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?

A list of predictor variables that will help us obtain a result on whether a customer is creditworthy and therefore, we can provide them with a loan. Not all variables are vital, and we will need to clean and trim down the data as well as to make sure that we don't skew the data. From the available data, the Credit-application-result will be our target variable and we will use specific classification models and compare them to acquire the best model classification for our data problem. Further, we will use this predicted data and score it with the data of new customers that have applied for loan and examine their creditworthiness based on our previous information we have from our own customers. The decision that needs to be made is if a loan should be provided to each of the 500 customers or not.

- What data is needed to inform those decisions?

We choose Credit-application-result as our target variable and a few predictor variables that are statistically significant or have a correlation of 0.7 and below to make sure our model is not bias. The Data on our previous customers and the results on whether the data they have provided did they receive a creditworthy loan issued or not.

We use this predictive model in the new data which contains the same information but on a new set of customers willing to apply for loan and our motive is to decide whether we sanction the loan to the new customers or not.

Data on customers income, purpose of the loan, age, length of current employment, installment per cent and payment status of previous credit.

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

We need to use a Binary model to help make these decisions as the target variable is a binary variable – Creditworthy or Non-Creditworthy. Our main aim is to find in a data that contains information on 500 customers of whether they are creditworthy or not.

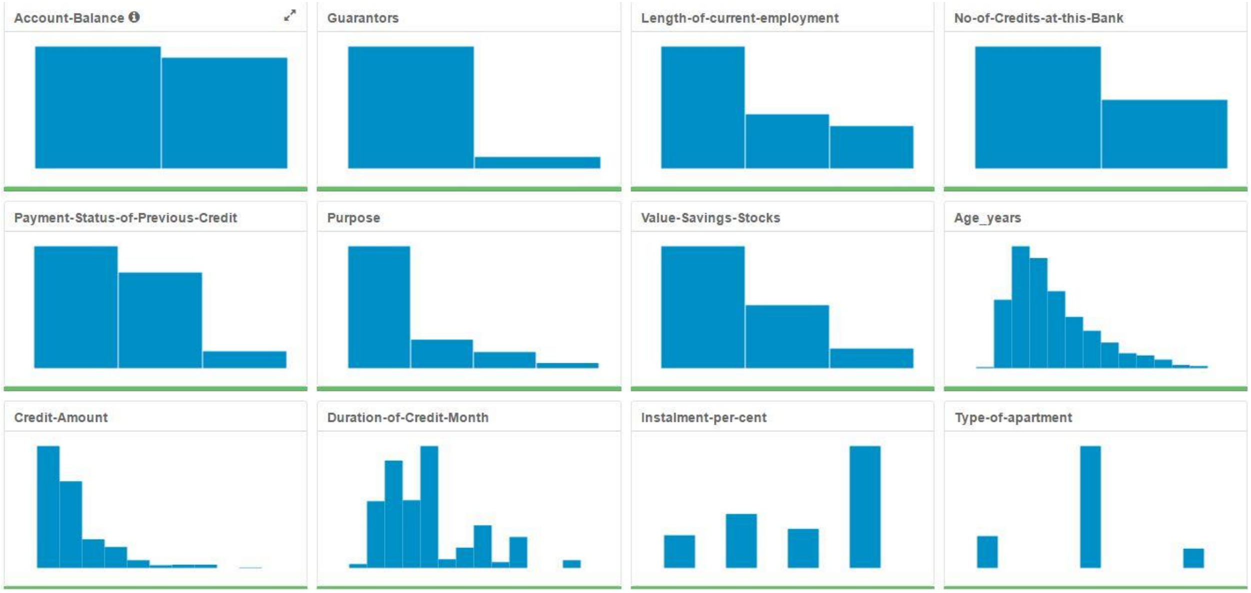
Step 2: Building the Training Set

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.

Variable **Age-years** had a few missing data in its column which we have imputed with median for that column. It is skewing the data more towards the left side (around age 25-35) as most of our data revolves around this age group.
Fields that had low variability were also removed such as
occupation - the more specific reason to remove Occupation and Concurrent credits is that both variables have just 1 value. And we would not learn anything from just 1 value so that is why we should remove them
Telephone – There isn't any significant meaning by imputing the number of phones a customer has. that the variable has no bearing on whether a customer is creditworthy or not.
No of dependents, **foreign worker** and **concurrent credits.**
Duration in current address also has a lot of missing values and not considering this as a predictor variable does not have any major impact on our model.
Most valuable available asset – again has very low variability, it had data that would be more distinguishable, then it would make sense to include this variable in the data set.

