Predictive Analytics of Car Accidents in Seattle

Chao Z.

1. Introduction

1.1. Background

Car accidents are common in the United States. The severity of them can vary from bumper-tail light collisions to life-threatening situations. It is estimated that in the United States 37,000 people die due to car accidents each year and 2.35 million are injured or disabled [1]. In year 2019, in the state of Washington, there were 103,347 total car accidents including 497 fatal ones, and in the city of Seattle, there were 10,315 total accidents and 22 fatal ones [2].

In addition to the cost of lives and injuries to the people involved, car accidents can also bring large financial costs. The total financial cost associated with car accidents is estimated to be $230.6 billion each year in the U.S. [1]. According to another article [3], the average cost associated with fatal accidents is about $1.4 million per death; the average cost associated with a non-fatal disabling injury is $78,900; and the average cost associated with non-disabling injuries and property damage is $8,900.

The causes of car accidents can be attributed to several factors such as road condition, weather condition, and human errors (e.g. speeding). Knowing the effects of different factors on the car accidents is a key to predicting and preventing potential accidents in the future.

1.2. Problem Description

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.

1.3. Interests

As mentioned above, car accidents can result in bodily injuries and financial costs. Therefore, parties who may be interested in this study include: (1) individuals, who may use the study to reduce to risk of running into potential accidents, (2) insurance companies, who may use the study to understand the cause of car accidents, and (3) governments, who may use the study to improve local infrastructures and build safer roads.

2. Data

2.1. Description

The data set used in the present study is provided through the IBM Applied Data Science Capstone Course [4]. The raw data set contains car accident information in Seattle from 2004 to 2020. The raw data set contains 194,673 entries of accidents with 38 descriptive features (columns).

2.2. Preliminary Data Preprocessing

The data set has some duplicate entries. We first drop the duplicated rows. This reduced the data set to 189,542 rows.

Some of the entries also contain Nan values. For example, two of the features that we consider as independent variables, road condition and weather condition, contain 5012 and 5081 Nan values, respectively. We will remove the Nan values for the exploratory data analysis and predictive modeling. In the exploratory data analysis session, we will mostly be focused on analyzing subsets of selected features of the data. Thus, in this phase, we will drop Nan values in each of the subsets selected for the analysis. In the later predictive modeling phase where we apply a machine learning approach, we will drop all Nan values in the dataset that is used by the machine learning algorithm. Additional data pre-processing will be applied.

Some of the columns in the data set contain redundant information or information that is not very relevant for the purpose of the present analysis, such as, "INCKEY", "COLDETKEY", and "SEVERITYDESC," etc., and therefore, are dropped. Some other data columns are not very clear. For example, the "SPEEDING" feature contains only "Y" and "nan" values, and it's not clear how many of the "nan" entries are actually speeding violations. Therefore, such columns are also dropped. In addition, the date and time information are converted into three additional feature parameters: "year", "dayofweek", and "hour". A full description of the feature parameters used in the present study are listed in Table 1. All other features of the original data set [4] that are not listed in Table 1 are dropped.

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

3. Exploratory Data Analysis

3.1. Severity of Accidents

The severity of the accident is analyzed. In the given data set, only two severity labels are present: 1 – property damage, which accounts for more than 2/3 of the total number accident cases, and 2 – injuries. A histogram plot showing the comparison of these two severity levels is shown in Fig. 1.

The case counts categorized by different severity levels are plotted over the years. It can be seen from Fig.2 that overall there is a decreasing trend in both the property damage type accidents and the injury type accidents. The number of property damage accidents in each year always remains no less than twice the number of injuries. Comparing year 2019 to 2005, the number of property damage dropped nearly half, and the number of injuries dropped about 30%.

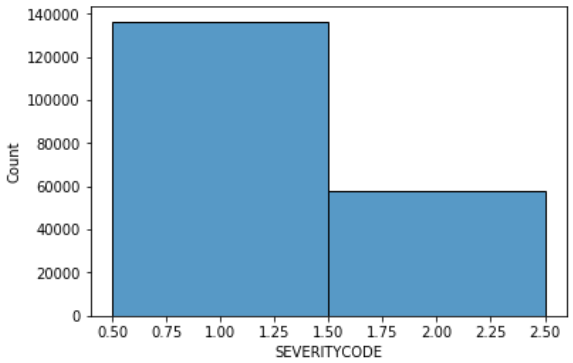


Fig. 1. Case count of severity levels (1 – property damage, 2 – injuries)

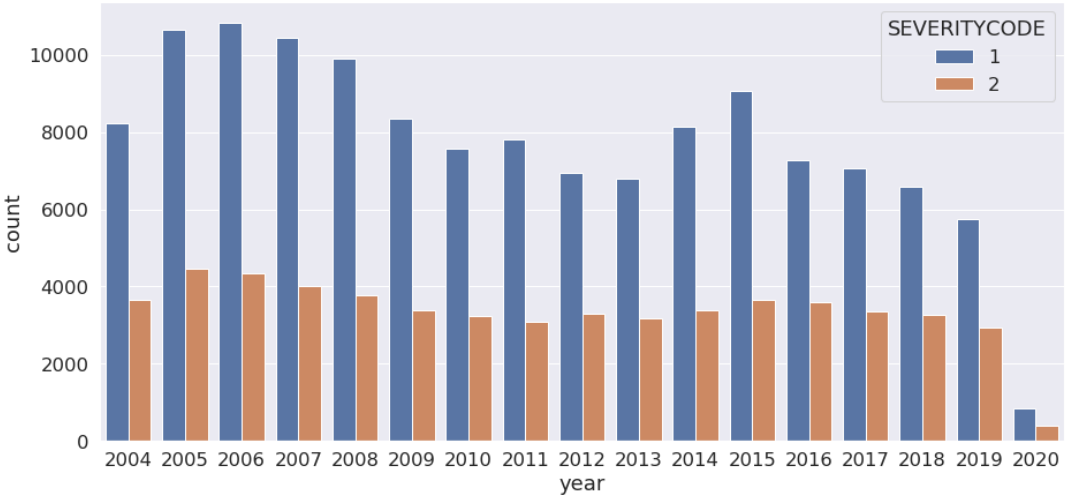


Fig. 2. Counts of severity levels over the years (1 – property damage, 2 – injuries)

The counts of severity labels are also plotted against day of week and hour of the day. Some trends are recognized. Overall, as shown in Fig. 3., the distribution of the accidents over a week is pretty flattened out; however, an interesting trend shows that through Monday to Friday the number of accidents slightly ramps up and drops through the weekend. Observing the case count over hours of the day reveals one important discovery as shown in Fig. 4. 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.

