Analysis of the Likelihood and Potential Monetary Impact of a Shareholder Lawsuit Against the Boston Beer Company

Brian Cullinan

5/6/22

[**Executive Summary** 2](#_Toc102739298)

[Introduction & Problem Statement 2](#_Toc102739299)

[Process Overview & Methodology 2](#_Toc102739300)

[Summary of Findings & Scorecard for Boston Beer Company: 3](#_Toc102739301)

[Data Sources 4](#_Toc102739302)

[General Methodology 4](#_Toc102739303)

[Specific Data & Feature Engineering Steps 5](#_Toc102739304)

[**Risk Assessment: Probability of Lawsuit** 8](#_Toc102739305)

[Introduction and General Approach 8](#_Toc102739306)

[Additional Feature Selection 9](#_Toc102739307)

[Model Selection 10](#_Toc102739308)

[Final Model Evaluation 14](#_Toc102739309)

[Risk Assessment Summary for Boston Beer Company 15](#_Toc102739310)

[**Severity Assessment: Potential Liability** 17](#_Toc102739311)

[Introduction and General Approach 17](#_Toc102739312)

[Additional Feature Selection 17](#_Toc102739313)

[Model Selection 19](#_Toc102739314)

[Final Model Evaluation 21](#_Toc102739315)

[Severity Assessment Summary for Boston Beer Company 21](#_Toc102739316)

[**Appendix** 22](#_Toc102739317)

[Citations 22](#_Toc102739318)

[Additional Files 22](#_Toc102739319)

[Code Repository 23](#_Toc102739320)

[Feature Collinearity Heat Map 23](#_Toc102739321)

# **Executive Summary**

## Introduction & Problem Statement

All publicly traded companies are required to disclose financial performance information to investors annually via a 10-K report. Additionally, they must disclose any corrections or changes to previous 10-K filings with a process known as a restatement. From time to time, these financial disclosures do not meet investor expectations, either due to normal fluctuations in business performance, alleged mismanagement, or a lack of transparency or fraud. These financial conditions can sometimes lead to a Shareholder Class Action (SCA) lawsuit. (Banasiewicz, 2015, pgs. 11-12). Although SCA lawsuits are rare, only impacting an average 3.5% of companies, when they do occur, they have the potential to be extremely costly. Since 1996, SCA lawsuits have cost companies over $87 billion in sum (Banasiewicz, 2015, pg. 102).

Given the potential high impact that an SCA lawsuit may have, it is best for companies to mitigate risk by taking preventative action should it be deemed necessary. This analysis was conducted for the Boston Beer Company (BBC) for this purpose and will quantify the risk of an SCA lawsuit against the BBC as well as estimate the potential monetary liability.

This analysis was conducted by reviewing data compiled from 10-K reports, stock performance, quality ratings, and securities for 348 companies in the “Consumer Staples” sector filed between 2009-2013. The original dataset contains 46 instances of companies who have faced an SCA lawsuit (13.2%) and 302 who have not (86.8%).

## Process Overview & Methodology

This analysis compiled for the BBC contains two assessments: 1) the predicted risk (probability) that the BBC will face an SCA lawsuit, and 2) the predicted severity of a settlement should the BBC face an SCA lawsuit. The first assessment (“Probability of a Lawsuit”) was modeled with a Voting Classifier model. The Voting Classifier model utilizes four different models (Random Forest (RF) Classifier, Logistic Regression (LR), GaussianNB, and ADA Boosting), with the results from RF and LR models weighted four times, which was tested to be the best performing model. A classification prediction is made through voting from the results of each of the models with weighting considered. Further discussion of how this model was chosen is provided in the “[Risk Assessment Summary for Boston Beer Compan](#_Risk_Assessment:_Probability)[y](#_v4mu79gst74t)” section.

For this assessment, the primary objective was to minimize false negative predictions (measured by recall score), where a company was predicted not to face an SCA lawsuit but did. A false negative is considered the worst possible outcome because if a company is not flagged as being at-risk, then they may miss the opportunity to take preventive actions. An outcome in which a company is classified as at-risk for an SCA lawsuit when they are not (false positive) is less adversely impactful since the downside is taking action to reduce risk.

Through validation with a test dataset, the Voting Classifier model was shown to be 55.7% accurate in predicting if a company would be sued (see: [Figure3](#_Figure3_-_Confusion)). The model was 66.7% accurate in correctly predicting the subset of companies who were sued. Recall that 13.2% of companies in this sector faced an SCA lawsuit, so the model is approximately five times more accurate than random guessing. By focusing on reducing false negatives, the model is cautious by design, which is illustrated by the bulk of the incorrect predictions being false positives (a company predicted to face an SCA lawsuit, but did not), which is likely to be a less adversely impactful false prediction. Of the 31 incorrect predictions, 27 (87.1%) were false positives. The cautious tendency of the model means that when it predicts a company will not face an SCA lawsuit, it is correct 88.6% of the time.

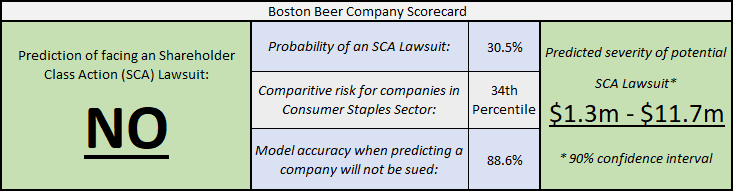
The second assessment (“Potential Liability”) predicts a penalty, often a settlement amount, that the BBC will be liable for should it face a successful SCA lawsuit. This prediction is limited to the settlement amount and does not include other related expenses such as legal fees.

