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 word limit)

Key Decisions:

Answer these questions

- What decisions needs to be made?

Due to the high number applications that occurred as a result of a scandal at another bank and the need of processing them in a short period of time, the most important decision to be made is to give a loan only to applicants who are Creditworthy.

- What data is needed to inform those decisions?

To inform this decision, the following data is needed: data on all past applications and the list of customers (applicants) that need to be processed in the next few days.

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

The model that we need to use is a Binary, since we want to see if our customers(applicants) are creditworthy (yes) or non-creditworthy (no).

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

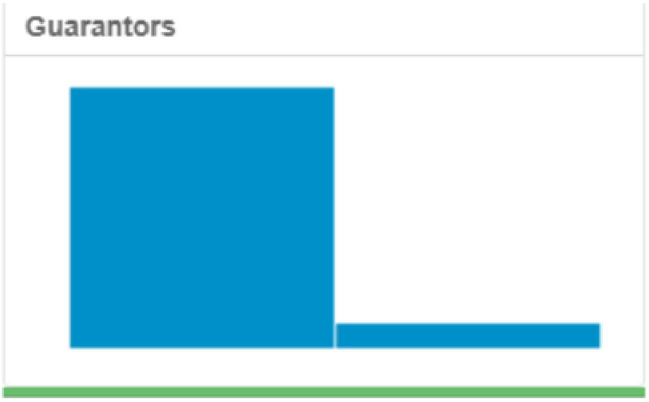
To achieve consistent results reviewers expect.

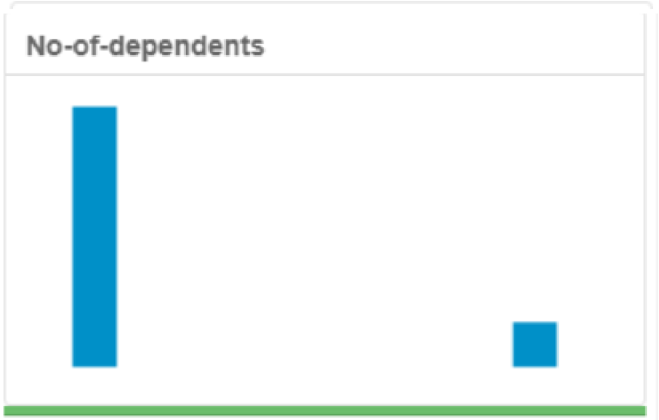
Answer this question:

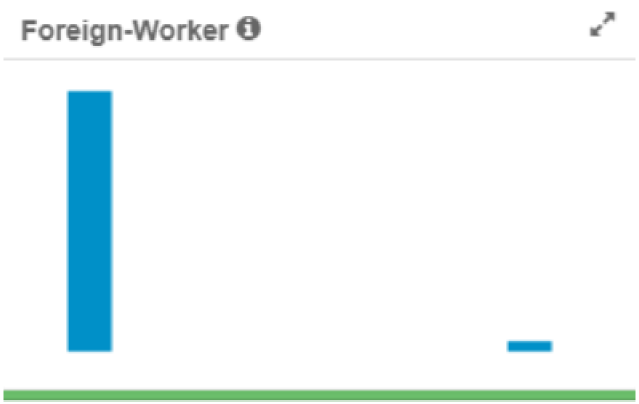
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

In my cleaning process, I have decided to remove the following fields:

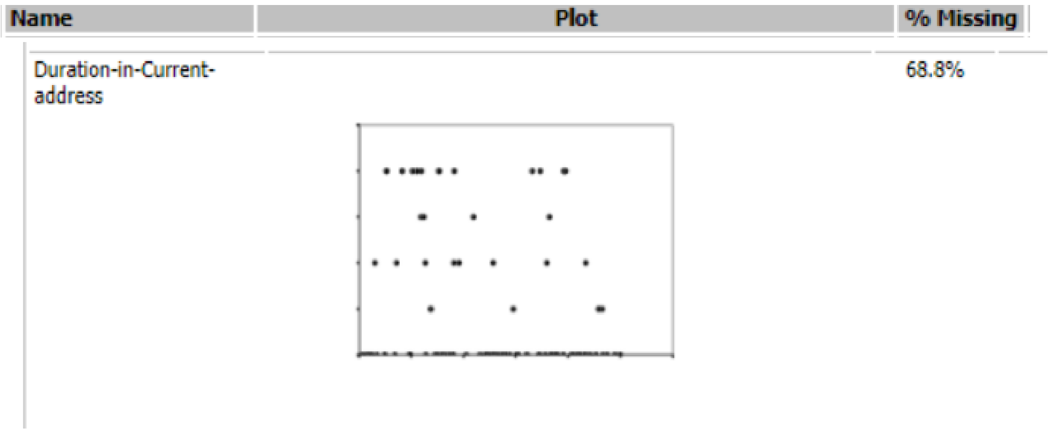
- Guarantors, Foreign-Worker, No-of-Dependents, because histograms show that data of these variables is skewed towards one type of data, which shows low variability.





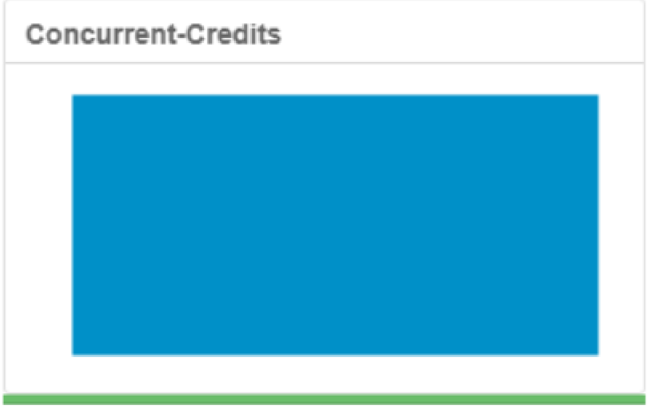


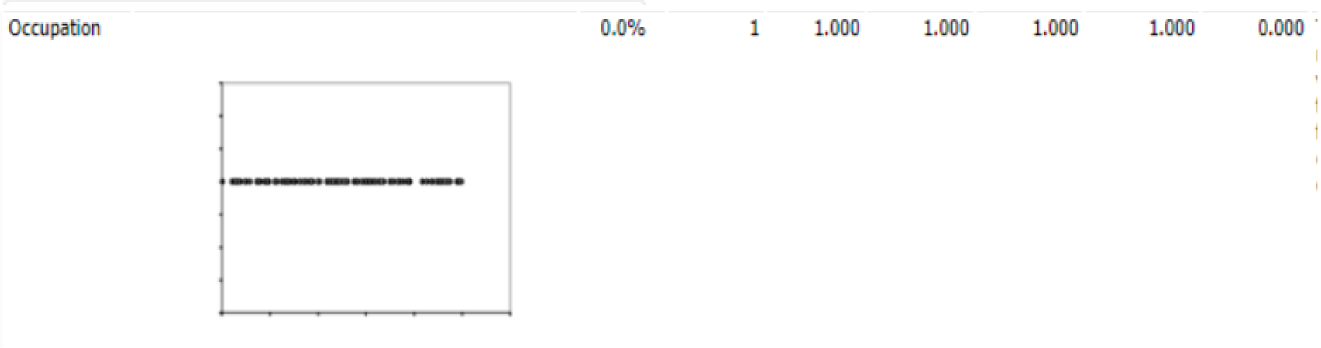
- Duration-in-Current-address because 69% of data is missing.



Name	Plot	% Missing
Duration-in-Current-address		68.8%

- Concurrent-Credits and Occupation because the data of these variables is completely uniform, showing a low variability.

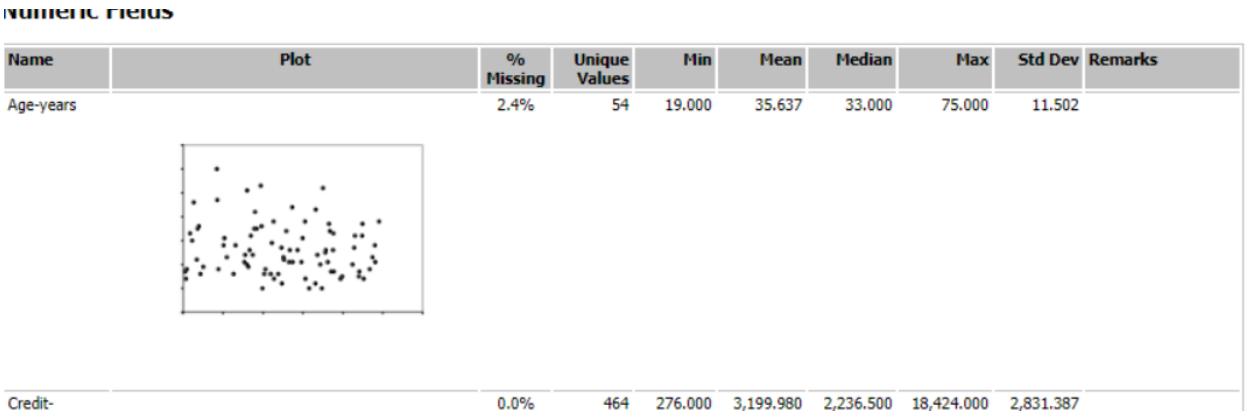




Occupation	0.0%	1	1.000	1.000	1.000	1.000	0.000

- Telephone because this variable no valuable data in order to use it for predicting the creditworthiness.