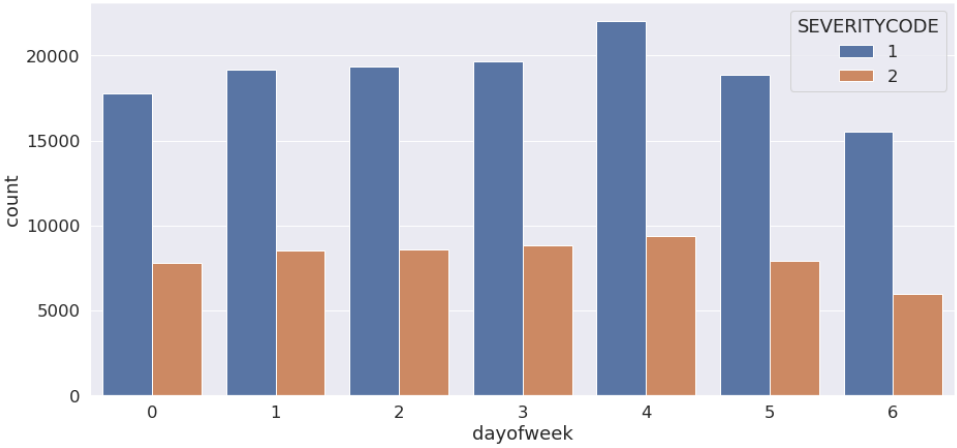


Fig.3. Counts of severity levels over a week (Monday starts with 0 on the x-axis)

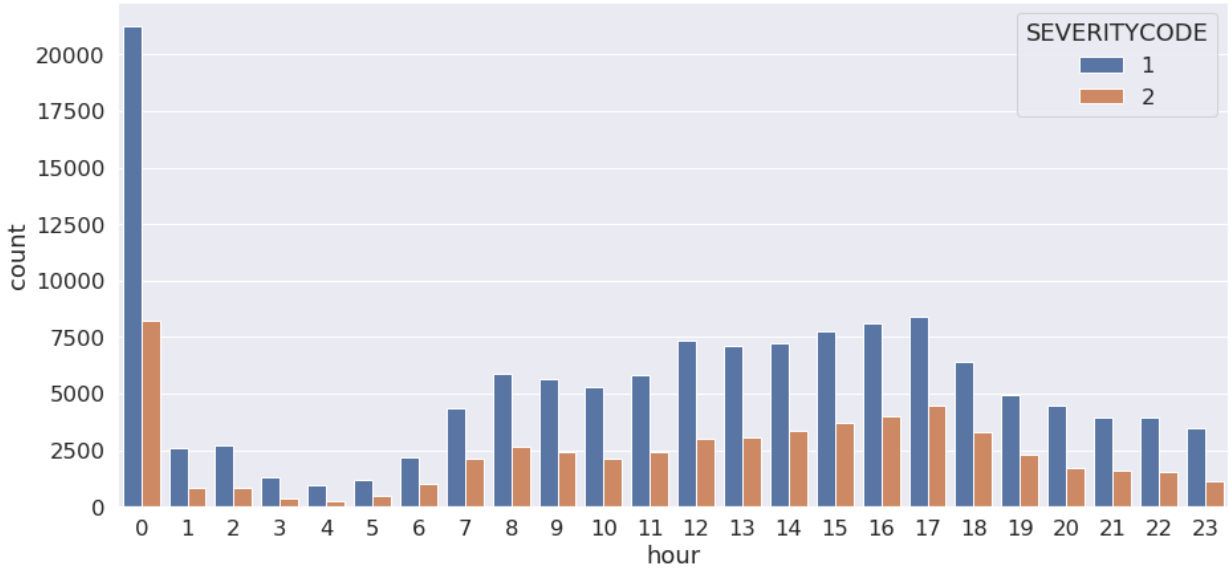


Fig.4. Counts of severity levels over a day (mid-night starts with 0 on the x-axis)

3.2. Collision Types

The counts of different collision types are also compared against years, day of week, and hour of day. From the figures showing the data vs. years (Fig.5) and the data vs the day of week (Fig.6), it can be observed that the top three collision types are always: “parked car,” “angles,” and “rear ended,” in every year, and throughout every day of the week. During most hours of the day, as shown by Fig.7,

these three types of collisions are also more dominant, with one small exception, that is in late night, from 22:00 to midnight, it appears the number of “rear ended” type of collisions decreases slightly. In addition, 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.

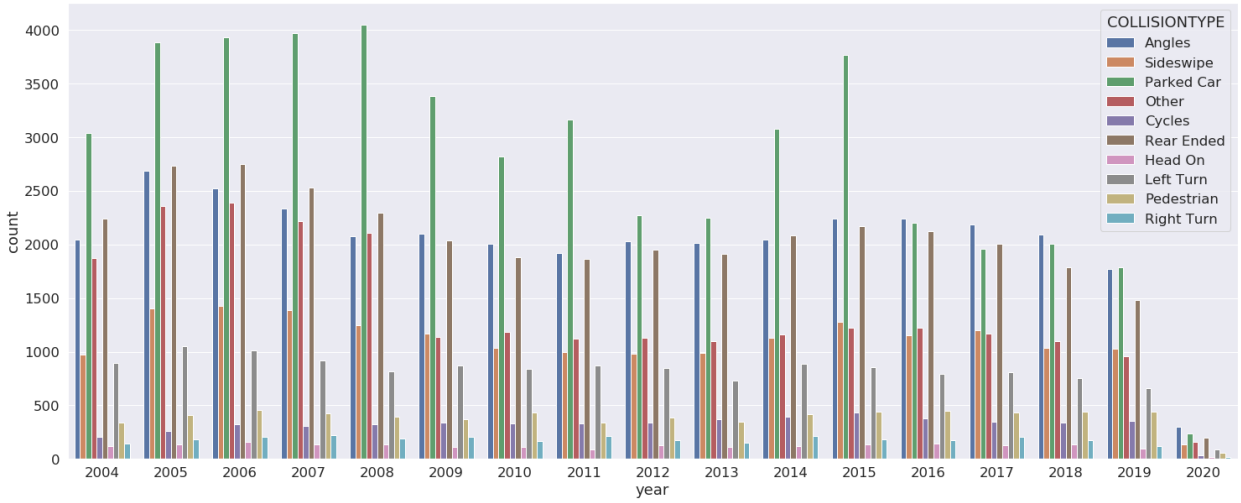


Fig.5. Counts of different collision types over the years

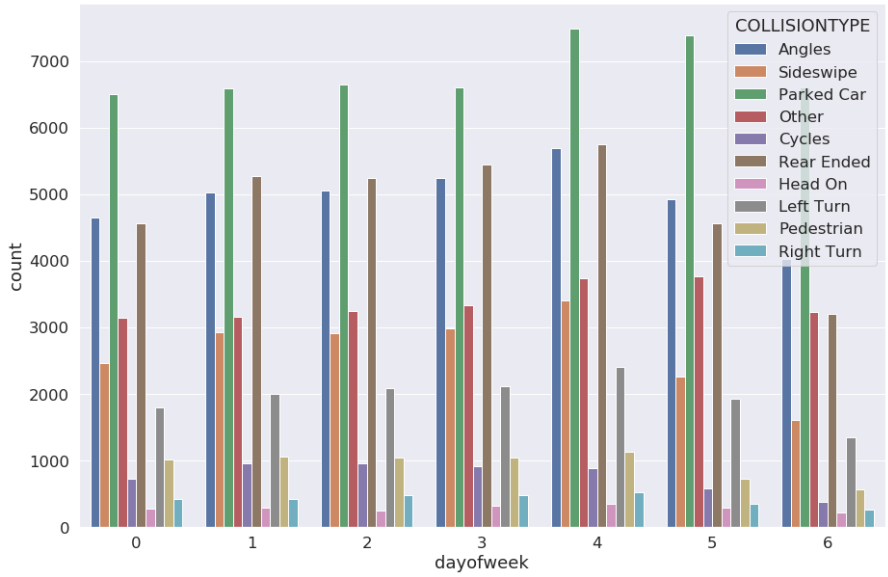


Fig.6. Counts of different collision types over the day of week (Monday starts with dayofweek = 0)

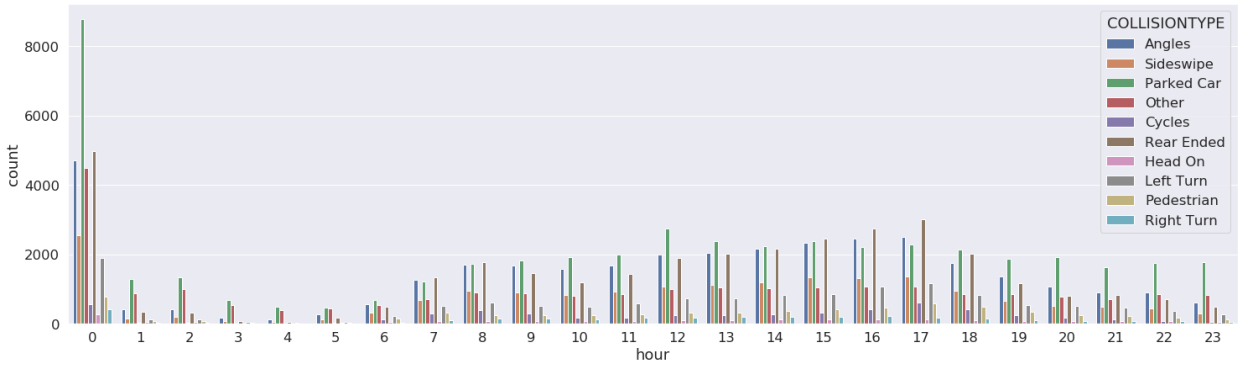


Fig.7. Counts of different collision types throughout the day (midnight starts with hour =0)

3.3. Weather

The influence of weather condition is also analyzed. First, the accidents counts in different severity categories (1 - property damage and 2 – injuries) are grouped based on the weather conditions. From Fig.8, it can be seen that most car accidents happen in three weather conditions: overcast, raining, and clear sky. Also it appears that in general there is no strong relation between the level of severity and the weather condition, as the ratio of severity level 1 vs. level 2 remains similar cross different weather condition categories. One exception is for the “Partly Cloudy” weather condition, under which there appears to be more injuries cases than property damage accidents. However, 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.

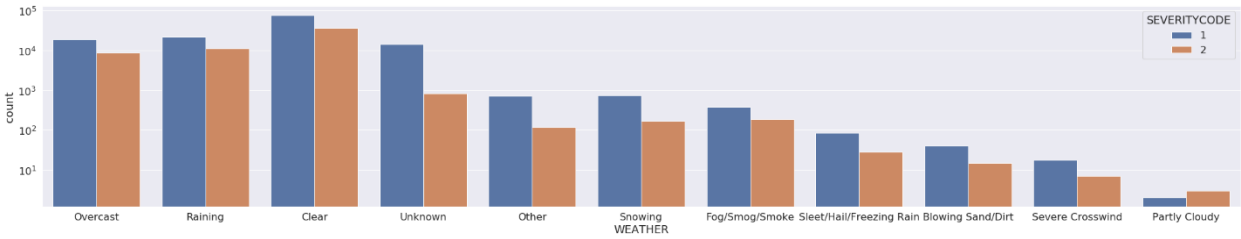


Fig.8. Severity label counts vs. weather conditions (Severity 1 = property damage; severity 2 = injuries. Note: y-axis is in log scale)

Secondly, the influence of weather condition on different collision types is analyzed. As can be seen in Fig.9., overall, the most common types of collisions cross all weather condition categories appear to be: “Angles,” “Parked Car,” “Rear Ended,” and “Other.” The percentage of “Cycles” drops significantly in some of the bad weather conditions such as snowing, fog, sleet, etc. This is most likely due to a reduced number of cyclers on the road in these weathers. Another finding that seems interesting is that the percentage of the most dangerous collision type, “Head On” collisions, also appear to drop significantly

in some of the extreme weather conditions such as “Fog/Smog/Smoke,” “Sleet/Hall/Freezing Rain,” “Blowing Sand/Dirt,” “Severe Crosswind.”

