Project: Predict Default Risk: Creditworthiness

Step 1: Business and Data Understanding

- **What decisions need to be made?**
  The objective is to identify whether the new customers who applied for loan are creditworthy to be extended one.

- **What data is needed to inform those decisions?**
  Historical data on past applications such as Account Balance and Credit Amount and list of customers are required in order to inform those decisions. This dataset is given to us.

- **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 Logistic Regression, Decision Tree, Forest Model, and Boosted Tree will be used to analyze and determine creditworthy of the new customers

Step 2: Building the Training Set

Here are some guidelines to help guide your data cleanup:

- *For numerical data fields, are there any fields that highly-correlate with each other? The correlation should be at least .70 to be considered "high".*
- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed*
- *Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called "low variability" and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.*
- *Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)*

After performing association analysis on the numerical variables, I found that there are no numerical variables which are highly correlated with each other, i.e. a correlation of higher than 0.7.

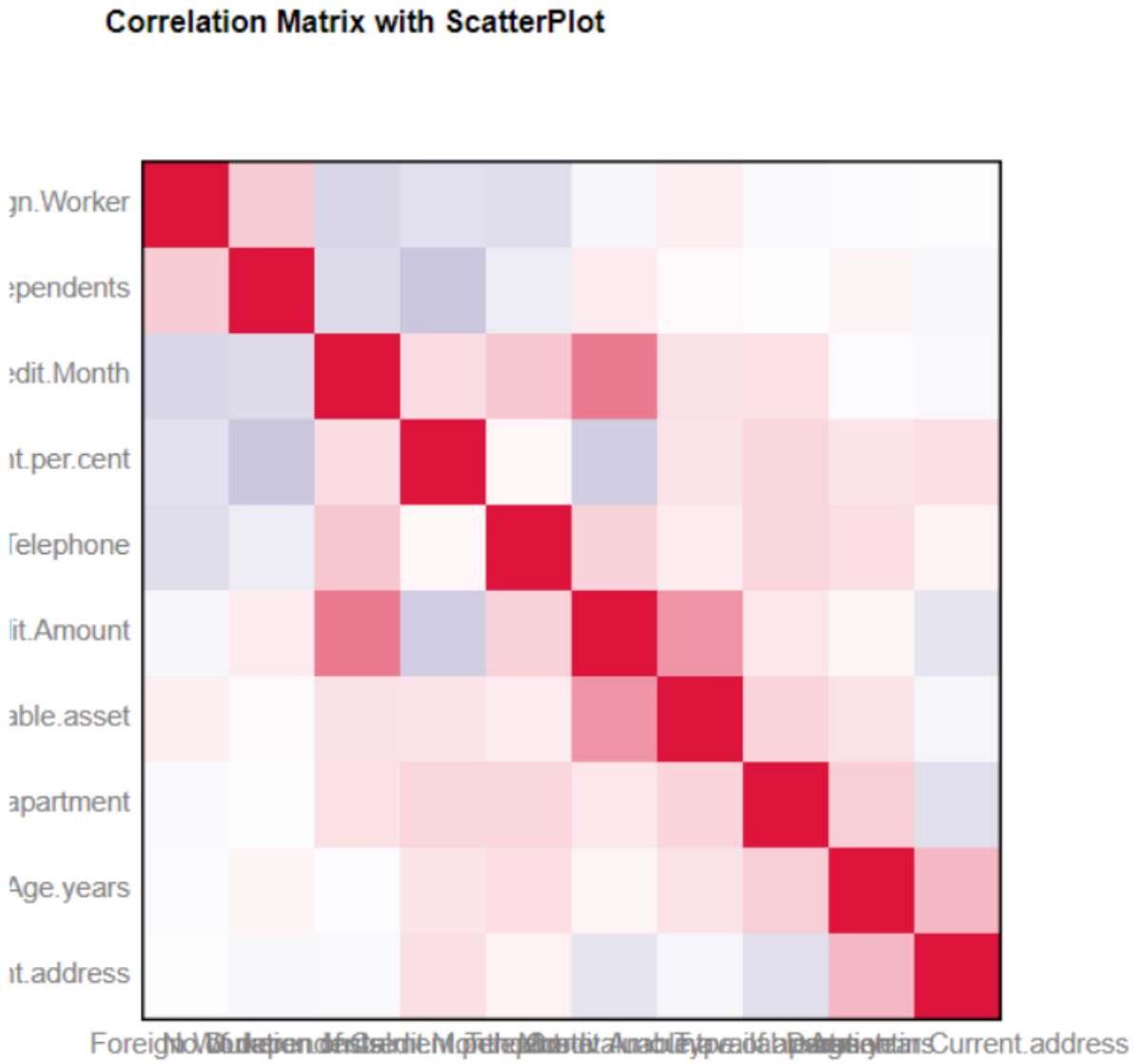


Fig: Correlation Matrix with ScatterPlot of numerical variables

Upon analyzing the data by summarizing all data fields, *Duration in Current Address* has 69% missing data. This field should be removed as this will skew our results. *Age Years* has 2% missing data, so it is appropriate to impute the missing data. The "median" age is used instead of the "mean" as the data is skewed to the left as shown below.

