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: The decision needs to be made is based on the given data, whether the new customers are Creditworthy to give a loan.

- What data is needed to inform those decisions?

Ans: The following data is needed.

1. Account-Balance
2. Duration-of-Credit-Month
3. Payment-Status-of-Previous-Credit
4. Purpose
5. Credit-Amount
6. Value-Savings-Stocks
7. Length-of-current-employment
8. Instalment-per-cent
9. Most-valuable-available-asset
10. Age-years
11. Type-of-apartment
12. No-of-Credits-at-this-Bank

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

Ans: The decision we need to make is whether a new customer is Creditworthy to get a loan or Not-Creditworthy. So the kind of model needed is a **Binary Classification Model.**

Step 2: Building the Training Set

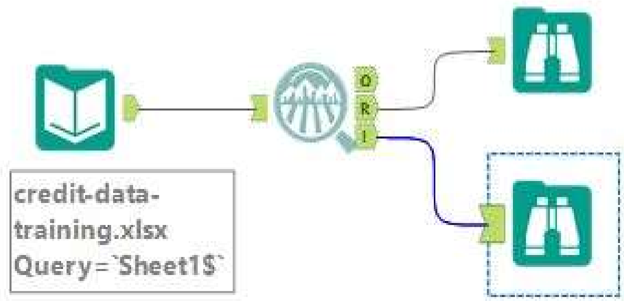
Build your training set given the data provided to you. The data has been cleaned up for you already so you shouldn't ***need to convert any data fields to the appropriate data types.***

Answer this question:

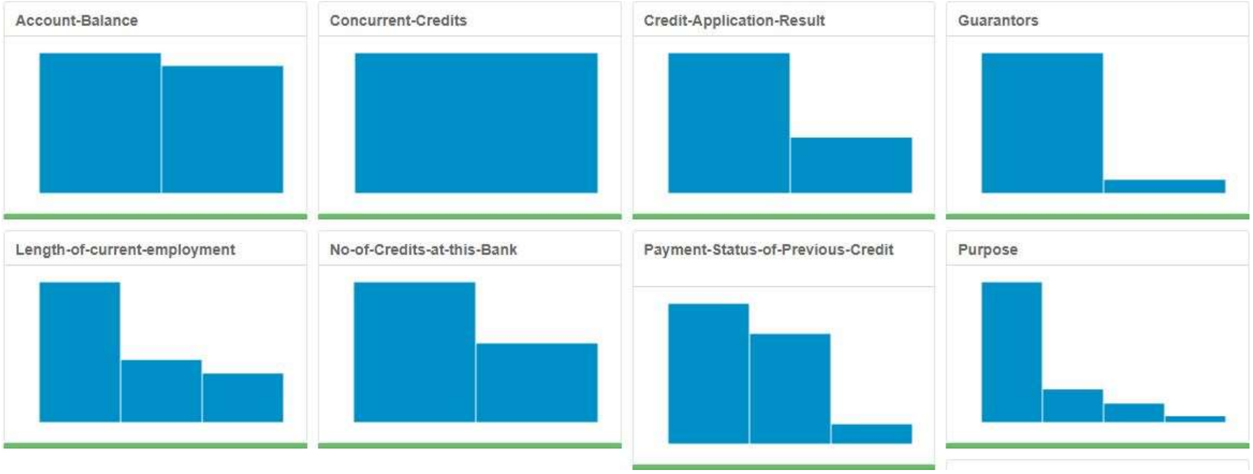
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

