Predicting Economic Health with Financial Indicators

Group 1 Executive Summary Report

Daniel Moultrie, Luke Everson, Elwood Olson, Douglas Byers
GENESIS 10 Dev10 Data Cohort

Introduction

Commented [EO1]:

For our capstone project, we decided to determine whether large-scale financial indicators can be used to predict a country's economic health on an individual scale. In other words, we want to know if abstract-seeming numbers like GDP and the total value of stocks traded are predictive of more tangible metrics that affect the everyday person, like a country's unemployment rate or level of income inequality. We wanted to know: does the stock market *really* have any connection to my individual quality of life?

Especially since several of us are new to the finance industry, these questions occurred to us naturally. The importance of indicators such as the consumer price index or a country's total reserves initially seemed vague and hard-to-grasp to us. However, over the course of this project, we gained an appreciation for the meaning of these figures, as well as their predictive insight.

In order to answer our questions, we decided to use data from the World Bank's open databank. Pulling a broad swath of economic, quality of life, and financial indicators from every country in the database, we created a sturdy collection of data about the connections between economics and finance over the past thirty years. With this data in hand, we applied machine learning models to help us answer the following guiding questions:

- How well can quality of life for individuals be predicted by large scale economic indicators?
- How well do large scale economic indicators predict wealth inequality?
- Can short-term economic data be used to effectively predict quality of life indicators?
- What are the most important economic indicators for predicting quality of life and wealth inequality?
- Can we accurately fill in missing economic data such as the Gini Index from the World Bank using financial indicators?

Research Process

The initial research stages were hindered by our individual group members' lack of familiarity with the financial sector. Initially, our plan was to take on a main problem similar to the problem we posed for this project regarding perceived economic health from large scale factors and how it relates to economic health on a more individualized scale, but we wanted to directly examine the stock market itself, including indices like the S&P 500, the Dow Jones, the NASDAQ, and similar indices on an international level. However, we soon realized that our lack of understanding of the financial sector and the stock market proved too large a roadblock to form any meaningful analysis from that data.

From here, we wanted to maintain the same basic subject, but look at it from a broader point of view. That led us to the problem and questions we settled on. Can large scale economic and financial factors be indicative of small-scale economic health and social well-being? This new main topic and the set of questions that came with it naturally brought us to the World Bank Databank as our main data source. With dozens of financial, social, and economic indicators from all over the world, the only problem the World Bank Databank posed was selecting the proper indicators to use as explanatory variables and choosing meaningful indicators to use as response variables.

Since the goal of our project was to examine the value of financial and economic factors in predicting variables associated with socioeconomic health and well-being of a country, we had to narrow down our choices for dependent variables to fit this topic. We eventually chose the Gini Index[1], Unemployment rate, Poverty rate, Income share held by the top 10%, and income share held by the bottom 10% as our economic response variables, with life expectancy at birth and infant mortality rate serving as the response variables for the more social quality of life and well-being side of the project.

Selecting our predictive variables posed more of a challenge purely because of the volume of data available from the World Bank. First, since we wanted to look at broad scale economic and financial factors, it was natural to choose some measure of GDP to be included in our analysis. We decided GDP per capita, as opposed to raw GDP, would be the best measure to use for our end goal. Also, since our initial project plan was to examine the stock market, we found variables that calculated the value of the stock market as a percentage of the GDP, allowing us to keep part of the original plan within our new proposal. With the goal of our project being to examine economic and financial health on a smaller scale, we found it logical to use how much money is readily available within a country's economy as a predictive variable, which led us to include broad money[2] and total reserves[3]. In addition to choosing variables that provide insight into how much money is readily available within an economy, we thought it would be useful to examine a variable that accounted for the spending power of the money in that economy, which prompted us to include inflation within our set of predictors.

Next, we thought we should incorporate government financial data into our predictors as well. Tax revenue and central government debt ended up being the measures we deemed would best fit within the scope of our questions. To round out the variable selection, we decided on international migrant stock and total exports. International migrant stock is essentially a proportion of the workforce made up of workers born in a foreign country. We thought international migrant stock would not only be a decent indicator of a healthy economy, since international workers often migrate for increased financial and economic opportunity, but also

[1] Gini index measures the extent to which the distribution of income (or, in some cases, consumption expenditure) among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus, a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.

[2] Broad money is the sum of currency outside banks; demand deposits other than those of the central government; the time, savings, and foreign currency deposits of resident sectors other than the central government; bank and traveler's checks; and other securities such as certificates of deposit and commercial paper.

[3] Total reserves comprise holdings of monetary gold, special drawing rights, reserves of IMF members held by the IMF, and holdings of foreign exchange under the control of monetary authorities.

would serve as potential insight into the social quality of life aspects of the analysis, since increased quality of life is typically the other major motivating factor in immigrating to an entirely new country. Exports as a percentage of GDP were chosen as our final predictive measure because they offer a tangible measure of how a specific economy operates, which is a realm so often dominated by obscure or hollow numbers.

With our variables selected, we wanted to use as large of a time frame as possible to be able to produce as many results as possible. Examining the data showed us that data before 1990 becomes very sparse, so our analysis includes all available data for countries registered in the World Bank Databank for the years 1990-2020 (2019 in the case of some variables, with 2020 data collection disrupted due to the ongoing Covid-19 pandemic).

As previously stated, all this data was selected from the World Bank Databank, which can be easily accessed through their free API. The documentation and process for calling that API and collecting the data itself into a useable form for the purposes of our analysis are outlined within our ETL report.

After the procedure of cleaning and processing our data, we ended up not having enough data to build meaningful models for Poverty rate, Income share held by the top 10%, and Income share held by the bottom 10%. From there, we carried out the initial steps of building our analysis for the remaining response variables: Gini Index, Unemployment rate, infant mortality rate, and life expectancy. However, we reached a point where we realized our analysis was creating too many steps and we needed to narrow down our focus. Since the Gini Index was our main predictive variable from the start, we began by deciding to keep all work regarding our Gini Index predictions. From there, we decided to keep unemployment and leave off our models and predictions for infant mortality rate and life expectancy since unemployment and the Gini Index are both directly related to the economy, rather than social demographics.

Desired Outcomes

Gini Index

After selecting the Gini Index as a valuable indicator of economic health for a country, we noticed that the Gini Index had numerous missing values for many countries throughout the years provided on the World Bank site. With knowledge of the use of models for filling in blank values for other indicators on the World Bank site, we determined that a valuable project outcome would be the prediction of the missing Gini Index values for the World Bank site. To create a machine learning model for the purpose of filling in these missing values, we set out to determine the predictive power of our previously selected financial indicators on the Gini Index.

Unemployment

Meanwhile, whereas many Gini Index values were missing from the World Bank data platform, the unemployment rate was available for the majority of countries for almost every year between 1990 and 2020. Since there was so much data available in this column, we decided that we would try building a machine learning model that predicted a country's future

unemployment rate based on data from the previous year. Specifically, we hoped that we would be able to use our model to predict global unemployment rates in 2021 based on financial data from 2020. To achieve this predictive effect, we staggered our dataset, pairing economic data from the year *n* with the unemployment rate from the year *n+1*.

Machine Learning Process

Before creating models, further cleaning and transformation of the data was necessary. This cleaning and transformation included cutting out rows and columns with too many missing values, standardization, and handling of outliers. To cut down on the overuse of missing or imputed data, we decided to remove any column or row in which more than half of the data was missing. After removal of columns and rows with more than half of their data missing, we were left with the listed inputs and responses for our model.

Input Variables	**Response Variables**
• Stocks traded, total value (% of GDP) • Inflation, consumer prices (annual %) • Tax revenue (% of GDP) • GDP per capita (current US$) • Broad money (% of GDP) • Total reserves (includes gold, current US$) • Exports of goods and services (% of GDP)	• Gini Index • Unemployment, total (% of total labor force) (year *n + 1*)

Now that the specific predictors were selected, we could begin the process of training and testing specific types of models. To be able to train and test models, the values for the predictors needed to be complete for each row. This caveat required us to either cut down the number of rows to only include the rows that had complete values, or to impute the missing values. We first attempted to only use the rows that had complete data for both the input and response variables. This process significantly cut down on the number of viable rows that could be used for the models. After consultation with our mentor, we decided to impute the missing values for our inputs using the K-Nearest Neighbors method. To perform K-Nearest Neighbors, we had to standardize our data.

Initially, we selected linear models, including Lasso and Linear regression, due to the continuous nature of our selected input and response variables. After running these models on our standardized dataset and receiving low accuracy on our test set, we determined that our data was not entirely linear and that there must be a better model for our data. This realization led us to the Random Forest Regression model. Unlike linear models, the Random Forest model accounts for inputs not being entirely independent of one another, which allows for greater accuracy. Also, this model can be run on unstandardized data which allows for easier interpretation of results.

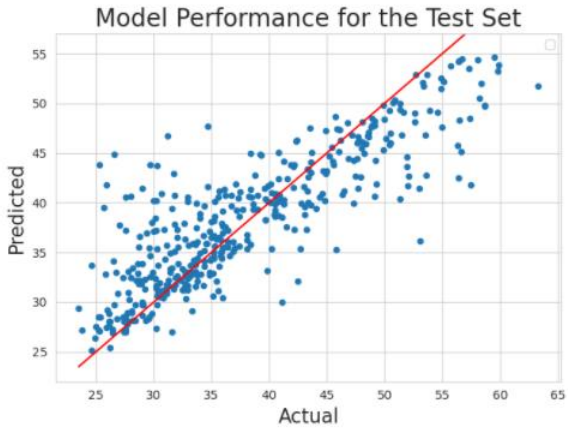
To use this model, we unstandardized our imputed dataset. We decided not to eliminate outliers from our dataset as there is not a systemic reason for these outliers. That is to say, these outlier points were directly measured by the World Bank and our model could derive value from these points. After inputting our final dataset into the Random Forest regressor, we found that the accuracy of predicting the outcomes for our test set was significantly better than our initial attempts with Linear models. This increase in accuracy led us to select the Random Forest model as our final model to predict our dependent variables.

Now that we had selected a final model type, the models needed hyper-tuning to increase model performance. To hyper-tune our model, we performed a random search and a grid search. To narrow down the fields to base our grid search on, we first performed the random search for all possible parameters. Using the outputs from our random search and advice from our mentor, we cut down the possible parameters and used a grid search to determine the best hyperparameters to use. We used the resulting hyper-tuned model provided by the grid search as the final model to base our predictions for both the Gini Index and the Unemployment Rate. To gather these predictions, we ran the model for the Gini Index on all rows in which the predictors were present and the Gini Index was missing. For the Unemployment Rate, we ran the model for every country to compare the model predictions to the values provided by the World Bank and provide our prediction for 2021.

Results

Gini Index

The graph below demonstrates how accurately our final Gini Index model predicted the test set:



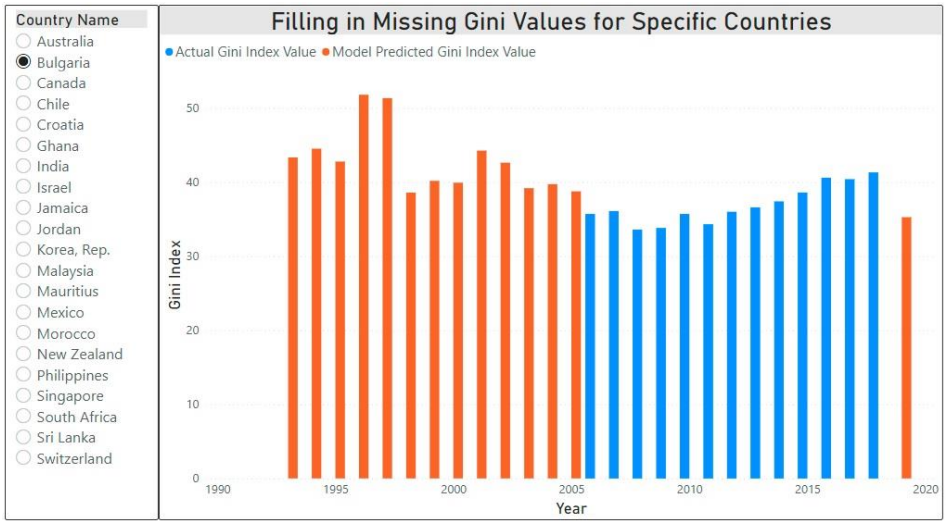
Here, the red line represents the points for which the actual and predicted values are equal. Points lying above the line are those for which the predicted value is higher than the actual value, and points lying below the line are those for which the actual value is lower than the predicted value.

Besides some scattered outliers, most of the predictions on the graph are within 5 index points away from the actual value. Overall, this model performed with a score of 0.71.

In the following chart, we've broken down the relative importance of each of the independent features for the model.

Features	Importance
• Stocks traded, total value (% of GDP)	• 0.1342
• Inflation, consumer prices (annual %)	• 0.0438
• Tax revenue (% of GDP)	• 0.1180
• GDP per capita (current US$)	• 0.3569
• Broad money (% of GDP)	• 0.0848
• Total reserves (includes gold, current US$)	• 0.1036
• Exports of goods and services (% of GDP)	• 0.1587

With an importance of about 36%, GDP per capita is the most important feature in the model. This result stands to reason, because wealthier and more developed economies are less likely to have severe income inequality problems. Yet, even though GDP is the most heavily weighted feature in the model, it is by no means the only variable contributing to the model's prediction. Much to the contrary, about 64% of the prediction is determined by the other features in the model. Hence, we can be sure that our model is making a broad-based inference and productively using all the data that we've fed into it.

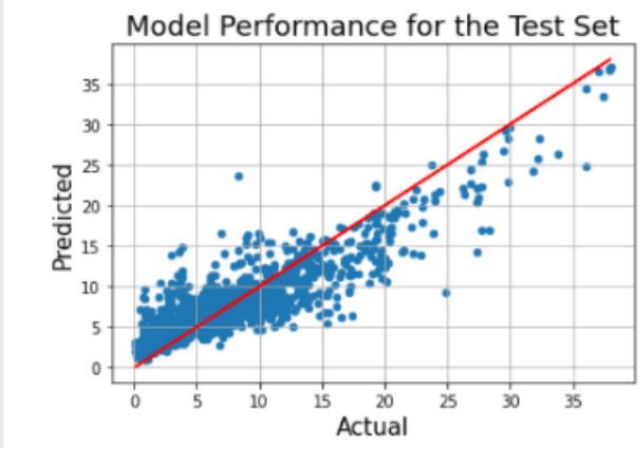


The graph above displays the ability of our model to fill in the missing values for the Gini Index on the World Bank site. In the example above the missing values for the country, Bulgaria, are filled in. These predicted values of the Gini Index would allow any viewer of the world bank site a more complete account of wealth inequality in a country from the years 1990-2020. These predictions could be used to supplement further research and provide users with a rough estimate of what income inequality looked like for a year and country in which the value was not provided.

While our model provides value in predicting the missing Gini Index value, it is not perfect. With an accuracy of 0.71, our model predicts the value accurately more often than not, but our model is susceptible to mistakes. As stated previously, our model only considers a select group of financial factors. This means that there are numerous other social and financial factors that could also play a role. Factors that we did not account for such as political structure, region, and organization of labor among others, could play roles in predicting the Gini Index of a country. Further research and testing over a longer time period could be done to account for all factors and provide the best predictions of the Gini Index.

Unemployment Rate

In the following graph, we've illustrated the performance of the model on the data that we set aside for testing purposes:



Once again, the red line represents points where the predicted value and the actual value are equivalent. Although our unemployment prediction model is imperfect, we can see that the most of our predictions are at least near to the actual value, especially for countries with unemployment rates below 20%. For unemployment rates higher than 20%, our model tends to be less effective, but this error is understandable given that such extreme rates of unemployment

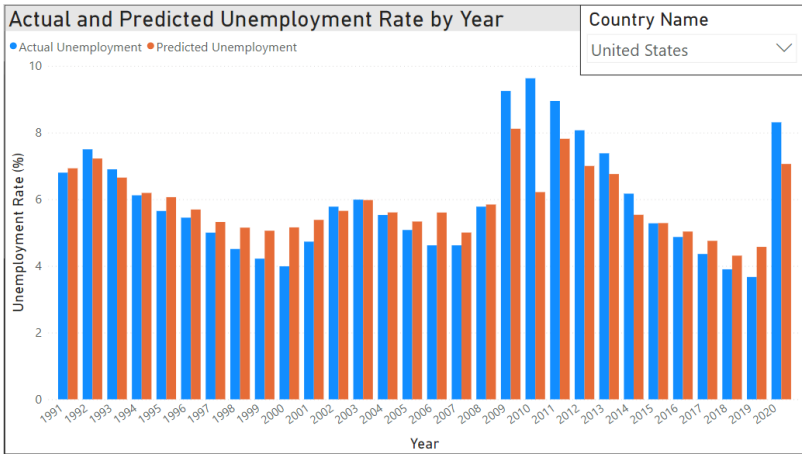
are rare and may be caused by non-financial factors, like natural disasters or political instability. Overall, our model performed on the test dataset with a score of 0.76.