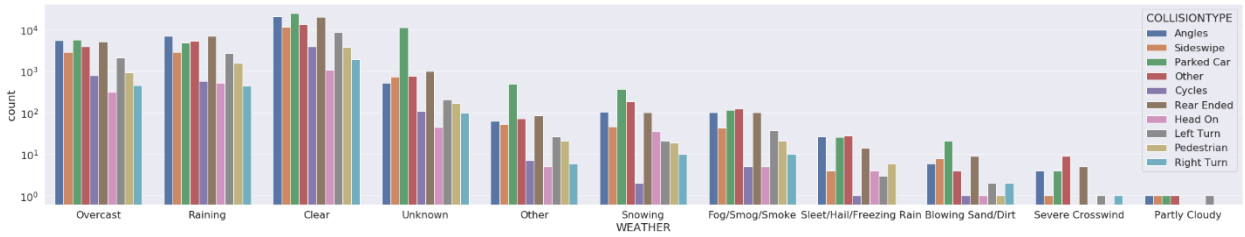


Fig.9. Accident type counts vs. weather conditions (note: y-axis is in log scale)

3.4. Road and Lighting Condition

The road condition is also analyzed. Figure 10 shows that (1) the wet and dry road conditions correspond to an overwhelmingly large percentage of accident counts compared to the other road condition categories, and (2) in all road condition categories the ratio between property damage case number and the injury case number is comparable to the ratio of total case counts of these two types of severity labels. The road condition does not appear to have a direct impact on the severity of an accident.

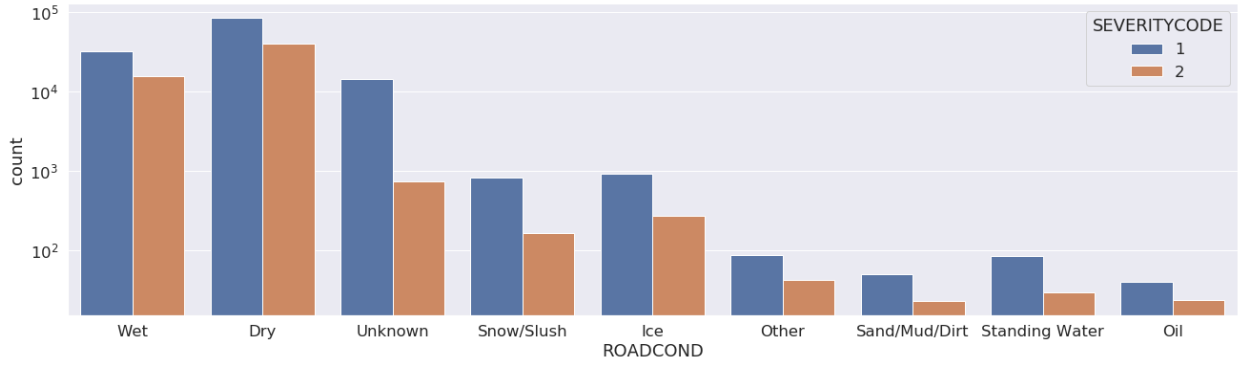


Fig. 10. Counts of severity levels vs. road conditions (note: y-axis is in log scale)

The percentage of collision types appears to be somewhat affected by which category of road condition they are in, as shown by Figs. 11. The proportions of right turn collisions are significantly reduced in the following road conditions: “Ice,” “Sand/Mud/Dirt,” “Standing Water,” and “Oil”. In addition, the numbers of rear ended cases is much greater than the number of collisions with parked cars when the road has standing oil or water, whereas under other road conditions the number of collisions with parked cars is usually comparable or even greater than the number of rear ended accidents.

Lighting condition of the road also plays a small, if not little, effect on the type of collisions. As shown in Fig. 12, when the lighting condition is “Dark – Unknown Lighting,” the sideswipe, left/right turn collisions are reduced to almost none. However, in other lighting condition categories, the ratio of different collision types remains about the same.

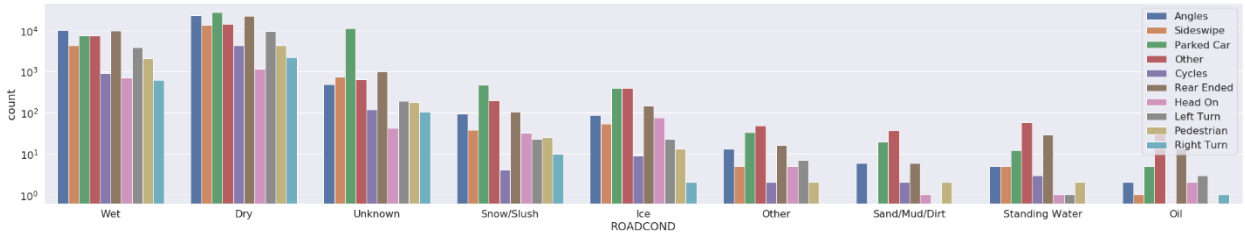


Fig. 11. Counts of collision types vs. road conditions (note: y-axis is in log scale)

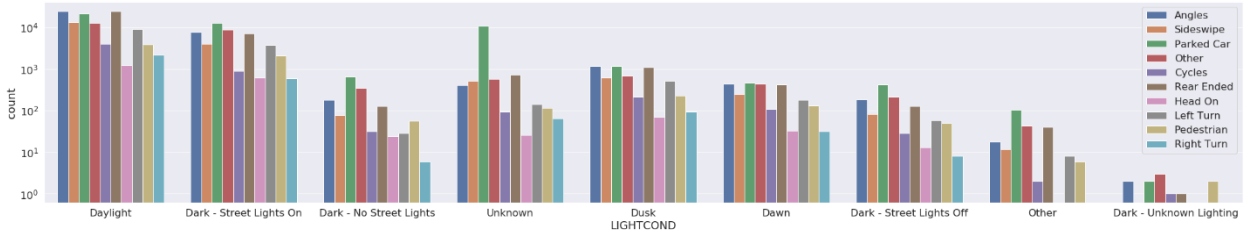


Fig. 12. Counts of collision types vs. road conditions (note: y-axis is in log scale)

3.5. Under Influence

Whether the person involved is under influence is also an interesting factor to analyze. From Fig. 13, a histogram plot of accident counts of whether or not under influence grouped by accident severity level, one can observe that whether the person involved is under influence or not does not have a strong influence on whether the collision is property damage type or injury type.

However, it appears 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.”

3.6. 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.

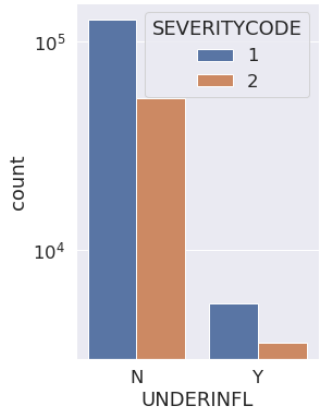


Fig. 13. Counts of under influence accidents grouped by severity level

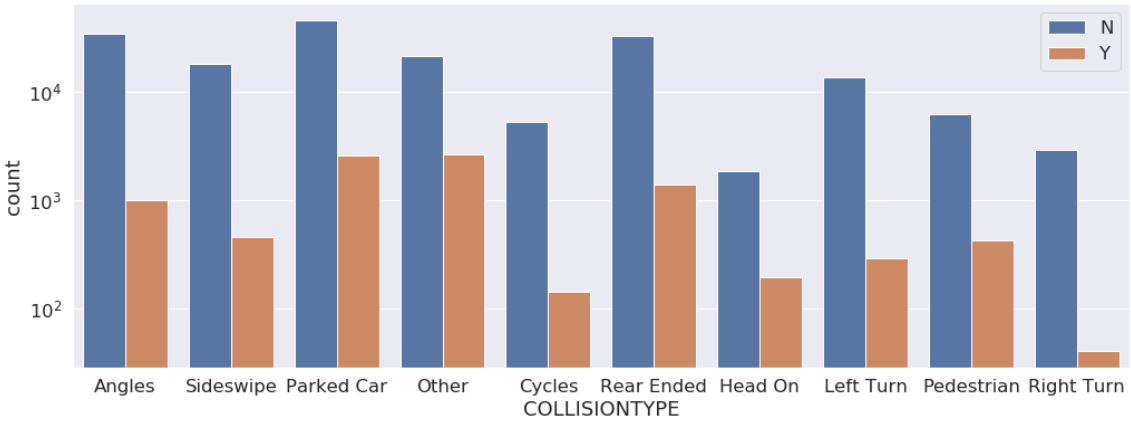


Fig.14. Counts of under influence accidents in different collision type categories

Lastly, the geographic features regarding the counts of persons and vehicles are studied. Overall, their distributions appear pretty scattered. No clear trend has been observed between the number of persons/vehicles and the geographic locations, as shown in Figs. 17.

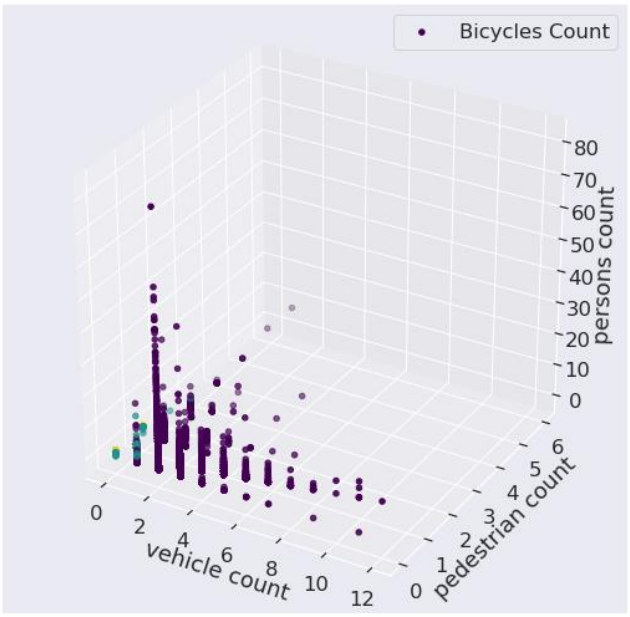


Fig. 15. Plot of Counts of Persons, Vehicles, Pedestrian, and Bicycles

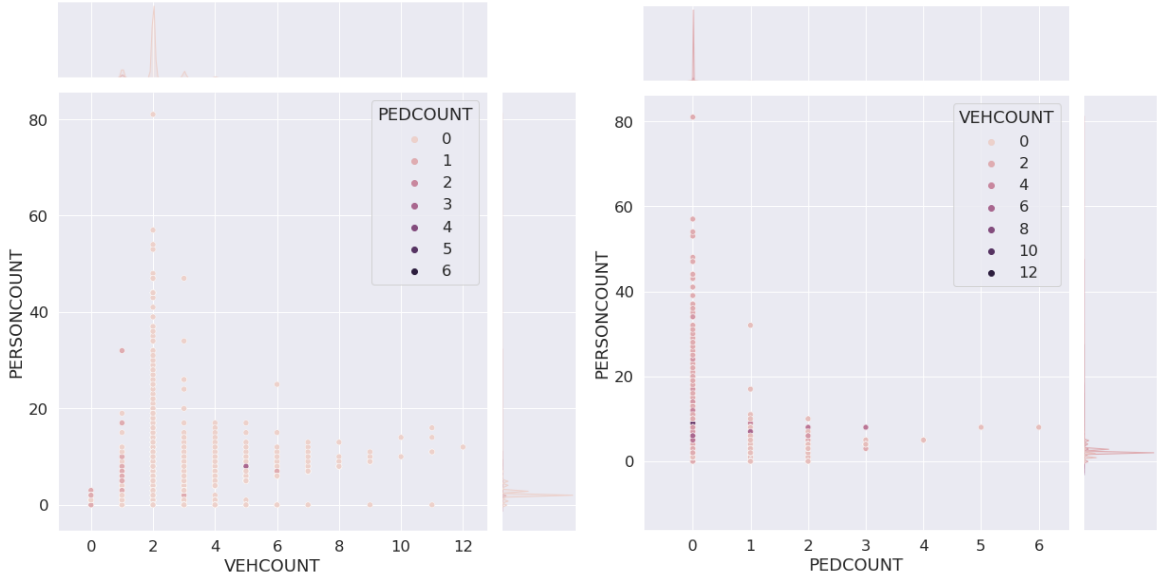
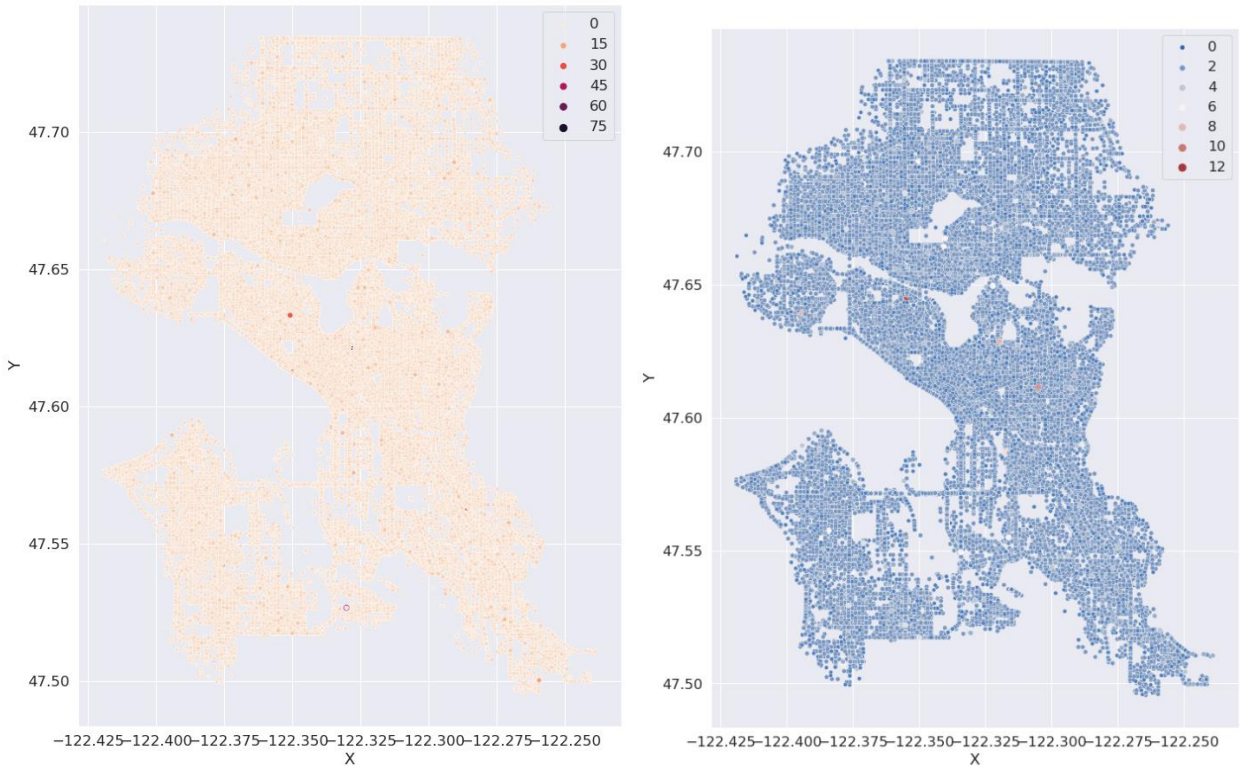


Fig. 16. Persons count vs. vehicle count, and vs. pedestrian count



(a) Persons count (b) Vehicle count

Fig. 17 Geographic features regarding the counts of persons and vehicles

4. Predictive Modeling

Predictive models are built on the data set using Sklearn library. From the exploratory analysis in the previous session, the following features are selected as independent variables in the predictive modeling: hour, day of week, weather, road condition, lighting condition, under influence. First, classification methods (decision tree, logistic regression, and K-nearest neighbor) are applied to model the relation between these independent variables and other parameters that are important characteristics of the accidents, which include: severity level, collision type, persons count, vehicle count, pedestrian count, and bicycle count. Secondly, we also attempted to predict the location of the accidents based on independent variables using polynomial regressions, but found that location turned out to be difficult to predict mainly because the relation to the independent variables is weak.

In this session the selected data set (corresponding to the aforementioned independent variables and accident features) is further pre-processed to get rid of any nan values. The “Other” and “Unknown” labels in some of the features (e.g. road condition) are merged. Further, one hot encoding is applied to the WEATHER, ROADCOND, LIGHTCOND features. The data set is then normalized using the standard scalar method before feeding into different predictive models. The data is split between testing set and training set using of testing set size of 0.15 (percentage).

4.1. 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. Below are the accuracy scores of different predicted features using the decision tree method:

```
training set accuracy of SEVERITYCODE =  0.6963596636937467
training set accuracy of PERSONCOUNT =  0.5877053998108916
training set accuracy of VEHCOUNT =  0.782340088420945
training set accuracy of PEDCOUNT =  0.9627213718024072
training set accuracy of PEDCYLCOUNT =  0.9706882011704275
training set accuracy of COLLISIONTYPE =  0.3315785438654775

testing set accuracy of SEVERITYCODE =  0.6981029614075737
testing set accuracy of PERSONCOUNT =  0.5846064731011512
testing set accuracy of VEHCOUNT =  0.7770255593367605
testing set accuracy of PEDCOUNT =  0.9630729128955181
testing set accuracy of PEDCYLCOUNT =  0.9699152849178191
testing set accuracy of COLLISIONTYPE =  0.2847368039968141
```

As can be seen in Fig. 18, the decision tree method almost does not predict any injury-type accidents. For the prediction of vehicle count and persons count features, the decision tree method also predicts poorly, as shown by Figs. 19 – 20.

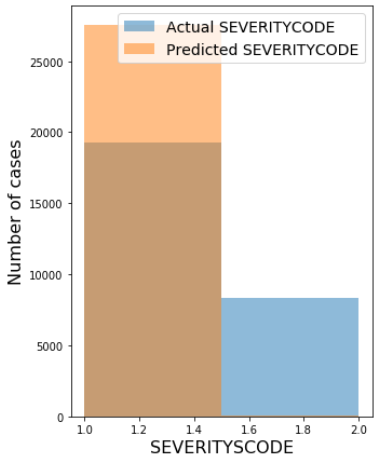


Fig. 18. Decision tree prediction on the severity level

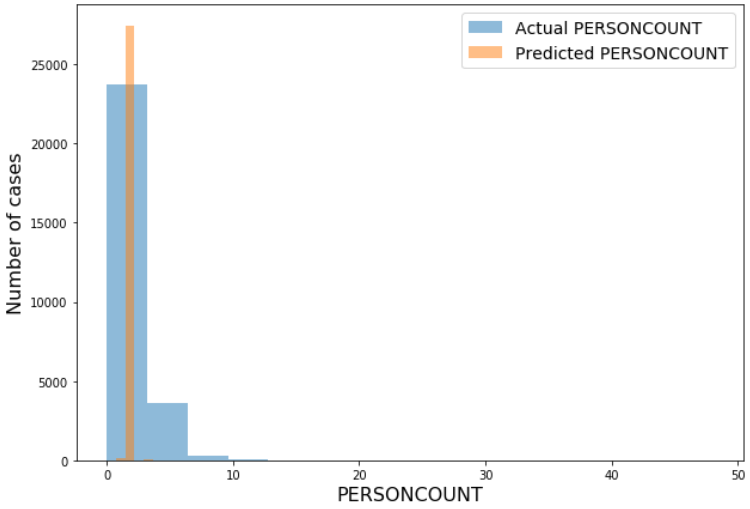


Fig. 19. Decision tree prediction of persons count

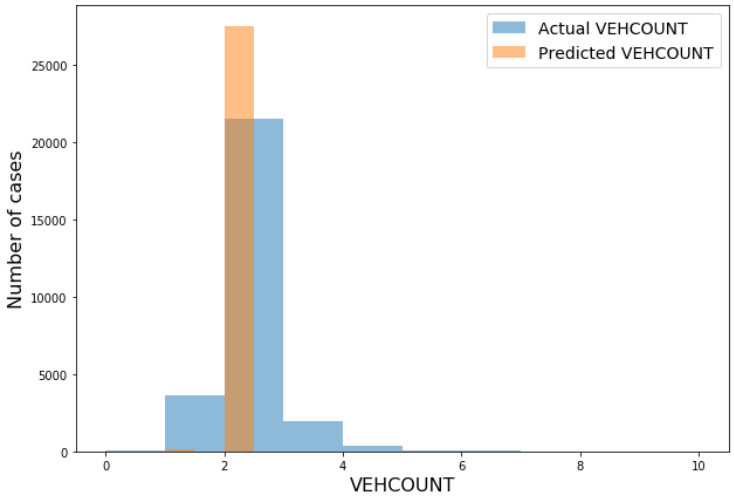


Fig. 20. Decision tree prediction of vehicle count

4.2. Logistic Regression

The logistic regression method is applied to study the data. Again, the collision type prediction capability is poor, although the logistic regression can correctly predict some of the common collision types to some extent (such as rear ended, parked car, and angles), as shown in Fig. 21. The accuracy scores for various features are summarized below:

```
training set accuracy of SEVERITYCODE =  0.6953502338299558
testing set accuracy of SEVERITYCODE =  0.6989718340453261
training set accuracy of COLLISIONTYPE =  0.28534921162249877
testing set accuracy of COLLISIONTYPE =  0.28600390992686986
training set accuracy of PERSONCOUNT =  0.5860954230661113
testing set accuracy of PERSONCOUNT =  0.5860907971906452
```

```
training set accuracy of VEHCOUNT =  0.7806406685236769
testing set accuracy of VEHCOUNT =  0.7785098834262545
training set accuracy of PEDCOUNT =  0.962644706243132
testing set accuracy of PEDCOUNT =  0.9631815219752371
training set accuracy of PEDCYLCOUNT =  0.9706818123738212
testing set accuracy of PEDCYLCOUNT =  0.9699152849178191
```

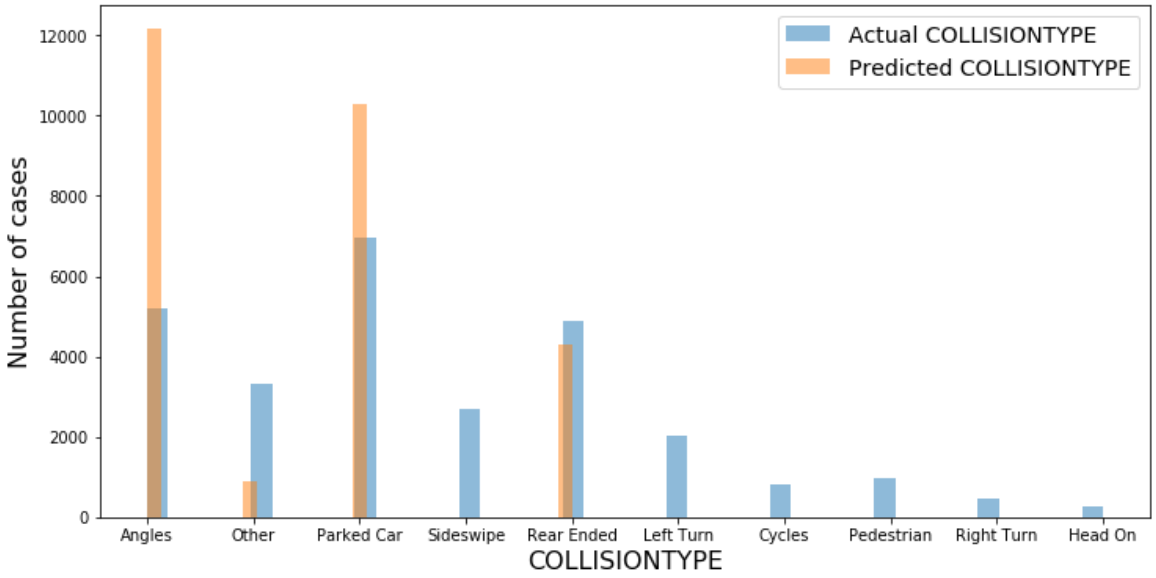


Fig. 21. Logistic regression method on the collision type prediction

4.3. K-Nearest Neighbor

As the previous two methods cannot predict an accurate severity level of the accidents, which is an important feature. The K-nearest neighbor method is applied to predict the severity of accidents. Different values of k are tried out to explore the best value to be used. The result of the predicted accuracy of SEVERITYCODE vs. number of neighbors (k) is shown in Fig. 22. The best value of k is 10, which results an accuracy of the 0.677964 on the predicted value of SEVERITYCODE. The histogram plot of the predicted SEVERITOCODE vs. the actual one is shown in Fig. 23, which shows great improvement over the previous two decision tree and logistic regression methods.

4.4. 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. The predicting capability, however, is generally poor. A few polynomial regression results are selected and shown in Fig. 24. Overall, this is also inline with the poor correlation between the features and the location data, as previously seen in Fig. 17.

