Applied Data Science Capstone Project

Cars Collision Project for Seattle, Washington, USA

Background and Introduction to problem

- 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.

- As of 2020, Seattle has a total metro area population of 3.4 million (www.macrotrends.net). The total number of personal vehicles in Seattle in the year 2016 hit a new high of nearly 444,000 vehicles.

- The project aims to predict how severity of accidents can be reduced based on a few factors.

Key stakeholders that will benefit from this project:

- The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors

- Car drivers themselves who may take precaution to reduce the severity of accidents

Understanding the Data

- The data used is the one provided by coursera as an example data for this project. The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding car accidents the severity of each car accidents along with the time and conditions under which each accident occurred. The data set used for this project can be found here.

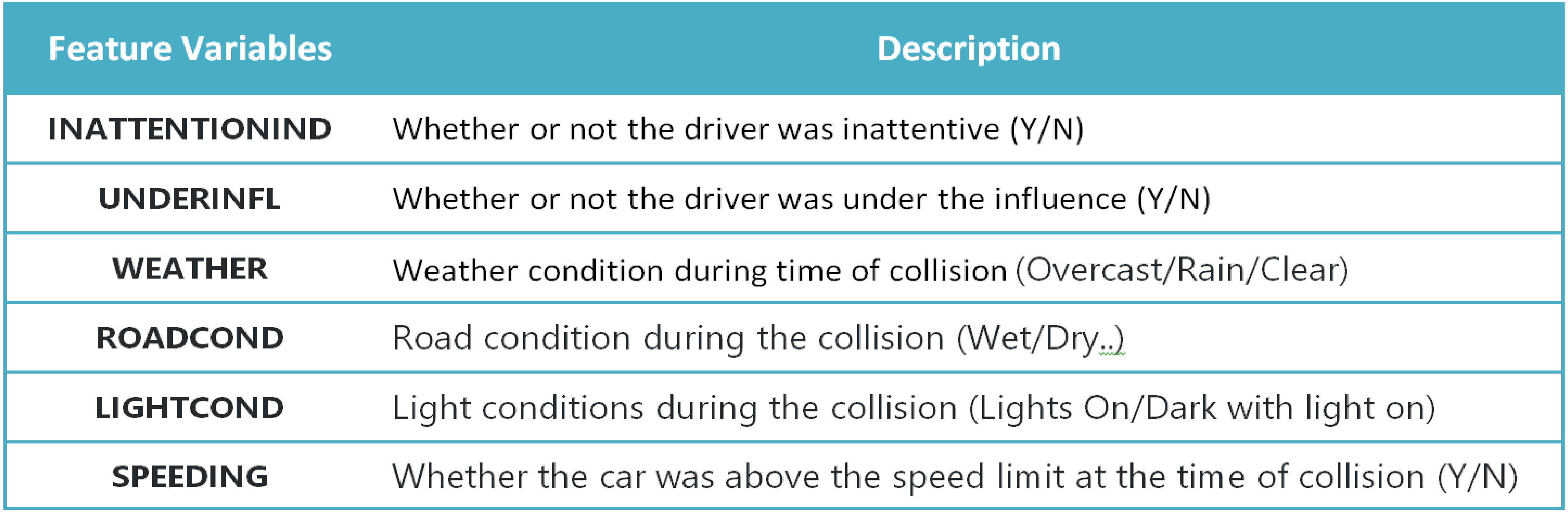
- There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature

- The models aim was to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Injury Collision) which were encoded to the form of 0 (Property Damage Only) and 1 (Injury Collision).
- Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and Under the influence. For lighting condition, Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was assigned 0, Mushy was assigned 1 and Wet was given 2.
- In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column.

