Predictive Analysis of Accident Severity

Rodrigo Moreira Ferreira

October 2020

Table of contents

1. Introduction

The Seattle government is concerned with the severity and number of accidents and wants to employ actions based on the analysis of historical data to alert drivers of the imminence of accidents. This study aims to predict the severity of an accident, given the locations, weather and road conservation. The analysis aims to determine a set of possible causes that contribute to the increase in the severity of accidents to allow preventive actions by road users. The targets audience of the project are drivers, rescue groups, police and insurance companys. It's expected to achieve a reduction in the number and severity of accidents to make drivers and passengers more secure.

2. About dataset

The data set is about collisions that occurred between 2004 and 2020 in the city of Seattle. The data set includes details of collisions provided by the Seattle Department of Transportation Traffic Management Division. It includes following fields:

Field	Description
OBJECTID	ESRI unique identifier
LATITUDE (X)	ESRI geometry field
LONGITUDE (Y)	ESRI geometry field
ADDRTYPE	Collision address type (Alley/Block/Intersection)
SEVERITYCODE	A code that corresponds to the severity of the collision (3 — fatality/2b — serious injury/2—injury/1 — prop damage/0 — unknown)
COLLISIONTYPE	Collision type
INCDTTM	The date and time of the incident
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
WEATHER	A description of the weather conditions during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)

3. Data Wrangling

Data Acquisition

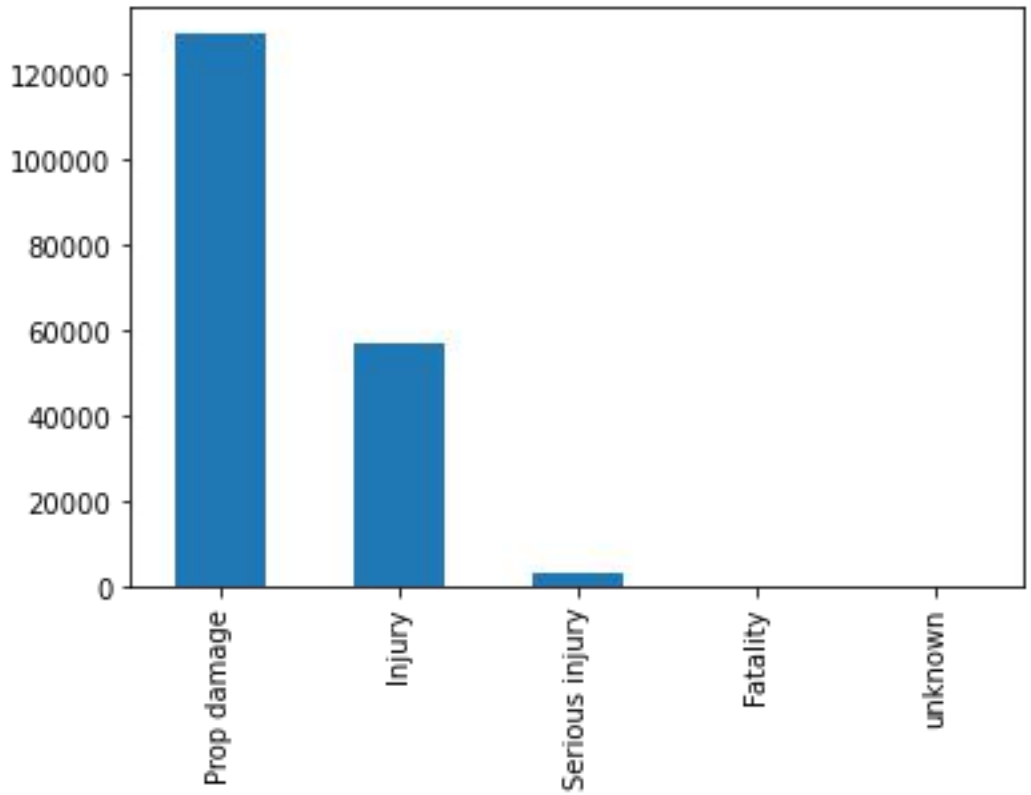
The data set was downloaded and loading the data to a the CSV file.

