Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

The goal is to know whether or not customers who have opted for a loan are creditworthy for an extended period. It is necessary to define whether the customer is credible or not.

Data on beyond packages consisting of Account Balance and Credit Amount and listing of clients to be processed are required so as to tell the ones decisions. To build the models, personal customer data and bank details are required.

The trouble needs a binary model to be solved. Decision tree, boosted tree, forest model and logistics regression are binary classification models that can be used to analyze and determine creditworthy of new customers based on old ones.

Step 2: Building the Training Set

When summarizing all statistics fields, Duration in Current Address has 69% lacking statistics and ought to be eliminated. While Age Years has 2.4% lacking statistics, it's far suitable to impute the lacking statistics with the median age. Median age is used rather than imply because the statistics is skewed to the left as proven below.

Figure 1 – Analysis of Training Set

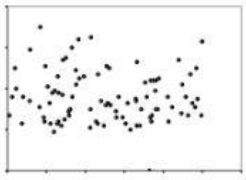
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	

Figure 2 – 2.4% Age-years missing values

In addition, Concurrent Credits and Occupation has one fee whilst Guarantors, Foreign Worker and No of Dependents display low variability wherein extra than 80% of the statistics skewed closer to one statistics. These statistics ought to be eliminated so as now no longer to skew our evaluation results. Telephone subject ought to additionally be eliminated because of its irrelevancy to the patron creditworthy.

Step 3: Training Classification Models

First it was created Estimation and Validation samples where 70% of dataset should go to Estimation and 30% of entire dataset should be reserved for Validation. The following models were created: Logistic Regression, Decision Tree, Forest Model, Boosted Model.

3.1. Logistic Regression (Stepwise)

The outcomes display that 6 variables are significant for the logistig regression model:

- account balance,
- purpose new car,
- credit amount,
- payment status of previous credit some problem,
- length of current employment,
- instalment per cent.

Relatório para o modelo de regressão Logistic LR

Resumo básico

Chamada:
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)

Desvios residuais (deviance residuals):

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

Coeficientes:

	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	.

Códigos de significância: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Parâmetro de dispersão para binomial assumido como 1)

Desvio nulo (null deviance): 413.16 em 349 graus de liberdade
Desvio residual (residual deviance): 328.55 em 338 graus de liberdade
R-quadrado de McFadden: 0.2048; critério de informação de Akaike: 352.5

Número de iterações do escore de Fisher: 5

Análise tipo II de testes de desvio (deviance)

Figure 3 – Stepwise logistic regression report

This model has an accuracy of 76%. The results when this model was tested against 30% of the data is shown in figure below.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR	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 LR

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

Figure 4 – Comparison report between train and test data set for logistic regression model

3.2. Decision Tree

Using Credit Application Result as the target variable, the overall accuracy for decision tree model is 79.1%.

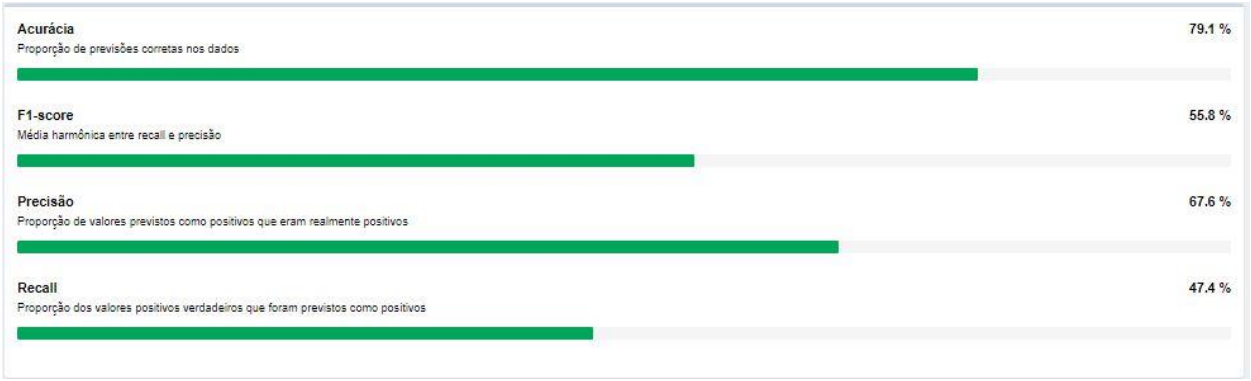
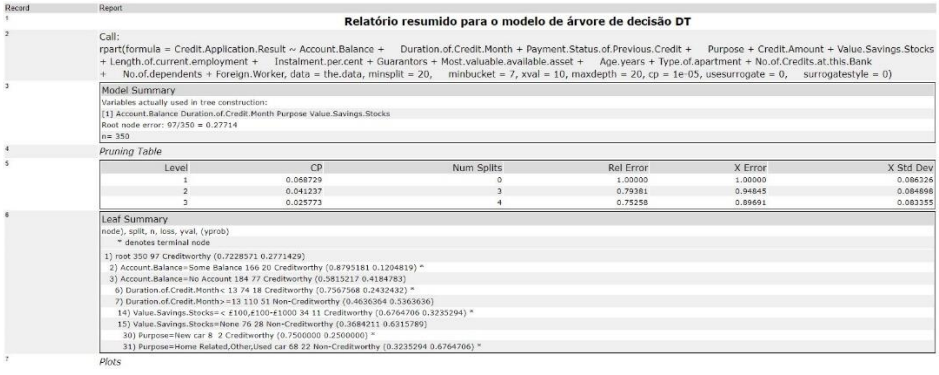


Figure 5 – Accuracy, F1-score, precision and recall for decision tree model

Below is a summary report for the decision tree model.



Record | Report

Relatório resumido para o modelo de árvore de decisão 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 + Guarantors + Most.valuable.available.asset + Age.years + Type.of.apartment + No.of.Credits.at.this.Bank + No.of.dependents + Foreign.Worker, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose 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	0.94845	0.084698
3	0.025773	4	0.75258	0.89691	0.083355

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)
30) Purpose=New car 8 2 Creditworthy (0.7500000 0.2500000) *
31) Purpose=Home Related,Other,Used car 68 22 Non-Creditworthy (0.3235294 0.6764706) *

Plots

Figure 6 - Summary report for the decision tree model

When the decision tree model is confronted with the 30% of test data, the accuracy is 74.67%.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT	0.7467	0.8304	0.7035	0.8857	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 DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Figure 7 – Comparison report between train and test data set for decision tree model

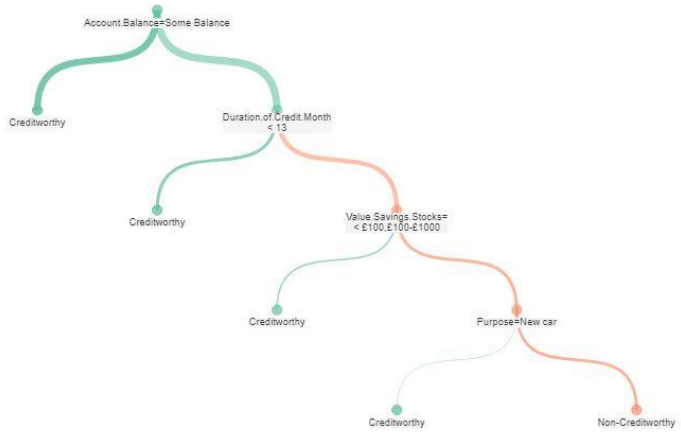


Figure 8 – Decision tree model plot

3.3. Forest Model

Record	Report
1	*Resumo básico*
2	Chamada: 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 + Guarantors + Most.valuable.available.asset + Age.years + Concurrent.Credits + Type.of.apartment + No.of.Credits.at.this.Bank + Occupation + No.of.dependents + Foreign.Worker, data = the.data, ntree = 500, replace = TRUE)
3	Tipo de floresta: classification Número de árvores: 500 Número de variáveis consideradas em cada divisão: 4
4	Estimativa out-of-bag (OOB) da taxa de erro: 22.3%
5	Matriz de confusão:
6	

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.079	233	20
Non-Creditworthy	0.598	58	39

Figure 9 – Basic summary of forest model

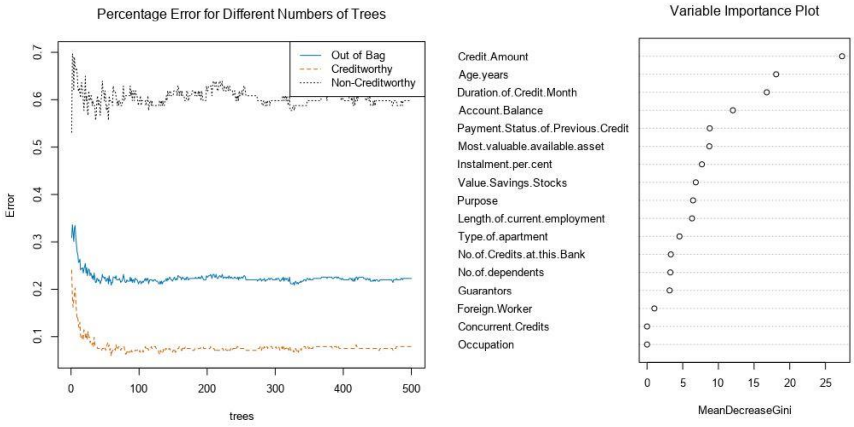


Figure 10 – Percentage error for different number of trees and variable importance

Important: Credit amount, age aears and duration of credit month are the 3 most important variables for forest model.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM	0.7800	0.8584	0.7706	0.9524	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 FM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Figure 11 – Comparison report between train and test data set for forest model

The overall accuracy for this model is 78% as show above.

3.4. Boosted Model

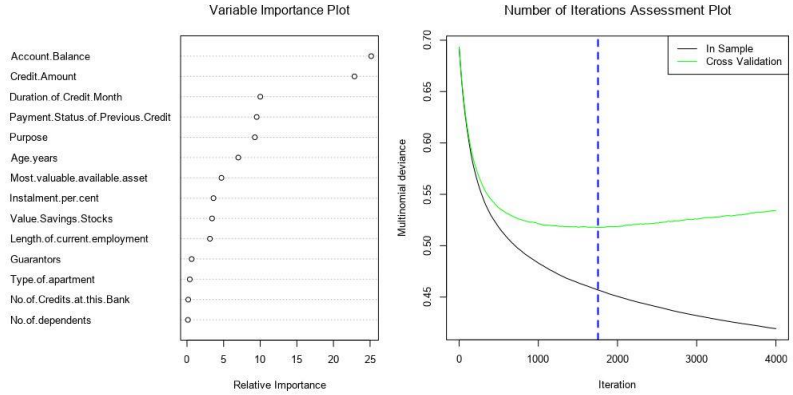
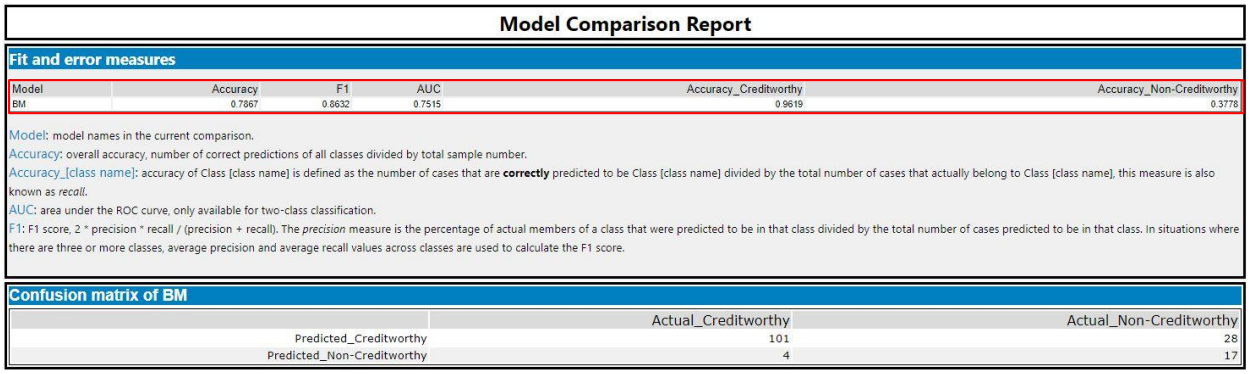


Figure 12 – Variable importance and number of iterations assessment per multinomial deviance

For this model account balance and credit amount are the most important variables by far.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM	0.7867	0.8632	0.7515	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 BM

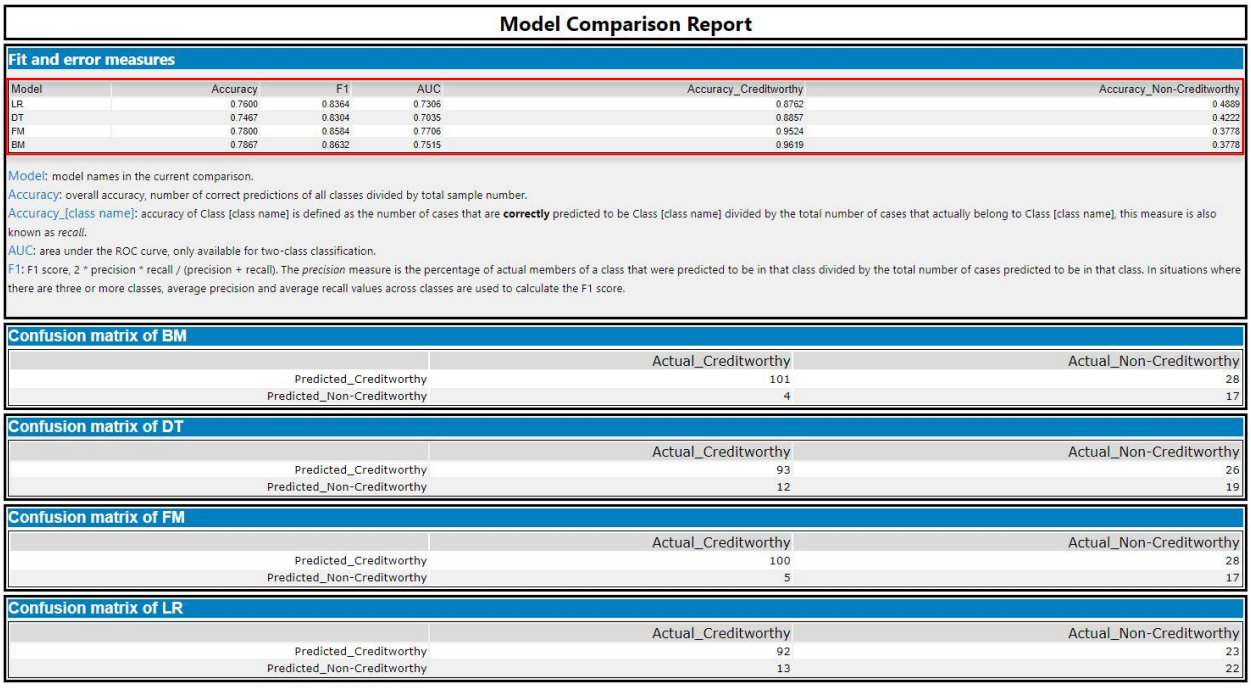
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Figure 13 – Comparison report between train and test data set for boosted model

The comparison report shows na overall boosted model accuracy of 78.67%.

Step 4: Writeup

It is time to decide on the best model and score new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as “Creditworthy”.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR	0.7600	0.8364	0.7306	0.8762	0.4889
DT	0.7467	0.8304	0.7035	0.8857	0.4222
FM	0.7800	0.8584	0.7706	0.9524	0.3778
BM	0.7867	0.8632	0.7515	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 BM

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

Confusion matrix of DT

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of FM

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	28
Predicted_Non-Creditworthy	5	17

Confusion matrix of LR

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

Figure 14 – Comparison report between models

Boosted model could be selected because it gives the very best accuracy at 78.67% towards validation set. But forest model is also a good choice for this prediction purpose. These models

are the only ones that presented a low bias. The following ROC curves shows the comparison between all of four models.

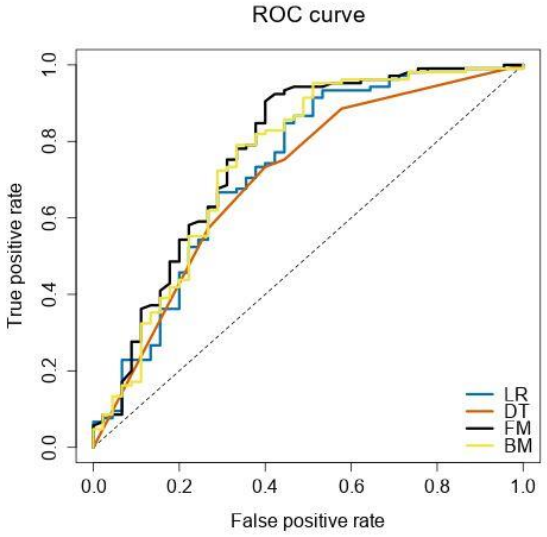


Figure 15 – ROC curve for the models

Although the high accuracy and low bias is present in the forest and boosted model, the **forest model** was chosen because it is the first to reaches the highest point in the ROC curve as it can be viewed at the figure above.

After deciding on the **forest model** to predict the creditworthyness of the brand new clients, the Alteryx workflow was adjusted by importing the new data, run a score with the forest model against it and then clean up the data to get an exact outcome (0 or 1) whether the person is creditworthy or not.

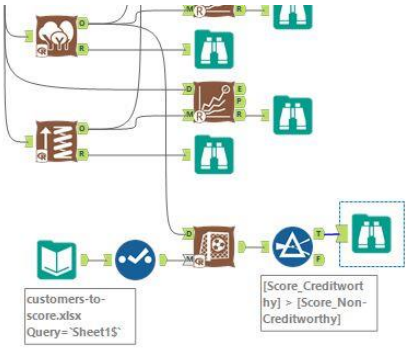


Figure 16 – Forest model was chosen

The end result is that **408 people of the brand new clients are creditworthy!**

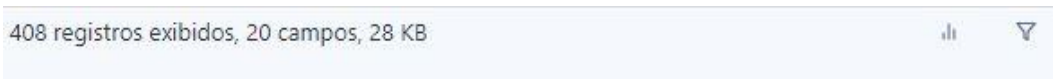


Figure 17 – Number of new clientes classified as creditworthy

Alteryx flow

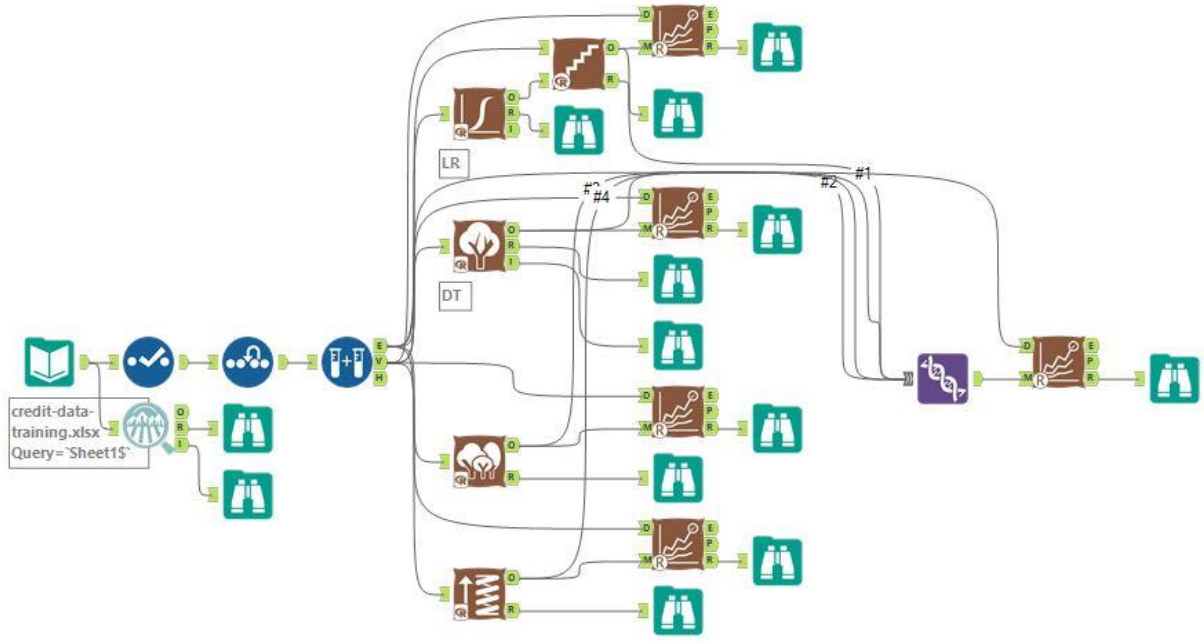


Figure 18 – Alteryx flow for predictiong default risk project

The figure shows the Alteryx workflow used to predict the number of new clients classified as creditworthy step by step.