- This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

Feature Selection

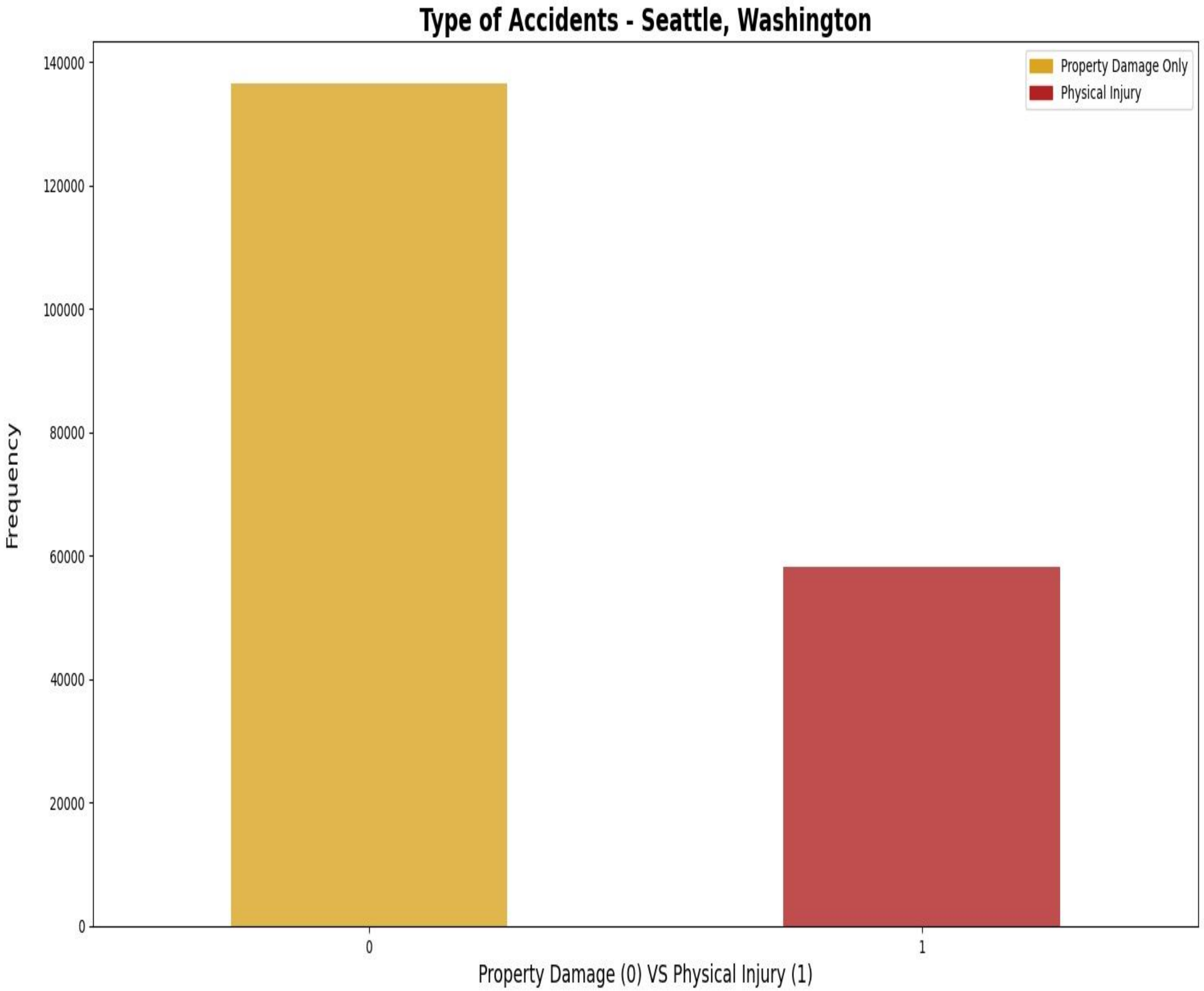
A total of 5 features were selected for this project along with the target variable being Severity Code

