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?
  The decision is whether the list of 501 applicants will get the loan or not.

- What data is needed to inform those decisions?
  All data related to these applicants for example:
  1. Financial status (salary, bank balance)
  2. Financial obligations (loans)
  3. Demographics
  4. Bank account details
  5. Demographics (age, job)

  Also data from the past about applicants whom their credits were approved for loans.
- What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
  Binary models will be considered to help us gain answers.

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)*

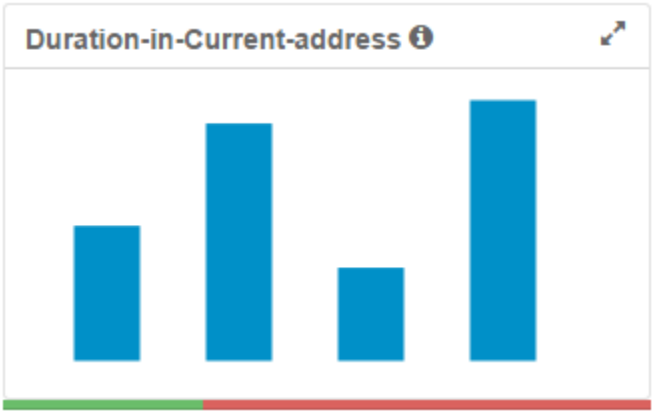
To achieve consistent results reviewers expect.

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.

    - Removed fields:

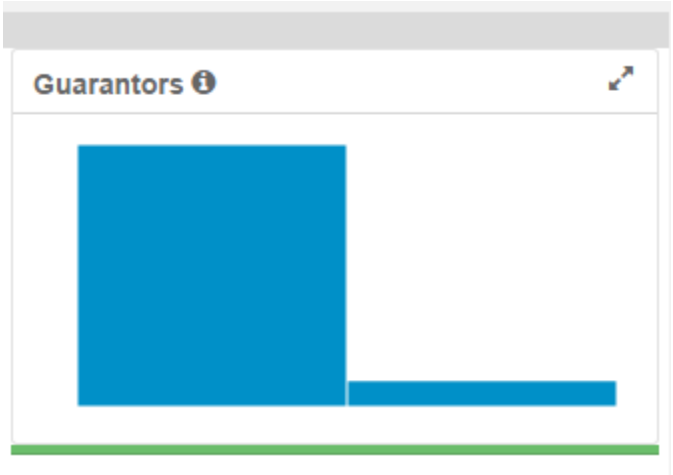
1- **Duration in current address**: Removed because almost 69% is missing :



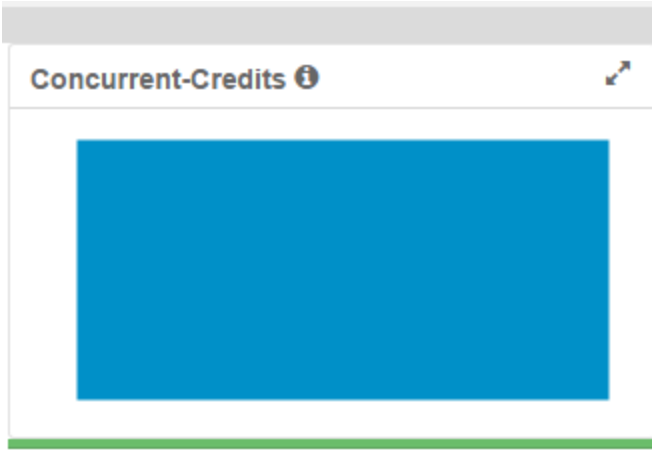
2- **Foreign Worker**: because of low variability, data is heavily skew towards (1.0 to 1.1) , and only 18 instances have the value of (2 to 2.1).

3- **Guarantors** : because of low variability

data is heavily skew towards (None) value.



4- **Concurrent Credits**: because of low variability
There are no other variations of the data, all values are (Other Banks/Depots).

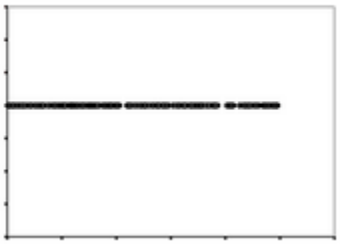


5- **No of dependent** : because of low variability
Data is heavily skew towards (1 to 1.1) value.

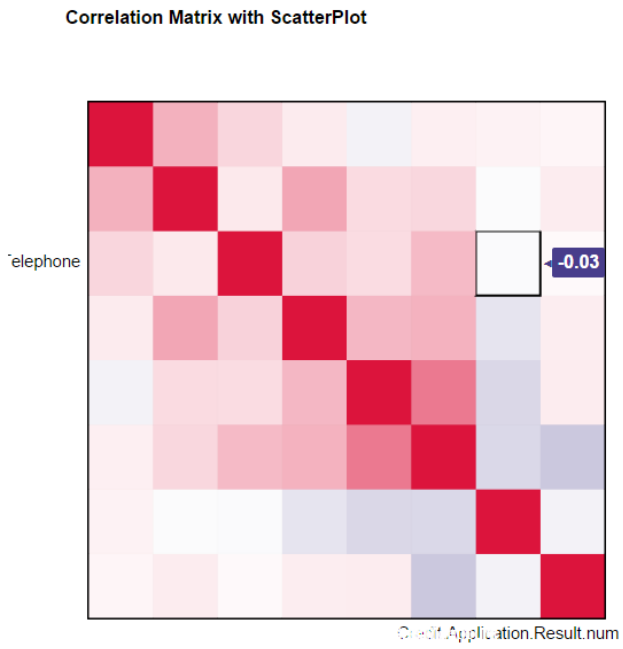
6- **Occupation :** because of low variability

There are no other variations of the data, all values are (1).

Occupation



7- **Telephone:** because it has no logical relationship with the creditworthiness of an individual. Also the correlation coefficient is -0.03.



- Imputed field:

Age years: Null field were imputed to median of age years which is (33), because the missing data is only 2%.

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.

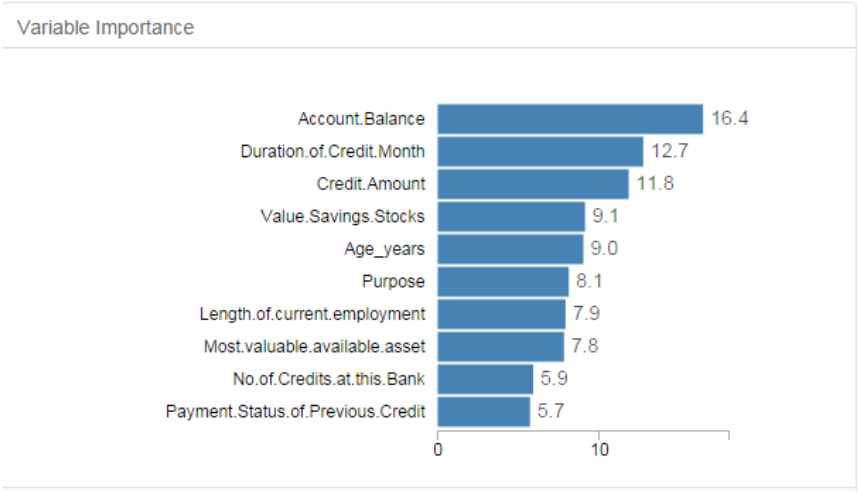
The significant predictor variables are:

- **Account Balance**
- **Duration of Credit Month**
- **Credit Amount**
- **Payment Status**
- **Purpose**
- **Length of current employment**
- **Most valuable available asset**
- **No of credits at this bank**

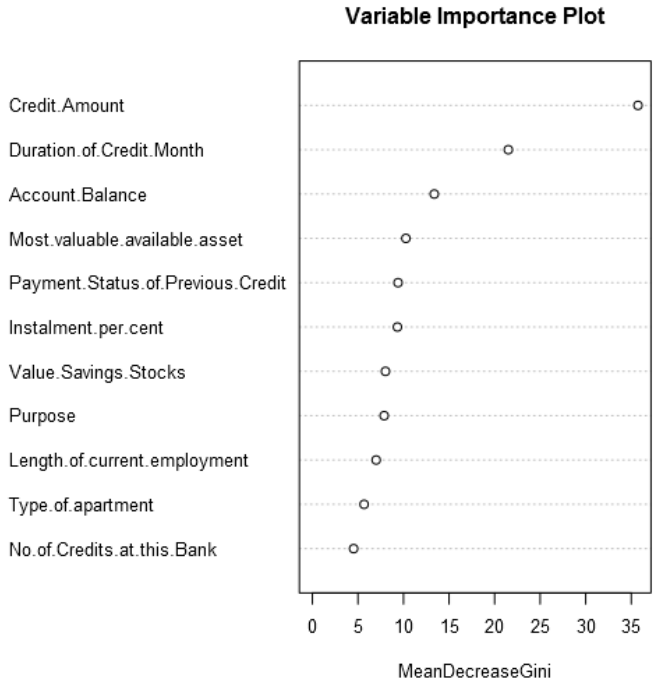
However the most important predictor variables are :

- **Account Balance**
- **Duration of Credit Month**
- **Credit Amount**

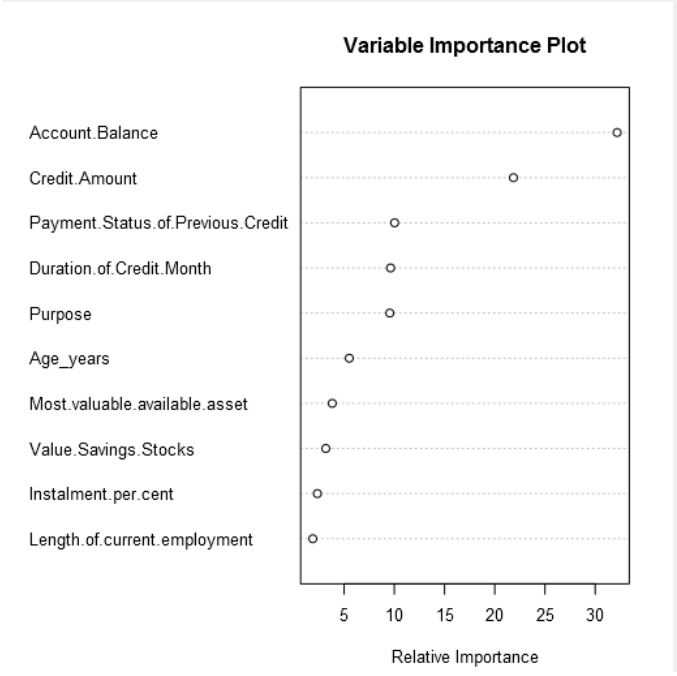
Please see the following bar chart, which shows the variable importance using the Decision Tree Model.



Moreover, the forest model is showing the variable importance plot, see the chart below:



Also, the Boosted Model's plot of variable importance shows almost the same set of 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?

Model	Accuracy
LG_result	0.7800
DT_result	0.6733
FM_result	0.8000
BM_result	0.7867

Confusion matrix of BM_result

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

Confusion matrix of DT_result

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	83	27
Predicted_Non-Creditworthy	22	18

Confusion matrix of FM_result

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	25
Predicted_Non-Creditworthy	5	20

Confusion matrix of LG_result

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	95	23
Predicted_Non-Creditworthy	10	22

In the logistic regression model:

Model is non-biased.

There is small difference the accuracies within "Creditworthy" and accuracies within "Non-Creditworthy", as 90% were predicted creditworthy which is actually creditworthy, however 48% only were correctly predicted (**Non-creditworthy**).

In the decision tree model:

Model is non-biased.

There is small difference the accuracies within "Creditworthy" and accuracies within "Non-Creditworthy", as 79% were predicted creditworthy which is actually creditworthy, however 40% only were correctly predicted (**Non-creditworthy**).

In the Forest model:
Also there is a bias toward (**creditworthy**) results, with a very high percentage of 95% were predicted creditworthy which are actually creditworthy, however 44% only were correctly predicted (**Non-creditworthy**).

In the Boosted model:
There is a bias toward (**creditworthy**) results, as 96% were predicted creditworthy which are actually creditworthy, however 37% only were correctly predicted (**Non-creditworthy**).

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

Step 4: Writeup

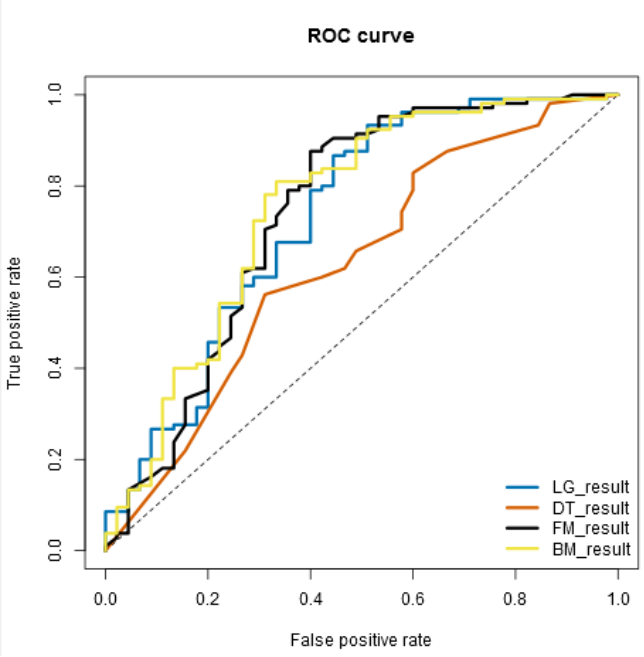
Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"

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

Considering the results in the previous step Forest Model, and Boosted Models, generated the highest percentages of accuracy with 95% and 96%, respectively.
Also when we look at the ROC Curve below, we will see that Forest Model and Boosted Models are higher than the other two models.
The graph clearly shows that Forest Model, and Boosted Model are the highest the most.



However when we check the total accuracy for both models, we will find that Forest Model has an accuracy of 80%, when Boosted Model has an accuracy of 78%.
Thus, I chose Forest Model to use.

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

- How many individuals are creditworthy?

412