Project Deliverable

Group Member: Daniel Clepper, Gaomingyu Fan, Darren Tung, Barry Yao, Julia Mengxuan Yu

1. Background and context to the problem statement.

American policy leaders aim to foster more inclusive and resilient economies in light of technology advancements that have made entrepreneurship increasingly accessible. Many Americans are choosing to start their own businesses for various reasons, but these "microbusinesses" (businesses with fewer than 10 employees, typically based online) often elude traditional economic data sources, posing a challenge for policymakers. Understanding where these microbusinesses are centrally located can help spur policy creation to incentivize growth, allocate resources such as capital and business development services, and focus recovery efforts in times of economic downturn. GoDaddy has collected microbusiness survey data over the past 4 years, and hosted a Kaggle Competition from December 2022 to June 2023 asking competitors to find advanced approaches to predict the microbusiness density (Microbusinesses per 100 people) for a given county.

1. Identification and description of the data set(s) along with their source.

Data Source: <https://www.kaggle.com/competitions/godaddy-microbusiness-density-forecasting/overview>

There are three datasets and we will merge them together.

**train.csv:** The training dataset has 7 features and 122266 observations. It has information about the raw count of microbusinesses in each county and the target variable microbusiness density.

**test\_csv. :** The test dataset contains 3 features and 25080 observations. Features include ID code, a unique identifier for each country using the Federal Information Processing System and the date of the first date of the month.

**census\_starter.csv:** This dataset has 6 features and 3143 observations. It has examples of useful columns from the Census Bureau's American Community Survey (ACS) at [data.census.gov](https://data.census.gov/). Features include the percentage of households in the county with access to broadband of any type, the percent of the population in the county over age 25 with a 4-year college degree, the percent of the population in the county born outside of the United States, the percent of the workforce in the county employed in information related industries, and the median household income in the county.

1. Proposed ML techniques you are proposing on applying to solve the problem

From an outlook of the dataset, several forecasting ML techniques could be applied such as ridge/lasso linear regression, random forest regressor, gradient boosting regressor, XGboost, SVR, and LightGBM. The strong temporal component that our dataset innately carries also speaks strongly of applications of time-series models. Given the variety of applicable techniques, there is no strong tell of which technique or combination of techniques are most effective until we are fully involved in a machine learning workflow. However, the plan is to start with a simpler model such as linear regression for baseline performance and work our way to the more complicated models, fine-tuning them based on cross validation results.