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?
  We need to decide if we it is safe for the bank to give loans to the individuals based on their data. We have to classify them as Creditworthy or Non-Creditworthy.

- What data is needed to inform those decisions?
  We have an excel file for these customers with various fields. We have to figure out which fields to be used as predictor variable based on various factors. I have considered 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, and No-of-Credits-at-this-Bank as the fields for the predictor variables.

- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  We will use binary model to help make these decisions because our target variable is binary and can be either Creditworthy or Non-Creditworthy.

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".

	Duration-of-Credit-Month	Credit-Amount	Instalment-per-cent	Most-valuable-available-asset	Type-of-apartment	Age_years_ImputedValu
Duration-of-Credit-Month	1	0.57398	0.068106	0.299855	0.152516	-0.064
Credit-Amount	0.57398	1	-0.28885	0.325545	0.170071	0.0693
Instalment-per-cent	0.068106	-0.28885	1	0.081493	0.074533	0.0392
Most-valuable-available-asset	0.299855	0.325545	0.081493	1	0.373101	0.08623
Type-of-apartment	0.152516	0.170071	0.074533	0.373101	1	0.3293
Age_years_ImputedValue	-0.0642	0.069316	0.03927	0.086233	0.32935	

As per the Pearson correlation table, none of the correlation is greater than 0.7. Hence, none of the fields are correlated to one another.

- Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed) and

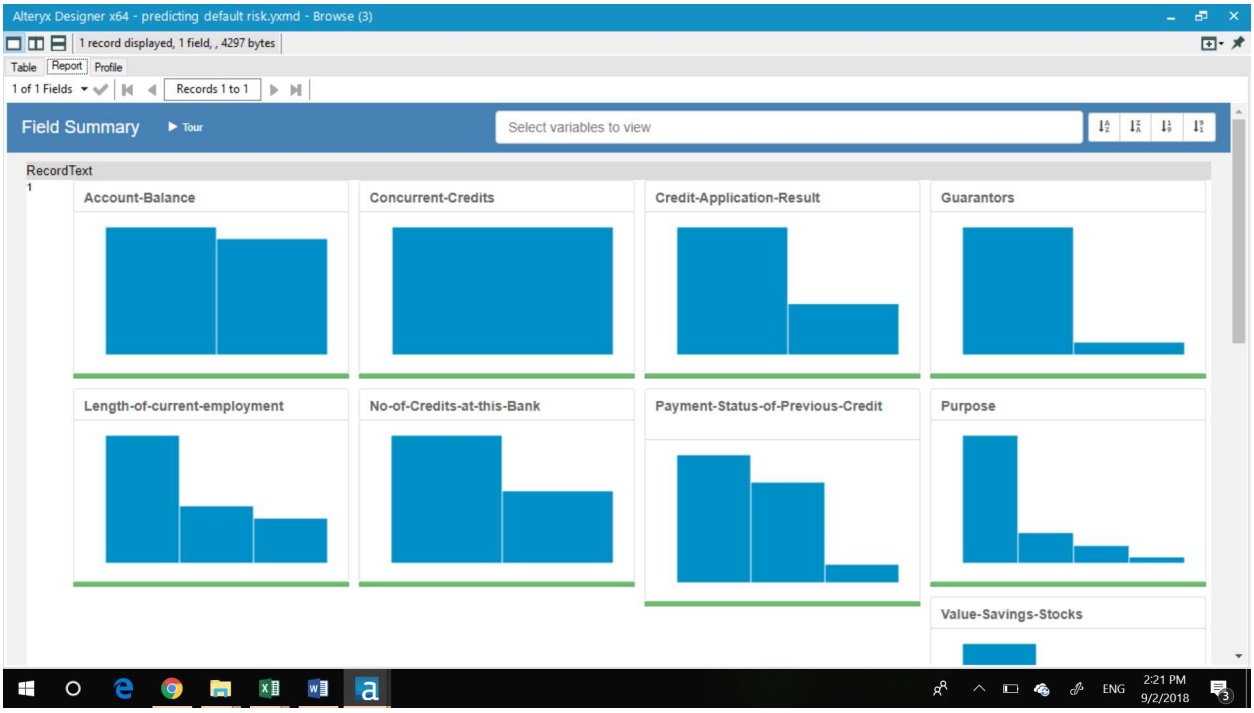
Yes, there was missing data in Age-years (2%) and Duration-in-Current-address (68.8%). Since the percentage of missing values is very high in Duration-in-Current-address, I have deleted this field. I have imputed the missing data in Age-years by the median value.

