Project: Creditworthiness

Ans Jayan

ansjayan@msn.com

# Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (250 word limit)

Identifying the creditworthiness of customers of a small bank. New customers will apply for loan. It is the job of the loan officer to identify the creditworthiness of these applicants. Using classification models, we can find the creditworthy customers and report to manager.

.

## Key Decisions:

Answer these questions

* What decisions needs to be made?

Is the loan applicant creditworthy or not.

* What data is needed to inform those decisions?

Data on all the past loan applicants and details of all new loan applicants.

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

We have to identify the creditworthiness. So, is the customer creditworthy or

Not, which is binary.

# 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 |

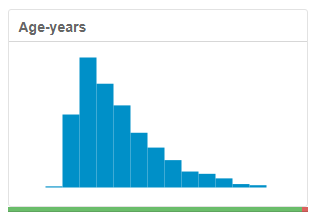
*To achieve consistent results reviewers expect.*

*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.

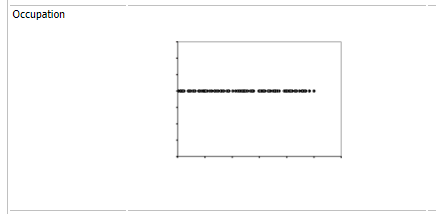
Fields Imputed :-

**Age-years** is imputed as it has 2% missing data. As it is right skewed median value is used for imputation as median value is less sensitive to skewness and outliers while mean will be pulled away from centre by extreme values.

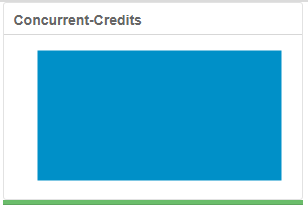


Fields Removed:-

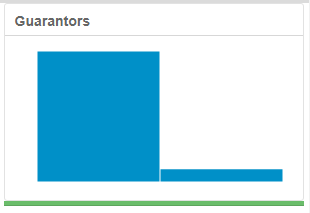
**Occupation** has only single value 1. Due to low variability, it is removed.



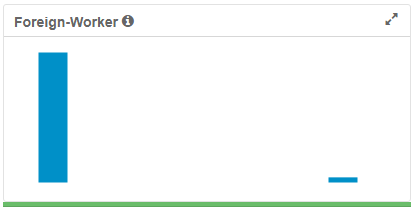
**Concurrent-Credits** has only one value Other Banks/Depts. Due to low variability, it is removed.



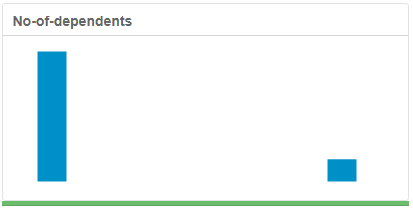
**Guarantors**: - Majority of values are None. Due to low variability, it is removed.



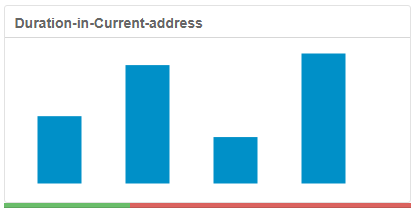
**Foreign-Worker**: - Majority of values are 1. Due to low variability, it is removed.



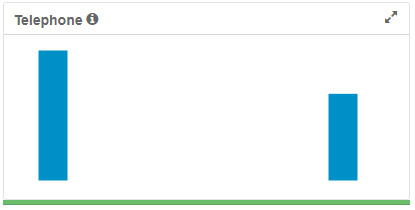
**No-of-dependents**: - Majority of values are 1. Due to low variability, it is removed.



