Capstone Final Report

Merrimack College

Data Science Capstone

Amol Gote

18th Dec 2020

# Executive Summary

The main objective of this report is to provide the findings related to building a machine learning model which will **predict the probability of** **shareholder class action litigation** and its **severity** for a company. From a business perspective a each of the company buys insurance to protects themselves from these class actions lawsuits, but from the coverage perspective these companies are not sure how much it should be. This model will help in predicting if the company will be sued and if sued what would be tentative settlement amount, based on which CIO of the company can make the decision about the appropriate insurance coverage.

The target company in this case is SJM (The J.M. Smucker Company), it is a food and beverage company, has a market capital close to $10 billion. The data used for the analysis comprised of the companies belonging to the sector associated with SJM sector. Analysis data comprised of annual filings done by the companies annually along with the restatements. Along with annual filings it also comprised of daily stocks trading and securities data for each of those companies. Besides that, it had standard and poor ratings data for ach of those companies.

The Business objective, in this case, was to identify if the company will be subject to shareholder class action litigation, and the resulting variables output must be “Yes” or “No”. There were 2 primarily two techniques that were used to build the model, so from the two techniques, best model with accuracy of close **84%** was chosen as recommended model.

Key financial indicators related to capital, assets, earnings, stock volatility (daily opening and closing) play a significant role in predicting the target variable. Below is the list of top 15 important variables which contribute significantly to predicting whether particular company will be subject class action litigation.

|  |  |
| --- | --- |
| Sr No | Description |
| 1 | Capital Surplus/Share Premium Reserve |
| 2 | Trading Volume - Daily |
| 3 | Price - Close - Daily |
| 4 | Current Assets - Other - Total |
| 5 | Receivables - Total |
| 6 | Price - Open - Daily |
| 7 | Working Capital (Balance Sheet) |
| 8 | Stockholders Equity - Total |
| 9 | Book Value Per Share |
| 10 | Common/Ordinary Stock (Capital) |
| 11 | Cash and Cash Equivalents - Increase/(Decrease) |
| 12 | Earnings Before Interest and Taxes |
| 13 | Price to earnings ratio (PE ratio) |
| 14 | Revenue - Total |
| 15 | Accumulated Other Comprehensive Income (Loss) |

# Data and Approach

The data available for analysis contains 17,416 observation, these include annual statements issued by publicly traded companies in North America as well as restatements. Since these companies belonging to various sectors and each sector litigation parameters could vary, hence filtered the data to contain only observation from assigned company sector. In summary data had 1768 variables for annual filings, stocks data had 76 variables, securities had 55 variables and ratings data had 6 variables. Assigned company is SJM (J. M. Smucker Company, also known as Smucker and Smucker's), the sector it belongs to is 30 (Consumer Staples). Filtering by the sector reduces the number of observations to 2323.

## Data preparation

* Filtered dataset with annual statement only and then added variables related to restatement to each of the annual statements, the restatement variables include whether there was restatement and by what magnitude the restatement was for.
* Following attributes in below mentioned were restated as mentioned in below table. The way variables were identified was first dataset with restatements only was created and then applied the logic for removing NA columns with threshold of 10-20%

|  |  |
| --- | --- |
| Variable | Description |
| at | Assets - Total |
| capx | Capital Expenditures |
| cogs | Cost of Goods Sold |
| dltt | Long-Term Debt - Total |
| epsfi | Earnings Per Share (Diluted) - Including Extraordinary Items |
| epspi | Earnings Per Share (Basic) - Including Extraordinary Items |
| ib | Income Before Extraordinary Items |
| ni | Net Income (Loss) |
| nopi | Nonoperating Income (Expense) |
| pi | Pretax Income |
| reuna | Retained Earnings - Unadjusted |
| seq | Stockholders Equity - Parent |
| teq | Stockholders Equity - Total |
| txt | Income Taxes - Total |
| wcap | Working Capital (Balance Sheet) |
| xint | Interest and Related Expense - Total |

* Post identifying the restatement variables and associated magnitude merged these variables with annual statement dataset.
* Now the analysis dataset contains annual statement variables as well as the restatement ones.
* Analysis dataset still contains multiple rows for single company. So, the data set was aggregated based on gv-key and ticket symbol for the company. Aggregation was mean/average based.
* Post merging restatement variables were again analyzed, due to aggregation and if the aggerated value was greater than 0.5 then they were marked as restated or else not.
* Stocks dataset then was merged with main dataset, from this dataset only certain analysis specific variables were picked which include as shown in below table

|  |  |
| --- | --- |
| Variable | Description |
| prccd | Price - Close - Daily |
| prchd | Price - High - Daily |
| prcld | Price - Low - Daily |
| prcod | Price - Open - Daily |
| trfd | Daily Total Return Factor |
| cshtrd | Trading Volume - Daily |

* These stocks variables were aggregated and then merged with main analysis dataset.
* Securities dataset was then merged with main dataset, from this only 2 variables were chosen *Monthly Total Return Factor* and *Dividend Rate – Monthly*.
* Post securities dataset, ratings dataset weas merged with main analysis dataset. Prior merging the dataset additional variable with numeric values was created in rating dataset, these numeric values are in decreasing order e.g. AAA = 100, BBB = 60 and finally not rated was given 0 value. While merging the rating dataset with main dataset additional variable was added indicating whether the rating has increased, decreased, or not changed, if none of these 3 then it was indicated as not rated.
* Finally target variables related dataset of SCA filings was loaded, in this case added the target variable litigated to the dataset under analysis. For each of the company record identify if has any SCA filing from the SCA filling dataset, if entry exists then mark for that company litigated = true or else if the record does not exist in SCA filing for that company then mark litigated attribute as false. Also, in case of litigation identify if there is any settlement amount and add same to main dataset. In case if multiple settlement amount exists for company then take the maximum settlement amount.
* On the final dataset following key financial ratios for analyzing company stock performance were added these include as mentioned in below table.

|  |  |
| --- | --- |
| Variable | Description |
| de\_ration | Debt to equity ratio |
| wc\_ratio | Working capital ratio |
| pe\_ratio | Pricing to earnings ratio |
| roe\_ratio | Return on Equity |

## Feature engineering

### NA value variables

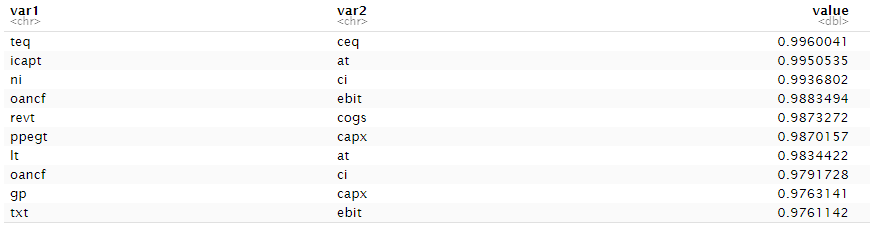
There are close to 1700+ variables in the consolidated dataset, so first step was to remove all columns with considerable amount of NA values. So, in this remove all the columns which contain 25% or greater NA or null values. This reduced the number of columns to close to 300

### Manual analysis for reducing number of predictive variables

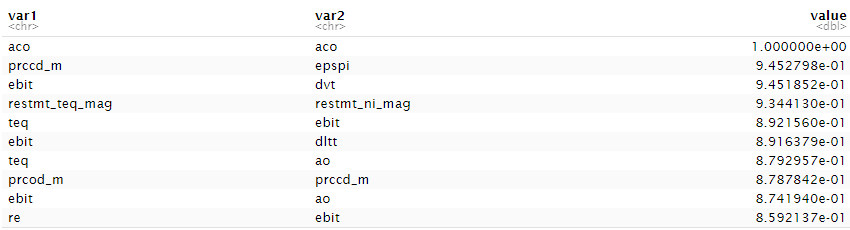
* Remove all the columns which are related to general information of the company like address, phone, URL, etc.
* There are lot of columns with hardly any variance, so identify columns which have values which are near to zero variance and remove them from dataset.
* Above steps reduced the number of columns from 300 to 220.
* Based on manual analysis of the columns and considering the target variable, removed the variables which logically are related to target variable, these reduce the variable to 147.

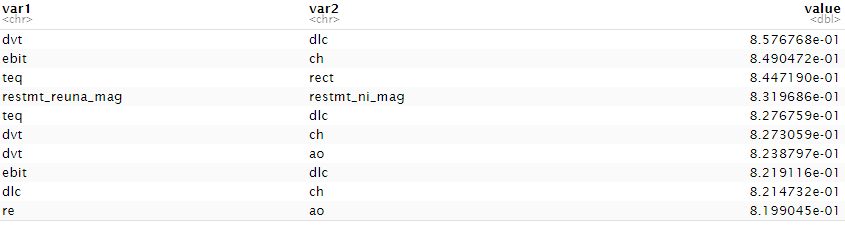
### Reducing Collinearity

To identify the variable correlation, built a correlation table like shown below. Based on the below table collinear variables pairs were analyzed and removed one variable from the pair to break the collinearity. Collinearity threshold that was considered was 0.9, any thing above attempt was made to break the collinearity pair.



Post collinearity removal the table looks something like this





Reducing the collinearity led to final 83 predictive variables. Some of the variable with high collinearity were ignored based on manual analysis and considering their business definitions.

### Encoding

Prior applying any encoding all the categorical variables are converted as factor. Then all the categorical variables were target encoded, there were very few categorical variables, these include standard and poor rating variable, rating increased/decreased variable and all restatement variables, these restatement variables indicate whether a particular variable was restated as part of restatement.

### Training and Test dataset set

Final dataset resulted in 333 observation, this was final dataset with all data merging, aggregation, encoding etc. This data set is then split to training and test dataset with ratio of 0.75/0.25, with 0.75 been training dataset and 0.25 been test dataset.

### Class imbalance

In this case target variable is whether company has been litigated or not, and if we look at the training dataset, the ratio of litigated or not (Yes/No) is 14%:86% which is clear sign of class imbalance. In case of class imbalance, the machine learning techniques tends to be more biased towards the majority class, causing bad classification of the minority class which is predictor in this case. So, in this case to reduce the imbalance R ROSE package was used and the strategy that was leveraged in this case of combination of both and up and down sampling. This effectively the ratio to 37%:63%

|  |  |  |  |
| --- | --- | --- | --- |
| **Litigated** | | | |
| Before | | After removing class imbalance | |
| No | Yes | No | Yes |
| 34 | 215 | 91 | 158 |
| 14% | 86% | 37% | 63% |

### Scaling variables

All the input variables were scaled to have equal range and/or variance. This ensures that equal weightage is given to all the input variables when running machine learning algorithms or techniques. In this case centering techniques used along with mean value of that variable.

# Detailed Findings

Since this is classification problem, and the predictor variable is categorical/qualitative. First approach was to build the model using logistic regression, analyze the performance, then alternatively use machine learning model random forest.

## Logistic regression

As part of the model building exercise using logistic regression, it had to go through multiple iterations. The first step was to identify the baseline model performance, once baseline is decided then various model improvements steps were taken, these steps included primarily reducing the number of dimensions by identifying less significant variables.

### Steps

1. As part of the step all the variables post feature engineering were used to perform regression analysis, this is resulted in a model with accuracy of 69.88% with below confusion matrix.

