Capstone Project

Merrimack College

March 2021

Taryn Popplewell

Data Science Program

Executive Summary

In 2019 there were a reported 74 securities class action lawsuits in the United States with a combined value of $2 billion. The reported average settlement amount was $27.4 million though four of the 74 suits filed in 2019 resulted in settlement values greater than $100 million (Bulan and Simmons). With such high settlement values, securities class action lawsuits are one of many risks publicly traded companies must plan for.

In an attempt to understand its own risk of experiencing a settlement class action, Mohawk Industries has requested a likelihood and severity estimation study be conducted. Using data available from Standard & Poor's (S&P) Compustat database, Mohawk data was grouped with companies in the same consumer discretionary sector to assess and compare the risks it faces of being filed against.

Key Findings

Several models were run to assess settlement likelihood. Their predicted likelihood of Mohawk experiencing a settlement can be viewed in the chart below. Logistic regression was selected as the final model and produced the highest likelihood estimate of .7171, or about 72% likely.

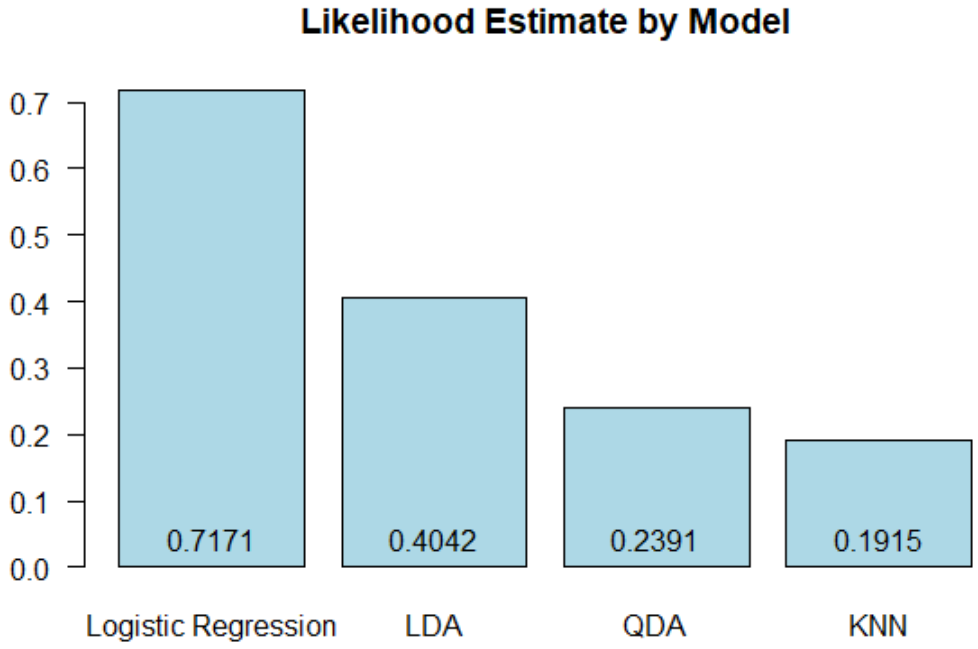

Figure 1. Likelihood Estimate by Model

To predict the severity or settlement amount of a settlement class action suit, several different models were run to determine this value. Linear regression was selected as the final model with a predicted settlement severity of about $24.2 million. A 95% confidence interval was built around this estimation and produce a range of value from $10 million to $38 million.

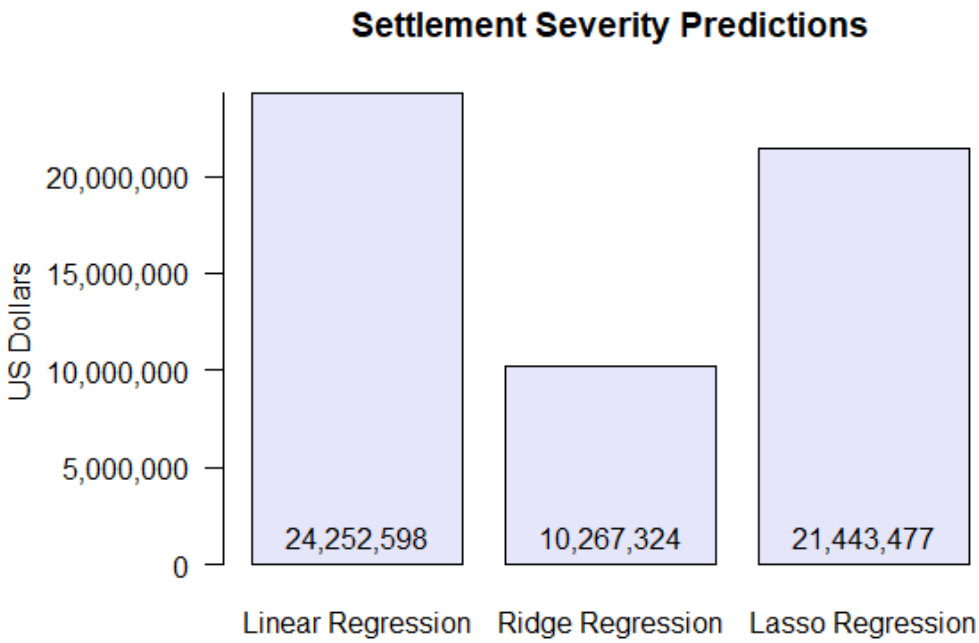

Figure 2. Severity Estimations by Model in US Dollars

Recommendations

Settlement data from 2005 to 2015 was used to determine likelihood and severity, though this information was scant against the total number of companies used in the study. Adjusting the research data for outliers left 155 companies total, 40 of which experienced a settlement class action between the aforementioned dates.

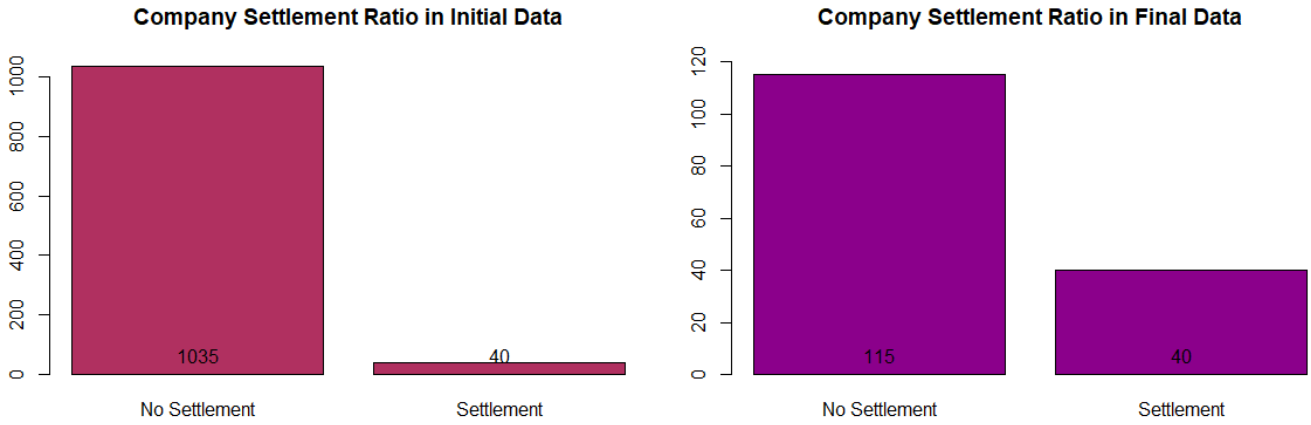

Figure 3. Non-settlement to settlement counts in initial vs final data sets

Data used to build the research data sets spanned January 2009 to September 2013, making the newest available data used for research at least seven years old. Reconducting this study with a more recent data set could result in more accurate or representative likelihood and severity estimations.

Detailed Procedures and Findings

Data Due Diligence and Feature Engineering

The data used in this study was scattered across four different spreadsheet extracts (named Fundamentals, Stocks, Securities, and Ratings) sourced from the Compustat database, which is well-known resource for fundamental financial and price information dating back to 1950 for both active and inactive publicly traded companies (Baker Library|Bloomberg Center). Data from these spreadsheets was dated from 2010 to 2014. There was an additional spreadsheet of Securities Class Actions from an unknown source with records spanning between the years of 2005 to 2015. The five spreadsheets contained information for a variety of different industries and sectors. To focus the data, only companies sorted into the same "gsector" field (Consumer Discretionary), as Mohawk Industries were included in the final dataset.

The collective field count across all sources surpassed 1,900 and required considerable culling. Due to limited knowledge of the financial industry, fields with descriptions that contained the word 'total' were given preference to be maintained. A small amount of research was conducted to identify some of the more commonly used financial health assessment ratios and determine whether they could be reproduced using the fields available. Fields related to Standard and Poor's (S&P) ratings of individual companies were also maintained where possible. Other fields that had descriptions that sounded familiar were also earmarked for possible inclusion. The final dataset required that each company be represented by a single row of data. To achieve this, special care was taken to aggregate each individual spreadsheet before merging them all together.

Once fields were flagged to be included in analysis, the datasets were evaluated to determine the how they should be merged. The goal of this effort was to keep the number of NAs introduced by coercion to a minimum. The Fundamentals data set was identified as the largest contributor of fields to what would be the final analysis-ready data set, so it was used as the backbone for all data merges.

