Potential Data Science Projects

Lina Lavitsky

1. Understanding Foodies in Different NYC Boroughs (with help from Yelp!)

The Problem: Do people who live in different boroughs/parts of boroughs in New York value different things in their food establishments (aka does a 5-star restaurant in Brooklyn have different features on average than a 5-star restaurant in the Bronx)?

- Interesting for potential restaurant owners considering where to open a new location.

Data: Yelp provides the data as part of their Yelp Dataset Challenge in nested json format. I also found the same data in a csv file, including columns such as address, restaurant name, average review, various indicator variables ("Wifi", "Delivery", "Parking", "Good for Kids"), and many other fields you can find on a Yelp review page.[1]

Hypothesis: I think there will be a difference by boroughs and am curious to see what that is! For example, do 5-star restaurants in Park Slope tend to be more "Kid Friendly" with "Outdoor Seating" than other neighborhoods?

1. https://raw.githubusercontent.com/vc1492a/Yelp-Challenge-Dataset/master/Prepped%20Data/output.csv

2. Can You Predict Stock Market Performance Using News Headlines?

The Problem: Is there a relationship between stock market performance and news headlines or does the stock market function independently? If there is a relationship, does it have any significant predictive power? Is the stock market more sensitive to bad news or good news?

- Interesting information for investors/financial analysts/traders.

Data: A dataset on Kaggle, where stock market performance is given a 1 or 0 for whether stocks are up or down on the day. There is 8 years worth of data, and each day has the top 25 headlines on the day.[1]

Hypothesis: I think that "bad" news may have a bigger predictive power for the stock market than "good" news, but I'm skeptical about whether either relationship will prove very predictive (many other factors drive stock returns).

1.https://www.kaggle.com/aaron7sun/stocknews

3. Predicting Movie Tastes Using Demographic Information

The Problem: Can you use certain demographic information about a user (i.e. age, job, gender, occupation, zip code) to predict what genre of movies that user may like?

- Would be useful information for Netflix, Hulu, or Amazon Video

Data: A Movielens dataset, which includes three files: movies.data, ratings.dat, users.dat.[1]

- Movies.Dat has columns: MovieID::Title::Genres
- Users.Dat has columns: UserID::Gender::Age::Occupation::Zip-code
- Ratings.Dat has columns: UserID::MovieID::Rating::Timestamp

Hypothesis: There will be some interesting relationships between age/occupation/gender and what movies people prefer.

1.http://files.grouplens.org/papers/ml-1m-README.txt