CLASSIFYING AND PREDICTING OCCUPATIONAL INCIDENT SEVERITY

Author(s)
Ms. Fatemeh Davoudi Kakhki, Iowa State University, Ames, IA
Dr. Steven Freeman, Iowa State University, Ames, IA
Dr. Gretchen Mosher, Iowa State University, Ames.

Abstract
The focus of the study is to build a predictive model and assess its performance in classifying the financial severity of occupational injuries in agribusiness bulk storage facilities (i.e. grain elevators). The data specifically look at food processing and feed milling operations within these facilities. The severity of occupational injuries is determined by the total dollar amount incurred on medical costs, indemnity costs and other expenses in workers' compensation claims. The data is available from an agribusiness insurance provider in Midwest USA. First, the most important independent variables that affect the total cost of claims are extracted from the original dataset. The claims cost variables are then applied as input factors in constructing a classification decision tree and random forests trees with the claims classified as severe and non-severe. Claims over ten thousand dollars are considered severe while those with zero to ten thousand dollars are classified as non-severe. For the purpose of balancing model overfitting and prediction accuracy, the data is partitioned to training, validation and test sets. The results show that the decision tree and random forests trees have accuracy rates of 94% and 93% respectively in predicting that a future claim will be classifies as severe or non-severe based on characteristics of the injury. In addition, incident location and injured worker demographics do not have a significant effect on predicting claim severity. The presented model identifies higher injury risk groups and prevalent causes of incidents in work environments, allowing a more focused intervention effort in the food processing and feed milling sectors. In addition, it is applicable in forecasting cost of future claims and identifying factors that contribute to escalation of claims costs.
Keywords: Classification Decision Tree, Random Forests Trees, Occupational Injuries

Introduction

Occupational incidents are a major problem in agribusiness industries. The data from a major insurance company includes more than 6,000 incidents in the food processing and feel milling sectors which incurred loss over 18 million dollars from 2008 to 2016. In the U.S., workers' compensation coverage has been in place for more than 100 years (Baldwin & McLaren, 2016). There are three main types of workers' compensation claims: medical only, temporary disability, and permanent disability, among which the greatest costs are imposed by permanent disability. The most common claims are "medical only" even though they represent a small share of the overall payments. Sources of workers' compensation insurance consist of private insurance carriers, state funded, or self-insured (Baldwin & McLaren, 2016).

The purpose of this study is to investigate the application of classification as a machine learning technique in predicting the financial severity of occupational incidents in food processing and feed milling in grain elevators. The confusion matrix is used to assess model prediction accuracy. Finally, the classification tree is used to explain the factors that lead to severe and non-severe incidents. This will contribute to identifying higher injury risk groups and determining more prevalent causes of incidents in work environments to focus on intervention efforts.

Decision Trees for Classification Purposes

A decision tree is a commonly used methodology for building classification systems based on multiple covariates for the development of a predictive model for a target variable (Lu & Song, 2015). Decision trees are among the most popular predictive analytics techniques among practitioners due to being relatively straightforward to build and understand, as well as handling both nominal and continuous inputs (Abbott, 2014). Other advantages of decision tree classification methods include the support for multi-level classification and nonlinear classification capability. Some important examples of decision trees include random forest and gradient boosted trees as they are considered as the best classifiers (Cui, Chen, He, &

Chen, 2015). Tree algorithms simply split the dataset hierarchically and can be applied as a replacement for logistic or multiple regression and ANCOVA (Lavery & Mawr, 2012). According to SAS Institute (2016), in classification trees where the response variable is categorical, the decision criteria for choosing the best split is the likelihood ratio chi-square and node splitting is based on the LogWorth statistics which is defined as [-log10(p-value)]; where the p-value is calculated so that it takes into account the number of different ways splits can happen. The calculation includes an unadjusted p-value, which supports input variables with many levels, and the Bonferroni p-value, which favors input variables with small number of levels. The optimal split is the one that maximizes the LogWorth.

