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?

The objective is to classify the new customers either as creditworthy or non-creditworthy using the dataset of old customers.

- What data is needed to inform those decisions?

The data that we have from first dataset “old costumers” which contains whom bank has provided the loan to. Then, using the new data of the new 500 customers covered by the same variables of the old dataset will classify the customers into creditworthy and non-creditworthy.

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

Since we trying to classify customers into two categories. Binary models will use to make the decisions.

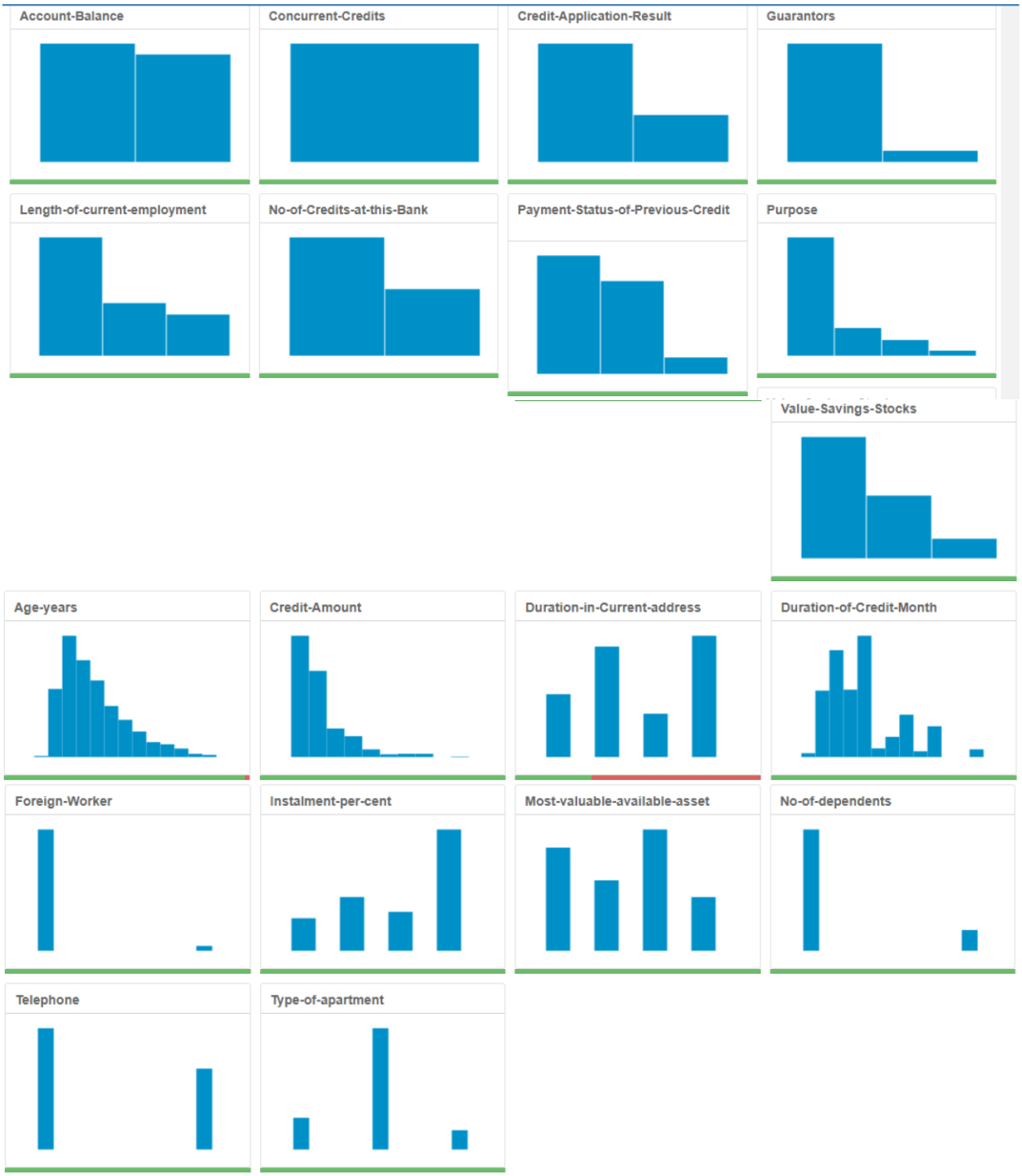
Step 2: Building the Training Set

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:

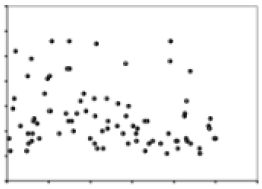
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

First of all, I Used field summary tool to provide a landscape of all variables, the report histograms below shows this:



Occupation	One data type only
Concurrent credit	One data type only
Telephone	Not relevant to classification
Duration in current address	69% missing data
No of dependents	Low variability
Foreign worker	Low variability
Guarantors	Low variability

The variable Age-years has just 2% missing data so it is appropriate to impute the missing data with the median age.

