Predicting Road Accidents : An Analysis of the impact of Road Condition and Weather to Car Accidents in Seattle USA

Marc Jerrone Castro

August 20, 2020

1. Introduction

1.1 Background

Every year, approximately 6 million car crashes occur in the U.S. where 72% of these crashes result in property damage, 27 % resulting in injuries ranging from bruises to life-changing disabilities, and 6% resulting in death. Although most of these crashes are caused by alcohol, speeding, and general recklessness - there is the underlying factor of weather and road conditions that may have escalated these avoidable accidents. Factors such as road traction, rain, lighting, and even the type of traffic in the area may play a role in predicting these accidents and could potentially serve as a tool to avoid them.

1.2 Problem

Collision data which captures road conditions and general weather might assist in identifying key or significant factors in crashes within the Seattle Area. This project aims to identify correlations between these factors and road crash severity and predict the probability of a crash occurring and it's severity with respect to these factors.

1.3 Interest

Not only the Seattle City Government, but city governments of areas with similar climates and conditions may apply a predictive model to warn drivers of the potential harms they may encounter on the road. Local news stations may also give out warnings about these factors and relay potential car crash severities such that local drivers would be more attentive given these road conditions. Finally, the primary targets of this study are the drivers whereas the resulting model from this project may further be developed as a real-time application that calculates risk and potential dangers given information such as weather, time, and road condition.

2. Data Acquisition & Learning

2.1 Data Source

Seattle **c**ar accident data such as car crash severity, type of street, weather, and other information from 2004 up to this year can be found from the following link:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The data contains 194,673 entries with 38 columns containing information such as the severity of the accident, address type, the weather, road lighting and other information such as number of people involved and exact date and time of accident.

2.2 Data Cleaning & Feature Selection

Although the data contains a lot of useful information, we limited our scope to some portion of the data. For this project, we limited our study to the following features:

Table 1. Selected Features from the Seattle Collision Dataset.

FEATURE	DESCRIPTION	DATA TYPE
'ADDRTYPE'	Indication of area where accident occur with values varying from alley,block, or intersection	CATEGORICAL STRING
'SPEEDING'	Whether or not speeding was a factor in the collision. (Y/Null)	BOOL
'HITPARKEDCAR'	Whether or not the collision involved hitting a parked car. (Y/N)	BOOL
'INCDTTM'	The date and time of the incident.	DATETIME
'UNDERINFL'	Whether or not a driver involved was under the influence of drugs or alcohol. (Y or 1 / N or 0)	BOOL
'WEATHER'	A description of the weather conditions during the time of the collision.	CATEGORICAL STRING
'ROADCOND'	The condition of the road during the collision.	CATEGORICAL STRING
'LIGHTCOND'	The light conditions during the collision.	CATEGORICAL STRING

'SEVERITYCODE'	A code that corresponds to the severity of the collision	INTEGER (with a mixture of STRING)

Note that a number of these features had minor issues regarding their formatting. First, the booleans within the dataset were either represented by Y or N. Ideally, we would like our data to be represented as integers such that our algorithm could handle it. Therefore all booleans were transformed into either 1 or 0. This would benefit all of our boolean data while significantly improving the quality of the data of 'UNDERINFL' data points within this feature were a mixture of Y/N and 1/0.

The next issue within our dataset was the presence of dates and time. Although informative, we would only like to extract specific information within this feature. Specifically, we are interested in the specific hour of the accident and specific month. Thus we parse the months and hours within this feature and drop the other elements. We then group these values based on a specific range whereas we group all accidents occurring as such:

Table 2. Binning of Months of Car Accidents.

MONTHS	**BIN**	**DESCRIPTION**
NOVEMBER TO FEBRUARY	0	Cold Season in Seattle with temperatures dropping a significant amount.
MARCH TO MAY	1	Warmer temperature with days begin to become longer than nights
JUNE TO AUGUST	2	Moderately Humid and warm temperatures - generally perceived as the rainy season.
SEPTEMBER TO OCTOBER	3	High Humidity season

A similar method was performed for the hours of each accident, whereas the following scheme was used to bin them:

Table 3. Binning of Hours of Car Accidents.

MONTHS	BIN	DESCRIPTION
12:00AM to 4:59AM	0	Early Morning
5:00AM to 10:59 AM	1	Morning
11:00AM to 2:59AM	2	Noon
3:00PM to 5:59PM	3	Afternoon
6:00PM to 8:59PM	4	Night
9:00PM to 11:59PM	5	Late Night

The third issue we had to address with the data were the categorical variables and our newly binned features. First we addressed the issue of missing values in our categorical values. For the purpose of consistency, all missing values were simply replaced by the most common value within their series. Afterwards, we identified that vague values existed within some of the categorical variables - whereas we opted to remove all data points with a value of 'Others' while opting to replace all 'Unknowns' with the most common value. As these features are represented by a string, we first address these variables by generating one-hot encodings for each data point and removing an extra column to accommodate the errors caused by the dummy variable trap.

Lastly, we performed some simple data cleaning on the severity code as it was not purely composed of integer values. We rescaled the values such that instances with a SEVERITYCODE of 2b were rescaled as 3 while instances with a SEVERITYCODE of 3 were rescaled as 4. As these codes represent increasing intensities of injuries/collision there is no need to convert them to one hot encodings and dummy variables.

3. Methodology
4. Predictive Modeling
5. Conclusions
6. Sources
   https://www.driverknowledge.com/car-accident-statistics/