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

1. What decisions needs to be made?
   The main decision to be made is determining whether or not a customer is creditworthy to give loan to. To determine who is creditworthy, the analyst could make use of a model and historical data.

2. What data is needed to inform those decisions?
   The company would need to have historical data of the previous applicants and their creditworthiness. The company would also need the same fields in the new applicants. Most importantly the data should be adequate and relevant to carry out the analysis.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
   Since we would need to determine if a customer is creditworthy or not, a binary model should be the kind of model that would produce the best result.

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 average 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:*

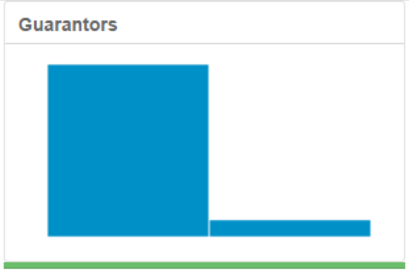
Variable	**Data Type**
Credit-Application-Result	String
Account-Balance	String
Duration-of-Credit-Month	Double
Payment-Status-of-Previous-Credit	String
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	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

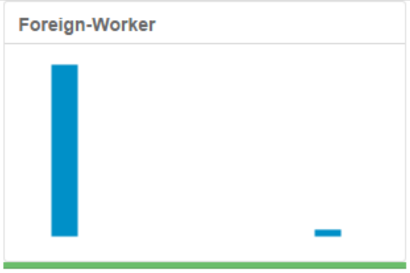
To achieve consistent results reviewers expect.

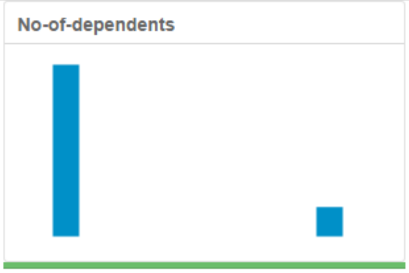
Answer this question:

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

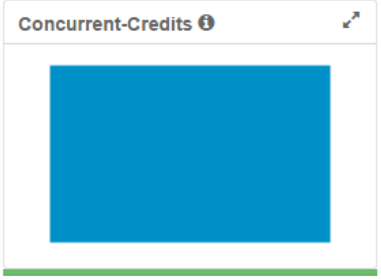
The first fields I decided to remove are ‘guarantors’, ‘foreign worker’, ‘no of dependents’ as these fields have low variability.



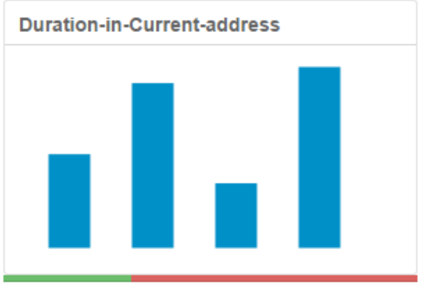


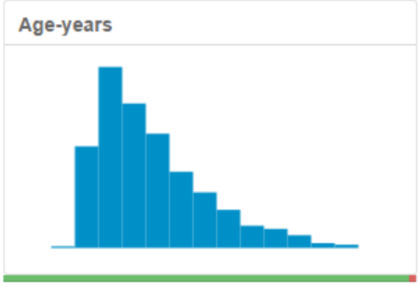


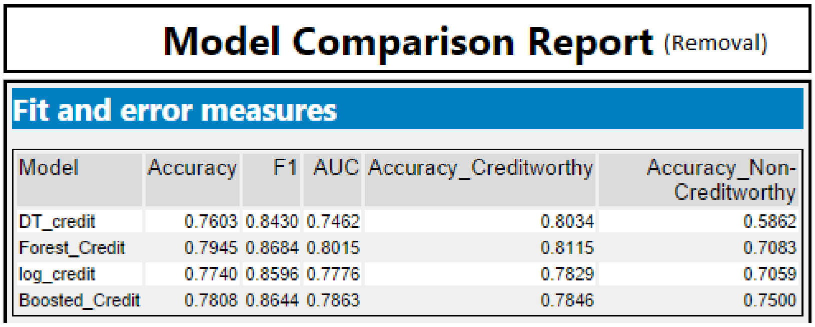
Then ‘Concurrent credits’ and ‘occupation’ were also excluded for having only one type of data. ‘Telephone’ field was removed as there is no logical reasoning for including the variable.



I removed the field ‘duration in current address’ as it has 69% missing data and imputation would skew data to one value. ‘Age years’ field on the other hand, had only 2% missing data and so imputing it is a better idea. The average imputation produce better end results.



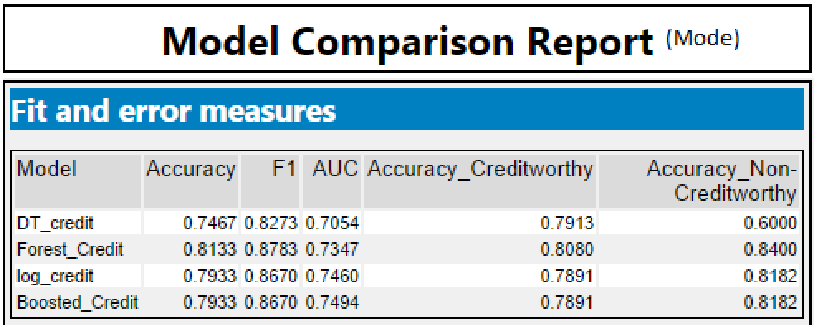




Model Comparison Report (Removal)

Fit and error measures

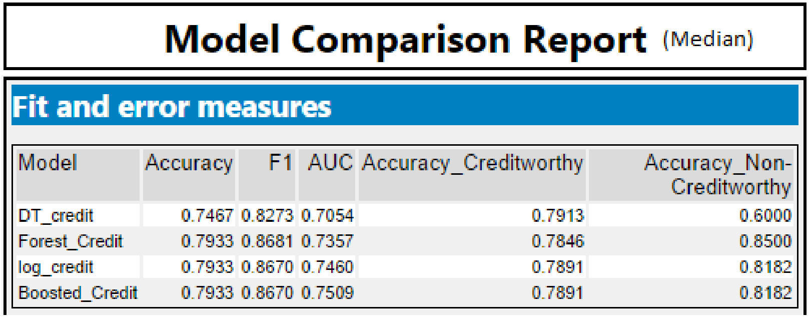
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7603	0.8430	0.7462	0.8034	0.5862
Forest_Credit	0.7945	0.8684	0.8015	0.8115	0.7083
log_credit	0.7740	0.8596	0.7776	0.7829	0.7059
Boosted_Credit	0.7808	0.8644	0.7863	0.7846	0.7500



Model Comparison Report (Mode)

Fit and error measures

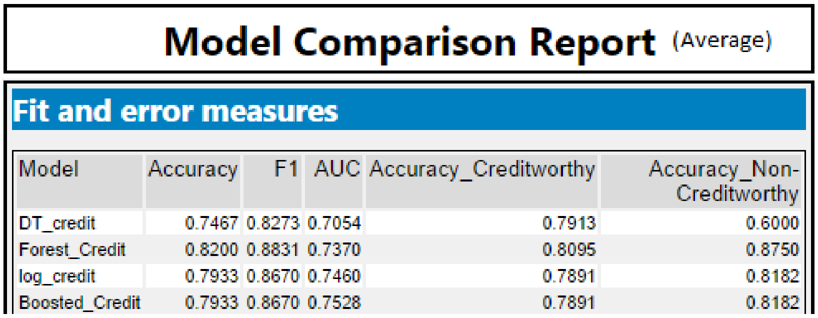
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Credit	0.8133	0.8783	0.7347	0.8080	0.8400
log_credit	0.7933	0.8670	0.7460	0.7891	0.8182
Boosted_Credit	0.7933	0.8670	0.7494	0.7891	0.8182



Model Comparison Report (Median)

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Credit	0.7933	0.8681	0.7357	0.7846	0.8500
log_credit	0.7933	0.8670	0.7460	0.7891	0.8182
Boosted_Credit	0.7933	0.8670	0.7509	0.7891	0.8182



Model Comparison Report (Average)

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Credit	0.8200	0.8831	0.7370	0.8095	0.8750
log_credit	0.7933	0.8670	0.7460	0.7891	0.8182
Boosted_Credit	0.7933	0.8670	0.7528	0.7891	0.8182

Step 3: Train your Classification Models

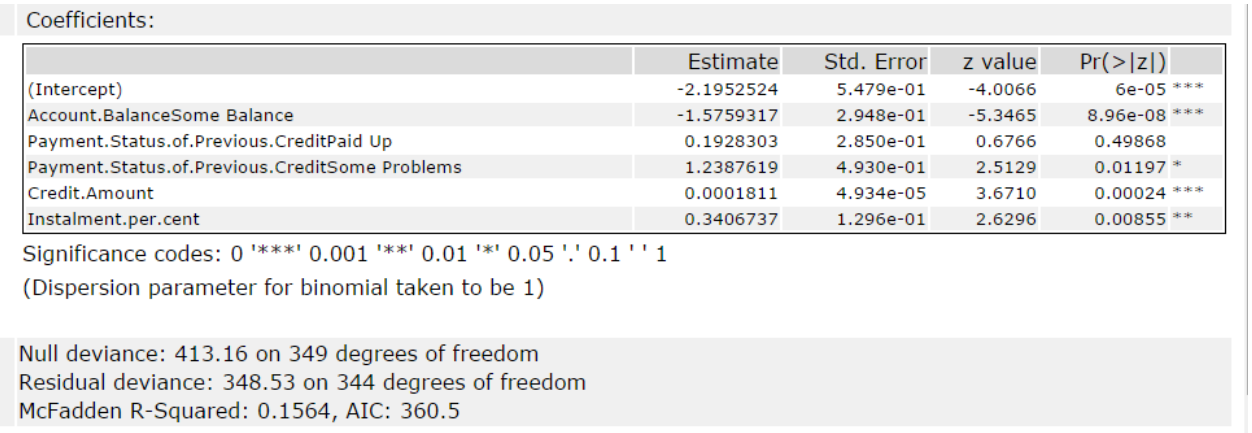
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1. You should have four sets of questions answered. (500 word limit)

Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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

A. Logistic Regression

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



Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1952524	5.479e-01	-4.0066	6e-05 ***
Account.BalanceSome Balance	-1.5759317	2.948e-01	-5.3465	8.96e-08 ***
Payment.Status.of.Previous.CreditPaid Up	0.1928303	2.850e-01	0.6766	0.49868
Payment.Status.of.Previous.CreditSome Problems	1.2387619	4.930e-01	2.5129	0.01197 *
Credit.Amount	0.0001811	4.934e-05	3.6710	0.00024 ***
Instalment.per.cent	0.3406737	1.296e-01	2.6296	0.00855 **

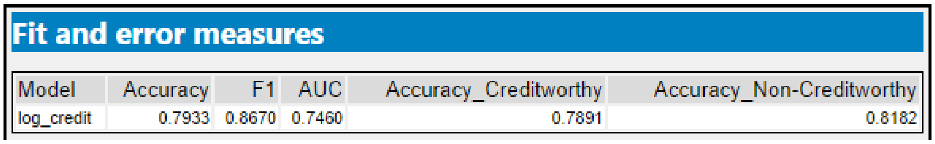
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 348.53 on 344 degrees of freedom
McFadden R-Squared: 0.1564, AIC: 360.5

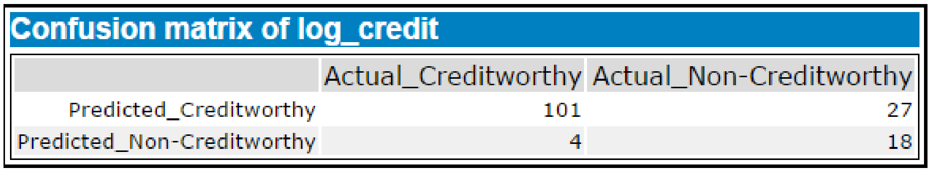
All variables listed in the table are statistically significant based on the p-values except for ‘payment.status.of.previous.creditpaid.up’ as it has a p-value greater than 0.05. However for this model, the lowest p-values which are denoted by 3(*)’s are ‘account balance Some balance’ and ‘Credit amount’ followed by ‘Installment per cent’ with 2(*)’s.

2. 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?



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
log_credit	0.7933	0.8670	0.7460	0.7891	0.8182



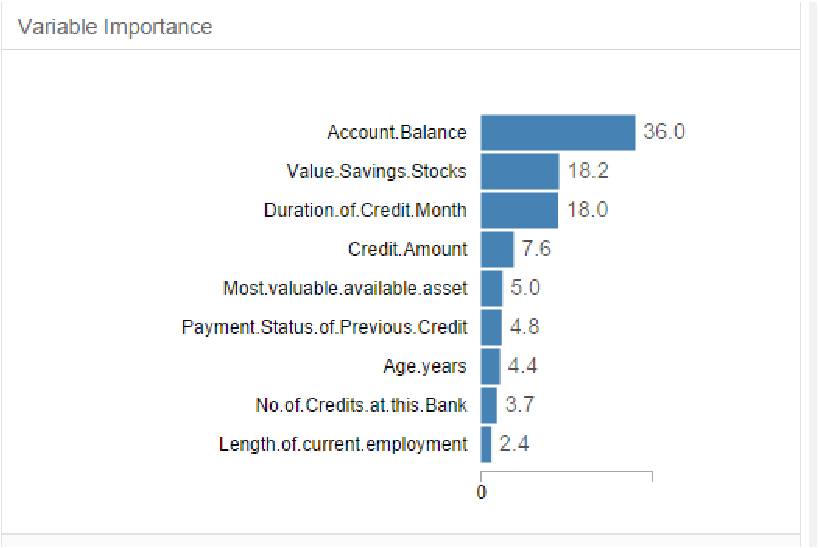
Confusion matrix of log_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

The overall accuracy is shown to be 0.7933. We can see that there is a slight bias towards the non-creditworthy prediction.

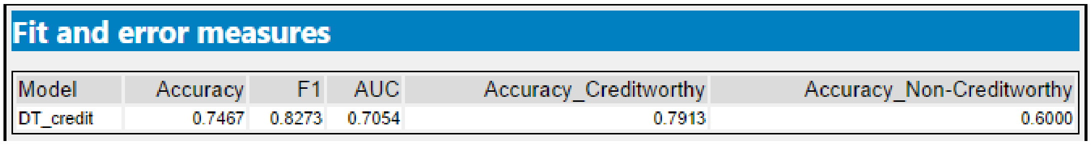
B. Decision Tree

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



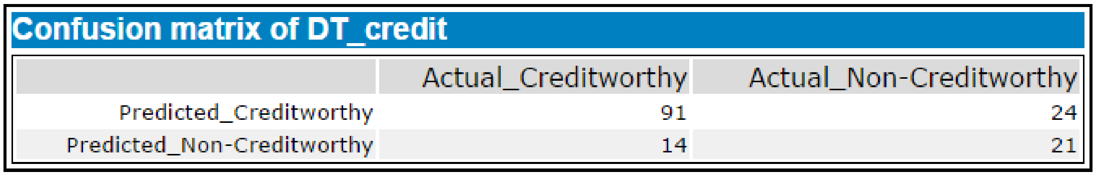
As seen from the variable importance diagram, there are 3-4 variables that are most important. ‘Account balance’ is notably the most important variable with a 36% followed by ‘value savings stocks’ at 18.2% then ‘duration of credit month’ at 18.0%. ‘Credit amount’ still stands out as an important variable with a 7.6%.

4. 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?



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.7913	0.6000



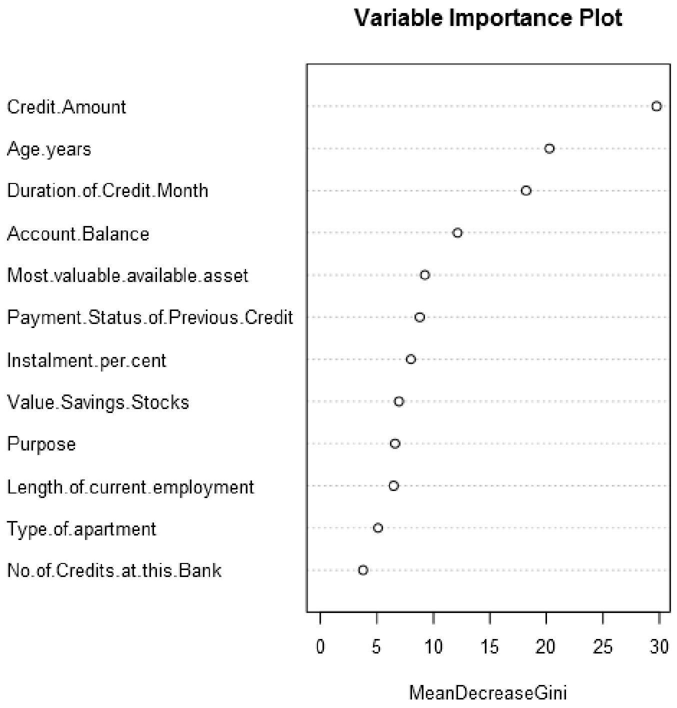
Confusion matrix of DT_credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

This model has an overall accuracy of 0.7467 and is highly biased towards creditworthy.

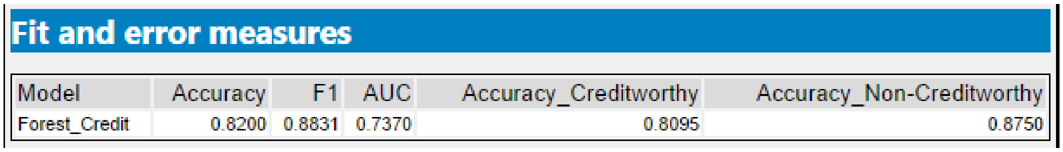
C. Forest Model

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



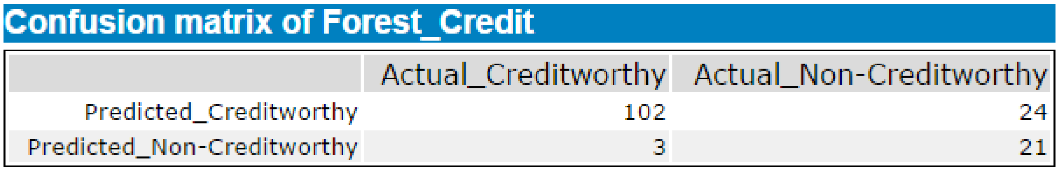
In this model, the most significant variable is ‘credit amount’ followed by ‘age years’ then ‘duration of credit month’ and lastly ‘account balance’

6. 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?



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Forest_Credit	0.8200	0.8831	0.7370	0.8095	0.8750



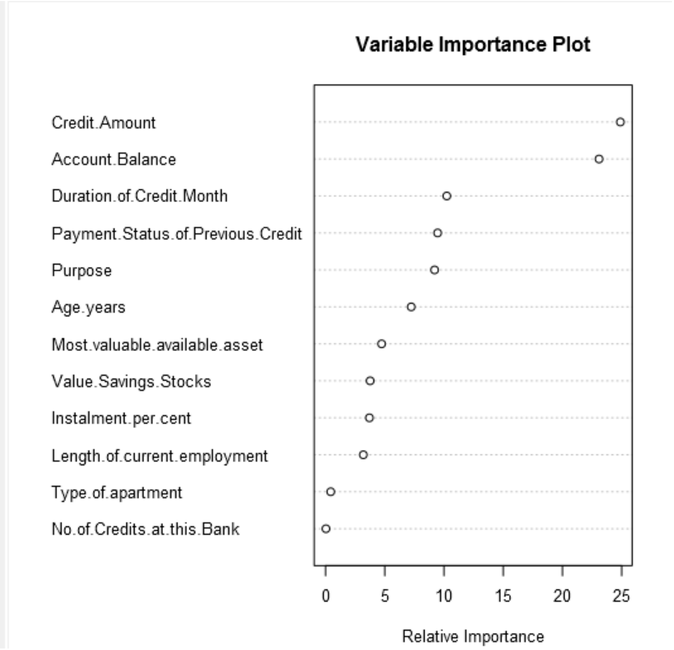
Confusion matrix of Forest_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

The overall accuracy of this model is 0.8200 and it shows bias towards non-creditworthy but still the gap is not as much as decision tree.

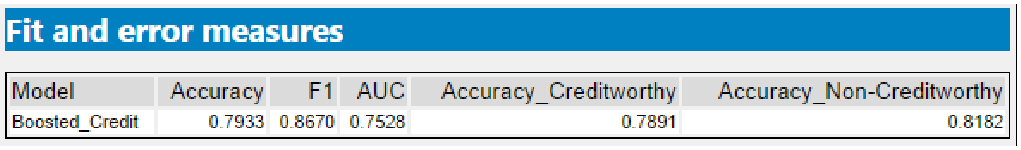
D. Boosted Model

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



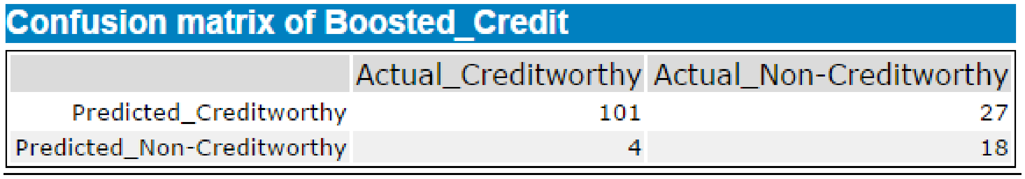
The model shows 2 variables as the most significant. 'Credit Amount' being the highest followed by 'account balance'. Starting from 'duration of credit month' the importance of each variable starts reducing.

8. 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?



Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Credit	0.7933	0.8670	0.7528	0.7891	0.8182



Confusion matrix of Boosted_Credit

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

The overall accuracy of this model is 0.7933 just like the logistic regression. There is still a slight bias towards the non-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"

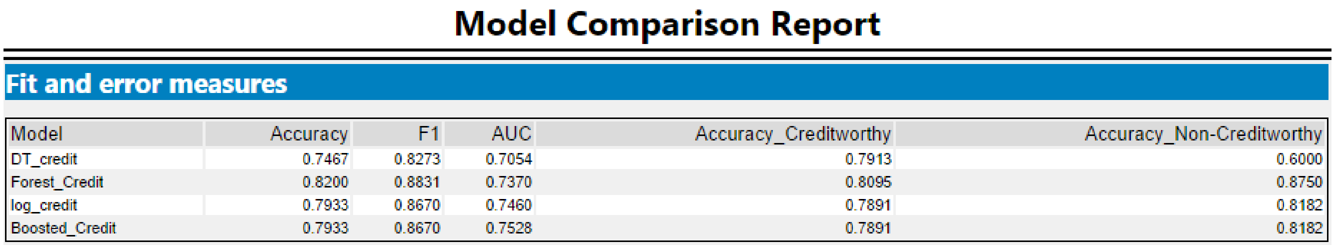
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

Answer these questions:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within “Creditworthy” and “Non-Creditworthy” segments
    c. ROC graph
    d. Bias in the Confusion Matrices

Note: Remember that your boss only cares about prediction accuracy for Credityworth and Non-Creditworthy segments.

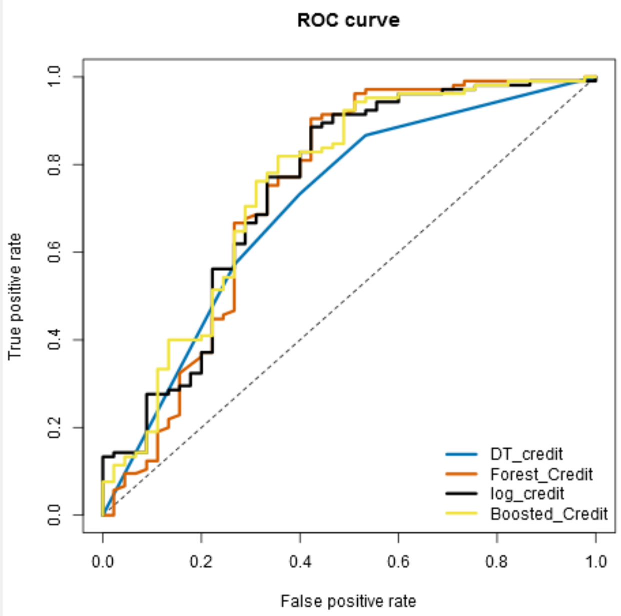
My conclusion is that forest model is the best model to be used in this problem. In terms of overall accuracy, the forest model is the most accurate. In both creditworthy and non-creditworthy segments, the forest model does better than other models. The Creditworthy segment has 80.95% accuracy while the Non-Creditworthy segment has the accuracy of 87.50%. Despite that, the forest model has some significant bias towards the non-creditworthy unlike logistic regression and boosted model as these two models have the least amount of bias but with lower accuracies. Now looking at the ROC Graph, neither of the models have a very high true positive rate. However, the boosted model is shown to have the least number of false positives. AUC shows that the forest model is only third best at determining the true positive rate with boosted model being number one followed by the logistic regression. In the end, I will still stick to the forest model as the AUC gap is not that high.



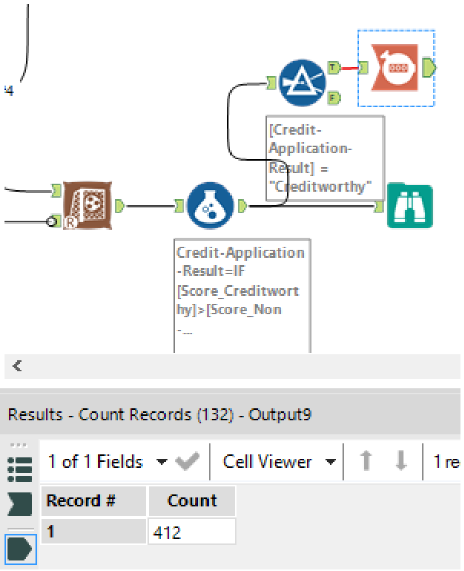
Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_credit	0.7467	0.8273	0.7054	0.7913	0.6000
Forest_Credit	0.8200	0.8831	0.7370	0.8095	0.8750
log_credit	0.7933	0.8670	0.7460	0.7891	0.8182
Boosted_Credit	0.7933	0.8670	0.7528	0.7891	0.8182



2. How many individuals are creditworthy?
Based on the selected model, there are 412 applicants that are creditworthy. The other 88 are not creditworthy.


[Credit-Application-Result] = "Creditworthy"
Credit-Application-Result=IF [Score_Creditworthy]>[Score_Non-...
1 of 1 Fields
Cell Viewer
1 re

Results - Count Records (132) - Output9

Record #	Count
1	412

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.