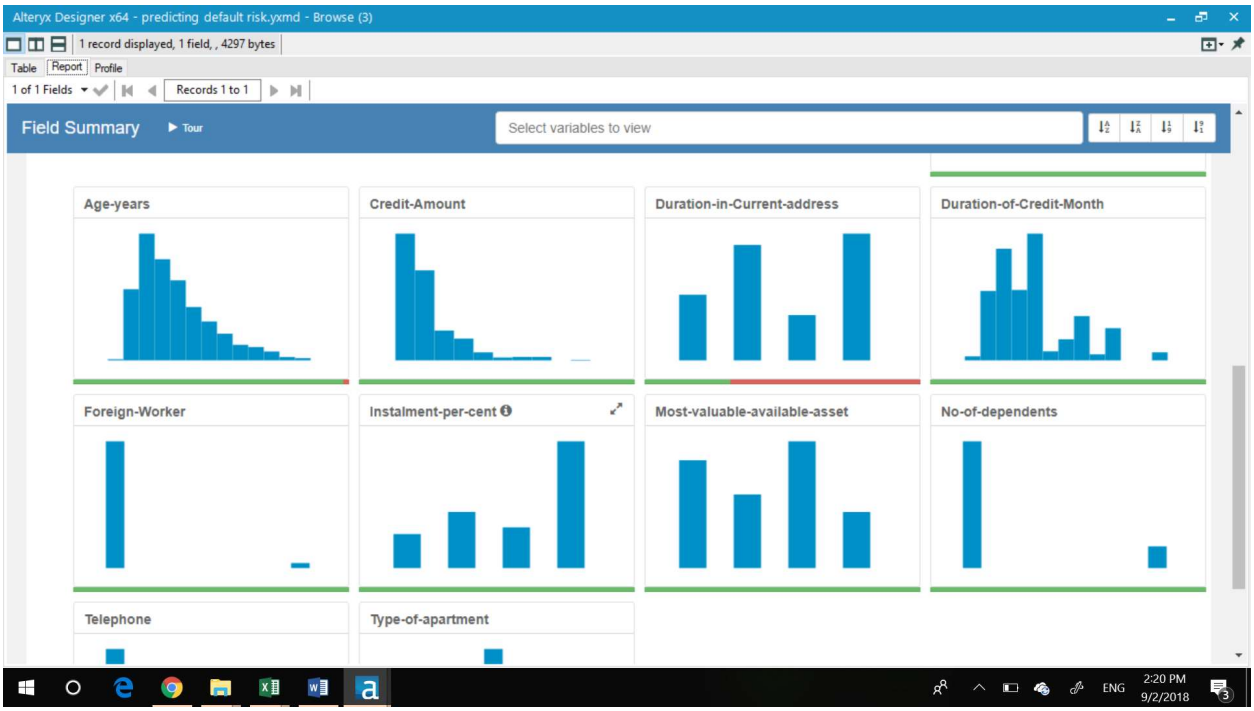
- Are there only a few values in a subset of your data field? Does the data field look very uniform (there is only one value for the entire field?). This is called “low variability” and you should remove fields that have low variability. Refer to the "Tips" section to find examples of data fields with low-variability.

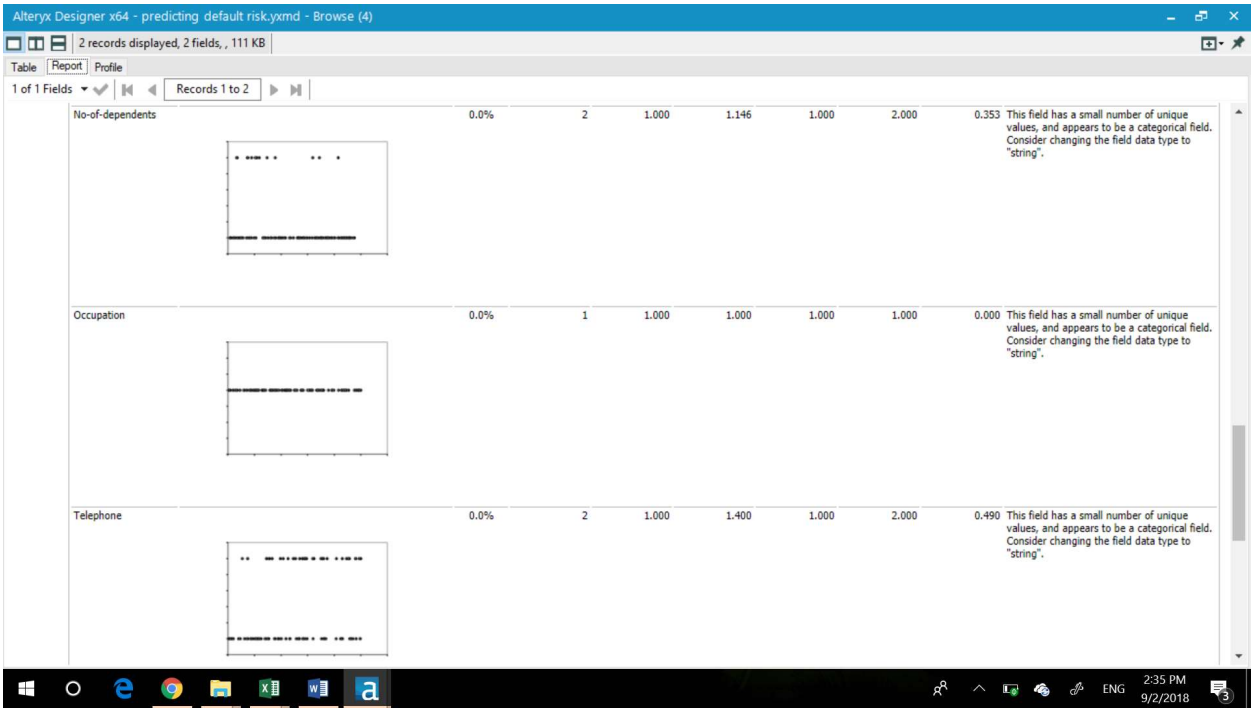
As per the field summary below, the Concurrent-Credits have low variability as it has jut one value. Data is entirely uniform and there is no other variations of the data. Similar is the case in Occupation.

