Project: Creditworthiness

Business and Data Understanding

Key Decisions:

- **What decisions needs to be made?**

Due to a financial scandal that hit a competitive bank last week, there has been a sudden influx of new people applying for loans at the bank. All of a sudden there are nearly 500 loan applications to process this week.

The bank sees this new influx as a great opportunity and wants to figure out how to process all of these loan applications to determine if customers are creditworthy to give a loan to.

- **What data is needed to inform those decisions?**

1. Data on all past applications: credit-data-training.xlsx
   - o Data has already been cleaned but will still require check on missing data
2. The list of customers that need to be processed in the next few days: customers-to-score.xlsx

The columns used are:

Credit-Application-Result	Length-of-current-employment	Type-of-apartment
Account-Balance	Instalment-per-cent	No-of-Credits-at-this-Bank
Duration-of-Credit-Month	Guarantors	Occupation
Payment-Status-of-Previous-Credit	Duration-in-Current-address	No-of-dependents
Purpose	Most-valuable-available-asset	Telephone
Credit-Amount	Age-years	Foreign-Worker
Value-Savings-Stocks	Concurrent-Credits	

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

Based on the Predictive Methodology Map to determine the appropriate analytical technique:
Step 1: The business problem is to predict outcome
Step 2: It is data rich since there are past data
Step 3: It is classification
Step 4: To get a binary outcome - to loan or not to loan

Hence it will be binary classification model.

Building the Training Set

The data provided has been cleaned up and there are a total of 20 fields for this data set. Shown below are the interactive output and the report generated from the *Field Summary* tool.

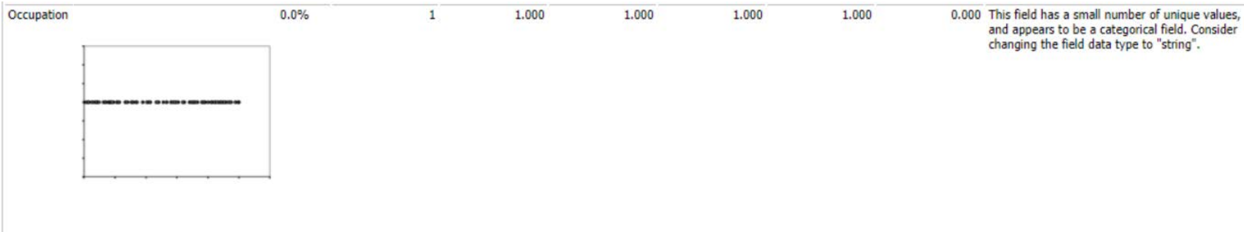

Figure 1 Field Summary (Report)

Figure 2 Field Summary (Interactive Output)

The following are the actions taken based on the health of the fields:

1. Fields with missing data
   a. **Duration-in-Current-Address** field has 69% missing data so it will be removed.
   b. **Age-years** has only 2.4% missing data hence will impute the missing data with the median age.

2. Fields that have low variability (only one value for the entire field)
   a. **Concurrent-Credits** and **Occupation** fields have low variability as it has just one value so it will be removed.
   b. **Guarantors** field have 457 instances of 'None' and just 43 instances of 'Yes' hence this field is heavily skewed to one type of data. This is considered as low variability so it will be removed. This is the same for the **Foreign-Worker** and **No.-of-dependents** fields and they will be removed as well.

3. **Telephone** field will be removed as there is no logical reason for including the variable.

Next, the *Imputation* tool is used to replace null fields with median value for the **Age-years** field as taking the median would be a better measure of central tendency in this situation. Follow by the *Summarize* tool to get the average of Age Years which is 36 (rounded up).

Record #	Avg_Age-years
1	35.574

Figure 3 Average of Age Years

The clean data set now have 13 columns.