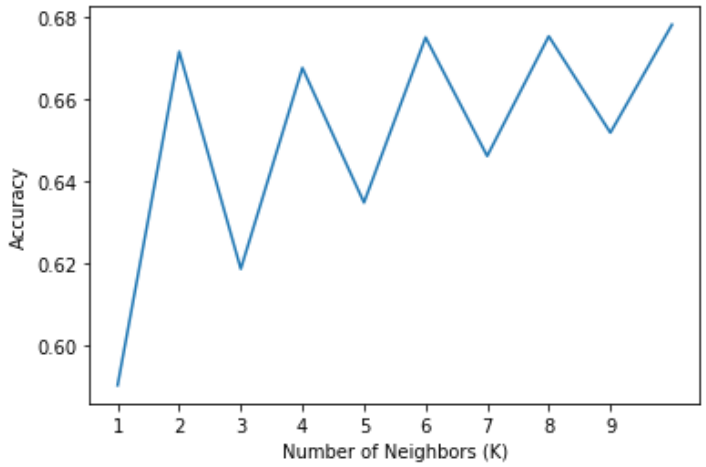


Fig. 22. Prediction accuracy on the SEVERITYCODDE over the number of neighbors

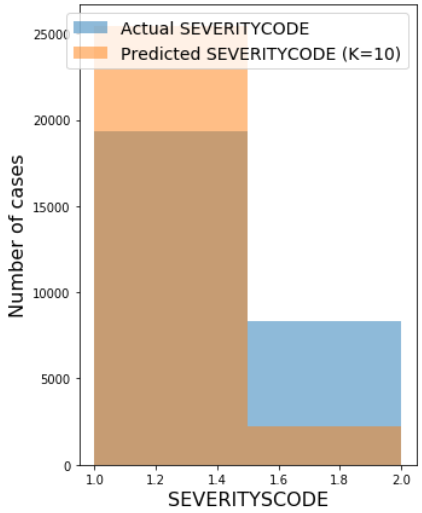


Fig.23. Predicted SEVERITYCODE vs actual one using K-nearest neighbor method

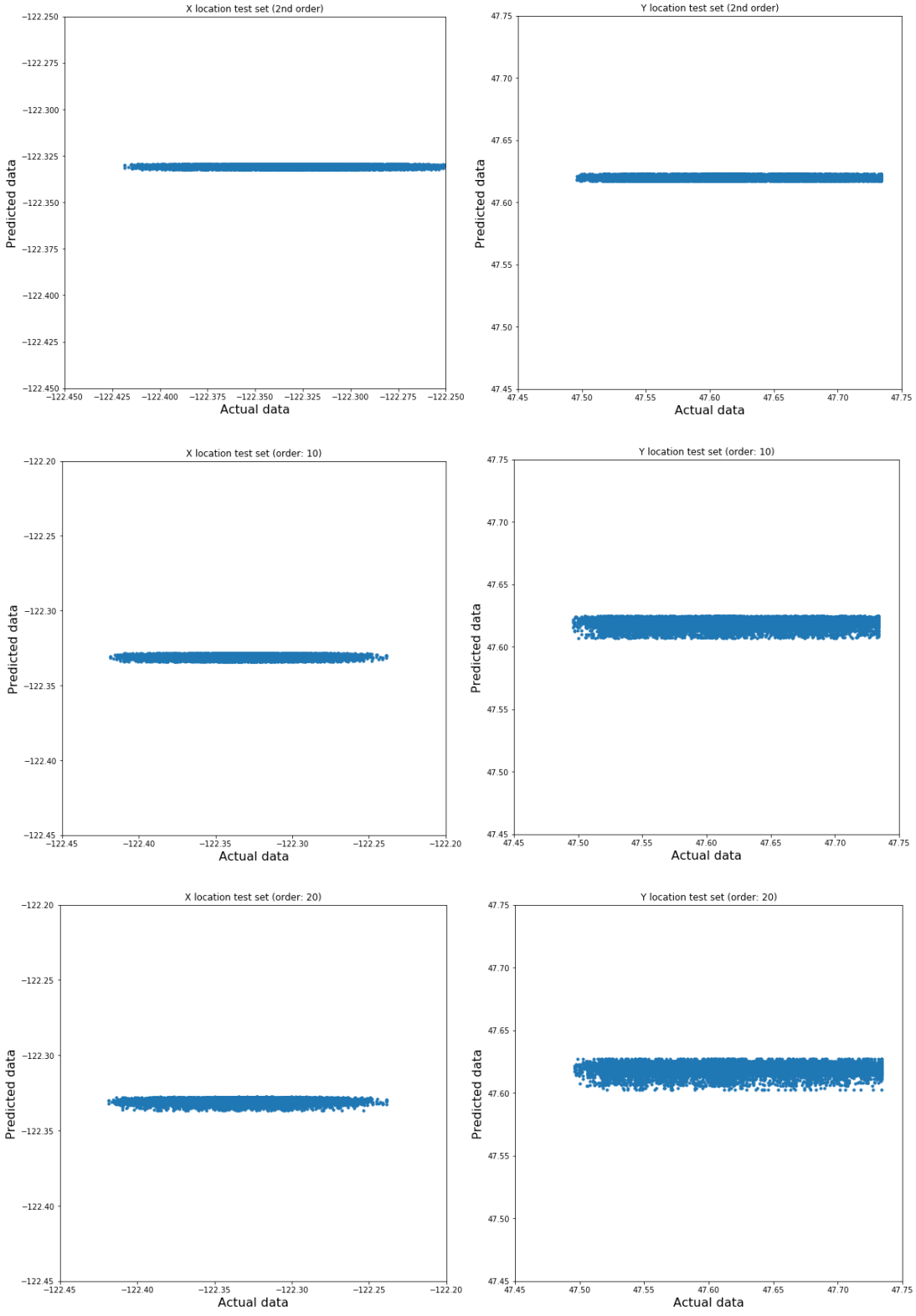


Fig. 24. Attempt using different orders of polynomial regression to predict X, Y location

5. 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.

Reference

[1] Shubhankar Rawat, "USA Accidents Data Analysis," *Medium.com*, Feb. 21st, 2020 https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02

[2] "2019 Washington State Car Accident Statistics & Reports," https://www.colburnlaw.com/seattle-traffic-accidents/

[3] "What Is the Average Cost of a Car Accident," Aug. 26th, 2020, https://www.theintelligentdriver.com/2020/08/26/what-is-the-average-cost-of-a-car-accident/

[4] "Collision Data in Seattle" *IBM Applied Data Science Capstone on Coursera*, https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv