**Capstone 2 Milestone Report**

**Problem:** People are still being killed in train-related accidents in the year 2017. We believe that this is an unacceptable problem.

**Client:** Regulators and Railroad Executives. Regulators seek to prevent harm to their constituents and Railroad Executives want to limit the impact that accidents have on operations. Less deaths = less downtime.

**Data Set:** The Federal Railroad Administration’s Office of Safety Analysis 2010-2017 Dataset of All US Accidents.

The dataset cleansing process is as follows:

1. Removing columns that did not contain any values
2. Removing columns that we felt would have little impact on our analysis
3. Removing columns that provided the same information as other columns, but in a different format.
4. Using Isolation Forests & Local Outlier Factors to detect and remove outliers
5. Encoding the appropriate categorical variables
6. Performing feature selection to slim down the number of categorical variables so we can perform PCA
7. Convert all categorical variables into indicator variables
8. Perform PCA to pick out the most important variables
9. Perform Random Oversampling to balance out the classes

**Potential Data Sets:** We could combine the dataset above with the poverty dataset from the previous capstone. It could be that areas with higher level of poverty contain more train-related deaths due to less safety features installed.

**Initial Findings:**

We ran into a lot of Out of Memory Errors due to the sheer amount of unique values in the categorical variables. Although we only had 50 variables, after encoding them and converting them into dummy variables, it quickly ballooned into 3,488 features.

Only after few a iterations of carefully justifying which variables to keep were we finally able to perform our Outlier Detection techniques and perform PCA.

After finally performing PCA, we were able to determine the most important variables.

|  |  |
| --- | --- |
| **Variable** | **Explained Variance Ratio** |
| Month of Incident | 48% |
| Age of Person Reported | 35% |
| Day of Incident | 11% |
| Year of Incident - 4 Digits | 2% |
| Hour of Incident | 2% |
| Minute of Incident | 1% |
| Type of Person\_A | 0% |

The top 6 variables explain 99% of the variability in the data.