Project: Creditworthiness

Step 1: Business and Data Understanding

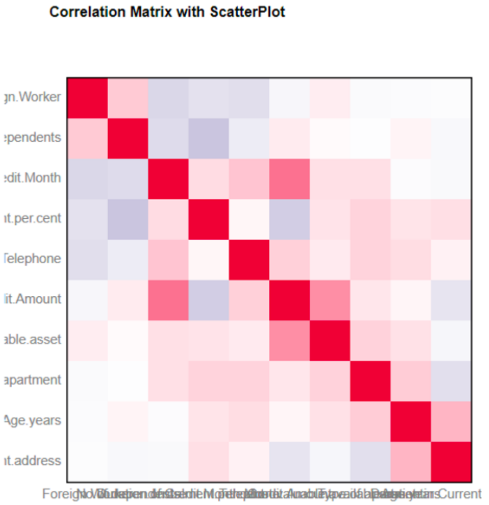
1. What decisions needs to be made?
    a. The objective is to determine whether new customers are creditworthy to the loan they applied.

2. What data is needed to inform those decisions?
    a. Data on past applications such as “Account-Balance” and “Credit-Amount” are needed to inform the decisions.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
    a. Binary classification models like “Logistic Regression”, “Decision Tree”, “Forest Model”, and “Boosted Model” are required to make these decisions.

Step 2: Building the Training Set

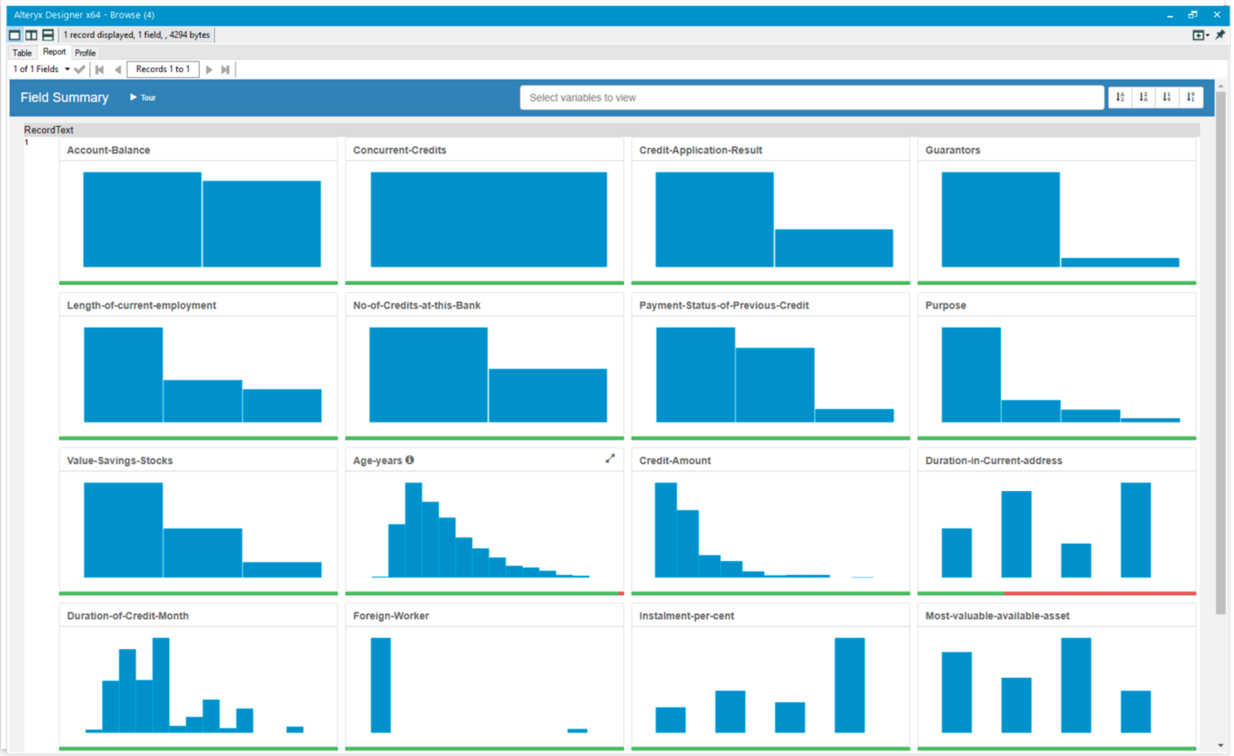
1. For the numerical data fields, an association analysis is performed to see the correlation. There are no numerical variables that are highly correlated with each other (correlation greater than 0.7).

