Final Report: Week 3 Version

Introduction

Background

Seattle is Washington State's largest city, with home to a large tech industry with Microsoft and Amazon headquartered in its metropolitan area. As of 2020, it 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. In one South Lake Union census tract, the car population has more than doubled since 2010 (www.seattletimes.com). The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability. Worldwide, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads and an additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.

Business Problem

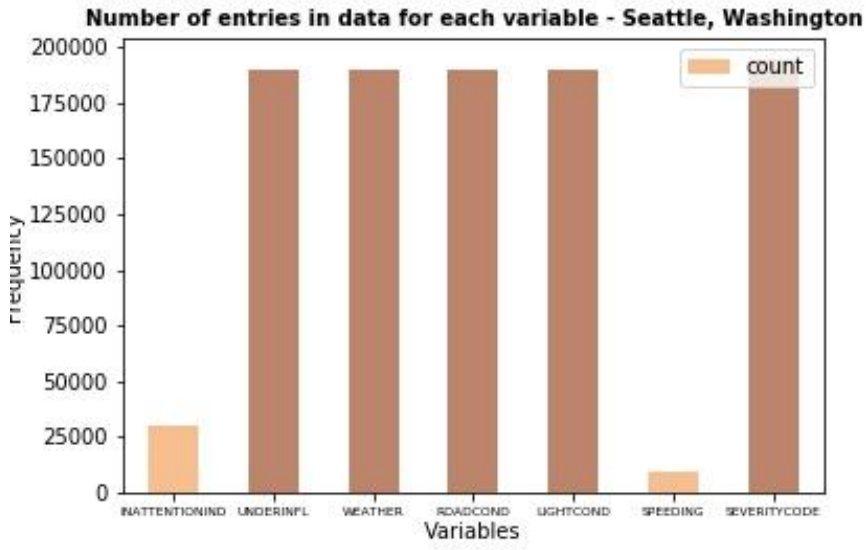
The world as a whole suffers due to car accidents, including the USA. National Highway Traffic Safety Administration of the USA suggests that the economical and societal harm from car accidents can cost up to $871 billion in a single year. According to 2017 WSDOT data, a car accident occurs every 4 minutes 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.

Understanding Data

Data Cleaning

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. First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there. These columns included pedestrian granted way or not, segment lane key, cross walk key and hit parked car.

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. As for Weather Condition, 0 is Clear, Overcast is 1, Windy is 2 and Rain and Snow was given 3. 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

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

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)

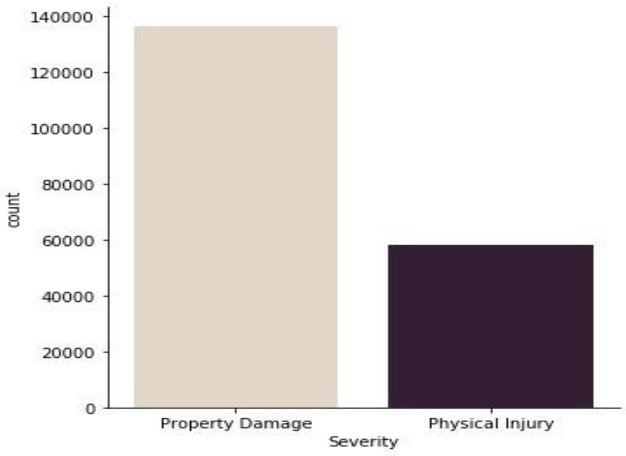
Methodology

Data Collection

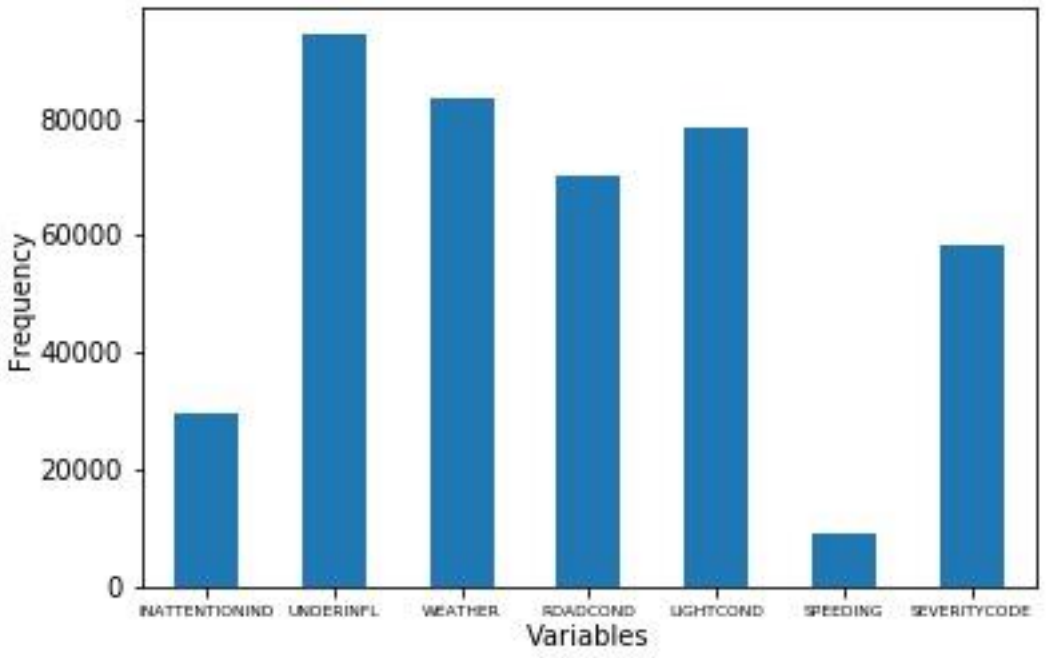
The dataset used for this project is based on car accidents which have taken place within the city of Seattle, Washington from the year 2004 to 2020. This data is regarding car accidents the severity of each car accidents along with the time and conditions under which each accident occurred.

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 Model Selection

The machine learning models used are Logistic Regression, Decision Tree Analysis. 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.

Results

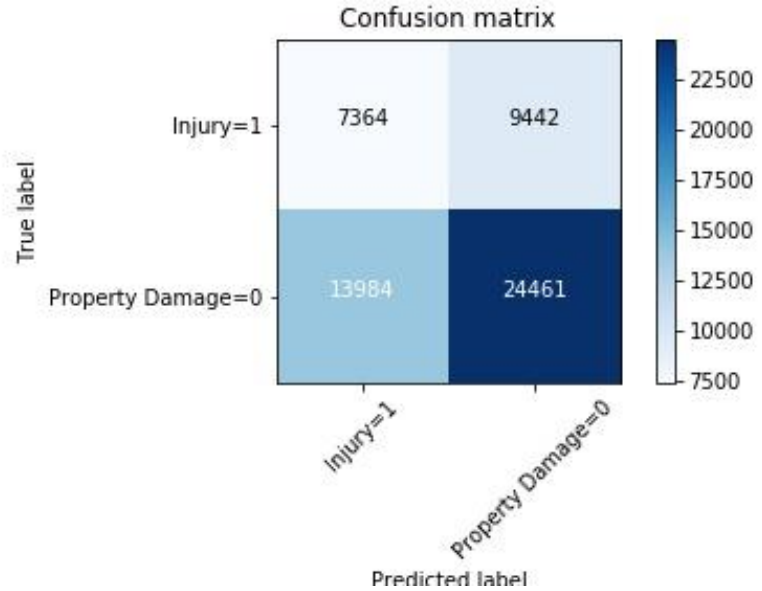
Decision Tree Analysis

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	support
0	0.64	0.72	0.68	33903
1	0.44	0.34	0.39	21348
accuracy			0.58	55251
macro avg	0.54	0.53	0.53	55251
weighted avg	0.56	0.58	0.56	55251

Confusion Matrix



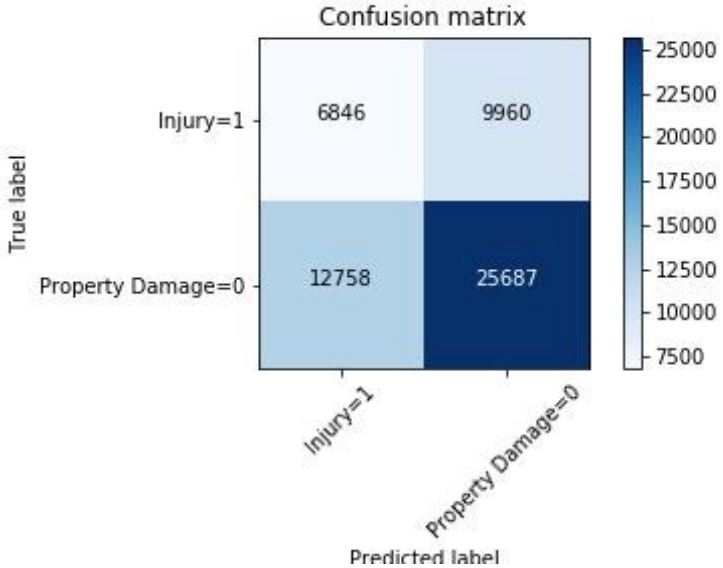
Logistic Regression

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.

Classification Report

	precision	recall	f1-score	support
0	0.72	0.67	0.69	38445
1	0.35	0.41	0.38	16806
accuracy			0.59	55251
macro avg	0.53	0.54	0.53	55251
weighted avg	0.61	0.59	0.60	55251

Confusion Matrix



Discussion

Average 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. The f1-score shown above is the average of the individual f1-scores of the two elements of the target variable i.e. Property Damage and Injury. When comparing the f1-scores of the three models, we can see that k-Nearest Neighbor has the highest f1-score meaning that it has a higher precision and recall of the other two models. Whereas, the Decision Tree model's f1-score is the lowest of the three at 0.56. Lastly, the f1-score of the Logistic Regression is at 0.60 which can be considered as an above average score. However, the average f1-score doesn't depict the true picture of the models accuracy because of the different precision and recall of the model for both the elements of the target variable. Hence, it is biased more towards the precision and recall of Property Damage due to its weightage in the model.

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. The highest precision for Property Damage is for Logistic Regression, whereas for Injury it is the Decision Tree. The Precision is calculated individually above in order to understand how accurate the model is at predicting Property Damage

and Injury individually. For the Decision Tree the precision of 0 is 0.64 and for 1 it is 0.44 which is fairly good. As for the Logistic Regression model, for 0 it is at 0.72 and for 1 it is 0.35. Lastly, for the k-Nearest Neighbor at 0 it is 0.93, which is highly accurate, however for 1 it is 0.08, extremely low. In terms of precision, the best performing model is the decision tree.

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. The highest precision for 0 is when using the k-Nearest Neighbor model at 0.70 as for 1 it is the Logistic Regression model at 0.41. The recall for both Property Damage and Injury is almost identical for the Decision Tree and k-Nearest Neighbor model. As for the Logistic Regression, the recall for Property Damage is 0.67 and for Injury it is 0.41. The recall for Property Damage and Injury is the most balanced in terms of being good for both the outputs of the target variable.

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.