Capstone Project 1 Proposal
Predicting Wind Farms Impact on Home Values

Questions for Mentor:

Capstone:

1. For approach is a classification algorithm the best option?
2. If I use the Zillow dataset, how do I get the data without using the UI (screen scraping?)?

Career

3. Seems like a lot of data science positions require a doctorate. Has that been your experience?

Next week's mentor call

4. I'll be in Northern Michigan with limited access to phone/internet. Could we skip next week and do a one hour call on Sept 26th?

Problem:

There are many factors that influence a home's value such as the region's economy, condition of the home, proximity to shopping & other attractions, school district rating, etc…, but a wind farm may be a greater factor on home values than one, or even a combination of several of these factors.

Imagine you just opened your favorite news feed and the local news is that a wind farm is coming to your area. As a homeowner, should you be concerned about your home's value? As a town board member should you raise concerns about property tax revenue (due to a decrease in home values) at the next meeting?

We propose to use data from various sources on wind farms and home values to build a machine learning model for predicting the impact of windfarms on home values. The model will calculate the year-over-year changes in home value for several years prior to building the wind farm and the year-over-year changes subsequent to the wind farms arrival.

Client:

Home owners, prospective home buyers, and local government officials interested in tax revenue can use this model to predict the likelihood that a wind farm in their area will impact home prices. They can then take steps, such as decide to not move to an area, petition their city to not go forward with a wind farm, or in the case of city officials insist on revenue from the wind farm that will compensate for the loss in tax revenue from homes.

Dataset:

The wind farm data will be acquired from the US Wind Turbine database. This data contains numerous information about each wind mill in each wind farm; of interest for this study is the windmill Id, project id, location, installation date, and number of windmills in the project.

In addition, dataset/s containing information (location, year built, assessed value by year) about homes is needed for this study. The final datasets to use are being evaluated with leading candidates being Zillow and Redfin.

Dataset parameters:
● The geographic area of study will be the states of Illinois and Indiana
● The dataset timeframe for wind farm data is farms constructed in the years 2000 – 2010.
● The dataset timeframe for housing data is 3 years prior and 3 years after a wind farm's construction

● Houses that are within 25 miles of a wind farm will be evaluated

Approach:

Since the objective is to *predict a home's change in value* due to arrival of a wind farm, a regression algorithm will be used to build the model in a supervised manner. The algorithm will provide a prediction of the percentage change in home value.

The predictor variables will be…
● windmill (windmill id, project id, location, installation date, and number of windmills in the project)
● houses (location, year built, assessed value by year)

To avoid overfitting the models, data will be divided into a training set and a test set; using a ratio of 80 percent of the data in the training set and the remaining 20 percent in the test set. The model will be built with the training set, and then the test set will be used to evaluate the model by pretending that the test-set data points are unknown. This will enable the accuracy of the model to be evaluated by comparing the categories assigned to these test-set data points by the model, to the true categories derived from the training data. (*overfitting approach is based on recommendation from Lilian Pierson's, "Classification Algorithms Used in Data Science"*)

Deliverables:

1. Milestone report on the capstone project
   a. Data wrangling
   b. Data story
   c. Exploratory data analysis

2. Final report
   a. In-depth analysis
   b. Final paper
   c. Presentation/slide deck
   d. Online video or blog post