Project: Creditworthiness

Step 1: Business and Data Understanding

The bank typically receives 200 loan applications per week. Currently, the loan approval process is manual at best. Another competitive bank got hit by a scandal, and that lead to a sudden influx of 500 loan applications at our bank.

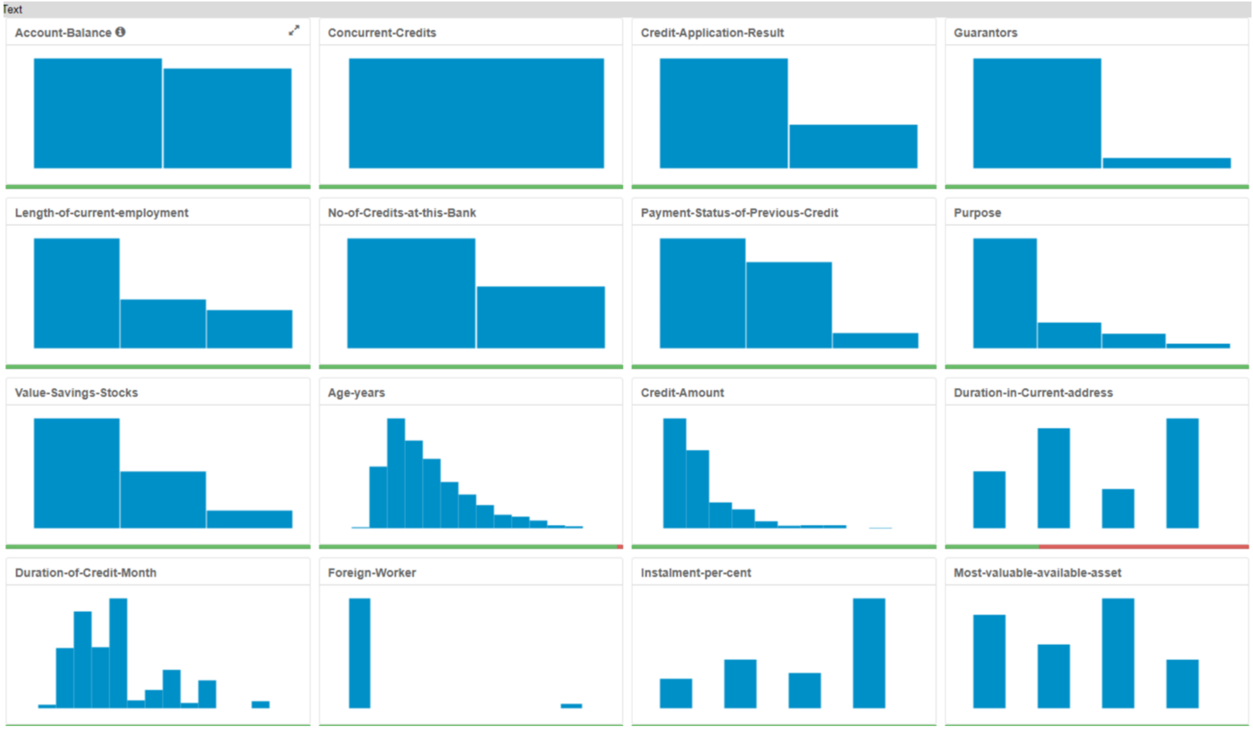
The manager wants me to figure out an approval process that would process the applications within a week.

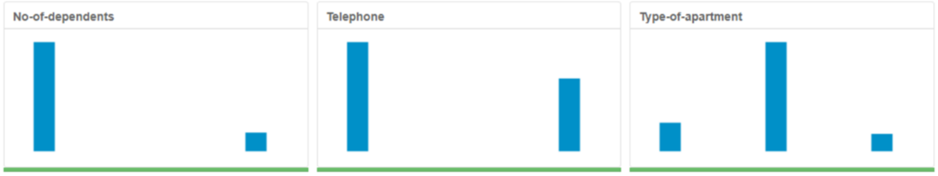
We need historical data from previous applications.

List of new applications to be processed.

We need to use a binary classification model to solve this problem, as the response variable will be either **yes**, approve the loan application or **no**, reject the loan application.

Step 2: Building the Training Set


Text
Account-Balance
Concurrent-Credits
Credit-Application-Result
Guarantors
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit
Purpose
Value-Savings-Stocks
Age-years
Credit-Amount
Duration-in-Current-address
Duration-of-Credit-Month
Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset



Duration in Current Address: This field is missing 69% of the data. Thus, we are going to drop this variable.

Concurrent Credit: There is only one unique value in this field. Hence, we will not consider this variable.

Credit Application Result: This is our target variable. The data is skewed towards yes. However, this could be a business reality.

Guarantors: This variable is heavily skewed towards none. Thus we will not consider this variable in the model.

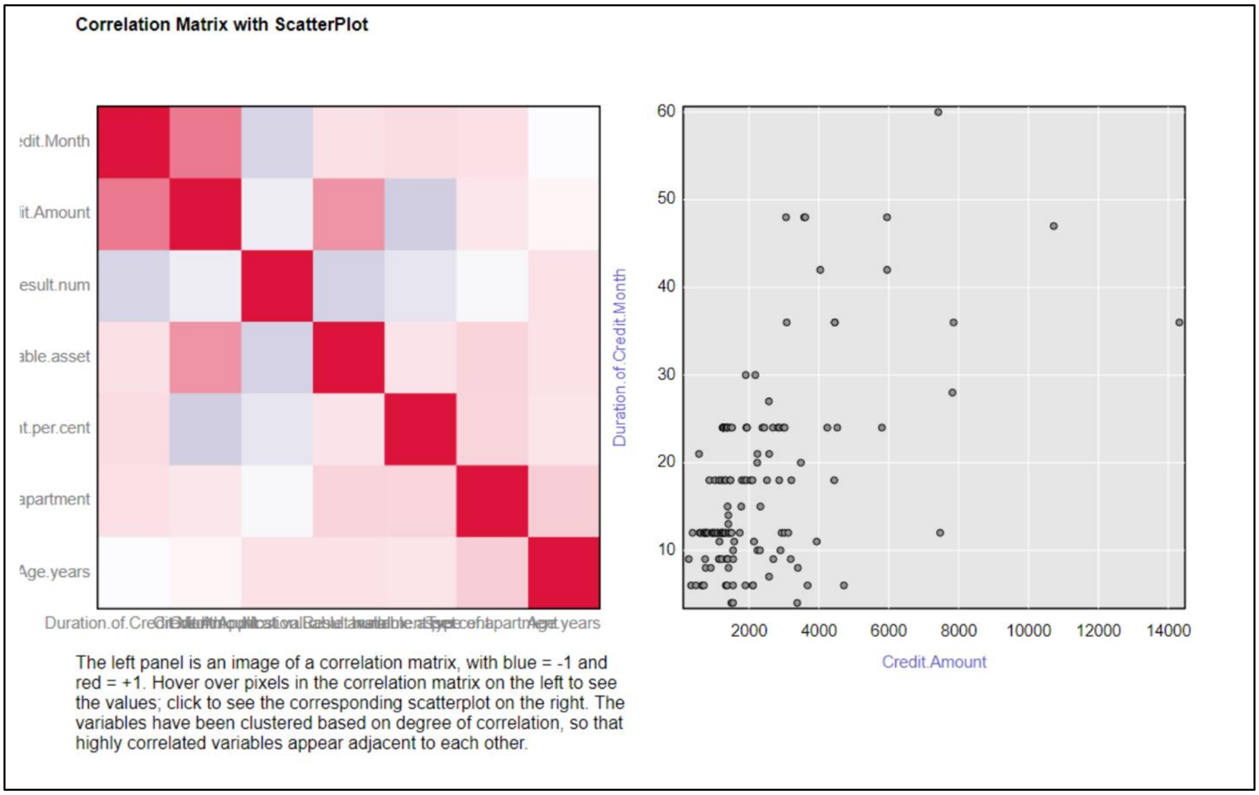
Foreign-Worker: This variable is heavily skewed towards none. Thus we will not consider this variable in the model.

No-of-dependents: This variable is heavily skewed towards none. Thus we will not consider this variable in the model.

Telephone: Telephone number plays no part in determining the creditworthiness of the candidate.

We will also check if there are any duplicate variables and investigate the correlation between the predictor variables.

We didn’t find any highly correlated predictor variable. The threshold used is 70%.



Correlation Matrix with ScatterPlot

The left panel is an image of a correlation matrix, with blue = -1 and red = +1. Hover over pixels in the correlation matrix on the left to see the values; click to see the corresponding scatterplot on the right. The variables have been clustered based on degree of correlation, so that highly correlated variables appear adjacent to each other.

Thus based on the data cleaning process we are going to proceed with following predictor variables :

Account-Balance, Duration-of-Credit-Month, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Value-Savings-Stocks, Length-of-current-employment, Instalment-per-cent, Most-valuable-available-asset, Age-years, Type-of-apartment, No-of-Credits-at-this-Bank.

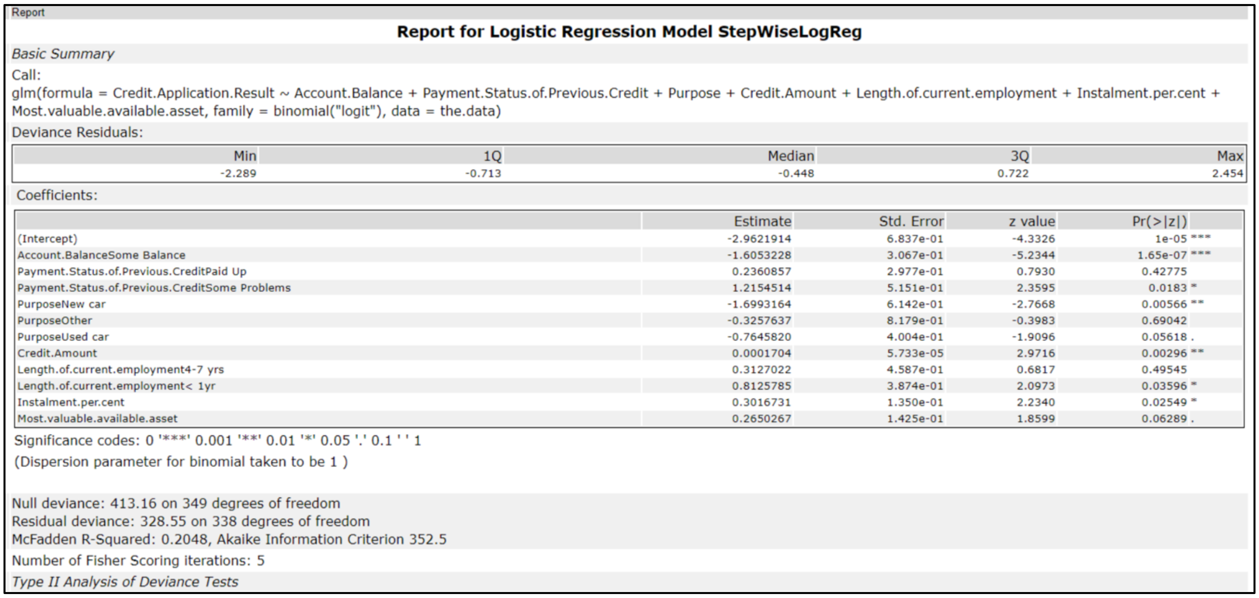
Step 3. Train your Classification Models

To choose best classification model for our dataset, I decided to test following Classification Models.

Stepwise Logistic Regression

The significant variables for this model are Account Balance, Payment Status, Purpose, Credit Amount, Length of Employment,

Installment per cent.



Report

Report for Logistic Regression Model StepWiseLogReg

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)

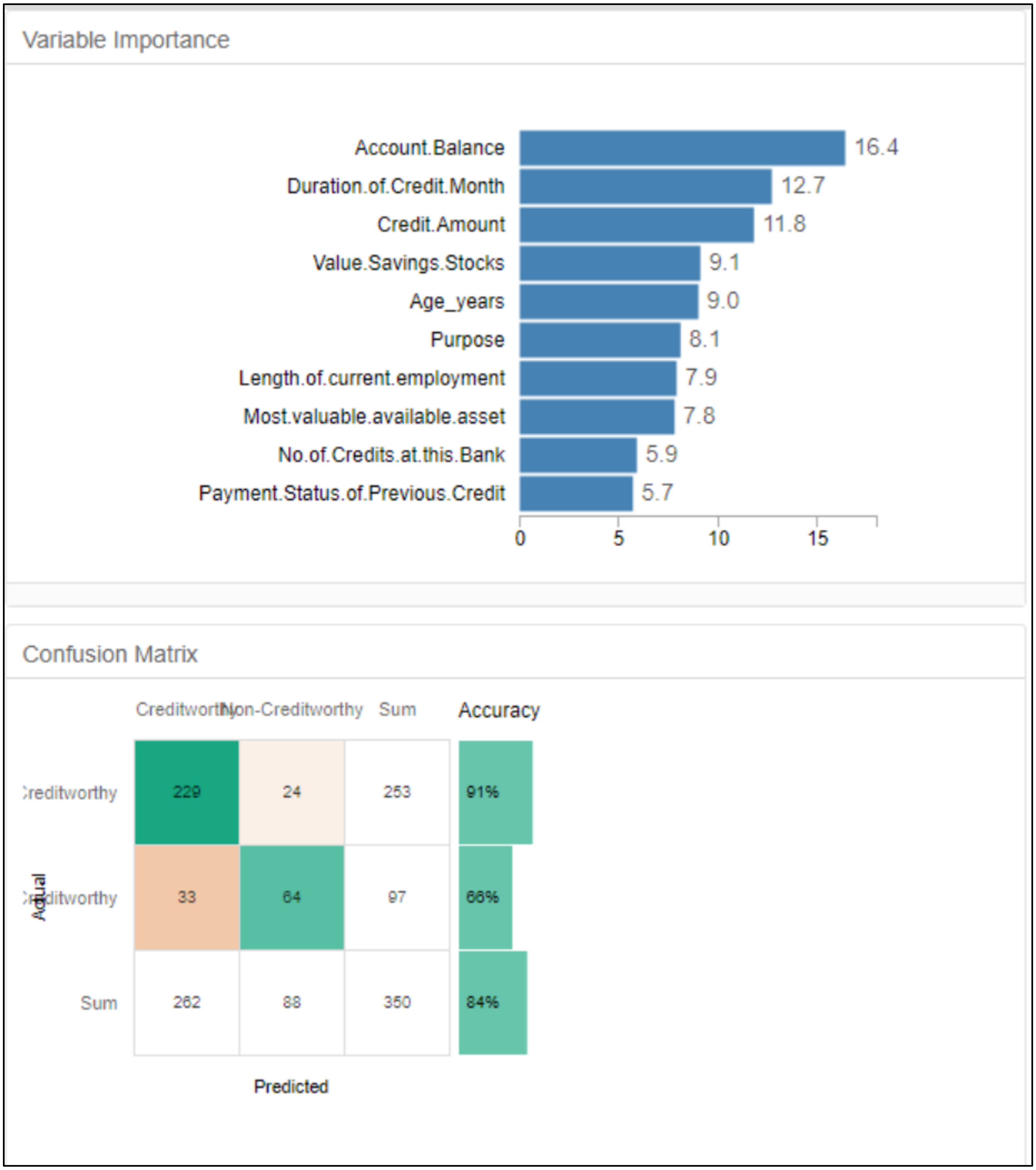
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Decision Tree

The significant variables for this model are Account Balance, Duration of Credit Month, Credit Amount, Value Saving Stocks.



Variable Importance

Account.Balance 16.4
Duration.of.Credit.Month 12.7
Credit.Amount 11.8
Value.Savings.Stocks 9.1
Age_years 9.0
Purpose 8.1
Length.of.current.employment 7.9
Most.valuable.available.asset 7.8
No.of.Credits.at.this.Bank 5.9
Payment.Status.of.Previous.Credit 5.7
0
5
10
15

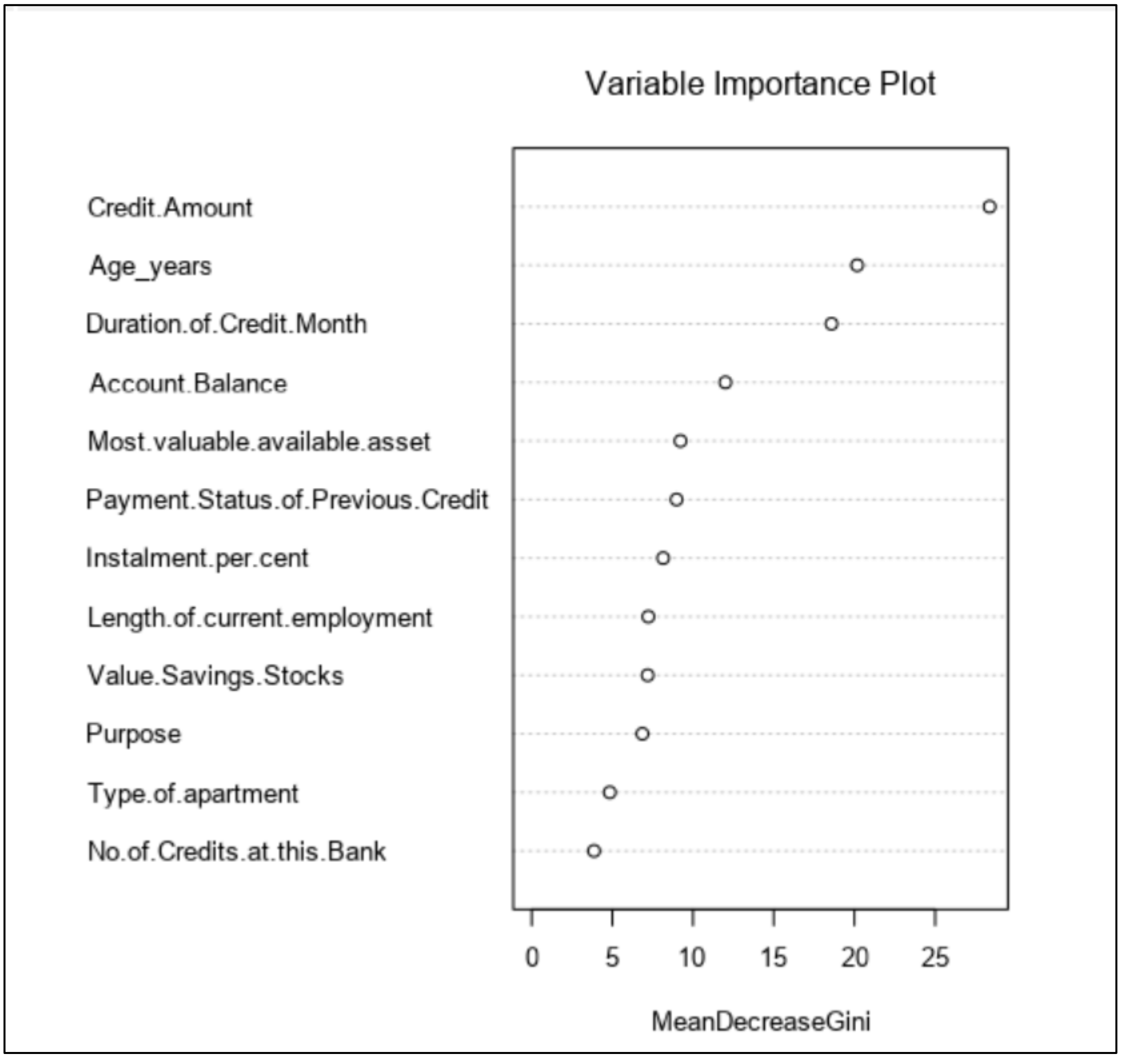
Confusion Matrix

Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
Creditworthy	229	24	253	91%
Non-Creditworthy	33	64	97	66%
Sum	262	88	350	84%

Predicted

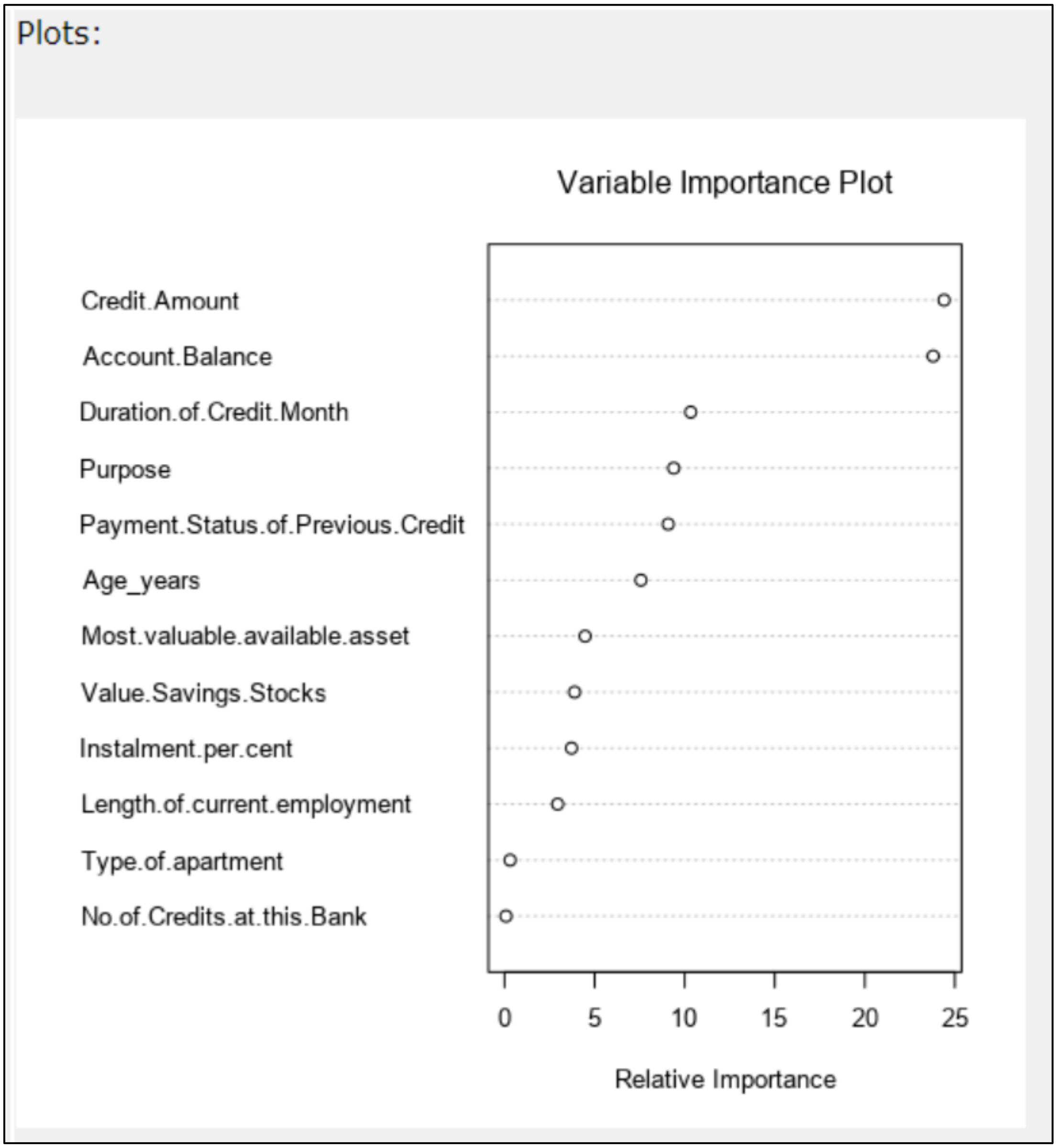
Random Forest

Following are the significant predictor variables for this model.



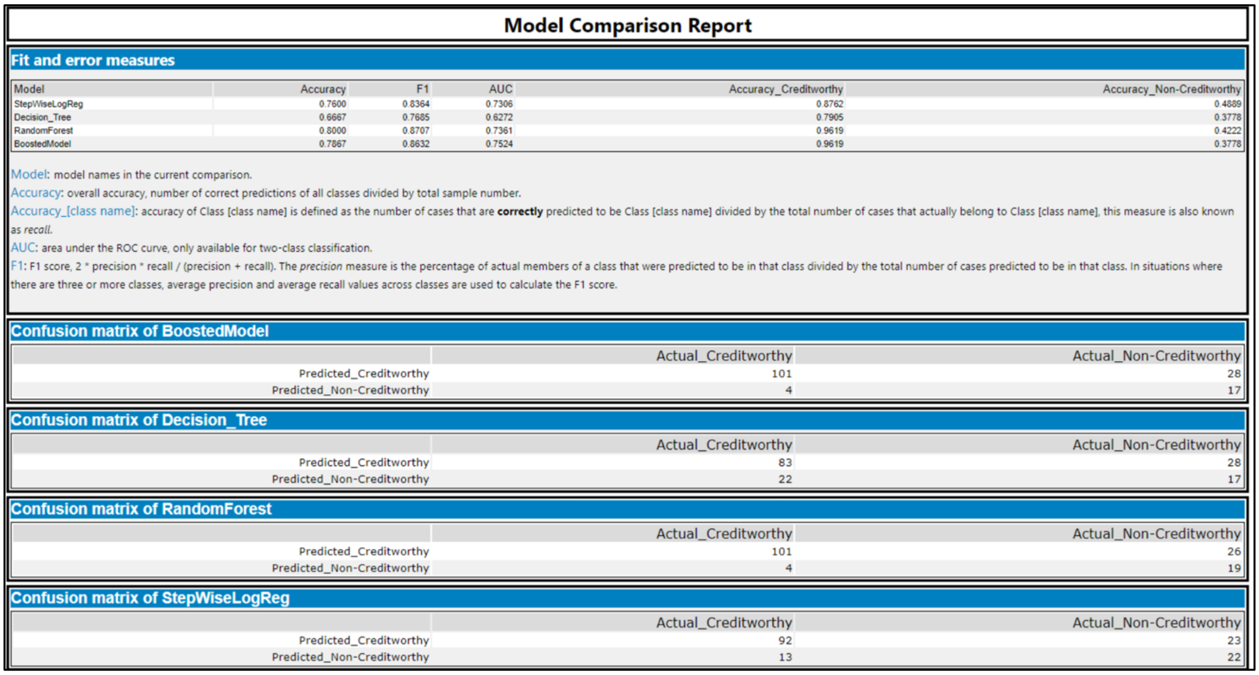
Boosted Model

Following are the significant predictor variables for the model.



Plots:

I used model comparison tool to validated and compare the models and to check if there are any biases in the model.

1. 

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
StepWiseLogReg	0.7600	0.8364	0.7306	0.8762	0.4889
Decision_Tree	0.6667	0.7685	0.6272	0.7905	0.3778
RandomForest	0.8000	0.8707	0.7361	0.9619	0.4222
BoostedModel	0.7867	0.8632	0.7524	0.9619	0.3778

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.

Confusion matrix of BoostedModel

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

Confusion matrix of Decision_Tree

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

Confusion matrix of RandomForest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Confusion matrix of StepWiseLogReg

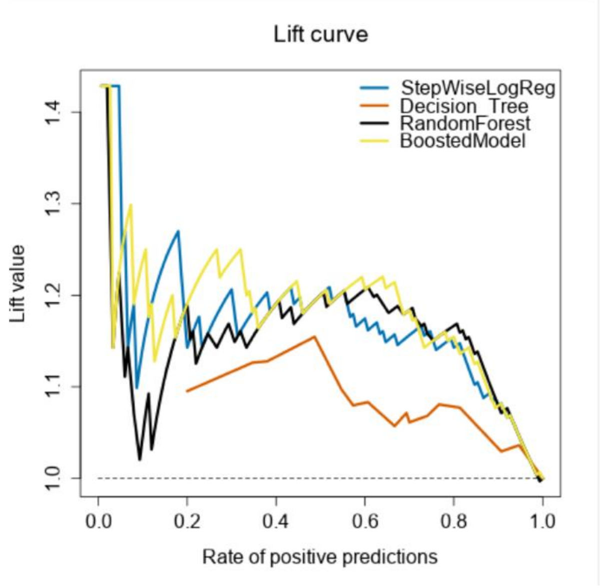
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

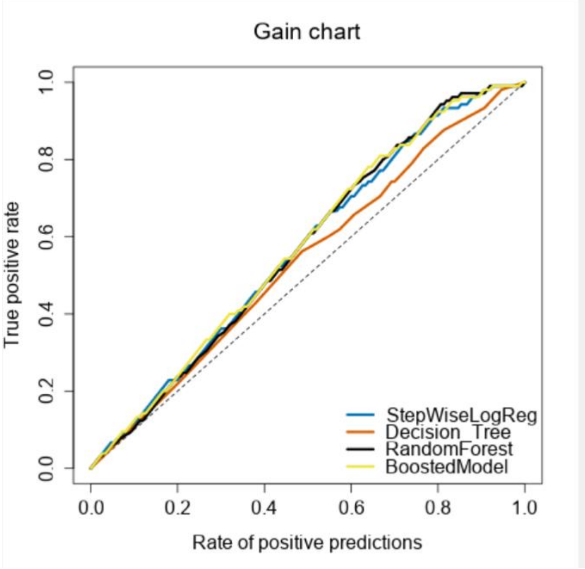
If we compare the accuracy of each model using the Model Comparison report above, we have Random Forest Model(0.80) with the best accuracy score followed by Boosted Model(0.78), then the Stepwise Logistic Model(0.76) then Decision Tree(0.67).

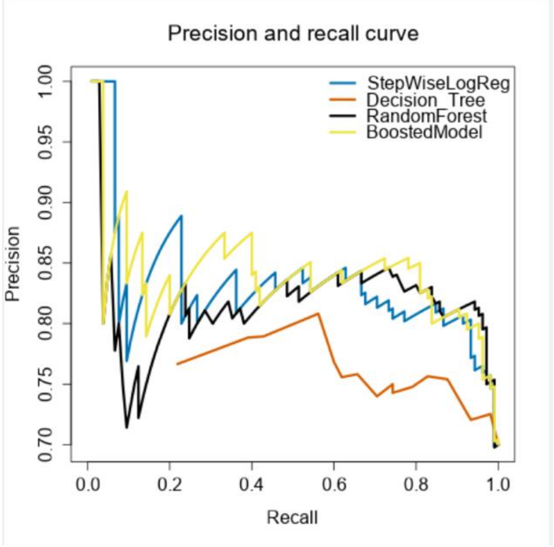
As we have discussed earlier, our dataset has more candidates who are creditworthy than non-creditworthy. Thus there is going to bias in our model to classify candidates as creditworthy.

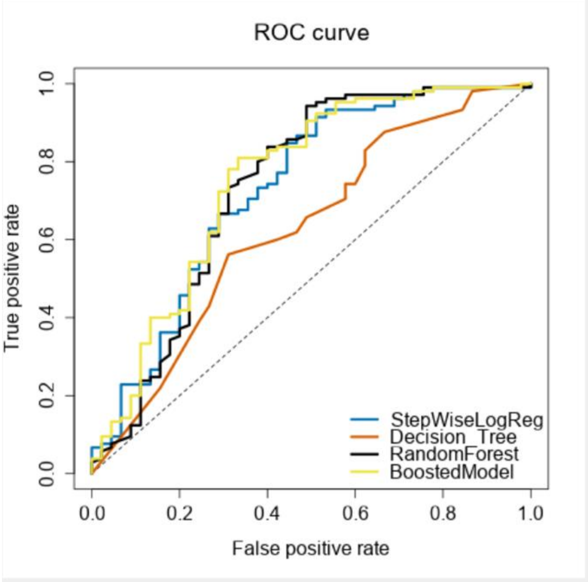
Our Manager is only concerned with our model's classification accuracy for Creditworthy and Non-Creditworthy segments.

Thus, when it comes to selecting the model, we give more weightage to the model's overall accuracy, the positive predictive value(True Positives) and negative predictive value(True Negatives).





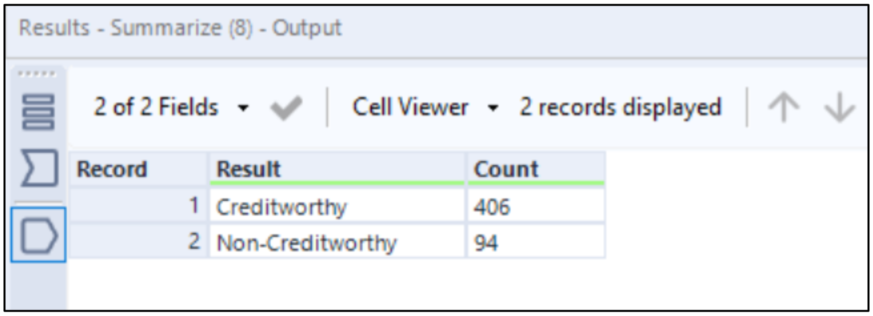




Step 4 WriteUp.

As discussed above we want a model with high overall accuracy, and ability to predict True Positive and True Negatives and the F1 score. Thus, I decided to go ahead with Random Forest Model

After scoring the model I got 406 candidates as credit worthy.



Record	Result	Count
1	Creditworthy	406
2	Non-Creditworthy	94

s