# Severity of traffic accident prediction using Machine Learning

Abhijit Anand Deshpande
University of Texas at Arlington
abhijit.deshpande@mavs.uta.edu

Ankit Ranjan
University of Texas at Arlington
ankit.ranjan@mavs.uta.edu

## Abstract

*The objective of this project is to predict the severity of traffic accidents using Machine Learning algorithms like Logistic Regression, Artificial Neural Network, and Support Vector Machine. Feature selection using Principal Component Analysis (PCA) is performed to show which features are more important than others to predict the severity of accidents on a scale of 1 to 4 where 1 is the least severe and 4 is the most severe, which consequently corresponds to 1 as the least delay in traffic and 4 as a significant delay in traffic. Performance on the above supervised learning techniques is compared and important features from PCA that are generally easier to obtain than others are identified to help scale the learnings of this project to any general traffic data set.*

## 1. Introduction

Motor vehicle crashes are a leading cause of death in the U.S., with over 100 people dying every day. More than 2.5 million drivers and passengers were treated in emergency departments as the result of being injured in motor vehicle traffic crashes in 2015. The economic impact is also notable: for crashes that occurred in 2017, the cost of medical care and productivity losses associated with occupant injuries and deaths from motor vehicle traffic crashes exceeded $75 billion. [1]

Traffic crash deaths resulted in $55 billion in medical and work loss costs in addition to the immeasurable burden on the victims' families and friends in 2018. [2]

Thus, it becomes imperative to stay ahead of the curve by predicting in advance where accidents might occur to avoid or minimize the losses accompanied by them, regardless of whether they are short or long delay.

The US-Accidents data set is used as it provides a wide number of features for accidents that occurred in 2016-2020, that can be used to predict the severity of accidents across all states in the US. Despite the large number of features, the aim of the project is to identify important features that are readily available across the world when data sets are compiled.

| Label | Total | Train | Test |
|---|---|---|---|
| 2 | 67568 | 47461 | 20107 |
| 3 | 28403 | 19759 | 8644 |
| 4 | 3212 | 2218 | 994 |
| 1 | 817 | 562 | 255 |

Table 1. Data set with 4 labels.

### 1.1. Data Set

There are a total of 70000 training samples and 30000 testing samples are chosen from the 3.5 million data set to ease the burden on the supervised learning technique: SVM which does not work well for too large data sets due to a time and computational constraint. The initial number of features is 49 and the breakdown of each class in the Train set and Test set and Total set are shown above. There are 4 classes in the data set which are 1, 2, 3 and 4. The objective of the design is to create a model that detects different severities of accidents. However, as traffic accidents of severity 3 and 4 are a matter of life and death i.e., their costs are very high, the cost of mis-classified actual positive (or false negative) is very high here in these

circumstances, even though they are grossly outnumbered by severity level 2.

**Average Economic Cost by Injury Severity or Crash, 2018**

| | |
|---|---|
| Death (K) | $1,659,000 |
| Disabling (A) | $96,200 |
| Evident (B) | $27,800 |
| Possible (C) | $22,800 |
| No injury observed (O) | $12,200 |
| Property damage only (cost per vehicle) | $4,500 |

Table 2: Estimates are the costs by severity of injuries, as defined in sections 2.3.4 through 2.3.6 of the *Manual on Classification of Motor Vehicle Traffic Accidents* (7th Edition) ANSI Standard D16.1-2007. [3]

## 1.2. Data Cleaning

Using the heatmap, the missing values were identified in the feature space. There were about 70% missing values in the features 'End_Lat', 'End_Lng' and 'Number' and thus these features were dropped. The rows where 'Weather_Timestamp' were missing corresponded to the rows of 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', and thus these rows were dropped. The remaining missing values were of features: 'Precipitation(in)', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)'. They were interpolated based on the weather at that month of the year. The synonymous and unstandardized strings in features like 'Weather_Condition', 'Wind_Direction' were replaced by a common string which helped in one-hot encoding them before they could be passed into models.

After each pass, the following heatmaps of missing and present values were generated as follows:

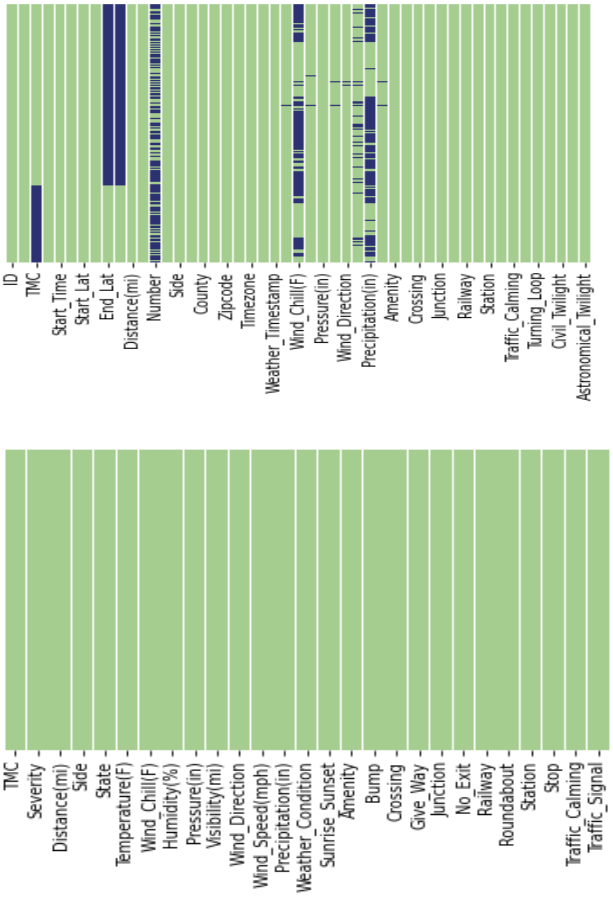


Figure 1. Each pass of data cleaning

Thus, the data was cleaned at this point and could be sent as input to the models used.

As an additional step, the Severity class combined as Severity 1 and 2 as class or Severity 0 while Severity 3 and 4 as Severity 1. This helps offset the imbalanced classes of Severity 1 and 4, which helps predict a better F1-Score.

## 2. Methods

The project performs predictive analysis on different supervised learning algorithms namely: *Logistic Regression, Artificial Neural Network, and Support Vector Machine.* Logistic Regression is the widely established algorithm of preference in this multi class case due to its scalability on larger datasets due to the accurate probabilities obtained by virtue of a large sample size for training. This results in faster training time and prediction as opposed to SVM where size of the dataset is a major drawback, despite SVM being more accurate.

## 2.1. Logistic Regression

It is a single perceptron model where the model is first fit on all features (with dummy variables included for the string features) and then fitted with15 most important features obtained from user defined PCA function.

These fitted functions are then then predicted against the test Severity class with all features and 15 features, respectively. This shows us which 15 features are most important to predicting the Severity of accidents. In decreasing order of importance, they are: 'TMC', 'Distance(mi)', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)', 'Amenity', 'Bump', 'Crossing', 'Give_Way','Junction', 'No_Exit'.

### 2.1.1 Observation

After training, the accuracy on testing set was 70.78 % and total training and testing time was 1.44 seconds, while with 15 PCs, the accuracy reduced to 68.13 % and total training and testing time was 0.33 seconds. Since the data was reduced from 3.5 million samples to 100,000 samples, the feature selection with 15 PCs help reduce training time and will be reasonably faster to implement for a much larger data set by keeping just the important features in the model.

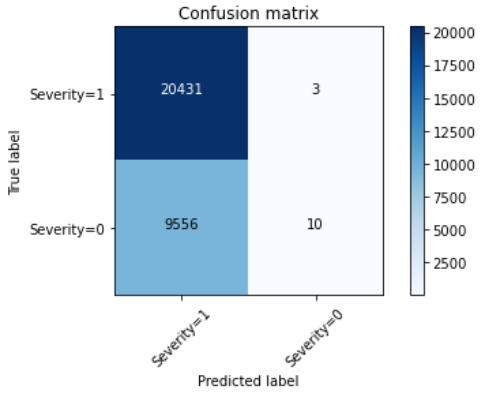


Figure 2. Confusion Matrix in Logistic Regression

## 2.2. Support Vector Machine

It is a model where nonlinear decision boundary is created by projecting hyperplane down to lower dimensions. The model is first fit on all features (with dummy variables included for the string features) and then fitted with15 most important features obtained from user defined PCA function.

Given computational limitations with SVM, the training sample size is only 7,000 and test sample size is 3,000.

These fitted functions are then then predicted against the test Severity class with all features and 15 features, respectively. This shows us which 15 features are most important to predicting the Severity of accidents. In decreasing order of importance, they are: 'TMC', 'Distance(mi)', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)', 'Amenity', 'Bump', 'Crossing', 'Give_Way','Junction', 'No_Exit'.

### 2.2.1 Observation

After training, the accuracy on testing set was 69.16 % and total training and testing time was 8.88 seconds, while with 15 PCs, the accuracy remained same at 69.16 % but total training and testing time reduced to 7.69 seconds.

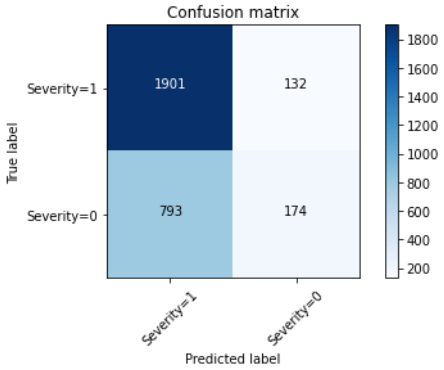


Figure 3. Confusion Matrix in SVM

# 3. Result

The most accurate model from SVM model and Logistic Regression model was the Logistic Regression model, however this may be attributed to the fact that SVM is trained on a smaller sample size.

## 3.1. Precision, Recall and F1-Score

The following score report is that of Logistic Regression and it shows that the F1-Score performs better than expected despite the imbalanced class 1 which is much lesser than to class 0.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.73 | 0.94 | 0.82 | 20543 |
| 1.0 | 0.65 | 0.23 | 0.34 | 9457 |
| accuracy | | | 0.72 | 30000 |
| macro avg | 0.69 | 0.59 | 0.58 | 30000 |
| weighted avg | 0.70 | 0.72 | 0.67 | 30000 |

Table 2. Score report in Logistic Regression.

The following score report is that of SVM and it shows a fall in F1-Score compared to Logistic Regression, which was the case for accuracy as well. This slight drop may be attributed to the fact that since both classes were imbalanced and further in SVM the training size was lesser than that in Logistic Regression, the evaluation scores below were naturally lesser as there was smaller data to learn from.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.71 | 0.94 | 0.80 | 2033 |
| 1.0 | 0.57 | 0.18 | 0.27 | 967 |
| accuracy | | | 0.69 | 3000 |
| macro avg | 0.64 | 0.56 | 0.54 | 3000 |
| weighted avg | 0.66 | 0.69 | 0.63 | 3000 |

Table 3. Score report in SVM.

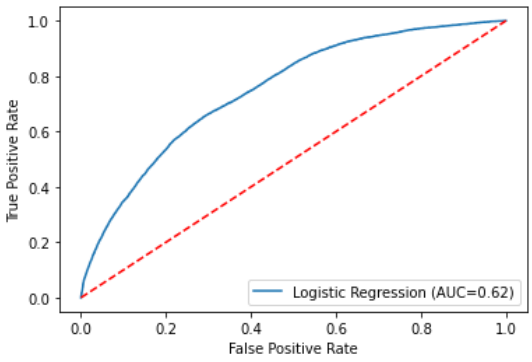


Figure 4. ROC and AUC for Logistic Regression

The above AUC value of 0.62 can be considered acceptable for now but can be improved by sampling larger data size with a technique called negative sampling, or another technique called Random-Over-Sampling to increase the grossly minoritized classes, or another technique called Random-Under-Sampling to decrease the grossly majorized class.

## 4. Conclusion

Costs associated (including opportunity loss due to death or impairment) traffic accidents are a huge problem in all countries, including the USA. A Logistic Regression model can do a basic prediction with decent accuracy and F1-Score. From the POIs (Points-of-interest) the following were found to be more important than other POIs (in decreasing order of importance): 'Amenity', 'Bump', 'Crossing', 'Give_Way','Junction', 'No_Exit'. The most important predictors were 'TMC', 'Distance(mi)', 'Temperature(F)','Wind_Chill(F)','Humidity(%)','Pressure(in)', 'Visibility(mi)'. These are all readily available data which can be translated to other data sets worldwide.

## 5. Future

The next step would be to add more features like vehicle condition, vehicle type, demographic information, and sober state of the driver, which is readily available information from hospitals (despite the sensitive nature, it may be parted with as it is to be used as a major cost reduction tool in traffic accidents) especially for severity 3 and 4, which are potentially fatal and linked with forensic reports. Additionally, other supervised machine learning techniques may be employed to evaluate and improve the recall (due to high cost of false negatives when classifying Severity 3 and 4) and F1- score.

## References


[1] https://www.cdc.gov/transportationsafety/costs/index.html

[2] *https://www.cdc.gov/transportationsafety/statecosts/index.html*

[3] https://injuryfacts.nsc.org/all-injuries/costs/guide-to-calculating-costs/data-details/

[4] Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights

[5] Summary on Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol.