Project: Creditworthiness

Step 1: Business and Data Understanding

The hypothetical bank I work for, has had a high inflow of loan applications in recent weeks due to a financial scandal from a competitor bank, and I am responsible for loan approvals at my bank. In previous times, we were approving our loan applications by hand but given the number of applications coming in now, we need to automate the process of approving them, our approval also has to be based on a system that ensures that the applicants we approve are creditworthy and would not default on the loan.

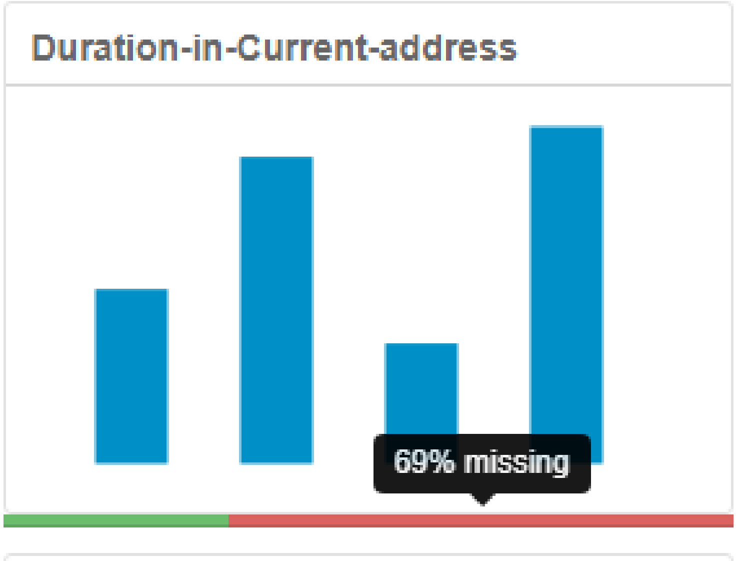
Key Decisions:

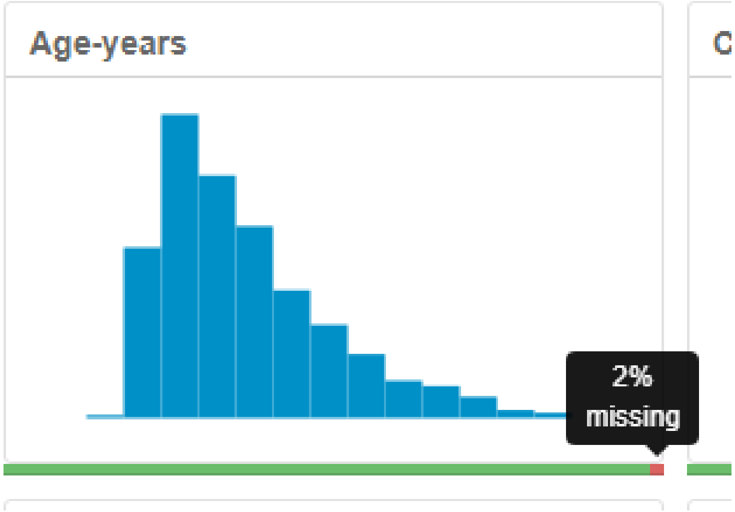
1. What decisions needs to be made?
    - We need to decide if a loan applicant is creditworthy or not.
2. What data is needed to inform those decisions?
    - Data on past loan applications we have had at the bank. This data can contain fields such as the income of the applicant, their employment status, level of education, monthly bank balance, credit card deficit, etc. and finally whether they paid back the loan or not. We also need data from the new loan applications we get. The data on past applications will then be used to train a classification model to make predictions on the new applications.
3. 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 classification model such as the logistic regression and decision tree models.

