Project 4: Predicting Default Risk

Structure of Report

1. Business and Data Understanding

a. Goal of project

- As loans officers at a young and small bank, we need to come up with an efficient solution to classify new customers on whether they can be approved for a loan or not. Using a series of classification models to figure out the best model and provide a list of creditworthy customers to the manager.
- Typycally getting 200 loan applications per week and approves them by hand, but now suddenly nearly 500 loan applications to process this week. This is a great opportunity and but needs to figure out how to process all of these loan applications within one week.
- 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.

b. Datasets overview

- `credit-data-training.xlsx` - This file contains all credit approvals from your past loan applicants the bank has ever completed [Datenbanken].
- `customers-to-score.xlsx` - This is the new set of customers that needed to score on the classification model.

c. Key Decisions:

- What decisions needs to be made?

> which application of new customers will be set as creditworthy and which will not.

- What data is needed to inform those decisions?

> All the data involves with applications need to be taken in to consider such as : `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` , `Telephone` , `Foreign-Worker`

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) need to be used to help make these decisions?

> We need to consider the models following to make decision
> - Logistic Regression with Stepweise
> - Boosted Model
> - Decision Tree
> - Forest Tree

2. Building the Training Set

Dataset 1 `credit-data-training.xlsx` **:**

The data set contains the data from the previous application. With this data, we create a model that is used to classify whether the applications from new customers are creditworthy or not.

	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	Gu
0	Creditworthy	Some Balance	4	Paid Up	Other	1494	£100-£1000	< 1yr	1	
1	Creditworthy	Some Balance	4	Paid Up	Home Related	1494	£100-£1000	< 1yr	1	
2	Creditworthy	Some Balance	4	No Problems (in this bank)	Home Related	1544	None	1-4 yrs	2	

At first we will have an overview about the summary of data in this dataset. Here we can see that, the given dataset had information of 500 customers.

With the summary of all data, it is clear to see that a 69% data of `Duration-in-Current-address` is missing. So we cannot consider this variable to model the classification for the creditworthness of applications. This variable will be removed from the dataset.

Another variable also missing is Age-years with around 2% missing. With a small missing data. For these data, we can generate the missing data with a predicted model, but to keep the problem not so complicatedm we can impute this missing data with mean of data `Age-years` .

Preparing Data to Modelling:

We split the data set into two parts: 70% for Estimation (for training the model) and 30% for Validation to help us verify that we are creating a useful model.

Dataset 2 `customers-to-score.xlsx` :

The dataset 2 includes information about the new customers. From this dataset, we will classify their credit applications with the help of prediction model into 2 type: `creditworthy` or `non-creditworthy`

	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	Dur Cu ad
0	No Account	9	No Problems (in this bank)	Home Related	2799	None	< 1yr	2	None	
1	No Account	12	No Problems (in this bank)	Home Related	2122	None	< 1yr	3	None	
2	No Account	24	Paid Up	Home Related	3758	£100-£1000	< 1yr	1	None	

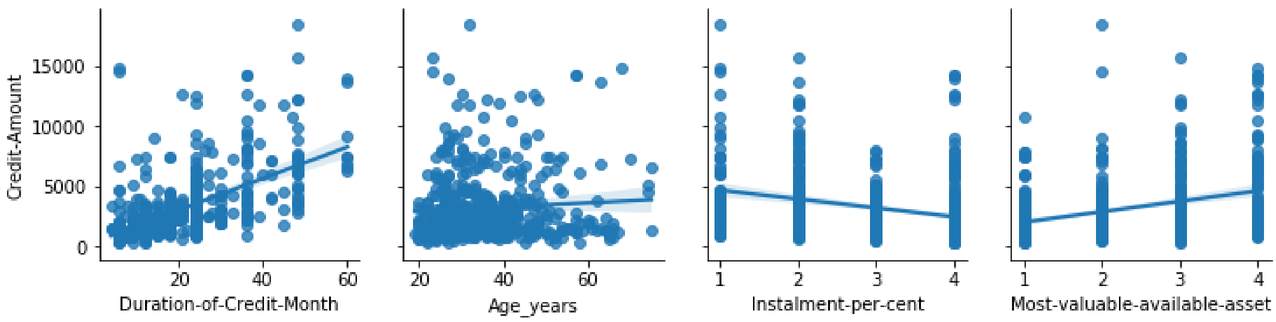
3. Train your Classification Models

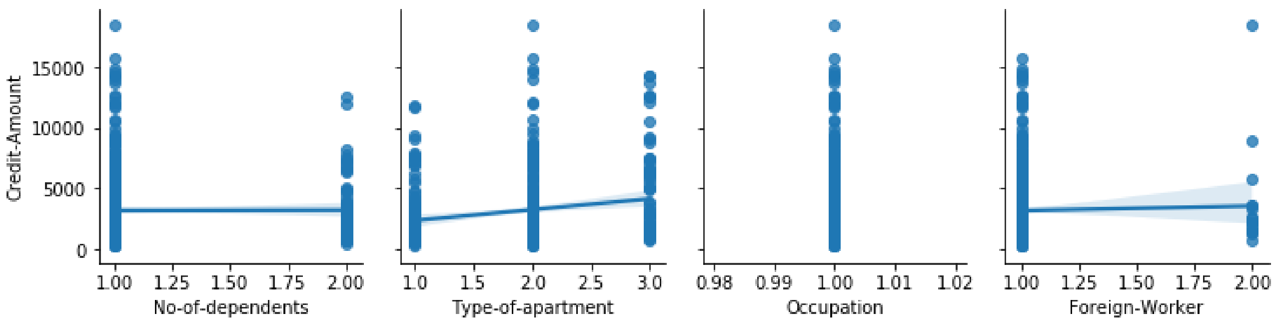
a. Logistic Regression

Data Analysis

At first we will use Logistic Regression to predict the binary outcome `Credit-Application-Result` by analyzing the `Credit-Application-Result` 's relationship with other predictor variables. The `pairplot` gives us an overview about the relationship between `Credit-Application-Result` and other variables from information of presious customers.

Relationship between Application Result and information of customers





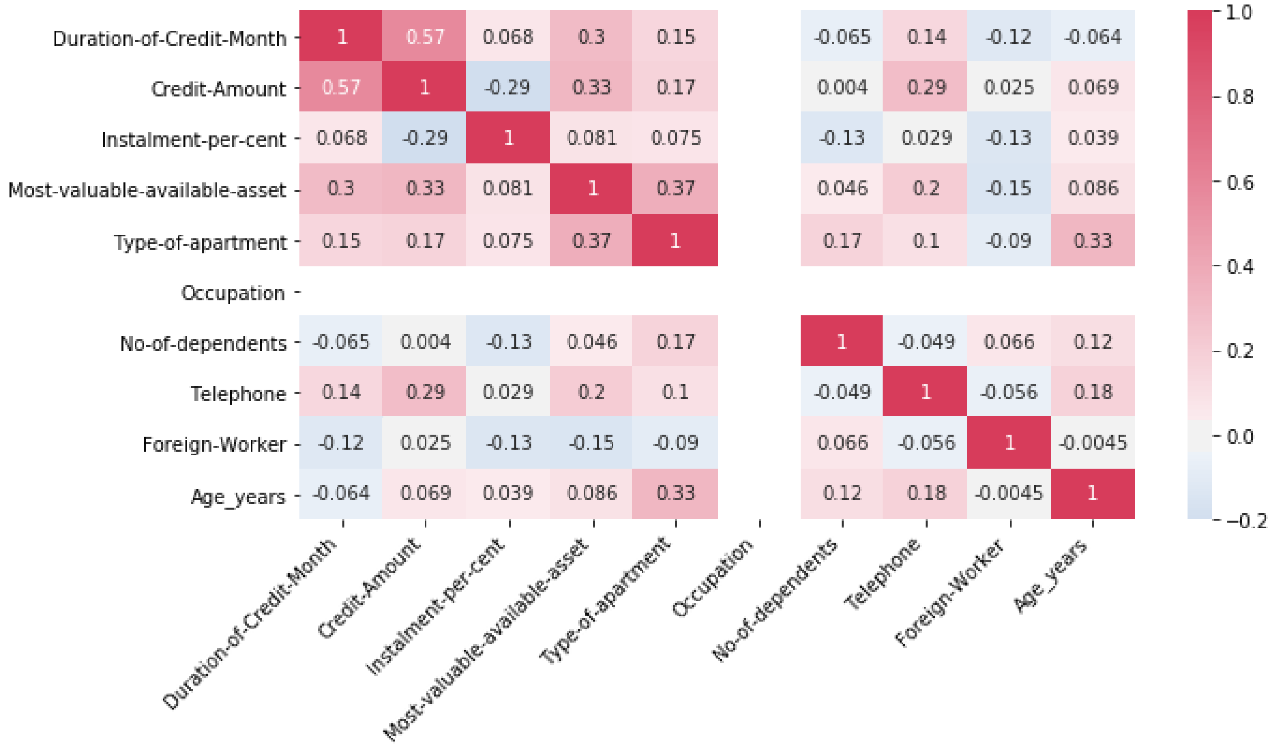
With the Person correlation we can see the most influencing factors on `Credit-Application-Result` with 3 statistically significant factors.

Pearson Correlation Analysis

Focused Analysis on Field Credit.Application.Result.num

	Association Measure	p-value
Duration.of.Credit.Month	-0.263971	0.0010984**
No.of.dependents	-0.214698	0.0083311**
Credit.Amount	-0.191578	0.0188487*
Most.valuable.available.asset	-0.135083	0.0993238.
Type.of.apartment	-0.130247	0.1121440
Instalment.per.cent	-0.110855	0.1768566
Telephone	0.102752	0.2108459
Foreign.Worker	0.093522	0.2549868

To ensure that all factors are not duplicated we can see the correlation matrix, which shows us a weak relationship between variables (all correlations are smaller than 0.7). So all variables can be used as predictor variables in Logistic Regression.



Duration-of-Credit-Month
Credit-Amount
Instalment-per-cent
Most-valuable-available-asset
Type-of-apartment
Occupation
No-of-dependents
Telephone
Foreign-Worker
Age_years
1 0.57 0.068 0.3 0.15 -0.065 0.14 -0.12 -0.064
0.57 1 -0.29 0.33 0.17 0.004 0.29 0.025 0.069
0.068 -0.29 1 0.081 0.075 -0.13 0.029 -0.13 0.039
0.3 0.33 0.081 1 0.37 0.046 0.2 -0.15 0.086
0.15 0.17 0.075 0.37 1 0.17 0.1 -0.09 0.33
-0.065 0.004 -0.13 0.046 0.17 1 -0.049 0.066 0.12
0.14 0.29 0.029 0.2 0.1 -0.049 1 -0.056 0.18
-0.12 0.025 -0.13 -0.15 -0.09 0.066 -0.056 1 -0.0045
-0.064 0.069 0.039 0.086 0.33 0.12 0.18 -0.0045 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Data Modelling

After that we run the Logistic Regression model with the target variables `Credit-Application-Result` and predictor variables. Using a technique `Stepwise regression`, we can get automatically the best predictor variables.

Report for Logistic Regression Model

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

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

Observe the report of Logistic Model, we can see the relationship between the target and predictor variable (with `p_value` < 0.05 then predictor variable is statistically significant). Another factor is that for this model `R-squared = 0.2048` which present a quite weak model.

Validation

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

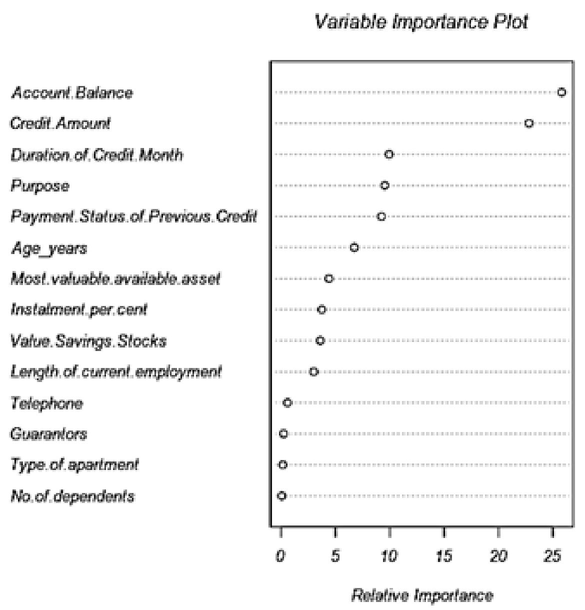
Confusion matrix of Logistic_Regression

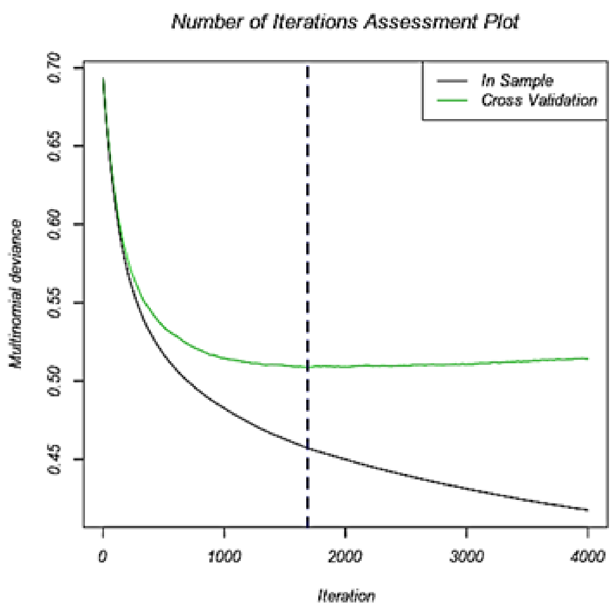
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

With the support of `Model Comparision` , we can see that the Accuracy of `Logistic Regression` for this prediction is 76% with 87,67% accuracy for Creditworthy but only 48.89% for Non-Creditworth.

b. Boosted Model

With Boosted Model, a machine learning technique for regression and classification problems, we can get the best predictor variables to predict `Credit-Application-Result` . With the Variable Importance Plot we can see that with this model, `Account Balance` , and `Credit-Amount` are 2 most important factors to predict the result of application for credit. Next one are `Duration-of-Credit-Month` , `Purpost` and `Payment-Status-of-Previous-Credit` .





Data Validation

To validate this Model we use `Comparison Method` to test the validation of the model:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7460	0.9619	0.3778

Confusion matrix of Boosted_Model

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

With the Model `Comparison Report` , the general accurancy of `Boosted Model` makes up 78.67% with 96.19% accuracy to predict Creditworthy, but only 37.78% for Non-Creditworthy.

c. Decision Tree

Decision tree Model is used in a predictive model to go from observations about the predictor variable to conclusions about the target variable. To predict `Credit-Application-Result`, the Decision Tree Model always choose to use the best predictor variables.

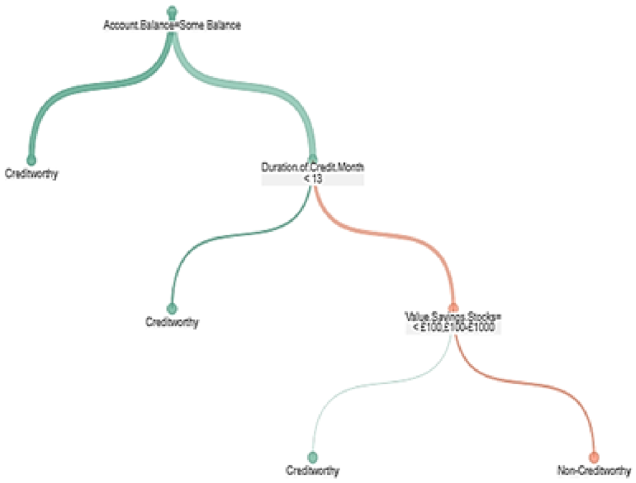
The Model Summary shows that variables used in tree construction are `Account-Balance`, `Value-Savings-Stocks`, and `Duration-of-Credit-Month`.

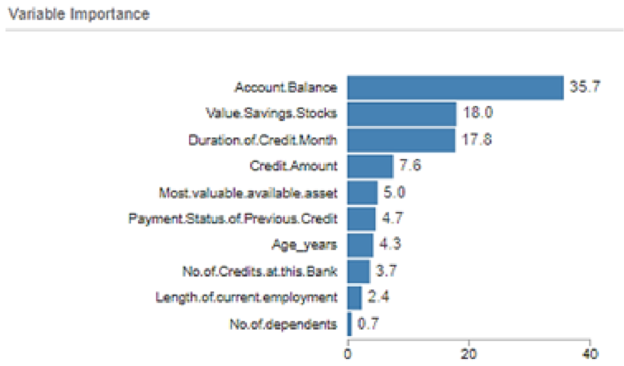
Model Summary
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350

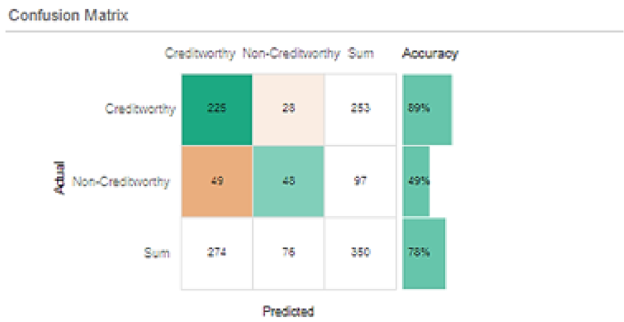
Pruning Table

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) *







Model Validation

With the `Comparison Report` , we can see `Decision Tree` model has 74,67% accuracy. The accuracy rate for Creditworthy is 86.67% and Non-Creditworthy is 46.67%.

Model Comparison Report

Fit and error measures

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

Confusion matrix of Decision_Tree

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

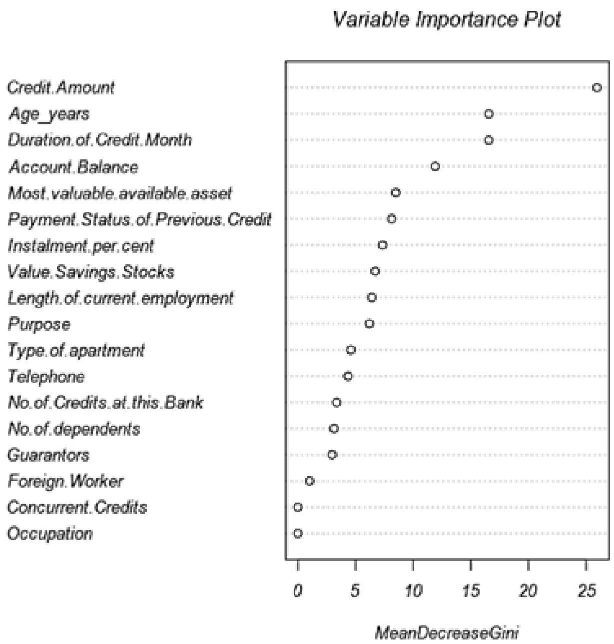
d. Forest Tree

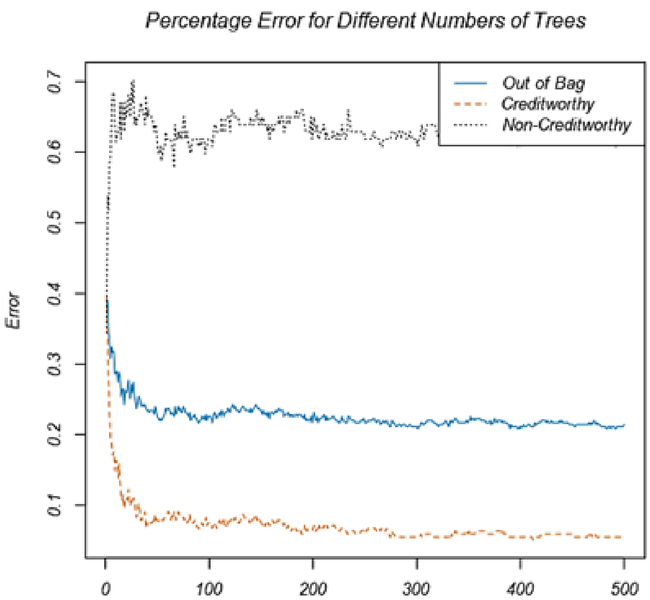
A forest model, a machine learning methods that predict a target variable using predictor variables having influences on the target variable. To predict `Credit-Application-Result` , the Forest Tree Model always choose to use the best predictor variables.

The Model Summary shows that most important variables used in Forest Tree Model are `Credit-Amount` , `Age-years` , and `Duration-of-Credit-Month` and `Account-Balance` .

