DSEG 6311 - Capstone Project

Executive Risk – Crane Co.

Matthew Purvis

Fall 2021

Table of contents

[Executive Summary 2](#_Toc87385784)

[**Likelihood Estimation** 2](#_Toc87385785)

[**Severity Estimation** 2](#_Toc87385786)

[Data and Approach 3](#_Toc87385787)

[**Summary** 3](#_Toc87385788)

[**Initial Dataset Restructuring** 3](#_Toc87385789)

[**Assessing the Variables to Use** 3](#_Toc87385790)

[**Aggregation of Multi-Year Reporting Data** 4](#_Toc87385791)

[**Review the Overall Population** 4](#_Toc87385792)

[**Weighting Variables** 5](#_Toc87385793)

[**Outlier Detection** 5](#_Toc87385794)

[Likelihood Assessment 6](#_Toc87385795)

[**Summary** 6](#_Toc87385796)

[**Pre-Modeling Steps** 6](#_Toc87385797)

[**Training the Models** 7](#_Toc87385798)

[**Testing the Models** 7](#_Toc87385799)

[Severity Risk Assessment 10](#_Toc87385800)

[Noteworthy Considerations 10](#_Toc87385801)

[Contributing Content: Industry Benchmarking 11](#_Toc87385802)

[**Likelihood** 11](#_Toc87385803)

[**Severity** 11](#_Toc87385804)

[Conclusion 11](#_Toc87385805)

[References 12](#_Toc87385806)

[Appendix A – Base Dataset Variables 13](#_Toc87385807)

[Appendix B – Binary and Multi-value Variables 14](#_Toc87385808)

[Appendix C – Initial Correlation Plot 15](#_Toc87385809)

[Appendix D – Severity Assessment Data – Before Outliers (in $ millions) 16](#_Toc87385810)

[Appendix E – Severity Assessment Data – After Outliers (in $ millions) 17](#_Toc87385811)

[Appendix F – Severity Assessment – Data Inputs and Calculation 18](#_Toc87385812)

[Appendix G – Important Features – Random Forest Model with SMOTE and hyper parameter tuning 19](#_Toc87385813)

[Appendix H – Cornerstone Research – Number of Filings by Industry2 20](#_Toc87385814)

# **Executive Summary**

Executive risk can be defined as the risk to a company being noncompliant with applicable laws and regulations relating to the timeliness, completeness, and accuracy of disclosures of past performance and future prospects, which for the context of this report result in shareholder class action litigation. This report provides an assessment of the level executive risk for Crane Co., which includes estimating and substantiating the likelihood and monetary impact (severity) of a shareholder class action litigation if it were to occur. The purposes of estimating the likelihood and severity are to provide management with a way to evaluate if the cost of insurance to defend Crane Co. in a shareholder class action suit is justifiable. Using publicly available data that includes financially reported and amended values from companies’ annual 10-k filings and historical settlement data in the Industrials GIC industry, below is a summary of the resulting findings.

## **Likelihood Estimation**

This report takes and analytical approach to obtain a probability estimate by using the publicly available data, performing many data transformation steps, and ultimately determining the appropriate data feature engineering steps prior to testing and evaluating the best model for predicting the ultimate probability estimate. It is important to note that many considerations were taken to establish variables that may be driving the likelihood of a shareholder suit occurring for companies throughout this process, which include potentially volatile reporting of key financial values, as well as significantly different restated financial values reported by entities in their respective annual 10-k filings. These considerations take into account the overall goal of the analysis, which directly tie to the timeliness, completeness and accuracy over the disclosure of past performance as reported by these companies.

Some of the pre-modeling approaches taken include splitting the data into training and test splits, as well as using synthetic minority oversampling (SMOTE) and hyper parameter tuning to optimize the overall efficacy of the model. The most effective model was a random forest model using SMOTE and hyper parameter tuning, which predicted the overall probability for Crane Co. to experience a shareholder class action suit to be at 12.1%. This was driven primarily by the volatility in the size of a company’s workforce (i.e., the number of reported employees), the volatility of reported versus restated working capital, and other important measures of a company’s overall financial health, specifically its reported return on total assets and return on total equity. The process used to substantiate the above probability estimate is covered in more detail in the report below.

## **Severity Estimation**

A straightforward method was used to estimate the severity in terms of total dollars (expressed within a 95% confidence interval) for Crane Co. in the event a shareholder class action suit was to occur and result in a settlement. It is important to state that a major factor for determining the overall risk to Crane Co. was to use the mean market value of the companies from the historical SCA filings and settlements data who had settled shareholder class actions. Based on this, using Crane Co.’s market value of $3,913 million, the overall financial risk to Crane Co. was determined to range, with 95% confidence, between $9.78 and $31.88 million. The process used to calculate and substantiate the severity estimate is contained the report below.

All additional work can be found on the following jupyter notebook provided with the reports:

* *CapStone\_01\_DataCleanup\_EDA.ipynb – Data preparation and transformation*
* *CapStone\_02a\_Likelihood\_Modeling.ipynb – Pre-modeling, training and testing models, post-modeling evaluations*
* *Capstone\_02b\_Likelihood\_SupportingVisuals.ipynb – Supporting visualizations used to analyze important features captured in prediction*
* *CapStone\_03\_Modeling\_Severity.ipynb – Severity analysis process with supporting visualizations*

# **Data and Approach**

## **Summary**

Before any analysis and modeling can be completed, an approach of how to use the data must be completed, as this is the strategic approach that ultimately creates the story for the overall results of the analysis. The datasets used in these reports consists of the fundamentals and SCA filings and settlements dataset. The fundamentals dataset contains financial data related to publicly traded company’s’ annual 10-k filings, which include any restated filing amounts where applicable. The SCA filings and settlements dataset provide a list of companies that have been named as a defendant in a shareholder class action suit, as well as the amount each company has paid out in damages for a settlement or court determination, where applicable (no payment would be provided if no settlement occurred or if the suit was dismissed by the court).

## **Initial Dataset Restructuring**

The first step in data preparation was to restructure the dataset to differentiate amounts stated in a company’s original 10-k filing with the restated amounts (if any). This resulted in the duplication of each numeric column where a restatement could potentially occur. The way in which this was completed was to use the use the ‘*datafmt’* column, where “STD” signifies amounts reported in a company’s originally filed 10-k and “SUMM\_STD” signifies amounts reported because of a restated filing. The “STD” and “SUMM\_STD” values were added to the prefix of each numeric column to avoid confusion of where the numeric values were derived (e.g., ‘*STD\_at’* = total assets associated with an originally filed 10-k; *‘SUMM\_STD\_at’* = total assets associated with a restated filing). The original dataset consisted of 3,974 rows and ultimately was reduced to 2,057, which highlighted that 1,917 of the rows were associated with a re-stated filing. As a final step to this first step, a ‘*restatementflag*’ column was generated for 1,917 of the final 2,057 columns to identify when company have some sort of restatement occur for a given quarter.

## **Assessing the Variables to Use**

One of the strategic paths of this analysis was to focus primarily on columns in which a restated amount had been adjusted. The result of filtering out the use of these columns is to isolate the potential association between 10-k restatements and shareholder class actions suits filed. As a result, of the 900 plus columns, 32 columns showed to have had a restated amounts occur within this sample of company filings. The 32 columns included very key financial values, such as net income, cost of goods sold, sales, total assets, total equities, etc. In addition to these columns, some other key financial columns not related to restatements were used, as well as creating some manually generated columns to potentially create additional value, which include the *suitflag*, *roa*, and *roe*, and *dte* columns. *See Appendix A –Dataset Variables for a full list of the variables used and their associated definitions.*

It is important to note that by assessing these columns for use removed the need to do an extensive review of which columns should be removed due to being primarily null and not useful. However, a null review for missing values was completed with only two rows showing as nulls. These two rows were removed, as they did not materially alter the overall results of the analysis.