Next, it's important to understand the relative importance of each of our independent variables in predicting the unemployment rate, as tabulated in the following chart:

Features	Importance
• Stocks traded, total value (% of GDP)	• 0.1397
• Inflation, consumer prices (annual %)	• 0.0642
• Tax revenue (% of GDP)	• 0.2814
• GDP per capita (current US$)	• 0.1976
• Broad money (% of GDP)	• 0.1067
• Total reserves (includes gold, current US$)	• 0.0787
• Exports of goods and services (% of GDP)	• 0.1314

From this table, we can see that tax revenue is the most important predictor of a country's unemployment rate. Since employed people are the most important contributors to the tax base, this result is reasonable. Another important feature is GDP per capita, probably because higher-earning countries tend to have more jobs to go around. As was the case for our Gini index model, our unemployment model takes all its features into account, meaning that it's productively interpreting all the information we're feeding into it.

The following chart depicts a side-by-side comparison of the actual unemployment rate for a particular year and the predicted value of the unemployment rate that year over the course of our entire dataset that has both an actual and predicted value, with the blue bars representing the actual unemployment and the orange illustrating the predicted value. As stated earlier, the predicted values are based on the predictive variables from the previous year; the 1991 prediction is based on the 1990 indicators, and so on. The graph below shows the values for the United States.



As we can see, the model deviates farthest from the actual value in the years following large scale societal events, including the economic recession of the late 1990s-early 2000s, the panic surrounding Y2K, the economic recession of 2008-2009, and the global effects of Covid-19 in 2020.

The effect that an event like Covid-19 can have on unemployment is just one of many examples that shows that there are severe limitations to purely economic and financial predictors when looking to quantify larger socioeconomic health. However, the accuracy and consistency the predictions depict outside of those larger scale events show that this model, or one similar, can serve as a roadmap for further analysis.

Conclusion

After settling on a project topic; determining the scope of our research; collecting the appropriate data; deciding on a machine learning method through much deliberation and trial and error; and tuning the models to better predict our response variables, our final models utilized large scale economic indicators to predict wealth inequality and unemployment. The model for the Gini Index (measure of wealth inequality) has an accuracy score of 0.711 and the model for total unemployment has a score of 0.764. These numbers indicate that the models are good at predicting their respective response variables, however it is recommended for both models that research into non-economic indicators be conducted for more accurate predictions. The final models can be used to answer our original exploratory questions as follows:

- How well can quality of life for individuals be predicted by large scale economic indicators?
    - Based on the two models described in this summary, the quality of life for individuals can be predicted fairly well by large scale economic indicators as shown by the accuracy scores of the two models of 71% and 76% respectively. In

the United States, for example, our model is very accurate at predicting unemployment rates, except when large-scale socially disruptive events occur. Although our models cannot foretell major economic discontinuities, they're effective at predicting how things will turn out when business continues as usual.

- How well do large scale economic indicators predict wealth inequality?
    - While our model for predicting the Gini Index had a promising accuracy score of 0.711 for our test set, it was determined that financial factors and other economic data alone are incapable of properly predicting wealth inequality in a variety of different political and socioeconomic regions.
- Can short-term economic data be used to effectively predict quality of life indicators?
    - Unemployment rate, which is one measure of quality of life, can be fairly accurately predicted with short term financial data. (Again, the model has a score of about 76%.) Our model fares especially well for countries having an unemployment rate lower than 20%.
- What are the most important economic indicators for predicting quality of life and wealth inequality?
    - The most important economic indicator for predicting quality of life was determined to be Tax Revenue (% of GDP) which accounted for 28.14% of the predictive power for our unemployment model. The most important economic indicator for predicting wealth inequality was GDP per capita (current US $) which accounted for 35.69% of the predictive power for our unemployment model.
- Can we accurately fill in missing economic data such as the Gini Index from the World Bank using financial indicators?
    - We were able to fill in missing Gini index values with our machine learning model roughly 70% accuracy. Using these predictions to fill in missing values on the World Bank site would provide a more complete picture of wealth inequality for countries between 1990 and 2020.

Recommendations

As previously stated in the results section, our models are not perfect. We recommend further research into enhancing the models using more indicators that could also be predictive of Gini Index and Unemployment such as social and political factors rather than simply using financial factors. Future models that are more accurate in predicting our response variables would be more valuable in successfully completing our intended outcomes.

We also recommend the expansion of the usage of models by the World Bank to predict future outcomes and fill in previous missing values. Our project displays the value of a model in fulfilling these goals. With more time to enhance the accuracy of the models and expand the amount of input and response variables, the World Bank could become a more complete resource that could provide further insight into the financial, social, and political demographics of all countries.