Joshua Osko
Coursera Datascience Capstone
7 September 2020

Predicting Car Accident Severity Based on Traffic Conditions

Introduction

Automobile accidents are a responsible for over $800 billion [1] and over 30,000 casualties [2] in the United States each year. Identifying factors which most contribute to fatal accidents could help political officials better formulate policies which could reduce loss of life and save billions of dollars. In this paper, we will build a predictive supervised machine learning model to predict the potential severity of an automobile accident based on existing traffic conditions.

Data

For this paper, we will examine the data included in the example dataset provided by the Coursera Applied Data Science Capstone course. This dataset contains 194,673 accident observations and records 38 attributes for each accident.

Most of these columns are things which are clearly not causal to the accident and thus we will not consider as impactful in determining whether or not an automobile accident will be fatal. Here is the list of attributes to remove: X, Y, OBJECTID, INCKEY, COLDETKEY, REPORTNO, LOCATION, STATUS, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SEVERITYCODE.1, SEVERITYDESC, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT', INCDATE, INCDTTM, SDOT_COLCODE, SDOT_COLDESC, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR, PEDROWNOTGRNT.

We will look to build a feature set useful for accurately predicting how severe an automobile accident will be given existing traffic conditions. The following features will be used to predict automobile accident severity:

Feature	Description
JUNCTIONTYPE	The type of junction where the accident occurred
INATTENTIONIND	Indicates cause of accident related to inattention
UNDERINFL	Whether or not the driver was under the influence
WEATHER	The weather conditions when the accident occurred
ROADCOND	The road conditions when the accident occurred
LIGHTCOND	The lighting conditions when the accident occurred
SPEEDING	Whether or not speeding was a cause of the accident

The target column SEVERITYCODE will be predicted based on these predictor variables.

The remaining attributes require a deal of cleaning to properly prepare it for training the model. All categorical variables will be converted into numerical integer values for model training. The following actions will be taken:

Feature	Action

JUNCTIONTYPE	Convert all values to a 0 if not intersection related or a 1 if intersection related
INATTENTIONIND	Convert all values to a 0 for cause of accident not due to inattention or a 1 if it was due to inattention
UNDERINFL	Convert all values into a 0 for not under the influence or a 1 for under the influence
WEATHER	Convert all values into a 0 for bad weather or a 1 for good weather
ROADCOND	Convert all values into a 0 for bad road conditions or 1 for good road conditions
LIGHTCOND	Convert all values into a 0 for poor visibility, 1 for moderate visibility, or 2 for good visibility
SPEEDING	Convert all values into a binary 1 for speeding or 0 for not speeding

Following these preprocessing steps, the remaining rows with empty cells will be removed as we will not make any further assumptions of the missing data due to uncertainty and risk to disrupting desirable model results.

References:

[1] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013
[2] https://www-fars.nhtsa.dot.gov/Main/index.aspx