Predicting Severity of Traffic Accidents in Philadelphia, PA

Leanne Chan
llchan@design.upenn.edu

Yangyi Shi
yangyis@seas.upenn.edu

Xinran Liu
xrliu@seas.upenn.edu

Abstract

Traffic accidents are costly. This project identifies accident hot-spots based on current data and runs a predictive model to predict under which circumstances an accident would occur in these hot-spots, and if so, the severity of the accident. It then compares the performance of different algorithms for multi-classification such as SVM, random forest, and multinomial logistic regression. Finally, suggestions for practical implementations of the model are discussed.

Methodology

1 – Fix Class Imbalance:

-Class Weights
-Upsampling
-Undersampling

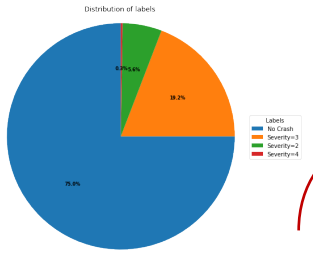


Figure 2 Distribution of Labels are imbalanced

Dependent Variable
Severity 0:
No accidents
Severity 2-4:
Accidents; 4 being the highest severity.

2 – Learning Algorithm:

-Multinomial Logistic Regression
-AdaBoosted Decision Tree
-Random Forest
-Alternative ensembles

Data Pre-processing

Pre-processing 1 - Cluster Analysis

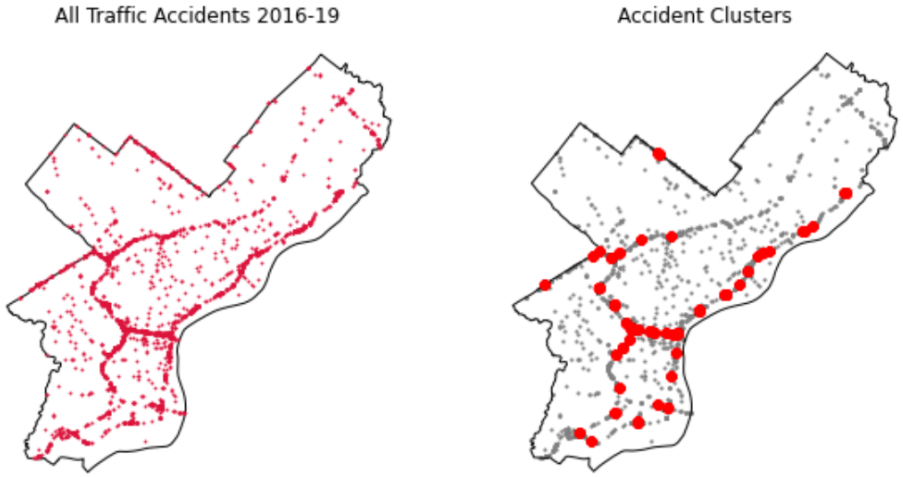


Figure 1 Traffic Accidents and Clusters in Philadelphia

Pre-processing 2 - Negative Sampling

For each accident in a cluster, three additional 'no-accident' points (Severity = 0) were randomly generated.

Why?

The existing data only represents accident incidences, but we need Instances of no accidents as well.

Penn Engineering

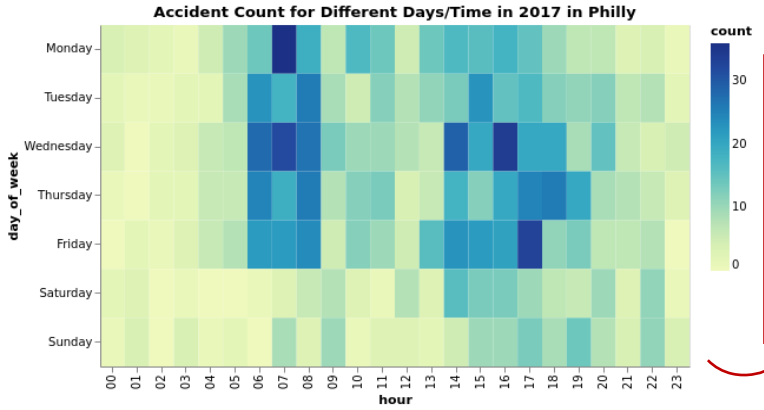
Predicting Severity of Traffic Accidents in Philadelphia, PA

Leanne Chan
llchan@design.upenn.edu

Yangyi Shi
yangyis@seas.upenn.edu

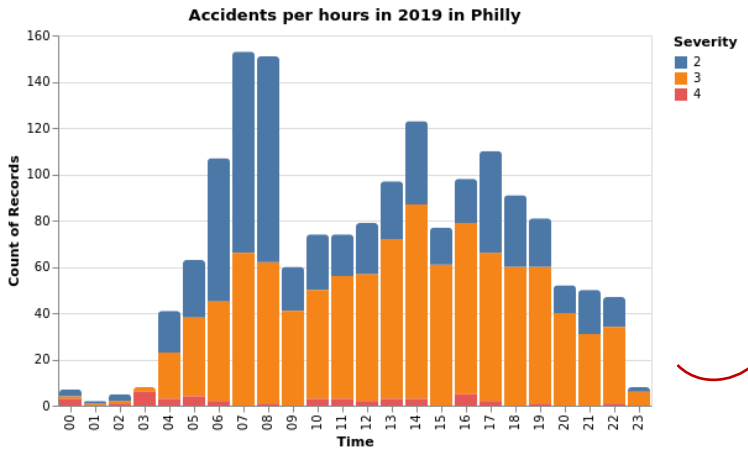
Xinran Liu
xrliu@seas.upenn.edu

Data Exploration



The highest accident counts occur during the morning (6-8am) and evening (4-6pm) rush hours.

Figure 3 Accident Count For Different Days vs Time



There is a clear class imbalance with a majority of the accidents being classified as '3' for the year 2019.

Figure 4 Severity Distribution of Accident per hours

Results

Algorithm	Tot Acc	Avg Acc	False Neg
LogReg	0.203	0.455	0.0
AdaBoost	0.575	0.507	0.291
RF	0.813	0.542	0.560
Voting	0.753	0.411	0.687

Random Forest and the voting classifier: Highest accuracy but have a high false negative rate, almost all the predictions are 0 (no accidents).

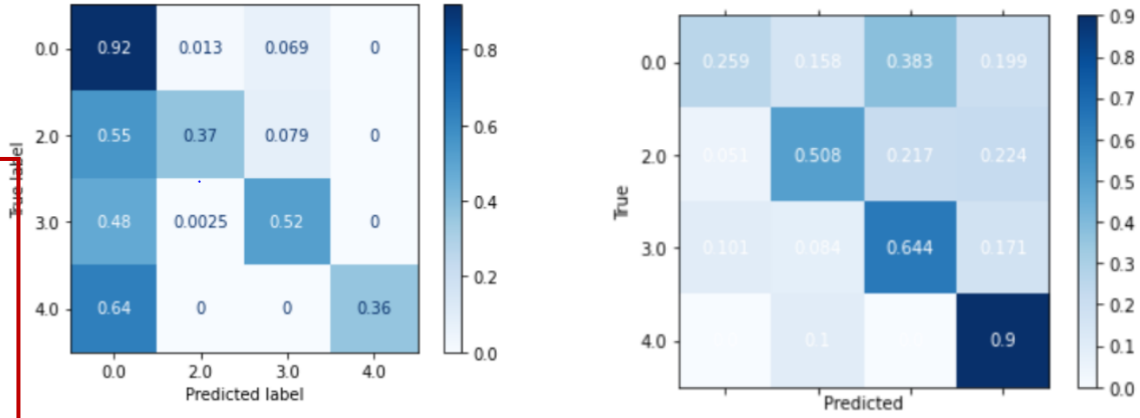


Figure 5 Confusion Matrix of random forest classification (left) and easy ensemble based on RF (right).

False negatives are costly as they endanger lives. Models with a lower false negative are better in the context of accident prediction. The balanced bagging and easy ensemble classifiers provide the best trade off between average accuracy and false negative rate, thus are the best models for accident prediction.

Penn Engineering