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?

Ans – We have to take a decision on whether or not to approve loans of each of the 500 loan applications that have been received by the bank (in an unusually short period of time). For this we have to classify each of them as either “creditworthy” or “non-creditworthy” using a predictive model which can classify applications based on past data.

- What data is needed to inform those decisions?

Ans - Firstly we need information on past applications from customers inorder to train the model. Then to use that model to make classification we would need information about the current 500 applicants. We will basically feed the information about them to the model and come up with a classification.

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

Ans - We need a binary model here because we need to classify each applicant as 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”.

No, based on the correlation matrix, there were no fields that had such high correlation.

- 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)*

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

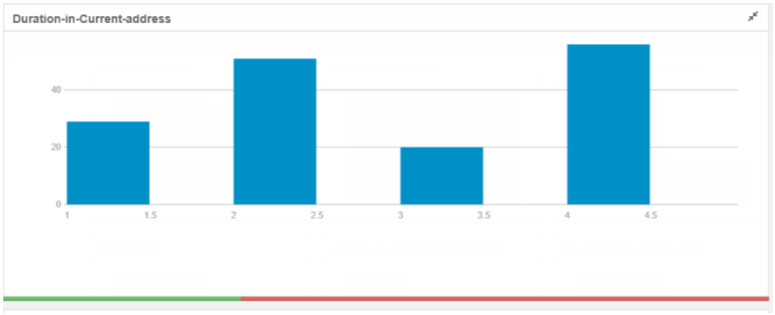
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

Answer this question:

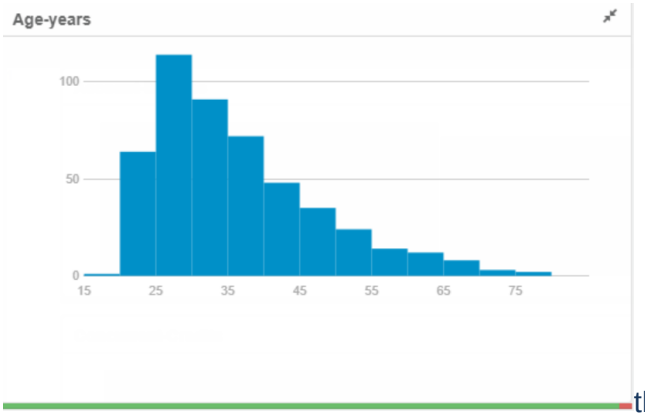
- In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.

The duration-in-current-address field was removed on account of very high percentage of null values. (68%)



