Aparna Shastry

Irvine, USA

Data Science Project Ideas

Feb 14, 2018

Overview

This document discusses 4 ideas for a data science project. The scope of document is to define a problem and give a link to the dataset

Ideas

1. **Predict Hospital Readmission Rate/ Time between admissions and Costs:** Using the data from over Medicare - certified 4,000 hospitals, across various csv files, predict the following:

a) What is the readmission rate within 30 day period
b) What is the expected time between admission
c) Number of hospital acquired conditions
d) Estimation of medicare - medicaid payments for next year

The results from first 3 points will help people to understand the quality of care. The last one would help allocating the budget.

Dataset: https://data.medicare.gov/data/hospital-compare (Download csv flat files button)

Note: This dataset was earlier used by a springboard career track student. It is also used in an exercise. However, that will not be a deciding factor to choose or drop this, because this is interesting and perhaps can add value.

2. **Bollywood Movie Production Cost/ Box office Revenue estimation:** The Indian film industry produces thousands of movies of which 43% are Bollywood movies (https://en.wikipedia.org/wiki/Bollywood). Hollywood movies are around 500 per year. Estimation of Production Costs and Box office revenue would be both interesting and challenging. Firstly, Indian films have more songs, more cultural diversity and hence they are different from Hollywood films, I would guess more features can be identified. As we want to take data over a few years, we will need to account for India's explosive population/inflation numbers. The data/statistics is not easily available in India for any topic, there is scope for web-scraping from different websites, starting with wikipedia to trailer/review sites. One can do NLP to correlate plot and revenues. Hence this is one of a kind data science problems in my view!

Dataset: To be scraped, starting point:
https://en.wikipedia.org/wiki/Lists_of_Bollywood_films

3. **Predict Average time for Building Permit issuance:** To build any structure, the permit from city administration is necessary. A model to predict the delay from application to issuance would serve the contractors to plan better. Cities like Los Angeles, New York and San Francisco have open data on building permits. Trulia conducted a study in 2016 and found that building permit delays are the main cause of supply not catching up with demand. It would be interesting to see if there is any one factor that influences the decision of issuing the permits by correlating with census data as well.

Example Dataset:
https://data.sfgov.org/Housing-and-Buildings/Building-Permits/i98e-djp9

Perhaps along with, population/building price characteristics of the city (yet to explore how to extract city specific data from https://www.census.gov/)

4. **Predict the Crash Severity in traffic violations:** It would be interesting to analyze and draw insights from Traffic violations data obtained from a city. As the dataset would contain traffic violations, collisions without any damage, collisions with all sorts of damage, this dataset is richer than the data containing crashes alone. The dataset is downloaded from Montgomery County, Maryland

   Dataset: https://catalog.data.gov/dataset/traffic-violations-56dda

5. **Predict the Crash Severity in motor crashes:** It would be of help to people if from the vehicle info, driver info (like age, gender) and weather/season conditions, if one could come up with a predictive model to predict whether there would be injury to people in case a crash happens at a particular location (?). This is a tricky question because the missing information of those that did not collide would give insights to train our predictive model. In the absence of "No crash" records, it is not going to be an interesting data science problem. It could be a good data analysis problem.

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

Note: This problem is documented here for the records. It was considered before coming up with the above 3.

Next Steps

Select one of the above to do at this point and submit a proposal

For those who are interested to know how I came up with ideas:

Approach Taken So Far

1. My mentor for this Springboard intermediate Data Science course has been insisting on doing a relatively unheard of data science project that would be of use to a set of people or organization. Tried to sit back and think of a problem, couldn't come up with an interesting problem that would make an impact.
2. Hence, got some experience on the data science flow by playing with a toy dataset
3. Over last two months, browsed a few completed capstone projects, kaggle problems.
4. Started reading Introduction to Statistical Learning book
5. Read a few motivational articles by Springboard alumni/TowardsDataScience/KDNuggets blog, spoke to some other data scientists I know. One important conclusion: Results from a data science project should beat commonsense baseline results.
6. Once I got the intuitive feel for what makes a good project, scribbled what I thought would make a good project, and how they can fit into various machine learning model types.
7. Any ML problem would be either predicting qualitative or quantitative. Qualitative seems more interesting to me, which I view as predicting time or money in their various forms.
8. Just relaxed and let the ideas flow. Became more clear with documentation. There are some more too.