Project: Creditworthiness

Step 1: Business and Data Understanding

- **What decisions needs to be made?**
  We need to identify the customers those are creditworthy for getting a loan.

- **What data is needed to inform those decisions?**
  To inform those decisions, data such as account balance, list of customers and credit amount is required.

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

Step 2: Building the Training Set

There are no variables which are highly correlated with each other i.e. correlation of higher than 0.7. This is concluded using association analysis

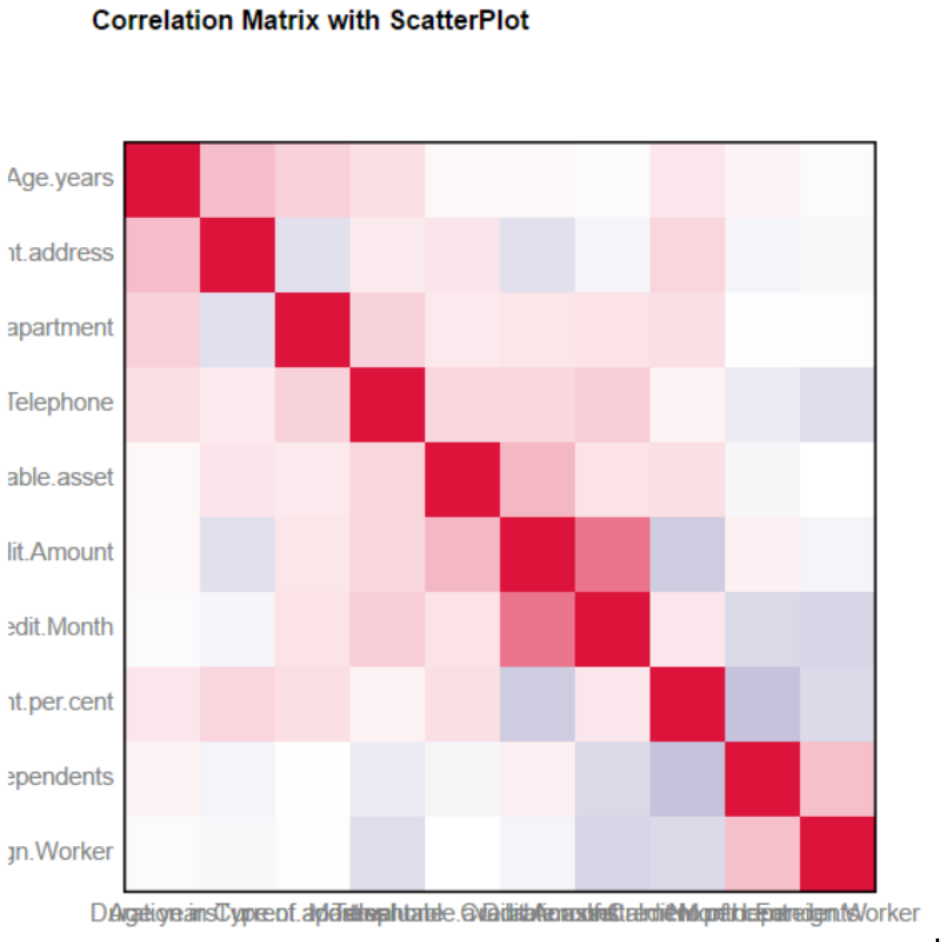

Fig: Correlation matrix of variables

Summarizing data fields gives us results as follows:

1. Duration in Current Address has 69% missing data, therefore it needs to be removed.
2. Concurrent Credits and Occupation has one value. So it needs to be removed.
3. Guarantors, Foreign Workers and No of Dependents show low variability as more than 80% of data is skewed towards one data, therefore it needs to be removed.
4. 2% data is missing in Age Years, it is appropriate to impute missing data with median age.
5. Median age is used instead of mean as data is skewed to left as shown below.
6. Telephone field is also removed due to its irrelevancy to customer creditworthiness.

Fig: Field Summary of all variables

Step 3: Train your Classification Models

a) **Logistic Regression (Stepwise)**

Account Balance, Purpose, Credit Amount are top 3 most significant variables with p-value of less than 0.05. Here, Credit Application Result is used as target variable.

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

Fig: Summary report for Stepwise Model

Accuracy for Creditworthy at 80.0% is higher than Non-Creditworthy at 62.9% and overall accuracy is 76.0%.

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

Fig: Model comparison report for Stepwise logistic model

b) Decision Tree

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

Fig: Decision Tree, Variable Importance and Confusion Matrix

Credit Application Result is used as the target variable.

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

Fig: Model Comparison Report for Decision Tree

c) Forest Model

Credit Application Result is used as the target variable.

Credit Amount, Age Years and Duration of Credit Month are the 3 most important variables.

Overall accuracy is 80.0%.

This model is not biased as the accuracies for creditworthy and non-creditworthy are 79.1% and 85.7% respectively, which are comparable.

