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

1. What decisions needs to be made?
   ANSWER: Out of a list of new 500 loan applicants, We need to decide , the number of people who are creditworthy.

2. What data is needed to inform those decisions?

ANSWER: We need Financial records of existing customers and of that of the possible clients. Besides , we need Income data, previous credit history , default history , net worth etc and other such inputs which impact a person's creditworthiness.

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

ANSWER: We can use a Binary models like a Logistic regression & non binary models like decision trees , boosted model & Forest model to help make these 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.***

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 average 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:

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

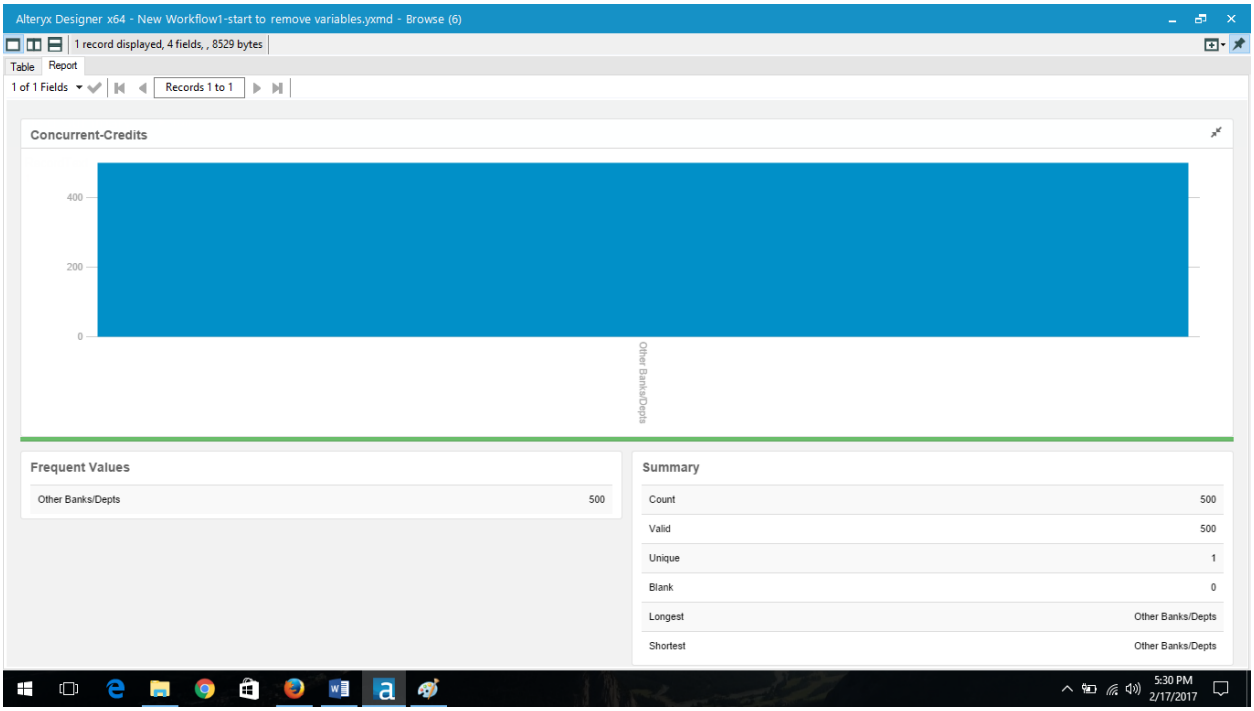
ANSWER:

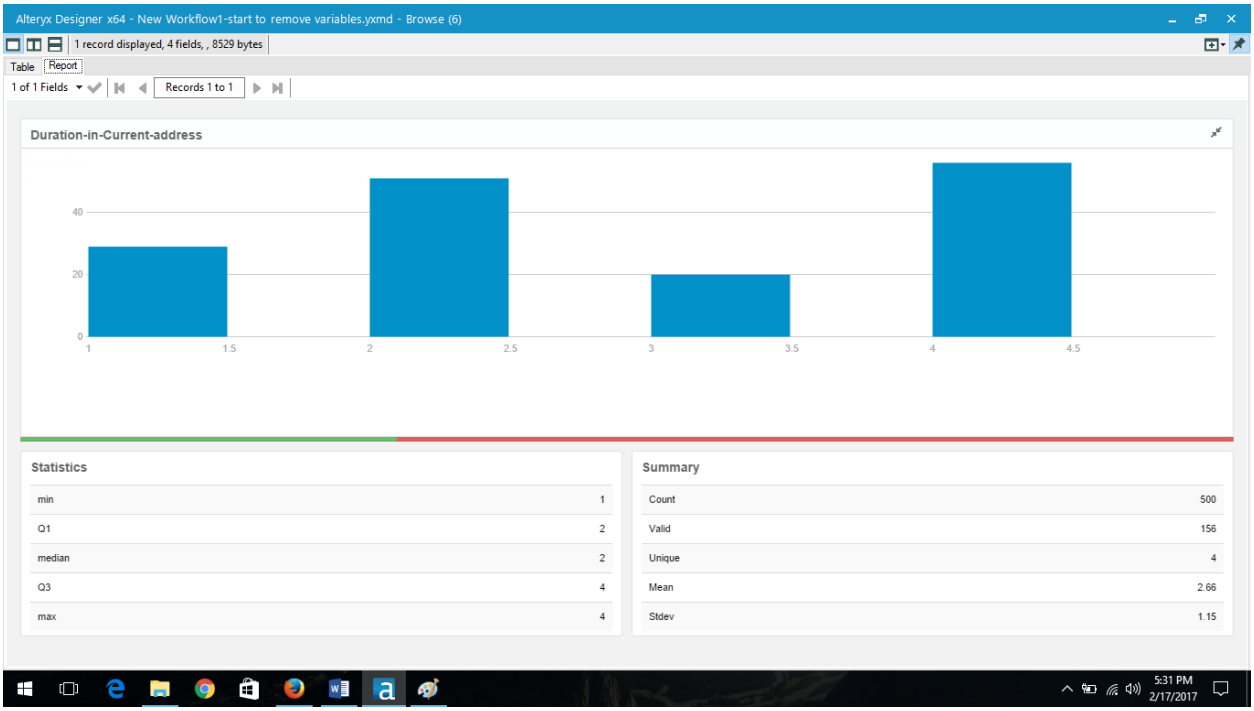
PREDICTOR VARIABLE REMOVED	REASON FOR REMOVAL
Duration in Current Address	*Too much missing data*
Foreign Worker	*Logical & Variable results are too skewed in one way.*
Concurrent credits	*Only one value in the variable*
No. of dependents	*Logical. Has no impact on a person's creditworthiness.*

