SEATTLE CAR ACCIDENT SEVERITY PREDICTION

Introduction

Context

Road traffic accidents are a major source of human and economic hardship. Road accidents occur every day. Some are severe, some leave people injured and some leave properties like cars damaged. It is estimated that road traffic accidents cost the United States' economy ~ $810 billion per year, including costs due to property damage, legal costs and associated medical bills. It is therefore important that we understand the factors influence the likelihood of a road traffic accident occurring at a given location, as well as those which influence the severity of the accidents.

Problem

Some of the most common factors which influence the likelihood and severity of accidents include:

- The Weather
- Road conditions
- Light conditions.

Additional factors which may influence the severity of accidents are driving at high speed, driving with consumption of alcohol etc.,

Target Audience

The target audience for this work will be city planners and emergency service responders. By understanding the key factors that influence the severity of accidents, it will be possible for local authorities to prevent or reduce the Severe or Fatal accidents in future by taking appropriate preventive measures.

A model that predicts the severity of an accident based on local and geographical conditions will help the emergency service handlers to prioritize the allocation of emergency service resources based on the information available at the time of accident reported.

Data

Data Collection

Data is obtained from all road traffic accidents recorded in the Seattle municipal area between Jan 2004–Aug 2020 by the Seattle Department of Transport (SDOT). Data is available in Seattle Open Data portal and saved as CSV.

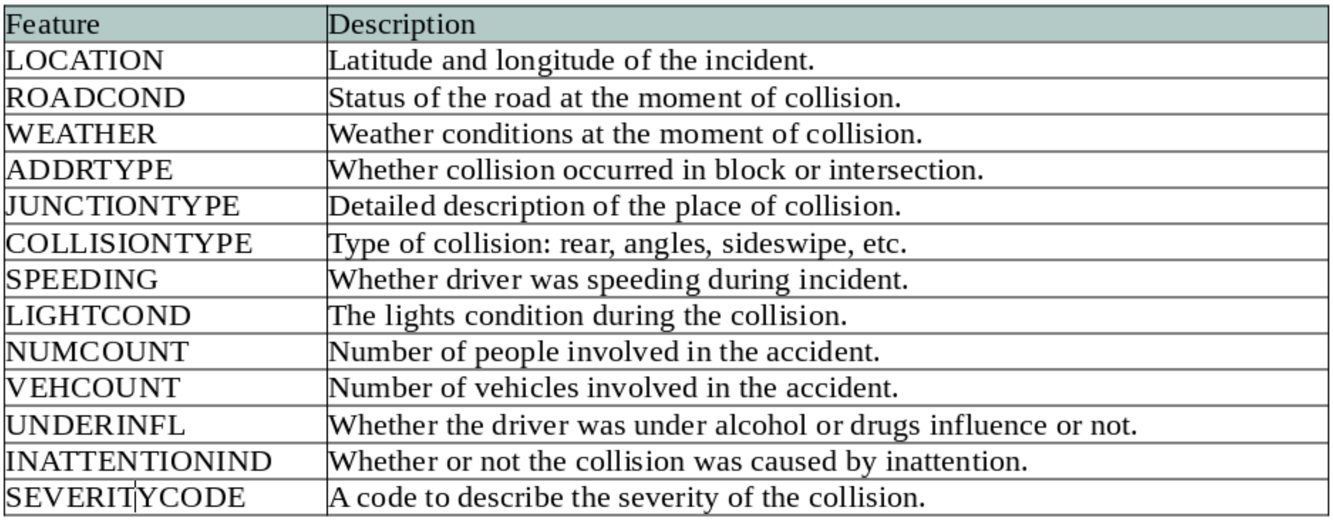
http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0

Dimensions of Data: The Dataset contains 221738 rows (accidents) and 40 columns (attributes)

The target variable is SEVERITYCODE which, in its original form, takes the values 0, 1, 2, 2b or 3. The definitions of these severity codes are as follows:

- **0:** Unknown
- **1:** Property/vehicular damage
- **2:** Minor injury
- **2b:** Serious injury
- **3:** Fatality

Some of the main features of the data and their description are as follows:



Feature	Description
LOCATION	Latitude and longitude of the incident.
ROADCOND	Status of the road at the moment of collision.
WEATHER	Weather conditions at the moment of collision.
ADDRTYPE	Whether collision occurred in block or intersection.
JUNCTIONTYPE	Detailed description of the place of collision.
COLLISIONTYPE	Type of collision: rear, angles, sideswipe, etc.
SPEEDING	Whether driver was speeding during incident.
LIGHTCOND	The lights condition during the collision.
NUMCOUNT	Number of people involved in the accident.
VEHCOUNT	Number of vehicles involved in the accident.
UNDERINFL	Whether the driver was under alcohol or drugs influence or not.
INATTENTIONIND	Whether or not the collision was caused by inattention.
SEVERITYCODE	A code to describe the severity of the collision.

Data Cleaning

Remove the data with unknown information in the Target variable

The predefined target variable in the data determines the Car Accident Severity. However, there are few rows in the data frame with 'SEVERITYCODE = 0' which means an accident with "Unknown" Severity. We cannot use these accident data with unknown information to predict the Car Accident severity. So, these rows should be dropped.

Relabel the Target Variable

The Target Variable "SEVERITYCODE" is having values (0, 1, 2, 2b, 3). It contains categorical values. So, this must be converted into numerical format. We have already dropped the rows with code value "0". So, we are left with (1, 2, 2b, 3) Relabel the codes from (1, 2, 2b, 3) to (1, 2, 3, 4)

Remove Columns with unnecessary information

Columns containing descriptions and identification numbers that would not help in the classification are dropped from the dataset to reduce the complexity and dimensionality of the dataset.

Finding missing values and handling them

Empty boxes, ‘Unknown’ and ‘Other’ were values considered as missing values. These were replaced with NA to make the dataset uniform. remove columns with more than 20% values missing.

Balancing the Data

Most of the accidents in the Data involve either property damage or minor injuries. Very few accidents involve serious injuries and Fatalities.

If we train this model, the model will be biased. If we train a model to predict severity using a dataset with majority of them having one particular outcome, then the model will become biased. To avoid this issue, we need to resample the data.

Down sampling SEVERITYCODE 1, 2 and 3 to match the number of samples in SEVERITYCODE 4.

Encoding Categorical columns and creating dummy Variables

Machine Learning model should be trained only on numerical data. So, convert all Categorical Columns to numerical format by creating Dummy Variables.

Feature Selection

The final feature set used to predict SEVERITY CODE, includes following columns:

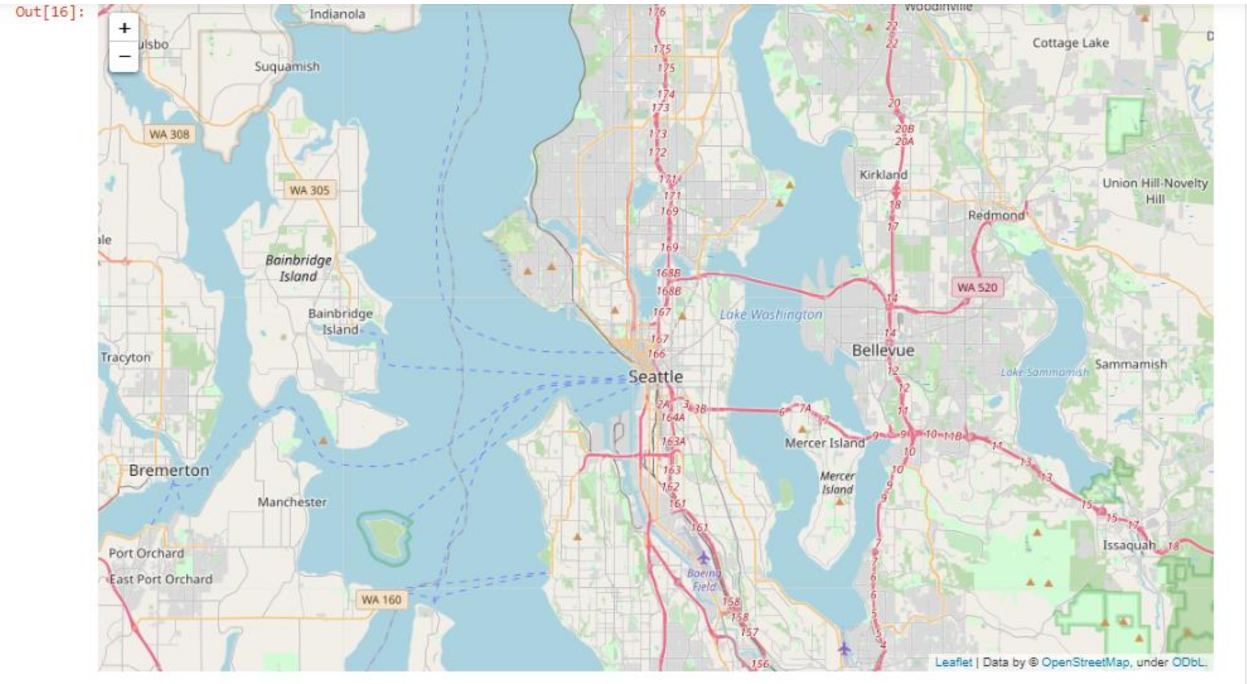
- **X, Y,** the latitude, and longitude of the accident
- **PEDOCUNT, PEDCYCLCOUNT, VEHCOUNT**, which indicate how many pedestrians, cyclists or vehicles are involved
- **INATTENTION ID, UNDERINFL, HITPARKEDCAR, SPEEDING,** which indicate if irresponsible driving is a factor
- Data from **WEATHERCOND**, which are expanded to 11 columns describing weather conditions using one-hot encoding
- Data from **LIGHTCOND**, which are expanded to 8 columns using one-hot encoding
- Data from **ROADCOND**, which are expanded to 5 columns using one-hot encoding
- **block, intersection**, which indicate the type of road on which an accident occurs

Visualizing the Data

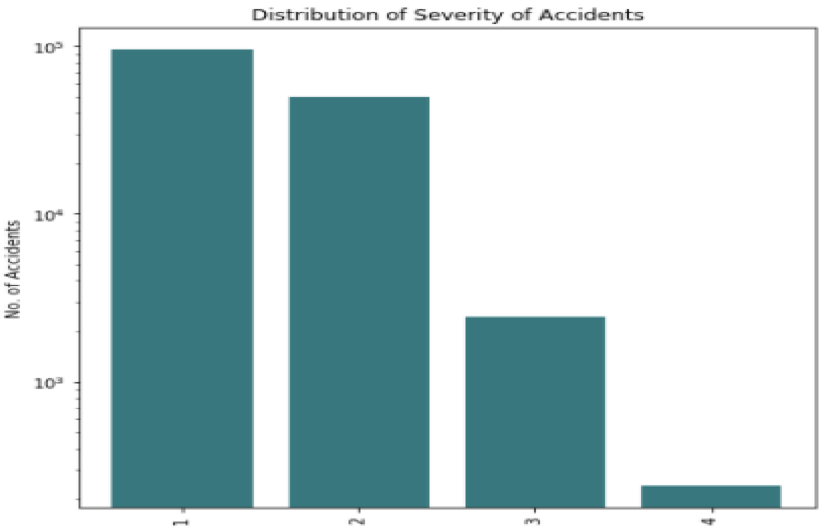
We can plot some of the key features of the dataset to gain an intuitive understanding of the Seattle car accident database.

We begin by mapping out the locations at which accidents occur, as well as the average accident severity at different locations in the city.

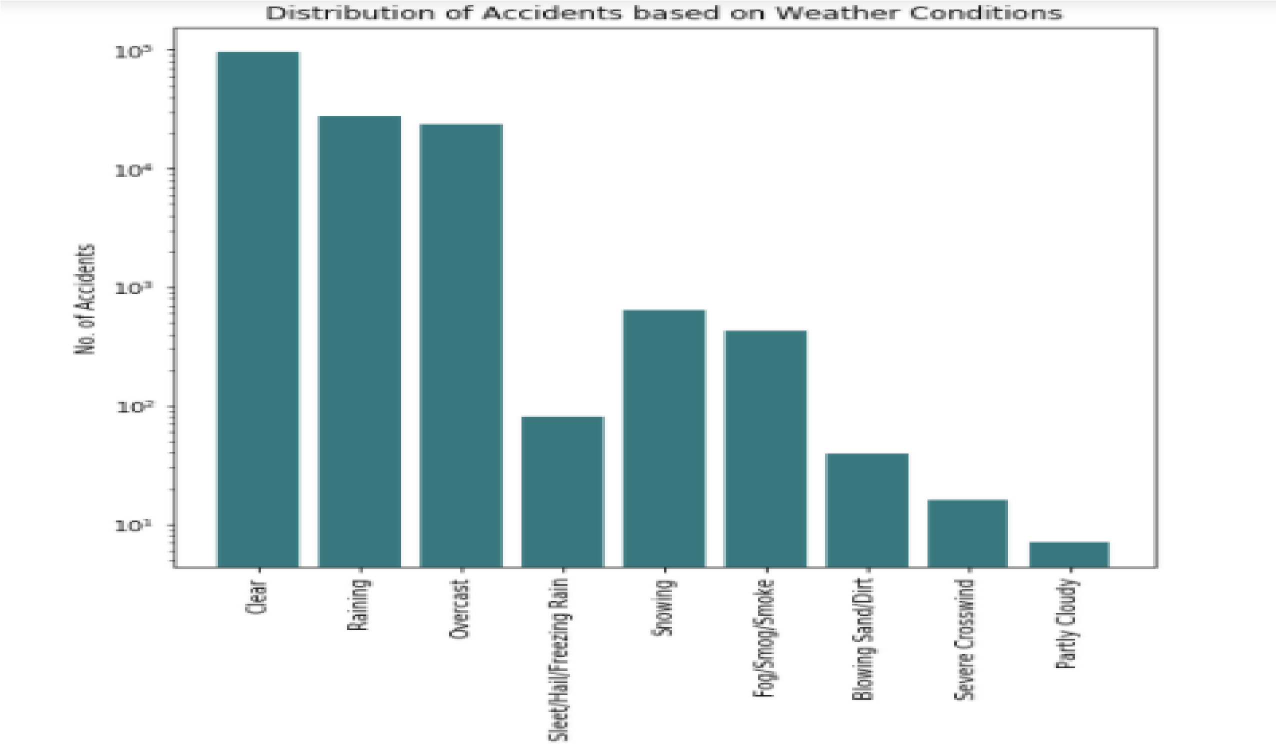
A map of the Seattle metropolitan area was drawn-up.

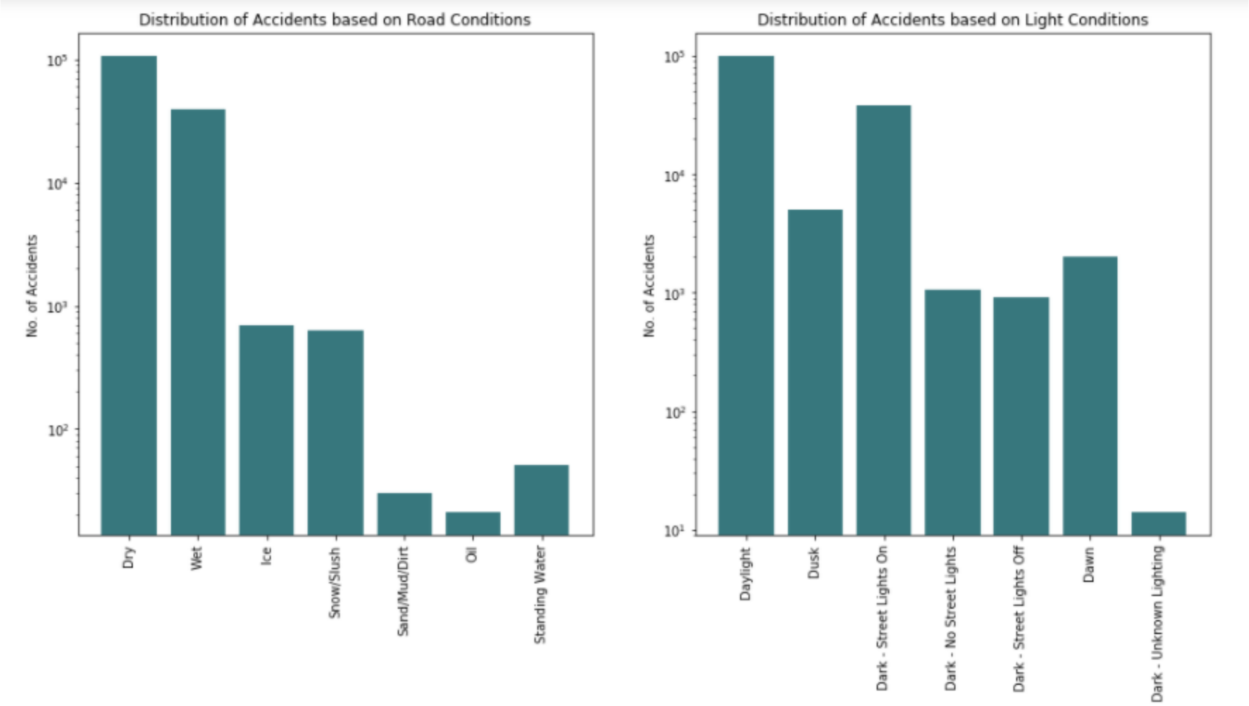


Distribution of Accidents based on Severity of an Accident



Distribution of Accidents based on Weather, Road and Light Conditions:

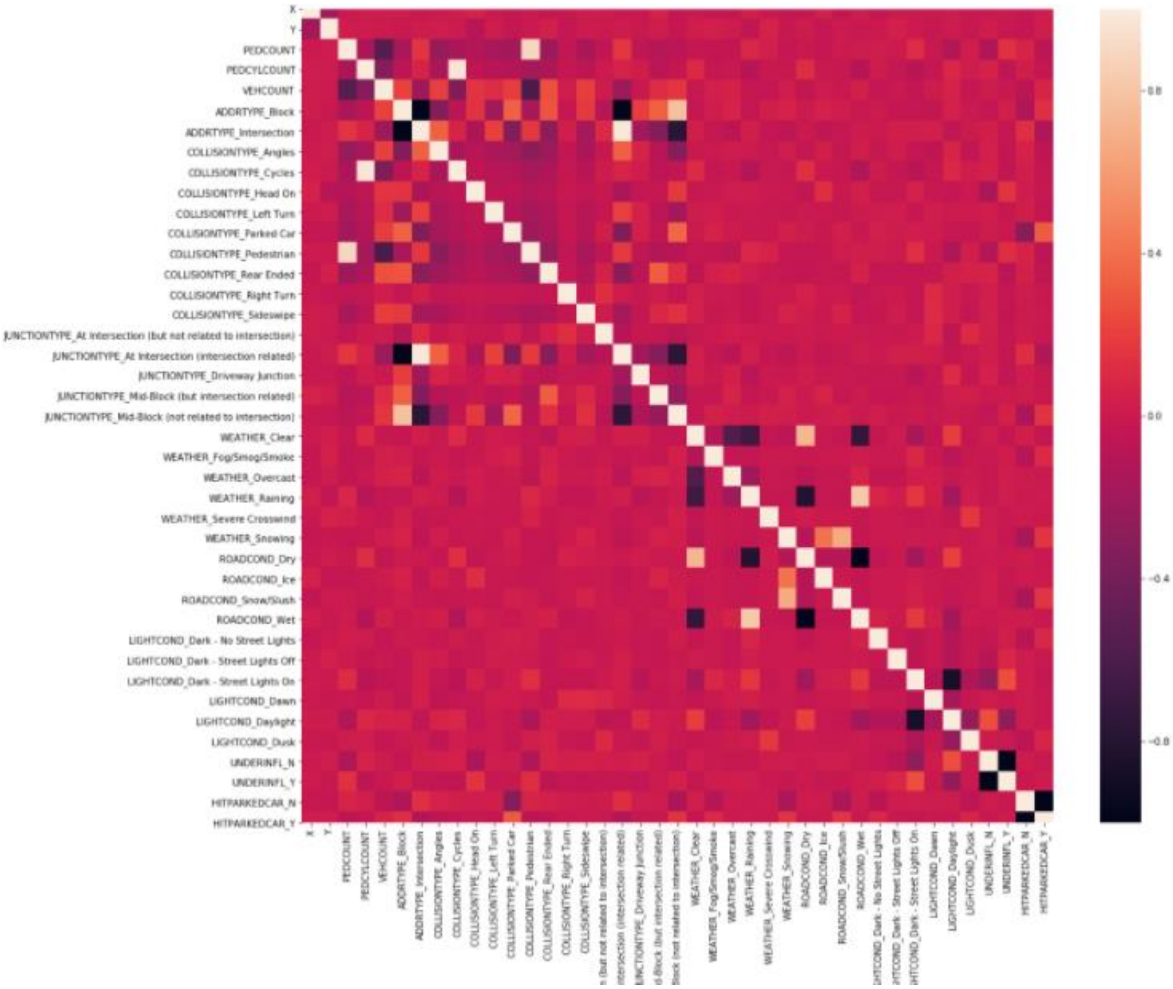




- Most accidents (75.6%) occurred in clear or overcast (i.e. dry) weather conditions. The remaining 24.4% took place either in severe conditions (such as severe winds) or during periods of precipitation (rain, snow, fog, etc.).
- Road conditions at the time of each accident. Clearly the road conditions are related to the prevailing weather at the time (e.g. if there is rain, the roads are likely to be wet), however conditions are not wholly determined by the weather. For instance, 61 accidents occurred on roads where oil was present.
- The light conditions at the time of each accident. 62.6% accidents occurred during daylight hours, while 26.2% of accidents occurred at nighttime in areas with streetlights (i.e. urban areas). The remaining 11.2% of accidents include those which happened at dawn/dusk, or on roads with no/faulty streetlights.

Correlation Matrix

The correlation matrix also provides information about which types of collisions correlate most strongly with the severity of an accident



Modelling and Evaluation

We now finally have a clean, balanced, and standardized dataset for the Seattle area. Categorical variables have been converted to numerical variables using standard data processing techniques. We are finally ready to begin building and testing models for predicting SEVERITYCODE from our chosen feature set.

The five models which will be built, tested, and compared are:

- Decision Tree
- Random forest
- KNN
- Logistic Regression
- Support Vector Machine (SVM)

Results

Model	Accuracy	Precision	recall	f1-score	Jaccard Index
Decision Tree	70%	0.71	0.70	0.69	0.70
Random Forest	70%	0.71	0.70	0.70	0.70
KNN	73%	0.73	0.73	0.74	0.73
Logistic Regression	75%	0.76	0.75	0.74	0.75
SVM	75%	0.75	0.75	0.75	0.75

From the above results Logistic Regression and Support Vector Machines gives the best Accuracy.

Conclusion

Although both has similar accuracy we prefer SVM, as it has key advantage of being able to return a ranked list of the most significant features in terms of their influence on the accident severity code (provided a linear mapping kernel is used).

The SVM model highlights that accidents involving pedestrians and multiple vehicles often have severe consequences, as do those in which excess speed is a factor. By identifying the ranking of the major causes of accident severity in this manner, it is hoped that town/city planners will be able to design new road infrastructure and target the introduction of traffic calming measures where they are most needed.

Future Work

- In future, the model could be improved to predict the accident severity on a continuum running from 1–4, rather than simply predicting a binary accident severity of 0 (minor) or 1 (major).
- In future, it may be worth revisiting this work and modelling the accident data, to see if the features which best predict accident severity have changed over time.