Predicting Car Accident Severity

Problem Statement

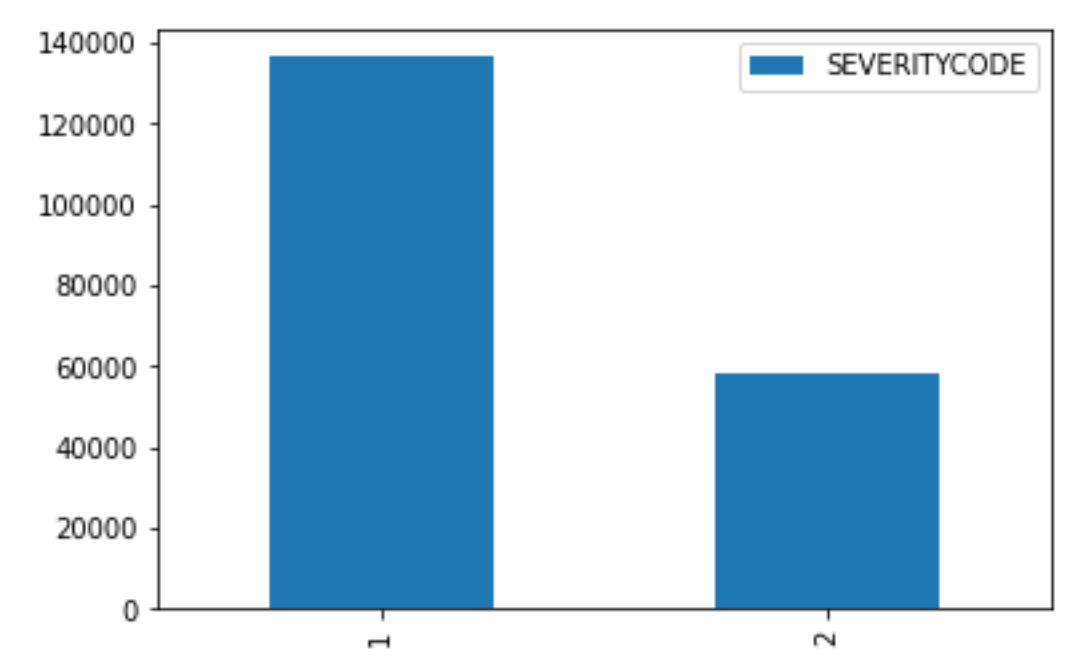
- The problem is to predict the severity of the collision which may happen at the intersection or mid-block of a segment.
- It is a classification problem where a suitable algorithm will be trained on the data set.
- The prediction will be made through a machine learning model trained on the Seattle collision dataset which has been recorded by the Traffic Records of the Seattle police department since 2004.

Data Description

- The dataset is provided by the Seattle Police Department traffic records which consists of the severity of collisions which may happen at the intersection or mid-block of a segment.
- https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0
- Each case has been assigned a **severity code** label
  - 0 which means unknown
  - 1 which means proper damage
  - 2 which means injury
  - 2b means serious injury
  - 3 is for fatality
- The data has 37 input attributes
  - Location of accident, collision type, No. of vehicles involved including bicycles, passengers and pedestrians.
  - Total number of injuries, serious injuries and fatalities
  - Date, time and the location of the collision
  - Road, weather and light conditions during the collision
  - Notable attributes: whether the driver involved was under any drug or alcohol influence. Whether speeding and inattention was a factor, whether a pedestrian right of way was granted, and whether the collision involved hitting a parked car
  - Miscellaneous attributes: give details about collision code both at local and provincial level with description, lane segment and crosswalk keys.

Initial Findings: Unbalanced Dataset

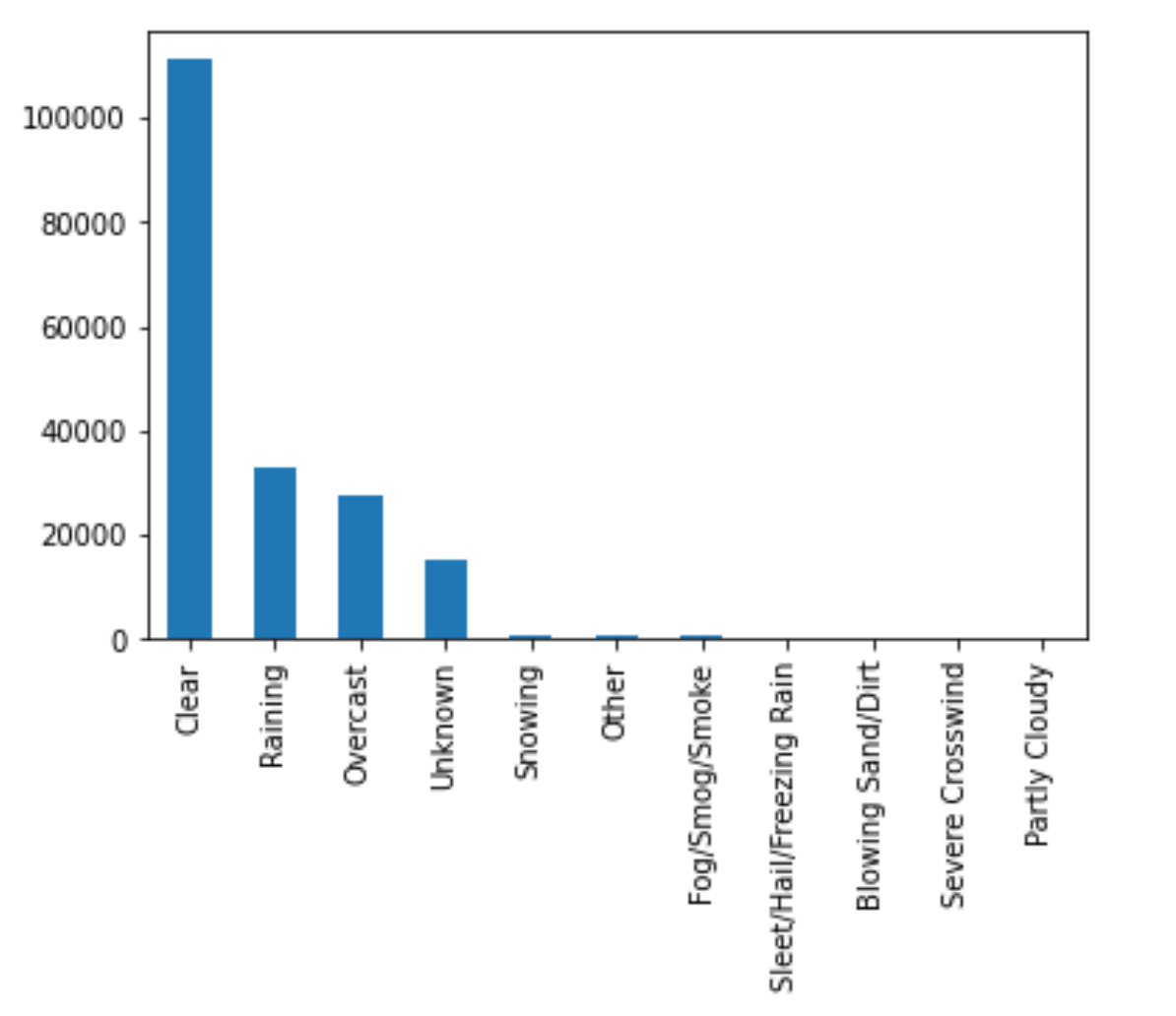
- Originally 5 labels listed in the meta information of the dataset, however the data contained **only 2 labels**. 1(no injury) and 2 (injury)
- Out of a total of 194673 cases, 136485 were class 1 and only 58188 were class 2



- This is unbalanced data sets which will require various data wrangling techniques such as under-sampling, over-sampling to reduce the bias in the model.

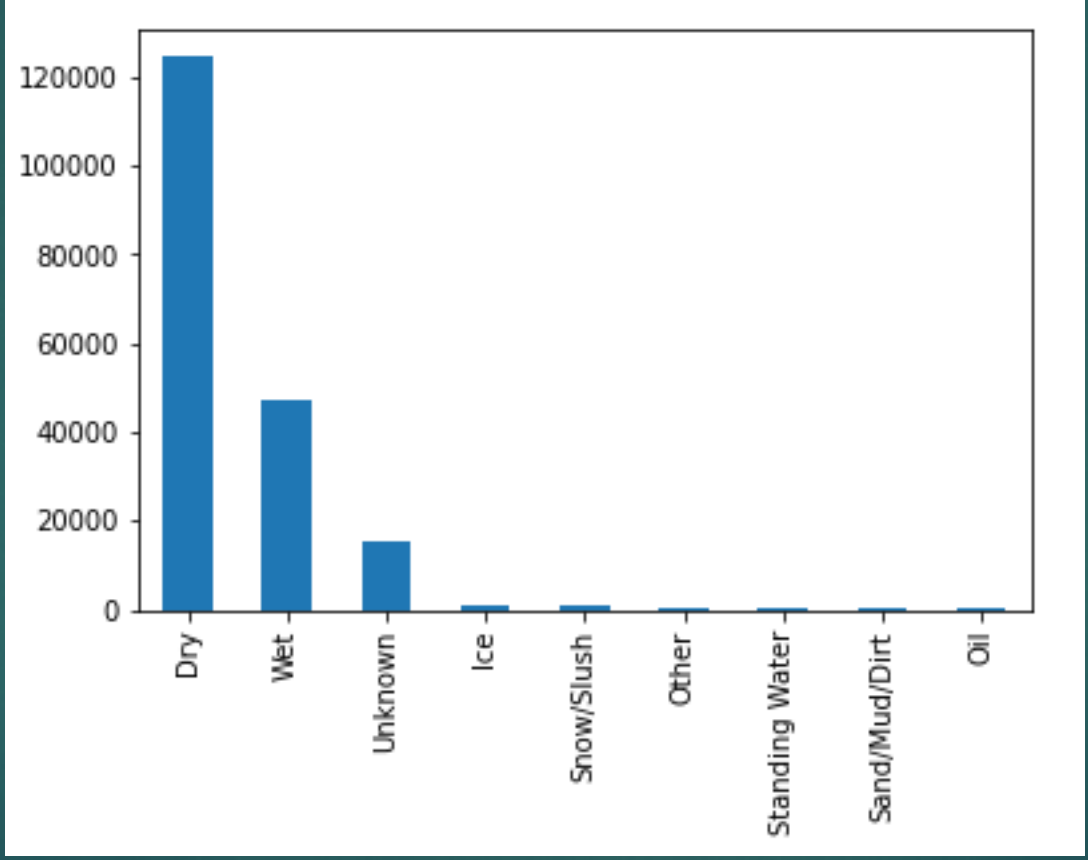
Effects of WEATHER Conditions

- A counterintuitive outcomes as the larges number of collisions took place under clear weather



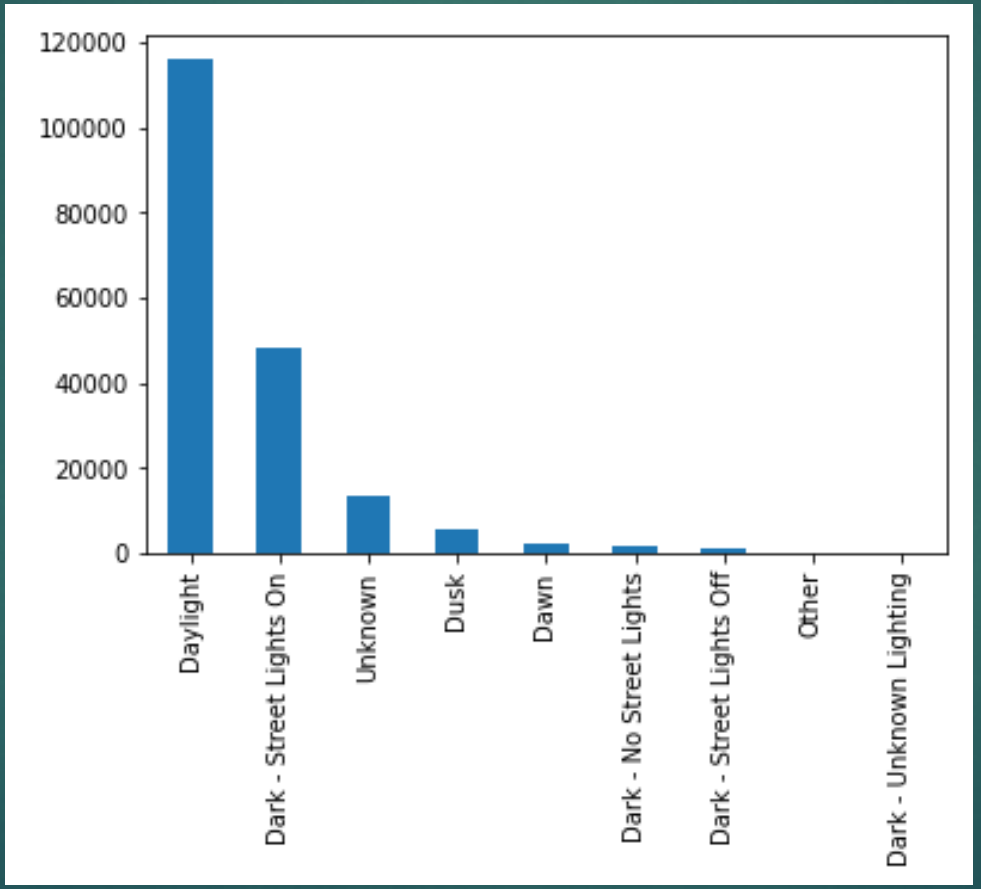
Effects of Road Conditions

- Most of the collisions took place under dry road conditions



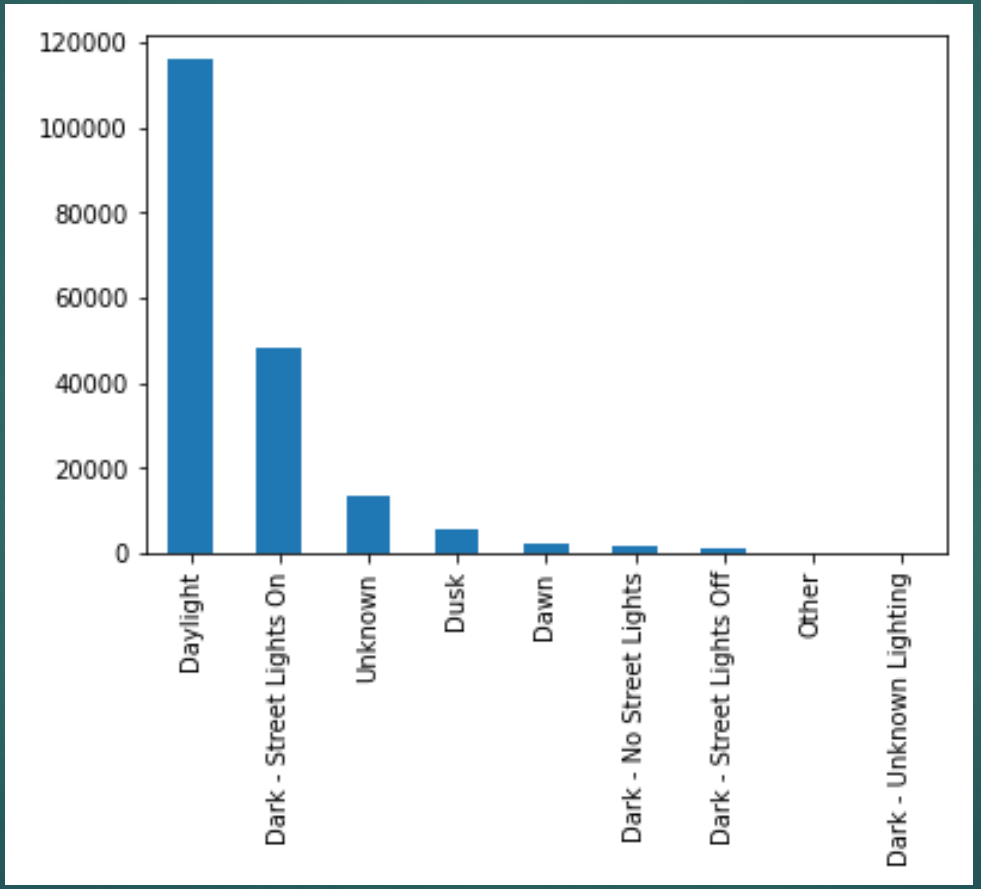
Effects of Light Conditions

- Most of the collisions took place during daylight as opposed to the common misconception that likelihood of a collision will be more in the dark.



Effects of Light Conditions

- Most of the collisions took place during daylight as opposed to the common misconception that likelihood of a collision will be more in the dark.



Effects of Speeding/Driver under influence/Inattention

- A sizeable number of 9333 cases where driver was speeding.
- Around 29805 cases reported driver inattention as one of the reasons for collision.
- Another notable factor was driver under influence of alcohol and/or drugs which was reported in 9121 cases.

Feature Engineering

Some Observations

- A feature sub-set was selected after careful analysis of the dataset.
- Various attributes which were used for mostly data logging purposes and did not show direct influence on predicting the severity of the collision were dropped.
- The features with the most missing values were also dropped, one important example is that of SPEEDING attribute which seems significant but has values only for 9333 cases.
- INATTENSIONIND was also available for 29805 times, however it deemed important to be kept

Feature Engineering

Feature Name	Type	Description
ADDRTYPE	Text	Collision address type: Alley, Block, Intersection
COLLISIONTYPE	Text	Collision Type
HITPARKEDCAR	Text	Whether or not the collision involved hitting a parked car. (Y/N)
PEDCYLCOUNT	Double	number of bicycles involved in the collision
VEHCOUNT	Double	number of vehicles involved in the collision
PERSONCOUNT	Double	total number of people involved in the collision
INATTENTIONIND	Text	Whether or not collision was due to inattention
UNDERINFL	Text	Whether or not a driver involved was under the influence of drugs or alcohol.

Data Preparation

- For the selected features set, the rows with missing values were removed.
- In order to reduce the model bias due to un-balanced data set, under-sampling technique was used.
- The categorical values of attributes were replaced by dummy/indicator variables
- The dataset was split in 80% training and 20% testing

Model Selection

- All the traditional classification algorithms were used to train the model such as Logistic Regression, SVM classifier, Decision Tree and K-Nearest Neighbors.

	Precision	Recall	Accuracy	F1-score
Logistic Regression	60%	60%	60%	60%
SVM	67%	66%	66%	66%
Decision Tree	**68%**	**67%**	**67%**	**67%**
KNN	62%	62%	62%	62%

Conclusion

- Data analysis and addressing any problems such as missing values and unbalanced data set must be resolved before model training
- Out of all the ML algorithms, Decision Tree performed the best, the second best was SVM followed by KNN.
- Logistic Regression was the least accurate method in this case, the regularization factor helps in over fitting but after trying a number of values, the best one was selected.