Project: Creditworthiness

Step 1: Business and Data Understanding

Key Decisions:

- What decisions needs to be made?

  Due to a financial scandal that hit a competitive bank last week, we suddenly have an influx of new people applying for loans for the bank. And we need to evaluate the creditworthiness of these new loan applicants.

- What data is needed to inform those decisions?

  The data needed for building models will come from ***credit-data-training.xlsx***, and the columns are :
  Credit-Application-Result
  Account-Balance
  Duration-of-Credit-Month
  Payment-Status-of-Previous-Credit
  Purpose
  Credit-Amount
  Value-Savings-Stocks
  Length-of-current-employment
  Most-valuable-available-asset
  No-of-Credits-at-this-Bank
  Type-of-Apartment
  Instalment-per-cent
  Age-years

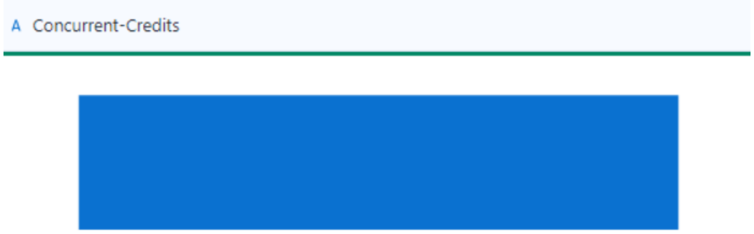
  The model will be used to evaluate the creditworthiness of the applicants from ***customers-to-score.xlsx***.

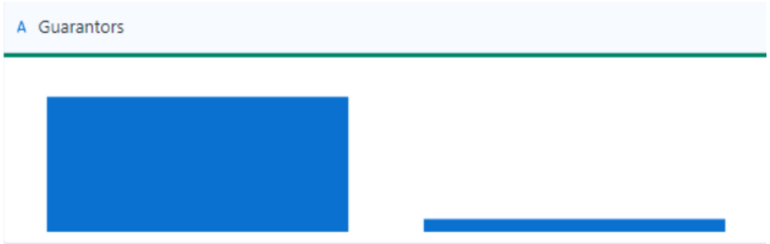
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

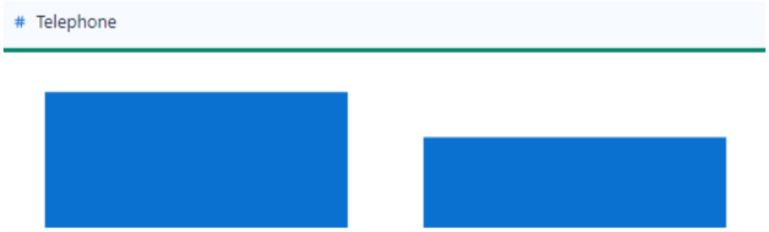
  Based on the methodology map, the model we need will be a binary classification model.

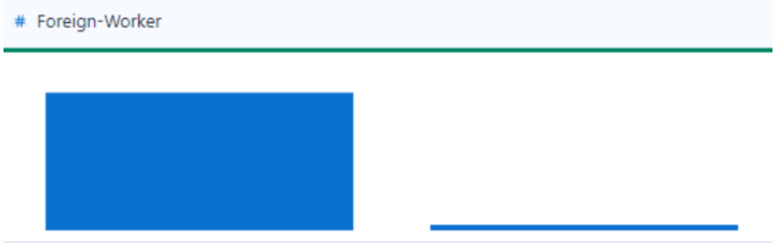
Step 2: Building the Training Set

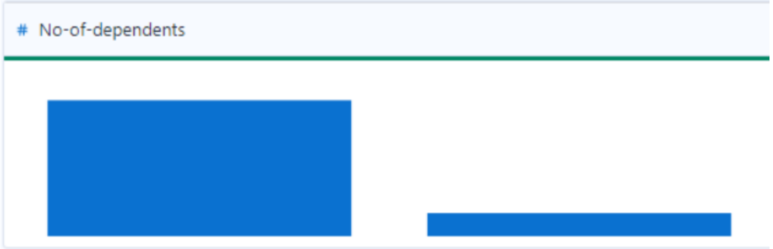
- In your cleanup process, which fields did you remove or impute?
  Fields with low-variability are removed:

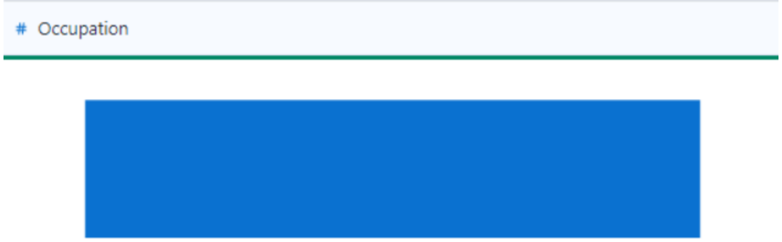




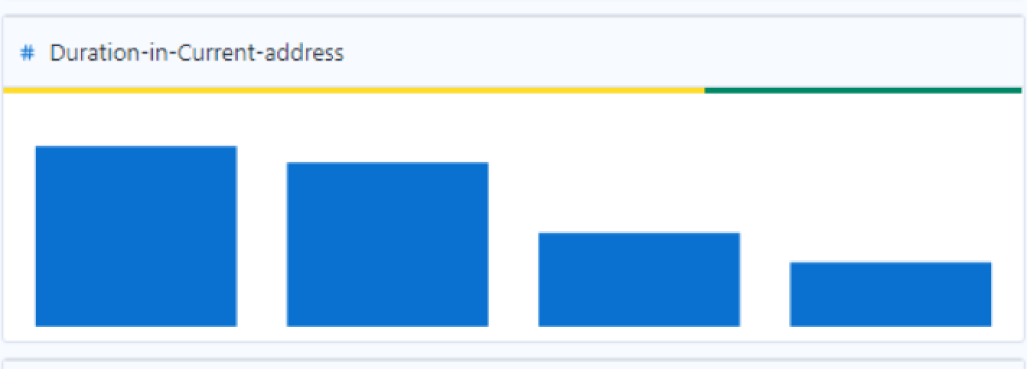




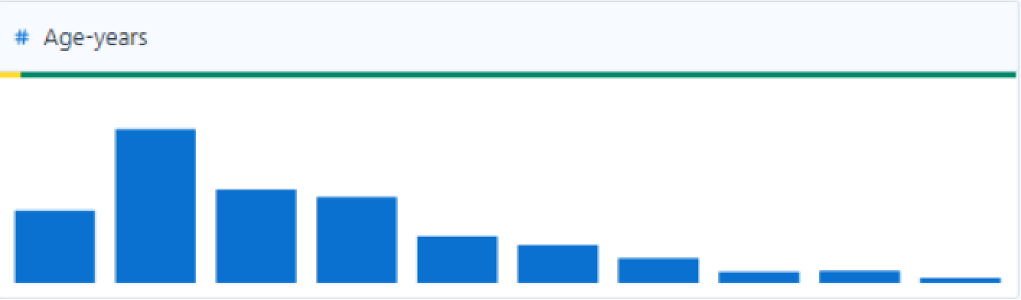




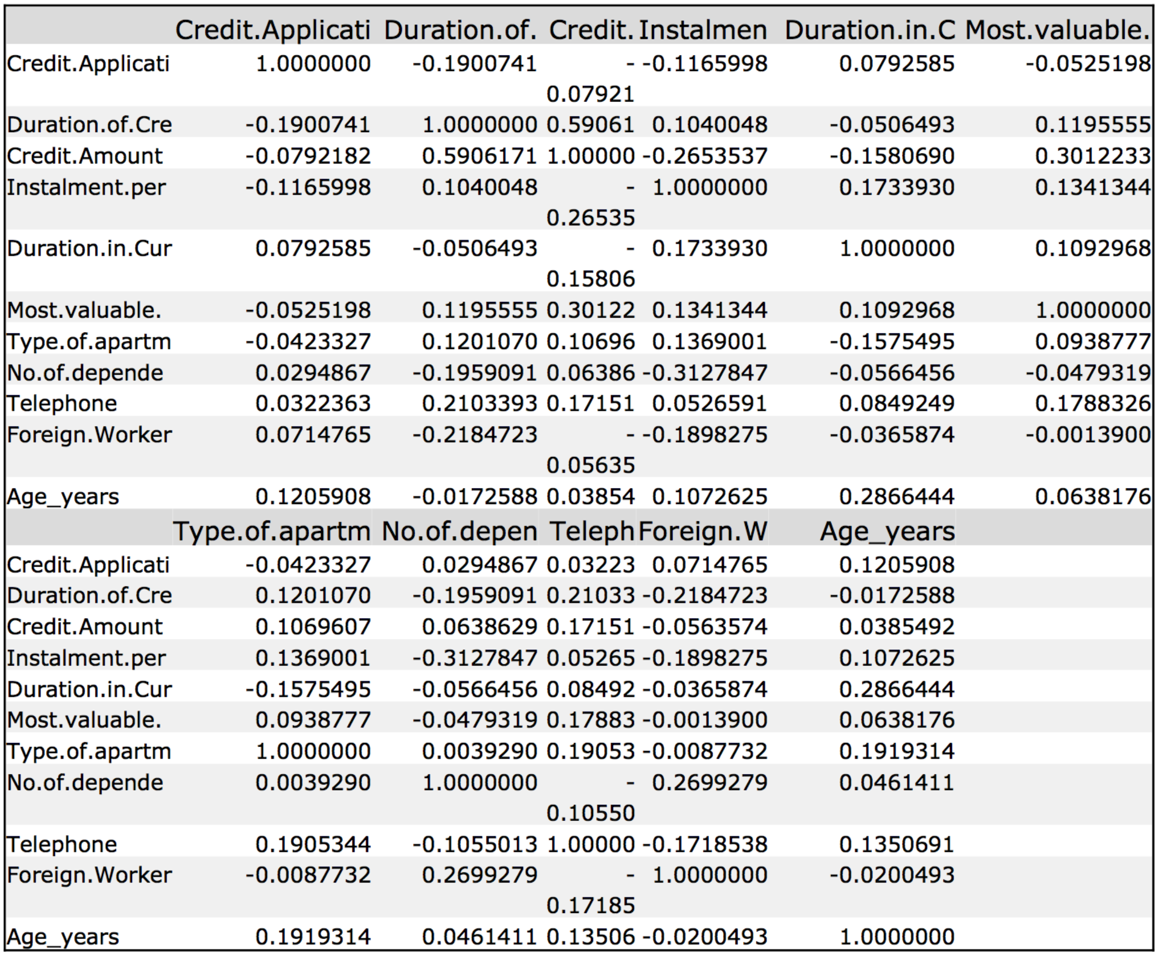
Duration-in-Current-address has a lot of missing data is removed.



Age-years has some missing data, I will substitute missing data here with the average age (35.637) of the dataset.



Full Correlation Matrix



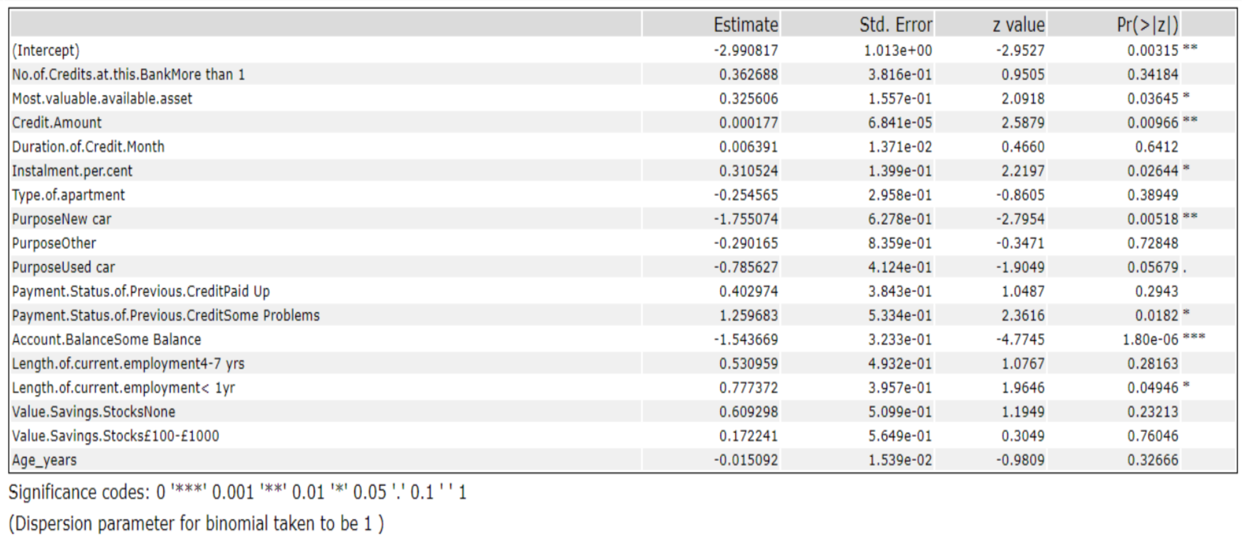
	Credit.Applicati	Duration.of.	Credit.	Instalmen	Duration.in.C	Most.valuable.	Type.of.apartm	No.of.depen	Teleph	Foreign.W	Age_years
Credit.Applicati	1.0000000	-0.1900741	-0.07921	-0.1165998	0.0792585	-0.0525198	-0.0423327	0.0294867	0.03223	0.0714765	0.1205908
Duration.of.Cre	-0.1900741	1.0000000	0.59061	0.1040048	-0.0506493	0.1195555	0.1201070	-0.1959091	0.21033	-0.2184723	-0.0172588
Credit.Amount	-0.0792182	0.5906171	1.00000	-0.2653537	-0.1580690	0.3012233	0.1069607	0.0638629	0.17151	-0.0563574	0.0385492
Instalment.per	-0.1165998	0.1040048	-0.26535	1.0000000	0.1733930	0.1341344	0.1369001	-0.3127847	0.05265	-0.1898275	0.1072625
Duration.in.Cur	0.0792585	-0.0506493	-0.15806	0.1733930	1.0000000	0.1092968	-0.1575495	-0.0566456	0.08492	-0.0365874	0.2866444
Most.valuable.	-0.0525198	0.1195555	0.30122	0.1341344	0.1092968	1.0000000	0.0938777	-0.0479319	0.17883	-0.0013900	0.0638176
Type.of.apartm	-0.0423327	0.1201070	0.10696	0.1369001	-0.1575495	0.0938777	1.0000000	0.0039290	0.19053	-0.0087732	0.1919314
No.of.depende	0.0294867	-0.1959091	0.06386	-0.3127847	-0.0566456	-0.0479319	0.0039290	1.0000000	-0.10550	0.2699279	0.0461411
Telephone	0.0322363	0.2103393	0.17151	0.0526591	0.0849249	0.1788326	0.1905344	-0.1055013	1.00000	-0.1718538	0.1350691
Foreign.Worker	0.0714765	-0.2184723	-0.05635	-0.1898275	-0.0365874	-0.0013900	-0.0087732	0.2699279	-0.17185	1.0000000	-0.0200493
Age_years	0.1205908	-0.0172588	0.03854	0.1072625	0.2866444	0.0638176	0.1919314	0.0461411	0.13506	-0.0200493	1.0000000

Using 0.7 as the benchmark for high correlation, there seems to be nothing of high correlation with the numerical data fields.

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.