Also, *Concurrent Credits* and *Occupation* has one type of value while *Guarantors*, *Foreign Worker* and *No of Dependents* show low variability, meaning, more than 80% of the data skewed towards one data. These data should be removed in so that our analysis is not skewed.

Telephone field should also be removed as logically it is irrelevant to the customer being creditworthy or not.

Fig: Field summary of all variables.

Step 3: Train your Classification Models

First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

A. Logistic Regression (Stepwise)

Choosing *Credit Application Result* as the target variable, I found that *Account Balance*, *Purpose* and *Credit Amount* are the top three significant variables with a p-value of less than 0.05.

Report for Logistic Regression Model LR_Stepwise_PDR

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, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Fig: Summary Report for Stepwise Logistic Regression Model

The overall accuracy is around 76.0% for the Stepwise Logistic Regression model. Individual accuracy for creditworthy is higher than non-creditworthy at 80.0% and 62.9% respectively. The model is slightly biased towards predicting a given data point as non-creditworthy.

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Stepwise_PDR	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of LR_Stepwise_PDR

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

Fig: Model Comparison Report for Stepwise Logistic Regression Model

B. Decision Tree

Choosing *Credit Application Result* as the target variable, the following, *Account Balance*, *Value Savings Stocks,* and *Duration of Credit Month* are the top three significant variables. The overall accuracy of this model is 74.7%.