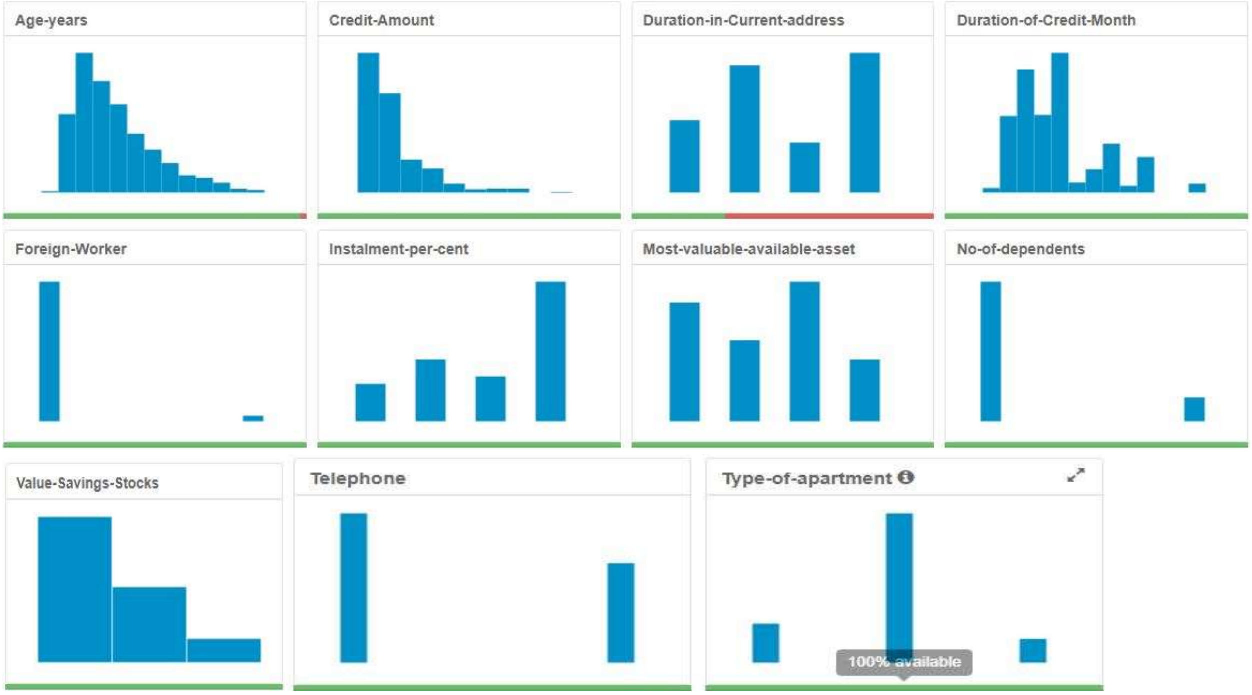
Ans:

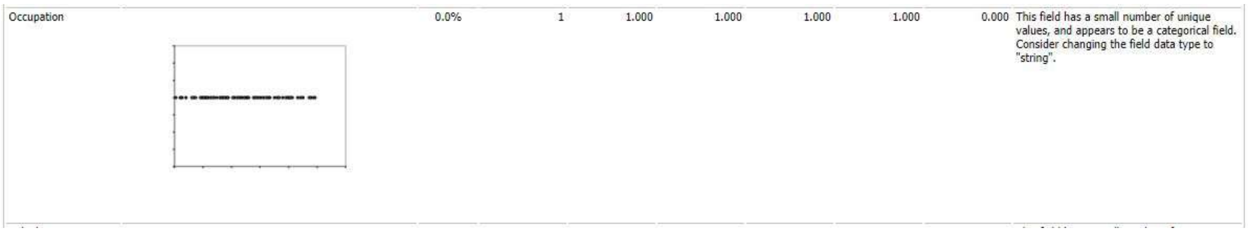
Step 1: For building the Training set and also to find the details (missing / low variance) about the data, I started visualization of data using Field Summary tool.



And here is the result of the Field Summary tool.





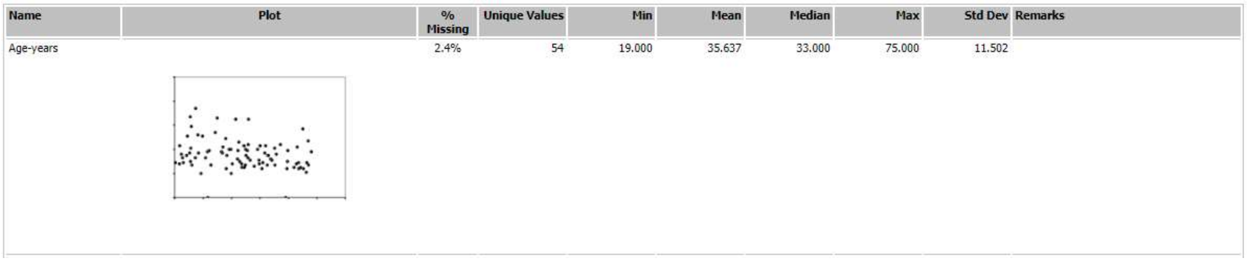


Occupation		0.0%	1	1.000	1.000	1.000	1.000	0.000	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".

Then I decided to remove the following fields for the following reasons.

1. Concurrent Credit	-	Low Variability (Only other banks/dept.)
2. Guarantors	-	Low variability
3. Duration-in-Current-Address	-	Too many Missing data (69% missing)
4. Foreign-Worker	-	Low Variability
5. No-of-dependents	-	Low Variability
6. Telephone	-	Incorrect/Irrelevant data
7. Occupation	-	Low variability (Only 1’ S)

Step 2: Next I decided to impute Age-years field. I thought of imputing the missing values with the Median of the non-zero values.



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	

Step 3: Train your Classification Models

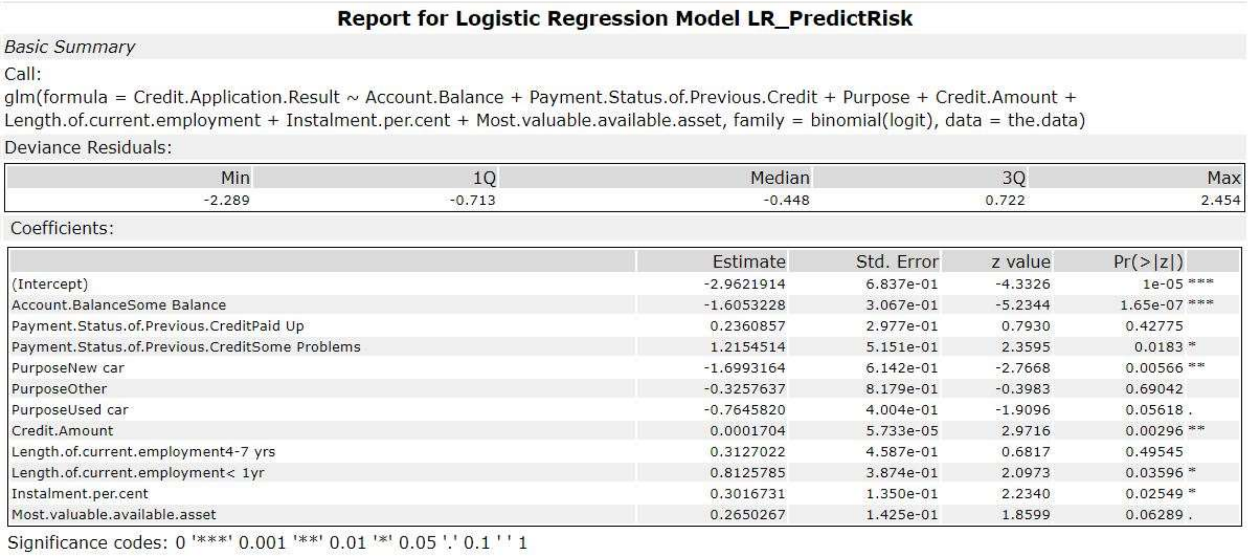
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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

Ans:
3.1 P - Values table for the Logistic Regression



Report for Logistic Regression Model LR_PredictRisk

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:

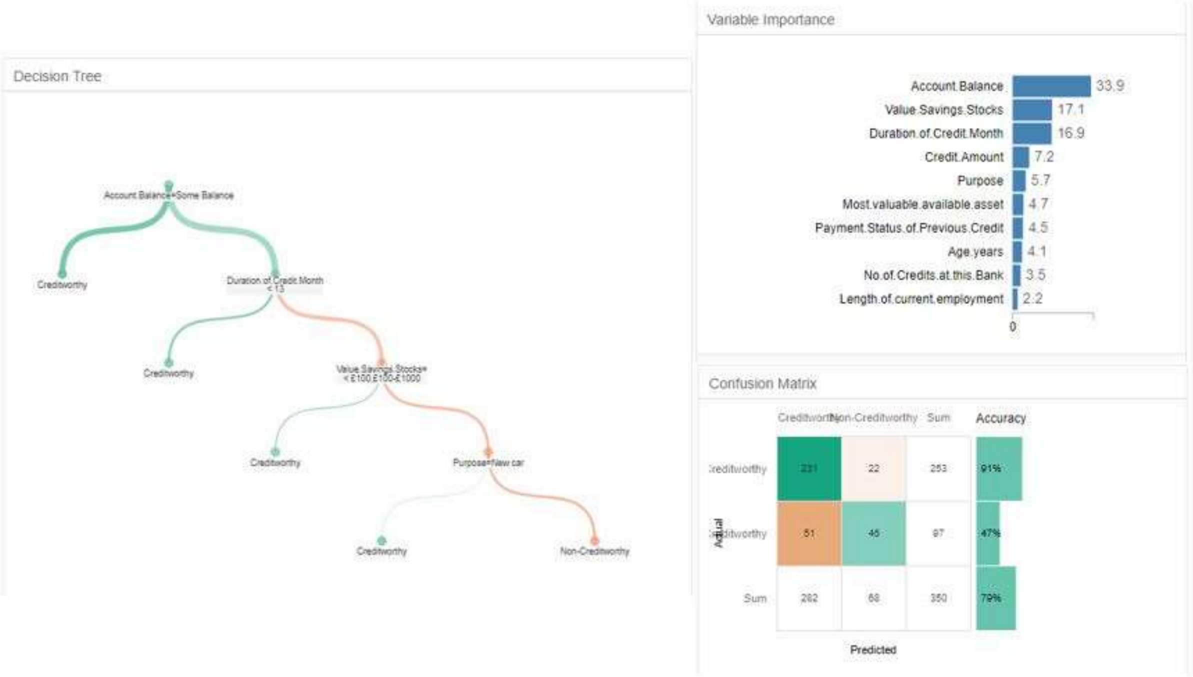
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01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

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

The Significant Predictor variables for Logistic Regression are:

- ✓ Account-Balance
- ✓ Payment-status-of-Previous-Credit
- ✓ Purpose
- ✓ Credit-Amount
- ✓ Length-of-current-employment
- ✓ Instalment-per-cent

3.2 Decision-Tree Summary



Confusion Matrix

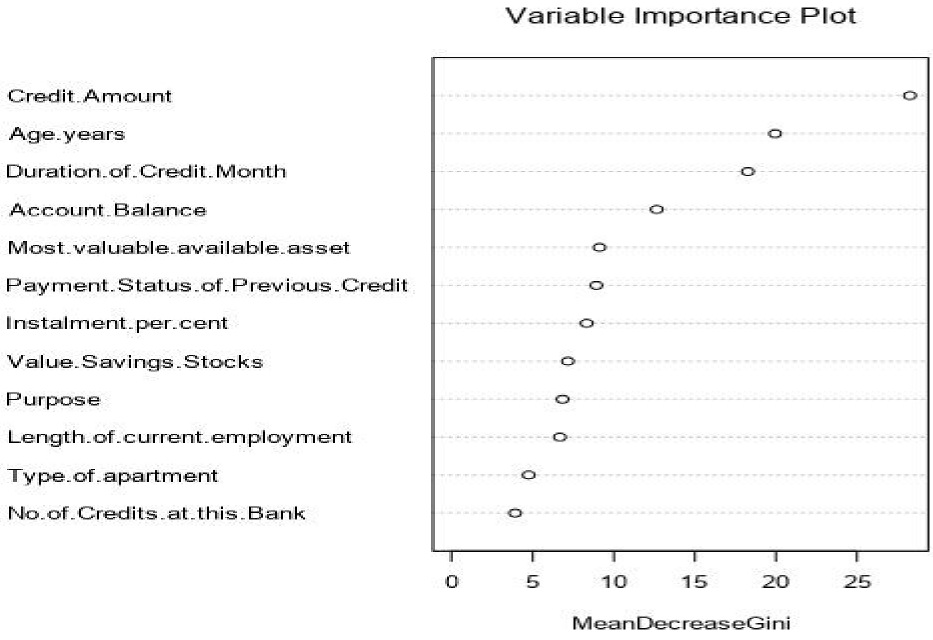
Actual	Creditworthy	Non-Creditworthy	Sum	Accuracy
reditworthy	231	22	253	91%
ditworthy	51	46	97	47%
Sum	282	68	350	79%

Predicted

The Significant Predictor variables for Decision-Tree are:

- ✓ Account-Balance
- ✓ Value-Savings-Stocks
- ✓ Duration-of-credit-month

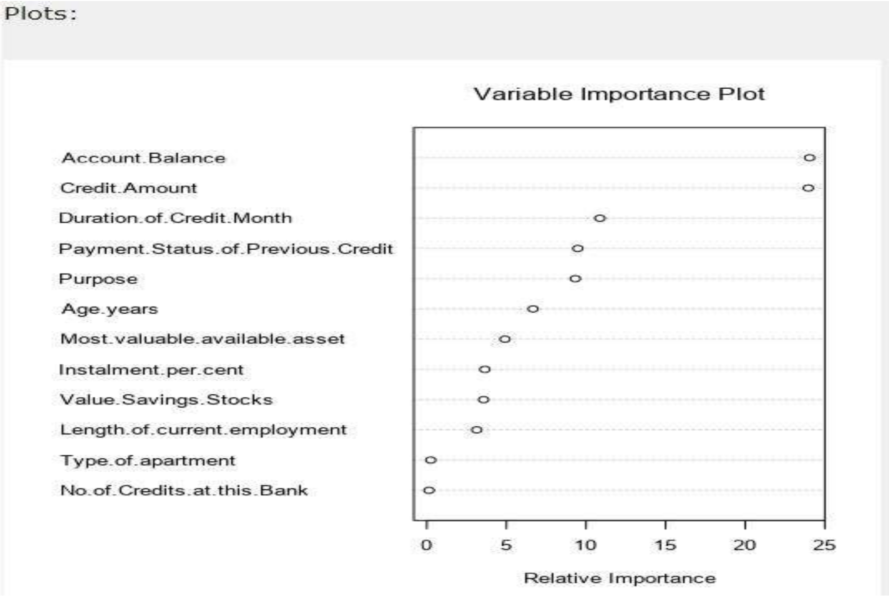
3.3 Forest Model



The significant Variables for Forest Model are:

- ✓ Credit-Amount
- ✓ Age-years

3.4 Boosted Model

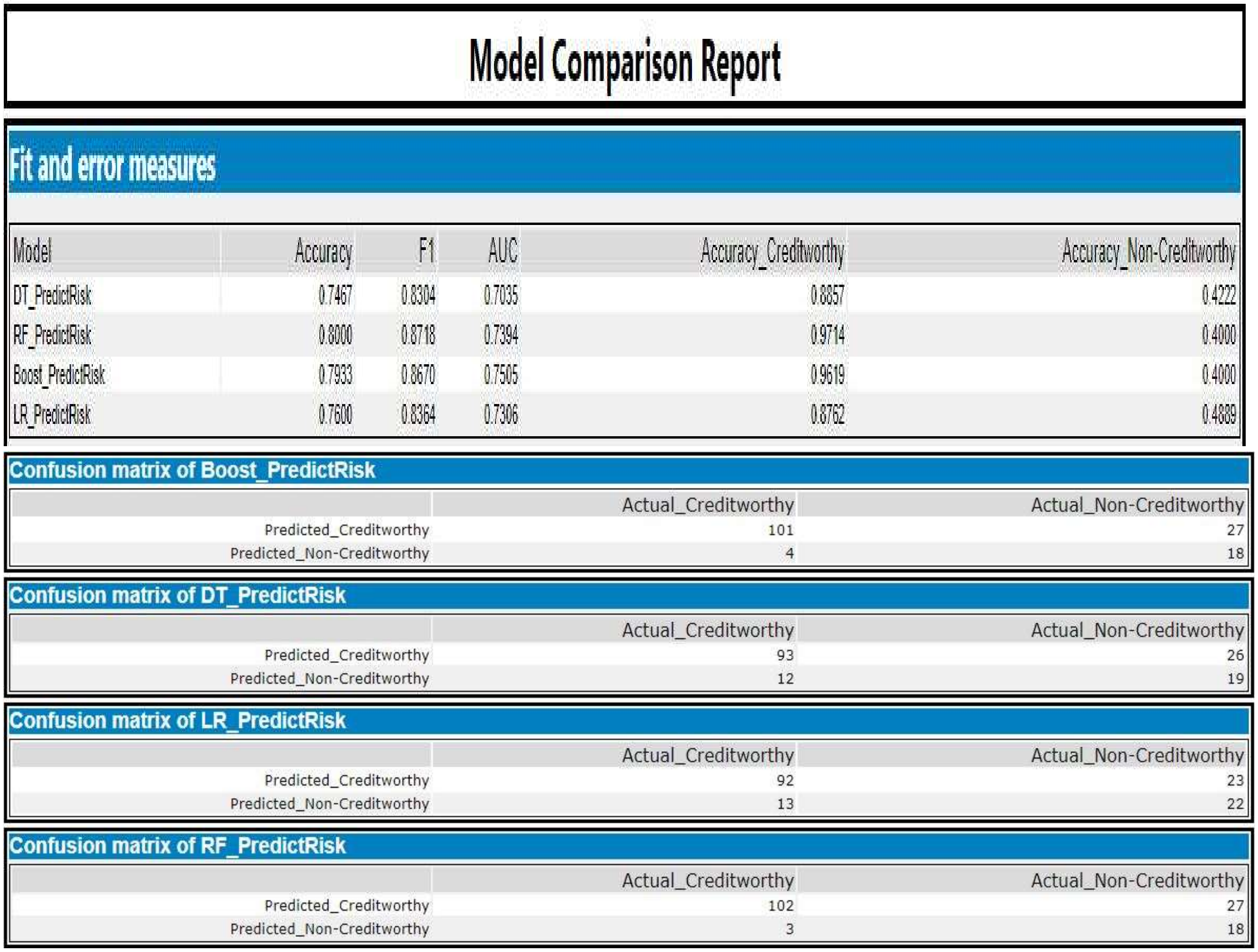


The Significant Variables for Boosted Model are:

- ✓ Account-Balance
- ✓ 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?

Ans: I used Model Comparison Tool to compare the overall accuracies of all the models and their confusion matrices. Here is the report of Model Comparison.



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
DT_PredictRisk	0.7467	0.8304	0.7035	0.8857	0.4222
RF_PredictRisk	0.8000	0.8718	0.7394	0.9714	0.4000
Boost_PredictRisk	0.7933	0.8670	0.7505	0.9619	0.4000
LR_PredictRisk	0.7600	0.8364	0.7306	0.8762	0.4889

Confusion matrix of Boost_PredictRisk

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

Confusion matrix of DT_PredictRisk

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

Confusion matrix of LR_PredictRisk

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

Confusion matrix of RF_PredictRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

From the above report, the accuracies of the various models are:

❖ Stepwise Model	-	0.7600
❖ Decision-Tree Model	-	0.7464
❖ Forest-Model	-	0.8000
❖ Boosted-Model	-	0.7933

Yes. There are bias in the models.

- ➢ In the Forest Model, there are 102 records were predicted as Creditworthy, actually they are Creditworthy also, but I we have 27 records predicted as Creditworthy are actually Non-Creditworthy. Similarly 3 records which are actually Creditworthy were predicted as Non-Creditworthy, but 18 records were actually Non-Creditworthy predicted as Non-Creditworthy.

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

Step 4: Writeup

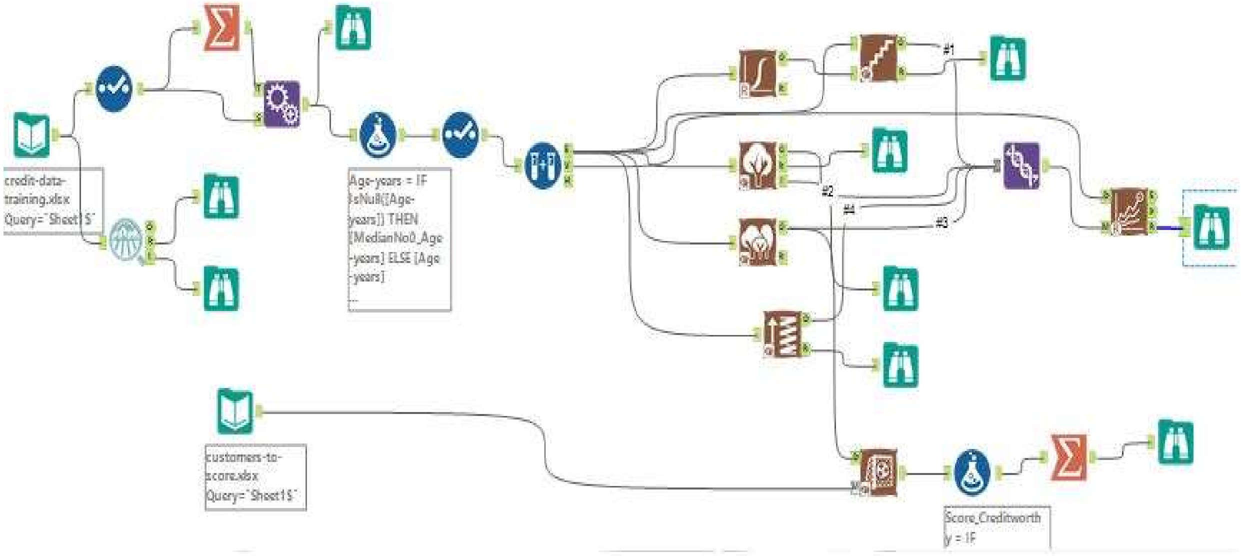
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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)

Ans: The following are the steps I followed to find the best model and score the new customers dataset.

1. First, I cleaned up the dataset by removing the 7 columns which are having missing values, irrelevant data & low variability.
2. Then I imputed the missing values for the field Age-years with the median of the non-missing values.
3. Then, I have created samples with estimation 0f 70% whereas the remaining 30% were validation data.
4. Next I added and configured various classification models one by one. First, I go with Logistic Regression model with Stepwise technique.
5. Then I added and configured Decision-Tree model followed by Forest model.
6. Finally added and configured Boosted model.
7. Combined the output of all the 4 models via Union Tool, and compared the accuracies and Confusion matrices with the help of Model Comparison Tool.

Here is my Alteryx Workflow of the entire thing:


credit-data-training.xlsx
Query=`Sheet1$`
Age-years = IF IsNull([Age-years]) THEN [MedianNo0_Age-years] ELSE [Age-years]
...
#1
#2
#3
#4
customers-to-score.xlsx
Query=`Sheet1$`
Score_Creditworthy = IF

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

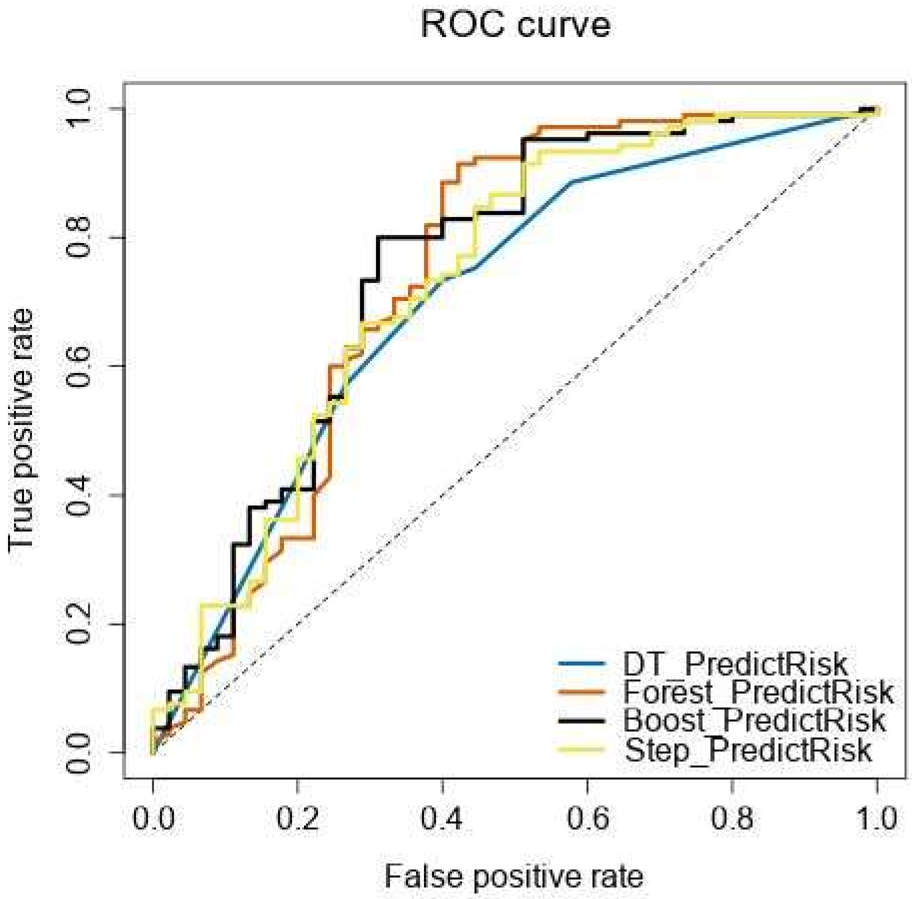
Ans: After comparing the accuracies of all the models for the validation dataset, I came to the conclusion that the **FOREST MODEL** is the best over other models, because the Forest model has an accuracy 0.8000 which is higher than that of others.
Hence, I decided to use Forest-Model to score the customers_to_score dataset.

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

Ans: For the Forest Model the accuracy of Creditworthy is 0.9714 which is greater than that of all the other models, also the accuracy of Non-Creditworthy is 0.4000 which is same as that of Boosted model but lower than Logistic and Decision Tree model. But the overall accuracy of Forest-Model is higher than all others, I decided to choose Forest-Model to score my new dataset.

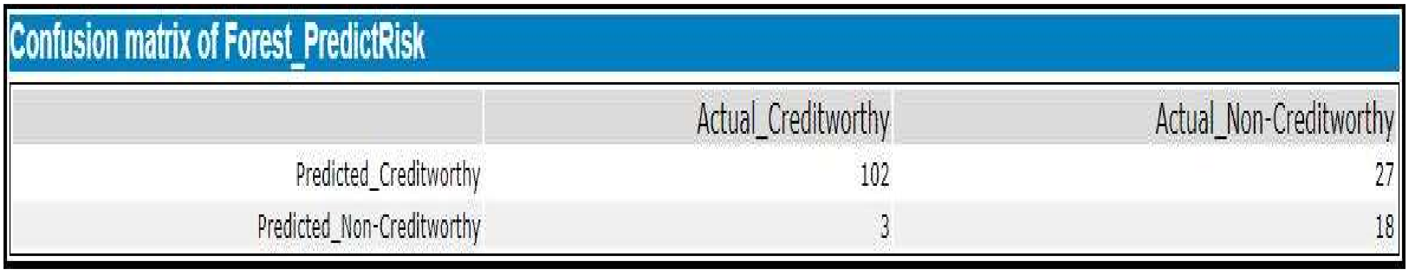
- ROC graph

Ans: In the ROC graph, the RED line indicates the Forest model, which implies it is performing better than others.



- Bias in the Confusion Matrices
-

Ans: There are 102 records which are predicted as Creditworthy are actually Creditworthy, also, 27 records which are predicted as Creditworthy are actually Non-Creditworthy.
And 3 records, actually Creditworthy were predicted as Non-Creditworthy, and 18 records actually Non-Creditworthy were predicted as Non-Creditworthy.



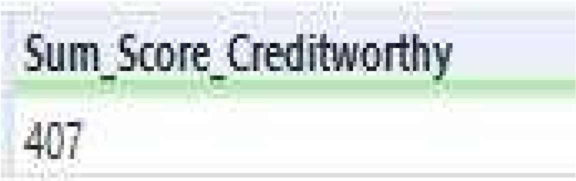
Confusion matrix of Forest_PredictRisk

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

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

- How many individuals are creditworthy?

Ans: For calculating the total number of Creditworthy, the customers-to-score dataset is given to the M side of score tool, using Forest model it is scored. The total number of Creditworthy is calculated as **407.**



Sum_Score_Creditworthy
407

<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.