Projection of GDP Growth in Conjunction with Internet Accessibility

Phillip Harmon

***Abstract*—A machine learning model has been created to predict a nation's GDP growth in conjunction with an change in internet speeds.**

All materials can be found in the project repository at https://github.com/pharmon0/Harmon_ECGR5105_FinalProject.

I. Introduction and Motivation

It is irrefutable that widespread access to the internet has been a major boon to developed societies. This much is certain. The aim behind this project was to develop a model to observe and predict this relationship.

II. Approach

To tackle this problem, data was gathered pertaining to national GDP growths and internet speeds from various countries. This data was then fed into a machine learning model which trained itself to the data. This model was then able to take in internet speed data and give out a value, indicating the change in GDP due to these changes in internet speed.

III. Dataset and Training Setup

A. Dataset Selection

Two datasets were selected to serve as the source of information for training this model. The first was a set of internet connectivity information sourced from Ookla's internet speed test service at https://www.speedtest.net [1]. This dataset included internet upload speed, download speed, and latency information labelled by country and year. It also included some extra metadata such as the number of devices tested, the number of tests in each sample, the country's average population as of 2005, and the country's internet speed ranking among all nations. The data in this dataset spanned the years 2020 to 2022, broken down quarterly, and broken down by fixed connections and mobile connections.

The second dataset listed annual GDP change by country sourced from World Bank national accounts data and OECD National Accounts data files (https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG) [2]. This dataset included annual change in GDP for a number of countries over the range of 1965 to 2020. Each entry in the dataset was labelled by country and contained a number of metadata values such as the 3-letter country code, the name of the economic indicator (GDP US$), and its ID code.

Both of the selected datasets were sourced from the dataset library at https://www.kaggle.com.

B. Feature Selection and Dataset Cleaning

Both of the selected datasets from Section III-A contained unnecessary data for the training of this model. These extra data points needed to be purged from the sets so that the training could be performed. Additionally, both datasets contained missing data entries. These missing entries then also needed to be removed.

A Python script was developed to match data entries in the two sets, drop unnecessary features, and remove entries with missing data. This script then output a single dataset '.csv' file containing only the country name as a string, the boolean mobile vs. fixed connection indicator, the yearly change in internet upload speed, the yearly change in internet download speed, the yearly change in internet latency, and the yearly change in the nation's GDP. All data in this cleaned set was only for the year 2020, as this was the only overlapping year in the two datasets. All of the annual change in internet quality data was obtained by taking the differences of the 2021 data and the 2020 data by quarter. In the end, the final cleaned data contained 6 features, including the GDP label as well as the country name, and there were 1220 samples in the dataset.

The first ten entries of the final dataset were placed in Figure 1 for example.

	Mobile	Country	Upload	Download	Latency	GDP Growth
0	0.0	Afghanistan	428.0	356.0	-20.0	-1.934778
1	0.0	Albania	5327.0	7265.0	-6.0	-3.311239
2	0.0	Algeria	-133.0	1447.0	-15.0	-5.480992
3	0.0	Andorra	55214.0	57985.0	3.0	-11.956058
4	0.0	Angola	439.0	-742.0	15.0	-4.040510
5	0.0	Antigua and Barbuda	147.0	909.0	-31.0	-15.973510
6	0.0	Argentina	3735.0	3610.0	-8.0	-9.905235
7	0.0	Armenia	4755.0	4156.0	-6.0	-7.600000
8	0.0	Australia	2468.0	14392.0	0.0	-0.284839
9	0.0	Austria	3096.0	11766.0	-4.0	-6.259035

Fig. 1. The Final Cleaned Dataset

C. Data Preprocessing

Before being used for training a model, the cleaned dataset was put through some standard data pre-processing measures to ensure a quality training experience for the model.

The first of these measures was data normalization. The continuous-range values from the dataset (download speed,

upload speed, latency, and GDP changes) were all normalized through min/max scaling, implemented by subtracting the minimum of each column from each entry within that column, and then dividing each entry by the difference of the maximum and minimum of that column. This ensures all data to have a consistent range between zero and one.

In addition to this, the data was divided by a train-test split. Eighty percent of the data was used for training the model, while the remaining twenty percent was only used to validate the model. This split was performed after shuffling the dataset, so that there would not be any structural interference with the training and validation.

Finally, the country names were discarded from the set, as they are irrelevant to the training, and the GDP data was extracted as the ground truths. The remaining four-featured data would serve as the inputs to be trained on. With this complete, the data was ready for training.

D. Model Selection and Training

The selected model to be trained on this internet access and economic data was a fully-connected artificial neural network. The specific architecture chosen was a network with four hidden layers. These hidden layers contained sixteen, thirteen, nine, and four layers respectively. Each of these hidden layers used a hyperbolic tangent activation function to allow the model to train against non-linearities. Additionally, there is a final output layer which consists of a single linear node with no activation function. A diagram of this neural network architecture can be seen in Figure 2.

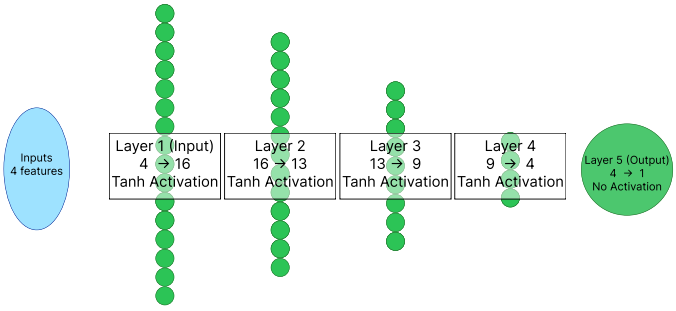

Fig. 2. The Selected Neural Network Architecture

Since there were four input features (mobile indicator boolean, upload speed change, download speed change, and latency change), and each node contained a bias value, in addition to its scaling factors, the entire network contained a total of four-hundred seventy-two trainable parameters.

IV. Results and Analysis

By experimentation, using the Adam optimizer, it was discovered that a quite small learning rate was required to ensure the training would converge. The final learning rate selected was 1×10^{-6}. Unfortunately, this small of a learning rate also neccesitates a large number of epochs for the training. According to the graph in Figure 3, the mean square error seems to converge around the 70,000th epoch, but the model was run for an entire 100,000 epochs to ensure proper convergence over various test conditions.

The final training took three minutes and thirty-four seconds to complete all 100,000 epochs. This resulted in a final training loss of 9.16×10^{-3}, a final validation loss of 6.06×10^{-3}, and a final loss over the whole dataset of 8.54×10^{-3}, all in terms of mean square error.

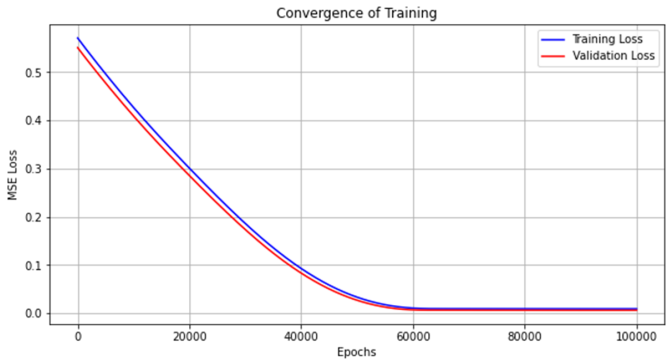

Fig. 3. Model Training Convergence

At the end, the model was used to show the output change in GDP for some 'naively independent' input data. This data was naive in that for each test, one of the inputs was given values in the range of the original dataset, while the other features were filled with zeros. With these 'naively independent' inputs, the output indicates the change in GDP for a nation corresponding to only a change in one of the internet metrics at a time. This does not give a full picture of the model, but it allows for some characterization of it. These output plots can be found in Figures 4, 5, and 6.

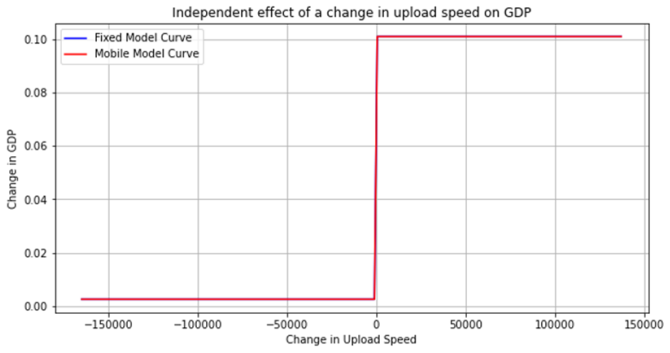

Fig. 4. GDP Growth due to a 'Naively Independent' Change in Upload Speed (kbps)

All materials for this project, including code, images, dataset files, and documentation, have been placed in a dedicated Github repository located at https://github.com/pharmon0/Harmon_ECGR5105_FinalProject.

V. Lessons Learned

The ideal planned outcome of this project was to produce a graph of National GDP growth by country over a span of time, and to project it further according to predicted internet

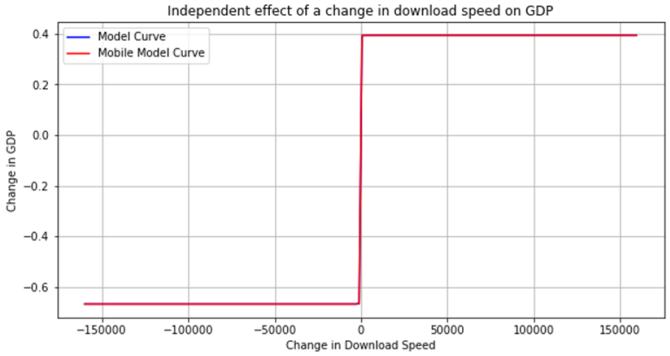

Fig. 5. GDP Growth due to a 'Naively Independent' Change in Download Speed (kbps)

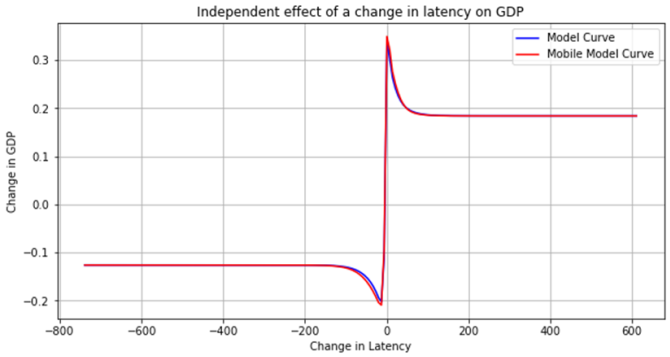

Fig. 6. GDP Growth due to a 'Naively Independent' Change in Latency (ms)

access growth by year. Unfortunately a number of things got in the way of this ideal.

One of those areas that needed improvement in this project was data selection. Quite simply, the selected datasets only shared a single overlapping year. Ideally more overlapping years would have been used to create a better model, and to help plot the relationship mentioned above.

Additionally, and unfortunately, time was a concern in completing the project. There was plenty of time allotted to do better work on this project, but as the sole member of the project team was sick for two weeks of the allotted five, all time commitments, including this project, were impacted heavily. Those two weeks of time could have provided sufficient time to find more data to supplement the project, and they could have also given the head-start to have the model trained with sufficient time remaining to produce better analyses at the end.

Finally, it was originally planned that this project would compare the neural-network results against a linear regression model and a support vector regression model. Once again due to time constraints, these components of the project were not completed in time to be analyzed for this report.

On the positive side of things, the quite clean convergence of the model was a pleasant surprise. This means that there exists a clean relationship between the GDP and internet access, which is a positive sign for the base concept of this project.

REFERENCES

[1] D. Angelides, *Speedtest data by Ookla*, Oct 2022. [Online]. Available: https://www.kaggle.com/datasets/dimitrisangelide/speedtest-data-by-ookla

[2] Z. M, *GDP annual growth for each country (1960-2020) new*, Nov 2021. [Online]. Available: https://www.kaggle.com/datasets/zackerym/gdp-annual-growth-for-each-country-1960-2020