Ideas for Capstone Project 1

Ahrim Han

2018/11/1

1. Identification of the schools that need the most supports

Overview:

The California Assessment of Student Performance and Progress (CAASPP) is the computer-based assessment to measure the skills of students against the same academic standards in the same way. With these results, we can get insights into student progress and helping schools put them to use improving teaching and learning. To broadening educational opportunities, it is important to analyze the various impacts on the low scores and find the exact causes. These results can be used to effectively identify under-performing schools that need the most of efforts and budgets to provide consulting and support collaboration.

Suggestion:

I plan to focus on analyzing multiple data sources to find more informed and granular causes that impact the 1) test scores, 2) school rankings, and 3) school ratings (10 scales) of schools.

I plan to use the following data.

- Zillow data (i.e., house prices)
- school financial data
- teacher data (such as education level, salaries, years of teaching experience, and student ratios)
- demographics (racial diversity, gender, age, population), income, and education level for residents

--- Updated on November 1, 2018 ---

Problems (Questions to solve) //to be developed

~~Turtlerock?~~

Interesting factors to be more deeply considered:

- History of scores
- Teachers qualifications

Dataset:

2017 CAASPP Statewide Student Data Summary

https://www.cde.ca.gov/ta/tg/ca/caaspp17datasummary.asp

Demographics, Income, Education of California

https://statisticalatlas.com/state/California/Overview

California Department of Education

https://www.cde.ca.gov/ds/dd/

California Public Schools (including Demographics of teachers)

https://www.ed-data.org/state/CA

Zillow dataset (House price)

https://www.zillow.com/research/data/

Comments from community mentor:

- Pros: Prediction of score outcomes for a school (school district, zip code based area, so on) can be interesting with unique features for each school

My notes and concerns:

- The analysis in various perspectives → expected to gain new insights
- More than two datasets can be combined → can practice data wrangling skills

- Any other ideas? Is there anything that your team is concerned about?

2. Influential review prediction

Overview:

Reviews and ratings of customer experiences are a big influence in determining the success of a local business. Yelp has been one of the most popular Internet rating and review sites for local businesses. Although many review sites have appeared and disappeared, Yelp has worked hard to maintain credible reviews, which have become a key factor in Yelp's continued growth. Therefore, it is important to examine how Yelp manages the review platform and find insights into how a good review can impact the growth of local businesses.

Suggestion:

I plan to analyze reviews of successful businesses (for example, more than 4 star-rating with more than 1,000 reviews) to investigate important factors that determine influential reviews. Many factors, such as review length, positive word count, elite member reviews, and exposure to top-ranked searches, including high-resolution photos and well-organized menu listings and corresponding pictures, affect the business revenue or review accumulating speed. With this insight, I can predict the success of a business or guide new business owners to get influential reviews.

Dataset: Yelp

https://www.yelp.com/dataset/download

Comments from Farrukh:

- Pros: Yelp data looks good because you can use 1) natural language processing and 2) statistics techniques (blending and mixture data is always good)

3. Fake review detection

Overview:

Product reviews are very important references when a consumer decides to purchase the product. Fake reviews harm the credibility toward companies and make to lose potential customers. Therefore, to maintain fair reviews, many companies, such as Amazon, focus efforts on finding fake reviews and give penalties on incentive and paid reviews.

Suggestion:

I plan to investigate the characteristics of fake reviews (e.g., unverified purchases and suspicious reviewers having patterns of only compliments or deleted reviews). With the insights gained, I plan to provide a fake review detection model that companies can use to filter fake or suspicious reviews.

Dataset: Amazon customer review

http://jmcauley.ucsd.edu/data/amazon/

https://s3.amazonaws.com/amazon-reviews-pds/readme.html

https://snap.stanford.edu/data/index.html#amazon

Comments from community mentor:

- Cons: Manual labeling of reviews (True or False) needed, which is very time consuming