Projection of GDP Growth in Conjunction with Internet Accessibility

Phillip Harmon, 801177304

- Project Manager
- Lead Engineer
- Data Analyst
- Secretary

Project Summary

This project aims to develop a machine learning model to analyze and predict growth of a nation's gross domestic product according to the availability and growth of internet access within that country. This model will allow for the projection of timeframes for developing nations becoming major players in the world economy based on their growing internet infrastructure and userbase. Additionally, this will give insight into the importance of internet access to all people.

Selected Datasets

- *Speedtest Data by Ookla*
  - Contains data on internet speeds across many countries, broken down by quarters, for the years 2020, 2021, and 2022
  - Data Source: World Bank National Accounts Data and OECD National Accounts Data Files. (https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG)
  - Dataset: https://www.kaggle.com/datasets/dimitrisangelide/speedtest-data-by-ookla
- *GDP annual growth for each country (1960-2020) NEW*
  - Contains annual GDP data for many countries from the years 1960 through 2020.
  - Data Source: Ookla's Speed Test Service (https://www.speedtest.net)
  - Dataset: https://www.kaggle.com/datasets/zackerym/gdp-annual-growth-for-each-country-1960-2020

Training and Evaluation Plan and Expected Results

The model will be a regression model whose output will be a GDP value for a given country. The current plan for training this model is to perform training with a standard gradient descent regression and a support vector regression model with kernelization, taking the best results of the two. The gradient descent provides a more algorithmically robust training procedure, where SVR can better handle nonidealities such as dependence and nonlinearity within the input data.

Both trainings will involve appropriate input scaling and feature selection. One important note is that the data from the two datasets will need to be merged according to year, while removing data from the GDP dataset that are not represented in the Ookla dataset. The primary example of this is that the GDP dataset spans many more years than the Ookla dataset, and it is

currently unknown if all nations present in the GDP dataset are identified within the Ookla dataset.

Due to the nature of the data being analyzed, it seems desirable that all of the data be used in training, as all of the datapoints present should hold significant value to the model. With this in mind, it is planned that cross-validation will be used to evaluate the model, rather than a regular train-test split.

Following the training of the model, sample data will be fed into the model in order to generate projections of the following handful of years for an as-of-yet undetermined set of countries.