Capstone Project-
Car Accidents Severity Analysis

Presentation

1. Business Understanding

1. BUSINESS UNDERSTANDING

Road accidents are one of the major causes of death and disability around the world. Reasons for road accidents can be environmental conditions such as weather, traffic on road, type of road, speed ,light conditions etc. This paper addresses the in-depth analysis that identifies as the contributory factors behind the road accidents and the quantification of the factors that affect the frequency and severity of accidents based on the crash data available. The severity of each accident can be predicted quite accurately with various classification machine learning algorithms. This can ultimately help government, traffic police, medical institutions, individual drivers and the insurance companies by getting useful insights of the accident severity regarding the causes and consequences of the accidents. The Machine Learning model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city. The model will predict the accident severity with various supervised machine learning algorithms i.e.

- **Algorithm A. Logistic regression**
- **Algorithm B. The K-Nearest Neighbours (KNN) algorithm**
- **Algorithm C. Decision Tree**
- **Algorithm D. Random Forest**

And finally, the accuracy of each algorithm will be plotted to check which algorithm performs better.

2. Data Understanding

2. DATA UNDERSTANDING

The data used for this project was collected by the SDOT traffic management Division and Seattle Traffic Records Group from 2004 to present. It was downloaded from the link shared in the IBM Applied Data Science Capstone course. The data consists of 38 independent variables and 194,673 rows. The dependent variable, "SEVERITYCODE", contains numbers that correspond to different levels of severity caused by an accident from 1 to 2. Severity codes are as follows:

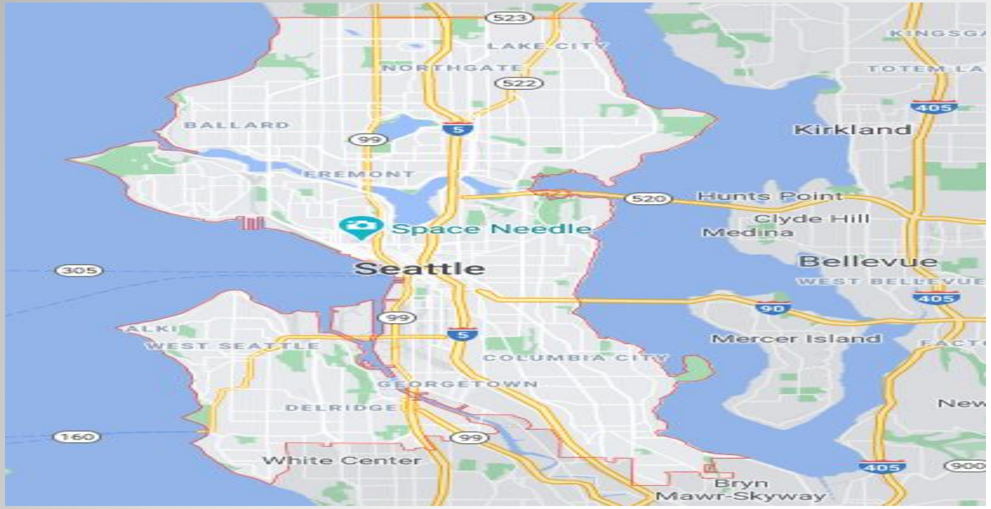
1: Property Damage Only Collision

2: Injury Collision

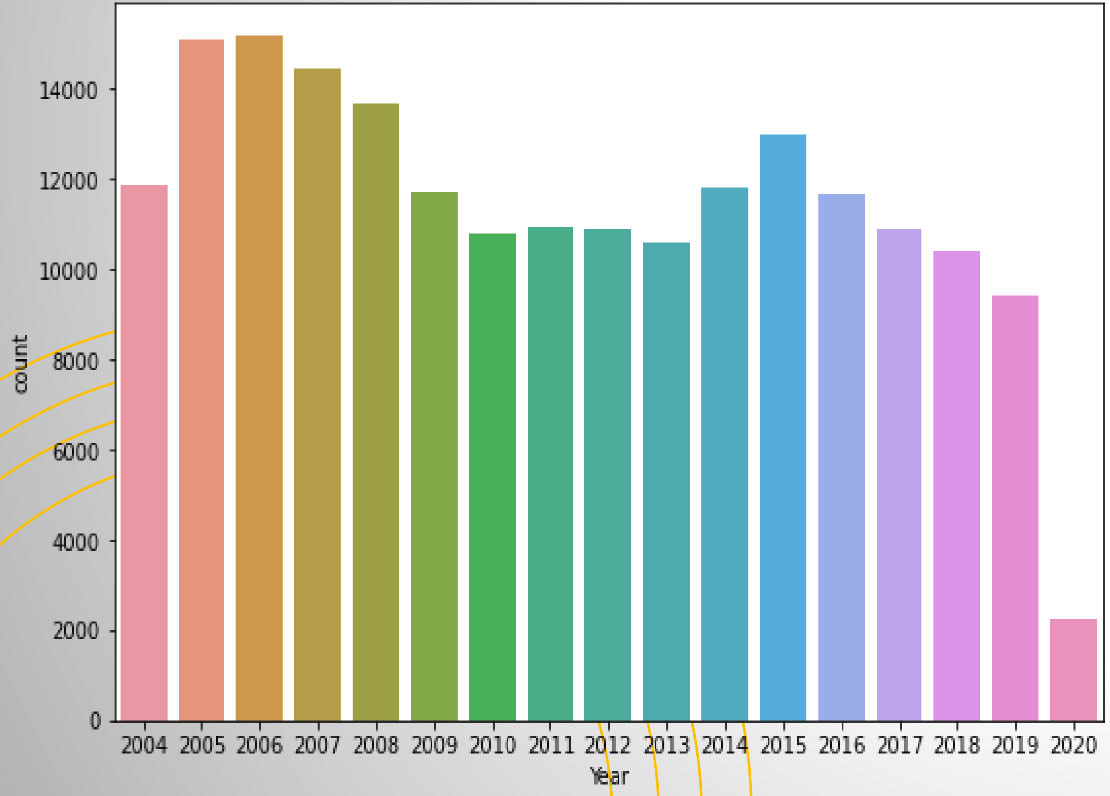
Furthermore, as there are null values in some records, hence the data has been pre-processed.

The link for the Metadata is mentioned below:

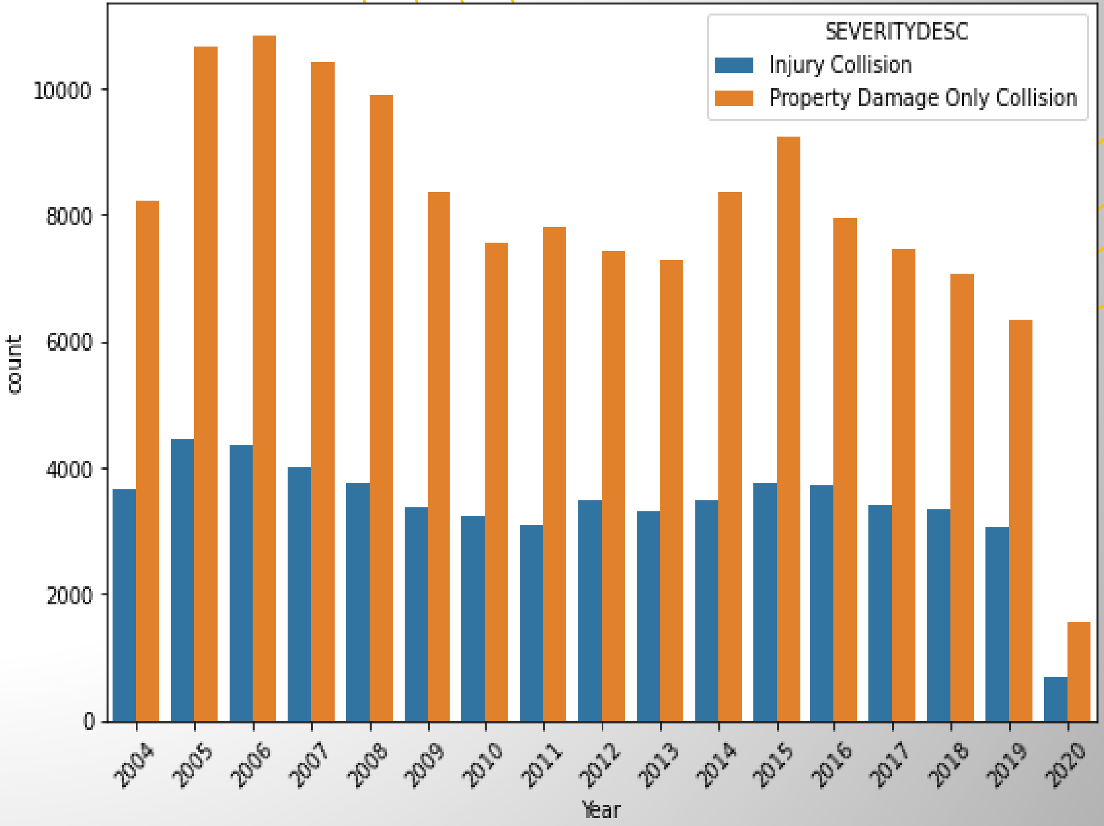
https://github.com/Engineer00/Coursera_Capstone/blob/master/Scripts/Metadata.pdf



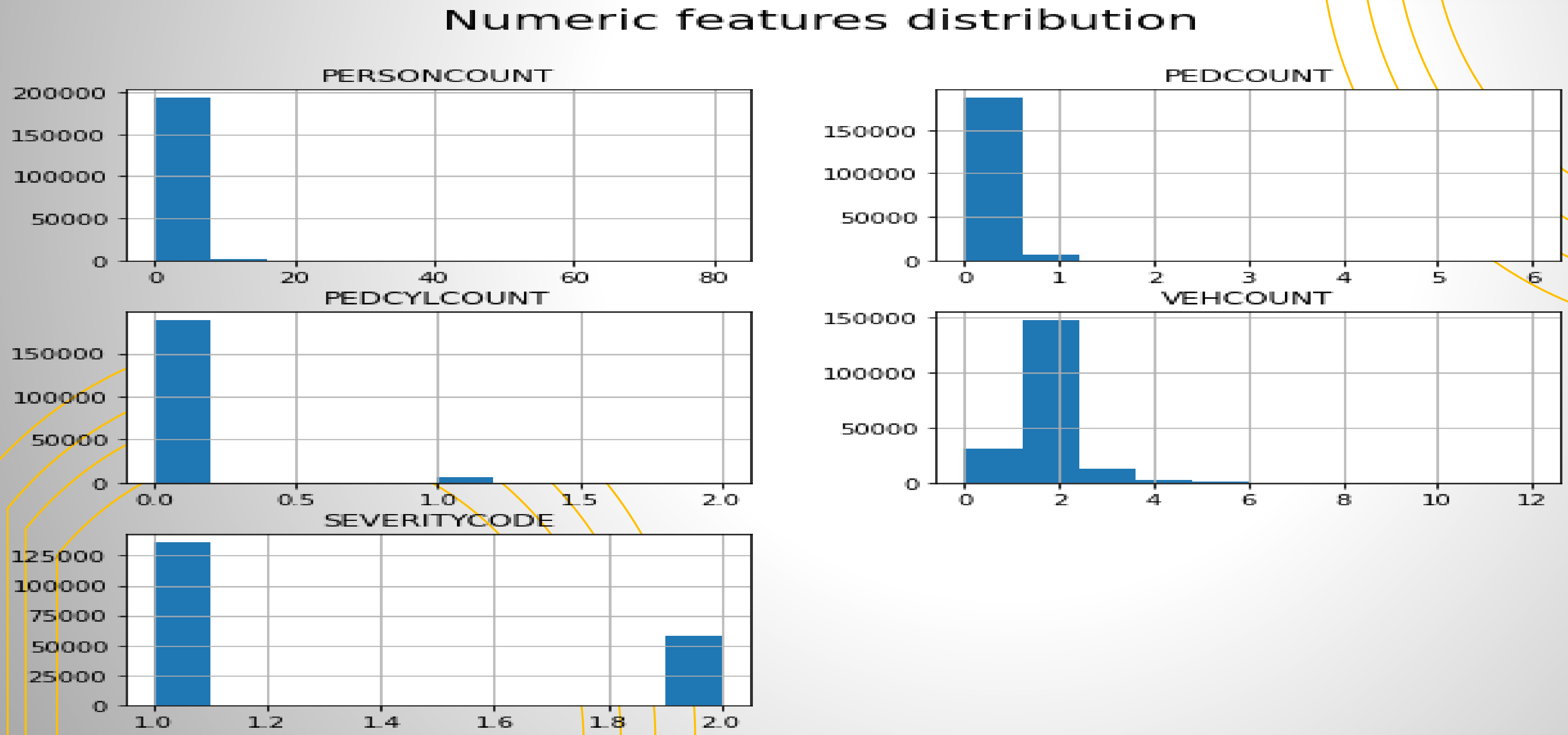
Car accidents in Seattle by Year



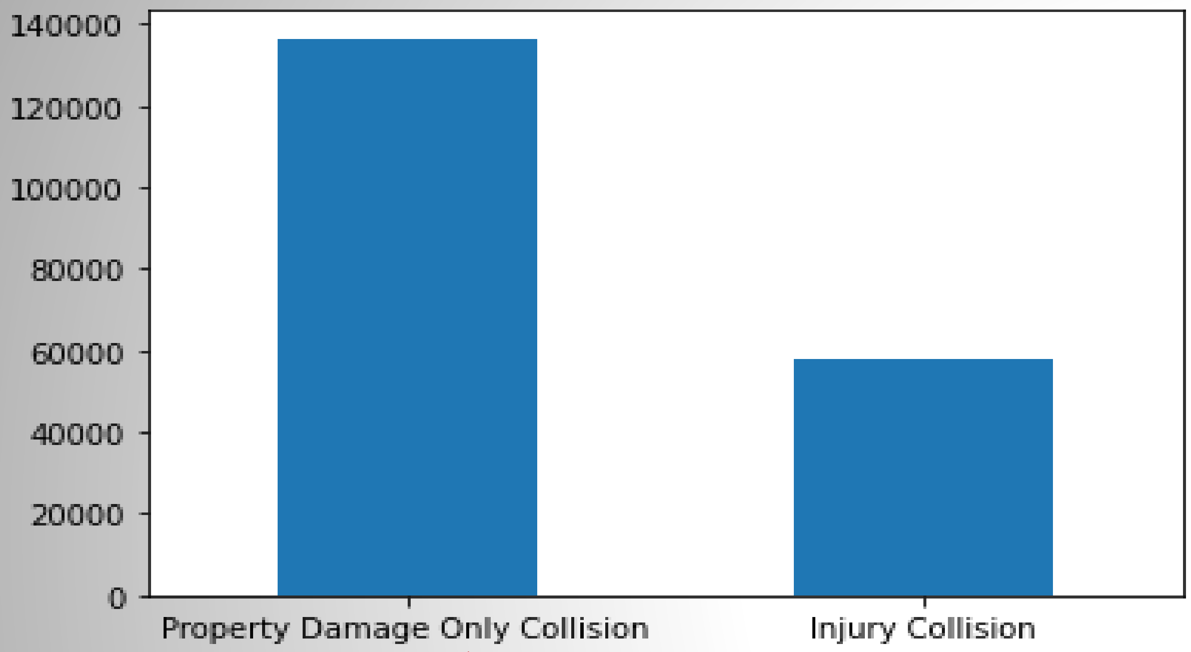
Car accidents in Seattle by Year & type



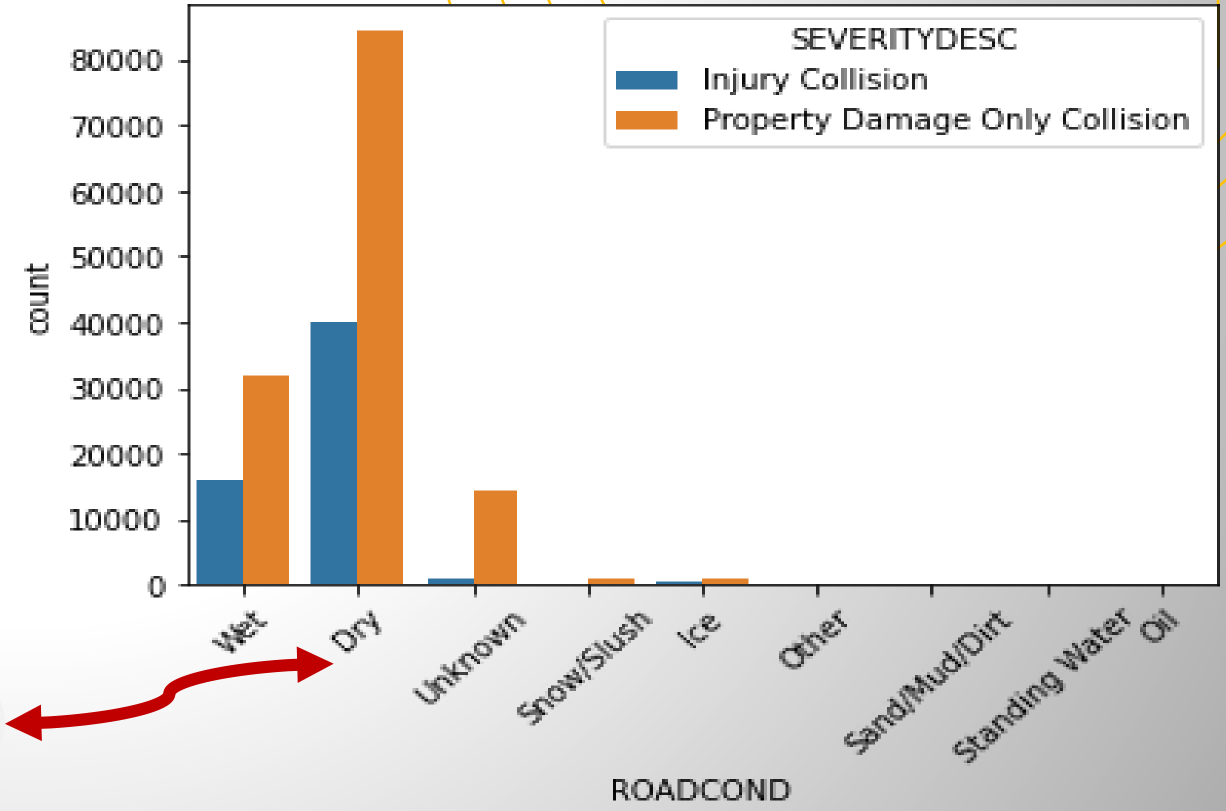
NUMERIC FEATURES VISUALIZATION



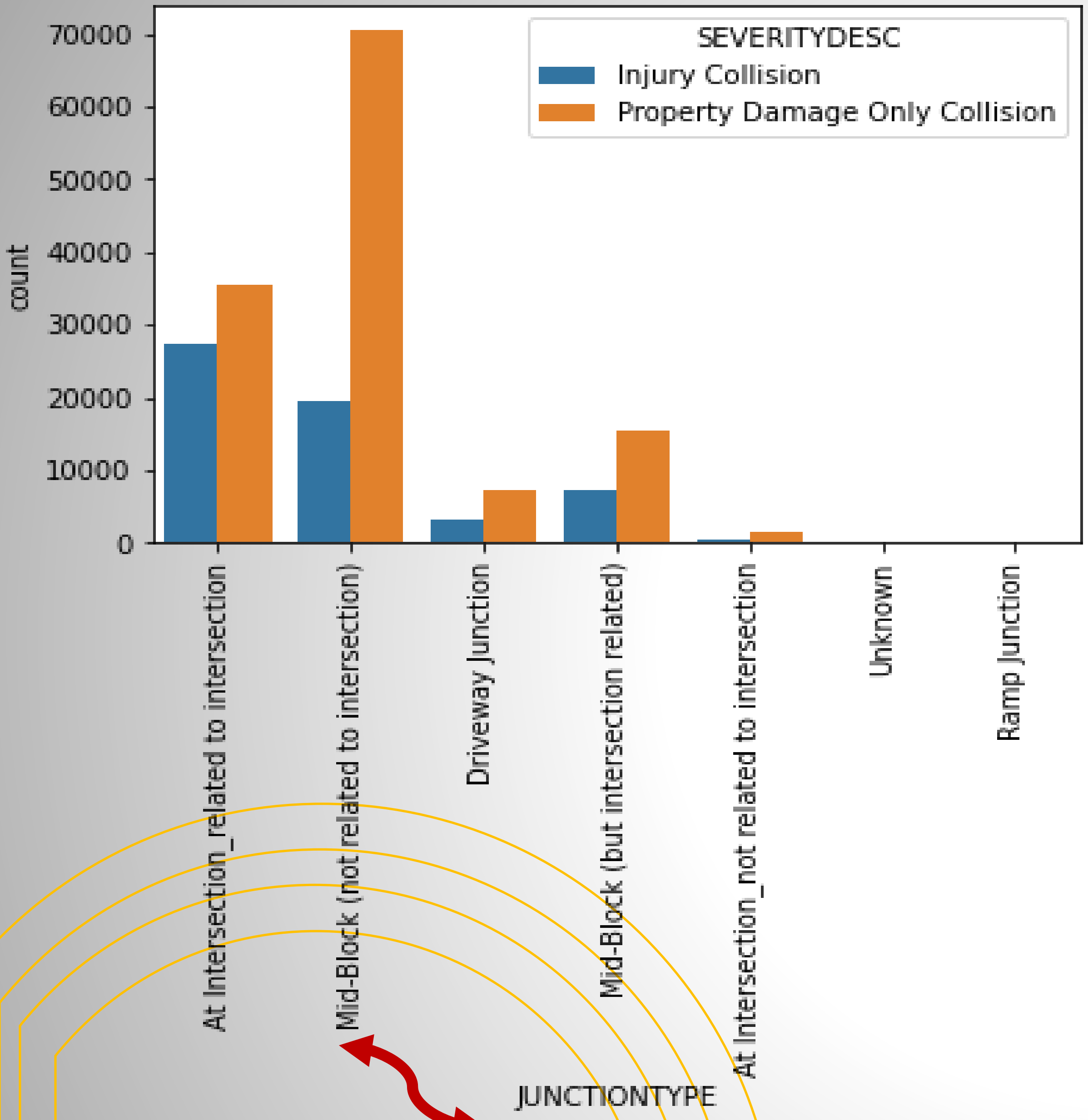
CATEGORICAL FEATURES VISUALIZATION



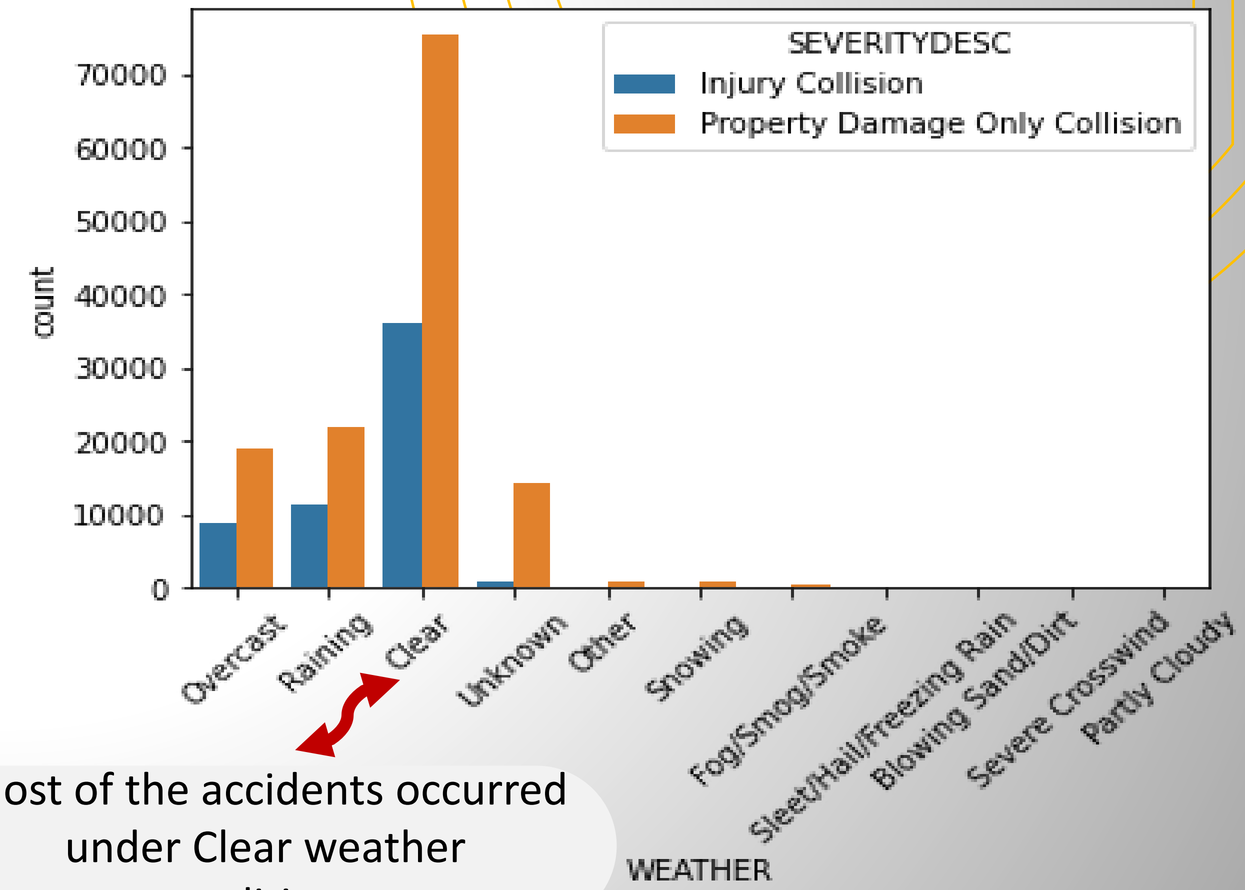
Most of the accidents that occurred involved “Property Damage Only Collision”



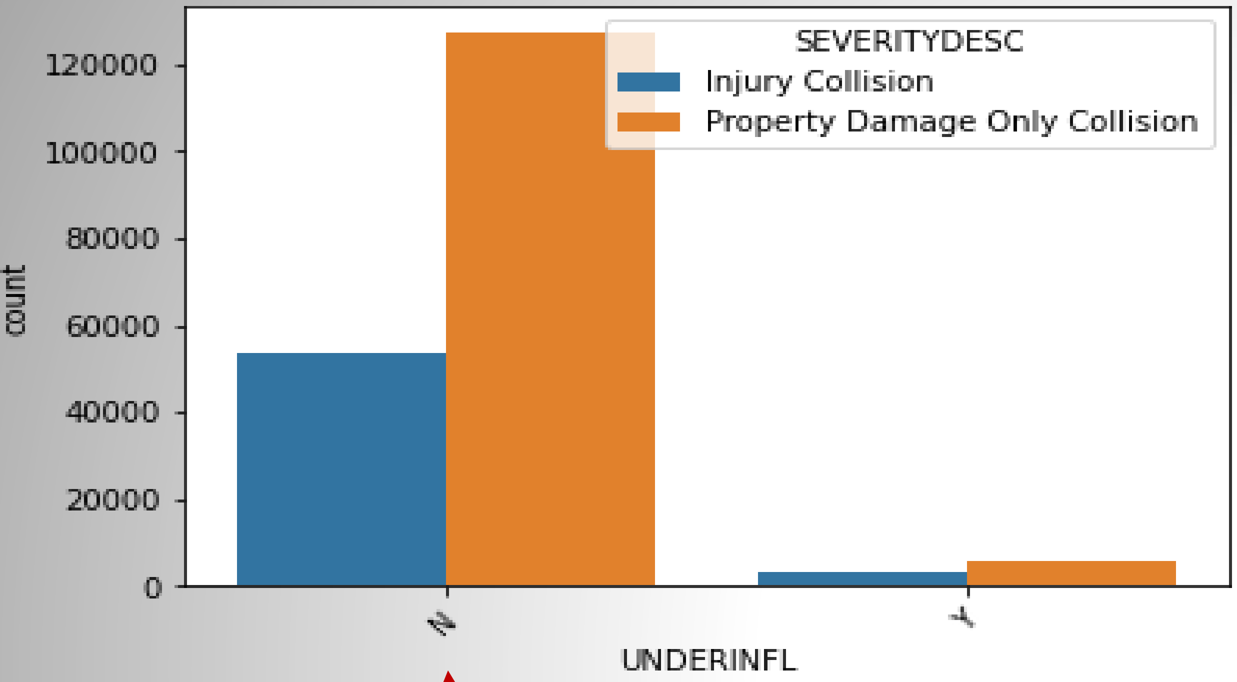
Most of the accidents happened under the Dry road conditions

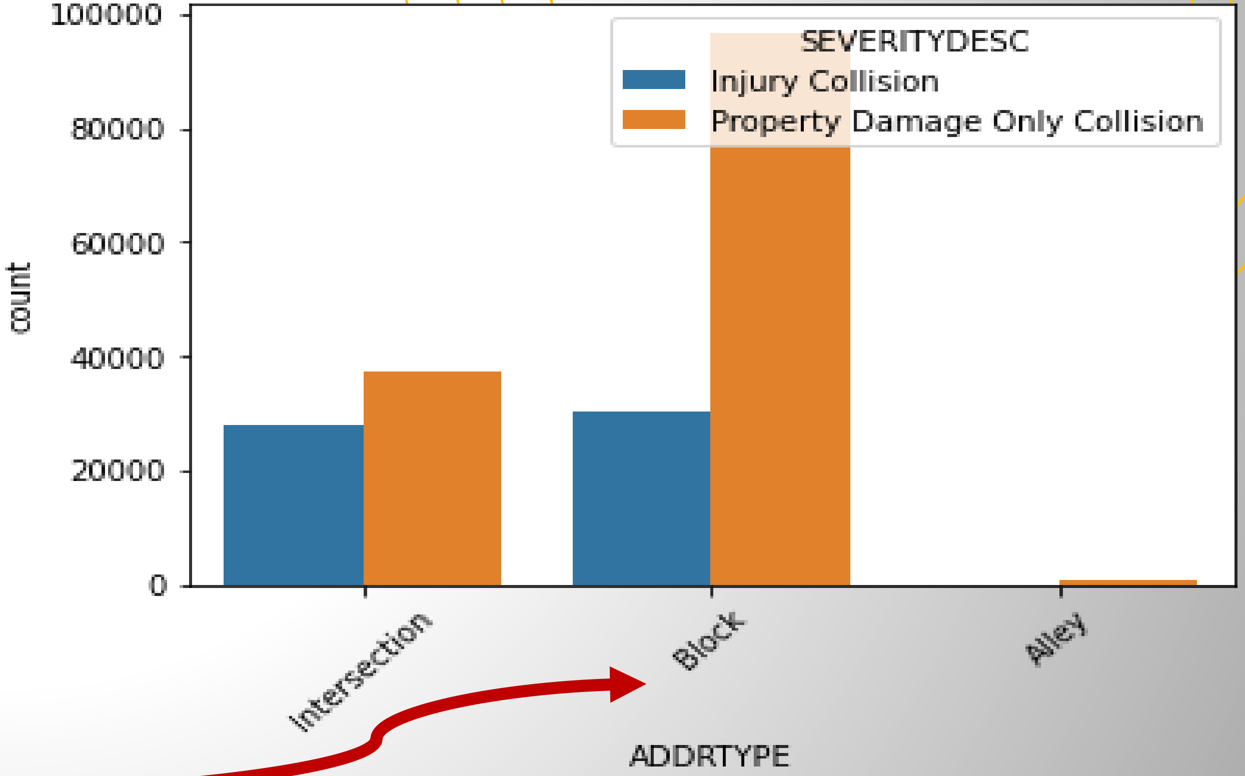


The major junction type involved was “Mid-Block”



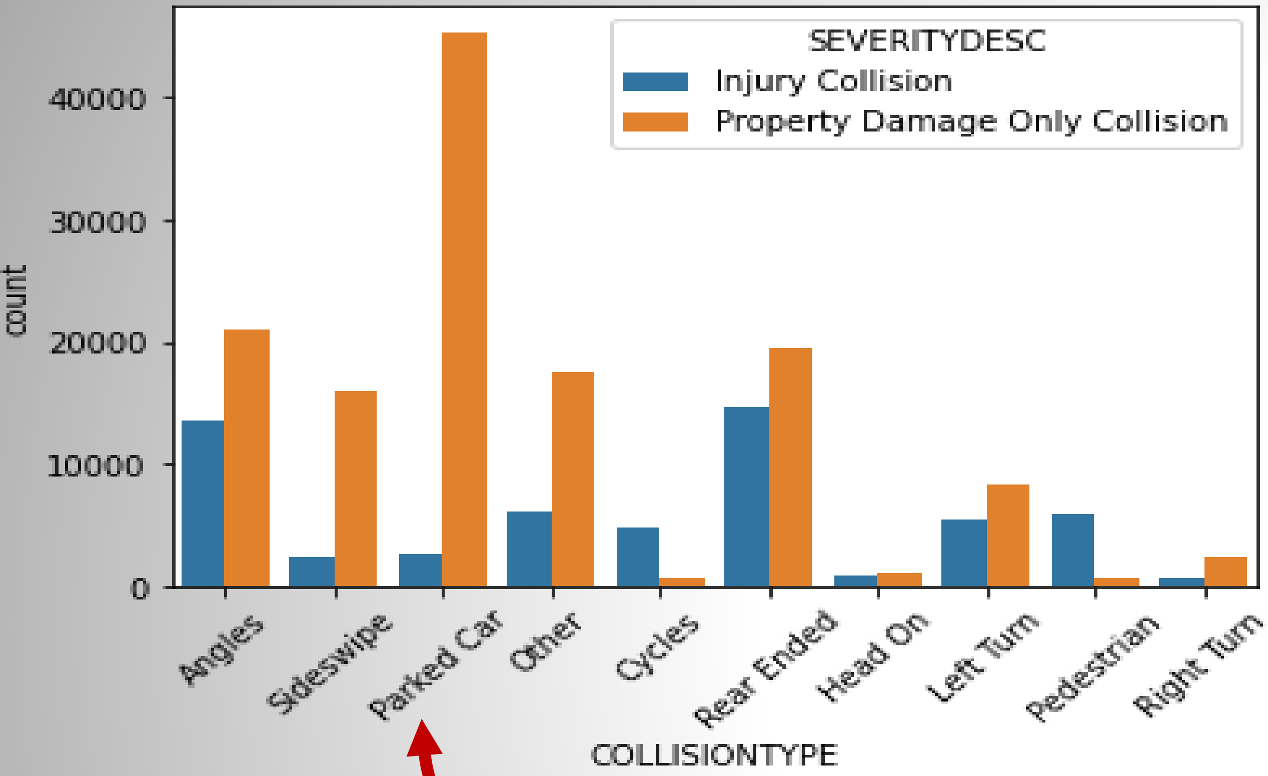
Most of the accidents occurred under Clear weather conditions.





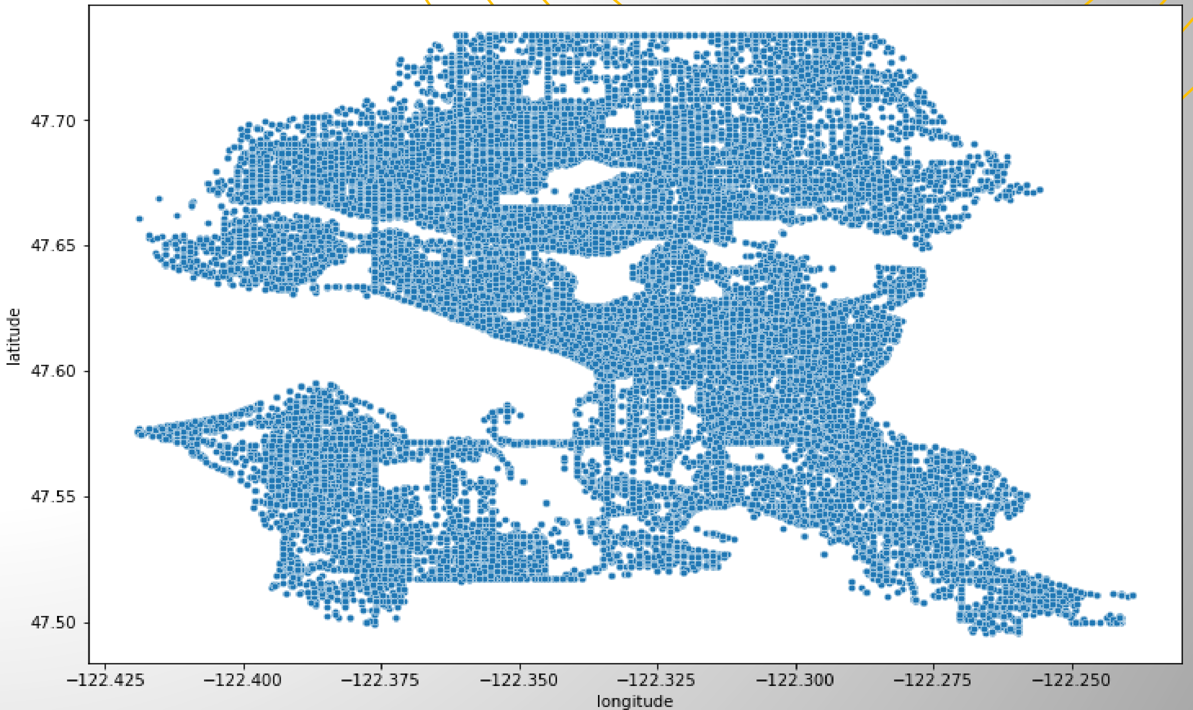
Most of the accidents occurred without the “Under influence of Alcohol/Drugs” factor

Property damage only collision has the major address type as the “Block”



The major collision type included “Parked Car”

Scatter plot of accident coordinates



3. Data Preparation & Modeling

*_After cleaning the data, features selected for Machine Learning application are mentioned in the following slide_

Selected Features for applying Machine Learning Algorithms

“SEVERITYCODE”

The Target Class(Variable)

1 longitude

2 latitude

3 PERSONCOUNT

4 PEDCOUNT

5 WEATHER

6 COLLISIONTYPE

7 LIGHTCOND

8 ROADCOND

9 ADDRTYPE

10 SDOT_COLDESC

11 HITPARKEDCAR

12 VEHCOUNT

13 PEDCYLCOUNT

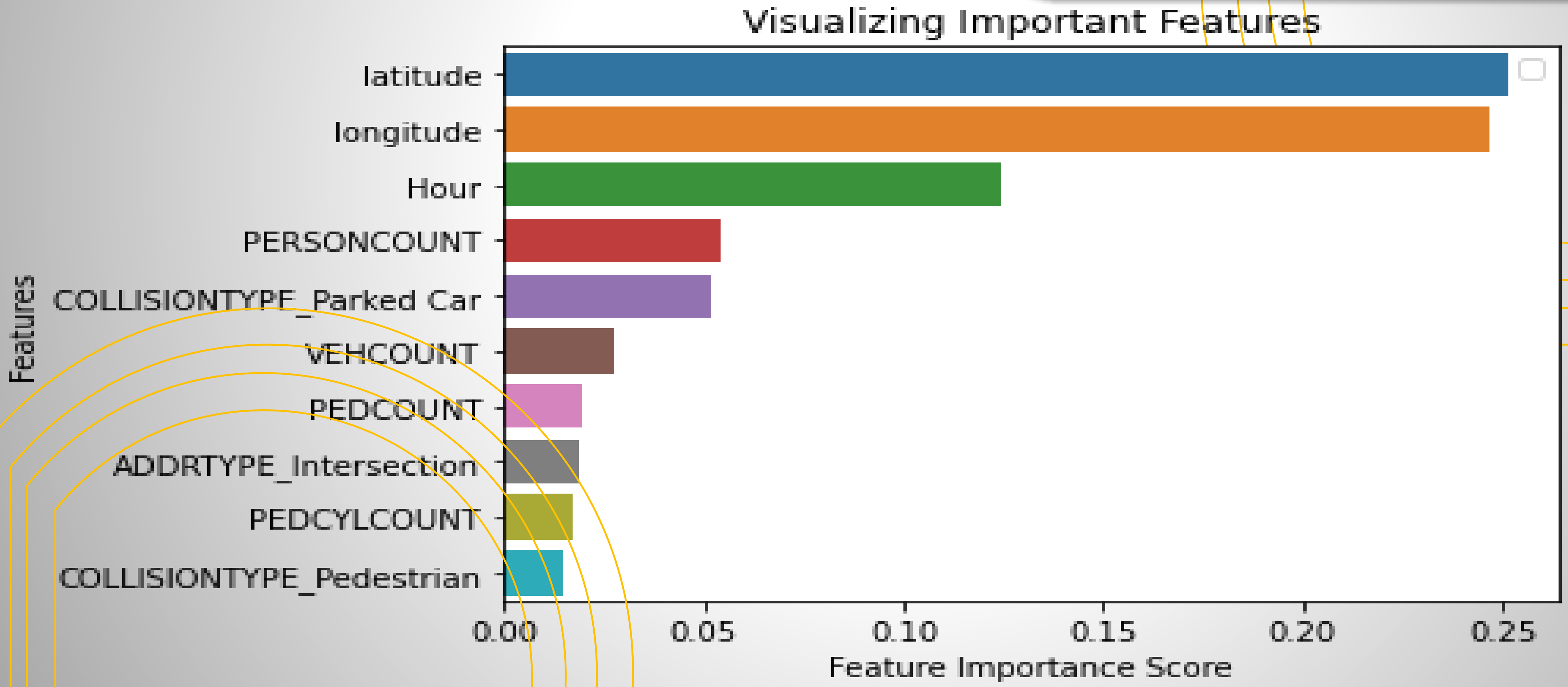
14 Hour

Machine Learning Algorithms Selection

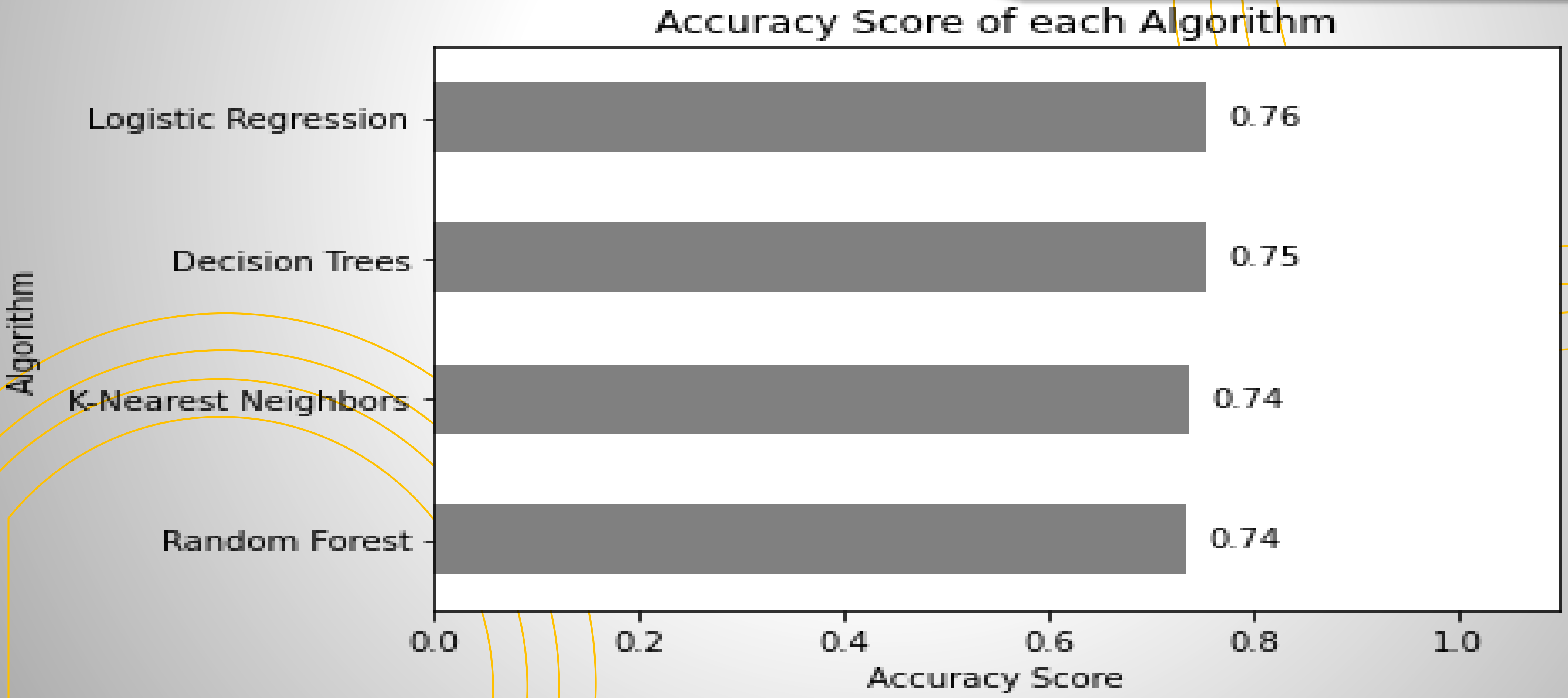
1. Logistic Regression
2. K-NN Neighbors
3. Decision Tree Algorithm
4. Random Forest Algorithm

4. Evaluation

FEATURES IMPORTANCE SCORE



ACCURACY SCORE OF ALGORITHMS



DEPLOYMENT

For the deployment phase as it can vary from project to project a simple pdf report has been generated.

SUMMARY

- Seattle road accidents data has been analysed in order to get useful insights.
- The data contains multiple attributes e.g. accident severity, collision type, coordinates of the incident, date and time of the incident, weather and road conditions, address types, no of persons injured and property damage and many other attributes.
- There are two accident severity types mentioned in the dataset i.e.
  - Property damage only collision(1)
  - Injury collision(2)
- All the mandatory Cross-industry standard process for data mining CRISP-DM phases are covered in this report which contains the following:
  - Business Understanding
  - Data Understanding
  - Data Preparation
  - Modelling
  - Evaluation
  - Deployment
- In the Modelling phase, four algorithms were selected where the target class was “accident severity”.
- Based on the predictions, “Logistic Regression” relatively performed better among the others having the accuracy percentage of approx.76%.

CONCLUSION

Based on the selected dataset(features) for this capstone project which includes mainly, coordinates, hour, person count and the collision type, it can be concluded that these particular classes have a somewhat impact on whether or not travelling along the Seattle roads could result in property damage (class 1) or injury (class 2). In this study, the technique of association rules with a large set of accident data to identify the reasons of road accidents were used. The results show that this model could provide good predictions against traffic accident with approx. 76% correct rate. It should be noted that due to the constraints of data and research condition, there are still some factors, such as engine capacity, traffic flows, gender, age of the driver, attaining the missing data etc. that are not considered in this model and can be taken into account for future study. The results of this study can be used in vehicle safety assistance driving and provide early warnings and proposals for safe driving, hence help in reducing the number of accidents.

Thank You

SPEED
SQL
STORAGE
IT
AZURE
BYOD?
BIG DATA
CLOUD
SECURITY
PATTERNS
DEVELOPERS
APPS
PLATFORM
WEB
DEVICES
MEMORY