SafeTree

Predicting Expected Return to Work Following an Injury

Kellie Heom, Morgan Jacobus, Vishnu Tanguturi, Leo (Mu-Fan) Weng

Summary

SafeTree is a web application that uses a **machine learning** algorithm to estimate median return to work times of employees suffering from work related injuries. Using a **random forest** algorithm, SafeTree builds a **regression model** from past incidents and summarizes past findings and predicted results on a United States map.

What is the Problem?

Work related injuries are a cause of significant stress and economic burden to employees and employers respectively. Predicted figures for recovery time will enable scheduled check-ins for employers and employees as well as better allocation of healthcare resources. Hopefully SafeTree leads to reduced cost of time away from work, reduced stress levels for injured employees, and **improved employer-employee relationships and industry productivity.**

Our Approach

1) Display the historical work related injuries data in a **choropleth** that would allow the user to interactively search in a web application what the median day away from work is by state, industry, predictor, and attribute.
2) Apply user-specified parameters to random forest regression algorithm to **predict the expected days away from work**. Diagnostic charts included in the visualization summarize model accuracy.
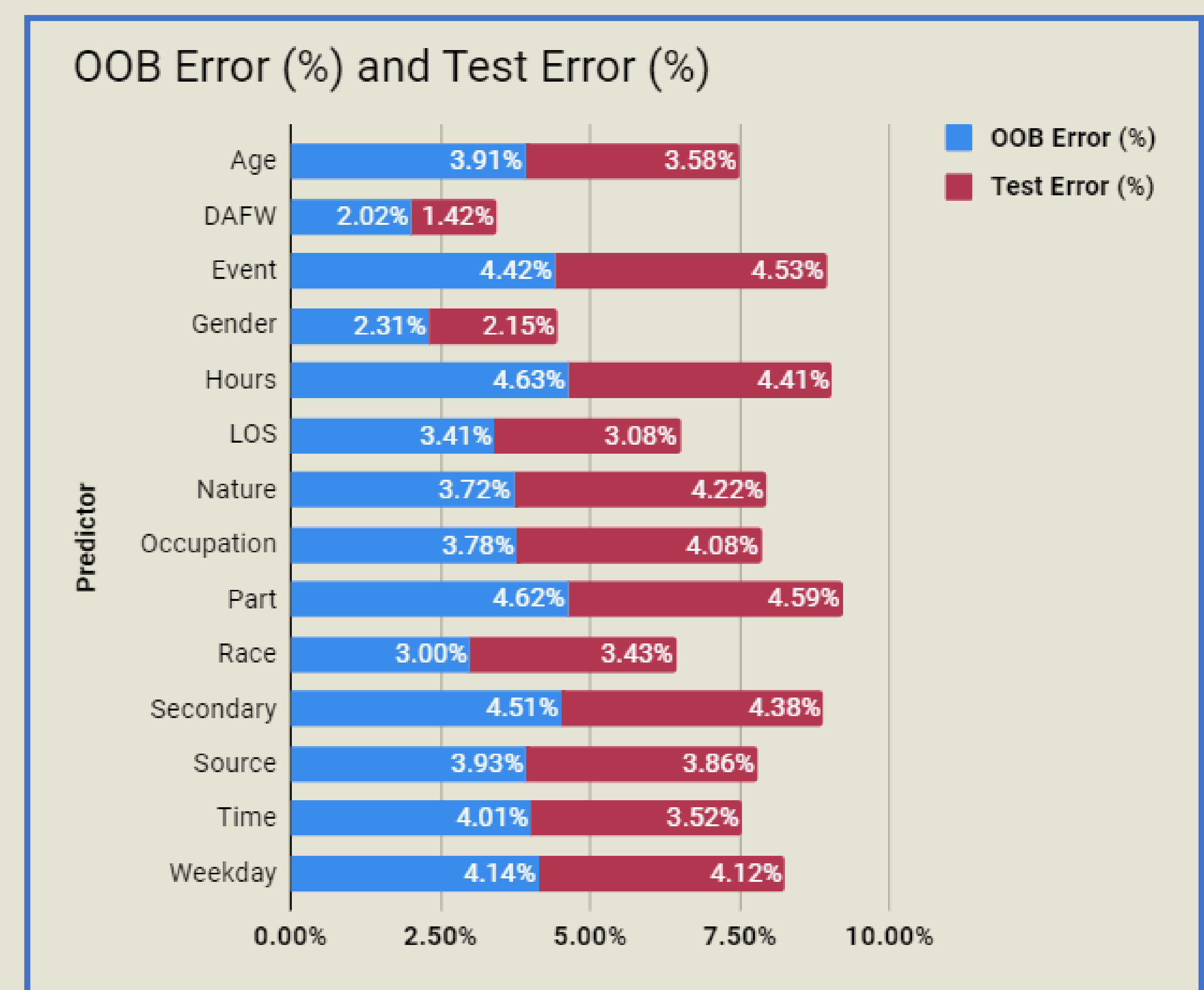
What is Random Forest?

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.

We trained 2/3 of the data with initial experiments based on data subset of Accommodation and Food Services in Georgia, then extrapolated across all industries/states. We used randomForest and XgBOOST to compare both models' test errors and evaluate the importance of predictors.
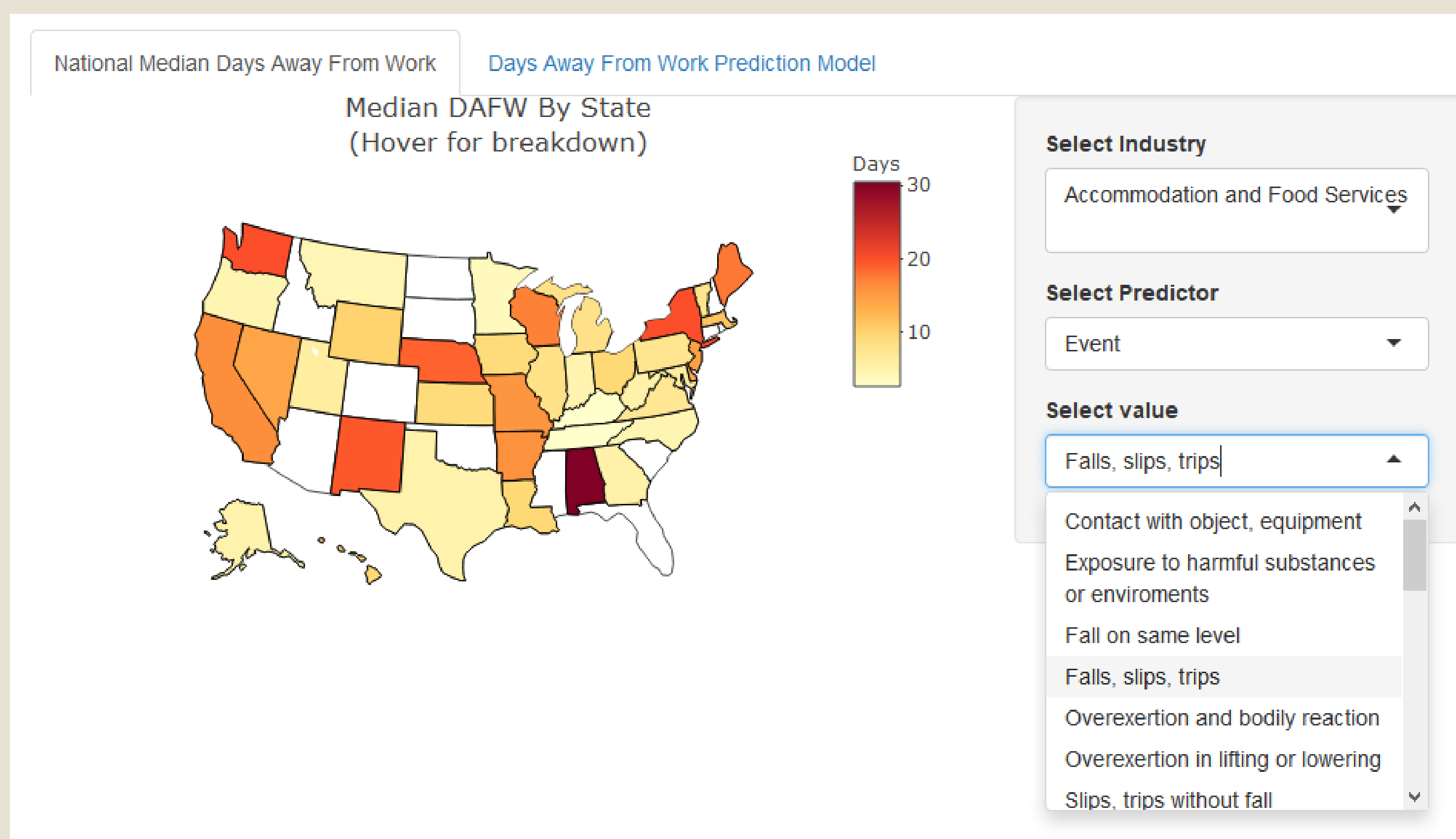
The application is run entirely through the R programming language. R packages such as Shiny, dplyr, randomForest, and party allow us to create interactive GUI's that integrate our data. The code is all housed in a single file and can be published to the web for the general population to access. We obtained the data from the Bureau of Labor Statistics website. The data is contained in a series of 25 text files, one of which contains the documentation. These 25 files were 5.47 GB in size.

OOB Error (%) and Test Error (%)

Predictor	OOB Error (%)	Test Error (%)
Age	3.91%	3.58%
DAFW	2.02%	1.42%
Event	4.42%	4.53%
Gender	2.31%	2.15%
Hours	4.63%	4.41%
LOS	3.41%	3.08%
Nature	3.72%	4.22%
Occupation	3.78%	4.08%
Part	4.62%	4.59%
Race	3.00%	3.43%
Secondary	4.51%	4.38%
Source	3.93%	3.86%
Time	4.01%	3.52%
Weekday	4.14%	4.12%

Solving the Problem

Our interactive visualization allows users explore industry and employee attribute combinations that yield minimal days away from work. By quickly interpreting a color-coded map of the US, employees can find jobs that ensure a **high standard of safety** and employers can find areas in need of improvement in their industry to **minimize the lost days** employees may face. No existing application easily visualizes the BLS data or predicts days away from work. Our approach makes a complicated data set easily available and interpretable to anyone.