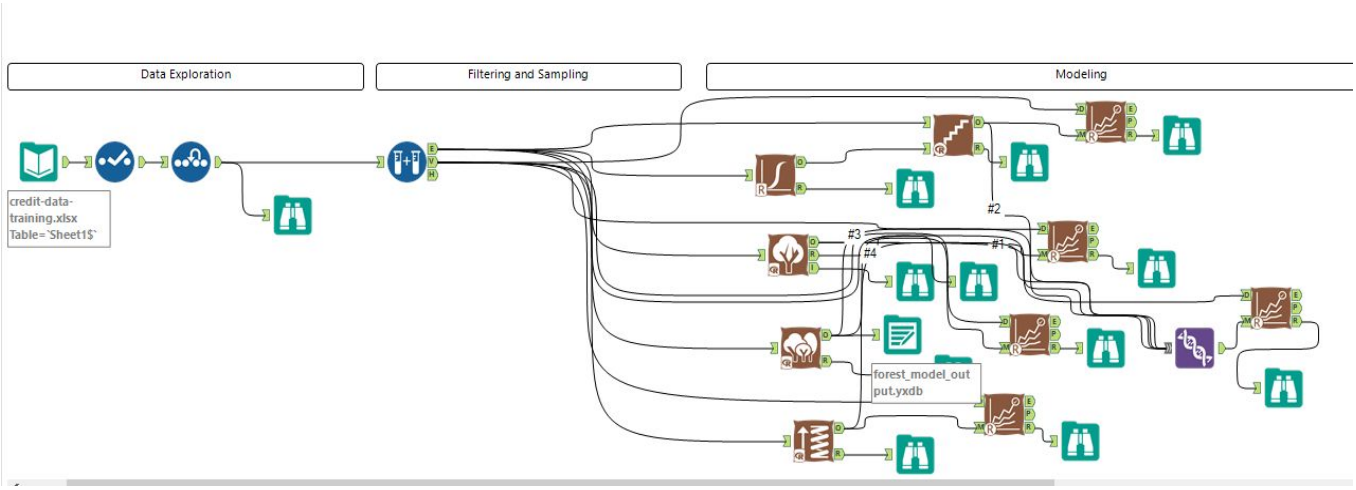
Fit and error measures

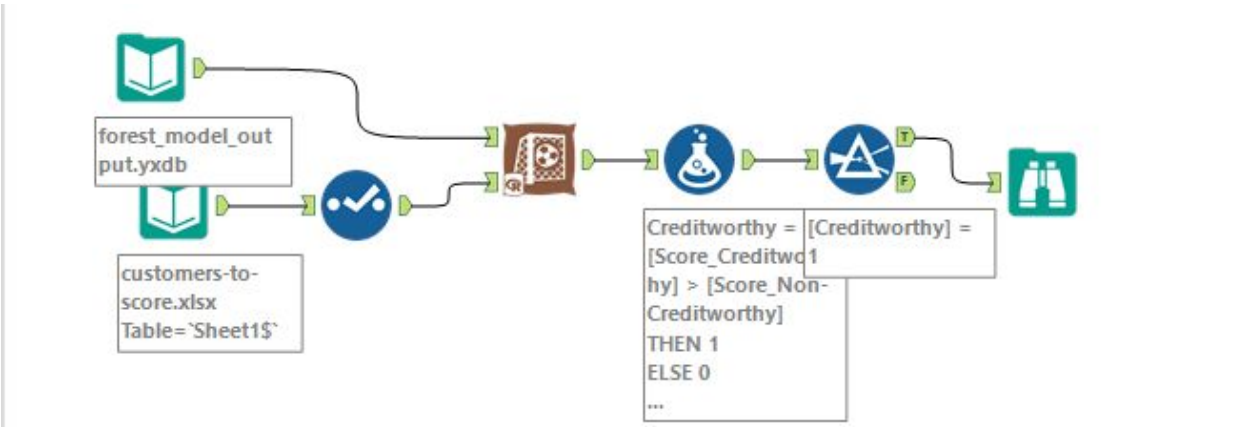
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_tree	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Model	0.8000	0.8707	0.7419	0.7953	0.8261
Boosted_Model	0.7867	0.8632	0.7524	0.7829	0.8095
Stepwise	0.7600	0.8364	0.7306	0.8000	0.6286

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 predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)



There are **415 creditworthy customers** using forest model to score new customers.





Results - Workflow - Messages

0 Errors | 0 Conv Errors | 0 Warnings | 2 Messages | 3 Files | All

Designer x64	**Started running at 03/22/2018 22:36:35**
Input Data (2)	500 records were read from "C:\Users\hpp\Downloads\customers-to-score.xlsx"(`Sheet1$`)
Input Data (3)	1 records were read from "C:\Users\hpp\Desktop\udacity -projects\forest_model_output.yxdb"
Score (1)	500 records were scored.
Filter (9)	415 records were True and 85 were False
Browse (6)	415 records
Designer x64	**Finished running in 14.2 seconds**