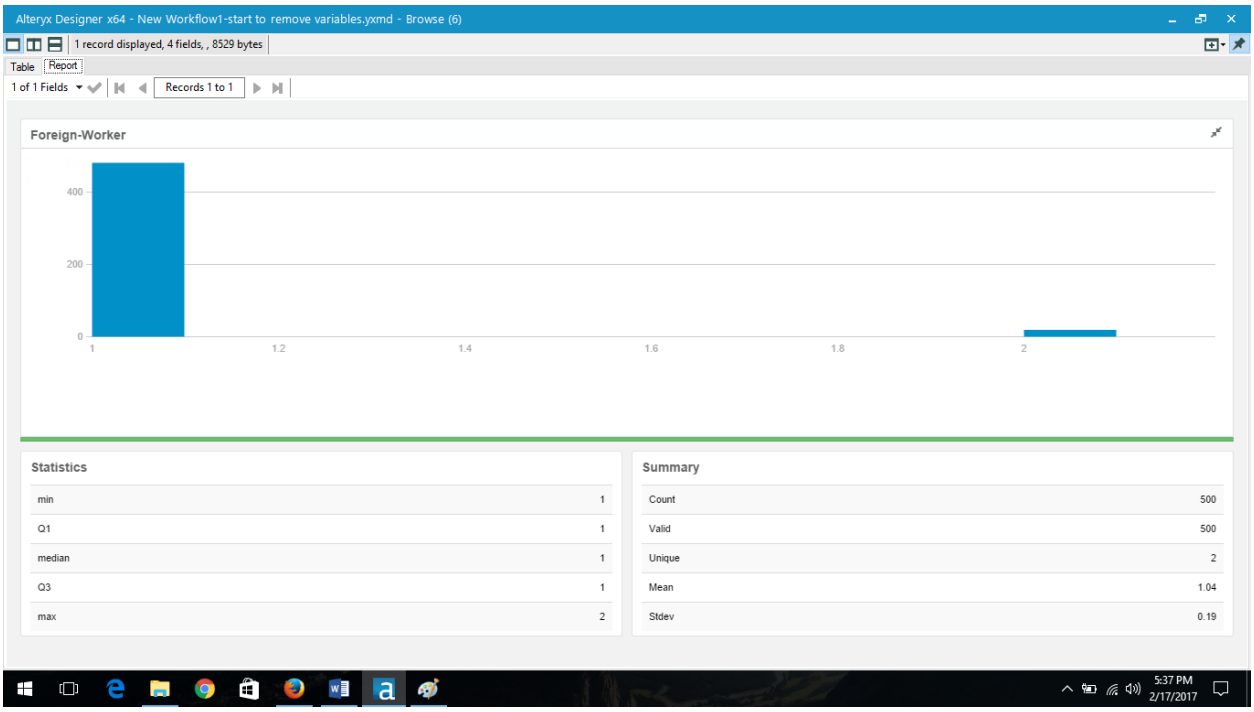
Guarantors	*Variable results are too skewed in one way.*
Telephone	*Logical. Has no impact on a person's creditworthiness*
Occupation	*Only one value in the variable*

I also imputed the missing data in the "Age-years" field as only 2% data was missing and imputation would be the better option rather than removal, as in this case it would help build a model that would be closer to the reality. For this to happen, the missing Age fields would have to be filled in with the "average with null "values derived using the summarize tool and planted in the table using the formula tool.

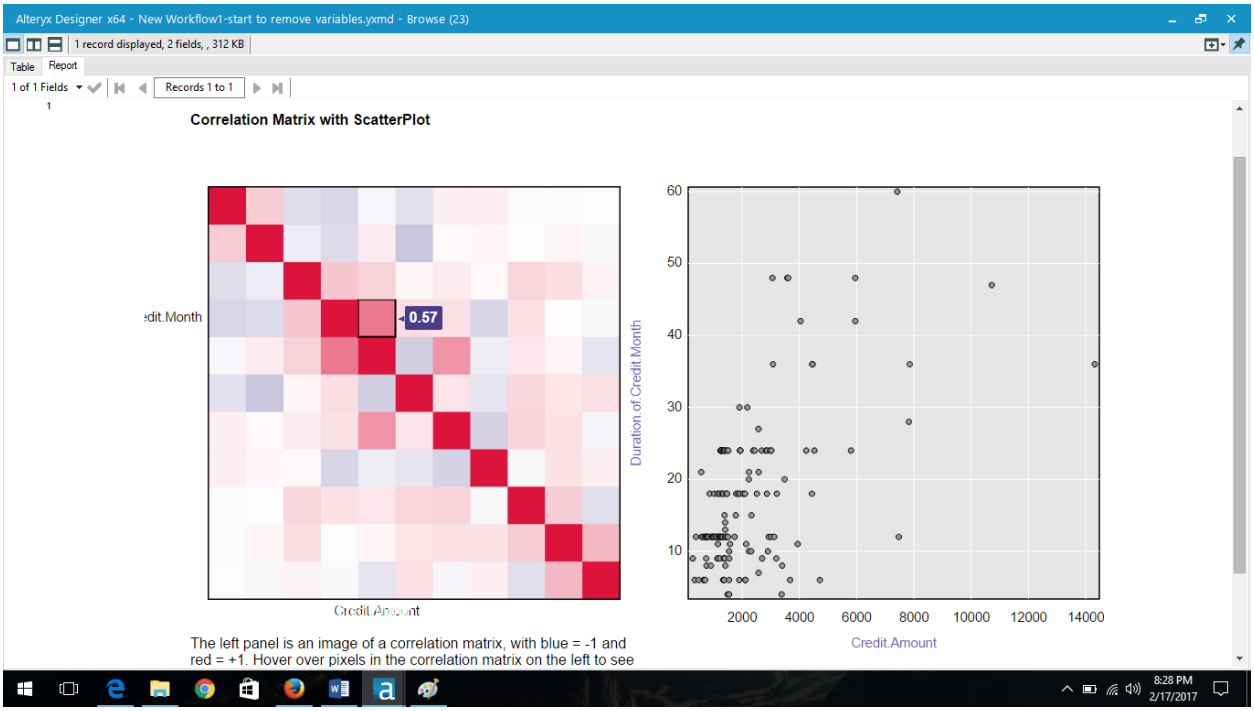
Below are 3 sample visualizations that support the above answer:







I did not find numerical fields that highly correlate with each other(More than 0.70), Here is the scatterplot image below:



The maximum co-relation was between the fields “Credit amount” and Duration of credit month”. It is 0.57, as can be seen from the image above.

