Yingying Jiang

501212156

Supervisor’s name

May 14, 2023

Abstract

The problem is to develop a predictive model with the concentration on the injury severity of drivers in traffic collisions occurring on county and local roadways within Montgomery County. As for the theme, the researcher decides to choose classification and regression learning methods.

To be more specific regarding research problem, the researcher plans out 4 main research questions based on the summary of research questions listed above. Initially, what factors contributed to the injury severity in traffic collisions in Montgomery County? Secondly, how could a predictive model be effectively developed in order to forecast the injury severity level of drivers involved in traffic collisions? Furthermore, how do different modeling techniques and data preprocessing methods effect the accuracy of the predictive model? Lastly, how could the local policy makers, law enforcers, and the other potential stakeholders from Montgomery, Maryland utilize this model for enhancing road safety and reduce the injury severity of the drivers from traffic accidents?

The data used by the researcher is named as “Crash Reporting - Drivers Data” published by Montgomery County Government of Maryland, which can be sourced from the U.S. Government’s Open Data (data.gov). According to data.gov, this dataset provides detailed information on drivers involved in traffic collisions in Montgomery County, and the data was collected via the Automated Crash Reporting System (ACRS) of the Maryland State Police. (Data.gov, 2023) The raw data contains 159,357 rows and 41 columns. The data type of most columns can be converted to categorical data type, such as “Route Type”, “Collision Type”, “Weather”, “Surface Condition”, and “Light”.

The researcher proposes the following methodologies and techniques to tackle down each question listed from the second paragraph of this abstract. For the tech stack of this research project, Python and Jupiter Notebook are chosen considering the rising popularity among industrial practitioners specially in the data science field. Before conducting the research, the researcher suggests that there is a pressing need for data cleaning during data preparation considering the data is not in the well-maintained condition because of the inconsistencies occurred in several columns. Therefore, the researcher is implementing the following techniques in the data preprocessing stage: removing data columns with too many missing values, removing attributes with low variance, reducing highly correlated columns, balancing class attribute, feature selection, and processing PCA - Principal component analysis for Dimensionality Reduction. As for predictive modeling, Logistic Regression and Naive Bayes classifier are selected for investigating the first research question proposed by the researcher. As for the second and third research questions, Classification Tree and k-NN (k-Nearest Neighbours) are chosen by the researcher.

Reference List:

data.montgomerycountymd.gov. (2023, May 13). *Crash reporting - drivers data*. data.gov. https://catalog.data.gov/dataset/crash-reporting-drivers-data