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?**

Determine whether the new customers to the bank should be approved for a loan or not based on data about their finances, employment, assets, etc.

- **What data is needed to inform those decisions?**

We would require data on all past applications and their credit application result, data about the assets the persons have in case they default from their payments, their financial status in terms of debt, savings, length of employment, age.

Based on the past data, we would build and train a model to be compared with internal sample data and use the information given (needed) about the new customers to try to predict the outcome of creditworthiness.

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

The decision is a binary one – yes or not (1 or 0) – if the customers are creditworthy or non-creditworthy.

In this problem we have to predict the outcome, we are data rich, it is a classification and therefore a binary solution is demanded.

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

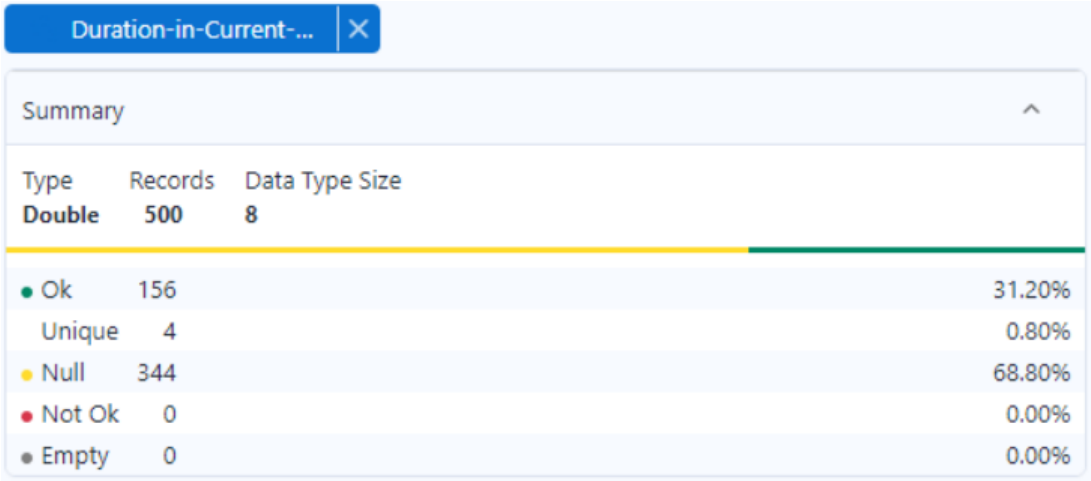
No-of-dependents	Double
Telephone	Double
Foreign-Worker	Double

To achieve consistent results reviewers expect.

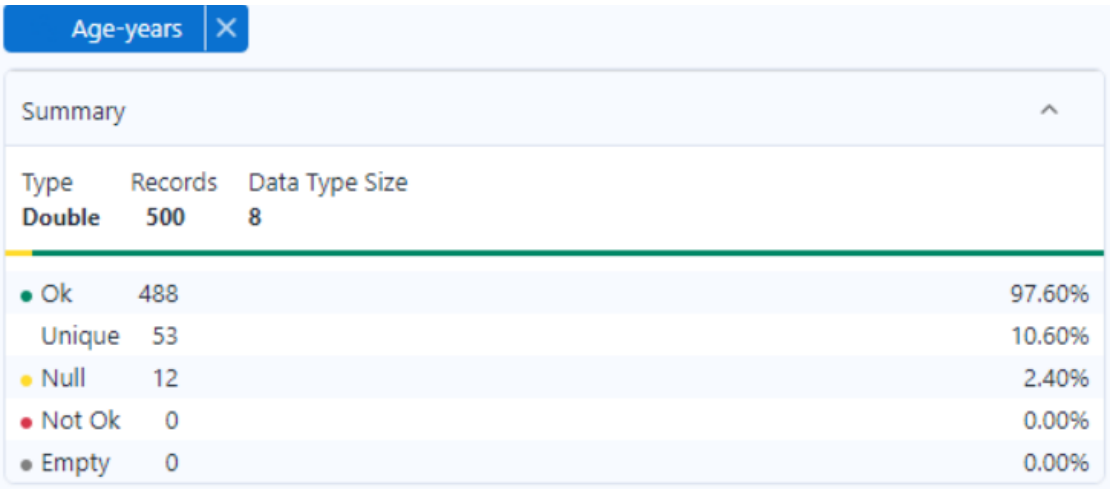
Answer this question:

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

I removed the ‘duration in current address’ field because most of the data was missing and it did not have a significant impact on our predictions.



On the other side, I imputed the ‘age-years’ data with the average of 35.637, because I do not want to erase the customer entirely from the analysis and the age is an important factor to take into consideration even though it does not hold a grand impact on the analysis.



I, furthermore, excluded:

- Concurrent – credits – because of its uniformity data (only one value)
- Foreign-worker – low variability
- No-of-dependents – low variability
- Telephone – not necessary
- Guarantors – low variability
- Occupation – uniformity

+ ‘Duration-in-current-address’

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.**

According to their shown p-value and significance, the predictor variables are the ones who have the lowest p-value or the highest number of stars.

Logistic regression:

For Logistic regression, the most important predictor variables consist of:

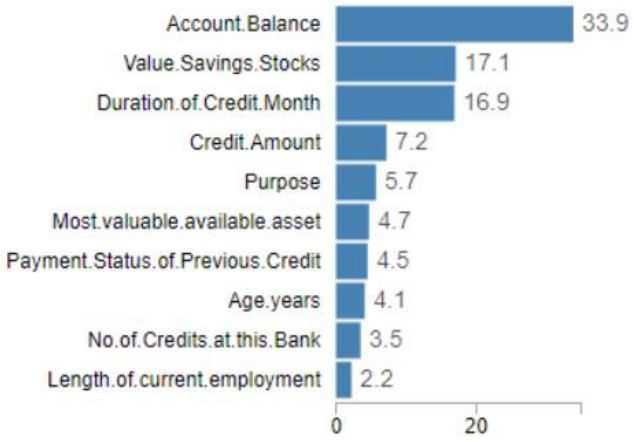
- Account Balance – it is important for customers to have a reserve of money to pay the credit if their income diminishes
- Credit Amount – the total sum that has to be paid is definitely a key factor
- Purpose – people tend to pay their credit if they get a new car (or not)
- Length of current employment – if it is under one year, the variable has an impact
- Instalment per cent
- Payment status

	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	.

Decision Tree:

The most important predictor variables for decision tree are:

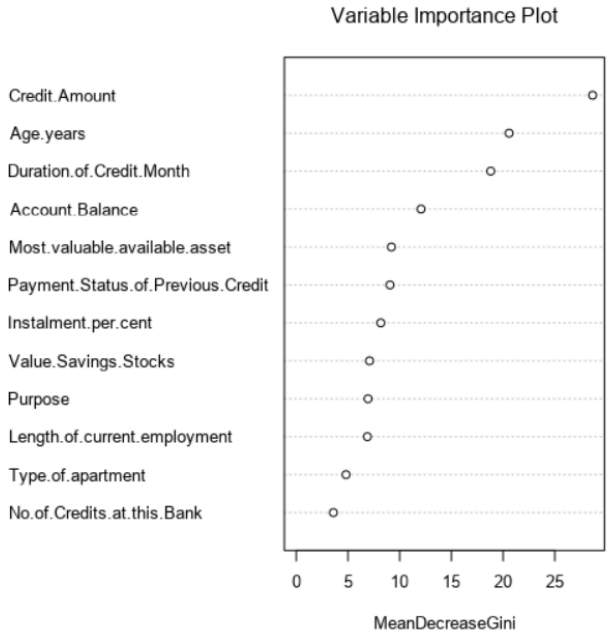
- Account Balance
- Value Savings Stocks – if the customer has other valuable assets which can be liquidated into cash
- Duration of credit month
- Credit Amount
- Purpose



Forest model:

The most important predictor variables for the forest model are:

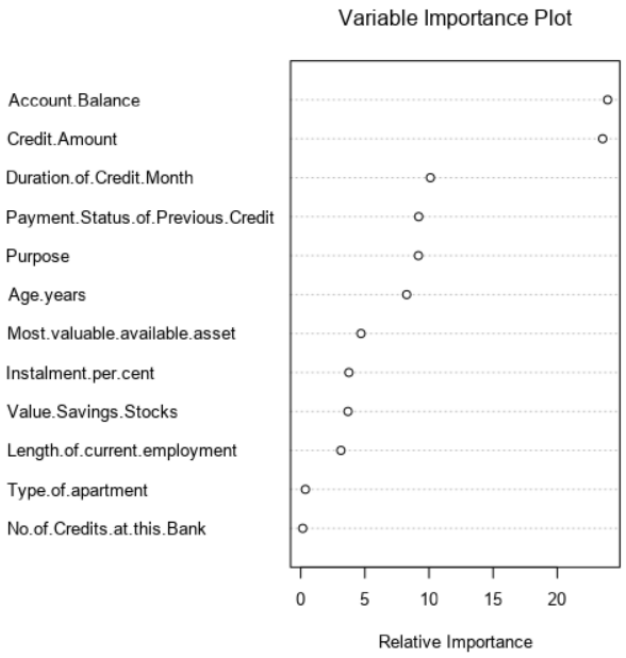
- Credit Amount
- Age Years
- Duration of credit month
- Account Balance
- Most valuable available asset



Boosted model:

The most important predictor variables for the forest model are:

- Account Balance
- Credit Amount
- Duration of credit month
- Payment Status
- Purpose



It is interesting to discover how each method considers different variables to determine the best formula so that the model created is as accurate as possible in predicting the data it worked with in the first place. Some methods are widely known for trying to tailor the model's accuracy based only on the primary data; this is why we have to compare all the methods and give them a validation test with existing data.

Based on this validation test and comparison, we will choose the most efficient and accurate model for our needs.

- **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 models' accuracies vary between 74% and 80% so they all score a high value in predicting the desired outcome and they can all be taken into consideration as valid models.

However, only Forest and Boosted Model attain the percentage of 86% on F1 index which shows the balance between Precision and Recall.

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
log	0.7800	0.8493	0.7333	0.8857	0.5333
forest	0.7867	0.8644	0.7389	0.9714	0.3556
boosted	0.7933	0.8670	0.7539	0.9619	0.4000
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222

Confusion matrix of Decision_Tree

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	26
Predicted_Non-Creditworthy	12	19

Confusion matrix of boosted

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

Confusion matrix of forest

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	102	29
Predicted_Non-Creditworthy	3	16

Confusion matrix of log

	Actual_Creditworthy	Actual_Non-Creditworthy
Predicted_Creditworthy	93	21
Predicted_Non-Creditworthy	12	24

To uncover the bias of each model, we use the confusion matrix and study the numbers shown and use the formulas below:

PPV = true positives / (true positives + false positives)
NPV = true negatives / (true negatives + false negatives)

Decision Tree:
PPV = 93 / (93 + 26) = 0.78
NPV = 19 / (19 + 12) = 0.61

Decision Tree		Boosted		Forest		Logistic Regression	
93	26	101	27	102	29	93	21
12	19	4	18	3	16	12	24
0.781513	0.612903	**0.789063**	**0.818182**	**0.778626**	**0.842105**	0.815789	0.666667

The most unbiased models are Boosted and Forest model. They display a balanced predicting efficiency which does not favor one outcome value over the other.

We can also see that the models tend to predict with a much higher accuracy the 'creditworthy' customers. Therefore, all models have an inclination towards considering the customers creditworthy, even when they are not.

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**

From my point of view, the Boosted Model scores better than Forest Model:

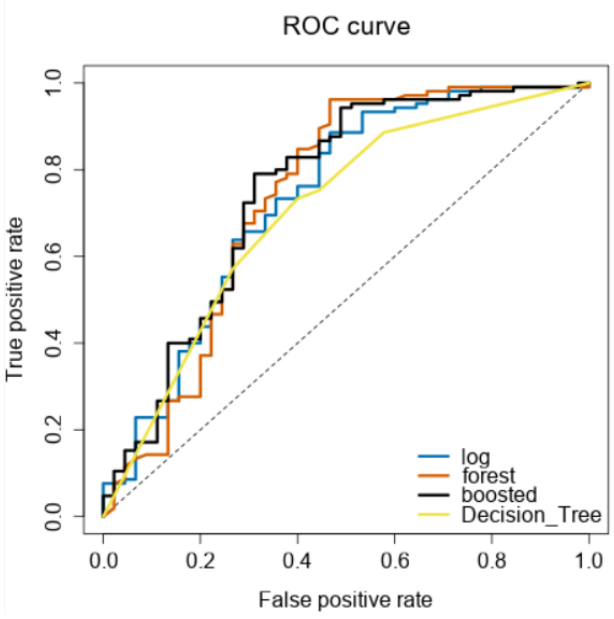
- Accuracy: Boosted 79.3% > 78.6% Forest
- F1: Boosted 86.7% > 86.4% Forest
- AUC: Boosted 75.4% > 73.9% Forest

Fit and error measures

Model	Accuracy	F1	AUC	Accuracy_Creditworthy	Accuracy_Non-Creditworthy
log	0.7800	0.8493	0.7333	0.8857	0.5333
forest	0.7867	0.8644	0.7389	0.9714	0.3556
boosted	0.7933	0.8670	0.7539	0.9619	0.4000
Decision_Tree	0.7467	0.8304	0.7035	0.8857	0.4222

Even though the Forest model scores better by a very small margin on Accuracy_Creditworthy (97% > 96%), Boosted has less bias on Accuracy_Non-Creditworthy (40% > 35%). As a result, the Boosted model has a higher overall accuracy.

Analyzing the ROC curve, it justifies our decision in choosing between Forest or Boosted as the most valid models to train the data and predict the outcomes. The biggest AUC (Area Under the Curve) is given by the Boosted model, which in the ROC curve can be visualized as being the line that gets first on the top.



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

- **How many individuals are creditworthy?**

  Using the Boosted Model, we reach the number of 441 creditworthy customers.
  Using the Forest Model, we reach the number of 412 creditworthy customers.

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.