**Predicting Accident Severity for the City of Seattle**

Phillip Asberry

September 21, 2020

**1. Business Understanding**

* 1. **Background**

The city of Seattle is the seat of King County Washington and the largest city of the state of Washington. With a population of 608,660 (according to the 2010 census) and a metropolitan area of 4 million, Seattle is one of the fastest growing regions in the US. As growth continues, traffic is a major concern.

Traffic pattern analysis and optimization is critical for the efficient flow of goods and services. With this information, Seattle can better facilitate the needs of its citizens. One element of traffic analysis is accident avoidance and mitigation. Accidents are an unfortunate part of any urbanized region. Along with the human cost, accidents can exacerbate traffic which can impede emergency personnel response. Predicting when and where potential accidents can happen would help the region of Seattle maintain growth encouraging commerce to proceed unencumbered.

* 1. **Problem**

A combination of weather conditions, season, day of the week, location, and vehicles involved can help make the determination of an accident’s severity. This project aims to predict the likelihood of severe accidents that could lead to traffic disruption.

* 1. **Interest**

This project would potentially be of interest to all parties associated with regional transportation in the Seattle metropolitan. With better understanding SDOT could predict when to deploy assets such as salt trucks. The mayors of the region could better lobby the state and federal governments for funds to improve infrastructure.

**2. Data Understanding**

**2.1 Sourcing Data**

The data for this project was provided by the SDOT Traffic Management Division, Traffic Records Group. This data set includes all types of collisions from the year 2004 to present.

**2.2 Data Analysis and Feature Selection**

Upon inspection of the data, the data was incomplete in many columns. The choice was made to simply drop underrepresented columns and fill in missing data with the mean or mode where appropriate. In addition, all columns that provide descriptions and codes were dropped as well. The intent for the feature set was to focus on a small but expressive set of data points.

Columns **SPEEDING, PEDROWNOTGRNT, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND** and **INTKEY** were dropped from feature consideration because they only showed up in at most 3% of the 194,673 records.

**ADDRTYPE, and JUNCTIONTYPE** were dropped in favor of keeping the **X** and **Y** coordinates of the accident.

**JUNCTIONTYPE**, **SDOT\_COLDESC**, **SDOT\_COLCODE**, **ST\_COLCODE** and **ST\_COLDESC** were viewed by the project as descriptive columns that provided annotations for the row. Beyond documentation, they were not deemed important features and were dropped.

The following features described in subsequent paragraphs composed the feature matrix X.

The columns **X** and **Y** represent the geographical coordinates of the incident. The data was scaled via Min/Max normalization with the missing data replaced with the mode of each column. The assumption to use the most popular **X, Y** tuple for accidents without coordinates stemmed from the assumption most accidents in the city probably happened in a rather common area for all accidents.

The assumption to use the mode for missing values was used throughout the dataset.

**VEHCOUNT, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT** were treated in the same fashion as the coordinate columns **X** and **Y**. The features were Min/Max scaled and null values were replaced with the mode.

**VEHCOUNT** described the number of vehicles involved in the accident.

**PERSONCOUNT** described the total persons involved in the accident.

**PEDCOUNT** described the total pedestrians involved in the accident.

**PEDCYLCOUNT** described the total number of bicycles involved in the accident.

The **INCDATE** column contains timestamp data about the incident. The date information from the column was parsed and split into to new columns **‘Incident Month’** and **‘Incident Day of Week’** representing the month of the year and the respective day of the week the accident occurred. The valid month values are January, February, March, April, May, June, July, August, September, October, November, and December. The valid days of the week are Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday.

The **ROADCOND** column represents a categorical list of conditions for the roads at the time of the accident. The list of values was Dry, Wet, Unknown, Ice, Snow/Slush, Other, Standing Water, Sand/Mud/Dirt and Oil. The decision was made to simplify the model by segregating the road conditions into two categories, Fair or Hazardous. This decision was made to lessen the dimensionality of the matrix features. All missing or unknown values were substituted with the mode.

The **WEATHER** column was also reduced categorically to two options, Clear and Inclement. The previous categories were Clear, Overcast, Unknown, Other, Partly Cloudy, Raining, Snowing, Fog/Smog//Smoke, Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, or Severe Crosswind. The values Clear, Overcast and Partly Cloudy were assigned to Clear, the others were delegated to Inclement weather. All unknown and null values were assigned to the mode of the column values.

The **LIGHTCOND** column contained the values Daylight, Unknown, Dusk, Other, Dark – Street Lights On, Dark – Street Lights, Dark – Street Lights Off, and Dark – Unknown Lighting. With the goal of dimensionality reduction, the values were reduced to two values, Day and Night. All values Daylight, Dusk, Dawn, were consider Day and all values prefixed with Dark were considered Night. All other unknown and null values were assigned to the mode.

**UNDERINFL** indicated whether any party involved were under the influence of alcohol. These values originally composed of a combination of 0 and No for negative cases and 1 and Yes in positive cases. The values were streamlined into either Yes or No. If a value was not present, the value was No. The assumption being if a driver were involved in an accident while intoxicated, it would have been noted. Driving under the influence in America is a serious offence.

**3. Modeling**

**3.1 Methodology**

Using pandas *get\_dummies()* to one-hot encode the categorical data described prior, a 194,673 x 36 feature matrix was the result.

The project philosophy for model building lay in a 90% split for Training and Dev sets and a 10% test set. Within the 90% Training/Dev split, the data was further composed into 3 folds. Using K-fold cross-validation 1/3 of the data served as a Dev/Validation set while the other 2/3 were used for model training.

The main reason for this approach to training our models is the relatively small size of our dataset. Dedication of at least 128,400 training and 64200 dev/cross-validation sets per fold were deemed adequate for training. The approximately 19,400 left for testing provided a decent sample to validate the models.