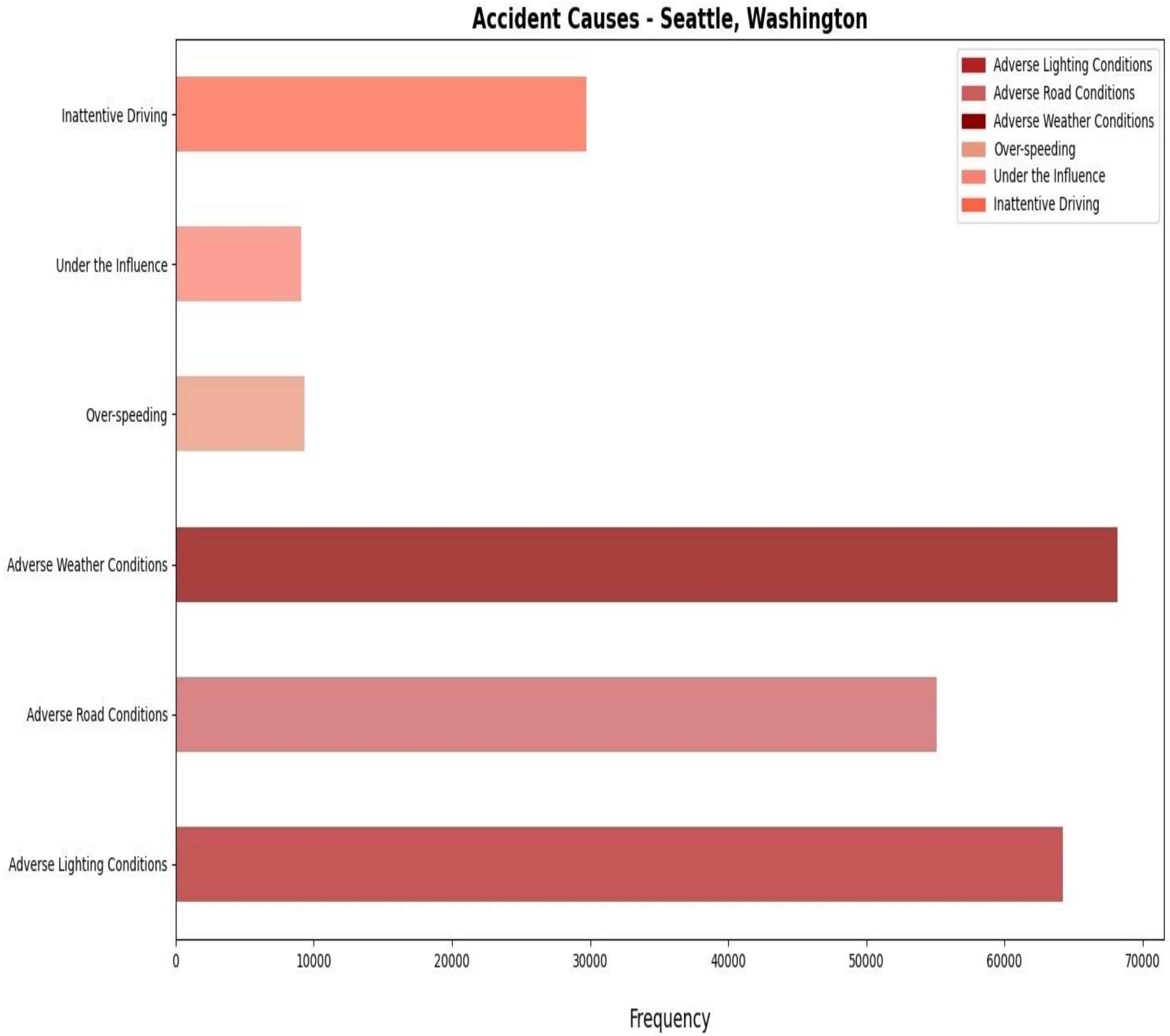


Feature Variables	Description
INATTENTIONIND	Whether or not the driver was inattentive (Y/N)
UNDERINFL	Whether or not the driver was under the influence (Y/N)
WEATHER	Weather condition during time of collision (Overcast/Rain/Clear)
ROADCOND	Road condition during the collision (Wet/Dry..)
LIGHTCOND	Light conditions during the collision (Lights On/Dark with light on)
SPEEDING	Whether the car was above the speed limit at the time of collision (Y/N)

Data Exploration: To develop deeper understanding of the data

Property damage caused by the accidents are more than the ones which caused physical injuries. Adverse Weather conditions followed by adverse lighting and adverse road conditions are the top 3 reasons for causing accidents

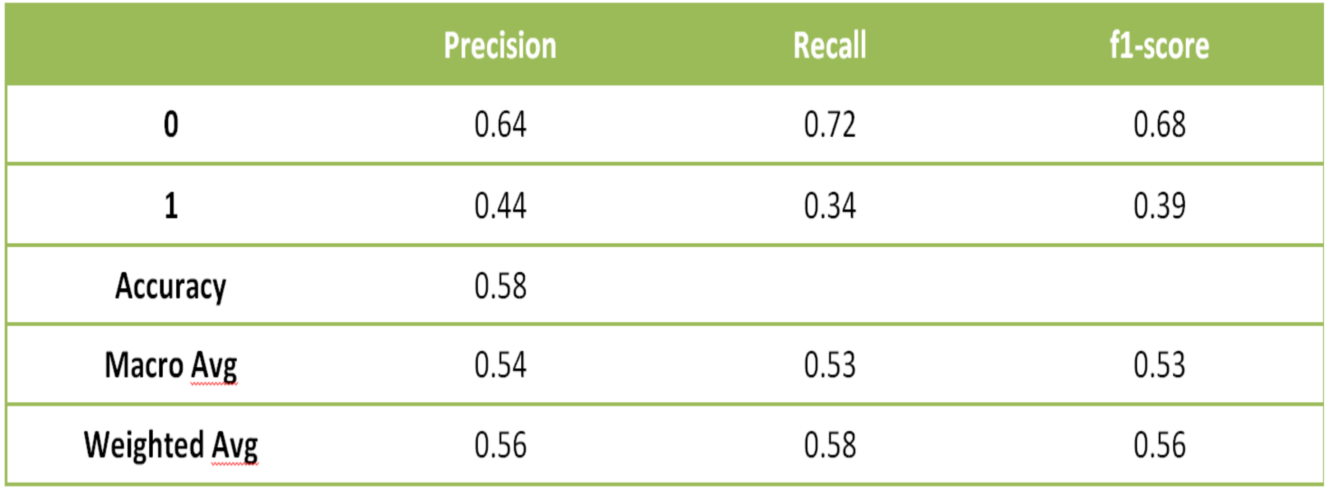




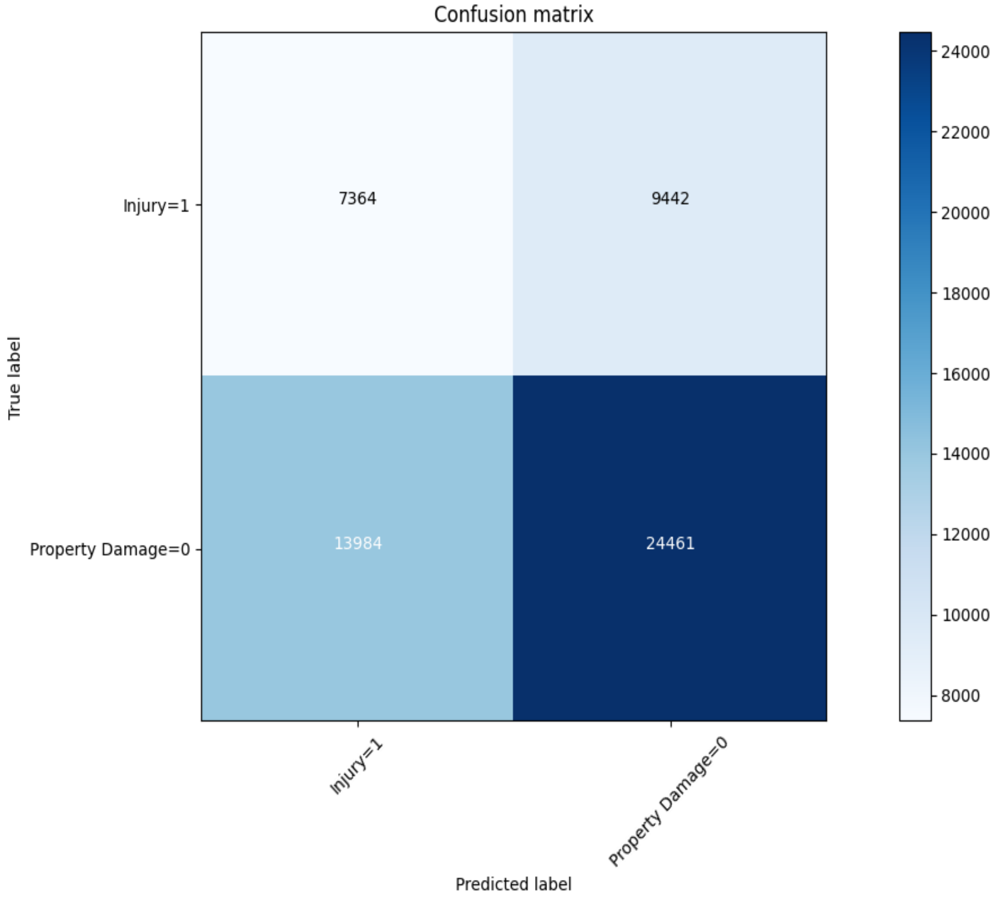
Machine Learning Model for the project (1/3)

The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor.

Decision Tree Analysis



	Precision	Recall	f1-score
0	0.64	0.72	0.68
1	0.44	0.34	0.39
Accuracy	0.58		
Macro Avg	0.54	0.53	0.53
Weighted Avg	0.56	0.58	0.56



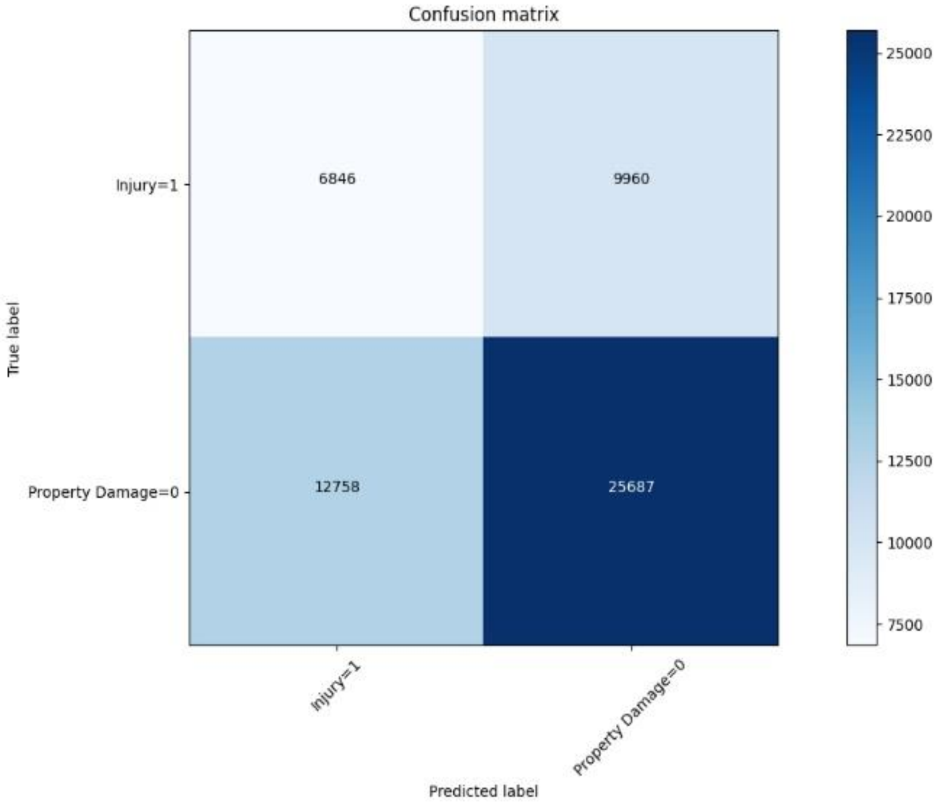
The criterion chosen for the classifier was ‘entropy’ and the max depth was ‘6’. The post-SMOTE balanced data was used to predict and fit the Decision Tree Classifier.

Machine Learning Model for the project (2/3)

The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor.