**Duration-in-Current-address**: - has 69% missing values. So, it is removed.



**Telephone**: Owning a telephone is now common. This does not give any creditworthiness details to us.



# 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.*

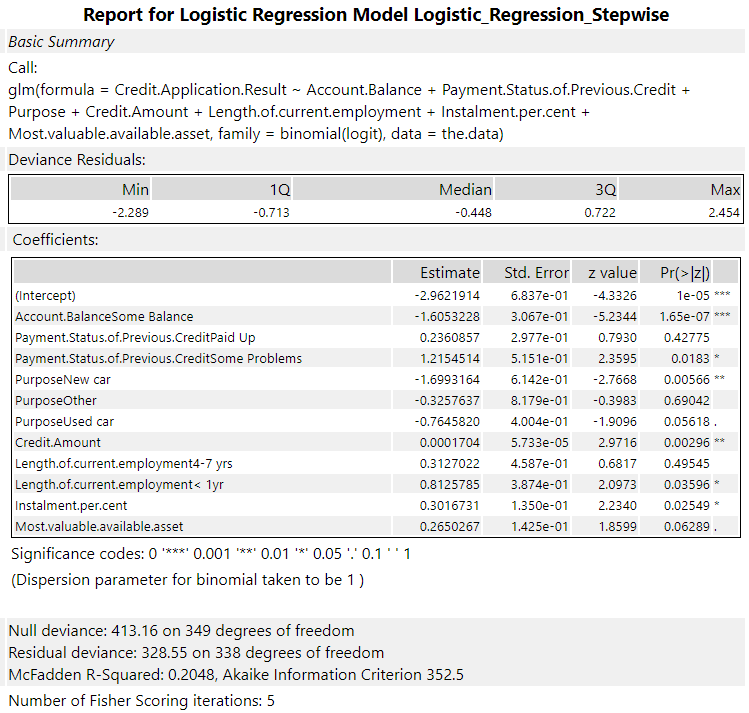
*Create all of the following models: Logistic Regression, Decision Tree, Forest Model, Boosted Model*

*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.

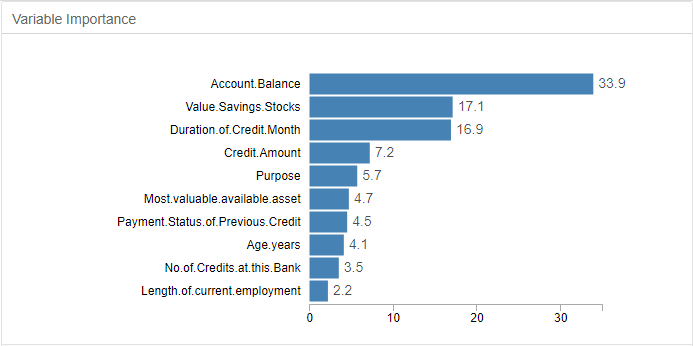
Logistic Regression

The significant variables are Account.BalanceSome Balance, Payment.Status.of.Previous.CreditSome Problems, PurposeNew car, Credit.Amount, Length.of.current.employment< 1yr, Instalment.per.cent



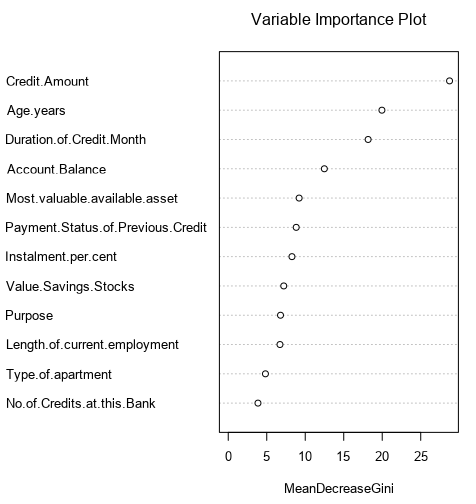
Decision Tree

The most important 3 variables are Account.Balance, Value.Savings.Stocks, Duration.of.Credit.Month.



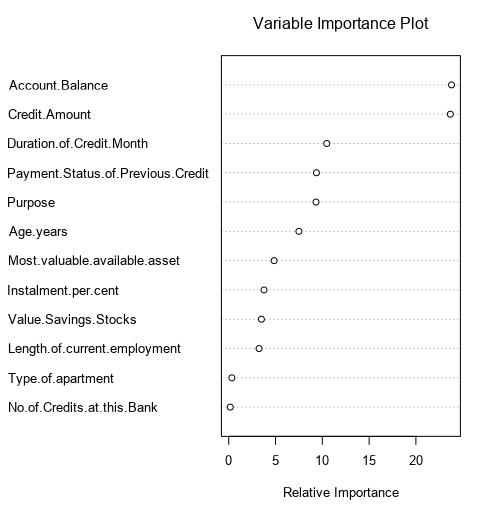
Forest Model

The most important 3 variables are Credit.Amount, Age.years, Duration.of.Credit.Month



Boosted Model

The most important 3 variables are Account.Balance, Credit.Amount, Duration.of.Credit.Month.

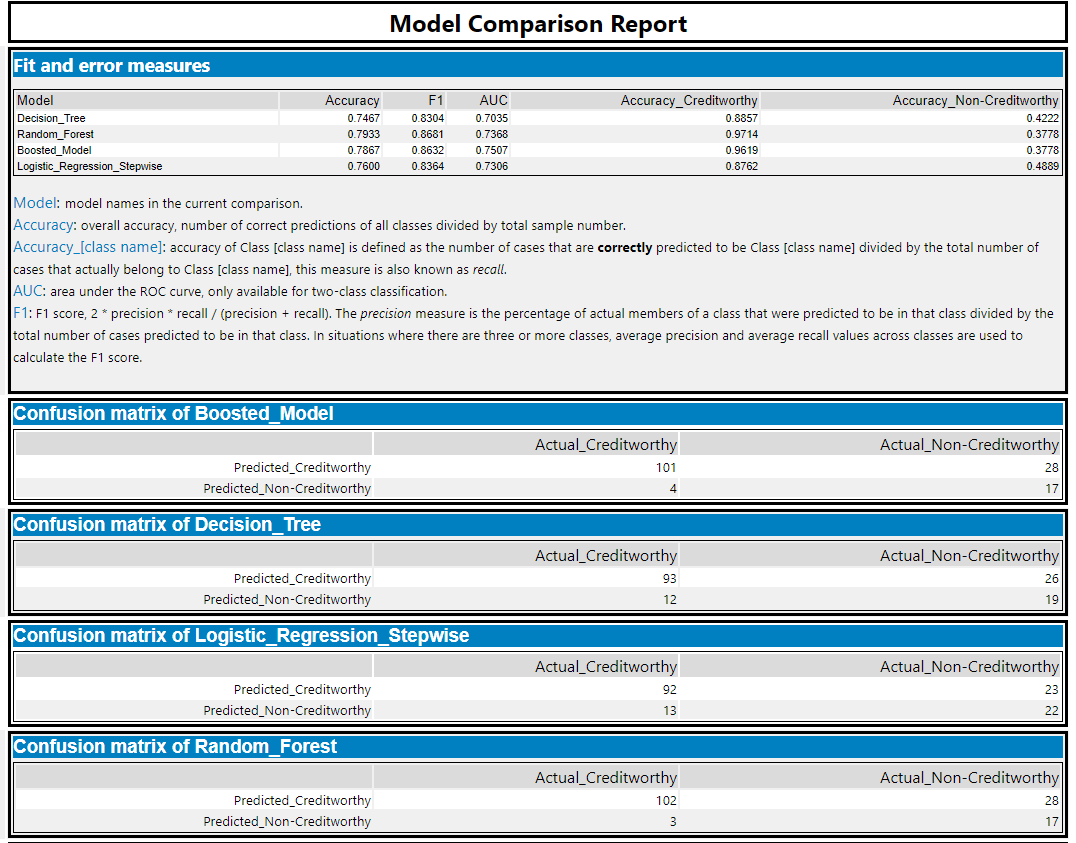


* Validate your model against the Validation set. What was the overall percent accuracy? Show the confusion matrix. Are there any bias seen in the model’s predictions?

From model comparison report the accuracy of logistic regression model is 76%, Decision Tree model is 74.67%, Random Forest model is 79.33%, Boosted model is 78.67%.

Count of Actual credit worthy is more than actual non-credit worthy, which means we have an imbalanced data. Accuracy measure alone is not that good for imbalanced data. The calculated PPV and NPV values are not same, so bias exists in the model’s prediction. For Random Forest model bias is medium, boosted model has less bias, logistic regression model and decision tree model has high bias.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic Regression**  TP = 92  TN = 22  FP = 23  FN = 13  Actual yes = 105  Actual no = 45  Predicted yes = 115  Predicted no = 35  Total = 300 | **Decision Tree**  TP = 93  TN = 19  FP = 26  FN = 12  Actual yes = 105  Actual no = 45  Predicted yes = 119  Predicted no = 31  Total = 300 | **Random Forest**  TP = 102  TN = 17  FP = 28  FN = 3  Actual yes = 105  Actual no = 45  Predicted yes = 130  Predicted no = 20  Total = 300 | **Boosted Model**  TP = 101  TN = 17  FP = 28  FN = 4  Actual yes = 105  Actual no = 45  Predicted yes = 129  Predicted no = 21  Total = 300 |
| **Accuracy**  (TP + TN)/Total | 0.38 | 0.3733 | 0.3966 | 0.3933 |
| **Misclassification Rate**  (FP + FN)/Total | 0.12 | 0.1266 | 0.1033 | 0.1066 |
| **True Positive Rate/Sensitivity/Recall**  TP/actual yes | 0.8762 | 0.8857 | 0.9714 | 0.9619 |
| **False Positive Rate**  FP/actual no | 0.5111 | 0.5777 | 0.6222 | 0.6222 |
| **True Negative Rate/Specificity**  TN/actual no | 0.4888 | 0.4222 | 0.3777 | 0.3777 |
| **Precision/PPV**  TP/predicted yes | 0.8 | 0.7815 | 0.7846 | 0.7829 |
| **Prevalence**  Actual yes/Total | 0.35 | 0.35 | 0.35 | 0.35 |
| **NPV**  TN/(TN + FN) | 0.6285 | 0.6129 | 0.85 | 0.8095 |



*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:

I choose to use the random forest model.

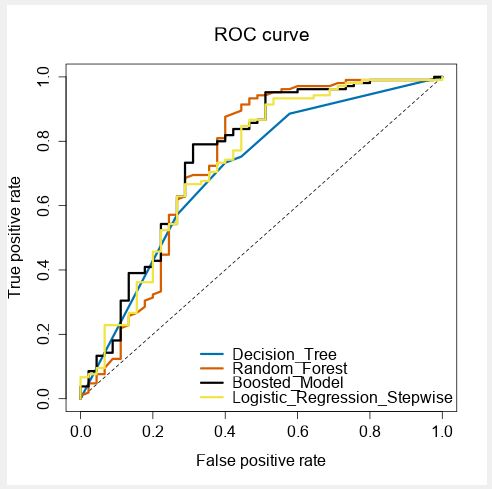
* + Overall Accuracy against your Validation set

Random Forest model has 79.33% overall accuracy.

* + Accuracies within “Creditworthy” and “Non-Creditworthy” segments

97.14% and 37.78%

* + ROC graph



The random forest model reaches the TPR fastest and reaches the top with AUC 0.7368.

* + Bias in the Confusion Matrices.

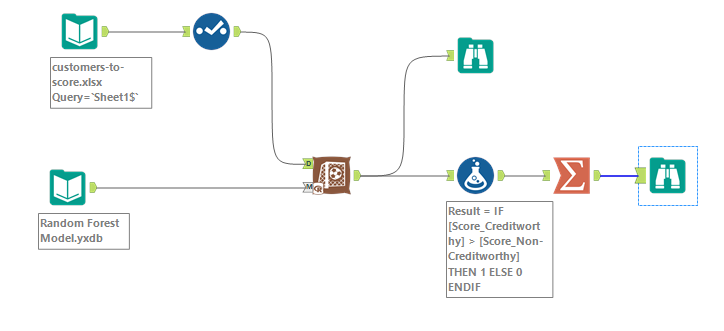
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Logistic Regression**  TP = 92  TN = 22  FP = 23  FN = 13  Actual yes = 105  Actual no = 45  Predicted yes = 115  Predicted no = 35  Total = 300 | **Decision Tree**  TP = 93  TN = 19  FP = 26  FN = 12  Actual yes = 105  Actual no = 45  Predicted yes = 119  Predicted no = 31  Total = 300 | **Random Forest**  TP = 102  TN = 17  FP = 28  FN = 3  Actual yes = 105  Actual no = 45  Predicted yes = 130  Predicted no = 20  Total = 300 | **Boosted Model**  TP = 101  TN = 17  FP = 28  FN = 4  Actual yes = 105  Actual no = 45  Predicted yes = 129  Predicted no = 21  Total = 300 |
| **Accuracy**  (TP + TN)/Total | 0.38 | 0.3733 | 0.3966 | 0.3933 |
| **Misclassification Rate**  (FP + FN)/Total | 0.12 | 0.1266 | 0.1033 | 0.1066 |
| **True Positive Rate/Sensitivity/Recall**  TP/actual yes | 0.8762 | 0.8857 | 0.9714 | 0.9619 |
| **False Positive Rate**  FP/actual no | 0.5111 | 0.5777 | 0.6222 | 0.6222 |
| **True Negative Rate/Specificity**  TN/actual no | 0.4888 | 0.4222 | 0.3777 | 0.3777 |
| **Precision/PPV**  TP/predicted yes | 0.8 | 0.7815 | 0.7846 | 0.7829 |
| **Prevalence**  Actual yes/Total | 0.35 | 0.35 | 0.35 | 0.35 |
| **NPV**  TN/(TN + FN) | 0.6285 | 0.6129 | 0.85 | 0.8095 |

