Capstone Project
The Prediction of severe accidents with Collision Data
(Case Study: Seattle)

Table of Content

Natascha Hey
09/16/2020

Introduction: Business Problem

Objective

The project's objective is to examine the probability getting into a car accident and to predict the severity of this accident.
Especially car insurance providers and car manufacturer can profit from this application. The manufacturer would install this application in the vehicle which warns drivers about certain conditions and hazardous areas. Therefore, the driver can reconsider his plans to take a specific route, or drive at all. The application can also provide support for the police and ambulance. If severe accidents are likely to happen on specific days, certain areas or junction types, those public institutions can plan accordingly.

How can machine learning help?

Machine learning helps to assess the highest risk factors of a severe accident. Are natural circumstances (e.g. bad weather and road conditions) responsible for more severe accidents, or is it the inattention of a driver or him being under influence of drugs a greater danger?
Do severe accidents happen more often at certain junctions, or even on highways? And are the accidents more severe when pedestrians are involved? Does the severity of an accident decrease with the number of people in the car, since the driver feels a greater responsibility for all occupants?
Since we have a binary problem (Is the accident severe or not?) we can use KNN, SVM, logistic regression or a Decision Tree algorithm to predict the outcome.
(see Methodology)

Data

The following dataset provides all collisions published by the Seattle Police Department from 2004 to present:

Explanation to Data

Data

Based on the definition of our problem, the following attributes will be used for data analysis:

- Locations or hazardous areas throughout Seattle
- Weather/Road/Light Conditions
- Is the accident due to Human failure? (e.g. influence of Alcohol, inattention)
- The number of participants involved in the accident (person and vehicle count)
- Are pedestrians or cyclists involved
- Time of the Accident (Month and daytime)

Methodology

Data preparation

It is our goal to find attributes which help us determine the probability of an accident to be severe or not.

First, we extract the data from a csv file and store it into a data frames using pandas and numpy. Then we analyze the size of the data and examine each column. Here we want to achieve that all data can be used by machine-learning algorithms, which can only process numerical data. Hence, we have to check the type of data in each column.

For instance, in column 'UNDERINFL', we find the unique data entries ‘N’,’Y’,’0’,’1’. One way to deal with it is to map all strings to integer: ‘N’:0 and ‘Y”:1. After transforming all columns to the float or string, one can continue to group those columns.
If the data is categorial, the first step is to create dummies of those categories. For instance, ‘WEATHER’ has 11 different variables like ‘rain’ and ’clear’ which should be converted to dummies.

Data visualization

After preprocessing the data, bar plots can help to find trends and patterns. With matplotlib.pyplot bar graphs can be easily created. For better comparison, we can use subplots to show similar datasets next to each other.
To visualize the results of KNN I used the mlxtend.plotting library.
An easy way to plot the traffic data of accidents on the map of Seattle is to work with ‘folium’. It creates an interactive map and can add markers to accidents.

Classification

Since we have a binary problem (Is the accident severe or not?) we can use KNN, SVM, logistic regression or a Decision Tree algorithm to predict the outcome.
With F1-Score, Jaccard index or LogLoss we can evaluate the accuracy of our model and can adapt it, by either collecting more data from other sources, or choosing different columns to train our model.
The Confusion Matrix can provide a visual review of the performance of each algorithm. If way more Severe accidents happened but they were predicted as not severe, one can check those data points and find patterns in those data. Then rearrange the dataset (include or dismiss columns) and optimize the outcome for the test set.

Data Analysis

The Data consists out of 194673 samples and 38 columns

Columns :

['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC','SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']

Groups: Locations or hazardous areas

Conditions caused by Nature

Human Failure

Count accident participants

Time of the Accident

Unnecessary columns (to be dropped)

Target column

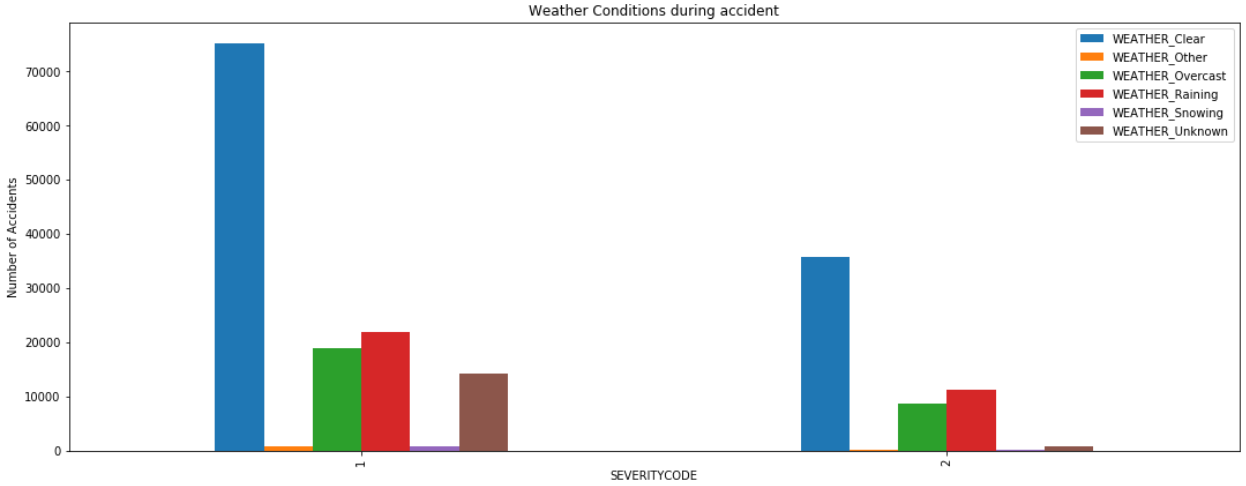
Dataset is not balanced

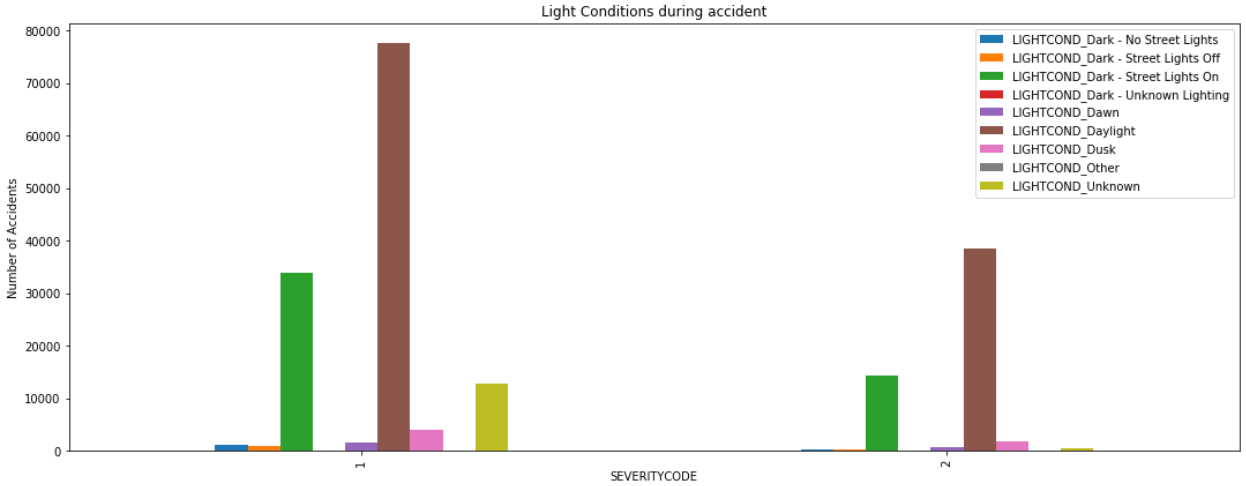
1 136485 (not severe accidents)

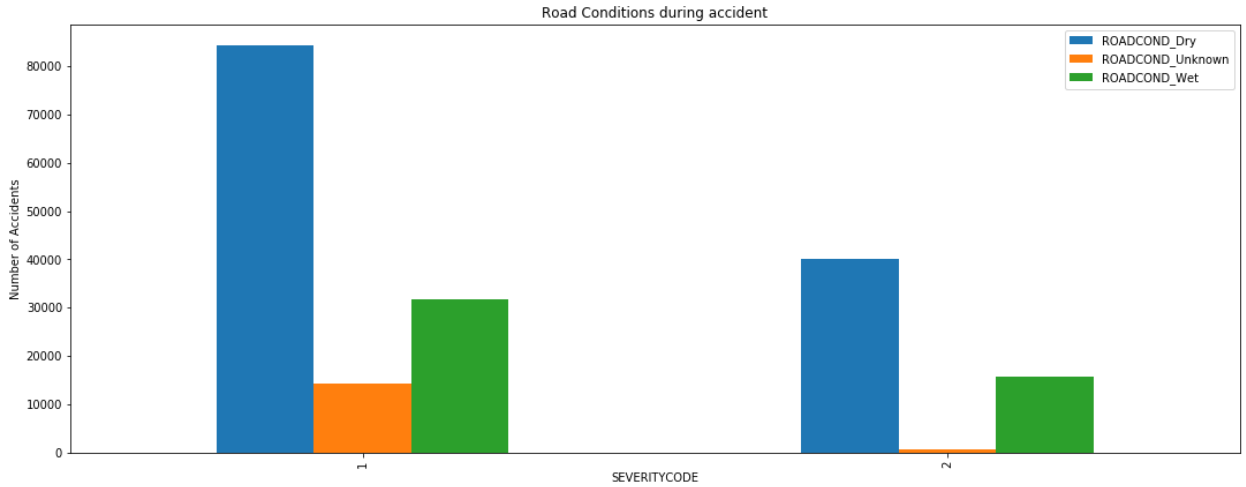
2 58188 (severe accidents)

Accidents caused by bad weather conditions

'WEATHER', 'ROADCOND', 'LIGHTCOND'





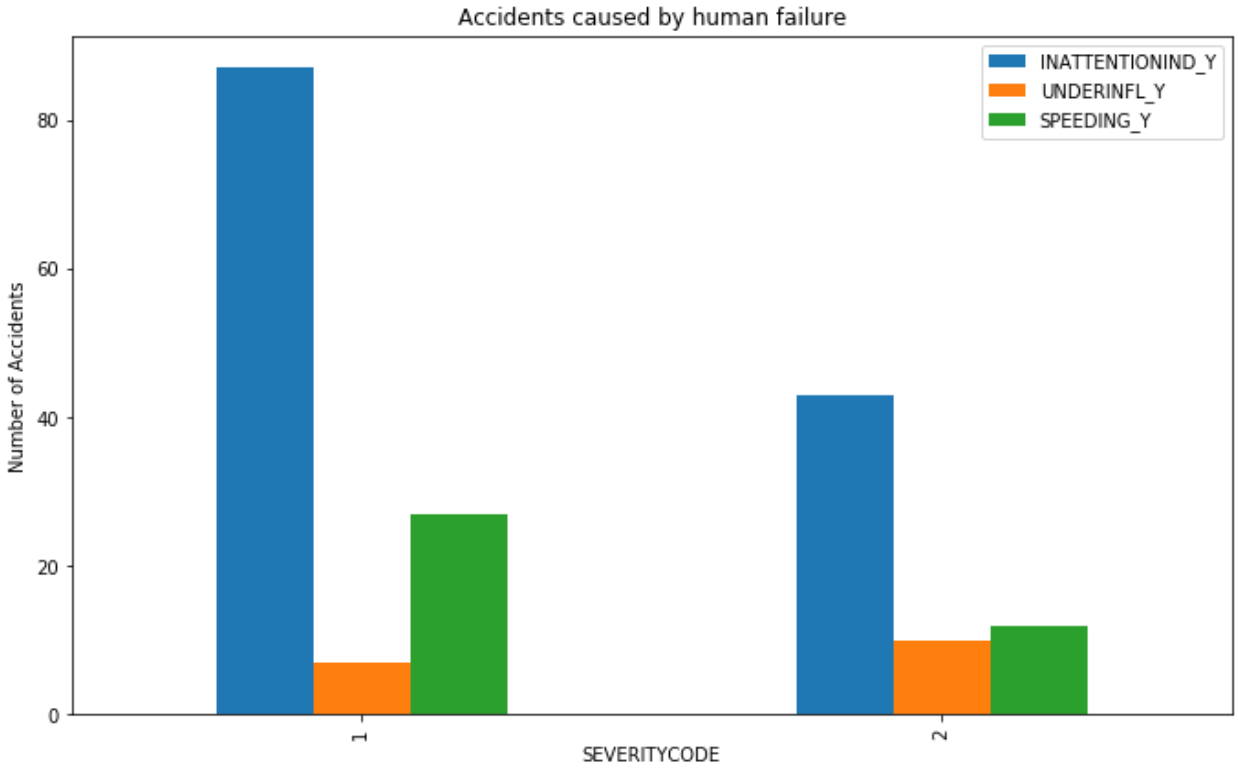


Weather, light and road conditions are critical for driving safe. Therefore, those three attributes are an obvious first choice for analyzing the frequency of severe and not sever accidents.
The ratio of severe accidents increases especially for the weather conditions ‘rainy’ and ‘overcast’, while the values of ‘unknown’ weather conditions almost vanishes. Hence, these attributes will be important to include in the machine learning algorithm.
Unknown light conditions also vanish almost completely when it comes to severe accidents, whereas the ration of such increases at night with streetlamps on and off.
In regard to the road conditions, severe accidents have a larger ratio on wet roads than on dry roads. Unknown road conditions also decrease drastically.
Those trends will be helpful to optimize the classification algorithm.

Accidents caused by human failure

'INATTENTIONIND', 'UNDERINFL', ‘SPEEDING’

First note, that the ‘UNDERINFL’ data column had to be preprocessed due to an irregularity in data. There were integers of 0 and 1 but also string of Y and N. For the data to be accurately evaluated, the strings had to be mapped by {‘Y’:0,’N’:1}.

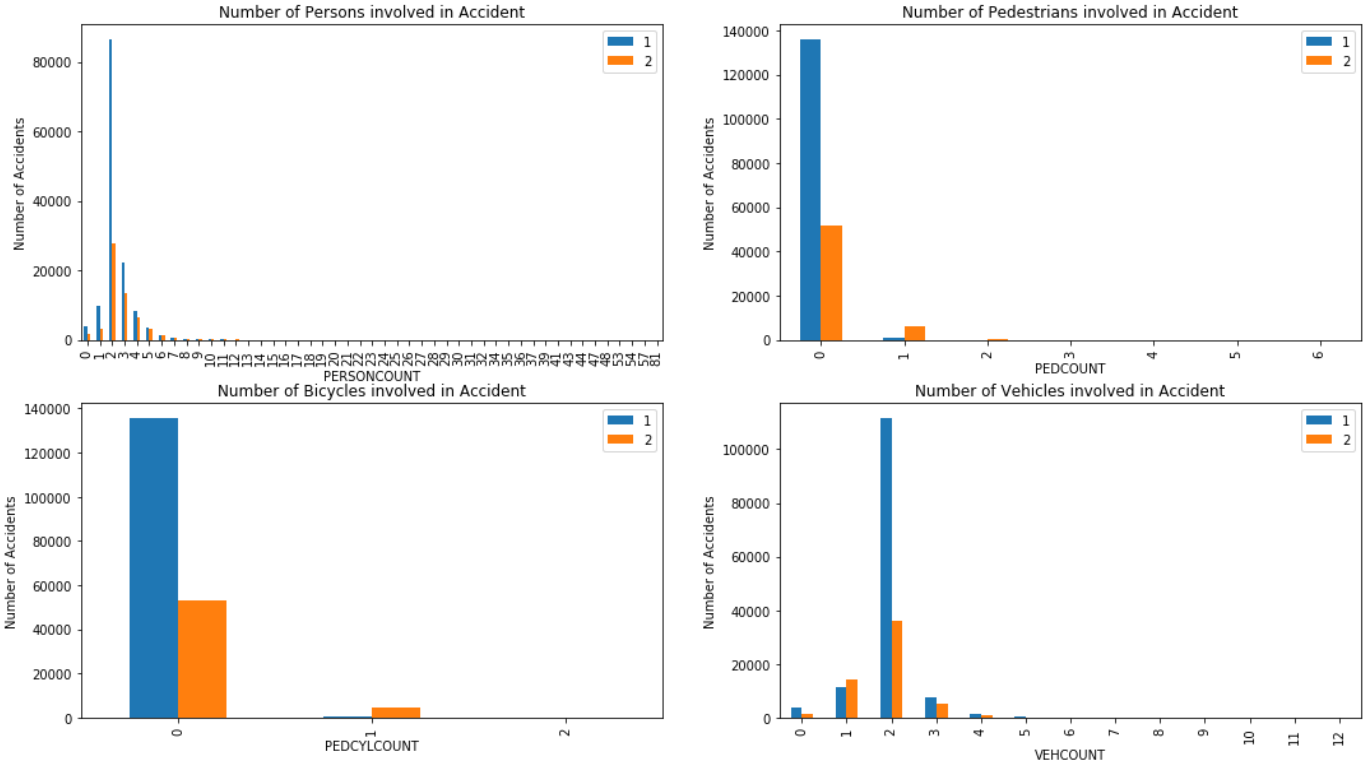


There is a reason why driving and being under influence is forbidden. The same is valid for writing messages on the phone or being distracted by other things. Thus, accidents caused by inattention, being under influence or speeding can lead to a good approximation of when an accident is severe or not.

On a randomly chosen sample of the data, the plot shows clearly that the ratio of severe accidents for speeding and driving under influence increases drastically. Hence, those three columns are good indicators for the severity of an accidents and can be used to add to the clustering data.

Count of Participants in Accident

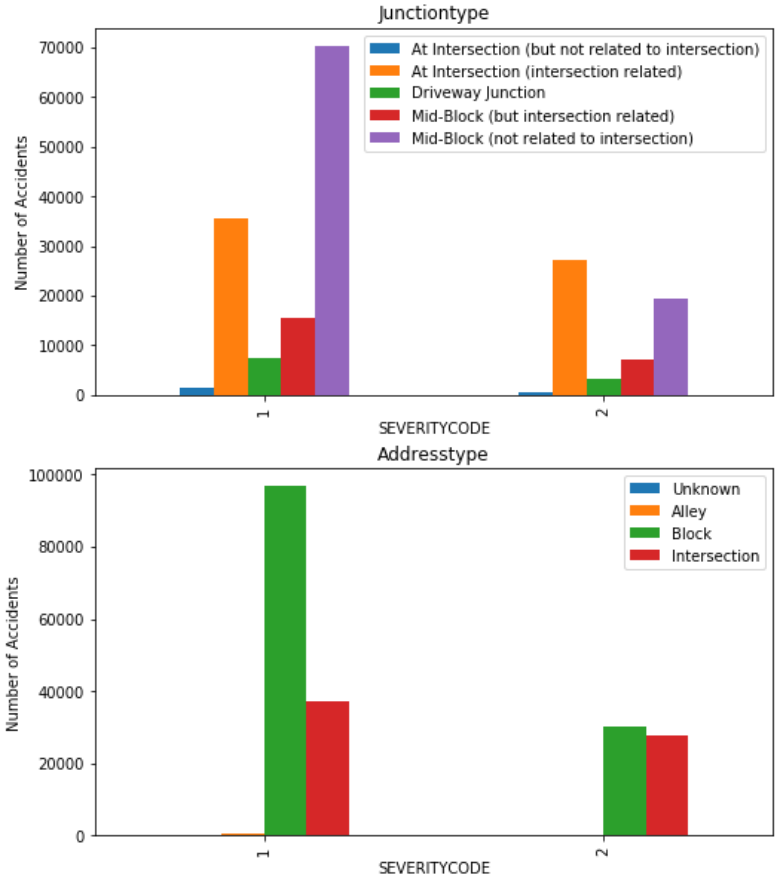
'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT'



Pedestrians and cyclists are the most vulnerable road users. When they are involved in an accident, it is likely to be a sever crash. Hence, it is an obvious criteria contributing to a good model to estimate the probability of a severe crash.

When the number of persons involved in an accident increases, the ration of severe accidents also increases. The same is true of the count of pedestrians, cyclists and vehicles. Given the data, most accidents are severe as soon as one or road users, which are not other vehicles are included. However, for the count of vehicles the ratio of severe accidents rises a lot for accidents including 3 cars.

Address types
‘ADRESSTYPE’, ‘JUNCTIONTYPE’



Accidents are more likely to happen at intersections. Especially severe accidents can be caused at junctions when, for instance, a driver ignores a red light and either the vehicle itself gets hit by another vehicle driving at full speed, or the driver hits a pedestrian crossing the walkway. In any case, the consequences are severe.

The attribute ‘Junctiontype’ shows that the ratio of severe accidents at intersections increases a lot, while accidents in mid-blocks are less likely to be severe. Accidents in mid-blocks related to intersections, on the other hand, have a higher probability to cause physical injuries.

This result is additionally demonstrated in the second graph where the ‘Addresstype’ is plotted. Accidents at intersections are far more likely to be severe than at blocks.

Both of the factors are valuable and should be included in the model.

Locations

Some locations in Seattle are more prone to have severe accidents. I chose two different methods to analyze the location of an accident. The first one includes the street addresses given by the column ‘LOCATION’, which is of type string. And the second one evaluates the geographical data given by the columns ‘X’ and ‘Y’ of type float.

1) ‘LOCATION’

To find the most common places for accidents to happen, I converted all elements from the location attribute in a list of strings and extracted the most common ones.
(The format is (‘string’, count))

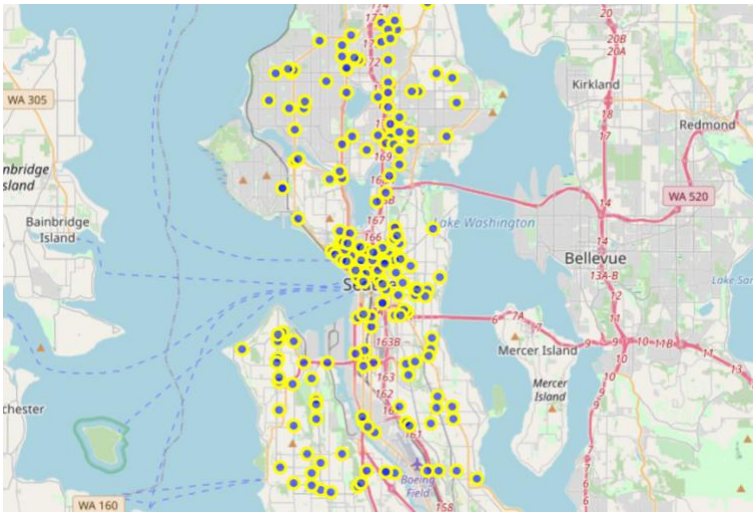
```
('ST', 28030), ('AVE', 27716), ('AND', 24104), ('BETWEEN',
16491), ('S', 15619), ('NE', 10026), ('N', 9381), ('SW', 8016),
('E', 6375), ('NW', 6121), ('W', 4429), ('WAY', 3504), ('PL',
1809), ('1ST', 858), ('DEAD', 835), ('END', 835), ('15TH', 781),
('RP', 687), ('8TH', 670), ('DR', 665), ('5TH', 603), ('4TH',
581), ('3RD', 580), ('6TH', 574), ('KING', 547), ('35TH', 535),
('2ND', 532), ('LAKE', 522), ('12TH', 510), ('45TH', 494),
('17TH', 494), ('14TH', 491), ('24TH', 481), ('42ND', 479),
('20TH', 479), ('30TH', 464), ('39TH', 458), ('23RD', 441),
('JR', 440), ('M', 439), ('L', 439), ('32ND', 428), ('16TH',
421), ('9TH', 407), ('25TH', 400), ('36TH', 391), ('22ND', 389),
('11TH', 383), ('65TH', 383), ('26TH', 380), ('46TH', 371),
('34TH', 370), ('28TH', 366), ('41ST', 363), ('40TH', 363),
('RAINIER', 358), ('50TH', 357), ('AURORA', 356), ('38TH', 356),
('WR', 355), ('7TH', 352), ('31ST', 345), ('SPOKANE', 344),
('ER', 344), ('BLVD', 344), ('BEACON', 342), ('37TH', 342),
('47TH', 337), ('18TH', 335), ('44TH', 334), ('13TH', 330),
('21ST', 318), ('OFF', 299), ('43RD', 297), ('ROOSEVELT', 296),
('80TH', 295), ('19TH', 291), ('75TH', 289), ('10TH', 288),
('48TH', 283), ('29TH', 277), ('70TH', 276), ('EAST', 275),
('55TH', 274), ('85TH', 268), ('GREENWOOD', 264), ('RD', 264),
('FREMONT', 264), ('27TH', 262), ('PARK', 254), ('MERCER', 254),
('WASHINGTON', 253), ('LINDEN', 252), ('NB', 247), ('BR', 244),
('33RD', 241), ('AV', 239), ('DENNY', 239), ('ON', 232),
('62ND', 232)
```

Most common strings are of course St and Ave following later by certain numbers. After examining streets and avenues with high traffic volumes and there lengths, 1st, Aurora

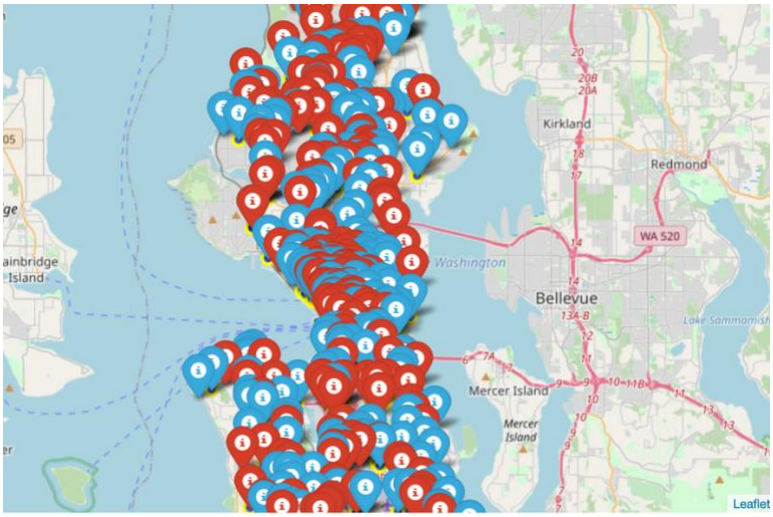
or Beacon Avenue appear to be prone for car accidents. The string ‘Dead End’ also appears often in the data and leads to the conclusion, that many accidents do happen in dead ends. Additionally, considering directions in the street name, many sever accidents appear to happen in the south of Seattle.

2) ‘X’,’Y’

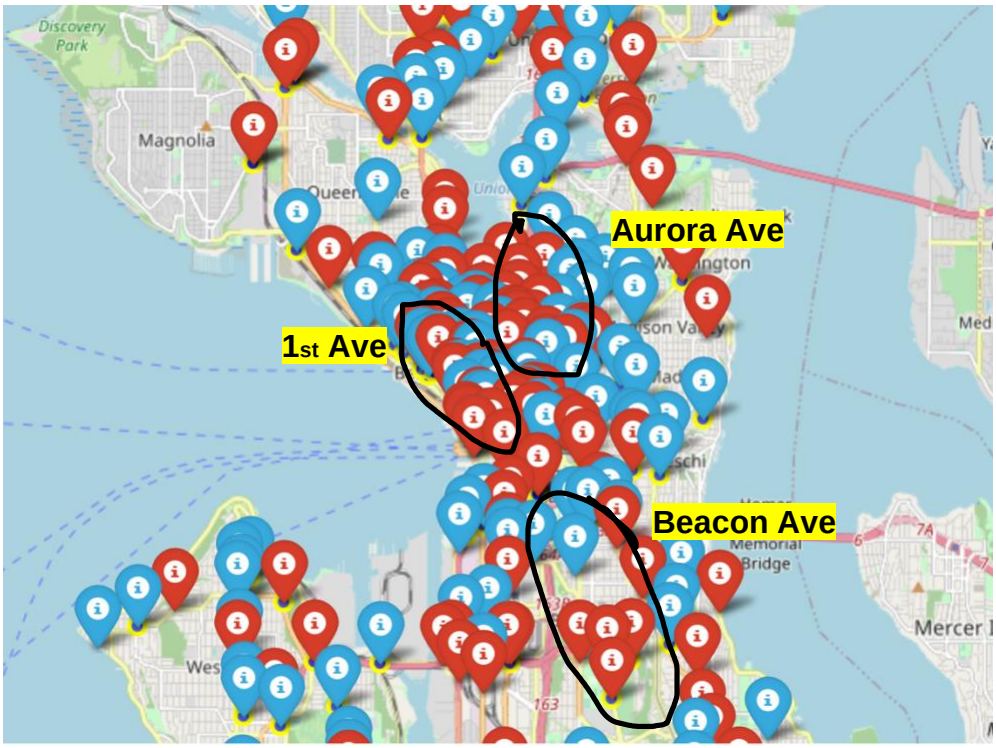
An easy way to plot the traffic data of accidents on the map of Seattle is to work with ‘folium’. It creates an interactive map and can add markers to accidents:



The accidents displaced on the graphic are just a subset of the actual dataset, but it helps to get a feeling of where accidents are more likely to happen. Now I labeled the markers. The blue labels are not severe accidents, and the red are severe ones.



Here we see that there are some areas with more red labels than others. We can zoom in further and can maybe find 1st, Beacon and Aurora Avenue from our string evaluation before.

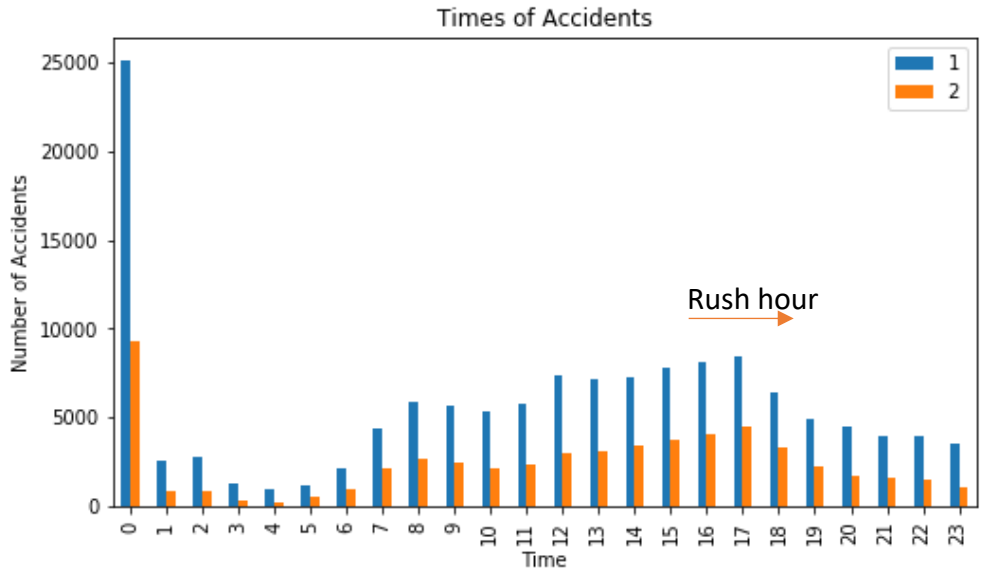


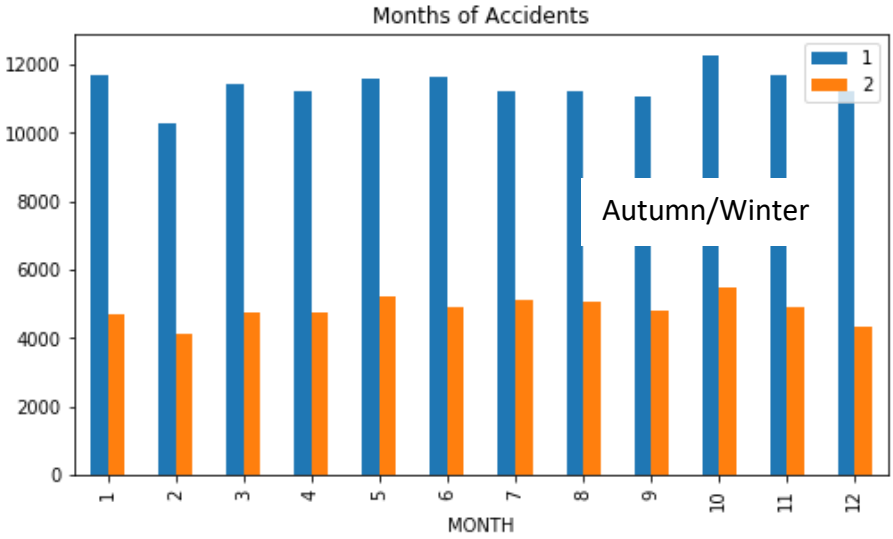
A KNN clustering will examine the clusters prone for severe accidents in the following section.

Time of Accident

‘INCDTTM’

To use the data for further evaluation, it first needed to be converted into Panda datetime data frame.





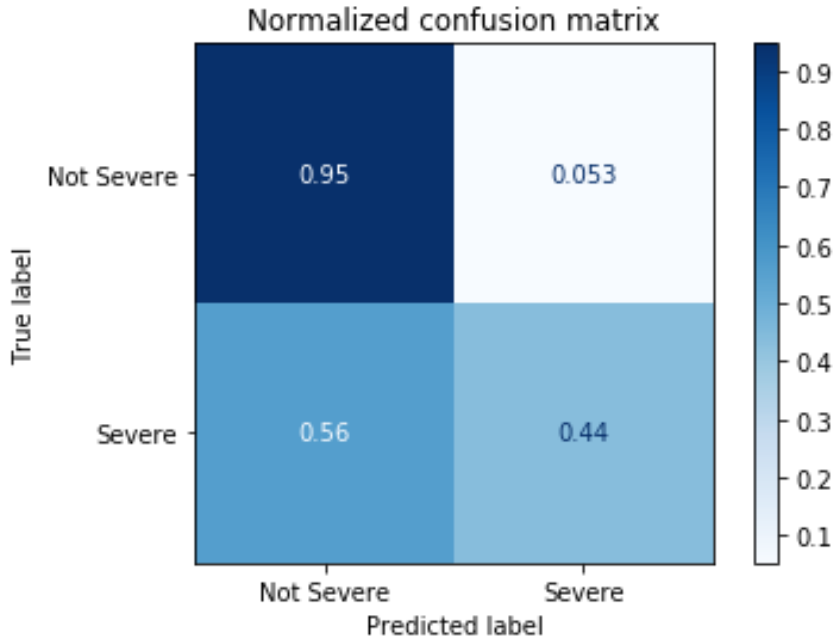
During rush hour the traffic volume is highest. Hence, it is expected to see a higher number of accidents happening around 5PM. The ratio of severe accidents also proves to be here during rush-hours, because people might want to get home as fast as possible. Or they do not pay attention to the road because their minds are still at work.

The month of the accident also plays a role. Since weather and light conditions are worse in Autumn and Winter severe accidents are more likely to happen, which is indicated by the bar graphs.
Daytime data and data of the specific month shall be therefore included in the model.

Results and Discussion

Clustering without Location

After running the KNN, SVM and Logistic Regression algorithm on a training set of size 2400 and a test set of size 640, the KNN delivers the best result for k = 4 with a F1 score of 0.803 and a Jaccard index of 0.82.



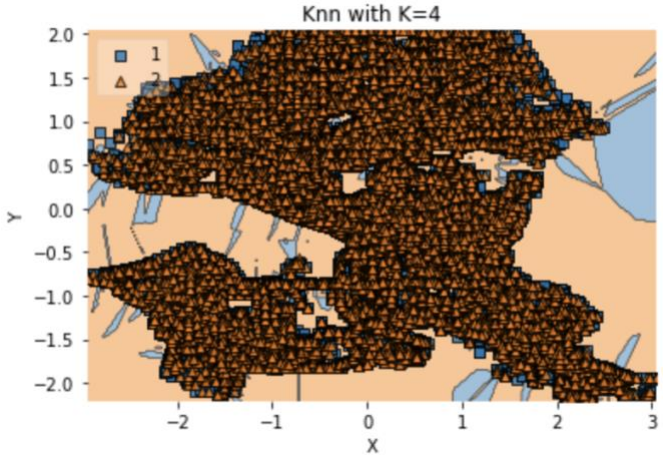
It is very good in predicting not severe accidents. However, it labels many data as “not severe” even though the accident was severe. Several problems in the data causes this inaccuracy:

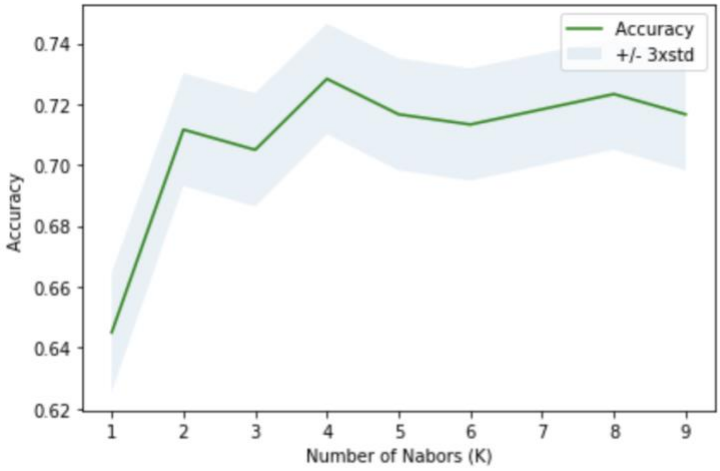
1) Many attributes which could strongly indicate the risk of a severe accident like ‘SPEEDING’ or ‘INATTENTION’ are labeled as unknown. In such cases the police could neither verify nor falsify if the driver broke the law.
   Therefore, to get a more accurate result, the speed of the car before the accident had happened and the allowed speed limit should be included into the dataset. Furthermore, street types such as highways or roads should be another categorical variable in the dataset.
2) The dataset was in general unbalanced with a ratio 13:5 of not severe accidents to severe accidents.

Nevertheless, for the given dataset the algorithm performs well given the dataset.

Clustering of location on the map

In the data analysis section, the map showed that there were specific regions which had more severe accidents than others. Clustering the regions with the KNN algorithm showed K = 4 with the optimal accuracy, which corresponds to the observations on the map.






```
The best accuracy was with 0.7283333333333334 with k= 4
```

Conclusion

As suggested, the most important features which contribute to the probability of severe accidents are hazardous areas like junctions, weather conditions, human failure, number of accident participants and the time of the accident. Additionally, the location given by geographical coordinates can also contribute to determine whether or not a severe accident is likely to happen. Therefore, to provide the best estimation of danger for a driver, it is important to analyze…

… his upcoming route with the most dangerous zones given by a KNN clustering.

… if the driver is not under influence and is reminded not to speed.

… the weather, lighting and road conditions.

… the time of the day and the year. For if it is late and autumn or winter, severe accidents are more likely to happen.

Opportunity:

To warn a driver before starting his ride, applications can be developed for cities to take those collision statistics into account. Those warning systems could even be integrated in GPS devices.