CAR ACCIDENT SEVERITY REPORT

INTRODUCTION

There is a huge impact on society due to traffic accidents where there is a great cost of fatalities and injuries. In recent years, there is an increase in the researcher's attention to determine the significant effect on the severity of the injuries to the drivers which is caused due to the road accidents. Accurate and comprehensive accident records are the basis of accident analysis. The effective use of accident records depends on some factors, like the accuracy of the data, record retention, and data analysis. There are many approaches applied to this scenario to study this problem.

A recent study illustrated that the residential and shopping areas are more hazardous than rural areas. As predicted, the frequencies of the casualties were higher near the residential zones possibly because of the higher exposure. A study revealed that the casualty rates among the residential areas are classified as relatively deprived and significantly higher than those from relatively affluent areas.

Accidents have become very common these days. Nearly 1.25 million people die in road accidents every year. On an average, there are 3,287 deaths in a day. Moreover, 20 - 50 million people are injured or disabled annually. Road traffic accidents rank as the 9th leading cause of death and accounts for 2.2% of total deaths globally. In this contest of better severity of the accidents, machine learning and neural techniques have been used in analysis. These techniques are helpful to reduce accidents.

Car accidents are one of the common types of collision occurring globally everyday by analysing the different factors which cause the collision. In this section, let's discuss the data capstone project on Car accident severity.

BUSINESS UNDERSTANDING¶

Car collisions or car accidents are one of the types of road accidents. According to Corrigan [1], despite collecting large quantities of traffic data, Transportation Departments of all levels are unable to use such data to good effect. A start-up called ODN was found in 2015 which could predict when and where accidents are most likely to happen. Officials could use such information to direct safety efforts at the stretches of road where the impacts could be the biggest. In the context of this research, few of the developed countries like US, UK governments could use the information generated from a prediction system with a Neural Network predicting the accident severity and use this information to enhance the laws to build safer roads for the future. In this project, we are dealing with all the possible ways to reach the destination by overcoming car accident severity with the different critical traffic conditions on the journey. By prediction, car accident severity improves the traffic safety measures. And implements the traffic rules accordingly by governments to improve accident severity.

OBJECTIVES - Solution to the problem

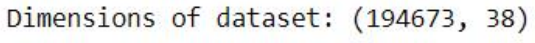
The objectives of this capstone project are mainly the following:

1. Gather a comprehensive database of road accident statistics for built up roads with factors that affect road safety which have been provided by the database.
2. Analyse data for the factors, which can impact accident rates (e.g., light conditions, weather, road surface conditions, types of junctions, etc.)
3. Determine the type of road classes with highest and lowest amount of accident rates from analysing tables of road accident statistics and charts created from the database UK-2019 accident set.
4. Suggest appropriate measures for the factors and the road class determined the most dangerous for improving car accident severity.

FEATURE SELECTION

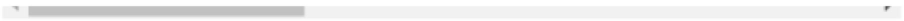
Since the study focuses on environmental conditions of the accidents, we can narrow down the dataset to ‘WEATHER’, ‘ROADCOND’, and ‘LIGHTCOND’.

We begin by importing main libraries followed by loading data file and printing the size of the dataset.



```
Dimensions of dataset: (194673, 38)
```

We can view the columns and first five rows of the dataset to get an idea of the data we are dealing with.



The target variable, ‘SEVERITYCODE’, is described by ‘SEVERITYDESC’. Let’s see how many different codes we have.

So, we have two severity codes: 1 for property damage only collision and 2 for injury collision.

We then narrow down our dataset to the features of interest, namely: ‘WEATHER’, ‘ROADCOND’, ‘LIGHTCOND’.

Handling Missing Data

The dataset consists of raw data so there is missing information. First, we will search for question marks and replace them with NANs. Then we will replace all NAN values with the most frequent data from each attribute. In addition to that, we are going to group some types of the features together if they are related to each other.

From the results above, it can be seen that we are missing 5081 weather data, 5012 road condition data, and 5170 light condition data. This missing information needs to be addressed.

Let’s also explore the different types of each feature to see if we can group them together.

Weather conditions can be grouped as follows:

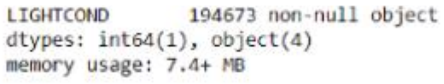
SevereWeather: Raining, Snowing, Sleet/Hail/Freezing Rain, Fog/Smog/Smoke, Blowing Sand/Dirt, Severe Crosswind
Overcast: PartlyCloudy and Overcast
Unknown: Other

Road conditions can be grouped as follows:

IceOilWaterSnow: Ice, Standing Water, Oil, Snow/Slush, Sand/Mud/Dirt
Unknown: Other

Light conditions can be grouped as follows:

Dark-No-Light: Dark — No Street Lights, Dark — Street Lights Off, Dark — Unknown Lighting
Dark-With-Light: Dark — Street Lights On
DuskDawn: Dusk, Dawn
Unknown: Other



```
LIGHTCOND        194673 non-null object
dtypes: int64(1), object(4)
memory usage: 7.4+ MB
```

Methodology

In this section of the report, exploratory data analysis, inferential statistical testing, and machine learnings used are described.

Data Visualization

Number of accidents are plotted against each environmental factor (feature) with percentage of each type of each feature to understand the impact of each factor.

First let's see the impact of **weather conditions.**

We can see from the graph above that majority of the accidents happened in clear weather. I was expecting to see more accidents in severe weather. We need more information on 'Unknown' weather conditions as the percentage should not be neglected particularly for accidents that caused property damage only.

Let’s now see the impact of **road conditions.**

We can see from the graph above that majority of the accidents happened on dry roads. I was expecting to see more accidents on wet or icy, snowy, oily roads! We also need more information on ‘Unknown’ road conditions as the percentage should not be neglected particularly for accidents that caused property damage only.

And finally let’s examine the impact of **light conditions.**

It can be seen from the graph above that majority of accidents happened during the day with daylight. This also was not as I expected! Again, we need more information on ‘Unknown’ light conditions as the percentage should not be neglected particularly for accidents that caused property damage only.

Machine Learning Model Selection

The preprocessed dataset can be split into training and test sub datasets (70% for training and 30% for testing) using the scikit learn “train_test_split” method. Since the target column (SEVERITYCODE) is categorical, a classification model is used to predict the severity of an accident. Three classification models were trained and evaluated, namely: K-Nearest Neighbor, Decision Tree, and Logistic Regression.

We will start by defining the X (independent variables) and y (dependent variable) as follows.

X data needs to be converted to numerical data to be used in the classification models. This can be achieved by using Label Encoding.

```
9        0          0          2
```

It is always better to normalize the features data.

Model

It’s time to build our models by first splitting our data into training and testing sets of 70% and 30% respectively.

K Nearest Neighbor (KNN)

KNN is used to predict the severity of an accident of an unknown dataset based on its proximity in the multi-dimensional hyperspace of the feature set to its “k” nearest neighbors, which have known outcomes. Since finding the best k is memory-consuming and time-consuming, we will use k=25 based on.

Decision Tree

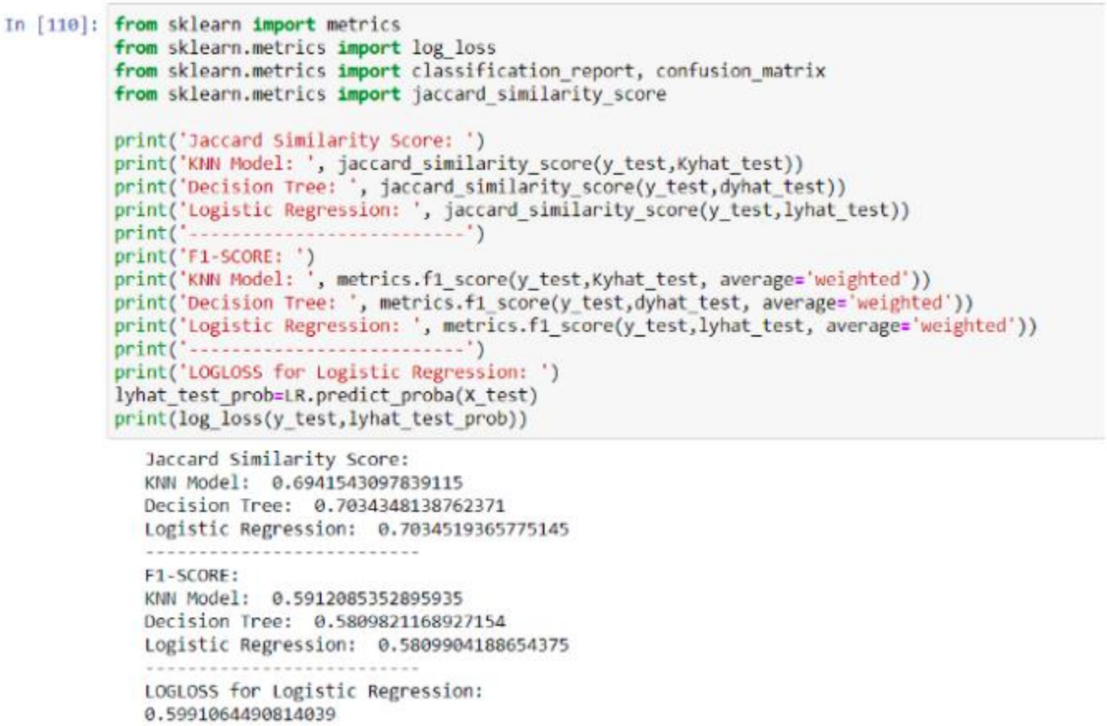
A decision tree model is built from historical data of accident severity in relationship to environmental conditions. Then the trained decision tree can be used to predict the severity of an accident. Since finding the maximum depth is also memory and time consuming, will use max_depth=30 based on.

Logistic Regression

Logistic Regression is useful when the observed dependent variable, y, is categorical. It produces a formula that predicts the probability of the class label as a function of the independent variables. An inverse-regularisation strength of C=0.01 is used as in.

Results (Model Evaluation)

Accuracy of the 3 models is calculated using these metrics: Jaccard Similarity Score, F1-SCORE, and LOGLOSS (with Linear Regression).



```
In [110]: from sklearn import metrics
          from sklearn.metrics import log_loss
          from sklearn.metrics import classification_report, confusion_matrix
          from sklearn.metrics import jaccard_similarity_score

          print('Jaccard Similarity Score: ')
          print('KNN Model: ', jaccard_similarity_score(y_test,Kyhat_test))
          print('Decision Tree: ', jaccard_similarity_score(y_test,dyhat_test))
          print('Logistic Regression: ', jaccard_similarity_score(y_test,lyhat_test))
          print('-------------------------')
          print('F1-SCORE: ')
          print('KNN Model: ', metrics.f1_score(y_test,Kyhat_test, average='weighted'))
          print('Decision Tree: ', metrics.f1_score(y_test,dyhat_test, average='weighted'))
          print('Logistic Regression: ', metrics.f1_score(y_test,lyhat_test, average='weighted'))
          print('-------------------------')
          print('LOGLOSS for Logistic Regression: ')
          lyhat_test_prob=LR.predict_proba(X_test)
          print(log_loss(y_test,lyhat_test_prob))
```

```
Jaccard Similarity Score:
KNN Model:  0.6941543097839115
Decision Tree:  0.7034348138762371
Logistic Regression:  0.7034519365775145
-------------------------
F1-SCORE:
KNN Model:  0.5912085352895935
Decision Tree:  0.5809821168927154
Logistic Regression:  0.5809904188654375
-------------------------
LOGLOSS for Logistic Regression:
0.5991064490814039
```

Discussion

First the dataset had categorical data of type ‘object’. Label encoding was used to convert categorical features to numerical values. The imbalanced data issue was ignored because there was a problem installing imbalanced-learn to use imblearn.

Once data was cleaned and analyzed, it was fed into three ML models: K-Nearest Neighbor, Decision Tree, and Logistic Regression. Values of k, max depth and inverse-regularisation strength C were taken from [6]. Evaluation metrics used to test the accuracy of the models were Jaccard Similarity Index, F-1 SCORE and LOGLOSS for Logistic Regression.

It is highly recommended to solve the data imbalance problem for more accurate results.

Conclusion

The goal of this project is to analyze historical vehicle crash data to understand the correlation of environmental conditions (weather, road surface, and lighting conditions) with accident severity. Vehicle accident data from the City of Seattle's' Police Department for the years 2004 until present were Used. The data was cleaned, and features related to environmental conditions were selected and analyzed. It was found that majority of accidents happened in clear weather, dry roads, and during daytime which wasn't what I expected. Machine learning models; K-Nearest Neighbor, Decision Tree and Logistic Regression were used to predict the severity of an accident based on certain environmental conditions. The models used were also evaluated using different accuracy metrics.