DATA SCIENCE CAPSTONE PROJECT

SEATTLE TRAFFIC ACCIDENT SEVERITY – A CASE STUDY

By Kate Pickering, September 2020

Outline

INTRODUCTION

Can we use simple accident data to build a model to predict accident severity?

DATA

The Seattle Police Department collects accident information, and makes it publicly available for analysis

CONTENTS

The data is available in two files, a CSV with the raw data, and a PDF file with the metadata

INITIAL DATA EXPLORATION: WHAT DATA IS AVAILABLE?

There is data on number of vehicles, people, pedestrians, cyclists involved, as well as inforation about accident severity, and involvement of driver inattention and speeding

Accident Severity Data

This data contains two possibilities, property damage or injury. This is what we would like to predict.

Accident Severity?

What input data can we select from?

Input Data

PERSON COUNT

This is the number of people involved in the accident. Ranges from 0 to 81.

PEDESTRIAN COUNT

This is the number of pedestrians involved in the accident. Ranges from 0 to 6.

CYCLIST COUNT

This is the number of pedestrians involved in the accident. Ranges from 0 to 2.

Input Data

VEHICLE COUNT

This is the number of vehicles involved in the accident. Ranges from 0 to 12.

DRIVER INATTENTION

This was a factor in 29805 accidents

SPEEDING

This was a factor in 9333 accidents.

Methodology

Since we are trying to model something with two potential outcomes, binary logistic regression is selected

Methodology

Split the 194673 lines of data into training and testing data sets

Use 80% for model training

Use 20% for model testing

Model Evaluation

Jaccard score of 0.75186

Model Evaluation

R^2 0.7512 using training data set

R^2 0.7518 using testing data set

Model Evaluation

Precision for property damage: 0.75

Precision for injury: 0.81

Recall for injury: 0.22

Model Evaluation

F1-Score 0.7

Discussion

The model is decent at predicting accident outcomes, but is skewed to classifying accidents as property datage over injury

Discussion

There are more factors involved in accident outcome than the variables used in our model

Conclusion

There are many laws and regulations attempting to limit driver inattention and speeding, which is validated by the data