New York City Traffic Collision Data Science Analysis

A Capstone project submitted in partial fulfillment of the requirements for the

Springboard Data Science Intensive program

Springboard

August 2017

Statement of the Problem

Traffic collisions and fatalities have been increased ever since 2012. The is a concern not only for the New York State Department of Transportation (NYSDOT) but also for the Major of New York City. The Major along with the NYDOT has provided us with traffic data captured by the NYPD for analysis. They want the following questions answered so they can work together to deploy the most effective safety measures.

1. What day of the week do most accidents occur?

2. What time of the day do most accidents occur?

3. What is the most frequent contributing factor in the cause of accidents?

4. What zip code do most accidents occur?

5. Predict the number of possible collisions likely to happen for next year

6. Predict the number of possible fatalities likely to occur for the next year

The Clients

New York State Department of Transportation (NYSDOT)

NYSDOT is responsible for coordinating and developing comprehensive transportation policy for the State; coordinating and assisting in the development and operation of transportation facilities and services for highways, railroads, mass transit systems, ports, waterways and aviation facilities; and, formulating and keeping current a long-range, comprehensive statewide master plan for the balanced development of public and private commuter and general transportation facilities.

Major of New York City

The Major of New York City is working with the NYDOT in support of Vision Zero. The goal of the Vision Zero Action Plan is the City's foundation for ending traffic deaths and injuries on our streets. The city has become nationally and internationally recognized as a leading innovator in safe street designs. The NYDOT has made major engineering changes and since 2005 fatalities have decreased by 34%, twice the rate of improvement at other locations. It involves better designs and regulations are already making our streets safer, and we will expand these efforts. We will bring more resources to enforcement and public outreach. In Albany, we will seek control over the City's speed limits and use of enforcement cameras.

Making New York the world's safest big city will require more than government policy and programs - It will take citizen action from the grassroots up. It demands the participation by the State legislature and lawmakers, industries, companies and authorities that operate large numbers of vehicles. Vision Zero’s goal is to engage every New Yorker to join the public conversation on street safety and to do his or her part to safely share the roads.

Both the NYDOT and the Major’s office plan to use the results of our analysis to improve existing road and traffic regulations resulting in improved road way designs to make streets safer.

The Data

The data being studied was obtained from the New York City Open Data project. The Mayor’s Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT) partner to form the Open Data team. As a hub of analytics in the City, MODA advocates for the use of Open Data in citywide data analytics and in the community. DoITT manages the technical operations with City agencies and our vendor partner Socrata,

ensuring that technological capabilities are always evolving to better meet user needs. Agencies are the data owners and have Open Data Coordinators who serve as the primary point of contact with the Open Data team.

The data set were collected by the New York Police Department (NYPD). It consists of over 1 million motor vehicle collisions records in New York City for a period of 5 years (2012 – 2017).

The data set were obtained from the New York Police Department (NYPD) at the link below.

https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

Data Attributes

The data consist of the following 29 attributes;

1	DATE,
2	TIME,
3	BOROUGH,
4	ZIP CODE,
5	LATITUDE,
6	LONGITUDE,
7	LOCATION,
8	ON STREET NAME,
9	CROSS STREET NAME,
10	OFF STREET NAME,
11	NUMBER OF PERSONS INJURED,
12	NUMBER OF PERSONS KILLED,
13	NUMBER OF PEDESTRIANS INJURED,
14	NUMBER OF PEDESTRIANS KILLED,
15	NUMBER OF CYCLIST INJURED,
16	NUMBER OF CYCLIST KILLED,
17	NUMBER OF MOTORIST INJURED,
18	NUMBER OF MOTORIST KILLED,
19-23	CONTRIBUTING FACTOR VEHICLE (1-5),
24-29	VEHICLE TYPE CODE (1-5)

The Approach

Our approach to the problem is to follow the six process stages of the Cross Industry Standard Process for Data Mining (CRISP-DM). The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology.

1. Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the Decision Model and Notation standard can be used.

2 . Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

3. Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.

4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

5. Evaluation

At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

6. Deployment

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that is useful to the customer. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (e.g. segment allocation) or data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. Even if the analyst deploys the model it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

Deliverables

1. Code using for data wrangling, data analysis and visualization
2. Presentation deck providing a project overview and highlights
3. This Capstone document
4. Recommendations

References

http://www1.nyc.gov/site/visionzero/the-plan/letter-from-the-mayor.page

https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95