1.Business Problem: - Car accidents are one of the major cause of unnatural death in the world and with more n more number of cars in the road every year, making the condition even worse. Every minute a person dies due to car crash, innovation in the field of road type and standard in not excelling, we are using same kind of roads and signs that were used hundreds of years before. Even though we have models to predict weather but the impact of same on road and connecting with car accidents are missing. I am resolving this big issue with my project which revolves around predicting severity or prone of having a car accident based on attributes, so the driver can be warned and to drive cautiously. My benefactor of this model would be car driver and Public Development Authority of Seattle.

2.Data: - 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. Feature to be used in this model:-
INATTENTIONIND,UNDERINFL,WEATHER,ROADCOND,LIGHTCOND,SPEEDING

3.Methodology: -

3.1.Data Collection: - Dataset used in this project is car accident data in Seattle from 2004 to 2020.This data contain the details of Car accident happened on what circumstances and under what condition.

3.2.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. The factor which had the greatest 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.

3.3.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.

4.Result: -

4.1 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.

4.1.1 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

4.2 Logistic 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.

4.2.1 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

4.3 k-Nearest Neighbor: - k-Nearest Neighbor classifier was used from the scikit-learn library to run the k-Nearest Neighbor machine learning classifier on the Car Accident Severity data. 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.

4.3.1 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. Discussion: - 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.

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. In

retrospect, when comparing these scores to the benchmarks within the industry, it can be seen that they perform well but not as good as the benchmarks. These models could have performed better if a few more things were present and possible. A balanced dataset for the target variable More instances recorded of all the accidents taken place in Seattle, Washington Less missing values within the dataset for variables such as Speeding and Under the influence More factors, such as precautionary measures taken when driving, etc.