Random Forest Trees

The Random Forests (RF) method is a machine learning technique that is useful in prediction problems (Bharathidason & Venkataeswaran, 2014). The RF method has a set of characteristics that makes it advantageous (Polley, Goldstein, & Briggs, 2011). As a powerful data driven method, random forest is non-parametric, has high predictive accuracy, and determines variable importance which contributes to better understanding of the individual role of each input factor (Rodriguez-Galiano, Mendes, Garcia-Soldado, Chica-Olmo, & Ribeiro, 2014). RF trees consist of a collection of arbitrary simple trees used to determine the final outcome. According to (Grömping, 2009), RF trees are random since a subset of the observations is used to build each individual tree, and also each split within each tree is created based on a subset of input variables, not all. As a large number of trees is made, the overall prediction of the forest is the average prediction of all individual trees. In classification, the ensembles of simple trees vote for the most popular class while in regression problems, the responses are averaged to obtain an estimate of the dependent variable. Applying the RF method will significantly improve the prediction accuracy (Hill & Lewicki, 2007). Using RF trees, the input variables that are significant in predicting the response variable are also identified.

Materials and Methods

Predictive modeling is the adopted methodology for this research. Predictive modeling is the use of data to forecast future events by relying on capturing relationships between explanatory variables and predicted variables from past events and applying them to predict future outcomes (Frees, Derrig, & Meyers, 2014). Although predictive modeling is heavily dependent on statistics, the major fundamental difference is that in statistics, a model is used to test a set of hypotheses, while in predictive analytics, data mining is done by building nonparametric and distribution-free models. (Abbott, 2014). There are various techniques in applying predictive modeling in a given dataset. In this research, classification via random forest trees and classification trees is done. The target variable is claim severity which has the binary classification of severe (S) and non-severe (NS). The reason for such classification is that the severity of claims in insurance analytics is determined based on the total dollar amount which is incurred on medical costs, indemnity costs, and other relevant expenses. For claims with the total amount between zero to ten thousand dollars, the level is considered non-severe (NS) and claims with cost over ten thousand dollars are considered as severe (S).

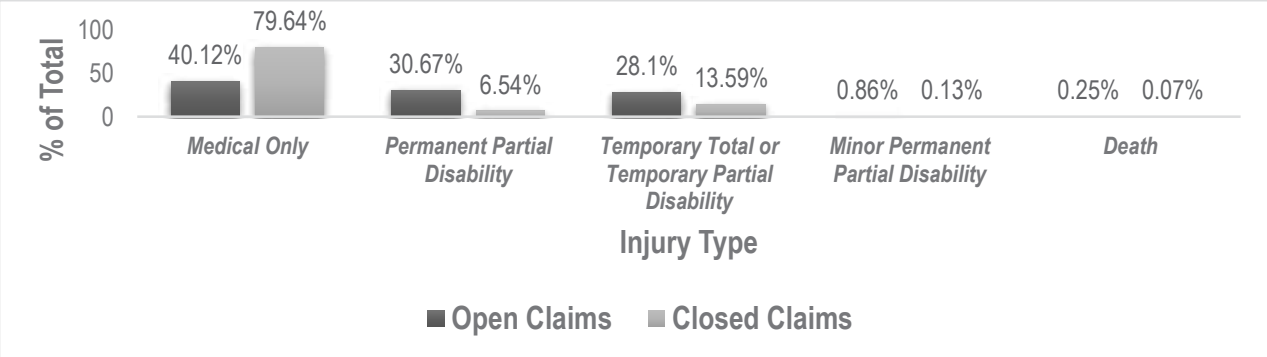
According to Mattew (2016) in classification methods, confusion matrix is the basis of the predictability power of the model. A confusion matrix shows the correct and incorrect number of cases classified under a defined target. It is used to calculate the accuracy of the prediction (See Table 1).

Table 1: Confusion Matrix

Test	Predicted N (Negative)	Predicted P (Positive)
Observed N (Negative)	True Negatives (TN)	False Positives (FP)
Observed P (Positive)	False Negatives (FN)	True Positives (TP)

Data Processing

The data shows the insurance company has had a loss of 18 million dollars over eight years. The amount is paid on both claims that are closed, and those open which will continue to cost for the parties involved. In more than 6000 incidents, 87% have closed claims, and 13% of claims are open. However, 60% of the total amount incurred is paid on open claims and only 40% on closed claims. In this study, all claims (both open and closed) are analyzed and predictive models are built to forecast the probability of a claim ending in severe or non-severe based on workers' compensation information. Injuries are categorized in five groups: medical only, permanent partial disability, temporary total or temporary partial disability, minor permanent partial disability and death. The distribution of types of injuries in this dataset for closed and open claims are shown in Figure 1.



Response and Input Variables

The variable of interest in this study is the outcome of a claim which is determine by the total amount paid on expenses, medical costs and indemnity costs. Looking at this total amount, claims are categorized as severe (S) and non-severe (NS). This research focuses on application of data mining in predicting that a claim will be classified as either S or NS based on the workers' demographics, incidents location, and characteristics of the injury in open claims dataset. The variables that are used as input variable from the original dataset are shown in Table 2.

Table 2: Description of the Independent Variables in the Dataset

Input Variable	Variable Type	Input Variable	Variable Type
Age of Worker	Continuous	Injured Body Part	Categorical
Tenure of Worker	Continuous	Cause of Injury	Categorical
Type of Injury	Categorical	Nature of Injury	Categorical
Occupational Class Code	Categorical	Injured Body Group	Categorical
Incident State	Categorical	Cause Group of Injury	Categorical
Gender of Worker	Categorical	Nature Group of Injury	Categorical

Partitioning Data

Data for this analysis is divided into three parts: training set, validation set, and test set. The training set includes 60% of the data points. This set is used to fit the model of interest and estimate model parameters. The validation set includes 20% of the data points. The model fitted to the training set is applied into the validation set to assess the predictive ability of the model that is useful for selecting the better model. Finally, the test set, includes 20% of data points that have not been used in the training or the validation sets, and is used to assess the generalization error of the final model. The decision about the usefulness of a predictive model is made based on the performance of the model in the test set only.

Results & Discussions

Two types of analyses are done. First, a decision tree was built aiming at classifying the binary S/NS response with all independent variables from Table 2. Second, the RF method was applied to classify the response. The results on all training, validation and tests are presented. Finally, both models are compared and assessed.

The Classification Decision Tree Summary Analysis

The results from this analysis indicate that the most influential predictor of a claim severity is type of injury. Claim status (open/closed), nature of injury, cause group of injury, and injured body group are other important factors respectively. The details of the analysis are shown in Table 3. The accuracy of the predictive model is determined from the confusion matrix for each set. The model did well on all training, validation and test sets. The overall accuracy rate of the test set indicates that the model can correctly classify and forecast future claims severity in almost 94% of the cases based on injury type, claim status, nature of injury, cause group of injury and injured body group.

Table 3: Analysis Details of Decision Tree

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.5785	0.6099	0.619	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.6843	0.7154	0.7212	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1903	0.1822	0.174	∑ -Log(p[j])/n
RMSE	0.2378	0.2323	0.2247	√∑(y[j]-p[j])^2/n
Mean Abs Dev	0.1136	0.1137	0.1067	∑ |y[j]-p[j]|/n
Misclassification Rate	0.0752	0.063	0.0619	∑ (p[j]≠pMax)/n
Prediction Accuracy	92.48	93.7	93.81	(TN+TP)/(TN+TP+FN+FP)
N	3682	1253	1243	n of total observations in each set

Random Forests Trees Summary Analysis

In this part, the random forests tree method is applied to the data. The analysis indicates the most contributing factors in predicting the target of claim severity are type of injury and claim status (open/closed). Incident state, cause of injury, body part injured, and class code are less significant contributors. The analysis details are shown in Table 4. The prediction accuracy rates of the random forests trees on all training, validation and test sets are close to those of the decision tree analysis in section 0. Although the prediction accuracy is high, the variables that are selected as the most contributing ones have too many levels which makes interpretation of the trees difficult and tedious. In data mining, it is regular to repeat some steps many times to make the final decision of a model (Hidayatul Qudsi, Kartiwi, & Binte Saleh, 2017). Simple interpretation of a model is as important as its applicability. Thus, the random forests tree analysis is done again with the most contributing variables in the decision tree analysis method. Body group and cause group of injury have fewer levels and make the interpretation easier. The results show that the new model accuracy is a bit lower (training set 87.8%, validation set 87.15%, test set 88%).

Table 4:Analysis Details of Random Forests Trees

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.6944	0.5647	0.5683	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7833	0.6752	0.6762	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.138	0.2033	0.1972	$\sum$ -Log(ρ[j])/n
RMSE	0.1879	0.2439	0.238	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.108	0.1436	0.1385	$\sum$ |y[j]-ρ[j]|/n
Misclassification Rate	0.0378	0.075	0.0668	$\sum$ (ρ[j]≠ρMax)/n
Prediction Accuracy	96.22	92.5	93.32	(TN+TP)/(TN+TP+FN+FP)
N	3682	1253	1243	n of total observations in each set

Model Comparison

Both models perform well in predicting the severity level of the claim based on type of injury, claim status, injured body group, cause group of injury, and nature of the injury. The final decision tree shows the following:

- Open claims are all predicted to end as severe claims.
- Medical injuries have 0.78 probability of ending non-severe, while minor/major permanent or temporary partial/total disabilities or death have 0.97 probability of ending as severe.
- Medical injuries caused by burn or scald, heat or cold exposure, cut, puncture, strain, fall, trip and motor vehicles have almost 100% chance of having a total incurred cost of 0 to 10,000 dollars.
- Minor/major permanent or temporary partial/total disabilities with nature of amputation, inflammation, contusion, crushing, tear or strain have 85% chance of turning severe and costing over $10,000.
- Minor/major permanent or temporary partial/total disabilities with nature of dislocation, fracture, concussion, laceration, hernia, and carpal tunnel syndrome have equal chance of becoming severe or non-severe.
- Temporary partial/total disability Injuries in head and upper extremities body groups are more severe than those in trunk, neck, and lower extremities.

Conclusion

The initial intent of this study is to predict what factors in workers' compensation data affect the financial severity of the claims. The results of this study can be applied in identifying higher injury risk groups and more prevalent causes of incidents in work environments to focus intervention efforts in food processing and feed milling sectors. Future studies will focus on separate analyses of open versus closed claims and medical injuries versus other types of injuries to investigate any other significant patterns in workers' compensation claims.

References

Cui, Z., Chen, W., He, Y., & Chen, Y. (2015). Optimal Action Extraction for Random Forests and Boosted Trees. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 179-188). ACM.

(1993-2017). Retrieved from MEDCALC: https://www.medcalc.org/manual/roc-curves.php

Abbott, D. (2014). Applied Predictive Analytics- principles and Techniques for Professional Data Analyst. Indianapolis: John Wiley & Sons, Inc.

Appel, R., Fuchs, T., Dollar, P., & Perona, P. (2013). Quickly Boosting Decision Trees-Pruning underachieving Features Early. Proceedings of the 30th International Coonference on Machine Learning. Atlanta, Georgia: JMLR:W&CP .

Baldwin, M. L., & McLaren, C. F. (2016). Workers' Compensation: benefits, Coverage, and Costs(2014Data). National Academy of Social Insurance.

Bharathidason, S., & Venkataeswaran, C. J. (2014). Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees. International Journal of Computer Applications, 26-30.

Frees, E. W., Derrig, R. A., & Meyers, G. (2014). Predictive Analytics in Actuarial Science. Cambridge University Press.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 308-319.

Hidayatul Qudsi, D., Kartiwi, M., & Binte Saleh, N. (2017). Predictive Data Mining of Chronic Diseases Using Decision Tree: A Case Study of Health Insurance Company in Indonesia. International Journal of Applied Engineering Research, 1334-1339.

CLASSIFYING AND PREDICTING OCCUPATIONAL INCIDENT SEVERITY

10

Hill, T., & Lewicki, P. (2007). Statistics: Methods and Applications. Retrieved from http://documents.software.dell.com/statistics/current/textbook

Hill, T., & Lewicki, P. (2007). Statistics: Methods and Applications. Tulsa, OK: Dell. Retrieved from http://documents.software.dell.com/statistics/current/textbook

Lavery, R., & Mawr, B. (2012). An Animated Guide: Regression Trees in JMP® & SAS® Enterprise Miner™. Proceedings of the 25th Annual Conference of Northeast SAS Users Group (NESUG), (pp. 11-14).

Lu, Y., & Song, Y. (2015). Decision Tree Method: Applications for Classification and Prediction. Biostatitics in Psychiatry, 130-135.

Mattew, D. G. (2016). Data Mining and Machine Learning Algorithms for Workers' Compensation Early Severity Prediction.

Polley, E. A., Goldstein, B. A., & Briggs, F. B. (2011). Random Forests for Genetic Association Studies. Statistical Applications in Genetics and Molecular Biology.

Rodriguez-Galiano, V., Mendes, M., Garcia-Soldado, M., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment, 189-206.

SAS Institute. (2016). JMP Statistical discovery From SAS. Retrieved from © SAS Institute Inc.: http://www.jmp.com/support/help/K_Nearest_Neighbor.shtml
The MathWorks, I. (2016). mathworks.com. Retrieved from www.mathworks.com