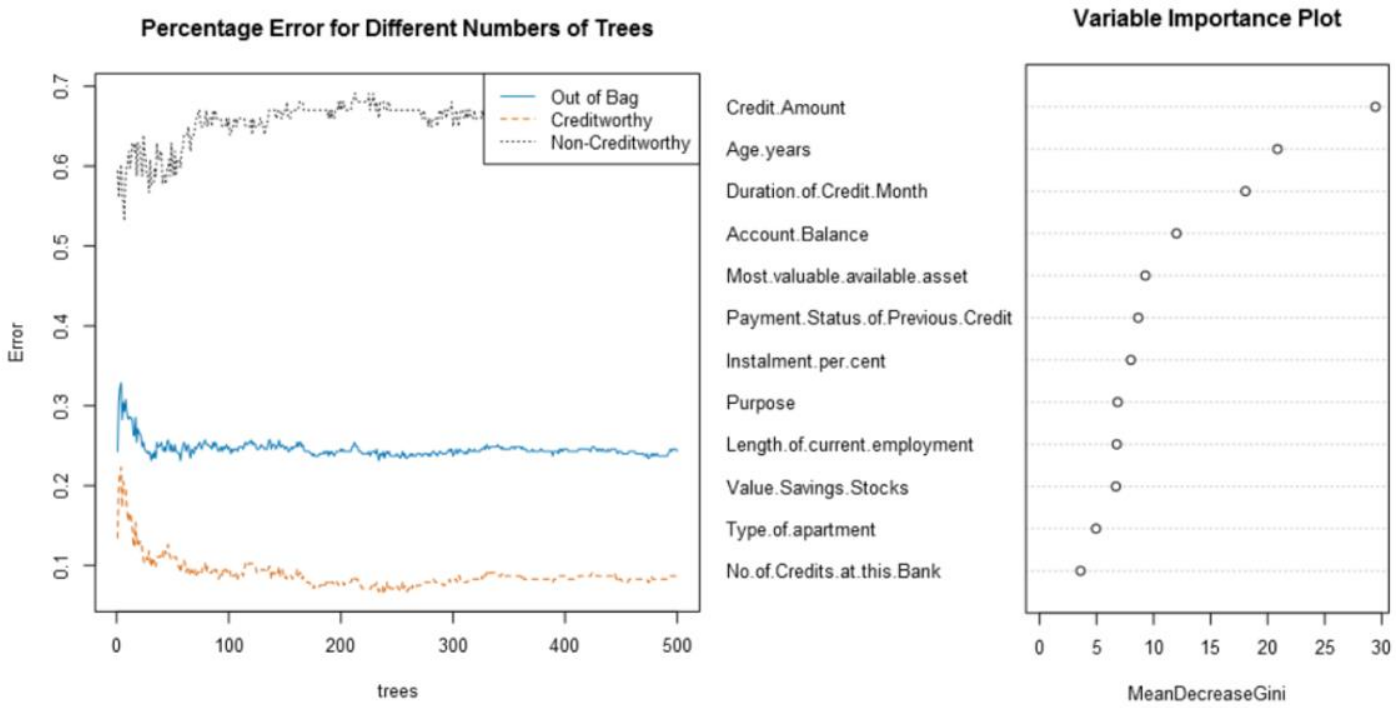

Fig: Percentage Error for Different Number of Trees and Variable Importance Plot

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

Fig: Model Comparison Report for Forest Model

d) Boosted Model

The most significant variables are Account Balance and Credit Amount.
Overall accuracy is 76.7%.
Accuracies for creditworthy and non-creditworthy are 76.7% and 78.3%.
It 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

