Nicholas Brower
Springboard DSCT – May 2022
Unit 24.5 Capstone Three: Project Ideas

1. Microbusiness Density Forecasting

Model type: Regression
Tags: time series, multivariate, historic, financial, demographic, education, business

Use changes in population, demographic, education, and financial data to predict changes in the microbusiness density of US counties. Predictions would leverage multivariate timeseries data to predict the number of microbusinesses per 100 adult residents of each county in the test set. The dataset provided by Kaggle would be augmented by US Census data, world bank data, historic economic data, local sales data, and potentially other sources to build a range of models based on the availability of data per locale.

Kaggle dataset description
https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/data

2. Philadelphia air quality forecasting

Model type: Regression
Tags: time series, live, environmental, weather

Use changes in daily and hourly air quality index sensor measurements combined with hourly and historic weather data to predict changes in air quailty for Philadelphia or the greater Philadelphia area. PM2.5, PM10, ozone, and other pollutant concentrations are available at regular intervals from a variety of sources. Possibly augment data with text mining of the social media accounts of fire departments of places determined by current wind speed and direction relative to the prediction location. Build a model that renders hourly predictions for the next 24 hours and for the next seven days.

IQAir offers detailed 48 hour data and 30 days of average data. A guide to using this data is available at the link below.
https://support.iqair.com/en/articles/4939698-how-can-i-access-historical-data-on-the-iqair-platform

An example page for Philadelphia
https://www.iqair.com/us/usa/pennsylvania/philadelphia

3. Book Review Rating Prediction

Model type: Regression
Tags: natural language processing

Use the contents of a written book review to predict the rating issued by the reviewer. Available data includes the book name, review date, number of comments, number of upvotes from other users, and the text of the review itself. Data may be augmented by reviews of the same book on other sites.

Primary dataset
https://www.kaggle.com/competitions/goodreads-books-reviews-290312/data