This assessment faced the significant challenge of a small number of positive observations (i.e., companies that faced an SCA lawsuit), with only 11 such examples, one of which (Monster Beverage Company) needed to be removed as an outlier. The low number of observations meant the model could only predict with a high degree of uncertainty (measured by Root Mean Squared Error (RMSE)). For the BBC, this results in a final dollar amount presented as a wide range, with 90% confidence that the actual amount would fall into that range. This assessment was modeled with a Partial Least Squares (PLS) model which outperformed several competing models and was chosen in part because the model tends to do well with sparse data.

## Summary of Findings & Scorecard for Boston Beer Company:

These models classify the BBC with the group predicted not to face an SCA lawsuit and gives the BBC a 30.5% chance of facing a lawsuit. The largest factor leading to the prediction of not facing a lawsuit is the strong overall market performance, reflected in the fact that BBC stock value increased 377% from $47.53 to $226.69 over the observation period. A secondary factor was that the BBC did not file a restatement. Factors which negatively affected the BBC prediction were high expenditures on property, plant, and equipment (PPE) as well as high administrative costs. Should the BBC face an SCA lawsuit, the model asserts a 90% confidence that the final settlement amount would be between $1.3 and $11.7 million. (0.2% to 2.1% of median annual revenue).

### *Figure1 - Boston Beer Company Risk Scorecard*



**Data Overview**

## Data Sources

The following data sources, originally sourced from the U.S. Securities and Exchange Commission’s (SEC) Electronic Data Gathering, Analysis, and Retrieval (EDGAR) database, were used throughout the analysis. Each of these sources was filtered to only the companies in the “Consumer Staples” sector.

1. “Fundamentals” - General financial data from 10-K reports and restatements
2. “Stocks” - Daily stock trading, performance, and dividend data
3. “Securities” - Data on monthly trading of company stock
4. “Ratings” - Various credit ratings

## General Methodology

The source data provided multiple observations per company at various intervals (daily, monthly, annually). Ultimately, since this analysis is seeking information regarding whether a company faced an SCA lawsuit and not the number of lawsuits or when the company faced a lawsuit, one observation per company was used to model the overall likelihood that an individual company would face an SCA lawsuit and what the subsequent potential liability would be. Therefore, the objective in data preparation was to join all of the data sources and condense to one observation per company. This resulted in a single data set compiled from all of the sources and aggregated to a single observation per company.

First, the source data provided over 2,000 variables that could be used as predictive features. The breadth of this data is too large for the predictive models and required feature trimming to a smaller number of the most meaningful variables. Additional detail on this process is provided in the [Data & Feature Engineering](#_Specific_Data_&) section.

Once feature engineering was completed, the next step was to aggregate the data without losing valuable information. For example, for some variables, a measure of central tendency would be appropriate (e.g., median), and in other cases a measure of variability would be more meaningful (e.g., standard deviation). To address this, both the median and standard deviation for each variable were captured.[[1]](#footnote-1) In each assessment described below, additional detail will be provided on how these versions of the features were compared and which was ultimately used.

Since many features in the datasets were skewed and highly variable due to the wide range of company size and financial performance among the companies included, the data was standardized using a z-score scaling approach so that all continuous variables were replaced with their z-score.[[2]](#footnote-2) This condensed the spread of values for a given feature, and better suited scenarios such as comparing companies with orders of magnitude difference in their metrics (e.g., a company valued in the hundreds of billions of dollars with a company valued in the millions).

Finally, for all features selected that had null values, null values were imputed using [Scikit Learn’s Iterative Imputer](https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html). Other simple imputation methods such as filling the missing values with the mode or median were considered but given the high variability from company to company in the dataset it was determined that this would add inaccurate data and instead it was decided that the best imputation strategy would be to utilize a method that relied upon more advanced modeling.

## Specific Data & Feature Engineering Steps

All the source datasets, which were combined into the final modeling dataset, required a robust feature selection process. For each dataset, this process was broken down into a few phases:

1. Phase 1: Remove Obviously Unrelated or Unhelpful Features
2. Phase 2: Remove Duplicate Features
3. Phase 3: Remove Redundant Features
4. Phase 4: Remove Low-Value Features Using Modeling Techniques

In Phase 1 for all datasets, any features that were clearly not helpful for modeling a prediction were removed. Examples of these include addresses, phone numbers, websites, dates, and general classification codes. Next, any feature that had over 20% null values was removed from the dataset, because a feature with that number of null values is unlikely to be helpful or accurate in modeling. This step resulted in between 21 to 82 percent of features being removed, depending on the dataset (see [table1](#_Table1:_Features_removed)).

### *Table1: Features removed from each dataset due to over 20% null values*

|  |  |  |
| --- | --- | --- |
| Dataset | # Of Features Removed | % Of Features Removed |
| Fundamentals | 1,463 | 82% |
| Stocks | 37 | 48% |
| Securities | 23 | 41% |
| Ratings | 3 | 21% |

The final step in this phase was to remove any feature that either only had one value, or where a very high percentage (above 90%) were one of two values. This was done because the low variance of these features meant that they would be poor predictors. An example of this is the “Level of Consolidation” feature from that Fundamentals dataset which had the same value (“C”) for all companies.

For Phase 2, all the datasets were assessed for duplicate features, meaning that the exact same feature appeared in more than one dataset. Since the Fundamentals dataset was processed first, this phase primarily applied to the Stocks, Securities, and Ratings dataset. For example, the securities dataset contains data on stock prices, but that data was also available in the Stocks data, so it was retained in the Stocks dataset and dropped from the Securities dataset.