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

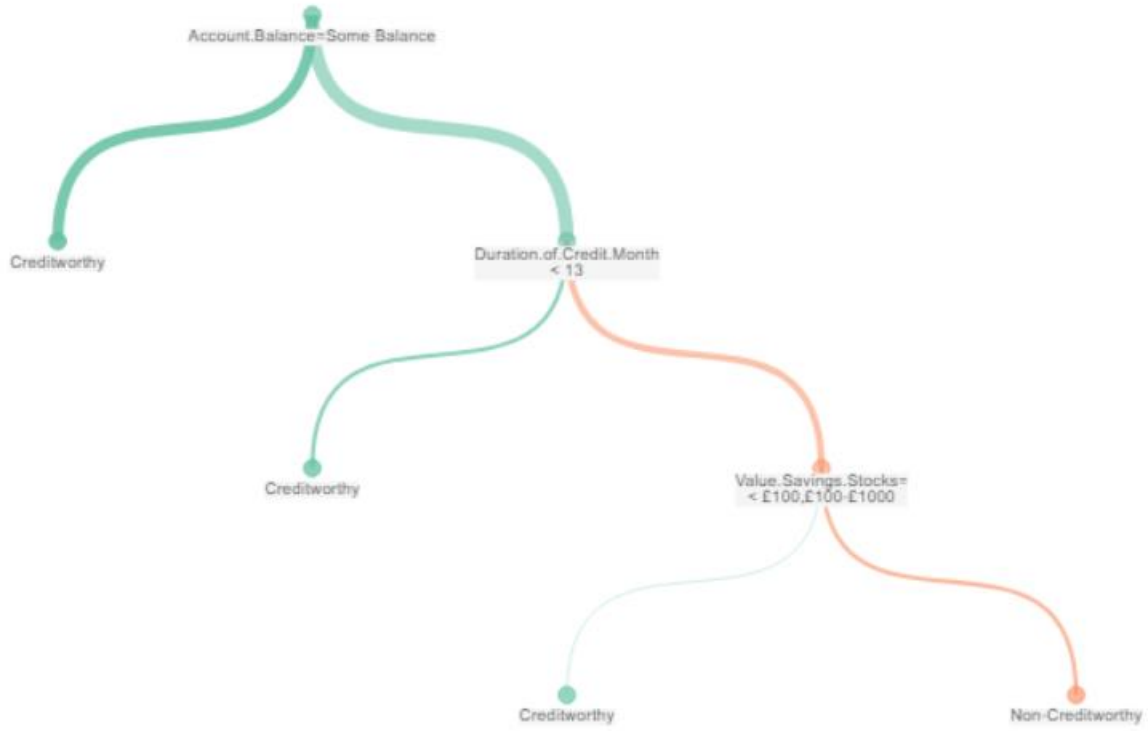


Fig: Decision Tree

Model Comparison Report

Fit and error measures

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

Confusion matrix of DT_PDR

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

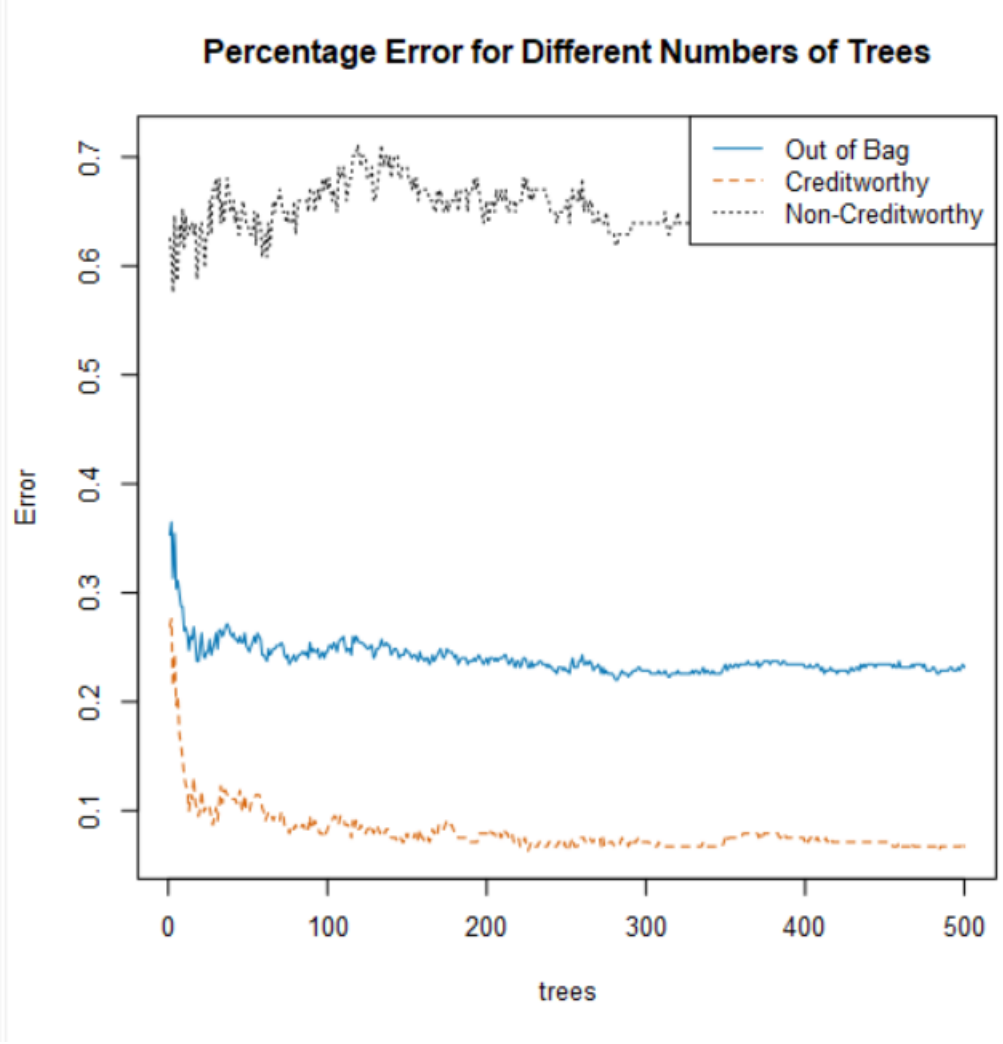
Fig: Model Comparison Report Confusion Matrix for Decision Tree

C. Forest Model

Choosing *Credit Application Result* as the target variable, *Credit Amount*, *Age Years* and *Duration of Credit Month* are the top three significant variables.

Overall accuracy is 80.0%. The model seems to be even as the accuracies for creditworthy and non-creditworthy are 79.1% and 85.7% respectively. There seems to be no bias in predicting either of Creditworthy or Non-Creditworthy.

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3



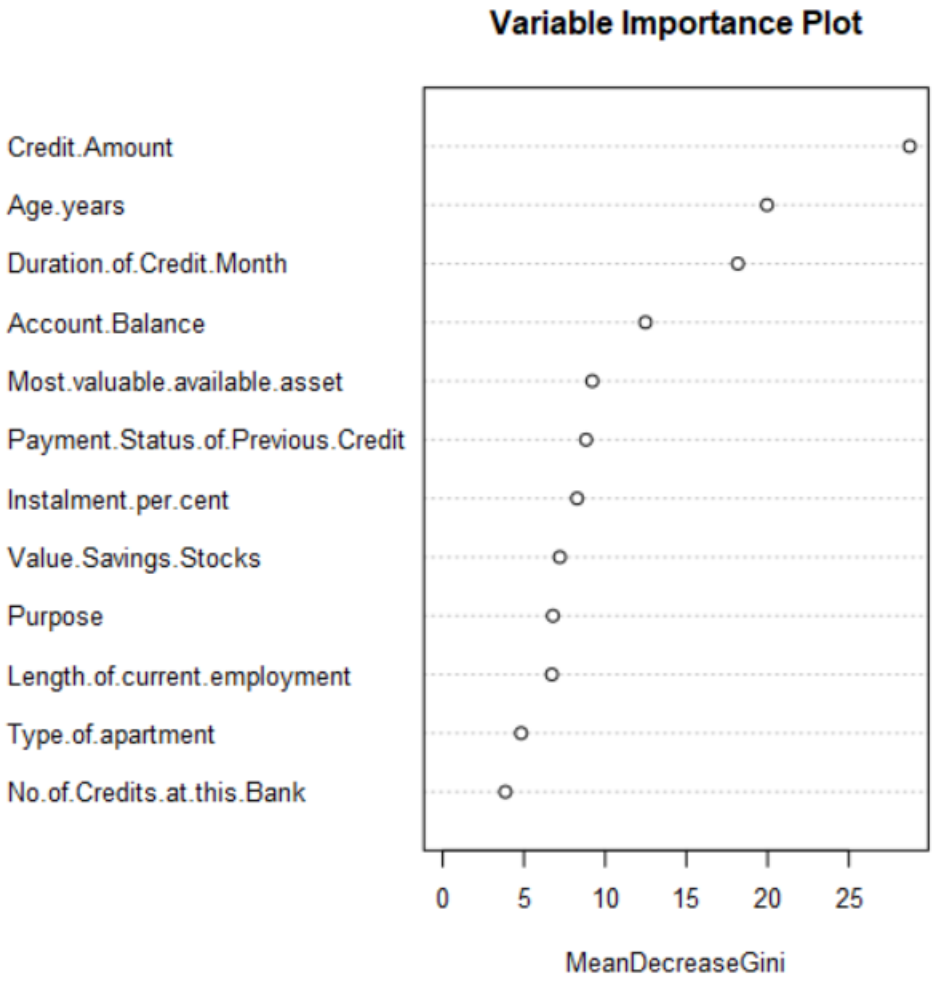


Fig: Percentage Error for Different Number of Trees and Variable Importance Plot

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FT_PDR	0.7933	0.8681	0.7368	0.9714	0.3778

Confusion matrix of FT_PDR

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Fig: Model Comparison Report for Forest Model

D. Boosted Model

From the *Variable Importance Plot, Account Balance* and *Credit Amount* are the most significant variables. The overall model accuracy is 76.7%. The accuracy for creditworthy and non-creditworthy are 76.7% and 78.3% respectively. This seems to be unbiased in predicting the credit-worthiness of customers.

Loss function distribution: Bernoulli
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 2036

Plots:

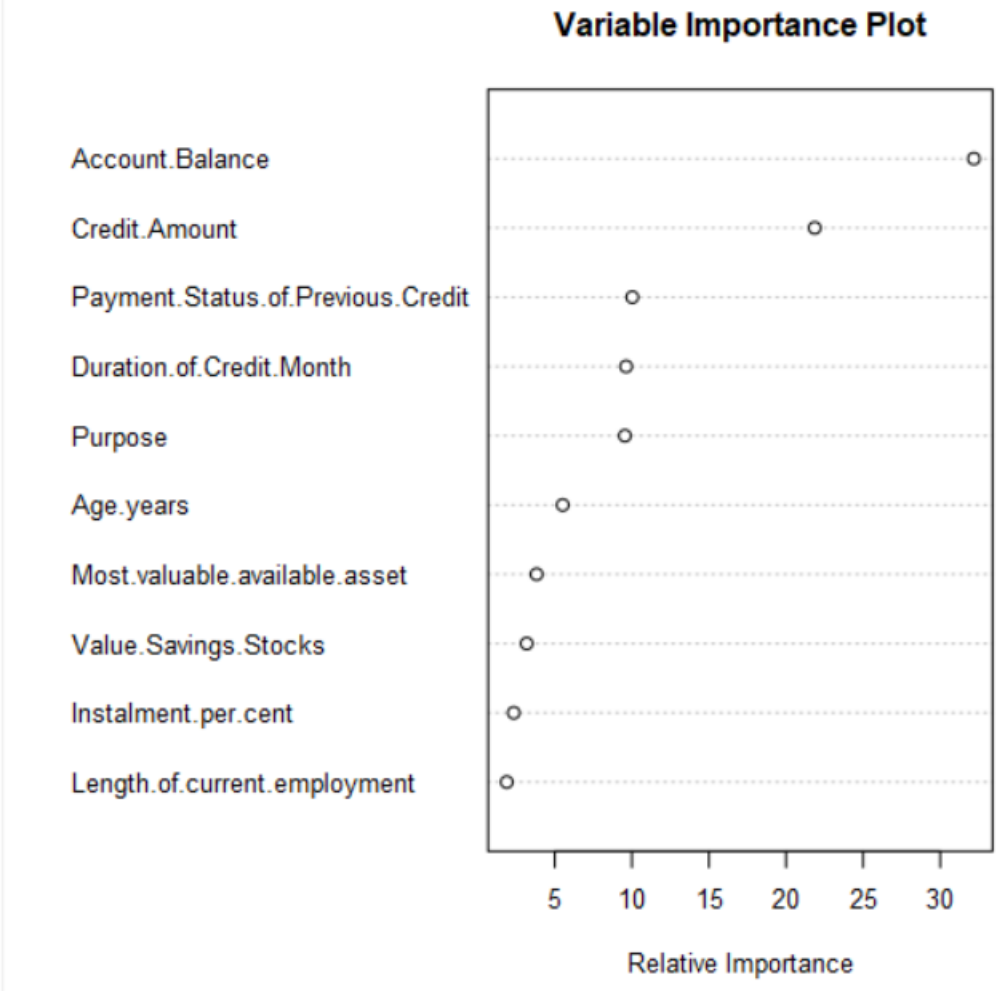


Fig: Variable Importance Plot for Boosted Model

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
BM_PDR	0.7867	0.8632	0.7524	0.9619	0.3778

