LITIGATION LIKELIHOOD & SEVERITY
MOHAWK INDUSTRIES

Classification & Regression Analysis Report

KHALED SHARAFADDIN

July.02.2021

EXECUTIVE SUMMARY

The financial dataset represents business information for over 2,600 US and international corporations worldwide from 2009 to 2014 fiscal years. Information is collected from multiple sources including companies' financial data, government ratings, securities and the SCA filing and settlements public data for companies that have been historically subject to litigation. The goal of this report is to identify and examine the potential risk factors associated with shareholders legal actions against MOHAWK INDUSTRIES. Understanding the risk factors of litigation will better prepare the organization financially and legally by mitigating these threats in the short and long term, and potentially enroll in lawsuit insurance programs suitable for the business.

Over the course of this analysis, methods used for exploratory data analysis and aggregation, feature selection, data imputation and normalization techniques, hyperparameter optimization, and implementation of several algorithms for classification and regression problems will be discussed. This report evaluates the performances of K-Nearest Neighbors, Random Forest, and Support Vector Machine algorithms on classifying MOHAWK INDUSTRIES likelihood of litigation. The report also provides severity estimation of lawsuit amount in U.S dollars, by assessing the accuracy performances of Lasso and Ridge regressions, Gradient Boosted Machine, Bagging and Random Forest methods.

Using the Random Forest classifier as the primary model for binary classification of litigation likelihood, the model performs 90% precision, followed by SVM and K-NN with 88% and 79% respectively. The rate of true positive instances for Random Forest achieved 90.3%, followed by 82% for SVM and 80.4% for K-NN. While the accuracy of the model is important, the sensitivity metric must not be overlooked because correctly identifying the positive instances of litigation likelihood is vital for MOHAWK INDUSTRIES's ability to manage legal risks. Contrary to K-Nearest Neighbor, Random Forest and Support Vector Machine predict that MOHAWK INDUSTRIES will NOT be subject to litigation.

Tree based models like Boosted Trees and Random Forest obtained lower test errors than Lasso and Ridge regularizations as regressors. Most likely due to the small dimensions of the data, as it contains only 37 companies. In terms of lawsuit settlement amount cost, MOHAWK INDUSTRIES is forecasted to pay a range between $20,000,000 and $57,000,000 dollars, based on the estimates obtained by the regression models. When taking into consideration the cumulative average estimate for Random Forest, Bagging, GBM, Lasso and Ridge regressors, the mean severity evaluation is $33,663,115 with median $27,441,548 US dollars.

MOHAWK INDUSTRIES is encouraged to focus on risk factors such as long and short deferred tax assets, average total returns, total investments, capital gains and losses, market value, debt, equity, credit rating, stock prices, inventory and equity, retained earnings, marketable securities, and cash flow. Moreover, it's critical to mitigate executive threats including illegal behaviours, inadequate controls, negligence and ineffective decisions. This analysis must take into account the limitations of the financial data available, weaknesses and strengths of the machine learning models, and the utilization of more complex predictive algorithms such as Artificial Neural Networks, PCA for dimensionality reduction and model stacking strategies in future analyses.

DATA & APPROACH

Feature Selection & Extraction

The Fundamentals, Securities, and Ratings files contain a large number of variables, all of which are not important in predicting whether a particular company will be sued or useful in determining the range of the settlement amount a company is likely to pay. Predictors from the three files are selected on the basis of their contribution to the overall predictive outcomes. Variables that contain information related to the company's revenues, expenses, profitability, stock prices, government debt ratings and loans are selected.

Feature selection steps include removing one class predictors such as the *'consol'* variable that consists of only category 'C'; and filling empty multi-class categorical predictors with new alternatives, such as updating the S&P Domestic Long Term Issuer Credit Rating *'splticrm'* and Accounting Standard *'acctstd'* variables from empty string class to 'Unspecified'. Other categorical variables will not be included in the final dataset because of their nature as identifiers to a particular industrial sector or groups, such as CIG group *'ggroup'*, CIG industries 'gind', and Stock Ownership Code 'stko'. In addition, many predictors contain values that might not be useful in determining the target outcomes as they contain many NULL or 0 values. For instance, variables such as Natural Resources at Cost 'fatn', treasury stock *'tstkp'*, and ADR ratio *'adrr'* all of which have either NULL or 0 values, and hence are removed as potential candidates. Similarly, some features have severe class imbalance which could affect training the machine learning model negatively. Oil and gas method *'ogm'* has two factor levels, where one category represents 99% of the records. Hence, these features will also be removed from the dataset.

Missing Data Treatment

Missing values must be handled properly in order to reduce the chances of drawing inaccurate inferences about the data. There are many variables where the proportion of missing values is large that it becomes unhelpful as predictors. Variables with 25% or more missing values are eliminated. The 25% threshold is reasonable because the dataset has an ample amount of predictors and records where the predictive power will not be severely affected. Moreover, the more missing data, the higher reliability on the imputation algorithm will be, and reducing this reliability is also sensible.

Multivariate imputation by chained equations (MICE) has been used to deal with the remaining missing data by focusing on one predictor at a time. The algorithm uses classification and regression trees (CART) as well as predictive mean matching (pmm) imputation methods to generate 'plausible' values. pmm method has the advantage of preserving the non-linear relationships of the predictors. The final dataset used in creating the predictive model will contain no missing values.

Factor Levels Reduction & Dummy Variables

It might be beneficial to reduce multiple factor levels in a particular variable, for the purpose of simplifying the model while avoiding a loss of its predictive power. Variables containing government ratings of organizations have been reduced based on their ordered variations. S&P Domestic Short and long Term Issuer Credit Ratings *'splticrm'*, *'spsticrm'*, and S&P Quality Ranking, *'spcsrc'*, are ordinal variables that have been converted from rating ranges A+ to D into 'High_rating', 'Med_rating', and 'low_Rating'.

Once all the feature engineering steps above have been completed, non-numeric variables in the dataset are converted into dummy variables, including the response variable ‘*if_sued’*, after being converted into binary class ‘No’ or 0 and ‘Yes’ or 1. Converting factors to dummy variables may speed up training the model, and some statistical methods depend on the Euclidean distance between data points. Therefore, having numeric variables is essential.

Feature Normalization

Machine learning algorithms such as K-Nearest Neighbor among many others require data scaling. normalizing the data to have mean=0 and standard deviation=1 reduces the impact of features with large values and puts them on the same scale with the features with smaller values (see formula below). This process helps reduce error rate, and speeds up the learning process.

$$x' = \frac{x - \bar{x}}{x_{max} - x_{min}}$$

Graph: Normalization Formula

Data Aggregation & Merging

Data sources from the fundamentals, ratings, and securities files are consolidated and aggregated into a unified dataset suitable for predictive modeling analysis. Due to the existence of multiple values of stocks, years, loans and other predictors, it’s critical that these data points are summarized in a way that’s reflective as an average value per company. In other words, each particular company must have only one aggregated variable representing the average value among all predictors. Figure.A below shows an example of the distribution of stock prices and monthly total returns. The median and the median absolute deviation (MAD) are used for highly skewed data points, while the mean and standard deviation is used for normally distributed data. The median absolute deviation is more robust and resilient to extreme high or low values and outliers than the standard deviation. On the other hand, categorical variables are aggregated by utilizing the most frequently found factor level.

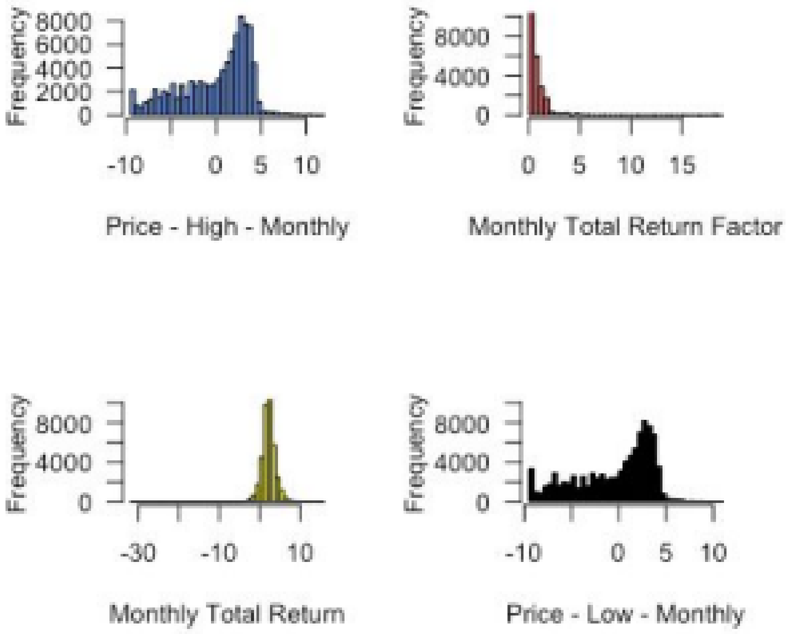


Figure.A: Distribution of Stock and Total Returns - Monthly

The final step of preparing the complete dataset is to create the independent variables '*if_sued*' as a binary classification outcome, and '*settlement_amount*' as a continuous dollar amount outcome U.S dollars, using the SCA filings settlement list of companies that have been sued historically. '*Settlement_amount*' is based on the maximum value found and will not be used as a predictor of whether the company will be subject to litigation. '*If_sued*' is initially set to 'Yes' for every company in the SCA filings and settlements. The target company, MOHAWK INDUSTRIES, has been removed from the dataset, in order to be predicted by the machine learning models. Finally, the data is merged together into one complete dataset using the global company key *'gvkey'* and the ticker symbol *'Ticker'*. The final dataset contains 903 distinct companies and 229 predictor variables.

Class Imbalance

The ratio of companies that have been sued vs. those that didn't is 18% to 82% of the dataset respectively. The disproportionate number of class instances might provide inaccurate evaluation by favoring the majority class instance *'Not Sued'*. To solve this problem, SMOTE has been implemented on the full dataset. This method is to artificially generate new examples of the minority class using the nearest neighbors of these cases.

Dimensionality Reduction

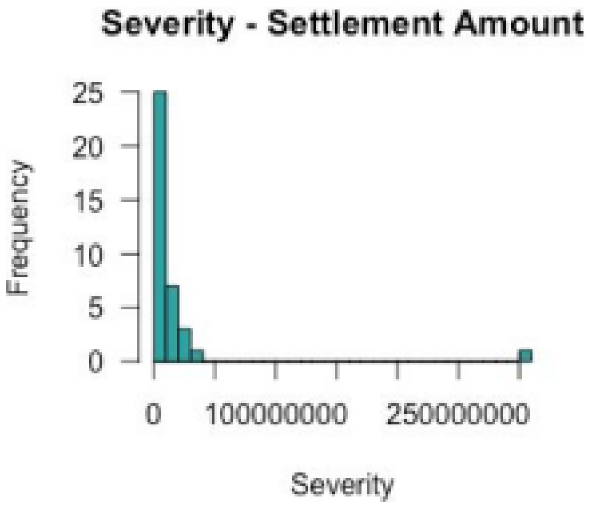
Using 229 predictors in the dataset is relatively large. Selecting a subset of these variables may increase the accuracy of the model, reduces overfitting, and enables the model to train faster. Two distinct datasets are created, each with a different number of predictors. The first set will use the 229 predictors, while the second set will take into consideration the elimination of pairwise correlations with the largest mean absolute value above the 80% threshold, in order to reduce multicollinearity. The datasets will use Random Forests, KNN, and Radial Support Vector Machine classification models to predict whether MOHAWK INDUSTRIES will be subject to legal actions. Random Forest, Bagging, Gradient boosted machine, L1 &L2 regularizations will be used to predict the settlement amount in U.S dollars.

Sampling Method

The datasets use K-Fold cross validation, with 10 folds. 9 Folds will be used for training the model while 1 fold will be used for testing purposes. This method allows for test error rate to be averaged across K errors, and derives the estimation accuracy of the model to perform higher.

Severity Estimation Sampling Size

The dataset that will be used to predict the severity estimation comprises only 37 companies. The histogram below shows the distribution of the severity cost for all companies. Two problems arise here: the number of instances in the set is very small which could reduce the accuracy of most statistical algorithms including OLS regression; the second problem is the existence of an outlier value of $303,000,000 for General Motor corporation (GM). It's unclear whether this is due to incorrect input or in fact GM paid this large cost. This outlier can affect the slope of the regression line greatly. Given the small dimension of the dataset, it's not advisable to remove the outlier, as it might provide useful insights and correlations between variables. Tree based algorithms such as random forest and ensemble methods like gradient boosted machine, as well as Lasso and Ridge regression algorithms are robust to outliers and therefore will be used for severity predictions.



Histogram: Distribution of Severity Estimation

DETAILED FINDINGS

Part I. Lawsuit Binary Classification (Yes, No Outcomes)

I. K-NEAREST NEIGHBORS

K-Nearest Neighbors classifies variables based on similarity measures. The algorithm uses euclidean distance to calculate the similarities between the variables. The classification occurs based on a majority vote of its neighbors. The dataset has been normalized and all predictor features are converted into numeric data types.

Optimal Value of K

The K-NN model has been trained using a variety of values for K. The statistical model achieved a high level of accuracy where K=5 on the testing dataset. After training the model, prediction estimation on the testing dataset resulted in a final accuracy of 79%. Confidence interval of 95% shows that the true accuracy estimation is between 75.4% and 83.4%.

Figure B below shows the Receiving Operating Characteristic curve, which plots the true positive rate (Sensitivity) as a function of the false positive rate (1-Specificity) for different cut-off points of a parameter. ROC achieved 79% which is good. Kappa rate is also measured, which calculates the percentage of data values that were classified correctly and then adjusts the values for the measure of agreement that is expected due to mere chance. The kappa level achieved is 58% which is acceptable. The sensitivity of the model resulted in 80.4% accuracy- a critical metric that must be used to identify the true positive rate which reflects whether K-NN identifies companies that might be subject to legal actions.

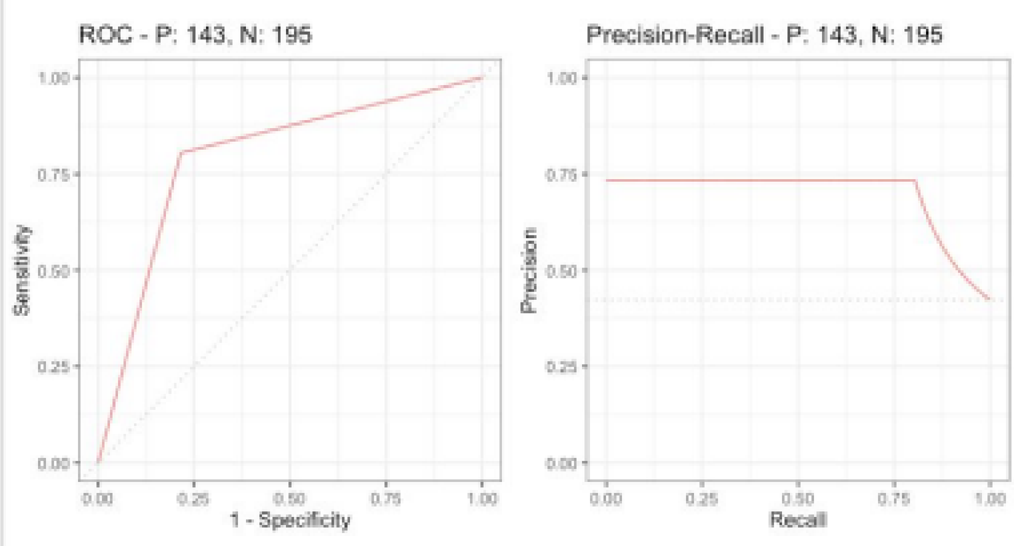


Figure B: K-NN ROC Curve for Prediction on Testing Dataset

II. SUPPORT VECTOR MACHINE

Support Vector Machine algorithm constructs a hyperplane in N-dimensional space that distinctly classifies the data points. A good separation between class instances is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class.

Tuning Support Vector Machine

SVM is applied on the predictors in the dataset in order to find a large functional margin that will minimize the error rate of the prediction. The dataset is split using repeated K-Fold cross-validation, with 10 folds. This will be useful in creating an accurate litigation risk prediction by training SVM and testing its accuracy.

SVM has two hyper-parameters that work by trying different value combinations to find optimal classification separation. Gamma parameter defines how far the influence of a single training example reaches, with low values indicating ‘far’ and high values meaning ‘close’. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors. Figure C shows different tuning parameters for gamma and C parameters, with best values C=100 and gamma=0.1.

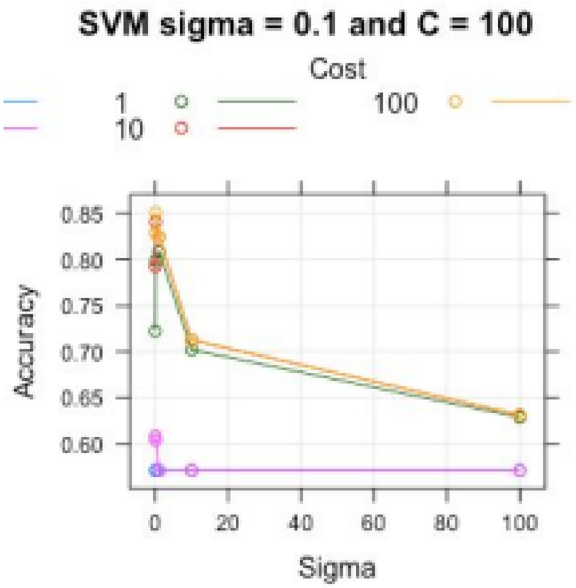


Figure C: SVM Model Hyper-Parameters

The final radial SVM model is tuned and built using C=100 and gamma=0.1 providing an 88.4% accuracy level, 82.0% sensitivity and 93.3% specificity on the testing dataset. Graph D represents the AUC at approximately 87%, indicating a high degree of separability between companies that are at high risk vs. low risk of lawsuit.

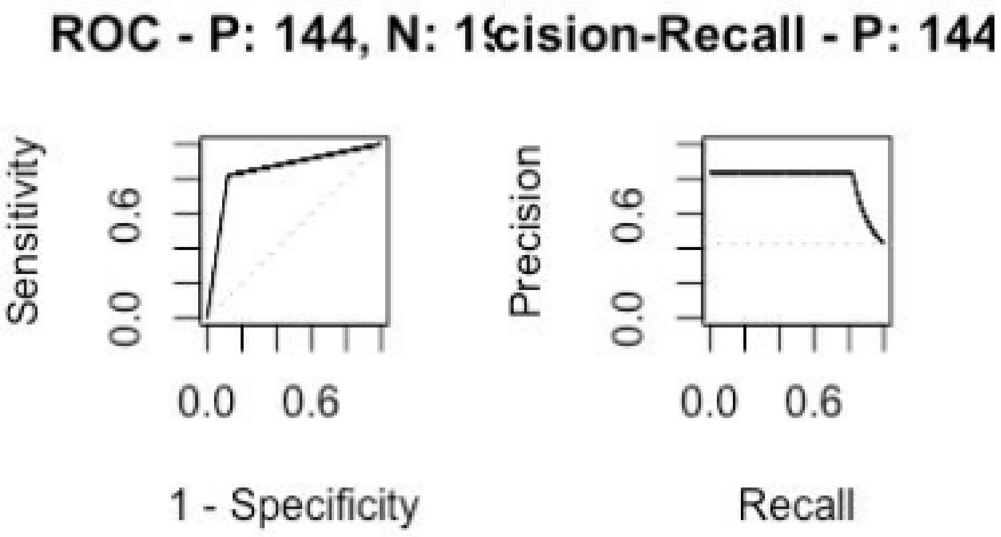


Figure D: ROC and AUC for Radial SVM

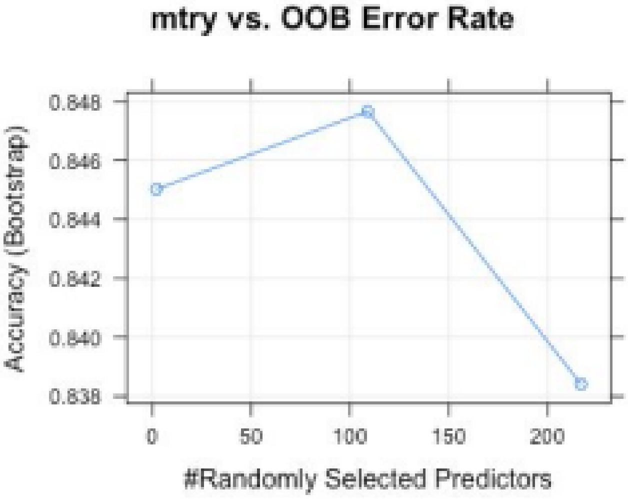
III. RANDOM FOREST

Random Forests is a supervised machine learning technique, where it averages multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This is important because the financial dataset contains a large number of predictors and has high variability.

Tuning Random Forest

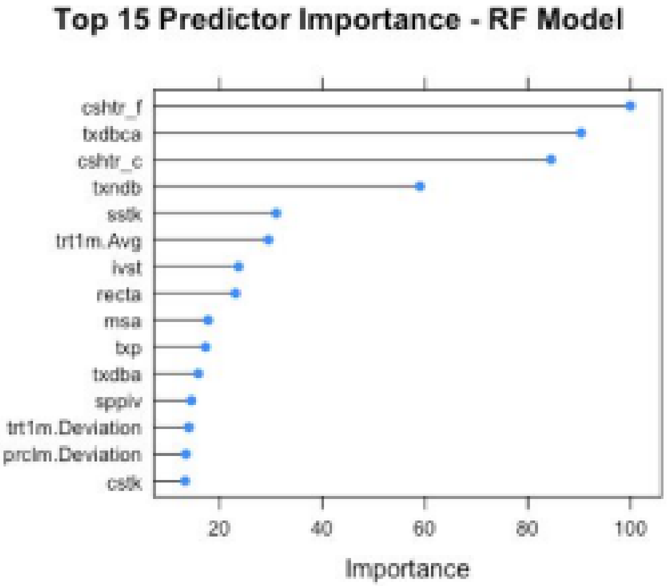
The algorithm is applied on the predictors including, but not limited to common shares traded, stock sales, and funds. The features will be utilized to predict the target variable *if_sued*, to the end goal of finding the highest accuracy level possible. The dataset is split using repeated K-Fold cross-validation, with 10 folds.

Before implementing Random Forest, it's recommended to tune the model by choosing the complexity parameters associated with the optimal resampling statistics such as the number of variables available for splitting at each tree node (mtry), as well as the number of trees (ntree). Graph E below shows the best mtry value = 109 based on the out of bag estimate, which is the mean prediction error on each training sample S, using only the observations that were not used in training the model.



Graph E: Number of Mtry vs. OOB Error Rate

The final random forest model is tuned and built using mtry=109 and ntrees=500 providing an accuracy level of 90.5%. The plot below provides a summary of the most important features in determining litigation risks. Notice that traded shares, long and short deferred tax assets, average monthly total return, total short-term investments, and capital gains and losses play major roles in predicting the company's risk of lawsuit.



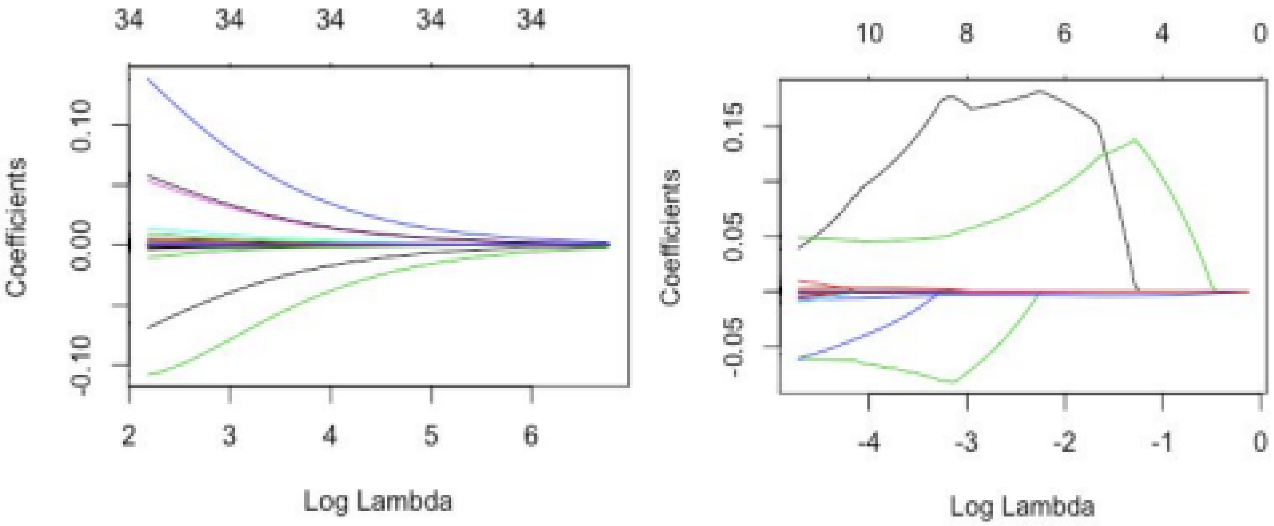
Plot: Variable Importance for Random Forests Model

Part II. Lawsuit Severity (Settlement Amount)

Lasso & Ridge Regularizations

The standard linear regression model performs very poorly due to the large multivariate dataset where the number of predictors is larger than the sample size. Alternative, lasso and ridge regularizations are utilized. These techniques impose penalty on the predictors by shrinking the coefficient values towards or equal to zero. Ridge regression shrinks the coefficients close to zero for predictors with least contribution to the predictive ability of the model by finding an optimal value for the constant lambda. On the other hand, lasso regression removes some features altogether by shrinking the coefficients to zero, which serves as a feature selection method as well.

Plot A below shows the function of log of lambda vs. the coefficients. When the log(lambda) = 8, all the coefficients are zero. As lambda value gets smaller, the coefficients increase and the sum of squares of the coefficients increase until lambda is near zero, where the coefficients are un-regularized. On the other hand, lasso regression in plot B imposes penalty on the absolute value of the coefficients where some of them will become exactly zero. The top of the plot indicates how many non-zero predictor variables are in the model. For instance, there are 6 predictors when log(lambda) = -2.



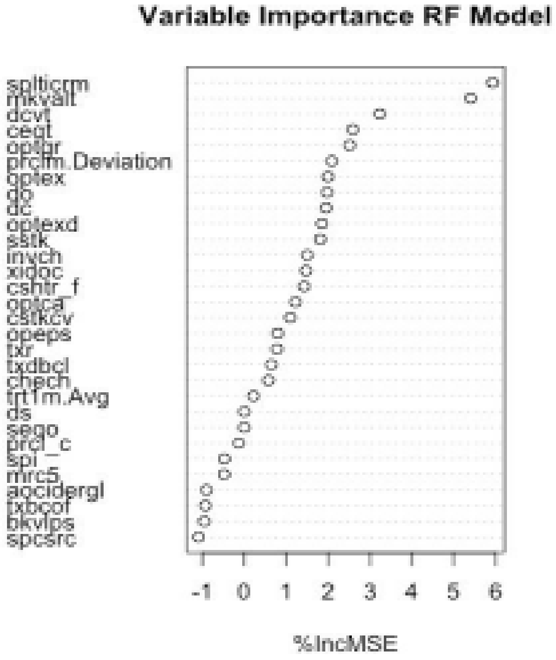
Plot A. Ridge Regularization Plot B. Lasso Regularization

GBM, Bagging, and Random Forest

Ensemble techniques such as bagging and Gradient Boosting Machine combine the results of multiple models on sub-samples of the dataset. The technique attempts to correct the errors of the previous model, where combinations of weak learnings form a collective decision in order to increase the overall results. This becomes particularly useful for the small dataset at hand.

Tree based models such as Random Forest are powerful algorithms resulting in high accuracy levels, interpretable, and can handle non-linear relationships in the dataset. In tree based models, samples are split into homogenous subsets based on the input variables. Overfitting, nonetheless, is a key challenge, particularly when the sample size of the dataset is very small. To avoid overfitting, tree size has been reduced.

The test errors for Random Forest, and the ensemble techniques, Bagging and GBM are smaller than lasso and ridge regularizations, potentially due to the high non-linear and complex relationships between the response variable, settlement amount, and the predictors. Graph below shows the most important predictors of settlement amount (severity) for random forest models. Market value, debt, equity, credit rating, inventory, common stocks and many other variables contribute to the overall prediction.



Plot: Variable Importance for Random Forests Model

VALIDITY AND RELIABILITY ASSESSMENT

Part I. Lawsuit Binary Classification (Yes, No Outcomes)

K-Nearest Neighbors

The K-NN statistical model is non-parametric and intuitive, requiring a homogenous type of predictors, and therefore, scaling the dataset is critical. While K-NN does not require training, finding a good value for K requires trial and error. The best model performed 79.3% accuracy level, misclassifying 70 instances out of 338. While a 79.3% accuracy estimate is moderately high, 70 misclassifications are costly nonetheless. Future implementations may include adding more data points, and reducing the dataset dimensions to achieve better precision.

While the K-NN classifier does not output probabilities, the model has predicted that MOHAWK INDUSTRIES will indeed be subject to litigation. This result suggests that the organization must take into account the accuracy of the model under K=5, accuracy level of 79%, as well as the number of features in the dataset. K-NN's accuracy is degraded as the dimensionality of the data increases, causing the model to perform lower. The sensitivity rate of 80.4% indicates a high accuracy in identifying the risk factors of litigation.

Support Vector Machine

Radial SVM model is chosen because of its effectiveness with high dimensional data, and uses the regularization parameter C, allowing it to reduce overfitting. The two datasets used to create the models both estimated that MOHAWK INDUSTRIES will NOT be subject to legal actions. Given that the algorithm does not estimate the probabilities of class instances because it uses the hyper-plane separation boundaries, it will not be possible to provide the probability of a lawsuit. In future implementations, it's recommended to improve on the SVM model by using Platt scaling to generate the probabilities of SVM's scores and fitting additional cross-validation on the training dataset.

Random Forest

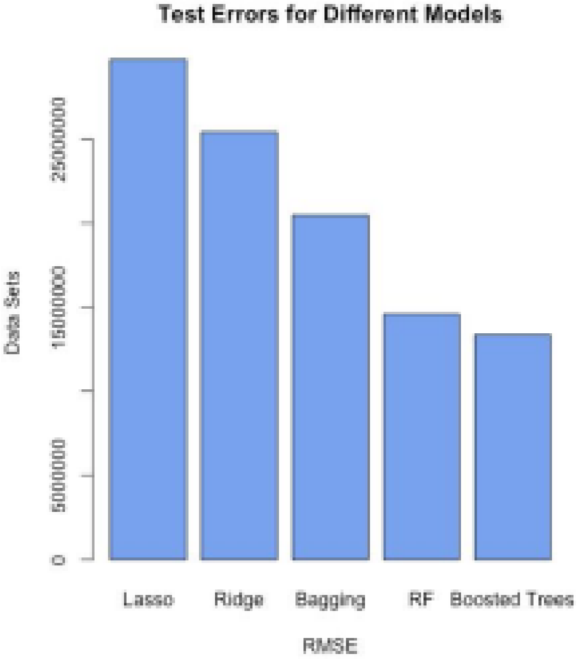
Random Forest depends on the number of trees, and number of variables selected at each node, and therefore, a trial and error is required. The larger the number of trees, the less overfitting will occur. The algorithm performed well with 90.5% accuracy level on the test dataset. Moreover, the model's ability to correctly identify the true positive rate (sensitivity) and correctly identify the true negative rate (specificity) were 90.28% and 90.67% respectively. The results suggest a high level of accuracy estimation, and has higher reliability of predicting litigation risks than K-NN and SVM algorithms.

Random Forest model estimates an average 55.6% probability that MOHAWK INDUSTRIES will NOT be subject to litigation. The average probability is estimated when applying the model to two distinct datasets, one containing 229 risk factors, and the other set containing only 62 factors that take into account removing high multicollinearity. In both datasets, the probabilities that the response variable 'if_sued'==0 are 52.2% and 59.1%. One of the advantages of using Random Forest is its ability to handle high dimensional dataset, robust to outliers and it works best when model interpretability is not critical to the organization.

Part II. Lawsuit Severity (Settlement Amount)

The small size of the data sample discussed in the previous sections is problematic in terms of model learning and outcome validity. There are only 37 companies that are used to train and test the model, a ratio that is not sufficient for most predictive algorithms.

The bar plot below shows the test error rate for the different machine learning algorithms used to predict the severity on the test dataset. Notice that random forest and boosted tree models provide the least Root Mean Square Error (RMSE) rate.

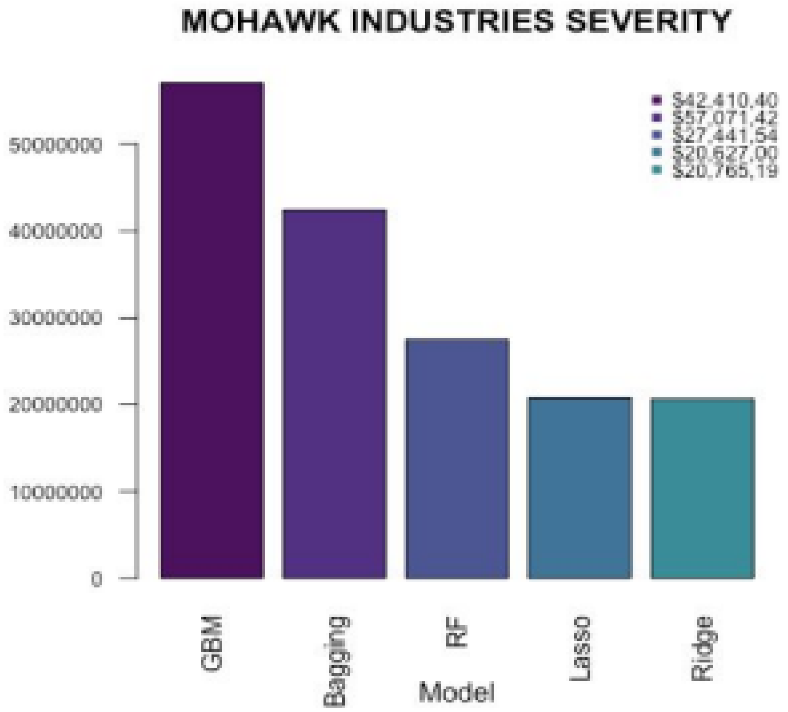


Barplot - Test Error for Different Models

Lasso's regularization performed relatively better than ridge, providing lower RMSE test error. Lasso has predicted that MOHAWK INDUSTRIES would pay approximately $20,627,000 in litigation settlement, while Ridge estimated $20,765,195. The performance of both regularization techniques provide plausible estimation of severity given that the model complexity is reduced, and features selection is performed.

Below is a summary of the severity predictions that MOHAWK INDUSTRIES would potentially pay for legal action settlements, per model used. The values range from $20,000,000 to over $57,000,000. Gradient boosted machine predicted the highest settlement amount to be above $56,000,000 dollars, while Lasso predicted over 20,000,000 dollars in severity. If taking into account the average settlement amount from all models, the predicted median severity would be approximately $27,441,548 U.S dollars in litigation losses. To validate this average, the ratio of market value to settlement amount is calculated for all companies in the dataset. The mean ratio across all companies is 0.02776% and the median is 0.01030%. These values reflect the percentage of settlement amount in relation to the fiscal market value of the company, providing a big picture insight into the range of severity.

There is a strong positive correlation of 85% between market capitalization and the ratio of settlement amount to that market value. MOHAWK INDUSTRIES has approximately $5186.407 billions dollars in fiscal market value, while the average severity calculated is $27,441,548 U.S dollars. The ratio of market value to severity for MOHAWK INDUSTRIES is estimated around 0.0189%, which is within the range of the median ratio across all companies.



Plot: Settlement Amount Prediction per Model (U.S Dollars)

CONCLUSION & RECOMMENDATIONS

Part I. Lawsuit Binary Classification (Yes, No Outcomes)

The below table shows class prediction for each of the ML algorithms. When the averaging of the results from the three models is taken into consideration, it's plausible to state that all models estimate that MOHAWK INDUSTRIES will NOT be subject to legal actions, with the exception of the K-Nearest Neighbor algorithm.

Accuracy	79.00%	88.40%	90.50%
Sensitivity	80.40%	82.00%	90.30%
Specificity	78.50%	93.30%	90.60%
ROC/AUC	79.00%	87.00%	88.40%
MOHAWK INDUSTRIES Prob. of Litigation (Yes)	TRUE	FALSE	44.40%
MOHAWK INDUSTRIES Prob. of Litigation (No)	FALSE	TRUE	55.60%

Table: Metrics of KNN, SVM, RF

Random Forest is able to achieve the highest prediction results of litigation risk for the binary classification problem at hand, which has large implications for MOHAWK INDUSTRIES. Since the analysis is more concerned with finding the true positive rate, as it conveys severe financial and regulatory risks, random forest sensitivity along with accuracy should be taken into consideration as reliable outcomes can be derived from it. Risk factors such as short term total investments, retained earnings, stock sale, total returns, marketable securities, cashflow, and capital expenditures among many others must be reviewed if necessary to reduce the chances of legal actions.

Part II. Lawsuit Severity (Settlement Amount)

The final dataset used to predict the settlement amount did not have enough instances to yield more accurate and valid severity results. However, averaging variations of algorithms is a good approach in order to produce an acceptable generalization as a baseline for additional further analysis in the future. MOHAWK INDUSTRIES is advised to review and focus on risk factors including its market value losses and gains, amount of debt, equity, cashflow, income tax, increase and monitor government credit rating, inventory, and stock data. The settlement amount that the company would pay ranges from $20 to $57 million US dollars, with an accumulative mean of $33,663,115 and median $27,441,548 US dollars.

Future analysts using this dataset are recommended to perform the analysis with Deep Neural Networks to increase classification and regression accuracies, and PCA to reduce multicollinearity and feature dimensions. A better understanding and knowledge of the available data is critical to derive sound conclusions and recommendations.

REFERENCES

Errickson, Josh. “Multiple Imputation.” *Umich*, dept.stat.lsa.umich.edu/~jerrick/courses/stat701/notes/mi.html. Accessed 27 June 2020.

Glen, Stephanie. “Median Absolute Deviation.” Statistics How To, www.statisticshowto.com, 26 Aug. 2019, www.statisticshowto.com/median-absolute-deviation.

“SMOTE Function | R Documentation.” RDocumentation, www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE. Accessed 28 June 2020.

“RBF SVM Parameters.” Scikit-Learn.Org, scikit-learn developers, scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html. Accessed 28 June 2020.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning : Data Mining, Inference, and Prediction (PDF) (Second ed.). New York: Springer. p. 134.

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

“Model Selection.” Amazonaws, rstudio-pubs-static.s3.amazonaws.com/54576_142b58255f8944c990c5663290d28517.html. Accessed 29 June 2020.

kassambara . “Penalized Regression Essentials: Ridge, Lasso & Elastic Net.” STHDA, sthda.com, 11 Mar. 2018,

www.sthda.com/english/articles/37-model-selection-essentials-in-r/153-penalized-regression-essentials-ridge-lasso-elastic-net.

Z., Zygmunt. “Classifier Calibration with Platt’s Scaling and Isotonic Regression - FastML.” Fastml.Com, 1 Aug. 2014, fastml.com/classifier-calibration-with-platts-scaling-and-isotonic-regression.

Kho, J., 2020. “Why Random Forest Is My Favorite Machine Learning Model. [online] Medium”. Towardsdatascience,

https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706 Accessed 28 June 2020.