SafeTree: Random Forest Regression for Predicting Expected Return to Work Following an Injury

Morgan Jacobus, Kellie Heom, Vishnu Tanguturi,
Mu-Fan (Leo) Weng
College of Engineering
Georgia Institute of Technology
Atlanta, GA, USA
{mjacobus,kheom3,vtanguturi3,mweng6}@gatech.edu

ABSTRACT

Work related injuries are a cause of significant stress and economic burden to employees and employers respectively. SafeTree will help people benchmark their recovery from work related injuries by predicting and visualizing average return to work time (RTW). By estimating days away from work, decision makers will be able to wisely allocate healthcare resources and reduce stress levels for injured individuals.

KEYWORDS

Random forests, regression, occupational injury

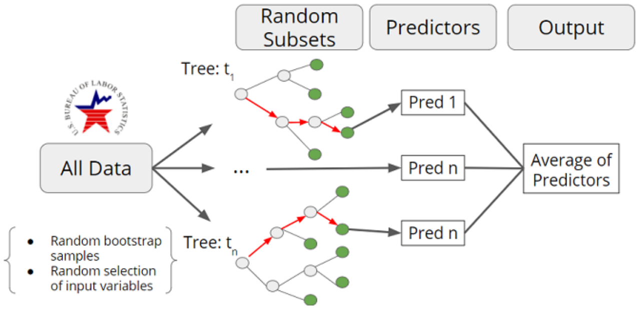


Figure 1: Random Forest Regression Model

ACM Reference format:
Morgan Jacobus, Kellie Heom, Vishnu Tanguturi, Mu-Fan (Leo) Weng. 2017. SafeTree: Random Forest Regression for Predicting Expected Return to Work Following an Injury . In *Proceedings of Data and Visual Analytics, Georgia Tech, 2017 (CX 4242)*, 7 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CX 4242, 2017, Georgia Tech

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Job related injuries are a cause of significant stress to those injured as well as economic burden to employers. It is estimated that in 2007 there were more than 5,600 fatal and 8,559,000 nonfatal occupational injuries which incurred costs of $6 billion and $186 billion. While most injury victims return to work the next day, nearly 30% of these injuries resulted in anywhere from 1-4 days away from work to even permanent disabilities (partial or total)[7]. The ability to estimate days away from work following occupational injury would greatly assist decision makers attempting to wisely allocate healthcare resources, as well as alleviate the stress of injured individuals by informing them about their expected recovery time, all based on other incidents of similar nature found in historical records.

SafeTree is a tool that uses a machine learning algorithm to estimate median return to work times of employees suffering from work related injuries. At the core of SafeTree is a random forest algorithm, as described by Brieman[2], to build a regression model that learns from a library of incidents to relate characteristics, such as age, gender, state, nature of incident, part of body, to days away from work. SafeTree also contains a visualization of the existing dataset on a United States Map.

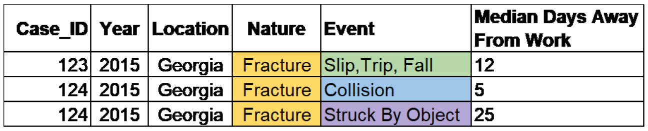
2 METHOD

Currently, there is no interactive system to visualize return to work. Previous studies [8][13][11] identify factors that contribute to safe work environments, however, they are limited to small cohorts of data. SafeTree contains a user-friendly interface to clearly visualize data from Bureau of Labor Statistics (BLS) and model results. This tool reveals good and bad locations or industries to work in by allowing users to explore days away from work across the country. The insights revealed will also allow employers to assess work safety and make improvements if necessary. The prediction functionality will allow employers and employees to agree on an appropriate recovery time and address our goal of alleviating stress related to work related injuries.

2.1 Data Collection and Cleaning

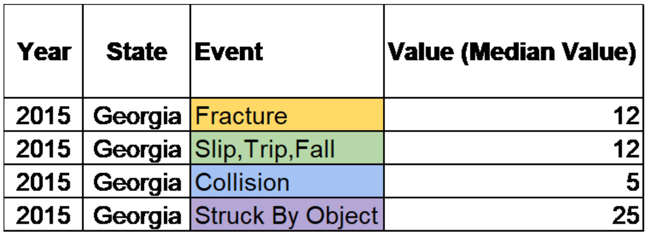
We accessed our data via FTP from the BLS database entitled "Nonfatal cases involving days away from work: selected characteristics (2011 forward)". We chose this method over using the less scalable Public Data API that requires continuously accessing every ID in the dataset. The database files contain summary data characterizing occupational injuries and illnesses based on survey data from both the employers and employees injured. BLS also provided breakouts by demographics for these survey data, such as age, category, event, and occupation.

The database was structured into three parts: a time series file, mapping files, and data files. The files were loaded into an SQLite database and joins were performed using the mapping files. We found the BLS data did not include the survey response data as individual records. Rather it was provided in a rollup of all survey responses containing those characteristics. To illustrate, we expected to find data in the format shown in Figure 2, however, the data was presented in the format shown in Figure 3.



Case_ID	Year	Location	Nature	Event	Median Days Away From Work
123	2015	Georgia	Fracture	Slip,Trip, Fall	12
124	2015	Georgia	Fracture	Collision	5
124	2015	Georgia	Fracture	Struck By Object	25

Figure 2: Expected and Desired Data Format



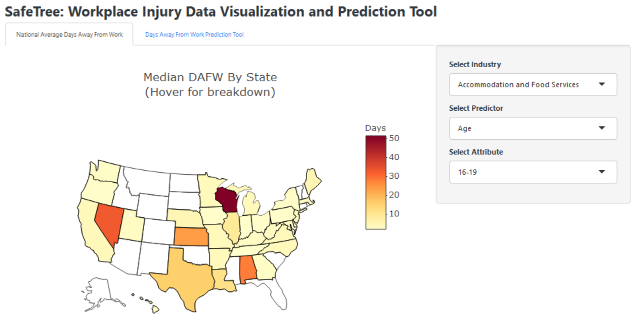
Year	State	Event	Value (Median Value)
2015	Georgia	Fracture	12
2015	Georgia	Slip,Trip,Fall	12
2015	Georgia	Collision	5
2015	Georgia	Struck By Object	25

Figure 3: Raw Data Format from BLS

In Figure 3, each row represents a time series observation of an aggregated value based on the selected characteristic and time frame. The first row represents the median value of days away from work for all fractures in the state of Georgia for the year of 2015, for example. This is not the desired output for RF regression, so we applied data cleaning techniques using OpenRefine and created separate columns for each category (age, part of body, etc.). Further, the industries in the raw database were too specific. To provide SafeTree with enough data to be tested on, we aggregated many records by categorizing the raw data into the much broader NAICS Industry categories. Following data transformation, our data resembles that shown in Figure 2, a format that can be implemented in RF algorithm.

2.2 Data Visualization

Prior to applying the RF model, we visualized the raw data by exporting results from the SQLite query and using the R and the shiny package. Our results are shown in Figures 4.



SafeTree: Workplace Injury Data Visualization and Prediction Tool

Select Industry: Accommodation and Food Services

Select Predictor: Age

Select Attribute: 16-19

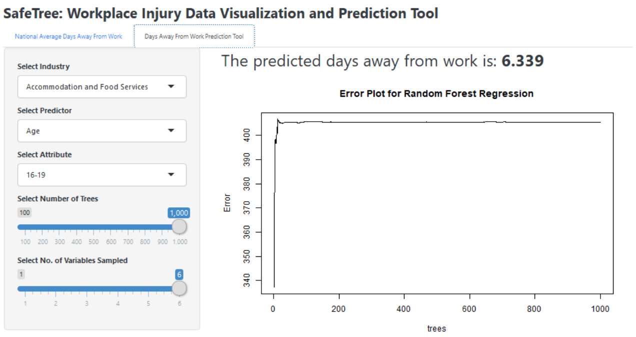
National Average Days Away From Work
Days Away From Work Prediction Tool
Median DAFW By State
(Hover for breakdown)
Days
50
40
30
20
10

Figure 4: Median RTW time by state, shown on a US map

We chose to build our visualization in R due to the wide availability of packages, its simplicity with manipulating large amounts of data, and its ability to display user-friendly filters. (No need to implement a choropleth map from scratch by choosing R over D3.) Using R allowed us to focus our efforts toward the RF model and visualization techniques rather than implementing the whole server-side technology framework from scratch.

2.3 User Interface

There are two main features of our user interface. First, we display the cleaned dataset for users to explore by applying filters. Second, we have built a "calculator" that predicts RTW based on indicated attributes. Users can switch between the two tabs shown in Figure 4 to change functionalities.



SafeTree: Workplace Injury Data Visualization and Prediction Tool

Select Industry: Accommodation and Food Services

Select Predictor: Age

Select Attribute: 16-19

Select Number of Trees: 100 – 1,000

Select No. of Variables Sampled: 1 – 6

The predicted days away from work is: **6.339**

National Average Days Away From Work
Days Away From Work Prediction Tool
100 200 300 400 500 600 700 800 900 1,000
1 2 3 4 5 6
Error Plot for Random Forest Regression
Error
340
350
360
370
380
390
400
trees
0
200
400
600
800
1000

Figure 5: RTW time prediction tool, along with error plot of the regression

2.3.1 *Dataset Visualization.*

As shown in Figure 4, an interactive US map shows the average RTW for each state on a color scale. Users can choose to filter the data by industry, predictor, and value. Options for industry include those listed in the NAICS, such as Construction, Retail Trade, and Public Administration. The predictors are each names of attributes that contribute to RTW in our model, such as the victim's age, gender, or occupation, the event that occurred, the nature of the incident, the part of body injured. Values are the specific descriptors that characterize the incident. For example, the predictor "age" has value options of "0-18", "19-30", and so on.

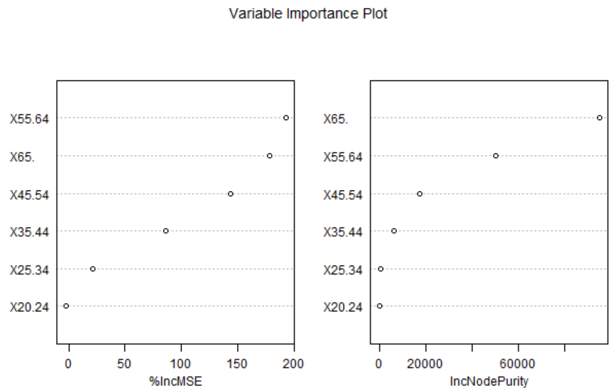


Figure 6: Variable importance plot, which is included under the RTW time prediction tool in the previous figure

2.3.2 Random Forest Regression. As shown in Figure 5, users specify a set of characteristics on the front-end, and we apply the trained RF algorithm to predict a RTW and send the corresponding data to R to produce a visualization. For each predictor, we trained the RF algorithm on 2/3 of the entire dataset. For each predictor, we constructed an individual random forest. This individualization was necessary because the BLS data was aggregated and not given in a more convenient record-by-record format. After constructing the random forest, we applied rudimentary boosting techniques (more erring on the side of data cleaning) to enhance the accuracy while using the testing data (remaining 1/3 of dataset). We initially wanted to use the XgBoost algorithm, however the dataset being only one predictor at a time and not being in record format made it difficult to accurately boost the forest. Successful boosting methods included a weighting of the attribute importances, and we were able to reduce the average error by 1.5%. However it was not practical to apply these methods in the final application framework, because the computing time increased significantly without significantly improving the statistically significance of the prediction values. Therefore, we simply trained the random forest model and predicted output based on the choices selected by the user. In addition to the prediction, we output the importance of each attribute as well as a plot depicting our model's error changing over the number of trees used in constructing the random forest. A sample plot is shown in Figure 6.

3 EXPERIMENTS AND EVALUATION

We designed experiments to test the accuracy of our RF model. First, we measured the accuracy of our model using two metrics, out-of-bag (OOB) error and test error. Both metrics summarize the accuracy of the RF model in predicting RTW, but OOB uses data points randomly selected from the training test, and test error uses data points in the testing set, which it has not "seen" before. We repeated this test for all attributes.

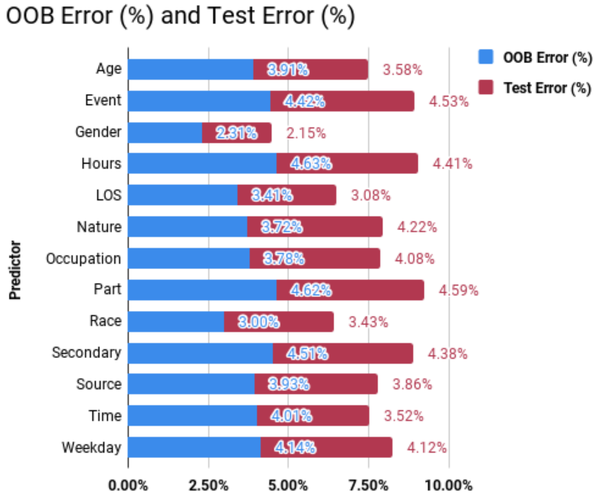
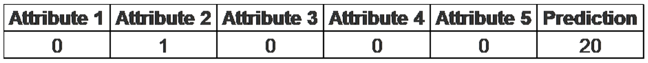


Figure 7: Table of OOB error and test error from our predictions

The results of our experiments indicate that OOB and test error was <5% for all predictors, as shown in Figure 7. Interestingly enough, race and gender were found to be the most accurate predictors of RTW (with the lowest error around 3%). Another observation (not pictured) is that when RTW itself was used as a predictor, albeit a larger range of values rather than precise values which are more useful to the user, it yields even lower error (about 2%). This is an obvious result because RTW should be the best predictor of RTW. However, it does (in a way) serve as a "positive control" and indicate to us the "lower limit" of error. On average the other predictors are only 2-3 percentage points away from the "ideal" predictor of RTW, which means the other predictors perform pretty well.

In another experiment that we ran, we used a different package, randomForestSRC, to re-develop our model. This package has a higher model development ability. With greater computing ability came greater potential for error: the flaws in the dataset we were working with became clearer. Here, it was even more important to have our data in continuous values, and our lack of data in record format was very hard to make up for. We attempted to clear this hurdle with data refactoring. A lot of work was done, and the best possible format was the following:



Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Prediction
0	1	0	0	0	20

Figure 8: Example input to random forest predictor

However, this process changed our data to be more suitable for a classification random forest rather than a regression model, and this negatively affected our accuracy. Having an unsatisfactory format of the data stemmed from the fact that the original dataset held an aggregated set of values, rather than a record by record split. If it was the latter, then there would not have been a need for a random forest to be generated for each predictor, rather a single random forest would take into account all predictors simultaneously, which would be better for a regression model. Originally, we thought the data could be refactored into this type, however it was inherently impossible to do so, and there was not any way we could work around this situation. Despite these challenges, our prediction model does work, unfortunately however, none of the predictions are to be taken with statistical significance, aside from a few models using gender as the main attribute which are significant.

4 PROJECT TIMELINE

4.1 Project Timeline

All tasks are equally distributed between team members. We used OpenRefine for data cleaning, SQLite for database creation, and R for algorithm creation as well as visualization. Use of Big Data technologies was discussed, but decided against as the scale is manageable on SQLite. We also considered using Tableau for the final application but decided to make use of R and its useful packages to create the application. Figure 9 provides a look at our project timeline. The dark blue bars represent major tasks while the light blue bars represent more specific tasks.

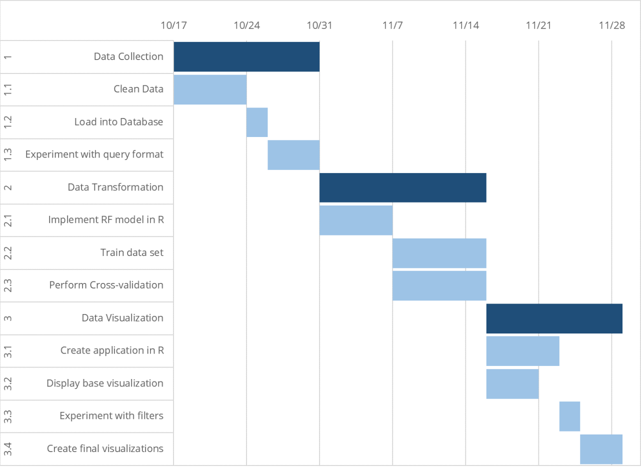


Figure 9: Project Timeline (Gantt Chart)

5 CONCLUSION AND DISCUSSION

By creating the SafeTree application, we have achieved our goals. SafeTree has made the dataset on RTW time and work injuries easily available and interpretable to users through its visualization tool. The random forest predictive tool also provides estimates for RTW time that employers and employees can agree on, which increases transparency following job related injuries. Hopefully, SafeTree will reduce stress and costs related to work related injuries and improve occupational safety America.

In the future, the model can be improved by utilizing a dataset that makes it easier to apply each predictor simultaneously. That way users can maintain a higher statistical significance in the output of the predictor model. Our high prediction errors, as seen in Figure 4, reflect the nonideal format of the data. No matter how we tried refactoring, it was impossible to obtain better accuracies. Also, the range of values that the predictor outputs is heavily constricted. The RF is poor at extrapolating for inputs outside the training set bound.

6 RELATED WORK

6.1 Occupational Injuries

Mackenzie et. al conducted a study in 1998 that followed 312 individuals with lower extremity injuries and attempted to identify predictors of when an employee returns to work after an injury[8]. This study found age, education, poverty status, and availability of support as significant predictors of lengthy RTW[8]. Numerous studies highlight the importance of age and gender in predicting RTW [3][9]. In a study by Vredenburgh, employer practices of training, communication, and rewards for reporting safety hazards were identified as factors that contributed to safe work environments[13]. However, only case studies of hospital employees were included in this study. Jobs across various industries and all demographic features need to be taken into consideration for SafeTree to achieve the same reliability in predictions.

6.2 Random Forests

The random forest (RF) algorithm is a type of ensemble learning that uses groupings of decision trees that vote to determine the most popular outcome. RFs regressors are widely used because they maintain accuracy of outcome values as the complexity of the model increases. This characteristic of RFs places an upper bound on generalization error and prevents overfitting[2]. The accuracy of RFs is achieved by introducing randomness during selection of variables at each node while growing the tree. Additionally, RFs uses bagging to randomly select training sets for the trees to grow on[2]. A depiction of the model is found in Figure 1.

In constructing RFs, it is important to maintain balance in the design and implementation of trees. In one study, the Random Subspace (RS) method was applied to forests of increasing complexity without reduction in general accuracy[5]. The RS method would be useful for designing an effective RF regression model for high amounts of data while maintaining accuracy. However, RS is mainly used for handling redundancy (e.g. image recognition). Due to a wide variety of circumstances for injuries, it might be difficult to categorize issues together using the RS method. In another study by Segal[10], RFs were created with constraints on the node size and number of node splits. Other algorithms to guide the implementation of SafeTree include Bayesian networks, SVMs, or k-NN [6][12].

6.2.1 Bagging. Bagging is a procedure for improving predictor accuracy by generating a sequence of learning sets from the original set by taking random subsets[1]. When used for regressors, the final result is the average result from the learning sets; when used for classifiers, the results from each learning set are aggregated by voting. Bagging might be useful to SafeTree because it improves accuracy of regression by artificially increasing the size of the learning set and produces an output based on the best of multiple results. Bagging in RFs produce continuous estimates of the generalization error as well as enhanced accuracy[2].

6.2.2 Boosting. Boosting is a method of improving the accuracy of a predictor that aggregates a group of moderately accurate predictors. In particular, the AdaBoost algorithm starts with a weak learning algorithm and maintains a set of weights over each member of the training set. Each round, the weak learner forms a hypothesis and the weights across the set are increased for the incorrectly classified examples. As a result, the algorithm improves in accuracy by focusing training on the hard examples in the set. In the end, all the hypotheses are aggregated by computing a weighted average, giving more weight to the hypotheses with the smallest errors[4]. Regular boosting can be extended to multiclass classification by reducing it to a larger binary problem. SafeTree will be constructed using RFs, which applies the concepts of boosting.

REFERENCES

[1] Leo Breiman. 1996. Bagging predictors. *Machine learning* 24, 2 (1996), 123–140.

[2] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (01 Oct 2001), 5–32.

[3] F Curtis Breslin and Peter Smith. 2005. Age-related differences in work injuries: A multivariate, population-based study. *American journal of industrial medicine* 48, 1 (2005), 50–56.

[4] Yoav Freund, Robert Schapire, and Naoki Abe. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* 14, 771-780 (1999), 1612.

[5] Tin Kam Ho. 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* 20, 8 (1998), 832–844.

[6] Taghi M Khoshgoftaar, Moiz Golawala, and Jason Van Hulse. 2007. An empirical study of learning from imbalanced data using random forest. In *Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on*, Vol. 2. IEEE, 310–317.

[7] JPAUL LEIGH. 2011. Economic burden of occupational injury and illness in the United States. *The Milbank Quarterly* 89, 4 (2011), 728–772.

[8] E. J. MacKenzie, J. A. Morris, G. J. Jurkovich, Y. Yasui, B. M. Cushing, A. R. Burgess, B. J. DeLateur, M. P. McAndrew, and M. F. Swiontkowski. 1998. Return to work following injury: the role of economic, social, and job-related factors. *American journal of public health* 88 (1998).

[9] Glenn S Pransky, Katy L Benjamin, Judith A Savageau, Douglas Currivan, and Kenneth Fletcher. 2005. Outcomes in work-related injuries: A comparison of older and younger workers. *American journal of industrial medicine* 47, 2 (2005), 104–112.

[10] Mark R Segal. 2004. Machine learning benchmarks and random forest regression. *Center for Bioinformatics & Molecular Biostatistics* (2004).

[11] Harry S. Shannon, Mayr Janet, and Haines Ted. 1997. Overview of the relationship between organizational and workplace factors and injury rates. *In Safety Science* 26 (1997). https://doi.org/10.1016/S0925-7535(97)00043-X

[12] Kirsten Vallmuur. 2015. Machine learning approaches to analysing textual injury surveillance data: a systematic review. *Accident Analysis & Prevention* 79 (2015), 41–49.

[13] Alison G Vredenburgh. 2002. Organizational safety: which management practices are most effective in reducing employee injury rates? *Journal of safety Research* 33, 2 (2002), 259–276.