Project: Creditworthiness

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?

Ans: In order to first understand the data, we first need to find out what to predict. From the project details we can easily identify analyzing the data of 500 loan applicant and systematically evaluate the creditworthiness of these applicants. We need to train the dataset with different classification models and score the dataset of 500 loan applicants against this training set. We need to find the answer of several important question in order to successfully completed these tasks: Which columns are the most important to keep for the training set? Which columns to drop? What actions to be taken against the missing data remove/impute?

- What data is needed to inform those decisions?

Ans: The columns we need to train the dataset are: Account Balance, Duration of credit month, Payment Status of previous credit, purpose, credit amount, value savings stocks, length of current employment, Installment percent, Most valuable available asset, type of apartment, No of credits at this Bank, Age-years.

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

Ans: Non-Binary model is used when there are more than two decisions to make. Here we need to predict a loan applicant is creditworthy or not that means only two possibilities in this case. Binary model will be a perfect fit for these decisions to be made.

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.***

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)

Note: *For the sake of consistency in the data cleanup process, impute data using the median of the entire data field instead of removing a few data points. (100 word limit)*

Note: *For students using software other than Alteryx, please format each variable as:*

Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

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

Ans: In the cleanup process, we have imputed ***Age-years*** field because of its missing values. We have imputed the missing values with median. Median imputation will work better because it is a number that is already present in the data set and is less susceptible to outlier errors as compared to mean imputation. From Figure 2 we can see, the missing data percentage of ***Age-years*** column.

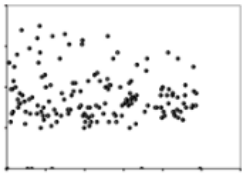
Name	Plot	% Missing	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
Age-years		2.4%	54	19.000	35.637	33.000	75.000	11.502	

Figure 2: Summary of *Age_Years* field

The other fields are removed for below reasons:

- `Duration in current address` should be removed due to too much missing data;
- `Occupation` and `concurrent Credits` should be removed because the entire dataset only has one observation;
- `Foreign workers`, `Guarantors`, and `number of dependencies` should be also removed because of the low variability in the dataset.
- The project description clearly states that the field `Telephone` should be removed since it doesn't contribute to the classification of a creditworthy people.

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

*Answer these questions for **each model** you created:*

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

Ans: Most significant predictor variables are: *Account Balance* and *Credit Amount*.

- 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? (Logistic Regression)

Ans: From figure 3 we can see, overall percent accuracy is 0.78 which is pretty low. The reason behind this is the low rate of *Accuracy_Non-Creditworthy* which is 0.4889. It undermines the high rate *Accuracy_Creditworthy* of 0.9048.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889

Figure 3: Fit and Error Measure of Logistic Regression

Confusion matrix of LR_Credit

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

Figure 4: Confusion Matrix of Logistic Regression

From figure 4 we can see the exact number of Predicted values. When predicting creditworthy the actual creditworthy was 95 and actual non-creditworthy was 23. On the other hand, while predicting non-creditworthy the actual creditworthy was 10 and actual non-creditworthy was 22. Its actual creditworthy number is higher than actual non-creditworthy number which was also indicated through above percentages.

The model is correctly predicting creditworthy individuals at a rate of 0.9048 and correctly predicting non-creditworthy individuals at a rate of 0.4889. It indicates the model is biased towards correctly predicting creditworthy individuals because the rate is way higher than the other.

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

Ans: From Figure 5 we can see, *Credit Amount, Duration of Credit Month, Account Balance* are the most important predictor variables with the value of 23.9, 20.5 and 20.3 respectively