Record #	Name
1	Credit-Application-Result
2	Account-Balance
3	Duration-of-Credit-Month
4	Payment-Status-of-Previous-Credit
5	Purpose
6	Credit-Amount
7	Value-Savings-Stocks
8	Length-of-current-employment
9	Instalment-per-cent
10	Most-valuable-available-asset
11	Type-of-apartment
12	No-of-Credits-at-this-Bank
13	Age-years

Figure 4 Clean data set number of columns

A correlation matrix was set up with the *Association Analysis* tool using **Credit-Application-Result** as the Target field and **Creditworthy** as the target level of interest. Looking through the correlation matrix and full correlation matrix shown below using 0.7 as the benchmark for high correlation, there seems to be nothing of high correlation.

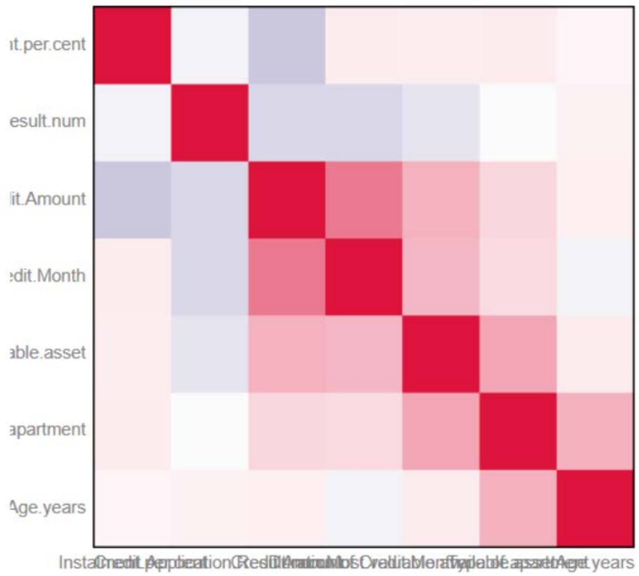

Figure 5 Correlation Matrix

Full Correlation Matrix

	Credit.Application.Result.num	Duration.of.Credit.Month	Credit.Amount	Instalment.per.cent	Most.valuable.available.asset	Age.years	Type.of.apartment
Credit.Application.Result.num	1.000000	-0.204317	-0.200990	-0.065345	-0.137917	0.056737	-0.021860
Duration.of.Credit.Month	-0.204317	1.000000	0.570441	0.079515	0.304734	-0.066319	0.153141
Credit.Amount	-0.200990	0.570441	1.000000	-0.285631	0.327762	0.068643	0.168683
Instalment.per.cent	-0.065345	0.079515	-0.285631	1.000000	0.078110	0.040540	0.082936
Most.valuable.available.asset	-0.137917	0.304734	0.327762	0.078110	1.000000	0.085437	0.379650
Age.years	0.056737	-0.066319	0.068643	0.040540	0.085437	1.000000	0.333075
Type.of.apartment	-0.021860	0.153141	0.168683	0.082936	0.379650	0.333075	1.000000

Figure 6 Full Correlation Matrix

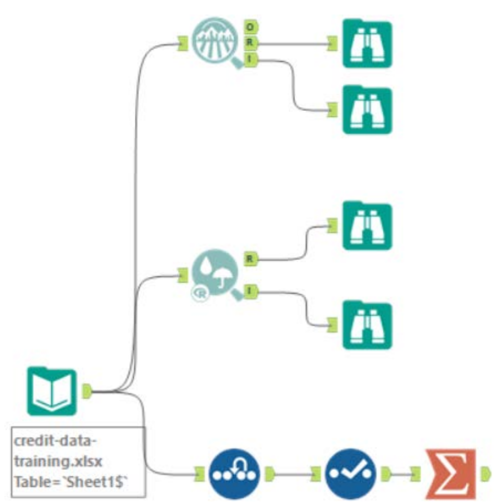

Figure 7 Alteryx flow (Prepare Data)

Train the Classification Models

Steps to create Estimation and Validation samples:

a. Use an *Input* tool to bring in the 'credit-data-training.xlsx' file.
b. Use a *Select* tool to set the 13 fields.
c. Use the *Create Samples* tool to create Estimation and Validation samples where 70% of the dataset go to Estimation and 30% of the entire dataset reserved for Validation and the random seed is set to 1.

1. Logistic Regression

Steps took to create the model – Logistic Regression:

a. Use the *Logistic Regression* tool and set the target variable as **Credit-Application-Result**.
b. Select all variables except **Credit-Application-Result** as predictor variables.
c. Add a *Stepwise* tool
d. Add a *Model Comparison* tool

Report for Logistic Regression Model stepwise_log

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

Figure 8 Logistic Regression Report

Based on the above report, the R-squared values is at 0.2048, which is quite low. Where the higher the value the better the model fits the data.

a. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

Based on the report shown above, the following are the most significant predictor variables:

Variable Name	P-Value	Significance Code
Amount.BalanceSome Balance	1.80e-06	***
PurposeNew car	0.00518	**
Credit.Amount	0.00966	**
Payment.Status.of.Previous.CreditSome Problems	0.0182	*
Length.of.current.employment< 1yr	0.04946	*
Most.valuable.available.asset	0.03645	*
Instalment.per.cent	0.02644	*

b. 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
stepwise_log	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_log

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

Figure 9 Logistic Regression Model Comparison Report

The validation data set was predicted quite well by this model with an overall accuracy of 76%. The creditworthy were predicted quite high at 88% and the non-creditworthy were tougher to predict at only 49%.

The confusion matrix shows 92 records that were predicted creditworthy that were actually creditworthy. Yet we had 13 records that were predicted non-creditworthy that were actually creditworthy.

The result shows a pretty good representation of where biases may occur. There are more non-creditworthy that were predicted creditworthy (23 records) than creditworthy that were predicted non-creditworthy (13 records).

2. Decision Tree

Steps took to create the model – Decision Tree:

a. Use the *Decision Tree* tool and set the target variable as **Credit-Application-Result**.
b. Select all variables except **Credit-Application-Result** as predictor variables.
c. Add a *Model Comparison* tool

Summary Report for Decision Tree Model credit_DT

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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age.years, data = the.data, minsplit = 20, minbucket = 7, usesurrogate = 2, xval = 10, maxdepth = 20, cp = 0)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Age.years
[3] Credit.Amount Duration.of.Credit.Month
[5] Instalment.per.cent Length.of.current.employment
[7] Most.valuable.available.asset No.of.Credits.at.this.Bank
[9] Payment.Status.of.Previous.Credit Purpose
[11] 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	0.0687285	0	1.00000	1.00000	0.086326
2	0.0412371	3	0.79381	0.92784	0.084295
3	0.0257732	4	0.75258	0.91753	0.083987
4	0.0206186	8	0.64948	0.92784	0.084295
5	0.0103093	9	0.62887	1.00000	0.086326
6	0.0017182	12	0.59794	1.06186	0.087894
7	0.0000000	18	0.58763	1.05155	0.087644

Figure 10 Decision Tree Model Report

Based on the above report, the root node error shows the percentage of how many of the data points were predicted incorrectly. The value for this model is pretty low at 28%.

a. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

The table shown below are the most significant predictor variables:

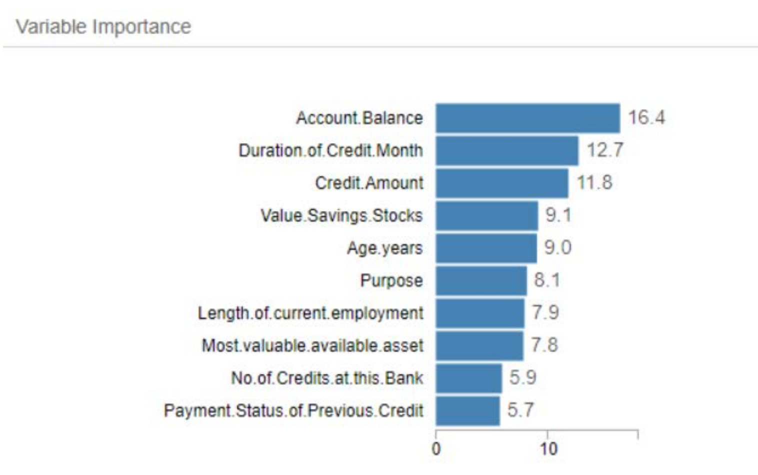

Figure 11 Decision Tree Variable Importance

Based on the confusion matrix, the overall accuracy is 84%. 91% of the creditworthy were classified correctly, while only 66% of the non-creditworthy were classified correctly. 9% of the creditworthy were actually incorrectly classified as non-creditworthy and 34% of the non-creditworthy are incorrectly classified as creditworthy.

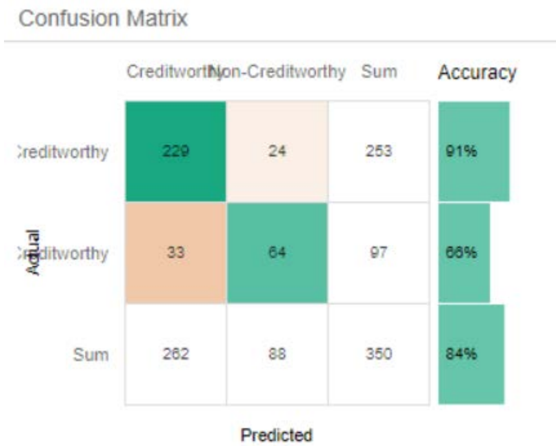

Figure 12 Decision Tree Confusion Matrix

b. 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
credit_DT	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 credit_DT

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

Figure 13 Decision Tree Model Comparison Report

The overall accuracy is 67% which is less than the logistic regression model. The creditworthy were predicted quite high at 79% and the non-creditworthy were tougher to predict at only 40%.

The confusion matrix shows 83 records that were predicted creditworthy that were actually creditworthy. Yet we had 22 records that were predicted non-creditworthy that were actually creditworthy.

The result shows a pretty good representation of where biases may occur. There are more non-creditworthy that were predicted creditworthy (27 records) than creditworthy that were predicted non-creditworthy (18 records).

3. Forest Model

Steps took to create the model – Forest Model:

a. Use the *Forest Model* tool and set the target variable as **Credit-Application-Result**.
b. Select all variables except **Credit-Application-Result** as predictor variables.
c. Add a *Model Comparison* tool

Report

Basic Summary

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 + Type.of.apartment + No.of.Credits.at.this.Bank + Age.years, data = the.data, ntree = 500, replace = TRUE)

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

OOB estimate of the error rate: 24%

Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.639	62	35

Figure 14 Forest Model Report

Based on the report shown above, the type of forest model is classification and there are 500 trees build for this model. The out of the bag estimate of the error rate is 24% which is pretty high. From the confusion matrix, we can see that creditworthy was predicted quite well at 9% but the non-creditworthy was quite bad at 64%.

a. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

The diagram shown below are the most significant predictor variables:

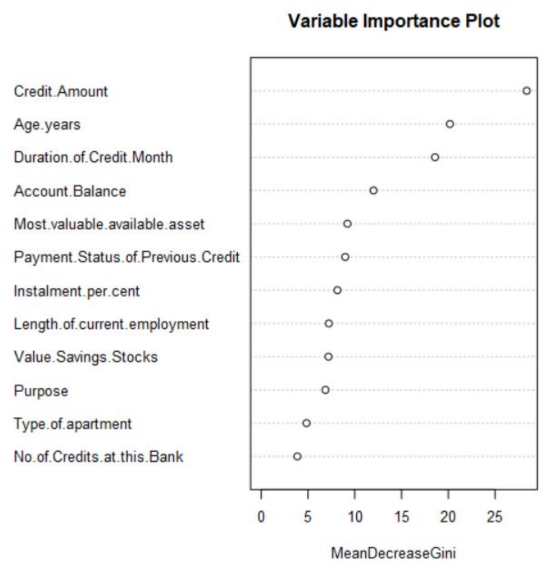

Figure 15 Forest Model Variable Importance Plot

b. 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
credit_FM	0.8000	0.8707	0.7361	0.9619	0.4222

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 credit_FM

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

Figure 16 Forest Model Comparison Report

The overall accuracy is 80% which is more than both logistic regression and decision tree model. The creditworthy were predicted quite high at 96% and the non-creditworthy were tougher to predict at only 42%.

The confusion matrix shows 101 records that were predicted creditworthy that were actually creditworthy. Yet we had 4 records that were predicted non-creditworthy that were actually creditworthy.

The result shows a pretty good representation of where biases may occur. There are more non-creditworthy that were predicted creditworthy (26 records) than creditworthy that were predicted non-creditworthy (19 records).

4. Boosted Model

Steps took to create the model – Boosted Model:

a. Use the *Boosted Model* tool and set the target variable as **Credit-Application-Result**.
b. Select all variables except **Credit-Application-Result** as predictor variables.
c. Under *model customization* options, select *Specify target type and the loss function distribution* & choose *Binary categorical*.
d. Add a *Model Comparison* tool

Report

Report for Boosted Model credit_Boosted

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 3940

Figure 17 Boosted Model Report

a. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

The diagram shown below are the most significant predictor variables:

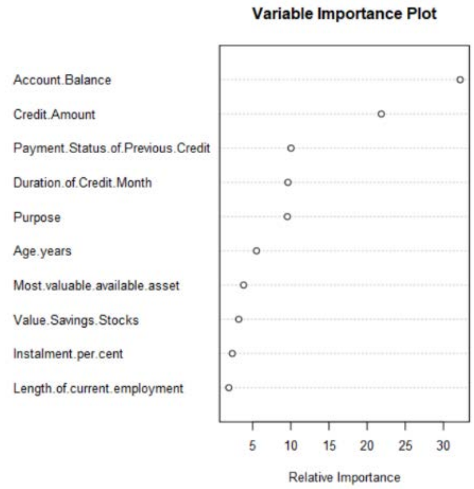

Figure 18 Boosted Model Variable Importance Plot

b. 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?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
credit_Boosted	0.7867	0.8632	0.7524	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 credit_Boosted

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

Figure 19 Boosted Model Comparison Report

The overall accuracy is 79%, the creditworthy were predicted quite high at 96% and the non-creditworthy were tougher to predict at only 38%.

The confusion matrix shows 101 records that were predicted creditworthy that were actually creditworthy. Yet we had 4 records that were predicted non-creditworthy that were actually creditworthy.

The result shows a pretty good representation of where biases may occur. There are more non-creditworthy that were predicted creditworthy (28 records) than creditworthy that were predicted non-creditworthy (17 records).

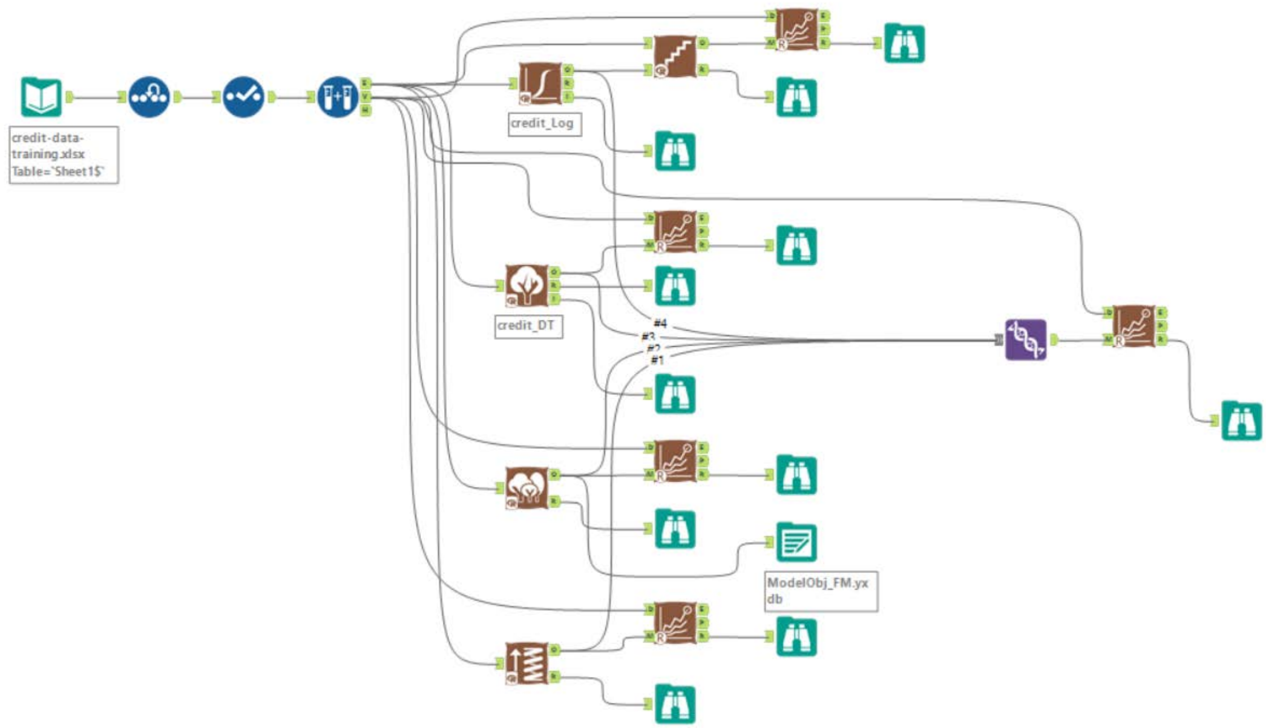

Figure 20 Alteryx Flow (Build Model)

Conclusion

Steps took to compare all 4 models:

a. Use the *Union* tool to union the 4 model objects together.
b. Add a *Model Comparison* tool.
c. Add an *output* tool to create a model object for the best model which is Forest Model.
d. Created a new canvas to score the model
e. Add in both model object output and Customer data set.
f. Bring in *Score* tool for the both input above.
g. Use the *Formula* tool to code into 1's and 0's.

Score_Creditworthy column	**Score_Non-creditworthy column**
If [Score_Creditworthy]>[Score_Non-Creditworthy] THEN 1 ELSE 0 ENDIF	If [Score_Non-Creditworthy]>[Score_Creditworthy] THEN 1 ELSE 0 ENDIF

h. Use Summarize tool to sum up the 1's for both creditworthy and non-creditworthy

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
credit_FM	0.8000	0.8707	0.7361	0.9619	0.4222
credit_DT	0.6733	0.7721	0.6296	0.7905	0.4000
credit_Log	0.7800	0.8520	0.7314	0.9048	0.4889
credit_Boosted	0.7867	0.8632	0.7524	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 credit_Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of credit_DT

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

Confusion matrix of credit_FM

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

Confusion matrix of credit_Log

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

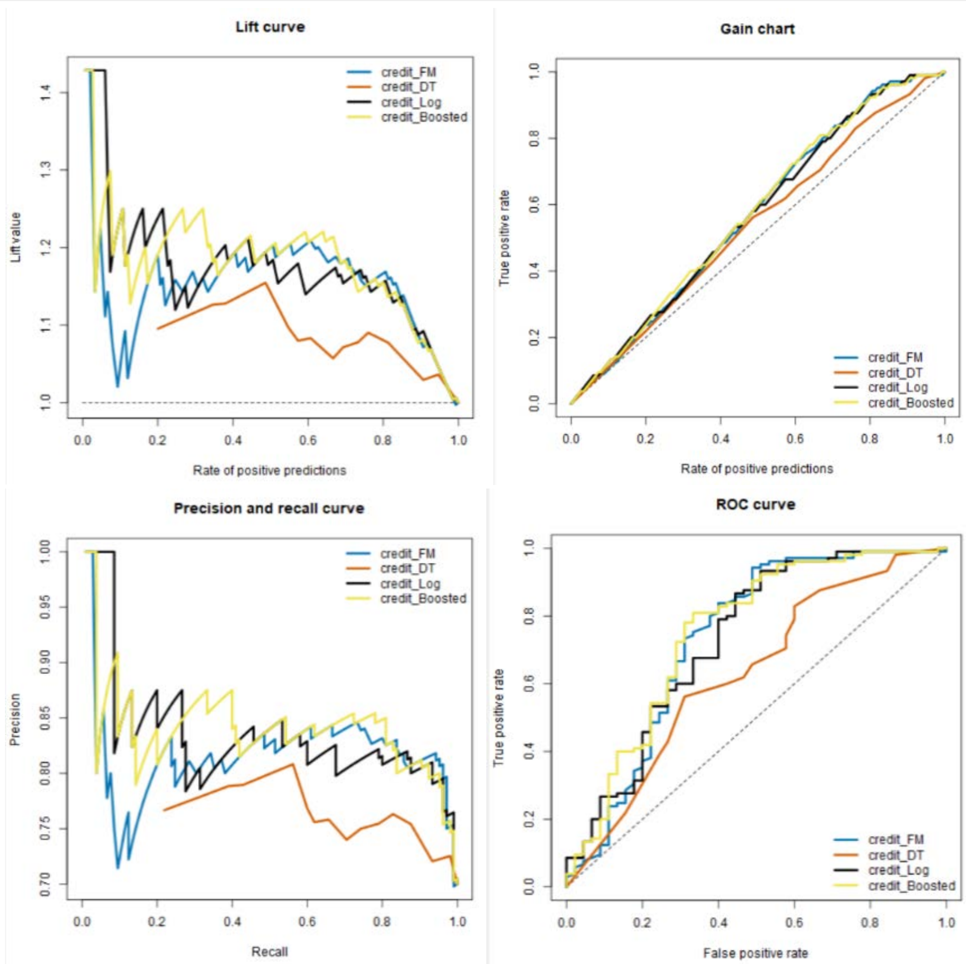

Figure 21 Model Comparison Report

The final model used for prediction will be the Forest model based on the model comparison report shown below

The Forest Model (80%) performed best, followed up by a Boosted model (79%) and coming in very close 3rd was the Logistic Regression model (78%). Under the confusion matrix, Forest model had 101 records that were predicted creditworthy that were actually creditworthy which is the highest compared to the rest of the models. It rank 2nd with 19 records that were predicted non-creditworthy that were actually creditworthy.

The ROC curve shows that the Forest model has the best overall true positive rates. It is one of the model that has the highest curve among all four models.

Based on the score on the new customers, there are 406 individuals that are creditworthy.

Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
406	94

Figure 22 Score Result

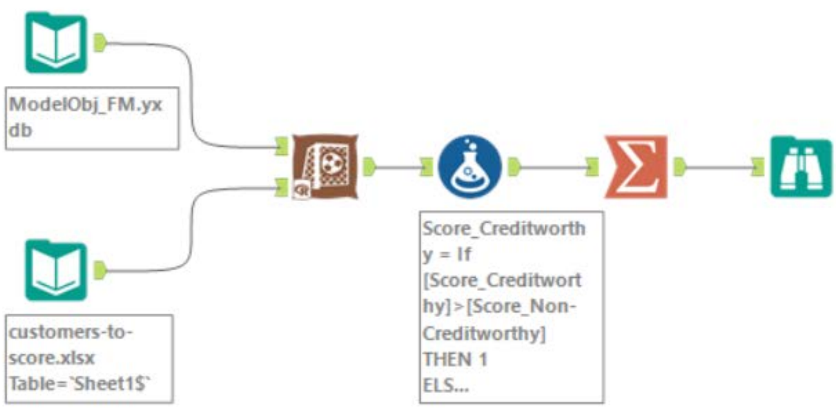

Figure 23 Alteryx workflow (Score Model)