During the data cleaning process, unnecessary columns were eliminated and all rows that do not contain enough data to complement the analysis were eliminated. Then the data types were converted to the appropriate formats.

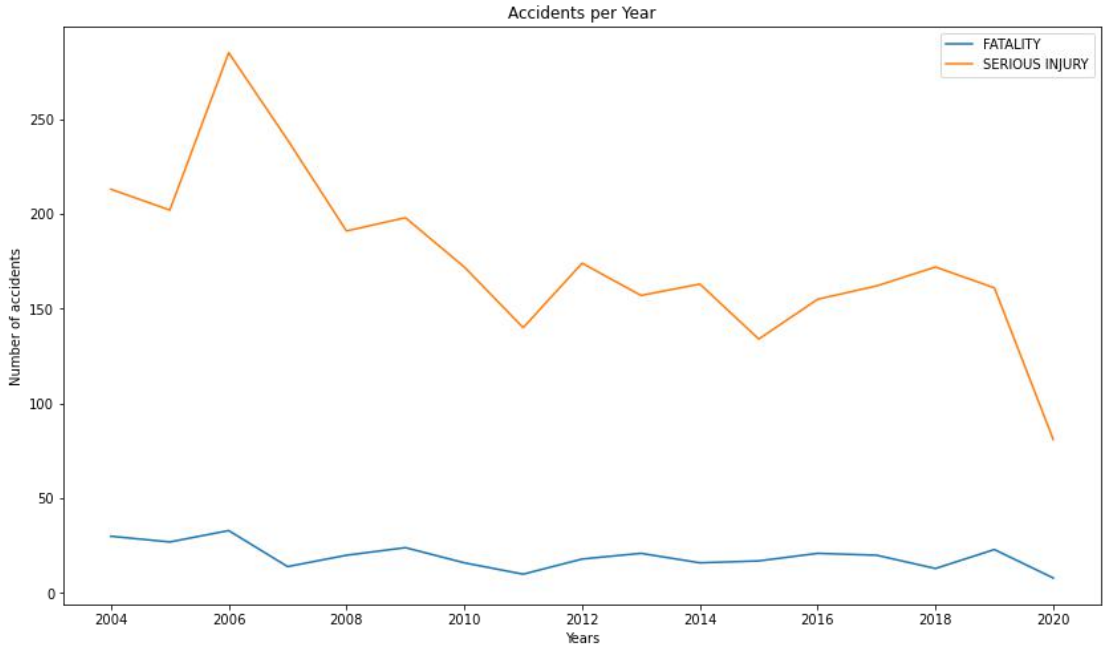
4. Data Visualization and Analysis

For the process of visualization and analysis of the data, histograms, line graphs and the plotting of the geographic location of the accident locations were used.
The results of the processing are presented below:

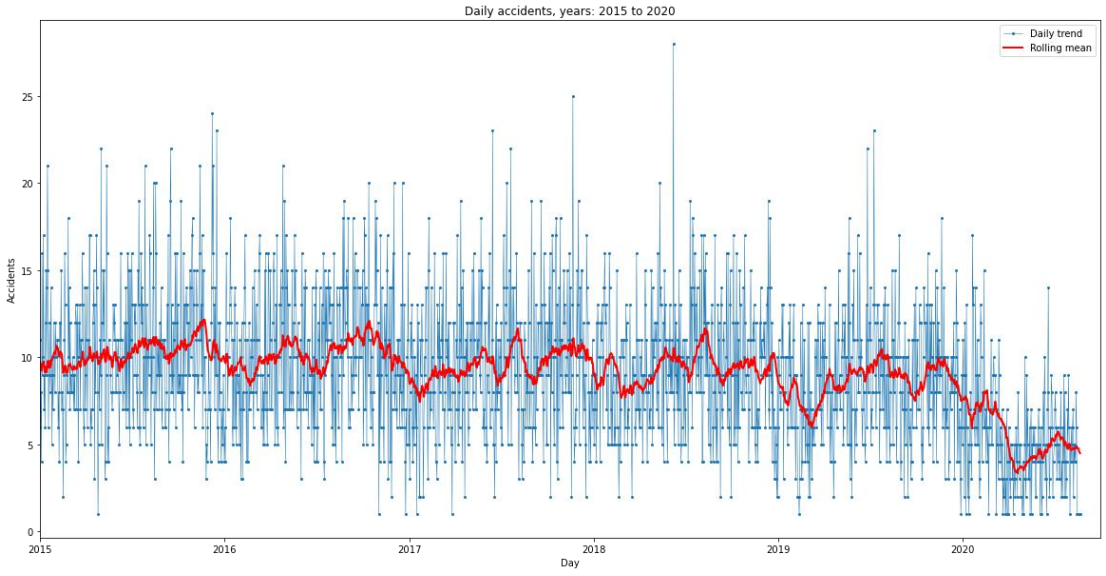
Accidents Severity



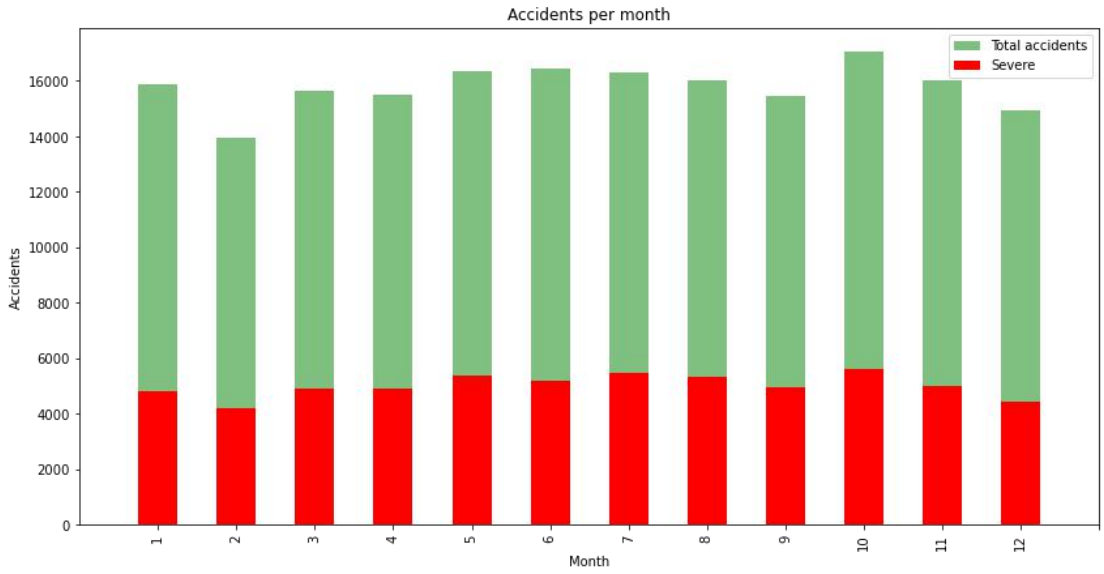
Fatality vs Serious injury



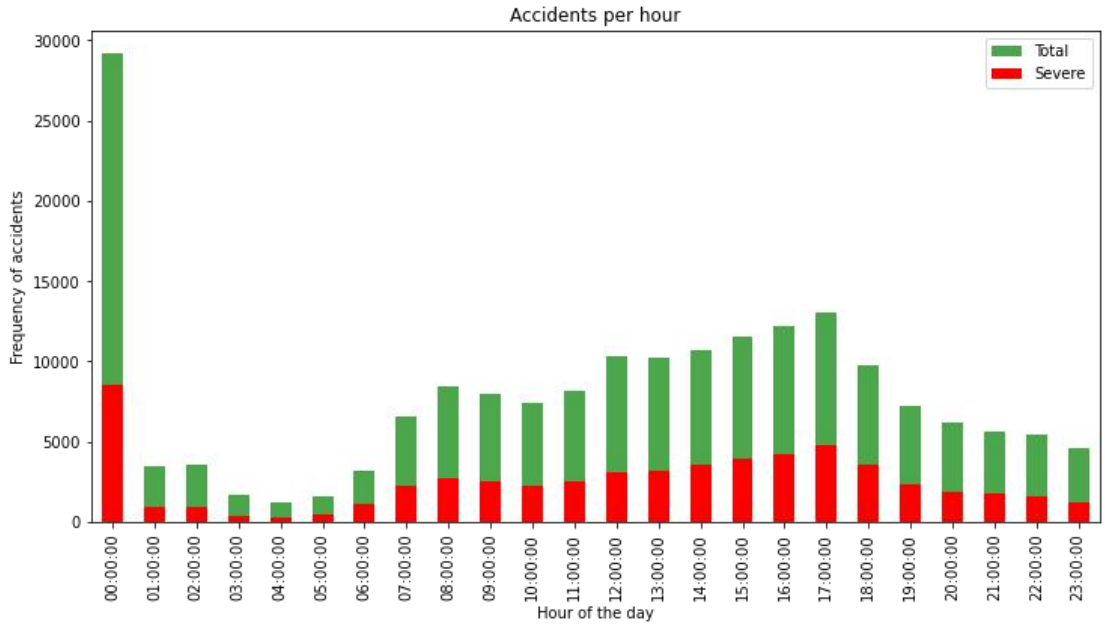
Accident chart by Year(2015 - 2020)



Accident chart by Month



Accident time



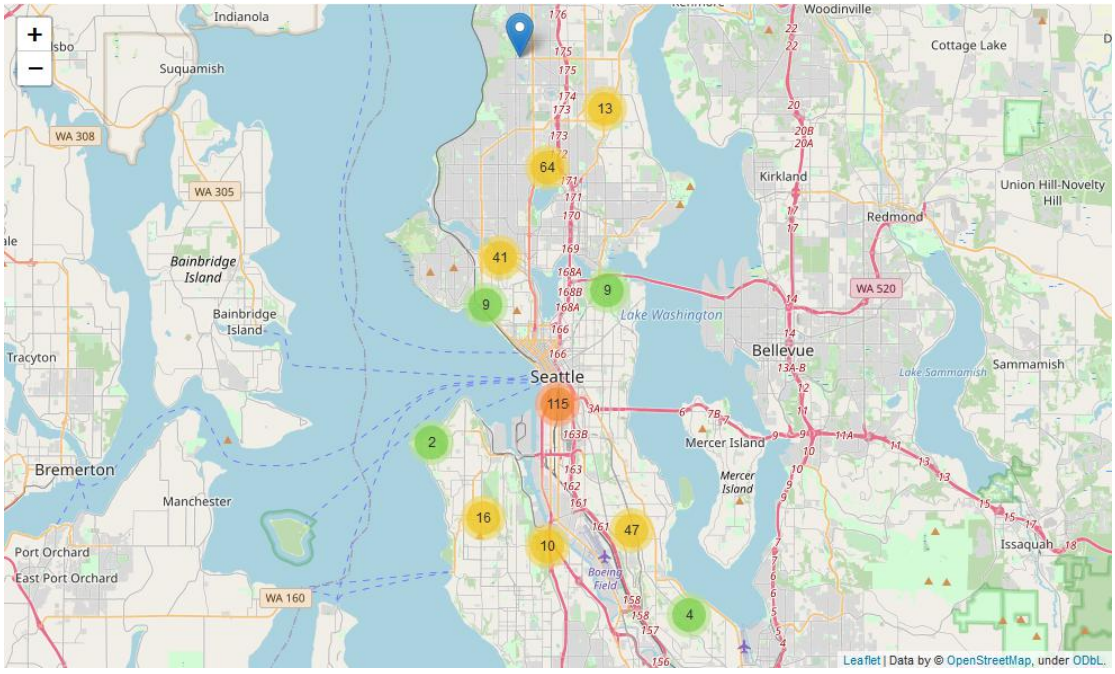
Considering the discrepancy in the frequency of accidents at 00:00, it is possible that there is a need to improve the collection of information.

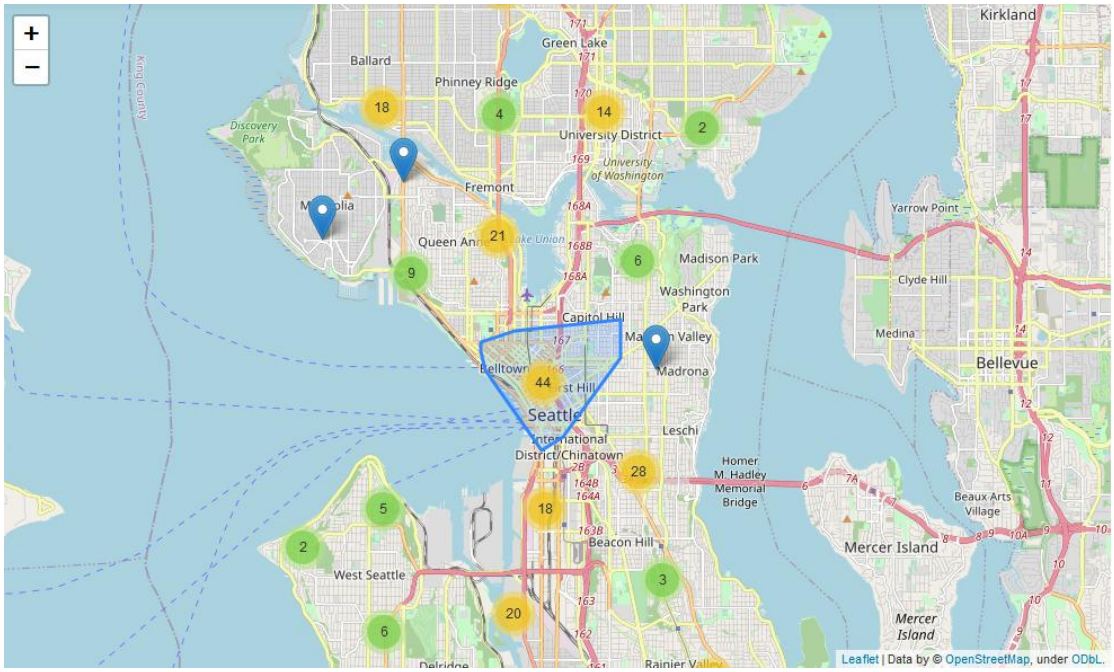
A correlation analysis of the variables used to assist the exploratory analysis was also included. The results of the processing are presented in the table below:

	ADDRTYPECODE	COLLISIONTYPE	WEATHER	ROADCOND	LIGHTCOND	SPEEDING	FATALITY	SERIOUS INJURY	INJURY
ADDRTYPECODE	1.000000	0.242859	0.085995	0.113349	0.086780	0.058708	-0.009948	-0.044538	-0.193674
COLLISIONTYPECODE	0.242859	1.000000	0.075280	0.096692	0.109841	-0.037295	-0.005615	-0.017902	-0.237198
WEATHERCODE	0.085995	0.075280	1.000000	0.658142	0.279622	0.022704	-0.004702	-0.024438	-0.095990
ROADCONDCODE	0.113349	0.096692	0.658142	1.000000	0.333520	0.042018	-0.013347	-0.028095	-0.121714
LIGHTCONDCODE	0.086780	0.109841	0.279622	0.333520	1.000000	0.044146	0.005350	0.003271	-0.092071
SPEEDINGCODE	0.058708	-0.037295	0.022704	0.042018	0.044146	1.000000	0.039670	0.042892	0.031258
FATALITY	-0.009948	-0.005615	-0.004702	-0.013347	0.005350	0.039670	1.000000	-0.005303	-0.027332
SERIOUS INJURY	-0.044538	-0.017902	-0.024438	-0.028095	0.003271	0.042892	-0.005303	1.000000	-0.082858
INJURY	-0.193674	-0.237198	-0.095990	-0.121714	-0.092071	0.031258	-0.027332	-0.082858	1.000000

Accidents Location

The map with the geographic positioning of the occurrences was used to demonstrate possible accident concentration points.



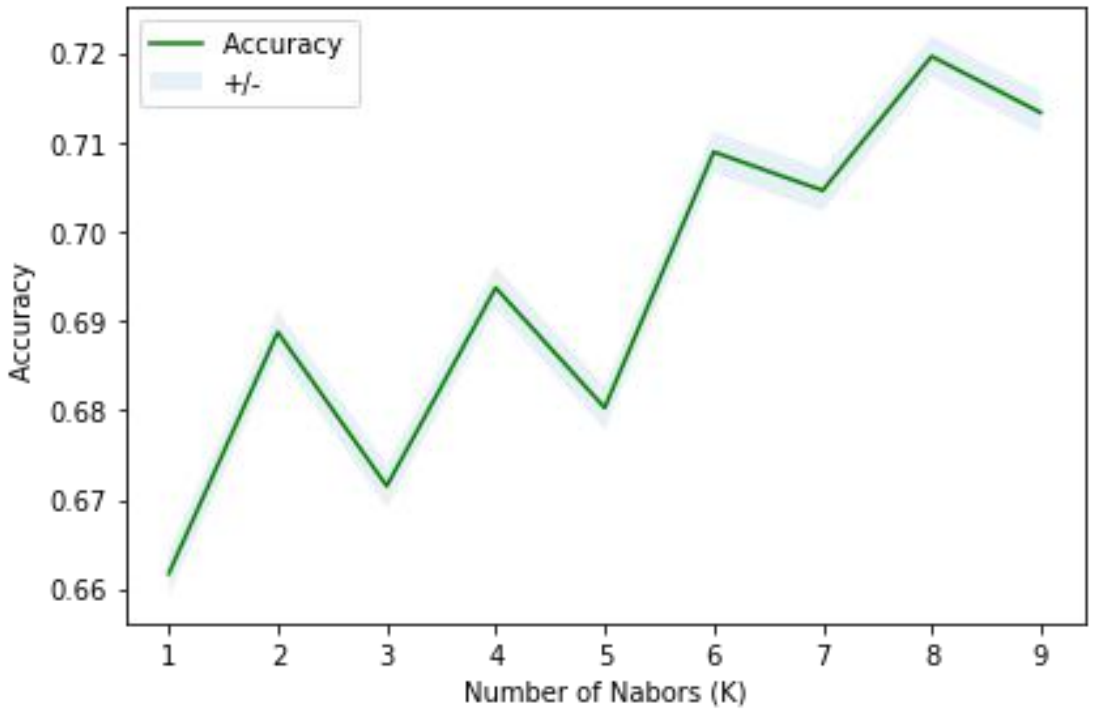


5. Classification

The K-Nearest Neighbor(KNN) algorithm was used for classification. The following characteristics of the data set will be used to make the predictions:

a. WEATHER;
b. ROADCOND;
c. LIGHTCOND;
d. SPEEDING;
e. ADDRTYPE;
f. COLLISIONTYPE.

The best accuracy was with 0.7195948619207132 with k= 8, as shown in the figure below:



The results are shown in the table below:

	Accuracy
Train	0.7268015404093691
Test	0.7195948619207132

6. Discussion of results

The positive correlation found in the SPEEDING and LIGHTCOND variables as a function of the severity of accidents indicates the need to maintain the conservation and lighting of the roads.

For this problem precision means the % of predicted severe accidents that were truly severe.

The use of a machine learning model for predicting the severity of an automobile accident reached an accuracy around of 0.726 for the training set and 0.719 for the test set.

7. Conclusion

From the analysis of the information obtained from the Seattle Department of Transportation Traffic Management Division we can conclude that the weather conditions, the state of conservation of roads, the lighting and the speed of traffic can have an aggravating impact in the case of automobile accidents.