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	

However, I decide to delete these variables and keep the Age-years variable.

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.
- 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?

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

1-Logistic – stepwise model

Which predictor variables are significant or the most important?
the most significant variables with p-value of less than 0.05 as shown below :

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	.

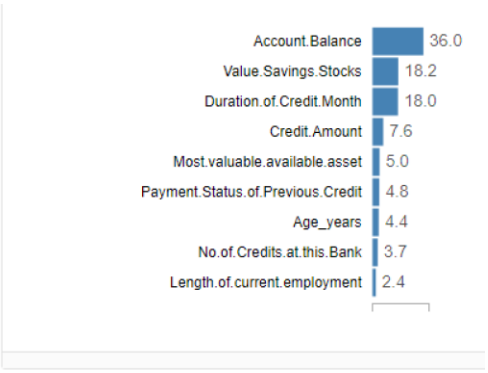
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial taken to be 1)

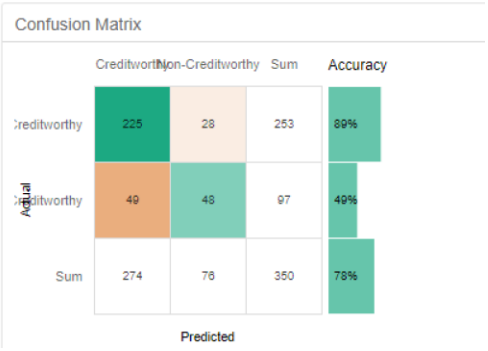
Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

The variables are: account balance, payment status of previous creditsome problems, purposenew car, credit amount, length of employment<1yr, instalment percent.

2-Tree model

Which predictor variables are significant or the most important?
Using the variable importance graph I found 3 top significant predictor variables for decision tree: Account balance, Value saving stocks, duration of credit month.

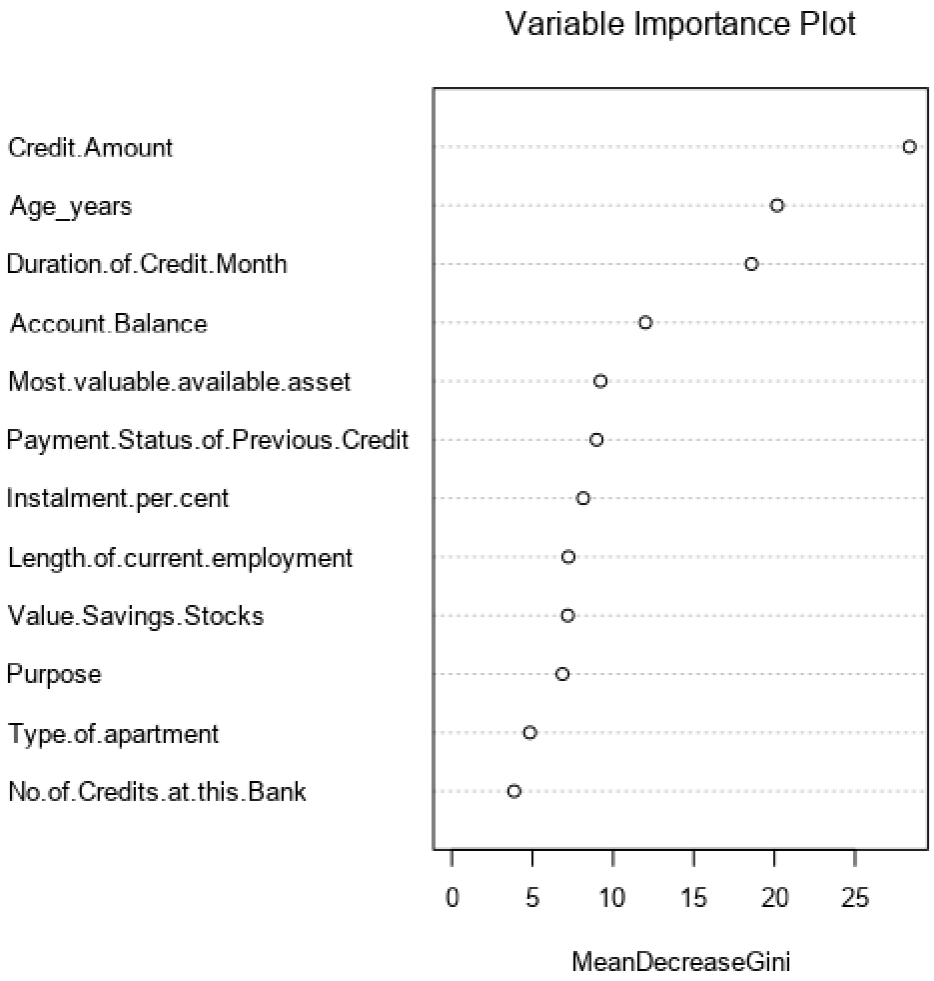




3- Forest model

Which predictor variables are significant or the most important?

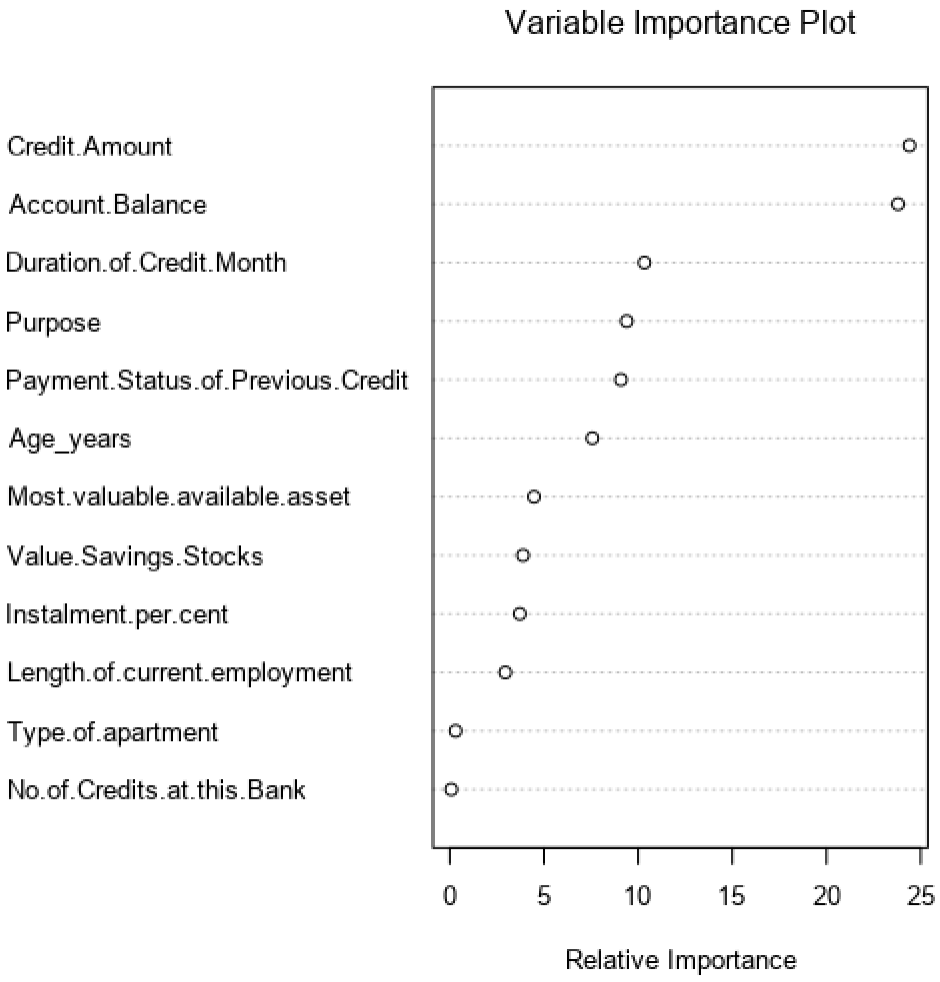
Using Credit Application Result as the target variables, Credit Amount, Age Years and Duration of Credit Month are the 3 most important variables.



4- Boost model

Which predictor variables are significant or the most important?

Using Credit Application Result as the target variables, Credit Amount, account.balance and Duration of Credit Month are the 3 most important variables.



Validation

Model comparison model

Validate your model against the validation set. What was the overall percent accuracy? Show confusion matrix. Are there any bias seen in the models prediction?

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
TREE	0.7467	0.8273	0.7054	0.8667	0.4667
forest	0.8000	0.8707	0.7361	0.9619	0.4222
boost	0.7867	0.8632	0.7524	0.9619	0.3778
X	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.

Model	*Accuracy*	*Accuray_creditworth*	*Accuracy_non_credirworth*
Logistic-stepwise	*76%*	*87%*	*48%*
Decision tree	*74%*	*86%*	*46%*
Forest	*80%*	*96%*	*42%*
Boost	*78%*	*96%*	*37%*

Overall, The Accuracy to predict "creditworthy"is better than "creditNonWorthy". The Cofusion matrix of all models summary shows also a very low accuracy for prediction of NonCreditworthy.

Confusion matrix of TREE

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

Confusion matrix of X

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

Confusion matrix of boost

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

Confusion matrix of forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

Note that model X = Logistic_stepwise model

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:

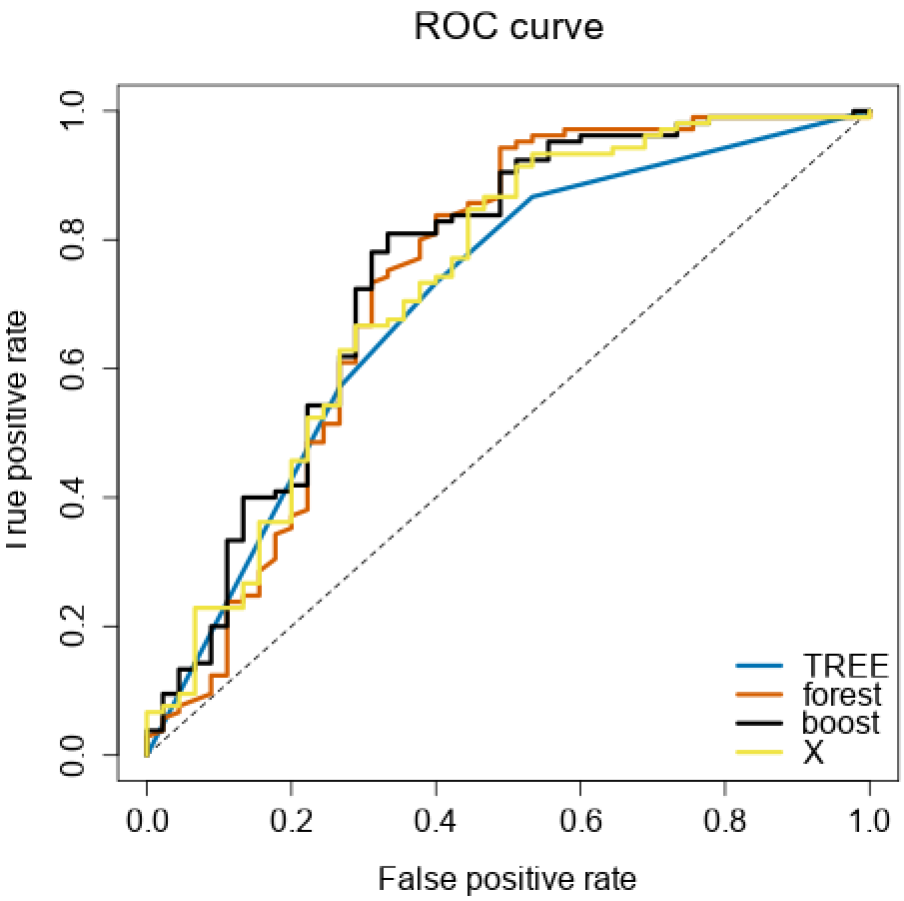
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - Overall Accuracy against your Validation set

Based on model comparison report, It appears that the forest model has the highest accuracy 80%.

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

Accuracy creditworthy rate = 0.9619 Being the high true positive rate and high.

- ROC graph



ROC shows that the forest model reached the positive rate fastest hence this also gives the good reason to select it.

- Bias in the Confusion Matrices

Confusion matrix of TREE		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of X		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

Confusion matrix of boost		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	28
Predicted_Non-Creditworthy	4	17

Confusion matrix of forest		
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	26
Predicted_Non-Creditworthy	4	19

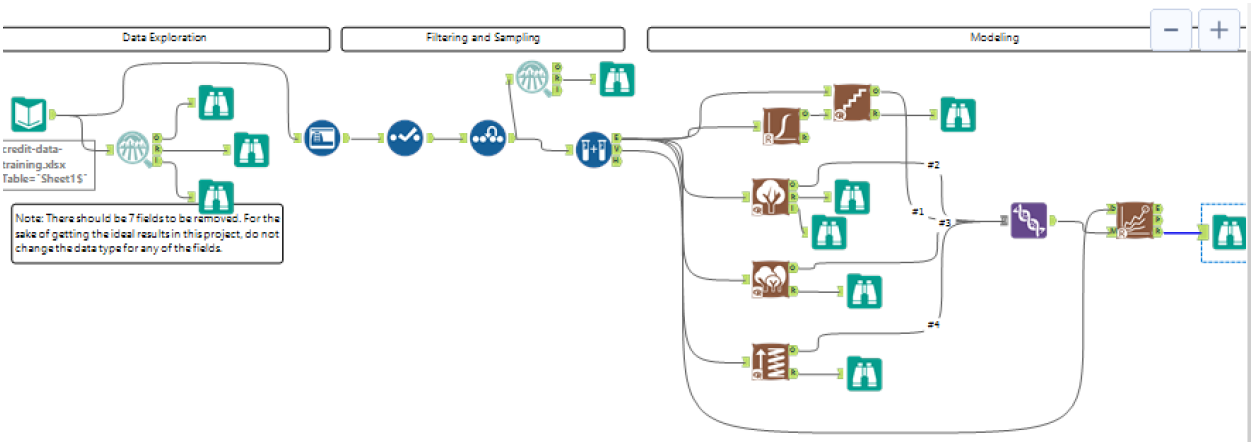
The accuracy difference between creditworthy and non-creditworthy are also comparable which makes it least bias towards any decisions

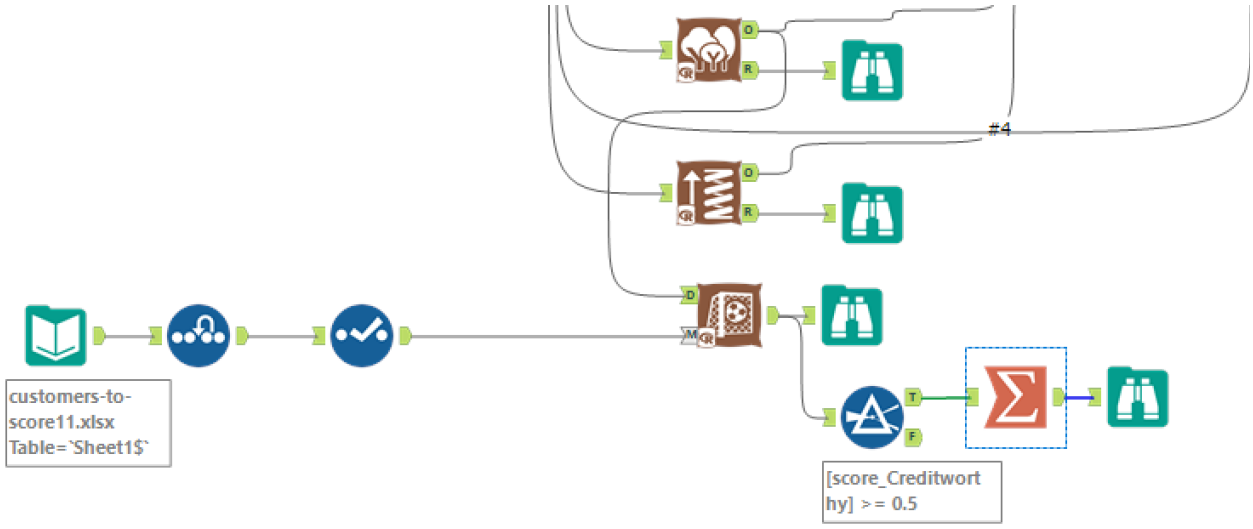
Note: Remember that your boss only cares about prediction accuracy for Creditworthy and Non-Creditworthy segments.

- How many individuals are creditworthy?

There are **408 creditworthy customers** using forest models to score new customers.

Workflow





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