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 that will **predict the probability of** **shareholder class action litigation** and its **severity** for a company. From a business perspective, each of the company buys insurance to protect themselves from these class actions lawsuits, but from the coverage standpoint, 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 the tentative settlement amount, based on which CIO of the company can decide 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. Analysis data comprised of annual filings done by the companies annually along with the restatements. Along with annual filings, it also comprised of stocks trading and securities data for each of those companies. Besides that, it had standard and poor rating data for each 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 output of the resulting variable must be "Yes" or "No". There were primarily two techniques that were used to build the model, so from the two techniques, the best model with an accuracy close to **86%** was chosen as a 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 the top 15 important variables that contribute significantly to predicting whether the company will be subject to 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) |

Table : List of important variables/prediction for class action litigation

For the target company (SJM), the final model result indicates that it **will not be litigated** by shareholder class action litigation and the **probability** of itbeen **litigated is 39%** and the **probability** of it **not** been **litigated is 61%**. In case if it gets litigated then based on existing settlement amount data, the litigation amount could range between **$498,595,595** and **$515,438,652.**

# Data and Approach

The data available for analysis contains 17,416 observations, these include annual statements issued by publicly traded companies in North America as well as restatements. Since these companies belong to various sectors and each sector litigation parameters could vary, hence the data was filtered to contain only observation from the assigned company sector. In summary, data had 1768 variables for annual filings, stocks data had 76 variables, securities had 55 variables and rating data had 6 variables. The 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

* Firstly, the dataset was filtered with annual statements only and then added variables related to restatement to each of the annual statements, the restatement variables include whether there was a restatement and by what magnitude the restatement was for.
* The following attributes mentioned below in the table were restated. The way these variables were identified was, firstly the dataset with restatements only was created and then applied the logic for removing NA columns with a 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 |

Table : Restated variables

* Post identifying the restatement variables and associated magnitude, merged these variables with the annual statement dataset.
* Now the analysis dataset contains annual statement variables as well as the restatement ones.
* The analysis dataset still contains multiple rows for a single company. So, the data set was aggregated based on gv-key and a ticker 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 was merged with the 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 dataset variables were aggregated and then merged with the main analysis dataset.
* The securities dataset was then merged with the main dataset, from this only 2 variables were chosen *Monthly Total Return Factor* and *Dividend Rate – Monthly*.
* Post securities dataset, rating dataset was merged with the main analysis dataset. Before merging the dataset additional variable with numeric value was created in the 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 the 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, the target variables related dataset of SCA filings was loaded, in this case, added the target variable litigated to the analysis dataset. For each of the company records, identify if it has any SCA filing from the SCA filling dataset, if the entry exists then mark for that company litigated equal to 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 the same to the main dataset. In case, if multiple settlement amount exists for the 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 the below table.

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

## Feature engineering

### Missing value variables

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

### Manual analysis for reducing the number of predictive variables

* Remove all the columns which are related to general information of the company like address, phone, URL, etc.
* There are a lot of columns with hardly any variance, so identify columns that have values that are near to zero variance and remove them from the dataset.
* The 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 were not related to the target variable, this reduced the number of variables 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 identified and then removed one variable from the pair to break the collinearity. The collinearity threshold that was considered was 0.9, anything above, an attempt was made to break the collinearity pair.

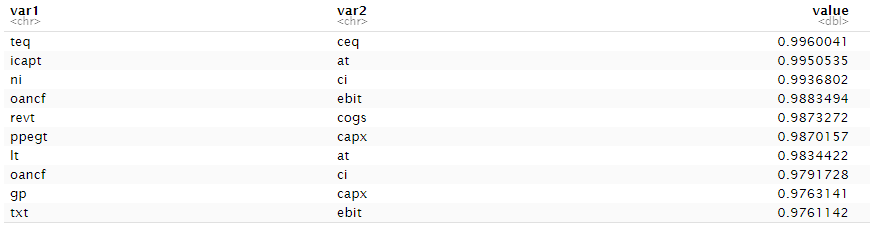
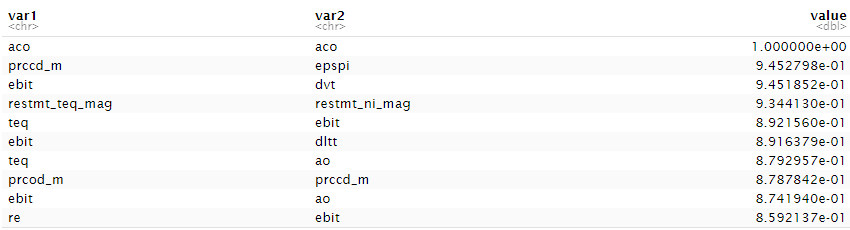


Figure : Collinear variables

Post collinearity removal the table looks something like this 

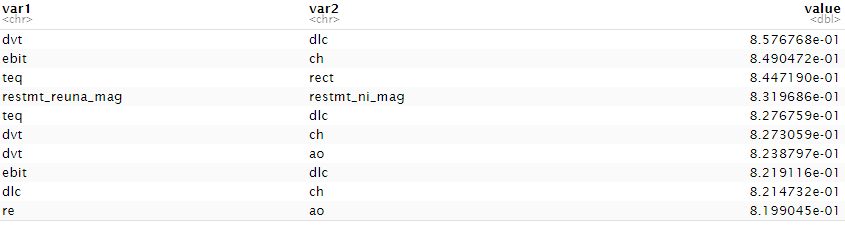


Figure : Remove collinear variables

Reducing the collinearity and other manual analysis led to the final 63 predictive variables. Some of the variables with high collinearity were ignored based on manual analysis and considering their business definitions. All the 63 variables have been listed in the appendix section along with their description.

### Encoding

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

### Training and Test dataset set

The final dataset resulted in 333 observations, this was the final dataset with all data merging, aggregation, encoding, etc. This data set is then split into training and test datasets with the ratio of 0.75/0.25, with 0.75 been the training dataset and 0.25 been the test dataset.

### Class imbalance

In this case, the target variable is whether the 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 a clear sign of class imbalance. In the case of class imbalance, the machine learning techniques tend to be more biased towards the majority class, causing bad classification of the minority class which is a predictor in this case. So, to reduce the imbalance R ROSE package was used and the strategy that was leveraged in this case was a combination of both 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% |

Table :Class Imbalance

### Scaling variables

All the input variables were scaled to have an 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, the centering technique was used along with the mean value of that variable.

## Approach

There are two business objectives, firstly to predict whether a particular company will face shareholder class action litigation (Yes/No), and secondly substantiate the likelihood severity of settlement amount.

The first business objective is a classification problem, and the predictor variable is categorical/qualitative. So, in this case, the first step would be to build a model using logistic regression, analyze the performance, then use a machine learning model random forest to compare the accuracy performance.

For the second business objective, there is a limitation from the data perspective as there are not that many companies in the data set that have a settlement amount. Hence the approach would be first creating a separate dataset from the dataset which was used to build a classification model, in this dataset, there would be only those companies which have been litigated and have settlement amount.

Post dataset creation add a variable as settlement ratio which would be the proportion of settlement amount by market value (capital) for that company. Once the dataset is been created identify the standard error for the settlement amount and calculate the mean of the settlement ratio. Now for calculating the target company settlement amount, multiply its market value with the average settlement ratio, this should give the settlement amount for that company. Using the standard error for the sample then calculate the confidence interval of the settlement amount for the target company.

# Detailed Findings

## 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 the baseline is decided then various model improvement steps were taken, these steps included primarily reducing the number of dimensions by identifying less significant variables.

### Steps

1. As part of this step, all the variables identified during feature engineering were used to perform regression analysis, this resulted in a model with an accuracy of 69.88% with below confusion matrix. The confusion matrix is a way to express how many of a classifier’s predictions were correct. In the confusion matrices below, the columns represent the true labels (observed) and the rows represent predicted labels. Diagonal values represent correct predictions, values in other cells represent instances where the classifier was incorrect, the row tells what the classifier predicted, and the column indicates what the actual value is.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 53 | 6 | 90% |
| Yes | 19 | 5 | 21% |
| **Accuracy** |  |  | **69.88%** |

Table : Confusion matrix for logistic regression with all variables

This step did not give the value of the variable coefficients properly, so combinations of multiple variables were tried to get proper p-values for each of the variables, so instead of all variables, filtered variables were provided as the next step. *(Please refer to the figures in Appendix section - Logistic regression first run – significant variables).* Hosmer-Lemeshow which is a goodness of fit test did not give an appropriate p-value.

1. Multiple iterations were performed till the model results started showing distinct significant variables values. In this case, the resulting confusion matrix was as below

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 64 | 7 | 90% |
| Yes | 8 | 4 | 33% |
| **Accuracy** |  |  | **81.93%** |

Table : Confusion matrix for logistic regression with selective variables

This gave the coefficients p-values indicating the significance of each of the variables as shown below, any p-value around 0.10 or less of is of prime significance.

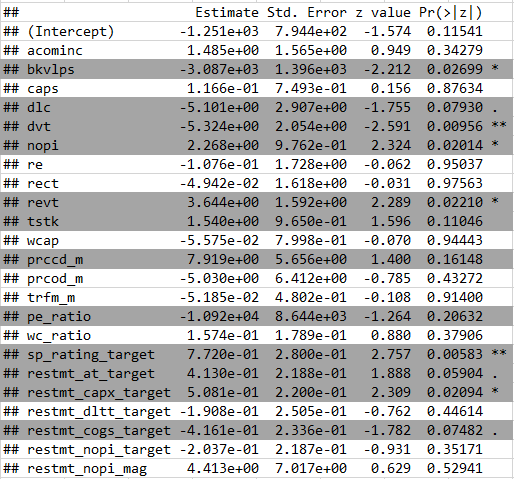


Figure : Significant variables - P-Values

Hosmer-Lemeshow goodness of fit test indicated a p-value of 0.0790 which is greater than 0.05. The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is less than 0.05

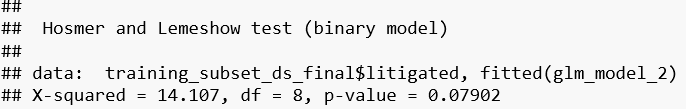


Figure :Hosmer-Lemeshow goodness of fit test

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 64 | 6 | 91% |
| Yes | 8 | 5 | 38% |
| **Accuracy** |  |  | **83.13%** |

Table : Confusion matrix with the final set of variables for logistic regression

Hosmer-Lemeshow goodness of fit test indicated a p-value of 0.08197 which has improved compared to previous runs.

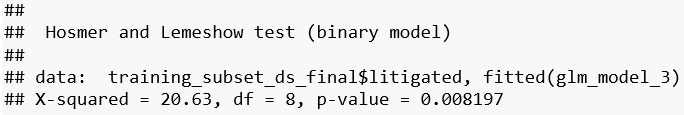


Figure :Hosmer-Lemeshow goodness of fit test

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

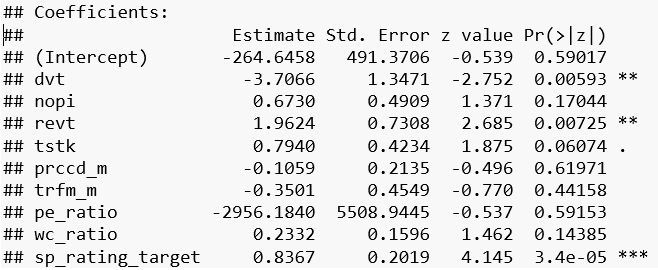


Figure : Final list of significant variables for logistic regression

## 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. The first step was to identify the right number of trees to be configured for the random forest, multiple iterations with 100, 300, and 500 trees were performed. Based on multiple runs the error rate stabilizes around 100 to 300 trees. The green color line captures the positive classification, and the red color line represents the negative classification. The black line represents the average classification.

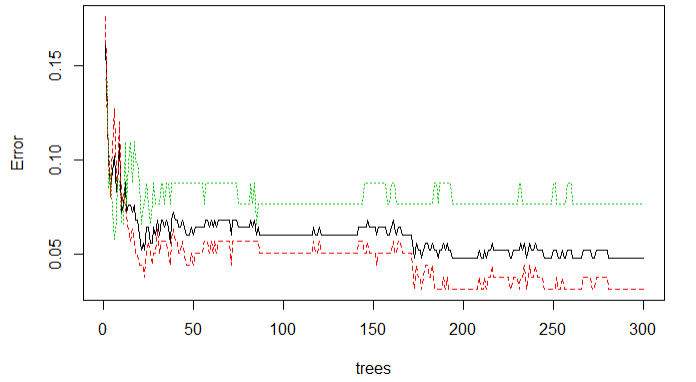
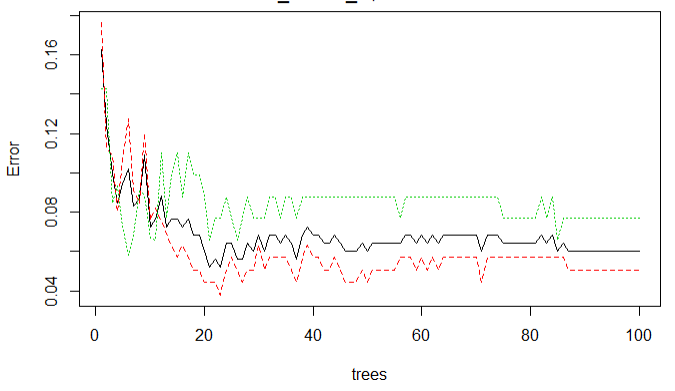


Figure : 100 and 300 Trees

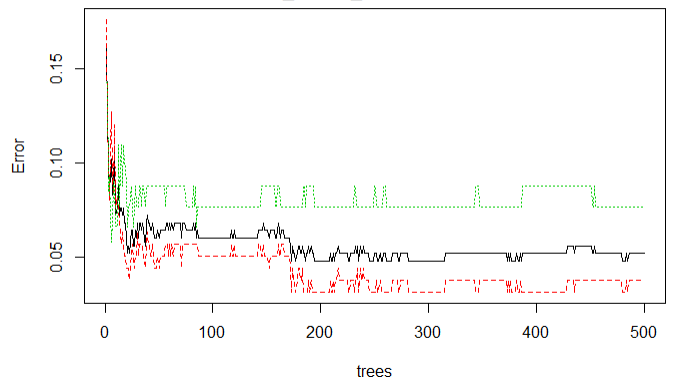


Figure : 500 Trees

Here is the accuracy with the number of trees = 100

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 67 | 7 | 91% |
| Yes | 5 | 4 | 44% |
| **Accuracy** |  |  | **85.54%** |

Table : Confusion matrix with the number of trees = 100

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 66 | 7 | 90% |
| Yes | 6 | 4 | 40% |
| **Accuracy** |  |  | **84.34%** |

Table : Confusion matrix with the number of trees = 300

Here is the accuracy with the number of trees = 500, this is identical to that of 300 trees.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 66 | 7 | 90% |
| Yes | 6 | 4 | 40% |
| **Accuracy** |  |  | **84.34%** |

Table : Confusion matrix with the number of trees = 500

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

|  |  |  |  |
| --- | --- | --- | --- |
|  | Observed | | Percentage correct |
| Predicted | No | Yes |  |
| No | 66 | 7 | 90% |
| Yes | 6 | 4 | 40% |
| **Accuracy** |  |  | **84.34%** |

Table : Confusion matrix for the random forest with selected variables

1. Below is the table comparing accuracies of various random forest iterations.

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

Table : Random forest number of trees and their accuracies

### Important Variables

Here is the list of important variables for the best accuracy random forest model. The importance value indicates, how much the model accuracy percentage decreases if we drop that variable, higher the value, greater the importance of the variable in the model. Trading volume is the most important variable.

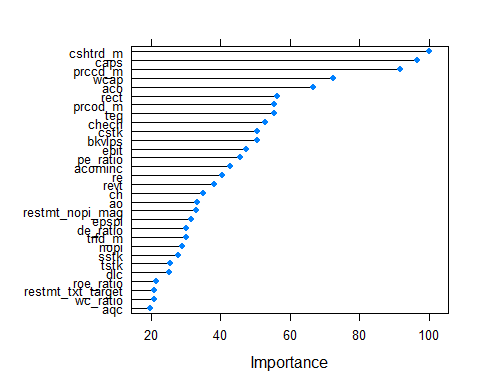


Figure : Random forest variable importance

|  |  |  |  |
| --- | --- | --- | --- |
| Sr No | Variable | Description | Importance Value |
| 1 | cshtrd\_m | Trading Volume - Daily | 100.00 |
| 2 | caps | Capital Surplus/Share Premium Reserve | 96.57 |
| 3 | prccd\_m | Price - Close - Daily | 91.75 |
| 4 | wcap | Working Capital (Balance Sheet) | 72.38 |
| 5 | aco | Current Assets - Other - Total | 66.79 |
| 6 | rect | Receivables - Total | 56.45 |
| 7 | prcod\_m | Price - Open - Daily | 55.60 |
| 8 | teq | Stockholders Equity - Total | 55.41 |
| 9 | chech | Cash and Cash Equivalents - Increase/(Decrease) | 52.81 |
| 10 | cstk | Common/Ordinary Stock (Capital) | 50.58 |
| 11 | bkvlps | Book Value Per Share | 50.51 |
| 12 | ebit | Earnings Before Interest and Taxes | 47.33 |
| 13 | pe\_ratio | Price to earnings ratio | 45.74 |
| 14 | acominc | Accumulated Other Comprehensive Income (Loss) | 42.68 |
| 15 | re | Retained Earnings | 40.57 |
| 16 | revt | Revenue - Total | 38.33 |
| 17 | ch | Cash | 34.97 |
| 18 | ao | Assets - Other | 33.25 |
| 19 | restmt\_nopi\_mag | Restement magnitude for Nonoperating Income (Expense) | 32.90 |
| 20 | epspi | Earnings Per Share (Basic) | 31.50 |
| 21 | de\_ratio | Debt to equity ratio | 30.15 |
| 22 | trfd\_m | Daily Total Return Factor | 30.13 |
| 23 | nopi | Nonoperating Income (Expense) | 29.12 |
| 24 | sstk | Sale of Common and Preferred Stock | 27.74 |
| 25 | tstk | Treasury Stock - Total (All Capital) | 25.68 |
| 26 | dlc | Debt in Current Liabilities - Total | 25.35 |
| 27 | roe\_ratio | Return on equity ratio | 21.50 |
| 28 | restmt\_txt\_target | Restatement for Income Taxes - Total | 20.99 |
| 29 | wc\_ratio | Working capital ratio | 20.94 |
| 30 | aqc | Acquisitions | 19.84 |
| 31 | csho | Common Shares Outstanding | 19.78 |