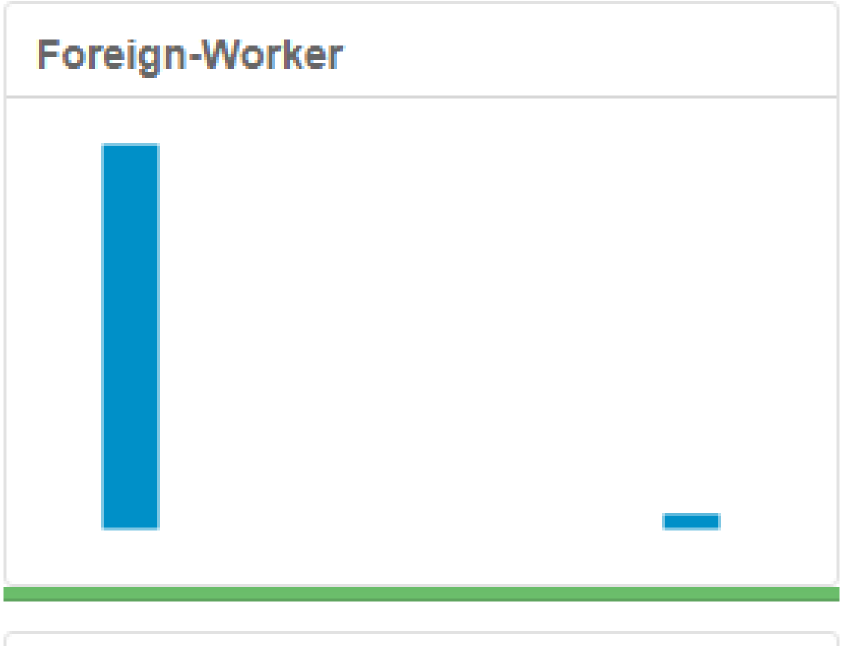
Step 2: Building the Training Set

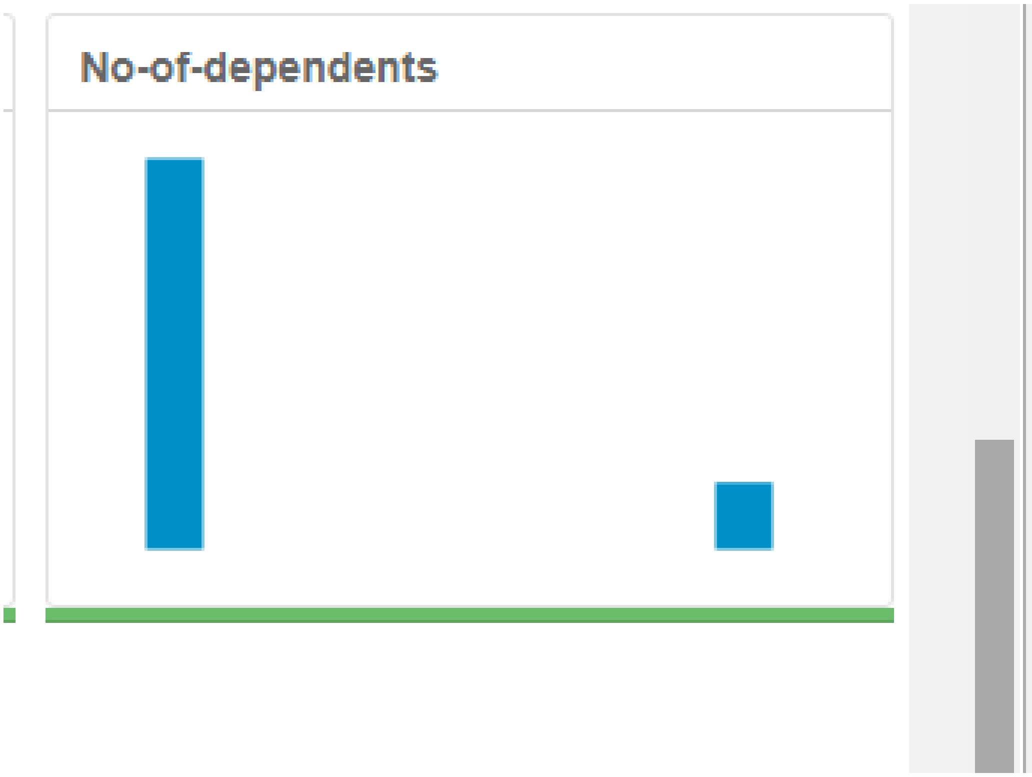
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

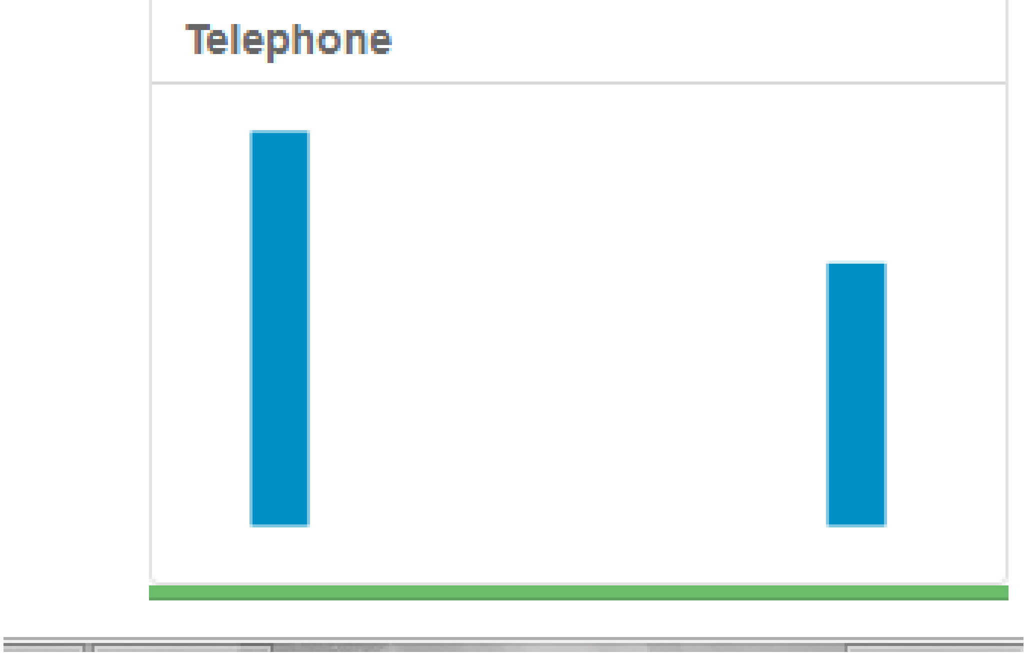
  Ans: I dropped the "duration in current address field" because it had almost 70% of its values missing. I also removed the 'Guarantors', 'concurrent credit, occupation, number of dependents, telephone and foreign worker fields, these have very small numbers of unique values of two and below. I imputed the Age field with the median age of 33, I chose to impute because the age field has only 2% of missing values and I chose to impute with the median instead of any other measure of central tendency e.g. the mean, because the frequency distribution of my data for the age field is skewed to the left hence the median is a better measure of central tendency in this case.

