ONYEKABA NZUBECHUKWU JUDE

Project: Creditworthiness

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/11a7bf4c-2b69-47f3-9aec-108ce847f855/project

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

- What decisions needs to be made?

  The decision to make is whether customers who applied for loan are creditworthy to be extended one or not.

- What data is needed to inform those decisions?

  Data on past applications such as:

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

- No of dependents
- Telephone
- Foreign Worker, are required in order to inform those decisions

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

  Binary classification models such as:
- logistics regression
- decision tree
- forest model
- boosted tree, will be used to analyze and determine creditworthy customers

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 words 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-	String

Credit	
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:

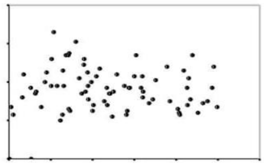
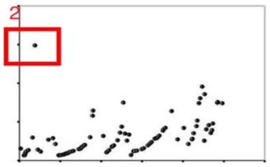
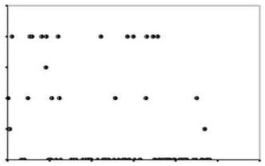
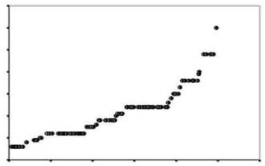
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

*My interactive output of field summary isn’t working. Checked, and it happened that it was a general bug in this version (2020.2) of Alteryx. I made use of the result end of the field summary tool

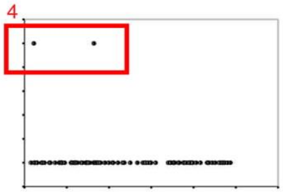
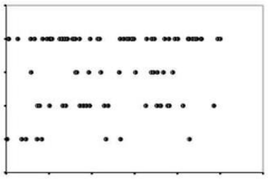
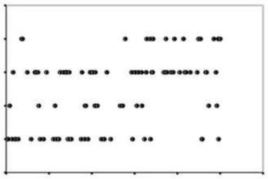
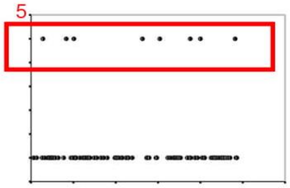
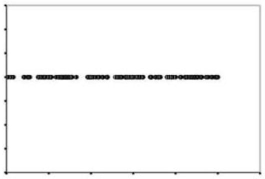
Record Report

1

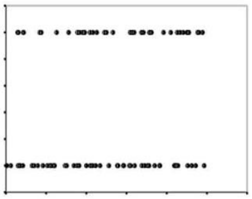
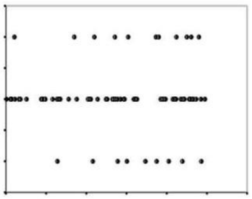
Numeric Fields

Name	Plot	1 % 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	
Credit-Amount		0.0%	464	276.000	3,199.980	2,236.500	2 18,424.000	2,831.387	
Duration-in-Current-address		3 68.8%	5	1.000	2.660	2.000	4.000	1.150	This field has over 10% missing values. Consider imputing these values. This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Duration-of-Credit-Month		0.0%	30	4.000	21.434	18.000	60.000	12.307	

Record Report

Name	Plot	% Missing	Unique Values 4	Min	Mean	Median	Max 4	Std Dev	Remarks
Foreign-Worker		0.0%	2	1.000	1.038	1.000	2.000	0.191	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Instalment-per-cent		0.0%	4	1.000	3.010	3.000	4.000	1.114	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Most-valuable-available-asset		0.0%	4	1.000	2.360	3.000	4.000	1.064	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
No-of-dependents		0.0%	5 2	1.000	1.146	1.000	2.000	0.353	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Occupation		0.0%	6 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".

Record Report

Name	Plot	% Missing	Unique Values 7	Min	Mean	Median	Max	Std Dev	Remarks
Telephone		0.0%	2	1.000	1.400	1.000	2.000	0.490	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Type-of-apartment		0.0%	3	1.000	1.928	2.000	3.000	0.540	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

2 String/Character Fields

Name	% Missing	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Account-Balance	0.0%	8 2	No Account	Some Balance	8 238	262	
Concurrent-Credits	0.0%	9 1	Other Banks/Depts	Other Banks/Depts	500	500	
Credit-Application-Result	0.0%	10 2	Creditworthy	Non-Creditworthy	10 142	358	
Guarantors	0.0%	11 2	Yes	None	11 43	457	
Length-of-current-employment	0.0%	3	< 1yr	1-4 yrs	97	279	
No-of-Credits-at-this-Bank	0.0%	12 2		More than 1	12 180	320	
Payment-Status-of-Previous-Credit	0.0%	3	Paid Up	No Problems (in this bank)	36	260	
Purpose	0.0%	4	Other	Home Related	15	355	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Value-Savings-Stocks	0.0%	3	None	£100-£1000	48	298	
Name	**% Missing**	**Unique Values**	**Shortest Value**	**Longest Value**	**Min Value Count**	**Max Value Count**	**Remarks**

Using the report output from the browse tool, I inspected the variables.

1. The numerical variables – age (years) had missing values of 2.4%. Since this wasn't much, I decided to fill it with the imputation tool. I inspected the distribution, and it was skewed to the right, so I filled with median. This is because median is least affected by skewness as compared to mode and mean.

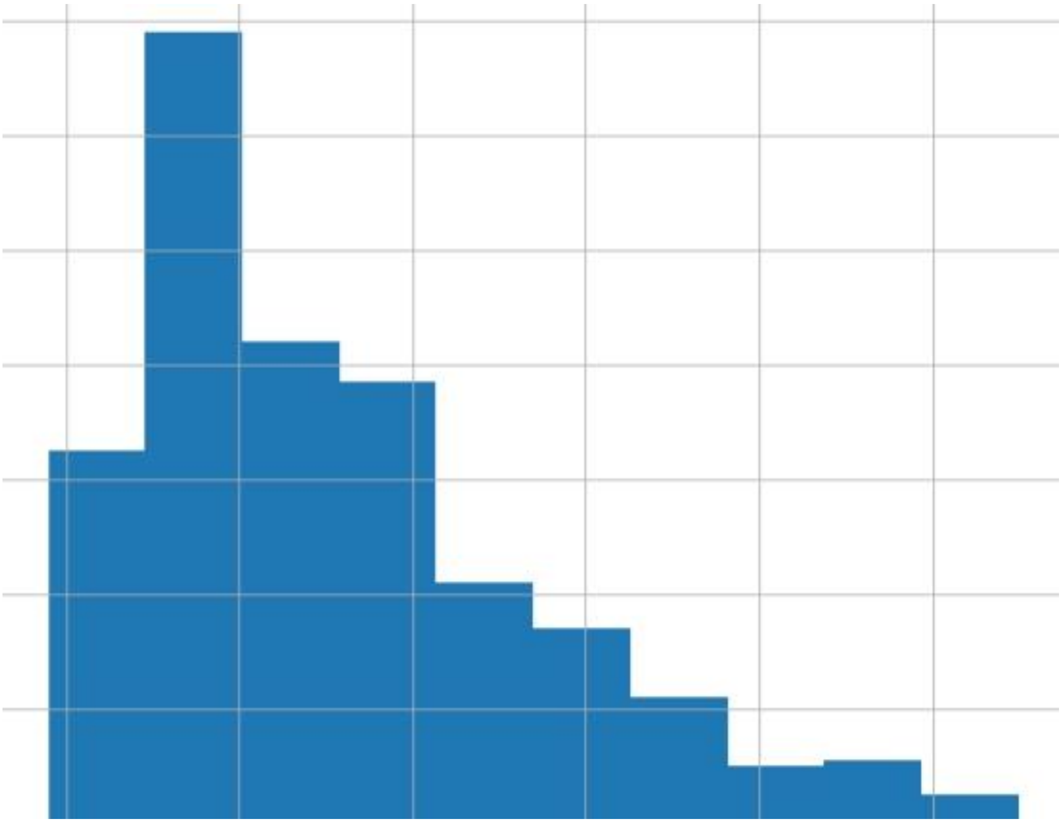


Fig 1: Distribution of Age-years

2. The numerical variable - credit-amount has an outlier. The outlier was 18,424,000. I decided to remove this outlier.
3. The numerical variable - duration in the current address has missing values as high as 68.8%. This was too much, so I decided to drop this column entirely.
4. The numerical variable - foreign worker upon interrogation was actually a categorical variable, with 1 and 2 as categories. From the report, there are only 2 categories with category 1 many times greater than category 2, so I decided to drop this column due to low variability. In this same way,
5. The numerical variable - number of dependent variable was a categorical variable with 1 and 2 as categories. I dropped this column because of low variability.
6. Occupation was also dropped because of low variability

7. Although telephone had two categories 1 and 2 and wasn’t skewed it had no logical correlation with Creditworthiness, so I dropped this column
8. Same for account-balance, two categories but normally distributed.
9. I dropped concurrent-credit, low variability, just one category, it won’t affect our model.
10. Credit-application-result was a close call. I kept it because the distribution wasn’t skewed too much.
11. The categorical variable guarantors was dropped because distribution was skewed to one category
12. No-of-credits at the bank was a close call. I decided to keep it for the same reason as 10.

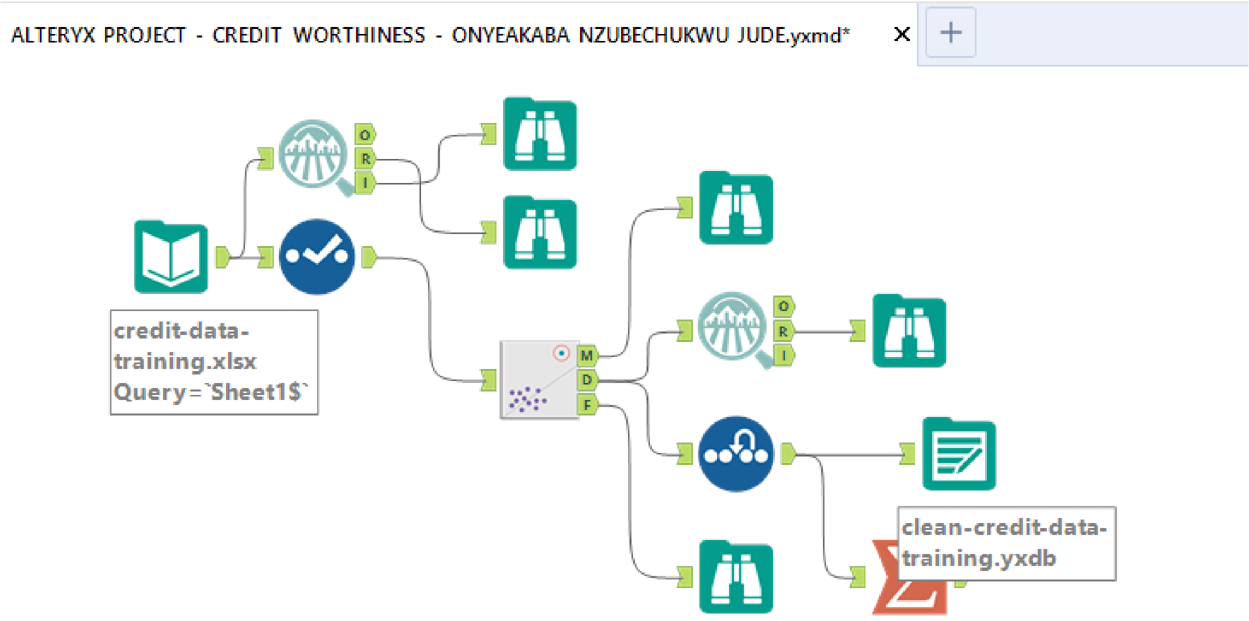

ALTERYX PROJECT - CREDIT WORTHINESS - ONYEAKABA NZUBECHUKWU JUDE.yxmd*
×
+
O
R
I
credit-data-training.xlsx
Query=`Sheet1$`
M
D
F
O
R
I
clean-credit-data-training.yxdb

Fig 2: Workflow to clean and prepare the training data

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

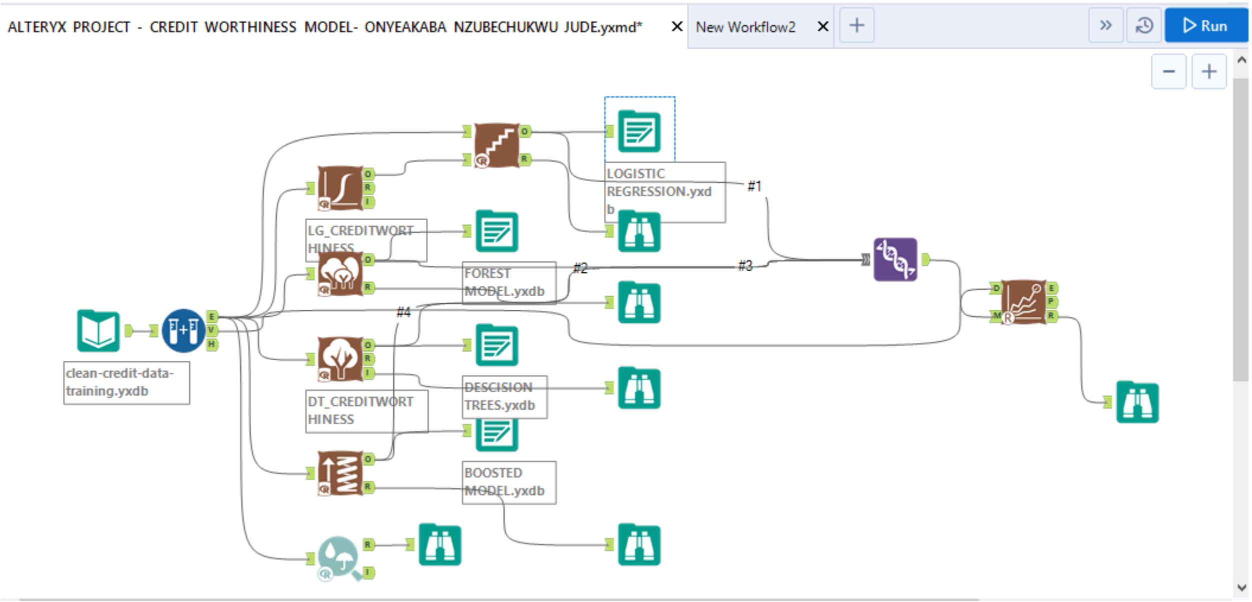
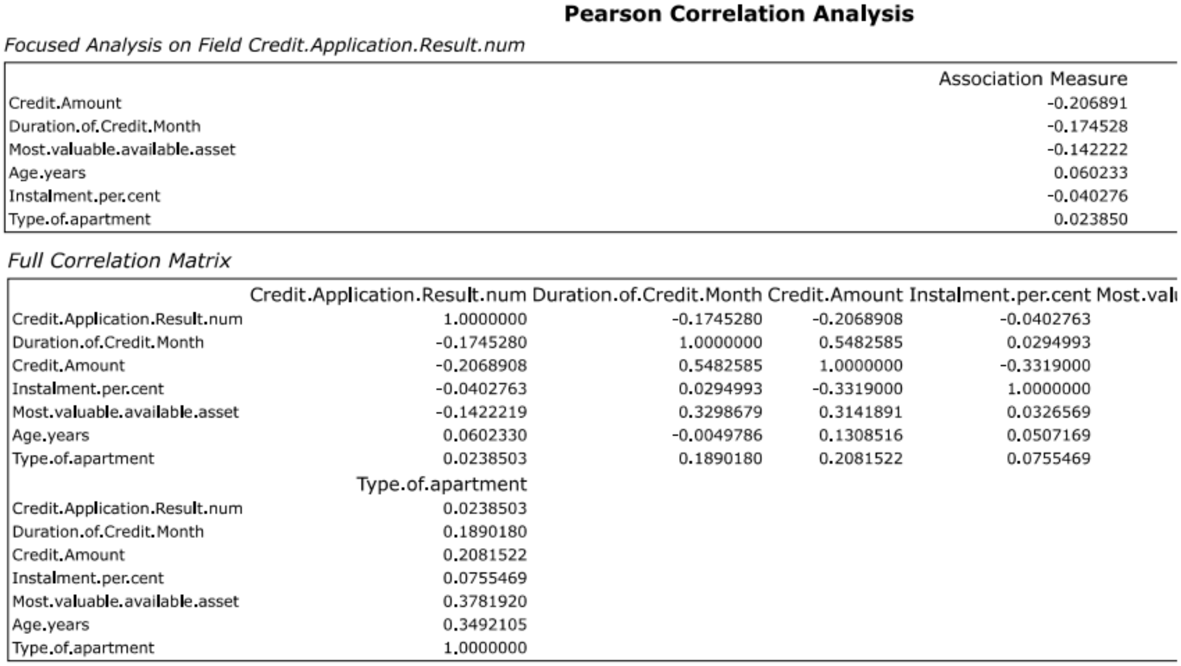


Fig 3: Workflow of model

*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.

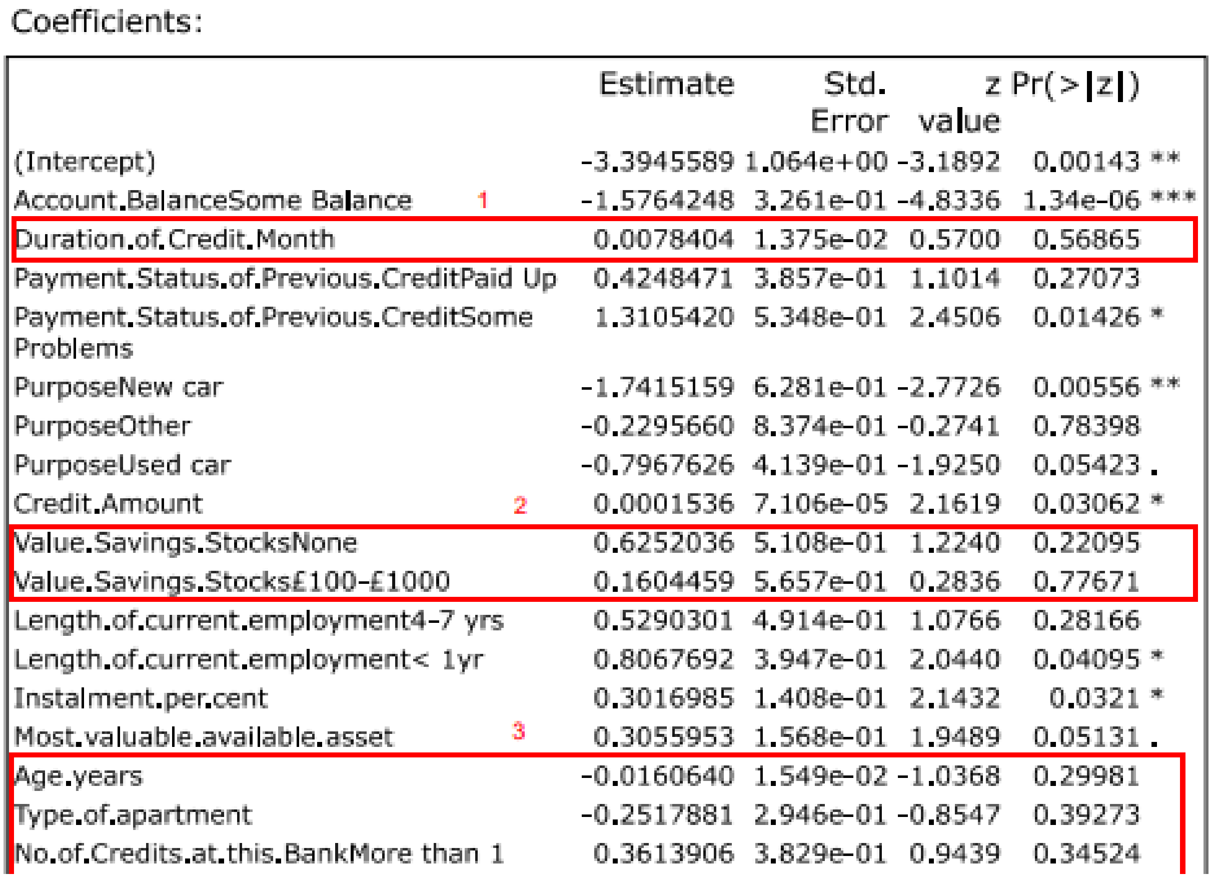


```
                              Pearson Correlation Analysis
Focused Analysis on Field Credit.Application.Result.num

                                                                          Association Measure
Credit.Amount                                                                       -0.206891
Duration.of.Credit.Month                                                            -0.174528
Most.valuable.available.asset                                                       -0.142222
Age.years                                                                            0.060233
Instalment.per.cent                                                                 -0.040276
Type.of.apartment                                                                    0.023850

Full Correlation Matrix
                              Credit.Application.Result.num Duration.of.Credit.Month Credit.Amount Instalment.per.cent Most.valu
Credit.Application.Result.num                     1.0000000               -0.1745280    -0.2068908          -0.0402763
Duration.of.Credit.Month                         -0.1745280                1.0000000     0.5482585           0.0294993
Credit.Amount                                    -0.2068908                0.5482585     1.0000000          -0.3319000
Instalment.per.cent                              -0.0402763                0.0294993    -0.3319000           1.0000000
Most.valuable.available.asset                    -0.1422219                0.3298679     0.3141891           0.0326569
Age.years                                         0.0602330               -0.0049786     0.1308516           0.0507169
Type.of.apartment                                 0.0238503                0.1890180     0.2081522           0.0755469
                                      Type.of.apartment
Credit.Application.Result.num                 0.0238503
Duration.of.Credit.Month                      0.1890180
Credit.Amount                                 0.2081522
Instalment.per.cent                           0.0755469
Most.valuable.available.asset                 0.3781920
Age.years                                     0.3492105
Type.of.apartment                             1.0000000
```

Fig 4: correlation matrix (association analysis)

1. Logistic

A. Without stepwise



Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.3945589	1.064e+00	-3.1892	0.00143	**
Account.BalanceSome Balance 1	-1.5764248	3.261e-01	-4.8336	1.34e-06	***
Duration.of.Credit.Month	0.0078404	1.375e-02	0.5700	0.56865	
Payment.Status.of.Previous.CreditPaid Up	0.4248471	3.857e-01	1.1014	0.27073	
Payment.Status.of.Previous.CreditSome Problems	1.3105420	5.348e-01	2.4506	0.01426	*
PurposeNew car	-1.7415159	6.281e-01	-2.7726	0.00556	**
PurposeOther	-0.2295660	8.374e-01	-0.2741	0.78398	
PurposeUsed car	-0.7967626	4.139e-01	-1.9250	0.05423	.
Credit.Amount 2	0.0001536	7.106e-05	2.1619	0.03062	*
Value.Savings.StocksNone	0.6252036	5.108e-01	1.2240	0.22095	
Value.Savings.Stocks£100-£1000	0.1604459	5.657e-01	0.2836	0.77671	
Length.of.current.employment4-7 yrs	0.5290301	4.914e-01	1.0766	0.28166	
Length.of.current.employment< 1yr	0.8067692	3.947e-01	2.0440	0.04095	*
Instalment.per.cent	0.3016985	1.408e-01	2.1432	0.0321	*
Most.valuable.available.asset 3	0.3055953	1.568e-01	1.9489	0.05131	.
Age.years	-0.0160640	1.549e-02	-1.0368	0.29981	
Type.of.apartment	-0.2517881	2.946e-01	-0.8547	0.39273	
No.of.Credits.at.this.BankMore than 1	0.3613906	3.829e-01	0.9439	0.34524	

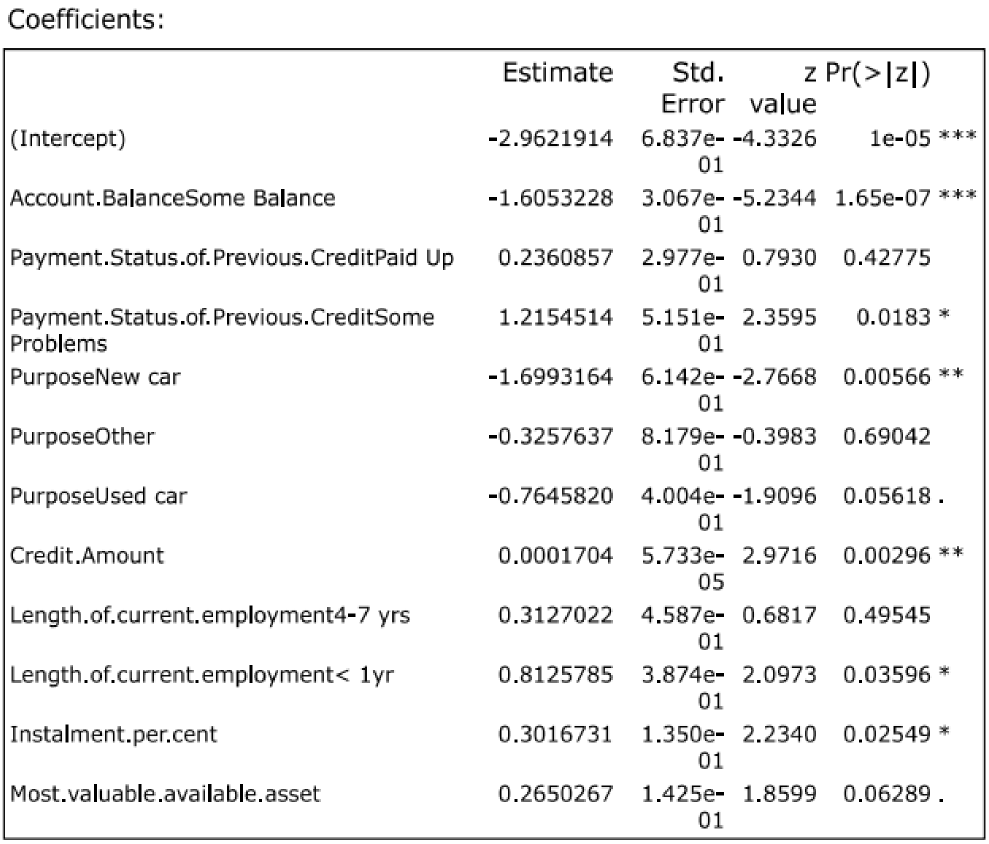
Figure 5: Coefficient and P-Value table of linear regression

From the result:

1. The p value of duration of credit month was greater than (>) 0.05, showing that it was statistically insignificant. I decided to drop this in my linear regression model
2. The value savings stocks none and 100-1000 categories also had p values > 0.05, dropped this as well
3. Age years, Type of apartment, no of credits at this bank more than 1 were also dropped due to been statistically insignificant

   From the stepwise regression, you would notice these variables were not included in the model.

B. With Stepwise

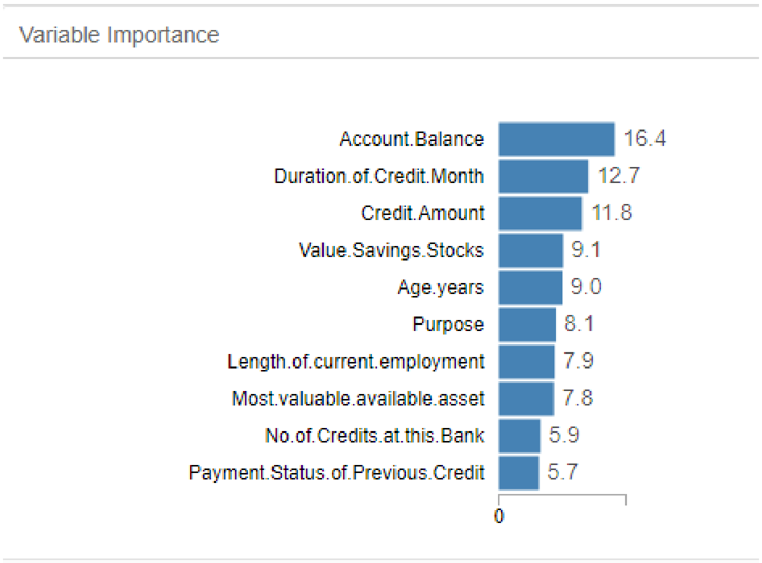


Coefficients:

	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	.

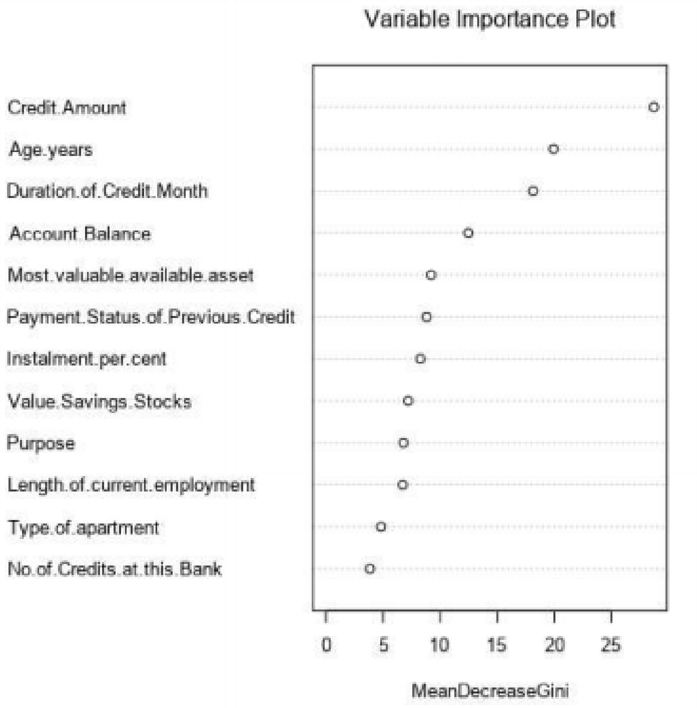
2. Decision Trees

From the result, Account Balance was the most important variable followed by duration of credit month. Surprisingly, payment status of previous credit was the least significant.



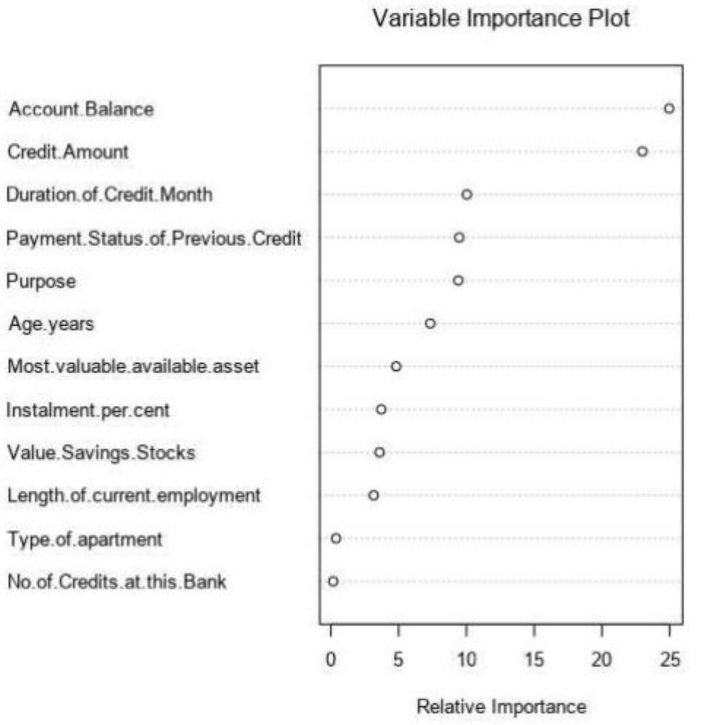
3. Forest Model

From the result, credit amount was the most important variable, followed by age in years. No of credits at the bank and telephone were the least significant.



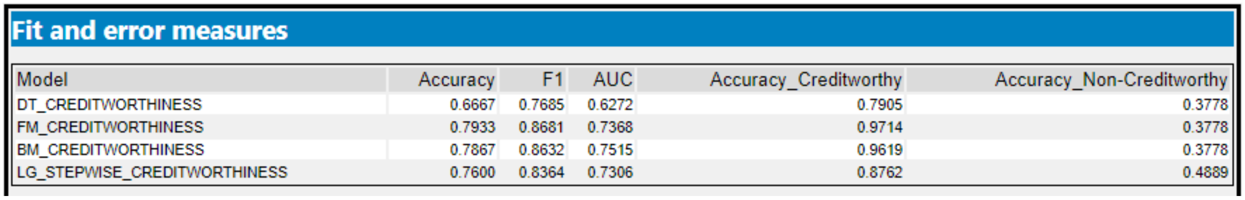
4. Boosted Model

From the result, Account Balance and Credit Amount are the most important variables, Telephone, Type of apartment and No of credits at this bank were the least important.



- Validate your model against the Validation set. What was the overall percent accuracy?

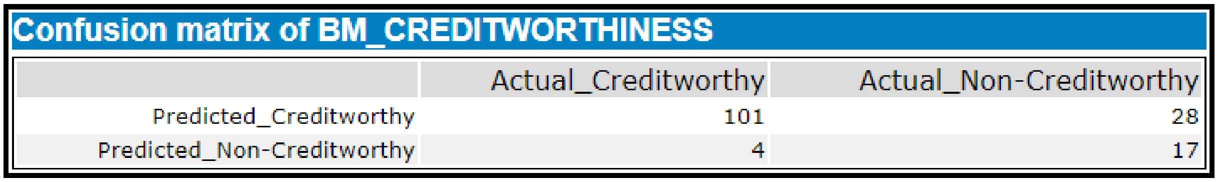
Forest Model achieved the highest accuracy with 79.33%, followed by Boosted Model of 78.67%, then Logistic Regression 76% and Decision Tree being the last with 66.67%



Fit and error measures

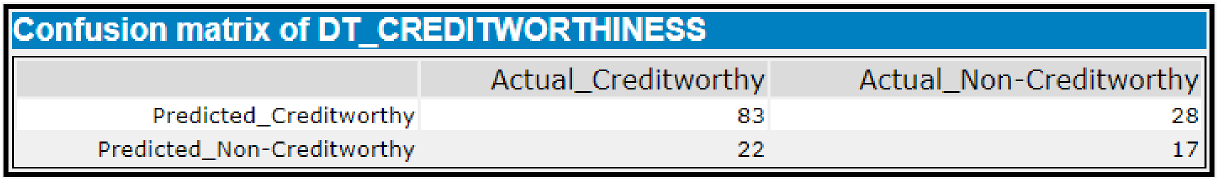
Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CREDITWORTHINESS	0.6667	0.7685	0.6272	0.7905	0.3778
FM_CREDITWORTHINESS	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CREDITWORTHINESS	0.7867	0.8632	0.7515	0.9619	0.3778
LG_STEPWISE_CREDITWORTHINESS	0.7600	0.8364	0.7306	0.8762	0.4889

- Show the confusion matrix. Are there any bias seen in the model's predictions?



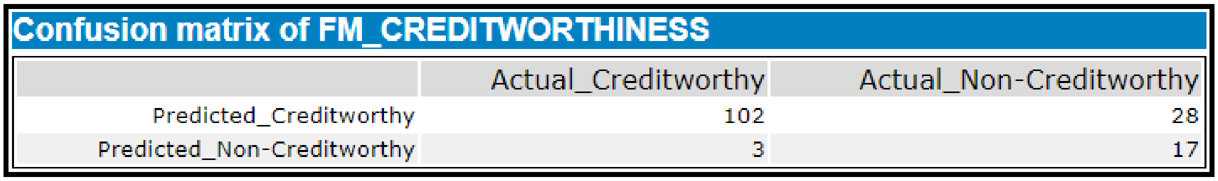
Confusion matrix of BM_CREDITWORTHINESS

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



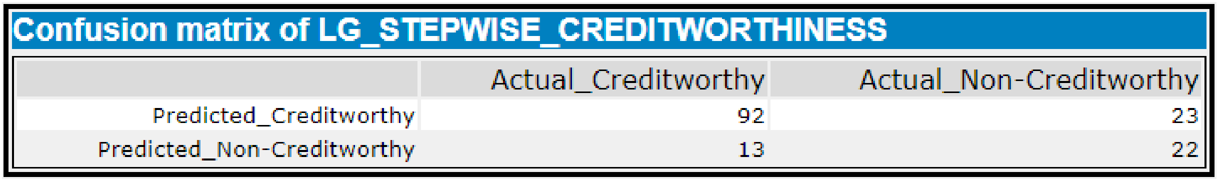
Confusion matrix of DT_CREDITWORTHINESS

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



Confusion matrix of FM_CREDITWORTHINESS

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



Confusion matrix of LG_STEPWISE_CREDITWORTHINESS

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

Model Name	Bias
Linear Regression	The model had a bias of 38.73 % in favour of credit worthy. This means the model is more likely to predict credit worthy than non credit worthy
Decision Tree	The model also had a bias of over 38.73% towards predicting credit worthy
Forest Model	The model had a bias of 59.36% in favour of credit worthy
Boosted Model	The model had a bias of 58.41% in favour of credit worthy

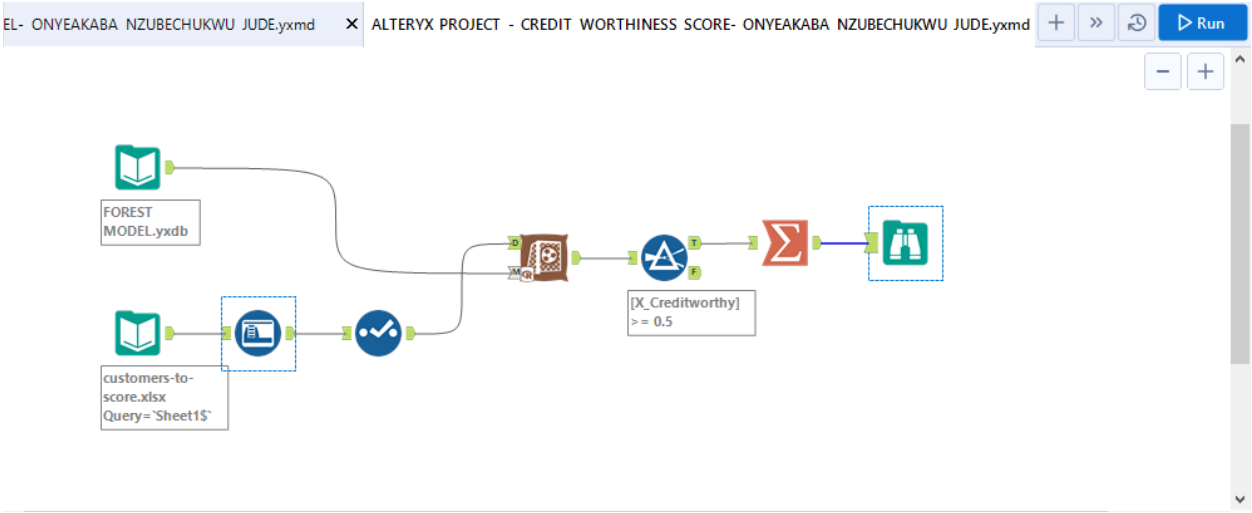
All the models had more difficulty is predicting non-creditworthiness i.e. a lot of customers who are not deserving are being labelled as creditworthy. Even the model which best predicts non-creditworthiness only has an accuracy of 48.89%. On the other hand, the models have a much easier times predicting creditworthiness, meaning its easier for the model to predict the deserving customers as credit worthy. The best models had an accuracy of 97.14% and the worst one does it with 79.05% accuracy in predicting credit worthy

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

Step 4: Writeup

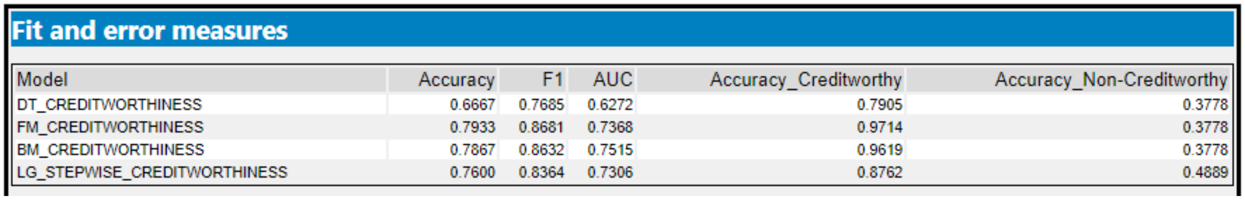
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"
Write a brief report on how you came up with your classification model and write down how many of the new customers would qualify for a loan. (250 word limit)

I connected the tools as in figure 3, after running the workflow, I studied the report from each browser tool and selected forest model because it had the highest level of accuracy. After applying this model using the score tool, and using the filter tool to select customers that scored above 50% (0.50),I used the summarise tool to count them – four hundred and twelve (412)


EL- ONYEAKABA NZUBECHUKWU JUDE.yxmd
ALTERYX PROJECT - CREDIT WORTHINESS SCORE- ONYEAKABA NZUBECHUKWU JUDE.yxmd
Run
FOREST MODEL.yxdb
customers-to-score.xlsx Query=`Sheet1$`
D
M
T
F
[X_Creditworthy] >= 0.5

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

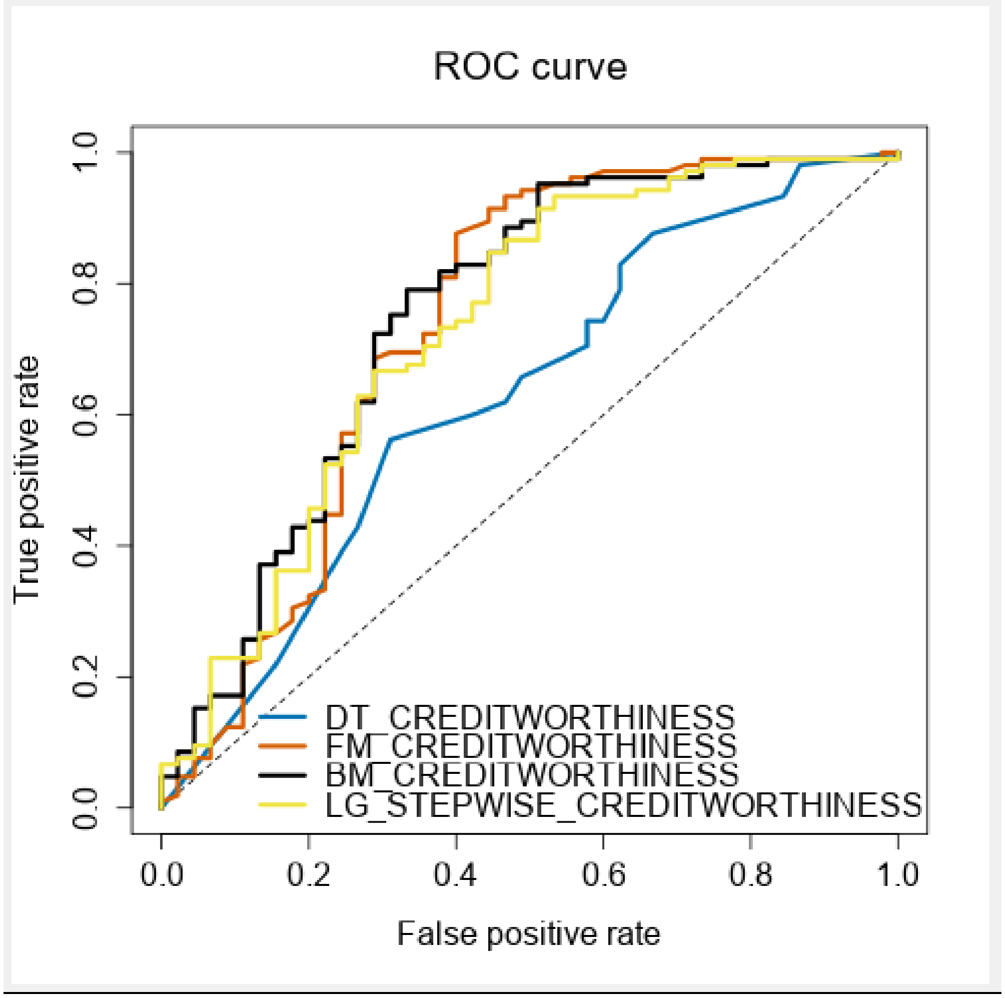


Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_CREDITWORTHINESS	0.6667	0.7685	0.6272	0.7905	0.3778
FM_CREDITWORTHINESS	0.7933	0.8681	0.7368	0.9714	0.3778
BM_CREDITWORTHINESS	0.7867	0.8632	0.7515	0.9619	0.3778
LG_STEPWISE_CREDITWORTHINESS	0.7600	0.8364	0.7306	0.8762	0.4889

- ○ ROC graph

An Ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate. From chart below we can see that the boosted model raises the fastest and is highest for most of the graph. Forest model and linear regression are also very high, but we not high enough. The decision tree has the lowest performance as it is lower than lines for other models.



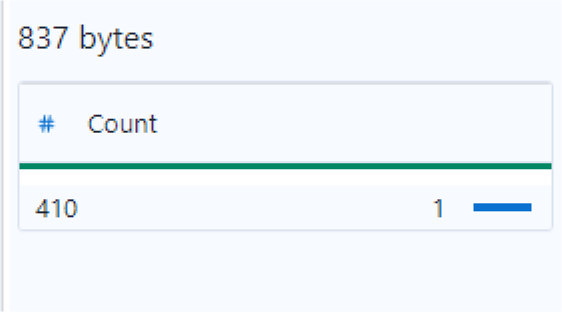
- ○ Bias in the Confusion Matrices

I selected the forest model because it had the highest accuracy in predicting creditworthy with 97.14% and 2nd highest in predicting non-creditworthy, 37.78%. It also had the highest overall accuracy of 79.33%

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

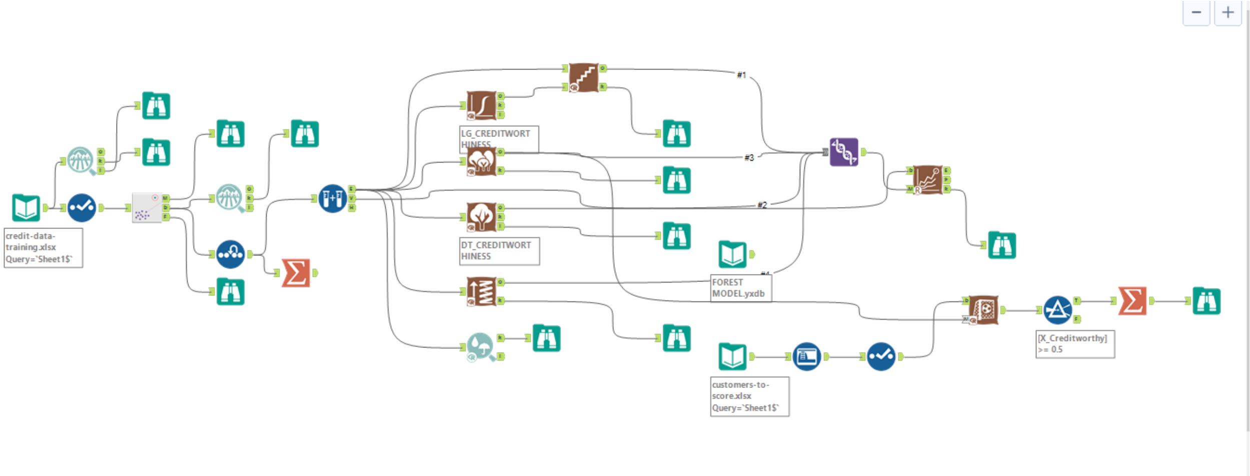
- How many individuals are creditworthy?

Four hundred and ten individuals were creditworthy. I achieved this by scoring the forest model with the customers to score sheet



837 bytes

# Count	
410	1



<u>Before you Submit</u>

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