Project: Creditworthiness

# 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?

Due to a financial scandal that hit a competitive bank, our bank suddenly has an influx of new people applying for loans instead of the competitive bank. Our bank typically receives about 200 loan application every week and the loans are approved by hand before the crisis that hit the competitor bank. Now our bank received about 500 loan applications this week, in a bid to quickly and accurately approve or deny loan applications, I have been asked to figure out how to process all the loan application within one week.

As a data analyst, I have decided to make a prediction of creditworthy customers so that they can be approved for a loan, using different prediction methods and choosing the best method ensures that a list of Creditworthy loan applicants are compiled and processed within a short time and with high accuracy.

* What data is needed to inform those decisions?

In order to accurately predict customers who are credit worthy, I need data on past loan applicants who have been approved and denied. The content of the data include:

* Credit application result: This is the outcome of previous loan application which is either creditworthy or not credit worthy.
* Account balance: This indicates whether the loan applicant has an account with the bank or not. This will indicate whether the bank has sufficient personal details about the loan applicant.
* Duration of credit month: This specifies the length of the loan. This indicates whether the repayment date for the loan is feasible with respect to the loan amount applied for.
* Payment status of previous loan: This specifies whether the loan applicant paid up the previous loan or defaulted. This tells bout the integrity of the loan applicant.
* Purpose of the loan: This specifies the reason for the loan application. This specifies whether the loan is for an investment, an asset or a liability.
* Loan value (amount): This specifies the loan amount the loan applicant requested.
* Value of savings: This specifies whether the loan applicant has current savings or not and the value of the savings if the applicant has.
* Length of current employment: This specifies the length of the loan applicant’s current employment. This indicates to an extent the income stability of the loan applicant.
* Duration in current address: This specifies how long the loan applicant has lived in the current address.
* Most valuable asset: This specifies the category of the loan applicant’s most valuable asset.
* Age: This specifies the age in years of the loan applicant.
* Type of apartment: This specifies the category of apartment of the loan applicant.
* Number of credits at the bank: This specifies the number of loan applications the applicant made at the bank.
* What kind of model (Continuous, Binary, Non-Binary, Time-Series) do we need to use to help make these decisions?

Model type:

The kind of model needed for this prediction is Binary model because the loan application has only two possible outcomes which are Creditworthy or Not 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.

In the dataset of previous processed loan application provided, there are two rows that contained dirty data which I had to cleanup.

1. Duration in current address: This column has about 69% null values of the total data. I had to drop the column because it is not useful to the model with about 345 of 500 data in the column is missing and trying to use imputation method (manufacturing data from the available data in the column) will add bias to the model.
2. Age in years: This column has about 2% null values of the total data. I replaced the null values with the median of the data in the column (imputation method). I could have deleted the entire field with null values in the dataset because 2% of the entire data is negligible and the data in the column will still be useful but I used imputation method because the imputation method did not affect the data distribution in the column and also it gave me more data with which I can train my model.

# 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.*

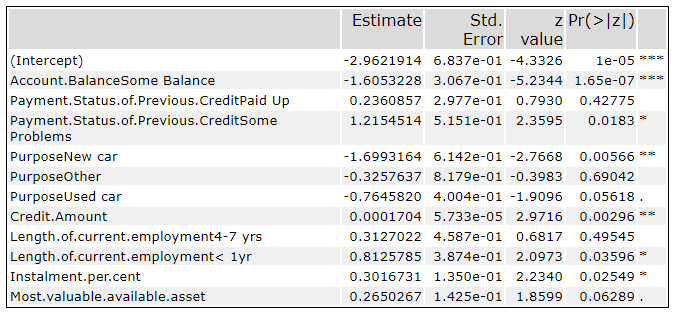
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

