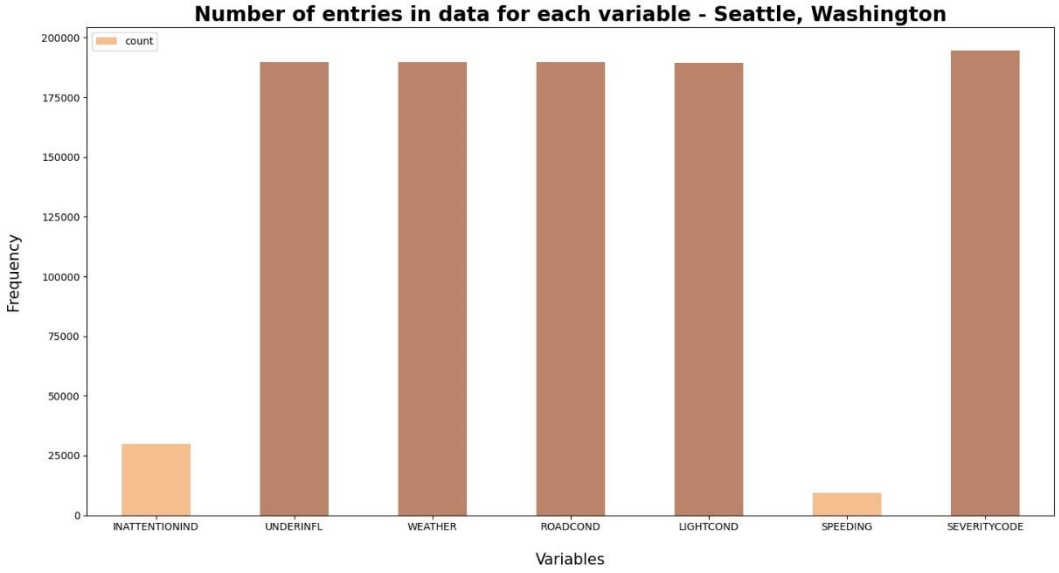
Data

The dataset used for this project is based on car accidents which have taken place within the city of *Seattle*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred. The model aims 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* (Physical Injury) which were encoded to the form of *0* (Property Damage Only) and *1* (Physical Injury). Following that, *0* was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either *Other* or *Unknown*, deleting those rows entirely would have led to a lot of loss of data which is not preferred.



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. Then the arrays were imposed on the original columns in the positions which had *Other* and *Unknown* in them. 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.

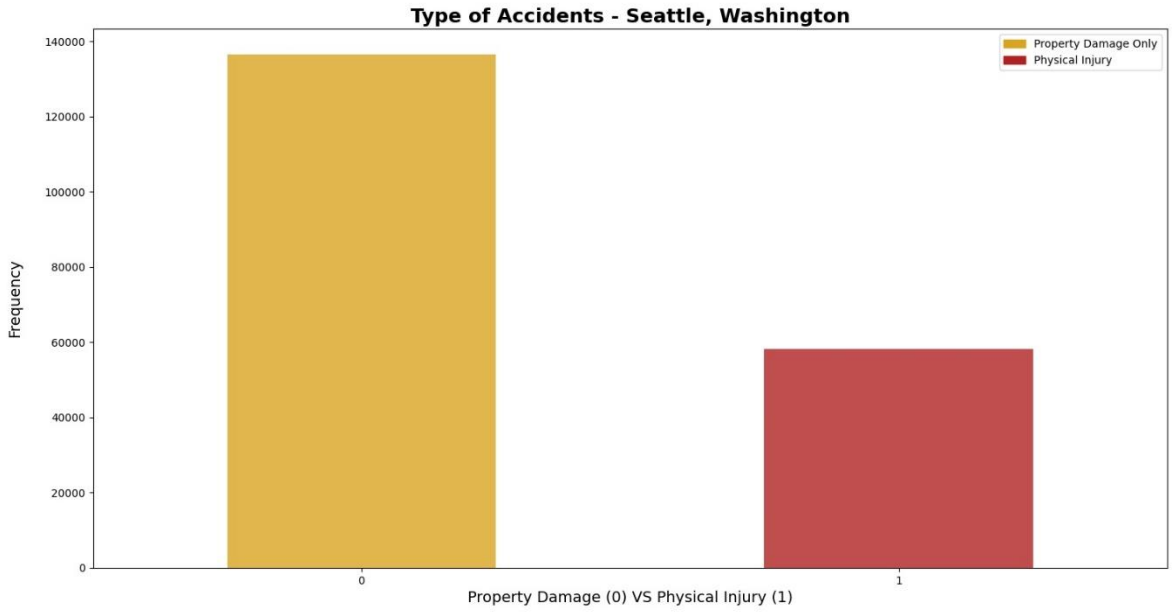
Feature Selection

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)

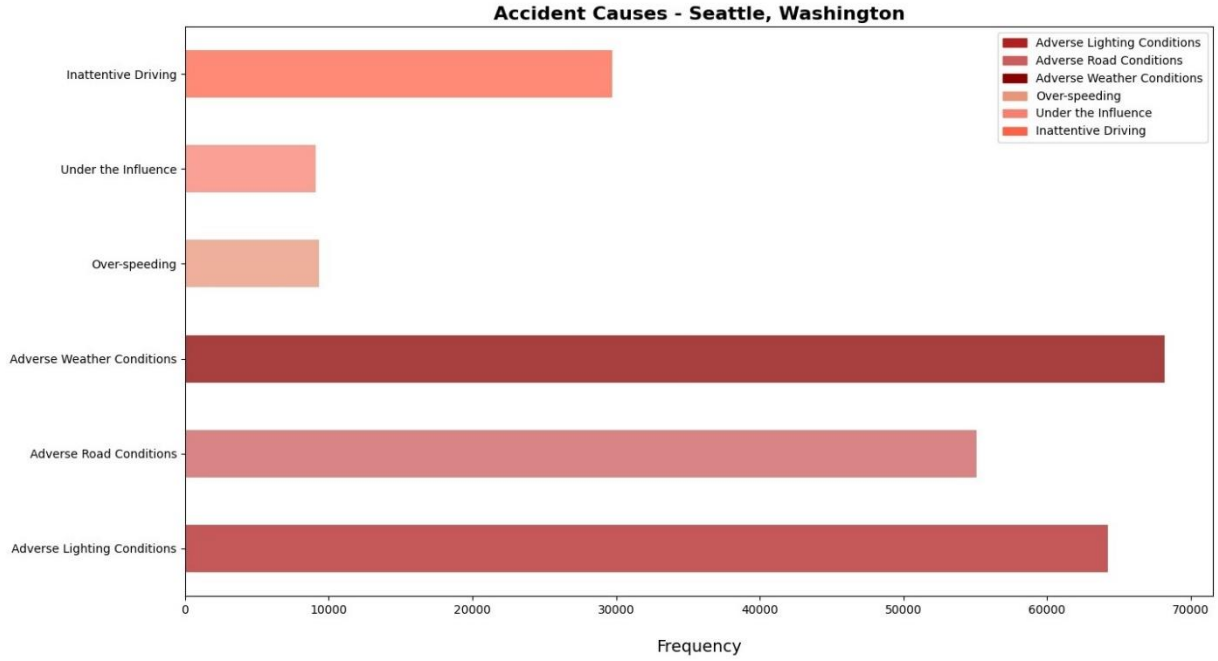
Methodology

Exploratory Analysis

Considering that the feature set and the target variable are categorical variables with the likes of weather, road condition and light condition being an above *level 2* categorical variables whose values are limited and usually based on a particular finite group whose correlation might depict a different image then what it actually is. Generally, considering the effect of these variables in car accidents are important hence these variables were selected. A few pictorial depictions of the dataset were made in order to better understand the data.



The above figure illustrates, after data cleaning has taken place, the distribution of the target variables between Physical Injury and Property Damage Only. As it can be seen that the dataset is *supervised* but an *unbalanced* dataset where the distribution of the target variable is in almost 1:2 ratio in favor of property damage. It is very important to have a balanced dataset when using machine learning algorithms. Hence, SMOTE was used from imblearn library in order to balance the target variable in equal proportions in order to have an unbiased classification model which is trained on equal instances of both the elements under severity of accidents.



As mentioned earlier, a number *0* as an element of an independent variable is supposed to depict the least probable cause of a severe accident. The graph above is supposed to depict all the non-zero values within each independent variable of the model and can be seen as the frequency of adverse conditions under which accidents took place. The factor which had most number of accidents under adverse conditions was adverse weather conditions while adverse lighting condition had the second most number of accidents caused by it. The factors which contributed the least to an instance of an accident are over-speeding and the driver being under the influence.

Machine Learning Models chosen

- **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable
- **Decision Tree Analysis:** 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 Neighbor:** K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance)

Results

The results of each of the three models had variations among them, one worked very well at predicting the positives accurately while the other predicted the negatives better.

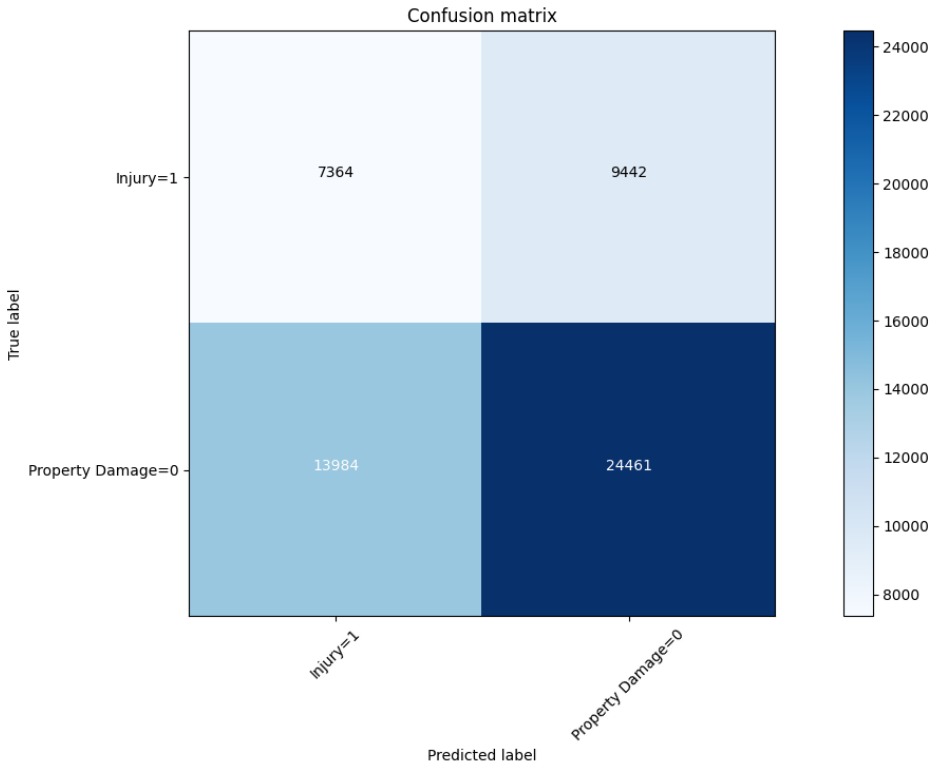
Decision Tree

The criterion chosen for the classifier was *entropy* and the max depth was *6*.

Decision Tree Classification Report

	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

Decision Tree Confusion Matrix



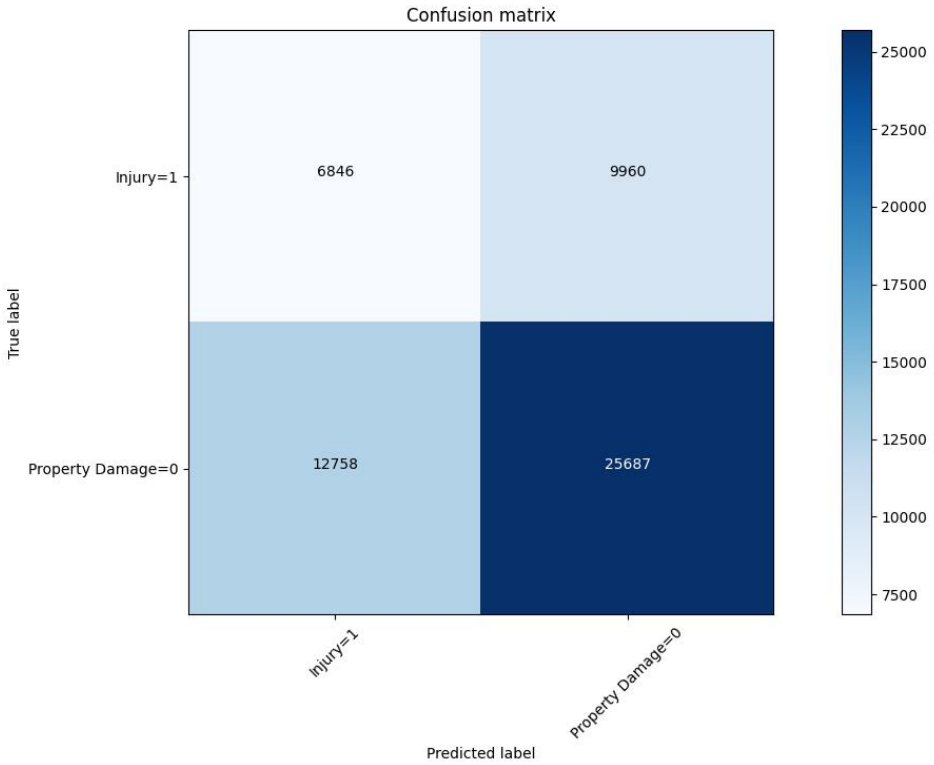
Logistic Regression

The C used for regularization strength was *0.01* whereas the solver used was *liblinear*.

Logistic Regression 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	0.60
Log Loss	0.68		

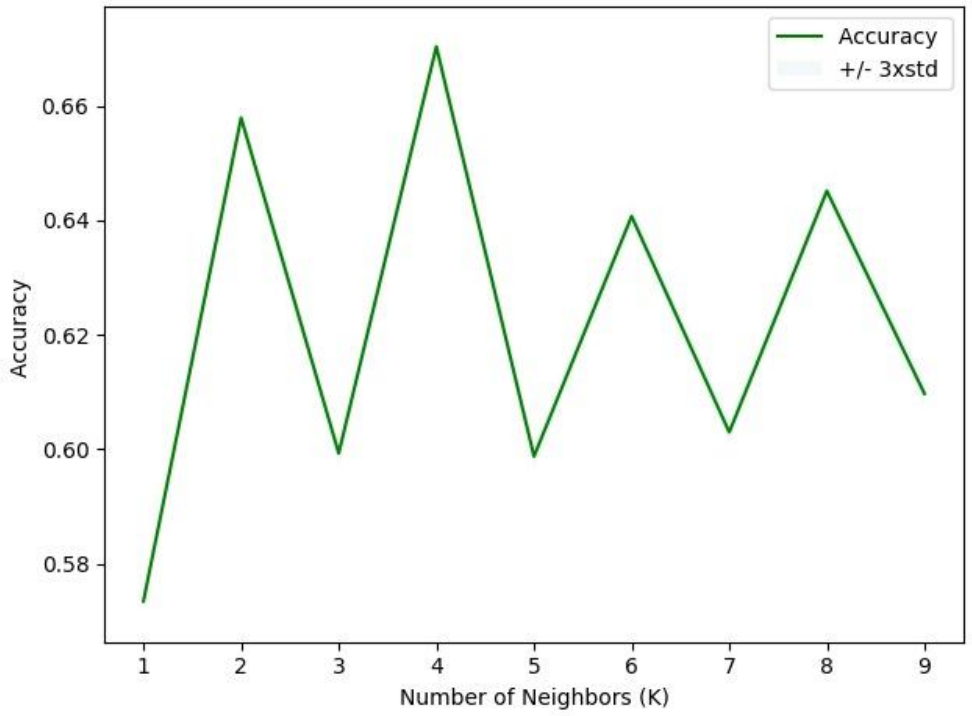
Logistic Regression Confusion Matrix



k-Nearest Neighbor

The best K, as shown below, for the model where the highest elbow bend exists is at *4*.

Choosing the best K



k-Nearest Neighbor Classification Report

	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

5. Model Accuracy

- **Precision:** Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant. Mathematically, it is calculated by dividing true positives by true positive and false positive
- **Recall:** Recall refers to the percentage of total relevant results correctly classified by the algorithm. In simpler terms, it tells how many relevant items were selected. It is calculated by dividing true positives by true positive and false negative
- **F1-Score:** f1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall is shown by

the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0

Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41
k-Nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

6. Conclusion

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 can be used side by side for the best performance.

7. Recommendations

After assessing the data and the output of the Machine Learning models, a few recommendations can be made for the stakeholders. The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area and could launch development projects for those areas where most severe accidents take place in order to minimize the effects of these two factors. Whereas, the car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.