Car Accident Severity

IBM Capstone Project

Prepared by: Yevgeniy Dobrynin

13-Oct-2020

Introduction

Statistically the most dangerous transport is car. Approximately 1.35 million people die each year as a result of road traffic crashes. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. Seattle city has an intensive traffic and as the result frequent collision accidents with fatalities and heavy injuries. From 2004 to 2020 it was recorded 78,596 injuries and 328 fatalities as the result of road accidents in Seattle city. Analysis of such data could reveal dangerous areas and other factors leading to high probability of accidents to occur.

Stakeholders who may benefit from the analysis are: drivers, insurance companies, hospitals and police.

Data Preparation

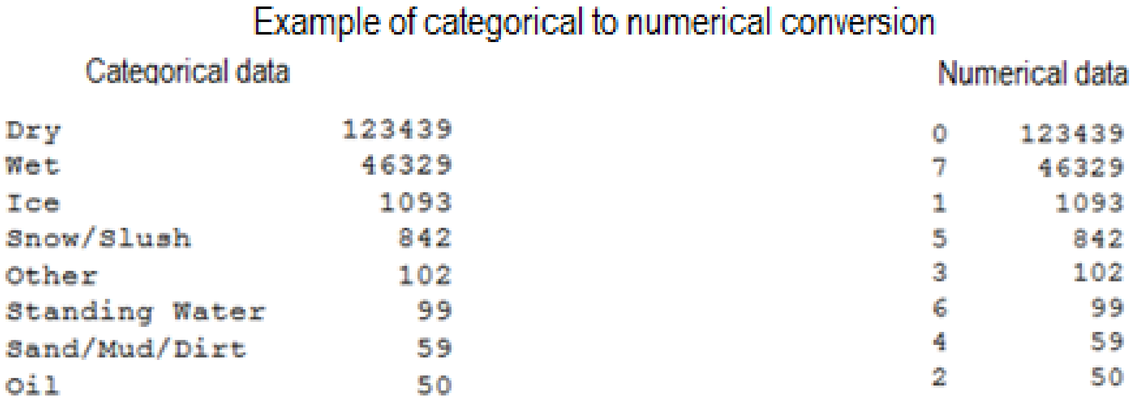
The information was provided by Seattle Police Department from 2004 to 2020 containing 221,737 records.
Database was obtained from the following link:
https://data-seattlecitygis.opendata.arcgis.com/datasets/collisions
Dataset contains 39 columns of accident records.

Using the Severity Code assigned to each accident we could identify accident severity

- 0: Unknown/no data
- 1: Property damage only
- 2: Minor injury collision
- 2b: Major injury collision
- 3: Fatality collision

Dataset in their original form is not suitable for analysis, thus the following preprocess were applied:

1. Drop columns that is not going to be used in analysis. 26 columns were dropped in total including unnecessary and redundant columns.
2. Data incompleteness. Around 15% of the data were missing one or more key features required for analysis. As the result of data cleaning and preprocessing such data were dropped from the dataset including NaN and Unknown values.
3. Convert categorical data to numerical. In order to apply Machine Learning models it requires to convert all categorical data to numeric. As an example, “ROADCOND” column contains strings that can’t be directly processed by Machine Learning Engines. Picture below demonstrates conversion from categorical to numerical string.
4. The same transformation were also applied to “SEVERITYCODE”, “WEATHER”, “LIGHTCOND” and “SPEEDING”



Example of categorical to numerical conversion

Categorical data

```
Dry                 123439
Wet                  46329
Ice                   1093
Snow/Slush             842
Other                  102
Standing Water          99
Sand/Mud/Dirt           59
Oil                     50
```

Numerical data

```
0     123439
7      46329
1       1093
5        842
3        102
6         99
4         59
2         50
```

5. Data balancing and standardization. The target variable for this study is SEVERITYCODE which represents severity of accidents contains the following values

1: Property damage only	-	113156
2: Minor injury collision	-	55584
2b: Major injury collision	-	2945
3: Fatality collision	-	328

This real-life representation of accidents outcome may bias the model; thus, the model has balanced and binarized to the following 2 categories.

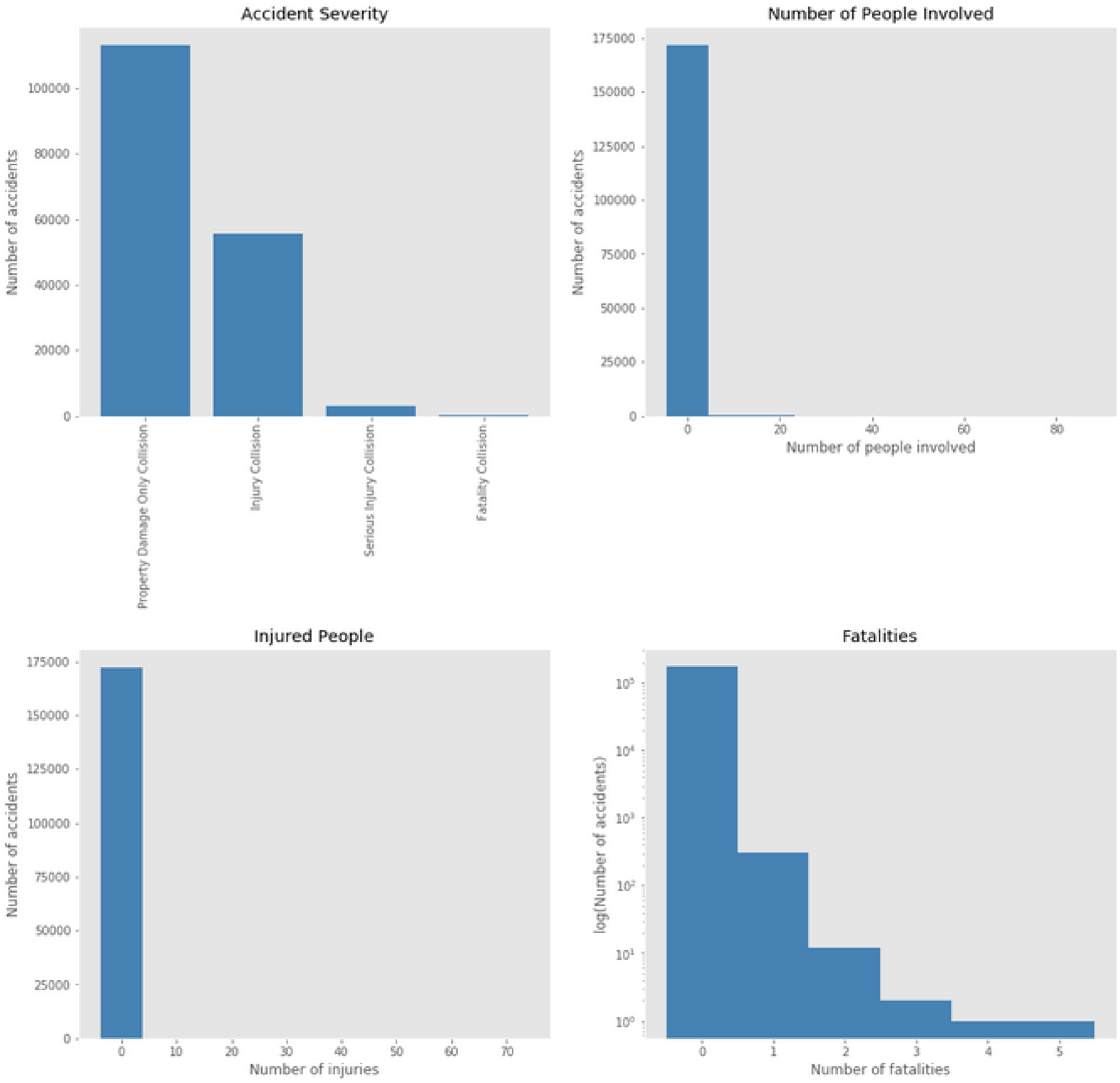
0 – Minor accidents that will include property damage only
1 – Major accidents that will include all type of injuries: minor, major injuries and fatality collisions

