Accident Severity Predictor

CEEMA

October 17th 2020

1. Introduction

1.1 Background

Seattle is a seaport city on the West Coast of the United States. Seattle is the largest city in both the state of Washington and the Pacific Northwest region of North America. According to U.S. Census data released in 2019, the Seattle metropolitan area's population stands at 3.98 million, making it the 15th-largest in the United States. As the population is higher the traffic is also higher and so do the accidents.

1.2 Problem

With the heavy traffic, driving skills, bad weather chances of happening an accident is very high. The chaos that an accident make will really damage the life of impacted people it can even cause death. If there is a way that can let people know the chances of getting into an accident and the severity of accident they can take some precautionary measure to avoid it. This project aims to build a model that predict the severity of accident.

1.3 Interest

This can be used by people who drive in the Seattle city and they can take decisions based on the conditions on a particular day to drive on a specific route or not.

2. Data acquisition and cleaning

2.1 Data sources

The dataset contains collision data collected from 2004 to 2020 from Seattle, WA, from GISWEB. Data consist of 194763 entries with 38 columns. With accident serverity as the predicting variable, features such as date and time of incident, collision type, junction type, location, weather condition, road condition, light condition, Car speeding, and under the influence of drug , inattentdence, pedestrian rights will be used for this analysis. This dataset is imbalanced with majority of the serverity in properties damage only collision, which makes up about 70% of the data, making classification more difficult. Some exploratory analysis is performed here in the notebook to understand data well.

2.2 Data cleaning

The Data collisions csv file downloaed had many missing data, because of lack of record keeping. However all rows were labelled meaning all cases have a severity code assigned to it. So I didn’t drop

any rows. Looking at the columns I understood those that may have an impact on determining the severity code could be
LOCATION','ROADCOND','WEATHER','JUNCTIONTYPE','LIGHTCOND','SPEEDING','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT'.

Then I found SPEEDING has lot of missing value compared to other columns so that need not be considered in the featureset.

PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT' were having no missing counts.

I have decided to replace missing values in ROADCOND,WEATHER,JUNCTIONTYPE and LIGHTCOND with the most frequent one's.

Then the categorical values are transformed to numerical ones for machine modelling.

2.3 Feature selection

After data cleaning there were oonly 8 rows in feature, initial data set had 38 columns. And in the cleaned dataset there were no missing values.

My final feature set includes

- 'ROADCOND'
- 'WEATHER'
- 'JUNCTIONTYPE'
- 'LIGHTCOND'
- 'PERSONCOUNT'
- 'PEDCOUNT'
- 'PEDCYLCOUNT'
- 'VEHCOUNT'

3.Exploratory Data analysis

Used bar plot and understood more accidents happened when lightcond was ‘Daylight’. The severity type more happened was Property Damage only collision

4. Modelling

There are two types of models, regression and classification, that can be used to predict. Regression models can be used for continuous target values , while classification models can be used to find the probabilities with discrete values. Also classification model determines the class label for an unlabeled test cases.

For our accident severity case its labelled data set and the target variable is categorical. So this can be modelled with classification and I chose Knn modelling technique.

5. Results

I have compared the different K values using Evaluation metrics accuracy evaluvation, jaccard_similarity_score and F1 score and found that 14 is the best K which gives .755 accuracy for jaccard_similarity_score and .72 for F1 score, which is an efficient one and close to 1.