Project Proposal: Analysis and Modeling of Vehicle Details in Car Crashes

01.22.2018

—

By Aparna Shastry

Overview

This document proposes a data science project using car crash data of State of Illinois over a few years. The dataset has details about the crash, vehicles involved and people involved over a few years. The dataset has millions of samples with 100+ factors to analyze and is expected to give some interesting insights.

Goals

The project goals at this point are,

1. First and Foremost goal of this project is to go through a project cycle and understand what it takes to do a good data science project. [Definition of "good" in this context is whether it is acceptable by an industry expert to consider me for a data science position in their team.]
2. Understand the correlation between the crash/person/vehicle data and Vehicle Make/Model/Types.
3. Come up with a predictive model to guess the make/model involved in the crash
4. Compare and Contrast the trends between State of Illinois and City of Chicago.
5. Find out what causes crashes in certain make/models, the so called "feature importance"

Like every data science project, this also has a risk of Scope Creep, and in this case, it might be good!

Possible Clients

The following type of people/group can make use of the outcome of this project, to make better decisions or take actions

1. Insurance companies, to decide the premium amounts on the vehicle insurance.
2. Car manufacturers themselves, those which are more prone to crashes, to know where is the scope for improving the design.
3. Car buyers who would want to avoid more crash causing/damage causing Make/Models until they knows that these manufacturers have taken measures to correct crash causing factors
4. Lawmakers perhaps, to ban certain models, beyond a threshold of car crashes

Data Source

This is where we got the data for this project:

https://github.com/stevevance/Chicago-Crash-Browser/blob/master/DATA.md

We plan to use a subset of this , i.e. years 2013-2015 .

Approach

This is the current plan of execution. It might evolve along the way.

1. Take just one year of data, say 2014.
2. Do the necessary data wrangling, like putting column names, ensuring they are consistent, Unknown fields are handled in a sensible manner.
3. Then familiarizing with the year 2014 data. The three tables, namely Crash, Vehicle and Persons sum up to 0.55 GB which is decent.
4. Conduct numerical and visual Exploratory Data Analysis (EDA) on the merged table, for just a single city, say Chicago. Eliminate columns which are irrelevant and copies of the same columns (in text/code form)
5. Repeat the above steps for the entire data (state of Illinois)
6. Conduct hypothesis tests as per the need.
7. Document the findings so far, draw statistical/practical inferences if any.
8. Fit Machine Learning predictive models to address the points mentioned in answer to question number 1.
9. Evaluate the model's predictive power.
10. Conclude with remarks/recommendations and scope for future work.

Milestones

I. Results on Make/Model questions with Chicago City Data

This milestone will have a IPython notebook demonstrating data wrangling, EDA and output/conclusions on the Chicago City Crash data over 3 years, 2013,2014 and 2015. Expected to take 4 weeks.

II. Results on all questions in the goals sections with the Illinois State Data

This is the final milestone with deliverables listed in the next section. Expected to take 4 more weeks after milestone 1.

Deliverables

1. IPython Notebook containing problem statement, concise and clear code with necessary and sufficient comments, outputs and conclusions
2. A Slide Deck containing visualizations, with minimal text
3. A Report in the form of an academic paper
4. A Blogpost, that is brief version of 3

Acknowledgment: Thanks to Steven Vance for extracting and keeping the IDOT data on his github.