Accident Severity Prediction in Seattle, Washington

Using KNN, SVM, Linear Discriminant Analysis, and Naïve Bayes

IBM Data Science Professional Capstone

Coursera

September 23, 2020

Table of Contents

Executive Summary

The objective of this project is to predict the severity of an accident in Seattle, Washington using machine learning models. Features such as road, weather and light condition were used for this prediction.

Predicting a categorical target data requires using a supervised classification. We used Decision Tree, Logistic Regression, Linear Discriminant, K-Nearest Neighbor, Naïve Bayes model and evaluated the result of the model using precision, recall and F1 score. We obtained the best result with Decision Tree algorithm with an accuracy of 0.75. The other models gave similar results within the range of 0.65 – 0.7. The worst result was Naïve Bayes with an accuracy of 0.65 and a precision of 0.52.

Introduction

Every day, millions of people commute using either their personal cars or public transportation for various reasons such as work, leisure etc. The urgency to get to our destination, lack of attention, driving under influence, road and weather condition, failure to obey traffic rules and regulations are one of the many reasons that causes accidents. This does not only lead to the loss of lives of the drivers and passengers involved, but this could also involve pedestrians and cyclist.

The benefits of being able to predict the severity of an accidents will not only forestall such mishaps but could help first responders send the right assistance to mitigate the damage and loss of lives of people. Emotional and socio-economic importance to the family, friends, and dependents of accident victims cannot be quantified.

In this project, we attempt to predict the severity of an accident using the machine learning algorithms such as Logistic regression, KNN, Linear Discriminant, and Naïve Bayes and Decision Tree. We used features such as road, light and weather condition. We evaluated the result of our prediction using F1 score, accuracy, precision and recall.

The prediction algorithm with no only help commuters determine whether it is safe to drive based on some certain conditions that are known in advance. Further, this model would also be beneficial to the first responders in Seattle, Washington in deploying the right personnel to accident scenes by knowing the severity of the accident ahead. Overall, this model could save lives and properties, and also reduce the cost of deploying by deploying the appropriate first responders.

Data

The accident data set of Seattle, Washington used in this project was provided by Coursera. The data consist of 37 attributes both numerical and categorical, and 194,673 records. The list below shows the 37 attributes provided in the data.

Data Columns: ['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
'INATTENTIONIND', 'UNDERINFL', '**WEATHER**', '**ROADCOND**', '**LIGHTCOND**',
'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']

For more details on the above attributes, see the metadata provided.

The **SEVERITYCODE** column is the target variable that contains the severity of the accident we are trying to predict. Based on this data, the severity can be categorized into bodily injury and property damage only collision.

To achieve the above goal, we select features that are known prior to the accident such as; **road, light and weather condition.**