Binarized and balanced result

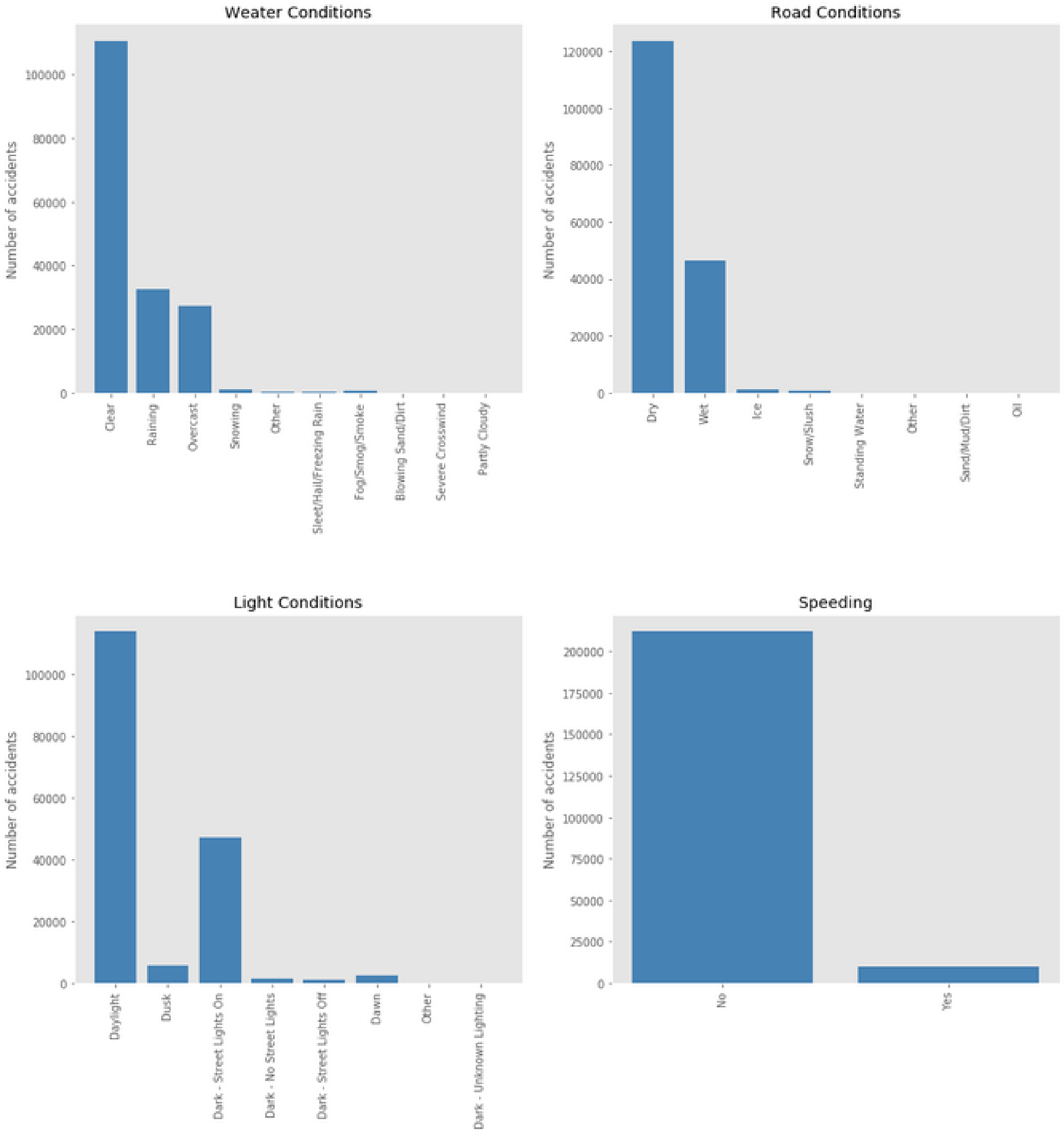
0 – 58587
1 – 58587

Plotting the data to see values distribution

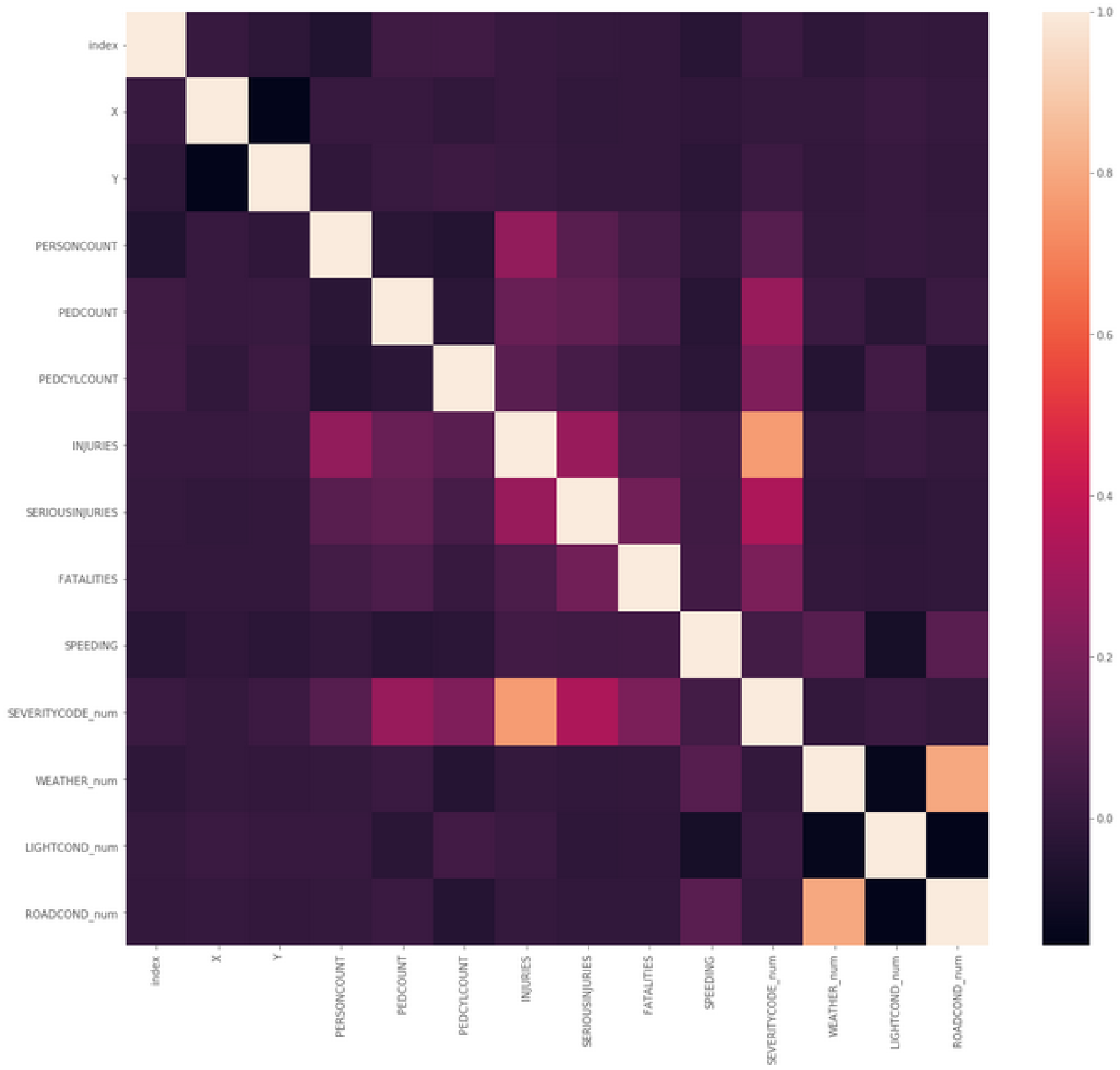
Visualize quantities of accidents and number of people involved



Visualize parameters that affect accident severity



Build correlation matrix to discover which columns need to be included for model building



Methodology and Model Preparation

The project data analysis was performed in Python Jupyter Notebook.
Data for road accident in Seattle city were cleaned and categorized for the features below to predict incident severity

Feature Categorization

LIGHTCOND:

Daylight = 5, Dark — Street Lights On = 2, Dark — No Street Lights = 0, Dusk = 6, Dawn = 4, Dark — Street Lights Off = 1 Dark — Unknown Lighting = 3 Other = 7

WEATHER:
Clear = 1, Raining = 6, Overcast = 4, Other = 3, Snowing = 9, Fog/Smog/Smoke = 2, Sleet/Hail/Freezing Rain = 8, Blowing Sand/Dirt = 0, Severe Crosswind = 7, Partly Cloudy = 5

ROADCOND:
Dry = 0, Wet = 7, Ice = 1, Snow/Slush = 5, Other = 3, Standing Water = 6, Sand/Mud/Dirt = 4, Oil = 2

SPEEDING:
Yes = 1, No = 0

Result Categorization

SEVERITYCODE:
1: Property damage only = 0, 2: Minor injury collision = 1, 2b: Major injury collision = 3, 3: Fatality collision = 3

Further SEVERITYCODE were binarized per the rule below

0 – Minor accidents that will include property damage only
1 – Major accidents that will include all type of injuries: minor, major injuries and fatality collisions

Binarizaton result was: 0 - 113156, 1 – 58587. The result was balanced to have equal ammoun of both outcome
Final result had the following structure: 0 – 58587, 1 – 58587.

Model Preparation

Features: WEATER, LIGHTCOND, ROADCOND, SPEEDING, PERSONCOUNT, PEDCOUNT, PEDCYLCOUN
WEATER – weather condition during the time of accident
LIGHTCOND – light condition on the road during the time of accident
SPEEDING – whether or not speeding was a factor of the collision
PERSONCOUNT – the total number of people involved in collision
PEDCOUNT - the number of pedestrians involved in the collision
PEDCYLCOUN – the number of bicycles involved in the collision

Result: SEVERITYCODE
0 – Minor accidents that will include property damage only
1 – Major accidents that will include all type of injuries: minor, major injuries and fatality collisions

K Nearest Neighbor

kNN models seek to categorize the outcome of an unknown data sample based on its proximity in the multidimensional hyperspace of the feature set to its “k” nearest neighbors, which have known outcomes. Establishing the value of “k” which optimizes the model’s accuracy (between 1 and the total number of samples in the dataset) is an empirical undertaking: if too-few neighboring datapoints are used, the model is susceptible to being dominated by noise, however if too many neighbors are included in the classification, the model risks losing all diagnostic power completely. kNN models were built for k=2–10 using the kNeighborsClassifier function from scikit learn. The model is optimised at k=7 at which the model correctly predicts SEVERITYCODE=0 61% of the time, correctly predicts SEVERITYCODE=1 65% of the time, and has F1 scores of 0.65 and 0.59 for both accident outcomes

Decision Tree

Decision tree models identify the key features on which the data can be partitioned (and the thresholds at which to partition the data) in the hope of arriving, after some iterations, at “leaves” which contain only accidents belonging to one target variable value (in this case, accident severity code). A decision tree model was trained on the data according to the “entropy” criterion and allowed to run until reach convergence. The decision tree correctly predicts SEVERITYCODE=0 63% of the time, correctly predicts SEVERITYCODE=1 67% of the time and has F1 scores of 0.67 and 0.62 for the two accident classifications.

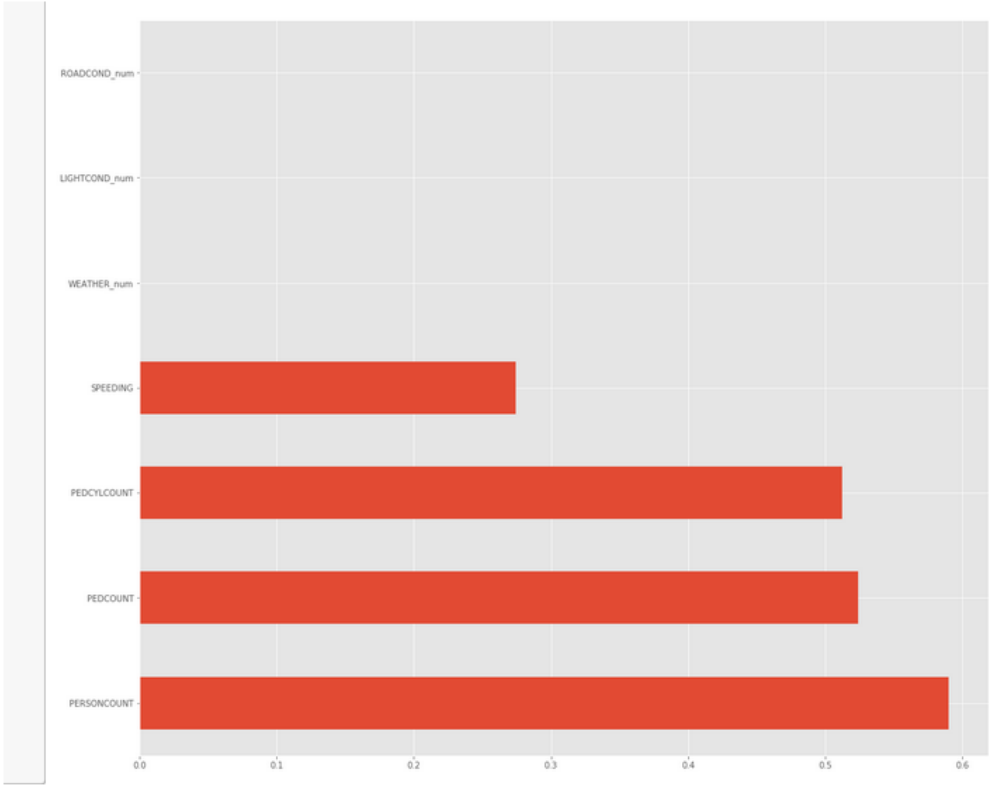
Logistic Regression

Converting the accident severity to a binary variable (0 for no/minor injuries, 1 for major injuries/fatalities) we can employ Logistic Regression techniques to attempt to classify accident outcomes based on the properties in the feature set. A Logistic Regression model was trained using an inverse-regularisation strength C=0.01 and tested on the testing subset. The Logistic Regression model correctly predicts SEVERITYCODE=0 accidents 59% of the time, correctly predicts SEVERITYCODE=1 accidents 75% of the time, and has F1 scores of 0.70 and 0.53 for the two outcomes.

Support Vector Machine

An SVM model was built using the scikit learn C-Support Vector Classification method (svm.svc), with a linear mapping kernel employed in order that the model could return a list of the features with the most diagnostic power for determining accident severity. The SVM model correctly predicts SEVERITYCODE=0 58% of the time, correctly predicts SEVERITYCODE=1 79% of the time, and has F1 scores of 0.71 and 0.48 for both accident outcomes

SVM model output showing the top ten features with the most influenced feature for determining accident severity ranking from least to most significant



Conclusion and Discussion

The study represents analysis of the data from Seattle Department of Transport to predict severity of accident based on available information of 221,737 accidents records from 2004 to 2020.

Four classes of models have been trained and evaluated: k-Nearest Neighbours, Decision Tree , Logistic Regression, Support Vector Machine. All the models showed almost similar result having F1 score 0.6-0.64 with the best performance of Decision Tree having average F1 scores of 0.64.

By knowing where condition when an accident happens and probability of having injuries to be evaluated which can benefit the following parties

a) Give an idea for drivers regarding the most dangerous places and conditions that they can optimize their travel to minimize risks of accidents
b) Allow Emergency Services to properly allocate their resources to minimize response time
c) Results could also be interested by Insurance Companies to evaluate insurance packages

This work highlights that machine learning technique can be used to analyze historical data and make reliably predictions for the accident's outcome. Model can be adapted to any road traffic network in any part of the world in which sufficient accident data are recorded.