For the Logistic Model, the most significant predictor variable is *Account Balance.*

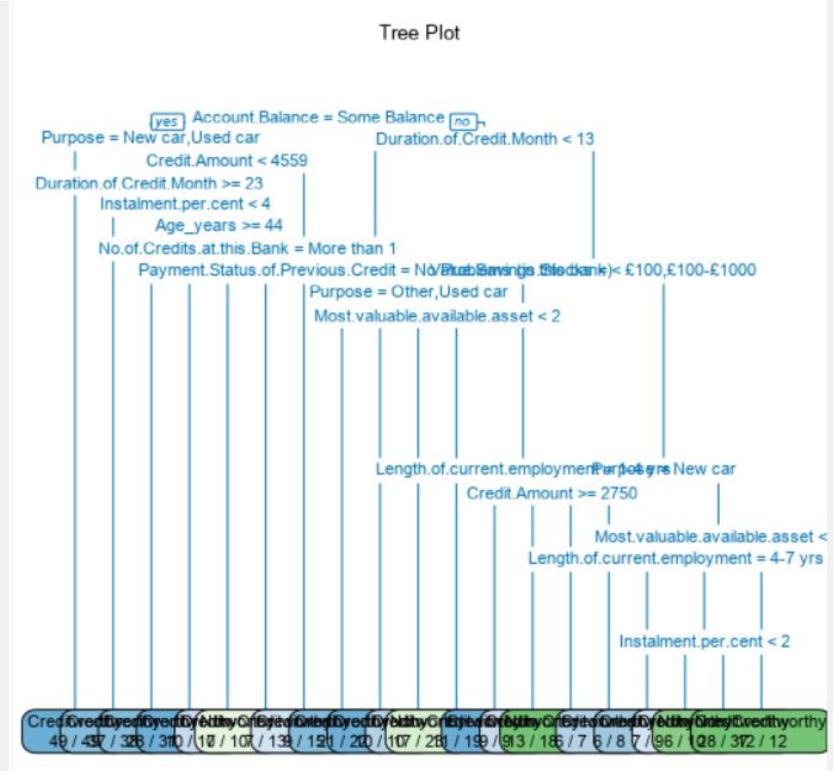


	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315 **
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645 *
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966 **
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644 *
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949
PurposeNew car	-1.755074	6.278e-01	-2.7954	0.00518 **
PurposeOther	-0.290165	8.359e-01	-0.3471	0.72848
PurposeUsed car	-0.785627	4.124e-01	-1.9049	0.05679 .
Payment.Status.of.Previous.CreditPaid Up	0.402974	3.843e-01	1.0487	0.2943
Payment.Status.of.Previous.CreditSome Problems	1.259683	5.334e-01	2.3616	0.0182 *
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06 ***
Length.of.current.employment4-7 yrs	0.530959	4.932e-01	1.0767	0.28163
Length.of.current.employment< 1yr	0.777372	3.957e-01	1.9646	0.04946 *
Value.Savings.StocksNone	0.609298	5.099e-01	1.1949	0.23213
Value.Savings.Stocks£100-£1000	0.172241	5.649e-01	0.3049	0.76046
Age_years	-0.015092	1.539e-02	-0.9809	0.32666

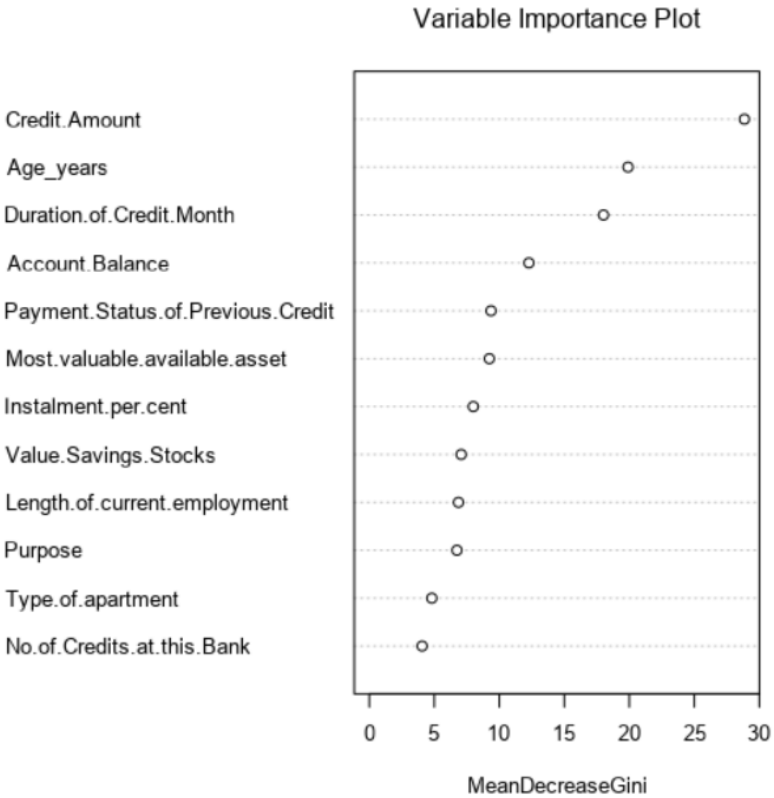
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

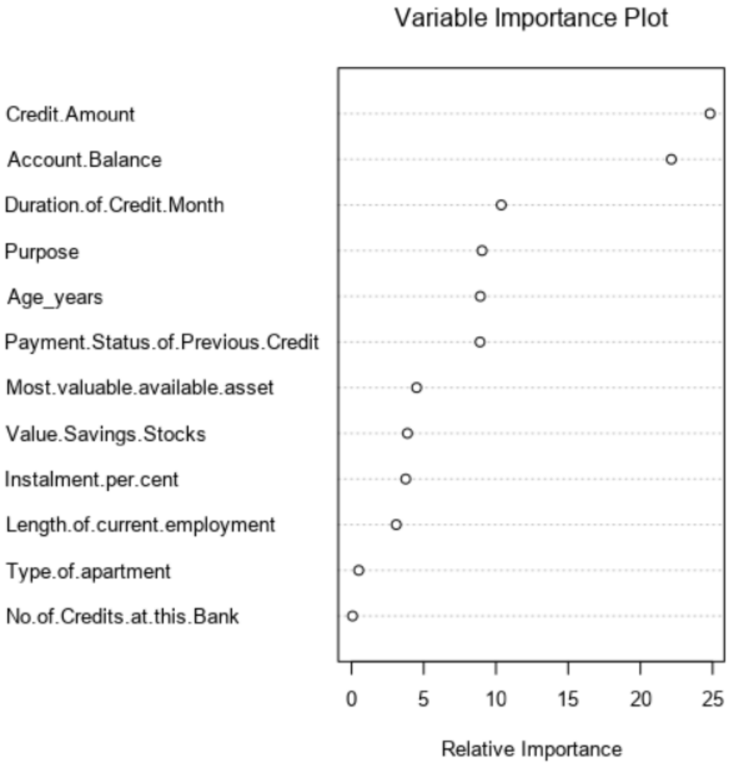
For the decision tree model, *Account Balance* is the most significant predictor variable.


Tree Plot
yes
Account.Balance = Some Balance
no
Purpose = New car,Used car
Duration.of.Credit.Month < 13
Credit.Amount < 4559
Duration.of.Credit.Month >= 23
Instalment.per.cent < 4
Age_years >= 44
No.of.Credits.at.this.Bank = More than 1
Payment.Status.of.Previous.Credit = No
£100,£100-£1000
Purpose = Other,Used car
Most.valuable.available.asset < 2
Length.of.current.employment
New car
Credit.Amount >= 2750
Most.valuable.available.asset <
Length.of.current.employment = 4-7 yrs
Instalment.per.cent < 2

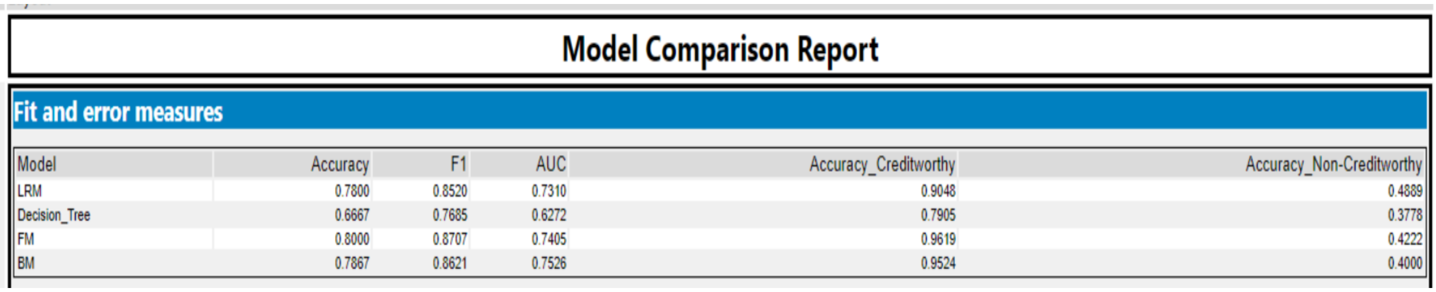
For the forest model, *Credit.Amount* is the most significant predictor variable.



For the boosted model, *Credit.Amount* is the most significant predictor variable.



- 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?



Model Comparison Report

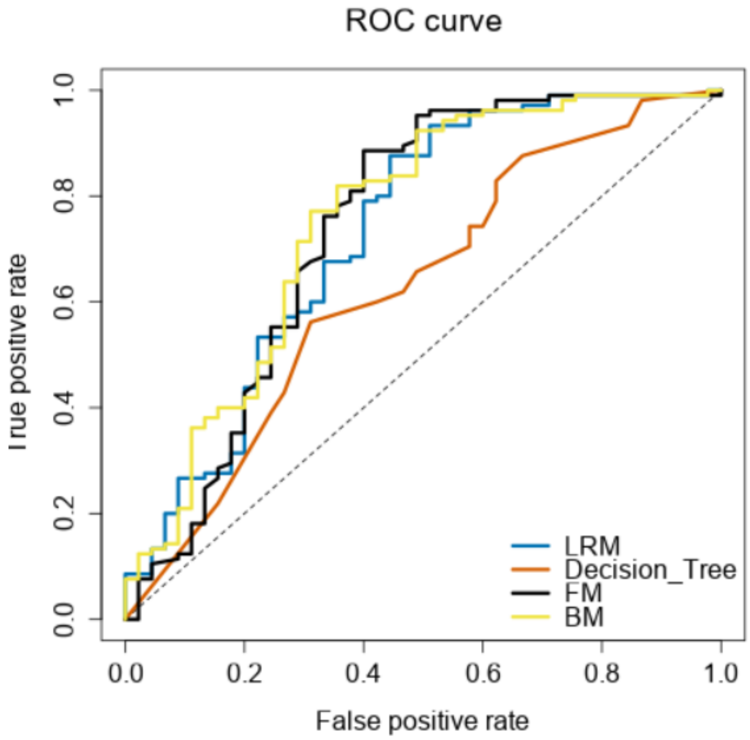
Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LRM	0.7800	0.8520	0.7310	0.9048	0.4889
Decision_Tree	0.6667	0.7685	0.6272	0.7905	0.3778
FM	0.8000	0.8707	0.7405	0.9619	0.4222
BM	0.7867	0.8621	0.7526	0.9524	0.4000

The model with the highest accuracy score is the forest model with 0.800.

The models appear to predict Creditworthy more accurately than Non-Creditworthy. It also looks like there are more applicants that are creditworthy and not.

Below is the ROC chart for the models:



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:

  The final model used for prediction will be the Random Forest model due to its highest overall accuracy of 0.800. The accuracy for predicting Creditworthy is 0.9619 and 0.4222 for Non-Creditworthy.
  The ROC plot shows the forest model is slightly better than the rest of the models.

- How many individuals are creditworthy?

  Based on the forest model, 408 individuals are creditworthy.