Project: Creditworthiness

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? - The objective is to identify whether customers who applied for loan are creditworthy to be extended one.

- What data is needed to inform those decisions? - Data on past applications such as Account Balance and Credit Amount and list of customers to be processed are required in order to inform those decisions

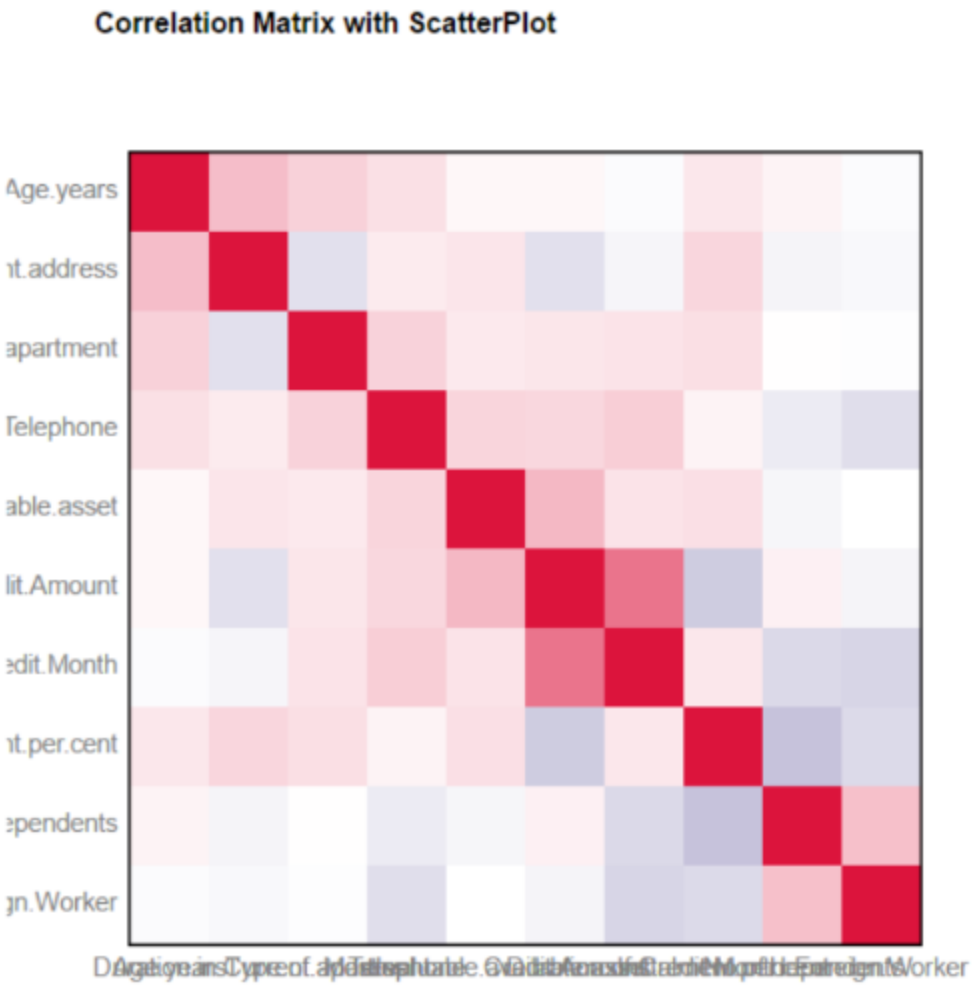
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions? - Binary classification models such as logistics regression, decision tree, forest model and boosted tree will be used to analyze and determine creditworthy customers

Step 2: Building the Training Set

To achieve consistent results reviewers expect.

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. - An association analysis is performed on the numerical variables and there are no variables which are highly correlated with each other, i.e. a correlation of higher than 0.7.



When summarizing all data fields, *Duration in Current Address* has 69% missing data and should be removed.
While *Age Years* has 2% missing data, it is appropriate to impute the missing data with the median age. Median age is used instead of mean as the data is skewed to the left as shown below.

In addition, *Concurrent Credits* and *Occupation* has one value while *Guarantors*, *Foreign Worker* and *No of Dependents* show low variability where more than 80% of the data skewed towards one data.
These data should be removed in order not to skew our analysis results.
Telephone field should also be removed due to its irrelevancy to the customer creditworthy.

Step 3: Train your Classification Models

- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables. – Logistic Regression (Stepwise)
- 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?

Using Credit Application Result as the target variables, Account Balance, Purpose and Credit Amount are the top 3 most significant variables with p-value of less than 0.05.

Report for Logistic Regression Model Stepwise_Logsitic

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, AIC: 352.5

Overall accuracy is around 76.0% while accuracy for creditworthy is higher than non-creditworthy at 80.0% and 62.9% respectively. The model is biased towards predicting customers as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise_Logsitic	0.7600	0.8364	0.7306	0.8000	0.6286

Confusion matrix of Stepwise_Logsitic

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

Decision Tree

Using Credit Application Result as the target variables, Account Balance, Value Savings Stocks and Duration of Credit Month are the top 3 most important variables. The overall accuracy is 74.7%.

Accuracy for creditworthy is 79.1% while accuracy for non-creditworthy is 60.0%. The model seems to be biased towards predicting customers as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000

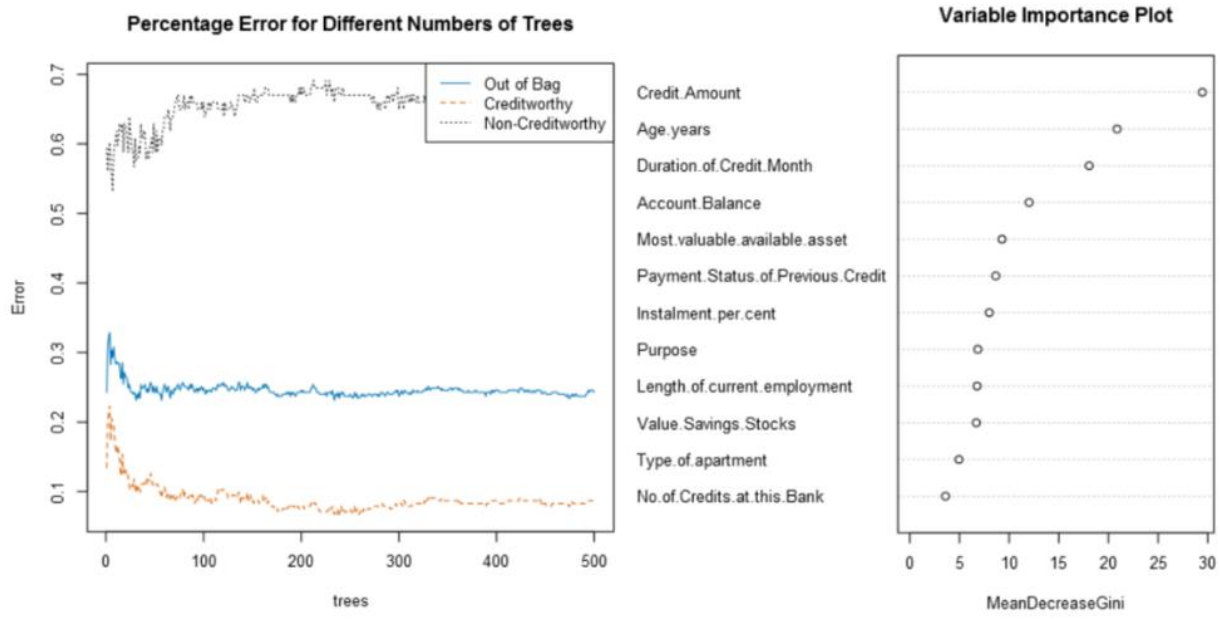
Confusion matrix of DT_Credit

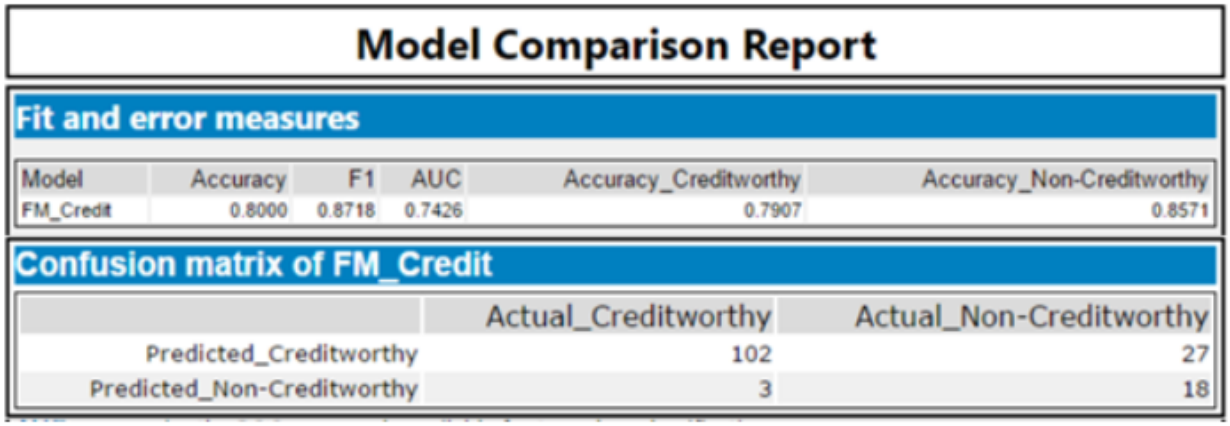
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Forest Model

Using Credit Application Result as the target variables, Credit Amount, Age Years and Duration of Credit Month are the 3 most important variables.

Overall accuracy is 80.0%. The model isn't biased as the accuracies for creditworthy and non-creditworthy are 79.1% and 85.7% respectively, which are comparable.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571

Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

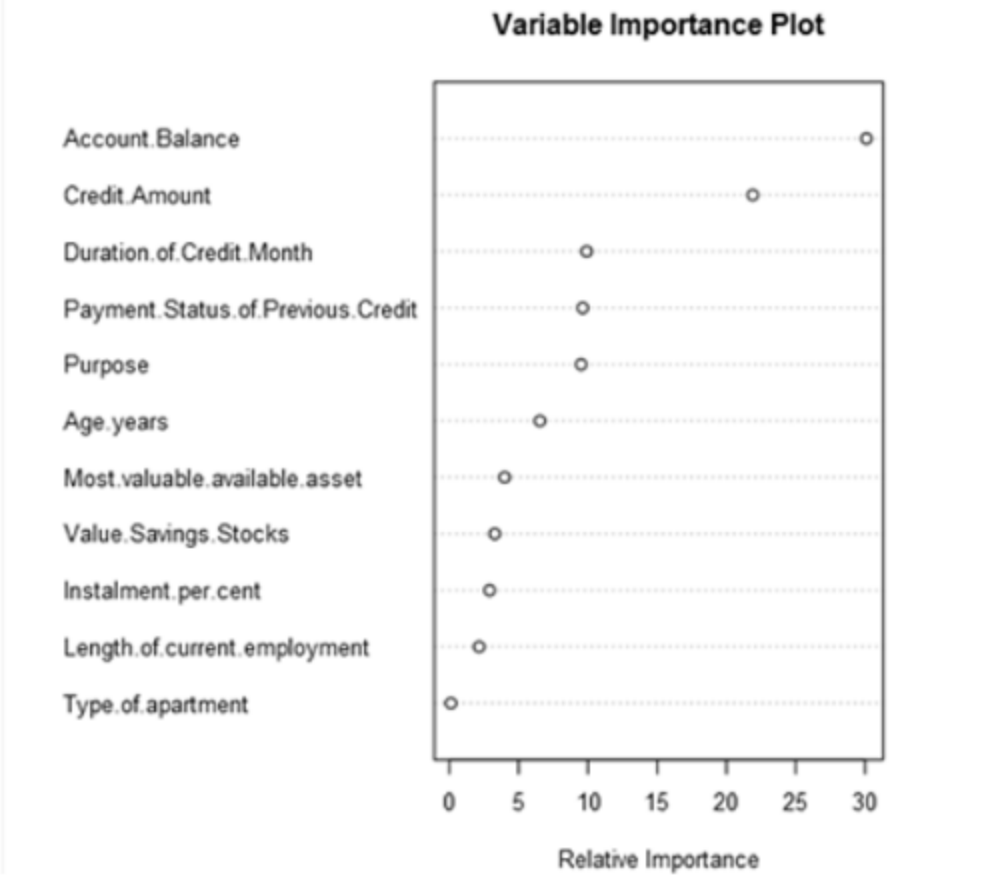
Boosted Model

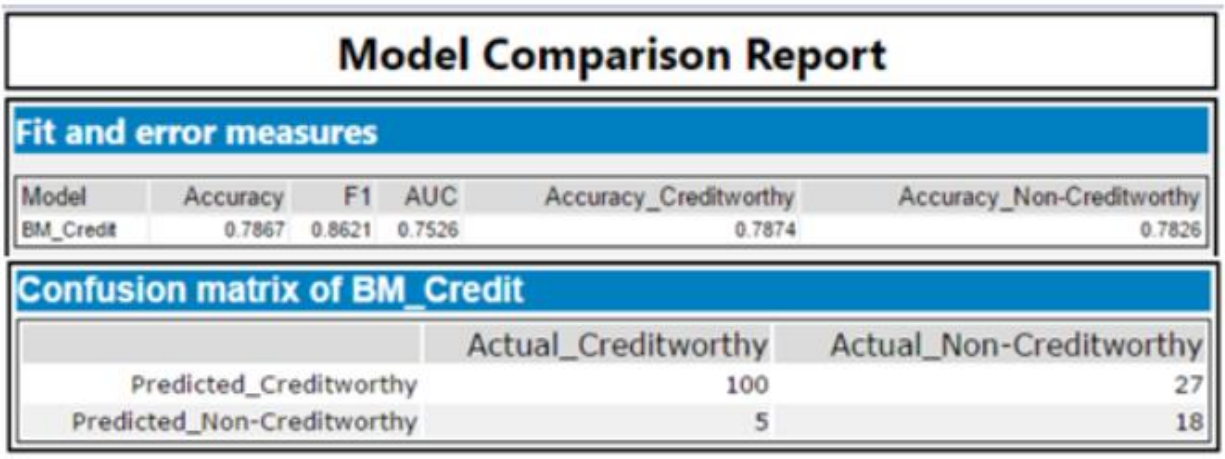
Account Balance and Credit Amount are the most significant variables from figure 10. Overall accuracy for is 76.7%. Accuracies for creditworthy and non-creditworthy are 76.7% and 78.3% respectively which indicates a lack of bias in predicting credit-worthiness of customers.

Report for Boosted Model BM_Credit

Basic Summary:

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2377





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Step 4: Writeup

- 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

- How many individuals are creditworthy?

.

Forest model is chosen as it offers the highest accuracy at 80% against validation set. Its accuracies for creditworthy and non-creditworthy are among the highest of all.

Forest model reaches the true positive rate at the fastes rate. The accuracy difference between creditworthy and non-creditworthy are also comparable which makes it least bias towards any decisions. This is crucial in avoiding lending money to customers with high probability of defaulting while ensuring opportunities are not overlooked by not loaning to creditworthy customers.
Tthere are **408 creditworthy cutomers** using forest models to score new customers.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.7467	0.8273	0.7054	0.7913	0.6000
FM_Credit	0.8000	0.8718	0.7426	0.7907	0.8571
BM_Credit	0.7867	0.8621	0.7526	0.7874	0.7826
Stepwise_Logsitic	0.7600	0.8364	0.7306	0.8000	0.6286

Confusion matrix of BM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DT_Credit

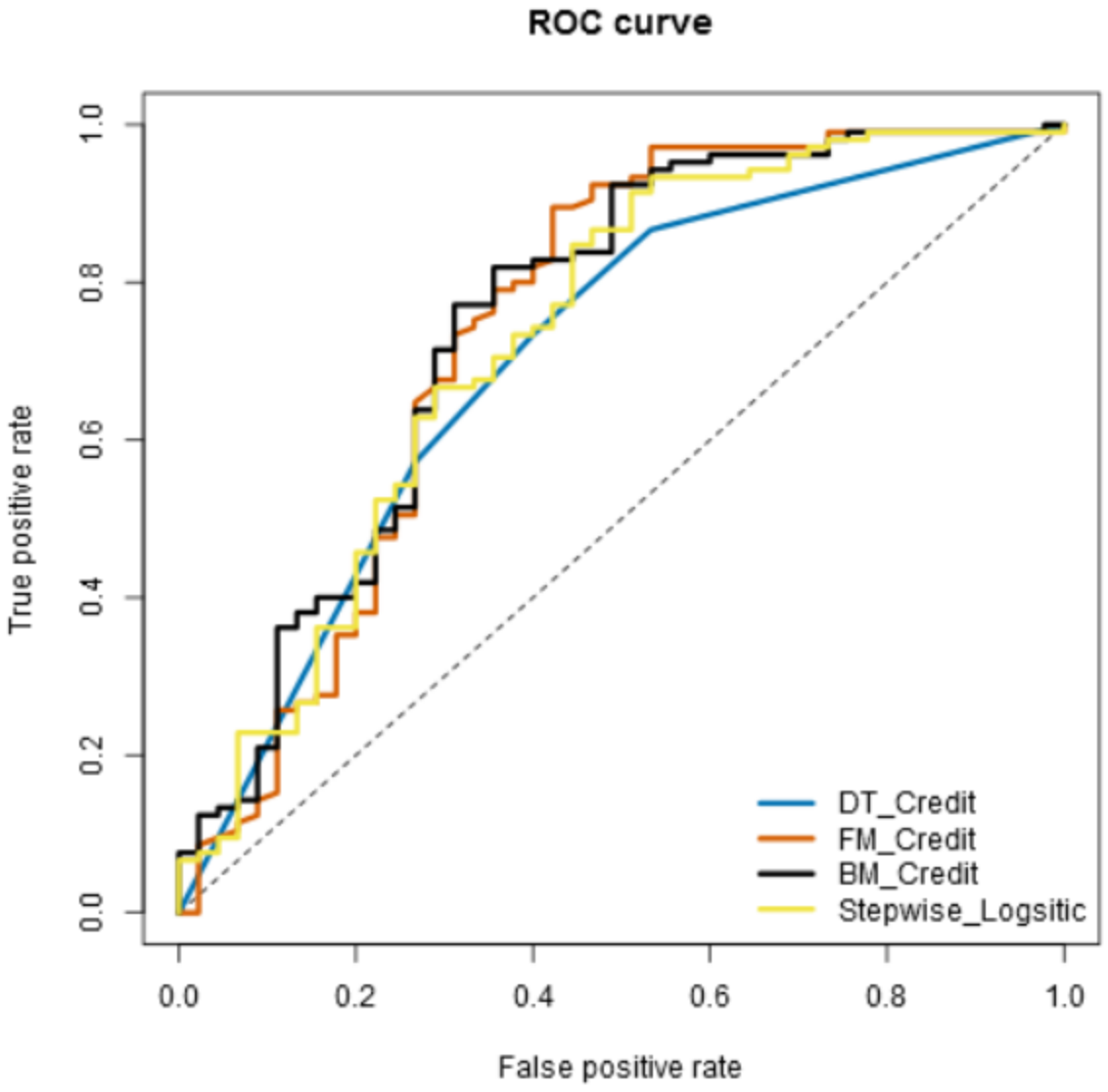
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

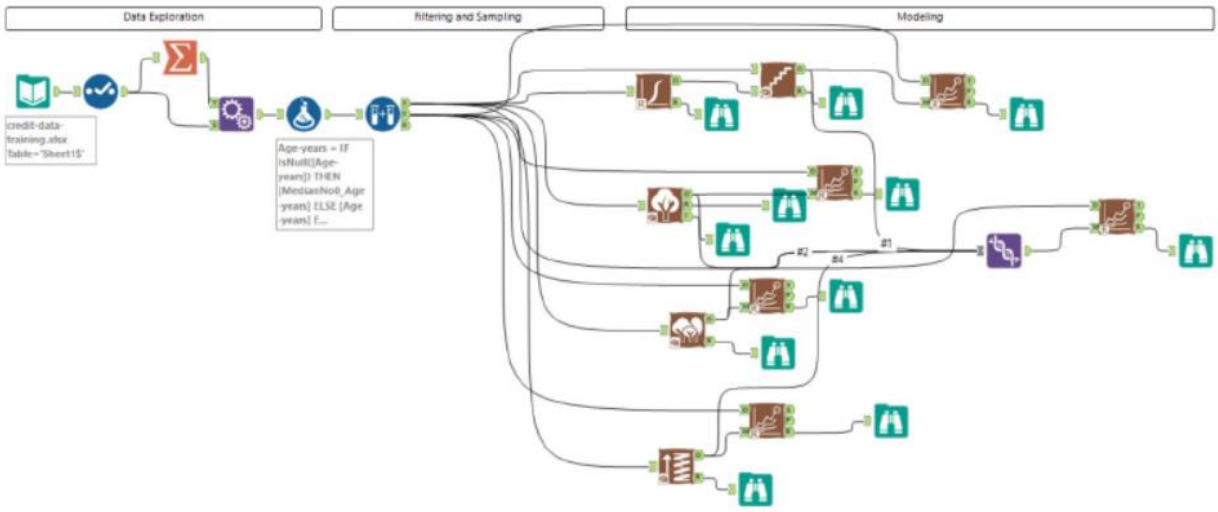
Confusion matrix of FM_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of Stepwise_Logsitic

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


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
DT_Credit
FM_Credit
BM_Credit
Stepwise_Logsitic


Data Exploration
filtering and Sampling
Modeling