Applied Data Science Capstone

Capstone Project - Car Accident Severity

OMID KARAMI

Applied Data Science Capstone

AUGUST 2020

3. Data

3.1 Data Description

In order for a project to be successful, it is important to use a manageable dataset to analyze it. For this purpose, all data must be available and also to achieve the desired result, with high accuracy and efficiency, its size should be large enough, and of course all pre-processing conditions, including data cleaning and normalization, should be considered. This project includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. (Timeframe: 2004 to Present). For further information you can visit the official website of Seattle Department of Transportation through the following link.

3.2 Data Cleaning

High quality data enables strategic and business systems to have a better insight of problem solving. Data cleaning is one of the most important steps in pre-processing stage. Eliminating unwanted, duplicate and irrelevant observations allows even simple algorithms to have significant insights into the data and achieve better results. Detect the outlier data and remove it, as well as handling missing data, help to producing quality data and model's performance.

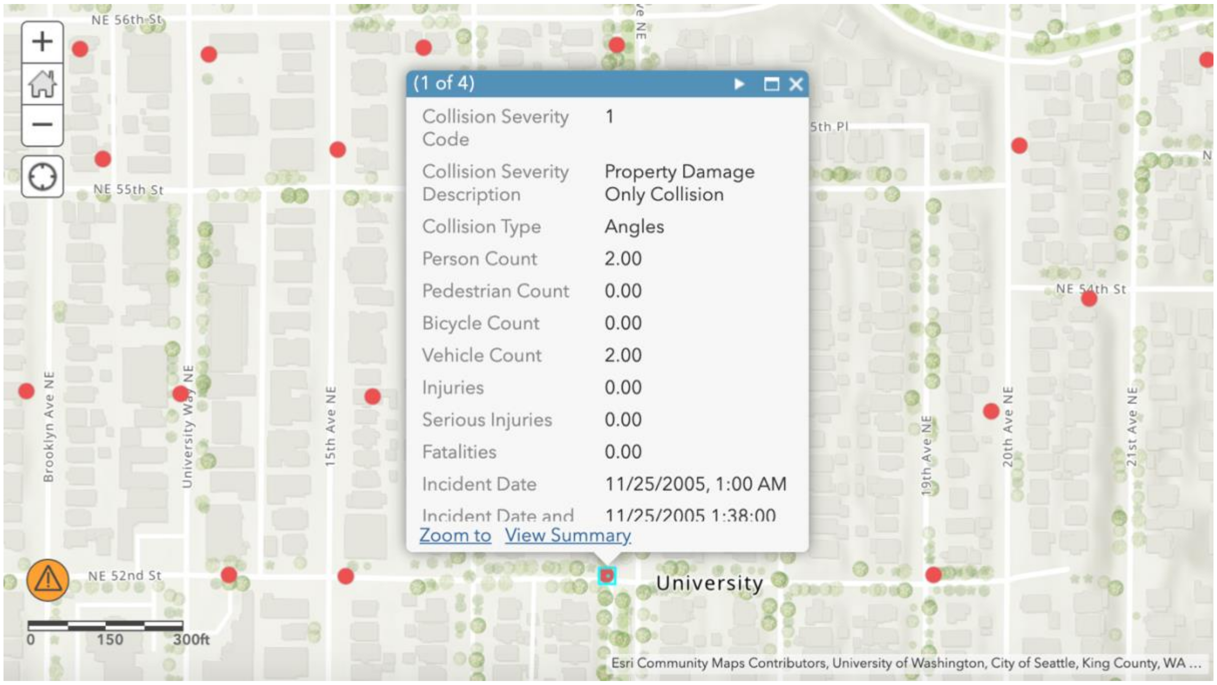
3.3 Feature Selection

The feature selection process in many cases enhances the performance of the machine learning model. This process involves reducing the number of input variables to reduce computational costs in developing a predictive model. In addition, the selection of features includes the selection of the most important and relevant features. In this research, the following features are key features in predicting the severity of accidents, which will be fully discussed in the methodology section.

- WEATHER: *A description of the weather conditions during the time of the collision.*
- PEDCOUNT: *The number of pedestrians involved in the collision.*
- PEDCYLCOUNT: *The number of bicycles involved in the collision.*
- VEHCOUNT: *The number of vehicles involved in the collision.*
- INJURIES: *The number of total injuries in the collision.*
- SERIOUSINJURIES: *The number of serious injuries in the collision.*
- FATALITIES: *The number of fatalities in the collision.*

Regarding to have a good insight of data exploration, I obtained information respecting the collision from the ArcGIS Online 1 that enables to connect people, locations, and data using interactive maps. This map and ocean of another map are available in ArcGIS Online website.

[1] https://www.arcgis.com/home/index.html



Collision Severity Code	1
Collision Severity Description	Property Damage Only Collision
Collision Type	Angles
Person Count	2.00
Pedestrian Count	0.00
Bicycle Count	0.00
Vehicle Count	2.00
Injuries	0.00
Serious Injuries	0.00
Fatalities	0.00
Incident Date	11/25/2005, 1:00 AM
Incident Date and	11/25/2005 1:38:00

Esri Community Maps Contributors, University of Washington, City of Seattle, King County, WA ...

Figure 1 - Pedestrian Count

References

Vatanavongs Ratanavaraha, Sonnarong Suangka, Impacts of accident severity factors and loss values of crashes on expressways in Thailand, IATSS Research, Volume 37, Issue 2, 2014, Pages 130-136, ISSN 0386-1112, https://doi.org/10.1016/j.iatssr.2013.07.001.

Fang Zong, Hongguo Xu, Huiyong Zhang, "Prediction for Traffic Accident Severity: Comparing the Bayesian Network and Regression Models", Mathematical Problems in Engineering, vol. 2013, Article ID 475194, 9 pages, 2013. https://doi.org/10.1155/2013/475194.

https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

https://www.arcgis.com/home/index.html