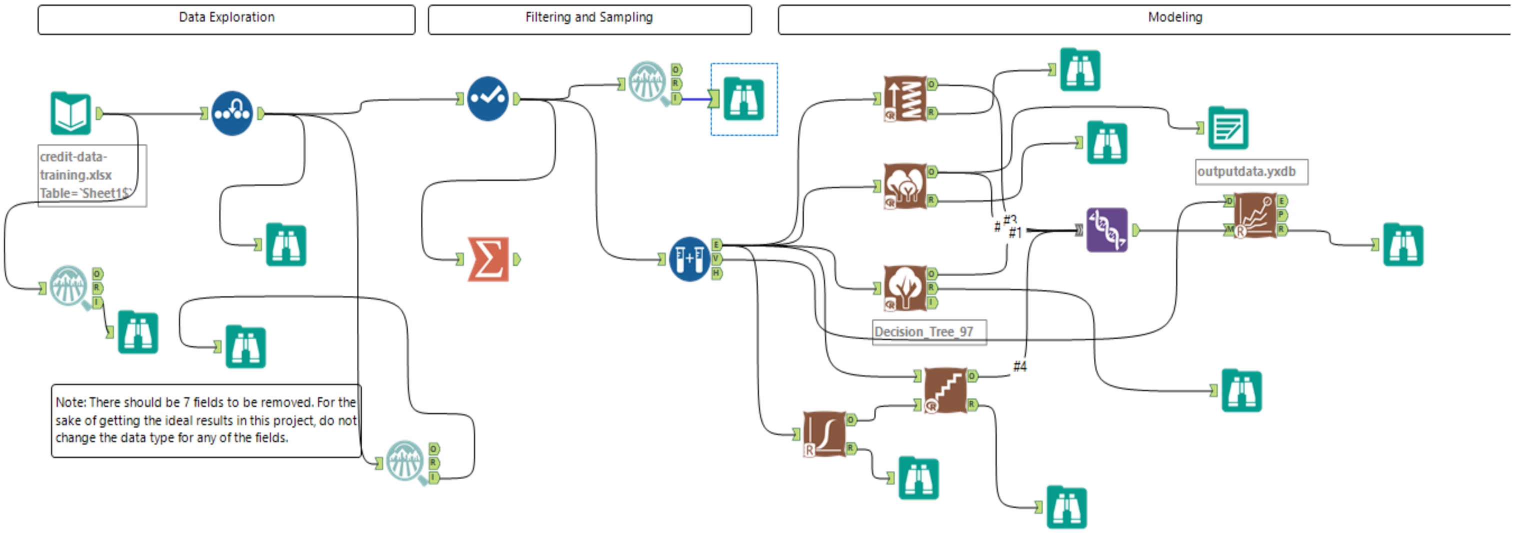


Step 3: Train your Classification Models

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.

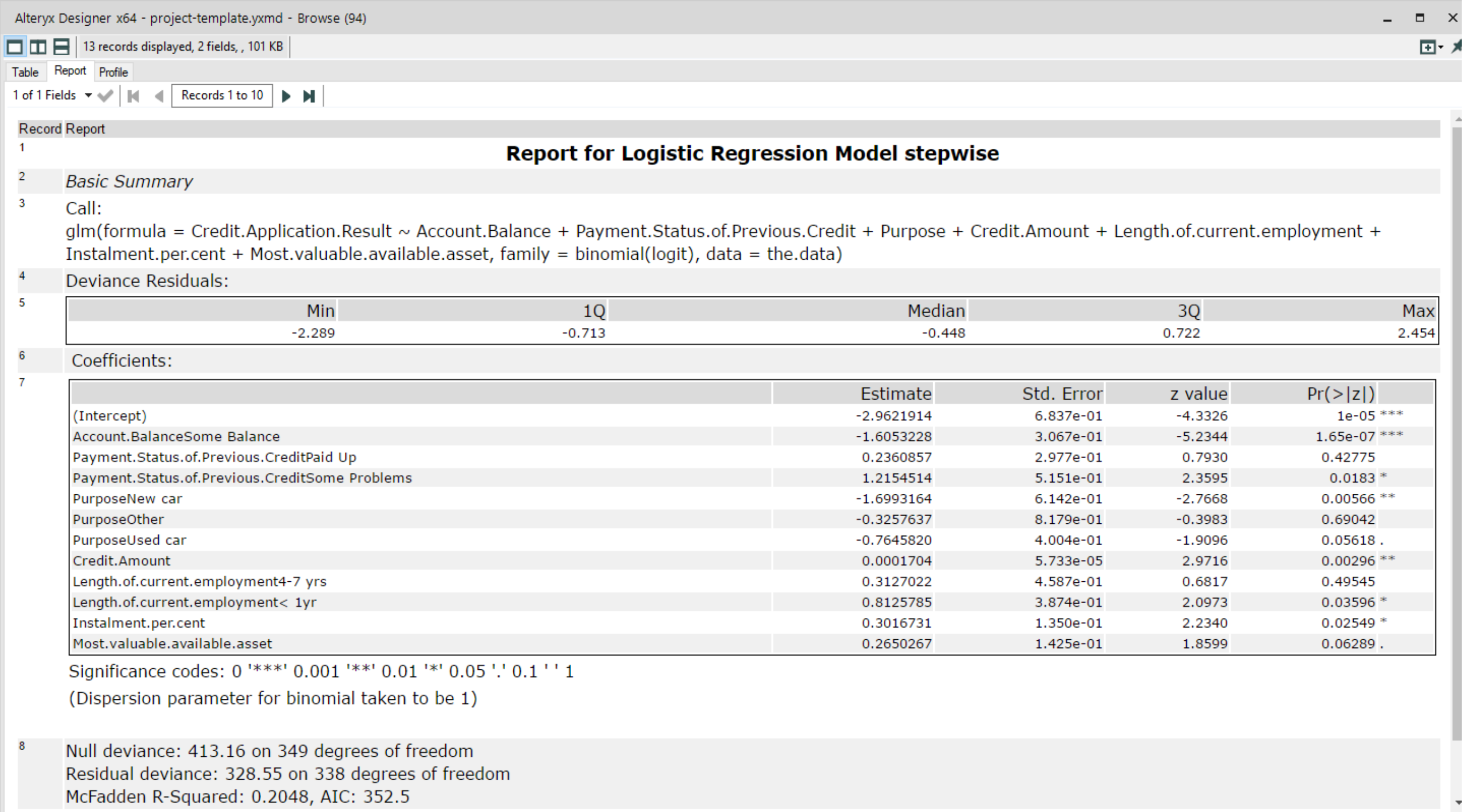
Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model

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



Note: There should be 7 fields to be removed. For the sake of getting the ideal results in this project, do not change the data type for any of the fields.

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


Alteryx Designer x64 - project-template.yxmd - Browse (94)
13 records displayed, 2 fields, , 101 KB
Table
Report
Profile
1 of 1 Fields
Records 1 to 10
Record Report

1 **Report for Logistic Regression Model stepwise**

2 *Basic Summary*

3 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)

4 Deviance Residuals:

5

Min	1Q	Median	3Q	Max
-2.289	-0.713	-0.448	0.722	2.454

6 Coefficients:

7

	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	.

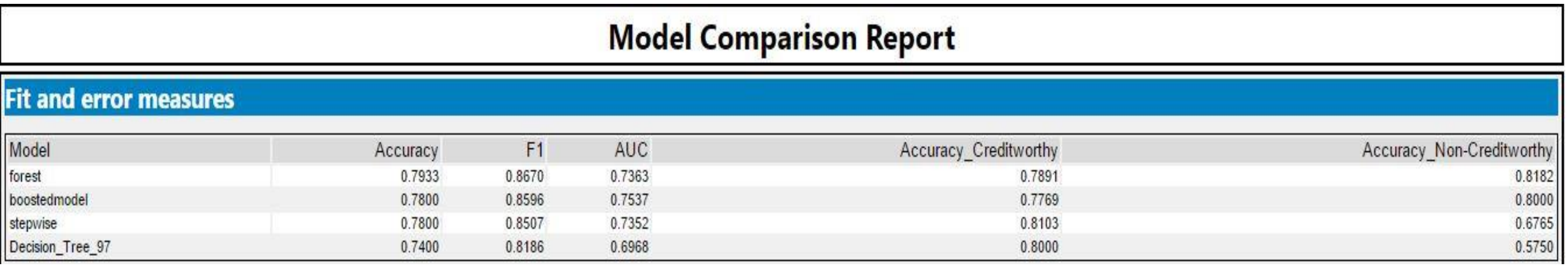
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial taken to be 1)

8 Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, AIC: 352.5

What we see above is the stepwise logistic regression model.
The model has by itself removed the unimportant predictor variables that were included in the estimation sample tool.
We can clearly say that the most important predictor variables based on p values are –
Account Balance - somebalance
Payment status of previous credit – some problems
Purpose- new car and used car
Credit amount
Installment percent
Most valuable available asset
Length of current employment <1year

- 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?



Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
forest	0.7933	0.8670	0.7363	0.7891	0.8182
boostedmodel	0.7800	0.8596	0.7537	0.7769	0.8000
stepwise	0.7800	0.8507	0.7352	0.8103	0.6765
Decision_Tree_97	0.7400	0.8186	0.6968	0.8000	0.5750

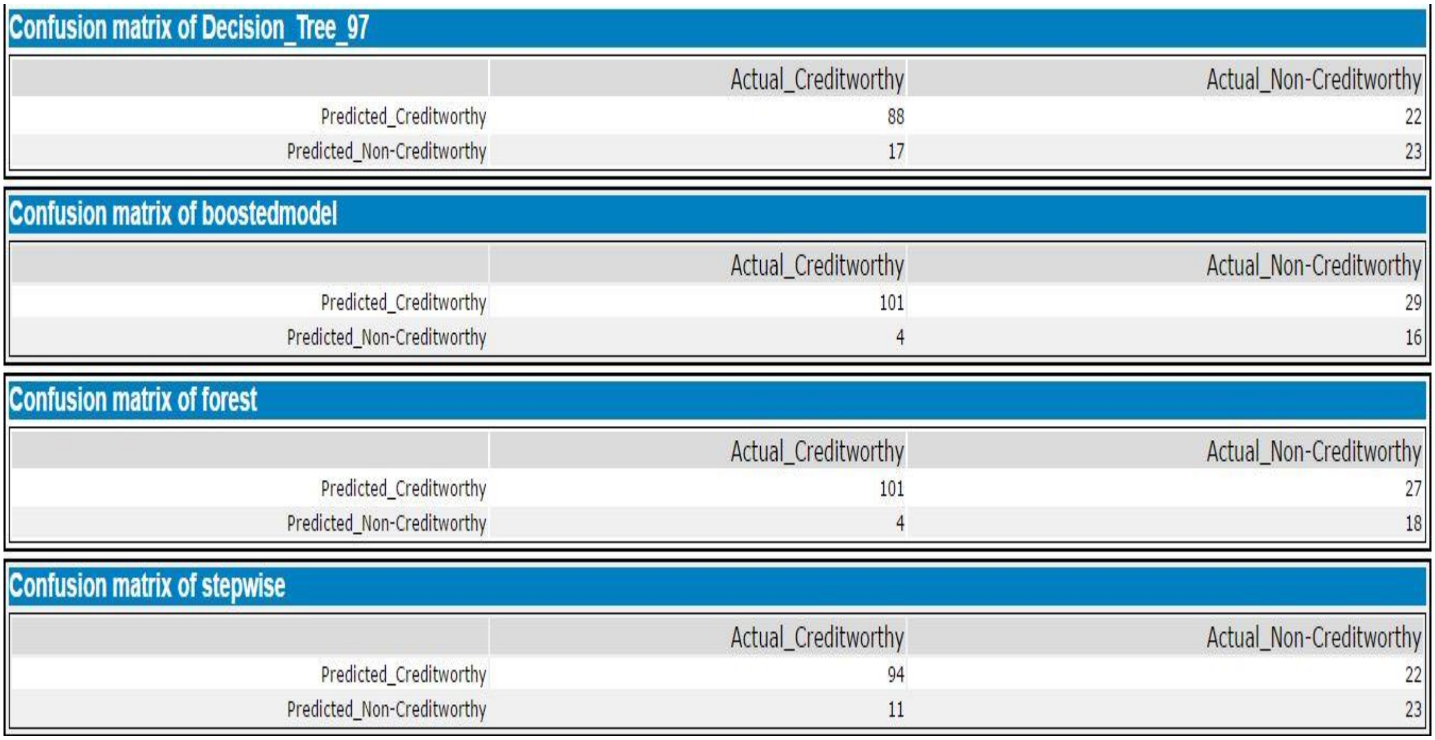
We have our four classification models – the boosted, the forest, the decision tree and the logistic regression models.

On using the validation data set along with our model comparison tool, it is easy to conclude that the Forest model has better accuracy when compared to the rest.

The decision tree has the lowest accuracy standing at only 74%.

