CAR ACCIDENT SEVERITY ANALYSIS

Seattle, Washington

(Applied Data Science Capstone)

Submitted by : K.S.L Pavan Kumar

Linkedin:https://www.linkedin.com/in/pavankodurupk/

Github : https://github.com/pavankoduru

Table of Contents

1. Introduction

1.1 Background

Seattle, also known as the Emerald city, 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.

1.2 Problem

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring

accidents. Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

1.3 Stakeholders

The target audience or stakeholders of the project is local Seattle government, police, rescue groups, and car insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

2. Understanding Data

2.1 Overview

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present.

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Little to no Probability (Clear Conditions)

1: Very Low Probability — Chance or Property Damage

2: Low Probability — Chance of Injury

3: Mild Probability — Chance of Serious Injury

4: High Probability — Chance of Fatality

2.2 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). There are some null values in some columns and we need to fill those values using with certain data filling methods like mean, median etc .

There is imbalance in the column called severity code. So, we can resample the given data to get balanced using Downsampling procedure.

2.3 Feature selection

After analyzing the data set, I have decided to focus on only four features, severity, weather conditions, road conditions, and light conditions.

Feature Variables	Description
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)
SEVERITYCODE	Accident severity condition (0,1,2,3,4)

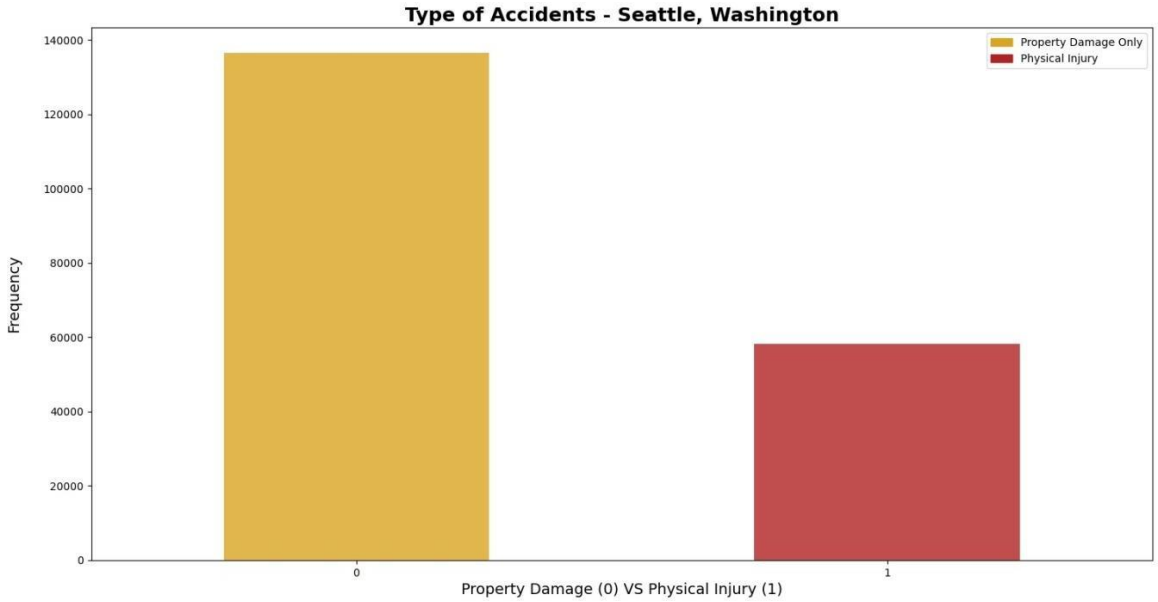
3. Methodology

3.1 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. The data set used for this project can be found here.

3.2 Exploratory Data 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, Resampling from library resample used 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.

3.3 Model Selection

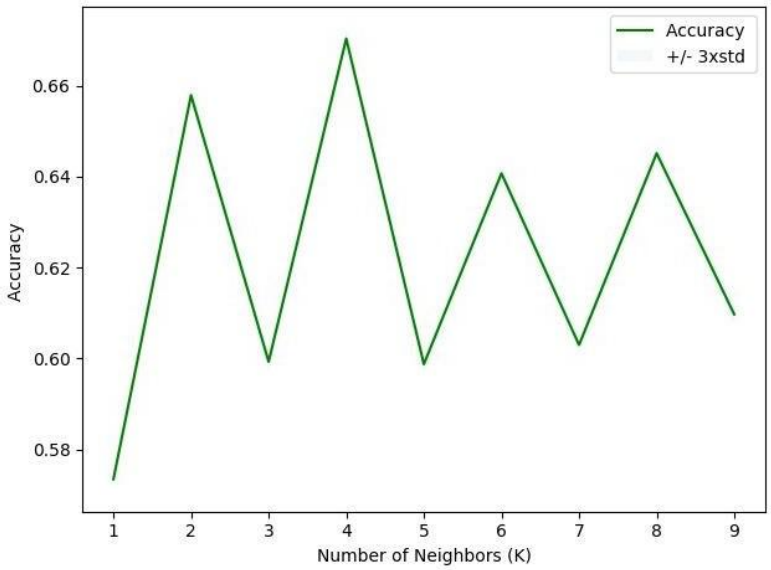
The machine learning models used are Logistic Regression, Decision Tree Analysis and k-Nearest Neighbor. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The Decision Tree Analysis breaks down a data set into smaller subsets 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. Results

4.1 K-Nearest Neighbour

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 balanced data was used to predict and fit the k-Nearest Neighbor classifier.

4.1.1 Best KNN Value



4.1.2 Accuracy scores

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.54	0.54	NA

4.2 Decision tree

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 ‘gini’ . The balanced data was used to predict and fit the Decision Tree Classifier.

4.2.1 Accuracy scores

Algorithm	Jaccard	F1-score	LogLoss
DTM	0.56	0.53	NA

4.3 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 balanced data was used to predict and fit the Logistic Regression Classifier.

4.3.1 Accuracy scores

Algorithm	Jaccard	F1-score	LogLoss
LR	0.55	0.52	0.67

5. Discussion

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.54	0.54	NA
Decision Tree	0.56	0.53	NA
LogisticRegression	0.55	0.52	0.67

5.1 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 Logistic Regression's f1-score is the lowest of the three at 0.52. Lastly, the f1-score of the Decision Tree model is at 0.53 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.

5.2 Jaccard coefficient

The Jaccard similarity index (sometimes called the Jaccard similarity *coefficient*) compares members for two sets to see which members are shared and which are distinct. It's a measure of similarity for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more

similar the two populations. Although it's easy to interpret, it is extremely sensitive to small samples sizes and may give erroneous results, especially with very small samples or data sets with missing observations. When comparing the Jaccard similarity of the three models, we can see that Decision Tree model has the highest f1-score meaning that it has a higher precision and recall of the other two models. Whereas, the k-Nearest Neighbor's f1-score is the lowest of the three at 0.54. Lastly, the f1-score of the Logistic Regression is at 0.55 which can be considered as an above average score.

6. Conclusion

When comparing all the models by their f1-scores,jaccard and logloss, 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 0.54. Moreover Decision Tree Model gives the best jaccard coefficient among other two models. 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.