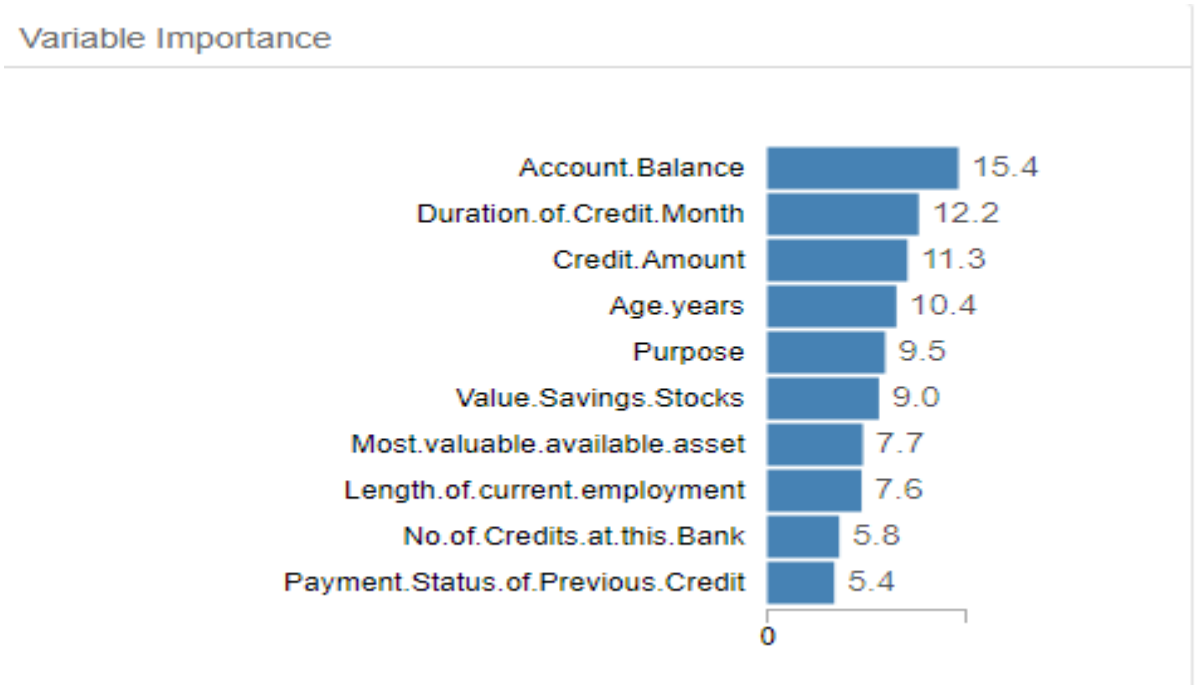


Figure 5: Variable Importance Chart of Decision Tree

- 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? (Decision Tree)

Ans: From figure 5 we can see, the overall accuracy is 0.6933 which is a bit lower. The reason behind this is the low accuracy rate of *Accuracy_Non-Creditworthy* which is 0.3778. The high accuracy rate of *Accuracy_Creditworthy* 0.8286 undermines by this low accuracy rate of non-creditworthy customers.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_Credit	0.6933	0.7909	0.6276	0.8286	0.3778

Figure 6: Fit and Error Measures Decision Tree

Confusion matrix of DT_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	87	28
Predicted_Non-Creditworthy	18	17

Figure 7: Confusion Matrix table of Decision Tree

From figure 7 we can see the exact number of Predicted values. When predicting creditworthy the actual creditworthy was 87 and actual non-creditworthy was 28. On the other hand, while predicting non-creditworthy the actual creditworthy was 18 and actual non-creditworthy was 17. Its actual creditworthy number is higher than actual non-creditworthy number which was also indicated through above percentages.

The model is correctly predicting creditworthy individuals at a rate of 0.8286 and correctly predicting non-creditworthy individuals at a rate of 0.3778. It indicates the model is biased towards correctly predicting creditworthy individuals because the rate is way higher than the other.

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

Ans: From the variable importance chart we can see the most important and significant variables are Credit Amount and Age-years. Followed by the other variables are: Duration of credit month, Purpose, Payment status of previous credit, Most valuable available asset etc.

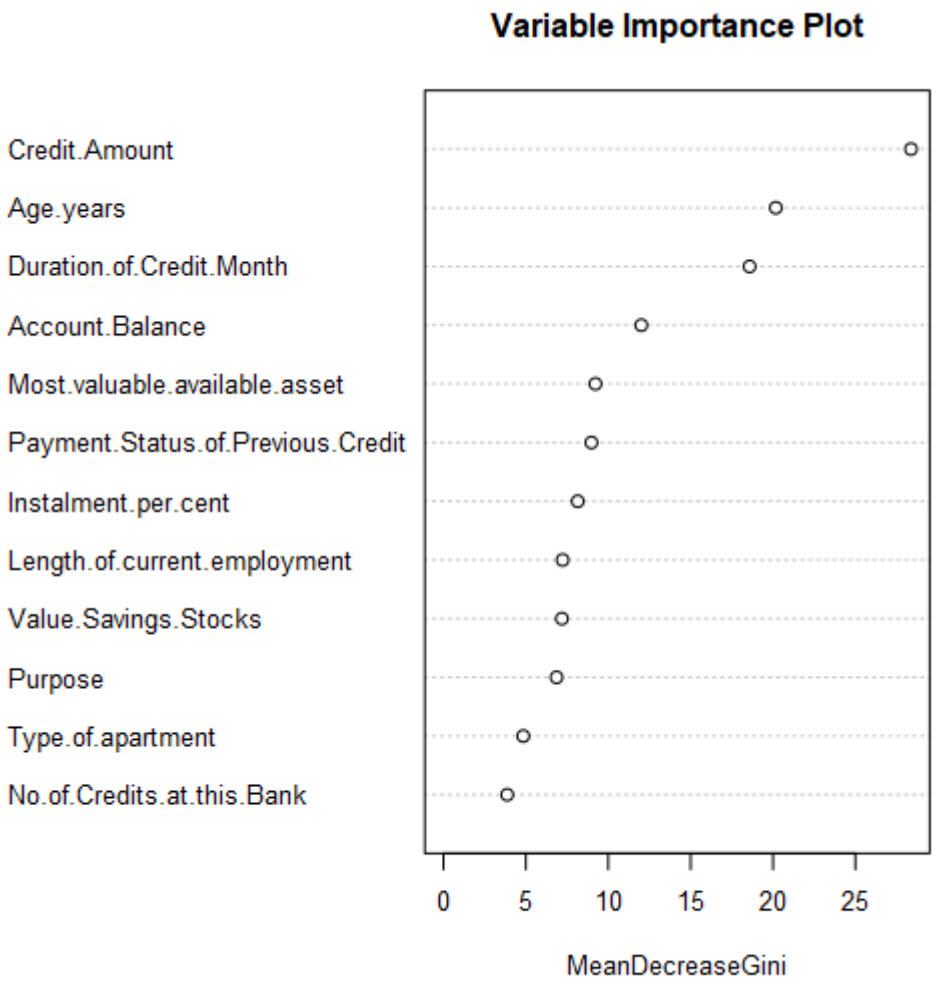


Figure 8: Variable Importance Chart (Forest Model)

- 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? (Forest Model)

Ans: From figure 9 we can see, the overall percent accuracy is 0.8. Which is pretty low. *Accuracy_Creditworthy* is a bit high 0.9619 but the *Accuracy_Non-Creditworthy* is very much low 0.4222. This low accuracy rate affected the overall accuracy rate.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_Credit	0.8000	0.8707	0.7361	0.9619	0.4222

Figure 9: Fit and Error Measures Forest Model

Confusion matrix of FM_Credit		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Figure 10: Confusion Matrix of FM_Credit

From figure 10 we can see the exact number of Predicted values. When predicting creditworthy the actual creditworthy was 101 and actual non-creditworthy was 26. On the other hand, while predicting non-creditworthy the actual creditworthy was 4 and actual non-creditworthy was 19. Its actual creditworthy number is higher than actual non-creditworthy number which was also indicated through above percentages.
The model is correctly predicting creditworthy individuals at a rate of 0.9619 and correctly predicting non-creditworthy individuals at a rate of 0.4222. It indicates the model is biased towards correctly predicting creditworthy individuals because the rate is way higher than the other.

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

Ans: From the variable importance chart we can see the most important and significant variables are Account Balance and Credit Amounts. Followed by the other variables are: Duration of credit month, Purpose, Payment status of previous credit, Age_years, Most valuable available asset etc.

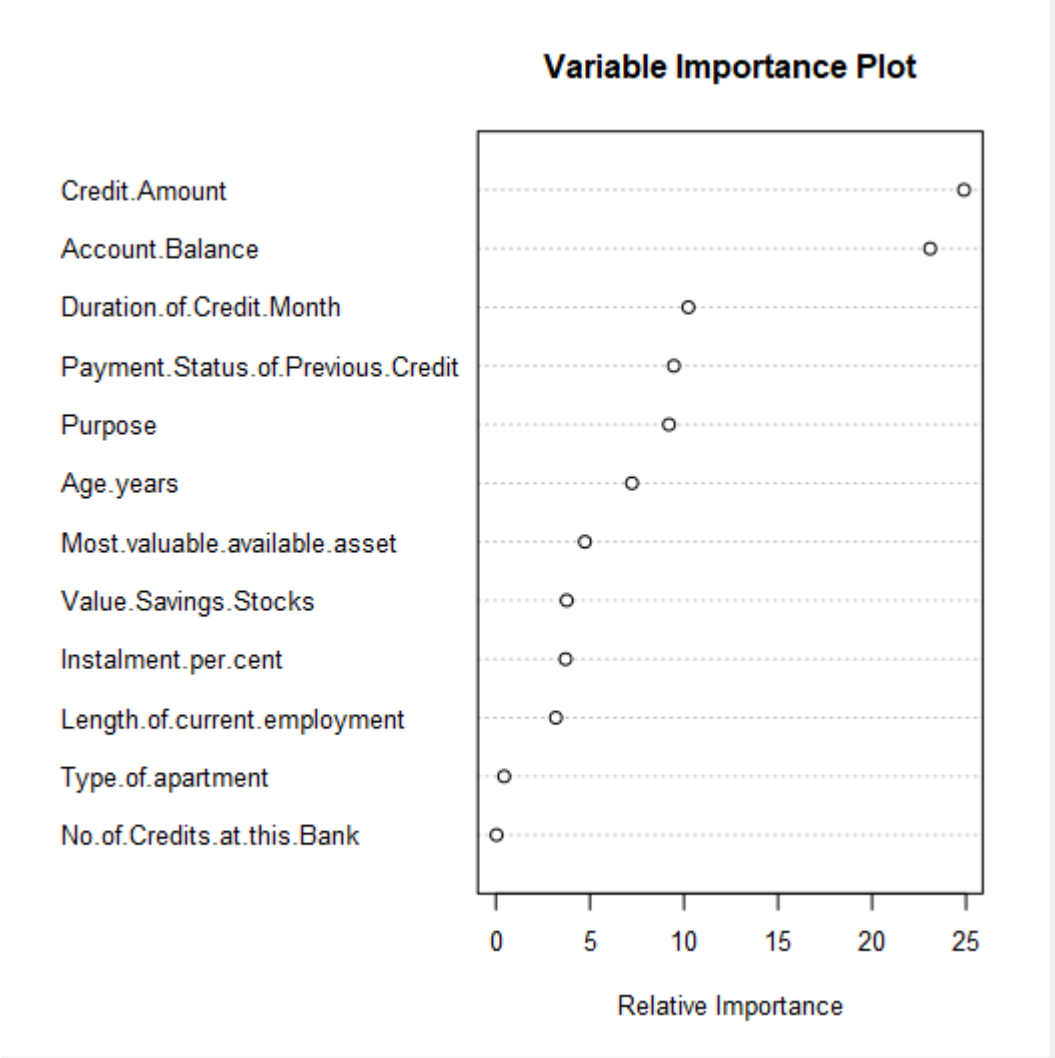


Figure 11: Variable Importance Chart (Boosted Model)

- 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? (Boosted Model)

Ans: After validating the Boosted Model, the overall accuracy rate is 0.7867. The rate is pretty low despite high accuracy rate (0.9619) of *Accuracy_Creditworthy.* The reason behind this is the low accuracy rate of *Accuracy_Non-Creditworthy* which is 0.3778.
The model is correctly predicting creditworthy individuals at a rate of 0.9619 and correctly predicting non-creditworthy individuals at a rate of 0.3778. It indicates the model is biased towards correctly predicting creditworthy individuals because the rate is way higher than the other.

Fit and error measures

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

Figure 12: Fit and Error measures Boosted Model

Confusion matrix of BM_Credit

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

Figure 13: Confusion Matrix Boosted Model

From figure 13 we can see the exact number of Predicted values. When predicting creditworthy the actual creditworthy was 101 and actual non-creditworthy was 28. On the other hand, while predicting non-creditworthy the actual creditworthy was 4 and actual non-creditworthy was 17. Its actual creditworthy number is higher than actual non-creditworthy number which was also indicated through above percentages.
The model is correctly predicting creditworthy individuals at a rate of 0.9619 and correctly predicting non-creditworthy individuals at a rate of 0.3778. It indicates the model is biased towards correctly predicting creditworthy individuals because the rate is way higher than the other.

You should have four sets of questions answered. (500 word limit)

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)

Answer these questions:

- 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:
    - Overall Accuracy against your Validation set
    - Accuracies within "Creditworthy" and "Non-Creditworthy" segments

- ROC graph
- Bias in the Confusion Matrices

Ans: Forest Model is more accurate than others if we compare all the modeling techniques. It's overall accuracy is 0.8000 is the highest of all. We can compare the figures from the figure:14.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
LR_Credit	0.7800	0.8520	0.7314	0.9048	0.4889
DT_Credit	0.7200	0.8056	0.6678	0.8286	0.4667
FM_Credit	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Credit	0.7867	0.8632	0.7524	0.9619	0.3778

Figure 14: Accuracy Rates Comparison Between Different Model

From figure 14, we can also compare the accuracy rates within Creditworthy and Non-Creditworthy segment. *Accuracy_Creditworhty* rate is highest for the Forest Model but *Accuracy_Non-Credtworthy* is not the highest. However, if we compare the overall accuracy rate Forest Model is the best performer.

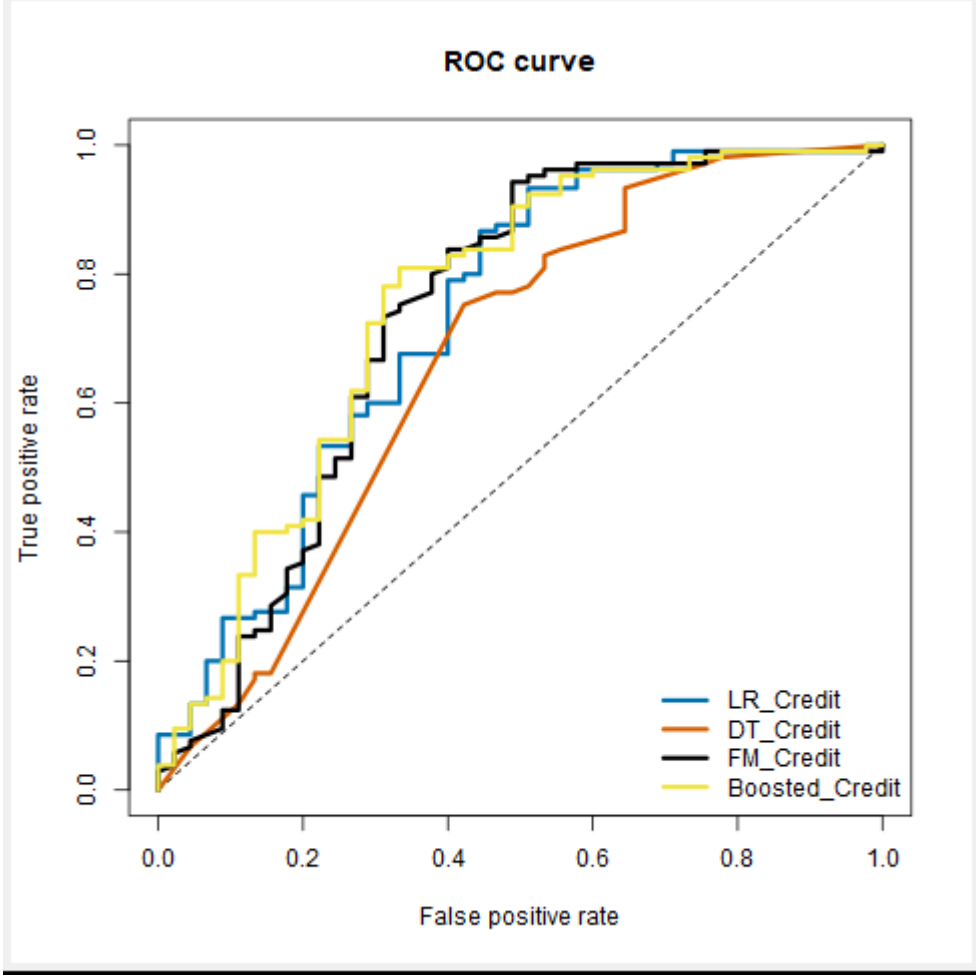


Figure 15: Comparative ROC Curve of different model

From the ROC curve of figure 15 we can see, the forest model represented through the black line is which maintains a good distance from the dotted line. It means, the classifier is doing a good job separating different classes. In order to use the quantify the performance of a classifier, we can use AUC (Area Under the Curve) which indicates the strength of the classifier. From figure 14 we can see, the AUC of Forest Model is 0.7361 which is second best behind Boosted Model still very much high. A very poor classifier has an AUC of around 0.5.

In case of bias, all the models are biased towards accurately predicting Creditworthy individuals than accurately predicting non-creditworthy individuals.

- How many individuals are creditworthy?

Ans: There are 406 individuals who are creditworthy.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.