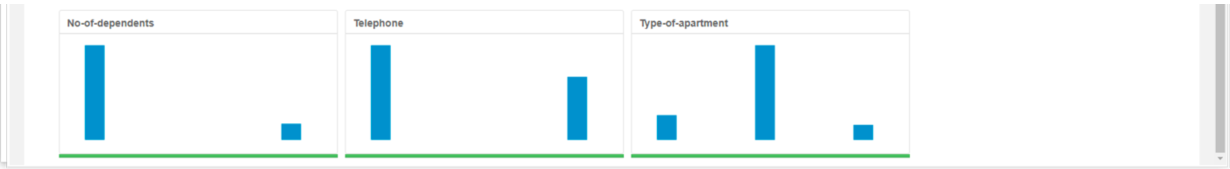


2. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
    a. A field summary tool was performed on all the variables to identify any missing values. Out of all the variables, “Duration-in-Current-address” had 69% missing data and “Age-Years” had 2% missing data.
    b. “Duration-in-Current-address” should be removed since it has the most missing data.

c. Since “Age-Years” has few missing data, it should be imputed with median age.

d. In terms of “Low Variability”, “Occupation” only had one value, “Concurrent-Credits”, “Guarantors”, “Foreign-Worker”, “No-of-dependents”, and “Telephone” has most of their data on one side indicating low variability. So, they should be removed.

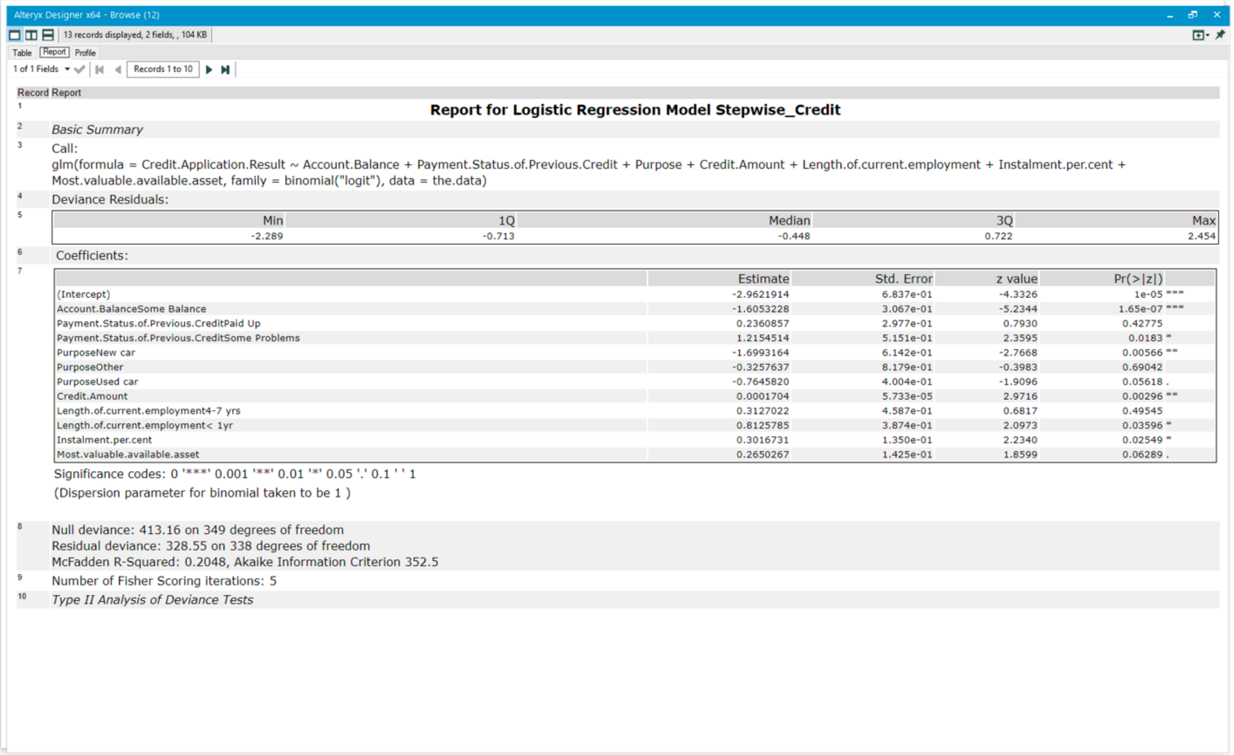




Step 3: Train your Classification Models

1. *Logistic Regression:*

a. **Stepwise:** The target variable used is “Credit-Application-Result” and selecting all the predictor variables expect the Credit-Application-Result. This resulted in the most significant variables with p-value less than 0.5 are: Account-Balance, Purpose, and Credit-Amount.


Alteryx Designer x64 - Browse (12)
13 records displayed, 2 fields, , 104 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 10
Record
Report

1 **Report for Logistic Regression Model Stepwise_Credit**

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

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

6 Coefficients:

7

	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 .

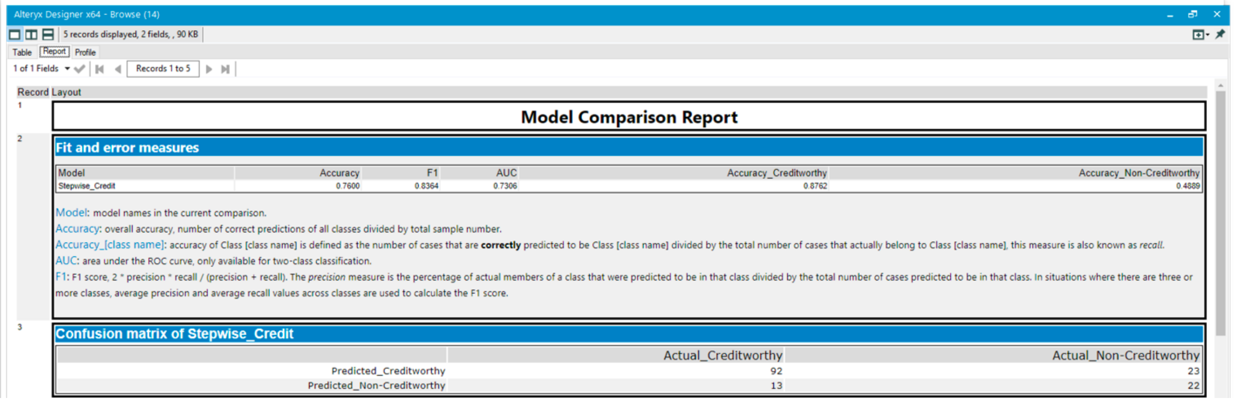
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

8 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

9 Number of Fisher Scoring iterations: 5

10 *Type II Analysis of Deviance Tests*

b. **Model Comparison:** The accuracy of the Stepwise model is 76%. Creditworthy is at 87%, higher than Non-Creditworthy which is at 48%.


Alteryx Designer x64 - Browse (14)
5 records displayed, 2 fields, , 90 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 5
Record
Layout

1 **Model Comparison Report**

2 **Fit and error measures**

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Credit	0.7600	0.8364	0.7306	0.8762	0.4889

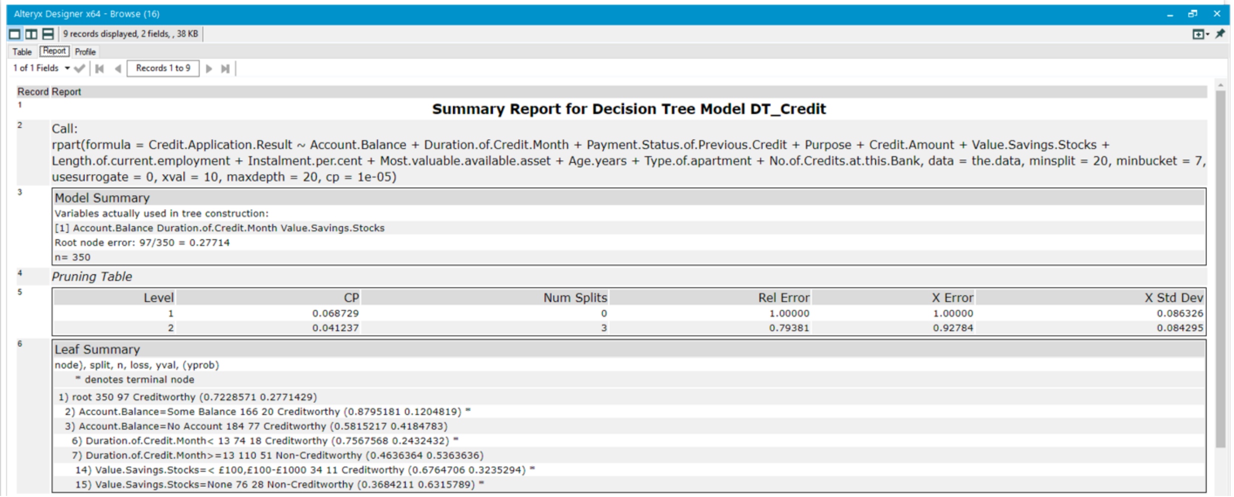
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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.

3 **Confusion matrix of Stepwise_Credit**

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

2. Decision Tree:

a. The “Root Node Error” is at 2.7%, which indicates that 2.7% were predicted incorrectly.


Alteryx Designer x64 - Browse (16)
9 records displayed, 2 fields, , 38 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 9
Record Report

Summary Report for Decision Tree Model DT_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 = 0, xval = 10, maxdepth = 20, cp = 1e-05)

