Benjamin Owusu Bediako

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

1. What decisions needs to be made?

Answer:
Understand the business problem by using a predictive analysis to list creditworthy customers (applicants) from the 500 influx loan applicants.

2. What data is needed to inform those decisions?

Answer:
Data to inform those decisions are:-

- Credit
- Family size
- Number of children
- Type of job
- Longest tenure at presence job
- Credit history
- Bank statement projecting cash flow
- Collateral available to secure the loan

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

Answer:
Binary model is needed to make those decisions because the business problem involves binary classification.

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.***

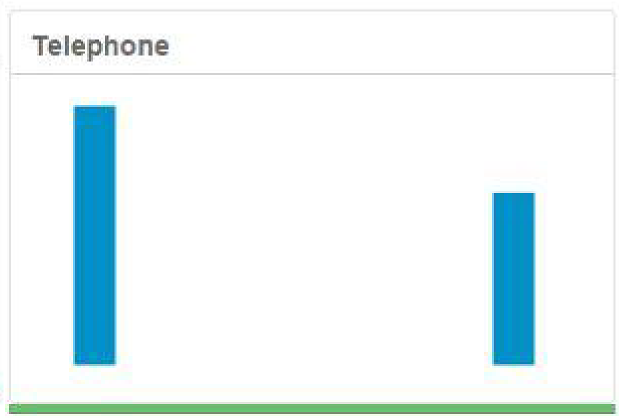
Answer this question:

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

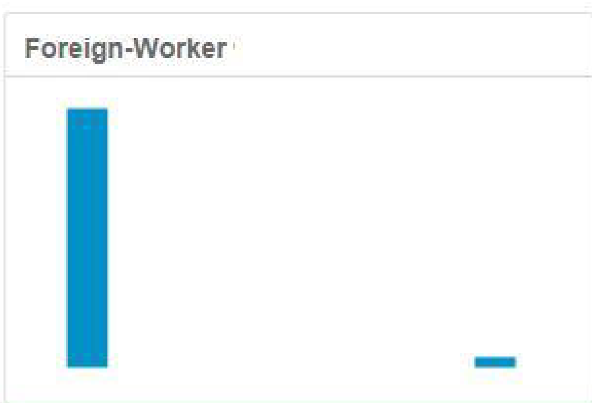
Answer:

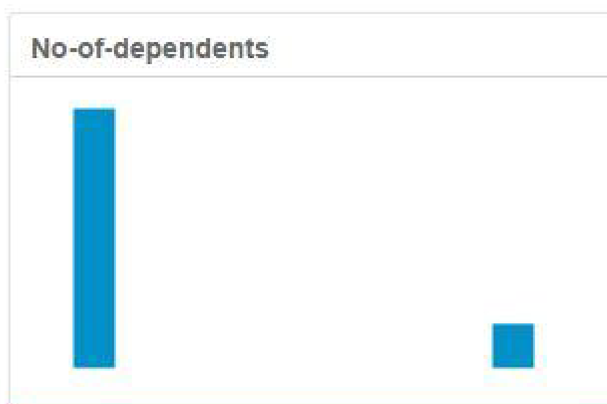
Duration in current address and **telephone** will be removed. Since the **duration in current** address has a higher percentage of null values (on data of 20 fields: numerical and non-numerical fields) upon running the field summary and the **telephone** showing information of loan applicant does not necessary mean the loan must be granted or not since that information might not be known.

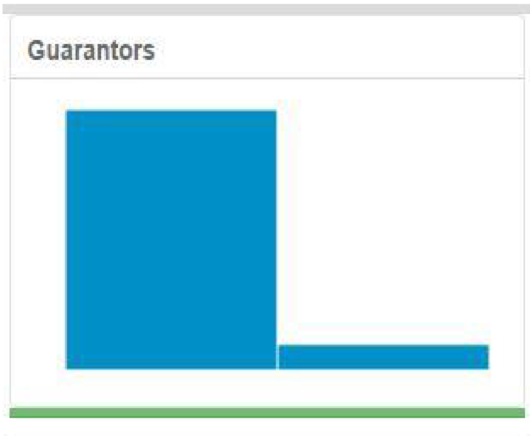
Below are the visualizations of them.

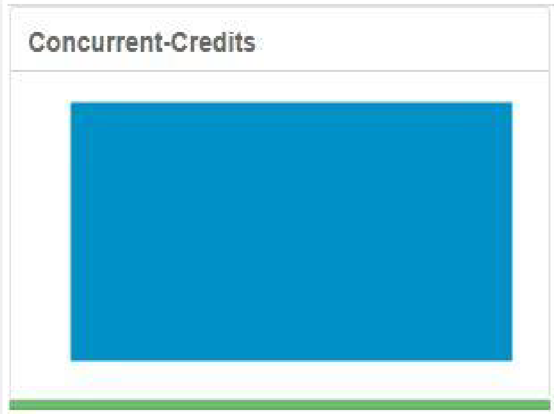


Again, foreign worker, number of dependents, guarantors and concurrent credits shown lack of consistency on data. Therefore their loan applications had to be refused. Below are the visualizations of them.

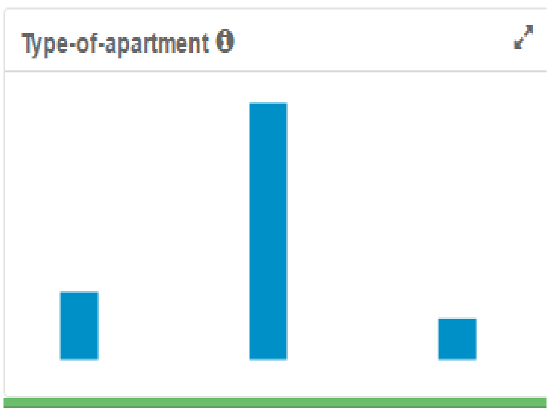






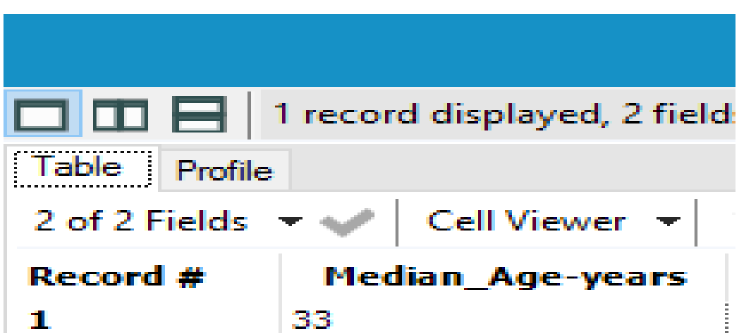


While type of apartment shows inconsistency on data therefore it must be removed.



After the running the models again, thirteen (13) variables shown more conceiving, therefore those will use.

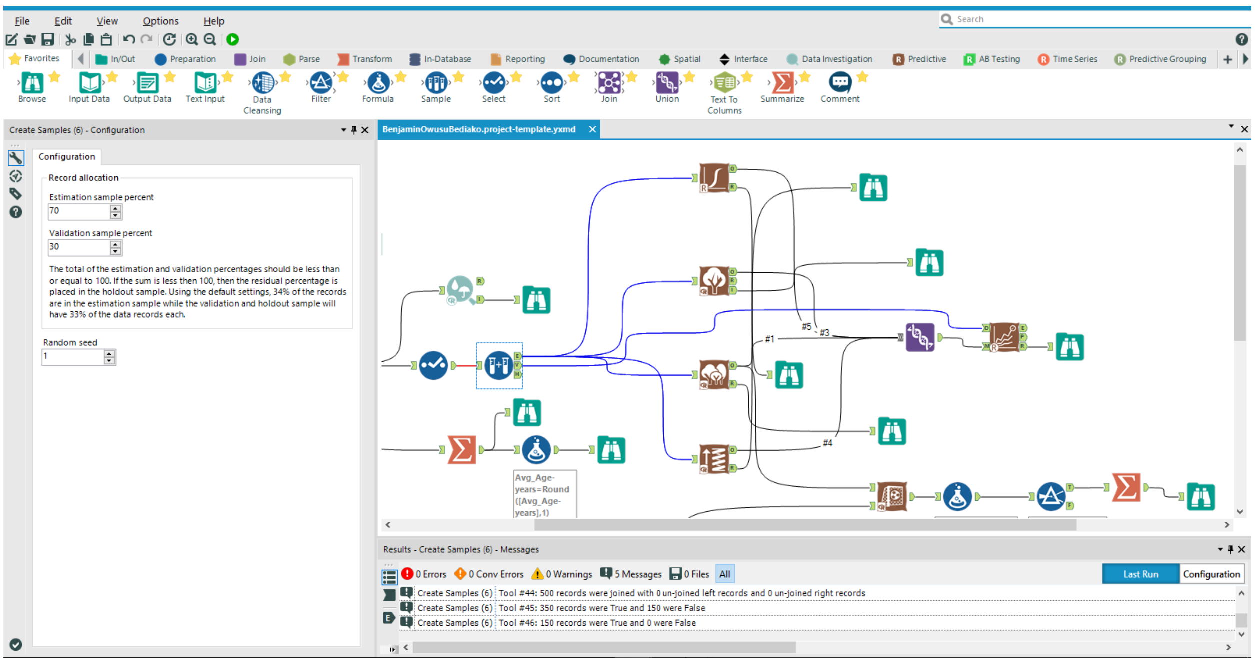
Furthermore, the age year had imputed missing values with the median of 33 which was best to take decision on loan applicants. Since missing data or values means there is no data value for variable, therefore imputing median value will fill in the missing data which are not available because median is best when data does exhibits some skewness. For example, a small number of very large values.



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.

Answer:
Showing the random was set to 1 as the question demands.



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

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

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

Answer:
The predictor variables maps to the 13 records (variables) and below are the results:

i. Logistic Regression

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

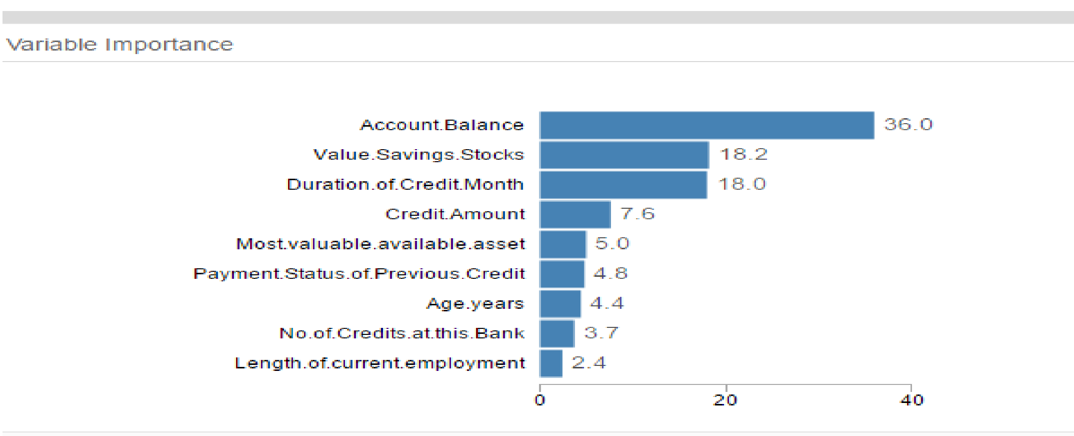
Coefficients:

	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 .

The most important predictor variables are: balance, purpose, payment status, credit amount, length of current employment, credit amount, instalment percent, most valuable available asset.

