Default Risk Prediction – Loan Approval

Step 1: Business and Data Understanding

- What decisions needs to be made?
  The decision need to be made is to identify customers who are creditworthy for loan approval.

- What data is needed to inform those decisions?
  Spread Sheet Column Names: 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 percent, most valuable available assent, age years, type of apartment, no of credits at this bank.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary model, since determine if a customer is creditworthy or not is a classification type of analysis where we are determine creditworthy or not.

Step 2: Building the Training Set

1. From figure 1. the Pearson correlation matrix, for the numerical data fields, there is no data fields that highly-correlate with each other, since the correlation should be at least 0.70 to be considered “high”.

Full Correlation Matrix

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

Figure 1. Correlation Matrix

2. Duration-in-current-address column has 69% missing data, so this column should be removed. Age years column has 2% missing data, since the numbers of missing data are much less, we should consider replacing the nulls with median of the column.

Figure 2. Field Summary for Numeric Variables

3. Low variability data column such as Guarantors, Foreign-Worker, and No-of-dependents should be removed, because the data field is heavily skewed to one type of data. The Occupation and Concurrent-Credits column should also be removed due to the data is entirely uniform and there is no other variations. Telephone should be removed because it is irrelevant to the analysis of customers' creditworthy.
4. In conclusion, 7 columns should be removed and 13 columns left: Concurrent-Credits, Guarantors, Duration-in-current-address, Foreign-Worker, No-of-dependents, Occupation, and Telephone.

Step 3: Train your Classification Models

Model 1: Logistic regression (Stepwise)

Record	Report
1	**Report for Logistic Regression Model Logistics_Stepwise**
2	*Basic Summary*
3	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)
4	Deviance Residuals:
5	(see table below)
6	Coefficients:
7	(see table below)
8	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
9	Number of Fisher Scoring iterations: 5
10	*Type II Analysis of Deviance Tests*

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)

Figure 3. Logistic Regression (Stepwise) Report

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Logistics_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 Logistics_Stepwise

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

Figure 4. Logistic Regression (Stepwise) Validation Report

The figure 3. report for the logistic regression shows the most significant predictor variables are Account Balance, Purpose, and Credit Amount. The p-value for the three variables are all <0.05 After validating the model, figure 4 shows the overall accuracy is 76%, while the model is good when predicting that a customer is creditworthy 80%, it is only 63% accurate when predicting that a customer is not creditworthy. Many creditworthy individuals would be denied to a loan as it classifies many creditworthy applicants as non-creditworthy. So this model is biased towards predicting individuals who are creditworthy, as it does not predict individuals who are not creditworthy nearly at the same level as those who are.

Model 2: Decision Tree

Figure 5. Decision Tree Model Report

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

Figure 6. Decision Tree Model Validation Report

Figure 5 shows the most significant predictor variables for the decision tree model are Account Balance, Value savings stocks, and duration of credit month.
From the figure 6, the overall accuracy of the Decision Tree model is 74.67%, accuracy for the creditworthy customers and non-creditworthy customers are 79.13% and 60% respectively. Biased towards non-creditworthy.

Model 3: Forest Model

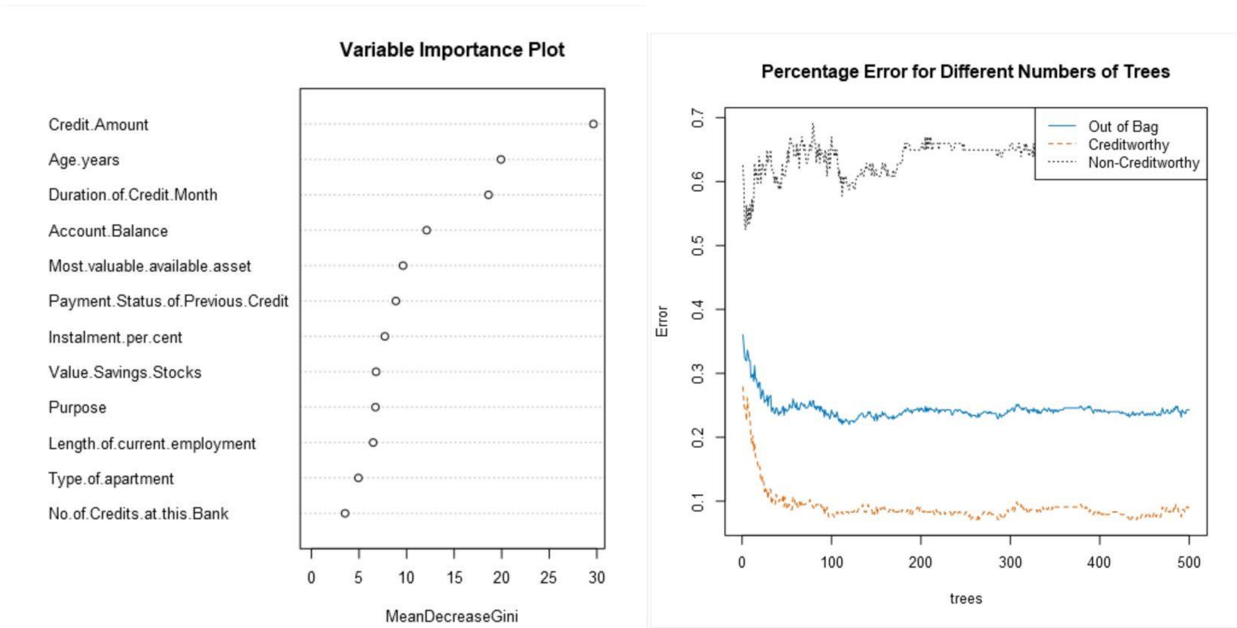


Figure 7. Forest Model Report

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.8067	0.8755	0.7392	0.7969	0.8636

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)

3

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

Figure 8. Forest Model Validation Report

For the Forest Tree model, figure 7 shows the most significant predictor variables are Credit Amount, Age, and Duration of credit month.

Figure 8 shows the model overall accuracy is 80.67%, accuracy for the creditworthy customers and non-creditworthy customers are 79.69% and 86.36% respectively, when a model predicted whether an individual was creditworthy or not at almost an equal percentage, that means it indicates little to no bias.

Model 4: Boosted Model

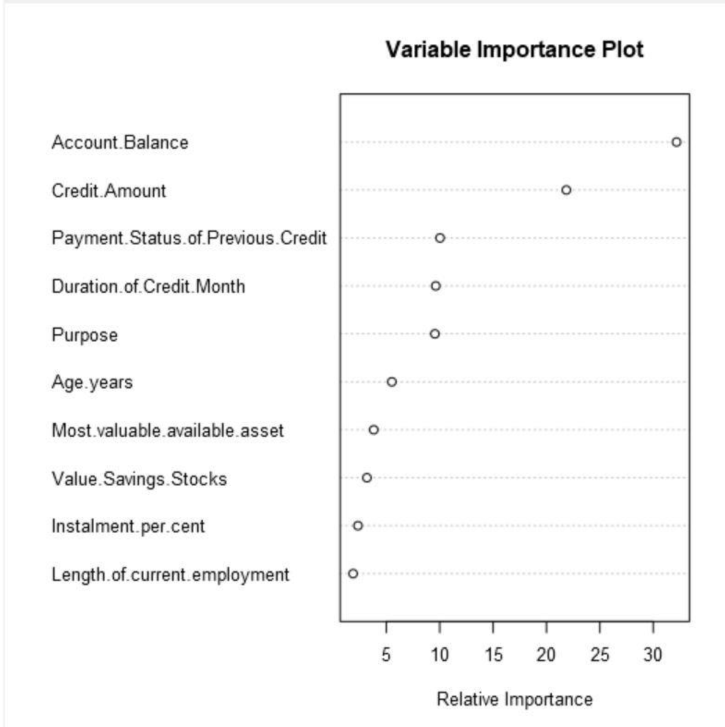


Figure 9. Boosted Model Variable Importance Plot

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

Figure 10. Boosted Model Validation Report

Figure 9 shows the most significant predictor variables for the boosted model are Account balance, Credit amount and Payment.
Figure 10 shows the overall accuracy for the model is 78.67%, and is not biased because the accuracy for the creditworthy and non-creditworthy customers are 78.29% and 80.95% repectively.

Step 4: Writeup

1

Model Comparison Report

2

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	0.8067	0.8755	0.7392	0.7969	0.8636
Boosted_model	0.7867	0.8632	0.7524	0.7829	0.8095
Logistics_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)

3

Confusion matrix of Boosted_model

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

4

Confusion matrix of Decision_Tree

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

5

Confusion matrix of Forest_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	26
Predicted_Non-Creditworthy	3	19

6

Confusion matrix of Logistics_Stepwise

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

Figure 11. Four Model Comparison

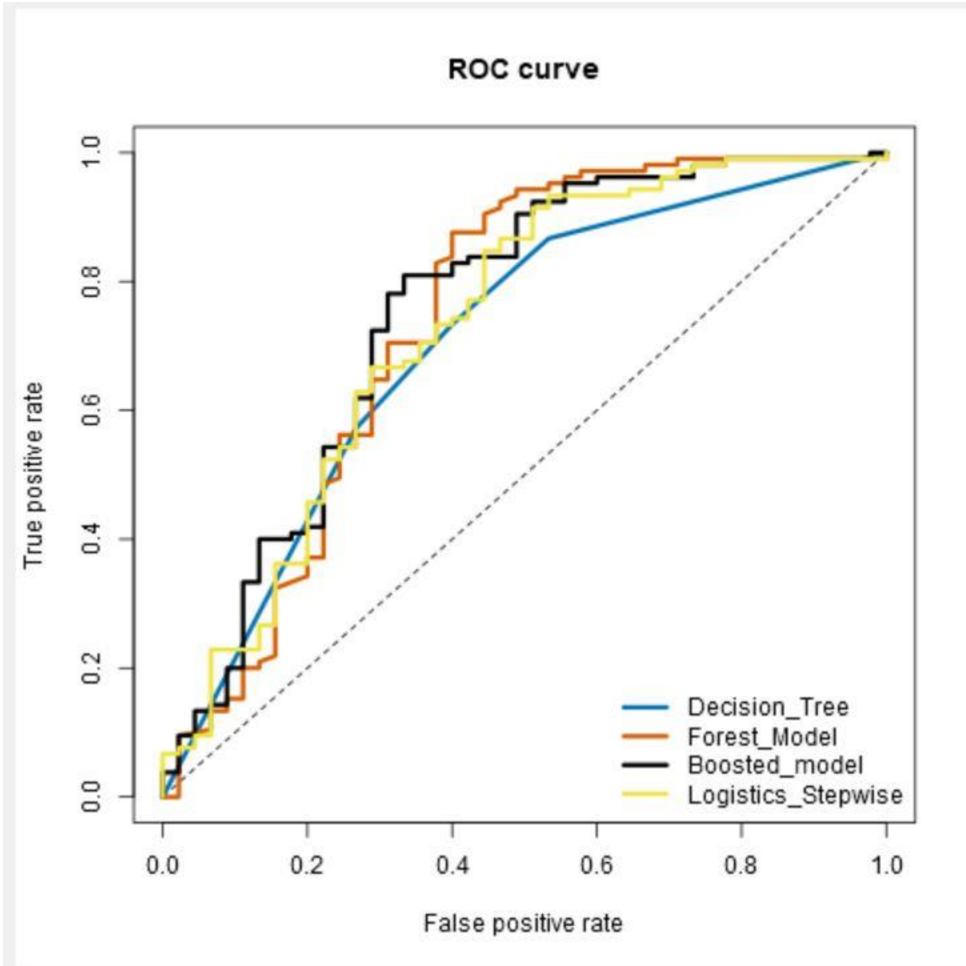
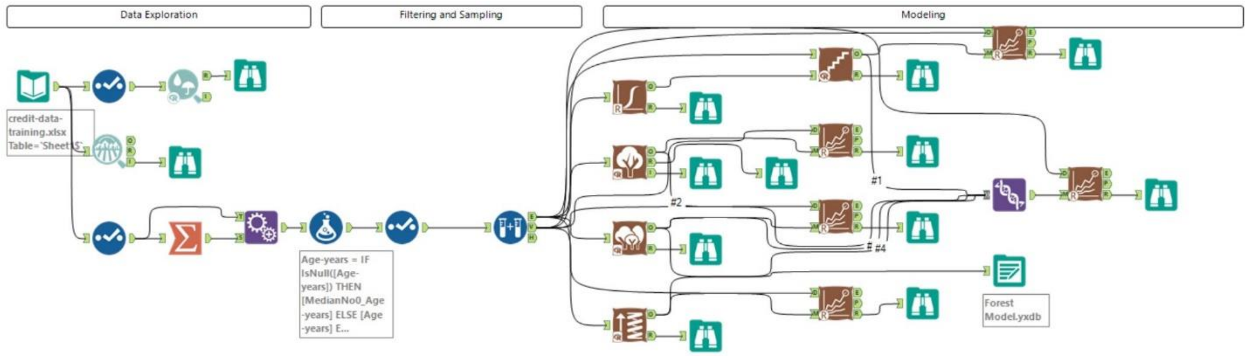


Figure 12. Four Model Comparison ROC Curve

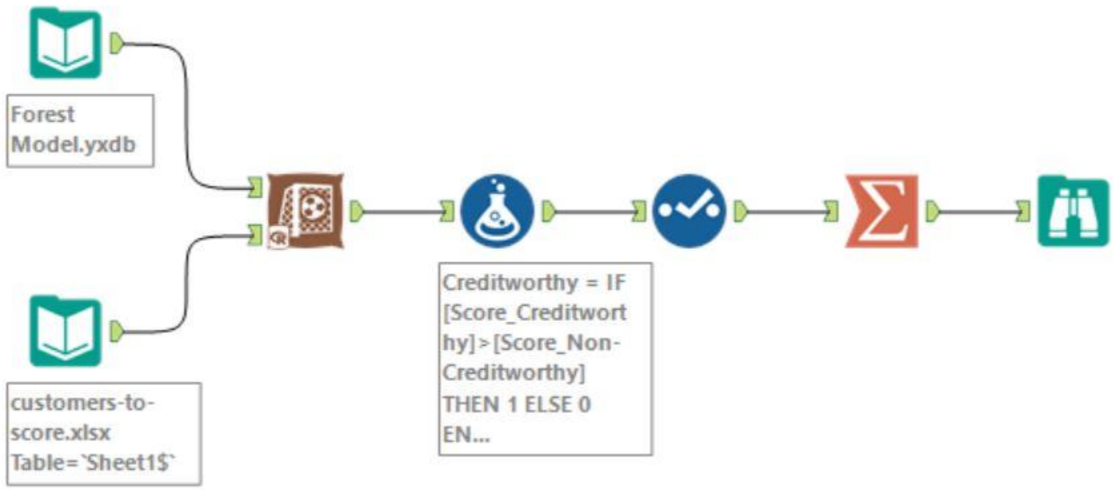
Figure 11 shows the best model is Forest model, because it has the highest overall accuracy which is 80.67%, and not biased towards creditworthy or non-creditworthy (79.69%, 86.36%).

Figure 12 indicates that the ROC graph for the Forest model is the highest line along the graph for most of the chart, and it rises the fastest of all models, which means that we are getting a higher rate of true positive rates vs false positives. The ideal ROC curve reach the top left corner, which means a high true positive rate and a low false positive rate.

After score the new customers with the forest model, there are 408 customers are score_creditworthy for a loan approval.



Data Exploration
Filtering and Sampling
Modeling
credit-data-training.xlsx Table=`Sheet1$`
Age-years = IF IsNull([Age-years]) THEN [MedianNo0_Age-years] ELSE [Age-years] E...
#1
#2
#4
Forest Model.yxdb



Forest Model.yxdb
customers-to-score.xlsx Table=`Sheet1$`
Creditworthy = IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN 1 ELSE 0 EN...