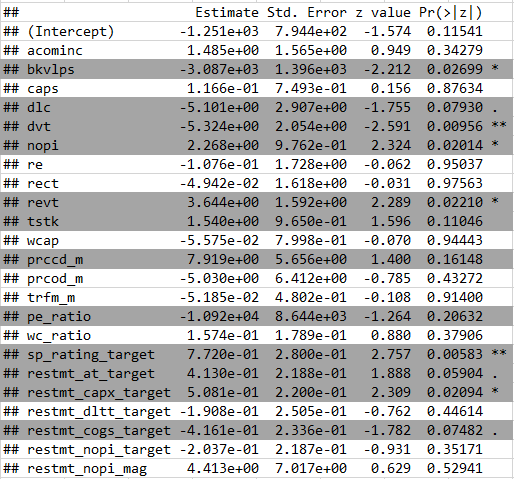
|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 53 | 6 |
| Yes | 19 | 5 |
| **Accuracy** | **69.88%** | |

This step did not give the variable coefficient value properly, so combinations of multiple variables were tried to give proper p-values for each of the variables, so instead of all variables, filtered variables were provided as next step

1. In this case the resultant confusion matrix was

|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 64 | 7 |
| Yes | 8 | 4 |
| **Accuracy** | **81.93%** | |

This gave the coefficients p-values indicating the significance of each of the variables as shown below



1. In this step only the variable with low p-value, which have higher significance on the target variables were chosen. This is results in accuracy improvement of the model to 83%

|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 64 | 6 |
| Yes | 8 | 5 |
| **Accuracy** | **83.13%** | |

Here is the list of final important variables

* Dividends - Total
* Revenue - Total
* Price - Close - Daily
* Price to earnings ratio
* Working capital ratio
* Treasury Stock - Total (All Capital)
* Standard and poor rating
* Nonoperating Income (Expense)
* Monthly Total Return Factor

## Random forest

In this approach as well baseline model performance was identified and then various steps and multiple iterations were performed to optimize the model accuracy

### Steps

1. First step was to identify right number of trees to configured for random forest, multiple iterations with 100, 300 and 500 trees were performed. Based on multiple runs the error rate stabilizes around from 100 to 300 trees.

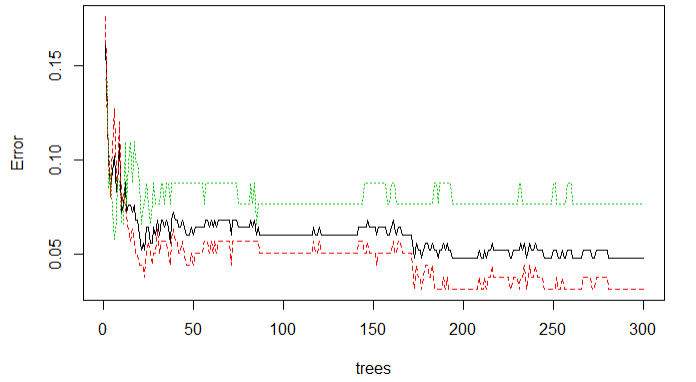
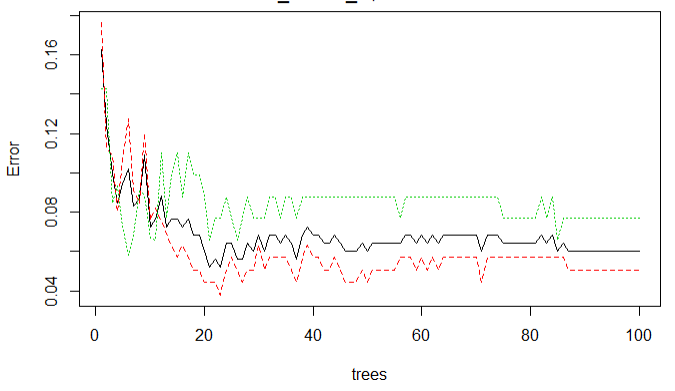


Figure 1: 100 and 300 Trees

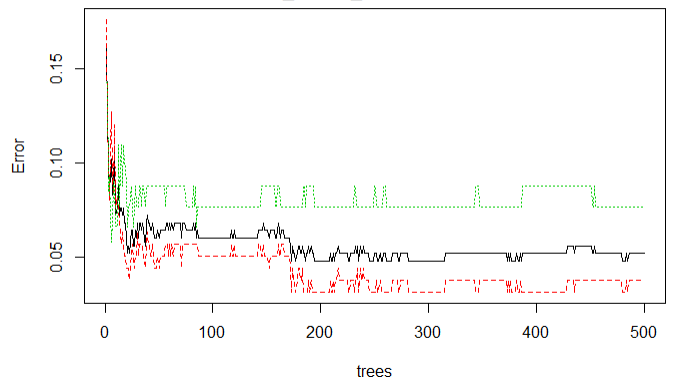


Figure : 500 Trees

Green color line captures the positive classification and red color line represent negative classification. Black like represent the average classification. Here is the accuracy with number of trees = 100

|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 67 | 7 |
| Yes | 5 | 4 |
| **Accuracy** | **85.54%** | |

Here is the accuracy with number of trees = 300. Based on accuracy percentage model with 100 number of trees performs marginally better than the model with 300 trees.

|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 66 | 7 |
| Yes | 6 | 4 |
| **Accuracy** | **84.34%** | |

Here is the list of important variables

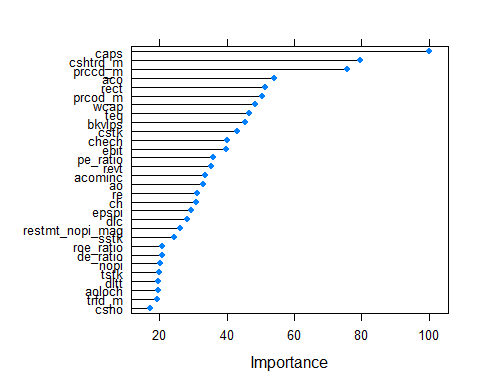


Figure : Random forest variable importance

1. As part of second steps instead of running all variables only significant variables were selected to for building the model, this did not result in any performance improvement. Here is the accuracy matrix

|  |  |  |
| --- | --- | --- |
|  | Observed | |
| Predicted | No | Yes |
| No | 66 | 7 |
| Yes | 6 | 4 |
| **Accuracy** | **84.34%** | |

1. Below is the table comparing various combinations of random forest iterations. AUC represents degree or measure of separability. It tells how much model is capable of classifying, higher the AUC, better the model is at predicting.

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No | Accuracy % | AUC | Configurations |
| 1 | 85.54217 | 88.63636 | Trees = 100 |
| 2 | 84.33735 | 86.42677 | Trees = 300 |
| 3 | 84.33735 | 85.98485 | Trees = 500 |
| 4 | 84.33735 | 79.54545 | Trees = 100 and Only important variables |

## Model comparison

Based on the 2 models’ ***Random forest is the clear winner*** and performs better from accuracy as well from AUC perspective.

Apart from these 2 model’s alternative models like gradient boosting which is better suited for classification and is based on tree model was tried for comparing with random forest performance. For logistic regression model comparison, alternative models using linear discriminant analysis (LDA) and lasso regression were used as well. Here is the comparison of other models

# Validity & Reliability Assessment

Evaluation of both models was done using

* ROC Curves and AUC
* Calibration curve

## Random Forest

Below is the ROC curve for the random forest, it shows the performance of the random forest model. A good classifier model will have the ROC curve hugging the top left corner, which is t happening in this case. AUC value of a good classifier must be between 0.8 to 0.9 in this it is 0.8863636, so the model is performing at par.



## Logistic regression



Above is ROC curve for logistic regression, AUC value here is 0.75 and does not fall in the range of 0.8 to 0.9 which indicates.

Below is the comparison of AUC values across all the models evaluated.

# Appendix