

# Project: Creditworthiness

## Step 1: Business and Data Understanding

### Key Decisions:

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

To come up with an efficient solution to classify 500 loan applications of new customers whether they can be approved or not.

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

- *Data to build Predictibe Models:* data on all past applications and the list of customers that need to be processed in the next few days.
- *Overall accuracy of the models* to choose the most efficient solution.

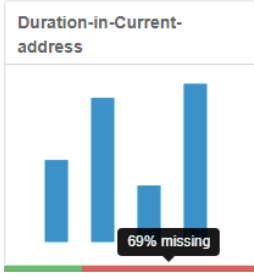
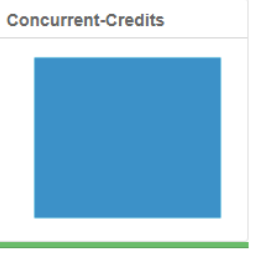
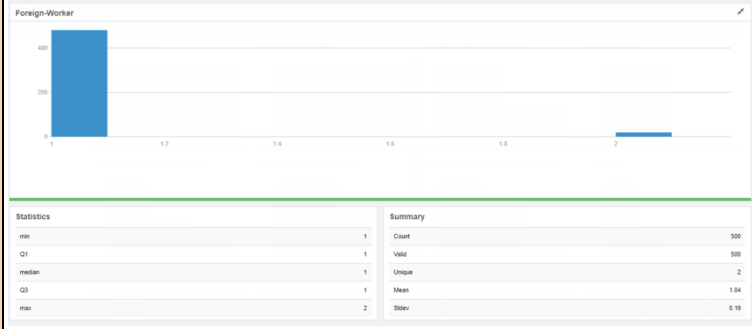
- **What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?**

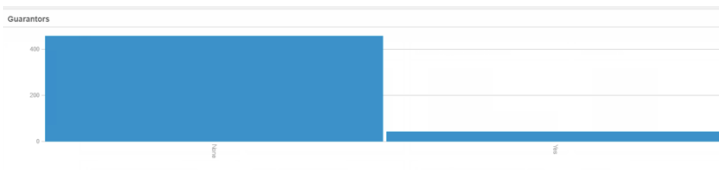

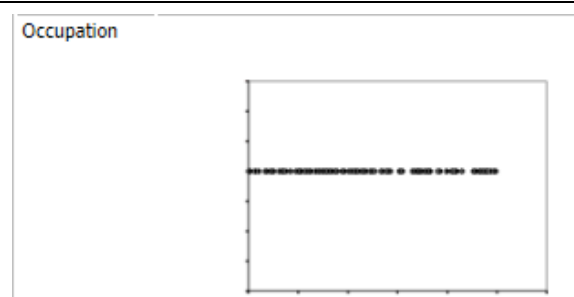
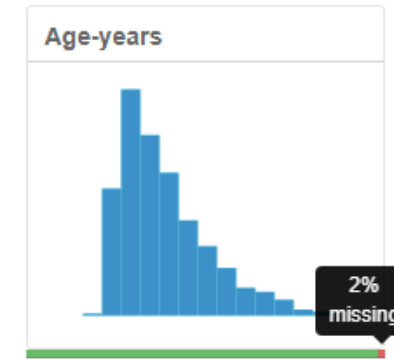
We need to use a Binary Predictibe Classification Models because a binary decision should be made (whether loans will be given to new customers or not).

## Step 2: Building the Training Set

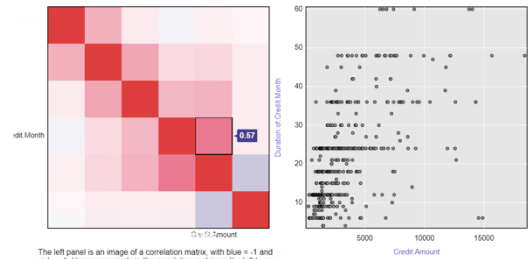
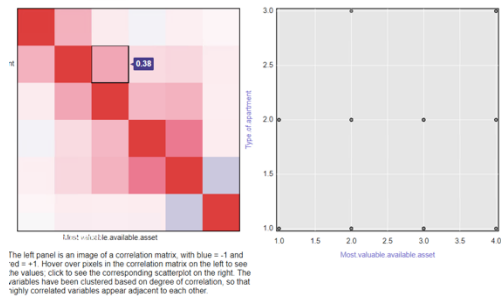
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

I have used a 'Field Summary' tool to find out the missing data in the set.

Field	Decision and Reason
Duration-in-Current-address	 <p><b>The field is removed.</b> About 70% of missing data in the field. Imputation will make a bias in data.</p>
Concurrent-Credits	 <p><b>The field is removed.</b> The field is very uniform. It includes only one value, so we can call it 'low variability' and exclude from the training set.</p>
Telephone	<p><b>The field is removed.</b> There is no logical reason for including the variable.</p>
Foreign-Worker	 <p><b>The field is removed.</b> The field is very uniform. It includes only one value, so we can call it 'low variability' and exclude from the training set.</p>

Guarantors	<div><div>Guarantors</div><table><tr><th colspan="2">Frequent Values</th><th colspan="2">Summary</th></tr><tr><td>None</td><td>487</td><td>Count</td><td>500</td></tr><tr><td>Yes</td><td>43</td><td>Valid</td><td>500</td></tr><tr><td></td><td></td><td>Unique</td><td>2</td></tr><tr><td></td><td></td><td>Blank</td><td>0</td></tr><tr><td></td><td></td><td>Longest</td><td>None</td></tr><tr><td></td><td></td><td>Shortest</td><td>Yes</td></tr></table></div>	Frequent Values		Summary		None	487	Count	500	Yes	43	Valid	500			Unique	2			Blank	0			Longest	None			Shortest	Yes	<p><b>The field is removed.</b></p> <p>The field is very uniform. It includes only one value, so we can call it ‘low variability’ and exclude from the training set.</p>
Frequent Values		Summary																												
None	487	Count	500																											
Yes	43	Valid	500																											
		Unique	2																											
		Blank	0																											
		Longest	None																											
		Shortest	Yes																											
No-of-dependents	<div><div>No-of-dependents</div><table><tr><th colspan="2">Statistics</th><th colspan="2">Summary</th></tr><tr><td>min</td><td>1</td><td>Count</td><td>500</td></tr><tr><td>Q1</td><td>1</td><td>Valid</td><td>500</td></tr><tr><td>median</td><td>1</td><td>Unique</td><td>2</td></tr><tr><td>Q3</td><td>1</td><td>Mean</td><td>1.15</td></tr><tr><td>max</td><td>2</td><td>Stdev</td><td>0.35</td></tr></table></div>	Statistics		Summary		min	1	Count	500	Q1	1	Valid	500	median	1	Unique	2	Q3	1	Mean	1.15	max	2	Stdev	0.35	<p><b>The field is removed.</b></p> <p>The field is very uniform. It includes only one value, so we can call it ‘low variability’ and exclude from the training set.</p>				
Statistics		Summary																												
min	1	Count	500																											
Q1	1	Valid	500																											
median	1	Unique	2																											
Q3	1	Mean	1.15																											
max	2	Stdev	0.35																											
Occupation	<div><div>Occupation</div></div>	<p><b>The field is removed.</b></p> <p>The field is very uniform. It includes only one value, so we can call it ‘low variability’ and exclude from the training set.</p>																												
Age-years	<div><div>Age-years</div></div>	<p><b>Data is imputed.</b></p> <p>As just 2% of data is missing, reasonable to impute the data.</p>																												

I have checked the correlation of the numerical data fields with each other with an ‘Association Analysis’ tool and have not found the correlation higher than 0.57



## Step 3: Train your Classification Models

### 1. Logistic Regression Results

- 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)	-3.0136120	1.013e+00	-2.9760	0.00292 **
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06 **
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124
Payment.Status.of.Previous.CreditSome Problems	1.2607175	5.335e-01	2.3632	0.01812 *
PurposeNew car	-1.7541034	6.276e-01	-2.7951	0.00519 **
PurposeOther	-0.3191177	8.342e-01	-0.3825	0.70206
PurposeUsed car	-0.7839554	4.124e-01	-1.9008	0.05733 .
Credit.Amount	0.0001764	6.838e-05	2.5798	0.00989 **
Value.Savings.StocksNone	0.6074082	5.100e-01	1.1911	0.23361
Value.Savings.Stocks£100-£1000	0.1694433	5.649e-01	0.3000	0.7642
Length.of.current.employment4-7 yrs	0.5224158	4.930e-01	1.0596	0.28934
Length.of.current.employment< 1yr	0.7779492	3.956e-01	1.9664	0.04925 *
Instalment.per.cent	0.3109833	1.399e-01	2.2232	0.0262 *
Most.valuable.available.asset	0.3258706	1.556e-01	2.0945	0.03621 *
Age.years	-0.0141206	1.535e-02	-0.9202	0.35747
Type.of.apartment	-0.2603038	2.956e-01	-0.8805	0.3786
No.of.Credits.at.this.BankMore than 1	0.3619545	3.815e-01	0.9487	0.34275

- 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?

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogR_Credit_Training	0.7800	0.8520	0.7314	0.9048	0.4889

Accuracy of the Logistic Regression model is 0.7800

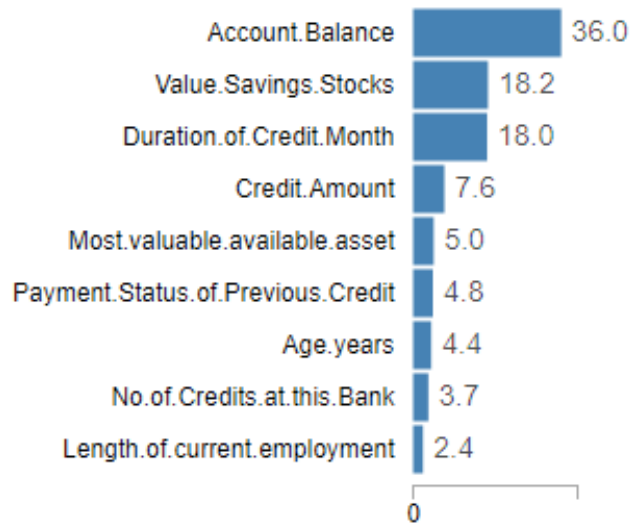
Confusion matrix of LogR_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Clearly seen that the model has a tendency to consider clients as 'creditworthy' with higher accuracy. About 9% of creditworthy clients were predicted as 'non-creditworthy', when 51% of non-creditworthy clients were predicted as 'creditworthy'.

## 2. Decision Tree Results

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

Variable Importance



- 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?

### Fit and error measures

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

Accuracy of the Decision Tree Model is 0.7467

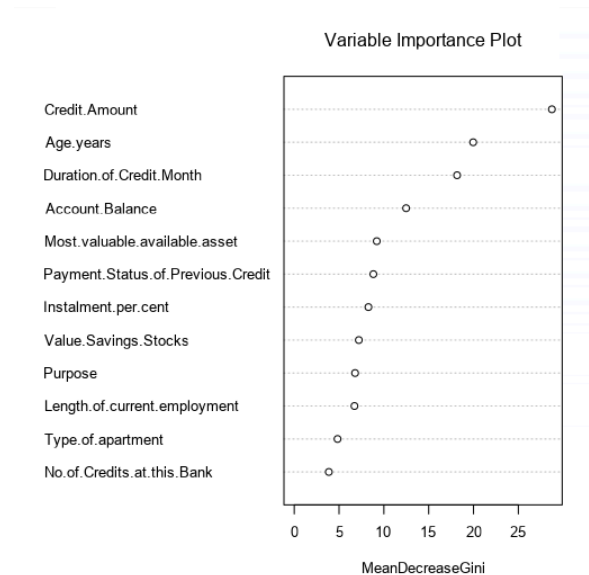
### Confusion matrix of DT\_Credit\_Training

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

The Decision Tree Model has a tendency to consider clients as 'creditworthy' with higher accuracy. 13% of creditworthy clients were predicted as 'non-creditworthy', when 53% of non-creditworthy clients were predicted as 'creditworthy'.

### 3. Forest Model Results

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



- 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?

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit_Training	0.7933	0.8681	0.7368	0.9714	0.3778

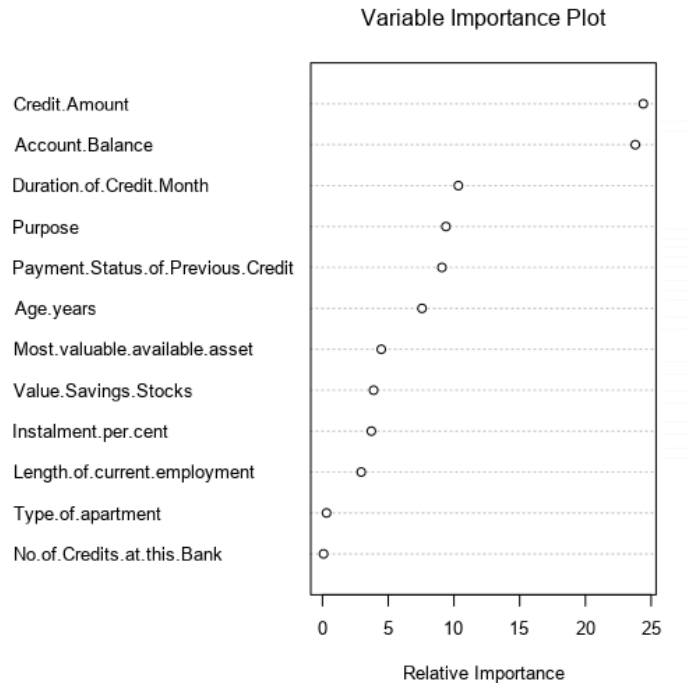
Accuracy of the Forest Model is 0.7933

Confusion matrix of FM_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

The Forest Model has a tendency to consider clients as 'creditworthy' with higher accuracy. 2% of creditworthy clients were predicted as 'non-creditworthy', when 62% of non-creditworthy clients were predicted as 'creditworthy'.

## 4. Boosted Model Results

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



- 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?

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit_Training	0.7867	0.8632	0.7524	0.9619	0.3778

Accuracy of the Boosted Model is 0.7867

Confusion matrix of BM_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

The Boosted Model has a tendency to consider clients as 'creditworthy' with higher accuracy. About 4% of creditworthy clients were predicted as 'non-creditworthy', when almost 62% of non-creditworthy clients were predicted as 'creditworthy'.

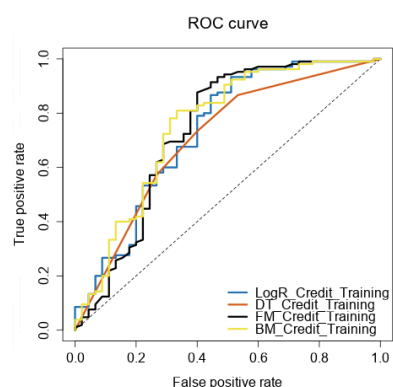


## Step 4: Writeup

According to overall accuracy against the validation set, **the Forest Model** has the highest F1 value, which is **0.8681**

Fit and error measures					
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LogR_Credit_Training	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit_Training	0.7467	0.8273	0.7054	0.8667	0.4667
FM_Credit_Training	0.7933	0.8681	0.7368	0.9714	0.3778
BM_Credit_Training	0.7867	0.8632	0.7524	0.9619	0.3778

Within Forest Model, there is the highest accuracy of correctly predicted cases for “Creditworthy” segment which is 0.9714. Accuracy of “Non-Creditworthy” segments is 0.3778.



As we see on the ROC curve graph the Forest Model reaches the highest point faster and remains higher more than other models, that proves that the Forest Model more accurately in separating ‘Creditworthy’ and ‘Non-creditworthy’ classes.

Confusion matrix of BM_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

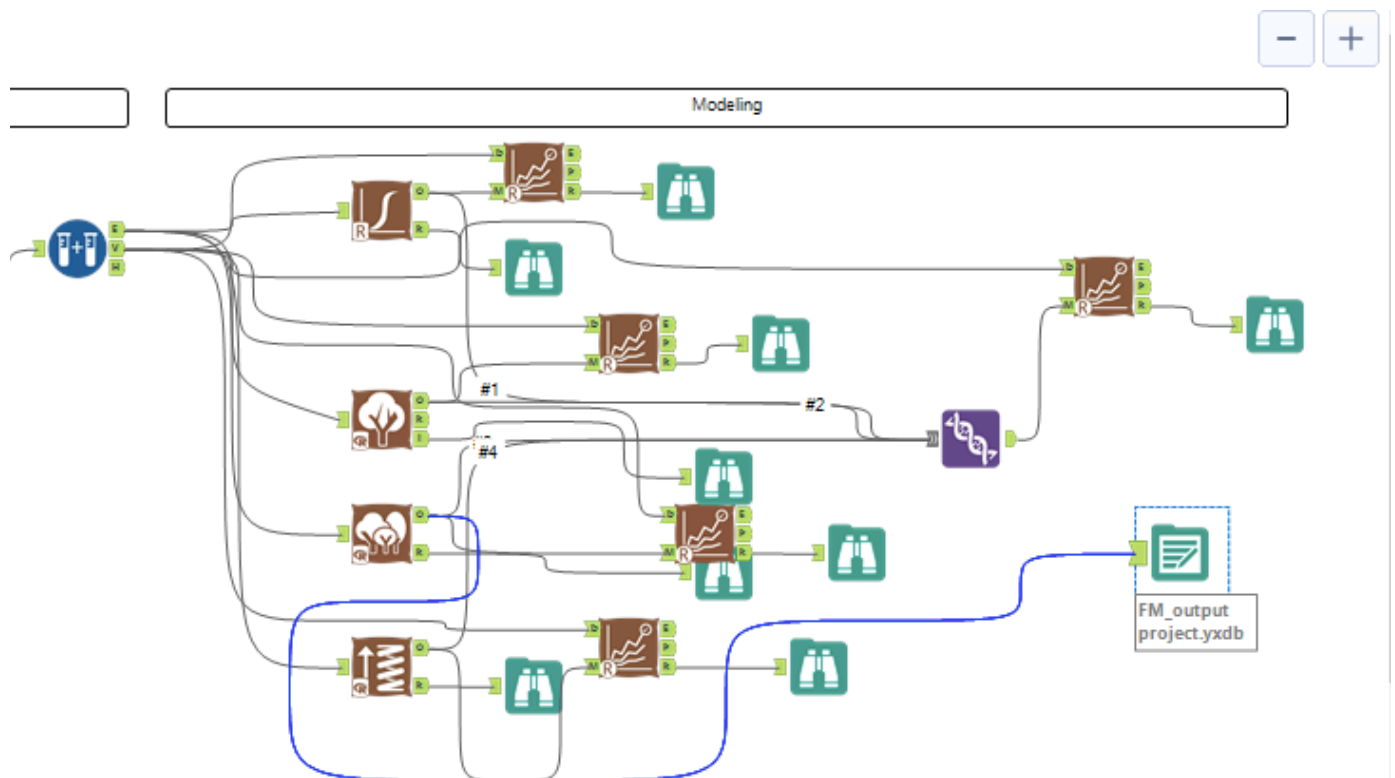
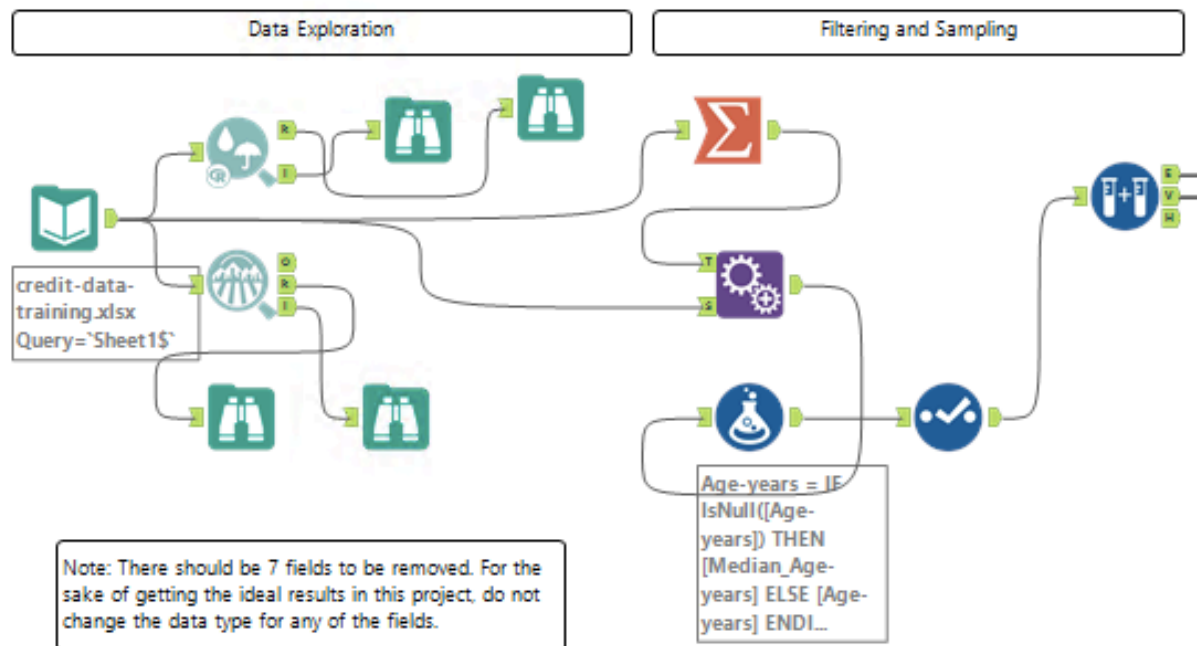
Confusion matrix of DT_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of FM_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of LogR_Credit_Training		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

**The Forest Model has the highest accuracy in comparison with other models. I recommend to use the Forest Model as the model to predict creditworthy customers.** As other models, Forest Model has a tendency to consider clients as ‘creditworthy’ more frequently. Just about 2% of creditworthy clients were predicted as ‘non-creditworthy’ and about 62% of non-creditworthy clients were predicted as ‘creditworthy’.

## A pic of my workflow



- How many individuals are creditworthy?

After applying the forest model to predict individuals who are creditworthy, **408 customers** would qualify for a loan.

