Predicting the Car Accident Severity

Pashkov Daniel

October 11, 2020

1. Introduction

1.1 Description of the background

A traffic collision, also called car accident one of the deadliest accidents in our life. Every year millions of people died on a road, for example in 2013 54 million people worldwide sustained injuries from traffic collisions. About 68,000 of these occurred in children less than five years old. Almost all high-income countries have high death rates, property damage as well as financial costs to both society and the individuals involved. Therefore, this is important problem for hundreds of countries in the world and, especially, for modern cities there number of individual transports increase every year.

1.2 Problem

Various factors contribute to the risk of collisions as well as the number of people involved in this collision including speed of operation, road environment, driving skills, impairment due to alcohol or drugs, and behavior, notably distracted driving, speeding and street racing.

This project aims to predict car accident severity based on this factors and dataset which used in this project must include this feature.

2 Data cleaning and feature selection

2.1 Data sourse

In project I used collisions data from Seattle provided by Seattle Police Department. Seattle population is about 750 000 and city is the center of 3 million urban area, also Seattle characterized wide range of weather condition and data collected over a long period: from 2004 to 2020, that's why Seattle data is perfect example for this project.

Data can be found in Collisions data set from Seattle government site here and the metadata which describe data set here

2.2 Data cleaning

Original data set have 221738 examples and 40 parameters. Our target parameters are "SEVERITYCODE" which include 5 values/codes however two codes are similar and be about ~2% of entire data, so this data was dropped. Another code means "unknown" values: about ~10% of set, this data also was dropped because it doesn't contain any information. After dropping NaN values entire data set contain 184167 examples: 30 % collision with injuries and 70% without injuries. This is imbalanced data set.

2.3 Feature selection

Upon examining the meaning of features, it was clear that need some redundancy.

First, in this data set there are lot of feature which contain different keys for police database such as "INCKEY" and "COLDETKEY": Unique key for the incident and Secondary key. This is the list of such features:

'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'INTKEY', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'COLLISIONTYPE', 'SDOT_COLCODE', 'SDOTCOLNUM', 'ST_COLCODE', 'SEGLANEKEY', 'CROSSWALKKEY'.

Second, there are duplicate columns: 'INCDATE' – data and 'INCDTTM' – data with time. 'SEVERITYDESC' and 'SEVERITYCODE' – 'severitycode' coded values in 'severitydesc'.

Third, columns 'STATUS', 'JUNCTIONTYPE', 'SDOT_COLDESC', 'ST_COLDESC', 'INATTENTIONIND', 'HITPARKEDCAR' doesn't include meaningful information for predicting, for example 'JUNCTIONTYPE' – describe which part of vehicle was broken.

In the result 17 feature was selected:

Feature	Description	Data preparation
'X', 'Y'	Longitude, Latitude	Remove NaN values
'ADDRTYPE'	Collision address type	
'PERSONCOUNT'	The total number of people involved in the collision	
'PEDCOUNT'	The number of pedestrians involved in the collision. This is entered by the state.	
'PEDCYLCOUNT'	The number of bicycles involved in the collision. This is entered by the state.	
'VEHCOUNT'	The number of vehicles involved in the collision. This is entered by the state.	
'INJURIES'	The number of total injuries in the collision. This is entered by the state.	
'SERIOUSINJURIES'	The number of serious injuries in the collision. This is entered by the state.	
'FATALITIES'	The number of fatalities in the collision. This is entered by the state.	
'INCDTTM'	The date and time of the incident.	Convert to datetime format, extract month, hour
'UNDERINFL'	Whether or not a driver involved was under the influence of drugs or alcohol.	Replace Y, N to 1,0
'WEATHER'	A description of the weather conditions during	Remove values with small number of examples (~10%)

	the time of the collision.	
'ROADCOND'	The condition of the road during the collision.	
'LIGHTCOND'	The light conditions during the collision.	
'PEDROWNOTGRNT'	Whether or not the pedestrian right of way was not granted. (Y/N)	Replace NaN values to 0 because other cells were filled as Y/1
'SPEEDING'	Whether or not speeding was a factor in the collision. (Y/N)	