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

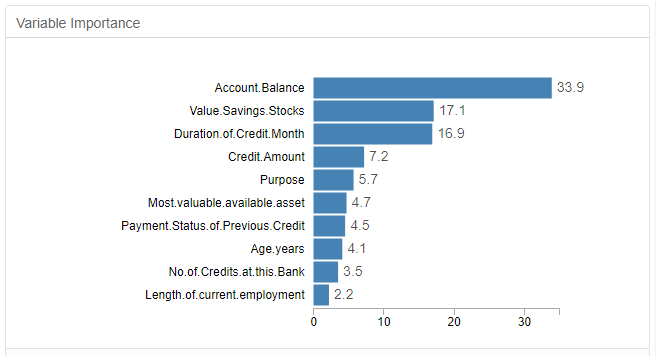
* Which predictor variables are significant or the most important? Please show the p-values or variable importance charts for all of your predictor variables.

For the models created, there were a few variables that showed significance to the target variable (Credit-application-result), for each of the four model, the variables are:

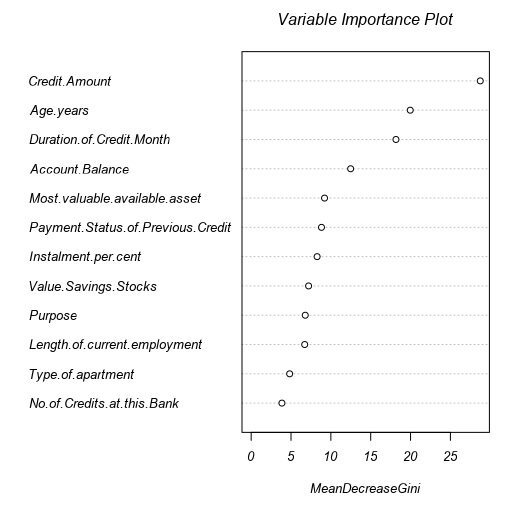
1. Logistic regression model: Account-balance, Payment-status-of-previous-credit, Purpose, Credit-amount, Length-of-current-employment, installment-per-cent and most-valuable-available-asset.



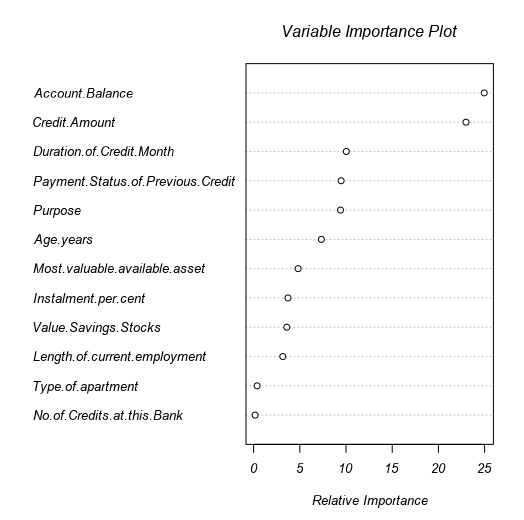
1. Decision tree model: Account-balance, Value-savings-stock, Duration-of-credit-month, Credit-amount, Purpose, Most-valuable-available-asset, Age-years, No-of-credit-at-this-bank and Length-of-current-employment.



