Recordable Incidents and Near-Misses in the Workplace

Final Project, IST 718
Kim Greene, Brandon Croarkin, T.S. Yeap, Amanda Sausville

Summary of findings and recommendations

The most significant finding in the data was that nearly 50% of incidents occur from employees who have been working for less than a year with the company. Through the data, we were also able to narrow down which injuries happen the most often (contusions), where they happen in the workplace (rig floor), and which companies are the worst offenders (Company D). With these findings, the most important recommendation would be to move inexperienced employees (0-1 years of employment) away from the rig floor, provide more thorough training and safety measures to new hires, and increase management during high demand times of year, especially in the summer months of June, July and August.

Specification.

Workplace injuries are an unfortunate result of the oil & gas industry that utilizes many powerful, yet dangerous equipment. Improper use of these machines can result in injuries that can permanently maim workers and result in losses in efficiency on these rigs. Preventing injuries is vital in increasing the livelihood of our workers and improving the efficiency, and thus profits, of all the rigs operated.

The goal of this project is to create a predictive model for Incidents and Near-Misses. With this model, the data can be used to find correlations in the data to show factors that are most associated with accidents or near-misses. The data can also show where the incidents take place geographically and which incidents occur the most. This knowledge of what leads to incidents can then be used to craft policies that can decrease these incidents moving forward.

The near miss and incident data sets are from the Intelex safety application that all companies (A - F) utilize to capture information around harmful (incident) or potentially harmful (near-miss) events. The application treats an incident differently than a near miss, therefore their data is not precisely aligned. However, there is some overlap in the dimensional data captured for both and that is what we will focus on in our analysis. Some elements of the data have been redacted or renamed for reasons of privacy: employee names and company names. All companies (A - F) are subsidiaries to a parent company that would like to evaluate these observations to possibly gain insights into how to avoid future incidents.

Observation

There are several different factors that can be concluded from the data in regard to the trends in near misses and incidents. These factors are: time of year, day of the week, how long the employees have been at the company, where injuries occur the most on the job site, which injuries occur the most and at which company.

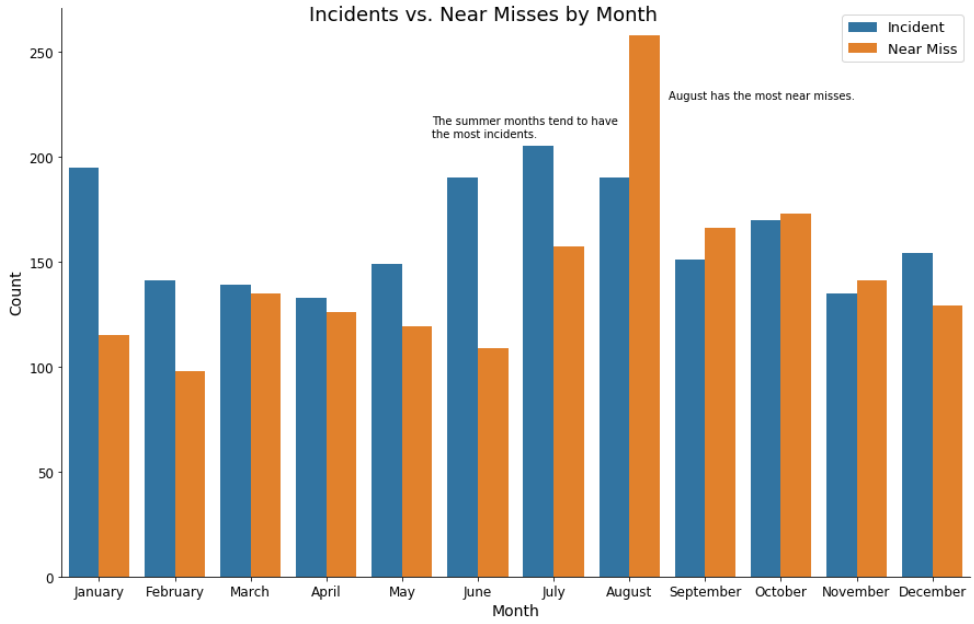

Figure 1: Bar plot of incidents vs. near misses by month

Injuries and Near misses happen most frequently in the summer months of June, July and August. Near misses almost double in August. January also sees a rise in injuries and near misses with the count equalling those of June.

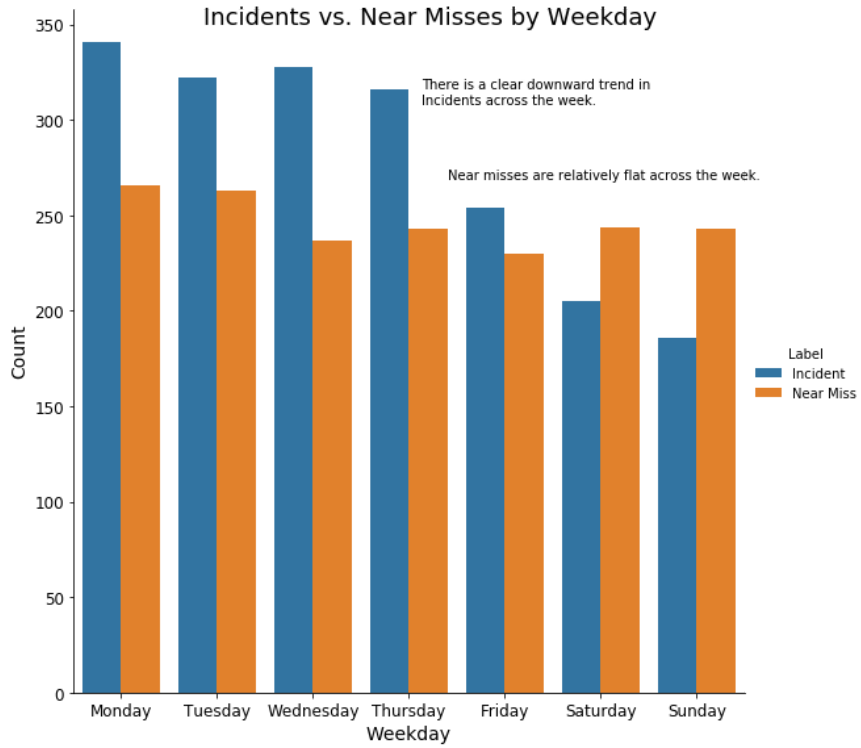

Figure 2: Bar plot of incidents vs. near misses by weekday

Looking at the days of the week that most injuries and near misses occur shows that are most often reported on Mondays. This is followed by Wednesdays and Tuesdays. Near misses were reported most frequently on Monday as well and the data remains relatively flat throughout the week but more were reported on Saturdays and Sundays.

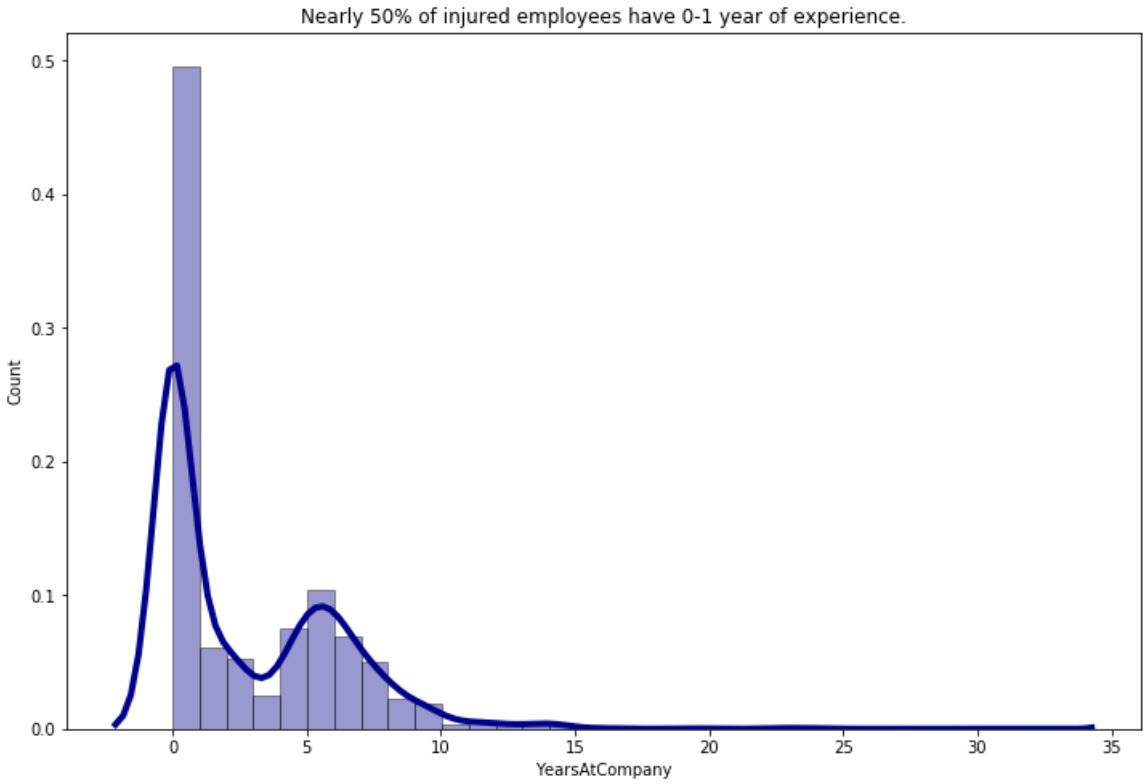

Figure 3: Histogram of years at company for employees who experienced injuries.

It was also found that employee duration played a significant factor in occurrence of injuries and near misses. It was reported that injuries happen most often within the first year of employment. After the first year of employment there is a significant drop in the numbers of reported injuries and near misses. Injuries and near misses reported within the first year make up nearly 50% of all cases.

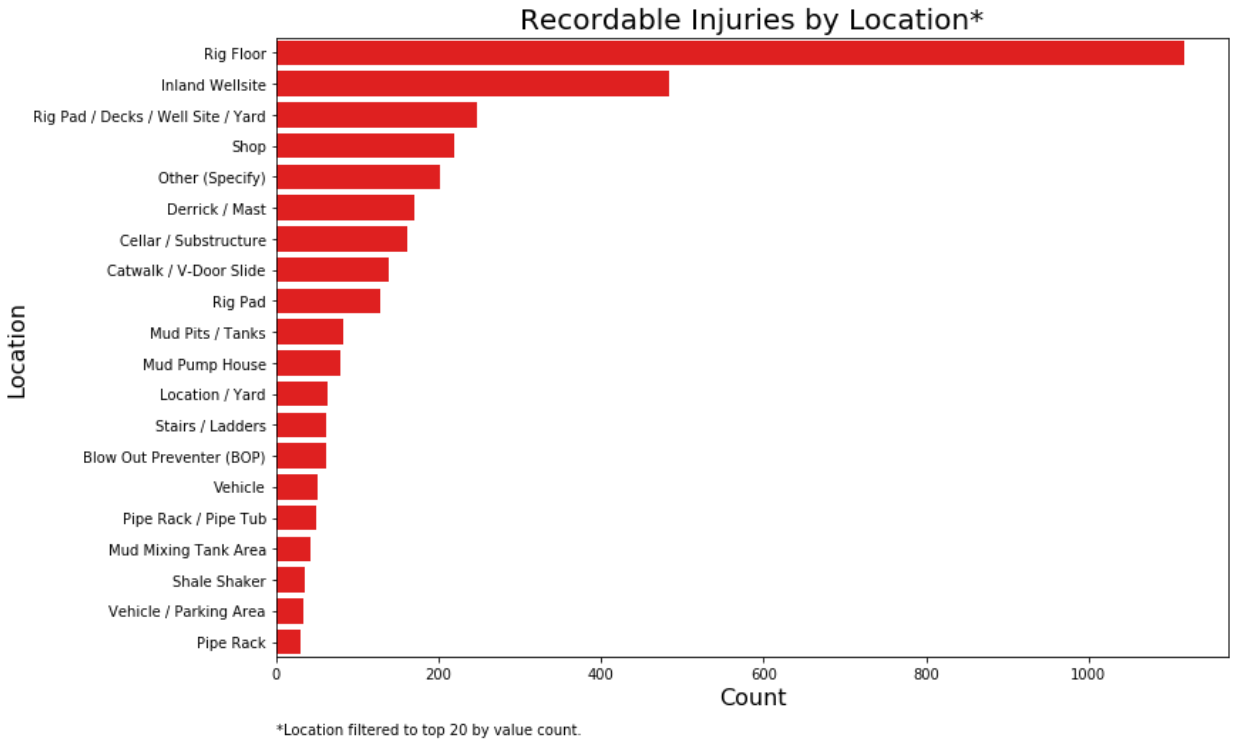

Figure 4: Bar plot of injuries by location

Most injuries and near misses happen on the rig floor. This is followed by the Inland Wellsite but the total of these injuries are half of injuries that happen on the rig floor. The most common injury reported is a contusion. This is closely followed by strains and then lacerations.