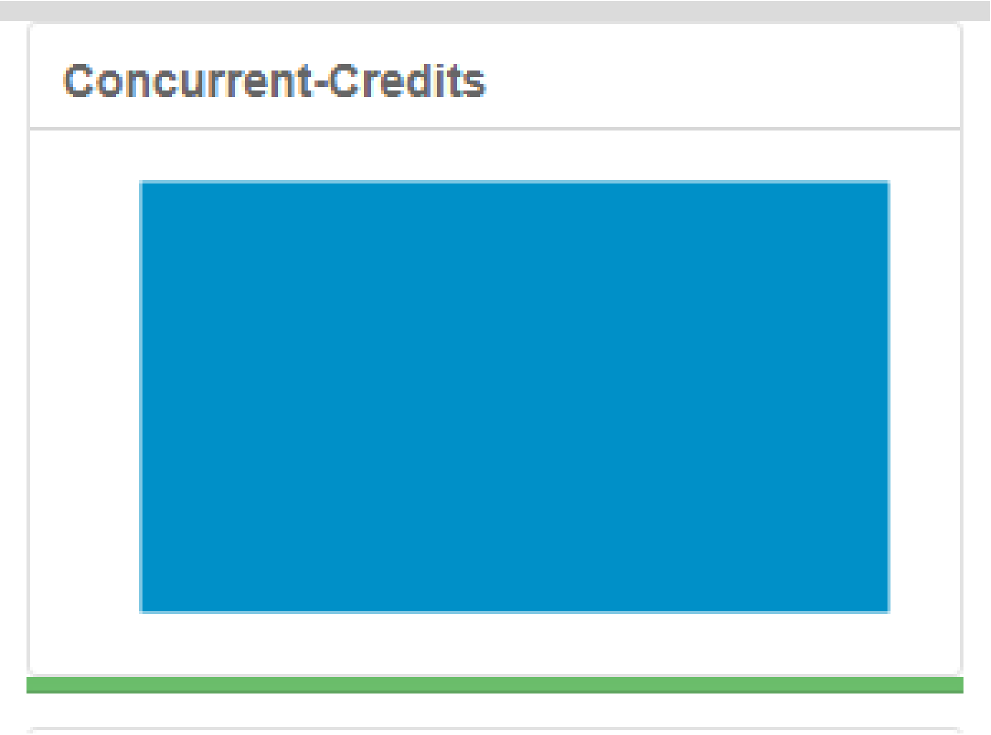


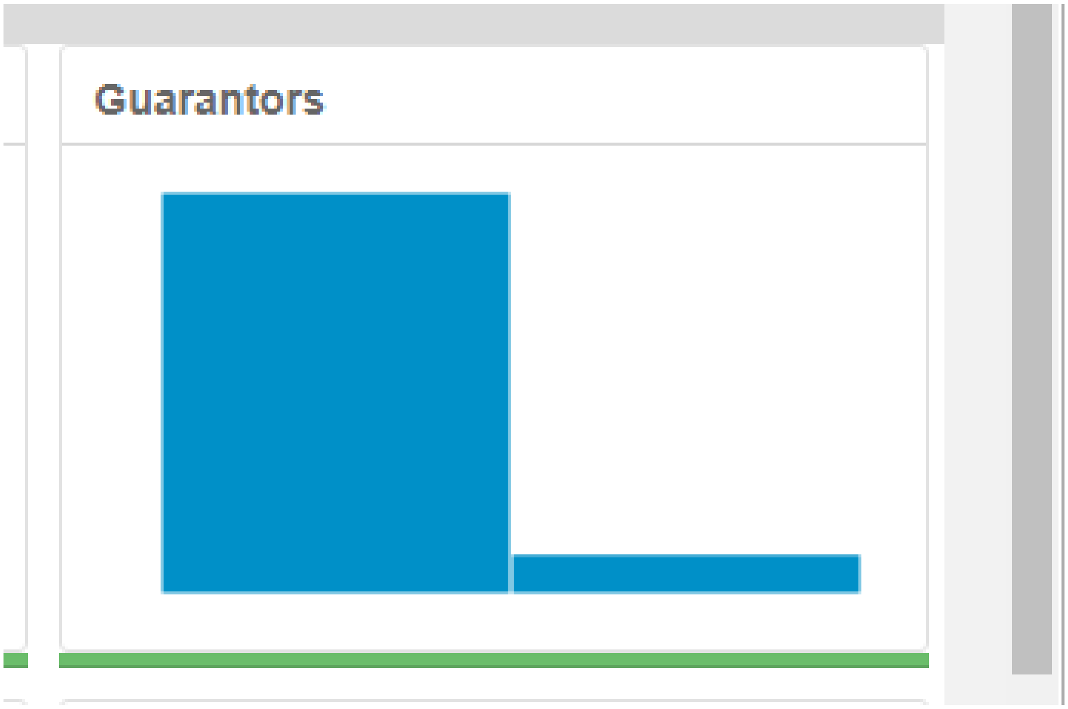

Age-years
C
2%
missing

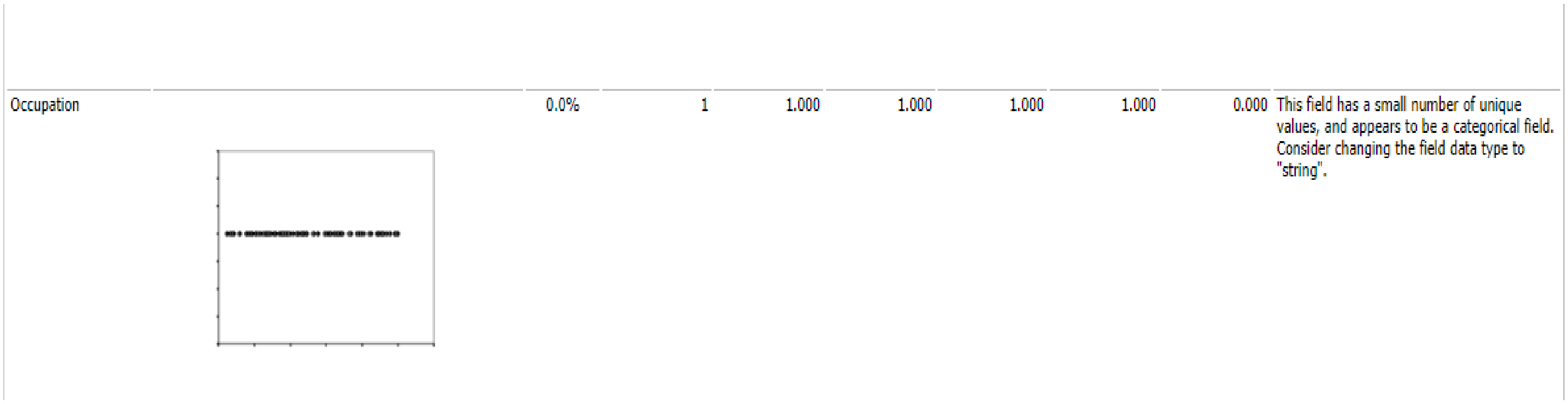

Foreign-Worker


No-of-dependents


Telephone


Concurrent-Credits


Guarantors



Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

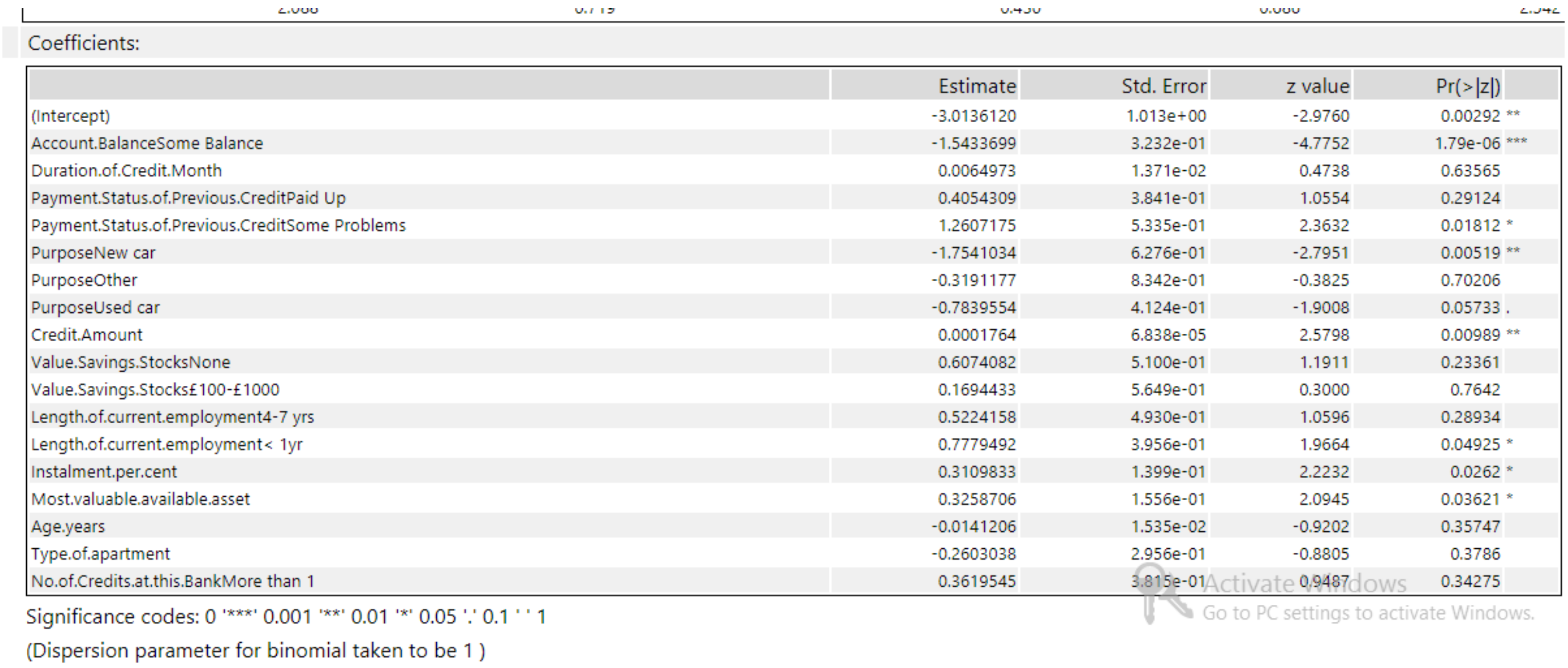
Step 3: Train your Classification Models

Logistic Regression Model

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

  ANS: Account Balance (some balance), Purpose (new car), credit amount, were the top 3 variables ordered by importance in my Logistic Regression model. Using the stepwise tool alongside the Logistic Regression Model, the top variables were still Account Balance (some balance), Purpose (new car) and credit amount

Logistic Regression Result without Stepwise tool



2.088 0.719 0.430 0.080 2.542

Coefficients:

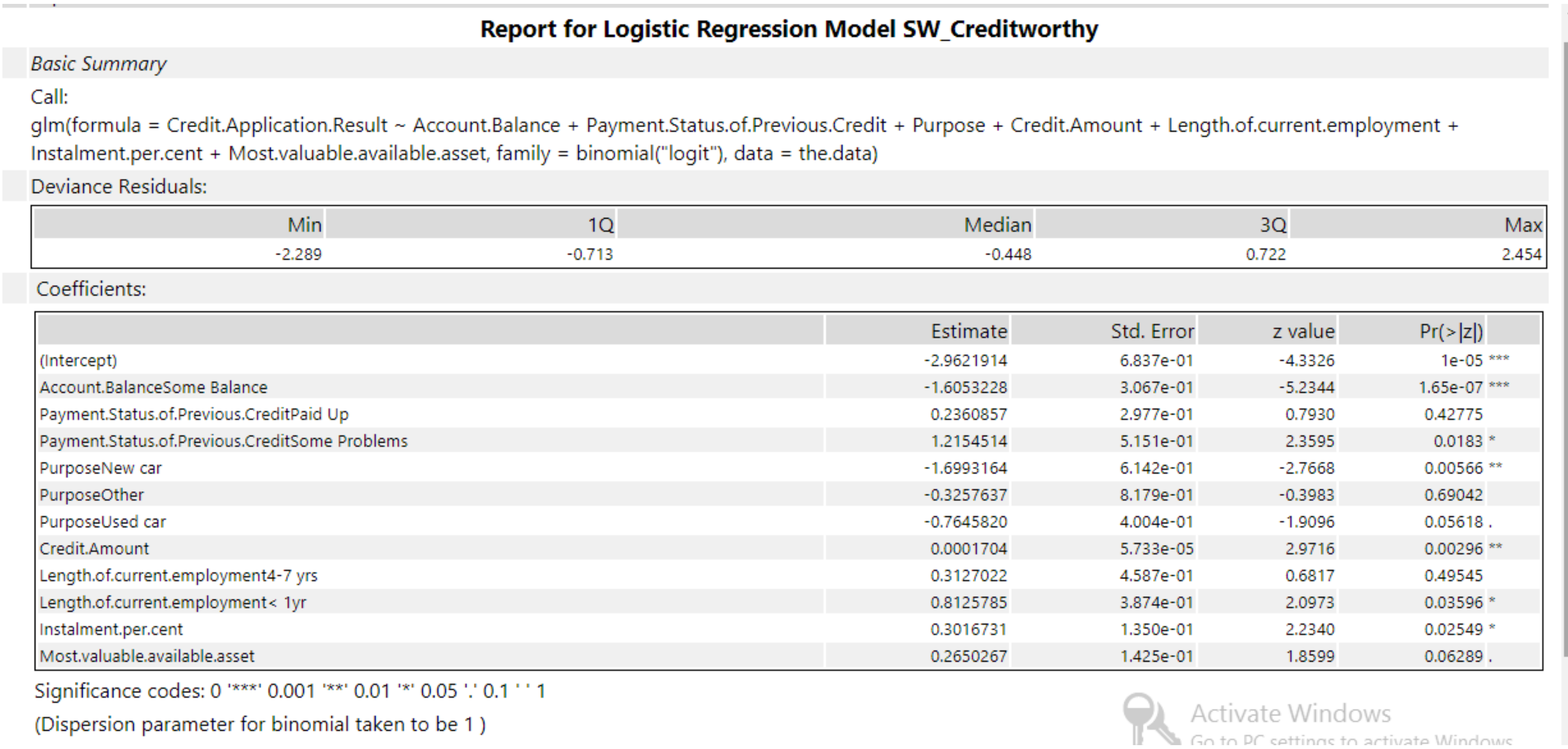
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.0136120	1.013e+00	-2.9760	0.00292	**
Account.BalanceSome Balance	-1.5433699	3.232e-01	-4.7752	1.79e-06	***
Duration.of.Credit.Month	0.0064973	1.371e-02	0.4738	0.63565	
Payment.Status.of.Previous.CreditPaid Up	0.4054309	3.841e-01	1.0554	0.29124	
Payment.Status.of.Previous.CreditSome 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	

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

(Dispersion parameter for binomial taken to be 1)

Activate Windows
Go to PC settings to activate Windows.

Logistic Regression Result with Stepwise tool



Report for Logistic Regression Model SW_Creditworthy

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	.

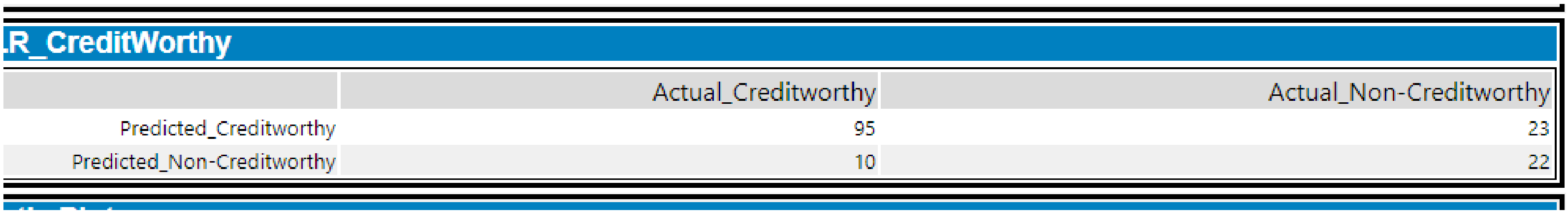
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

Ans: The overall accuracy of my logistic regression model was 78% without the stepwise feature. After using the stepwise tool, the accuracy dropped to 76%. The bias in the confusion matrix below is minimal. The Positive Predictive Values(PPV) = True Positives/ (True Positives + False Positives) = 95/ (95 + 23) = 0.80 while the Negative Predicted Values (NPV) = True Negatives/ (True Negatives + False Negatives) = 22/(22 + 10) = 0.68. With the PPV at 0.8 and the NPV at 0.68, there is a slight bias in the logistic model towards predicting the Creditworthy segment.

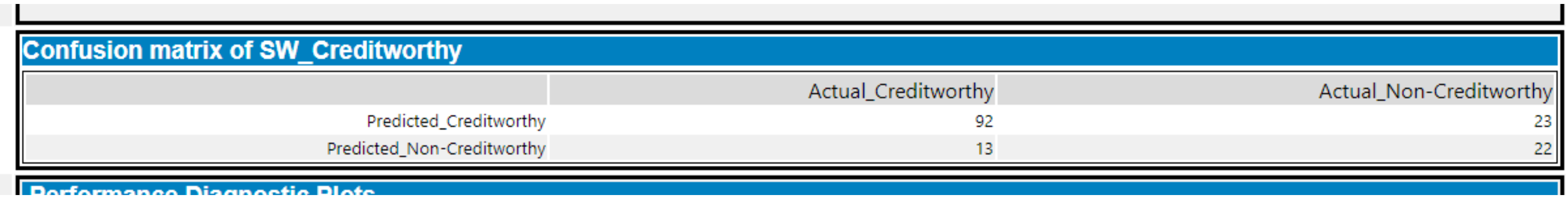
Confusion Matrix without Stepwise tool



R_CreditWorthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

Confusion Matrix with Stepwise



Confusion matrix of SW_Creditworthy

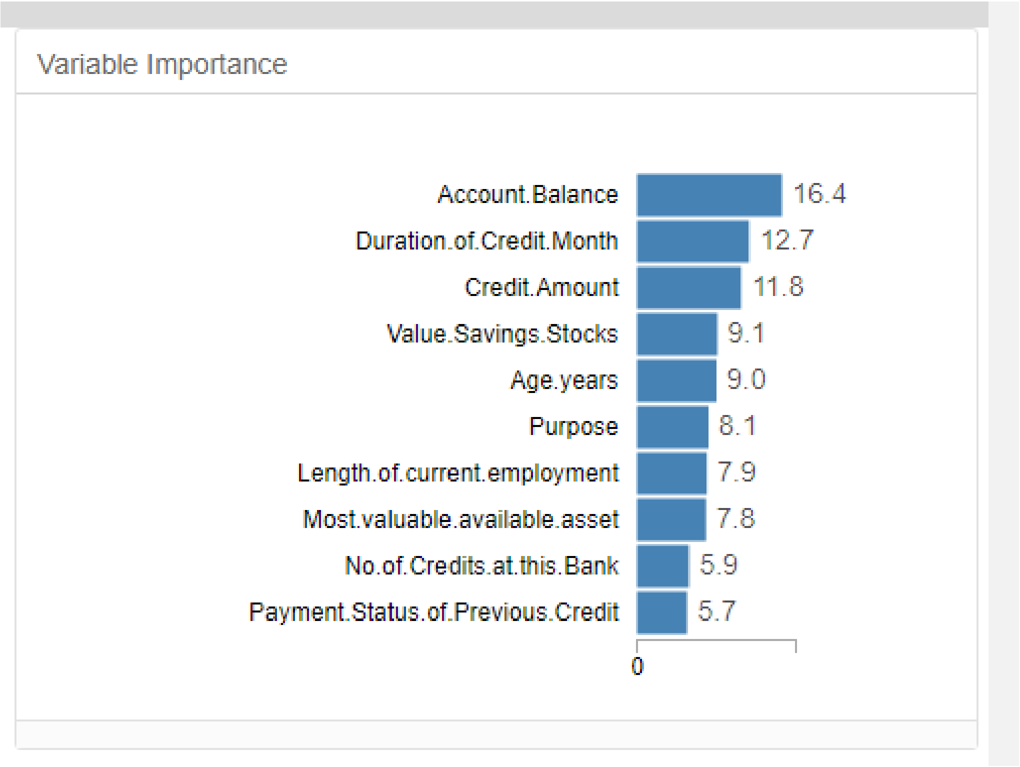
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Performance Diagnostic Plots

Decision Tree Model

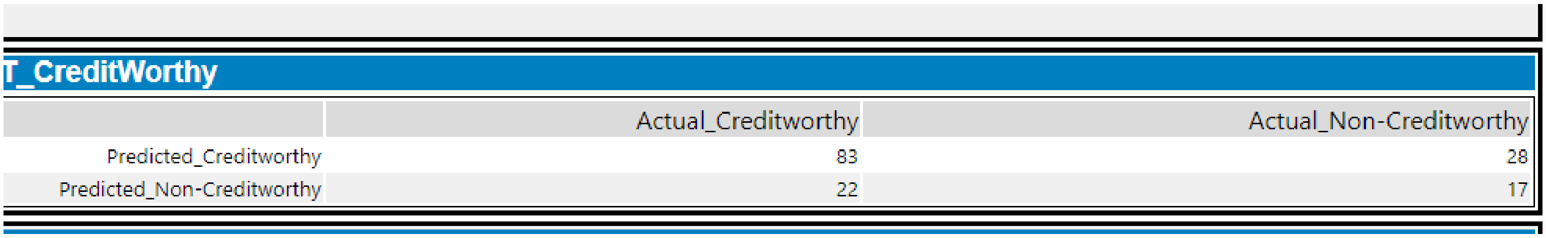
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

  ANS: Account balance, duration of credit month, credit amount, value saving stocks and age were the top 5 variables ordered by importance in my decision tree model.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

Ans: The overall accuracy of my decision tree model was 75%, from the confusion matrix. Its PPV is 0.74 and its NPV is 0.44, the model has a large bias towards the creditworthy segment since it predicts the creditworthy class far better than the non-credit worthy class.



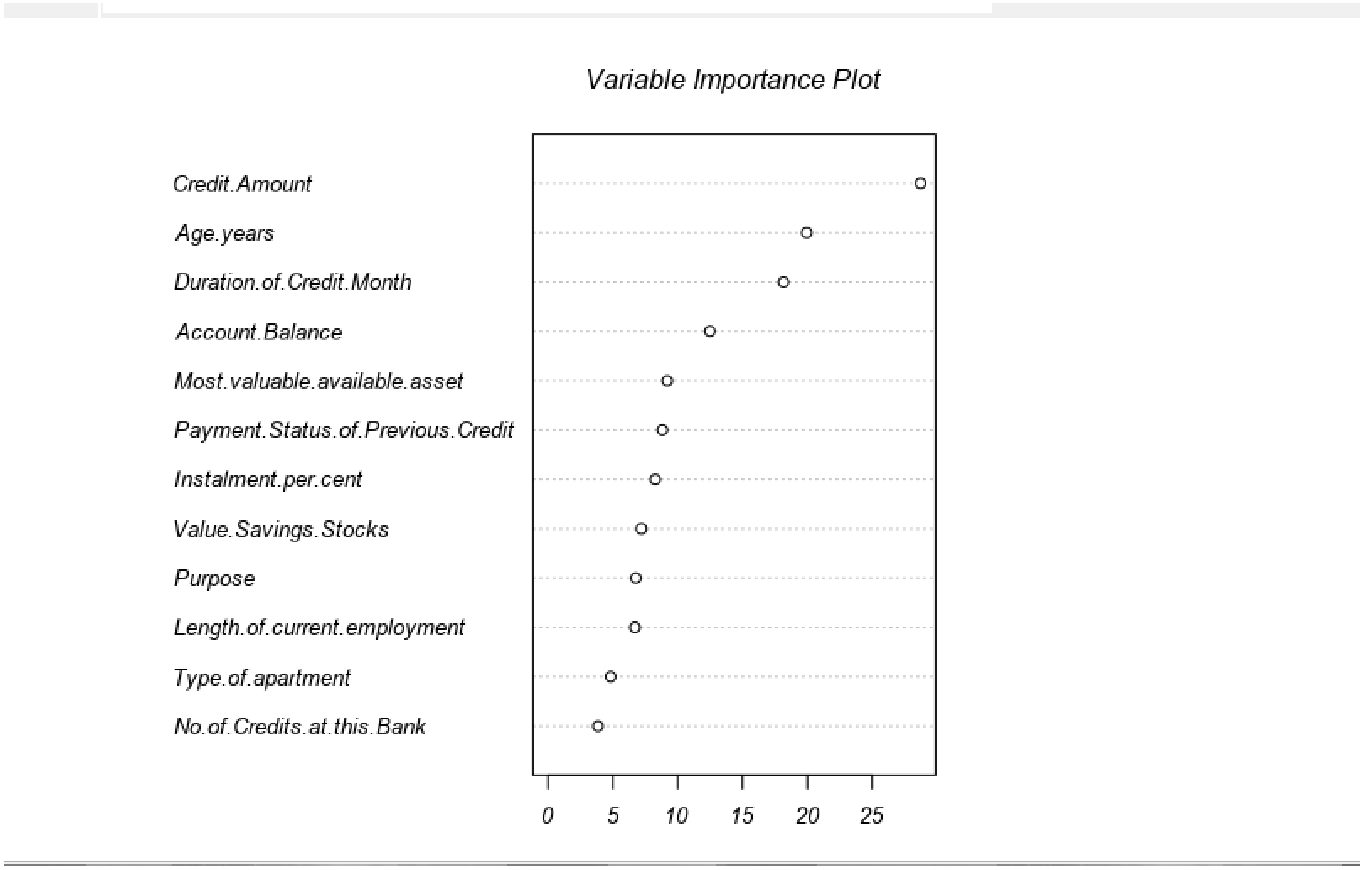
T_CreditWorthy

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

Forest Model

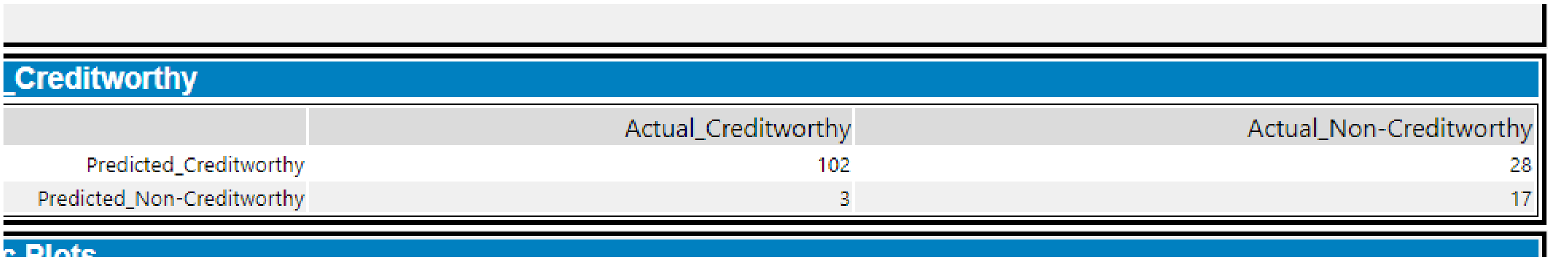
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

  ANS: Credit amount, age, duration of credit month, account balance, most valuable available asset were the top 5 variables ordered by importance in my forest model.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

Ans: The overall accuracy of my forest model was 79%. Its PPV is 0.78 and its NPV is 0.85, the forest model has a slight bias towards the non-credit worthy segment.



_Creditworthy

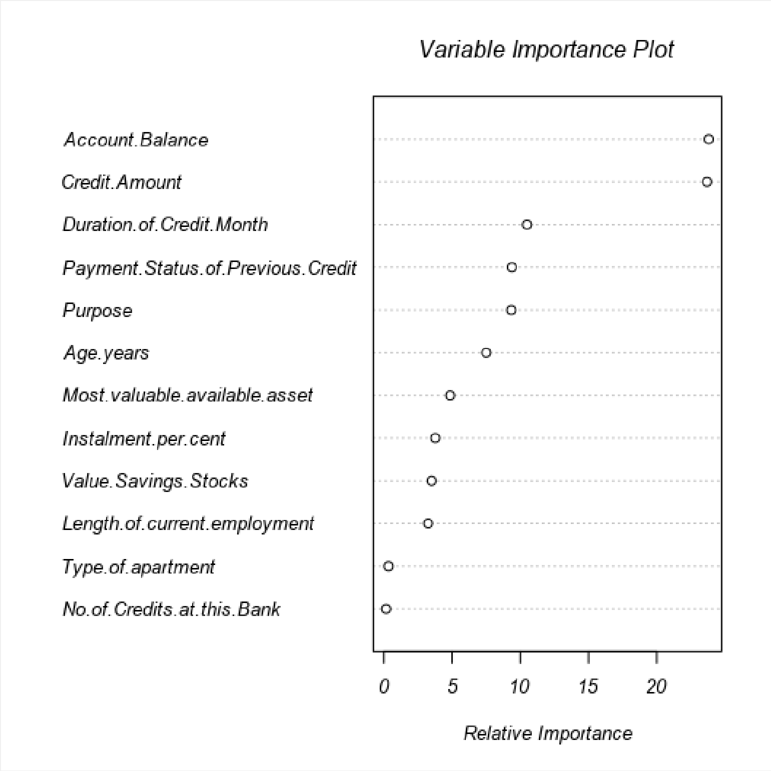
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

c Plots

Boosted Model

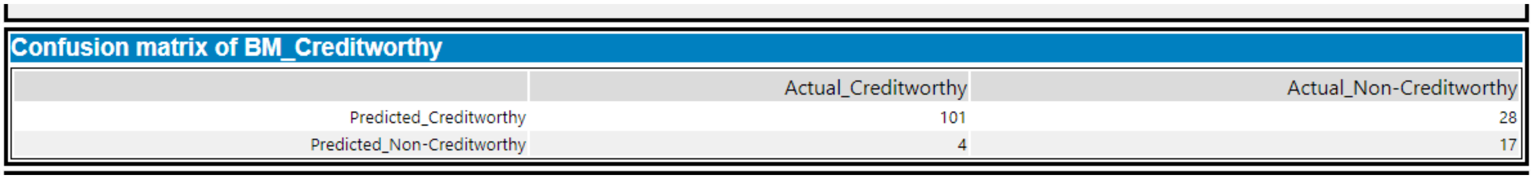
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

  ANS: Account Balance, Credit amount, duration of credit month, payment status of previous credit and purpose were the top 5 variables ordered by importance in my boosted model.



- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Is there any bias seen in the model's predictions?

Ans: The overall accuracy of my boosted model was 78%. Its PPV is 0.78 and its NPV is 0.81, this model has a very slight bias towards the non-credit worthy segment, of all the models it has the least bias.



Confusion matrix of BM_Creditworthy

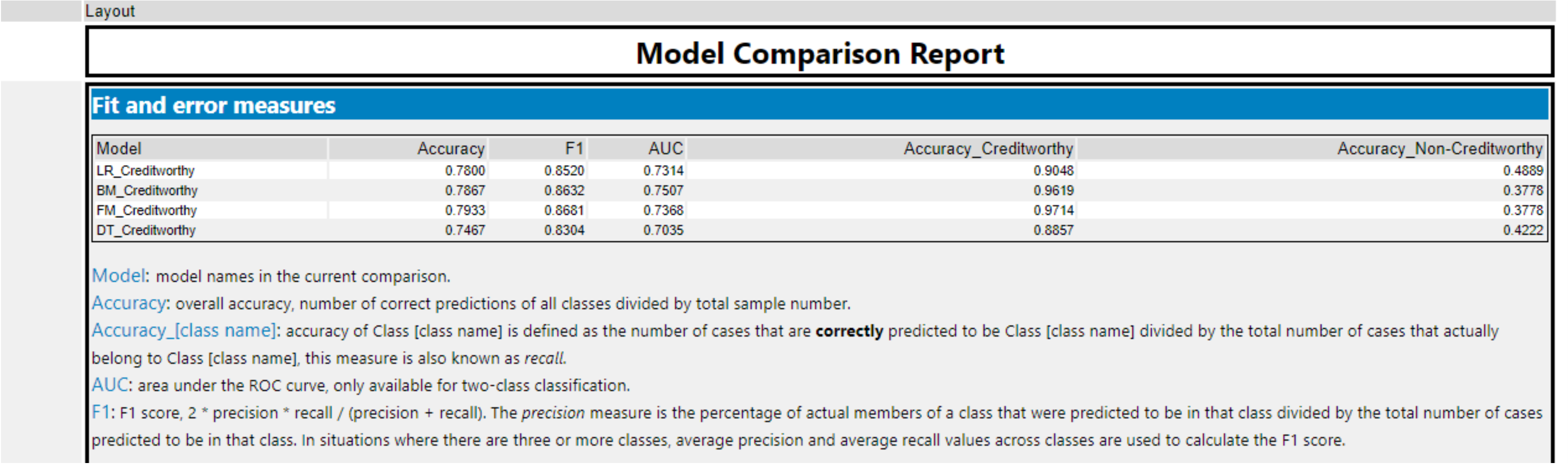
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Step 4: Write-up on Choice Model

If Score_Creditworthy is greater than Score_NonCreditworthy, the person is labeled as "Creditworthy"

Of all the models I tried out above, I chose the Forest Model.

It has the highest overall accuracy, at 79%, it also has the highest creditworthy accuracy, at 97% and a relatively good non-creditworthy accuracy of 37%. Its bias when it comes to predicting either of the segments is negligible with a Positive Predictive Values (PPV) at 0.78 and Negative Predictive Values (NPV) at 0.85 as stated above.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Creditworthy	0.7800	0.8520	0.7314	0.9048	0.4889
BM_Creditworthy	0.7867	0.8632	0.7507	0.9619	0.3778
FM_Creditworthy	0.7933	0.8681	0.7368	0.9714	0.3778
DT_Creditworthy	0.7467	0.8304	0.7035	0.8857	0.4222

Model: model names in the current comparison.

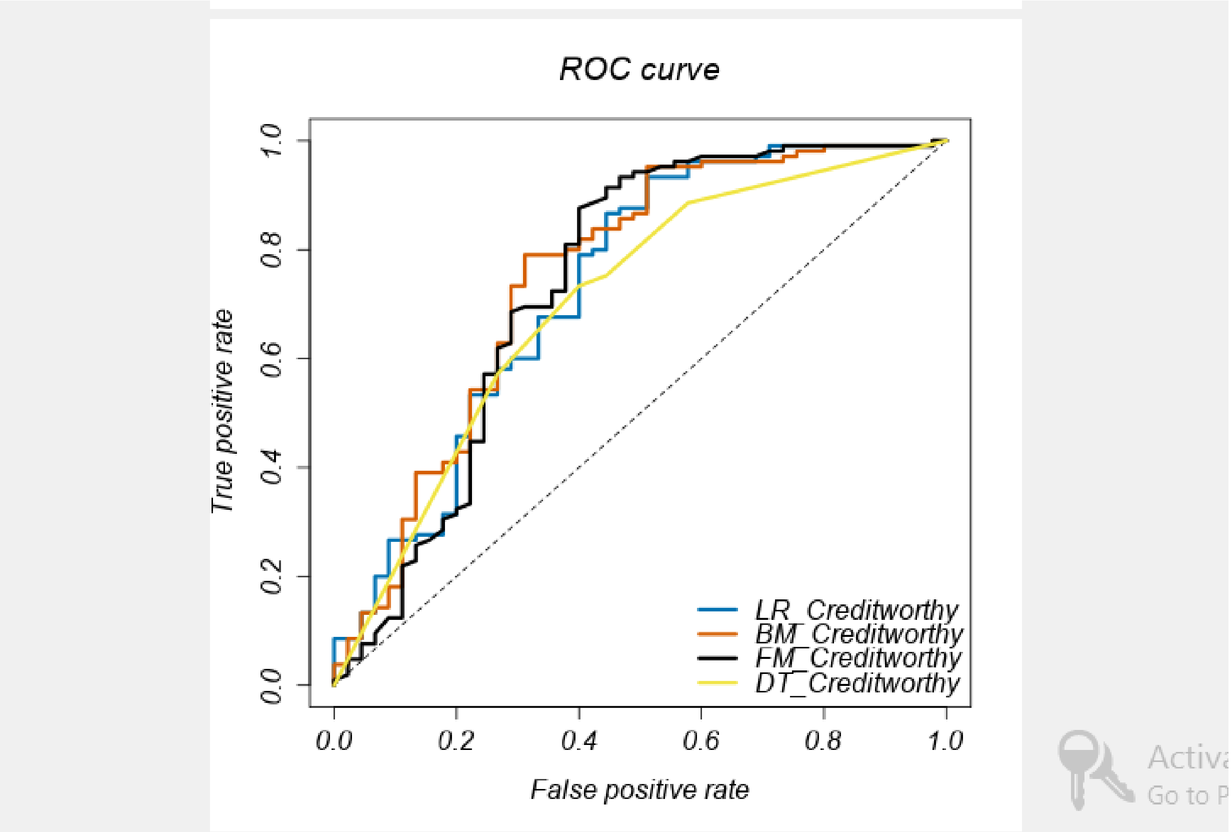
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall.*

AUC: area under the ROC curve, only available for two-class classification.

F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

The Receiver Operating Characteristic (ROC) curve graphically shows the tradeoff between the rate at which the model predicts True Positives versus the rate at which it predicts False Positives. From the image below we can see that the Forest model gets to the top first and it has the second highest area under the curve at 0.74, this shows it has the best highest true positive rate (that is, it will predict the best number of actual credit worthy applicants as credit worthy), given a particular number of false positives (that is, given a particular number of actual non-credit worthy applicant which are predicted as credit worthy).



- How many individuals are creditworthy?

  Ans: They are four hundred and eight (408) credit worthy individuals hence four hundred and eight (408) of our five hundred (500) new applicants, qualify for a loan.