Fundamentals
The Fundamentals file contained annual data for year 2009 to 2013. Most of the relevant companies in this file had data reported for at least four of the five available years. This data was aggregated so that each row represented a single company. Since there was usually more than one row of data per company, continuous fields were averaged across all available rows per company. Categorical variables did not vary over year by company. Fields containing mostly NAs or empty rows were ruled out, though one exception was made for the delete reason field. All of the fields selected for inclusion from this data set had an NA percentage fewer than 45%, regardless of whether they were continuous or categorical. The final aggregated Fundamentals data set came to include 1,100 rows composed of 50 fields. Fields from this data set were used to create two calculated fields; liquidity ratio (liqratio) and debt-to-income ratio (dtaratio). More detail on how these calculations were performed is available in Appendix B.

Securities
The Securities file contained monthly data from January of 2010 up through September of 2013. Many of the company codes in this file contained duplicate date under a slightly different ticker. For example, company A might be listed under ticker symbol "COMPA" as well as a second ticker symbol "COMPA.1" for the same time periods. To handle this kind of scenario companies with duplicate ticker symbols had their monthly data counted and compared. The ticker with the greatest number of months reported was retained for final reporting and the data for the alternate, less frequent ticker was discarded. In the instance when a company's duplicate tickers had an equal number of reported months, one ticker and the associated monthly data was selected at random for inclusion. Once each company was represented by a single ticker, the continuous variables reported each month were averaged to create a single row of data for each company. This decision was made to keep data cleansing uniformity since continuous variables were averaged in the Fundamentals file as well. The final aggregated Securities data set came to include about 2,100 rows composed of 10 fields.

Stocks
The Stocks file was the largest with over 4 million rows of data to consider across 76 fields. This file contained daily data about trading highs and lows, dividend activity, and other data points related to stock trading from January 2010 to September 2013. Because this data was recorded daily, there was a high volume of NAs. To aggregate the data, the last 30 days of activity for each company was averaged. The 30-day limit was implemented to avoid including too many zero/NA rows in the final aggregated continuous fields. This data set had the same issue as the Securities file, where company data was sometimes duplicated under a slightly varied ticker symbol. These duplicates were dealt with using the same method employed with the Securities data. The final aggregated Stocks data set came to include about 2,100 rows composed of 9 fields.

Ratings
The Ratings file contained monthly ratings assessments for each company from January 2010 to September 2013. Research was conducted to determine the ranking system for the academic-style ratings system of A/B/C/D etc., with many variations between each letter level. These rankings were then converted to numbers with NA values equated to 1 and each subsequent positive ranking equated to an increasing numerical value. There were no noticable issues with duplicated ticker symbols across available dates in this file. After each ranking variable was converted to a numerical equivalent the max date row for each company was selected to be used in analysis. The final aggregated Ratings data set came about 1,500 rows composed of 6 fields.

SCA Filings and Settlements
The SCA Filings and Settlements file differed from the other data sets in that it did not contain a unique company code to identify each company. This presented the problem of having to merge this data with the other records using a combination of the company filing name and ticker symbol instead of the unique "gvkey" field, which was available in all the previously mentioned Compustat data files. In this file and the previous four others, the company filing name field was cleansed using a series of regular expressions and trim functions. The resulting filing names had all punctuation removed, were converted to all capital letters, leading and trailing empty characters removed, and all spaces between words replaced with underscore characters. As an

example, the original filing name of “APPLE Computer, Inc.” was cleansed to “APPLE_COMPUTER_INC”. Words like “corporation” and “international” were also shortened to “corp” and “intl”. More detail on the text cleansing process for filing names can be found in Appendix B.

There were also instances where a company had multiple row entries in the SCA file. When this was encountered the settlement values available were summed into a single value. A new field was created to indicate whether a company experienced a settlement class action. If a company had a settlement amount greater than 0 then it was flagged as having a filing. Some companies had a settlement status of “Ongoing” but these were not flagged as having a filing since a final ruling had not been made within the date range of the data set, which spanned from 2005 to 2015. The final aggregated SCA Filings and Settlements data set came to 1,702 rows composed of 3 fields.

The five data sets were merged together, with the Fundamentals data set serving as the backbone and limiting factor for all gvkey joined onto it. Regardless of which file was selected as the backbone, NAs would have been introduced in the join, and using the Fundamentals file introduced the fewest. The SCA file was joined to the merged Compustat data set using the cleansed Filing Name field from both sources as well as ticker symbol. This join found 40 shared rows, meaning 40 of the roughly 1,100 companies in the final data set were flagged as having had some form of settlement.

Response Variables

The severity estimate response variable was created using the available settlement amounts from the SCA file. If a company did not have a settlement amount listed, its value was set to zero. The likelihood estimation required a binary 0/1 or Yes/No field. This was derived using the settlement amount. Any company with a settlement amount greater than 0 received a classification of 1 to indicate it had experienced a settlement class action. Any companies without a settlement amount received a classification of zero.

Normality Checks and Handling Outliers

A simple linear regression formula was run against the entire cleansed data set to assess normailty. Figure 4 below shows the residuals and QQ plot of the intial data set. We can see that residuals are not randomly distributed and that there are several outliers that could be affecting the normal distribution of the data.

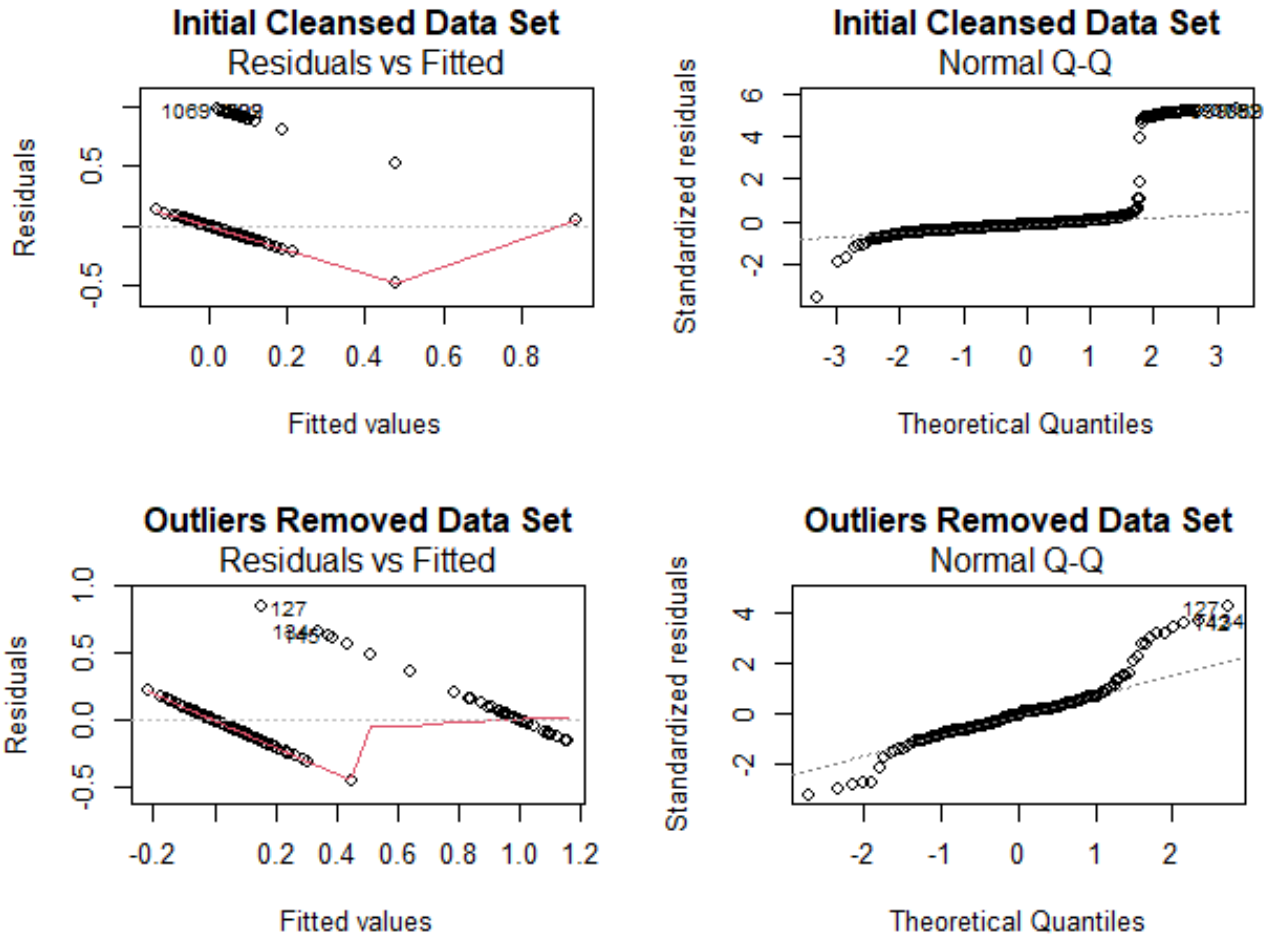

Figure 4. Initial Cleansed vs Outliers Removed Data Normality Checks of Likelihood Data Set

Because of the disproportionately low number of settlements within the cleansed data set, these records were set aside before removing outliers from the non-settlement data. This was done to prevent the number of settlements in the data set from dropping any lower. After outliers were removed issues persist with the residuals, but the Normal QQ plot is much improved. The outliers-removed data set will be the final data set.

Similar results are seen in Figure 5 below when checking the normality of the severity estimate data set. The only difference between the likelihood and severity data sets is the response variable. In the severity data set the residuals plot shows more randomness because it is not a categorical variable. See the before and after plots below of the severity estimate data sets.

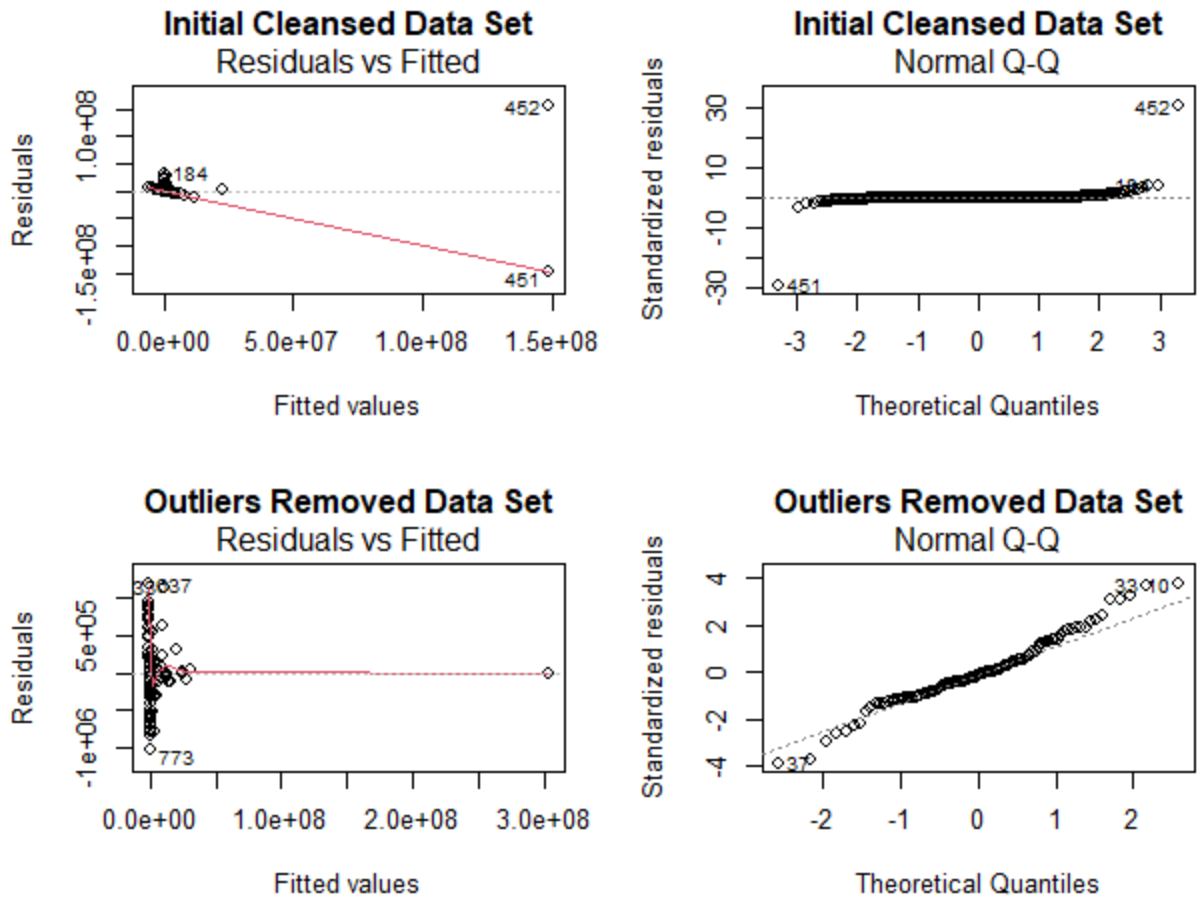

Figure 5. Initial Cleansed vs Outliers Removed Data Normality Checks of Severity Data Set

Methods and Findings

The evaluation of the Mohawk Industries data required both qualitative and quantitative supervised modeling. The likelihood estimation was a question of classification and used qualitative models like logistic regression. The severity estimate required a ballpark figure and companion confidence interval for a settlement amount in the event Mohawk were to experience a settlement class action. Regardless of whether estimating likelihood or settlement severity, each model used a train/test ratio of .7/.3 for validation purposes.

In the spirit of exploration, a handful of models were run for both estimations. The accuracy results of each model, its interpretability, and the actual results produced were considered before a final model was selected to make the requested likelihood and severity estimations.

Likelihood Estimation

The likelihood estimation process compared four different model results. Logistic regression (glm), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and K Nearest Neighbords (KNN) models were all run and compared. The correct classification rates produced by each model was used as the primary factor when determining which model to ultimately select for final estimations, alongside model ease of use and understanding. Figure 6 below displays the correct classification rates produced by the confusion matrices of each model. Figure 1 in the Executive Summary of this document displays the resulting likelihood estimate produced by each model.

The best fit for each model happened to be found using the same three variables. All three likelihood models used the last 30-day average number of shares outstanding from the Stocks file, the current S&P Quality Ranking (as of 2013) from the Ratings file, and the liquidity ratio which was calculated using fields from the Fundamentals file.

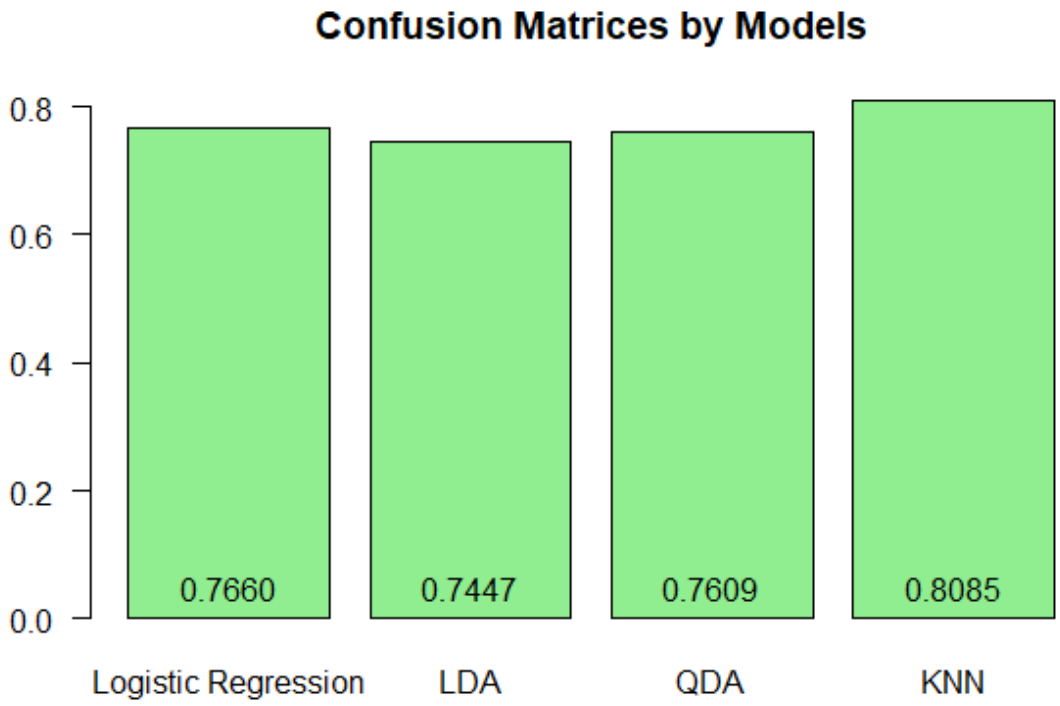

Figure 6. Correct classification rates produced by model confusion matrices.

Logistic Multiple Regression

Logistic multiple regression is designed to be used to make classification predictions with data that does not necessarily meet the usual normality assumptions required for linear regression. The data used in this study is not the picture of normality so this model can serve to work around some of those issues.

The model returned a confusion matrix showing a correct classification rate of about 0.7660. This means the model made a correct prediction against the test data about 76.6% of the time.

Using the coefficients produced by the glm model it was possible to produce a specific likelihood estimate for Mohawk Industries. This predicted value came to .5985, suggesting that Mohawk is about 60% likely to experience a settlement class action suit.

Linear Discriminant Analysis

The LDA model returned a confusion matrix that showed a correct classification rate of about 0.7447, or the model correctly predicted whether a company experienced a settlement about 74.5% of the time.

Using the coefficients produced by the LDA model it was possible to produce a specific likelihood estimate for Mohawk Industries. This predicted value came to .4042, suggesting that Mohawk is about 40% likely to experience a settlement class action suit.

Quadratic Discriminant Analysis

The LDA model returned a confusion matrix that showed a correct classification rate of about 0.7609, or the model correctly predicted whether a company experienced a settlement about 76.1% of the time. This is the same correct classification rate as the logistic regression mode.

The QDA model does not produce coefficients so a specific likelihood estimate was not created using this model. Instead, the QDA model was evaluated against the Mohawk data and classified it as a 0, or in the ‘not sued’ category. Using the above correct classification rate of .7609 we can estimate that according to QDA analysis, Mohawk has a 1-.7609 chance of being sued, or about 24%.

K-Nearest Neighbors

The KNN model produced the highest correct classification rate of .8085, or 80.9%. Similarly to QDA, coefficients are not produced with KNN models so the same combination of classification and classification rates will be used to reach a likelihood estimation. The KNN model classified Mohawk as 0/’not-sued’. Therefore, we estimate that the company’s likelihood of experiencing a settlement class action is 1-.8085, or about 19%.

Severity Estimate

The severity estimation process compared three different model results. Linear regression (lm), Ridge regression, and Lasso regression models were all run and compared. Figures 7 and 8 below respectively display the mean-squared errors (MSEs) of each model and the 95% confidence interval of the severity estimate. The resulting severity estimate produced by each model is available as Figure 2 in the Executive Summary of this document.

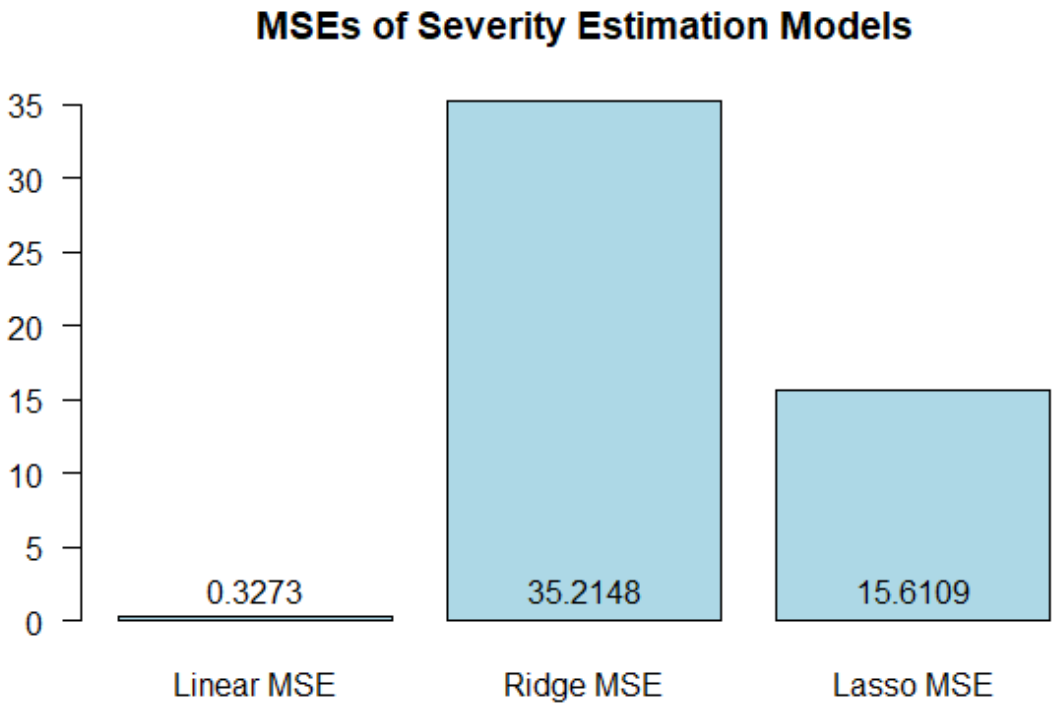

Figure 7. Mean-Squared Error of Severity Estimation Models

The linear regression model used the Reason for Deletion field, the Capital Expenditures field, the Net Income (Loss) field, the Total Market Value field, and the Total Short-Term Investments field, all of which originated in the Fundamentals file. The Ridge model used the Monthly Trading Volume field, the S&P Domestic Long Term Issuer Credit Rating field, the Monthly Low-Price field, and the Total Market Value field. Lasso regression models used all the same fields as the linear regression model except for the Short-Term Investments field which was removed.

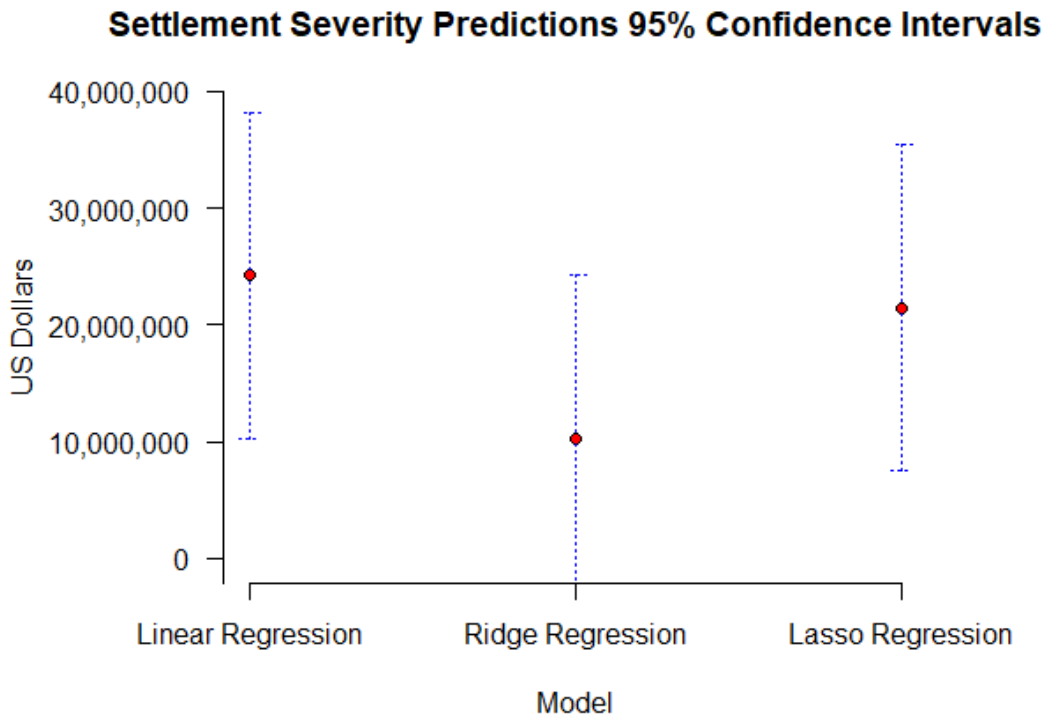

Figure 8. Severity 95% Confidence Intervals by Model in US Dollars
**Values below zero not displayed*

Linear Regression

The linear regression model returned an MSE of .3273, which is very good result. Using the coefficients produced by the model, the predicted severity estimate came to $24,252,598 with a 95% confidence interval from $10,308,620 to $38,196,575.

Ridge Regression

Ridge regression is a model that estimates its coefficients not by ordinary least squares, but by a 'ridge' estimator that is designed to reduce variance, though it does introduce bias. Ridge regression can also be used to lessen the effects of multicollinearity, which we didn't look very closely for in the normality checks. In this study the same set of variables were used for both Ridge and Lasso models. Cross validation was used on the training set to determine the best penalty value (lamba) before re-running the model against the test set for the final coefficient estimates.

The ridge regression model returned an MSE of 35.2, the highest MSE of the three. Using the coefficients produced by the model, the predicted severity estimate came to $10,267,324 with a 95% confidence interval from -$3,676,654 to $24,211,301, though in terms of settlement class actions suits the negative lower limit is equal to zero from a practical business stand point.

Lasso Regression

Lasso regression uses shrinkage and penalty values against coefficients both to create a regression model and to perform variable selection. Cross validation was used on the training set to determine the best penalty value (lamba) before re-running the model against the test set for the final coefficient estimates.

The lasso regression model returned an MSE of 15.6. Using the coefficients produced by the model, the predicted severity estimate came to $21,443,477 with a 95% confidence interval from $7,499,500 to $35,387,455.

Final Model Selection and Conclusion

Likelihood Model Selection

Despite KNN having the best classification rate, the logistic regression model is ultimately recommended to estimate the likelihood of a settlement class action suit against Mohawk Industries. This decision was made due to the glm model's correct classification rate, the relative consistency of prediction values across all qualitative methods and the glm model's placement among them, as well as for the interpretability of the glm model output. The formula used to create the glm estimation can be found in Appendix B.

Severity Model Selection

Linear regression was the clear choice for the severity estimation model. There was a rather wide range of MSEs produced across the three models, though the predicted severities were somewhat consistent. The consistency of severities alongside linear regression's low MSE singled it out as the best choice for the final severity estimation. Another point in linear regression's favor is its easy interpretability. Like the glm model, the coefficients produced by the linear model can be used to create a simple formula to calculate the severity estimate and confidence interval. This formula can be found in Appendix B.

Final Estimates and Conclusion

In conclusion, the final likelihood estimation for a settlement class action against Mohawk Industries is predicted to be about 72% with a severity estimate of $24.2 million in a 95% confidence interval ranging from roughly $10 million to $38 million. The highest date for non-settlement data used in this study was September of 2013. Mohawk's reported annual revenue for 2013 was $7.3 billion (Macrotrends), so the severity estimates reached while high, do not exceed the value of the company.

Recommendations for improvement include refreshing the provided dataset to include more up-to-date data to potentially improve model accuracy and reliability.

References

Macrotrends. (2020) Mohawk Industries Revenue 2006-2020 | MHK. Retrieved March 3. 2021 from https://www.macrotrends.net/stocks/charts/MHK/mohawk-industries/revenue

Bulan T., L., Simmons E., L. (March 2020). *Securities Class Action Settlements—2019 Review and Analysis.* Retrieved March 9, 2021 from https://corpgov.law.harvard.edu/2020/03/11/securities-class-action-settlements-2019-review-and-analysis/

Baker Library|Bloomberg Center. (ND). Harvard Business School. Retrieved March 3, 2021 from https://www.library.hbs.edu/Find/Databases/Compustat

Appendix A contains the data dictionary and fields selected for use in this study from the provided data files.

Appendix B contains all code used in this study.