CAPSTONE PROJECT
THE PREDICTION OF SEVERE ACCIDENTS WITH COLLISION DATA
(CASE STUDY: SEATTLE)

NATASCHA HEY

09/16/2020

INTRODUCTION: BUSINESS PROBLEM

- Objective:

  To examine the probability getting into a car accident and to predict the severity of this accident.

- Who can profit ?

  Car insurance providers

  Car manufacturers

  Car drivers and with them pedestrians and cyclists

- How can machine learning help?

  → Machine learning helps to assess the highest risk factors of a severe accident

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:

Dataset:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Description to data:

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

Attribute	Data type, length	Description
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • **Alley** • **Block** • **Intersection**
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.

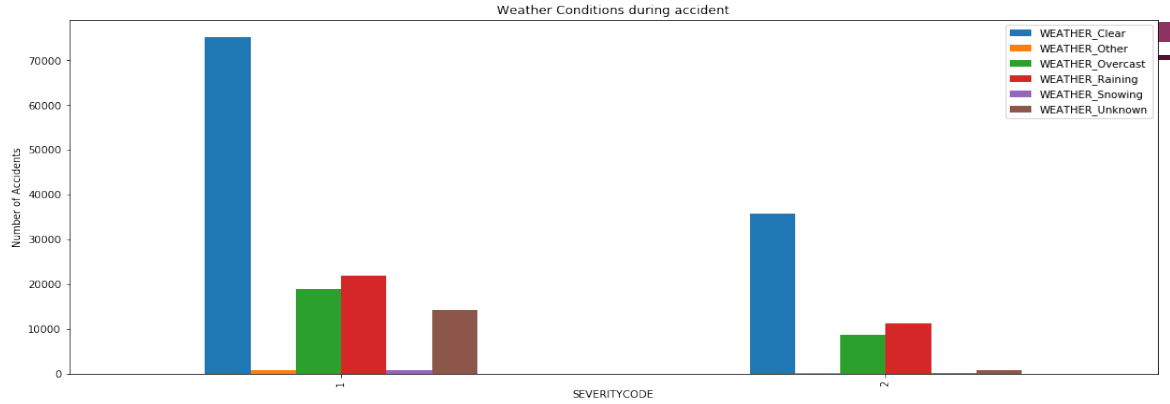
Attribute	Data type, length	Description
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.

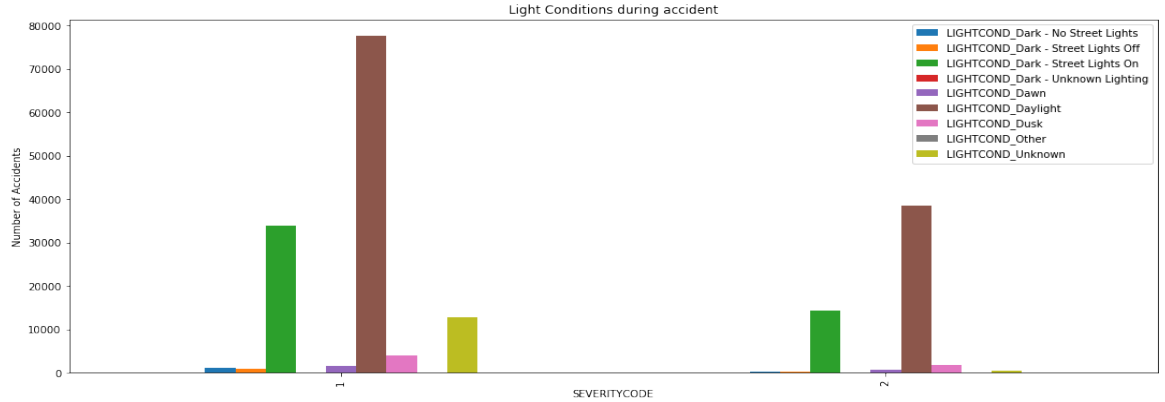
DATA

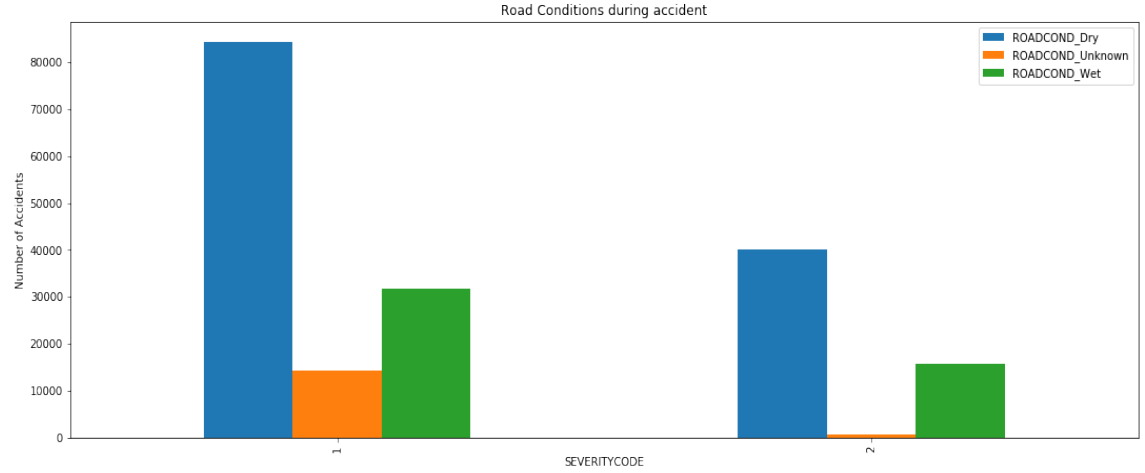
- Choice of Data by groups:
  - Target column: ‘SEVERITYCODE’
  - Locations or hazardous areas 'X', 'Y' 'ADDRTYPE' 'LOCATION' JUNCTIONTYPE'
  - Conditions caused by Nature WEATHER', 'ROADCOND', 'LIGHTCOND',
  - Human Failure 'INATTENTIONIND', 'UNDERINFL’, ‘SPEEDING’
  - Count accident participants 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT',
  - Time of the Accident 'INCDTTM',
  - Unnecessary columns (the rest to be dropped)
- Dataset is not balanced
  - 1 136485 (not severe accidents)
  - 2 58188 (severe accidents)

METHODOLOGY

- Data preparation
  - extract the data from a csv file and store it into a data frames using pandas and numpy
  - Clean data and drop all nan or convert to 0
  - Check datatype in each column. For instance, in column 'UNDERINFL', we find the unique data entries 'N','Y','0','1'. For the data to be accurately evaluated, the strings had to be mapped by {'Y':0,'N':1}
  - Create dummies for categorial data
- Data Visualization
  - Bar graphs with matplotlib.pyplot,
  - KNN clusters with mlxtend.plotting
  - Traffic data visualization on Seattle map with folium
  - Confusion Matrix with sklearn.metrics
- Modeling
  - Classification with KNN, SVM, Logistic Regression with sklearn
  - F1 and Jaccard Index for accuracy







DATA ANALYSIS

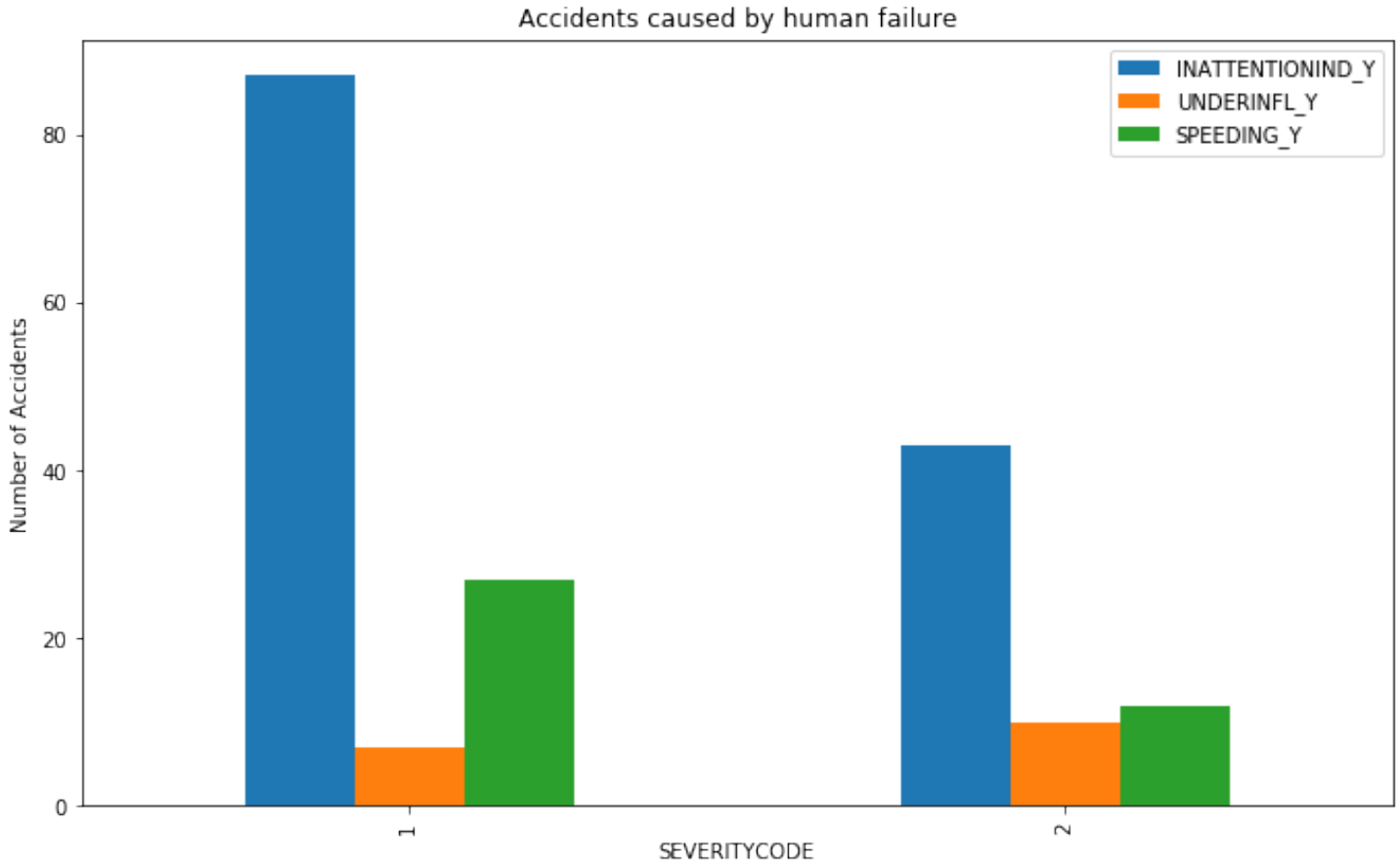
1) ACCIDENTS CAUSED BY BAD WEATHER CONDITIONS

SEVERE ACCIDENTS MORE LIKELY WHEN:

- WEATHER = ‘RAINY’, ‘OVERCAST’
- LIGHT CONDITIONS = ‘DARK’
- ROAD CONDITIONS= ‘WET’

2) Accidents caused by human failure

Severe accidents are more likely to happen when Speeding and under influence.



3) Count of Participants in Accident

- Personcount:

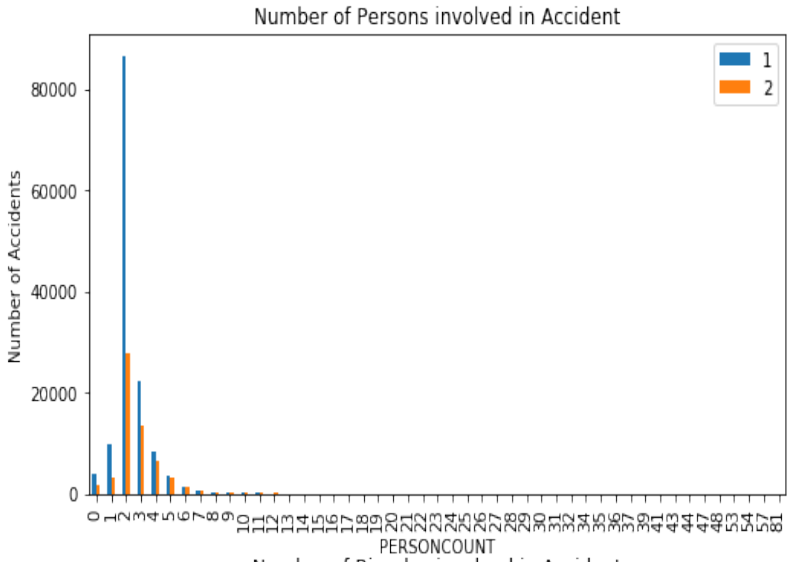
The more persons involved, the more likely it is for the accident to be severe

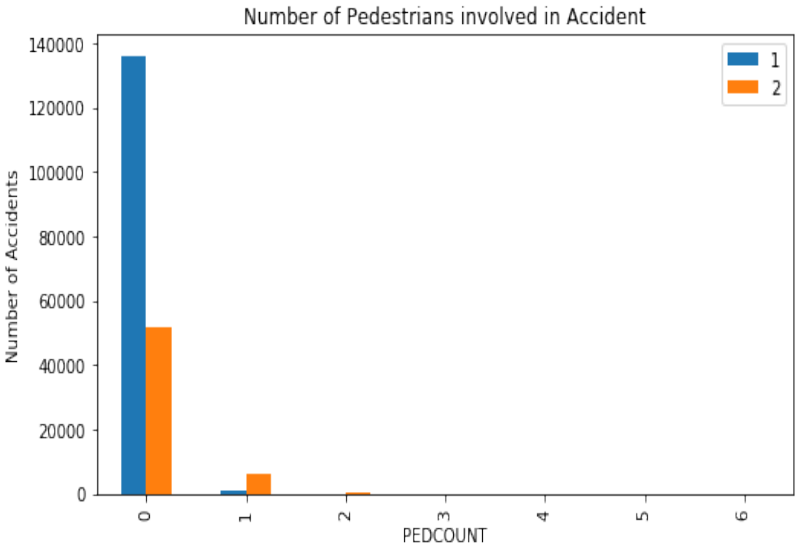
- Pedcylcount and Pedcount:

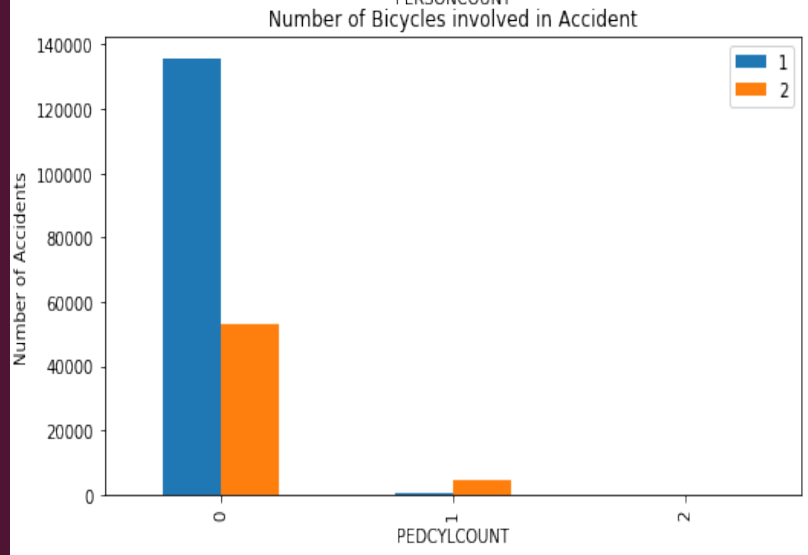
If pedestrians and bicycle riders are involved, the accident is likely to be severe.

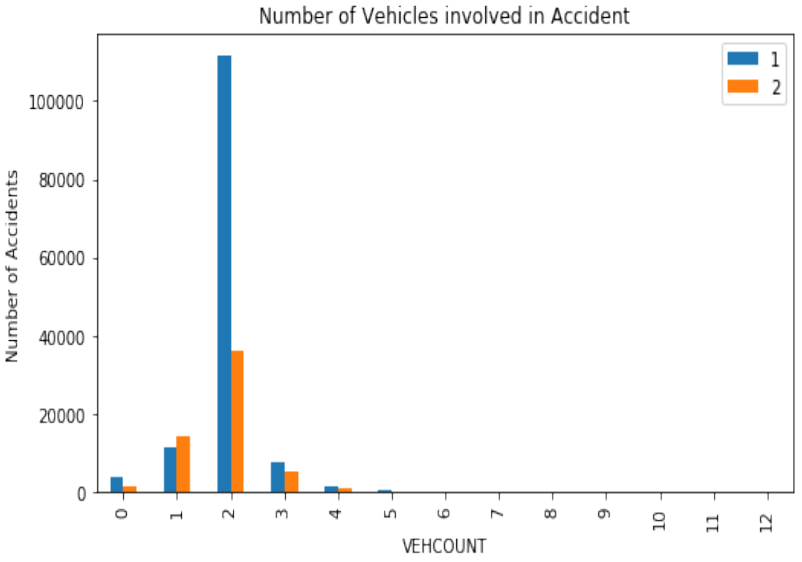
- Vehcount:

The accident is likely to be severe when more vehicles are involved



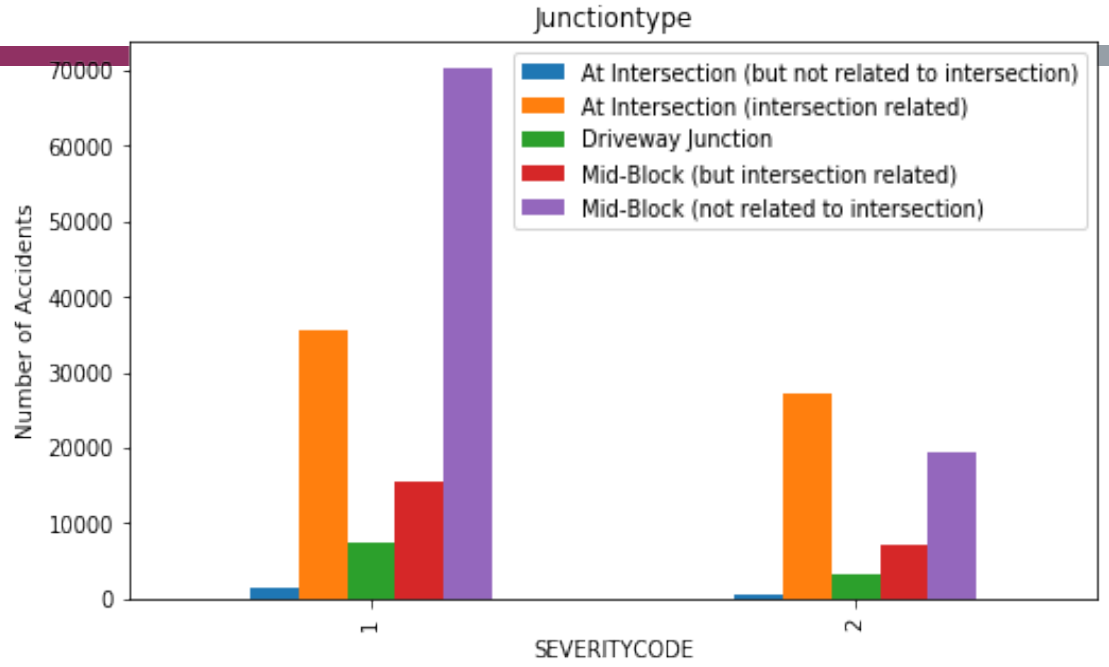


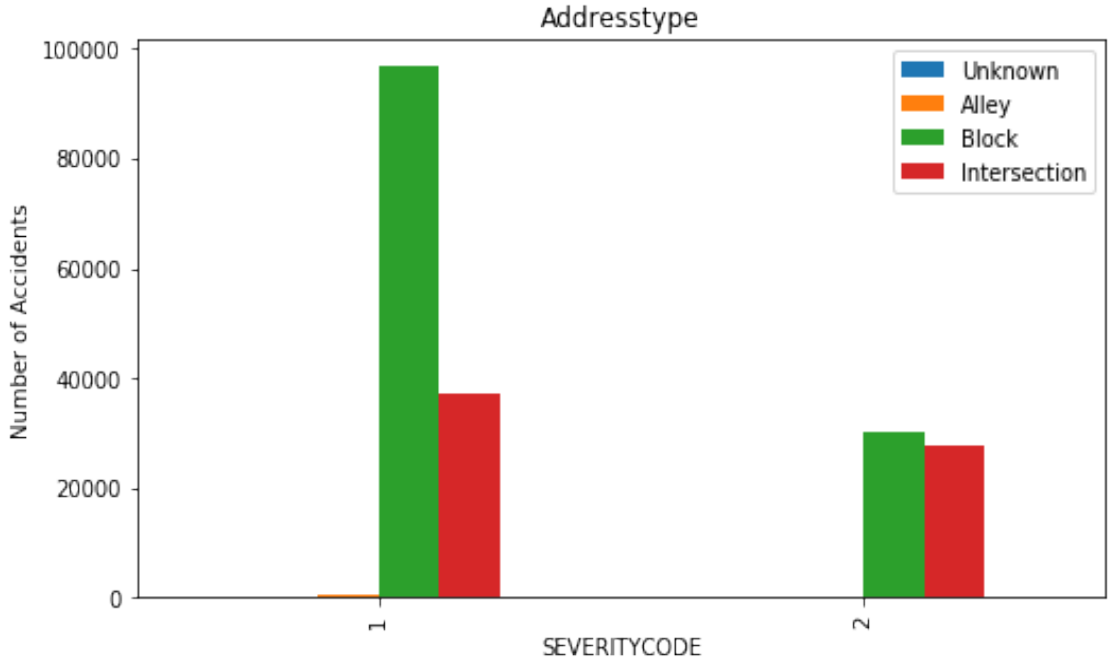




4) Address and Junction Types

Severe accidents are more likely to happen at Intersections





5) Locations

5.1) String Analysis

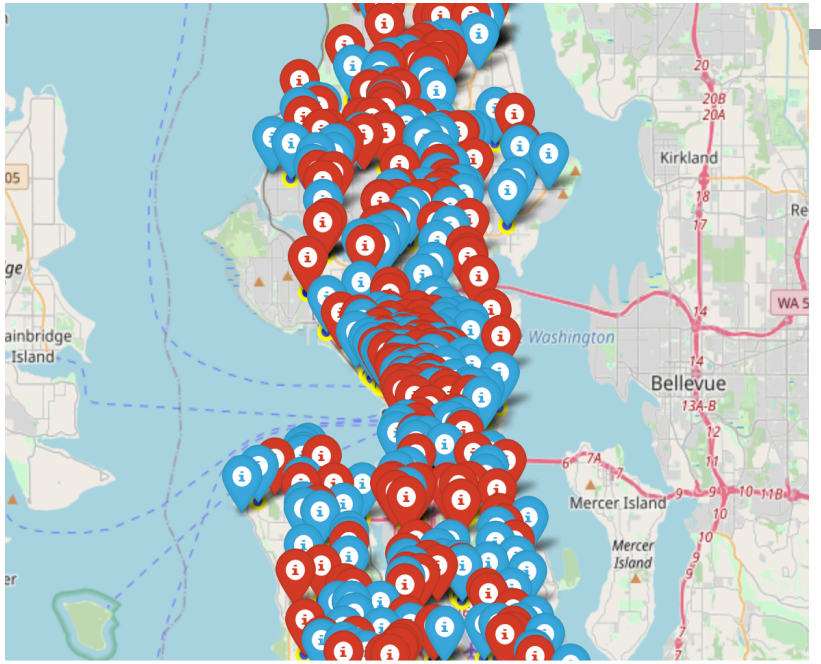
Many accidents occur

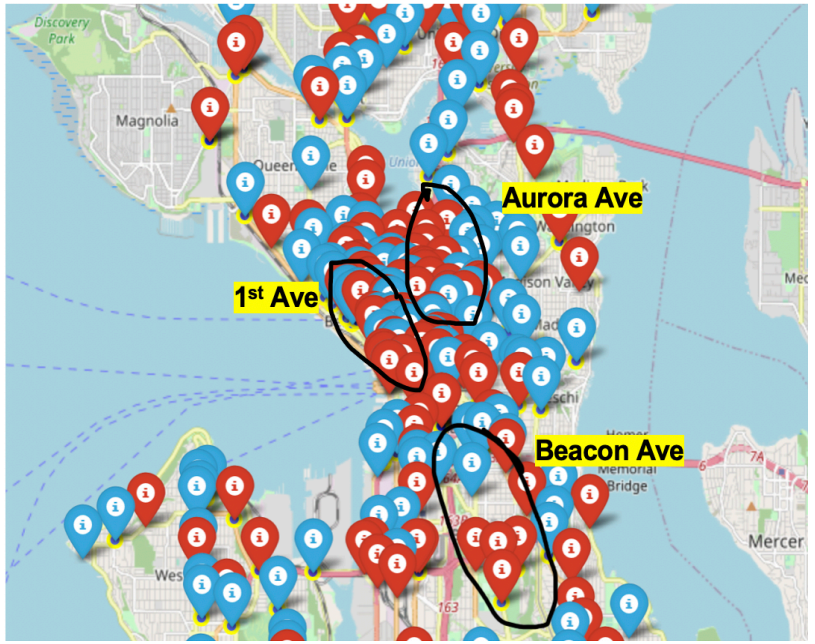
- In 'Dead Ends'
- Along 1st, Beacon and Aurora Avenue
- More in the South

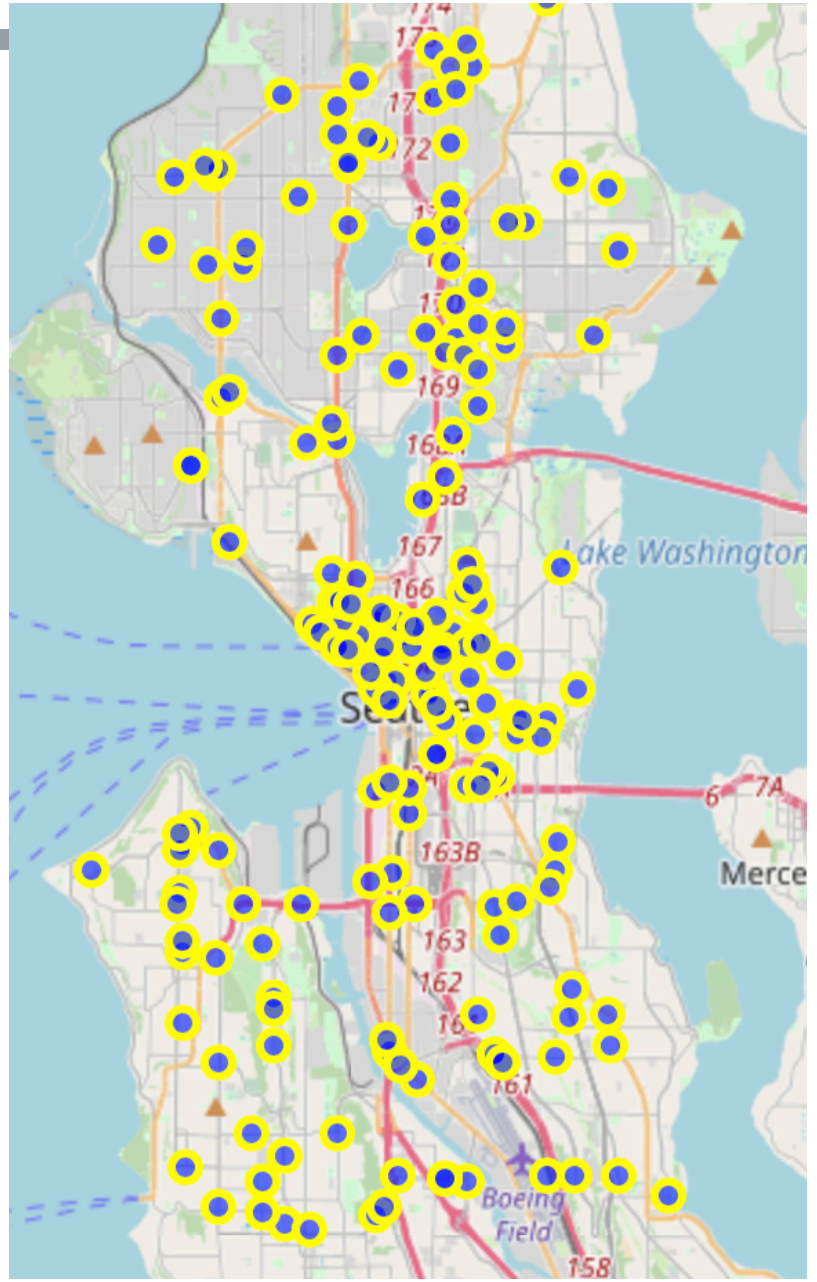
```
('ST', 28030), ('AVE', 27716), ('AND', 24104), ('BETWEEN',
16491), ('S', 15619), ('NE', 10026), ('N', 9381), ('SW', 8016),
('E', 6375), ('NW', 6121), ('W', 4429), ('WAY', 3504), ('PL',
1809), ('1ST', 858), ('DEAD', 835), ('END', 835), ('15TH', 781),
('RP', 687), ('8TH', 670), ('DR', 665), ('5TH', 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)
```

5) Locations

5.2) Spatial Data Analysis with Folium







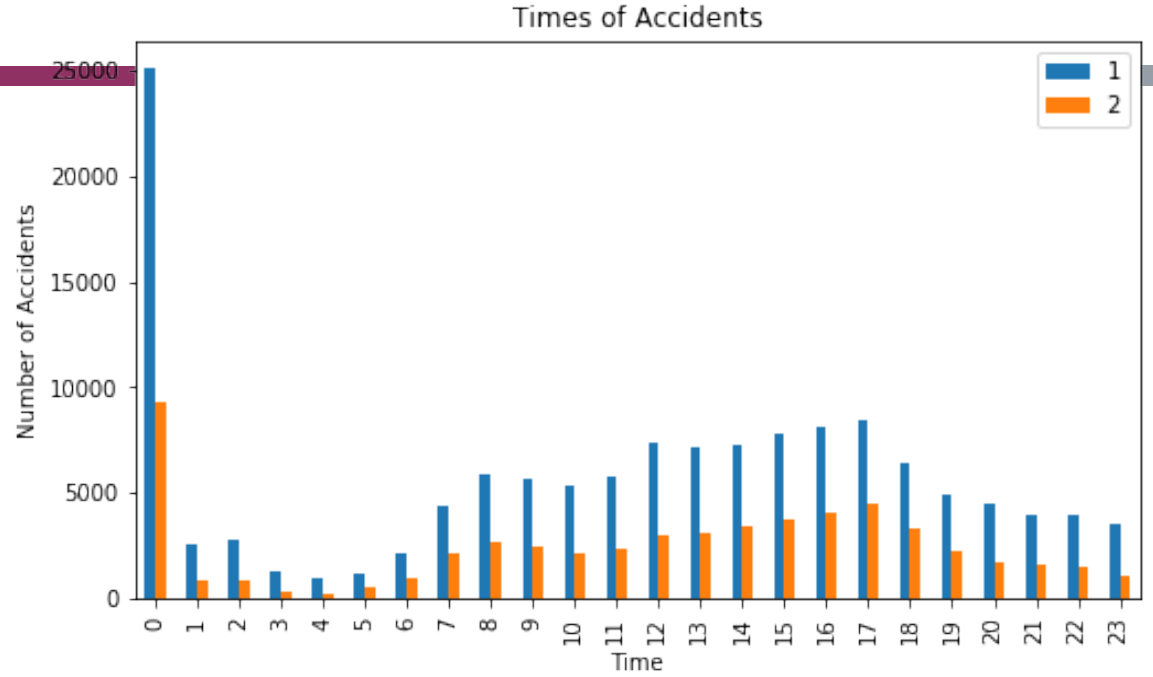
6) Time of accident

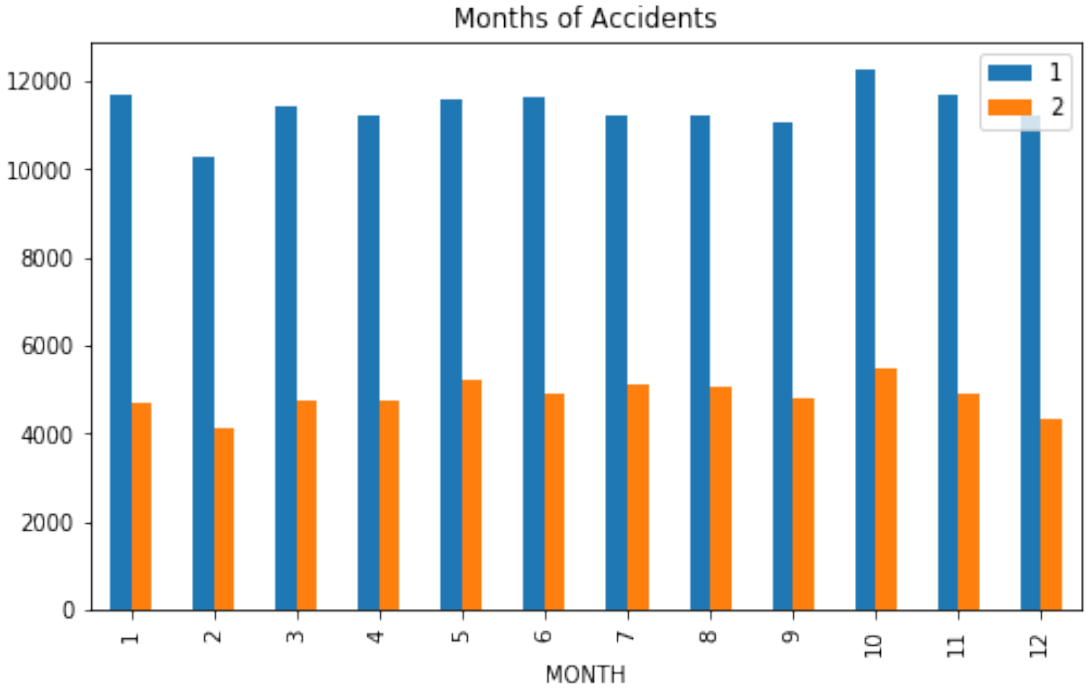
- Months:

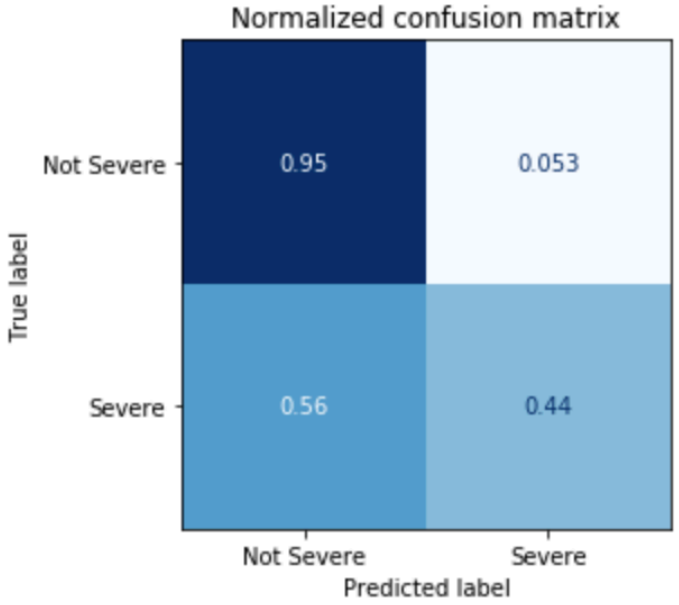
(Severe) accidents are more likely to happen in October

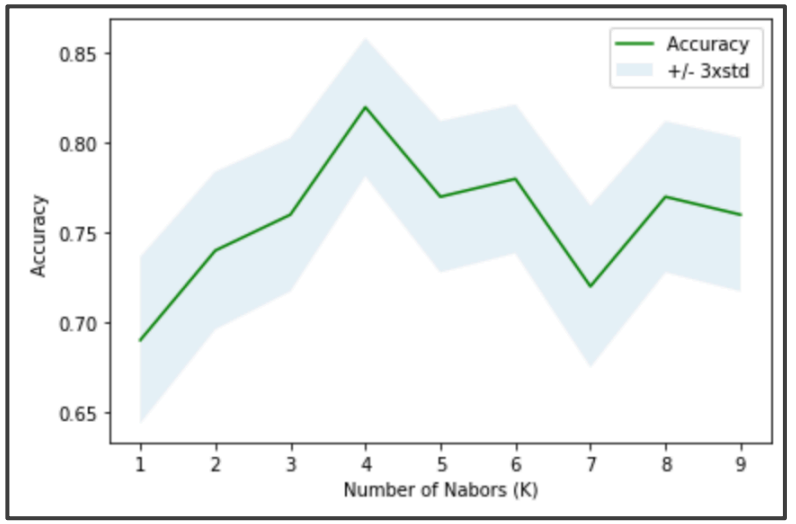
- Time:

Severe accidents are more likely to happen during rush-hour and at midnight

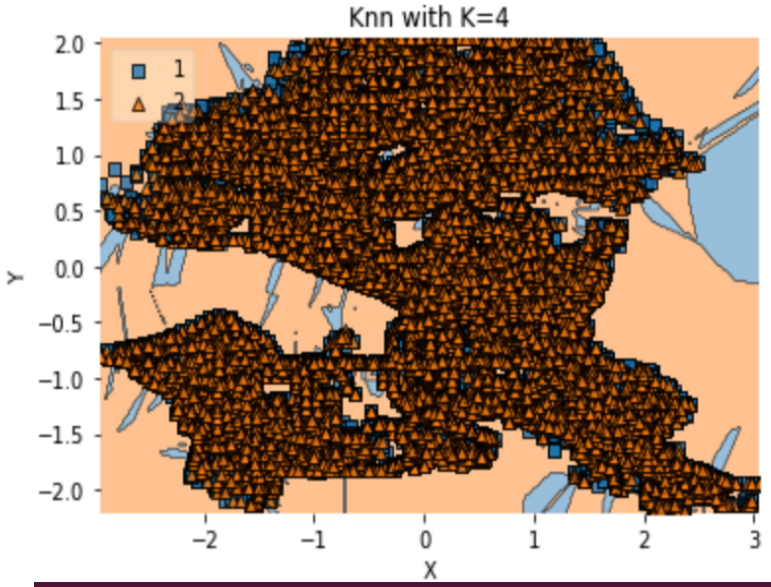








The best accuracy was with 0.82 with k= 4



RESULTS

- The **KNN** algorithm delivers the best result for **k = 4** with a **F1 score of 0.803** and a **Jaccard index of 0.82**.

DISCUSSION

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:

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.
- 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.

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.

<u>Opportunity:</u>

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.