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 manager needs to exploit the influx of new 500 loan applications by predicting which are the creditworthy customer based on past data of credit data in order to avoid wasting time on non-creditworthy customer and acquiring worthy customers.

- What data is needed to inform those decisions?

  To deicide creditworthy customers, we will use two datasets which are all past application data and data of customers who apply for a loan. The first dataset used to build the prediction model with the following data:
  - Account balance.
  - Age in years.
  - Credit application result.
  - Credit amount.
  - Duration of credit month.
  - Instalment per cent.
  - Length of current employment.
  - Most valuable available asset.
  - No of credits at this bank.
  - Payment status of previous credit.
  - Purpose.
  - Type of apartment.
  - Value savings stocks.

  Also, the second dataset is used to predict the number of creditworthy customers.

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

According to the description of the business problem, the binary model is fitting to make these decisions on coming loan applications since it is about evaluating new application's creditworthiness. Therefore, the application must be 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".
- 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.

Firstly, the concurrent-credits and occupation were removed because they have only one value in all records so, those are ambiguous fields that do not provide any insight to predict the worthy application.

Figure 1: Concurrent-Credits Distribution

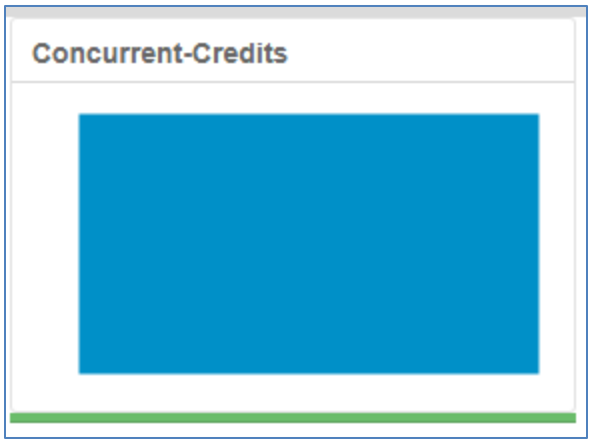
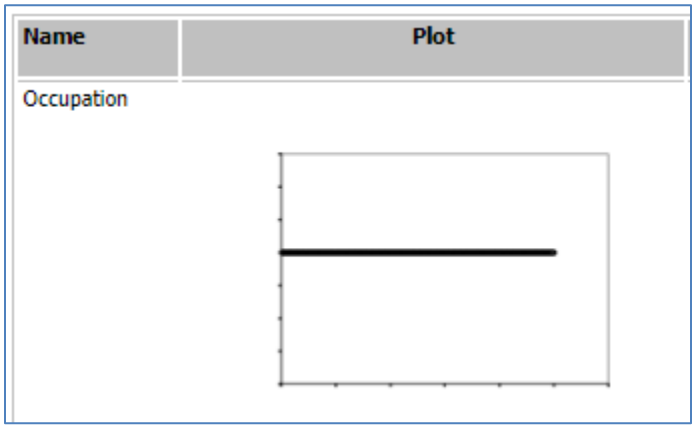


Figure 2: Occupation Distribution



Then, the guarantors, foreign-worker, and no-of-dependents are removed because of their skewed values to avoid bias in the result.

Figure 3: Guarantors Distribution

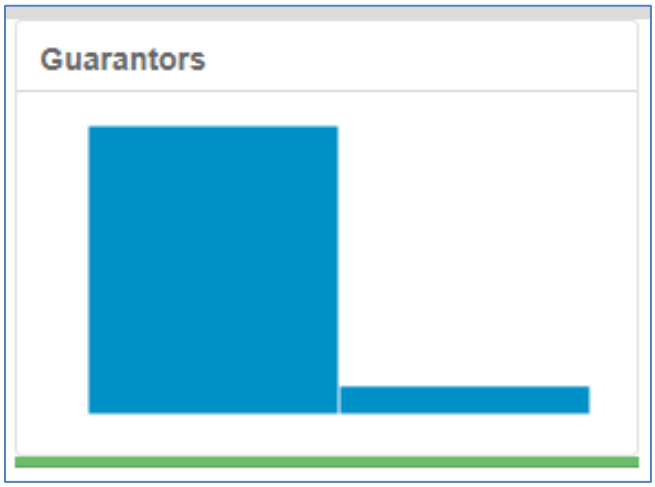


Figure 4: Foreign-Worker Distribution

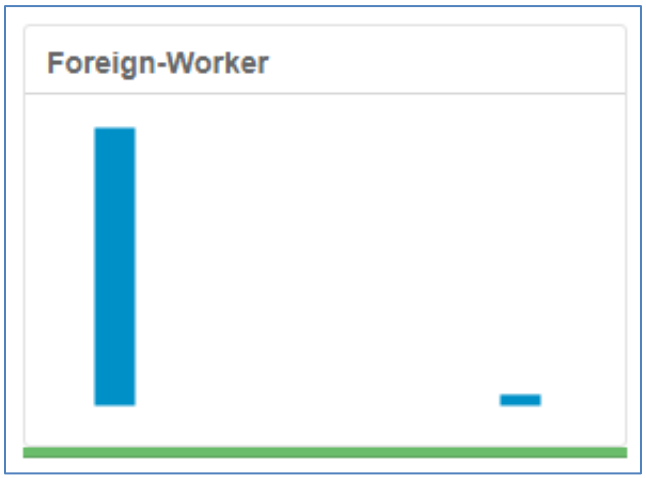
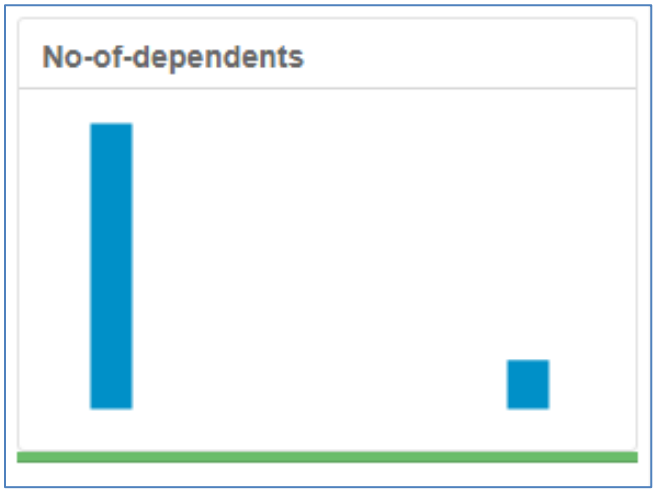
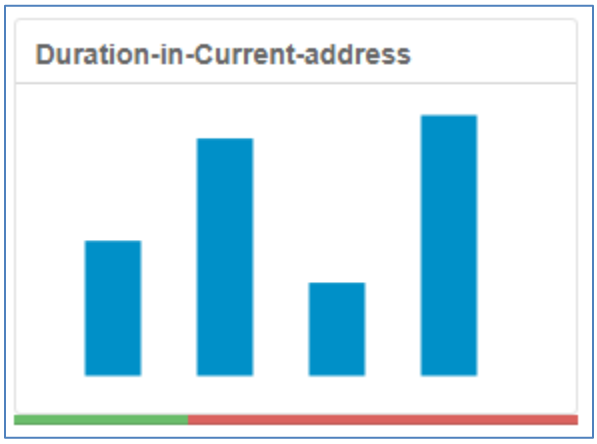


Figure 5: No-of-dependents Distribution



After that, the duration-in-current-address field was removed because most of the values are null. The next figure showed by red color in a tiny horizontal bar.

Figure 6: Duration-in-current-address Distribution



The telephone field deleted due to some reasons that it is informative field like ID. Also, the missing value of age-years imputed by the median of data to align the data distribution.

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?

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

A)

Figure 7: Coefficients of Logistic Regression

	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

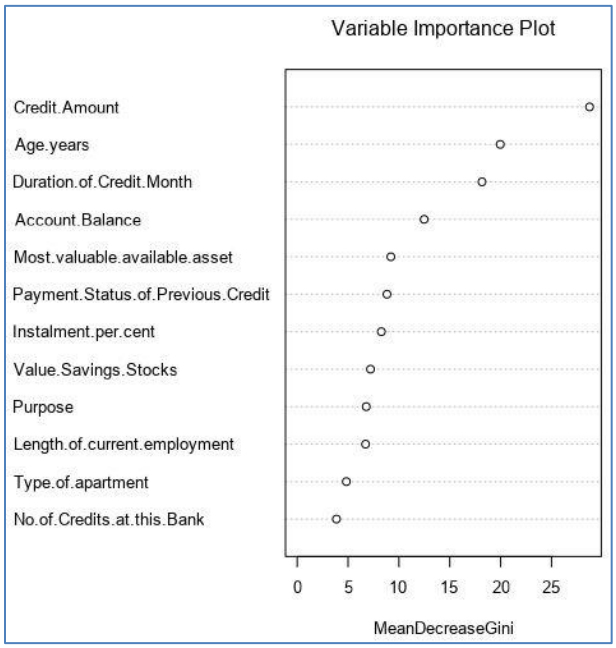
(Dispersion parameter for binomial taken to be 1)

According to above figure, the significant predictor variables of Logistic Regression model are Account-Balance, Instalment-per-cent and credit amount. The rest are part of variable which are some problems in payment status of previous credit, new car in purpose and < 1yr in length of current employment.

Figure 8: Variable Importance in Decision Tree

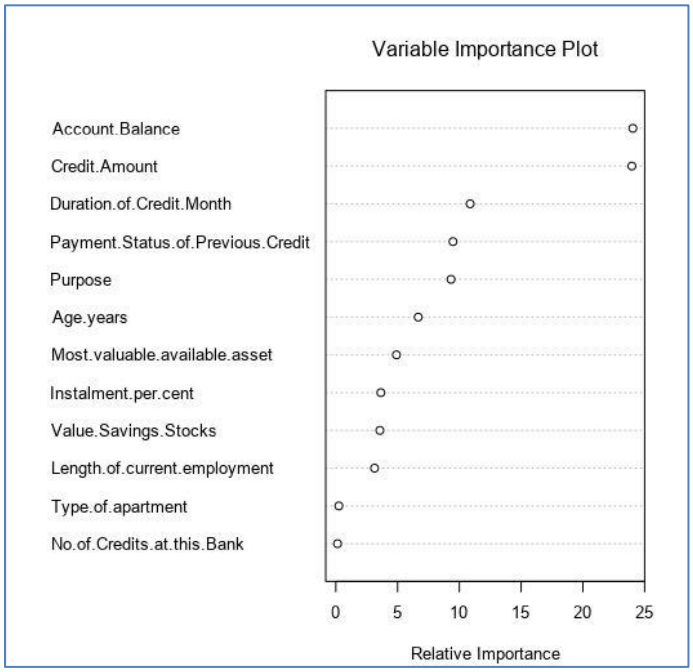
The most important predictor variables in Decision Tree model are account balance, duration of credit month, credit amount and value savings stocks as figure 8 has shown.

Figure 9: Forest Model



The most important predictor variables in Forest model are credit amount, age years and duration of credit month as figure 9 has shown.

Figure 10: Boosted Model



In figure 10, the important predictor variables of the boosted model are account balance and credit amount.

B)

Model Comparison Report

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
FM_1	0.7933	0.8681	0.7368	0.9714	0.3778
DT_1	0.6733	0.7721	0.6296	0.7905	0.4000
BM_1	0.7933	0.8670	0.7505	0.9619	0.4000
st_lr	0.7600	0.8364	0.7306	0.8762	0.4889

The overall accuracy of the models is Forest model has the highest accuracy that is **79.33%**, as well the Boosted model is 79.33%, then the stepwise of logistic regression is 76.00% and the last Decision Tree is 67.33%. As well, the Forest models has the highest accuracy of creditworthy segment while stepwise logistic regression has the highest accuracy of Non-creditworthy.

Figure 11: Confusion matrix of Models

Confusion matrix of BM_1

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	101	27
Predicted_Non-Creditworthy	4	18

Confusion matrix of DT_1

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

Confusion matrix of FM_1

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

Confusion matrix of st_lr

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

In figure 11, the models have predicted well in creditworthy while in non-creditworthy have same performance approximately and is not good like creditworthy. Because of a number of creditworthy samples are huge compared with a non-creditworthy sample. However, the Forest model has the highest accuracy in a creditworthy segment, as well the stepwise Logistic Regression model in the non-creditworthy segment. As a result, all models biased toward creditworthy because of difference between accuracy of creditworthy and non-creditworthy like the Forest model 97.14% for creditworthy and 37.78% for non-creditworthy.

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.

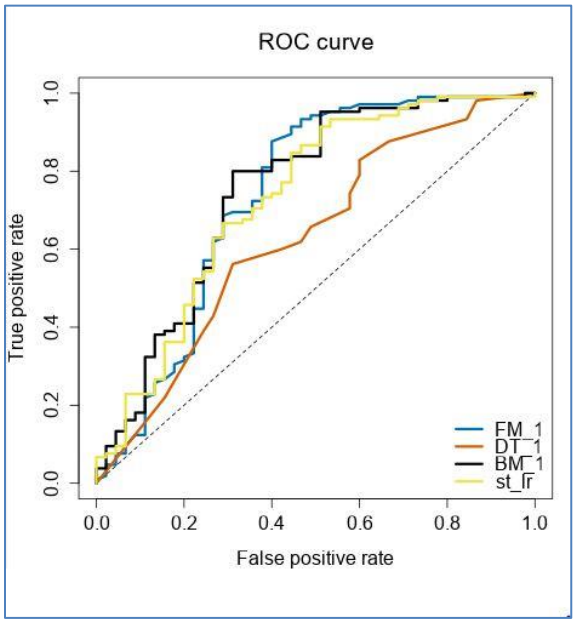
- How many individuals are creditworthy?

A & B)

We decide to use Forest Model for prediction that has the highest accuracy. Because of the overall accuracy of the model (79.33%), creditworthy (97.14%) and non-creditworthy (37.78%) segments as we explained in the previous step.

C)

Figure 12: ROC Curve of Models



The ROC curve represents the ability of a model to predict a true positive sample. So, the Forest model has the highest true positive rate among the other models, and the Decision Tree the lowest rate. As a result, the Forest model is more confident to predict the creditworthy segment.

The bias of the model in the confusion matrix is shown clearly because of a number of positive samples compared with negative in a dataset. Therefore, the accuracy of models in a positive segment is higher than negative. As the figure of confusion matrix in the previous step click here.

D)

The number of creditworthy from the new 500 loan application is **410** and **90** non-creditworthy. The process to get the predicted number of creditworthy are shown in the following figures(13-15).

Figure 13: Prepare Data (Cleaning)

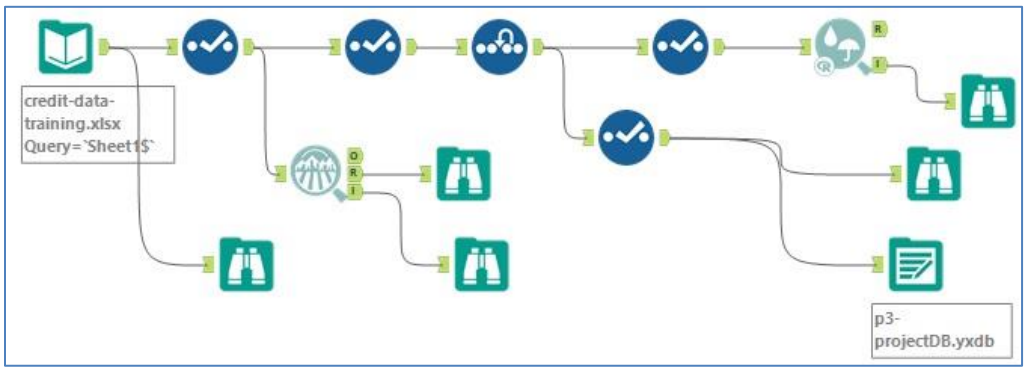


Figure 14: Validation & production model

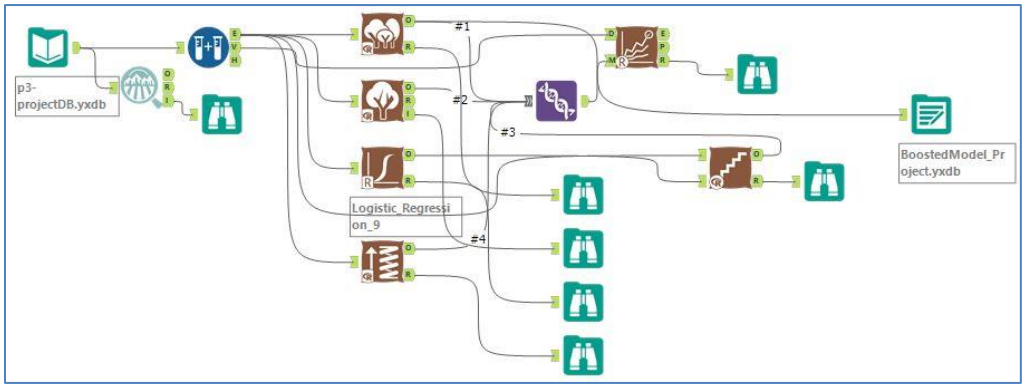
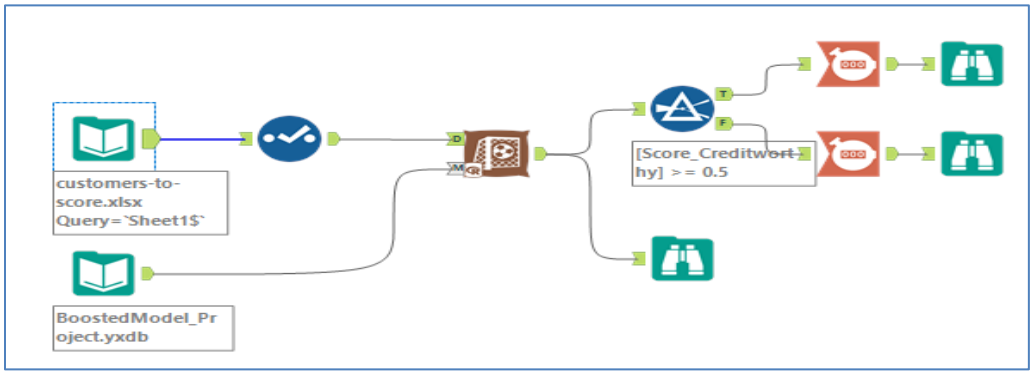


Figure 15: Prediction



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.