## **Aggregation of Multi-Year Reporting Data**

The next step in the analysis is to take a quantitative approach by determining the values that potentially get used in the model run. When determining the values used, it is important to first consider the overall goal of this analysis, which is to predict the likelihood a shareholder class action suit will occur and estimate the potential severity a shareholder class action suit entails due to untimely, incomplete, or inaccurate disclosures of past and future performance. Some questions this analysis considers to achieving the overall goal include:

* How volatile are the variances between the originally stated values and the restated values for each company over time?
* How volatile are the variances as a percentage of the originally reported value, as this incorporates additional relativity for large swings in reported values (i.e. over-reporting $100 in sales when sales are $1,000 (10% variance) versus $10,000 (1% variance) have significantly different impacts to a company).
* How volatile are the financially reported values for each company over time?

Using the questions above, the base dataset variables were labeled for each. Below are the prefixes/ suffixes used for the ‘*at’* variable *(Total Assets*), as these are the variables ran in the final model:

1. *‘StdDev\_at’* – Standard deviation of reported Total Assets values over time
2. *‘Vol\_at\_variance’* – Absolute value of the MAX or MIN variance over time
3. ‘*at\_PercentChange’* - Absolute value of the MAX or MIN variance over time expressed as percent of the total of the reported value

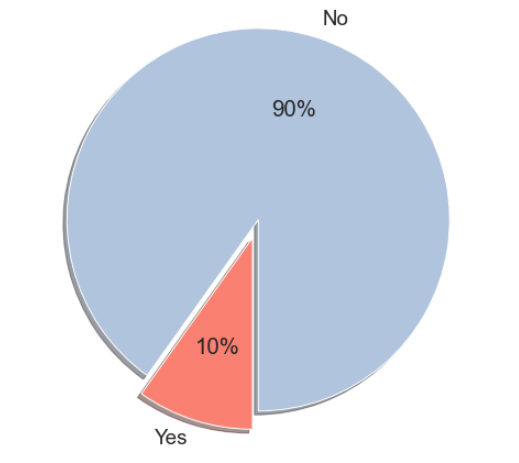
Some other tactics were taken but ultimately not used in the final model runs. For example, the mean reported values for each company over time were taken using the same logic in the third question above. However, the mean for each column was extremely correlated with the standard deviation values generated so were ultimately filtered out of the analysis prior to running the model. Another tactic used was to take the most recently filed values for each company. However, these values did not trigger significantly relevant results when attempting the model run, so were also not used in the final model runs.

## **Review the Overall Population**

Up to this point, the dataset has been transformed to contain each individual company on a row-by-row basis with the relevant categorical and numeric variables included. The next steps are to step back and ask the following questions:

* What percent of the total population has a suit been filed?

**Figure 1**: Percent of companies who have had a suit filed against them



* Is the existing ratio of companies in which a suit has been filed to the entire population sufficient to be able for a predictive or analytic model to achieve an outcome that provides insight and aids in the support of the overall goal of this analysis?

Figure 1 to the right displays the answer to the first question, which indicates that only 10% (57) of the 578 companies have had a shareholder suit filed against them during the time provided from the datasets used for this analysis. This is too low to provide a sufficient predictive analysis, so two approaches were considered. The approaches included:

* Taking a random sample of the remaining companies to remove any bias from the total population
* Using SMOTE to synthetically increase the number of companies in which a suit was filed, which raises the ‘Yes’ values to a 50% / 50% split with the No values

The latter approach was ultimately used in the modeling and will be addressed in the upcoming *Likelihood* section of the report. The next step from here is to assess the quantitative data by weighting the variables to be scaled appropriately for the model run and assessing what actions need to be taken for the outliers that are present.

## **Weighting Variables**

The scale for the numeric values is a very important consideration for this analysis. With the sample size being relatively small and the potential for significant differences in values amongst many different numeric continuous variables, it is important to weight the variables to an “apples to apples” comparison. This is completed, simply by the following formula for each column:

***Z-score*** *= Absolute Value( (Variable – Variable Mean) / Variable Standard Deviation )*

Each variable has now been converted to its relative Z-score, which specifically values each variable by how many standard deviations it is from its respective mean. For example, if the *‘Vol\_roe\_variance’* (calculated roe variance of originally stated and restated roe) equals 1, then this number is 1 standard deviation from the entire variable’s mean. This weighting technique helps with a couple things. It helps by reducing the significant swings between companies that may have substantially different values, which could ultimately impact the outlier analysis and filter out key companies that would help with the overall modeling. It also helps play a role in how the variables will interact and correlate when running the model.

## **Outlier Detection**

The final step before beginning the modeling portion of this analysis is to detect the existing outliers and decide what steps to take. Since the variables have been weighted using z-scores and are converted to how many standard deviations they are away from their respective mean, the first step in detecting the outliers is to isolate any values greater than 3 standard deviations from the mean. Once detected, two options are considered, which include:

1. Filtering the outliers from the sample where greater than 3 standard deviations from the mean
2. Capping the max z-score at 3, converting any value greater than 3 to equal 3

It was ultimately decided to cap the z-score to a maximum of 3 for two reasons. First, by capping the value, additional features from the company with an outlier may provide value and be relevant to the modeled results. Second, it would further reduce the sample, making the model potentially less reliable. These final steps conclude the initial approaches to transform the dataset for the model runs in the upcoming section. It is important to note that *‘gvkey’* column is how the modeled results and prediction outcomes will be traced back to the original dataset when final model interpretations are reviewed.

# **Likelihood Assessment**

## **Summary**

This section will highlight the steps taken to estimate and substantiate the likelihood of a shareholders class action litigation if it were to occur. Below are, at a high level, the steps taken:

1. Complete pre-modeling tasks, which include analyzing the binary and multi-value variables, assessing the correlations amongst the variables, and splitting the dataset into two datasets, train and test
2. Run the model on the training data, incorporating both the synthetic minority oversampling technique (SMOTE) and hyper parameter tuning to boost performance
3. Run the model on the test data, incorporating the same hyper parameter and SMOTE at the training data
4. Evaluate and select the model to use for this analysis
5. Using the selected model, predict the probability of a shareholder class action suit to occur for Crane Co.

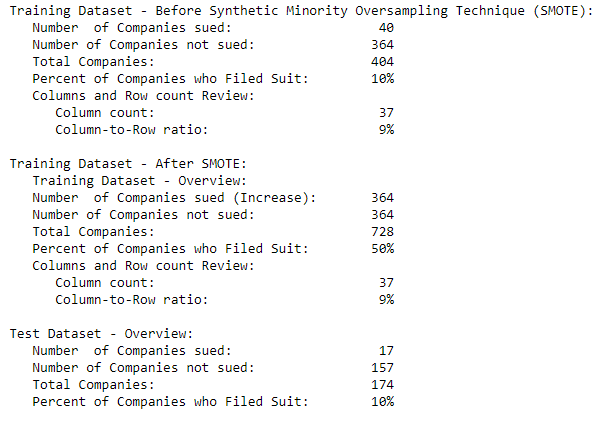
## **Pre-Modeling Steps**

The first step in preparing the data for the model is to ensure the binary and multi-value columns are converted to values that the model can interpret. All binary columns are converted to a format where the value is either 0 or 1. Dummy variables were generated for the multi-value columns. *See Appendix B – Binary and Multi-value Variables for variables included in the conversion.*

The next step in the pre-modeling process was to review the correlations between the numeric variables to ensure optimal results are modeled. It was determined to remove selected variables with a correlation greater than 65%. The purpose of removing the correlated variables is to prevent poor model performance. When assessing the highly correlated variables, there were several variables that had a significant correlation and were removed from the analysis. *See Appendix C – Initial Correlation Plot for overview of results.*

The method used to test the models used in this analysis is the train test split method. This split the dataset up into 404 and 1 74 companies in the training and test datasets, respectively, which was a 70% / 30% split. The split resulted in a total of 10% of companies having a positive suit indication for both the test and train splits. However, regarding the training data, due to the limitation of the overall row count (404 total rows) and low percentage of rows where there was an indication of a company being sued (10%), obtaining optimal results in the modeling was a challenge. To increase the row count and the number of companies in which a suit was filed, SMOTE was utilized. This resulted in an increased positive indication rate for a suit to be filed to from 10% (40 companies) to 50% (364 companies), which resulted in a new dataset with a total of 728 companies. To the right in *Figure 2* is a summary of the test and train datasets using SMOTE.

**Figure 2:** Overview of split datasets, which includes impacts before and after SMOTE utilized



## **Training the Models**

A total of 5 models were used for prediction and trained using either the original training dataset or the training dataset using SMOTE. The original training dataset (404 companies) was treated as a baseline dataset to assess the performance improvement of the modeling using SMOTE. In addition to using SMOTE, for both the logistic regression and the random forest models, hyper parameter tuning was conducted using grid search. The purposes of the hyper parameter tuning were to run iterations of the modeling using many combinations of the potential parameters available and assessing the model the performs the best. Below are each of the model runs used for this analysis:

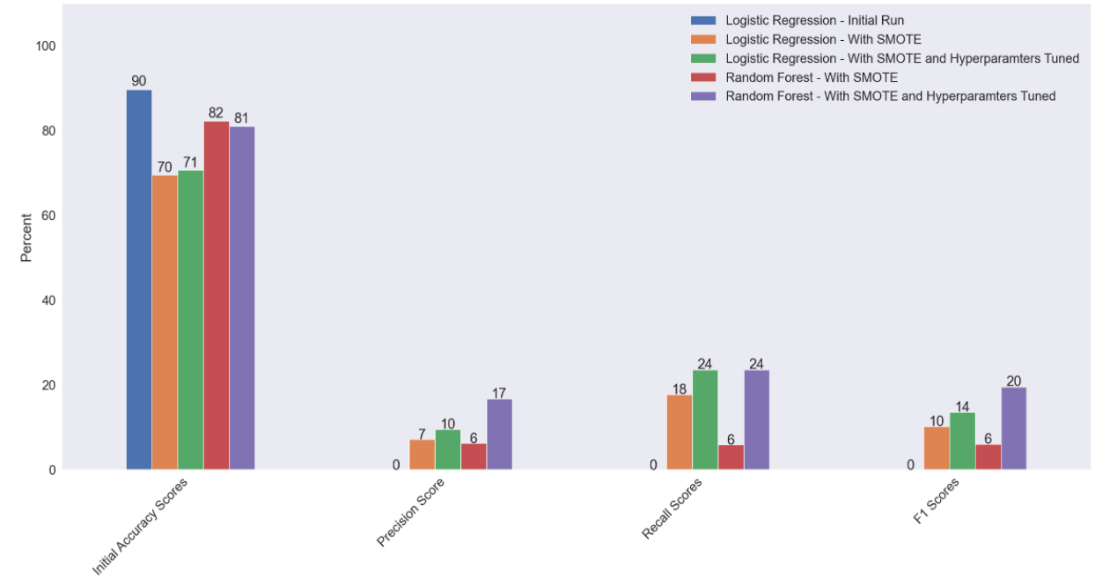
1. Logistic Regression – Original training dataset
2. Logistic Regression – using SMOTE
3. Logistic Regression – using SMOTE and Hyper parameter tuning
4. Random Forest Classifier – using SMOTE
5. Random Forest Classifier – using SMOTE and Hyper parameter tuning

## **Testing the Models**

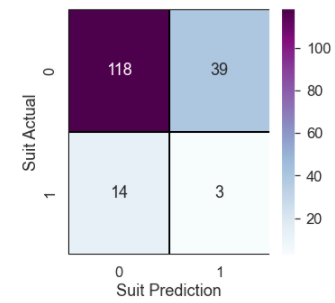
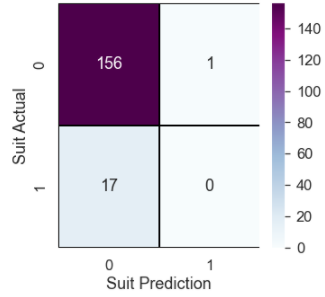
Figure 3 on the following page displays the overall results of the models using the testing data. This visual provides a simple view at a high level of the overall results of the modeling conducted on the testing data. Below are some of the observations made that impact the decision on which models will be selected and reviewed for validation:

* The initial logistic regression model run, which did not include the use of SMOTE resulted in the least reliable model, as the precision, recall and overall F1 scores resulted in 0
* Both models using tuned hyper parameters achieved better results in virtually all categories than models with no hyper parameter tuning
* Although the recall rate is equivalent for both the logistic regression and random forest models using SMOTE and hyper parameter tunings, the random forest model achieved better overall accuracy and F1 scores

**Figure 3:** Results of modeling using the test dataset



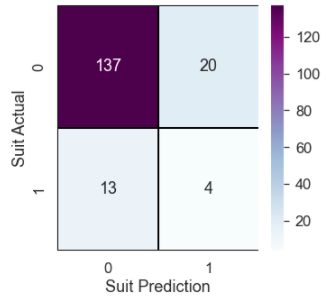
For the purposes of assessing the overall model performance, most of the assessment will be focusing on the recall, which measures the rate (in percent) a company was predicted to be sued versus the total number of companies that were actually sued. Below in figure 4 are the results of the first two model runs, the initial logistic regression model run (*left*) and the logistic regression model run with SMOTE (*right*). The logistic regression model ran significantly better with SMOTE rather than without, resulting with 17.7% accuracy (3 out of 17 companies predicted to be sued and were sued), while the original model resulted in 0.0% accuracy (*highlighted in red*).



**Figure 4:** Logistic Regression Model Confusion Matrices Without SMOTE (left) and With SMOTE (right)

1 = ‘Yes’; 0 = ‘No’

The best overall performing model was the random forest model using SMOTE and hyper parameter tuning. In addition to the recall being equivalent to the most effective logistic regression model (24%), the overall performance of the model (i.e., precision and accuracy) was also better than the logistic regression results. The random forest model with hyper tuned parameters and SMOTE also outperformed the initial random forest model with SMOTE. Below in *Figure 5* are the results of the random forest model using SMOTE with hyper tuned parameters. The overall accuracy of the model was 81%. With a recall of 24 %,, this outperforms the other random forest model by approximately 6%, along with having better precision and F1 scores.



**Figure5:** Best Performing Model – Random Forest with SMOTE

1 = ‘Yes’; 0 = ‘No’

The variables, which can be referenced in *Appendix G*, that made up the majority of the predictive power relate with a mix of variables related to the volatility of reported values over time calculated using the standard deviation, as well as being related to specific variances between reported and restated values. Based on these important predictor variables for companies who have had a shareholder class action suit filed, my model predicts that there is a 12.1% likelihood for Crane Co. to be named in one of these suits, which is driven by a number of features that are tied to the size of a company’s workforce and its overall financial health. Some of the key drivers found are included below (*See ‘Appendix G for a full list of the predictor variables obtained as a result of the modeling*):

* Volatility in the number of reported employees with a company over time was the most significant predictor
* Significant volatility in the value of originally reported versus restated working capital (wcap) expressed in dollars and as a percent to the originally reported value; Working capital is directly related with short-term liquidity
* Volatility in key financial metrics such as return on assets and return on equity over time, which relate specifically to the relationship between net income and total assets/ total equity
* Other variables with predictive power include:
  + The volatility of common shares used to calculate earnings per share (cshfd)
  + The volatility of extraordinary items and discontinued operations (xido)
  + Significant variances in the originally reported cost of goods sold value versus the restated value (cogs)

*The pre-modeling work and model evaluations can be found in the following notebook: CapStone\_02a\_Likelihood\_Modeling.ipynb.*

*Additional supporting visuals assessing predictor variables can be found in the following notebook: Capstone\_02b\_Likelihood\_SupportingVisuals.ipynb.*

# **Severity Risk Assessment**

The first step in the process to determine a severity estimate was to join the settlement amount field from the SCA filings and settlements dataset to the final transformed dataset that provides the companies in a row-by-row basis. The settlement amount field was merged to the existing dataset and assessed against the market value field provided in the fundamentals dataset. It is important to note that there was only a total of 15 settlement amounts to use for this analysis. Below is a summary of the steps taken to calculate the overall severity risk for Crane Co. (*see Appendix D – F for additional context and visualizations)*:

1. Review the overall data, which is displays the settlement amounts and market values by company (15 in total)
2. Filter outlier companies based on market value (two in total) from the dataset that will skew the results if estimated (*see appendix D and E for a view at the before and after view of the distribution of values when removing the outliers)*
3. For the remaining 13 companies:
   1. Calculate the mean of the market value ($1,460.54 million)
   2. Calculate the mean and standard error of the settlement amount ($7.78 and $2.10 million, respectively)
   3. Establish a confidence level of 95% using 1.96 as the confidence interval
4. Determine upper and lower bounds by using the following calculation, which resulted in a lower and upper bound of $3.65 and $11.90 million, respectively:

Lower Bounds = Mean settlement amount – (1.96 \* Std Error of the Mean)  
Upper Bounds = Mean settlement amount + (1.96 \* Std Error of the Mean)

1. Establish Crane Co.’s market value as the target variable ($3,912.94 million)
2. Calculate the target variable as a percent of the mean market value (268%)
3. Calculate the overall upper and lower-level exposure risk for Crane Co. to assume  
   ($15.2 and $31.88 million, respectively)

*\*Note: More detail on the calculation can be found in Appendix F, and the supporting code/ visualizations can be found in the following notebook provided with the submission: ‘CapStone\_03\_Modeling\_Severity’.*

# **Noteworthy Considerations**

Although the model output does provide an estimated likelihood that Crane Co. will have a shareholder class action file, it is important to note that the modeled results are based on the best results obtained but are imperfect and additional research should be conducted before any consideration is taken into making decisions solely off the output of the generated results. Some of the limitations with the model to consider include:

1. The model does contain a limited sample size and is based on historical behavior. Based on this, with limited historical data collected, future behavior has a potential to change significantly in a way the model cannot predict.
2. The model does not consider the wide array of business models and strategies for the companies within the industry, which can make company-to-company comparisons less reliable.
3. The model uses synthetic minority oversampling (SMOTE), which generates synthetic samples for the minority class, which in this case is the number of companies in the population in which a suit has been filed against. This means the companies used in the analysis run the risk to inaccurately reflect actual competitors or market conditions.

# **Contributing Content: Industry Benchmarking**

## **Likelihood**

According to Cornerstone Research’s ‘*Securities Class Action Settlement 2020 Year in Reviews*’ report, there has been, on average, 17 class action filings occur each year within the industrials GIC industry2 (*see Appendix H for the table obtained from Cornerstone Research’s report to reference).* There was a total of 578 companies provided in the original fundamentals dataset in the Industrials GIC industry. When considering the overall activity of filings as a percent to total, this would indicate that approximately 3% (17 / 578) of companies in this industry have had suits filed against them annually on average from 1997 to 2019. The prediction obtained from the random forest model does seem relatively high when benchmarked with the overall industry actvity.

## **Severity**

According to Cornerstone Research’s ‘*Securities Class Action Settlement 2020 Review and Analysis*’ report, the median settlement in 2020 was $10.1 million, while the average settlement doubled from $27.8 million in 2019 to $54.5 million in 2020. The average was driven by six mega settlements, which were equal to or greater than $100 million in 2020, ranging from $149 million to $1.2 billion1. It is important to note that the original data used in this analysis dates back to 2015, and losses have trended upward to date. The severity results of thise analysis show to be consistent (when considering the upward trends) with the findings from Cornerstone’s research in regards to similar loss ranges.

# **Conclusion**

Overall, the modeled results did provide the ability to predict with some level of success as to the likelihood and what is driving the potential for a shareholder class action suit to occur. However, taking into consideration the model limitations, in conjunction with both the room for error in the model’s overall prediction, which was correct only 24% of the time, and the industry benchmark of 3% of companies being sued annually on average, it is not recommended to use this model until much more diligence is done around this risk.

However, upon assessing the potential risk from a severity lens, the straightforward method to estimate Crane Co. potential dollar loss does provide some meaningful insight into the potential dollar loss it is exposed. In addition, the supplemental benchmarking research provided by Cornerstone offers some additional assurance in terms of what the overall losses look like for publicly traded entities.

Based on the findings in this analysis and due to the potential monetary impact shareholder class action litigation can have on Crane Co., it is recommended that Crane Co. purchase between $25 and $35 million in insurance coverage, depending on management’s risk appetite.

# **References**

1Cornerstone Research. (2021). *Security Class Action Settlements 2020 Review and Analysis*. Cornerstone Research, Inc. https://www.cornerstone.com/Publications/Reports/Securities-Class-Action-Settlements-2020-Review-and-Analysis.pdf

2Cornerstone Research. (2021). *Security Class Action Settlements 2020 Year in Review*. Cornerstone Research, Inc. https://www.cornerstone.com/Publications/Reports/Securities-Class-Action-Filings-2020-Year-in-Review.pdf

# **Appendix A – Base Dataset Variables**

|  |  |
| --- | --- |
| **gvkey** | **Global Company Key** |
| datadate | Data Date |
| most\_recent\_date | Most recent date in which a company filed a 10-k |
| tic | Ticker Symbol |
| conm | Company Name |
| GIC\_Industry | GIC industry in which the company is associated |
| GIC\_SubIndustry | GIC Subindustry in which the company is associated |
| suitflag | Flag a company (0 = No; 1 = Yes) if company had a shareholder class action suit filed against them based on SCA Filings and Settlement dataset |
| restatementflag | Flag a company (0 = No; 1 = Yes) if company experienced a restated filing |
| idbflag | International, Domestic, Both Indicator |
| at | Assets - Total |
| capx | Capital Expenditures |
| cogs | Cost of Goods Sold |
| cshfd | Common Shares Used to Calc Earnings Per Share - Fully Diluted |
| dltt | Long-Term Debt - Total |
| dp | Depreciation and Amortization |
| emp | Employees |
| epspi | Earnings Per Share (Basic) - Including Extraordinary Items |
| ni | Net Income (Loss) |
| ppent | Property, Plant and Equipment - Total (Net) |
| sale | Sales/Turnover (Net) |
| teq | Stockholders Equity - Total |
| txt | Income Taxes - Total |
| wcap | Working Capital (Balance Sheet) |
| xido | Extraordinary Items and Discontinued Operations |
| xint | Interest and Related Expense - Total |
| xsga | Selling, General and Administrative Expense |
| roa | Return on assets (Net Income / Total Assets) |
| roe | Return on equity (Net Income / Total Equity) |
| gla | Gain/Loss After-tax |
| dvt | Dividends - Total |
| invt | Inventories - Total |
| dte | Debt to Equity (Total Equity / Total Debt) |
| revt | Revenue - Total |
| chech | Cash and Cash Equivalents - Increase/(Decrease) |
| do | Discontinued Operations |
| exre | Exchange Rate Effect |

# **Appendix B – Binary and Multi-value Variables**

|  |  |
| --- | --- |
| **Column** | **Notes** |
| **Binary:** |  |
| idbflag | D = ‘Domestic’ converted to 0  B = ‘Both’ converted to 1 |
| Restatement Flag | 1 = ‘Yes’; 0 = ‘No’ |
| **Multi-Value** |  |
| GIC\_Sub\_Industry | Newly created columns:  'GIC\_Industry\_Aerospace & Defense' \*\*  'GIC\_Industry\_Industrial Conglomerates'  'GIC\_Industry\_Construction & Farm Machinery & Heavy Trucks'  'GIC\_Industry\_Electrical Components & Equipment' 'Industrial Machinery'  'GIC\_Industry\_Heavy Electrical Equipment' 'Construction & Engineering'  'GIC\_Industry\_Agricultural & Farm Machinery'  \*\*Removed due to logistic regression requirement to drop at least 1 dummy variable |
| GIC\_Industry | Removed and not used due to direct correlation to GIC\_Sub\_Industry column |

# **Appendix C – Initial Correlation Plot**

Chart, scatter chart

Description automatically generated

# **Appendix D – Severity Assessment Data – Before Outliers (in $ millions)**

Chart, histogram

Description automatically generated

# **Appendix E – Severity Assessment Data – After Outliers (in $ millions)**

Chart

Description automatically generated

# **Appendix F – Severity Assessment – Data Inputs and Calculation**

Table

Description automatically generated

# **Appendix G – Important Features – Random Forest Model with SMOTE and hyper parameter tuning**

Chart, bar chart

Description automatically generated

# **Appendix H – Cornerstone Research – Number of Filings by Industry**2

Graphical user interface, application

Description automatically generated