Predictive Analytics of Car Accidents in Seattle

Chao Z.

Introduction

- The total financial cost associated with car accidents is estimated to be $230.6 billion each year in the U.S.
- In this study, a car accident data set was analyzed to explore the effects of different contributing factors.
- Predictive models were built to predict the likelihood of an accident and the severity of it for certain given input conditions.

Data

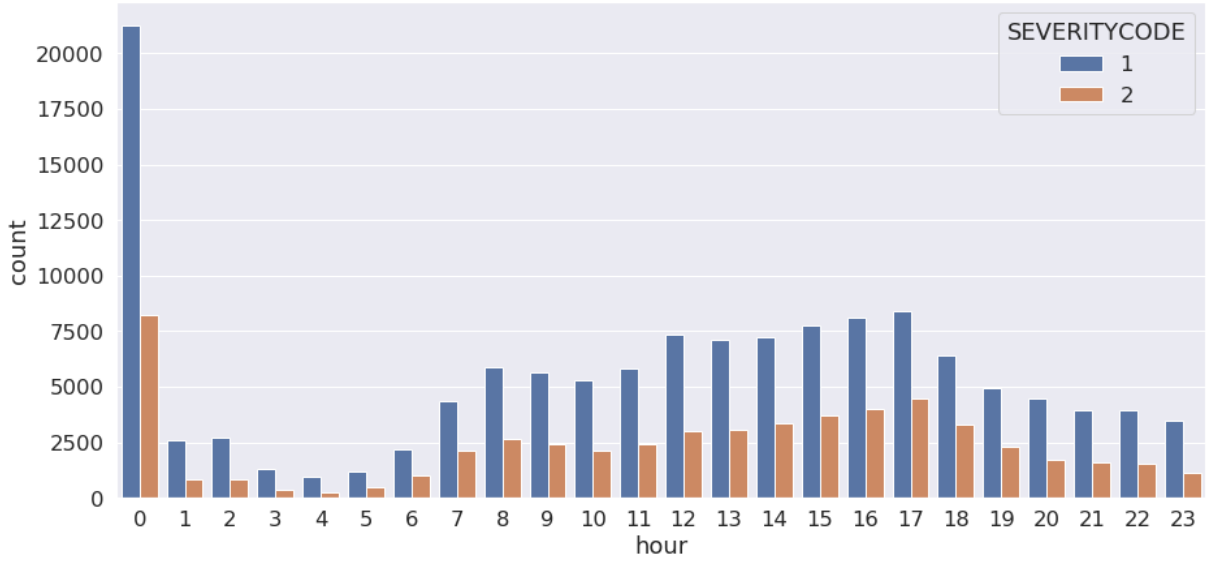
Table 1. Description of Feature Parameters in the Present Study

Type	Variable Names	Description
Time:	year	The year
	dayofweek	The day of the week
	hour	The hour of the day
Condition:	WEATHER	Weather
	ROADCOND	Road condition
	LIGHTCOND	Street lighting condition
Human error:	UNDERINFL	Whether the driver is under influence
Characteristics of the accident:	SEVERITYCODE	A code that characterizes severity of the accident
	COLLISIONTYPE	The type of the collision
	PERSONCOUNT	Number of person(s) involved
	VEHCOUNT	Number of vehicle(s) involved
	PEDCOUNT	Number of pedestrian(s) involved
	PEDCYLCOUNT	Number of bicycle(s) involved
Location:	X	Longitude of the location
	Y	Latitude of the location

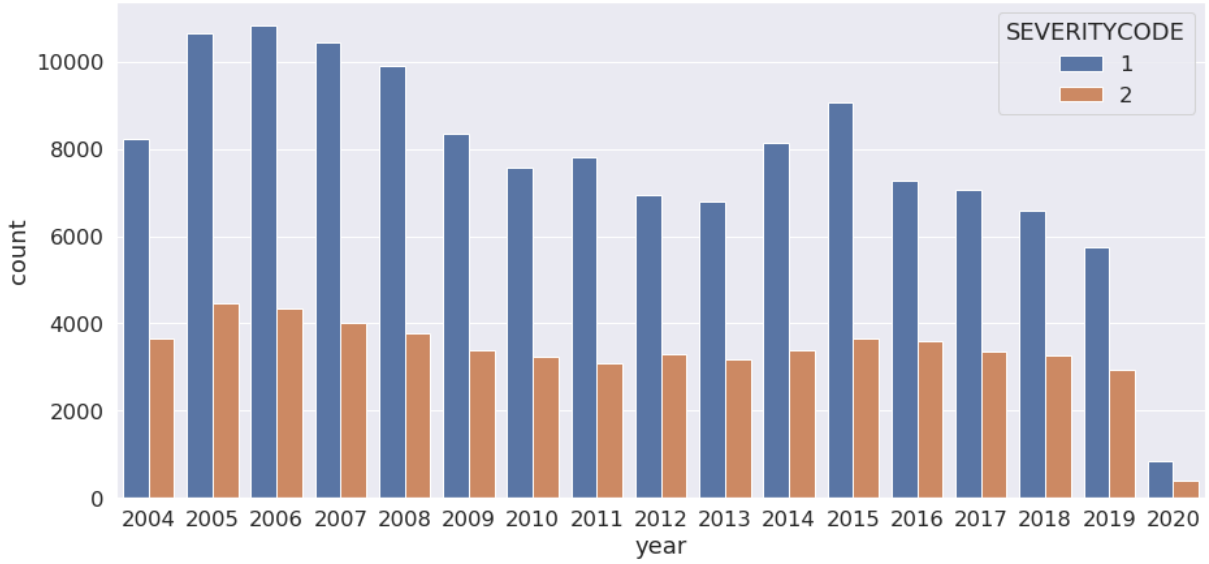
- Exploratory Analysis

Severity of Accidents

A large number of accidents, including both property damage and injuries, happen in the first hour into midnight. A second time period that appears to have a large concentration of accidents appears to be in the late afternoon.

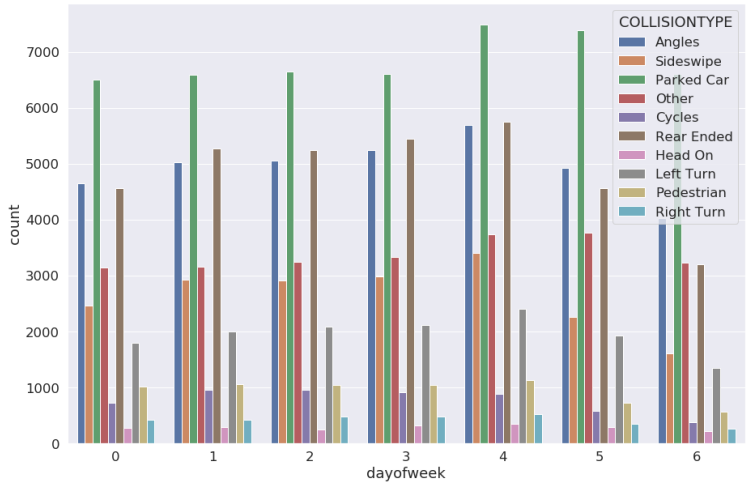


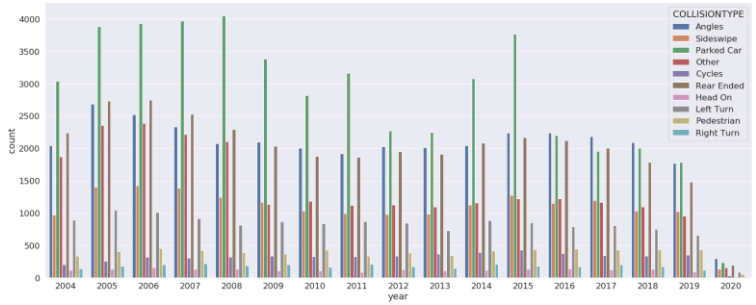
Overall there is a decreasing trend

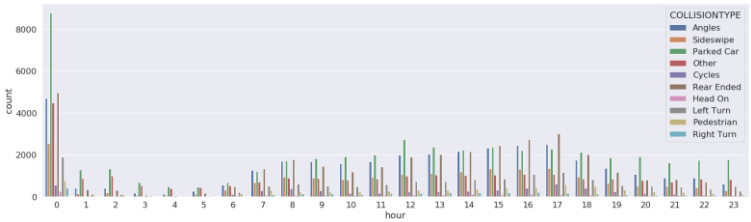


Collision Types

The “rear ended” type of collision counts generally skyrockets in the late afternoon around 16:00 – 17:00, and becomes the most dominant type of collisions in this time period, when typically people are trying to get home. During other hours of the day, collision with a parked car appears to be the most common collision type.