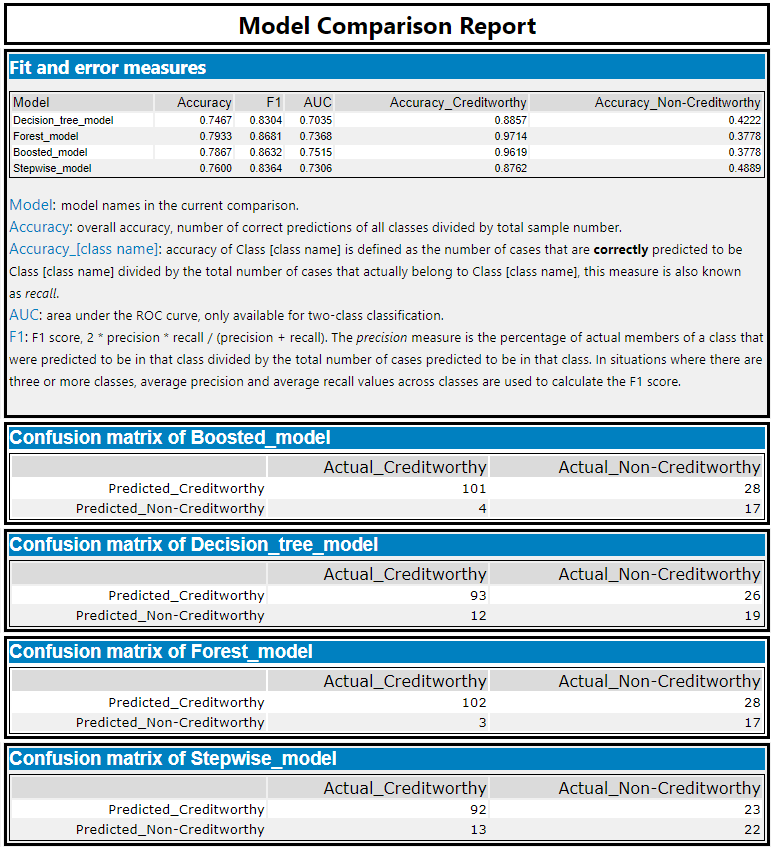
1. Forest model: Credit-amount, Age-years, Duration-of-credit-month, Account-balance, Most-valuable-available-asset, Payment-status-of-previous-credit, Installment-per-cent, Value-savings-stocks, Purpose, Length-of-current-employment, Type-of-apartment and No-of-credits-at-this-bank.



1. Boosted model: Account-balance, Credit-amount, Duration-of-credit-month, Payment-status-of-previous-credit, Purpose, Age-years, Most-valuable-available-asset, Instalment-per-cent, Value-savings-stocks, Length-of-current-employment, Type-of-apartment and No-of-Credit-at-this-bank.



* 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)*

# 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

Chosen model:

Having trained the dataset with four models and compared the result of each against the validation set of my dataset, I understood that the Forest model did the best job in predicting Creditworthy applicants, this is evident in the Fit and Error measures chart from the comparison tool which compared the accuracy of the four models.

Overall accuracy:

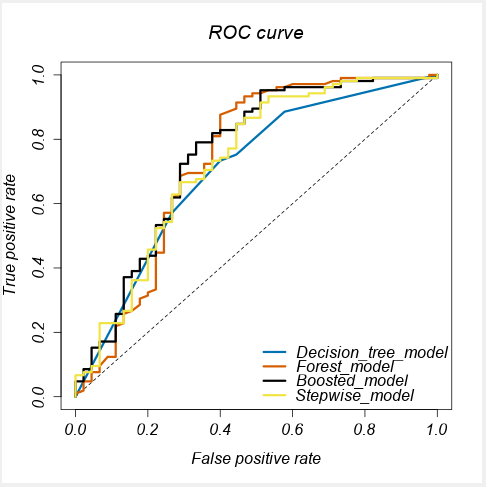
The Forest model showed the highest overall accuracy of 79.33% with the accuracy of predicting Creditworthy applicants at 97.14% while the accuracy of predicting Non credit worthy applicants is 37.78%.

Individual prediction accuracy:

The Forest mode has an accuracy error of 6.7% when working with the estimation set of the dataset.

ROC graph:

The ROC graph shows that of the four models, the Forest model did the best job predicting the Creditworthy applicants. The Forest model, of the four models is the model closest to the top left of the ROC graph.



Bias in Confusion matrix:

The Forest model has the least overall bias in the in classifying Creditworthy applicants in the confusion matrix when working with the validation set of the test dataset. The Forest model rightly classified 102 Creditworthy applicants which is the best classification among the four models and it makes sense to go with the model because we are only interested in loan applicants who are Creditworthy.

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

* How many individuals are creditworthy?

The number of creditworthy individuals that applied for loan is 408 from the 500 loan applicants, this indicates that 82% of the application will be approved.