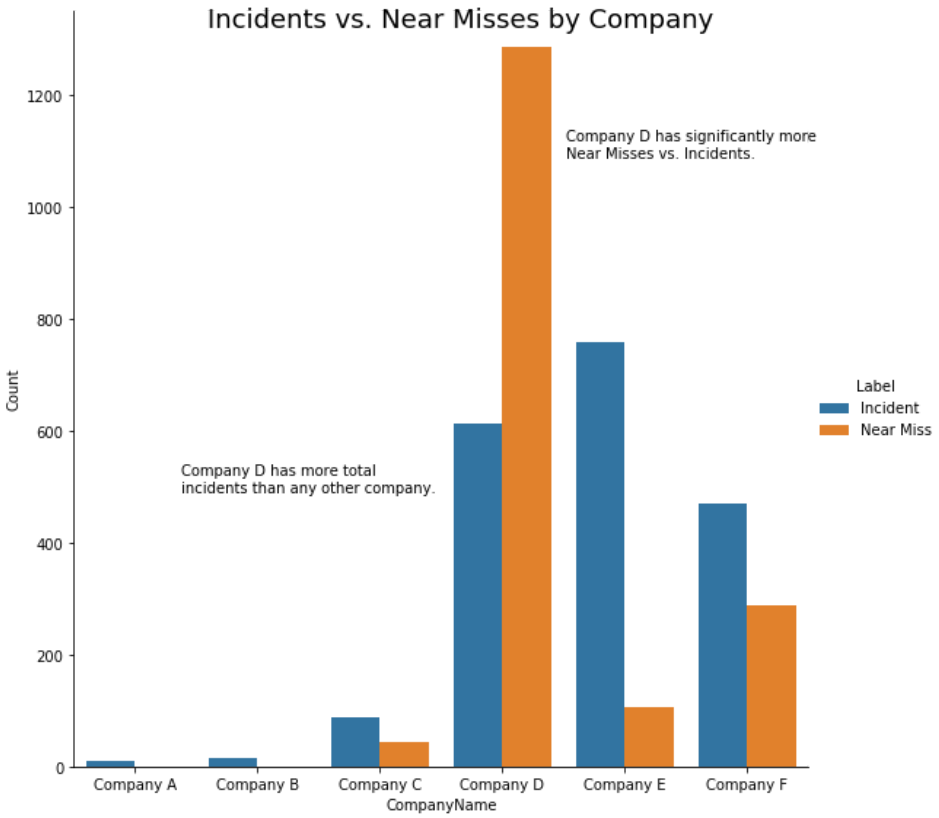

Figure 5: Bar plot of incidents and near misses by company.

Company D and E have most of the incidents. There has been a steady decline in incidents and near misses starting in 2015. The reports have since leveled off for these companies and has been steady from 2015-2018. Company F, in comparison, has a rising number of incidents starting in 2015.

Analysis

Text mining

Text mining was done in order to extract insight from the 'IncidentDescription' column of the data frame. In order to do this, each incident description was tokenized into its individual word components. This list of words was then filtered to remove stop words, numbers, and words less than 3 characters. The final term-document matrix (TDM) contains 5275 columns of words included with a value of 1 or 0 to flag whether the word is included in the description or not, respectively. This TDM would be used in modeling after merging with other significant predictors.

Figure 6: Word cloud for near misses.

Figure 7: Word cloud for incidents.

Also, we created WordCloud to show the words found in the 'IncidentDescription' column. To understand more which words appears the most in the 'IncidentDescription', we made a frequency distribution to make a comparison. The top 3 words found in near-misses were 'hand', 'pipe' and 'floor'; while the top 3 words found in real incidents were 'left', 'right' and 'back'. There were some recurring words in the top 10 frequency distribution list in between near-miss and real incident. For example, 'hand', 'pipe', 'floor', 'rig' and 'back'. The word 'almost' was also in the top 10 frequency distribution list for near-miss.

Sentiment Analysis

Sentiment analysis is the process of determining whether a piece of text is positive, negative, or neutral. It was used as part of the analysis step to gather information on the sentiment of each incident description to see how it differs between injury types. To do this, the SentimentIntensityAnalyzer was used from the nltk package to extract the positive, negative, neutral, and compound score of each incident description.

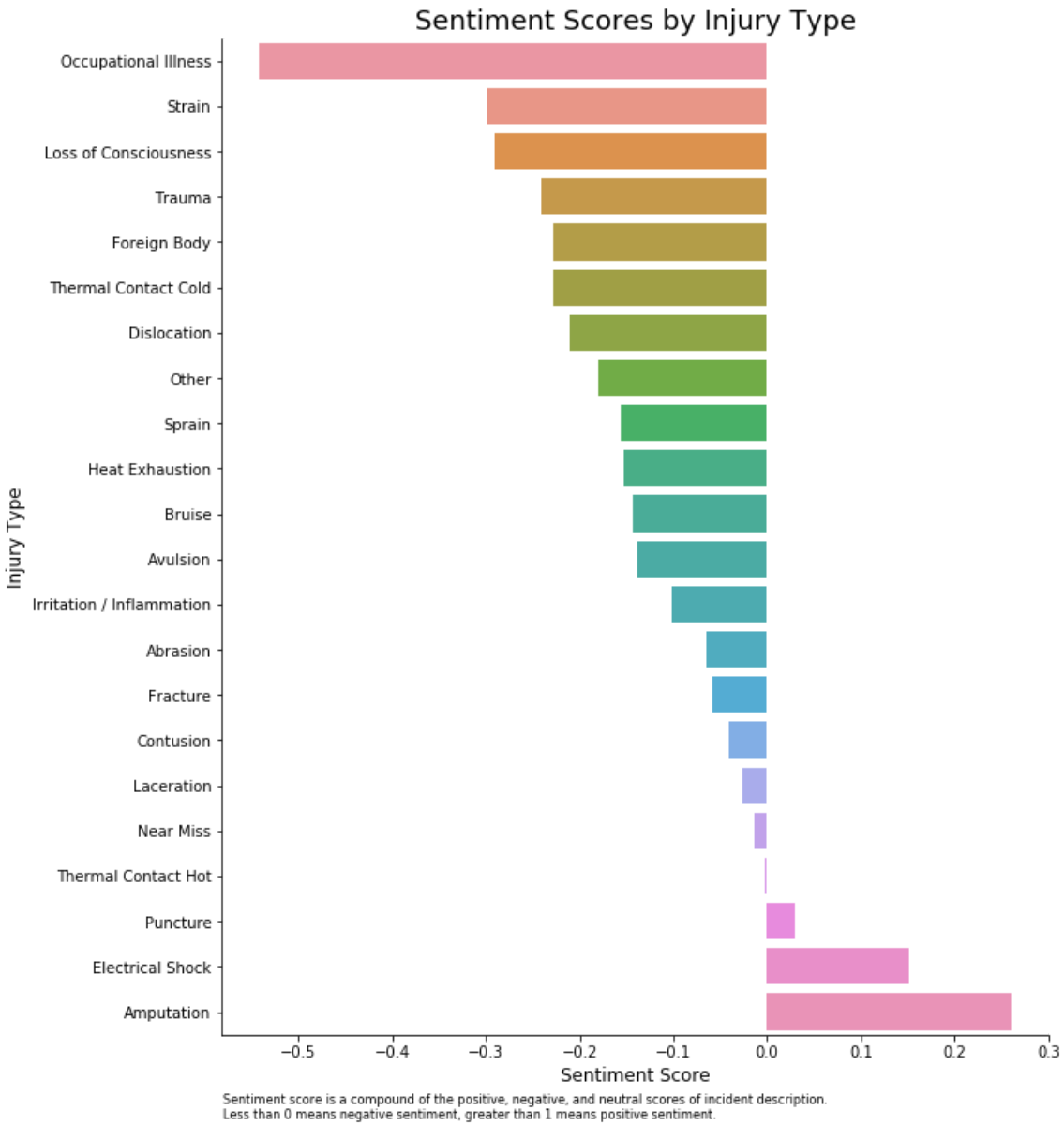

Figure 8: Sentiment scores by injury type.

From this analysis, it was found that ‘Occupational Illness’, ‘Strain’, and ‘Loss of Consciousness’ were the top 3 most negative sentiment injuries. Interestingly, ‘amputation’ was found to be the least negative. Some of these results were skewed though, as ‘Occupational Illness’ only had two occurences in the dataset with one containing the word ‘kill’ numerous times to refer to turning off a well. Additionally, the manner of these incident descriptions, which are written in a very straightforward, descriptive manner, do not naturally show much emotion. However, it was also found from these analyses that the incident descriptions for ‘Near Miss’ have a higher positive sentiment and lower negative sentiment than ‘Incidents’.

Random Forest

Random forests are a supervised learning algorithm that can be used for both classification and regress. This model is composed of numerous decision trees created on randomly selected data samples that take a vote from each tree to select the best solution for voting.

In this analysis, the random forest model was used to classify whether there was an 'Incident' or 'Near Miss' based on the column variables and the TDM. The random forest models were the primary machine learning model implemented due to their ability to handle wide datasets well and the relative importance or variables it is able to find. Thus, it is a very strong model, but also easily interpretable.

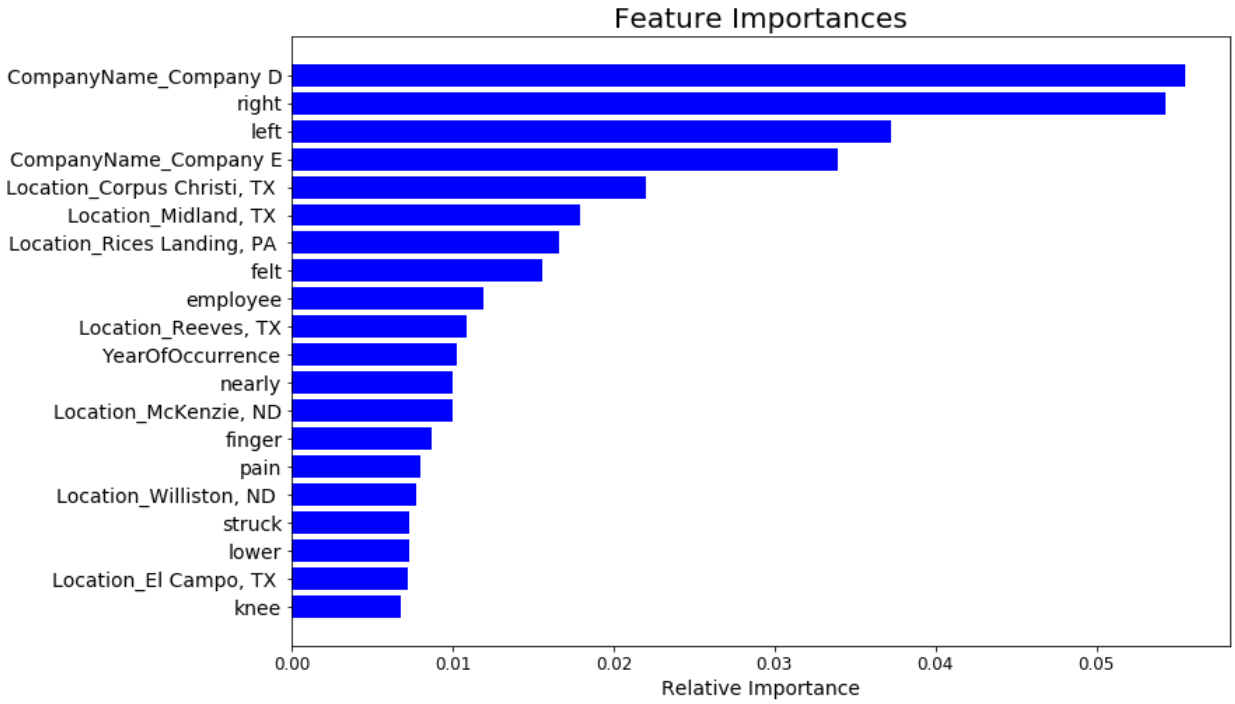

Figure 9: Feature importance from Random Forest model ran on combined TDM and incident dataframe.

Using the TDM and the incidents dataframe, the random forest model was able to achieve 96% accuracy in classifying whether there was an incident or a near miss. This is a substantial boost from the baseline of 53% of the occurrences being actual incidents. From this model, the most important features were found to be Company D, "right", and "left".

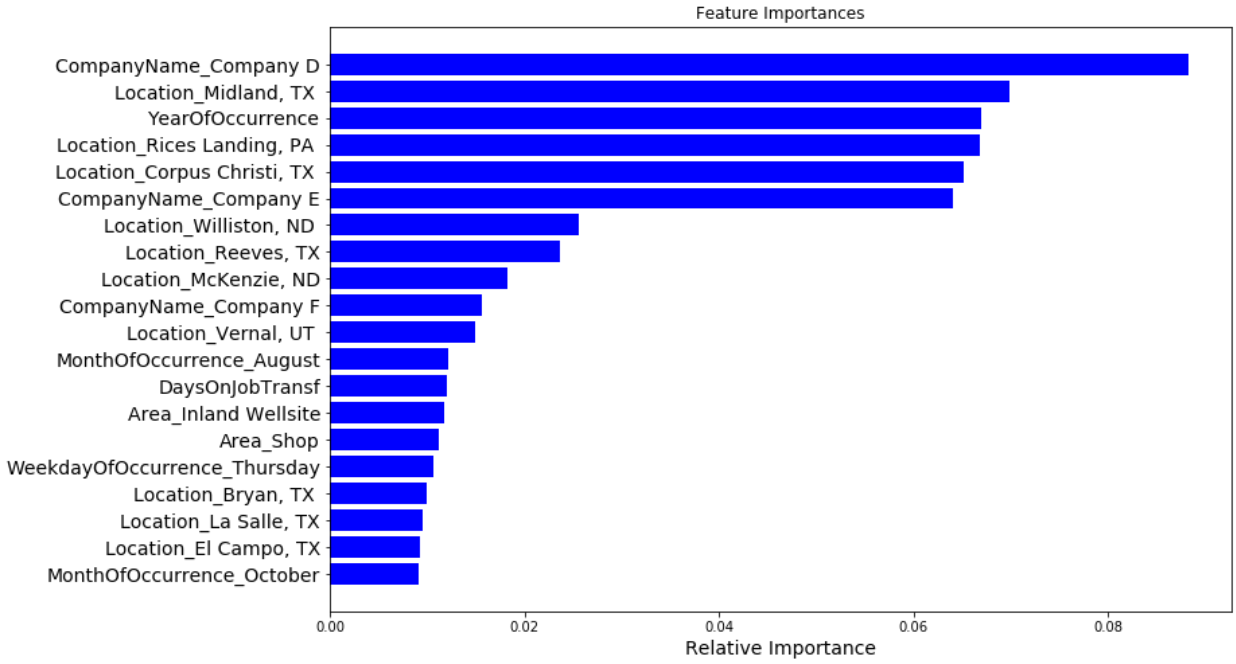

Figure 10: Feature importance from Random Forest model ran on incident dataframe.

When the model was rerun without the incident description TDM included, the accuracy dropped to 93%, which is still a very strong score. Company D was again found as a the top predictor, with Midland, TX and year of occurence being the other top predictors. From these analyses, it is clear that Company D is a clear offender in incidents.

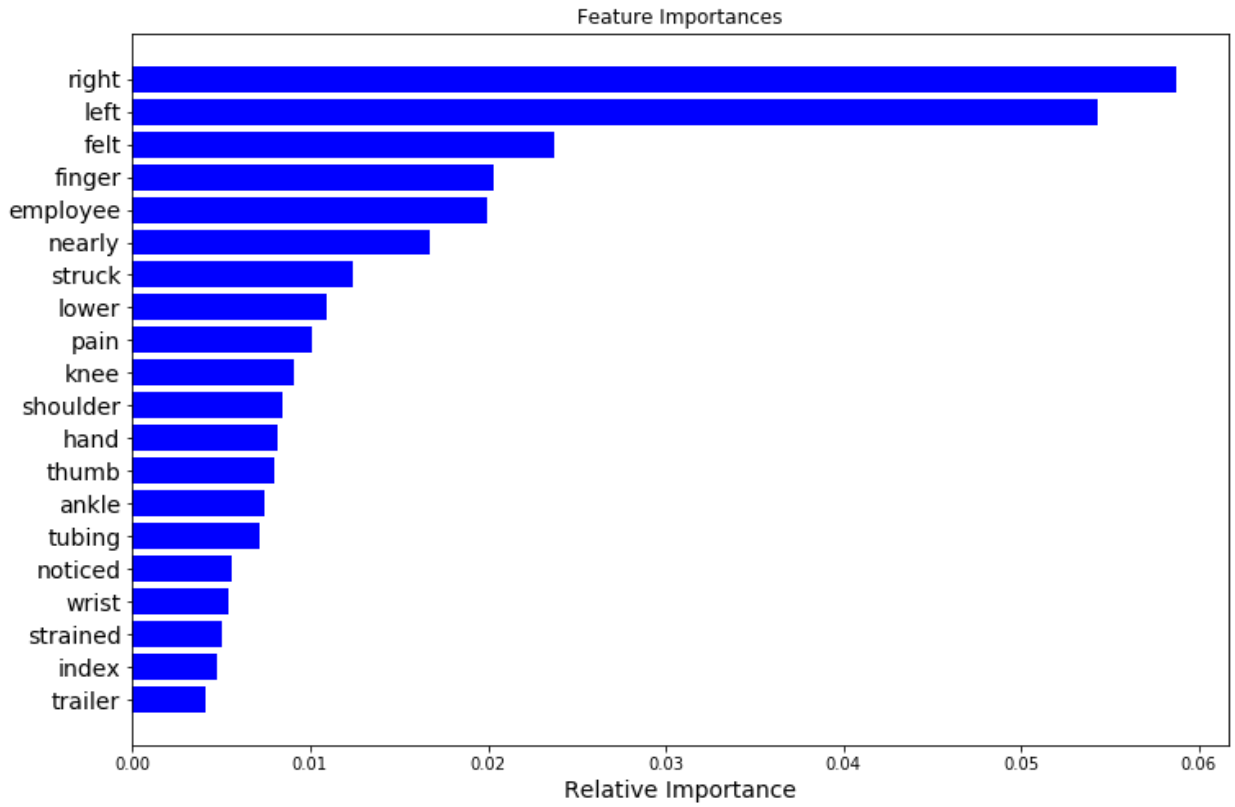

Figure 12: Feature importance from Random Forest model ran on TDM.

The final Random Forest model created was run on just the TDM to classify incidents and near misses. Using just the TDM, the random forest was able to achieve 93% accuracy. The top predictors were "right", "left", and "felt". The TDM and incidents dataframe were both very strong at predicting incident vs. near miss, but the best model was created when these two were merged to create one dataframe.

Naive Bayes

The Naive Bayes is a supervised learning algorithm that uses the probabilities of each attribute belonging to each class to make a prediction. The probability of a class value given a value of an attribute is called the conditional probability. To make a prediction, the probabilities are calculated of the instance belonging to each class and then used to select the class value with the highest probability. This model was chosen because it works well with categorical data, which this dataset was largely consisted of. It also served as a good comparison and sanity check to the random forest models.

The results of the analysis showed that Random Forest models were far superior to the Naive Bayes model, both in the overall accuracy obtained, but in the ability to interpret its results. The Naive Bayes model achieved an accuracy of only 81.5% at predicting incidents vs. near misses

on the combined TDM and incidents data frame. When applied to just the TDM, the Naive Bayes model achieved only 73% accuracy. Both of these were significantly worse than the performance of the Random Forest models.

From this, it is clear that the Random Forest model was the better algorithm for the given problem. Future iterations of this predictive model would likely focus on tuning this algorithm further.

Recommendation

The most important recommendation for the company would be to shift the new, inexperienced workers, (0-1) year, away from the rig floor. This is where nearly fifty percent of all actual incidents happen. Another recommendation for the rig floor would be to have more management oversight to help prevent near-misses and injuries from happening. This would be especially important during the summer months of June, July and August, as well as in January, when a majority of the incidents occur.

Many of the incidents occur early in the week. The company might want to lessen the work hours slightly on Mondays-Thursday to prevent injuries and near-misses. Shifting the work from earlier in the week, on Friday-Sunday when reports of injuries are less. Perhaps it would be helpful to take a deeper look into why there are more incidents that occur early in the week.

Company D had the most incidents and near-misses until 2014. Starting in 2015, there was a significant drop off in their incident reports. If they made any changes to how they were managing the work spaces or improvements to the safety regulations, it should be reviewed and documented. The data Company D could provide to the rest of the company can be used as a framework to lower incidents and near-misses across other the other companies. The information could be applied to Company F, as their injuries have been steadily rising from 2015-2018.

As a majority of the incidents and near-misses happen during the summer months and January and with employees of 0-1 year of experience, it might be helpful for the company to look at hiring cycles, price and influx of demand. If more injuries are occurring during these periods and involving inexperienced employees, having more insights on hiring and placement of employees during the first year could have a significant impact on lowering the number of incidents.

References

1. https://www.datacamp.com/community/tutorials/random-forests-classifier-python
2. https://blog.sicara.com/naive-bayes-classifier-sklearn-python-example-tips-42d100429e44
3. https://amueller.github.io/word_cloud/auto_examples/index.html#example-gallery

Appendix

- Company names were masked for confidentiality.