Project: Creditworthiness by Oluwatosin Amosu.

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250-word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

  The decision to be made is to know whether or not loans should be given out to each 500 applicants.

- What data is needed to inform those decisions?

  Data needed could be-

  1) Salaried or not: Whether the person is employed and receiving a salary or not helps us to identify creditworthiness.

  2) Income: A person's income signifies the loan amount that can be safely opted

  3) Amount of EMI Payments per month

  4) Repayment Tenure: the duration of the loan.

  5) Credit Amount: The amount of credit in view with the applicant's income can give us a sense whether they will default on their loans and many more.

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

  Since the problem of determining creditworthiness requires two answers. i.e. “Yes” or “No”, we will go for Binary classification model.

Step 2: Building the Training Set

Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn’t ***need to convert any data fields to the appropriate data types.***

The following steps are taken in building the training set:

STEP 1: Drag the input data tool to the canvas and select the credit data training set.

STEP 2: Use field summary tool to check the profile of each variable and to determine which variable(s) should not be consider as a potential predictor variable.

The results of the field summary are as follows:

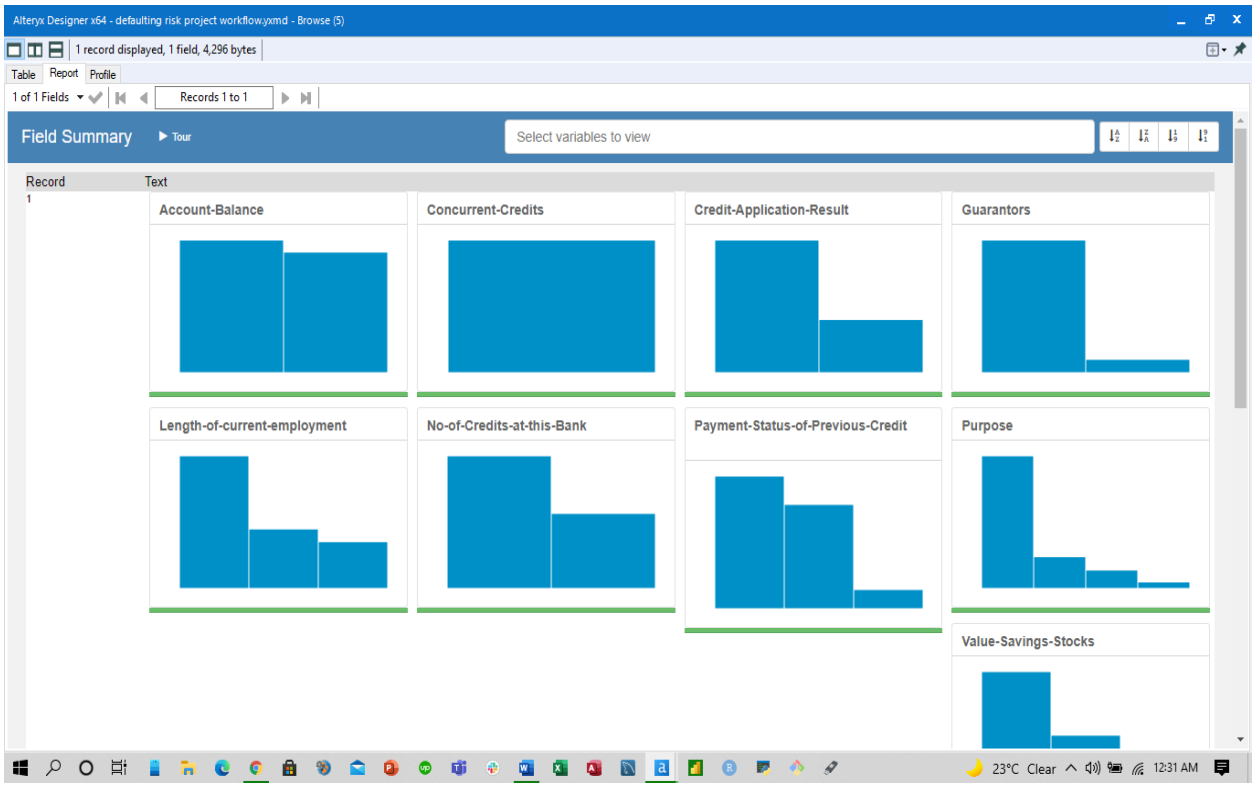


Figure 1a: The field summary result part 1 from the interactive report.

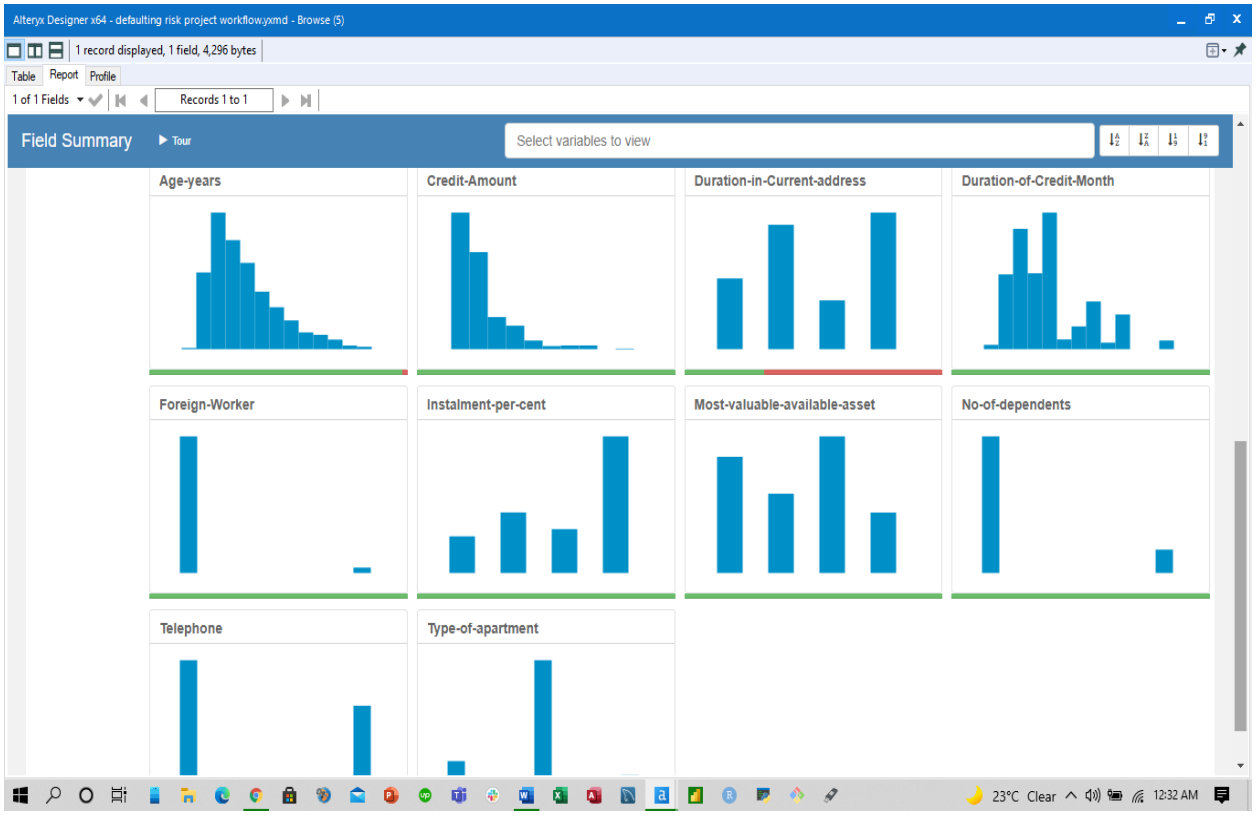


Fig 1b: The field summary result part 2 from the interactive report.

1. From Fig 1a, it seems variables like Account-Balance, Credit-Application-Result, Length-of-current-employment, No-of-Credits-at-this-Bank, Payment-Status-of-Previous-Credit, Purpose, Value-Savings-Stock, Instalment-per-cent, Most-valuable-available-asset, and Type-of-apartment are healthy. Variables with more than 4 categories are not recommended as good predictor variables as they make prediction difficult but instead of selecting it manually, a stepwise tool was used to automatically select the best combination of predictor variables. The goal here is to remove variables that aren't fit to be a potential predictor variable.
2. The **concurrent-credit** variable was removed because the data is entirely uniform and there is no variation of the data. This is a type of low variability.
3. The **Guarantors** variable was removed because the data skewed towards none as this is a variable with low variability.
4. From Fig 1b, **Duration-in-current-address** variable was also removed because 69% (more than 50%) of the records were missing. The best method to deal with variable with more than 50% missing data is to delete it.
5. Fig 1b shows that **foreign-worker** variable was removed because out of the 500 instances, 481 instances skewed to be a (1) foreign worker while 19 instances are two workers.
6. **No of dependent** variable was removed due to its low variability.
7. I removed the **telephone** variable because of its low variability. That is, applicants with more telephones will be creditworthy which will make the data to skew to their side.

N: B: **Occupation** variable was removed also because it wasn't displayed in the field summary results. When checked out from the raw dataset, it appears that same value was recorded for every applicant. In this case, such variable is not beneficial to our model.

STEP 3: **Select** tool was used to remove the justified predictor variables above.

STEP 4: **Association Analysis** tool was used to check predictor variables that are highly correlated to each other and also to the target variable. We are basically interested in variables with low correlation.

STEP 5: **Summarize tool** was used to find the median of the **Age Years** field ignoring zeros and was later appended using the Append Fields tool to the rest of the records.

STEP 6: A **formular tool** was used to impute the null cells with the median value. The median value was used because the data is "skewed right" (the tail at the right end), values are leaned towards one side and hence mean will not be appropriate.

STEP 7: Select tool was used to deselect the new field created called **MedianNo0_Age-years** as a result of summarize tool.

STEP 8: This is the final step in building the training data. At this step, the data is already cleaned and ready to modelling.

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

LOGISTIC REGRESSION ALGORITHM

In training and validating the logistic regression model, the following steps are followed:

STEP 1: Create samples tool was used and estimation was set at 70%, 30% for validation while setting the random seed to be 1.

STEP 2: The logistic regression model was created with the estimation dataset

STEP 3: Stepwise tool was used to select the best combination of predictor variables.

STEP 4: For the first model which is the logistic regression model as seen from the workflow below, and also from Fig 2b, it can be seen that there are 7 variables that are statistically significant excluding the intercept but the most important significant variables are **Account-Balance, Purpose**, and **Credit-Amount.** Also, the p-value for each variable is also shown in Fig 2b.

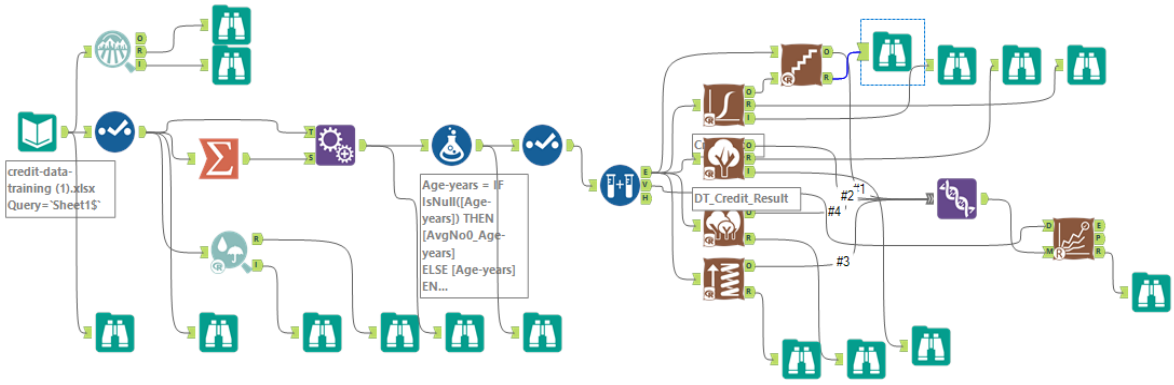


Fig 2a: Workflow to build the model, train the model and for the validation of the model.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.990817	1.013e+00	-2.9527	0.00315	**
Account.BalanceSome Balance	-1.543669	3.233e-01	-4.7745	1.80e-06	***
Duration.of.Credit.Month	0.006391	1.371e-02	0.4660	0.6412	
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	*
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	.
Credit.Amount	0.000177	6.841e-05	2.5879	0.00966	**
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	
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	*
Instalment.per.cent	0.310524	1.399e-01	2.2197	0.02644	*
Most.valuable.available.asset	0.325606	1.557e-01	2.0918	0.03645	*
Age.years	-0.015092	1.539e-02	-0.9809	0.32666	
Type.of.apartment	-0.254565	2.958e-01	-0.8605	0.38949	
No.of.Credits.at.this.BankMore than 1	0.362688	3.816e-01	0.9505	0.34184	

Figure 2b: Result of the logistic regression model with all potential predictor variables.

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 322.19 on 332 degrees of freedom
McFadden R-Squared: 0.2202, Akaike Information Criterion 358.2

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Figure 2c: R2 for the logistic regression model without the stepwise tool.

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

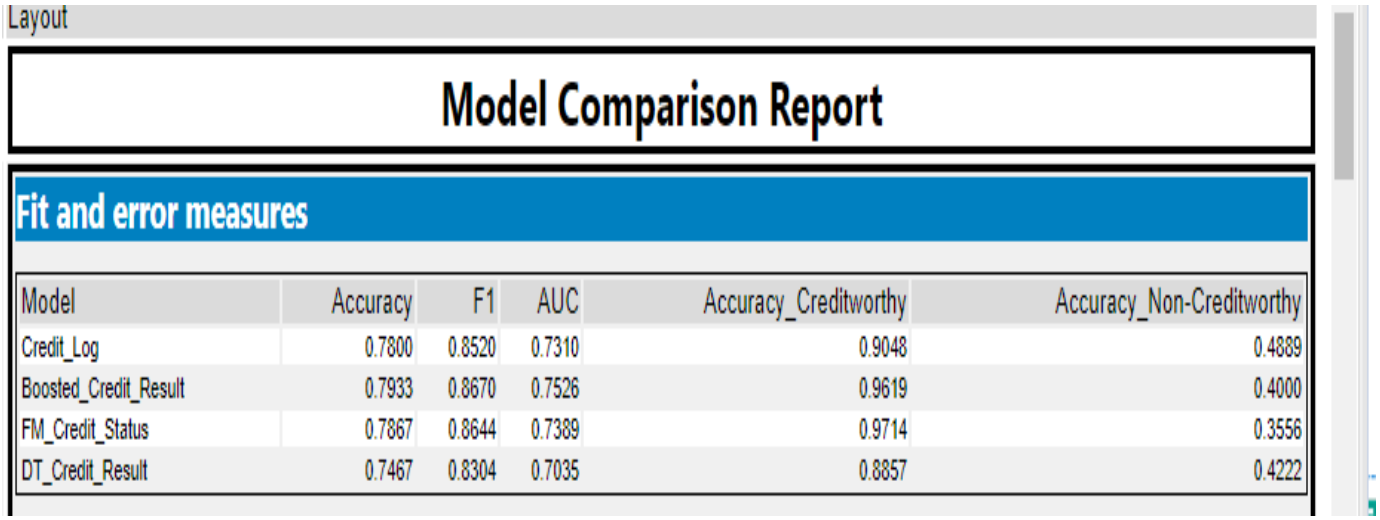
Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Figure 2d: R2 for the stepwise regression model.

N: B: It was observed that the logistic regression model has more R2 value thus more overall accuracy (0.774 as seen in Fig) when all potential predictor variables are used than when the stepwise tool was used (0.7600 as seen in Fig). Having said that, the logistic regression model was used to validate the model.

STEP 4: To validate the model for the logistic regression model, we check the overall percent accuracy which is 78%.


Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Credit_Log	0.7800	0.8520	0.7310	0.9048	0.4889
Boosted_Credit_Result	0.7933	0.8670	0.7526	0.9619	0.4000
FM_Credit_Status	0.7867	0.8644	0.7389	0.9714	0.3556
DT_Credit_Result	0.7467	0.8304	0.7035	0.8857	0.4222

Figure 3a: Accuracy of all the four classification models.

Discussing Bias

From the Fig 3b below, the classifier predicted 95 to be creditworthy and they were actually creditworthy in reality.

- The classifier predicted 22 applicants to be not-creditworthy and they were not creditworthy in reality.

There is total 150 predictions made by this logistic regression model and out of the 150 predictions, it predicted a total of 118 applicants to be creditworthy and a total of 32 applicants not to be creditworthy. Out of the 118 applicants it predicted to be creditworthy, 96 are actual creditworthy (right) and 23 are not creditworthy (wrong).

Using the Precision - Predicted Positive Value (PPV) and Recall Negative Predicted Value (NPV) to discuss the bias for the logistic regression, we have:

PPV= true positives \ (true positives + false positives)

NPV= true negatives \ (true negatives + false negatives).

The overall percent accuracy of the Logistic model is 78% which is strong.

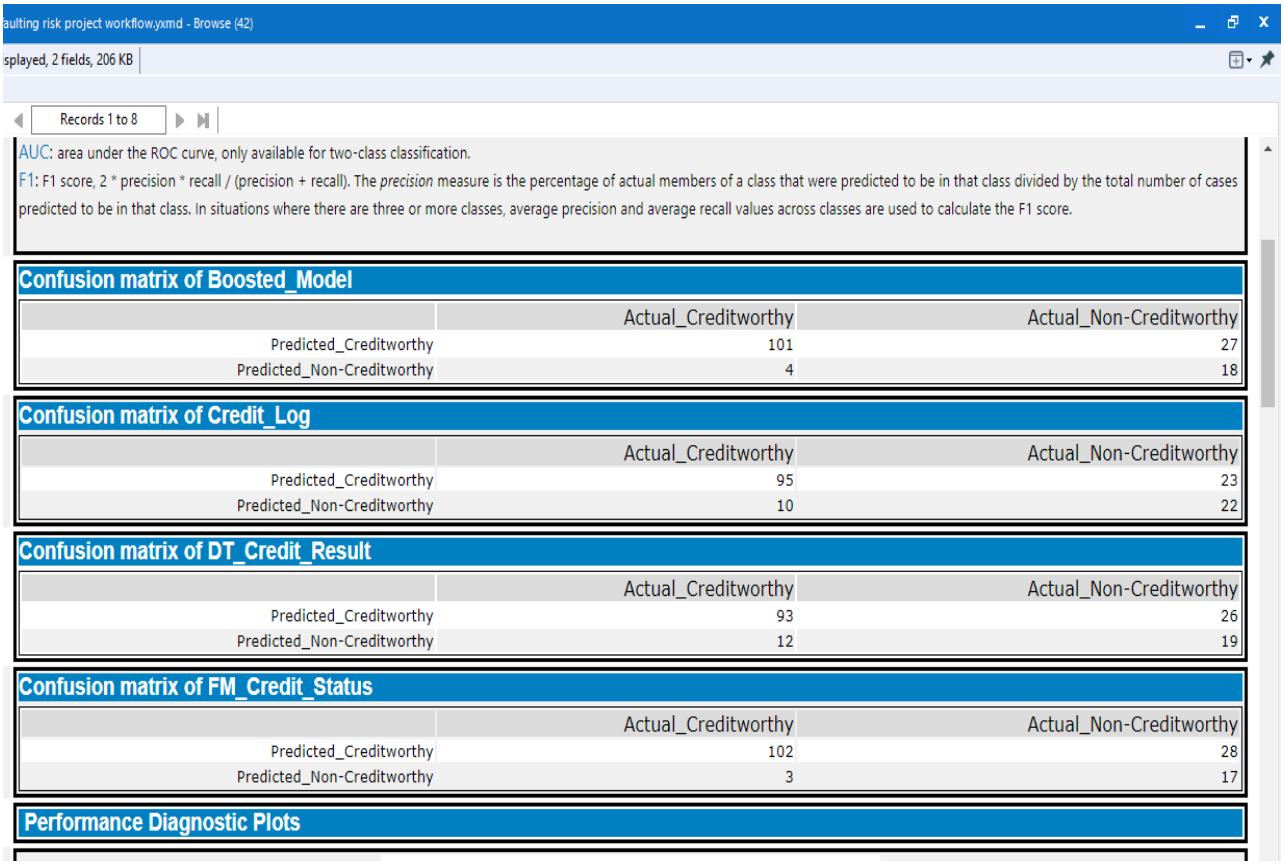


Fig 3b: model comparison results

PPV= true positives \ (true positives + false positives) = 95 / (95+23) =.81

NPV= true negatives\ (true negatives + false negatives) =22/ (22+10) = .68

So, after checking the confusion matrix there is bias seen in the model's prediction to non-Creditworthy due to much differences between the PPV and the NPV.

DECISION TREE ALGORITHM

In training and validating the decision tree model, the following steps are followed:

STEP 1 AND 2 was repeated from the logistic regression model

STEP 3: For the second model which is the decision tree model. Also, from Fig 4a, it can be seen that there are 10 variables of importance but the most important predictor variables are **Account-Balance, Value-Saving-Stock**, and **Duration.Of.Credit.MonthD.**

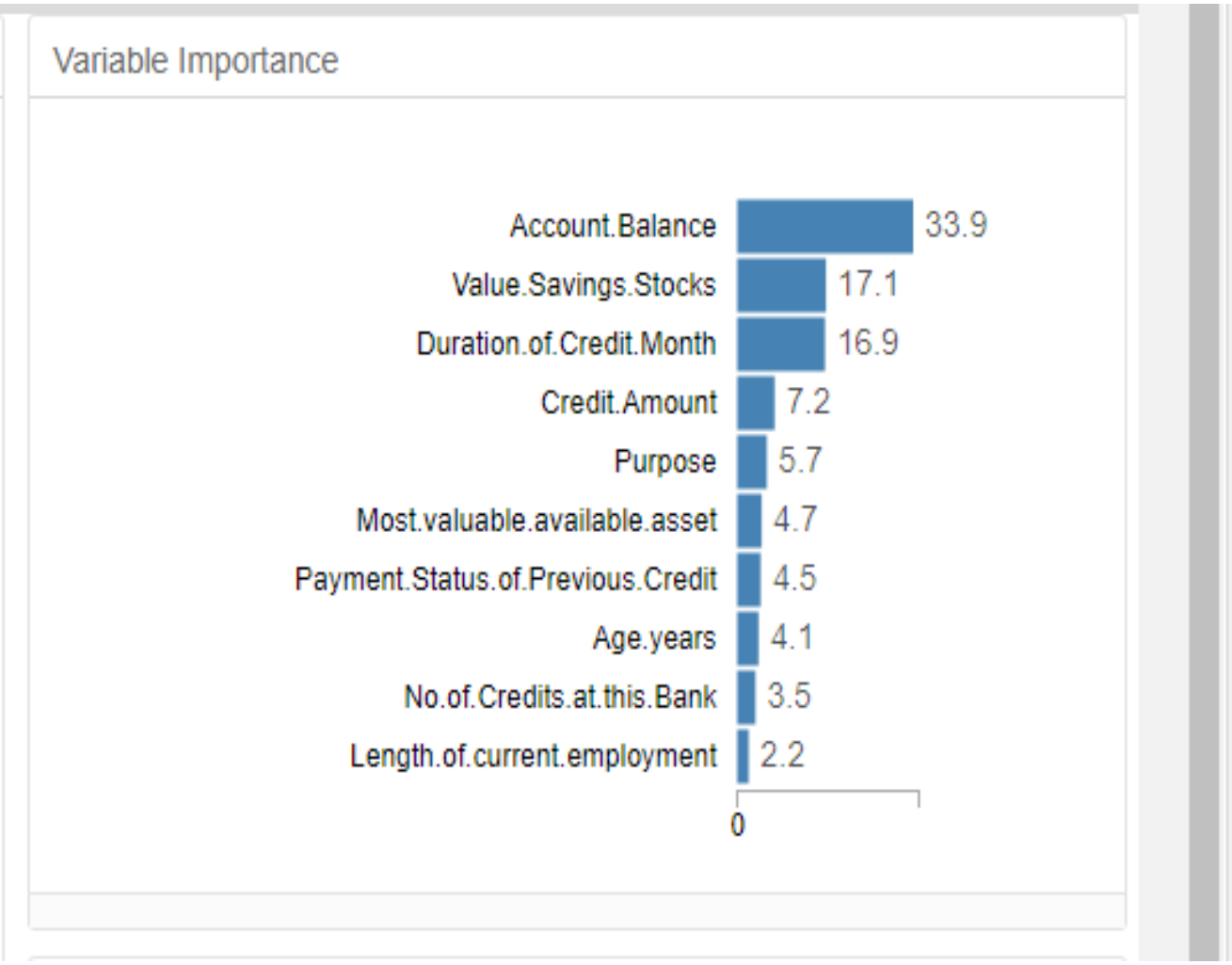


Fig 4a: Variable importance chart from the decision tree

STEP 4: To validate the model for the decision tree model, we check the overall percent accuracy which is 75% from Fig 3a

Discussing Bias from The Confusion Matrix

From the Fig 3b above, the classifier predicted 93 to be creditworthy and they were actually creditworthy in reality.

- The classifier predicted 19 applicants to be not-creditworthy and they were not creditworthy in reality.

There is total 150 predictions made by this logistic regression model and out of the 150 predictions, it predicted a total of 119 applicants to be creditworthy and a total of 31 applicants not to be creditworthy. Out of the 119 applicants it predicted to be creditworthy, 93 are actual creditworthy (right) and 26 are not creditworthy (wrong).

Using the Precision - Predicted Positive Value (PPV) and Recall Negative Predicted Value (NPV) to discuss the bias for the logistic regression, we have:

PPV= true positives \ (true positives + false positives)

NPV= true negatives \ (true negatives + false negatives).

The overall percent accuracy of the Logistic model is 75% which is strong.

PPV= true positives \ (true positives + false positives) = 93 / (93+26) =.78

NPV= true negatives\ (true negatives + false negatives) =19/ (19+12) = .61

So, after checking the confusion matrix there is bias seen in the model's prediction to non-Creditworthy due to much differences between the PPV and the NPV.

RANDOM FOREST MODEL ALGORITHM

In training and validating the forest model, the following steps are followed:

STEP 1 AND 2 was repeated from the logistic regression model

STEP 3: For the third model which is the forest model. Also, from Fig 5a, it can be seen that there are 12 variables of importance but the three most important predictor variables are **Credit.Amount, Age.years**, and **Duration.Of.Credit.Month.**

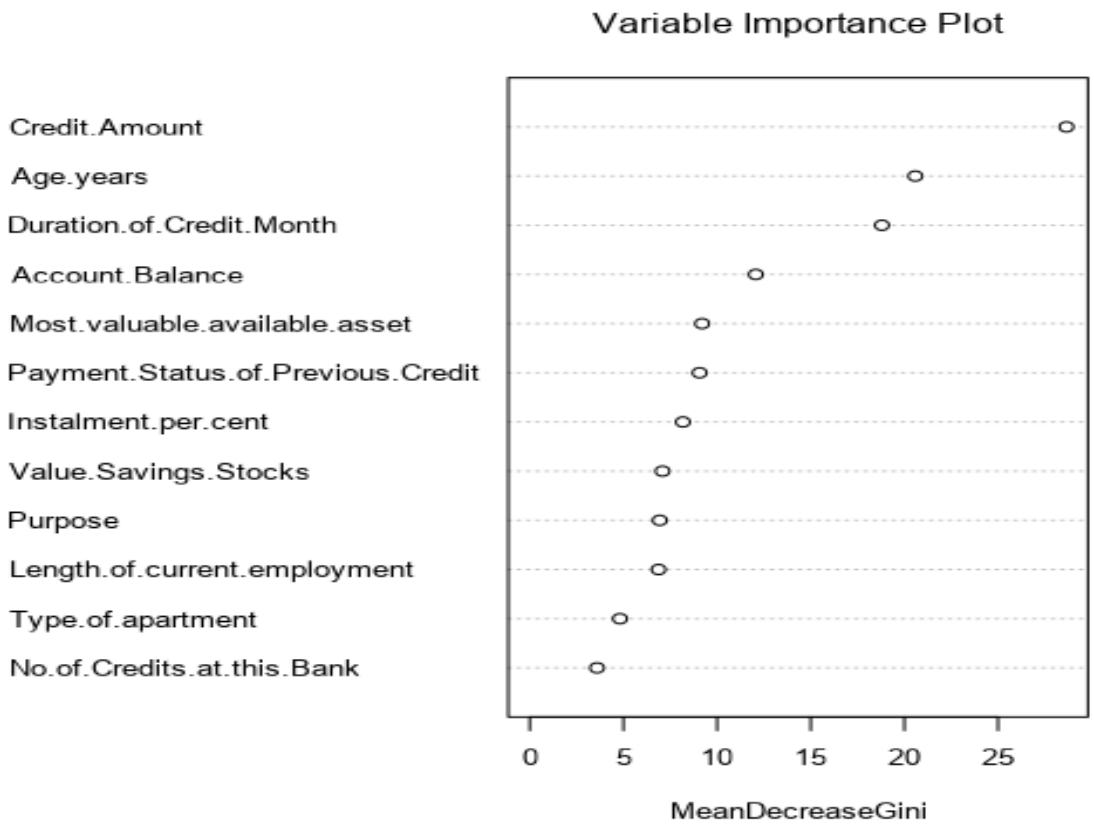


Figure 5a: Variable importance plot for the random forest model.

STEP 4: To validate the model for the random forest model, we check the overall percent accuracy which is 78% from Fig 3a

Discussing Bias from The Confusion Matrix

From the Fig 3b above, the classifier predicted 102 to be creditworthy and they were actually creditworthy in reality.

- The classifier predicted 17 applicants to be not-creditworthy and they were not creditworthy in reality.

There is total 150 predictions made by this logistic regression model and out of the 150 predictions, it predicted a total of 130 applicants to be creditworthy and a total of 20 applicants not to be creditworthy. Out of the 130 applicants it predicted to be creditworthy, 102 are actual creditworthy (right) and 28 are not creditworthy (wrong).

Using the Precision - Predicted Positive Value (PPV) and Recall Negative Predicted Value (NPV) to discuss the bias for the logistic regression, we have:

PPV= true positives \ (true positives + false positives)

NPV= true negatives \ (true negatives + false negatives).

The overall percent accuracy of the Logistic model is 78% which is strong.

PPV= true positives \ (true positives + false positives) =102 / (102+28) =.78

NPV= true negatives\ (true negatives + false negatives) =17/ (17+3) = .85

So, after checking the confusion matrix there is NO bias seen in the model's prediction to Creditworthy and it has the least differences between the PPV and NPV.

BOOSTED MODEL ALGORITHM

In training and validating the boosted model, the following steps are followed:

STEP 1 AND 2 was repeated from the logistic regression model

STEP 3: For the fourth model which is the boosted model. Also, from Fig 6a, it can be seen that there are 10 variables of importance but the three most important predictor variables are **Account.Balance**, **Credit.Amount,** and **Duration.Of.Credit.Month.**

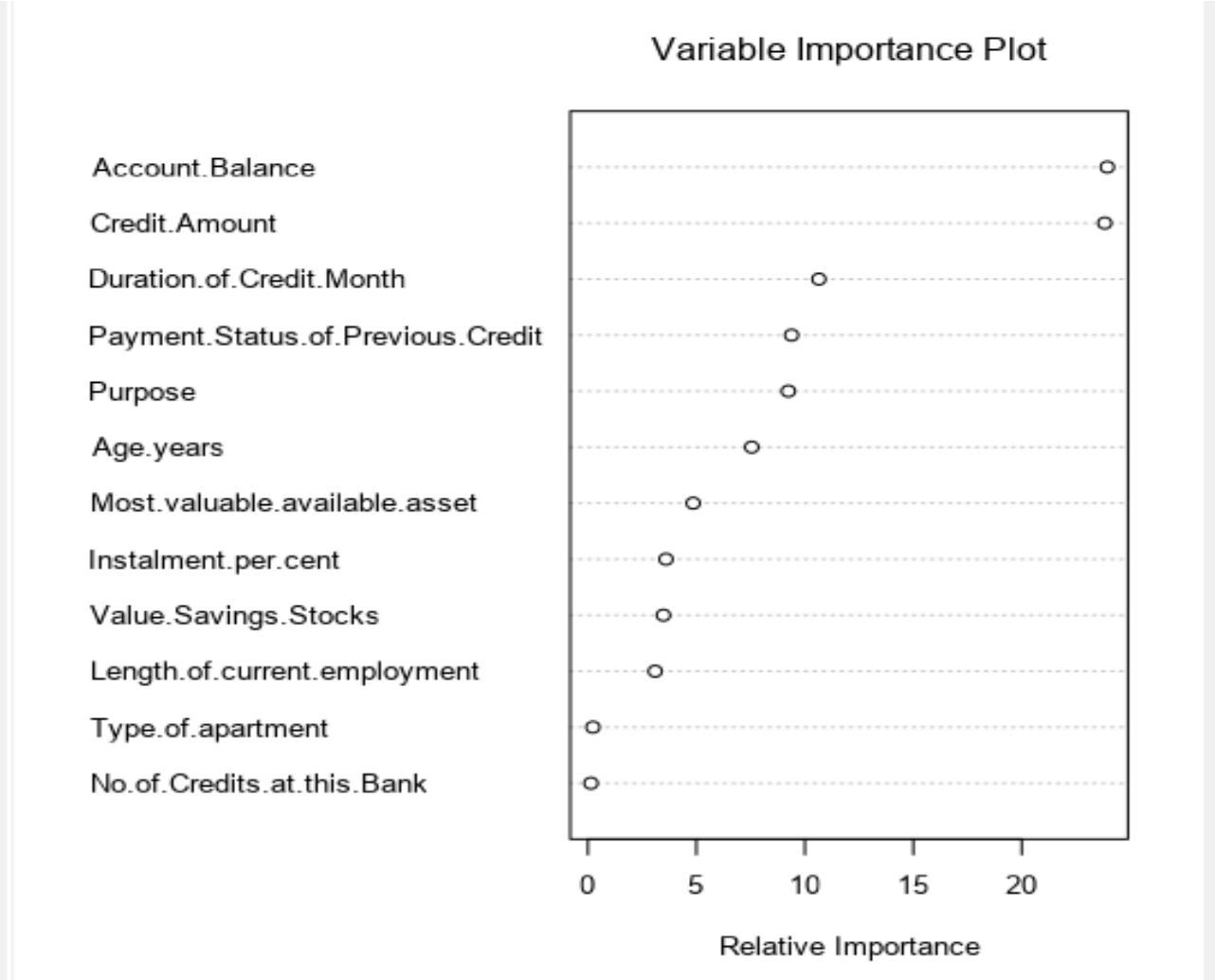


Fig 6a: Variable importance plot for boosted model

STEP 4: To validate the model for the random forest model, we check the overall accuracy which is 79% from Fig 3a and we will consider the variable importance model and the number of iterations assignment plot. The number of iterations assignment plot help us show the amount of variance or the deviance that is captured with more iterations. This

plot shows that as more trees are added, the prediction became smarter as seen below in Fig 6b.

Also, this model performs well with independent sample.

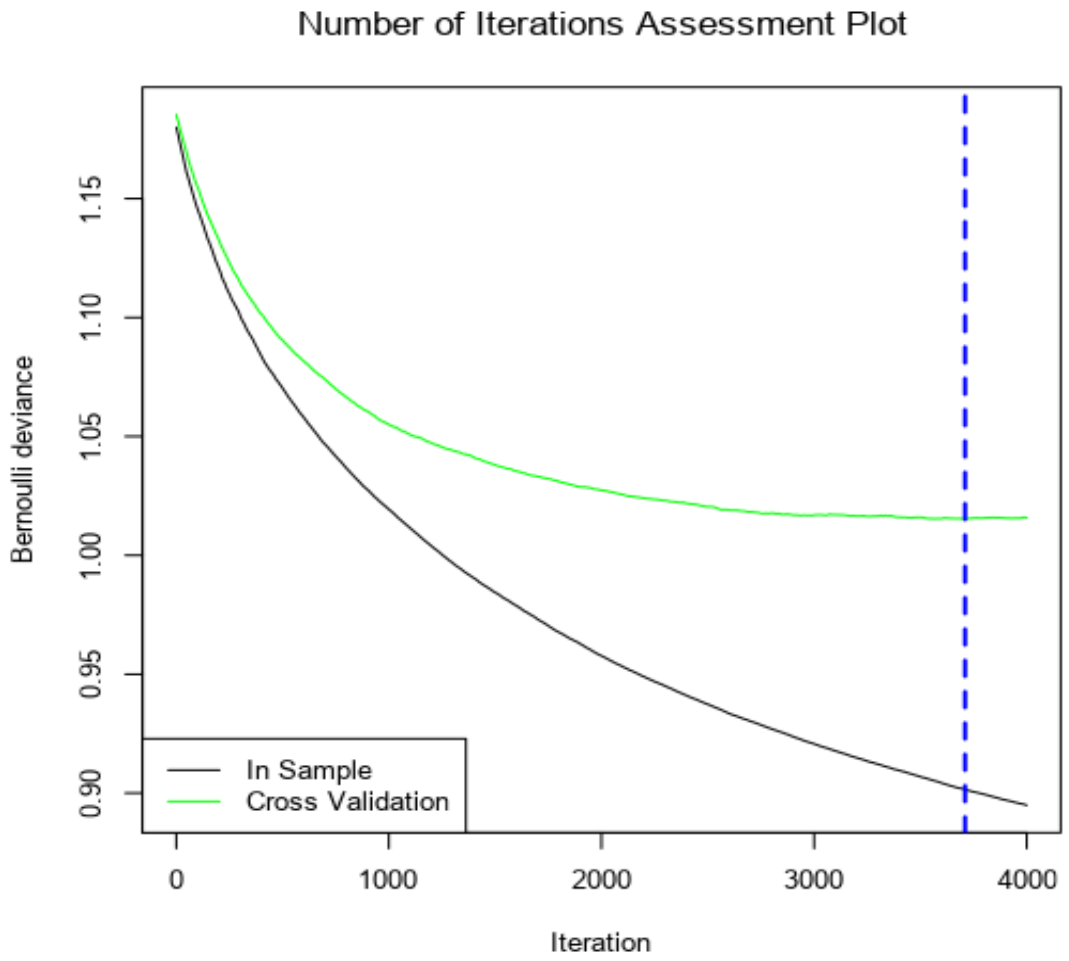


Fig 6b: number of iterations assignment plot

Discussing Bias from the Confusion Matrix

From the Fig 3b above, the classifier predicted 101 to be creditworthy and they were actually creditworthy in reality.

- The classifier predicted 18 applicants to be not-creditworthy and they were not creditworthy in reality.

There is total 150 predictions made by this logistic regression model and out of the 150 predictions, it predicted a total of 128 applicants to be creditworthy and a total of 22

applicants not to be creditworthy. Out of the 128 applicants it predicted to be creditworthy, 101 are actual creditworthy (right) and 27 are not creditworthy (wrong).

Using the Precision - Predicted Positive Value (PPV) and Recall Negative Predicted Value (NPV) to discuss the bias for the logistic regression, we have:
PPV= true positives \ (true positives + false positives)
NPV= true negatives \ (true negatives + false negatives).
The overall percent accuracy of the Logistic model is 79% which is strong.

PPV= true positives \ (true positives + false positives) =101 / (101+27) =.79

NPV= true negatives\ (true negatives + false negatives) =18/ (18+4) = .82

So, after checking the confusion matrix there seems not to be bias in the model's prediction to Creditworthy.

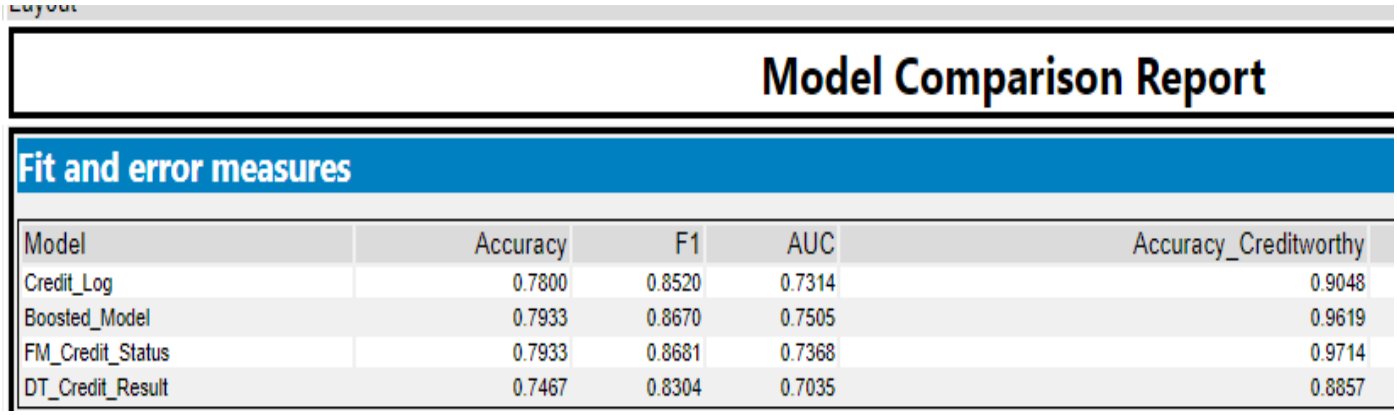
Step 4: Writeup

Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

To decide the best model to go with, we will explore the following techniques:

1. **Overall Accuracy determining Creditworthiness**



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy
Credit_Log	0.7800	0.8520	0.7314	0.9048
Boosted_Model	0.7933	0.8670	0.7505	0.9619
FM_Credit_Status	0.7933	0.8681	0.7368	0.9714
DT_Credit_Result	0.7467	0.8304	0.7035	0.8857

Figure 7a: Shows the model comparison report.

From the report, the forest and the boosted model has the highest accuracy of **0.7933.**

2. F1 score: The higher the F1 score, the better the model. The Forest Model has the highest F1 score of **0.8681** from Fig 7a.

3. ROC Curve

   Regarding the inference from the ROC curve for selection of the model, we should choose the model which reaches the top fastest and remains higher (above all) for the most part of the graph. Definitely, AUC will be used as well.

   From the figure below, the ROC curve can be quantified using the AUC. This means that the highest AUC will reaches the top fastest and remains higher (above all) for most part of the graph. The boosted model and the forest model did apparently well here also. Using the AUC score from fig 7a, boosted model took the lead.

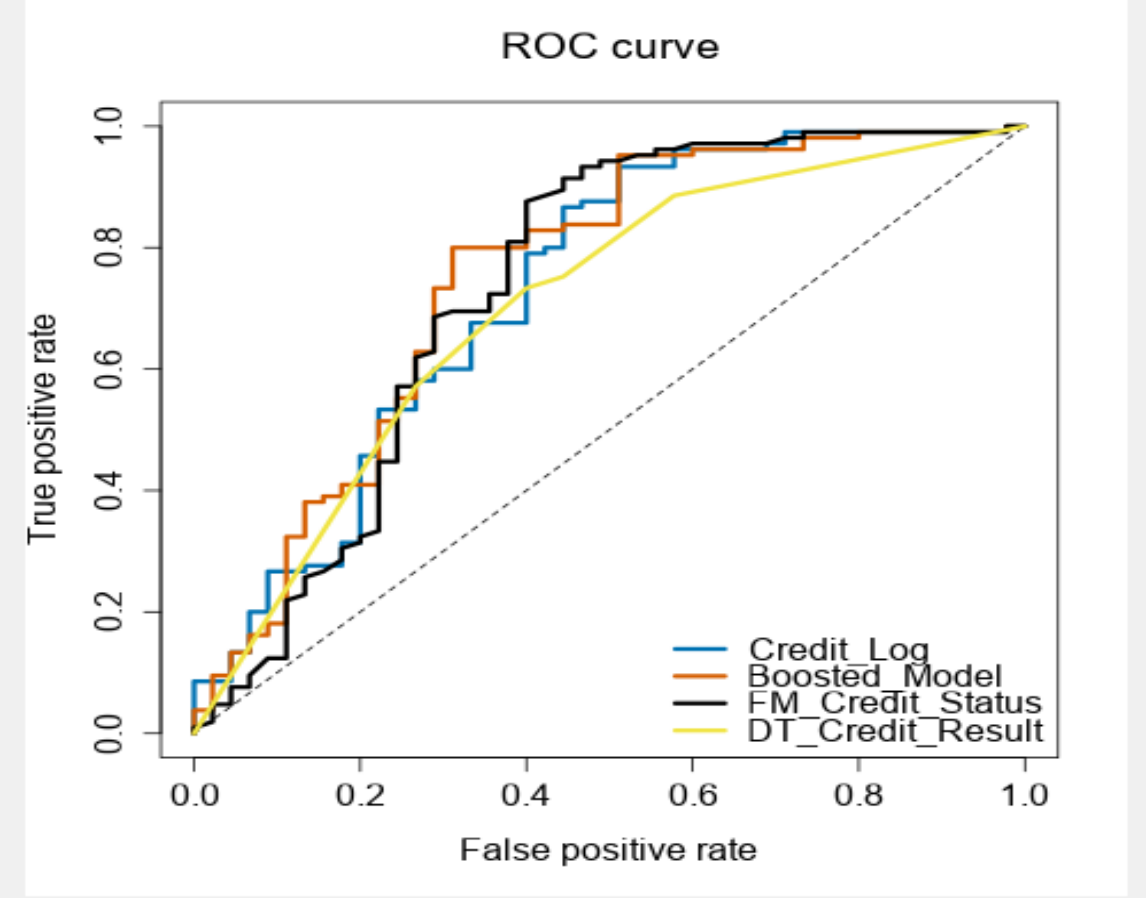


Fig 8.0: the ROC curve

4. Bias in the confusion Matrix

It seems the competition is between forest model and boosted model. So, we will consider the bias in the confusion matrix for both models.

As discussed earlier, for forest model, we have:

PPV= true positives \ (true positives + false positives) =102 / (102+28) =.78

NPV= true negatives\ (true negatives + false negatives) =17/ (17+3) = .85

The difference here is .07

And for boosted model, we have:

PPV= true positives \ (true positives + false positives) =101 / (101+27) =.79

NPV= true negatives\ (true negatives + false negatives) =18/ (18+4) = .82

The difference here is .03

The goal is to select the model with the least difference which is boosted model here.

STEPS IN SCORING THE NEW CUSTOMERS

STEP 1: The boosted model was outputted and saved to a file and a new workflow was opened.

STEP 2: In the new workflow, the input data tool was used to input the outputted file as well as the list of customers to score.

STEP 3: A score tool was then used by connecting the M node to the model and D to the customer. (No configuration required for the score tool).

STEP 4: A formular tool was then used with IF THEN statement to label the score_creditworthy if score_creditworthy is > than score_non-creditworthy and if not, it should label it as 0.

STEP 5: A select tool was used to rename the variable name to credit_application_result.

STEP 6: A filter is applied to credit_application_result not equal to 0 and lastly,

STEP 7: A sort tool is connected to the true side of the filter tool to sort the values in descending order after which a browse tool was added.

The filter tool only helps to see customers to which loans should be given out first. In total, there are 441 applicants out of 500 that the boosted model predicts they won't default on their loans.