Upon closely inspecting the confusion matrix we can see that the Accuracy for predicting the True creditworthy is done at almost 79% accuracy for all the models.

The only difference comes in predicting the true and actual Non-creditworthy for a customer in the validation dataset for the Decision tree and logistic regression models.



Confusion matrix of Decision_Tree_97

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	88	22
Predicted_Non-Creditworthy	17	23

Confusion matrix of boostedmodel

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	29
Predicted_Non-Creditworthy	4	16

Confusion matrix of forest

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

Confusion matrix of stepwise

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	94	22
Predicted_Non-Creditworthy	11	23

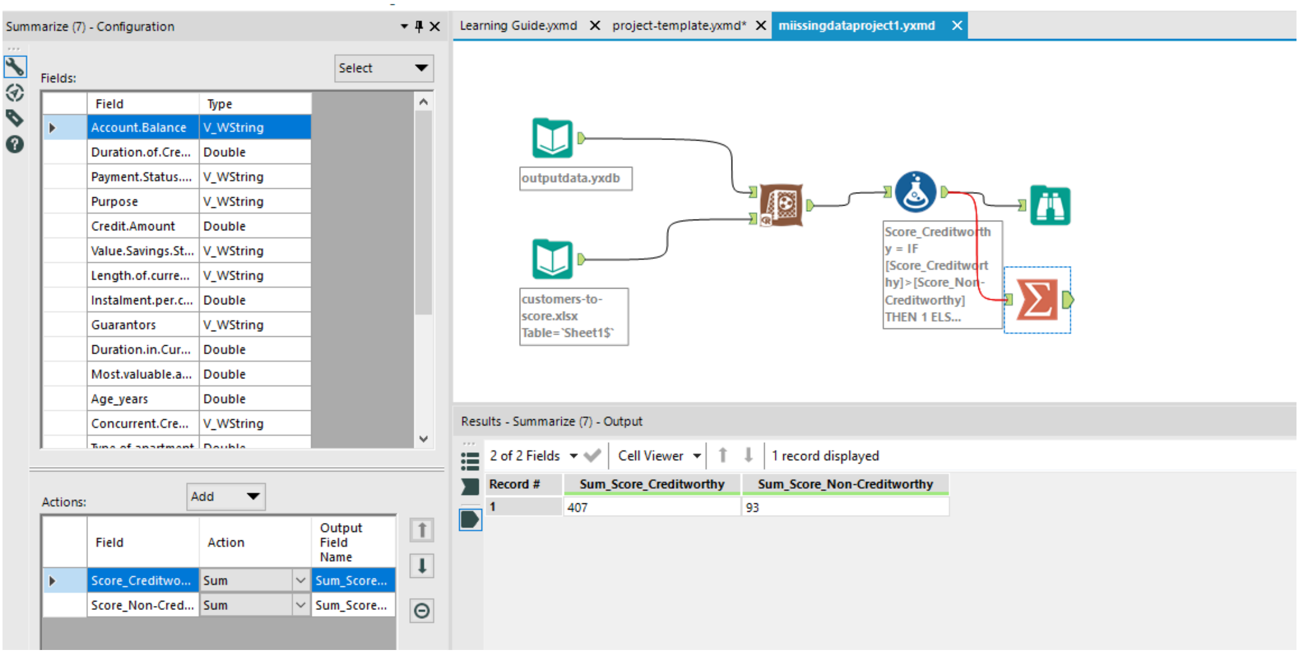
- Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model's predictions?

If the model predicts one of the classes with much higher accuracy than the other that means that the model is biased towards that class.
Let's start with **the decision tree model**. While this model is good at predicting the creditworthy with about 80% accuracy, it is still bad when it comes to predicting the non-creditworthy customers as its accuracy is 57%. Similarly, for the **model stepwise logistic regression**, the model accurately predicts the creditworthy with 81% and accuracy for the non-creditworthy is just about 68%. It might still be better than the decision tree model, but the false positives are still high. We will be accepting loan applications from customers who are non-creditworthy.
In other words, we would deny a loan to many creditworthy individuals as it classifies many creditworthy applicants as non-creditworthy. Then we might say that the above 2 models are biased towards predicting individuals who are creditworthy, as it does not predict individuals who are not creditworthy nearly at the same level as those who are. We have two such models.

The other two models **– the boosted and the forest models** have almost no bias as the accuracies for predicting the creditworthy and non-creditworthy is almost the same.

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"



Summarize (7) - Configuration

Fields: Select

	Field	Type
▶	Account.Balance	V_WString
	Duration.of.Cre...	Double
	Payment.Status....	V_WString
	Purpose	V_WString
	Credit.Amount	Double
	Value.Savings.St...	V_WString
	Length.of.curre...	V_WString
	Instalment.per.c...	Double
	Guarantors	V_WString
	Duration.in.Cur...	Double
	Most.valuable.a...	Double
	Age_years	Double
	Concurrent.Cre...	V_WString
	Type of apartment	Double

Actions: Add

	Field	Action	Output Field Name
▶	Score_Creditwo...	Sum	Sum_Score...
	Score_Non-Cred...	Sum	Sum_Score...

Learning Guide.yxmd
project-template.yxmd*
miissingdataproject1.yxmd
outputdata.yxdb
customers-to-score.xlsx Table=`Sheet1$`
Score_Creditworthy = IF [Score_Creditworthy]>[Score_Non-Creditworthy] THEN 1 ELS...

Results - Summarize (7) - Output

2 of 2 Fields | Cell Viewer | 1 record displayed

Record #	Sum_Score_Creditworthy	Sum_Score_Non-Creditworthy
1	407	93

Answer these questions:

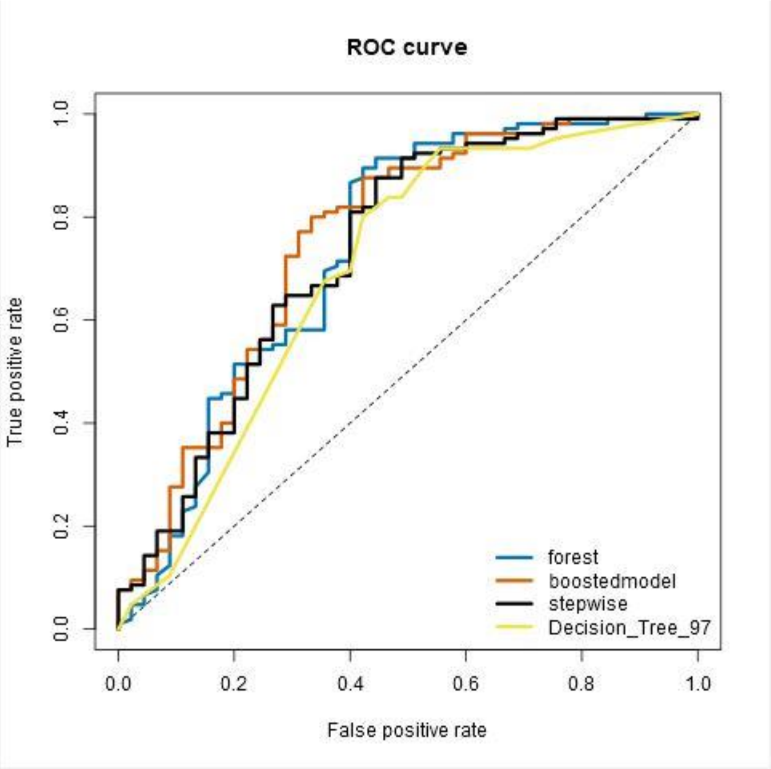
- Which model did you choose to use? Please justify your decision using **all** of the following techniques. Please only use these techniques to justify your decision:
    - ○ Overall Accuracy against your Validation set

The overall accuracy for the Forest model is the highest with 78% accuracy against the validation dataset and therefore we will use this model output to score to predict the credit worthiness for the rest of the new customers.

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

    For our chosen forest model, we have the accuracies for the creditworthy and non-creditworthy with 79% and 81% accuracy which is very preferable for our model prediction.

- ROC graph



we can see that the ROC graph for the Forest model is the highest line along the graph for most of the chart, and it rises the fastest of all models – meaning that we are getting a higher rate of true positive rates vs. false positives. We want a high rate of true positive vs. true negative rates because we do not want to extend loans to people who are not creditworthy. The ideal ROC curve hugs the top left corner, indicating a high true positive rate and a low false positive rate.

- Bias in the Confusion Matrices
  The selected model have one of the least differences between the accuracies of the two segments Creditworthy" and "Non-Creditworthy" with 79% accuracy in predicting the true creditworthiness and 81% in predicting the true non-creditworthy customers.

- How many individuals are creditworthy?

In the End we find that 407 customers are creditworthy and the rest 93 are non-creditworthy.