Phase 3, removing redundant features, was manually labor intensive and required close inspection of the data. All these datasets contain many similar features that describe nuanced measures within a particular area. This analysis refers to them as “sets”. An example of a set is provided in [Table2](#_Table2:_Example_of) below. These features all describe income related to operations. From this set, only “oancf” (Operating Activities - Net Cash Flow) and “opeps” (Earnings Per Share from Operations) were retained as predictor features and the remainder was dropped for the reasons described in [Table2](#_Table2:_Example_of) below.

### *Table2: Example of a Set from Fundamentals Dataset*

| **Variable Name** | **Description** | **Retained** | **Notes** |
| --- | --- | --- | --- |
| oancf | Operating Activities - Net Cash Flow | Yes |  |
| oiadp | Operating Income After Depreciation | No | Identified as a more granular subset measure to “oancf”. Therefore, all value would be expressed via “oancf” |
| oibdp | Operating Income Before Depreciation | No |
| opeps | Earnings Per Share from Operations | Yes |  |
| oprepsx | Earnings Per Share - Diluted - from Operations | No | Analysis showed that this is nearly always the same value as “opeps” |

While all datasets had some form of sets, this phase was most impactful to the Fundamentals dataset. This was the case for two reasons: first, the Fundamentals dataset contained many more sets of nuanced measures, and second, the Fundamentals dataset was processed first and for the remaining datasets Phases 1 and 2 removed many of the features that would have been otherwise removed in Phase 3.

The other feature engineering approach taken in Phase 3 was to create summary features composed of multiple source features. When possible, this allows for more information to be extracted from a smaller number of features. This process was used with both the Fundamentals dataset and the Stocks dataset. For the Fundamentals dataset, an example of this is the feature set that describes retained earnings. In this case, the three features were combined into a single summary feature.

### *Table3: Example of a Summary Feature from Fundamentals Dataset*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Retained** | **Notes** |
| rea | Retained Earnings - Restatement | Create summary feature and drop | Sum into a single retained earnings feature (“re\_total') |
| reajo | Retained Earnings - Other Adjustments |
| recta | Retained Earnings - Cumulative Translation Adjustment |

While this summarization process was used for all datasets, it was applied most extensively with the Stocks dataset. In fact, all features retained from the Stocks dataset were engineered in this way. This was the case because the dataset of daily stock performance was well positioned to glean more information from through summarization in order to measure fiscal performance over a prolonged period in a way that allowed for all companies to be compared. The primary incremental information in the Stocks dataset that is not provided by the other datasets is daily stock performance data. As discussed above, a primary objective of data processing and feature engineering was to condense all data into a single observation per company. Therefore, the challenge was to retain as much value as possible from daily data into a single aggregated observation. To do this, a single representative feature “Price - Close - Daily” was selected to be retained. “Price - Close - Daily” is considered to accurately represent the daily valuation of a company from stockholders’ perspective and captures any changes and volatility in stockholder opinion and pricing. The first step was to aggregate the daily reported prices and to summarize this feature by capturing the first, last, minimum, maximum, and standard deviation. From there, the following features were created:

### *Table4: Engineered Features from the Stocks Dataset*

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| st\_prc\_end | The last (most recent) value for stock closing price. |
| st\_per\_growth | The overall percent growth taken from the first and last closing prices. Contains negative values to show loss. |
| st\_per\_currentToMax | The percent difference from the current value to the all-time maximum value. |
| st\_per\_minToStart | The percent difference from the starting value to the all-time minimum value. |
| st\_volatility | A volatility measure in daily closing prices. Normalized as a percentage by dividing the standard deviation from the mean. |

At the end of Phase 4, there were 90 features remaining in total across all of the datasets. As described in the “[General Methodology](#_General_Methodology)” section above, in most cases, when the data was aggregated to a single observation per company, both median and standard deviation were retained. Therefore, these 90 features became 158 features in the overall prediction set. The majority of the retained features came from the Fundamentals dataset (74 of the 90 pre-aggregation), however at least one feature was retained from each of the source datasets.

### *Table5: Features by Dataset*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Global ID's/ Response Variables** | **Fundamentals** | **Ratings** | **Securities** | **Stocks** |
| Total # of Features | 3 | 146 | 1 | 6 | 5 |

Phase 4 was the final stage of the initial feature selection process and was also the first stage of the modeling portion of the analysis. Phase 4 was conducted separately for each Risk Assessment area (i.e., probability of lawsuit and magnitude of lawsuit). Therefore, Phase 4 will be described in the “*Additional Feature Selection*” section of each Risk Assessment area, below.

# **Risk Assessment: Probability of Lawsuit**

## Introduction and General Approach

The first risk assessment area predicts the probability that the BBC will face an SCA lawsuit. To make this prediction, data from the 347 other companies in the “Consumer Staples” sector were used to predict which of those 13.2% (46) companies would face an SCA lawsuit. The general approach for modeling was to consider many possible options for viable models for features and models to be used, to see which performed best using recall (primary) and accuracy (secondary). Feature selection has been largely addressed in the “[Specific Data & Feature Engineering Steps](#_Specific_Data_&)” section above, but the final stage of feature selection for this Risk Assessment area will be covered below (see “[Additional Feature Selection](#_Additional_Feature_Selection)”).

Following the same approach that was used for feature selection, for modeling, first many models were tested to see which had the most potential. The most promising models were then selected, refined, and reevaluated and a “winning” model (i.e., the one with the most potential measured by recall (primary) and accuracy(secondary)) was chosen.

## Additional Feature Selection

This stage of feature selection is referred to as “Phase 4”. In this phase, the remaining potential predictor features were evaluated using modeling techniques in order to identify which features would make good predictors. Two ensemble methods were applied, Random Forest and Gradient Boost, as they are particularly good at identifying feature importance. The recall statistic was used as the primary performance measure. Therefore, each of these models was tuned to optimize the recall statistic (more detail on tuning approach will be provided below as part of “[Model Selection](#_Model_Selection)”). Once the models were tuned for recall, they were fed all remaining features to show which were good predictors of whether a lawsuit would occur, as measured by Gini impurity decrease. In this case, Random Forest had a recall statistic twice that of Gradient Boost, and therefore Gradient Boost was dropped as a model to be used for feature selection.

Features identified using Random Forest, i.e., those with at least a 0.1 Gini impurity decrease were retained for model evaluation and all others were removed. One override was made to include “SP-Quality-Rating” (S&P Quality Rating - Current, from the Ratings dataset) so that there was at least one feature from all data sources in the model evaluation. The final feature set used for the model “Risk Assessment 1: Probability of a Lawsuit” includes 28 of the 162 original features (see [Figure2](#_Figure2_-_Features)).

### *Figure2 - Features Selected using Random Forest Model*

### 

## Model Selection

In order to choose the best model for predicting an SCA Lawsuit, an iterative process was used comparing many models in several rounds until the best model was identified. Models were trained with 80% of the total records. One challenge in working with this dataset was that there were few examples of companies which had been sued (positive class). Since the number of observations for the positive class were very low, modeling required class balancing to even out the positive and negative class for improved model training. Since the number of cases was already low, the approach to overcome this challenge was to create synthetic data to generate examples of companies who had been sued using an approach called SMOTE. This helped address the problem of having relatively few observations in that class. The models were then tested using a test dataset made up of a random sample of 20% of the overall records.

Overall model performance was measured with accuracy and recall scores, in addition to Receiver Operating Characteristic (ROC) curves with the Area Under the Curve (AUC) metric. Since there are few observations in the positive class (only 12 in the test set), and since the most important statistic is model accuracy in predicting this class, recall was the most important statistic because it is most sensitive to false negatives. In the later stages of model evaluation, where a very detailed comparison is called for, confusion matrices were also utilized.

In the first round, many models were selected for comparison and compared to see which had the most potential. The results of this phase are provided in [Table6.](#_Table6_-_SCA)

### *Table6 - SCA Lawsuit Prediction Model Performance for Round 1*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Consideration Reasoning** | **Acc.** | **Recall** | **Result** |
| Logistic Regression | Easy to interpret and uses regression techniques for classification | 0.614 | 0.416 | Evaluate Further |
| Random Forest (RF) Classifier | Ensemble method of many learners, that is generally a good classification model which provides clear insights into feature importance | 0.785 | 0.333 | Evaluate Further |
| Bagging Classifier | Similar to RF but all features are considered instead of random subset. | 0.714 | 0.250 | Eliminated: Overall Performance |
| Gradient Boosting Classifier | Ensemble method that is similar to RF, it is an attractive option for its ability to combine weak learners into strong learners | 0.742 | 0.333 | Eliminated: Similar Model to RF with lower performance |
| Ada Boost Classifier | Ensemble model that generally does well combining weak learners. Is sensitive to outlier data. (Add Citation) | 0.685 | 0.416 | Evaluate Further |
| K Neighbors Classifier | Simple/ intuitive model that offers a good comparison for the rest of the models by taking a different approach of making predictions based on the closest neighbors | 0.642 | 0.083 | Eliminated: Overall Performance |
| Linear Discriminant Analysis | A simple model that can be used provides easy to interpret results | 0.685 | 0.333 | Evaluate Further |
| Decision Tree Classifier | Very easy to interpret the results, meaning that it provides clear insights into why predictions were made | 0.700 | 0.333 | Evaluate Further |
| Gaussian NB | selected as a comparison model because the Naive Bayes model takes a different approach and assumes no relationships between features. | 0.357 | 0.750 | Evaluate Further |

In the second phase of model selection, each of the remaining models were tuned for recall. This was done using a grid search cross validation method that allows for many combinations of tuning hyperparameters to be run and to optimize the model for a specified statistic. The models were optimized for recall, which resulted in some accuracy scores decreasing compared to the first evaluation. For example, the accuracy of Random Forest decreased from 0.785 in the first round to 0.585 in the second round. However, the recall score increased from 0.333 to 0.416. This means that the overall accuracy of the model went down, but the accuracy in predicting the 12 positive class results improved. Given the specific methodology that this analysis followed, this is a favorable tradeoff. As noted in [Table7](#_Table7_-_SCA) below, after this phase only two models remained. However, in the final evaluation, a Voting Classifier will be added, and two models (ADA Boosting and GaussianNB) which were eliminated on their own would be utilized to see if they improve the results in the Voting Classifier.

### *Table7 - SCA Lawsuit Prediction Model Performance for Round 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall** | **AUC** | **Result** |
| Logistic Regression | 0.614 | 0.416 | 0.58 | Evaluate Further |
| Random Forest:  Tuning = Recall | 0.585 | 0.416 | 0.61 | Evaluate Further |
| LDA:  Tuning = Recall | 0.685 | 0.333 | 0.58 | Eliminated: Poor recall score. |
| Classification Tree:  Tuning = Recall | 0.628 | 0.250 | 0.49 | Eliminated: Tuning for recall, degraded performance. |
| ADA Boosting:  Tuning = Recall | 0.414 | 0.500 | 0.45 | Eliminated: Poor accuracy. Will experiment with as a member of the voting ensemble classifier |
| GaussianNB | 0.357 | 0.750 | 0.52 |

In the final phase of model selection, the remaining models were refined by once again revisiting selected features and fine-tuning the feature set to the specific model. As described in “[Additional Feature Selection](#_Additional_Feature_Selection)" above, for Random Forest, the feature importance function was once again used and any feature without at least a 0.1 Gini impurity decrease was removed. This only resulted in removing a single feature “rat\_spcsrc” (S&P Quality Rating, from the Ratings dataset).

For Logistic Regression, the refining process was more involved. Backward elimination was utilized, where the feature set was compared to the response variable (Boolean 0/1 for whether an SCA lawsuit occurred). The feature that had the least association with the response, as measured by the p-value, was eliminated. It was then removed from the overall set and the remaining features were rechecked to identify which had the highest p-value. This process continued until only the most significant features (measured by p-value) remained. Backward elimination resulted in removing 6 features from the predictor set ('at', 'lse', 'pi\_std', 'rest\_count\_of\_diffs', 'sstk\_std', 'xsga').

Once the feature set was refined for each of the models, the models were re-tuned for recall again using the gridsearchCV approach. Unfortunately, this did not improve the results for the models.

### *Table8 - SCA Lawsuit Prediction Model Performance for Round 3*

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall** | **AUC** |
| Logistic Regression | 0.542 | 0.416 | 0.58 |
| Random Forest | 0.542 | 0.416 | 0.53 |

Given the inability to improve the models with additional refinement alone and because no one model was clearly superior, the Voting Classifier model was used to determine whether predictions could be improved by tying the models together in a voting ensemble. In addition to evaluating Logistic Regression and Random Forest together, the remaining models which were eliminated on their own, but which had promising recall scores (GaussianNB and Ada Boosting Classification) were added to the ensemble. Many different combinations and weights were considered. Ultimately, a combination of Logistic Regression, Random Forest, GaussianNB, and ADA Boosting Classification, with Logistic Regression and Random Forest each weighted four times that of GaussianNB and ADA Boosting Classification, was identified as a good model.

### *Table9 - Final Model (Voting Classifier) Performance*

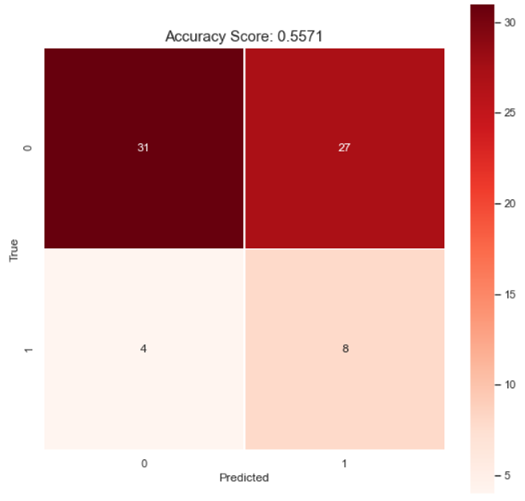
|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Recall** | **AUC** |
| **Voting Classifier Ensemble:**  Random Forest (4x weight),  Logistic Regression (4x weight),  GaussianNB (normal weight),  ADA Boosting (normal weight) | 0.571 | 0.667 | 0.59 |

## Final Model Evaluation

As noted above, the final model chosen was the Voting Classifier noted about due to its high recall score and overall accuracy. Other models were shown to have much higher accuracy, some in the 0.75 to 0.80 range, but this was almost always achieved to the detriment of a recall score. The reason for this is traced back to the class imbalance of the data. The training set was balanced using SMOTE to even out the positive and negative classes, but that method would be inappropriate for the test set, and therefore an imbalance remained. As a result, high accuracy could be achieved by nearly always predicting that a company would not face an SCA lawsuit.

This analysis instead prioritizes avoiding false negative results, and therefore was constructed to prefer predicting the positive class that a company would be sued. While it is true that the model over-predicts this event, it was done so for the reason that a company could better understand the possible risk and take corrective action. The final predictions from this model are shown in [Figure3](#_Figure3_-_Confusion)

### *Figure3 - Confusion Matrix from Voting Classification Model in Predicting SCA Lawsuit*



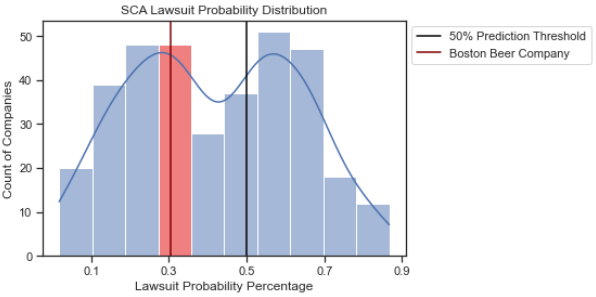
As shown in [Figure3](#_Figure3_-_Confusion) above, when the model predicts that a company will not be sued, it does so accurately, correctly making that prediction 88.6% of the time (31 out of 35 predicting the negative class). For the 12 companies in this test set that have faced an SCA lawsuit in this sector, 8 were correctly identified, which translates to the 66.7% recall score. Most of the model's inaccuracy can be attributed to the 27 companies who were predicted to face an SCA lawsuit but who did not. These companies account for 87% of the incorrect predictions in the model (27 out of 31 incorrect predictions). This reflects the cautious nature of the model and allows for being able say with confidence that when the model predicts that a company will not face SCA lawsuit they in fact will not.

One of the main challenges in producing this model was that there were few examples of companies which had been sued. Overcoming this challenge required creating synthetic data to generate examples of companies who had been sued. While this is a valid approach to the problem of sparse data, in the end the lack of true observations may have contributed to some limitations of the model which prevented higher performance.

## Risk Assessment Summary for Boston Beer Company

The model predicts that the Boston Beer Company (BBC) is more likely than not to face a SCA lawsuit, assessing that level of risk at 30.5%. As noted above, the model is designed to be cautious in making predictions. When the model predicts that a company will not be sued, as it has for the BBC, it is correct 88.6% of the time. The predicted probability of a SCA lawsuit puts BBC at the 34th percentile for risk level out of the companies in the “Consumer Staples” sector.

### *Figure4 – Risk Comparison for BBC with Other Companies in Consumer Staples Sector*



The major drawback of a Voting Classifier model comprising several models, and of some of the individual models, is that the results can be more difficult to interpret for one observation. This is the trade-off to the prediction improvements provided by the Voting Classifier. However, given that Random Forest has 40%[[3]](#footnote-3) of the voting share of the final model, interpreting the results from Random Forest gives a reasonable indication of the factors that led to the prediction that the BBC would not face a SCA lawsuit. [Table10](#_Table10_–_Most) below shows the features that had the largest impact on the predicted probability of a SCA lawsuit. The features highlighted in green were favorable in that they lowered the probability of a lawsuit, whereas the features in red did the opposite.

By far, the largest favorable factor was that over the observation period from 2010 to 2013, the value of BBC’s stock increased dramatically from $47.53 per share to $226.69 per share, and that the stock had been highly profitable for investors (*“Stock-Per-Growth” and “Earnings-Share-Diluted”*). Among other positive factors was that BBC has not filed a restatement as a correction to the annual 10-K reports (*“Number-Of-Restatements”*).

The factors which increased the probability of a SCA lawsuit included high capital and administrative expenditures, which were far greater than the median value for companies in the “Consumer Staples” sector (*“Capital-Expenditures” and “Administrative-Expense*”). Another factor, which is not practically controllable, is that companies with a higher gross profit tend to have a higher probability for a SCA lawsuit (“Gross-Profit”).

### *Table10 – Most Impactful Features for BBC’s Risk Assessment*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Impact to Prob.** | **Sector Median** | **BBC Value** | **Interpretation** |
| Stock-Per-Growth | -7 | 0.2 | 3.77 | BBC stock increased 377% from $47.53 to $226.69 |
| Earnings-Share-Diluted | -5 | 0.37 | 4.6 | Shows that BBC stock value has been highly profitable for investors |
| Funds-From-Ops(SD) | -2.8 | 4.61 | 2.01 | The amount of cash flow generated by business operations. The low S.D. shows there is less volatility in cash flow from core business. |
| Dilution-Avail(SD) | -2 | 10.38 | 8.69 | Lower S.D. than the sector median shows less volatility for BBC in net income. |
| Number-Of-Restatements | -1.3 | 1 | 0 | BBC has not filed a restatement |
| Capital-Expenditures | 1.6 | 39.18 | 147.82 | Shows that BBC has high expenditures on property, plant, and equipment |
| Sale-Of-Stock | 1.7 | 1.23 | 1.29 | Selling more stock than the median may indicate that seeking funding and decrease stock value. |
| Administrative-Expense | 2.2 | 60.19 | 210.11 | Administrative costs have increased from 174.849 to 270.262 between 2010 and 2013 |
| Stockholders-Equity-Total | 2.3 | 103.93 | 214.92 | Unavoidable risk that larger companies tend to be more prone to an SCA lawsuit |
| Gross-Profit | 2.9 | 102.8 | 319.24 | Unavoidable risk that larger companies tend to be more prone to an SCA lawsuit |

# **Severity Assessment: Potential Liability**

## Introduction and General Approach

The second assessment area predicts the severity of a potential SCA lawsuit in the form of a predicted settlement amount that the BBC would be liable for. To make this prediction, data from the same 347 other companies in the “Consumer Staples” sector was used, however as noted in the “[Risk Assessment: Probability of Lawsuit](#_Risk_Assessment:_Probability)” section, only 13.2% (46) companies had faced a lawsuit. Of those 46 companies, 35 (76.1%) did not have a settlement amount, most frequently due to a case that was on-going or dismissed, leaving only 11 (23.9%) companies for this modeling exercise.

Additionally, there was one company (Monster Beverage Company) who paid a significant settlement amount, presumably due to an extensive company history of lawsuits including SCA lawsuits in addition to various other class action lawsuits (CSPdailynews, 2014). The settlement amount for this company was an extreme outlier, and needed to be removed, leaving only 10 (21.7%) companies for the modeling. Therefore, the modeling exercise was forced to only use 10 of the 347 companies (2.8%) in the “Consumer Staples” sector. This was a limitation for severity modeling, and it should be noted that the lack of observations impacts the accuracy of the model, including for predictions made for the BBC.

With the accuracy limitation noted, the same general approach for modeling was followed. First, many prospective features and models were considered. The severity assessment utilized the same feature selection steps previously addressed in the “[Specific Data & Feature Engineering Steps](#_Specific_Data_&)” section above, but the final stage of feature selection for this Risk Assessment area will be addressed below (see “[Additional Feature Selection](#_Additional_Feature_Selection)”).

Following the same model selection approach that was used for the “[Risk Assessment: Probability of Lawsuit](#_Risk_Assessment:_Probability)” section, first many models were tested to see which had the most potential, with the “winning” model (i.e., the one with the most potential as measured by RMSE) chosen from that evaluation.

## Additional Feature Selection

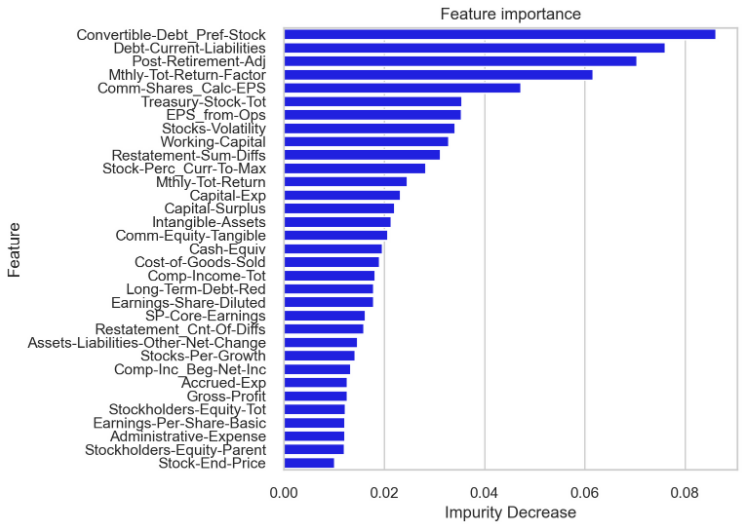
Similar to the [Risk Assessment: Probability of Lawsuit](#_Risk_Assessment:_Probability) section, a fourth phase of feature selection was conducted for the severity assessment. In this phase, the remaining potential predictor features were evaluated to select the best predictors for the severity of a potential SCA lawsuit.

In the [Risk Assessment: Probability of Lawsuit](#_Risk_Assessment:_Probability) section, the modeling was classification based, predicting one of two possible outcomes. This is worth noting because oftentimes in the classification model the version of a feature using standard deviation from aggregation proved to be a better predictor than the same feature using median from aggregation. In this assessment, the modeling is regression based to predict a continuous variable. Given that, the standard deviation is less useful to predict an increase or decrease in severity. For this reason, all the standard deviation feature versions were dropped, and the median version was retained.

Next, the feature set was evaluated by checking the correlation with the response variable, and features with a very low correlation were removed. After this, a backward elimination was attempted, but ultimately not utilized. The reason that this approach was not used is due to the data itself. After much analysis, it was discovered that a problem with backward elimination can occur when using a dataset that is made up of a small number of observations, many potential features, and especially features that have a high degree of collinearity. (StatsModel/GitHub) All of those conditions exist in this dataset, with the collinearity stemming from the data containing many features that ultimately roll up to measures of a company's financial performance. For this reason, backwards elimination was not utilized.

Instead, much like the first assessment, the features were evaluated using the Random Forest. All remaining features were fed to a Random Forest Regression model and tuned to minimize the Mean Squared Error. From there, feature selection used the feature importance function, retaining features showing a Gini impurity decrease over 0.1.

### *Figure5 - Features Selected using Random Forest Model*



The final step in this phase of feature selection was to check the feature collinearity. As has been previously stated, since all features in some way are measures of a company's financial performance, there is an inherent degree of collinearity. The process taken was to review the feature pairs with the highest degree of collinearity and remove the one with the lower Gini impurity decrease. (see: [Feature Collinearity Heat Map](#_Feature_Collinearity_Heat)) For reasons that will be discussed in the modeling section below, this process was ultimately abandoned and the final feature set used was the one described above from the Random Forest feature importance method.

## Model Selection

The primary challenge in this modeling exercise was the small number of observations. As was previously stated, there were only 11 companies to use as observations for model building, with one of those needing to be removed as an outlier. The implications of building models with only 10 observations are that there is a high degree of error and uncertainty in the models. This is important to keep in mind throughout the modeling and scoring sections of this report and will ultimately be expressed in the form of a final estimated 90% confidence interval for the potential severity for the BBC with a wide range.

Like modeling a prediction of an SCA lawsuit, for modeling the severity of a potential SCA lawsuit an iterative process was used comparing many models until the best model was identified. Given the low number of observations, the models were not able to build and test with the traditional 80%/20% split of training/test records, and instead a 60%/40% split was made so that there were four test records. Overall model performance was measured with RMSE, which provides a more interpretable form of the mean squared error (MSE).

In the initial model evaluation, many models were selected for comparison. Due to the fewer number of models considered than in the previous assessment, and due to these models, all having more unique set up requirements than the classification models considered in the last assessment, more time was spent tuning the models in this initial evaluation. As was mentioned in the feature selection section above, one aspect of model building was removing features that had a high degree of collinearity. However, since the models that were part of the initial evaluation individually may perform better or worse with this step taken, all models were tested with and without collinear features. In the end, the models selected to evaluate further (Random Forest (RF) Regressor and Partial Least Squares (PLS)) were models that can handle collinear features and performed best with them included. Therefore, the step to drop collinear features was not pursued and the model evaluation table shown below ([Table11](#_Table11_-_SCA)) is based on the feature set from the Random Forest feature importance step. The results of the initial model evaluation are provided in [Table11](#_Table11_-_SCA) as follows:

### *Table11 - SCA Lawsuit Prediction Model Performance for Round 1*

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Consideration Reasoning | RMSE | Result |
| Multiple  Regression | Simple model that is easy to interpret, but prone to overfitting. Requires extensive feature selection to avoid collinearity | 10,085,443 | Eliminated |
| Ridge Regression | Typically, better than Multiple-Regression at avoiding overfitting. | 16,199,201 | Eliminated |
| Lasso Ridge | Has feature selection as part of the model building process. Also good at avoiding fitting. | 34,856,804 | Eliminated |
| PLS Regression | Generally, can perform well with a high number of collinear features, and with few observations. | 5,036,729 | Evaluate Further |
| Random Forest Regression | An ensemble method based on building many trees from a random selection of features means that it can handle many features and determine which are best for prediction. | 5,508,572 | Evaluate Further |
| Bagging | Another ensemble method with many similarities to Random Forest. Is generally better than Random Forest at avoiding overfitting but comes at the cost of interpretability. | 5,858,117 | Eliminated |

The second model evaluation aimed to improve the selected model performance by fine tuning hyperparameters, and for the RF Regressor further reducing the feature set using another round of the feature importance function. Unfortunately, this change degraded the performance for Random Forest. PLS was not able to be improved upon from the initial evaluation, but remained the best performing model, and was selected as the final model.

### *Table12 - SCA Lawsuit Prediction Model Performance for Round 2*

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Result** |
| PLS Regression  (Round2) | 5,036,729 | Selected |
| RF Regressor  (Round2 - Reduced Feature Set) | 5,617,928 | Eliminated |

## Final Model Evaluation

The PLS Regression was selected as the best model because it produced the lowest RMSE by a considerable margin. Given the model’s strength with sparse observations and collinear data, it is not surprising that this model outperformed the rest. Recall that this model was training with few observations and therefore even though it was the best performing model, the RMSE is still relatively high and confidence in the overall accuracy somewhat low. In other words, should more data be collected in the future, the model selection steps should be revisited.

## Severity Assessment Summary for Boston Beer Company

The selected model predicts that should the BBC face an SCA lawsuit which is not dismissed, that the projected severity would be $6.5 million. This figure is solely the predicted settlement amount and does not include legal fees and other expenses related to defending the company from an SCA lawsuit.

Due to the small size of the dataset, this point estimate was translated into a 90% confidence score showing that the model predicts with 90% confidence that settlement amount would be between $1.3 and $11.7 million. Of the set of 5 (4 test records plus BBC), BBC is the median value and is plotted according to the point estimate. The plot below shows how the predicted value aligns with the other predictions (from the test set only) made by the model.

*Figure6 – Predicted vs. Actual Value for Test Set and BBC*

Chart, scatter chart

Description automatically generated

For the plot above, given that the actual value for BBC is not known, the predicted point estimate was used as the actual value placing BBC on the correct prediction line. However, it is known that this estimate has a margin of error, which is illustrated by the other predictions over and under the correct prediction. On this limited test set, the model was equally divided in over and under predicting a settlement amount. This shows why it is more important to consider the entire confidence range provided instead of simply using the point estimate.

# **Appendix**

## Citations

Banasiewicz, A.D. (2015) The ecosystem of executive threats: A conceptual overview. Risk Management,

Vol 17 no. 2, 2015.

Banasiewicz, A.D. (2015). Quantifying executive threats: shareholder litigation

Int. J. Business Competition and Growth, Vol. 4, Nos. 1/2, 2015

Monster Settles 6-Year Lawsuit with Investors. (2014, Apr. 21). CSPdailynews.com. Retrieved April, 30 2022 from

<https://www.cspdailynews.com/beverages/monster-settles-6-year-lawsuit-investors>

“Jolespin”. (2018) OLS results return NaN for P>|t| field with larger dataset. StatsModel/ Github. <https://github.com/statsmodels/statsmodels/issues/4831>

Thailappan, D (2021, June 01) AdaBoost : A Brief Introduction to Ensemble Learning. Analytics Vidhya.

<https://www.analyticsvidhya.com/blog/2021/06/adaboost-a-brief-introduction-to-ensemble-learning/#:~:text=Advantages%20and%20disadvantages,-Coming%20to%20the&text=The%20accuracy%20of%20weak%20classifiers,it%20needs%20a%20quality%20dataset>.

## Additional Files

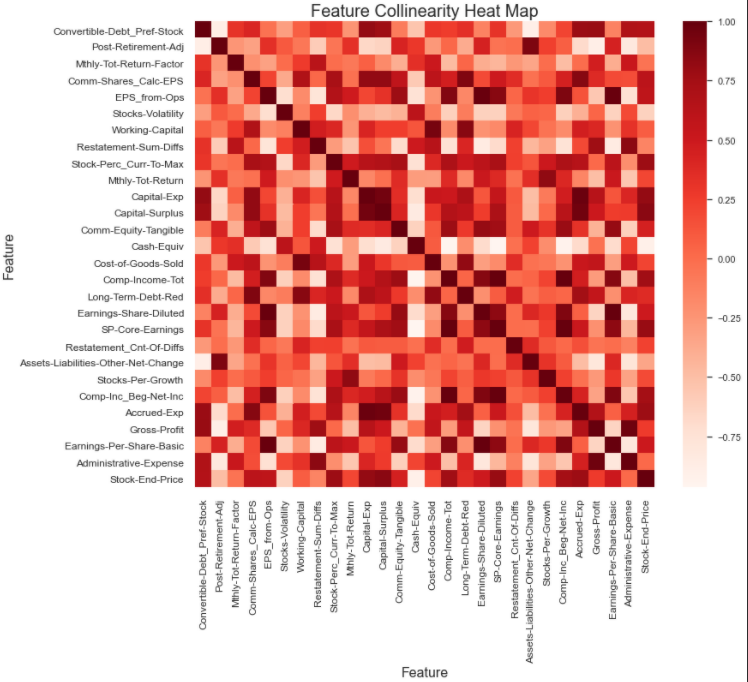
1. Data used for modeling (Compiled Data described in report): “Clean\_Modeling\_Data.xslx”
2. Source Code (Python in notebook format):
   1. Data Clean-Up/Feature Engineering: “Capstone-Date-Prep.ipynb”
   2. Risk Assessment: Probability of Lawsuit: “Capstone\_LS\_Modeling.ipynb”
   3. Severity Assessment: Potential Liability: “Capstone\_Severity\_Modeling.ipynb”

## Code Repository

https://github.com/bcullinan32/CapstoneProject

## 

## Feature Collinearity Heat Map



1. Median was selected over mean due to the skewness of the data. [↑](#footnote-ref-1)
2. Z-score measures the number of standard deviations from the median. [↑](#footnote-ref-2)
3. The voting classifier model gives Random Forest and Logistic regression each 40% of the vote, and GaussianNB and ADA Boost each get 10% of the vote. [↑](#footnote-ref-3)