Predicting Default Risk (Creditworthiness)

Business and Data Understanding

- What decisions needs to be made?
  The decisions that needs to be made are whether a new customer could be approved for a loan or not.

- What data is needed to inform those decisions?
  Data on all past applications and list of customers that need to be processed in the next few days.

- 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, decision tree, forest model, boosted model) are needed to help make these decisions

Building the Training Set

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

  Duration in Current Address has 69% of the data missing. Since fields with a lot of missing data should be removed this variable has been removed.

  The histography of the variable **Guarantors, Foreign-worker and No-of-dependents** shows that majority of the data is heavily skewed towards one type of data. Also, **Concurrent Credits** and **Occupation** have that are entirely uniform and there are no other variations of the data. All these variables have been removed due to low variability.

  Telephone does not have any predictive ability to the credit application result, so this field should also be removed.

  Age-Years has 2% of the data missing. The missing data of this variable has been imputed using the median, 33 of the entire data field.

Train your Classification Models

Due to the update of the Model Comparison tool, I calculated the accuracy for creditworthy and noncreditworthy using the confusion matrix. (Formula has been shown)

1. Logistic Regression (Stepwise)

For this logistic regression (stepwise) model, Account Balance, Payment status of Previous Credit, and Purpose are three of the most significant variables. The overall accuracy is 76%. Using the confusion matrix,

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 92/(92+23) = 0.8, 80% while

accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 22/(13+22) = 0.6286, 62.86%

The model seems to be slightly biased towards predicting customers as non-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, Akaike Information Criterion 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	0.8762	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 Stepwise

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

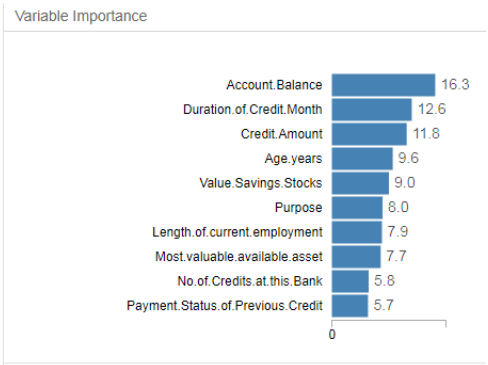
2. Decision Tree

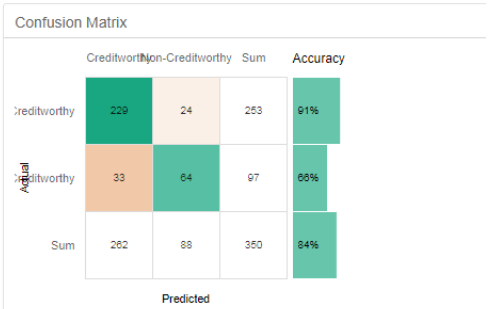
For this decision tree model, Account Balance, Duration of Credit Month, and Value Saving Stocks are three of the most significant variables. The overall accuracy is 67.33%. Using the confusion matrix,

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 83/(83+27) = 0.7545, 75.45% while

accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 18/(22+18) = 0.45, 45%

The model seems to be biased towards predicting customers as non-creditworthy.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.6733	0.7721	0.6296	0.7905	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 Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

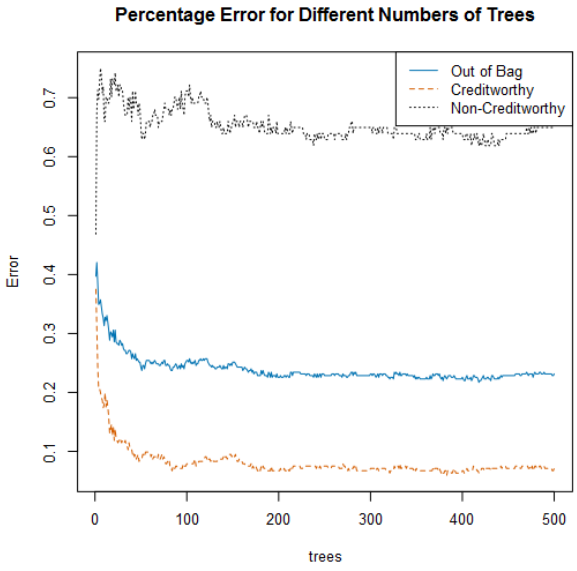
3. Forest Model

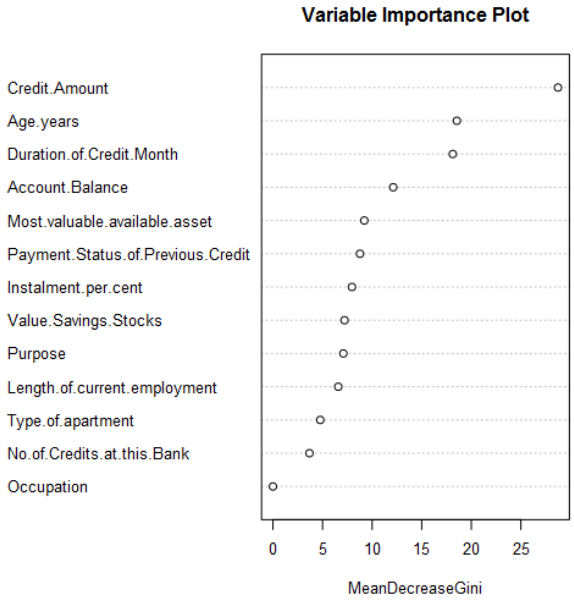
For this forest model, Credit Amount, Age Years and Duration of Credit Month are three of the most significant variables. The overall accuracy is 80.0%.

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 102/(102+27) = 0.7907, 79.07% while

accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 18/(3+18) = 0.8571, 85.71%

Since accuracies for creditworthy and non-creditworthy are comparable 79.07% and 85.71% respectively, this model isn't biased.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8000	0.8718	0.7419	0.9714	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 Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

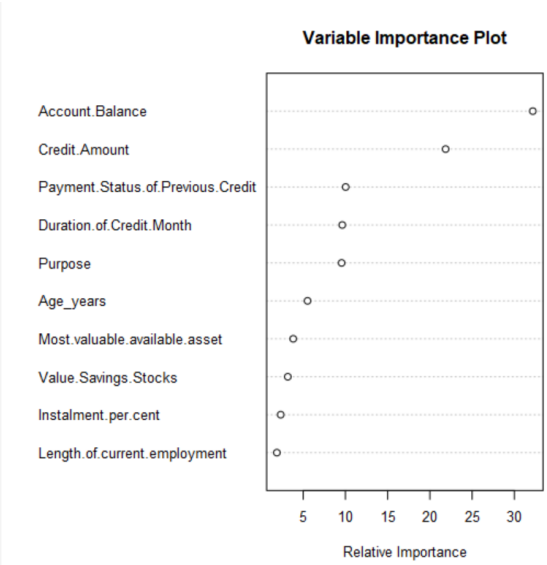
4. Boosted Model

For this boosted model, Account Balance, Credit Amount and Payment status of Previous Credit are three of the most significant variables. The overall accuracy is 78.67%.

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 101/(101+28) = 0.7829, 78.29% while

accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 17/(4+17) = 0.8095, 80.95%

Since accuracies for creditworthy and non-creditworthy are comparable 78.29% and 80.95% respectively, this model isn't biased.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7513	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 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_Model

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

Writeup

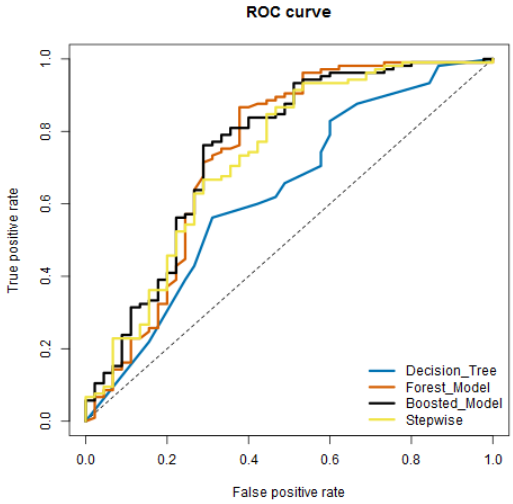
- Which model did you choose to use?

  Forest Model has been chosen since it has the highest accuracy of 80% among all four classification models. Also accuracies for creditworthy and non-creditworthy are among the highest of all.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.6733	0.7721	0.6296	0.7905	0.4000
Forest_Model	0.8000	0.8718	0.7419	0.9714	0.4000
Boosted_Model	0.7867	0.8632	0.7513	0.9619	0.3778
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Forest model reaches the top true positive rate the quickest and overall the highest the most.



Using the confusion matrix of the Forest Model,

accuracy for creditworthy = actual creditworthy / (predicted creditworthy) = 102/(102+27) = 0.7907, 79.07% while

accuracy for non-creditworthy = actual non-creditworthy / (predicted non-creditworthy) = 18/(3+18) = 0.8571, 85.71%

Accuracies for creditworthy and non-credit worthy are 79.07% and 85.71% which shows a lack of bias in predicting whether customers are creditworthy or not.

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of Stepwise

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

- How many individuals are creditworthy?
  There are 417 creditworthy new customers that we could approve for a loan and 83 non-creditworthy customers that should not be approved for a loan.

Record #	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	417	83

Alteryx Workflow