Predicting Default Risk

Classification Models

# *Table of Contents*

[***Table of Contents*** 1](#_Toc67989329)

[***Predicting default risk*** 2](#_Toc67989330)

[Business and Data Understanding 2](#_Toc67989331)

[Key Decisions: 2](#_Toc67989332)

[Building the Training Set 2](#_Toc67989333)

[Missing Data 3](#_Toc67989334)

[Clean-up 3](#_Toc67989335)

[Field Summary 4](#_Toc67989336)

[Train Classification Models 5](#_Toc67989337)

[Logistic Regression: Stepwise 5](#_Toc67989338)

[LR Model accuracy & Confusion Matrix 5](#_Toc67989339)

[Decision Tree 6](#_Toc67989340)

[DT Model Accuracy & Confusion Matrix 6](#_Toc67989341)

[Random Forest Model 7](#_Toc67989342)

[RF Model Accuracy & Confusion Matrix 7](#_Toc67989343)

[Boosted Model 8](#_Toc67989344)

[Boosted Model Accuracy & Confusion Matrix 8](#_Toc67989345)

[Writeup 9](#_Toc67989346)

[Model Reports 11](#_Toc67989347)

[Linear Regression, Stepwise report 11](#_Toc67989348)

[Decision Tree 12](#_Toc67989349)

[Random Forest 13](#_Toc67989350)

[Boosted Model 14](#_Toc67989351)

[Alteryx Workflows 15](#_Toc67989352)

[Resources 16](#_Toc67989353)

[Websites 16](#_Toc67989354)

[Udacity Knowledge 16](#_Toc67989355)

# *Predicting default risk*

## Business and Data Understanding

The analysis is to determine if customers are creditworthy for a new loan. The team typically gets 200 loan applications per week and approves them all by hand.

Due to sudden increase, there are nearly 500 loan applications to process this week due to a financial scandal that hit a competitive bank last week. Your manager sees new influx is a great opportunity to figure out to process all these loan applications in one week. I need to systematically evaluate the creditworthiness of these new loan applications.

### Key Decisions:

To be able to make an informed decision if customers are creditworthy, I need to have the following:

* Data on past loan applicants
* List of customers that need to process.

We have been provided two data sets:

* Credit-data-training contains approvals from the past loan applicants the bank has processed.
* Customers-to-score contains the new customers that I need to score on the classification model.

I will create a Binary classification model, to help make the decision if the customers are creditworthy or not.

## Building the Training Set

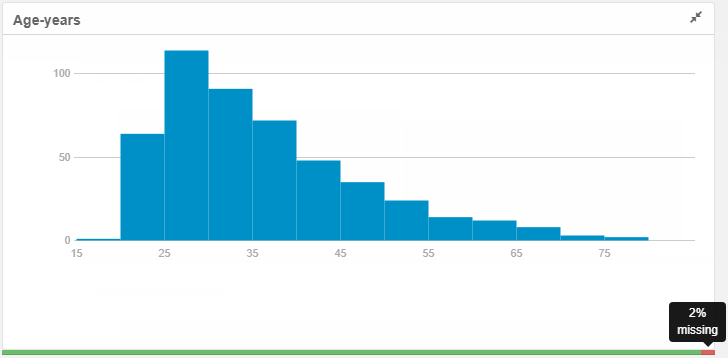
First, I need to explore the data to see if any columns that should be removed or imputed.

Using Alteryx to explore the data using the field comparison tool.

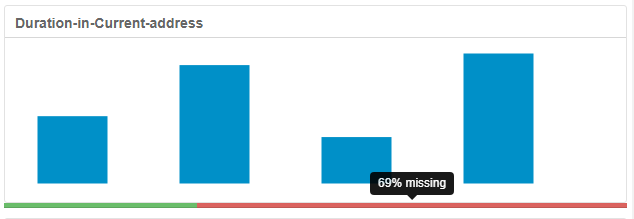
|  |
| --- |
|  |

### Missing Data

* Age-years 2.4% missing.
  + Impute the miss data using the median age of 33 from the entire field.



* Duration-in-Current-address 68.8% missing.
  + This field will be removed, as there is a lot of missing values.



### Clean-up

There are 3 fields that are heavily skewed towards toward one type of data, these should be removed.

* Guarantors: is heavily towards ‘None’
* Foreig-Worker: heavily towards ‘1’
* No-of-dependents: heavily towards ’1’

|  |  |  |
| --- | --- | --- |
|  |  |  |

2 fields have low variability, no variations and is entirely uniform, there is no variation and should be removed.

* Concurrent-Credits, has only 1 unique value.
* Occupation also contains only 1 unique value.

|  |  |
| --- | --- |
|  |  |

The last field that would have no significance for the model would be Telephone, this can also be removed.

### Field Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Field Category | Percent Missing | Unique Values |
| Age-years | Numeric | 2.4 | 54 |
| Credit-Amount | Numeric | 0 | 464 |
| Duration-in-Current-address | Numeric | 68.8 | 5 |
| Duration-of-Credit-Month | Numeric | 0 | 30 |
| Foreign-Worker | Numeric | 0 | 2 |
| Instalment-per-cent | Numeric | 0 | 4 |
| Most-valuable-available-asset | Numeric | 0 | 4 |
| No-of-dependents | Numeric | 0 | 2 |
| Occupation | Numeric | 0 | 1 |
| Telephone | Numeric | 0 | 2 |
| Type-of-apartment | Numeric | 0 | 3 |
| Account-Balance | String | 0 | 2 |
| Concurrent-Credits | String | 0 | 1 |
| Credit-Application-Result | String | 0 | 2 |
| Guarantors | String | 0 | 2 |
| Length-of-current-employment | String | 0 | 3 |
| No-of-Credits-at-this-Bank | String | 0 | 2 |
| Payment-Status-of-Previous-Credit | String | 0 | 3 |
| Purpose | String | 0 | 4 |
| Value-Savings-Stocks | String | 0 | 3 |

|  |  |
| --- | --- |
| Result after imputing Age-years.    Age-years validation 35.574 rounded up = 36 |  |

## Train Classification Models

Creating the estimation and validation samples where 70% of the data set will be used for estimation and the remaining 30% is reserved for validation.

4 Models will be created: Logistic Regression, Decision Tree, Forest Model, Boosted Model.

Stepwise will be used, to improve the efficiency as this tool will automate finding the best predictor variables.

Target Variable in all models is Credit Application Result

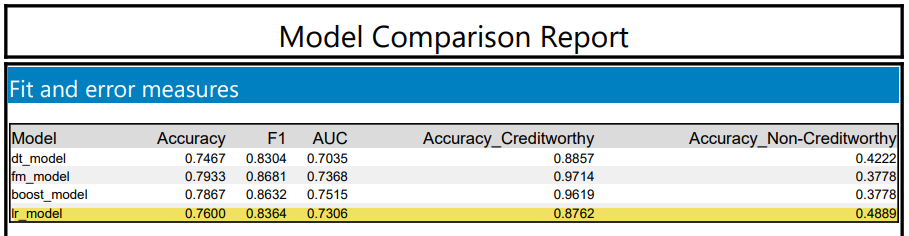
### Logistic Regression: Stepwise

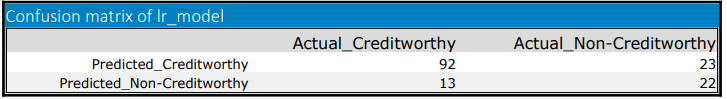
The top 3 predictor variables that statistically significant with p-values of less that = 0.05.

|  |  |
| --- | --- |
| Variable | p-value |
| Account balance | -2.41e-08 |
| Purpose | 0.00665 |
| Credit amount | 0.00167 |

For full report [Reference figure 1](#_Linear_Regression,_Stepwise)

#### LR Model accuracy & Confusion Matrix





* Credit accuracy = actual\_creditworthy / (predicted\_creditworthy)

= 92 / (92+23) = 0.8 or = 80%

* Non-Credit accuracy = actual\_Non-creditworthy / (predicted\_Non-creditworthy)

= 22 / (13+22) = 0.6286 or = 62.86%

The linear regression model shows bias in predicting customers as non-creditworthy

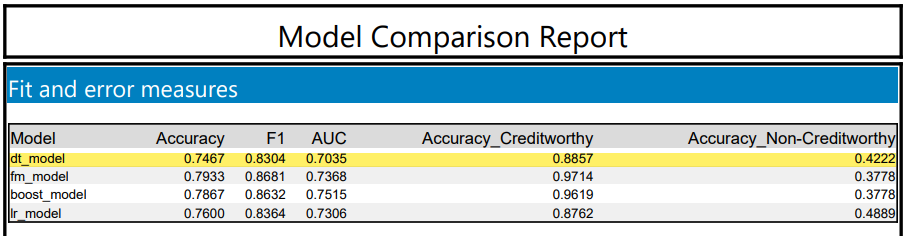
### Decision Tree

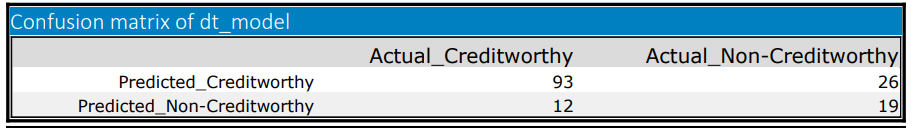
The top 3 predictor variables that statistically significant are:

|  |  |
| --- | --- |
|  | * Account Balance * Vale Savings Stocks * Duration of Credit Month |

Full report details [Reference figure 2](#_Decision_Tree)

#### DT Model Accuracy & Confusion Matrix





* Credit accuracy = actual\_creditworthy / (predicted\_creditworthy)

= 93 / (93+26) = 0.7815 or = 78.15%

* Non-Credit accuracy = actual\_Non-creditworthy / (predicted\_Non-creditworthy)

= 19 / (12+19) = 0.6129 or = 61.29%

Similar to the linear regression model, the Decision Tree shows bias in predicting customers as non-creditworthy.

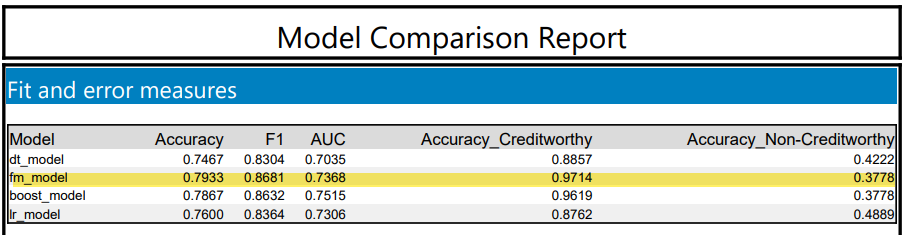
### Random Forest Model

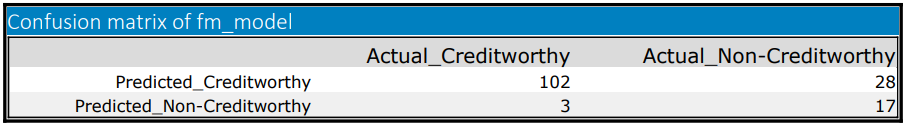
The top 3 predictor variables that statistically significant are:

|  |  |
| --- | --- |
|  | * Credit Amount * Age-Years * Duration of Credit Month |

Full report details [Reference figure 3](#_Random_Forest)

#### RF Model Accuracy & Confusion Matrix





* Credit accuracy = actual\_creditworthy / (predicted\_creditworthy)

= 102 / (102+28) = 0.7846 or = 78.46%

* Non-Credit accuracy = actual\_Non-creditworthy / (predicted\_Non-creditworthy)

= 17 / (3+17) = 0.85 or = 85.00%

The Random forest does not seem to build biased as the accuracies are quite close to each other.

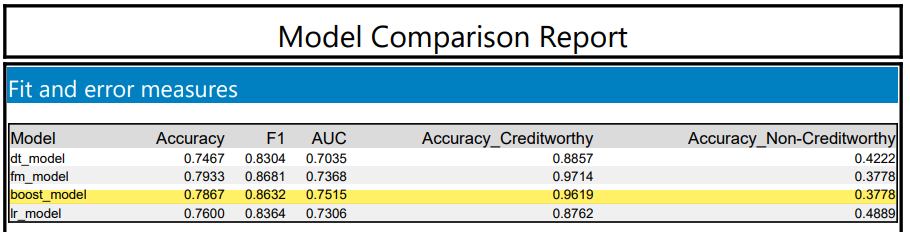
### Boosted Model

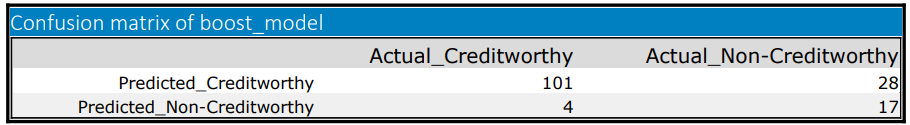
The top 3 predictor variables that statistically significant are:

|  |  |
| --- | --- |
|  | * Account Balance * Credit Ammount * Duration of Credit Month |

Full report details [Reference figure 4](#_Boosted_Model)

#### Boosted Model Accuracy & Confusion Matrix





* Credit accuracy = actual\_creditworthy / (predicted\_creditworthy)

= 102 / (102+28) = 0.7829 or = 78.29%

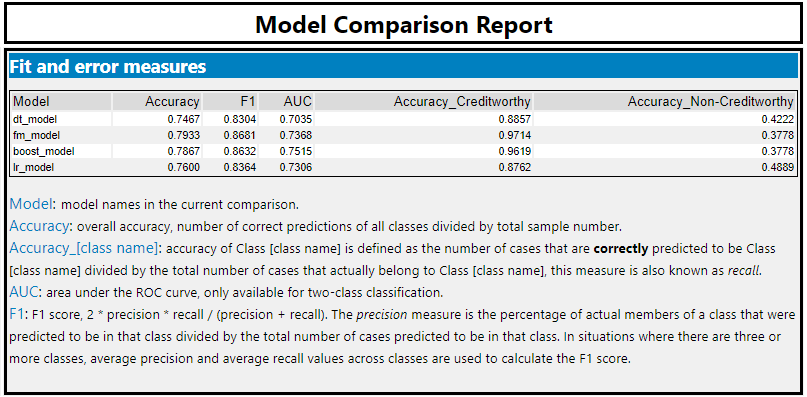
* Non-Credit accuracy = actual\_Non-creditworthy / (predicted\_Non-creditworthy)

= 17 / (3+17) = 0.8095 or = 80.95%

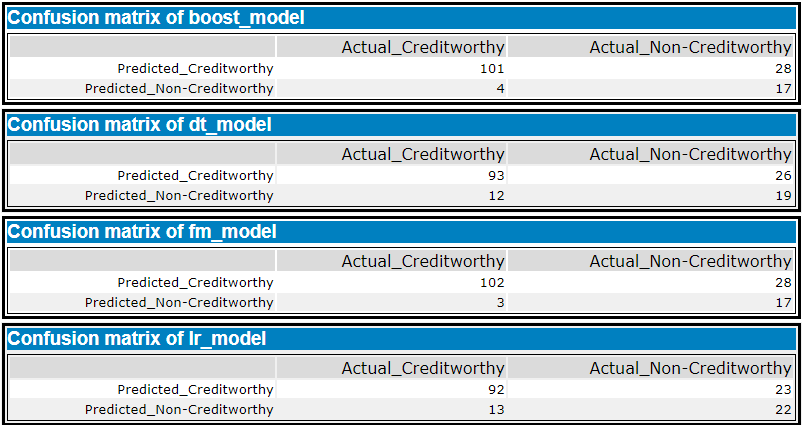
The Boosted Model does not seem to be biased as the accuracies are quite close to each other.

## Writeup

The Forest model has the highest accuracy of 79.33% from all the four models and has the highest F1 score of 86.81%. However, the AUC is lower than the next highest, the Boosted Model.

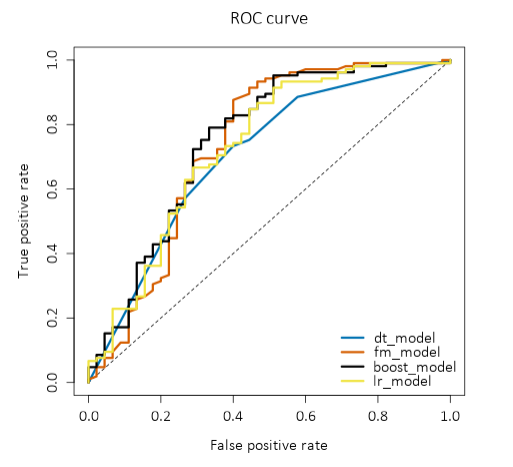


The Linear Regression (Stepwise) and Decision Trees seem to create bias towards classifying customers as Non-Creditworthy whereas the Forest and Boosted Models seems to be comparable in classify credit and Non-Creditworthy customers.



Model Accuracies,

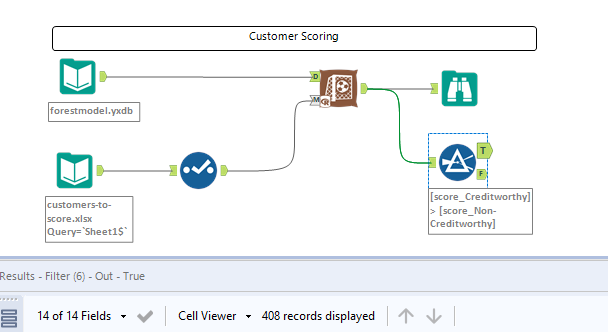
|  |  |  |
| --- | --- | --- |
| Model | Credit Worthy | Non-Credit Worthy |
| Logistic Regression | 80.00% | 62.86% |
| Decision Tree | 78.15% | 61.29% |
| Random Forest | 78.46% | 85.00% |
| Boosted | 78.29% | 80.95% |



Looking at the ROC curve, the Decision Tree performed the worst, whereas the Boosted and Forest model performed the best, these two models also reaches the true positive rate the fasted with the Forest model slightly leading.

The Forest model will be chosen to predict the classification for credit and Non-Credit worthy of the new 500 loan applicants.

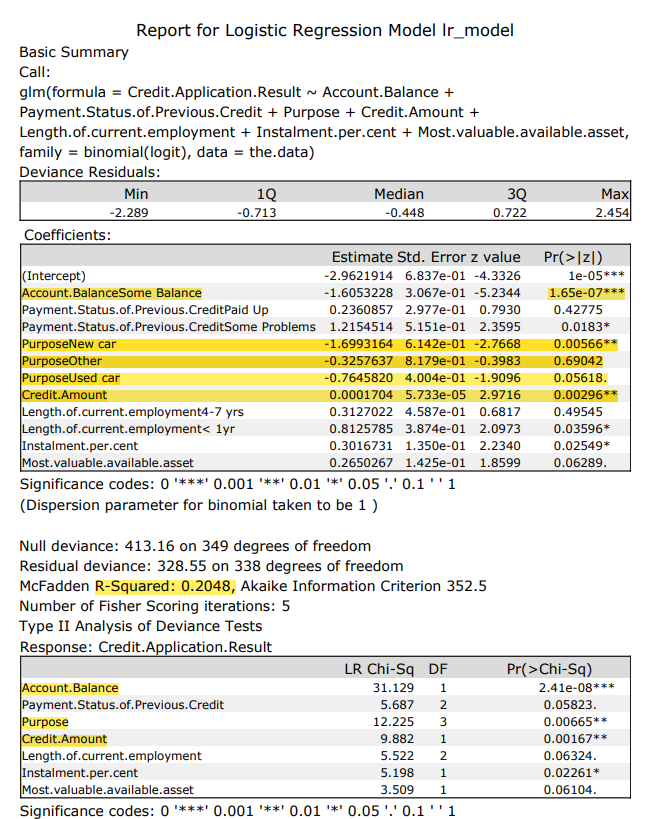
From the new loan applications 408 (98%) are credit worthy and should be approved for a new loan, 92 have been classified as Non-Credit worthy.



## Model Reports

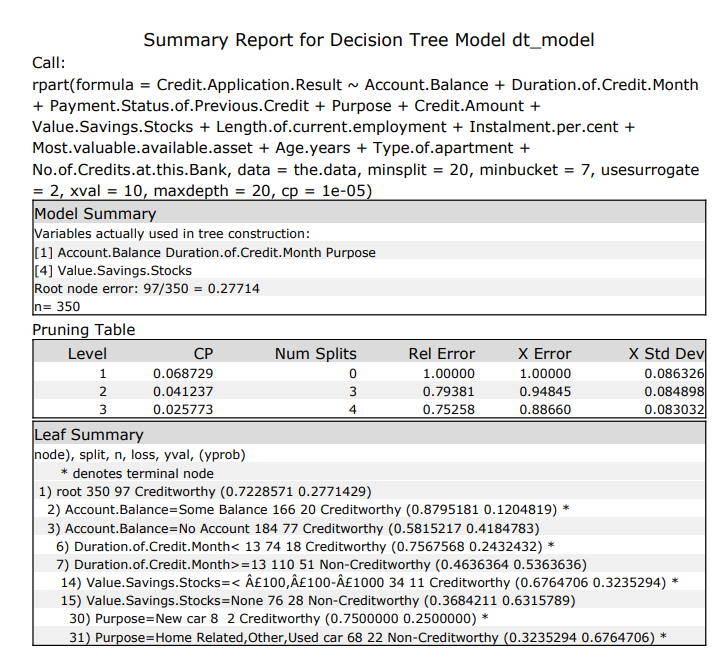
### Linear Regression, Stepwise report

*Figure 1*



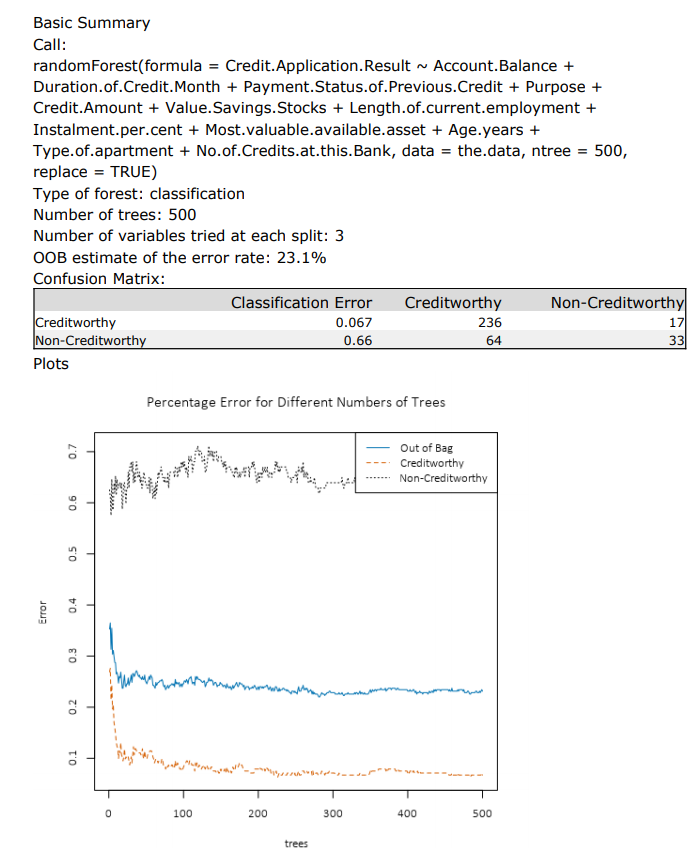
### Decision Tree

*Figure 2*



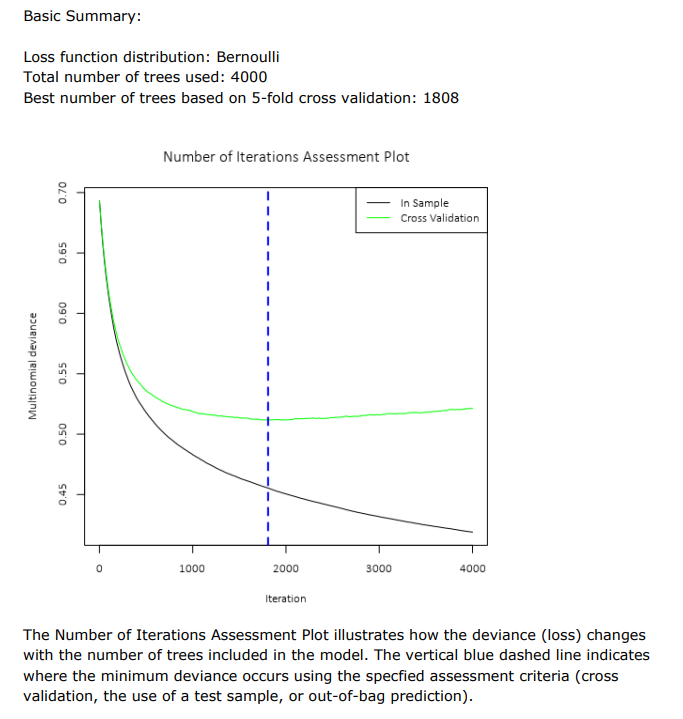
### Random Forest

*Figure 3*

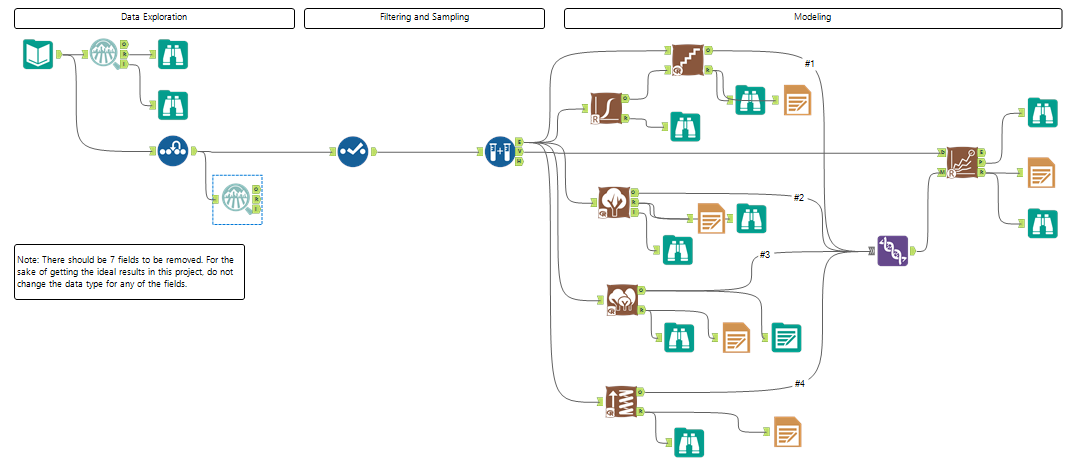


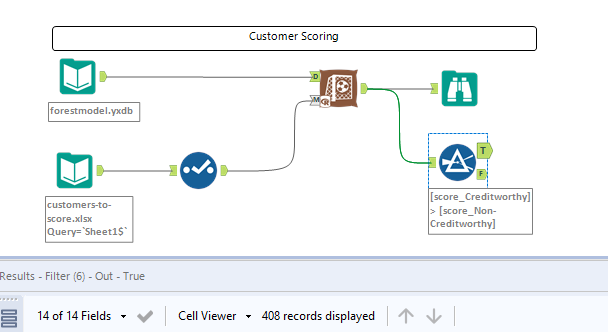
### Boosted Model

Figure 4



## Alteryx Workflows





## Resources

Here are some guidelines to help you clean up the data:

* Are any of your numerical data fields highly correlated 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: If you decide to impute any data field, for the sake of consistency in the data clean-up process, impute the data using the median of the entire data field.

#### Websites

<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

#### Udacity Knowledge

<https://knowledge.udacity.com/questions/483519>

<https://knowledge.udacity.com/questions/462675>

<https://knowledge.udacity.com/questions/473072>