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 data used for the analysis comprised of the companies belonging to the sector, it 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. In summary data had 1768 variables for annual filings, stocks data had 76 variables, securities had 55 variables and ratings data had 6 variables. Number of records for the assigned section were 2323.

The Business objective, in this case, was to identify if the company will be subject to shareholder class action litigation, so the **target variable** was **is-litigated**, and values could be “Yes” or “No”. Since this is classification problems, 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.**  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. Below is summary view of model performance.

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.

Clearly random forest performance is better from accuracy as well as AUC perspective. Below is the list of important variables which contribute significantly to predicting the target variable.

|  |  |  |
| --- | --- | --- |
| Sr No | Variable | Description |
| 1 | re | Retained Earnings |
| 2 | pe\_ratio | Price to earning ratio |
| 3 | ao | Assets - Other |
| 4 | epspi | Earnings Per Share (Basic) - Including Extraordinary Items |
| 5 | wcap | Working Capital (Balance Sheet) |
| 6 | caps | Capital Surplus/Share Premium Reserve |
| 7 | acominc | Accumulated Other Comprehensive Income (Loss) |
| 8 | ebit | Earnings Before Interest and Taxes |
| 9 | cstk | Common/Ordinary Stock (Capital) |
| 10 | cshtrd\_m | Trading Volume - Daily |
| 11 | revt | Revenue - Total |
| 12 | chech | Cash and Cash Equivalents - Increase/Decrease |
| 13 | aco | Current Assets - Other - Total |
| 14 | aoloch | Assets and Liabilities - Other - Net Change |
| 15 | prccd\_m | Price - Close - Daily |
| 16 | ch | Cash |
| 17 | prcod\_m | Price - Open - Daily |
| 18 | roe\_ratio | Return on equity ratio |
| 19 | teq | Stockholders Equity - Total |
| 20 | wc\_ratio | Working capital ratio |

Key financial indicators related to capital, assets, earnings, stock volatility (daily opening and closing) play a significant role in predicting the target variable.

# 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. Assigned company is SJM (J. M. Smucker Company, also known as Smucker and Smucker's), it is a food and beverage company and 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

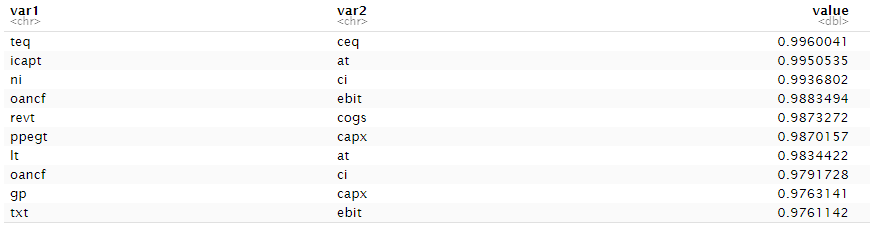
There are close to 1700+ variables in the consolidated dataset, so first step was to remove all columns with considerable amount of NA values. So, in this remove all the columns which contain 25% or greater NA or null values. This reduced the number of columns to close to 300

### Manual analysis for reducing number of predictive variables

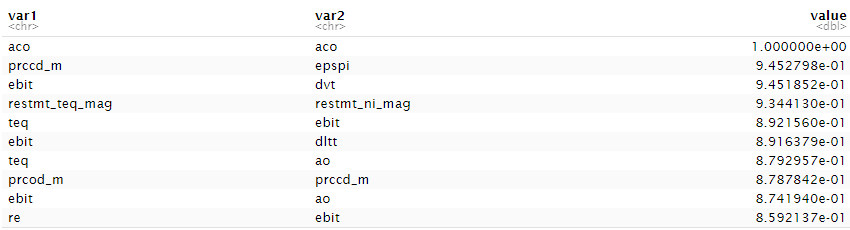
* Remove all the columns which are related to general information of the company like address, phone, URL, etc.
* There are lot of columns with hardly any variance, so identify columns which have values which are near to zero variance and remove them from dataset.
* Above steps reduced the number of 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.

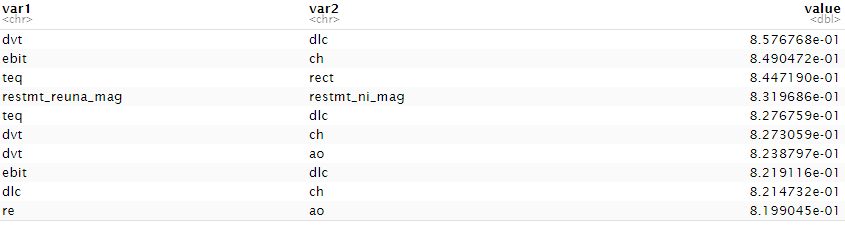
### Reducing Collinearity

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.

### 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 looks at the training dataset, the ration of litigated or not is

# Detailed Findings

# Validity & Reliability Assessment

# Appendix