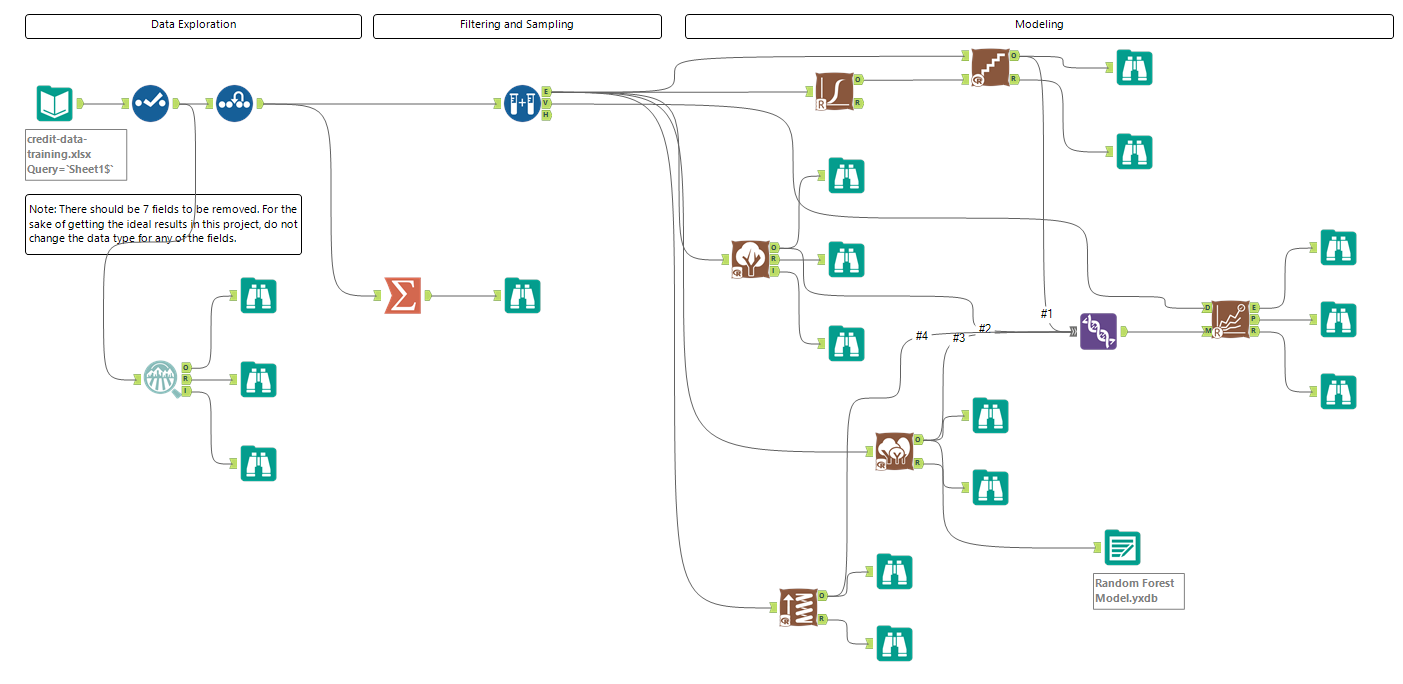
The random forest model’s PPV = 0.7846 and NPV = 0.85, there exist medium bias.

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

* How many individuals are creditworthy?

408 individuals are creditworthy.





**Before you Submit**

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/265/view) here. Reviewers will use this rubric to grade your project.