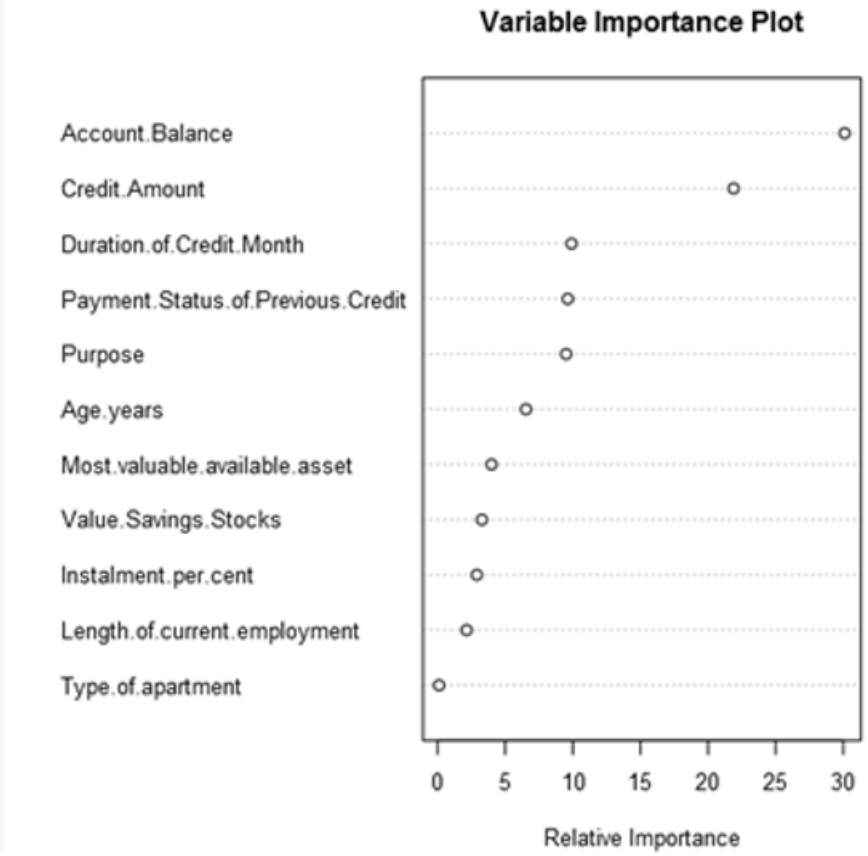

Fig: Variable Importance Plot for Boosted Model

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

Fig: Model Comparison Report for Boosted Model

Step 4: Writeup

We will choose Forest model as it offers the highest accuracy at 80% against validation set. It has highest accuracies for creditworthy and non-creditworthy.
There are **408 creditworthy customers** 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

Fig: Model Comparison Report for all 4 classification models

Forest model reaches the true positive rate at the fastest 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.

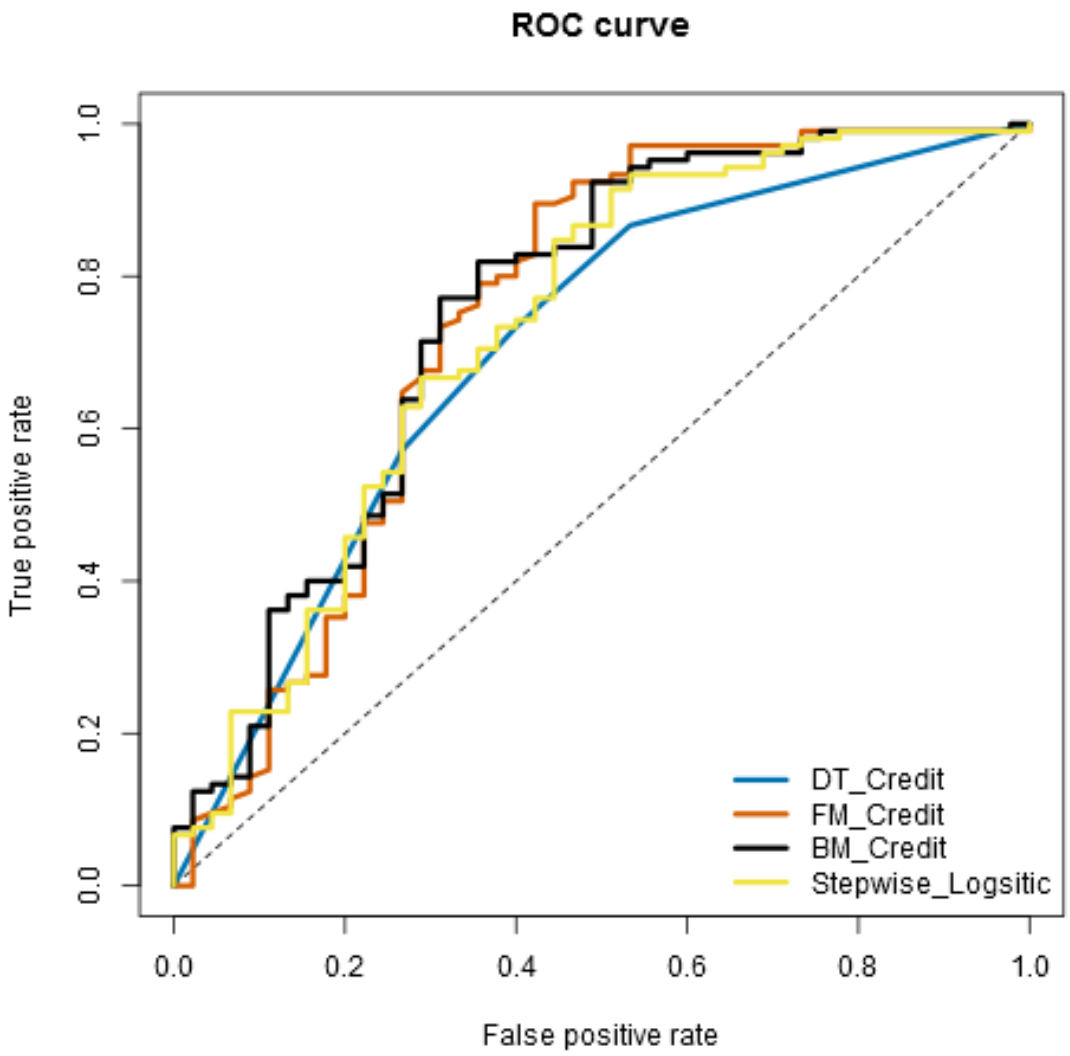

Fig: ROC curve for all 4 classification models

Alteryx Flow