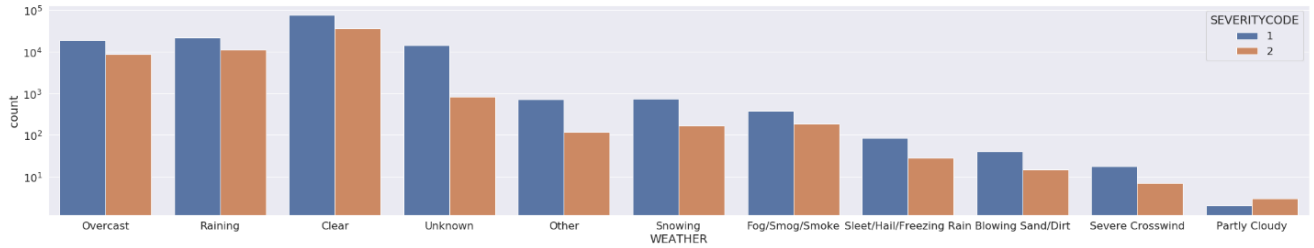


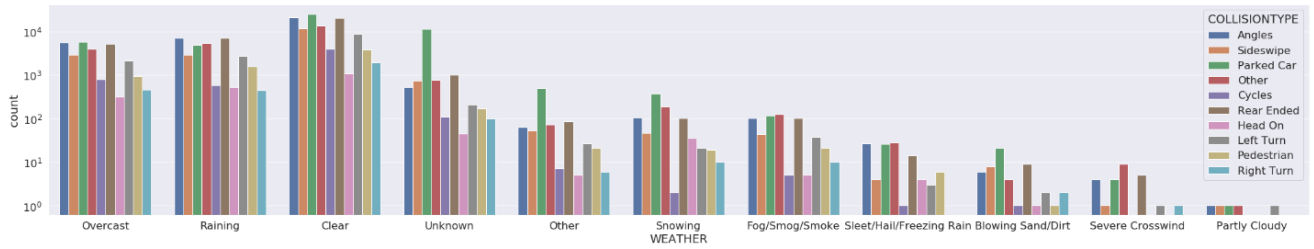




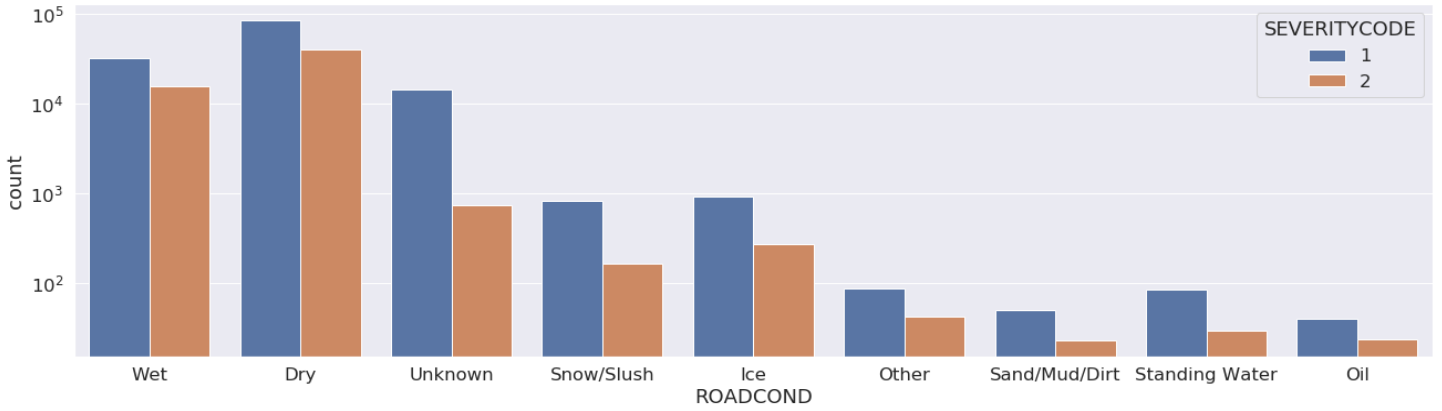
Weather

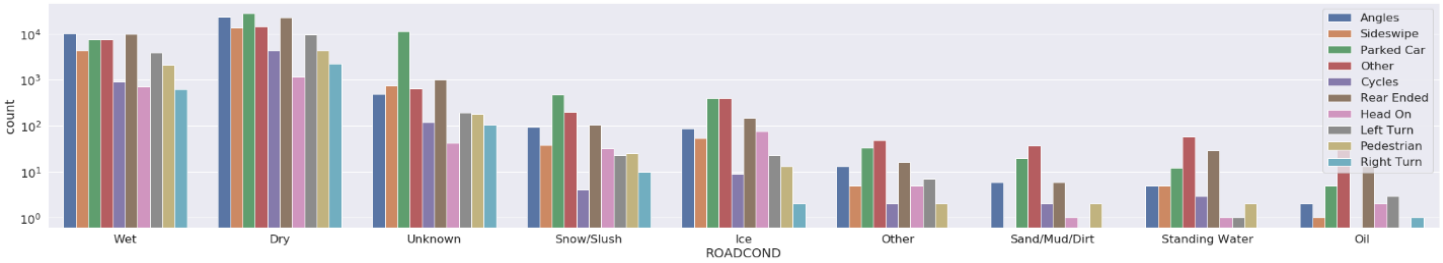
- the total number of accident cases in this weather condition category is very small, and thus it cannot be concluded that partly cloudy weather is related to more injuries than property damage accidents.

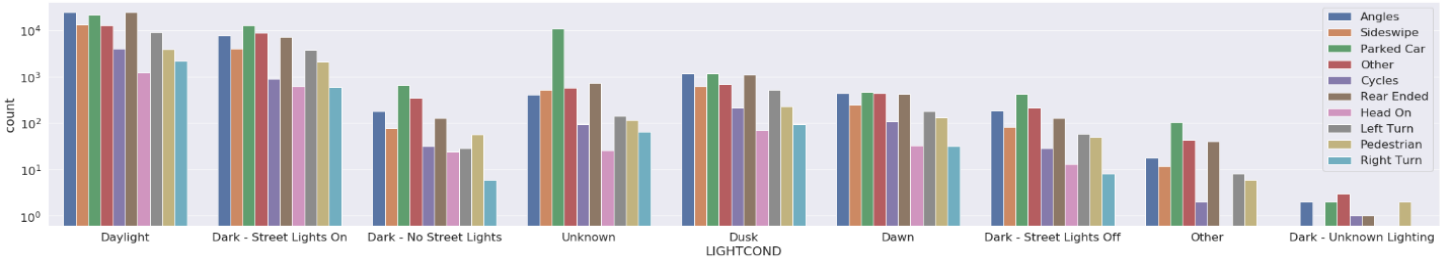




Road and Lighting Condition

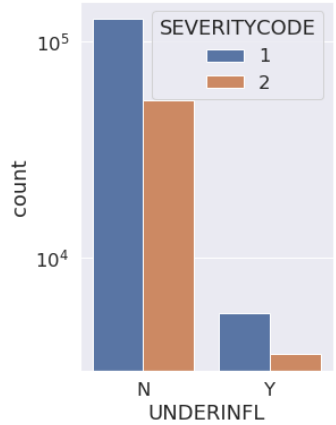


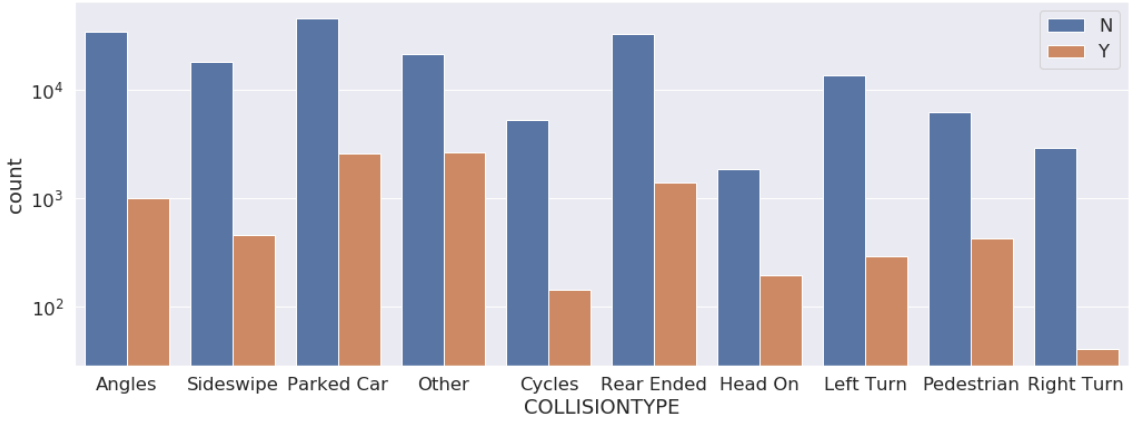




Under Influence

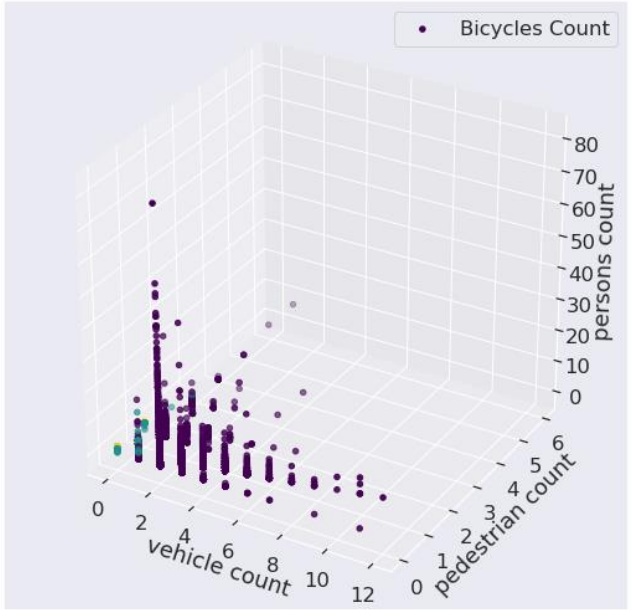
- Whether the person is under influence has some relation to the type of collisions. From Fig. 14, when the person is under influence, the counts of certain collision types are greater than others. The most common types of collisions when a person is under influence are: “Parked Car,” “Rear Ended,” “Angles.” The least common types of collision involving a person under influence are: “Right Turn” and “Cycles.”

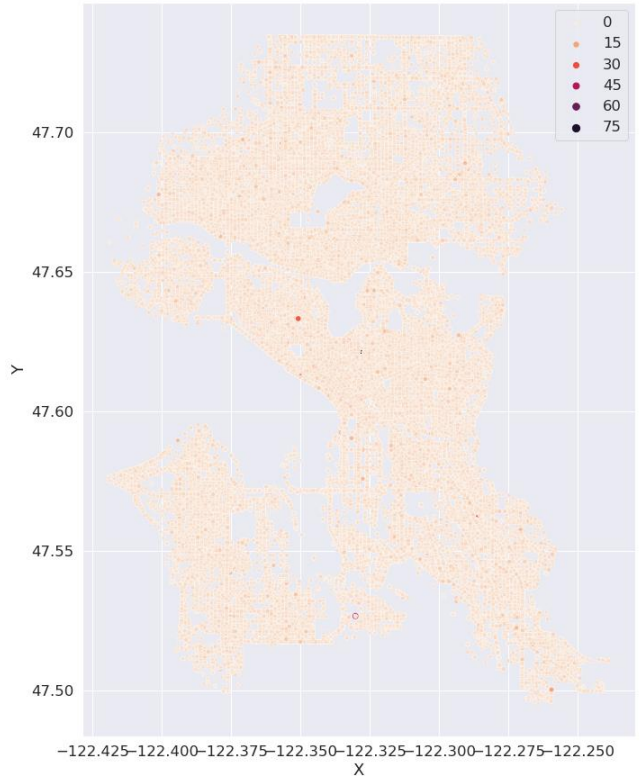




Counts of Persons, Vehicles, Pedestrian, and Bicycles

The data of persons count, vehicle count, pedestrian count and bicycle count are also analyzed. From Figs. 15 and 16, one observes that the vast majority of accidents involve less than 5 persons, 2 vehicles, and 0 pedestrian and 0 bicycles. An interesting finding is that some of the accidents that involve over 20 people only involve 1 to 3 vehicles. Another finding is that most of the bicycle-involved accidents appear to involve only one vehicle.

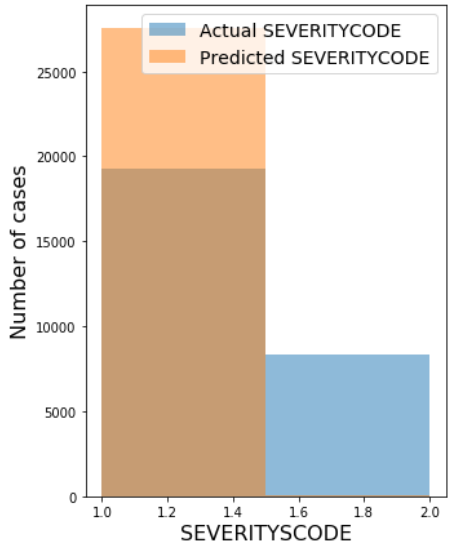


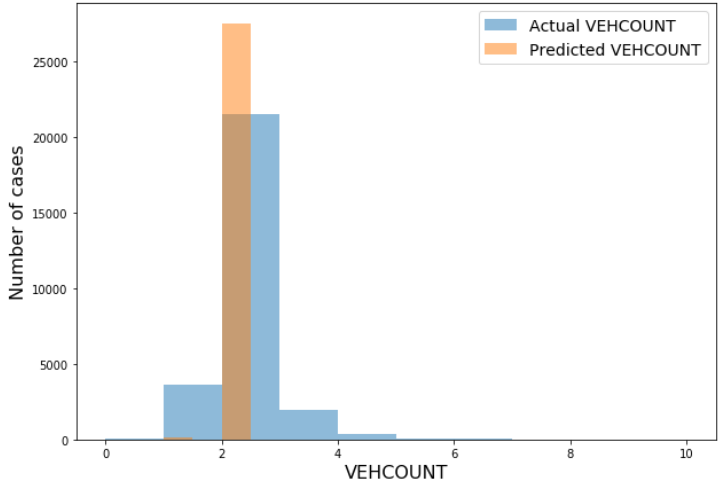


- Predictive Modeling

Decision Tree

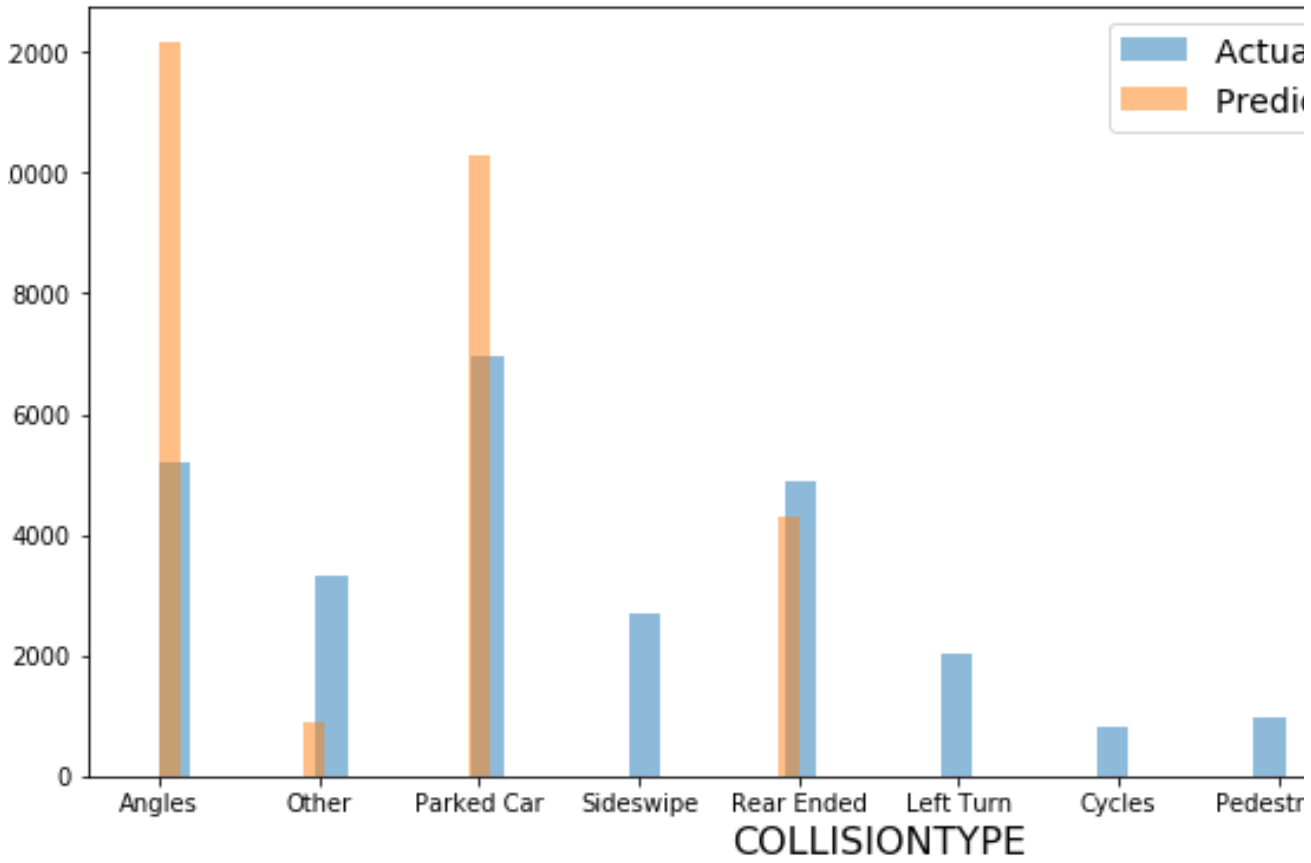
- Decision tree method generally poorly on predicting the severity level of the accidents. Features that are generally easier to predict are pedestrian count, bicycle count, and vehicle count





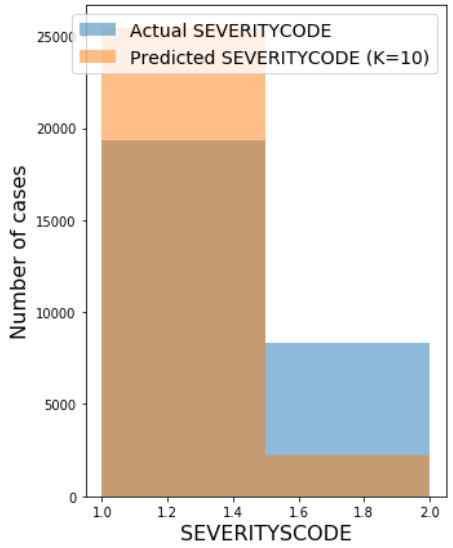
Logistic Regression

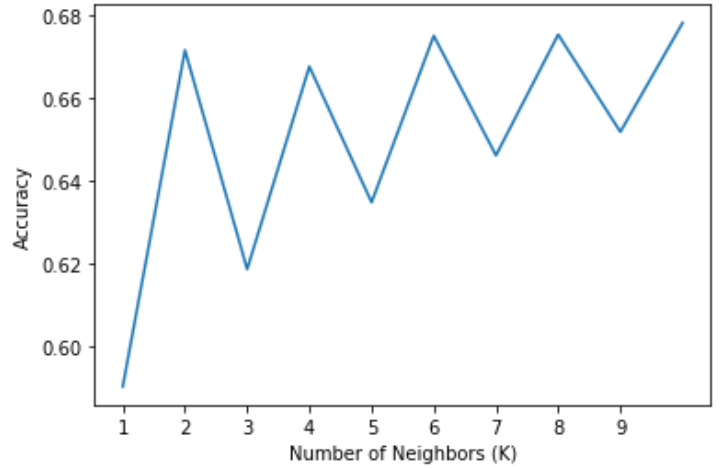
- the logistic regression can correctly predict some of the common collision types to some extent



K-Nearest Neighbor

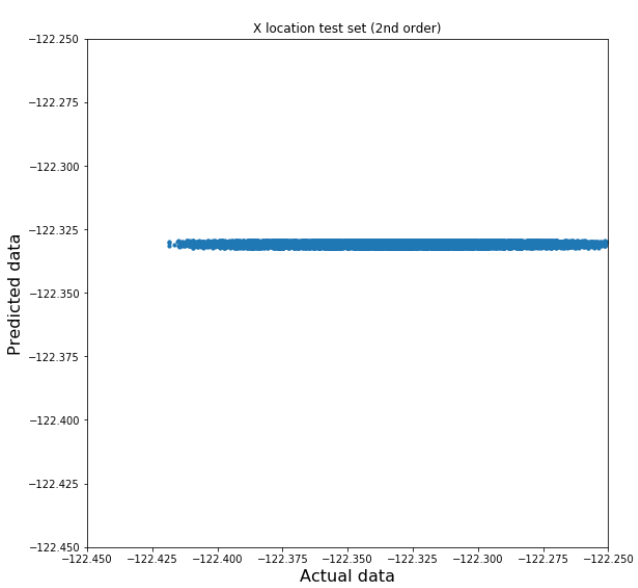
- As the previous two methods cannot predict an accurate severity level of the accidents, which is an important feature.

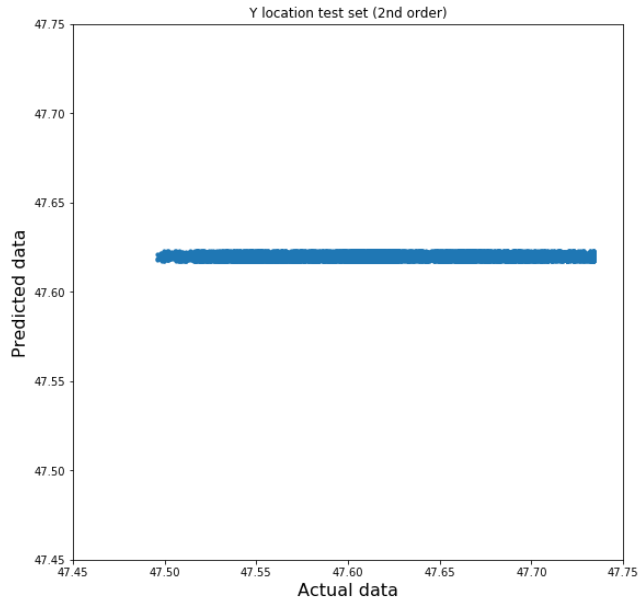




Polynomial Regression

- Polynomial regression was attempted to predict the X, Y locations of the accidents. Different orders of polynomials from 2 up to 20 are tested





- Conclusions

- In the present study, exploratory data analysis and predictive modeling were applied to study the data set representing car accidents in Seattle. In the exploratory analysis, it is found that over the years, there is declining trend of the number of accidents. A large number of accidents appear to happen one hour after midnight. Certain collision types dominate over others. Some collision types (e.g. rear-ended) are related to the hour of the day; some collisions also appear to have a weak correlation to the weather and road conditions. An overwhelming majority of accidents involve less than 5 people, 2 vehicles, 0 pedestrian and 0 bicycles.

- In the predictive modeling of the data, the severity level (represented by “SEVERITYCODE”), types of collisions, and persons count are generally hard to predict. The logistic regression model can only correctly predict some of the common collision types to some degree. The K-nearest neighbor method has an advantage on predicting the severity level over the decision tree and logistic regression methods. Location coordinates are not strongly correlated to other variables and are thus difficult to predict.