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?

  When I am making a Business Problem decision, I always refer to the Business Problem Flow Chart, that follows:

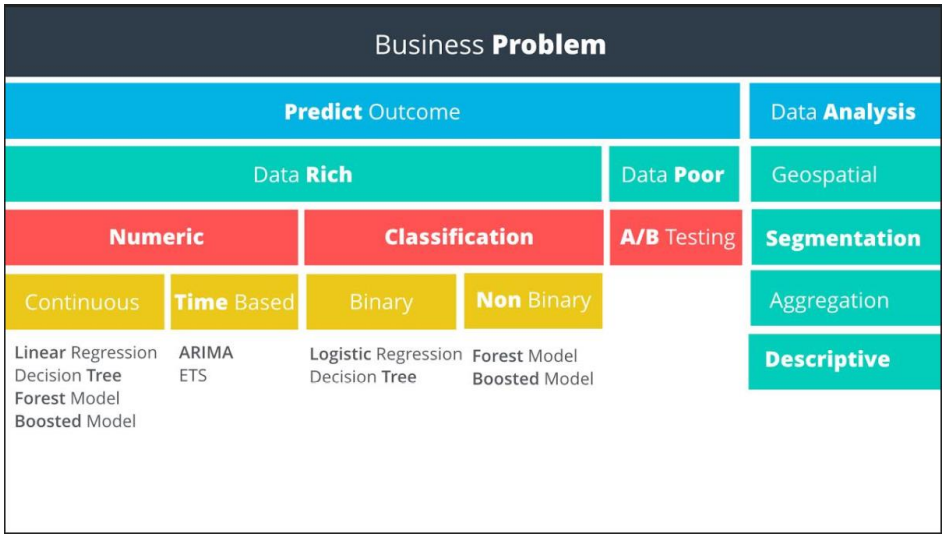


Figure 1: Business Problem Flow Chart
Source: Udacity.com (2018)

  Where one may see that we have to:

  - Predict an Outcome.
  - Is a Data Rich Problem.
  - The Outcome is a Categorical variable, so we must use a Classification model.
  - We have a Binary type of outcome (gets loan or not).
  - Ideally, we must construct a Logistic Regression or Decision Tree model. Better both and compare them.

- What data is needed to inform those decisions?

We need past data on decisions made for previous loan to construct our model, with relevant data that may influence (predictor variables) the outcome of the loan decision. Once we construct our model, we must evaluate it with the data that is going to be used to predict a future outcome.

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

  According to the business problem flow chart, ideally we need to use a Binary kind of model to predict the outcome of the loans (gets loan or not, it only has two (2) ways of going).

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)*

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

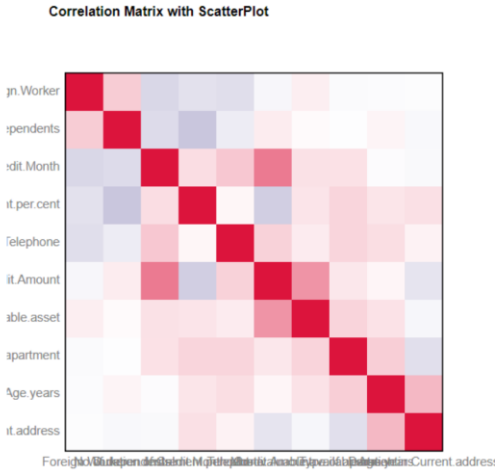
Purpose	String
Credit-Amount	Double
Value-Savings-Stocks	String
Length-of-current-employment	String
Instalment-per-cent	Double
Guarantors	String
Duration-in-Current-address	Double
Most-valuable-available-asset	Double
Age-years	Double
Concurrent-Credits	String
Type-of-apartment	Double
No-of-Credits-at-this-Bank	String
Occupation	Double
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

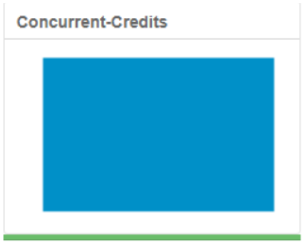
Answer this question:

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

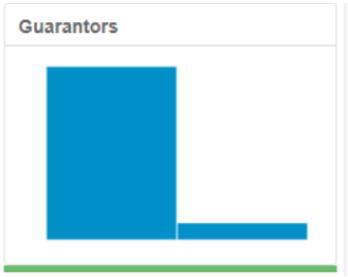
  We see that the variables don't share a correlation that is 0.7 or higher (analysis tool):



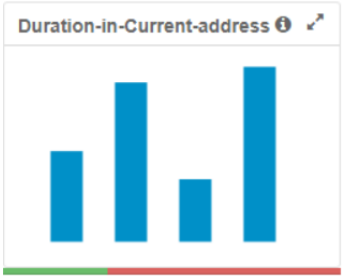
The following variables are to be removed, and the reasons, simplified:
Concurrent Credits: Low Variability, only one unique variable.



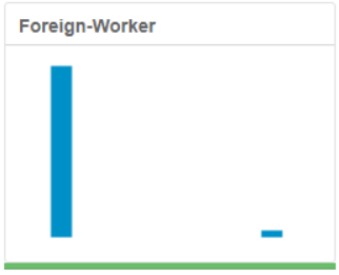
Guarantors: Low Variability, one variable overpass greatly the other.



Duration in Current Address: 69% of missing data, it is too much.

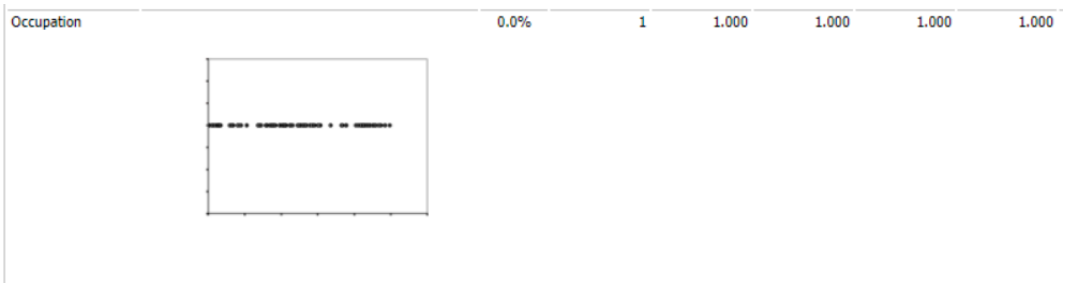


Foreign Worker: Low variability, one instance is way greater than the other.

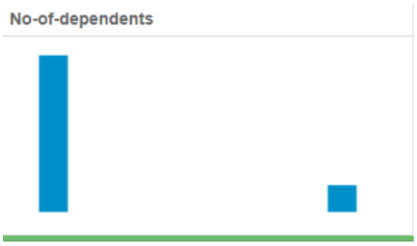


Telephone: Phone number are completely irrelevant.

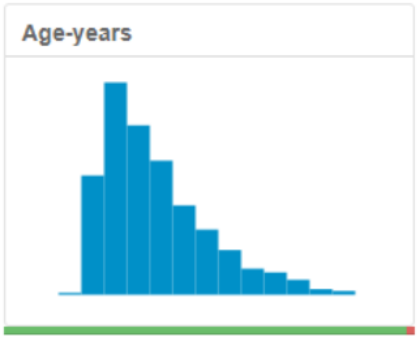
Occupation: Low Variability. One unique variable.



Number of Dependents: Low Variability.



Age-Years: This variable had 2% of data missing. Imputed with median (33).



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.

  Logistic Regression:

As we run Logistic Regression (using stepwise, to improve overall performance), we see the following chart:

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	.

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

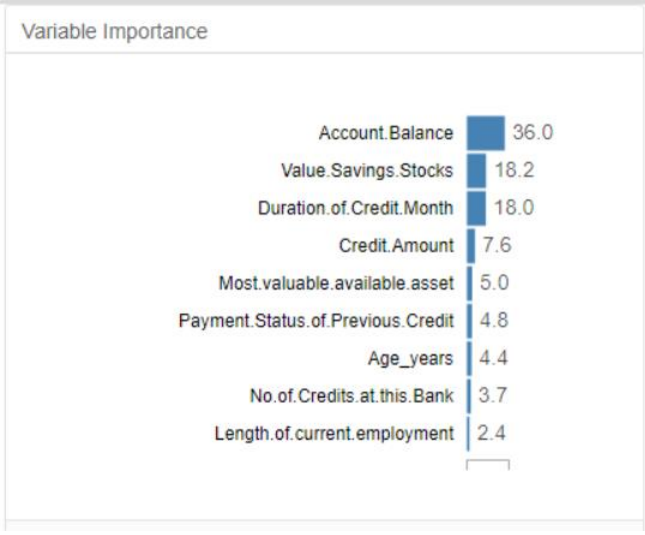
(Dispersion parameter for binomial taken to be 1)

Null deviance: 413.16 on 349 degrees of freedom
Residual deviance: 328.55 on 338 degrees of freedom
McFadden R-Squared: 0.2048, Akaike Information Criterion 352.5

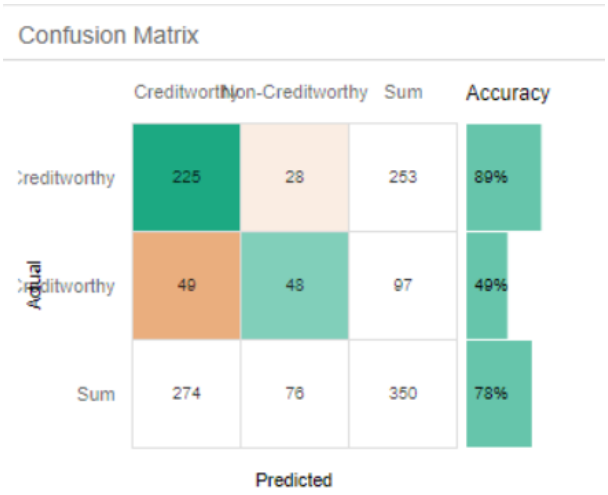
The “p-value” indicator, which shows the statistical significance, made strong assumptions on these variables: Account Balance, Payment Status, Purpose, Credit Amount, Length of Current Employment, and instalment per cent. The R-Squared value is 0.2048, which show a very weak model. The variables shown are supposedly optimal for the model.

Decision Tree:

Then we run the Decision Tree Model and observe the charts:



In this case, we see that the most relevant variables are account balance, value saving stocks and duration of credit.



By seeing this confusion matrix, the overall accuracy for this model is 78%, which is good, but it might be just better.

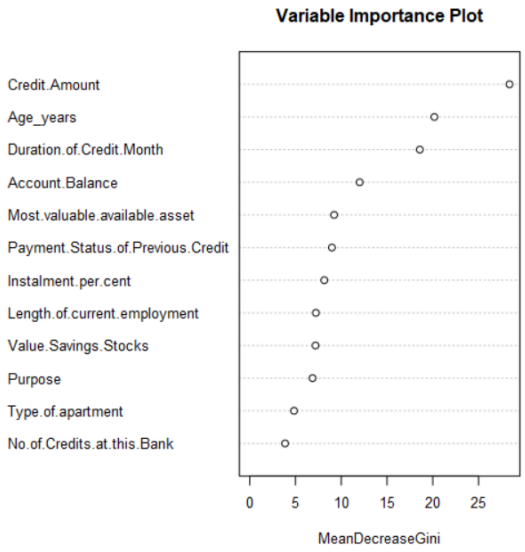
Forest Model:

By running the forest model one may see the error rates and importance variable, where one may start to compare to other models.

Type of forest: classification
Number of trees: 500
Number of variables tried at each split: 3
OOB estimate of the error rate: 36.3%
Confusion Matrix:

	Classification Error	Creditworthy	Non-Creditworthy
Creditworthy	0.087	231	22
Non-Creditworthy	0.639	62	35

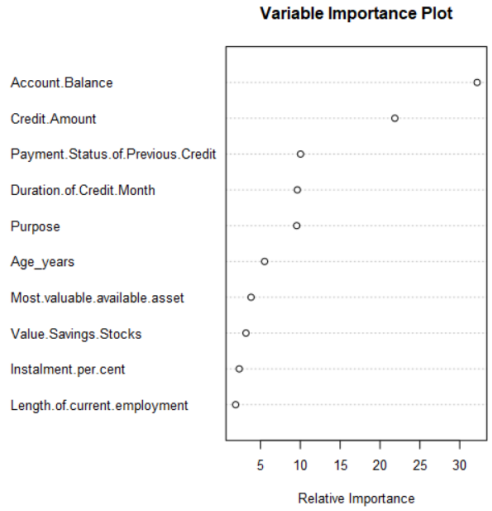
One may see that there is a High OOB estimate error (36.3%) which brings doubts to our model. But when observing the confusion matrix, there is very good predictions made on creditworthy clients with less than a 10% error, but there is terrible prediction rate on Non-Creditworthy clients, with over 50% of error rate.



When observing the importance variable plot, there is notable attention on the variable importance plot, where credit amount, age and duration of credit have the top 3, and account balance is placed 4th (compared to other models which was on the top).

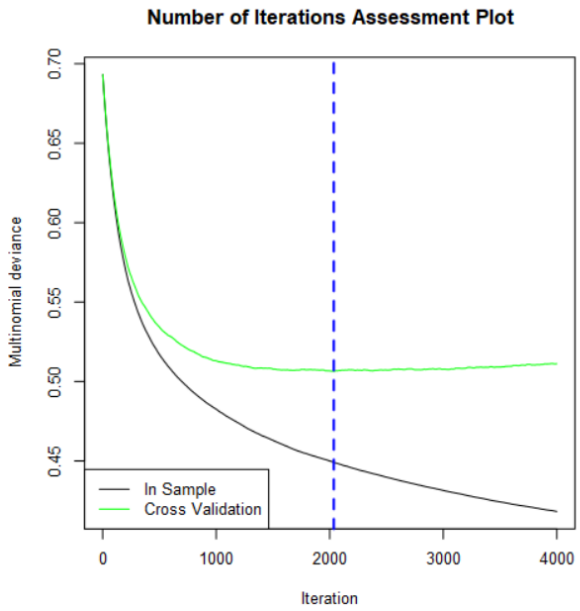
Boosted Model:

Now we run the boosted model, where we observe the variable importance plot and the number of iterations plot.



Account balance placed the first position in this plot, credit amount was second and payment status is third, worth saying that duration of credit places fourth.

The iteration assessment plot shows us strong that the model gets more powerful when it iterates more trees, but has some deviance.



Conclusion:

As seen in the past 4 models, there is a trend with most solid predictor variables, where they stand out:

1. Account Balance
2. Credit Amount
3. Duration of Credit
4. Payment Status of Previous Credit.

- 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?

When validating the given models, we use the union and insert them into the comparison tool to compare all of them. The forest tree had the most accuracy of all, and the boosted model was not so good.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
Decision_Tree	0.7467	0.8273	0.7054	0.8667	0.4667
Forest_Tree	0.8000	0.8707	0.7361	0.9619	0.4222
Boosted_Tree	0.7867	0.8632	0.7524	0.9619	0.3778
Step_Logistic	0.7600	0.8364	0.7306	0.8762	0.4889

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

Confusion matrix of Boosted_Tree

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

Confusion matrix of Decision_Tree

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

Confusion matrix of Forest_Tree

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

Confusion matrix of Step_Logistic

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

The forest model won with 80% of the Accuracy. The average accuracy obtained is 77.33%. It is to be remarked that the creditworthy accuracy was pretty strong in all of the models instead of the boosted one, but they all had sloppy performance in the non-creditworthy section.

All the models are biased towards creditworthy, because they have a difference greater than 10% between them.

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

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

- How many individuals are creditworthy?

Before you Submit

Answering Step 4:

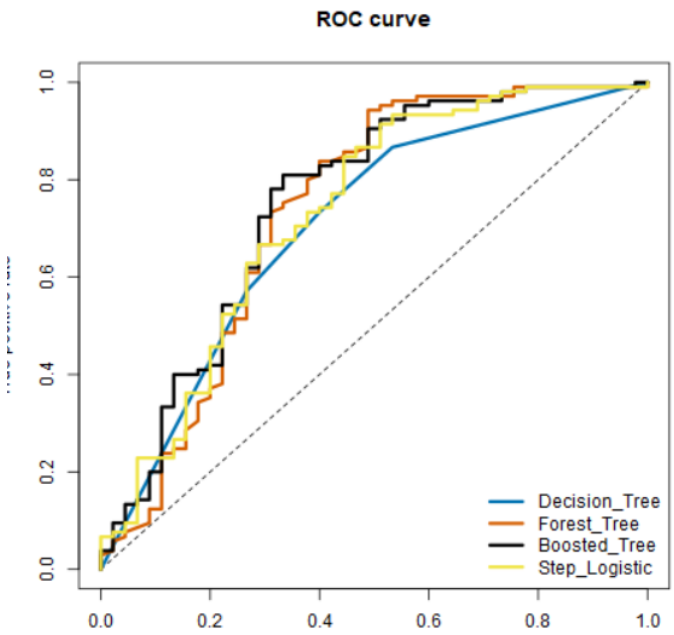
Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

The chosen model for predicting is the forest model, which has a 80% of accuracy, and the overall validation set had a 77.33% accuracy.

In the creditworthy segments, boosted model and forest model show a top performance percentage of over 96%, but they lower greatly in Non-Creditworthy predictions with 42% for forest model and 37% for boosted model (the two lowest in this aspect compared to decision tree and logistic regression). This indicates that in general forest model performs overall better than the rest.

The ROC curve shows us that the forest model has the best performance when predicting fast and accurate values, since it reaches number 1 first.

All four models, as seen in the comparison report in step 3, have biases towards Creditworthy because their difference is greater than 10% and Creditworthy has greater percentage than Non-Creditworthy accuracy.



The total customers creditworthy are: 406.

Final Alteryx: