

Predictive Analytics for Business Nanodegree

Project: Creditworthiness

Latifa M.Alyaeesh

Misk & Udacity

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.

Fortunately for you, you just completed a course in classification modeling and know how to systematically evaluate the creditworthiness of these new loan applicants.

For this project, you will analyze the business problem using the Problem Solving Framework and provide a list of creditworthy customers to your manager in the next two days.

You have the following information to work with:

1. Data on all past applications
2. The list of customers that need to be processed in the next few days

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Key Decisions:

Answer these questions

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

We need to made decision for determining if customer are creditworthy to give a loan.

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

We need aggregation 500-loan application in order to provide a list of creditworthy customers to manager in the next two days.

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

Binary classification model such as logistic regression, forest model, decision tree and boosted model will used to help make decision to analyzed and determine creditworthy customers.

Step 2: Building the Training Set

Answer this question:

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



Figure 1: Field Summery

The following are the actions taken based on the health of the field summary:

1- Field with missing data

- **Duration in Current Address:** The field has 69% missing data so it will be removed.
- **Age years:** The field has 2% missing data therefore will impute the missing data with the median age.

Hint: Median age used instead of mean as the data skewed to the left as shown above.

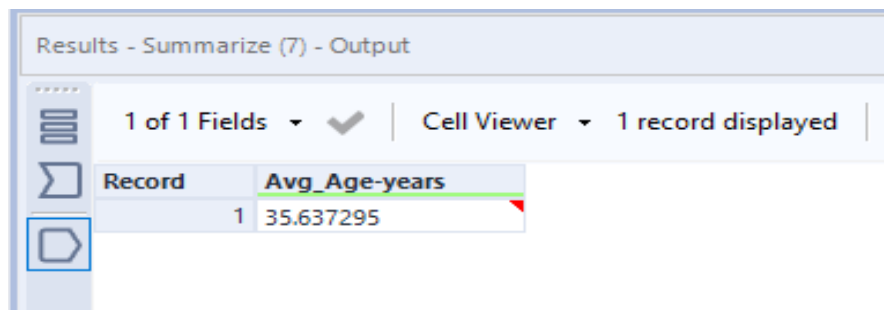
2- Field that have low variability

- **Concurrent Credits & Occupation:** the fields have low variability as it has just one value so it will be removed.
- **Guarantors:** the field has 457 instances of none and just 43 instances of yes therefore this field heavily skewed to one type of data. This is considered as low variability so it will be removed as well.

3- Field with no logical reason

- **Telephone:** The field must be removed because of its irrelevancy to the customer creditworthiness.

After that, the summarize tool to get the average of age years which is (35.637295), will be rounded up (36).



The screenshot shows a software interface titled "Results - Summarize (7) - Output". It features a sidebar with icons for list, folder, and document. The main area displays "1 of 1 Fields" with a checkmark and "Cell Viewer" showing "1 record displayed". Below this is a table with two columns: "Record" and "Avg_Age-years". The first row shows "1" under "Record" and "35.637295" under "Avg_Age-years".

Record	Avg_Age-years
1	35.637295

Figure 2: Average age years

An association analysis performed on the numerical variables and there are no that are highly correlated with each other. Moreover, for looking through the correlation matrix and full correlation matrix shown below using 0.7 as the benchmark for high correlation, it is means nothing of high correlation.

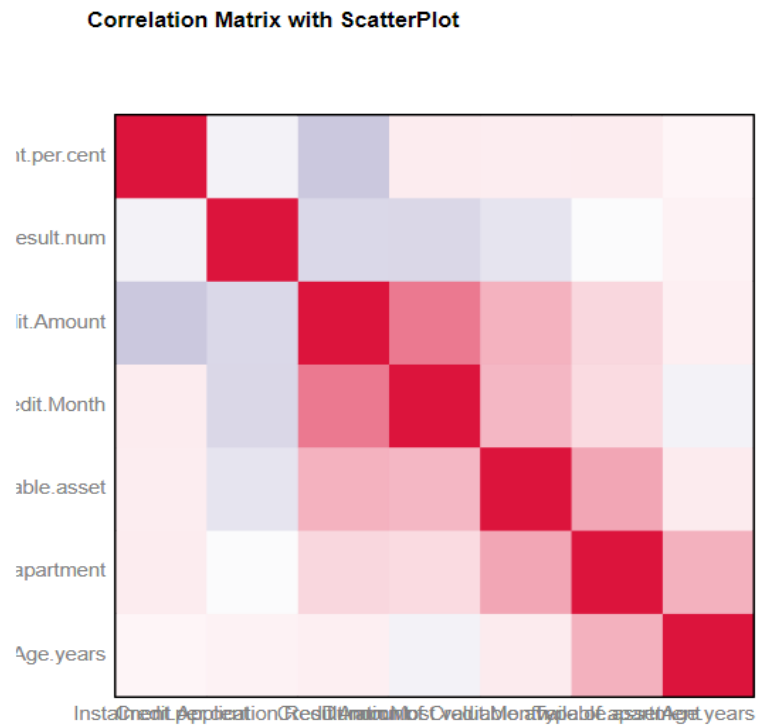


Figure 3: Correlation Matrix

Full Correlation Matrix

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

Figure 4: Full Correlation Matrix

- Alteryx Workflow:

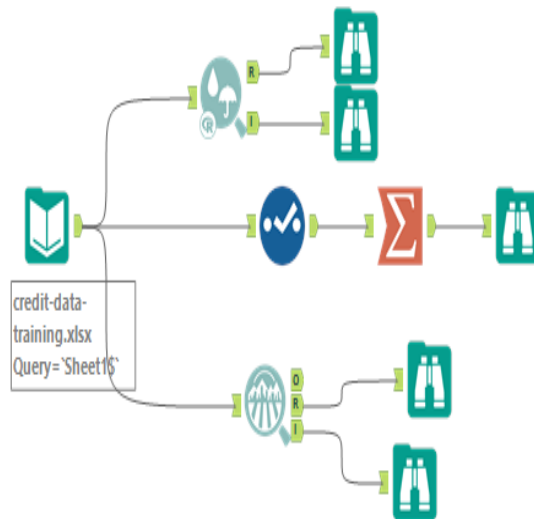


Figure 5: Prepare Data Workflow

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

1- Logistic Regression

Report for Logistic Regression Model Logistic_Regression

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

Figure 6: Logistic regression report

Based on the logistic regression report the McFadden R-Squared: 0.2048, which is quite low, where the higher the value the better the model fits the data.

Q1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

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 ,	

Figure 7: Coefficients table

Q2: 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
Logistic_Regression	0.7600	0.8364	0.7306	0.8762	0.4889
<p>Model: model names in the current comparison.</p> <p>Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.</p> <p>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 <i>recall</i>.</p> <p>AUC: area under the ROC curve, only available for two-class classification.</p> <p>F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The <i>precision</i> 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.</p>					
Confusion matrix of Logistic_Regression					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	92		23		
Predicted_Non-Creditworthy	13		22		

Figure 8: Logistic Regression Model Comparison Report

Accuracy of the Logistic Regression Model is 0.7600. The confusion matrix shows above 92 records that predicted creditworthy that is Actual_ creditworthy. While we had 13 records that predicted non-creditworthy which is Actual_creditworthy. The result shows above a good representation of where biases may occur. There are 22 records for creditworthy that were predicted non-creditworthy. Other side, there 23 records for non-creditworthy that were predicted creditworthy.

2- Decision Tree

Q1: 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 the most the significant predictor variables:

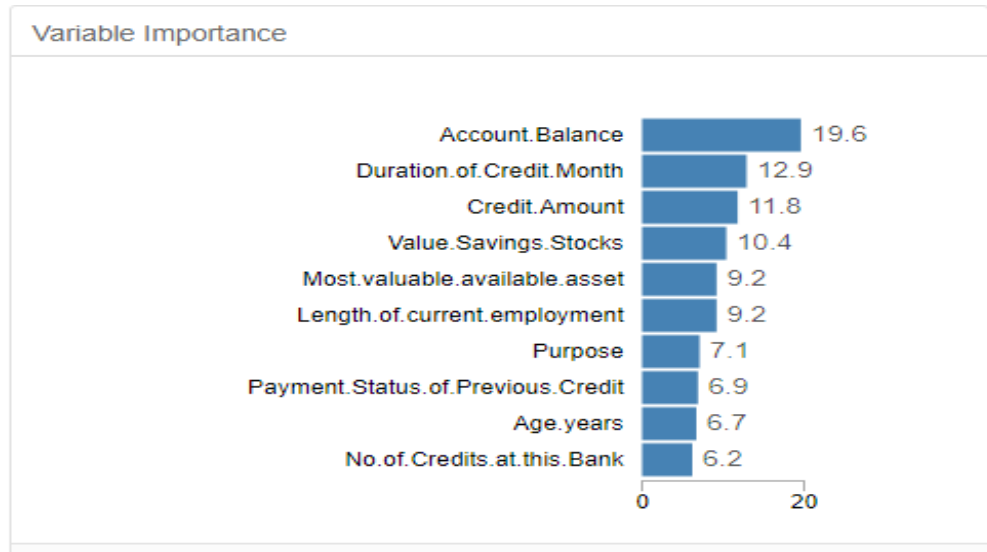


Figure 9: Decision Tree Variable importance

Q2: 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
Decision_Tree	0.6867	0.7892	0.6806	0.8381	0.3333
<p>Model: model names in the current comparison.</p> <p>Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.</p> <p>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 <i>recall</i>.</p> <p>AUC: area under the ROC curve, only available for two-class classification.</p> <p>F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The <i>precision</i> 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.</p>					
Confusion matrix of Decision_Tree					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	88		30		
Predicted_Non-Creditworthy	17		15		

Figure 10: Decision Tree Model Comparison Report

The overall accuracy is 0.6867, which is less than logistic regression model. The creditworthy predicted quite high at 0.8381 and the non-creditworthy were tougher to predict at 0.3333. The confusion matrix shows 88 records that predicted creditworthy that were Actual_Creditworthy. in other side, we had 17 records that predicted non-creditworthy that were Actual_Creditworthy.

The result shows above a good representation of where biases may occur. There are 30 records for non-creditworthy that predicted creditworthy and 15 records for creditworthy predicted non-creditworthy.

3- Forest Model

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 + Age.years + Type.of.apartment, 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	0.095	229	24
Non-Creditworthy	0.619	60	37

Figure 11: Forest Model

Based on the Forest Model report:

- Type of forest: **Classification**
- Number of trees: **500**
- Number of variables tried at each split: **3**
- The out of the bag estimate of the error rate: **24%** “ pretty high “

Q1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

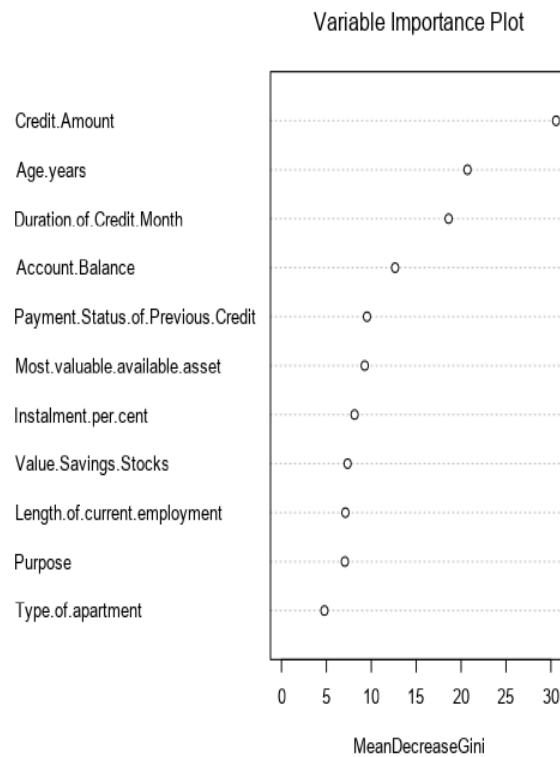


Figure 12: Forest Model Variable Importance Plot

Q2: 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
Forest_Model	0.8000	0.8707	0.7342	0.9619	0.4222
<p>Model: model names in the current comparison.</p> <p>Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.</p> <p>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 <i>recall</i>.</p> <p>AUC: area under the ROC curve, only available for two-class classification.</p> <p>F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The <i>precision</i> 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.</p>					
Confusion matrix of Forest_Model					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	101		26		
Predicted_Non-Creditworthy	4		19		

Figure 13: Forest Model Comparison Report

Based on the forest model comparison report the accuracy is 0.80. in other side, the creditworthy were predicted quite high 96% and the non-creditworthy predicted 42%. Based on the confusion matrix shows 101 records that were predicted creditworthy “ Actual_Creditworthy “. In other side we had 4 records that were predicted non-creditworthy “ Actual_Creditworthy “. In addition, there are more non-creditworthy that were predicted 26 records, also there are more 19 records for predicted_Non_Creditworthy.

4- Boosted Model

Report

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: 3947

Figure 14: Boosted Model Report

Q1: Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

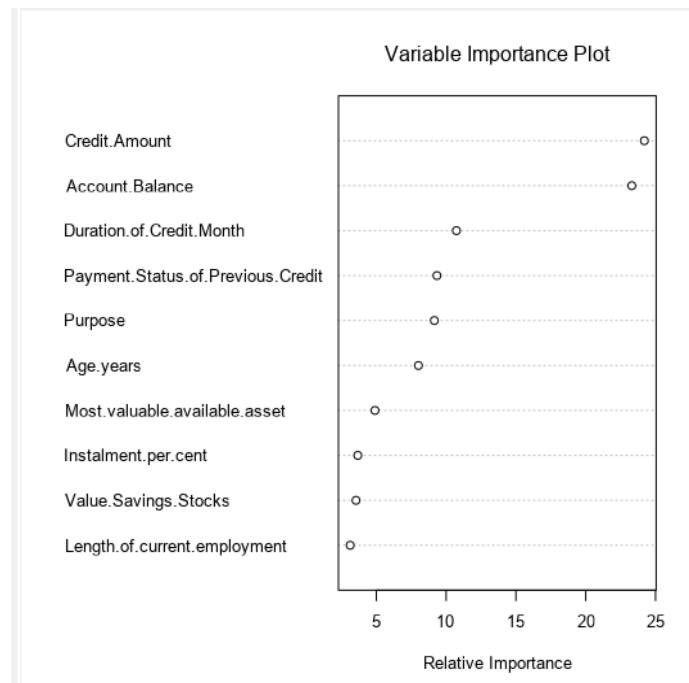


Figure 15: Boosted Variable Importance plot

Q2: 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
Boosted_Model	0.7933	0.8670	0.7532	0.9619	0.4000
<p>Model: model names in the current comparison.</p> <p>Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.</p> <p>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 <i>recall</i>.</p> <p>AUC: area under the ROC curve, only available for two-class classification.</p> <p>F1: F1 score, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The <i>precision</i> 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.</p>					
Confusion matrix of Boosted_Model					
	Actual_Creditworthy		Actual_Non-Creditworthy		
Predicted_Creditworthy	101		27		
Predicted_Non-Creditworthy	4		18		

Figure 16: Boosted Model Comparison Report

Based on the forest model comparison report the accuracy is 0.7933. In other side, the creditworthy were predicted quite high 96% and the non-creditworthy predicted 40%. Based on the confusion matrix shows 101 records that were predicted creditworthy “ Actual_Creditworthy “. In other side we had 4 records that were predicted non-creditworthy “ Actual_Creditworthy “. In addition, there are more non-creditworthy that were predicted 27 records , also there are more 18 records for predicted_Non_Creditworthy.

- **Altryx workflow:**



Figure 17: Build Model workflow

Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if $Score_Creditworthy$ is greater than $Score_NonCreditworthy$, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
 - Overall Accuracy against your Validation set
 - Accuracies within "Creditworthy" and "Non-Creditworthy" segments
 - ROC graph
 - Bias in the Confusion Matrices

Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

Model Comparison Report					
Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.6867	0.7892	0.6806	0.8381	0.3333
Forest_Model	0.8000	0.8707	0.7342	0.9619	0.4222
Boosted_Model	0.7933	0.8670	0.7532	0.9619	0.4000
Logistic_Regression	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, $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. 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.

Figure 18: Model Comparison Report

According to the overall accuracy against the validation set, the Forest Model preformed best, which is 80%. Its accuracies for creditworthy and non-creditworthy are among the highest of all.

Confusion matrix of Boosted_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

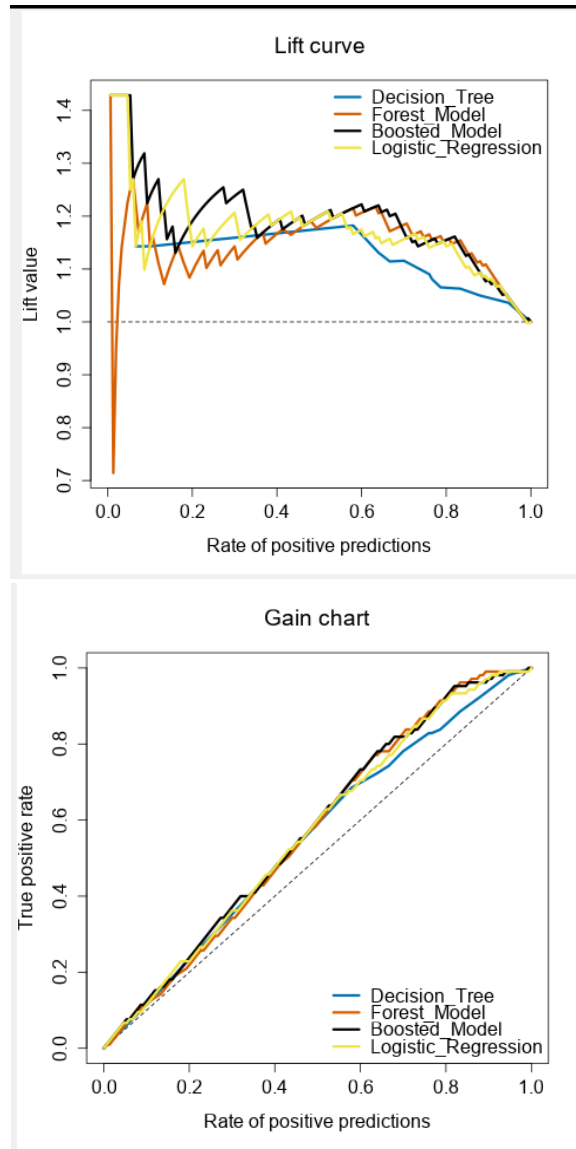
Confusion matrix of Decision_Tree		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	88	30
Predicted_Non-Creditworthy	17	15

Confusion matrix of Forest_Model		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of Logistic_Regression		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Figure19: Confusion Matrix

Based on the confusion matrix , forest model had 101 records that predicted creditworthy that “Actual_Creditworthy” which is highest compared to the rest of the model. In other side, forest model had 19 records that predicted Non-Creditworthy that “Actual_Creditworthy”.



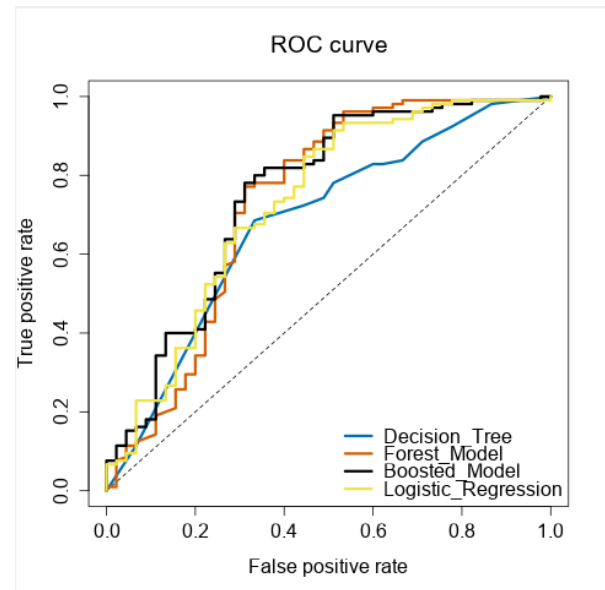
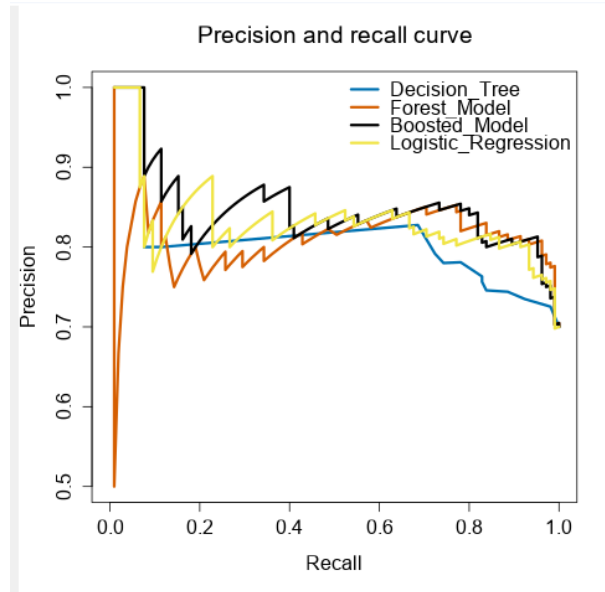
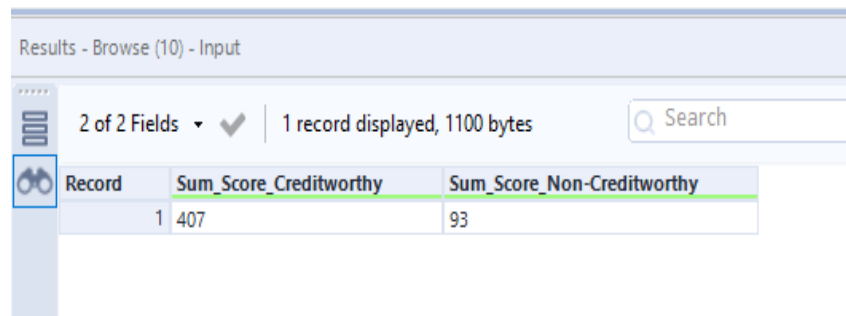


Figure 20: Model Comparison Repot

In addition, the ROC curve shows that forest model has the best overall true positive rates; it one of the model that has the highest curve among all four model.

- **How many individuals are creditworthy?**

Based on the score on the new customer, there are 407 individual are creditworthy.



The screenshot shows a data browser window titled "Results - Browse (10) - Input". It displays a table with 2 columns: "Sum_Score_Creditworthy" and "Sum_Score_Non-Creditworthy". The first record shows a value of 407 for creditworthy and 93 for non-creditworthy. The interface includes a search bar and a status bar indicating "1 record displayed, 1100 bytes".

Record	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	407	93

Figure 21: Sum Score

- **Altryx Workflow:**

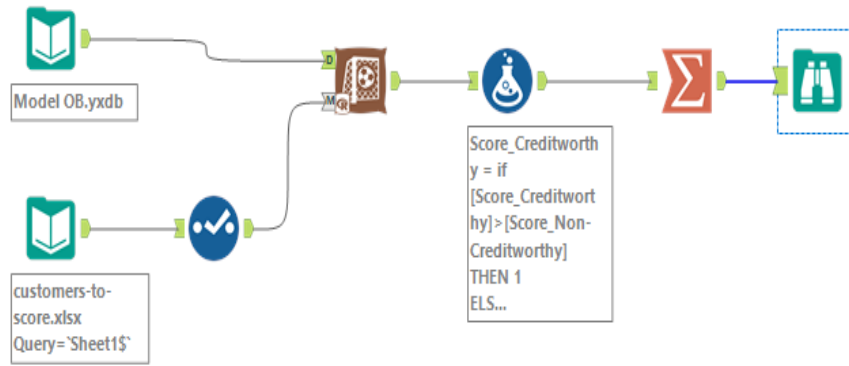


Figure 22: Score model Workflow