

Idea 1: Modular/manufactured Homes and its relation to affordable housing

The bay area is facing an affordability housing crisis where there aren't enough homes that require less than 30% of a family's household income. Traditionally, affordable housing is built using a combination of inexpensive construction materials and techniques; e.g. stucco for coating walls. However, a number of startups (Kattera, Blokable, RAD Urban, Kasita, Boklok, etc.) are looking into solving this issue using techniques in manufacturing and technology. I would like to know from the data if modular/manufactured homes could truly be a solution for affordable housing if housing regulations were to be relaxed.

Potential Client: Government agency

Thoughts: Floods temp. Housing;

Questions:

- Are manufactured/modular homes focusing on affordable housing at all? Or just high-end homes?
- What are the strong predictors of price for affordable housing vs. manufactured/modular housing? How different are they?

Sources:

<https://catalog.data.gov/dataset/manufactured-homes-survey>

<https://catalog.data.gov/dataset/housing-affordability-data-system-hads>

Idea 2: Outdoor air impact on exercise intensity in Beijing, China

Beijing is notorious for its outdoor air pollution, however, people are still willing to exercise outdoors and even run marathons in these not so ideal conditions. I'd like a better understanding of how much workouts are affected by external factors like the outdoor air quality (PM2.5, humidity, temperature, etc.).

Client: Government agency (public health)

Questions:

- Is it better to get exercise or stay inside due to pollution concerns at a certain PM2.5 threshold?
- Could you predict peak performance on a route based on the weather?
- Do people prefer to run or bike under certain conditions?
- Do people change their exercise duration or route depending on certain types of weather?

Sources:

<http://developers.strava.com/>

<https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>
<https://www.nytimes.com/video/world/middleeast/100000005705502/big-data-big-problems-how-stravas-heat-map-uncovers-military-bases.html>
<http://olivernash.org/2014/05/25/mining-the-strava-data/index.html>

Idea 3: Predicting box office revenues based on metadata from IMDB, or social media text data from Twitter

Movies with large revenues at the box office may be largely associated with how critics and the audience rate the movie, the actors involved and much more. Given the rich variety of potential features, it'd be interesting to determine what are the top three features.

Potential Client: Movie Directors

Questions:

- Are regression techniques more predictive than classification techniques for the problem?
- Could social media data contain stronger features than movie ratings?

Sources:

<http://www.imdb.com/>
<https://www.rottentomatoes.com/>
<https://medium.freecodecamp.org/whose-reviews-should-you-trust-imdb-rotten-tomatoes-meta-critic-or-fandango-7d1010c6cf19>
http://snap.stanford.edu/class/cs224w-2015/projects_2015/Predicting_Box_Office_Revenue_f_or_Movies.pdf
<http://cs229.stanford.edu/proj2011/YooKanterCummings-PredictingMovieRevenuesUsingImdbData.pdf>

Idea 4: Predicting occupancy in a room based on time-series data from sensors

IoT use in homes has proven in many cases to make people's lives more efficient than ever before. In this study, I'm hoping to use time series data to predict if someone is present in a room or not without the use of motion sensors.

Potential Client: NEST, IoT companies

Sources:

<https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+>

Idea 5: PEER structural damage identification contest

Identify structural damage from photos using CNNs

http://peer.berkeley.edu/news/2018/04/peer-hub-imagenet-phi-challenge-pre-announcement/?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+PeerCtr+%28PEER+News%29

Idea 6: Predicting best-selling dropshipping products

Dropshipping enables normal people to move (goods) from the manufacturer directly to the retailer without going through the usual distribution channels. Most websites like Shopify publish best practices on how to identify successful products to sell like bow ties. However, a lot of the explanations seem based on intuition. I would like to see data science principles applied to data from scraping various websites.

<https://trends.google.com/trends/>

<https://www.instagram.com/developer/>

<https://www.shopify.com/guides/what-to-sell/trending-products>

Finalize one capstone idea based on the feedback that you received on your initial ideas and your discussions with your mentor.

The proposal should address the following questions:

- What is the problem you want to solve?
- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?
- What data are you using? How will you acquire the data?
- Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
- What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Submission instructions:

1. Write your proposal in a Google Doc (1-2 pages) and submit the link via the "Submit" button. Make sure your mentor has permissions to comment on the document.
2. Work with your mentor to incorporate any feedback into later drafts and submit as many times as needed.
3. Once your mentor has approved your proposal, convert the doc to a PDF file.
4. Create a GitHub repository for this project (if you haven't done so already).
5. Add the PDF to your GitHub repository for this project.
6. Share the proposal with your peer community for feedback.

Note: All code and further documentation you write will be added to this repository.

Your project will be evaluated using this [rubric](#).

For pro runners and cyclists,