Model Summary

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

Pruning Table

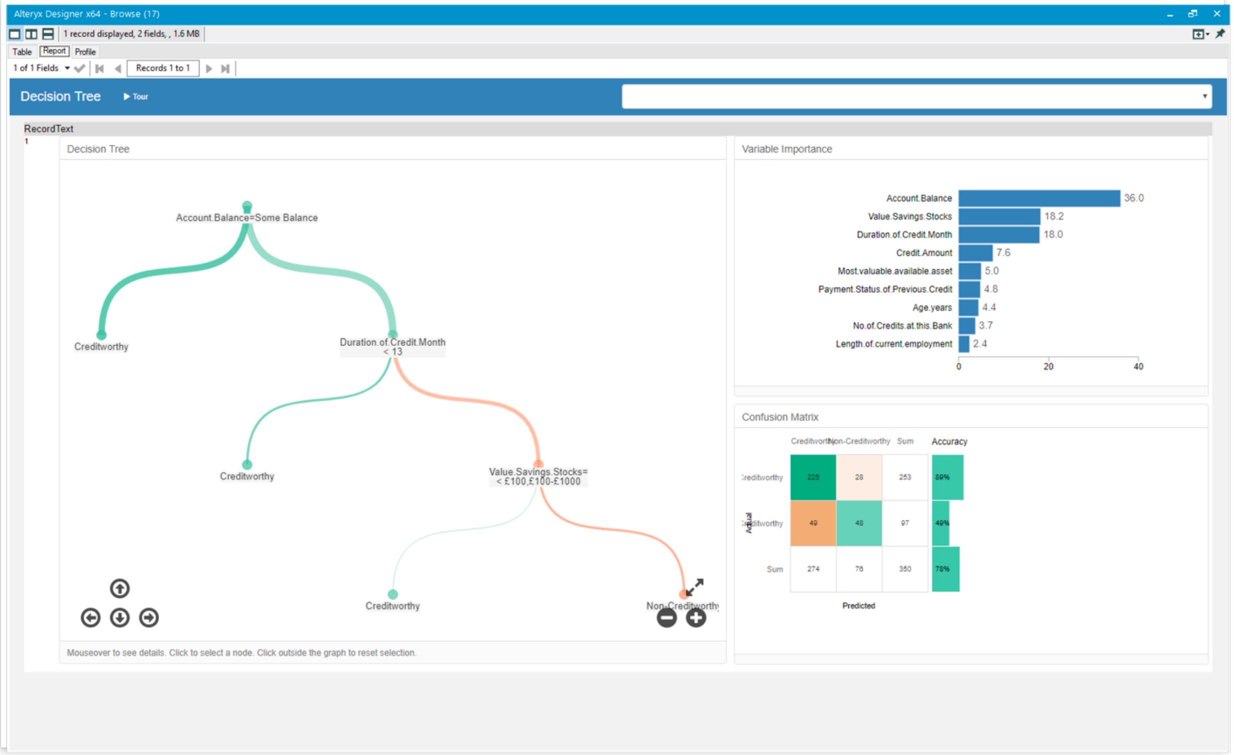
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.92784	0.084295

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) *
```

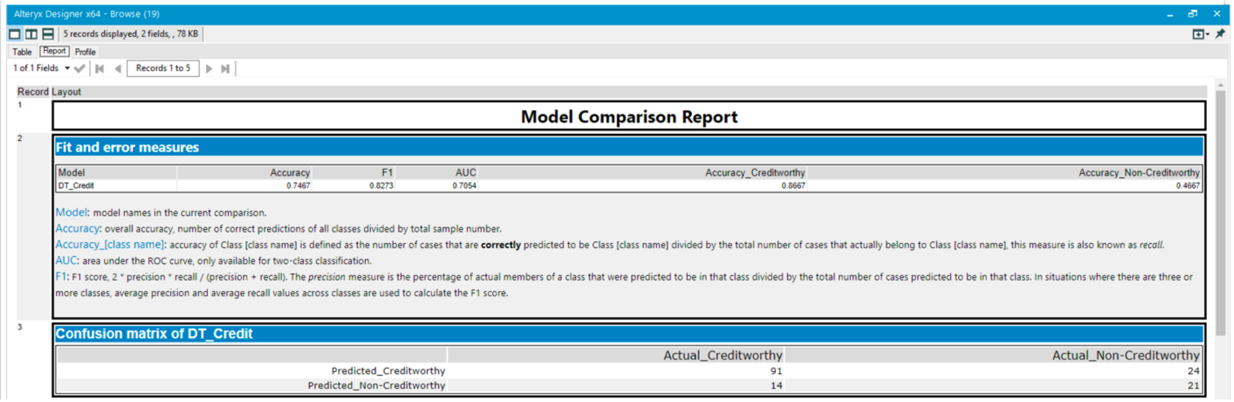
b. There are three variables that have significant importance: Account-Balance, Value-Saving-Stocks, and Duration-of-Credit-Month.

c. In the Confusion Matrix, the Creditworthy accuracy is at 89%, while the Non-Creditworthy accuracy is at 49%. However, the overall accuracy is at 78%.


Alteryx Designer x64 - Browse (17)
1 record displayed, 2 fields, , 1.6 MB
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Decision Tree
Tour
RecordText
Decision Tree
Account.Balance=Some Balance
Creditworthy
Duration.of.Credit.Month < 13
Creditworthy
Value.Savings.Stocks= < £100,£100-£1000
Creditworthy
Non-Creditworthy
Mouseover to see details. Click to select a node. Click outside the graph to reset selection.
Variable Importance
Account.Balance 36.0
Value.Savings.Stocks 18.2
Duration.of.Credit.Month 18.0
Credit.Amount 7.6
Most.valuable.available.asset 5.0
Payment.Status.of.Previous.Credit 4.8
Age.years 4.4
No.of.Credits.at.this.Bank 3.7
Length.of.current.employment 2.4
0
20
40
Confusion Matrix
Actual
Predicted

	Creditworthy	Non-Creditworthy	Sum	Accuracy
Creditworthy	225	28	253	89%
Creditworthy	49	48	97	49%
Sum	274	76	350	78%

d. **Model Comparison:** The accuracy of the Decision Tree in the Model Comparison tool is 76%. Creditworthy accuracy is at 86% and the Non-Creditworthy accuracy is at 46%.


Alteryx Designer x64 - Browse (19)
5 records displayed, 2 fields, , 78 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 5
Record Layout
1
2
3

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667

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 DT_Credit

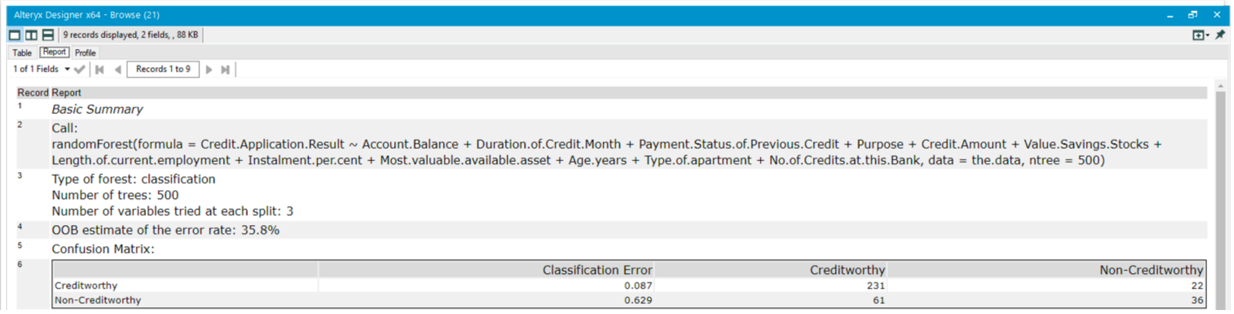
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

3. *Forest Model:*

a. The type of forest used is classification. The number of trees used, 500 (default). Number of splits, 3.

b. The Out of Body (OOB) estimation error rate is at 35.8%, which is quite high.

c. Compared to the OOB, the confusion matrix shows a better estimation error rate. Creditworthy at 8.7% and Non-Creditworthy at 62.9%



Alteryx Designer x64 - Browse (21)
9 records displayed, 2 fields, , 88 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 9

Record Report

1 *Basic Summary*

2 Call:

```
randomForest(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, ntree = 500)
```

3 Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

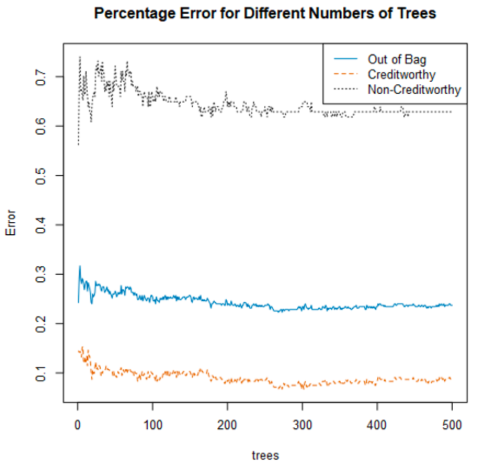
4 OOB estimate of the error rate: 35.8%

5 Confusion Matrix:

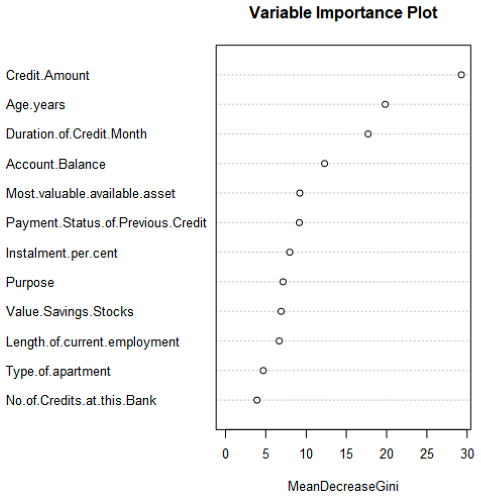
6

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.629	61	36

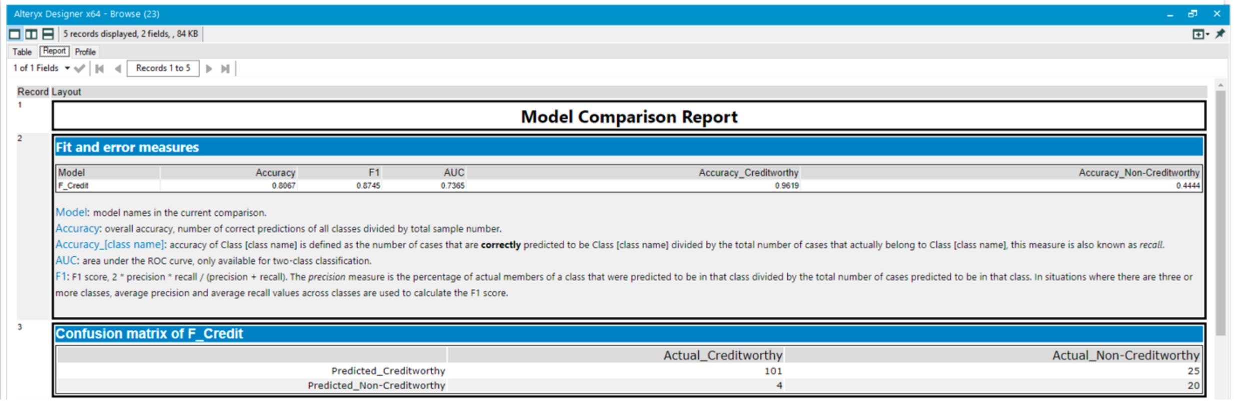
d. **Percentage error for different number of tress:**



e. **Variable Importance:** The most important variables are: Credit-Amount, Age-Years, and Duration-of-Credit-Month which have plots on the top right side of the graph.



f. **Model Comparison:** The accuracy for the Forest model is at 80%. The Creditworthy accuracy at 96% and Non-Creditworthy at 44%.


Alteryx Designer x64 - Browse (23)
5 records displayed, 2 fields, , 84 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 5
Record Layout
1
2
3

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
F_Credit	0.8067	0.8745	0.7365	0.9619	0.4444

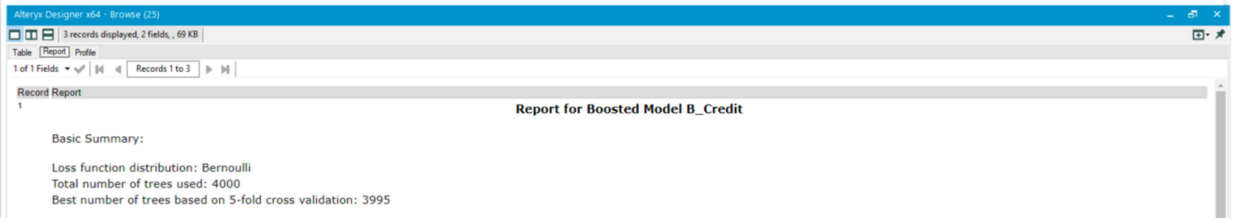
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided 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 F_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

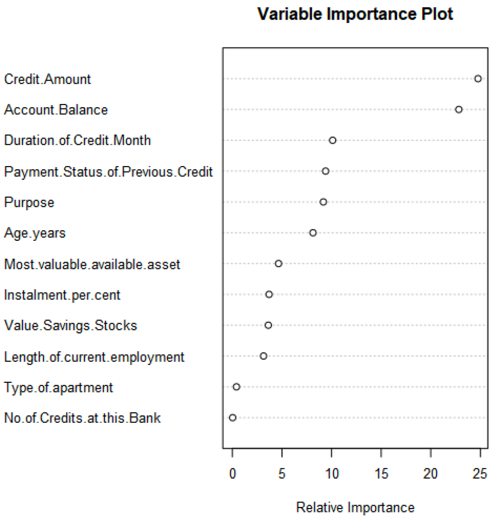
4. *Boosted Model:*

a. Since we are trying to predict whether to give the customers loan or not, there are only two outcomes, hence, Bernoulli.

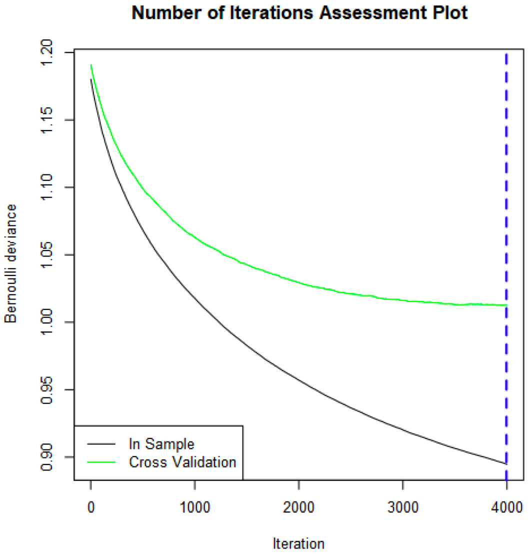

Alteryx Designer x64 - Browse (25)
3 records displayed, 2 fields, , 69 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 3

```
Record Report
1                         Report for Boosted Model B_Credit

     Basic Summary:

     Loss function distribution: Bernoulli
     Total number of trees used: 4000
     Best number of trees based on 5-fold cross validation: 3995
```

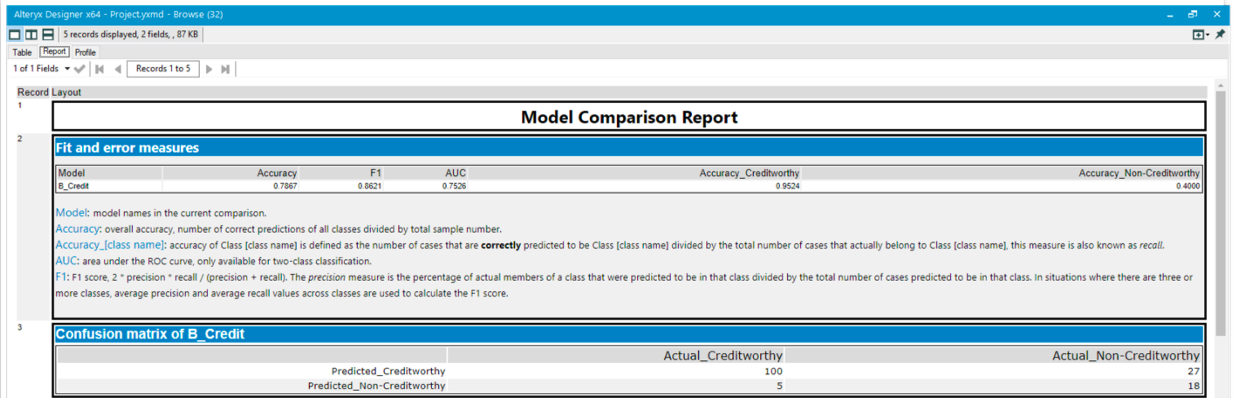
b. **Variable Importance:** The two most variables that are significant are: Credit-Amount and Account-Balance.



c. **Number of Iteration Assessment:**



d. **Model Comparison:** The accuracy for the Boosted model is at 78%. Creditworthy accuracy is at 95% and Non-Creditworthy accuracy at 40%. The accuracies between Creditworthy and Non-Creditworthy is most biased.



Alteryx Designer x64 - Project.yxmd - Browse (32)
5 records displayed, 2 fields, , 87 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 5
Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
B_Credit	0.7867	0.8621	0.7526	0.9524	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.

3

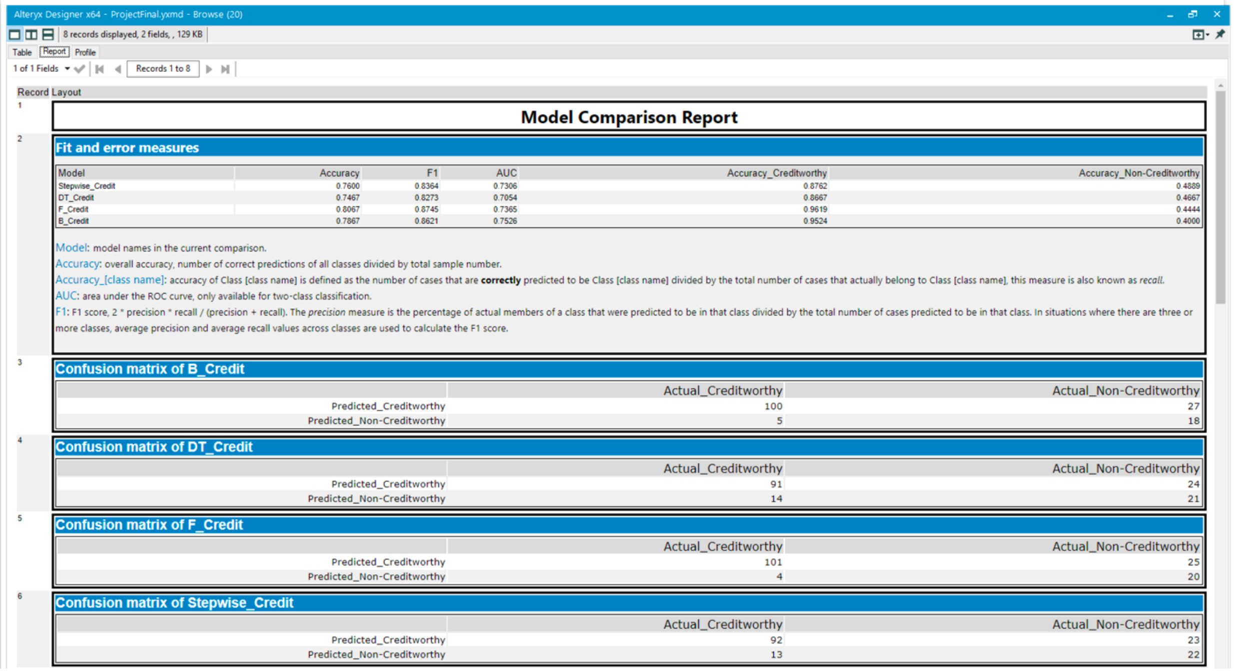
Confusion matrix of B_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Step 4: Writeup

1. Model Comparison:

a. Out of all the four models, Forest Model has the highest accuracy at 80%. It also has highest Creditworthy accuracy at 96% and the least Non-Creditworthy accuracy at 44%.



Alteryx Designer x64 - ProjectFinal.yxmd - Browse (20)
8 records displayed, 2 fields, , 129 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 8
Record Layout

1

Model Comparison Report

2

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Credit	0.7600	0.8364	0.7306	0.8762	0.4889
DT_Credit	0.7467	0.8273	0.7054	0.8667	0.4667
F_Credit	0.8067	0.8745	0.7365	0.9619	0.4444
B_Credit	0.7867	0.8621	0.7526	0.9524	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.

3

Confusion matrix of B_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

4

Confusion matrix of DT_Credit

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

5

Confusion matrix of F_Credit

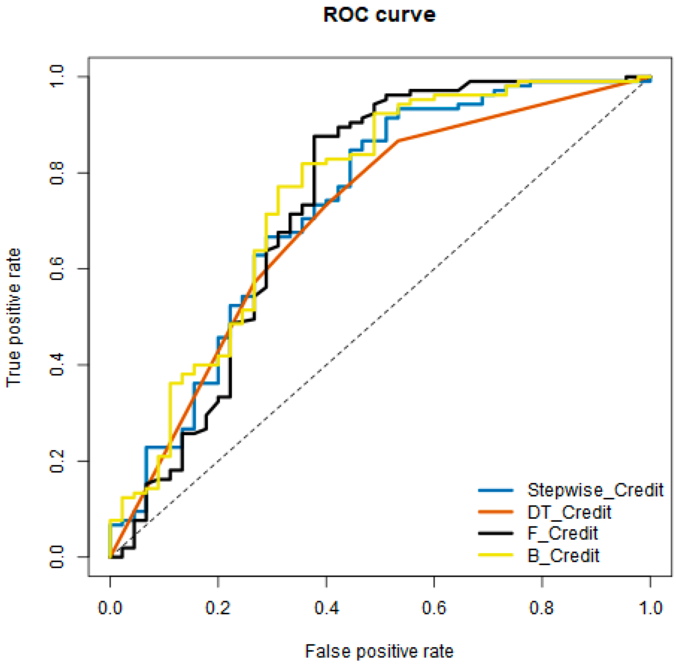
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	25
Predicted_Non-Creditworthy	4	20

6

Confusion matrix of Stepwise_Credit

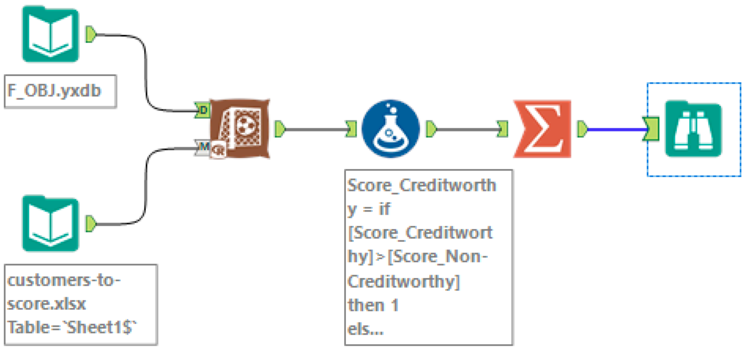
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

b. **ROC Curve:** The Forest Model was the first to reach the top in the ROC curve.

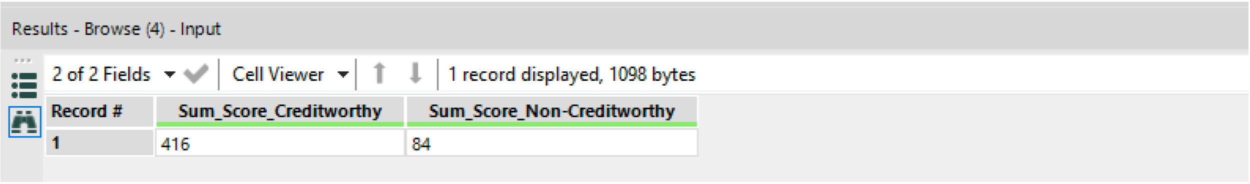


c. The accuracies between Creditworthy and Non-Creditworthy lead to the least bias in the Confusion Matrix.

d. There is a total of 416 creditworthy customers.



Score_Creditworthy = if [Score_Creditworthy]>[Score_Non-Creditworthy] then 1 els...


Results - Browse (4) - Input

1 record displayed, 1098 bytes

Record #	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	416	84