Confusion matrix of BM_PDR

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

Fig: Model Comparison Report for Boosted Model

Step 4: Writeup

Comparing all the 4 models used so far

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_PDR	0.7467	0.8273	0.7054	0.8667	0.4667
FT_PDR	0.7933	0.8681	0.7368	0.9714	0.3778
BM_PDR	0.7867	0.8632	0.7524	0.9619	0.3778
LR_Stepwise_PDR	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of BM_PDR

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

Confusion matrix of DT_PDR

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

Confusion matrix of FT_PDR

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of LR_Stepwise_PDR

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

Fig: Model Comparison Report for all 4 classification models

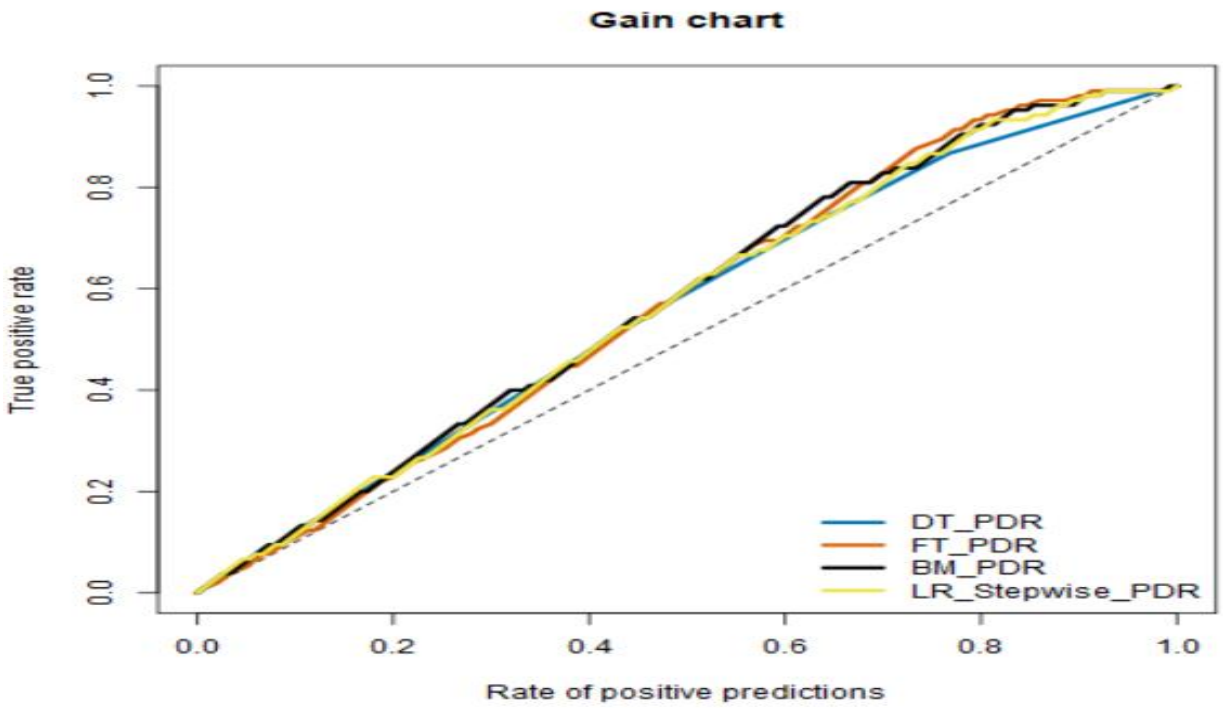


Fig: Gain chart for all 4 classification models

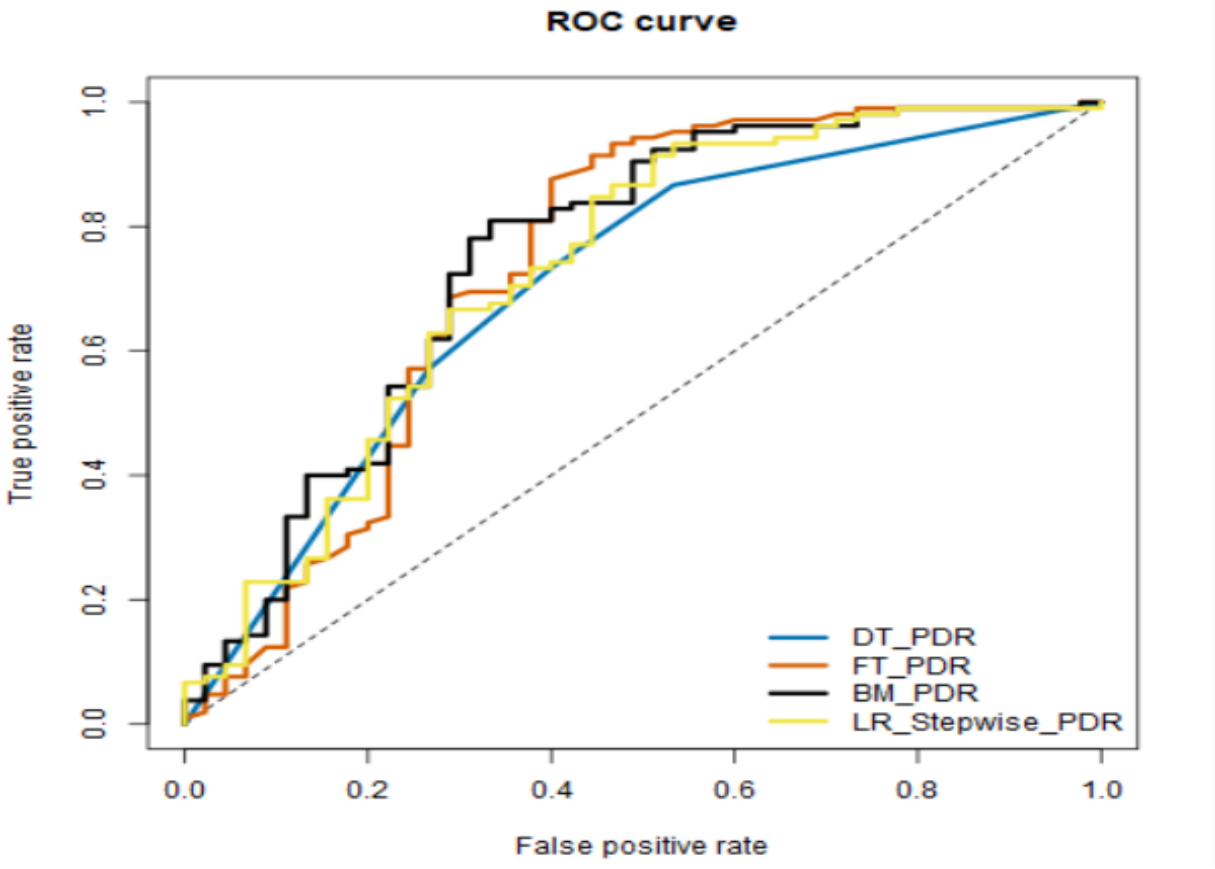


Fig: ROC chart for all 4 classification models

The Forest model is chosen as it shows the highest accuracy of about 80% against validation data. Additionally, the Forest model's accuracy in predicting the creditworthy and non-creditworthy is among high compared to the other three models.

The GAIN Chart and the ROC chart shows Forest model reaches the true positive rate the fastest. The models also have less bias when predicting between creditworthy and non-creditworthy categories. This low bias implies that the model is good enough to predict and avoid lending money to customers with high probability of defaulting while ensuring at the same time that the right customer is also approved of the loan.

Record #	Sum_X_Creditworthy	Sum_X_Non-Creditworthy
1	408	92

After scoring the remaining list of the new loan applicants, The forest model predicts that there are **408** creditworthy cutomers**.**

My Alteryx Workflow: