Predicting and Forecasting GDP Using World Bank Data

Capstone Presentation By Mitchell Meislin

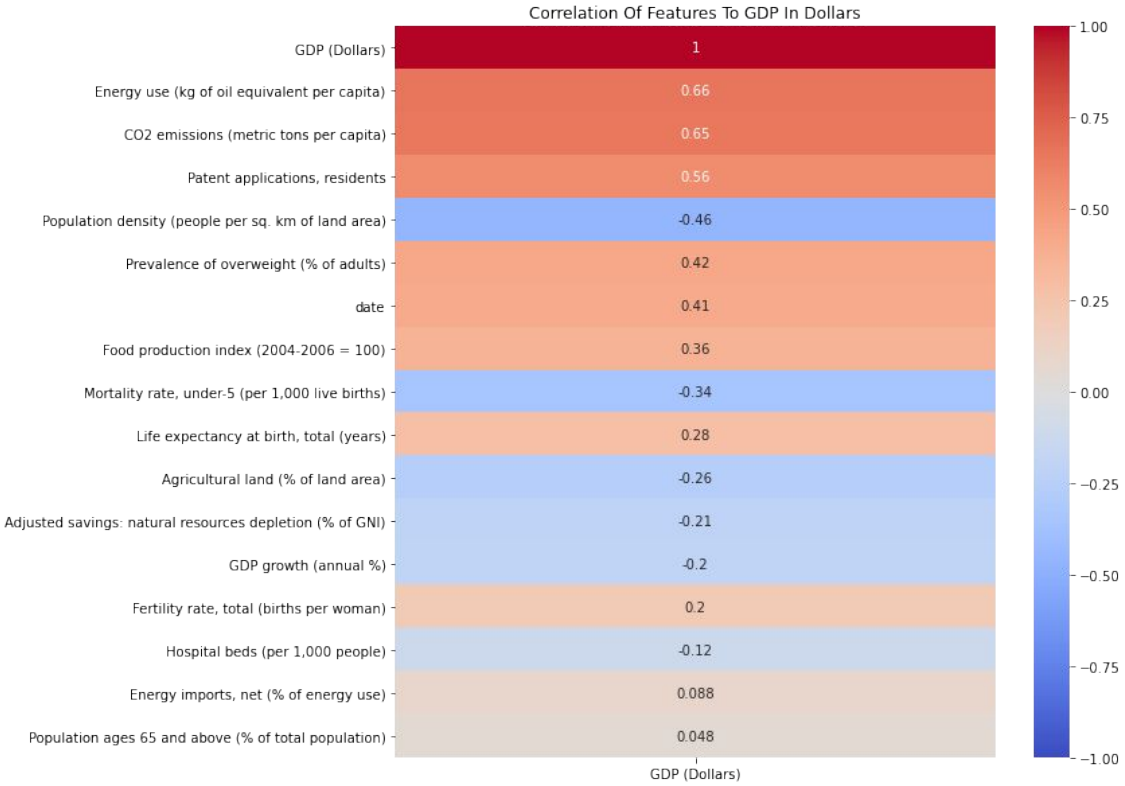
Guiding Question:

How accurately can I predict GDP of a country, without knowing the country name, based solely on political, environmental, financial, and health data? For the USA, how accurately can I forecast GDP by percent change and GDP in dollars?

Data Cleaning and EDA

- Downloaded a 16,000 row dataset from the World Bank with metrics for every country on political, environmental, financial, and health data each year since 1960
- This dataset was filled with an excessive amount of null values
- I cleaned and organized the data into 4 new datasets, each for a specific modeling purpose
- The 4 cleaned datasets were as follows:
    1. Top 5 Ranked Countries By GDP and Their Yearly GDP in Dollars
    2. Top 5 Ranked Countries By GDP and With Yearly Associated Political, Environmental, Financial, and Health data
    3. USA GDP and GDP Percent Change, Yearly Since 1961
    4. USA GDP and GDP Percent Change, Yearly Since 1980 With Associated Political, Environmental, Financial, and Health data

Political, Environmental, Financial, and Health Metrics Used To Predict GDP:



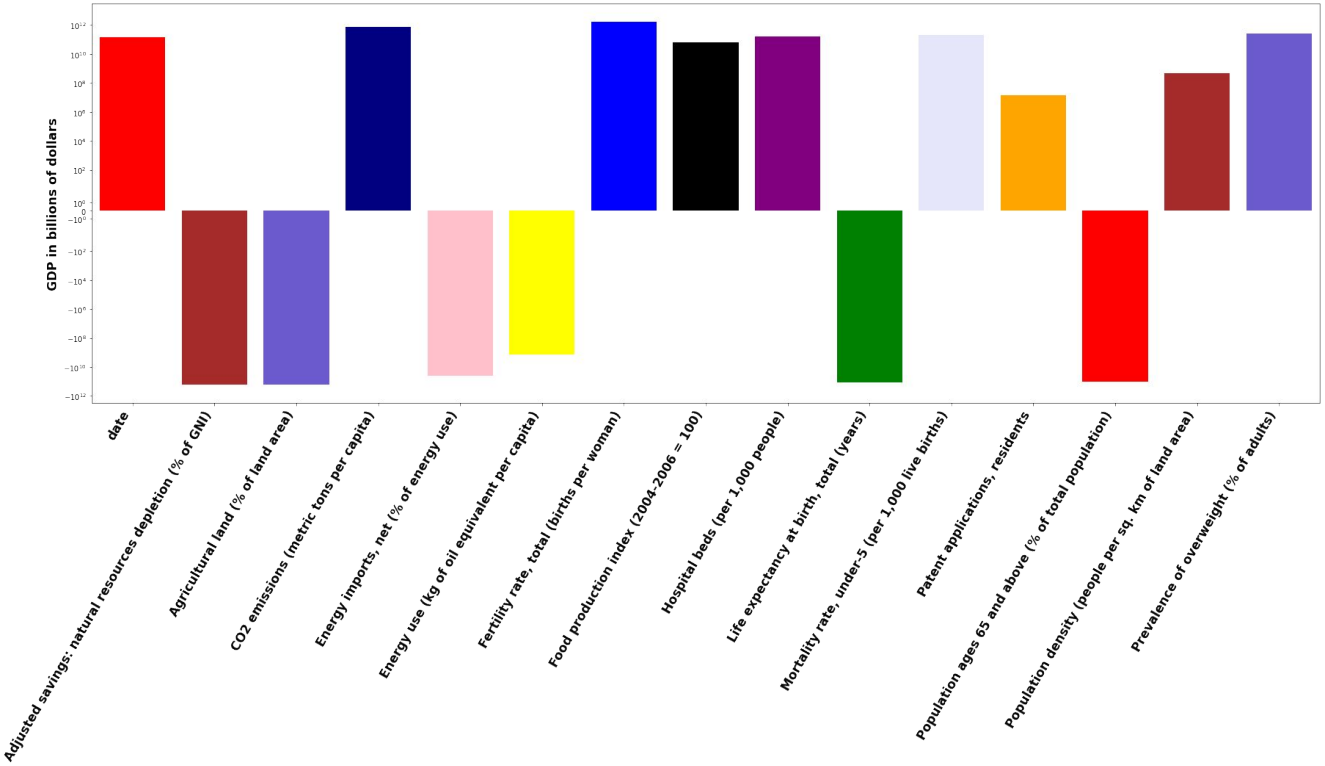
How accurately can I predict GDP of a country, without knowing the country name, based solely on political, environmental, financial, and health data?

- Used The Following Data set:
  - Top 5 Ranked Countries By GDP and With Yearly Associated Political, Environmental, Financial, and Health data
- I stripped away the country name and randomized country and year in training and testing data
- I predicted the unknown countries GDP with a generalized model based on associated political, environmental, financial, and health data.
- Models Used:
  - Basic Linear regression
  - Random Forest With Extra Trees
  - Neural Network

Basic Linear Model

- R2 score of 92% when predicting GDP values that have been withheld.
- Strong association between my predictions and the actual GDP values
    - Model is predicting based on political, environmental, financial, and health data.

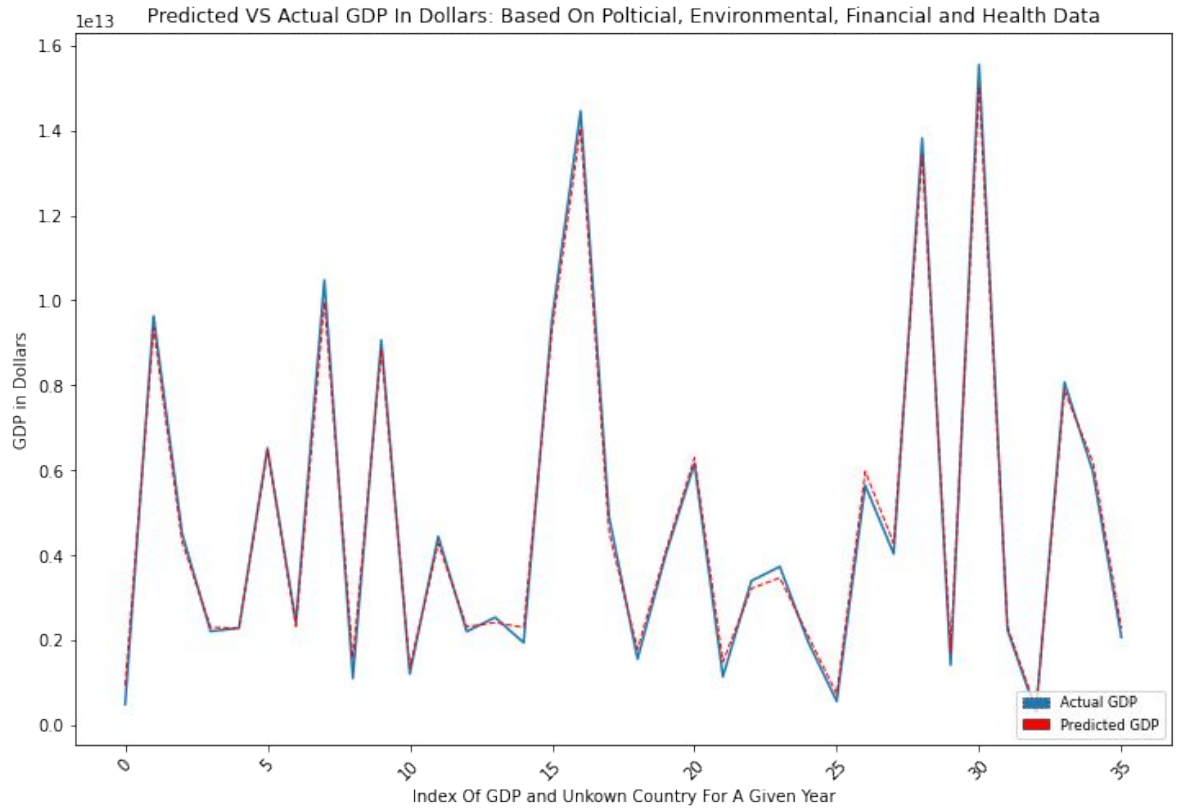
Basic Linear Model: A one unit increase in each element, results on average in a GDP change of the following:



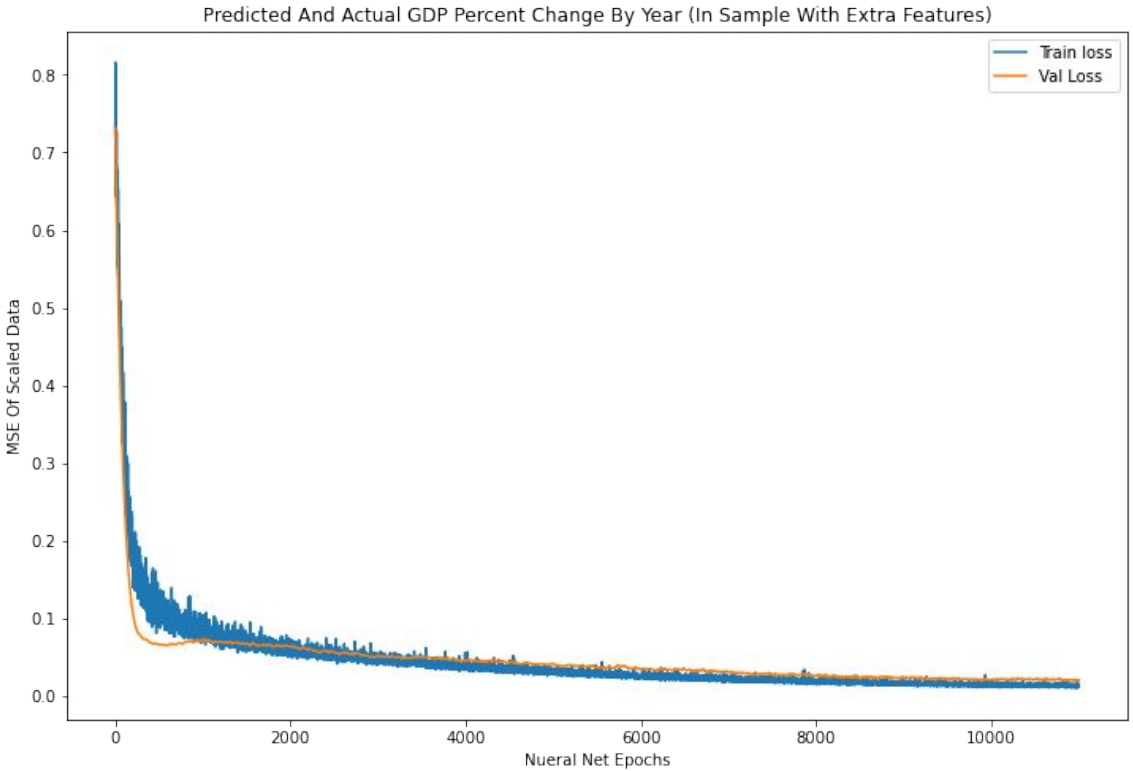
Extra Trees Model

- R2 score of 99.5% when predicting GDP values that have been withheld.
- Strong association between my predictions and the actual GDP values
    - Model is predicting based on political, environmental, financial, and health data.

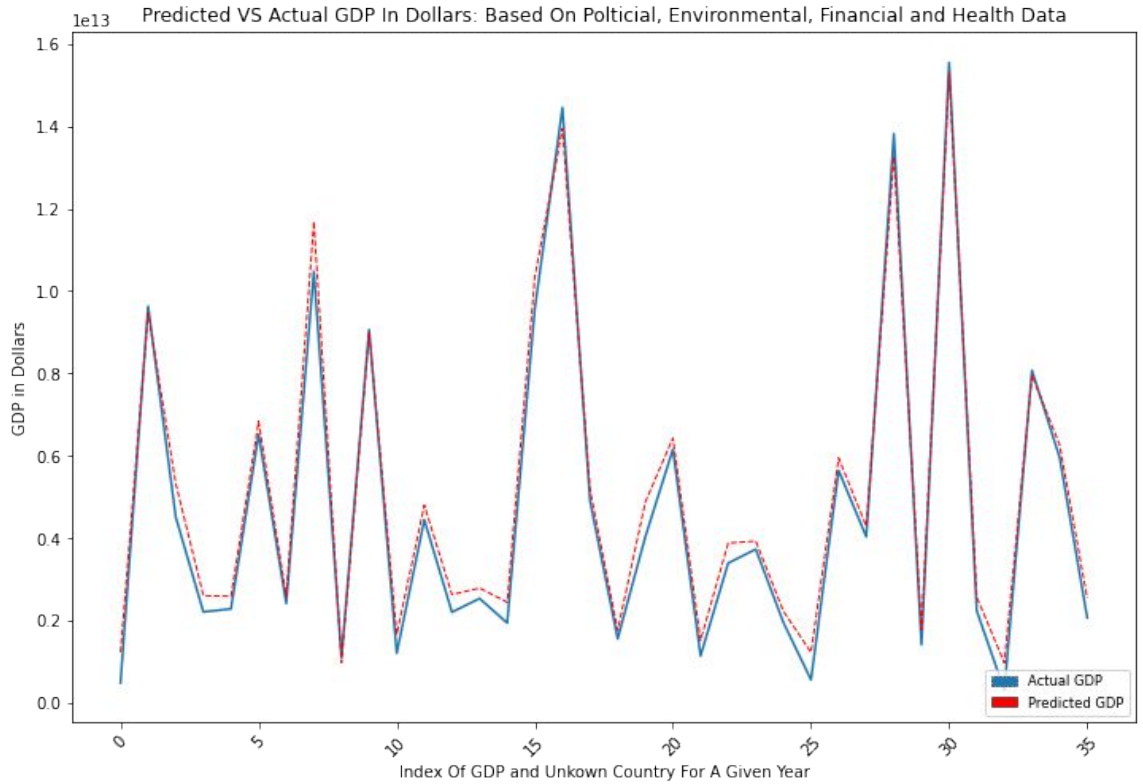
Extra Trees Predictions



Neural Net Model:
Progressive Improvements In Predictions During Training



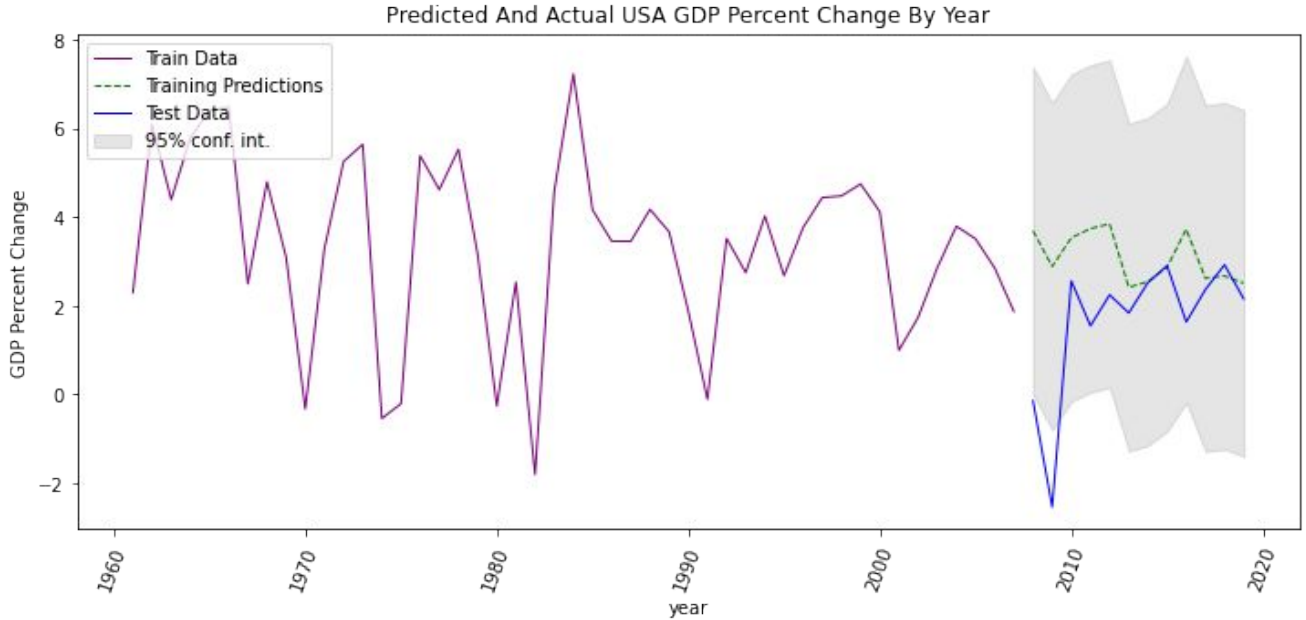
Neural Net Predictions



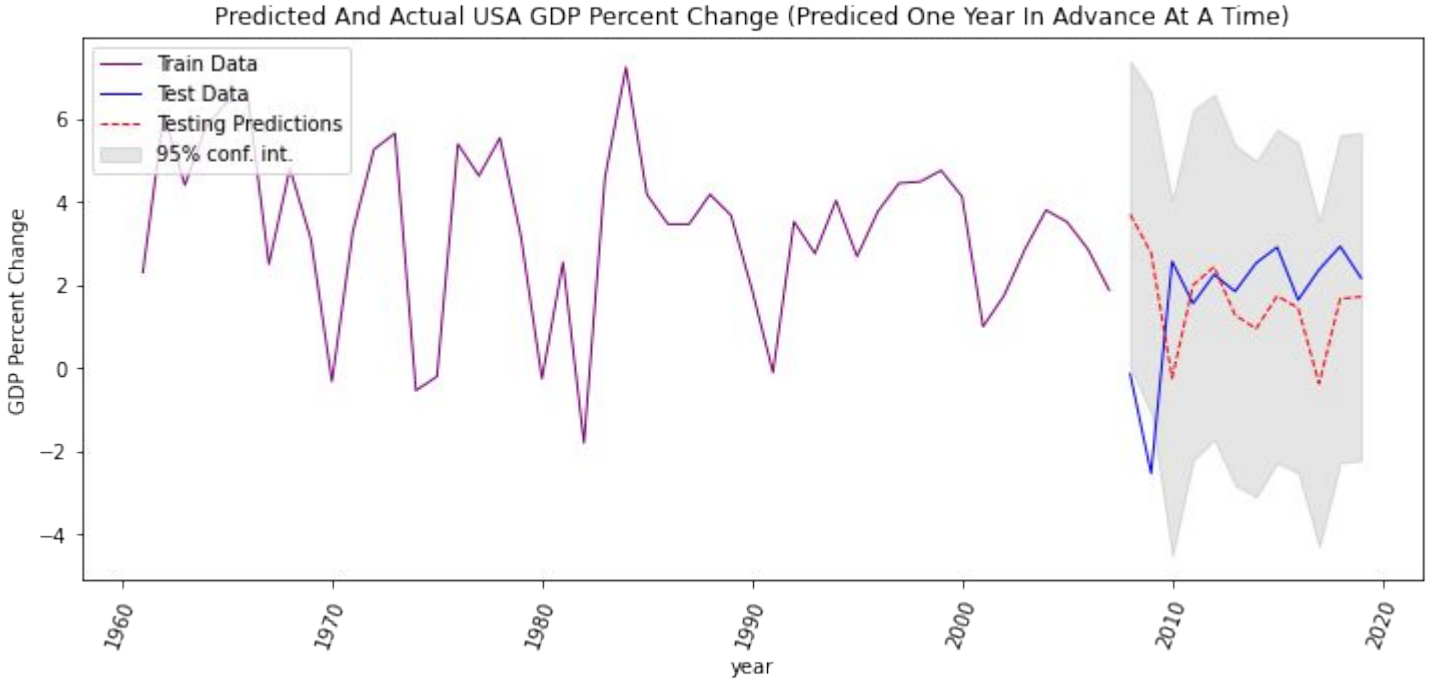
For the USA, how accurately can I forecast GDP by percent change and GDP in dollars?

- Used The Following Data set:
  - USA GDP and GDP Percent Change, Yearly Since 1961
- I predicted USA GDP in dollars and USA GDP Percent Change
- Models Used:
  - Sarima out of sample predictions (predicting final 20% of GDP data)
  - Sarima out of sample predictions (predicting 1 year ahead at a time for final 20% of GDP data)
  - Sarima In Sample Forecast With Added Yearly Political, Environmental, Financial, and Health Metrics
  - FB Prophet Model (In Sample)

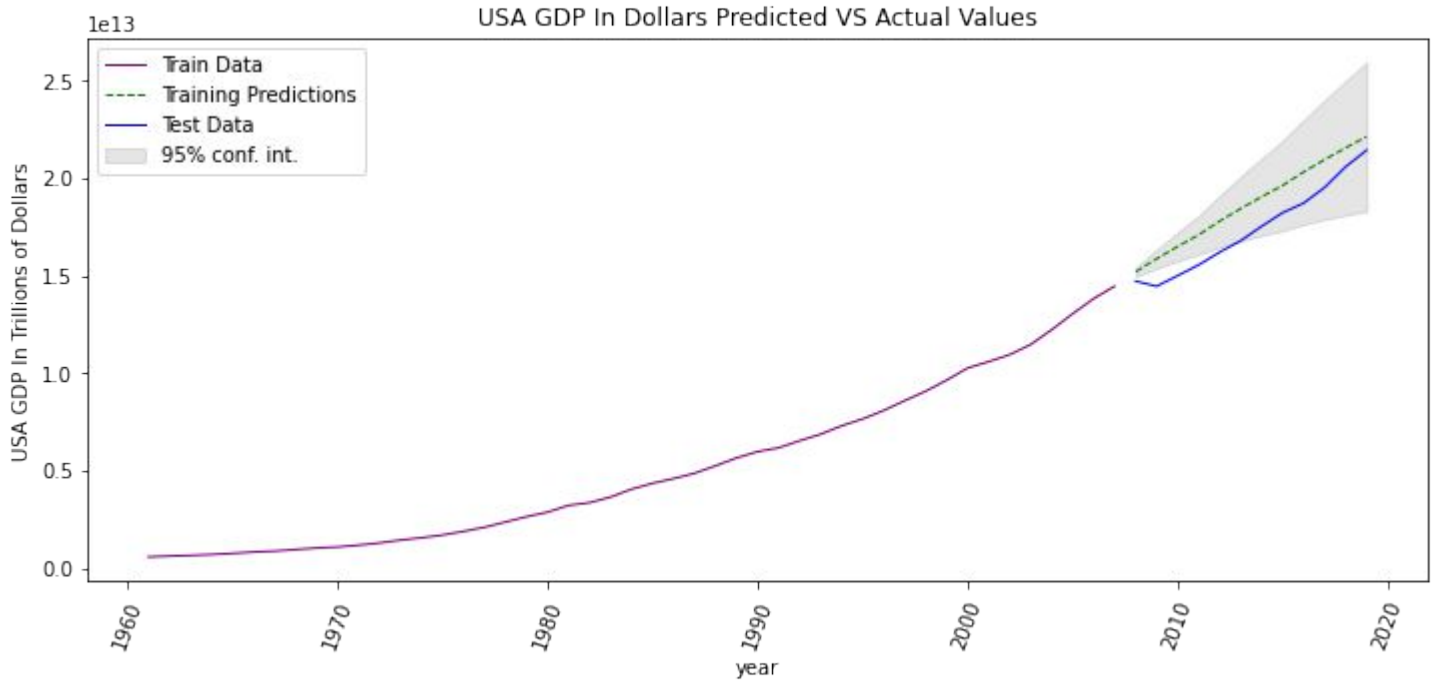
Sarima Out Of Sample Predictions:
Predicting Final 20% Of USA GDP Percent Change Values



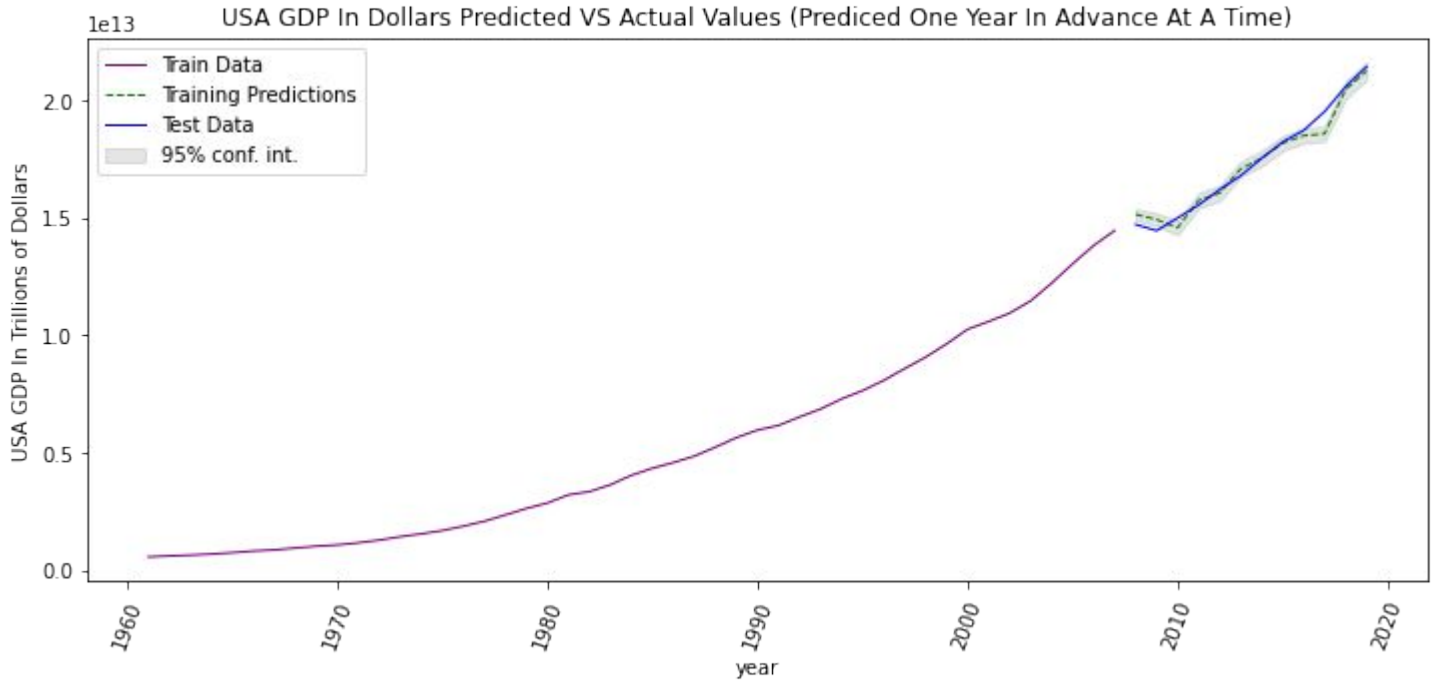
Sarima Out Of Sample Predictions:
Predicting 1 Year Ahead At A Time USA GDP Percent Change



Sarima Out Of Sample Predictions:
Predicting Final 20% Of USA GDP Values In Dollars



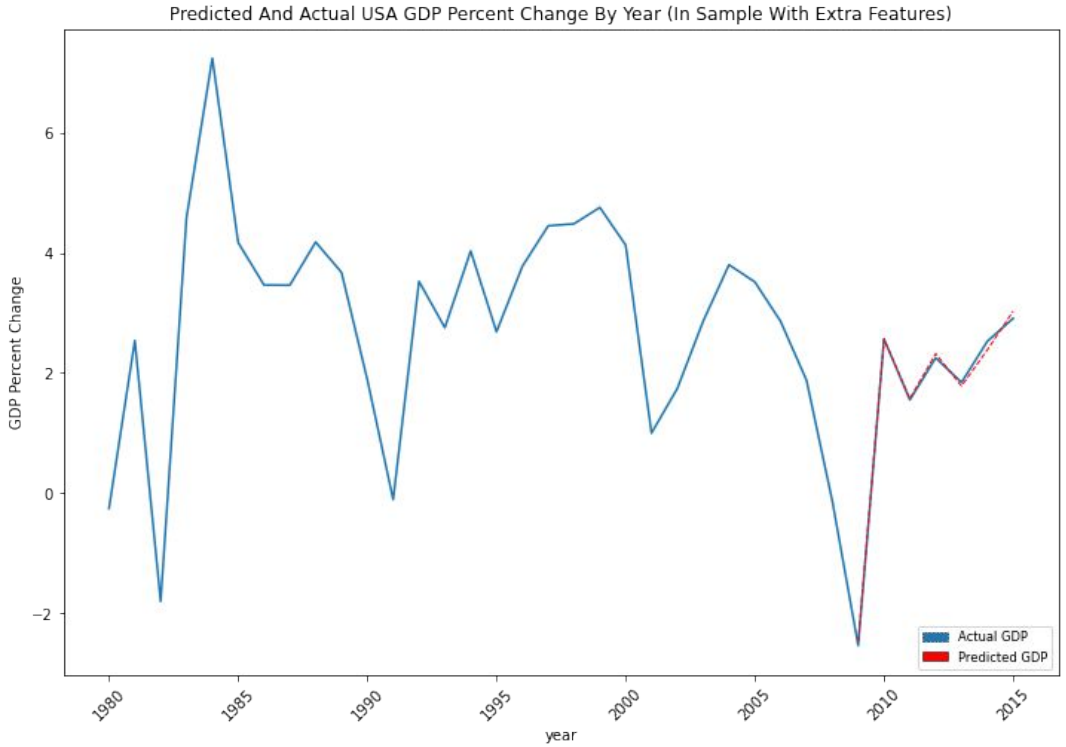
Sarima Out Of Sample Predictions:
Predicting 1 Year Ahead At A Time USA GDP In Dollars



Scoring My Forecasting Model From Previous Slide:

- USA GDP in 2020 was $20.93 trillion
  - My model from the previous slide is on average 379 Billion Dollars off of the actual GDP Value
- Strong association between my predictions and the actual GDP values
  - R2 Score of 97%

Sarima In Sample Forecast With Added Yearly Political, Environmental, Financial, and Health Metrics (USA GDP Percent Change)



Sarima In Sample Forecast With Added Metrics Coefficient Values

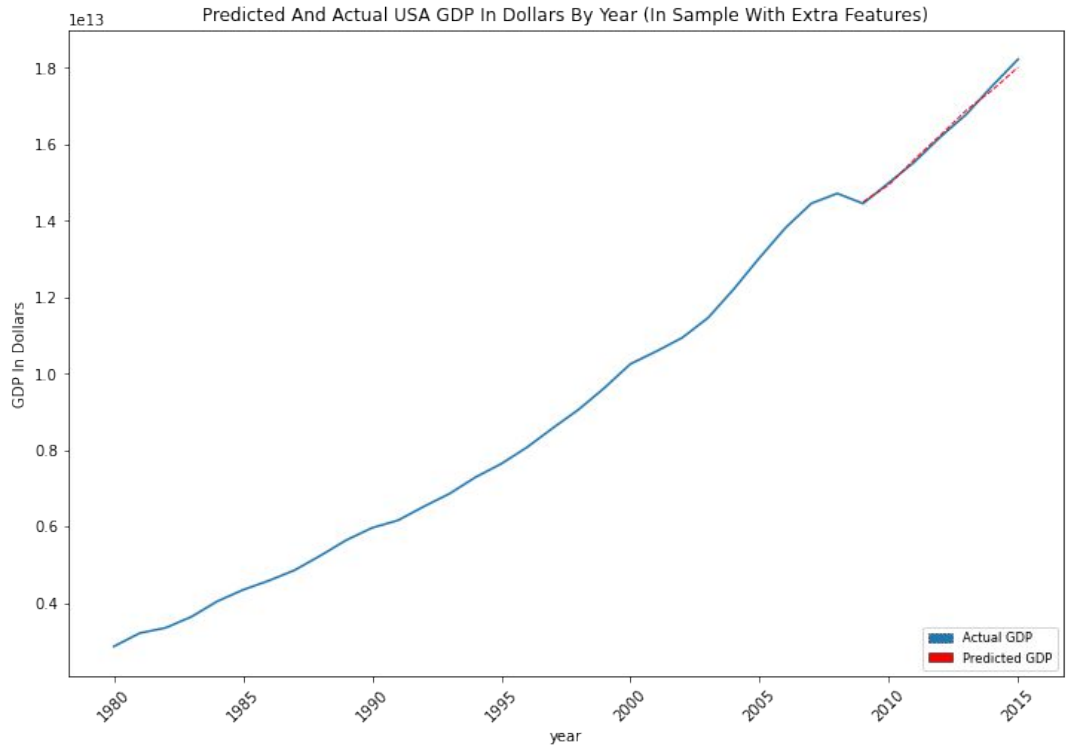
- One unit change in a feature leads to the associated percent change in GDP

	coef	std err
date	2.9432	0.523
Adjusted savings: natural resources depletion (% of GNI)	-1.5986	0.661
Agricultural land (% of land area)	-0.7751	0.390
CO2 emissions (metric tons per capita)	2.6090	1.233
Energy imports, net (% of energy use)	0.7973	0.085
Energy use (kg of oil equivalent per capita)	0.0017	0.004
Fertility rate, total (births per woman)	-6.6254	2.314
Food production index (2004-2006 = 100)	-0.4119	0.113
Hospital beds (per 1,000 people)	6.9617	6.306
Life expectancy at birth, total (years)	3.8372	1.587
Mortality rate, under-5 (per 1,000 live births)	-6.7974	3.143
Patent applications, residents	-5.755e-05	9.76e-06
Population ages 65 and above (% of total population)	8.9109	0.894
Population density (people per sq. km of land area)	-6.5723	1.363
Prevalence of overweight (% of adults)	-0.2378	0.768

Scoring My In Sample Forecasting Model From Previous Slide:

- My model from the previous slide is on average 8% off of the actual value of GDP Percent Change
- Strong association between my predictions and the actual GDP values
  - R2 Score of 99.7%

Sarima In Sample Forecast With Added Yearly Political, Environmental, Financial, and Health Metrics (USA GDP In Dollars)



Sarima In Sample Forecast With Added Metrics Coefficient Values

- One unit change in a feature leads to the associated change in GDP (Dollars)

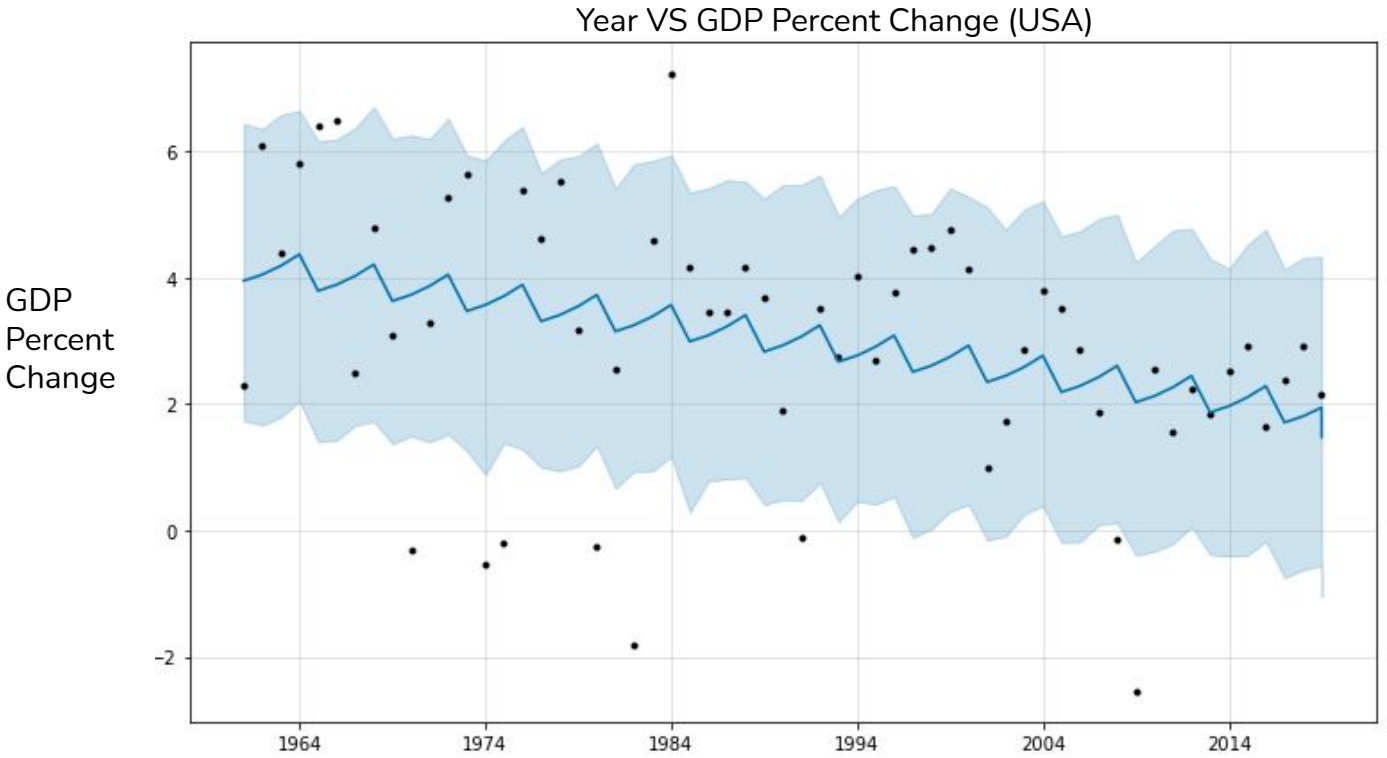
	coef	std err
date	6.559e+11	9.45e-13
Adjusted savings: natural resources depletion (% of GNI)	2.398e+11	5.43e-13
Agricultural land (% of land area)	1.833e+10	1.09e-12
CO2 emissions (metric tons per capita)	-4.586e+10	4.21e-13
Energy imports, net (% of energy use)	-2.516e+10	1.19e-11
Energy use (kg of oil equivalent per capita)	1.148e+09	1.84e-10
Fertility rate, total (births per woman)	1.146e+12	9.96e-14
Food production index (2004-2006 = 100)	-1.301e+10	2.56e-12
Hospital beds (per 1,000 people)	3.736e+11	2.09e-14
Life expectancy at birth, total (years)	1.809e+10	6e-13
Mortality rate, under-5 (per 1,000 live births)	1.227e+11	2.12e-13
Patent applications, residents	4.911e+06	7.21e-08
Population ages 65 and above (% of total population)	1.356e+11	2.94e-13
Population density (people per sq. km of land area)	1.138e+12	1.77e-13
Prevalence of overweight (% of adults)	-6.808e+11	6.27e-13

Scoring My In Sample Forecasting Model From Previous Slide:

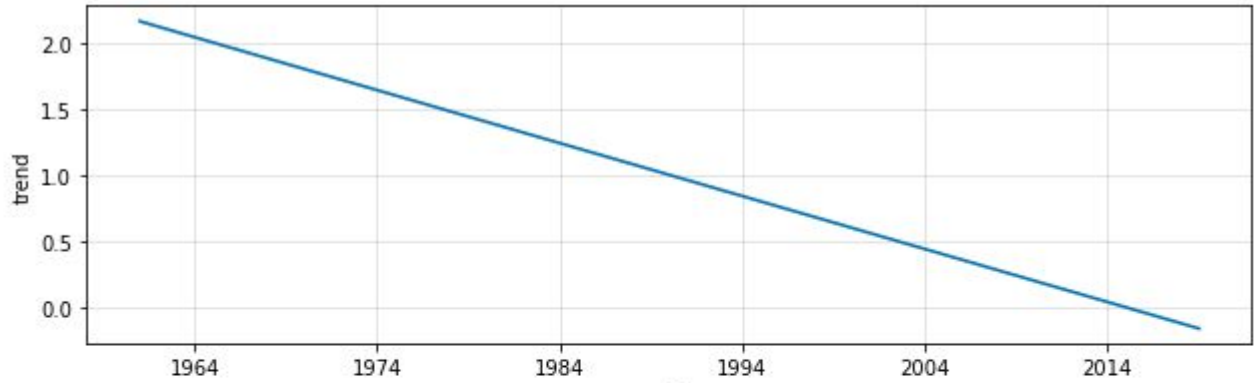
- USA GDP in 2020 was $20.93 trillion
  - My model from the previous slide is on average 104 Billion Dollars off of the actual GDP Value
- Strong association between my predictions and the actual GDP values
  - R2 Score of 99.3%

FB Prophet Model (In Sample)

- Predictions and confidence interval in blue. Displaying actual GDP percent change values in black



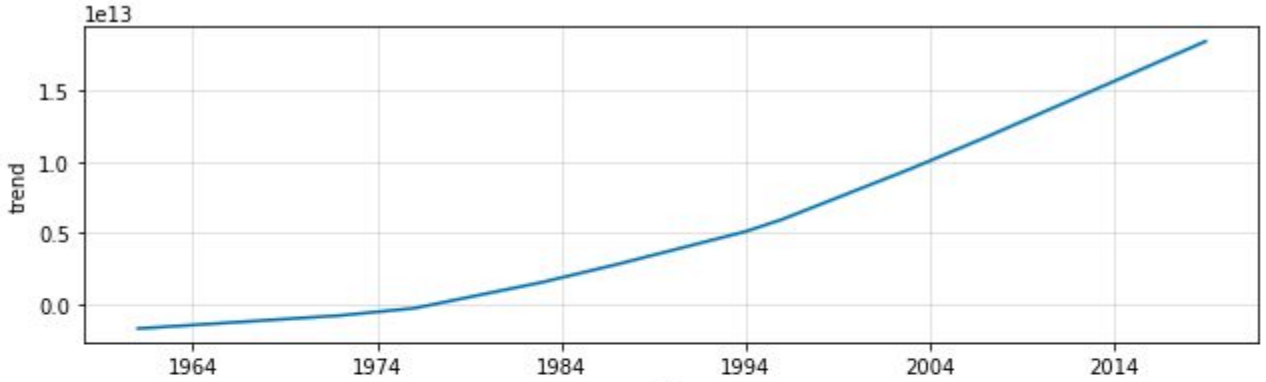
Detected Trend:
FB Prophet Model (In Sample) GDP Percent Change



FB Prophet Model (In Sample)

- Predictions and confidence interval in blue. Displaying actual GDP dollar values in black

Detected Trend:
FB Prophet Model (In Sample) GDP In Dollars



Conclusions:

- How accurately can I predict GDP of a country, without knowing the country name, based solely on political, environmental, financial, and health data?
  - Random forest with extra trees model
    - Achieved an R2 score of 99.5% (strong predictive power)
    - Achieved an average error of 270 Billion Dollars
  - Most Important Features to GDP Prediction
    - Fertility Rate, total births per woman
    - CO2 emissions (metric tons per capita)
    - Prevalence of overweight (% of adults)
    - Mortality rate, under-5 (per 1,000 live births)

Conclusions:

- For the USA, how accurately can I forecast GDP in dollars?
    - In Sample Forecast With Added Features
        - Achieved an R2 score of 99.3% (strong predictive power)
        - Achieved an average error of 104 Billion Dollars
            - In sample forecasts have limitations
    - Most Important Features to GDP Prediction
        - Fertility Rate, total births per woman
        - Population density (people per sq. km of land area)
        - Prevalence of overweight (% of adults)
        - Date
        - CO2 emissions (metric tons per capita)

Conclusions:

- For the USA, how accurately can I forecast GDP in dollars?
    - Out Of Sample Predictions 1 Year Ahead At A Time
        - Achieved an R2 score of 97% (strong predictive power)
        - Achieved an average Error Of 379 Billion Dollars

Conclusions:

- For the USA, how accurately can I forecast GDP by percent change?
    - In Sample Forecast With Added Features (USA GDP Percent Change)
        - Achieved an R2 score of 99.7% (strong predictive power)
        - Achieved an average error of average 8% off GDP percent change
            - In sample forecasts have limitations
    - Most Important Features to GDP Prediction (Percent Change)
        - Population ages 65 and above (% of total population)
        - Hospital beds (per 1,000 people)
        - Mortality rate, under-5 (per 1,000 live births)
        - Fertility rate, total (births per woman)
        - Population density (people per sq. km of land area)

Recommendations:

- To best predict GDP with a generalizable model I recommend using a random forest with extra trees
- My generalized model for predicting GDP of any country based on political, environmental, financial, and health data was created using the countries with the top 5 GDP's
    - To create a more truly generalized model I recommend including data from a myriad of countries with high, middle and low GDP's
        - This model would give more robust information about what metrics a country can focus on in order to increase GDP
- To best forecast future GDP I suggest predicting one year in advance at a time
    - Political, environmental, financial, and health data increase accuracy
        - For best predictive power I suggest creating forecasts for political, environmental, financial, and health metrics for each year in the future
        - I suggest using those forecasts as your features for your model