Also, I imputed the field “Age-Years” because this field had 2.4% missing data. I imputed the missing data using the Median of the entire Age-Years field, which is 33.



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	
Credit-		0.0%	464	276.000	3,199.980	2,236.500	18,424.000	2,831.387	

Step 3: Train your Classification Models

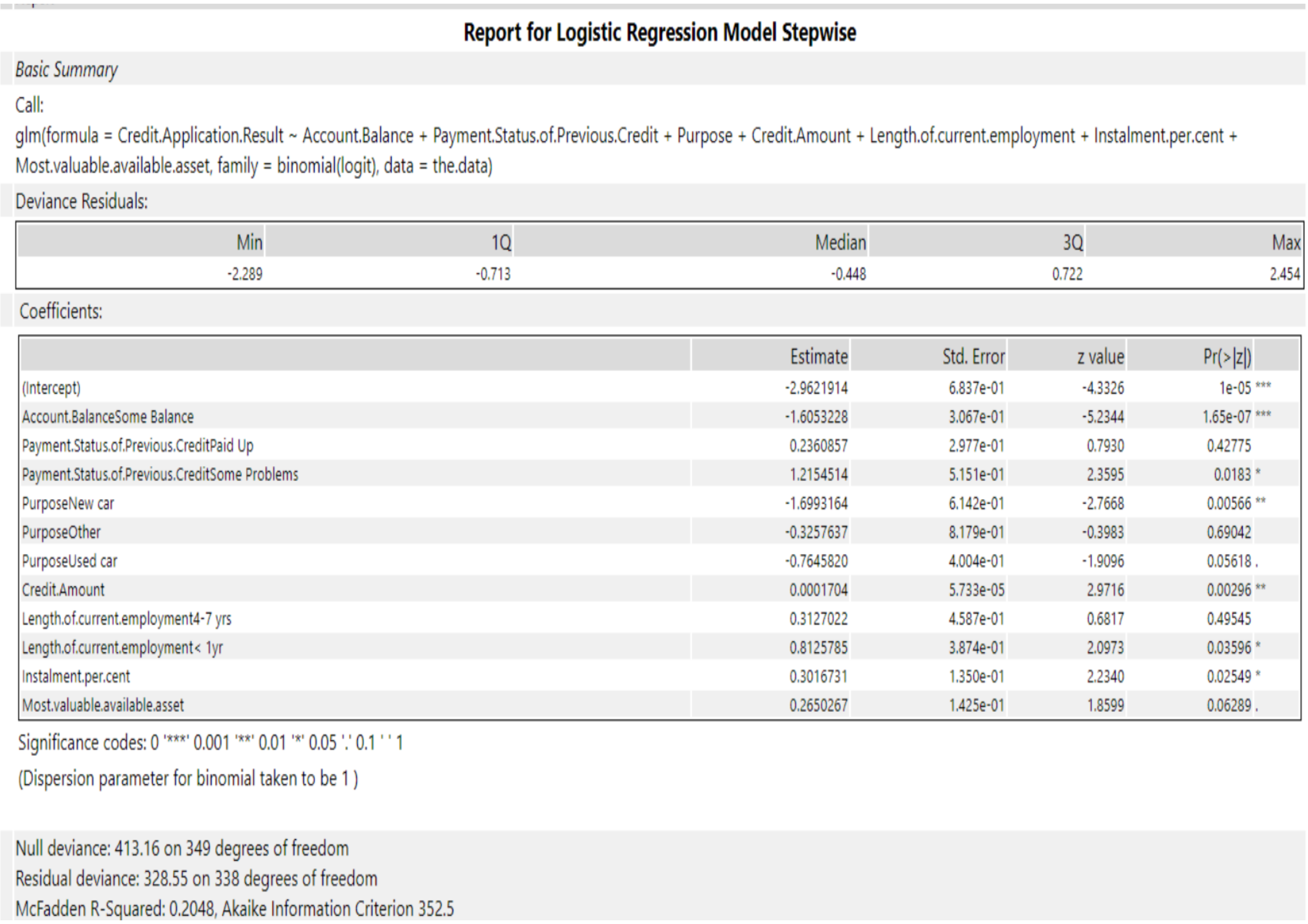
First, create your Estimation and Validation samples where 70% of your dataset should go to Estimation and 30% of your entire dataset should be reserved for Validation. Set the Random Seed to 1.

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 Model – the first three significant variables are Account Balance, Purpose and Credit Amount.



Report for Logistic Regression Model Stepwise

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:

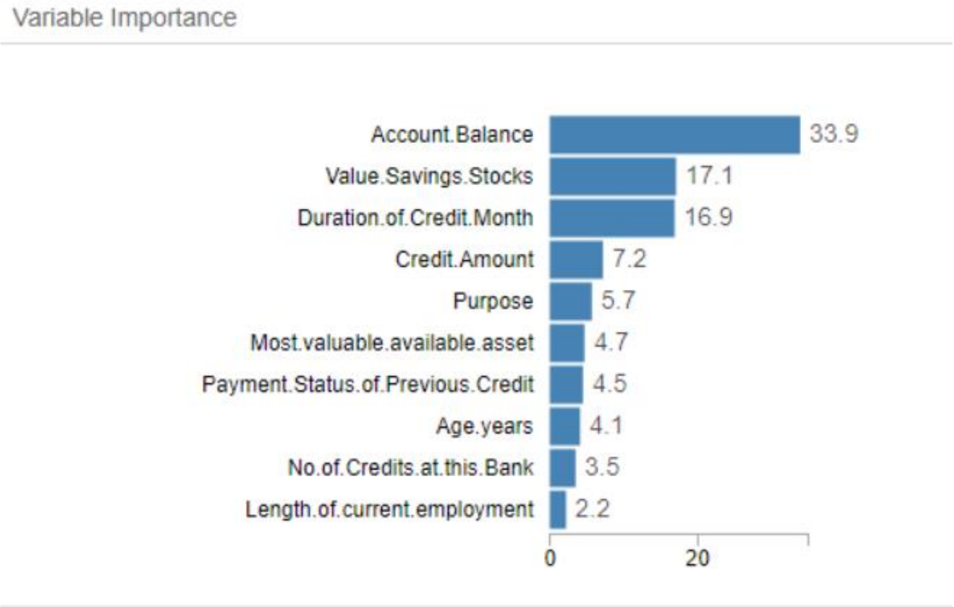
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.9621914	6.837e-01	-4.3326	1e-05 ***
Account.BalanceSome Balance	-1.6053228	3.067e-01	-5.2344	1.65e-07 ***
Payment.Status.of.Previous.CreditPaid Up	0.2360857	2.977e-01	0.7930	0.42775
Payment.Status.of.Previous.CreditSome Problems	1.2154514	5.151e-01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

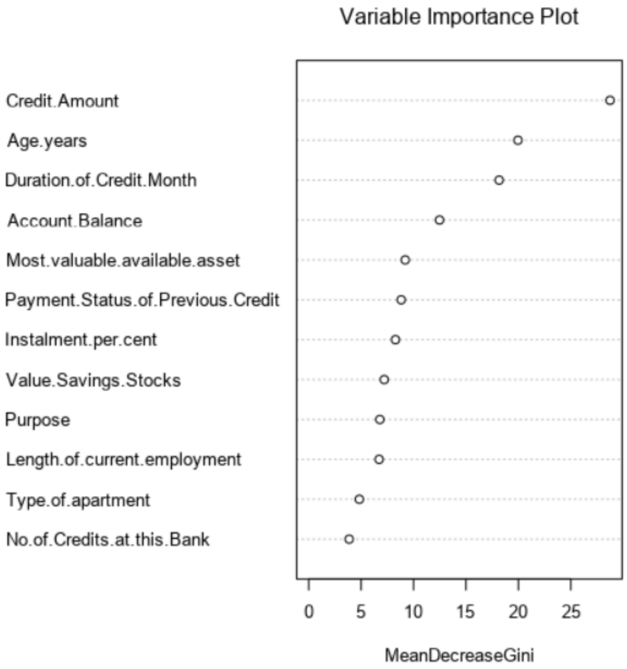
(Dispersion parameter for binomial taken to be 1)

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

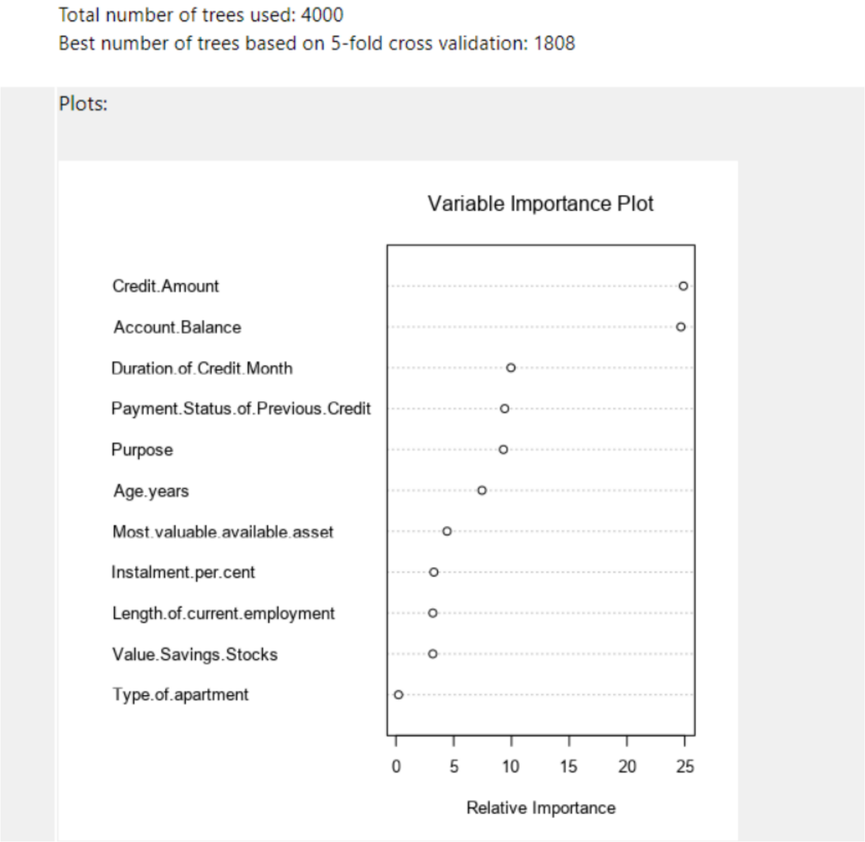
Decision Tree – the three most significant variables are Account Balance, Value Savings Stocks, and Duration of Credit.



Forest Model – the most significant variables are Credit Amount, Age Years, Duration of Credit.



Boosted Model – the most significant variables are Credit Amount, Account Balance, and Duration of Credit.

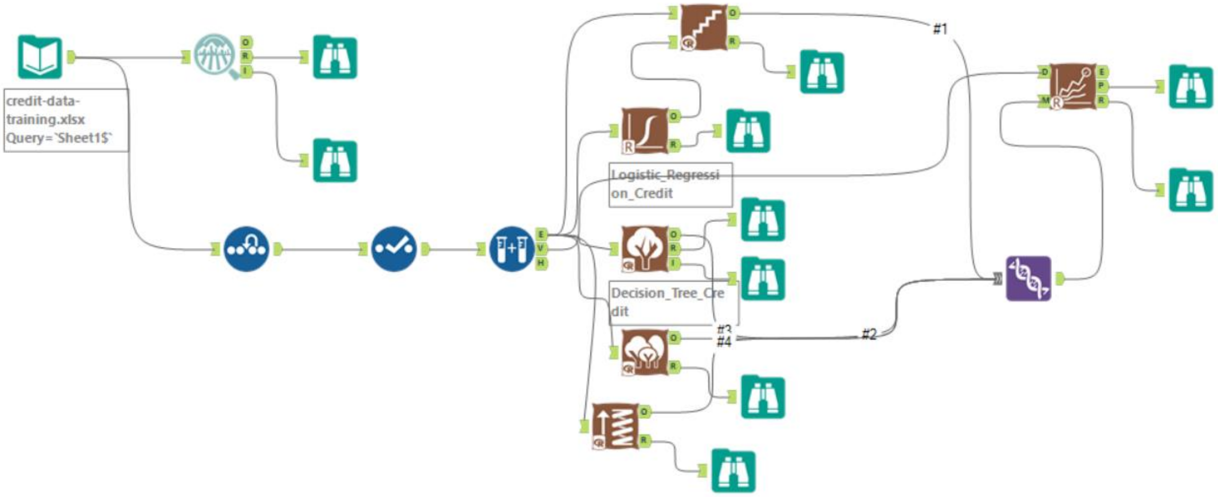


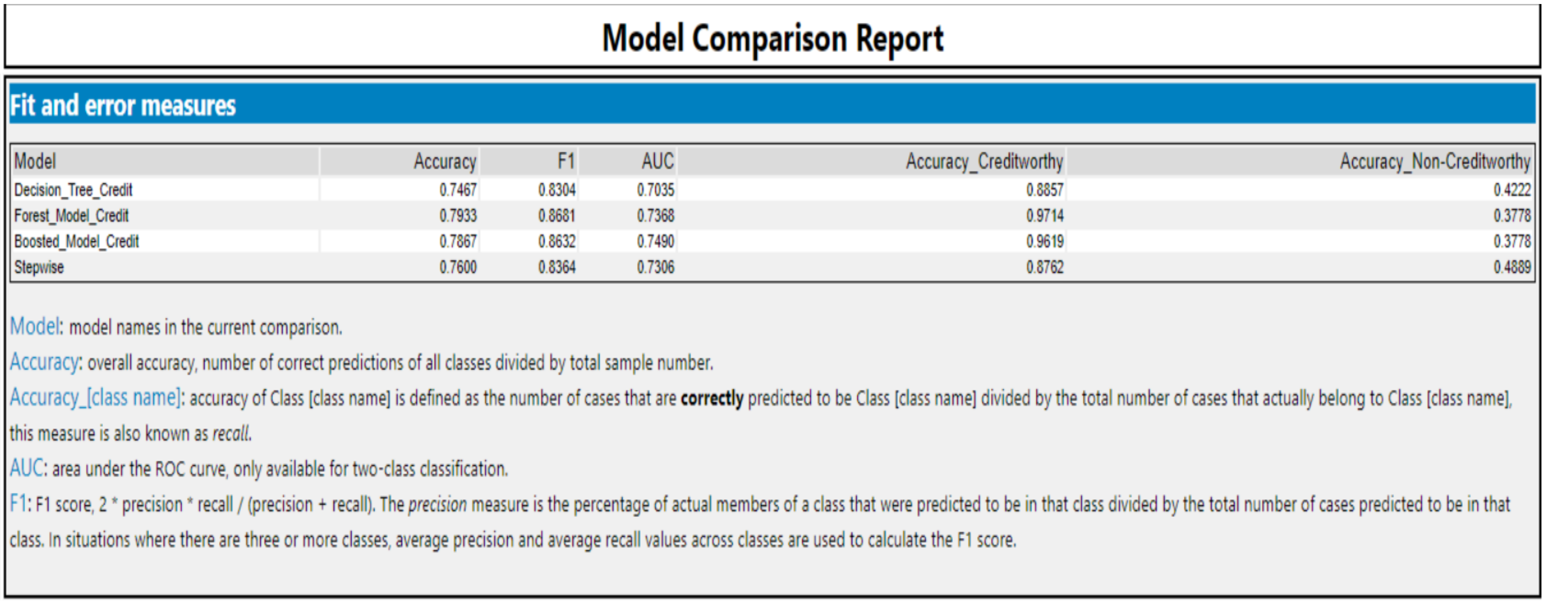
Total number of trees used: 4000
Best number of trees based on 5-fold cross validation: 1808

Plots:

- 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?

By using the Model Comparison Tool, we can validate our models against the Validation set.





Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree_Credit	0.7467	0.8304	0.7035	0.8857	0.4222
Forest_Model_Credit	0.7933	0.8681	0.7368	0.9714	0.3778
Boosted_Model_Credit	0.7867	0.8632	0.7490	0.9619	0.3778
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

Model: model names in the current comparison.

Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.

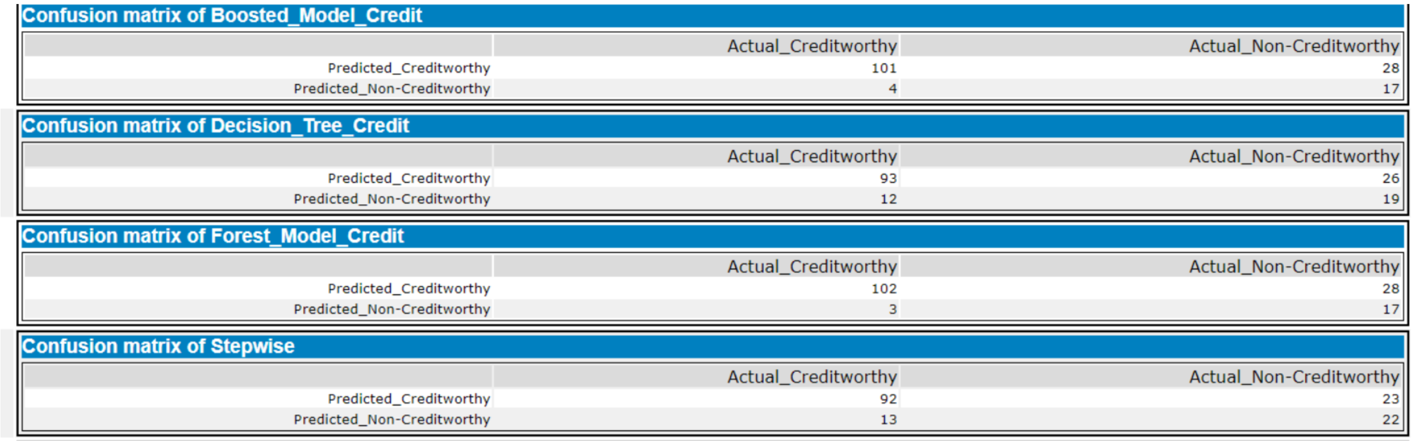
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.

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.

Based on the Model Comparison Report, the overall accuracy is:

- Decision Tree Model – 0.7467
- Forest Model – 0.7933
- Boosted Model – 0.7867
- Logistic Regression (Stepwise) Model – 0.7600



Confusion matrix of Boosted_Model_Credit

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

Confusion matrix of Decision_Tree_Credit

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

Confusion matrix of Forest_Model_Credit

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

Confusion matrix of Stepwise

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

Based on Confusion Matrix, in my opinion all models are biased to **creditworthy.**

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

Step 4: Writeup

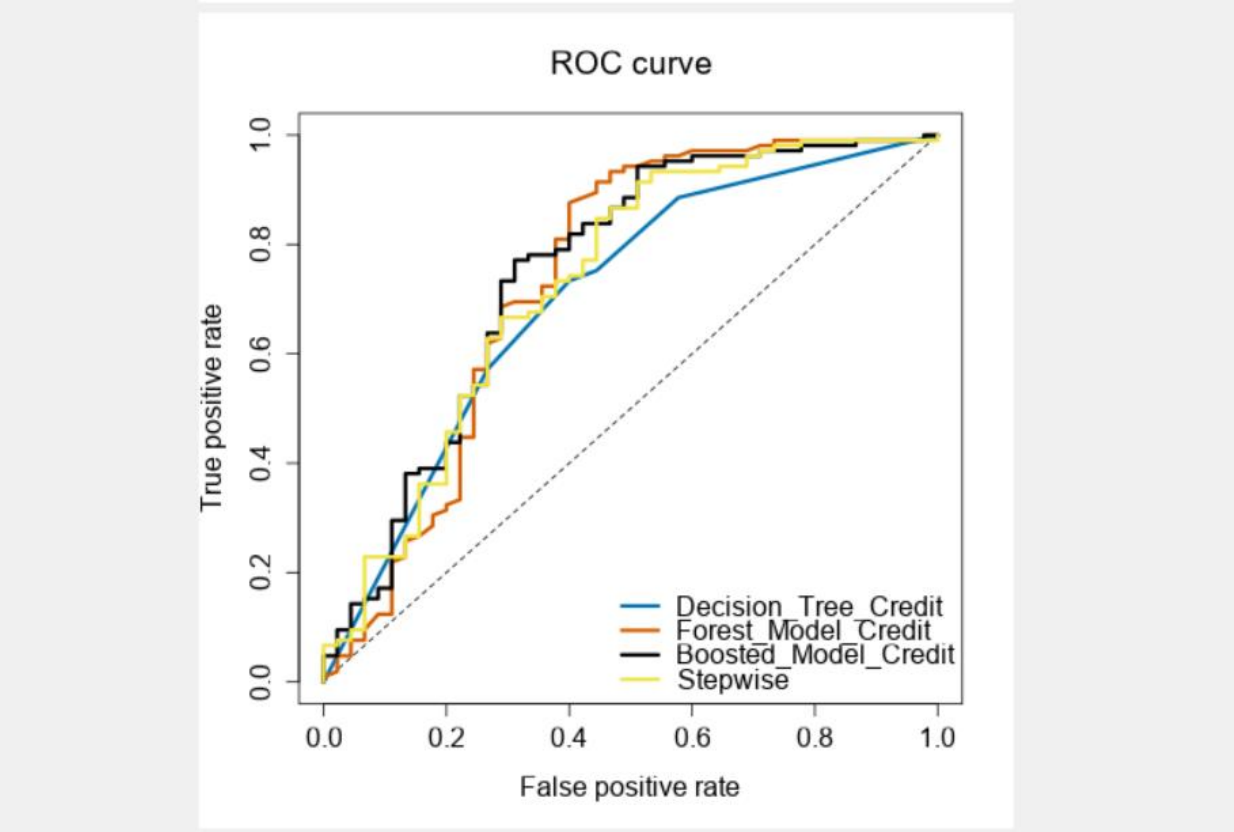
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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)

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

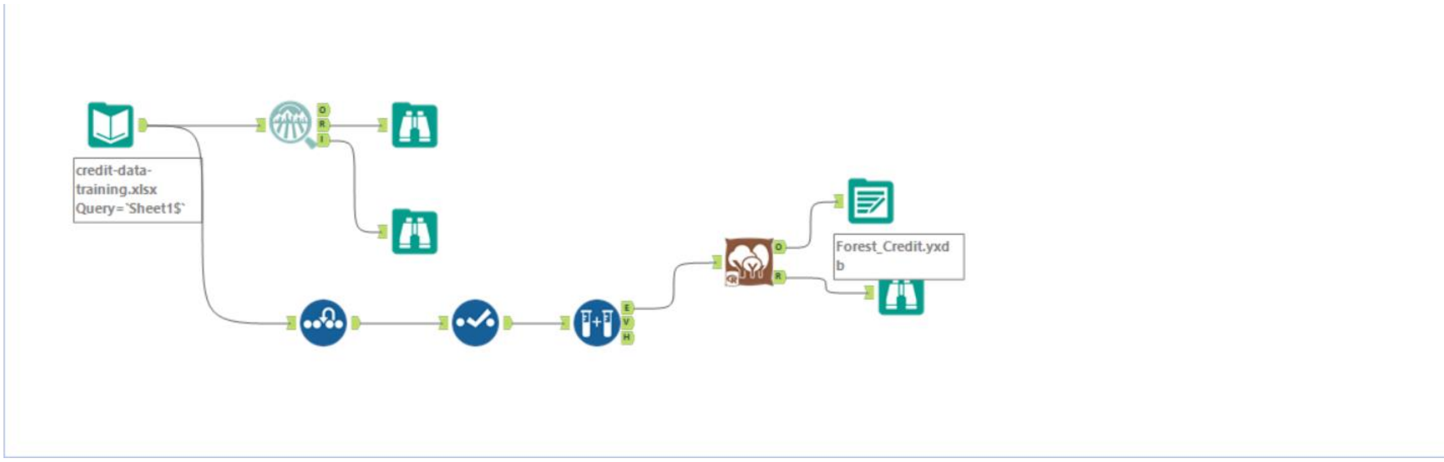
Based on the Model Comparison Report, I chose the **Forest Model.** This model has the highest overall accuracy against our validation set, which is **0.7933**. **Forest Model's** accuracy within "creditworthy" segment is also the highest, **0.9714**. Also, the **ROC Curve Graph,** shows hoe the **Forest Model** has a better performance as it reaches faster the top-left corner towards true-positive rate.

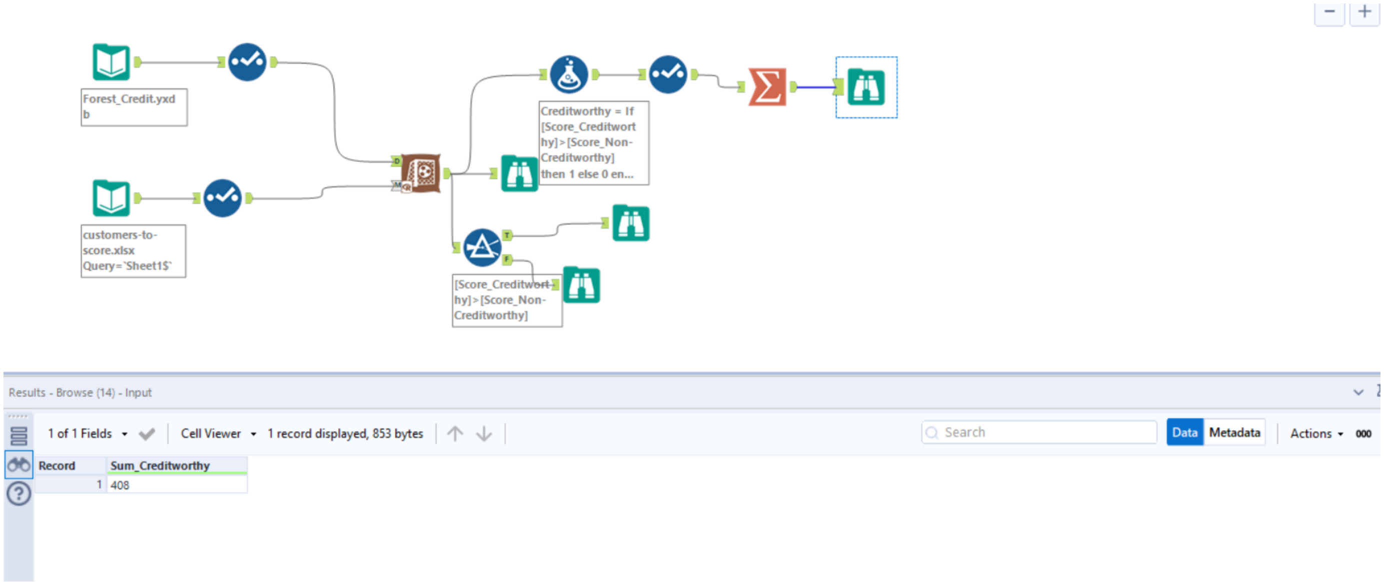


Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

After choosing the model, I have scored our new customers. As a result, I got **408** customers who are creditworthy.





Forest_Credit.yxdb
customers-to-score.xlsx Query=`Sheet1$`
Creditworthy = If [Score_Creditworthy]>[Score_Non-Creditworthy] then 1 else 0 en...
[Score_Creditworthy]>[Score_Non-Creditworthy]
T
F
Results - Browse (14) - Input
1 of 1 Fields
Cell Viewer
1 record displayed, 853 bytes
Search
Data
Metadata
Actions

Record	Sum_Creditworthy
1	408

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.