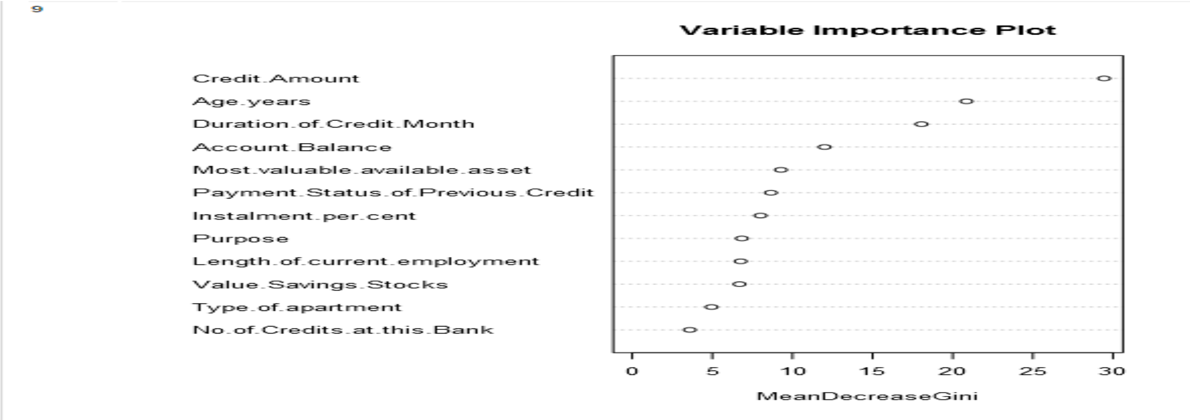
ii. Decision tree

The most important predictor variables are: balance, value of savings stocks, duration of credit month, credit amount, most valuable available asset, payment status of previous credit.



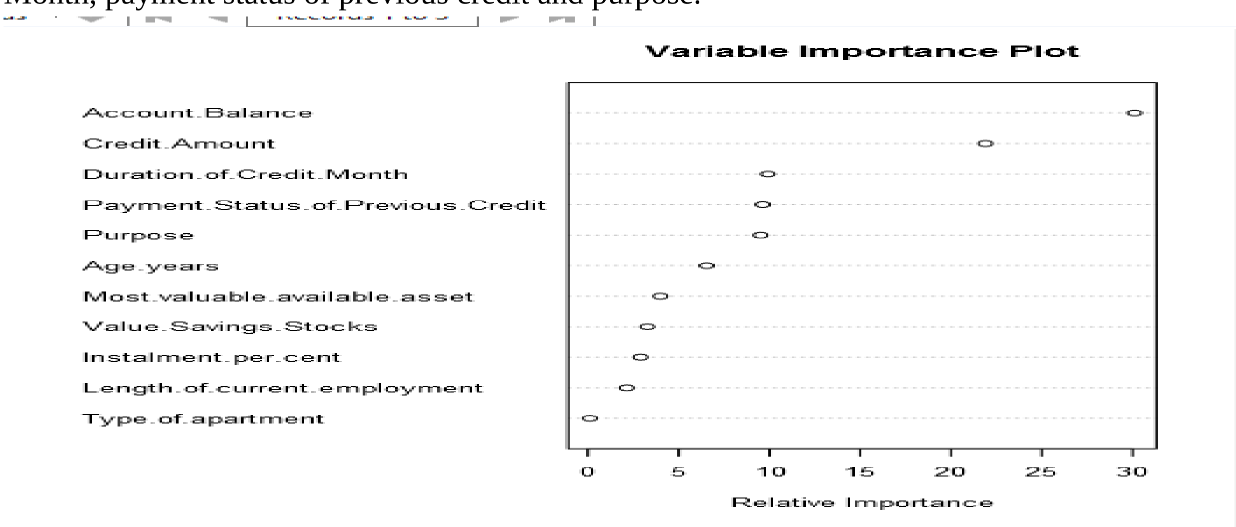
iii. **Forest Model**

The most important predictor variables are: - Credit Amount, Age Years, Duration of Credit Month and Account Balance.



iv. **Boosted model**

The most important predictor variables are: - Account Balance, credit amount, Duration of Credit Month, payment status of previous credit and purpose.



2. 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?

Answer these questions:

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set

Answer:

Forest model had a higher overall percent accuracy.

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

Fit and error measures

Model	Accuracy	F1	AUC
DT_Bank	0.7467	0.8273	0.7054
RF_Bank	0.8000	0.8718	0.7426
BM_Bank	0.7933	0.8670	0.7528
SW_Log	0.7600	0.8364	0.7306

Answer:

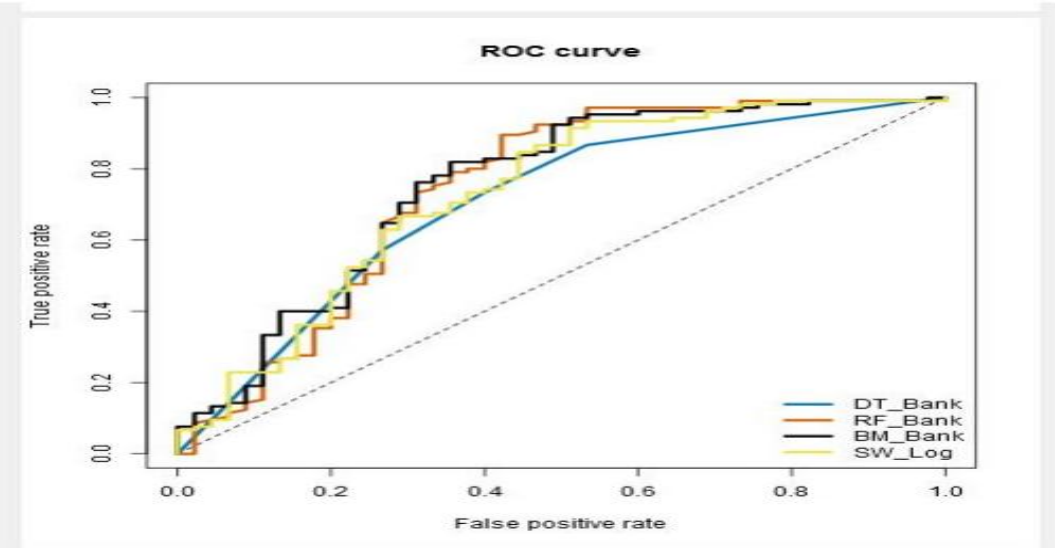
Logistic model predicted the best creditworthy while Forest model predicted the better accuracy non_creditworthy.

Accuracy_Creditworthy	Accuracy_Non-Creditworthy
0.7913	0.6000
0.7907	0.8571
0.7891	0.8182
0.8000	0.6286

c. ROC graph

Answer

The ROC curve asserts that Forest model had better true positive rate.



d. Bias in the Confusion Matrices

Answer:

Confusion matrix of X_Boosted

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	100	27
Predicted_Non-Creditworthy	5	18

Confusion matrix of X_Decision_Tree

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

Confusion matrix of X_Forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	27
Predicted_Non-Creditworthy	3	18

Confusion matrix of X_Logistic

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

Looking at the visualization of the confusion matrices above, it can be asserts that the confusion matrices all have a larger values on **predicted_creditworthy** of **actual_creditworthy** and **actual_non_creditworthy** showing bias**.** But since we are interested in scoring new customers based on the **actual_creditworthy**, the **confusion matrix of X_Boosted** has predicted_creditworthy and predicted_non-creditworthy of 100 and 5 respectively; the **confusion matrix of X_Decision_**Tree has predicted_creditworthy and predicted_non-creditworthy of 91 and 14 respectively; **the confusion matrix of X_Forest** has predicted_creditworthy and predicted_non-creditworthy of 102 and 3 respectively; and the **confusion matrix of X_Logistic** has predicted_creditworthy and predicted_non-creditworthy of 95 and 10 respectively. Therefore, the **confusion matrix of X_Forest model** predicts much better with the highest actual-creditworthy among the other confusion matrices models.

2. How many individuals are creditworthy?

Answer:
If a customer is creditworthy then the count record is 413.

Record #	Count
1	413

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.