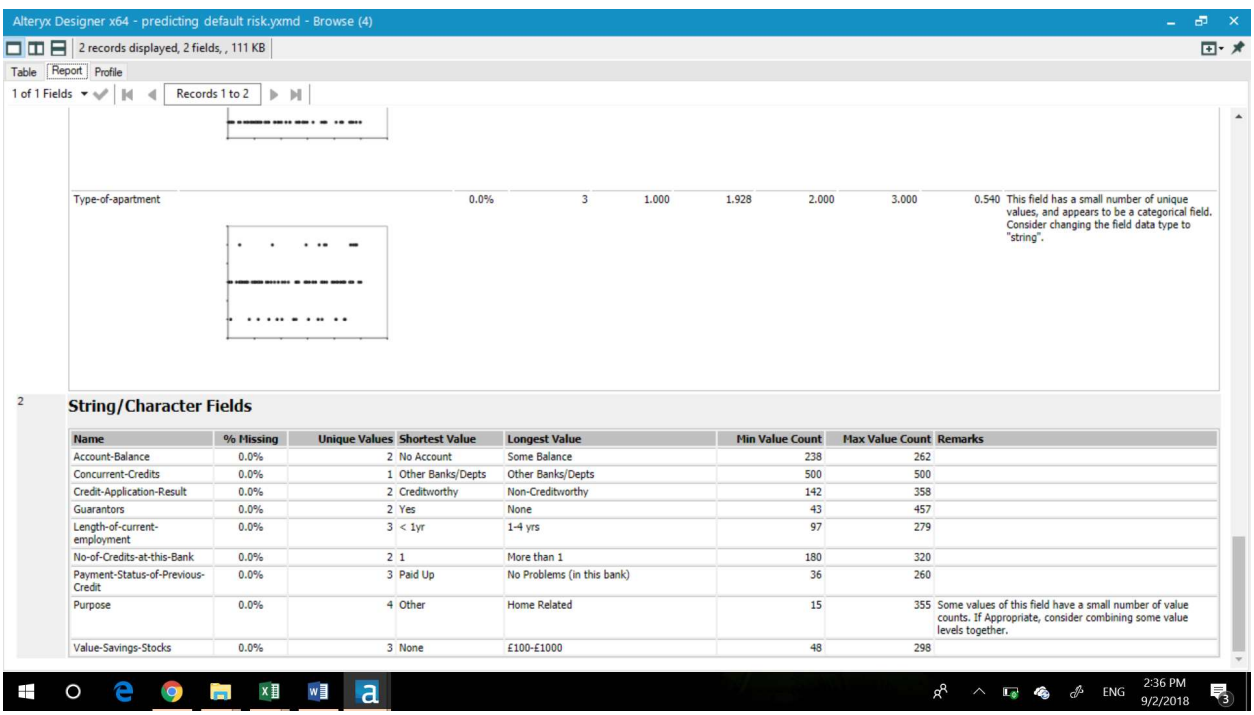
In Guarantors there are 457 instances of None and just 43 instances of yes. Here the field is heavily skewed to one type of data, and hence low variability. Similar is the case in Foreign worker and No. of dependents.

I will remove all these low variability fields. I will remove telephone field because it is not affecting creditworthiness of people.


Alteryx Designer x64 - predicting default risk.yxmd - Browse (3)
1 record displayed, 1 field, , 4297 bytes
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Field Summary
Tour
Select variables to view
RecordText
1
Account-Balance
Concurrent-Credits
Credit-Application-Result
Guarantors
Length-of-current-employment
No-of-Credits-at-this-Bank
Payment-Status-of-Previous-Credit
Purpose
Value-Savings-Stocks
ENG
2:21 PM
9/2/2018


Alteryx Designer x64 - predicting default risk.yxmd - Browse (3)
1 record displayed, 1 field, , 4297 bytes
Table
Report
Profile
1 of 1 Fields
Records 1 to 1
Field Summary
Tour
Select variables to view
Age-years
Credit-Amount
Duration-in-Current-address
Duration-of-Credit-Month
Foreign-Worker
Instalment-per-cent
Most-valuable-available-asset
No-of-dependents
Telephone
Type-of-apartment
ENG
2:20 PM
9/2/2018





- Your clean data set should have 13 columns where the Average of **Age Years** should be 36 (rounded up)

Average age is 35.57 years and clean data has 13columns

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.

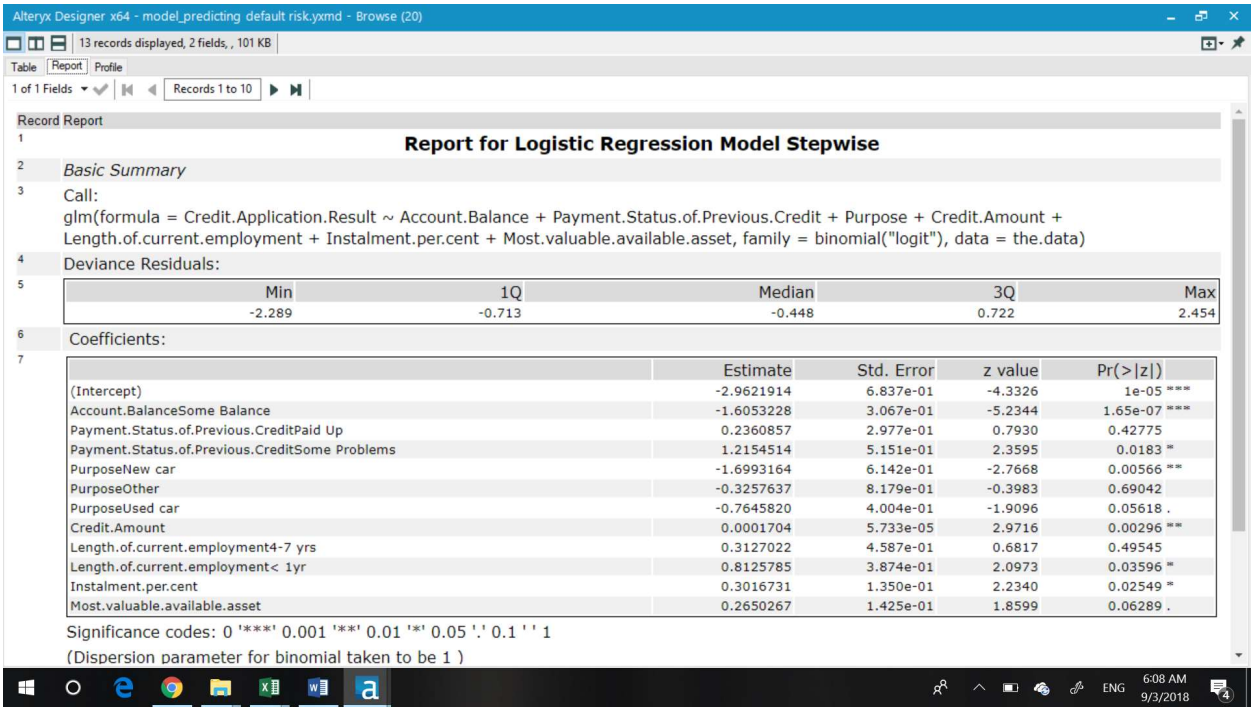
Step 3: Train your Classification Models

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

LOGISTIC REGRESSION

*Answer these questions for **each model** you created:*



Alteryx Designer x64 - model_predicting default risk.yxmd - Browse (20)

13 records displayed, 2 fields, , 101 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 10

Record Report

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)

ENG 6:08 AM 9/3/2018

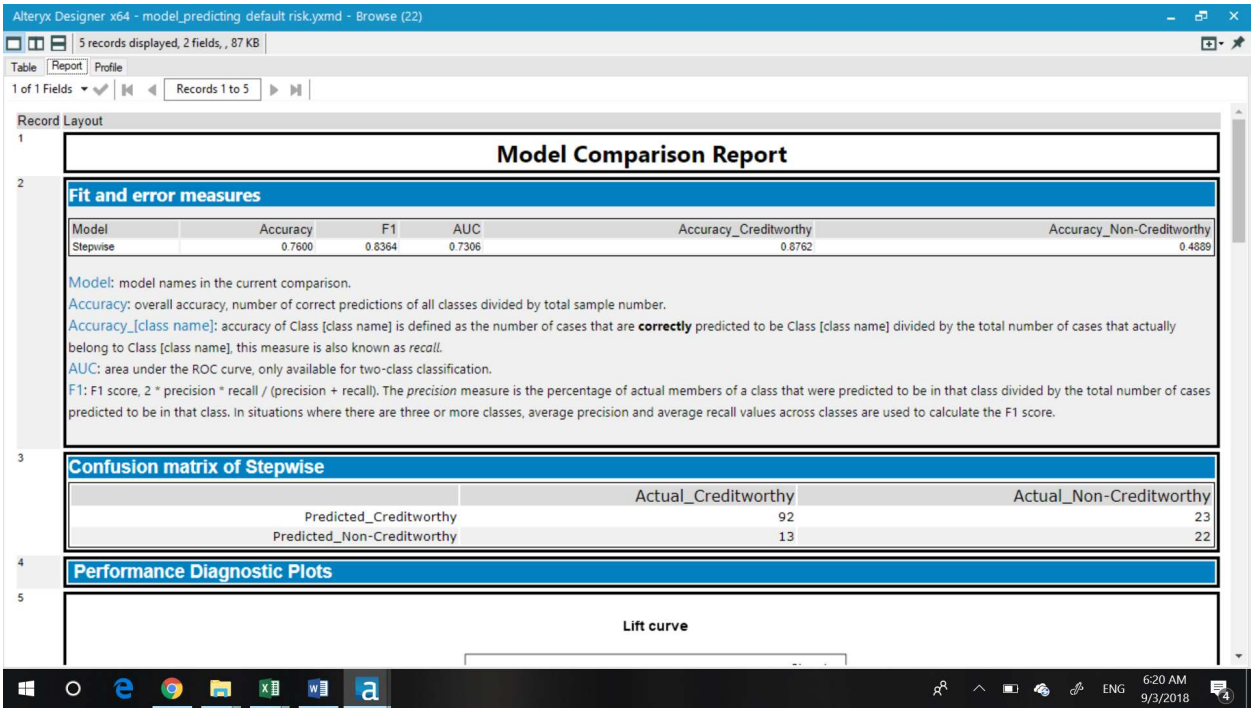
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

As per the charts shown above, the p-value of Account.BalanceSome Balance is the least and is the most significant with 3stars. The other significant variables in the order of significance are PurposeNewCar, Credit.Amount, Length.of.currentemployment<1yr, Installment.per.cent, Payment.status of Previous.CreditSomeProblems..

- 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?

After validating the model we can see the model comparison report. The overall accuracy is 76%. The accuracy_creditworthy is much higher 87.6%. However, the accuracy non-creditworthy is lower at 48.8%.

The accuracy on training set was 77.4%, so we did not see much reduction in accuracy on validation set(as it is 76%).



Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Stepwise	0.7600	0.8364	0.7306	0.8762	0.4889

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

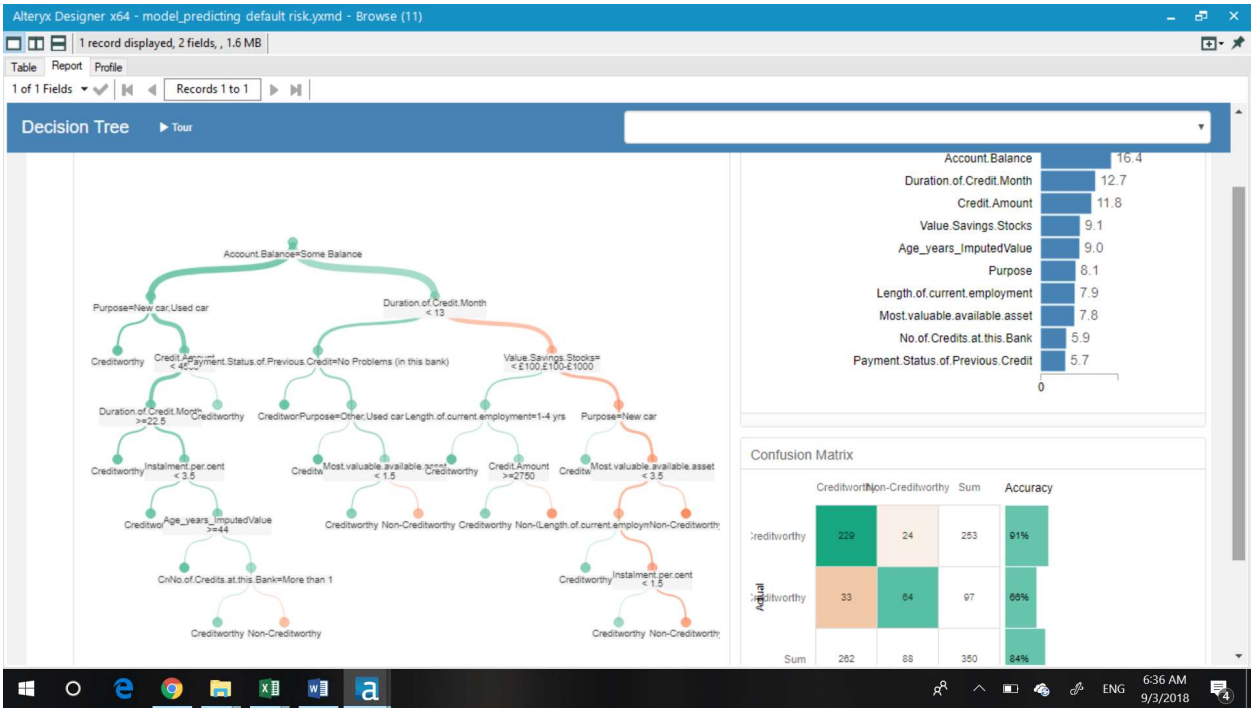
You should have four sets of questions answered. (500 word limit)

DECISION TREE 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.

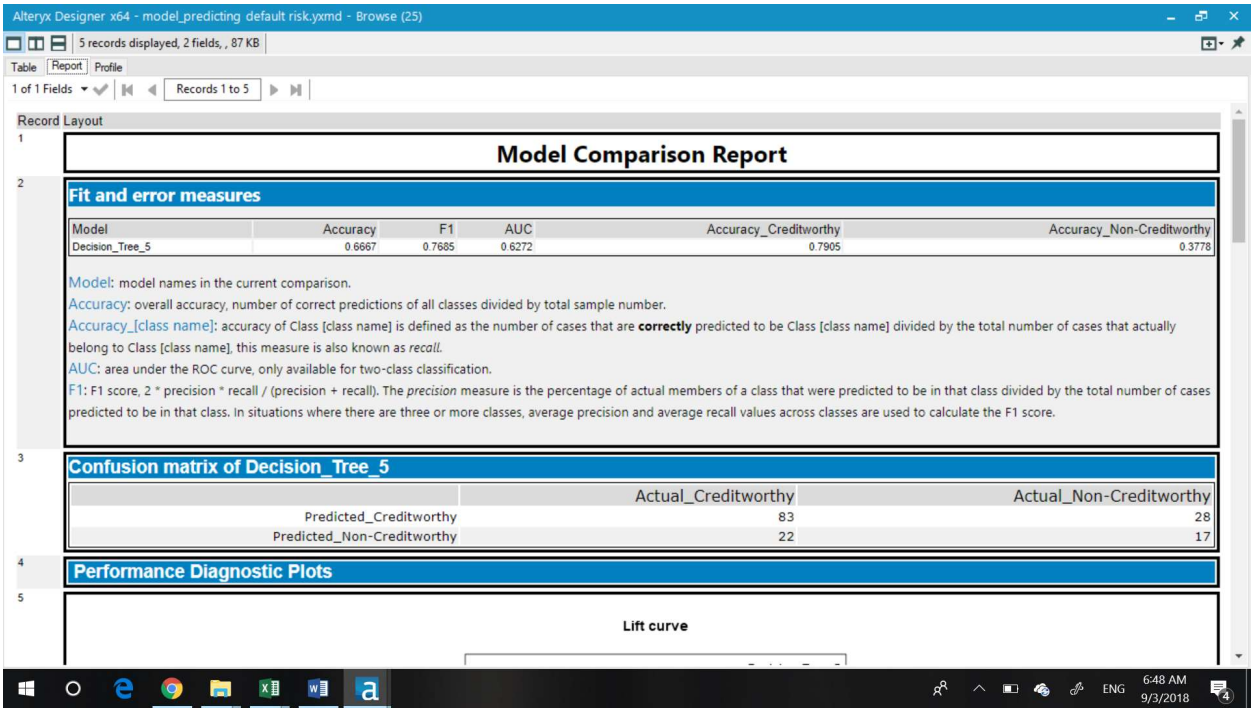
As per the decision tree output given below, the variables in the order of importance are Account Balance, Duration of Credit month, Credit Amount, Value Saving stocks, Age_years, Purpose, Length of current employment, Most valuable asset, No.of credits at this Bank, Payment status of previous card.



- 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?

After validating the model we can see that overall accuracy 66.7%. Accuracy creditworthy is 79%, and accuracy_non_creditworthy is 37.8%

The accuracy on training set was 84%, so we see much reduction in accuracy on validation set(as it is 66.7%). There are few misclassifications in the training set and more in the test set. So, there is high variance and we are overfitting the data.



Alteryx Designer x64 - model_predicting default risk.yxmd - Browse (25)

5 records displayed, 2 fields, , 87 KB

Table | Report | Profile

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
Decision_Tree_5	0.6667	0.7685	0.6272	0.7905	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.

3

Confusion matrix of Decision_Tree_5

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

4

Performance Diagnostic Plots

5

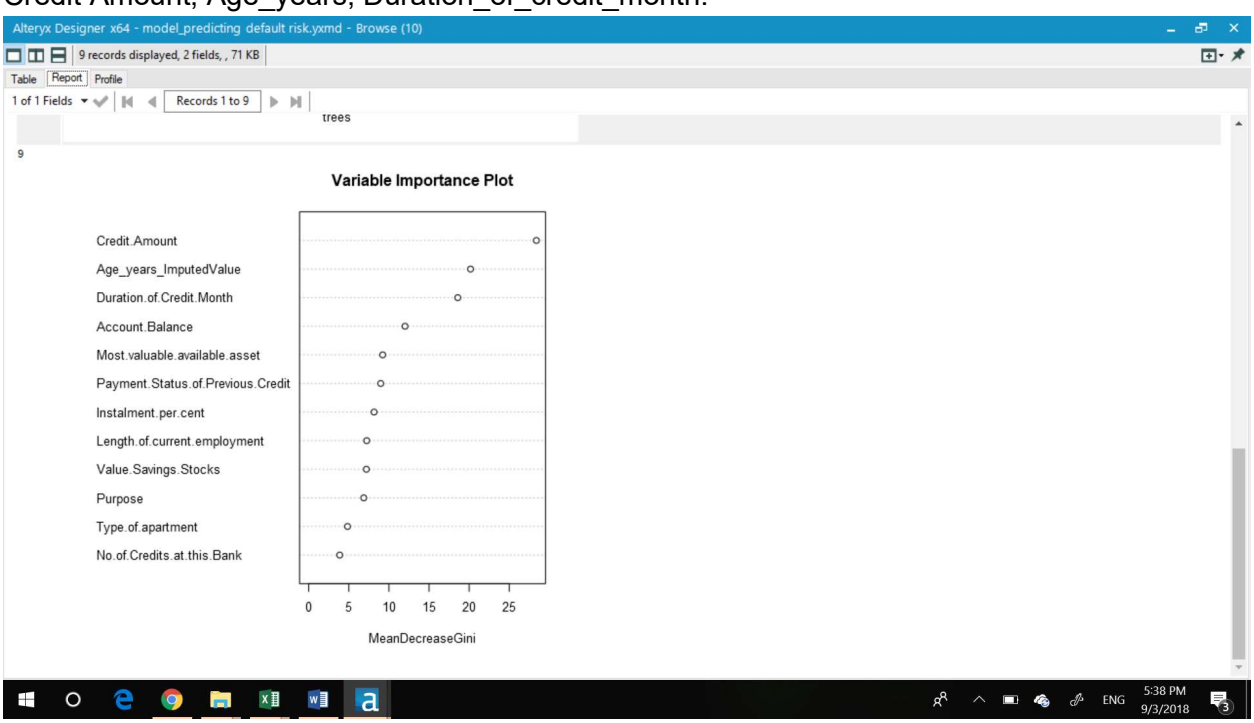
Lift curve

FOREST MODEL

*Answer these questions for **each model** you created:*

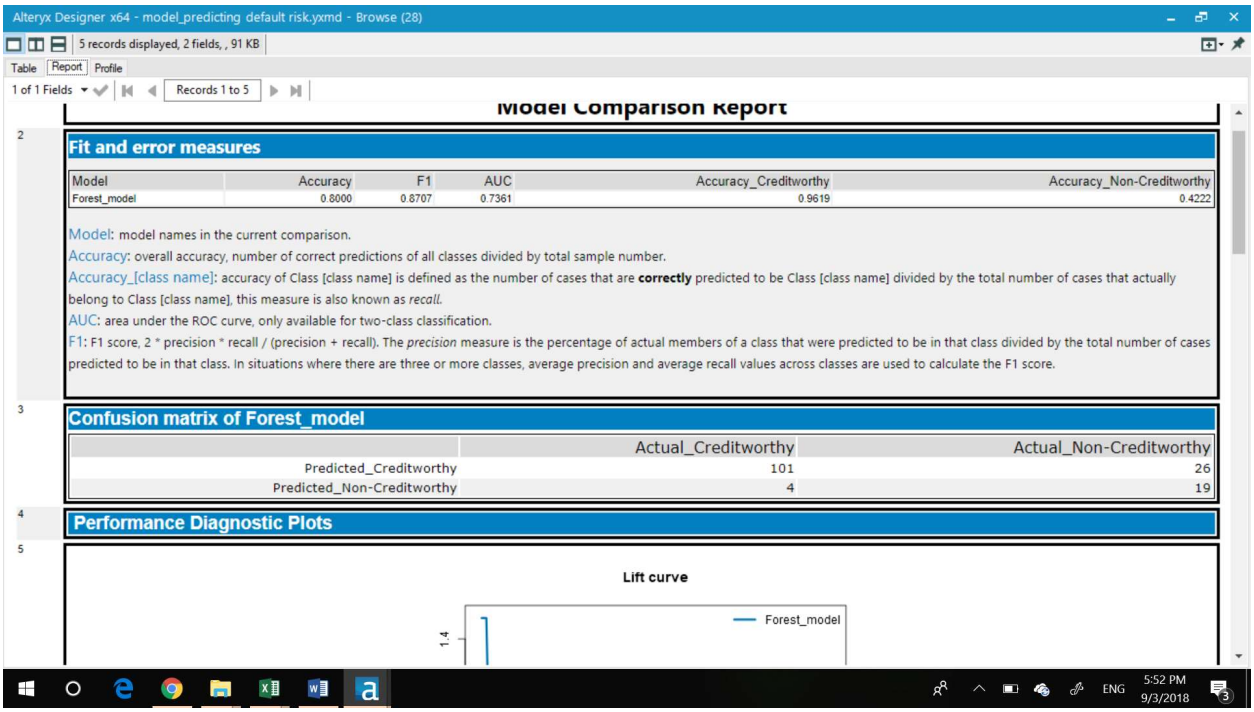
- Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

From the variable importance chart we can see that the variables in the order of importance are Credit Amount, Age_years, Duration_of_credit_month.



- 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 overall accuracy on validation set is 80%. Accuracy_Creditworthy is 96% and Accuracy non-creditworthy is 42%. The overall accuracy on estimation data was approximately 77%. So, the model has worked pretty well. The model is not overfitting at all. It is a case of low variance.

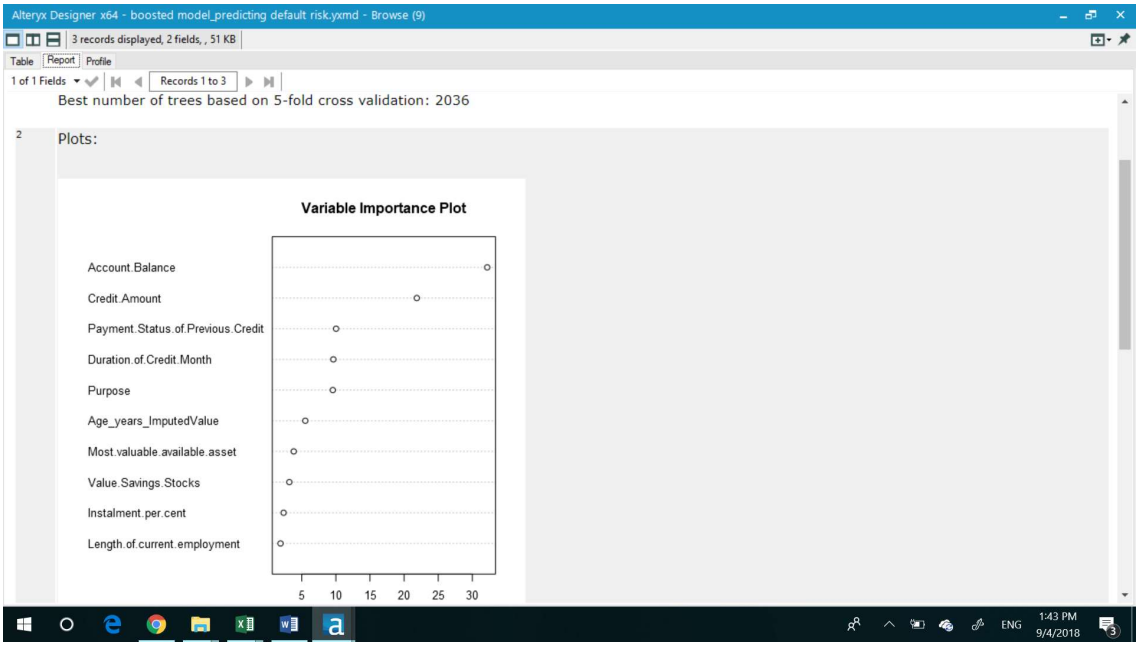


BOOSTED MODEL

*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.

The most important predictor variable in order of their importance 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?

The overall accuracy on validation set is 78.67%. The accuracy_creditworthy is 96.19%, and accuracy_non creditworthy is 37.8% . From the assessment plot on the estimation data we can see that at 2000 iteration loss is 0.4. Validation loss is almost 0.5 at 2000 iteration. It was decreasing till 2000itearation and after that it started to increase.

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:

Alteryx Designer x64 - model_predicting default risk.yxmd - Browse (15)

8 records displayed, 2 fields, , 168 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 8

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Boosted_Model	0.7867	0.8632	0.7524	0.9619	0.3778
Forest_model	0.8000	0.8707	0.7361	0.9619	0.4222
Decision_Tree_5	0.6667	0.7685	0.6272	0.7905	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.

Confusion matrix of Boosted_Model

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

Confusion matrix of Decision_Tree_5

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

ENG 2:14 PM 9/4/2018

Alteryx Designer x64 - model_predicting default risk.yxmd - Browse (15)

8 records displayed, 2 fields, , 168 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 8

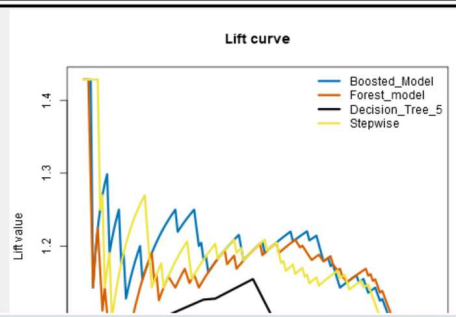
Confusion matrix of Forest_model

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

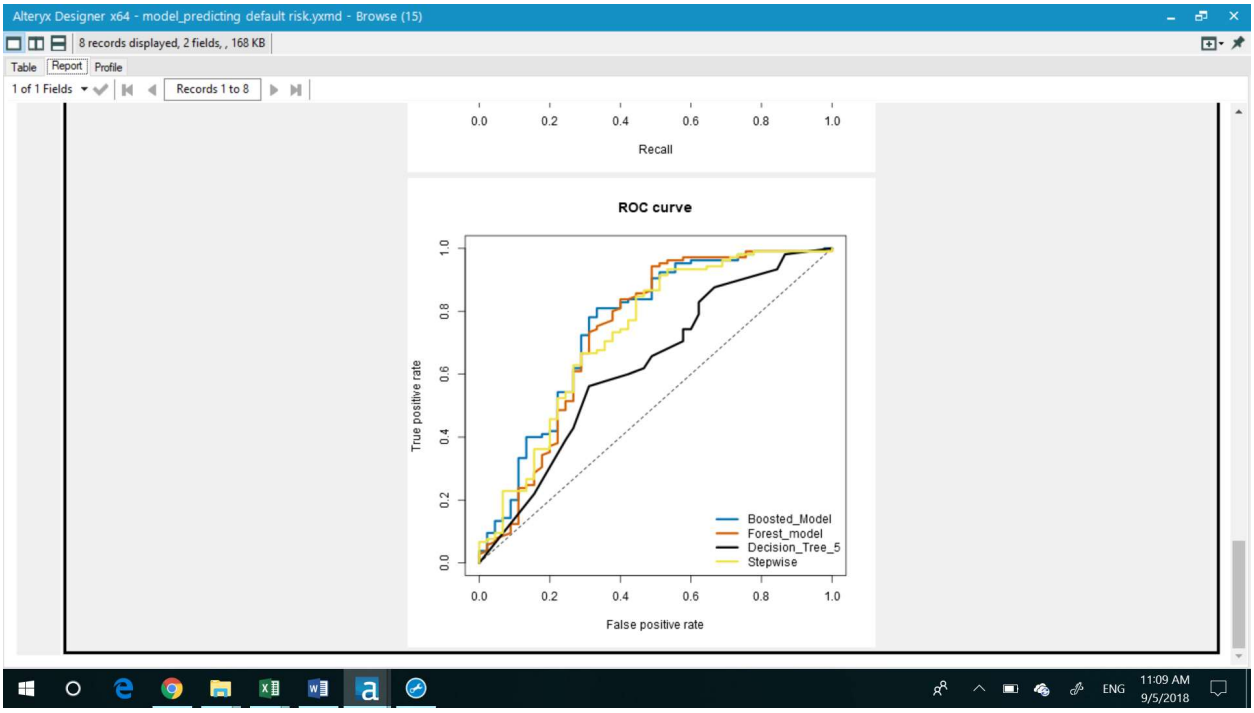
Confusion matrix of Stepwise

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

Performance Diagnostic Plots



ENG 2:14 PM 9/4/2018



- 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:

From the above Model Comparison Report I would select Forest Model based on the following:

- **Overall Accuracy against your Validation set**

The overall accuracy of Forest Model is 80% on validation set and it is higher than all other models

- Accuracies within “Creditworthy” and “Non-Creditworthy” segments

For Forest Model Accuracy for Creditworthy is highest at 96% and the accuracy of non-creditworthy is 42.2%. It is low, but towards the higher side as compared to all other models.

- ROC graph

From the ROC graph we can see that Forest Model reaches the highest level and reached there the fastest . Here Boosted Model seems to do pretty good too. However, it does not reach the highest level as compared to all others.

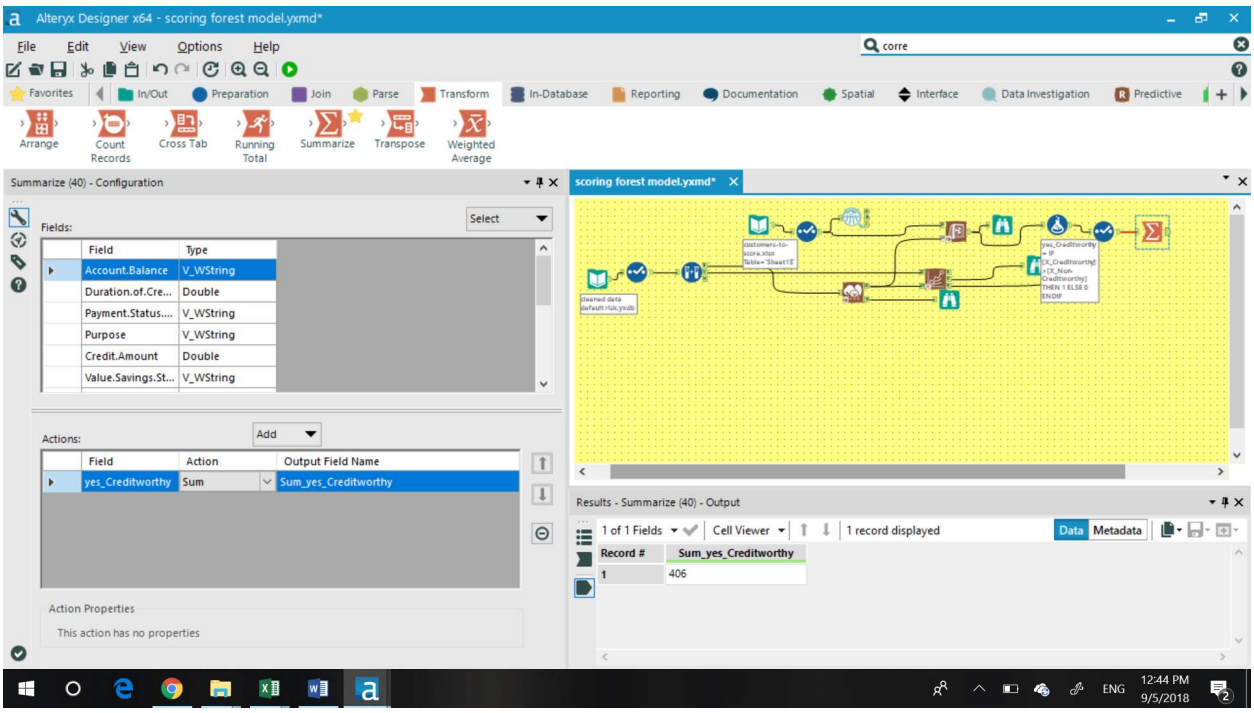
- Bias in the Confusion Matrices

Overall, we have few misclassification in the estimation set and few in the validation set. So, the bias and variance are low.

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

- How many individuals are creditworthy?

As per the Alteryx flow below, when we score the data on Forest Model we get 406 customers which are 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.