Predicting the Severity of Car Accidents Based on Road Conditions

Aryan Mistry

September 2, 2020

Introduction:

The problem that is to be solved is to determine the locations, times of day, and weather conditions that elicit the most severe car accidents in Seattle, so that they may be prevented in the future. The issue of automobile accident severity would be immediately useful to the Seattle Police Department as well as the Department of Public Works, and as such, they would be the audience for this problem. In determining which areas and times of day correspond with the most severe accidents, they would have it in their power to mitigate the circumstances that contribute to these accidents, namely, issuing more police officers in that area, improving roads, installing streetlights and relevant road signs, etc. On a larger scale, this issue serves as a model for road safety in metropolitan cities across America. Car accidents are the leading cause of death in the United States for the age group 1-54 years old, so this issue has major implications for its effect on human life. The ultimate goal would be to reduce the number of accidents in these areas over a given period of time, or at least reduce their average severity, and in turn reduce the fatality rate associated with these accidents.

Data:

The dataset that was used for this model was provided by Coursera and was sourced from the Seattle Department of Transportation Traffic Management Division, Traffic Records Group. It tracks traffic collisions recorded by the Seattle Police Department from the year 2004 to the present, and is updated on a weekly basis. It has 38 features and 194,673 data points.

Methodology:

Data Cleaning/Feature Selection:

As the data was jointly provided by the SPD and the SDOT, it contained a large number of logistical features such as report numbers, identification numbers, collision keys, and department codes. These were all dropped from the dataset as they were irrelevant to solving the problem. Data points that contained null or unknown values were also dropped. The remaining data points were almost entirely categorical, and thus it wasn't straightforward to determine which of the features correlated best with the target variable 'SEVERITYCODE', which listed a number from 1-3 based on the severity of the collision. A severity code of 1 indicated no injuries and minor property damage only, while a code of 3 indicated a collision that ended in a fatality. Of all the features, four were ultimately selected, namely 'ADDRTYPE', 'WEATHER', 'ROADCOND', and 'LIGHTCOND'. The first feature denotes whether the collision occurred at an intersection, block, or alley; the second denotes the weather conditions that day, e.g. clear, raining, sleet/hail, snow, etc; the third denotes the

condition of the road, e.g. whether it was wet, dry, oily, etc; the last denotes the visibility conditions at the time, e.g. dusk, dark with no streetlights, daylight, etc.

The features were loaded into a Pandas Dataframe, and were cleaned of null and unknown values. As mentioned before, the features were all categorical, and as such, they needed to be converted to numerical values in order to be used in a machine learning algorithm. To accomplish this, I used LabelEncoder from SciKitLearn's preprocessing package. LabelEncoder essentially assigned each unique category within a feature a numerical value, e.g. 'Alley' was given a value of 1, 'Block' a value of 2, and 'Intersection' a value of 3, and so on. The drawback of Label Encoding our categorical data is that it may be misinterpreted by an algorithm. For example, in the LIGHTCOND column, the category of "Dusk" is given a label of 6, and the category of "Dark" is given a label of 1. Obviously, the Dusk category does not hold 6 times more weight than the night category, but to the machine learning algorithm, this would be exactly the case. To rectify this, the data had to be normalized. I used SciKitLearn's StandardScaler package to do this. Lastly, I stored the cleaned data into two NumPy arrays, one for the features, and one for the target variable, ready to be used in the model.

Creating and Training a Machine Learning Model:

The data was then split into testing and training sets. Specifically, I employed an 85/15 split, in which 85% of the data was used for training the model and 15% for testing the model. As this was a classification task, I decided to go with a Support Vector Machine, or SVM, as my algorithm of choice. My kernel of choice was a

Radial Basis Function, or RBF, as it works well with multiclass classification. The cleaned NumPy arrays were fed into the model, and the predictions made by the model were stored in a separate array.

Results:

Since this project deals with multiclass classification, i.e. we are attempting to classify accidents as one of three severity scores given the road and light conditions, I decided to use F1-score as the primary evaluation metric for the SVM model. F1-score, although a less intuitive metric than accuracy, takes into account how the data is distributed, and then takes the harmonic mean of the machine learning model's recall and precision, and is thus a better metric for model evaluation in this case. Importing the F1-Score package from SciKitLearn's Metrics library, I compared the model's predicted severity scores to the actual, ground truth severity scores. The final score came out to 0.8077, which means, in simplified terms, that **the model was able to predict the severity of collisions with about 80.77% accuracy.** In other words, the features pertaining to address type, light conditions, road conditions, and weather were able to explain over 80% of the change within the data with regard to collision severity. Thus, these factors are statistically significant.

Discussion/Conclusion:

As mentioned before, the four features used in the model were indeed statistically significant, as they explained over 80% of the change within the data. Plotting each of these

features against the target variable, it becomes clear that a majority of collisions occur at intersections, during adverse weather, and due to poor road conditions. This begs the question of how these higher collision rates may be mitigated. When it comes to weather, it would seem useful to close roads or limit access to them during prolonged periods of snow or rain, thereby reducing traffic in those areas and lowering the chances of a collision. It would also seem prudent to establish safety measures at major intersections, such as CCTV cameras and radar guns, to discourage reckless driving in those areas, as well as station more police officers at those intersections. Hopefully, datasets such as these will help to further shed light on the nature of car accidents and what can be done to avoid them, so that we may look forward to a safer future of transportation.