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 4
OOB estimate of the error rate: 21.4%
Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.055	239	14
Non-Creditworthy	0.629	61	36





Validation

To validate the Forest Tree Model we use `Comparison Model`. With the Model Comparison report, we can see that the Forest Tree Model has a accuracy rate of 78.67%. Out of them, 96.19% predict accurately the application for creditworth, but only 37.78% for non creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Model	0.7867	0.8632	0.7519	0.9619	0.3778

Confusion matrix of Forest_Model

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

4.Conclusion:

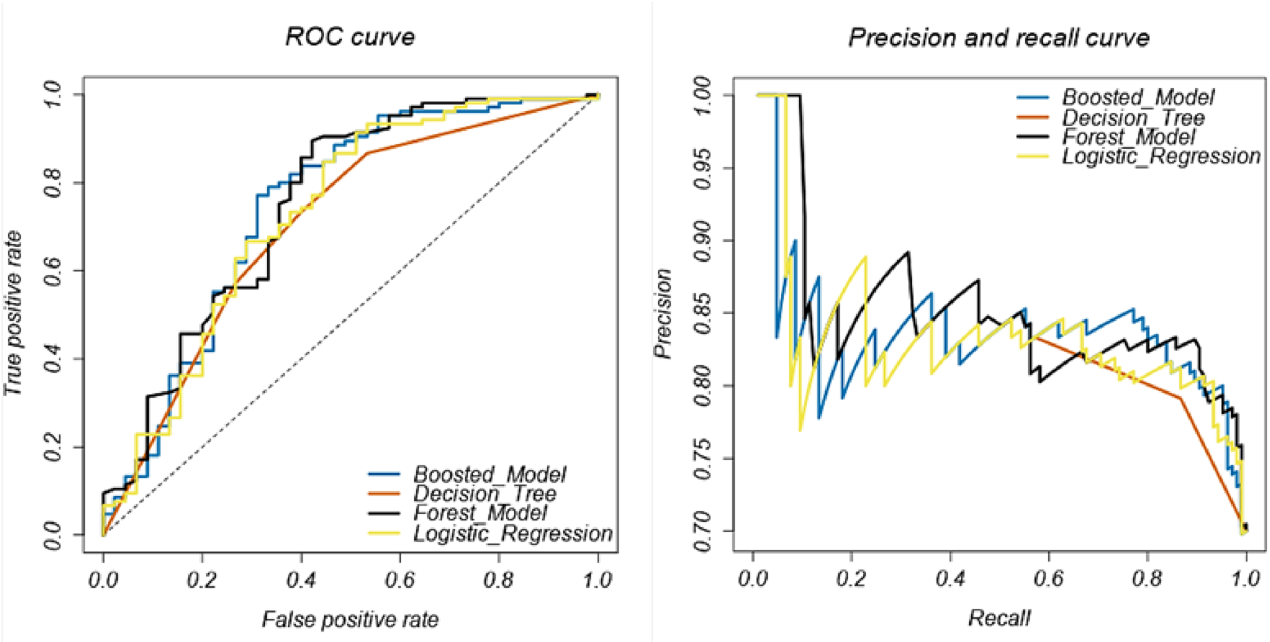
a. Choosing Model

After all, we will find which model is the best to predict the creditworthiness of applications from 500 new customers. With the `Comparison Model`, we will have the Report following:

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7460	0.9619	0.3778
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Model	0.7867	0.8632	0.7519	0.9619	0.3778
Logistic_Regression	0.7600	0.8364	0.7306	0.8762	0.4889



With the report, we can see that `Boosted Model` and `Forest Model` have the best rate of accuracy (78.67% for both). However, The ROC curve gives an expression that `Forest Model` is better when we use such models for prediction of creditworthiness.

b. Prediction of Application

After choosing model, we can use `Score Tool` in Alteryx to calculate the `Credit Application Result` .

	Account.Balance	Duration.of.Credit.Month	Payment.Status.of.Previous.Credit	Purpose	Credit.Amount
341	Some Balance	18	Paid Up	Used car	3049
280	No Account	36	Paid Up	Used car	3446
33	No Account	12	Paid Up	Home Related	1567

3 rows × 22 columns

And we can see that with `Forest Model`, from 500 new customers with given information in Dataset, we will classify into 2 category: `Creditworthy` and `Non-creditworthy` with the number like following:

	Application_Result	Count
0	Creditworthy	417
1	Non-Creditworthy	83

All of the process to choose a suitable model as well as to predict the application result follows the Alteryx Workflow: