Project: Creditworthiness

Step 1: Business and Data Understanding

- What decisions needs to be made?

I need to determine if customers are creditworthy to give a loan to, that is knowing how to systematically evaluate the creditworthiness of new loan applicants coming in.

- What data is needed to inform those decisions?

The data used is from the Credit-data-training dataset provided which contains all credit approvals from your past loan applicants the bank has ever completed. These were the variables used:

1. Account Balance
2. Duration of Credit month
3. Payment status of Previous Credit
4. Credit Amount
5. Purpose
6. Value-Savings-Stocks
7. Length-of-Current-employment
8. Installment per credit
9. Most-valuable-available assets
10. Age-years
11. Type of apartment
12. No of Credit at this Bank

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

The kind of model to use is a Binary model since it is a Yes or No answer: Creditworthy or not; Approved or not Approved.

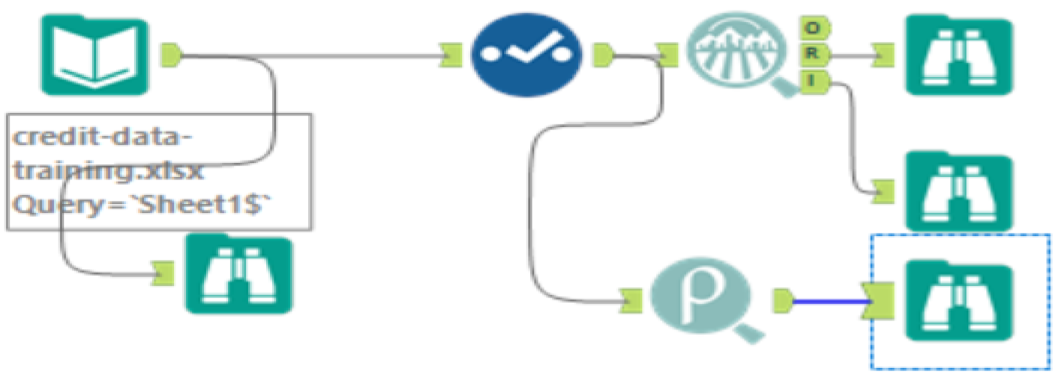
Step 2: Building the Training Set

*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.***

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

From the diagrams below, these decisions were taken:

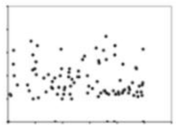
1. Concurrent Credits, Occupation, Guarantors, Foreign Worker and No of Dependents were removed because it shows low variability. This was done in order not to skew our analysis results.
2. Duration in Current Address has 69% missing data and was also removed.
3. While Age Years has few missing data, I decided to impute the missing data with the median age.
4. Finally, Telephone field was also removed due to its irrelevancy to the customer creditworthy.



<u>Figure above showing Alteryx workflow used to build data set</u>.

<u>Figure above showing the field summary of the data set</u>.

Record	Report
1	**Numeric Fields**

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	

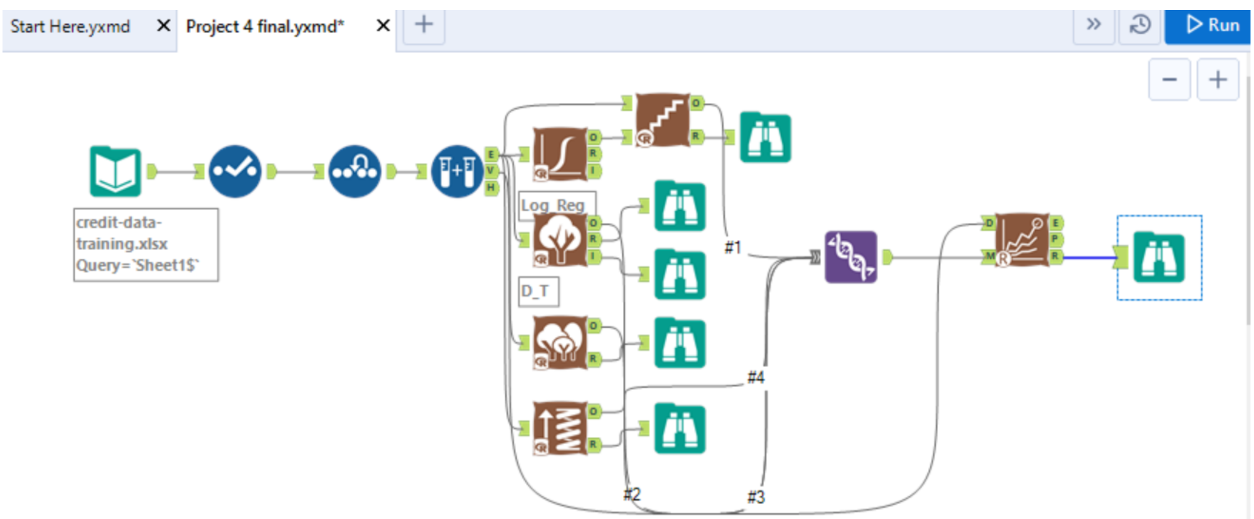
<u>Figure above showing the percentage of the missing data in the variable – Age-years</u>

FieldName	Duration-...	Credit-Amount	Instalm...	Durati...	Most-valua...	Age-years	Type-of-apa...	Occupation	No-of-de...	Telephone	Foreign-Worker
Duration-of-Credit-Month	1	0.57398	0.068106	[Null]	0.299855	[Null]	0.152516	[Null]	-0.065269	0.143176	-0.115916
Credit-Amount	0.57398	1	-0.288852	[Null]	0.325545	[Null]	0.170071	[Null]	0.003986	0.286338	0.025493
Instalment-per-cent	0.068106	-0.288852	1	[Null]	0.081493	[Null]	0.074533	[Null]	-0.125894	0.029354	-0.133411
Duration-in-Current-address	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]
Most-valuable-available-asset	0.299855	0.325545	0.081493	[Null]	1	[Null]	0.373101	[Null]	0.046454	0.203509	-0.146005
Age-years	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]	[Null]	[Null]
Type-of-apartment	0.152516	0.170071	0.074533	[Null]	0.373101	[Null]	1	[Null]	0.170738	0.101443	-0.089848
Occupation	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	[Null]	1	[Null]	[Null]	[Null]
No-of-dependents	-0.065269	0.003986	-0.125894	[Null]	0.046454	[Null]	0.170738	[Null]	1	-0.048559	0.065943
Telephone	0.143176	0.286338	0.029354	[Null]	0.203509	[Null]	0.101443	[Null]	-0.048559	1	-0.055516
Foreign-Worker	-0.115916	0.025493	-0.133411	[Null]	-0.146005	[Null]	-0.089848	[Null]	0.065943	-0.055516	1

<u>Figure above shows the person correlation table.</u>

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.



*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.

From the diagrams as shown below, the significant predictor variables for each of the models are:

1. Logistic Regression: Account Balance, Payment status of Previous Credit, Purpose, Credit Amount, Length-of-Current-employment and Installment per credit.

2. Decision Tree: Account Balance, Duration of Credit month and Value-Savings-Stocks

3. Forest Model: Credit Amount, Age-years, Duration of Credit month and Account Balance.

4. Boosted Model: Credit Amount, Account Balance, Duration of Credit month and Payment status of Previous Credit.

<u>P value table for Logistic Regression model</u>

Report

Report for Logistic Regression Model X

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:

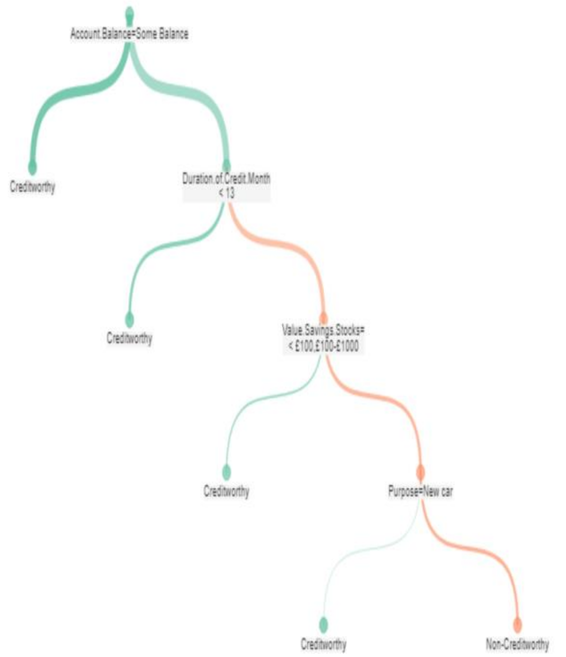
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05	***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07	***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775	
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01	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	.

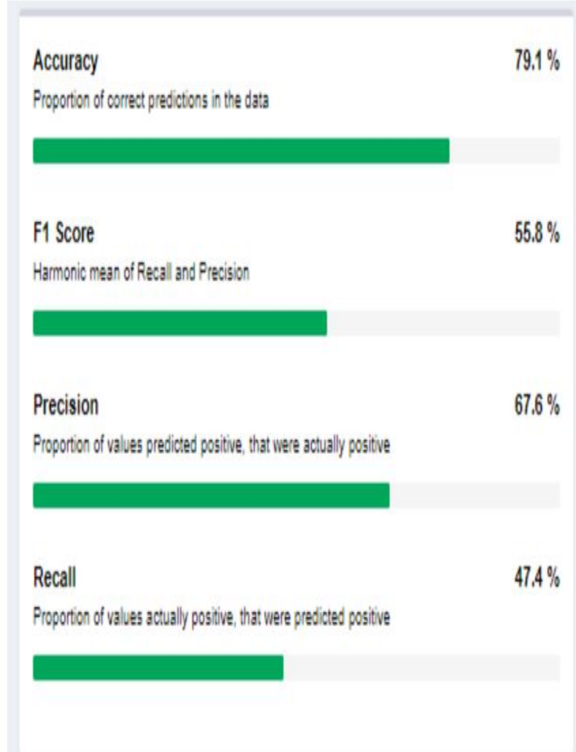
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

Number of Fisher Scoring iterations: 5

Type II Analysis of Deviance Tests

Decision Tree summary





Summary Report for Decision Tree Model D_T

Call:
rpart(formula = Credit.Application.Result ~ 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, data = the.data, minsplit = 20, minbucket = 7, xval = 10, maxdepth = 20, cp = 1e-05, usesurrogate = 0, surrogatestyle = 0)

Model Summary

```
Variables actually used in tree construction:
[1] Account.Balance Duration.of.Credit.Month Purpose Value.Savings.Stocks
Root node error: 97/350 = 0.27714
n= 350
```

Pruning Table

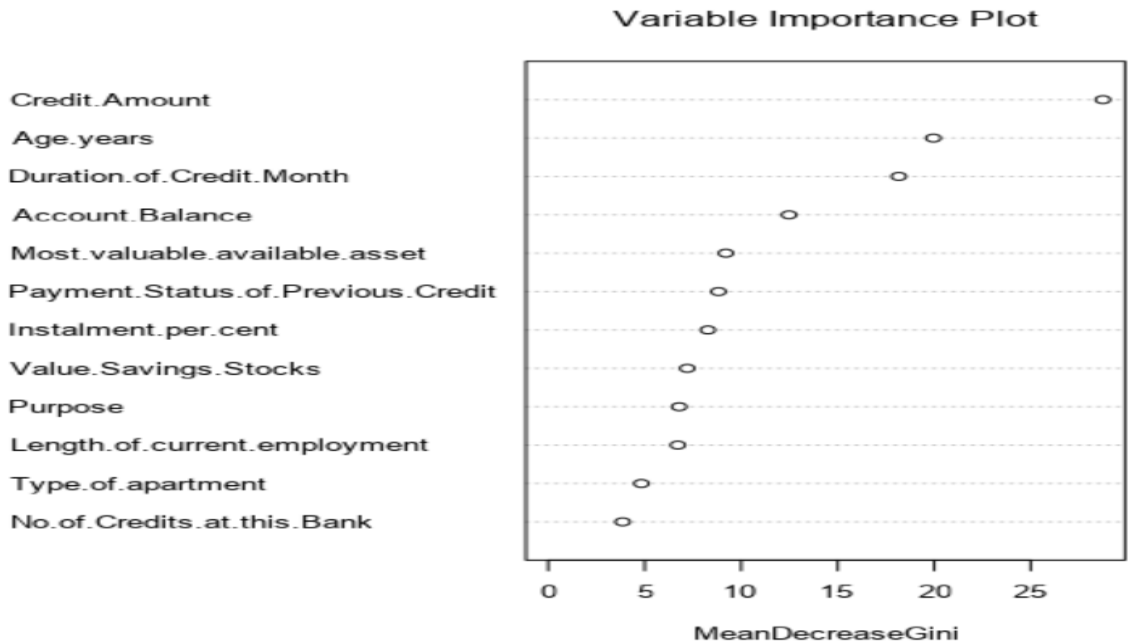
Level	CP	Num Splits	Rel Error	X Error	X Std Dev
1	0.068729	0	1.00000	1.00000	0.086326
2	0.041237	3	0.79381	0.94845	0.084898
3	0.025773	4	0.75258	0.88660	0.083032

Leaf Summary

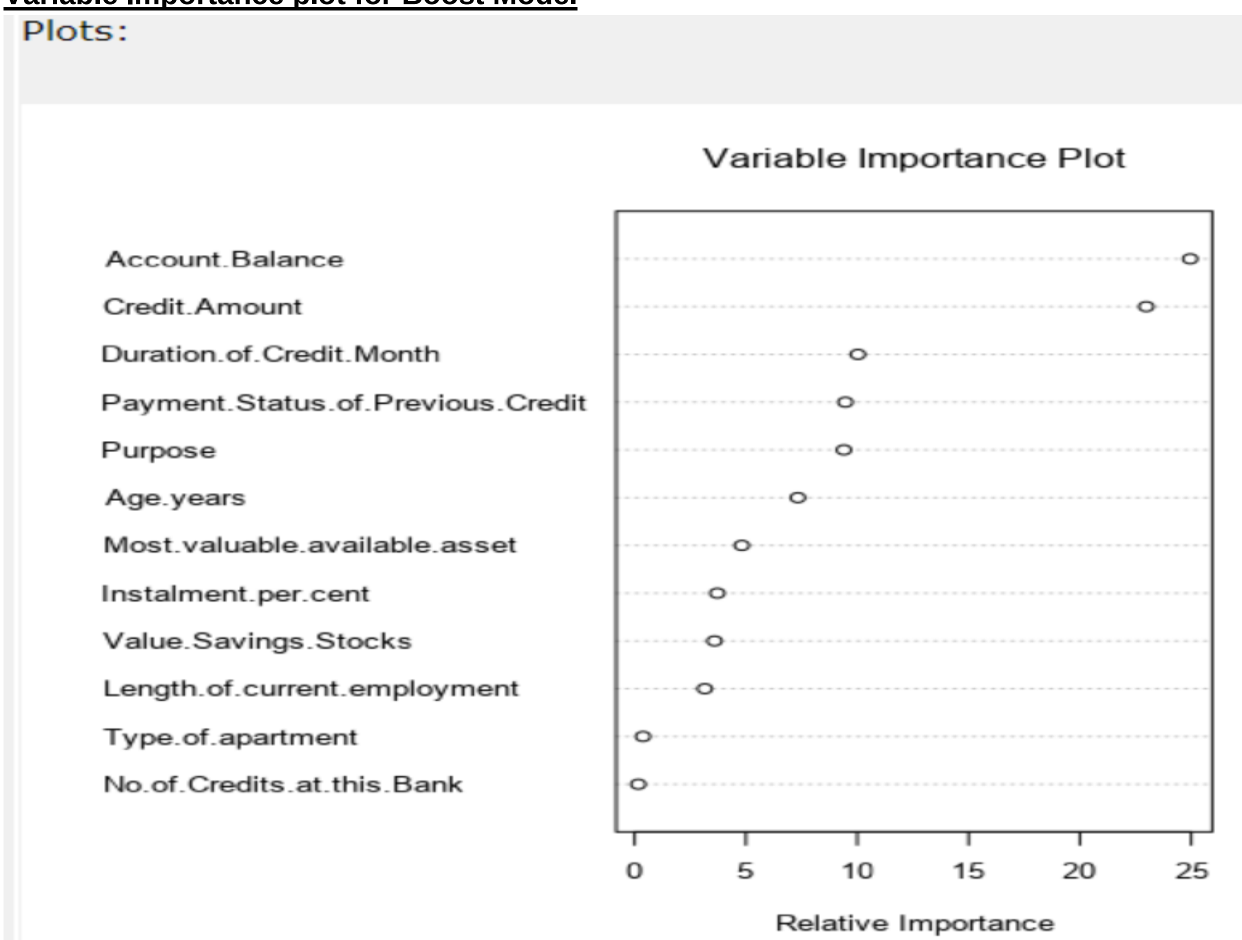
```
node), split, n, loss, yval, (yprob)
    * denotes terminal node
1) root 350 97 Creditworthy (0.7228571 0.2771429)
  2) Account.Balance=Some Balance 166 20 Creditworthy (0.8795181 0.1204819) *
  3) Account.Balance=No Account 184 77 Creditworthy (0.5815217 0.4184783)
    6) Duration.of.Credit.Month< 13 74 18 Creditworthy (0.7567568 0.2432432) *
    7) Duration.of.Credit.Month>=13 110 51 Non-Creditworthy (0.4636364 0.5363636)
     14) Value.Savings.Stocks=< £100,£100-£1000 34 11 Creditworthy (0.6764706 0.3235294) *
     15) Value.Savings.Stocks=None 76 28 Non-Creditworthy (0.3684211 0.6315789)
       30) Purpose=New car 8  2 Creditworthy (0.7500000 0.2500000) *
       31) Purpose=Home Related,Other,Used car 68 22 Non-Creditworthy (0.3235294 0.6764706) *
```

Plots

Variable Importance plot for Forest Model



Variable Importance plot for Boost 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?

The model comparison function was used to compare models to get their respective overall accuracy and confusion matrix as shown below:

Layout

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
X	0.7600	0.8364	0.7306	0.8762	0.4889
D_T	0.7467	0.8304	0.7035	0.8857	0.4222
F_T	0.7933	0.8681	0.7368	0.9714	0.3778
BT	0.7867	0.8632	0.7515	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 BT

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

Confusion matrix of D_T

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of F_T

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

Confusion matrix of X

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

Performance Diagnostic Plots

From the above diagram, the Logistic regression has an overall accuracy of 76%. The accuracy to predict Creditworthy is 87% while for Non creditworthy is very low at 48%. Also in the confusion matrix, it shows low accuracy for prediction of Non creditworthy customers showing that there is bias.

In the Decisions tree model, the overall accuracy is quite same as above but at 74%. There is good prediction for the Creditworthy customers but yet low accuracy for Non creditworthiness. Hence there is bias.

Both Forest and Boost model did better with an overall accuracy of 79% and 78% respectively, but yet there is bias in predicting the Actual Credit worthy and Non creditworthy customers.

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"

Steps I took to come up with the classification model:

1. I created the analytical data set, cleaned the data, removed 7 irrelevant columns and imputed the Age year with the median age
2. I created 70% estimation sample and 30% validation sample
3. I added and set up Logistic regression model with a Stepwise tool
4. I added and set up Decision tree model
5. I added and set up Forest model
6. I added and set up Boosted model
7. I joined all the models using the union tool and compared them all with a model comparison tool
8. Then used the best model to score and get the number of individuals that are creditworthy.

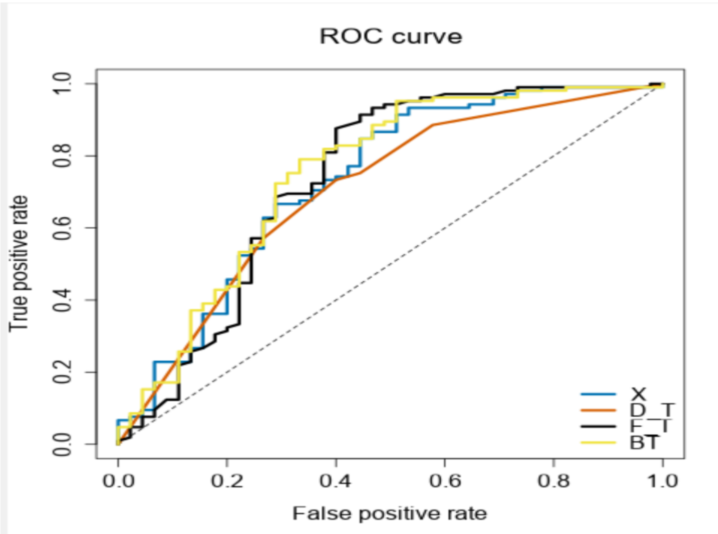
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set

From the model comparison report, Forest Model has the highest overall accuracy of 79%.

- Accuracies within "Creditworthy" and "Non-Creditworthy" segments

Also Forest model has the highest accuracy for the Creditworthy segment although it has a low accuracy for non credit worthy segment.

- ROC graph



When comparing the ROC curve of the four models, Forest tree and Boosted model performs better while Decision tree seems to perform worst.

- Bias in the Confusion Matrices

Confusion matrix of BT

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

Confusion matrix of D_T

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of F_T

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

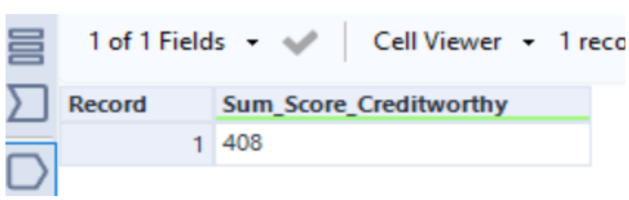
Confusion matrix of X

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

From the above confusion matrix, it is observed that the Forest tree model predicted more than the other models for creditworthy customers. It predicted 102 customers for Creditworthy and 28 for Non creditworthy customers.

Since the boss cares more about prediction accuracy for Creditworthy and Non-Creditworthy segments, Forest tree model will be used because it has the highest accuracy.

- How many individuals are creditworthy?



Record	Sum_Score_Creditworthy
1	408

408 individuals are creditworthy.

Altreyx workflow for Project: Creditworthiness