PREDICTING VEHICLE ACCIDENT SEVERITY IN SEATTLE, WASHINGTON

IBM Python Applied Data Science Capstone

The goal of this project is to use vehicle collision data recorded in the city of Seattle to predict whether an accident would cause property damage or personal injury.

By Mohit Mhapuskar
mmhapusk@stevens.edu
linkedin.com/in/mmhapusk
github.com/mmhapusk

Introduction: Business Understanding

1. Background

As we all know, Seattle is a major seaport city on the West Coast of the United States in the northwestern state of Washington. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. As of July 2016, Seattle was the fastest-growing major U.S. city, with a 3.1% annual growth rate.

Like any other fast-growing city in America, an increase in development coincides with an increase in population which in turn leads to more and more automobiles traversing through the city. As of 2016, the number of vehicles in Seattle hit a new high of 446,000. More automobiles mean a higher number of vehicle accidents take place. In 2019 alone, there were 10,315 total car crashes within the city, with about 2,612 of those crashes resulting in injuries and 22 proving fatal.

2. Solution

As you might have guessed then, our goal for this project is to analyze vehicle accident data from the city of Seattle and apply Machine Learning techniques to determine which attributes and conditions (for example, road condition and weather) can lead to an accident that just causes property damage or can lead to something more serious like a personal injury.

The findings and results of this project can then be used by the concerned stakeholders such as the Seattle Department of Transportation (SDOT) or the Seattle Police Department (SPD) to take improved measures in order to reduce the number of fatal accidents as well as total accidents overall.

The Data

1. Data Acquisition

The dataset that we will be using for this project is a collection of records of vehicle collisions that have taken place in Seattle city proper between the years 2004-present. The data has been recorded by the Seattle Police Department (SPD) and compiled together by the Seattle Department of Transportation (SDOT). The dataset itself can be found here, while the metadata for the dataset can be found here. The dataset contains a total of 194,673 records and 37 columns. The dataset will undergo extensive cleaning and engineering before it can be used for training our models.

The target variable we will be trying to predict is Accident Severity (represented in the dataset as **SEVERITYCODE**), which is a binary variable that classifies whether a particular accident involved Property damage only or Injury to a person as well.

2. Data Cleaning

The dataset was relatively raw and had to undergo a lot of cleaning, preprocessing and transformation before it could be used for the initial exploratory data analysis and eventually for modeling purposes.

2.1 Dropping Features

Firstly, there was the issue of irrelevant features. As mentioned before, our dataset consisted of about 37 columns, including our target variable. A lot of these attributes are either redundant or irrelevant in terms of modeling purposes and do not contribute towards predicting our target variable. Hence, we dropped these unnecessary variables.

Here is a list of variables we dropped from our dataset:

1. **KEYS (OBJECTID, COLDETKEY, REPORTNO, INTKEY, SEGLANEKEY, CROSSWALKKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM) :** These columns are keys or numbers that have been assigned by SDOT to uniquely identify certain records, incidents and locations. This data is irrelevant and not useful for modeling. We will only keep INCKEY as our unique identifier for our records.
2. **STATUS :** Irrelevant variable.
3. **INCDATE:** As the date of the incident is already present in INCDTTM, this is a redundant variable and therefore safe to drop.
4. **LOCATION:** The street addresses of the accidents. As we already have Latitude and Longitude co-ordinates which we can use to plot on a map, this variable is redundant and therefore can be dropped.
5. **ST_COLDESC, ST_COLCODE:** There are two collision code sets in the data to identify collisions, ST codes and SDOT codes. Since we are using the SDOT codes to identify the types of collisions in our EDA, we can drop the ST codes.
6. **VEHCOUNT, PERSONCOUNT, PEDCYLCOUNT, PEDCOUNT, PEDROWNOTGRNT, HITPARKEDCAR:** These attributes are descriptive statistics recorded after the accident has already taken place. Since we are more interested in the features which describe the conditions preceding the accident, these attributes are not that useful for modeling or EDA.

2.2 Data Transformations

Secondly, there was the issue of transforming the attributes into the correct formatting to make it easier for analysis. Our target variable of **SEVERITYCODE** is a binary class categorical variable that was initially encoded as '1' (Property Damage Only) and '2' (Personal Injury). This was changed to '0' (Property Damage Only) and '1' (Personal Injury), also known as one-hot encoding. This was done primarily because one-hot encoding is easier to interpret for the Machine Learning algorithms in Python's scikit-learn library. In a similar manner we performed one-hot encoding for the following variables:

1. **UNDERINFL:** This variable denotes whether the driver was under the influence of alcohol or drugs. Initially encoded as 'Y' (positive) or a missing value (negative). One-hot encoded to '0' (negative) and '1' (positive).

2. **INATTENTIONIND:** Variable denotes whether the driver was inattentive or not. Initially encoded as 'Y' (inattentive) or a missing value (attentive). One-hot encoded to '0' (attentive) and '1' (inattentive).

3. **SPEEDING:** Variable denotes whether the driver was over the speed limit or not. Initially encoded as 'Y' (over speeding) or a missing value (not speeding). One-hot encoded to '0' (not speeding) and '1' (over speeding).

The variables **ROADCOND** (Road Condition), **LIGHTCOND** (Lighting Conditions) and **WEATHER** had a lot of values which were simply labelled 'Unknown'. This is as good as a missing value and was therefore replaced by a blank value, to be dealt with later when we began dealing with missing values.

The variable **INCDTTM** (Incident Date Time) contained a timestamp for each accident, consisting of Year, Month and Day of Week. We would use the to_datetime() function in the pandas library to extract the year, month and day from the timestamp into separate columns. This would make it easier for us to analyze the data during EDA.

The variables **X** and **Y** which represent our geospatial coordinates were also renamed as **LONGITUDE** and **LATITUDE** respectively for our convenience.

2.3 Dealing with Missing Values

Thirdly, there was the issue of missing values in our dataset. Generally there are three ways to deal with missing values in data: (i) Drop the rows entirely, (ii) Substitute those values on the basis of other values in that column or (iii) Leave them as is (generally not recommended). After compiling a list of missing values for each attribute in our dataset, this is how we dealt with each of them.

1. **LONGITUDE, LATITUDE:** There were 5334 missing values for longitude and latitude. However, since this is geospatial data, substituting the missing values is overly complex and would have required the use of some sort of geography-based libraries. Since it was a negligible portion of the data, it simply made more sense to drop these rows.

2. **JUNCTIONTYPE:** There were 4199 missing values for junction type. Since this attribute is also location-based, it would not make logical sense to substitute the data. Like before, these rows also represented a miniscule portion of the data, so we just dropped these rows.

3. **ROADCOND, LIGHTCOND, WEATHER:** There were roughly 18000 missing values for these 3 attributes. Since these 3 are multi-class categorical variables, it made sense to replace the missing values on the basis of frequency of occurrence of the classes (provided there was a clear winner in each case). For Road condition, the most occurring class was 'Dry' and therefore the missing values were replaced with class 'Dry'. In a similar manner, the missing values in Light Condition were replaced with class 'Daylight' and the missing values in Weather were replaced by class 'Clear'.

4. **COLLISIONTYPE:** There were 4757 missing values for collision type. Since this is also a multi-class categorical variable, it made sense to replace the missing values on the basis of frequency of occurrence of each of the classes. However, there was no clear winner in this case, with the frequency of classes being fairly distributed. Since there already existed a collision class labeled 'Other', we substituted the missing values with 'Other'.

2.4 Feature Selection

After data preprocessing, this was the final list of features that were selected with which we could begin EDA.

Feature	Data Type	Description
Year	Date, Integer	Year of accident
Month	Date, Integer	Month of accident
Day of Week	Date, Integer	Day of accident
ADDRTYPE	Multi-class Object, String	Type of Address (Block, Intersection)
JUNCTIONTYPE	Multi-class Object, String	Type of Junction (Mid-Block, Driveway junction etc.)
LONGITUDE	Geodata, Float	Longitudinal Coordinates of Accident
LATITUDE	Geodata, Float	Latitudinal Coordinates of Accident
COLLISIONTYPE	Multi-class Object, String	Type of Collision (Angle, Rear Ended etc.)
WEATHER	Multi-class Object, String	Weather (Clear, Raining etc.)
ROADCOND	Multi-class Object, String	Road Condition(Dry, Wet etc.)
LIGHTCOND	Multi-class Object, String	Lighting Condition (Daylight, Dark etc.)
UNDERINFL	Binary-class Object, Integer	Whether driver was under the influence(0,1)
SPEEDING	Binary-class Object, Integer	Whether driver was over the speed limit (0,1)
INATTENTIONIND	Binary-class Object, Integer	Whether driver was inattentive (0,1)
SEVERITYCODE	Binary-class Object, Integer, Target Variable	Whether accident caused only property damage or Injury (0,1)

We selected these 15 attributes from our initial 37 to proceed forward with our Exploratory Data Analysis for now. We will apply some further transformations and reductions down the line during the Data modeling phase to make the data suitable for training and testing purposes.