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?
  Determine which customers are creditworthy out of 500 new loan applications.

- What data is needed to inform those decisions?
  -Data on all past applications
  -List of the 500 new customer applications with information to determine creditworthiness such as account balance, age, purpose of credit, etc.

- 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. Normally for a binary model, we would need to compare the Logistic Regression and the Decision Tree models and select the more accurate model.

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)*

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.

I removed Guarantors because it has low variability; the data is skewed towards “None”.
I removed Duration-in-Current-address because it has a lot of missing data.
I removed Concurrent-Credits because it has low variability; the data is uniform.
I removed Occupation because it has low variability; the data is uniform.
I removed No-of-dependents because it has low variability; the data is skewed towards “1”.
I removed Telephone because it is not a relevant variable to predict creditworthiness.
I removed Foreign-Worker because it has low variability; the data is skewed towards “1”.
I imputed Age-years with the median because only 2.4% of the values were missing so there is no need to remove the entire field. Median is better to use when the distribution is not symmetrical, as is the case with our data and age in general which is skewed to the left (younger) side. Using the median age of 33 insures there are the same number of people who are older than the median age as there are younger than it.

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?

Logistic Regression

Account-Balance, Payment-Status-of-Previous-Credit, Purpose, Credit-Amount, Length-of-current-employment, Instalment-per-cent, and Most-valuable-available-asset are significant predictor variables; the Stepwise model confirms this.

	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	.

