# Project: Creditworthiness

Shahrooz Govahi

## 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**: I am working in a small bank and I am responsible for determining a potential client is creditworthy to give a loan to or not. We have a huge amount of requests from new applicants and I should provide a report that shows which applications should be approved for giving loans to them.

2. What data is needed to inform those decisions?
   **Answer**: I need a dataset related to our previous clients and the result of their loan. I need different fields so I can build a model base on those fields and the target variable would be "Credit-Application-Result" base on previous clients financial behavior related to their loans. Some fields that could be helpful are like these ones:
   Number of Loans, Age, Credit Score, Previous Payment Behavior, Age of Credit Cards, Total Debt, Percentage of Total Credit Usage, Employment Status, Owning a Home, Level of Education, Job Title, Marital Status, Being Active on Professional Social Media, Range of Income and based on paying off previous loans and other financial factors each individual should be labeled as Creditworthy or Non-Creditworthy in a separate field. Then based on a trained model we can determine new applicants are similar to which category and then label them like our previous clients.

3. What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?
   **Answer**: Base on our labels that present in "Credit-Application-Result" variable, we have two labels: "Creditworthy" and " Non-Creditworthy". So, we need a binary classification model and that's a supervised learning problem. I will train this model base on important features we have in our dataset to classify new applications in one of two mentioned categories. Base on the result we can make decision the applicant is creditworthy or not and we can process the client's request for loan or reject it.

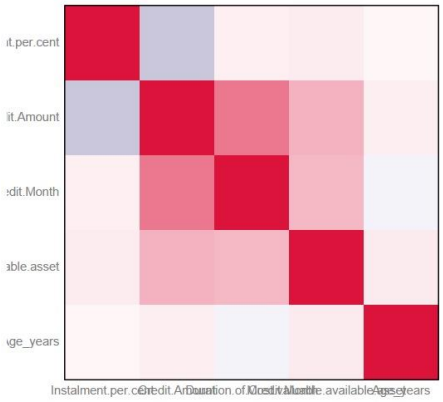
## Step 2: Building the Training Set

*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".*

  ***Answer**:* There is not any high-correlation between numerical data fields. The most correlated fields are "Duration-of-Credit-Month" and "Credit-Amount" and the correlation coefficient for these two variables are 0.57 and it is not high base on the determined threshold.




- *Are there any missing data for each of the data fields? Fields with a lot of missing data should be removed.*
  ***Answer**:* Yes, there is one field with 69% missing value: "Duration-in-Current-address" and I will remove it in next part.

- *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.*
  ***Answer**:* There are 6 fields that has low variability and I will discuss them in the next part.

- *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 average 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 |

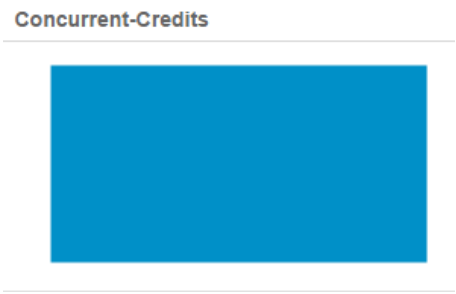
| | |
|---|---|
| Duration-in-Current-address | Double |
| Most-valuable-available-asset | Double |
| Age-years | Double |
| Concurrent-Credits | String |
| Type-of-apartment | Double |
| No-of-Credits-at-this-Bank | String |
| Occupation | Double |
| No-of-dependents | Double |
| Telephone | Double |
| Foreign-Worker | Double |

*To achieve consistent results reviewers expect.*

*Answer this question:*

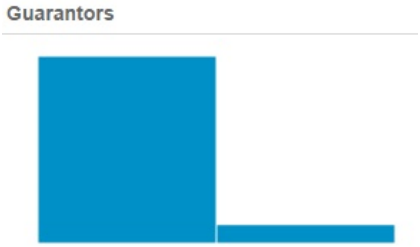
1. In your cleanup process, which fields did you remove or impute? Please justify why you removed or imputed these fields. Visualizations are encouraged.
   **Answer**: Between these 20 fields, I decided to remove these 7 fields:

   - **Concurrent-Credits**
     As it's shown in the histogram of this variable, there is only one value for this column: "Other Banks/Debts". So it has a very low variability and it doesn't help our classification model.

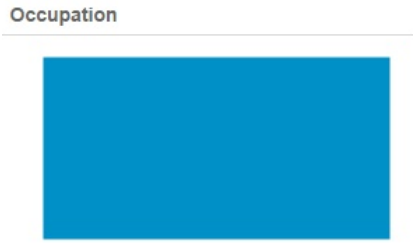



   - **Foreign-Worker**
     Base on the histogram of this variable, there are only two values available for this column: "1" and "2". The "1" category contains 481 instances and the "2" category contains 19 rows. This is a low variability field and it's not very helpful for making distinguish between cases.

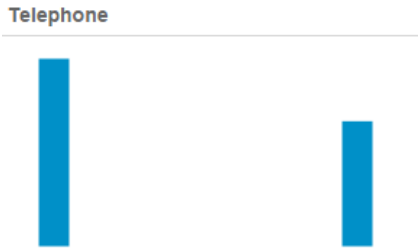

   - **Guarantors**
     This fields is the same as previous one and it has only two values: "None" and "Yes". The first value contains 457 instances and the second one contains 43 rows. So it has a low variability.



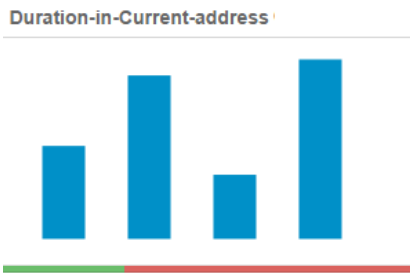
- **Occupation**
  This fields has only one value :"1". It means all loan appliers are in the same category and there is no distinction between them in this field. So it has a very low variability.



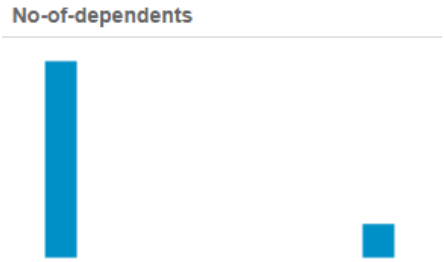

- **Telephone**
  This field has 2 values and it seems like it means an applicant has a telephone or not. Conceptually it's not a relevant variable for creditworthiness of a person. So I remove it.




- **Duration-in-Current-address**
  The red part of the underneath ribbon shows percentage of missing value in this field. It means 69% of this variable is missing in our data set. So it is not a helpful variable for using in our model.




- **No-of-dependents**
  This variable has two values. 1 and 2;
  The first value has 427 instances and the second one has 73 rows. It has low variability and very helpful for our model.




There is one field with 2% missing values in the dataset: **Age-years**. This amount of missing values is acceptable and I will assign Median of other values for these 2% missing ones. The reason of choosing median as a substitute value is that median is one of central tendency measures and it is more robust than mean and it is a better representative of all non-missing values. The median of these values is 33 years and I've assigned this number for missing 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.*

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

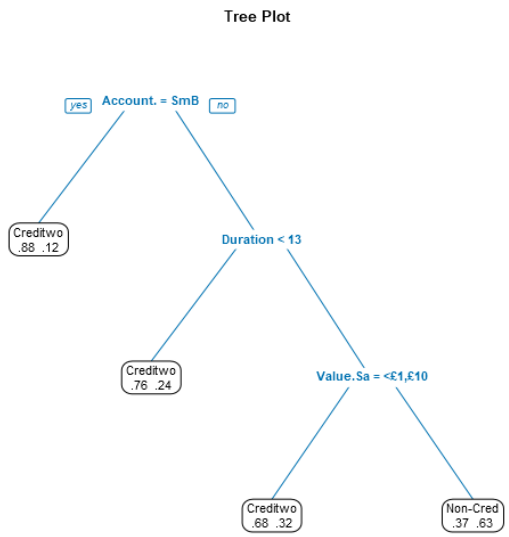
*Answer these questions for **each model** you created:*

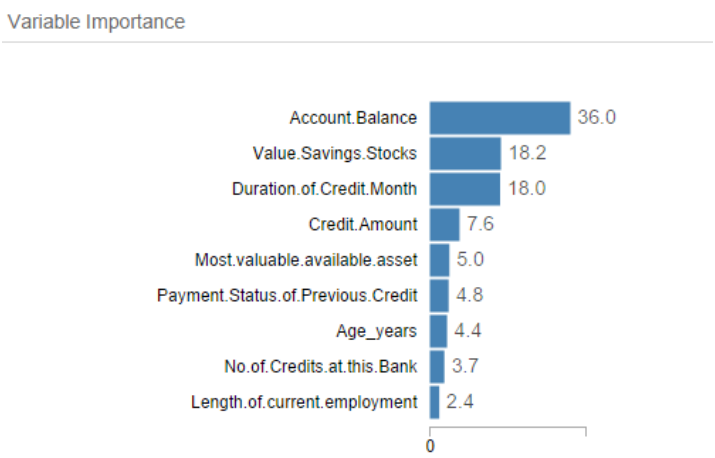
1. Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.
   **Answer**:

   **1.1) Logistic Regression Model**
   The most important variables for this model are:
   - "*Account-Balance*": "Some Balance" category is a significant predictor with P-value equal to 6.4*e-7. So, the P-value is roughly zero. It means this field is a very significant variable for this model.
   - "*Payment-Status-of-Previous-Credit*": "Some Problem" category is a significant predictor with P-value equal to 0.01. This value is less than 0.05 and so it's a significant value but it is a weaker predictor rather than the previous one.
   - *"Credit-Amount"* is a significant predictor with P-value equal to 0.02. This value is less than 0.05 and it is significant if we consider alpha as 0.05.
   - "*Length-of-current-employment*": "< 1yr" category is a significant predictor with P-value equal to 0.02. This value is less than 0.05 and so it's a significant value.
   - "*Installment-per-cent*" is another significant predictor with P-value equal to 0.01. This value is less than 0.05 and so it's a significant value.

   **1.2) Decision Tree Model**
   The most important variables for this model are:
   - "*Account-Balance*" is a significant predictor.
   - *"Value-Savings-Stocks"* is a significant predictor.
   - *"Duration-of-Credit-Month"* is a significant predictor.

   *Here is the decision tree plot and it shows how these predictors work and their threshold point.*
   *The variable importance is hereunder:*

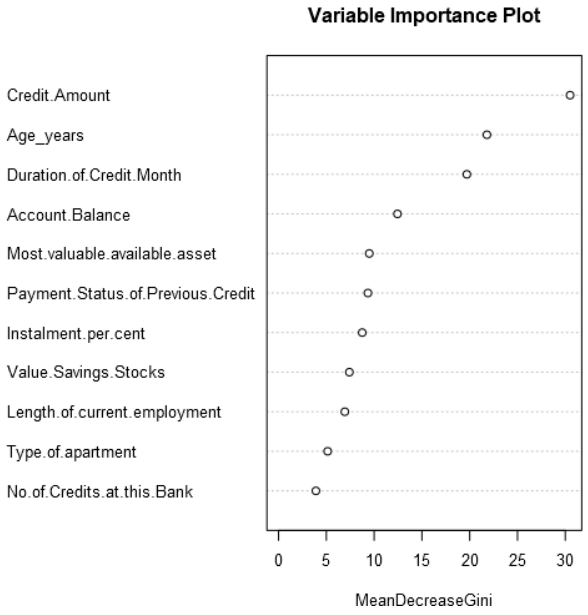


### 1.3) Forest Model

The most important variables for this model are:

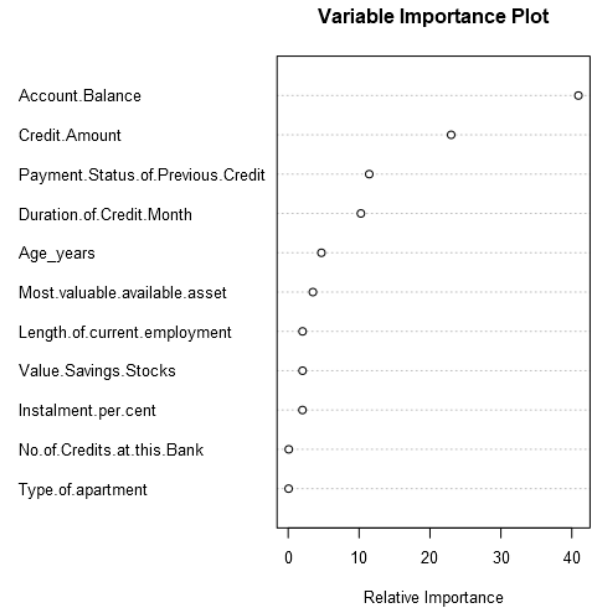
- *"Credit-Amount"*
- *" Age-years"*
- *"Duration-of-Credit-Month"*
- *"Account-Balance"*




### 1.4) Boosted Model

The most important variables for this model are:

- *"Account-Balance"*
- *"Credit-Amount"*
- *"Payment-Status-of-Previous-Credit"*
- *Duration-of-Credit-Month"*




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**: Here is the overall percent accuracy for all 4 models:

**Fit and error measures**

| Model | Accuracy | F1 | AUC | Accuracy_Creditworthy | Accuracy_Non-Creditworthy |
|---|---|---|---|---|---|
| Logistic_Reg_CreditWorth | 0.7800 | 0.8520 | 0.7314 | 0.8051 | 0.6875 |
| Dicision_Tree_Creditworth | 0.7467 | 0.8273 | 0.7054 | 0.7913 | 0.6000 |
| Forest_Model_Creditworth | 0.8133 | 0.8793 | 0.7403 | 0.8031 | 0.8696 |
| Boosted_Model_Creditworth | 0.7867 | 0.8632 | 0.7524 | 0.7829 | 0.8095 |

The confusion matrix for all 4 models are:

**Confusion matrix of Boosted_Model_Creditworth**

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

**Confusion matrix of Dicision_Tree_Creditworth**

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

**Confusion matrix of Forest_Model_Creditworth**

| | Actual_Creditworthy | Actual_Non-Creditworthy |
|---|---|---|
| Predicted_Creditworthy | 102 | 25 |
| Predicted_Non-Creditworthy | 3 | 20 |

**Confusion matrix of Logistic_Reg_CreditWorth**

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

I all models the number of “False Positives” are greater than the number of “False Negatives”. It means the number of Non-Creditworthy applicants that classified as Creditworthy is greater than number of Creditworthy applicants that classified as Non-Creditworthy. It is kind of a bias for all of our four models. The reason for this bias is that number of Creditworthy labels in the training dataset is 358 on the other hand the number of Non-Creditworthy labels is 142 instances. It creates a bias for the model to predict similar cases with those 358 instances as Creditworthy and there is a tendency to create False Positive for all models. This is a general observation.

The detail for each model is hereunder:

1. Overall accuracy of Logistic Regression model is 78%. This model is accurate in 80.51% of cases for predicting Creditworthy applicants, however it has a low accuracy for predicting Non-Creditworthy individuals: 68.75%. This model has a bias for predicting Non-Creditworthy more than best models.

2. Overall accuracy of Decision Tree model is 74.67%. This model is accurate in 79.13% of cases for predicting Creditworthy applicants, however it has a low accuracy for predicting Non-Creditworthy individuals: 60%. This model has the biggest bias for predicting Non-Creditworthy clients. It will lose many creditworthy clients because of this bias.

3. Overall accuracy of Random Forest model is 81.33%. This accuracy is the biggest one between these models. This model is accurate in 80.31% of cases for predicting Creditworthy applicants, and it has a very good accuracy for predicting Non-Creditworthy individuals: 86.96%. Overall this is the best available model in this project.

4. Overall accuracy of Boosted model is 78.67%. This model is accurate in 78.29% of cases for predicting Creditworthy applicants, and it has a good accuracy for predicting Non-Creditworthy

individuals: 80.95%. This model has the biggest bias for predicting Creditworthy clients. It will classify the most number of Non-Creditworthy clients as Creditworthy because of this bias and it potentially means losing money for lending to wrong individuals.

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

## Step 4: Writeup

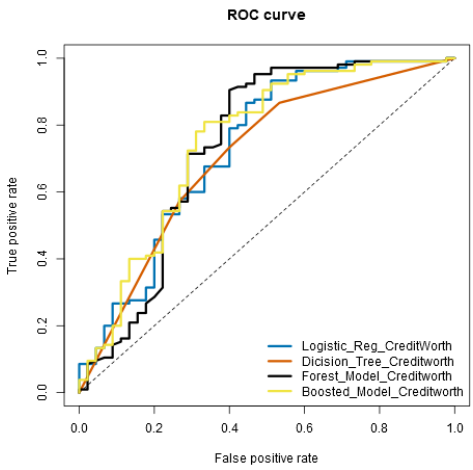
*Decide on the best model and score your new customers. For reviewing consistency, if Score_Creditworthy is greater than Score_NonCreditworthy, the person should be labeled as "Creditworthy"*

*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:*

1. Which model did you choose to use? Please justify your decision using only the following techniques:
    a. Overall Accuracy against your Validation set
    b. Accuracies within "Creditworthy" and "Non-Creditworthy" segments
    c. ROC graph
    d. Bias in the Confusion Matrices

    **Answer**: I will choose the "**Random Forest**" model between these 4 models, based on these arguments:
    a. It has the maximum overall accuracy 81.33%.
    b. It has the largest accuracy for Non-Creditworthy label prediction: 86.96% and the second largest accuracy for Creditworthy label: 80.31% and it has the biggest F1 score between these models.
    c. It has the second largest AUC for ROC graph between all ROC graphs: 0.7403



d. It has the least bias for “False Negative”. There are 3 Actual Creditworthy cases that classified as Non-Creditworthy ones. It means we will find the most individuals that worth to give loan to them with this model.

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

2. How many individuals are creditworthy?
   **Answer**: Base on the “Random Forest” model, there are **409** individuals that Creditworthy between all new applicants.