Logistic Regression

	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	0.60
Log Loss	0.68		



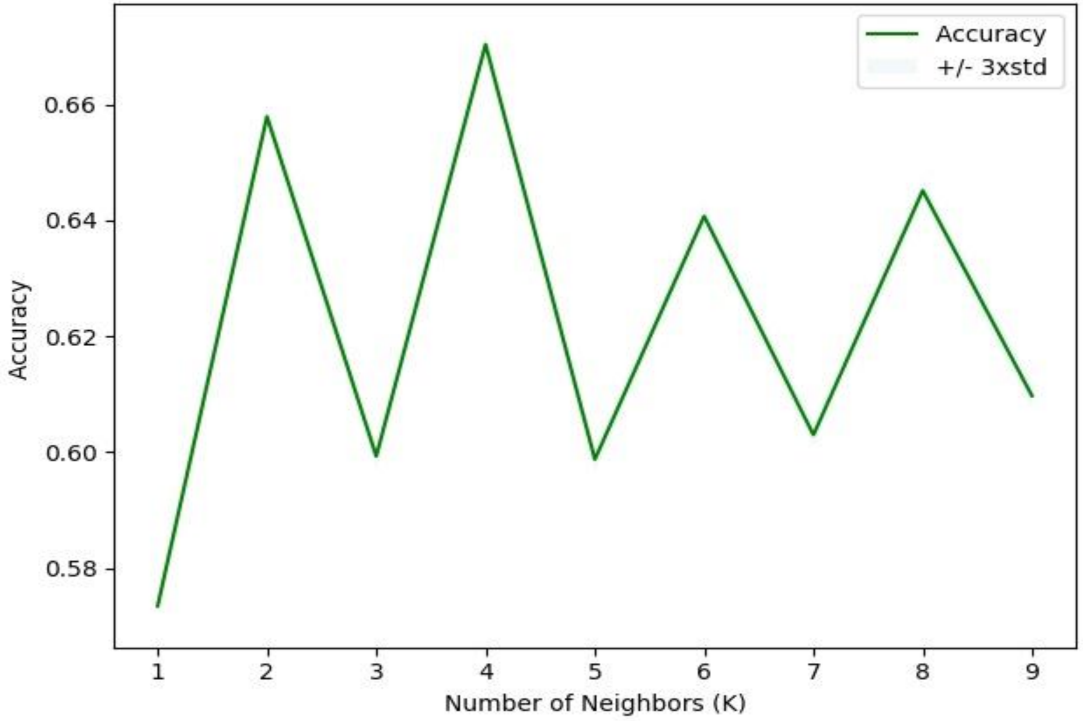
The C used for regularization strength was ‘0.01’ whereas the solver used was ‘liblinear’. The post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier.

Machine Learning Model for the project (3/3)

The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor.

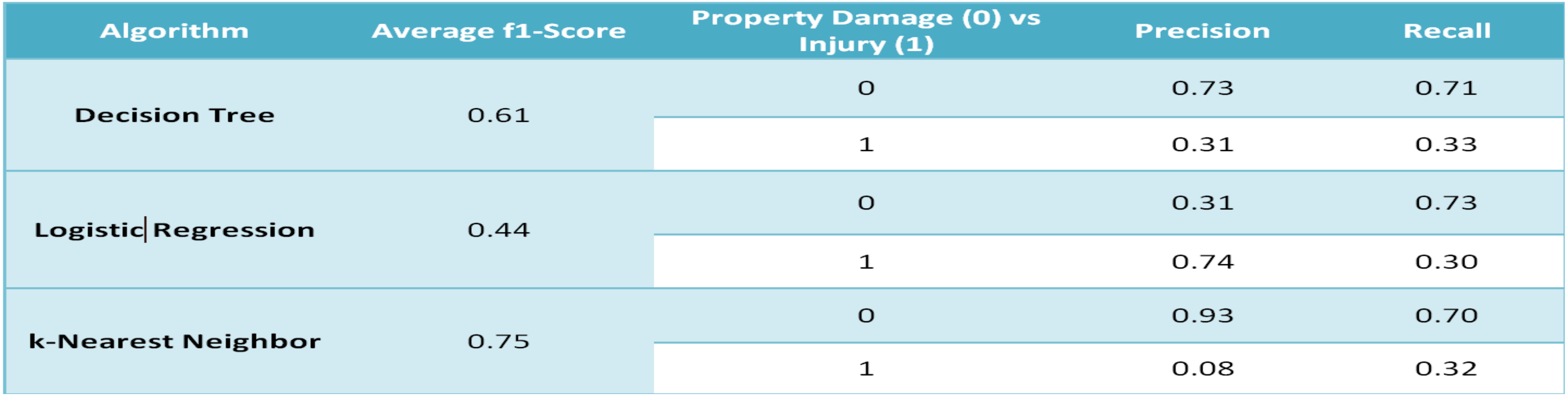
k-Nearest Neighbor

	Precision	Recall	f1-score
0	0.93	0.70	0.80
1	0.08	0.32	0.13
Accuracy	0.67		
Macro Avg	0.50	0.51	0.46
Weighted Avg	0.86	0.67	0.75



The best K, as shown below, for the model where the highest elbow bend exists is at 4. The post-SMOTE balanced data was used to predict and fit the k-Nearest Neighbor classifier

Discussion and Conclusion



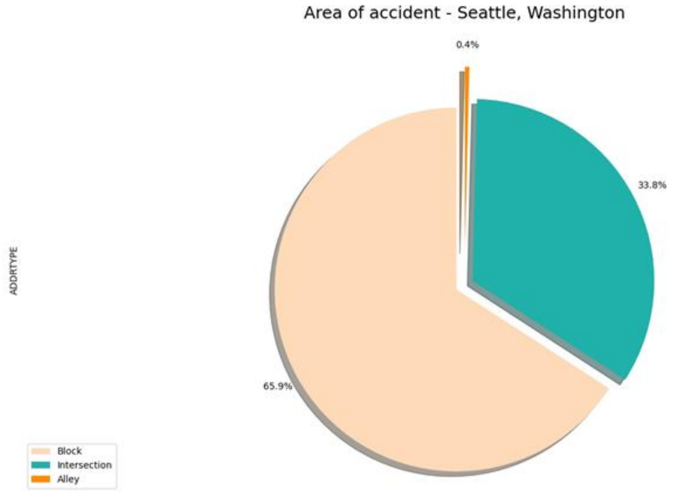
Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.61	0	0.73	0.71
		1	0.31	0.33
Logistic Regression	0.44	0	0.31	0.73
		1	0.74	0.30
k-Nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

When comparing all the models by their f1-scores, Precision and Recall, we can have a clearer picture in terms of the accuracy of the three models individually as a whole and how well they perform for each output of the target variable. When comparing these scores, we can see that the f1-score is highest for k-Nearest Neighbor at 0.75. However, later when we compare the precision and recall for each of the model, we can see that the k-Nearest Neighbor model performs poorly in the precision of 1 at 0.08. The variance is too high for the model to be selected as a viable option. When looking at the other two models, we can see that the Decision Tree has a more balanced precision for 0 and 1. Whereas, the Logistic Regression is more balanced when it comes to recall of 0 and 1. Furthermore, the average f1-score of the two models are very close but for the Logistic Regression it is higher by 0.04.

It can be concluded that the both the models (Logistic regression and Decision Tree) can be used side by side for the best performance.

Recommendations

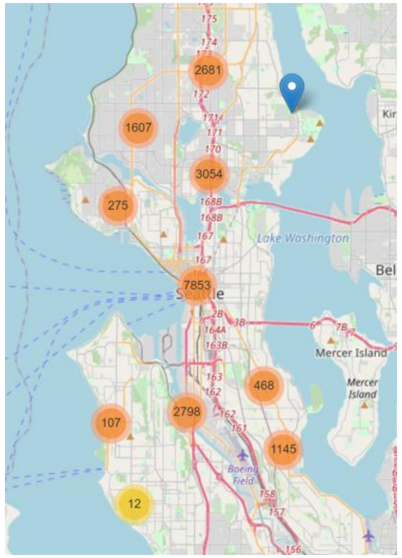
Public Development Authority of Seattle (PDAS)



Almost all of the accidents recorded have occurred on either a block or an intersection, the PDAS can take the following measures in response car accidents:

- Launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors
- Increased investment towards improving lighting and road conditions of the area which have high instances recorded
- Install safety signs on the roads and ensure that all precautions are being taken by people within the area

Car Drivers



A higher concentration of accidents can be mostly seen on the main roads of the city, specifically near the highway in the city center. The following steps can be taken by car drivers to avoid severe accidents:

- Be extra careful around the I-5 highway which goes through the city center since it has the highest proportion of accidents recorded of total seattle
- Most incidents occur under adverse weather, road and light conditions. Precautions should be taken under such circumstances, for e.g. driving slow on a wet road which may lead to loss of control

Thank You