Coursera

IBM Professional Data Science Certificate Capstone Project Report

Factors Effecting Accident Severity in the City of Seattle, WA

Chad Hackert

SEP 2020

https://github.com/CHackert/Coursera\_Capstone

*Introduction / Business Problem*

The US National Safety Council (NSC) has estimated that in the first six months of 2020 the total number of US motor-vehicle fatalities reached 18,300, motor-vehicle related injuries reached 2,086,000 and property damage was over $206 billion USD (ref1). Helping drivers and pedestrians be more informed of the most likely factors leading to an accident before travelling should substantially reduce these numbers and provide for safer and less stressful motor-vehicle travel experience. In addition to the travelling public City Government can use the results of this project to identify and correct high accident controllable infrastructure factors and impose laws directly related to traffic safety. Using data collected by the Seattle Police Department (ref2), this project aims to provide useful information regarding varying conditions and their effects on accident severity in the City of Seattle.

*Data*

*Data Overview*

The traffic accident data to be used for this project was collected by the Seattle Police Department (SDP) during routine accident reports and made publicly available by the Department of Transport (DOT) (ref2). This data set covers accidents in the City of Seattle from January, 2004 to May, 2020 and includes 194673 individual reports. The data collected by the SPD falls into 38 features (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, SDOTCOLNUN, SPEEDING, ST\_COLCODE, ST\_COLDESC, SEGLANEKEY, CROSSWALKKEY, HITPARKEDCAR) which have either administrative or accident related details. Using the accident detail data (ie. Severity, Location, Weather, more) this project will attempt to predict the most common factors effecting the severity of collisions.

*Data Cleaning, Preparation and Exploration*

While the data set is large and contains volumes of useful features not all of these features are causal to the accidents or relate to the ultimate severity. In order to examine the factors leading to severe accidents features that are administrative in nature or redundant were removed. Of the 190000+ reports many were missing data related to the severity of the accident reported on. Driver inattention, pedestrians not granted right of way and speeding data were incomplete in that the reports only indicated a positive result when completed. The negative results were added to the dataset using a replace function and assigning a 1 or 0 value. In addition, one column (“INTKEY”) which was a numerical value assigned to specific intersections in the city. This column was removed as over 2/3rds of the records did not have information included. After this, any rows in the spreadsheet which were missing data that could not be recovered were dropped. This resulted in 10527 records being removed from the file constituting a 5.4% loss of data.

An initial exploration of the data was completed using histograms to determine the number of accidents which occurred given a specific feature and grouped by the severity of the accident (injury accident vs property damage only accident).

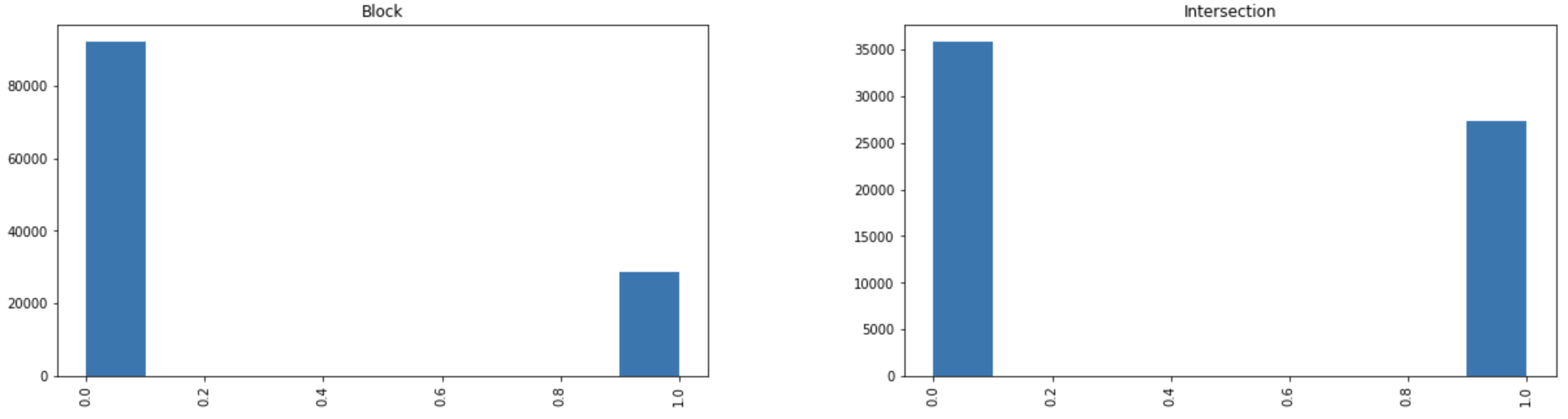


Fig 1. Histograms of the number of injury accidents vs the number of property damage only accidents (‘SEVERITYCODE’) grouped by their location (‘ADDRTYPE’).

The results of the histograms were interpreted with respect to their total count and the relation of injury to property damage only accidents to determine which features were to be selected for modeling. These features were then compared to information published by the US DOT (ref3) regarding most common causes for motor-vehicle accidents and filtered down to eight features. These are Address Type (intersection or on street), Collision Type, Driver Inattention, Impaired Driving, Weather, Road Conditions, Light Conditions and speeding.

REFERENCES

ref1: US National Safety Council - <https://injuryfacts.nsc.org/motor-vehicle/overview/preliminary-estimates/>

ref2: Seattle MV Accident Data - <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv>

ref3: DOT National Motor Vehicle Crash Causation Survey - <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811059>