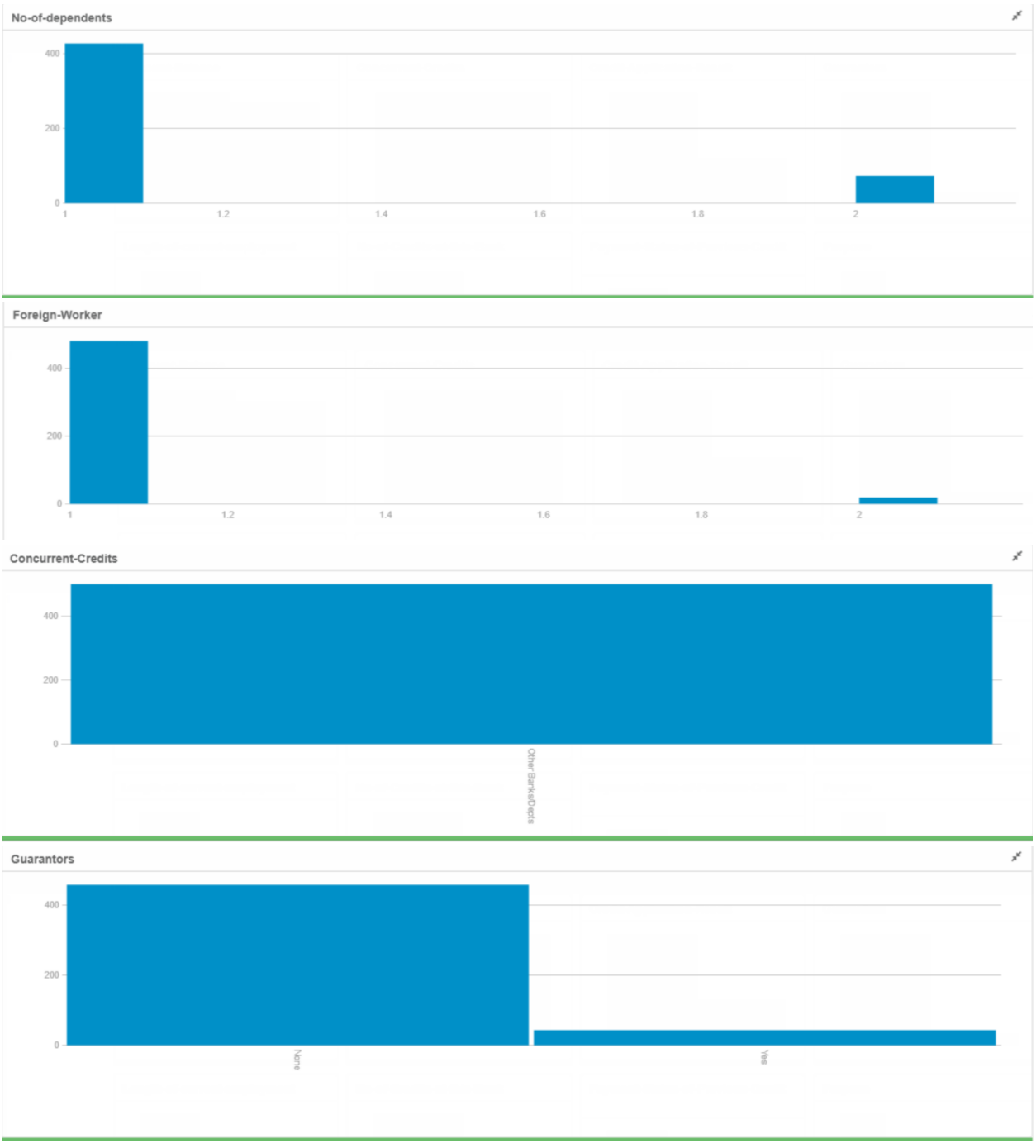
red part indicates the null values

The age years field was imputed because it had just 2.4% null values. The median value of the field was used to impute it.



the small red part indicates the null values

The following fields have been removed on account of “low variability” because either they have one value or are skewed to much toward much :- Guarantors, Concurrent Credits, occupation, no. of dependants, foreign workers.



The telephone field was removed because it was in no way contributing to the classification model.

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:*

Answers for Logistic Regression Model

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

In this model the predictor variables that are significant are – Account Balance, credit amount, instalment percent, purpose, length of employment and payment status. Below are the p values against variables in the second column from the right along with the stars in the rightmost column

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	.

- 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 of the model is 76% . Below you can see the confusion matrix.

Confusion matrix of logistic regression model

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

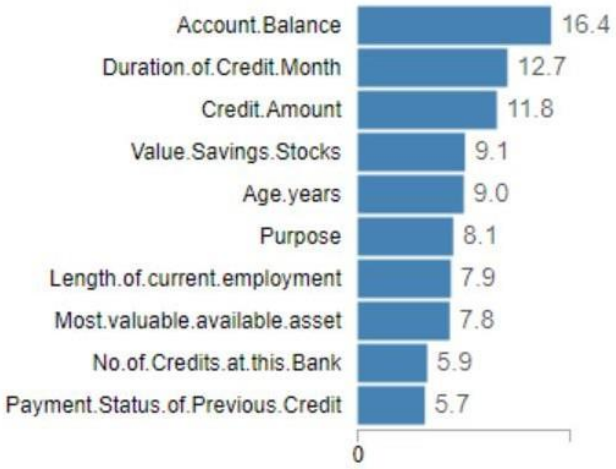
The model is biased towards creditworthy outcome because of the more than 10% difference in predicting accuracy of creditworthy and non creditworthy outcome.

ANSWERS FOR DECISION TREE MODEL

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

  ANS –

  The most important variable is the account balance variable followed by duration of credit month ,credit amount variable,value savings and age. Below is the variable importance chart.



- 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 percentage of this model is 66.67%. Here is the confusion matrix.

Confusion matrix of Decision_Tree_91

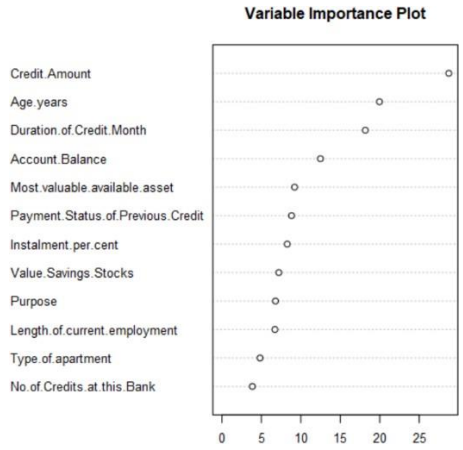
	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	28
Predicted_Non-Creditworthy	22	17

It seems the model is biased the creditworthy outcome because its accuracy in predicting the creditworthy outcome is far better than its accuracy in predicting non creditworthy outcome.

ANSWERS FOR FOREST MODEL

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

ANS -Please refer to the variable importance plot below. As per this plot the most important variables are in this order - credit amount, age years, duration of credit month, account balance and so on.



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?

ANS - The overall accuracy of the model is 79.33%. Below is the confusion matrix. The model is a lot biased towards predicting creditworthy outcome because of the considerable difference in accuracy of creditworthy outcome prediction and non creditworthy outcome prediction.

Confusion Matrix:

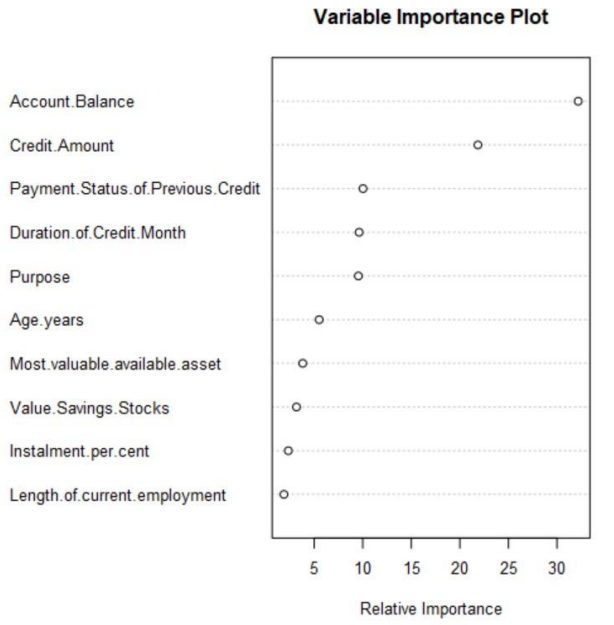
Confusion matrix of forest_model_1

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

ANSWERS FOR BOOSTED MODEL

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

ANS – As can be seen from the variable importance plot below, credit amount and account balance are by far the most important 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?

The overall percent accuracy is 78.67 percent. Below is the confusion matrix. As can be seen here the model is biased towards creditworthy outcome because its accuracy is lot better when it comes to predicting a creditworthy outcome than a non-creditworthy outcome.

Confusion matrix of boosted_model

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

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"

Ans – I've decided to use the forest model to score the values for the new customers. (By which I mean classify them as creditworthy or non 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)

Ans – I have chosen this model because not only does it have the highest accuracy but it also has the highest accuracy in predicting the creditworthy outcome. If a person gets incorrectly classified as creditworthy (even when he is not) the bank stands to lose a lot more than when a person who is creditworthy gets classified as non creditworthy. As per this model, 408 people are classified as creditworthy and would qualify for a loan.

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

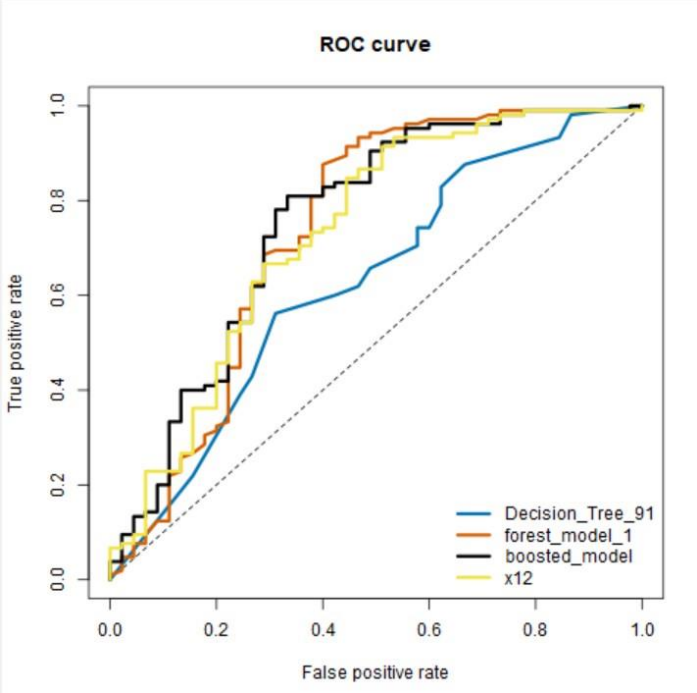
Ans – the overall accuracy against the validation set is 79.33 percent. This is highest among all models.

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

Ans - The accuracy within creditworthy segment was 97.14% which is the highest among all models. The accuracy against the non creditworthy segment was lower though 42% but it was second highest among all models. And there was a stronger emphasis on not classifying a customer as creditworthy when they actually aren't . Hence it is important to focus on the accuracy

    - ROC graph

ANS - As can be seen from the ROC Graph the forest model among all models has the maximum AUC and it is also the model that is most away from the diagnol indicating it is the best performing model



- Bias in the Confusion Matrices

Confusion matrix of Decision_Tree_91

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

Confusion matrix of boosted_model

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

Confusion matrix of forest_model_1

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	28
Predicted_Non-Creditworthy	3	17

Confusion matrix of x12

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

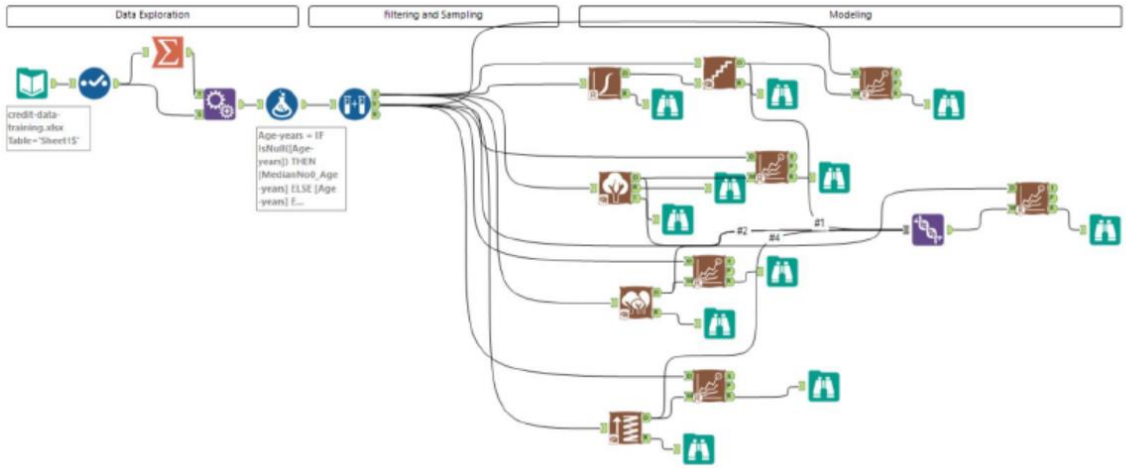
ANS – as can be seen in the confusion matrix above , the forest model has relatively less bias than all other models. And it most accurately predicts the actual creditworthy individuals

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

- How many individuals are creditworthy?

ANS - As per the model 408 individuals are creditworthy.

Alteryx Flow



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.