Problem

The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours in the state of Washington while Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. The project aims to predict how severity of accidents can be reduced based on a few factors.

Stakeholder

The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

Data Wrangling

1- Data Gathering:
   Fortunately, we gave have the data gathered in a csv file and we don't have to do web-scraping or any other means of collecting data
2- Data assessment: the data is messy and there are quality issues and tidiness issues so we have to detect these issues to clean our data.
   Detecting these issues manually by eyes on data in sheets software (excel, google sheets…etc.)
   And programmatically using pandas functions (info(), sample(), head(), tail(), value_counts()…etc.)

Data assessment

- Drop columns EXCEPTRSNCODE, EXCEPTRSNDESC, REPORTNO, STATUS, INCDATE, PEDROWNOTGRNT, SEGLANEKEY, CROSSWALKKEY, duplicate SEVERITYCODE, INCKEY, COLDETKEY, SDOT_COLCODE, ST_COLCODE, SDOTCOLNUM, OBJECTID (Tidiness issue)

- Encoding UNDERINFL to be Y/N or 0,1 (Quality issue)

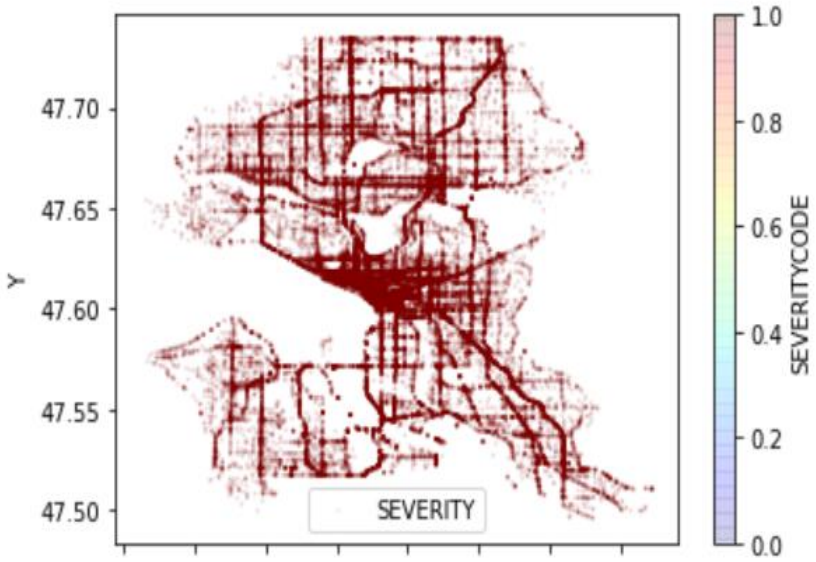
- Encoding in attention (0 = No, 1 = Yes) (Quality issue)

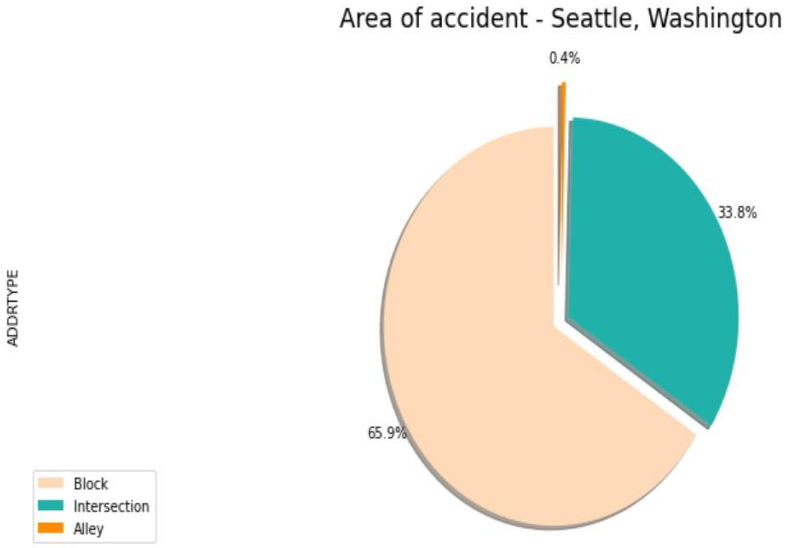
- Encoding Under the influence (0 = No, 1 = Yes) (Quality issue)
- Encoding Speeding (0 = No, 1 = Yes) (Preprocessing for ML)
- Encoding Light Conditions (0 = Light, 1 = Medium, 2 = Dark) (Preprocessing for ML)
- Encoding Weather Conditions(0 = Clear, 1 = Overcast and Cloudy, 2 = Windy, 3 = Rain and Snow Encoding Road Conditions(0 = Dry, 1 = Mushy, 2 = Wet) (Preprocessing for ML)
- Change type of INCDTTM to datetime (Quality)
- Rename Columns to proper names (Quality)

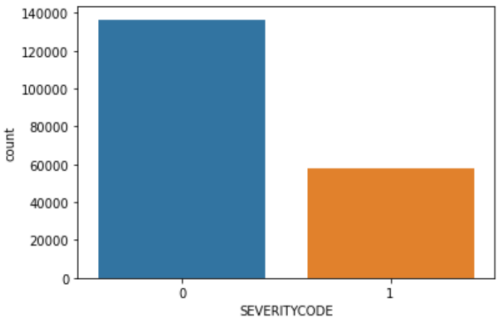
3- Data Cleaning: Fixing the above mentioned issues

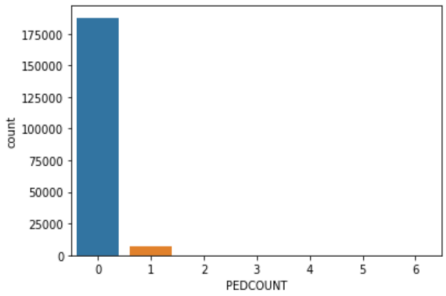
Data Analysis and Visualization

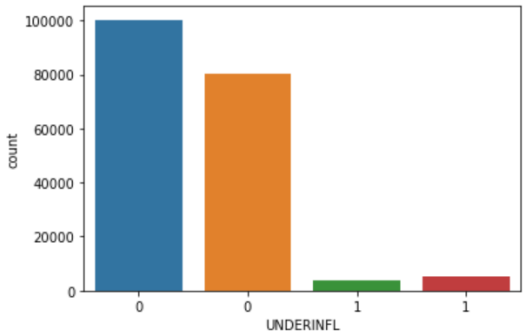
Analyzing our clean data and obtaining insights

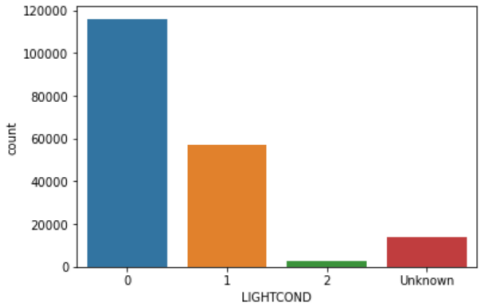












Insights

- Most of accidents cause car damage rather than human injury
- Most of accidents are vehicles crash rather than pedestrians hit
- Most of accidents happen while drivers where sober
- Most of accidents happen in a dry road
- Most of accidents happen in a clear weather
- Most of accidents happen in day-light
- Most accidents happen in block area

Machine Learning Model Selection

The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance). The reason why Decision Tree Analysis, Logistic Regression and k-Nearest Neighbor classification methods were chosen is because the Support Vector Machine (SVM) model is inaccurate for large data sets, while this data set has more than 180,000 rows filled with data. Furthermore, SVM works best with dataset filled with text and images.

Results

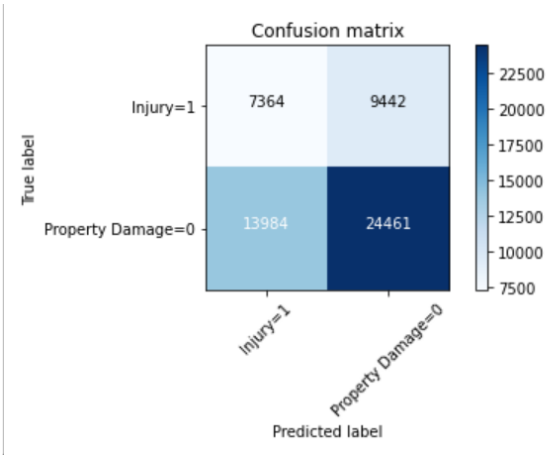
1. **Decision Tree Analysis**

Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘6’. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

Classification Report

	Precision	Recall	F1-score
0	0.64	0.72	0.68
1	0.44	0.34	0.39
Accuracy	0.85		
Macro Avg	0.54	0.53	0.53
Weighted Avg	0.56	0.58	0.56

Confusion Matrix



2.Logisitic Regression

Logistic Regression from the scikit-learn library was used to run the Logistic Regression Classification model on the Car Accident Severity data. The C used for regularization strength was '0.01'whereas the solver used was 'liblinear'. The post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier.

Classification Report

	Precision	Recall	F1-score
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.6

Confusion Matrix