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:*

Confusion matrix of BoostedModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DTree

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

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Confusion matrix of Stepwise_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	32
Predicted_Non-Creditworthy	13	13

1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

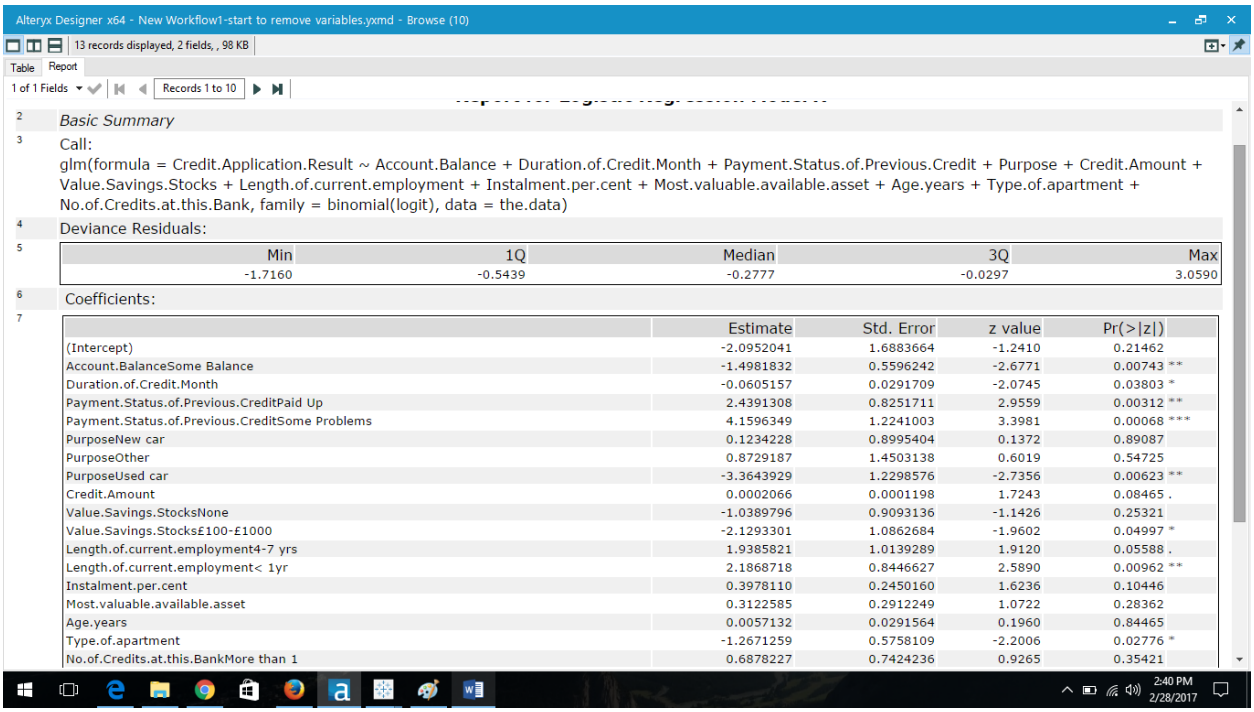
2. Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

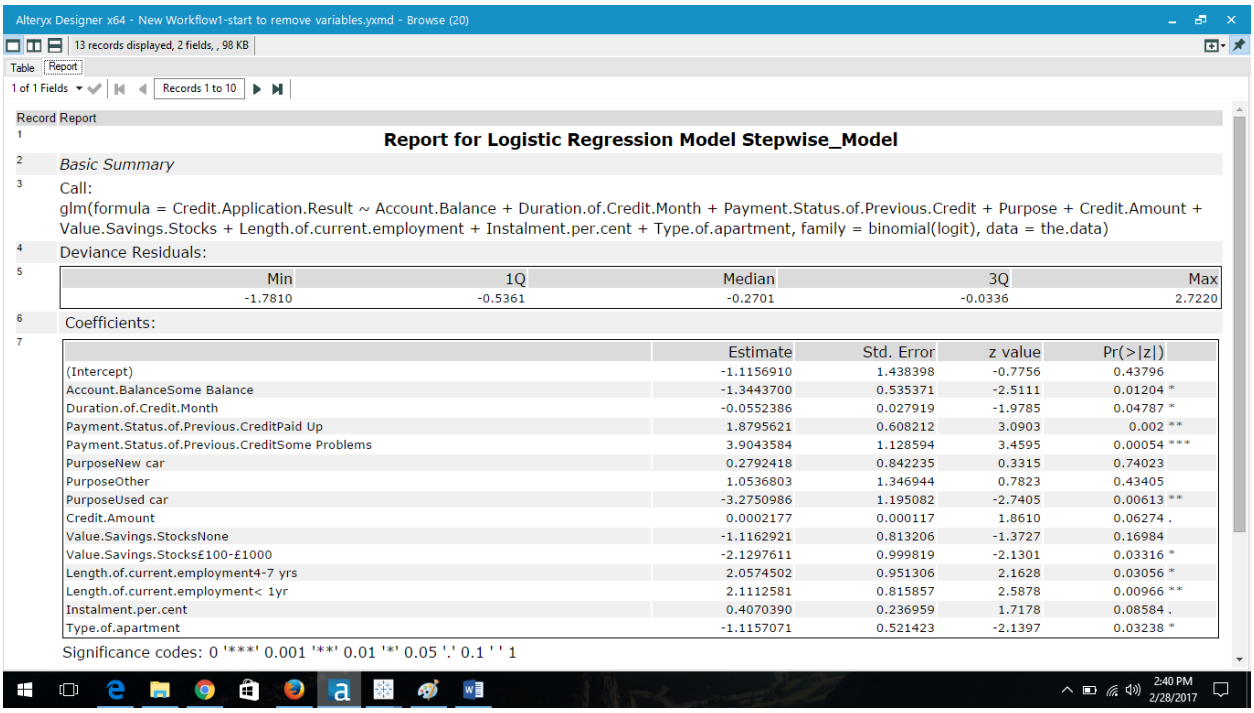
You should have four sets of questions answered. (500 word limit)

ANSWER:

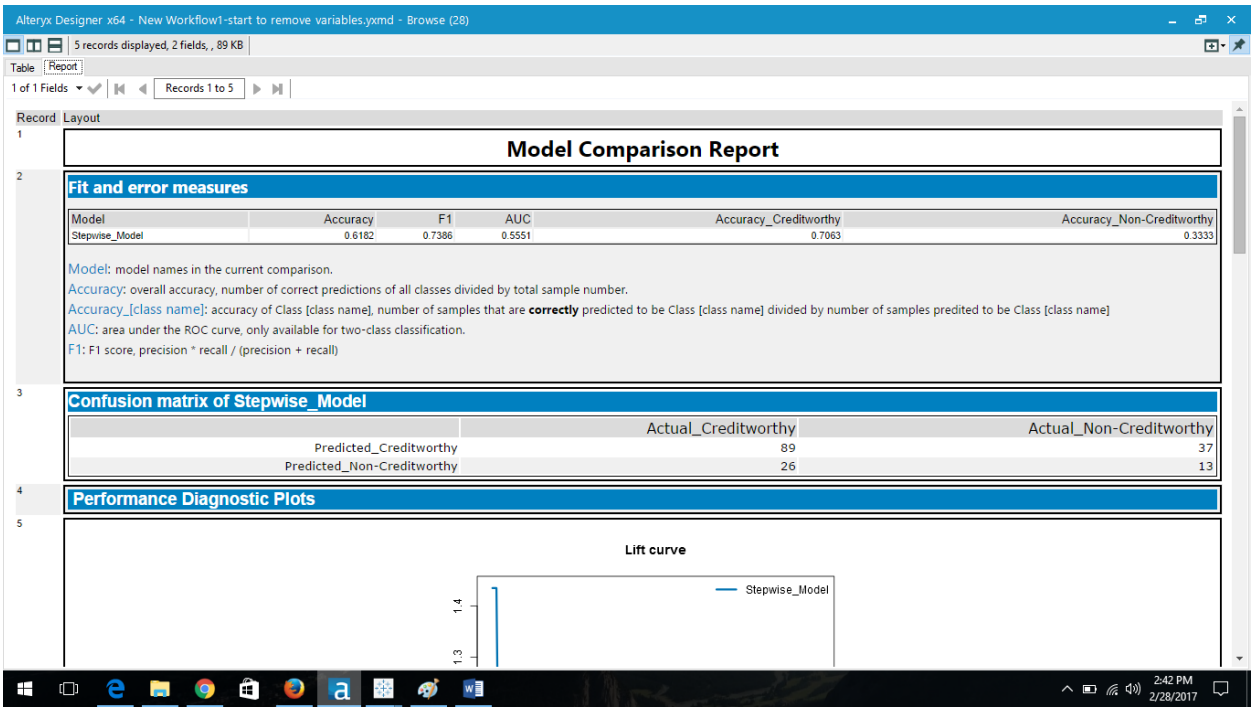
MODEL 1: LOGISTIC REGRESSION & STEPWISE LOGISTIC REGRESSION

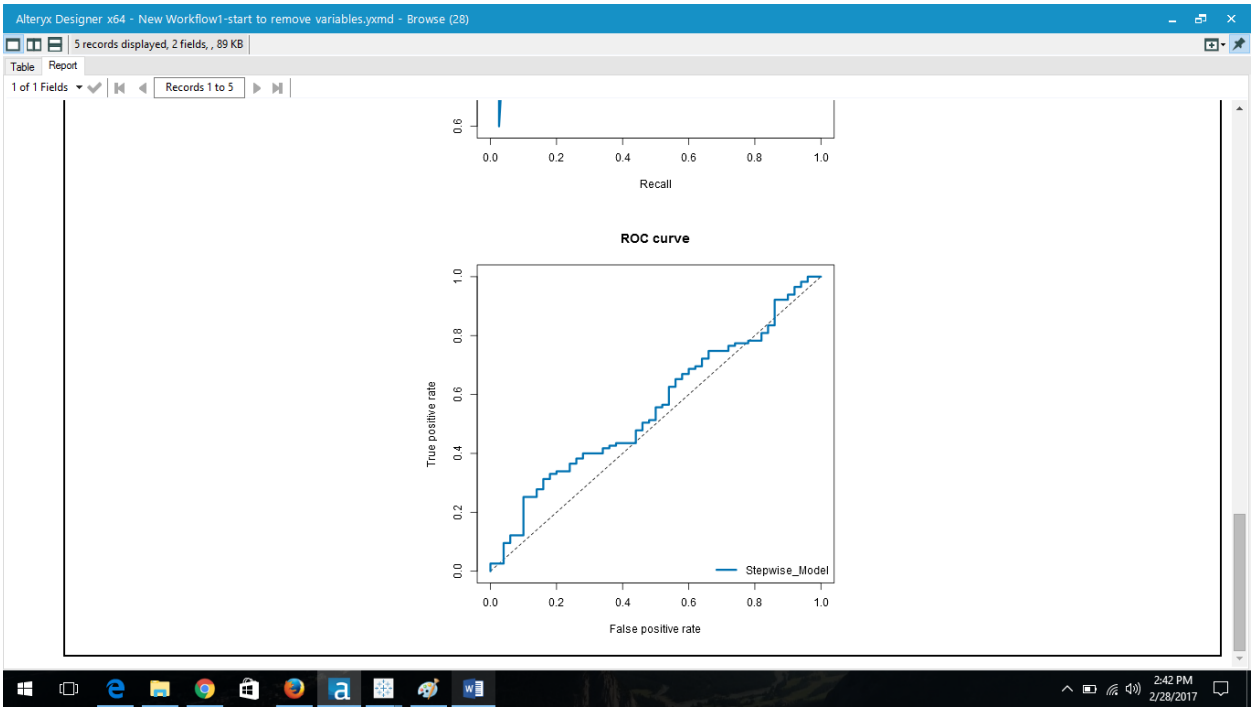
I first ran the logistic regression with all the possible predictor variables and I then ran the stepwise logistic regression tool so that it automatically selects the most significant variables. (Screenshots below)





Based on the model comparison report of the estimated dataset and the validation samples, the result is as follows:





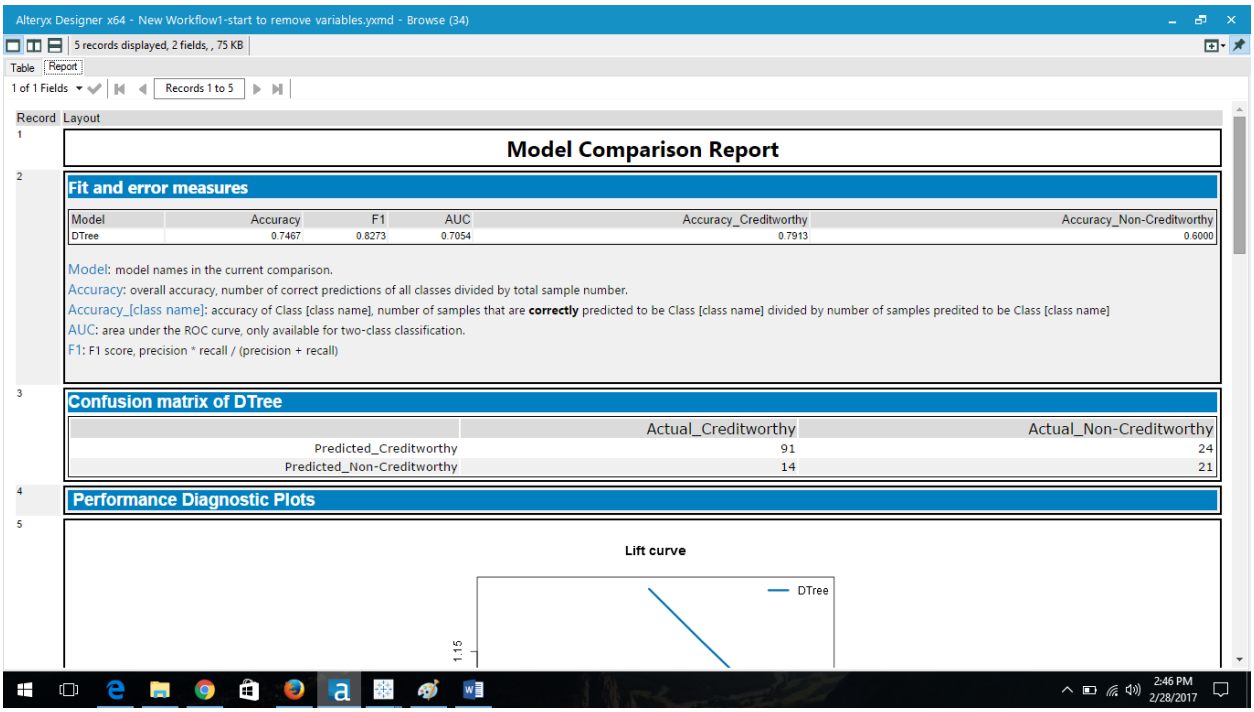
To summarize: Accuracy of the model = 61.82%
Creditworthy predicted accurately: 70.63%

MODEL 2: DECISION TREE

Most Important variables as per this model = Account balance, Credit Amount, Duration of credit month, Value of Savings stocks amongst others as seen below

Actual Decision tree

Confusion matrix



To summarize:

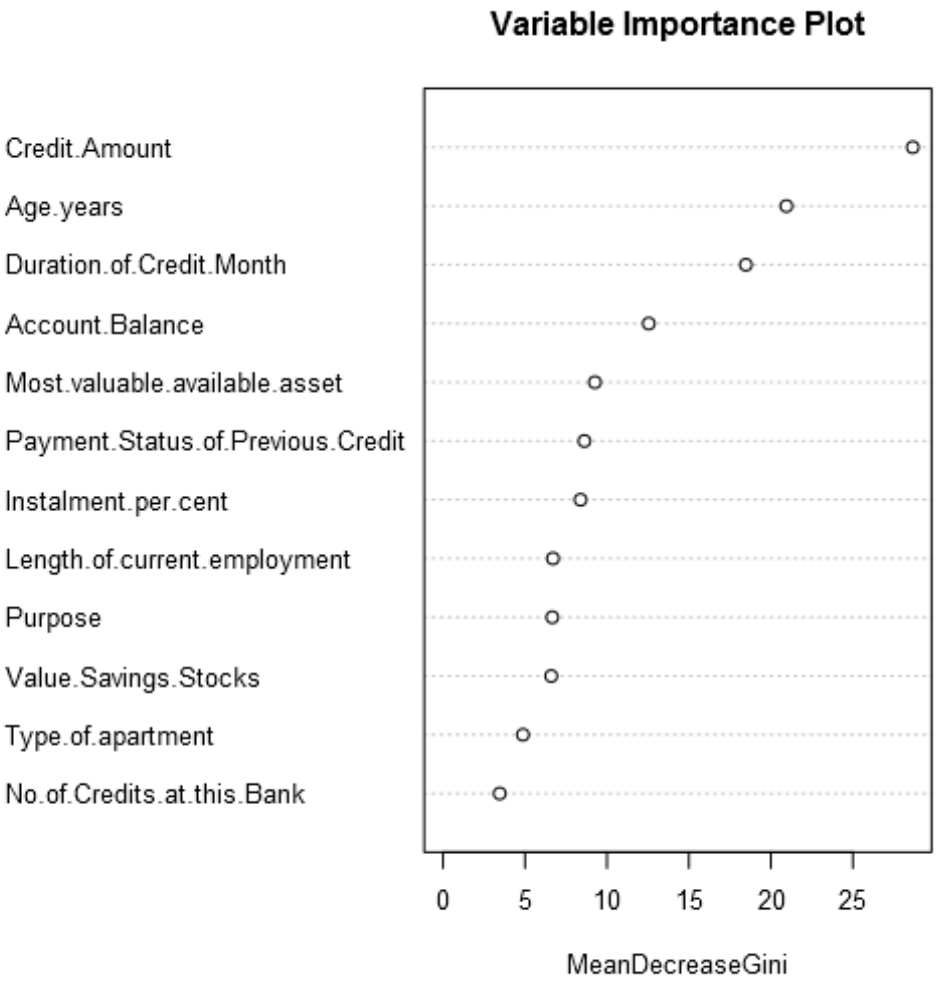
Accuracy of the model = 74.67%

Creditworthy predicted accurately: 79.13%

MODEL 3: Forest Model

Variables importance plot

Most Important variables – Credit Amount, Age years, Duration of credit month ,Account balance ,most valuable asset amongst others



Alteryx Designer x64 - New Workflow1-start to remove variables.yxmd - Browse (38)

9 records displayed, 2 fields, , 71 KB

Table | Report

1 of 1 Fields | Records 1 to 9

Record | Report

1 *Basic Summary*

2 Call:
randomForest(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, ntree = 500)

3 Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3

4 OOB estimate of the error rate: 36.6%

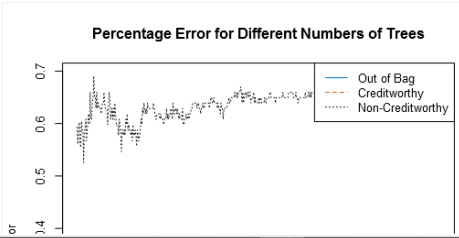
5 Confusion Matrix:

6

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.083	232	21
Non-Creditworthy	0.649	63	34

7 *Plots*

8



2:49 PM
2/28/2017

Alteryx Designer x64 - New Workflow1-start to remove variables.yxmd - Browse (40)

5 records displayed, 2 fields, , 89 KB

Table | Report

1 of 1 Fields | Records 1 to 5

Record | Layout

1 **Model Comparison Report**

2 **Fit and error measures**

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
ForestModel	0.8200	0.8831	0.7430	0.8095	0.8750

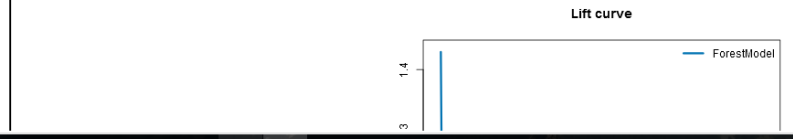
Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name], number of samples that are **correctly** predicted to be Class [class name] divided by number of samples predited to be Class [class name]
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, precision * recall / (precision + recall)

3 **Confusion matrix of ForestModel**

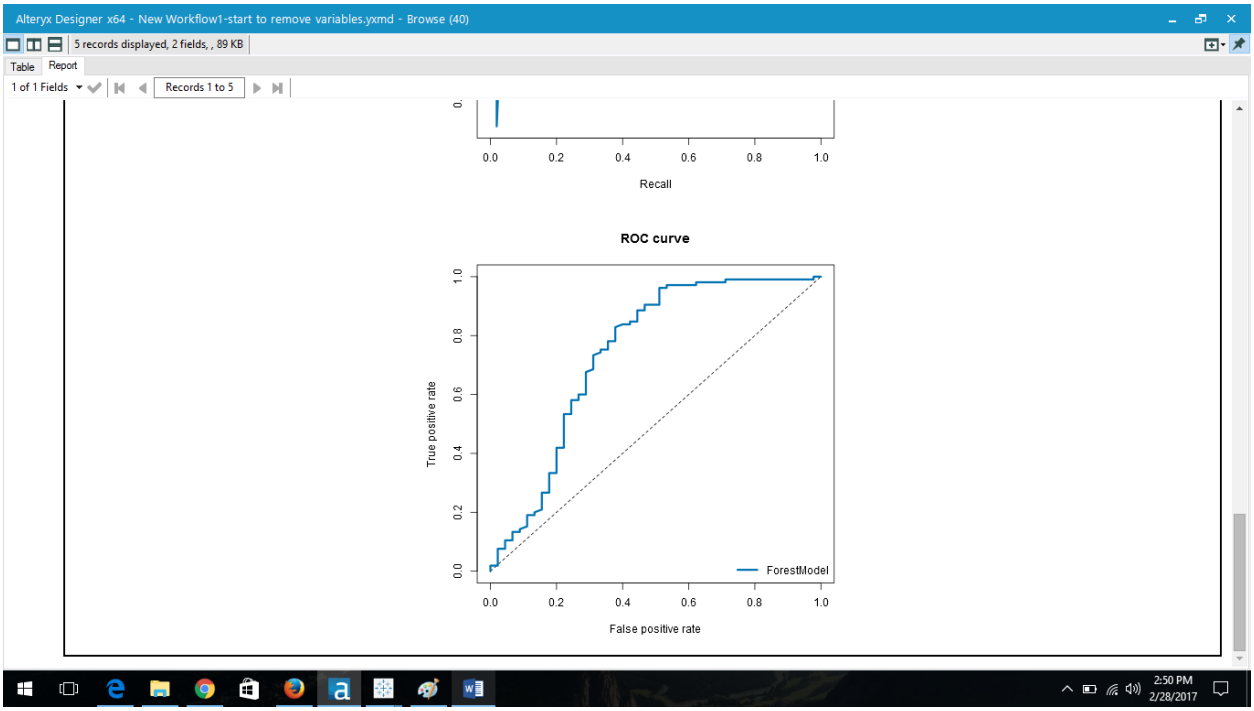
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

4 **Performance Diagnostic Plots**

5



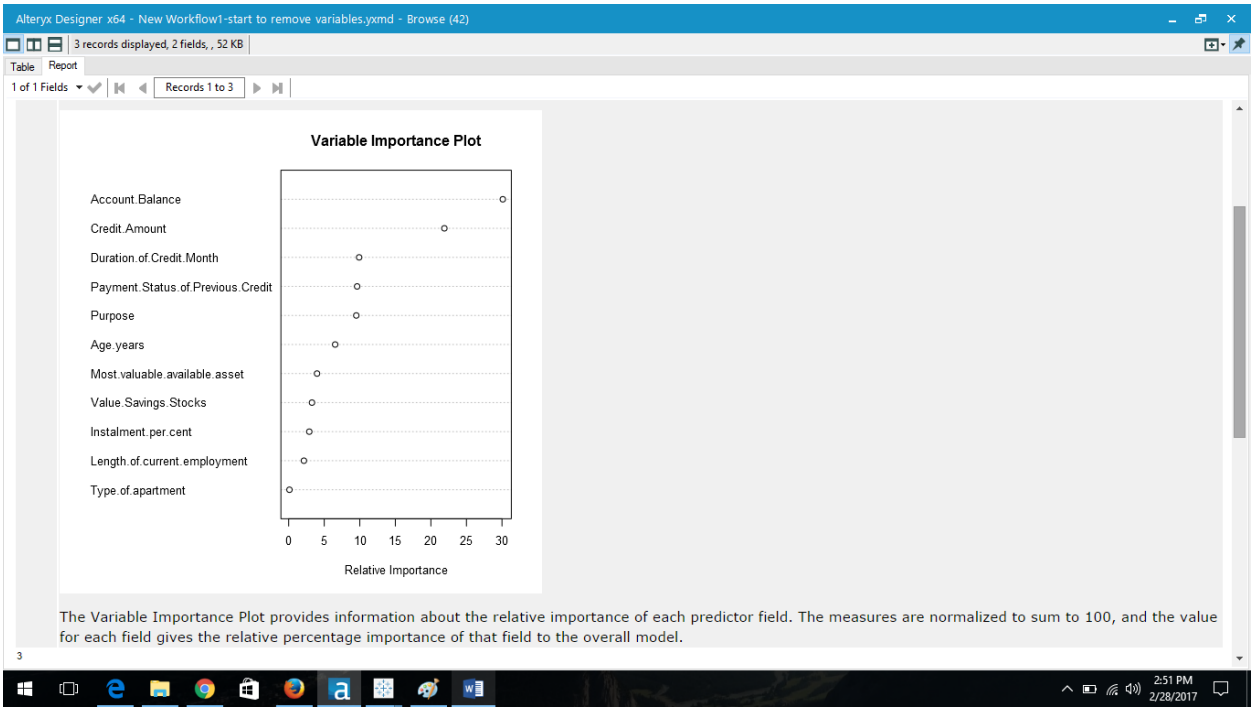
2:50 PM
2/28/2017



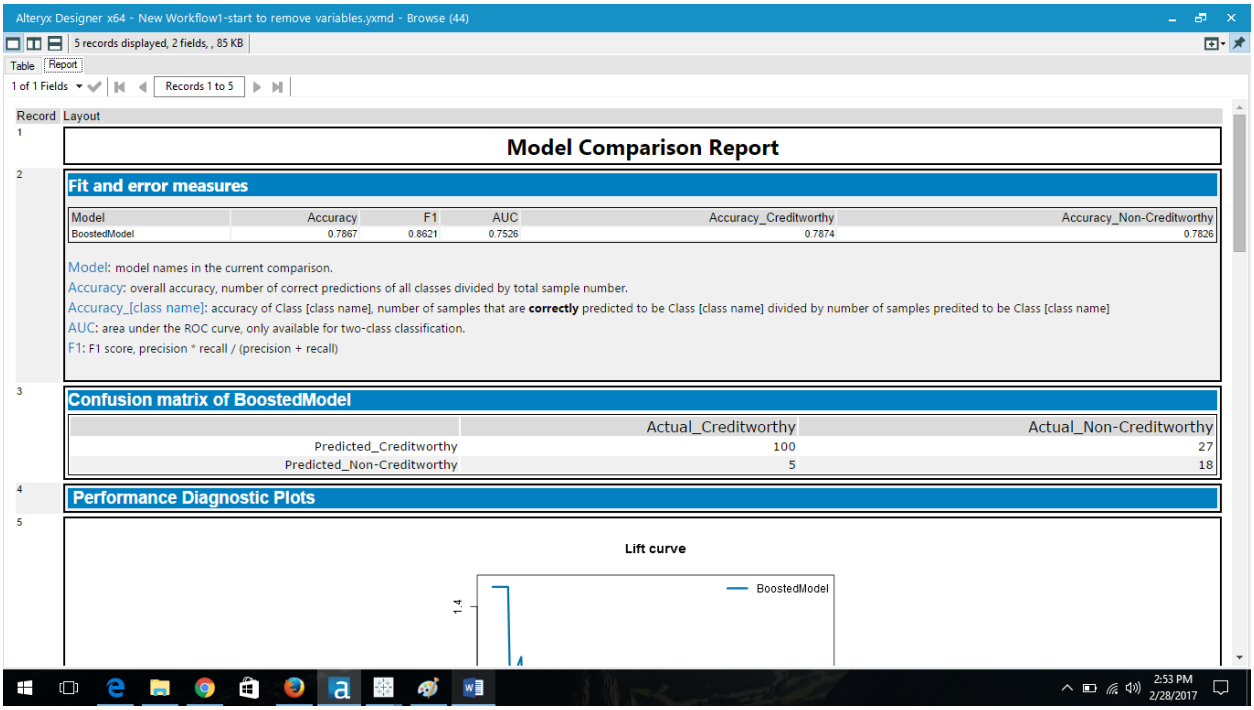
To summarize: Accuracy of the model: 82%
Creditworthy predicted accurately: 80.95%

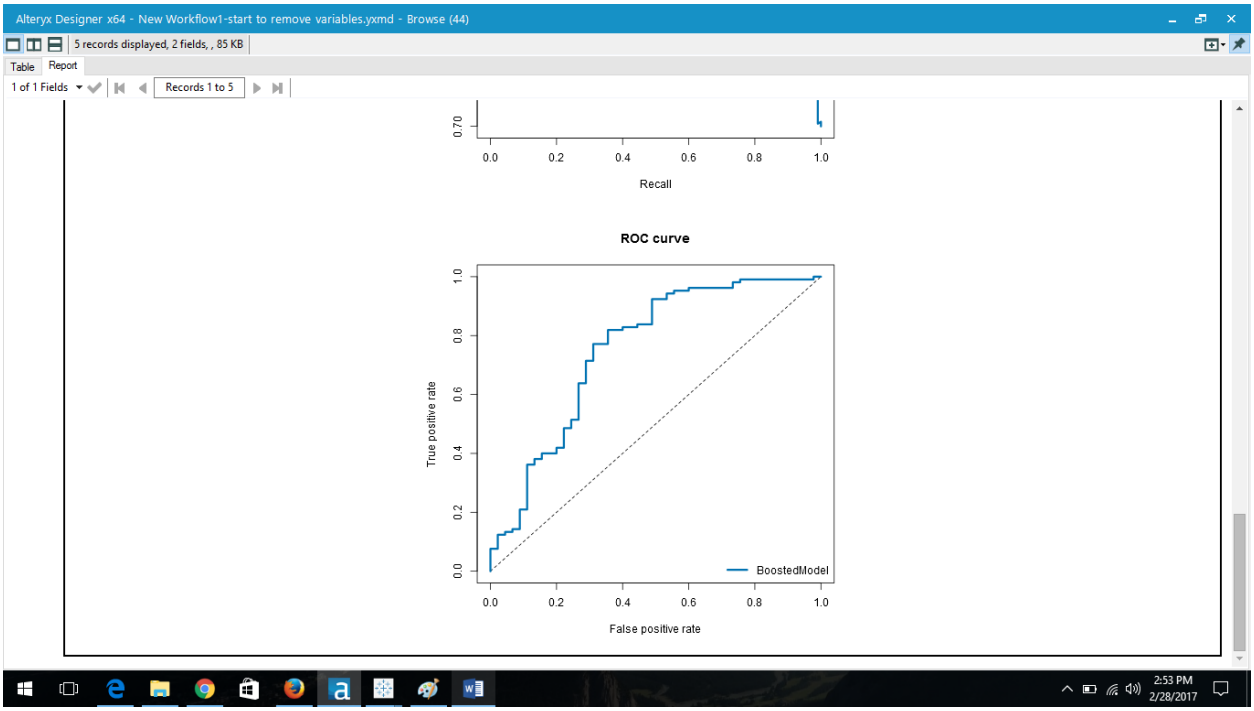
MODEL 4: Boosted Model

Variable importance plot
Most Important Variables: Account balance, credit amount ,Duration of credit month, Payment status of Previous credit ,Purpose amongst others



Model comparison report





To summarize: Accuracy of the model: 78.67%
Creditworthy predicted accurately: 78.74%

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:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    c. ROC graph
    d. Bias in the Confusion Matrices

ANSWER:

Let’s begin with the reports,

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
ForestModel	0.8200	0.8831	0.7430	0.8095	0.8750
Stepwise_Model	0.7000	0.8035	0.6142	0.7419	0.5000
DTree	0.7467	0.8273	0.7054	0.7913	0.6000
BoostedModel	0.7867	0.8621	0.7526	0.7874	0.7826

Confusion matrix of BoostedModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of DTree

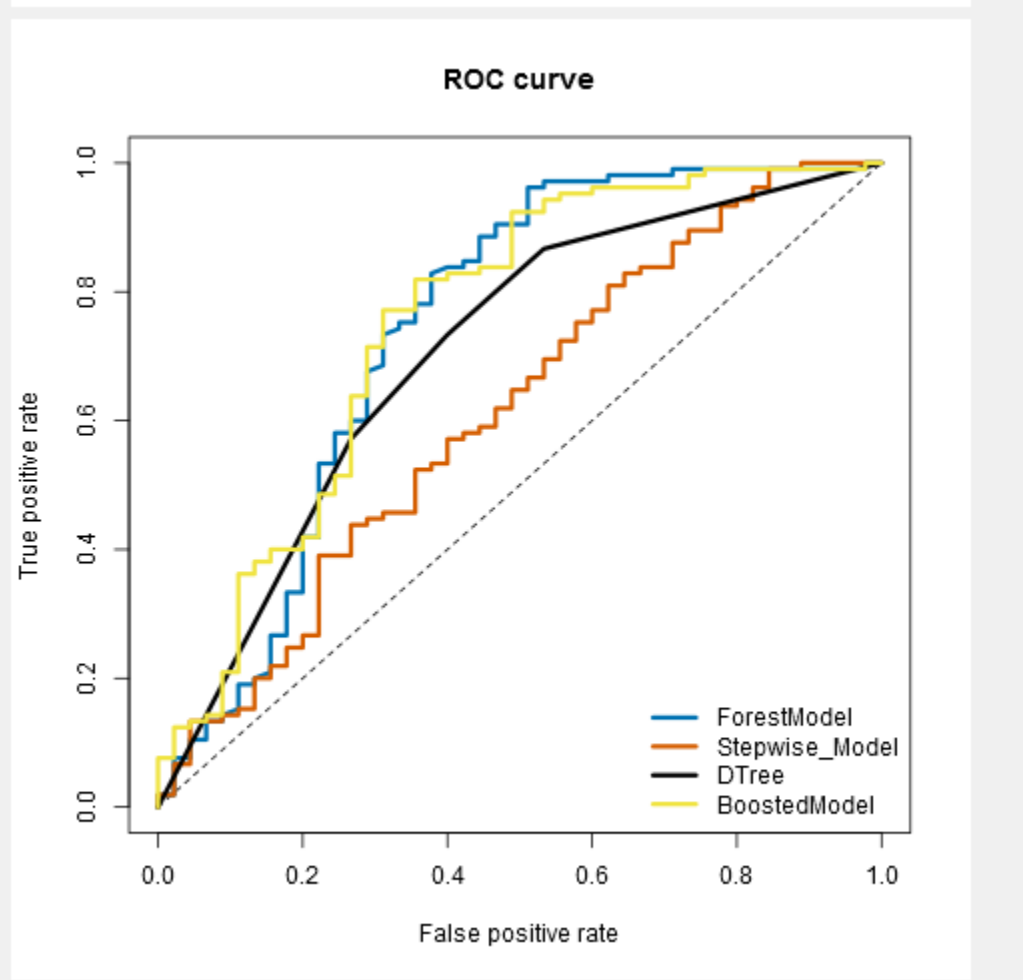
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	91	24
Predicted_Non-Creditworthy	14	21

Confusion matrix of ForestModel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	24
Predicted_Non-Creditworthy	3	21

Confusion matrix of Stepwise_Model

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	32
Predicted_Non-Creditworthy	13	13



Based on the above reports, Considering the overall accuracy of the model & the accuracy of the “Creditworthy” applicants validated, I conclude that the Forest Model is the best fit in this case. Though the ROC curve is marginally better of the given boosted model (derived visually) and also through the AUC values on the above report, the forest model AUC values & ROC curves are very close, considering all the 3 factors together- Overall Accuracy, Accuracy of Creditworthy predictions derived through the confusion matrices and the ROC curve, I believe the Forest Model best fits this case.

To summarize the Forest Model

Overall accuracy – 82.00%

Creditworthy predicted accurately: 80.95%

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

2. How many individuals are creditworthy?

When we score the above model, 406 applicants out of the 500 are determined to be creditworthy

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.