Table :Feature importance for random forest

## Model performance comparison with other techniques

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

Figure : Two main model accuracy comparison

Apart from these 2 model’s alternative models like gradient boosting which is better suited for classification and is based on the tree model was tried for comparing with random forest performance. For the 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

Figure : Accuracy comparison across various models

## ROC and AUC – Model performance measurement

AUC - ROC curve is a performance measurement technique used for classification problems. ROC is a probability curve and AUC represents the degree or measure of separability. It indicates how much is the model capable of classifying target variables. The higher the AUC, the better the model is at predicting. A good classifier model will have the ROC curve hugging the top left corner, which is happening in this case. AUC value of a good classifier must be between 0.8 to 0.9.

### Random Forest

Below is the ROC curve for the random forest. AUC value, in this case, is 0.8863636 which falls in between 0.8 and 0.9, which is an indicator that the model is performing satisfactorily.



Figure : ROC curve for random forest

### Logistic regression



Figure : ROC curve for logistic regression

Above is the 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 substandard model performance.

Below is the comparison of AUC values across all the models evaluated. ***Based on AUC values Random forest is the best performing model.***

Figure : AUC Comparison across models

## Target company classification (SJM)

Based on the best performance, the random forest is the recommended model. The recommended model indicates that the target company may not face class action litigation and there is a 39% probability of class action litigation been filed. Below is the classification probability matrix for the target company SJM.



Figure : Target company prediction probability

## Other findings

Based on the litigated companies’ data here are some common traits that could be an indication of class action litigation.

|  |  |  |
| --- | --- | --- |
|  | Understanding Litigation | Cornwell & Sample - Trial Lawyers Litigated | Understanding Litigation | Cornwell & Sample - Trial LawyersNot Litigated |
| Trading Volume - # Shares (Avg) | 1935593.3  High | 746803.9 |
| Capital Surplus/Share Premium Reserve (Avg) | $1680.7703 million | $874.1961 million |
| Working Capital (Avg) | $420.2020 million | $212.4778 million |
| Current Assets  (Avg) | $278.9319 million | $173.5446 million |
| Receivables - Total  (Avg) | $701.0267 million | $477.8728 million |
| Stockholders’ Equity (Avg) | $4942.408 million | $2193.186 million |
| Common/Ordinary Stock (Capital) | $257.9056 million | $186.4528 million |
| Book Value Per Share (Median) | $7.019325 | $3.343108 |
| Earnings Before Interest and Taxes | $1384.401 million | 670.237 million |
| PE Ratio | Negative | Positive |
| Retained Earnings  (Avg) | $ 4317.692 million | $ 1554.839 million |
| Revenue – Total  (Avg) | $20083.754 million | $5680.156 million |
| Cash (Avg) | $744.0797 million | $338.7759 million |
| Assets – Other (Avg) | $357.0753 million | $175.5579 million |
| Earnings Per Share (Basic) (Avg) | $ 0.9608556 | $1.5948895 |
| Restatement for Nonoperating Income (Expense) | $47.37 million | $(7029.75) million |

Table : Detail attributes across companies that are litigated and not litigated

Some clear indicators of company been litigated are high trading volume, higher revenue, negative PE ratio, earnings per share less than a dollar.

## Severity estimation

Based on the classification dataset, created a separate dataset, and filtered to include only rows which have litigation settlement amount, based on this got a very limited number of observations as below



Figure : Missing values for litigation settlement amount observations

There is a missing value over here. Market capital value is calculated by multiplying the price of a stock by its total number of outstanding shares. So, to calculate market value/capital for the GRO company brought in the following variables from the model dataset and replaced the missing value by calculating the market value.

1. outstanding shares (csho)
2. mean of “Price - Close – Daily” (prccd\_m).

Market value = outstanding shares (csho) \* mean of “Price - Close – Daily” (prccd\_m).

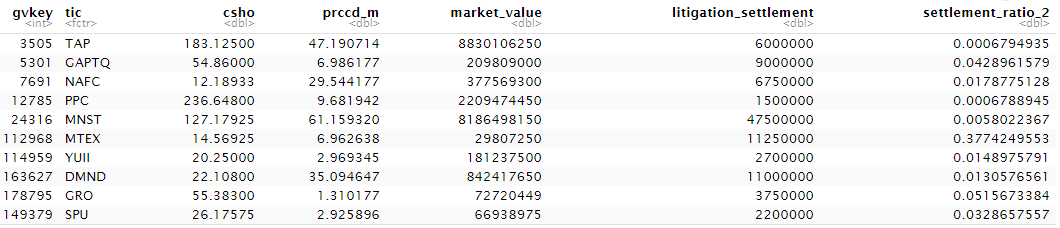


Figure : Addressed NA value for market value

Settlement ratio = Litigation settlement amount / market capital

The standard error for litigation settlement for the above sample size = $4,296,698.02

Mean of settlement ratio = 0.05577476

Target company settlement amount = market capital \* mean of settlement ratio

Estimated target company settlement amount = $507,017,124

Based on the standard error calculate the 95% confidence interval which will be ($498,595,595 and $515,438,652). So, in summary we are **95% confident** that the settlement amount for the target company is **between $498,595,595** and **$515,438,652, estimated target company settlement amount is $507,017,124.**

## Next steps

One of the biggest hurdles with this model building exercise was the amount of data that is available for the target company sector, in this case, the total number of observations was close to 333. For these 333 records they had to further split in training and test dataset, so effectively a smaller number of observations were available to train both models’ logistic regression and random forest. So, the next logical step would be to get more data for the assigned sector and retrain the models and verify the model performance again.

## Limitations

The models that were built with computing constraints of personal laptops, so the model was trained with a minimal set of records, but in a commercial approach, the model will be trained using commodity hardware or cloud-based approach where there would not be any computing constraints which will not provide any boundaries on the sample size to train the model.

The second limitation was the amount of data available for the litigation settlement amount, there were very few observations, due to which rudimentary statistical approach of calculating means and then the confidence interval was applied. As a general thumb rule for any model building exercise

number of observations = number of predictors \* 4.

In this case number of observations was even less than the number of predictor variables.

# Validity & Reliability Assessment

This section covers how the recommended model should be tested in future implementation. When new data arrives, the model needs to be periodically trained and deployed to the production system. Data tend to vary with time, so it is recommended to train the model periodically, this is sometimes referred to as continual learning, where the model gets trained and deployed continuously.

In this continual training approach, the same approach mentioned above of confusion matrices, ROC curves, and the AUC values will be used to validate the model and get the accuracy.

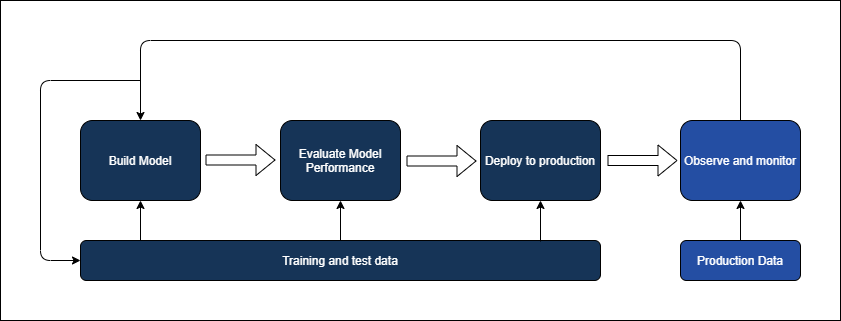


Figure : Continuous training of the model

# Appendix

## Predictive variables

Here is the list of predictive variables that have been used for building the models

|  |  |  |
| --- | --- | --- |
| Sr No | Variable name | Variable description |
| 1 | aco | Current Assets - Other - Total |
| 2 | acominc | Accumulated Other Comprehensive Income (Loss) |
| 3 | ao | Assets - Other |
| 4 | aoloch | Assets and Liabilities - Other - Net Change |
| 5 | aqc | Acquisitions |
| 6 | bkvlps | Book Value Per Share |
| 7 | caps | Capital Surplus/Share Premium Reserve |
| 8 | ch | Cash |
| 9 | chech | Cash and Cash Equivalents - Increase/(Decrease) |
| 10 | csho | Common Shares Outstanding |
| 11 | cshtrd\_m | Trading Volume - Daily |
| 12 | cstk | Common/Ordinary Stock (Capital) |
| 13 | de\_ratio | Debt to equity ratio |
| 14 | dlc | Debt in Current Liabilities - Total |
| 15 | dltt | Long-Term Debt - Total |
| 16 | dvt | Dividends - Total |
| 17 | ebit | Earnings Before Interest and Taxes |
| 18 | epspi | Earnings Per Share (Basic) - Including Extraordinary Items |
| 19 | gvkey | Global Company Key |
| 20 | litigated | Target Variable |
| 21 | litigation\_settlement | Litigation settlement amount |
| 22 | nopi | Nonoperating Income (Expense) |
| 23 | pe\_ratio | Pricing to earnings ratio |
| 24 | prccd\_m | Price - Close - Daily |
| 25 | prcod\_m | Price - Open - Daily |
| 26 | Re | Retained Earnings |
| 27 | rect | Receivables - Total |
| 28 | revt | Revenue - Total |
| 29 | roe\_ratio | Return on Equity |
| 30 | siv | Sale of Investments |
| 31 | sp\_rating\_target | Standard and poor rating |
| 32 | sstk | Sale of Common and Preferred Stock |
| 33 | teq | Stockholders Equity - Total |
| 34 | tic | Ticker Symbol |
| 35 | trfd\_m | Daily Total Return Factor |
| 36 | trfm\_m | Monthly Total Return Factor |
| 37 | tstk | Treasury Stock - Total (All Capital) |
| 38 | wc\_ratio | Working capital ratio |
| 39 | wcap | Working Capital (Balance Sheet) |
| 40 | restmt\_at\_mag | Indicating if restatement for - Assets - Total |
| 41 | restmt\_at\_target | Indicating if restatement for - Assets - Total |
| 42 | restmt\_capx\_mag | Restatement Magnitude for - Capital Expenditures |
| 43 | restmt\_capx\_target | Restatement indicator for Capital Expenditures |
| 44 | restmt\_cogs\_mag | Restatement Magnitude for - Cost of Goods Sold |
| 45 | restmt\_cogs\_target | Restatement indicator for Cost of Goods Sold |
| 46 | restmt\_dltt\_target | Restatement indicator for Long-Term Debt - Total |
| 47 | restmt\_epspi\_mag | Restatement Magnitude for - Earnings Per Share (Basic) - Including Extraordinary Items |
| 48 | restmt\_epspi\_target | Restatement indicator for Earnings Per Share (Basic) - Including Extraordinary Items |
| 49 | restmt\_ib\_target | Restatement indicator for Income Before Extraordinary Items |
| 50 | restmt\_ni\_mag | Restatement Magnitude for - Net Income (Loss) |
| 51 | restmt\_ni\_target | Restatement indicator for Net Income (Loss) |
| 52 | restmt\_nopi\_mag | Restatement Magnitude for - Nonoperating Income (Expense) |
| 53 | restmt\_nopi\_target | Restatement indicator for Nonoperating Income (Expense) |
| 54 | restmt\_reuna\_mag | Restatement Magnitude for - Retained Earnings - Unadjusted |
| 55 | restmt\_reuna\_target | Restatement indicator for Retained Earnings - Unadjusted |
| 56 | restmt\_teq\_mag | Restatement Magnitude for - Stockholders Equity - Total |
| 57 | restmt\_teq\_target | Restatement indicator for Stockholders Equity - Total |
| 58 | restmt\_txt\_mag | Restatement Magnitude for - Income Taxes - Total |
| 59 | restmt\_txt\_target | Restatement indicator for Income Taxes - Total |
| 60 | restmt\_wcap\_mag | Restatement Magnitude for - Working Capital (Balance Sheet) |
| 61 | restmt\_wcap\_target | Restatement indicator for Working Capital (Balance Sheet) |
| 62 | restmt\_xint\_mag | Restatement Magnitude for - Interest and Related Expense - Total |
| 63 | restmt\_xint\_target | Restatement indicator for Interest and Related Expense - Total |

## Files

### R Knitted file

Attached is the knitted R file.



### Model files

1. Classification input dataset – This is a raw dataset, prior rescaling, and class balancing.



1. Classification input dataset for Random forest – This is the final dataset for the random forest model; it has likelihood probabilities for the target variable.



1. Classification input dataset for logistic regression – This is the final dataset for the logistic regression model; it has likelihood probabilities for the target variable.



1. Severity estimation input dataset.



## Important R code snippets

### Logistic Regression

Below code snippets for running logistic regression first and final step

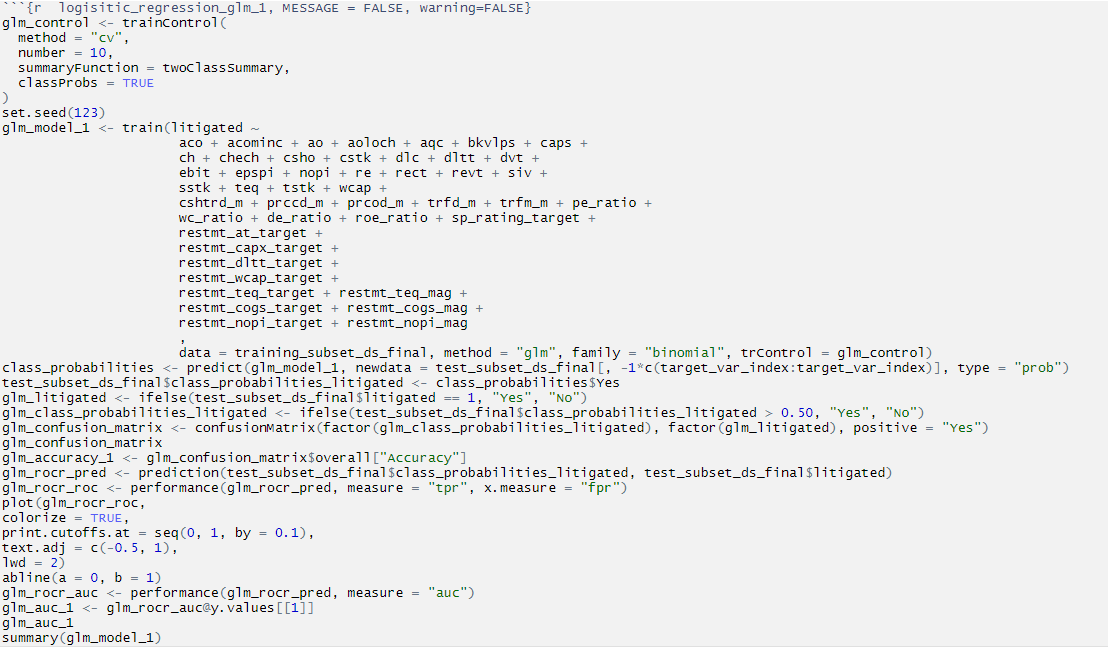


Figure : Logistic regression first run



Figure : Logistic Regression last run

### Random forest

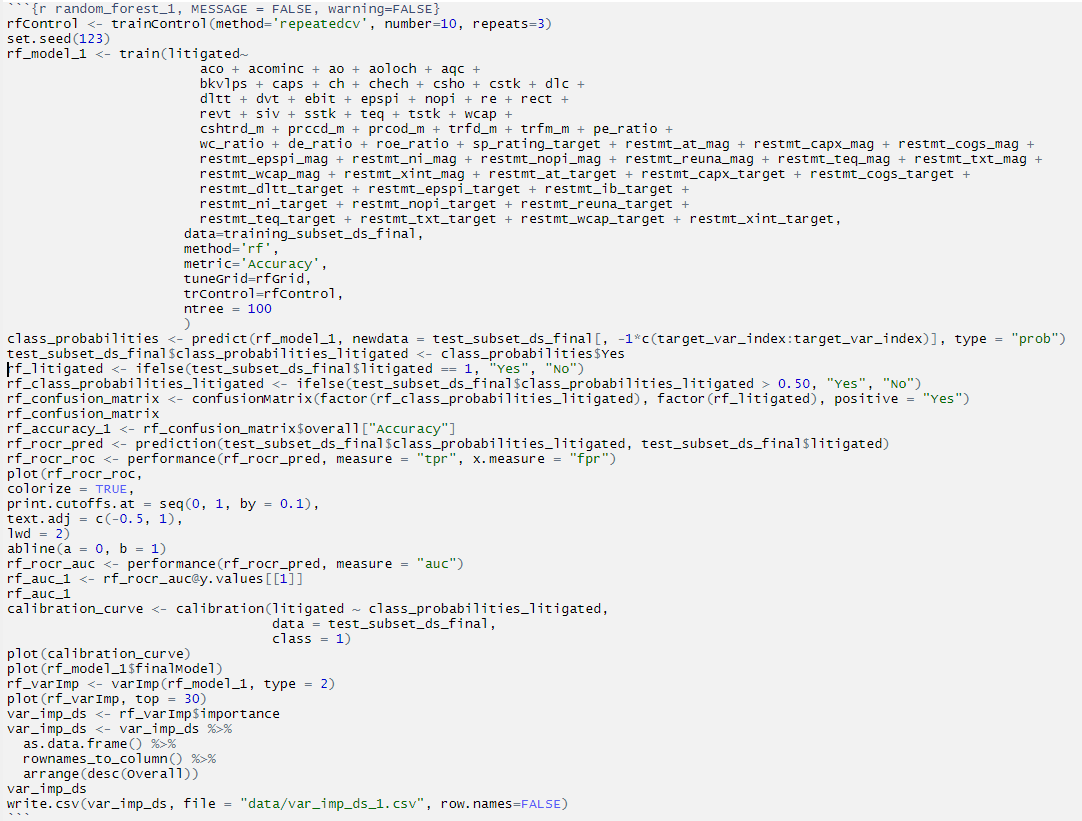


Figure : Random Forest

## Logistic regression first run – significant variables