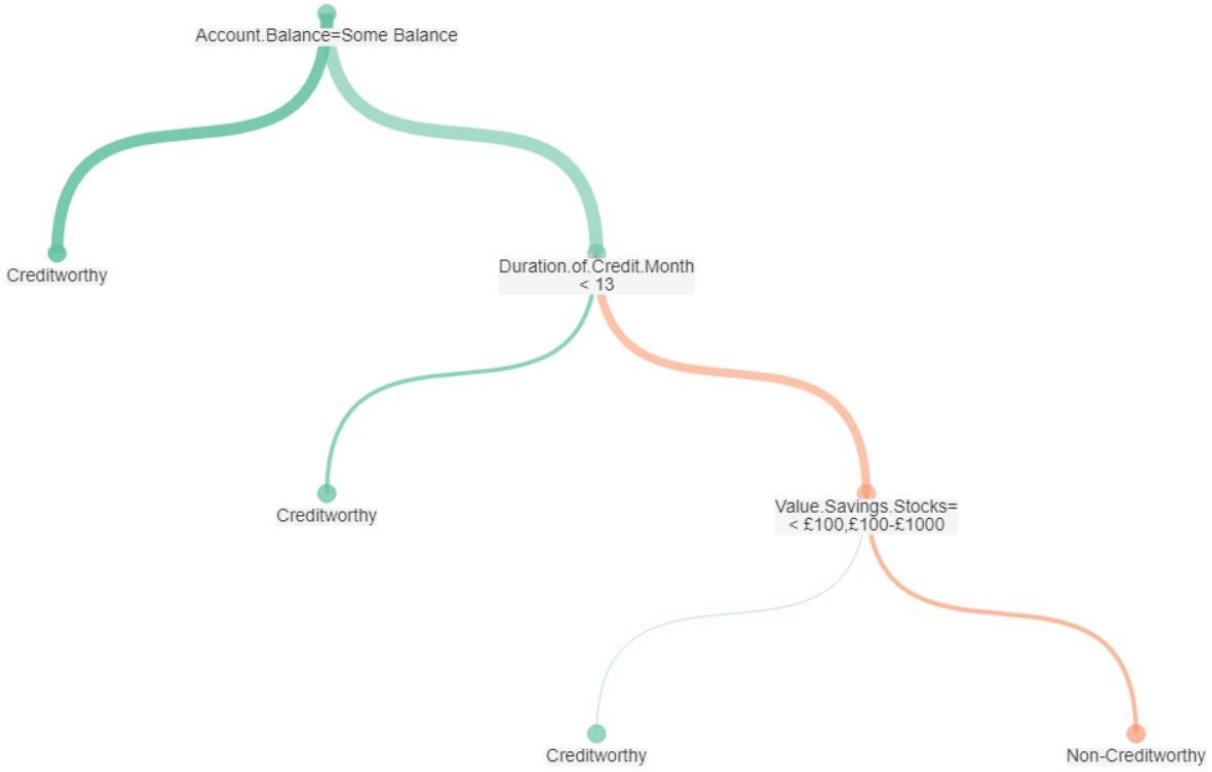
Confusion matrix of Logistic_Stepwise

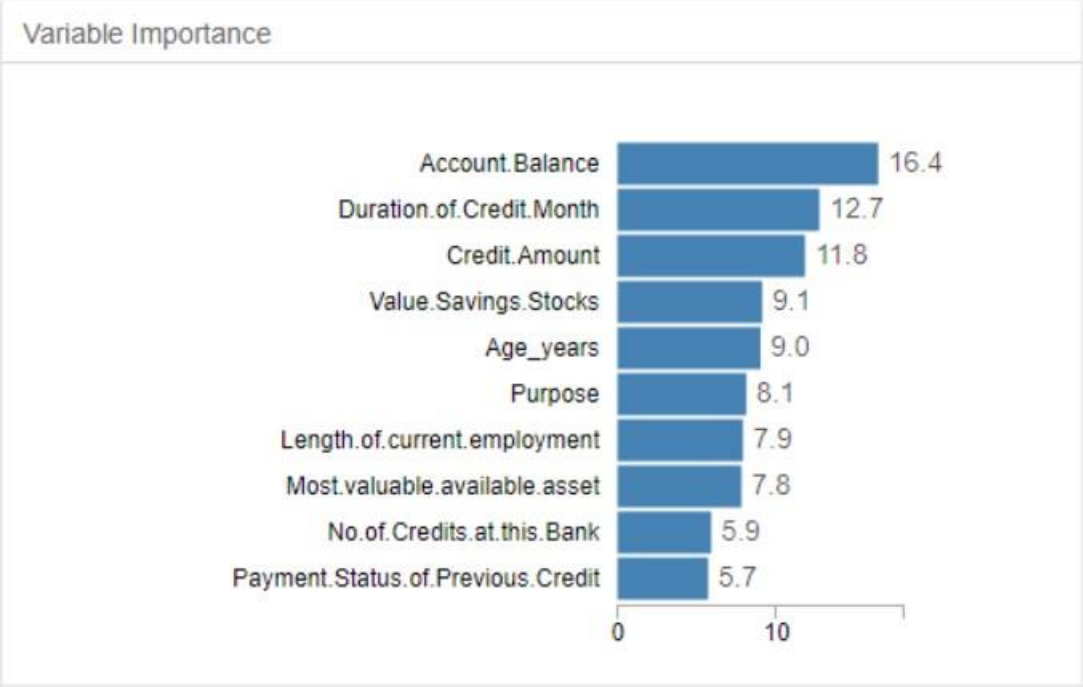
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	92	23
Predicted_Non-Creditworthy	13	22

The Logistic Regression model has an overall accuracy of 76%. It has an accuracy of 80% in predicting creditworthy individuals and an accuracy of 63% in predicting noncreditworthy individuals. The model is biased towards correctly predicting creditworthy individuals than noncreditworthy.

Decision Tree

The significant predictor variables are Account-Balance, Duration-of-Credit-Month, and Value-Savings-Stocks.





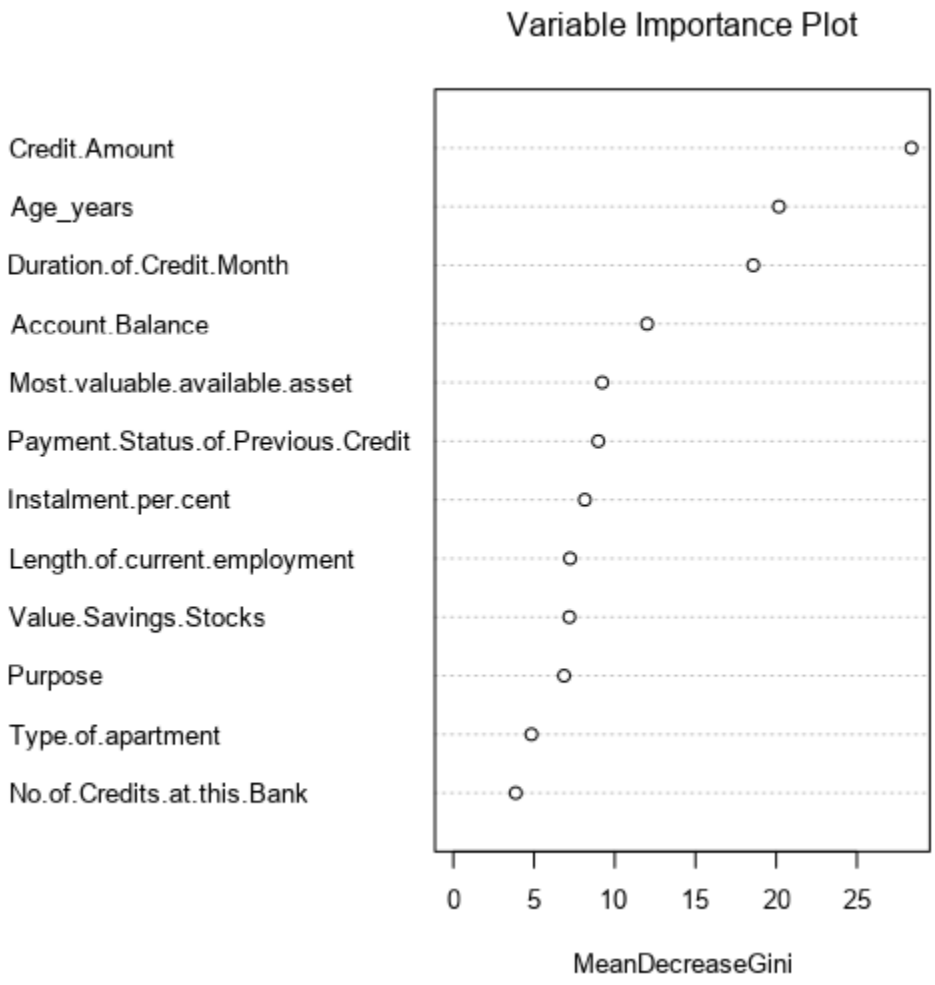
Confusion matrix of DT

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

The Decision Tree model has an overall accuracy of 75%. It has an accuracy of 79% in predicting creditworthy individuals and an accuracy of 60% in predicting noncreditworthy individuals. The model is biased towards correctly predicting creditworthy individuals than noncreditworthy.

Forest Model

The significant variables are Credit-Amount, Age_Years, and Duration-of-Credit-Month.



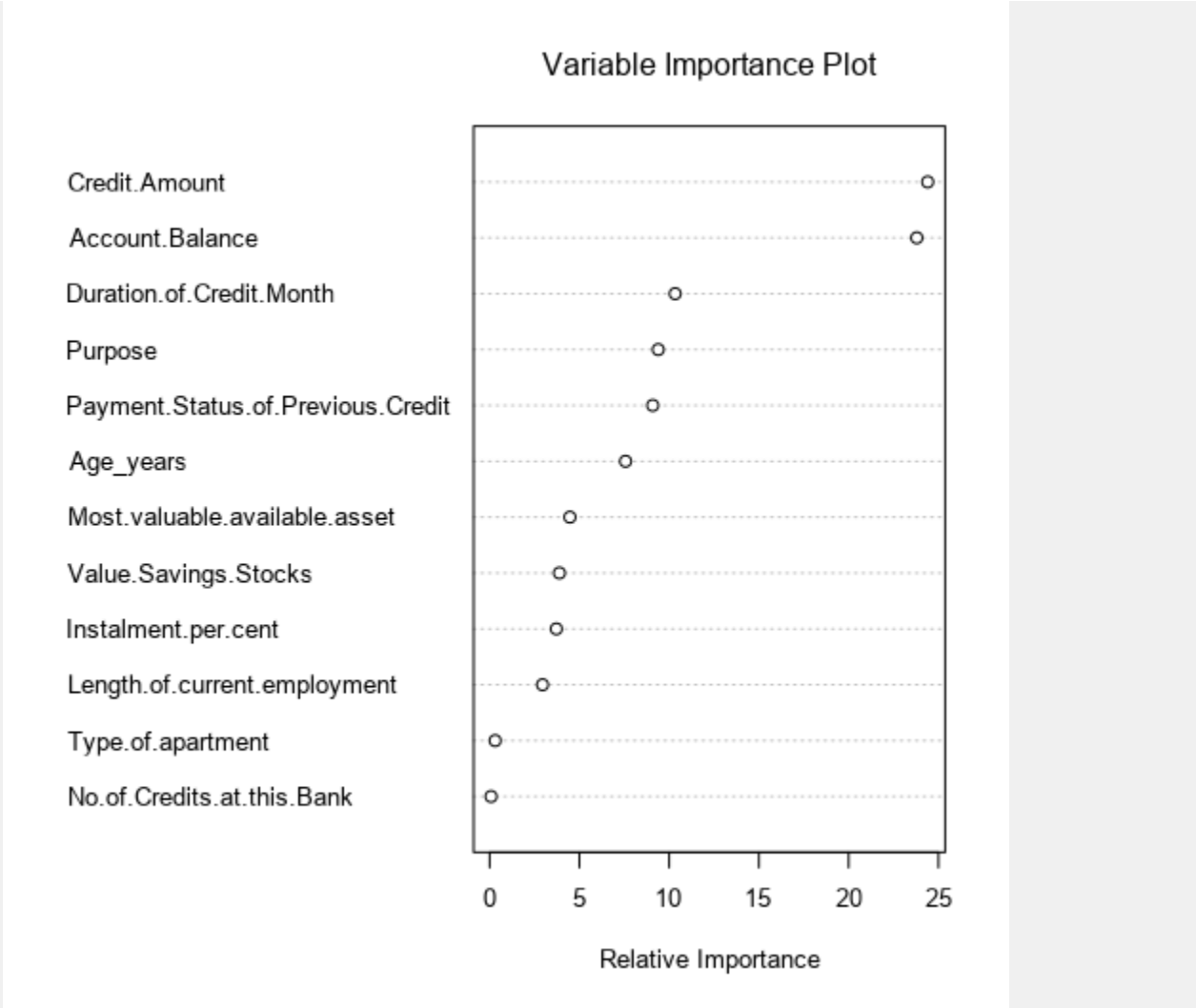
Confusion matrix of FM

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

The Forest Model has an overall accuracy of 80%. It has an accuracy of 80% in predicting creditworthy individuals and an accuracy of 83% in predicting noncreditworthy individuals. The model is not biased.

Boosted Model

The significant variables are Credit-Amount and Account-Balance.



Confusion matrix of Boosted_Creditworthiness

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

The Boosted Model has an overall accuracy of 79%. It has an accuracy of 78% in predicting creditworthy individuals and an accuracy of 81% in predicting noncreditworthy individuals. The model is not biased.

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

Step 4: Writeup

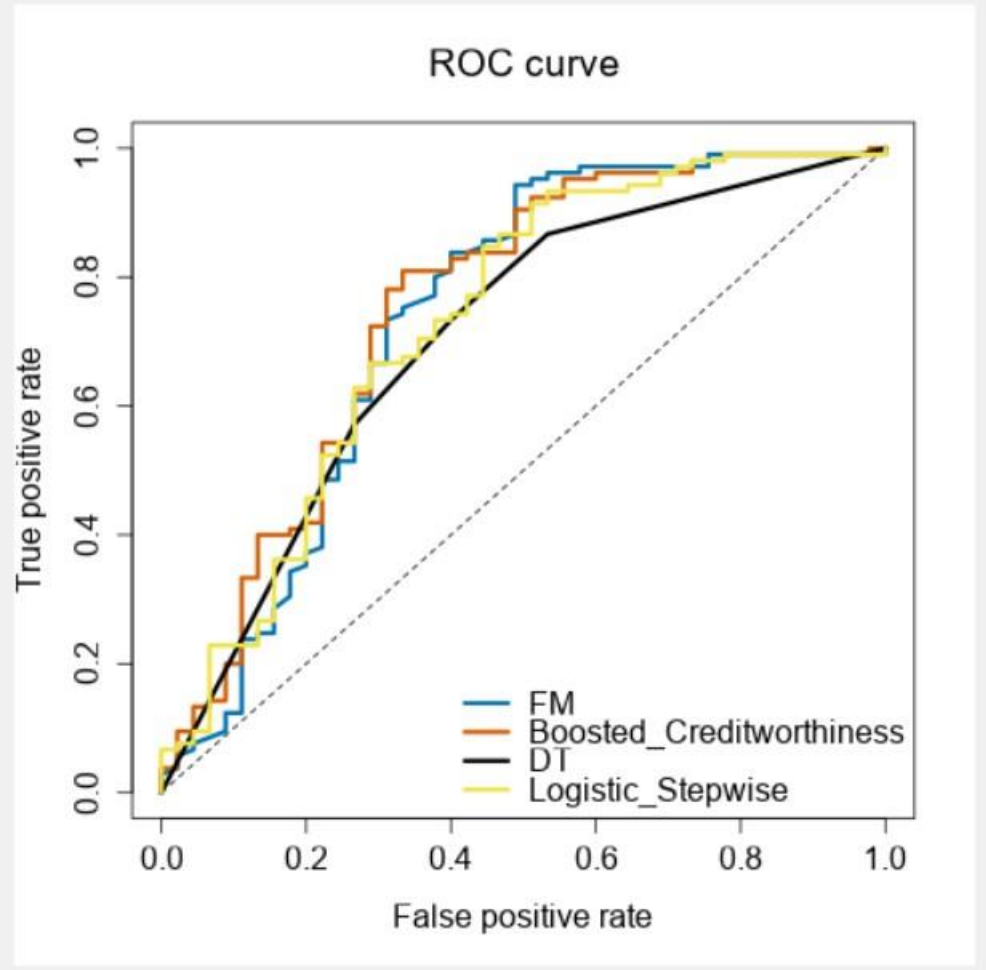
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

I chose the Forest Model because it had the highest Overall Accuracy against my Validation set, the highest accuracy within the "Creditworthy" segment (PPV), the highest accuracy within the "Noncreditworthy" (NPV) segment, and it also had the highest F1 score. The Forest Model was one of the two unbiased models. In the ROC graph, the Forest and Boosted Models performed the best while the Decision Tree model performed the worst. The Forest Model reaches the highest point on the ROC graph, so it has the highest true positive rate.



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

- How many individuals are creditworthy? 406