Abstract:

In this final project, I use NASDAQ economic data as well as stock market data as predictors using a random forest to predict GDP per capita. I found 6 predictors that allowed the random forest to predict GDP per capita with a 94.17% accuracy. I then explore why those predictors might be predictors for GDP per capita. Finally, I normalize the data to compare GDP per capita and the stock market data (as well as the 6 predictors in a separate analysis) and find that the stock market does not have as much of an influence as what people believe (in this specific paper).

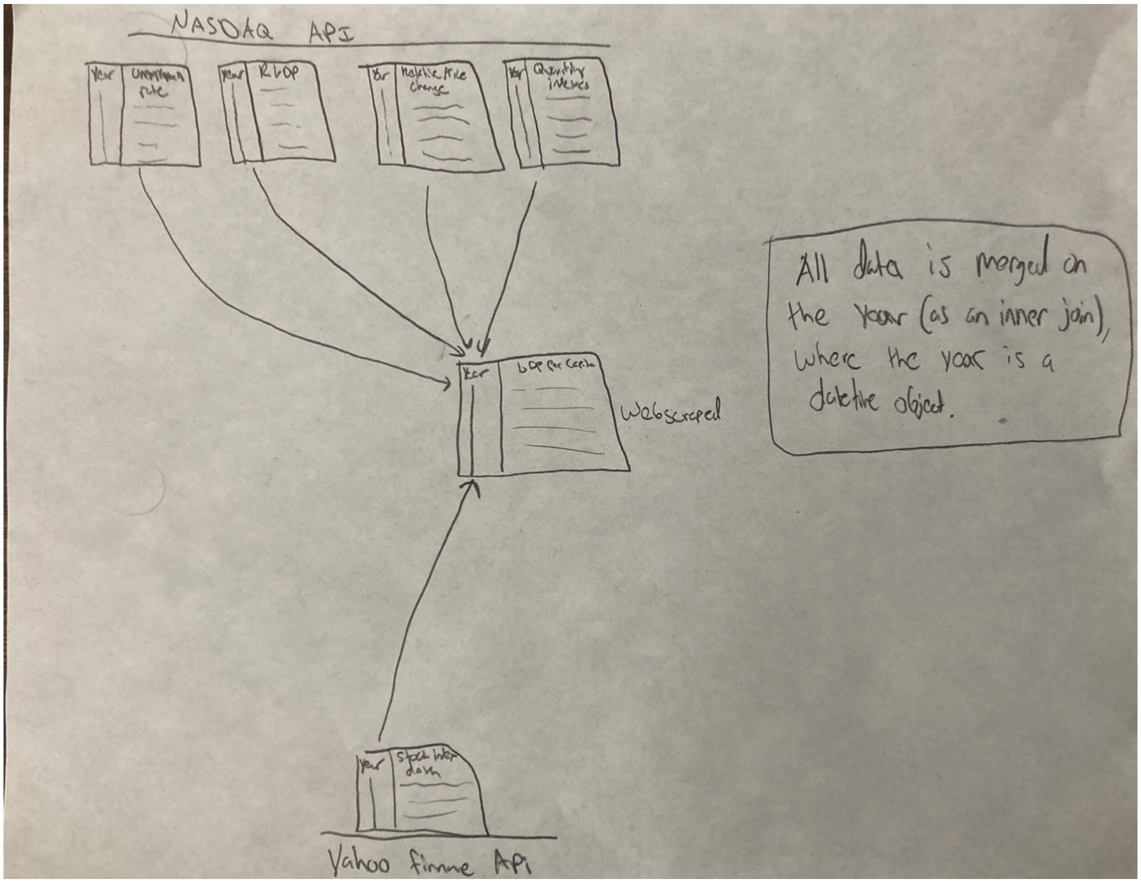
Background:

I take three different data sources (NASDAQ API for economic data, GDP Per Capita web scraped data, and Yahoo Finance stock data), and create a meaningful analysis. I want to see what effect economic data and stock data has as a predictor for gdp per capita. The stock data uses three different indexes that cover the US economy -- the NASDAQ composite, S&P 500, and Dow Jones Industrial Average. The NASDAQ API comes with four separate datasets on the US economy:

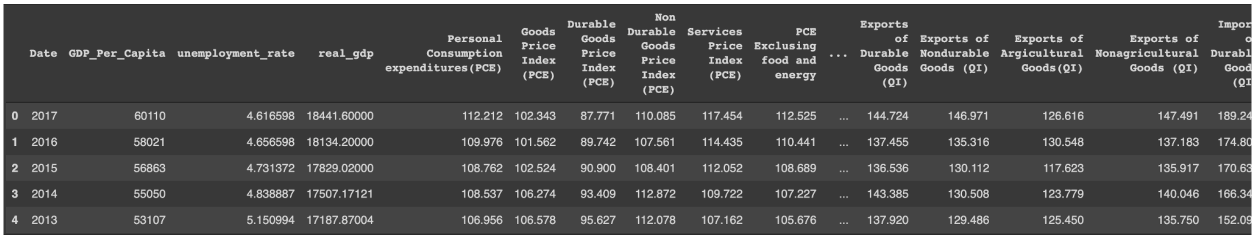
- The first data set is the unemployment rate in the US over time
- The second data set is the real GDP in the US over time.
- The third data set is the Personal Consumption Expenditures (which are indicators of the US economy)
- The fourth data set is the exports and imports of the US

Each dataset is then merged on the year (I converted the datetimes into yyyy format), to create a dataset of 24 features that I can then use 23 of these features as predictors for GDP per capita in the next homework assignment. I am only using US data for the whole project (including US GDP per capita)

Combined Data System:



In this data system, I web scrape the data to get the GDP per capita. I then use the Yahoo Finance API to download the stock data . Finally, I use the NASDAQ API to get all the different economic data. For all data, I get the data from 1960- 2017 and merge the data on the year. The data frame looks like below (it is too large for one screenshot):

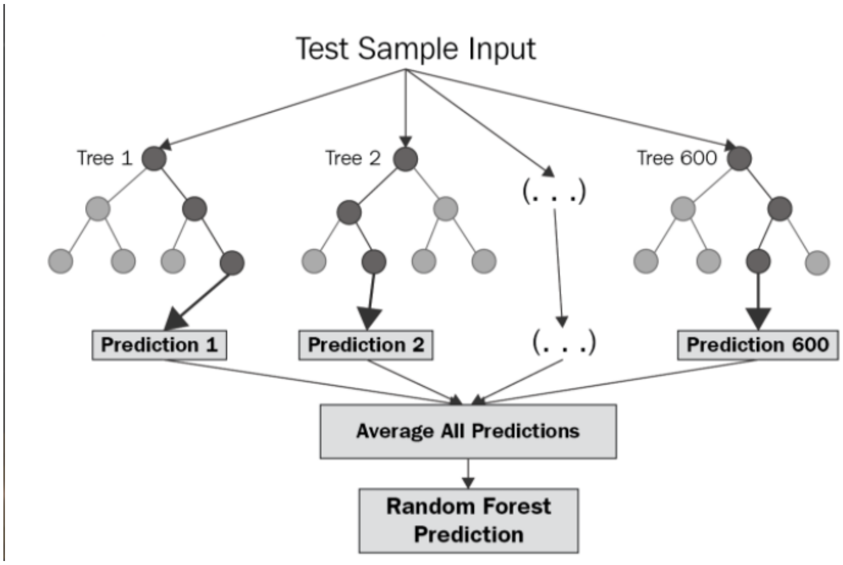


	Date	GDP_Per_Capita	unemployment_rate	real_gdp	Personal Consumption expenditures(PCE)	Goods Price Index (PCE)	Durable Goods Price Index (PCE)	Non Durable Goods Price Index (PCE)	Services Price Index (PCE)	PCE Exclusing food and energy	...	Exports of Durable Goods (QI)	Exports of Nondurable Goods (QI)	Exports of Argicultural Goods(QI)	Exports of Nonagricultural Goods (QI)	Impor o Durabl Good (QI
0	2017	60110	4.616598	18441.60000	112.212	102.343	87.771	110.085	117.454	112.525	...	144.724	146.971	126.616	147.491	189.24
1	2016	58021	4.656598	18134.20000	109.976	101.562	89.742	107.561	114.435	110.441	...	137.455	135.316	130.548	137.183	174.80
2	2015	56863	4.731372	17829.02000	108.762	102.524	90.900	108.401	112.052	108.689	...	136.536	130.112	117.623	135.917	170.63
3	2014	55050	4.838887	17507.17121	108.537	106.274	93.409	112.872	109.722	107.227	...	143.385	130.508	123.779	140.046	166.34
4	2013	53107	5.150994	17187.87004	106.956	106.578	95.627	112.078	107.162	105.676	...	137.920	129.486	125.450	135.750	152.09

Modeling:

To figure out the biggest predictors of gdp per capita given our features, I decided to use a Random Forest Regression. Random Forests are much better than decision trees because they are less prone to overfitting. Given all the data is numerical, it works perfectly with Skikitlearn's

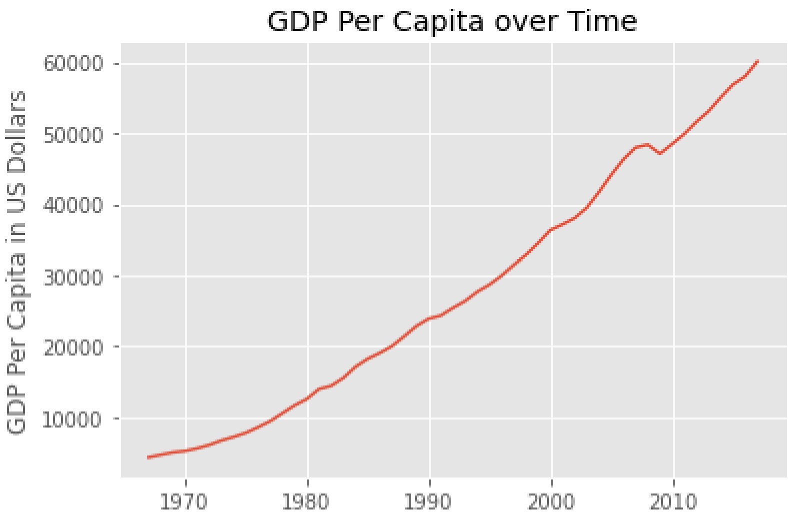
Random Forest Regression. The model works by making decision trees and then averaging the predictions at the end.



For this specific model, I decided to to a train-test-split at 30% testing data (70% training). This is so the algorithm has something to test the accuracy on. I used random_state= 15 to guarantee the same results when the grader runs the algorithm. I found for this dataset that 7 trees that had a maximum of 5 features produced the best results.

Analysis and Results

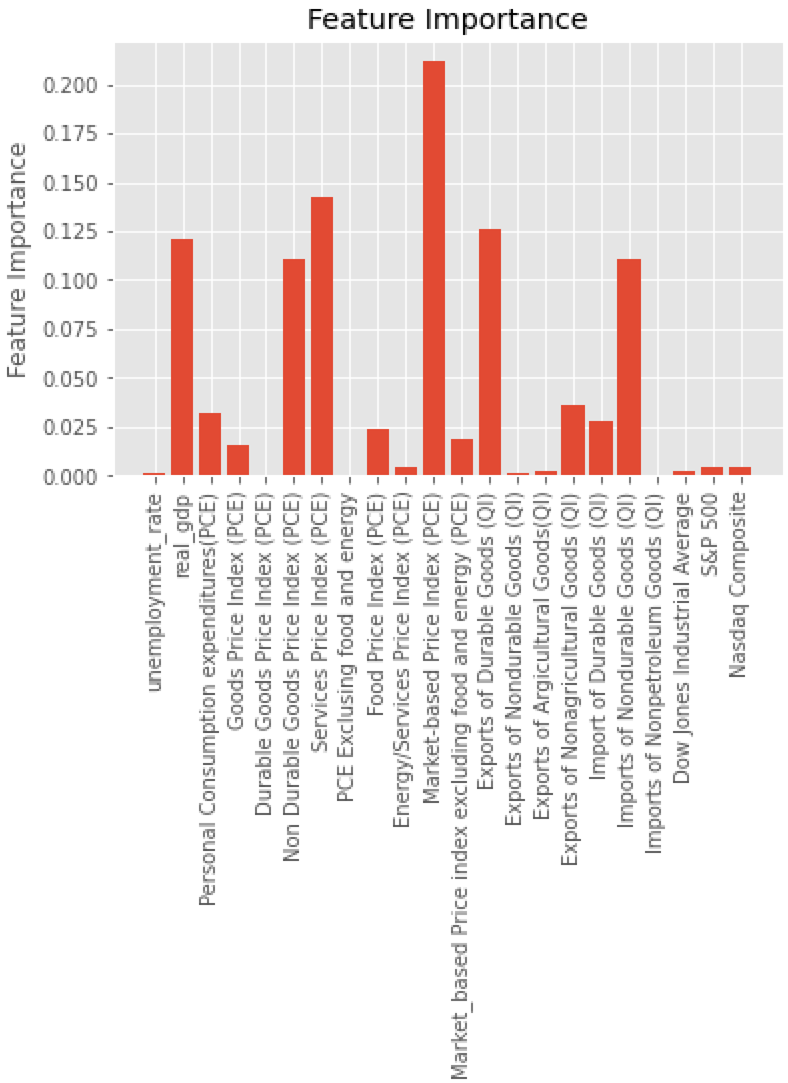
I first plotted US GDP per capita over time to get a better understanding of the data:



The data tells us that over time, the US has steadily grown in GDP per capita. However, what is driving this growth? After making the features on the random forest regressor model predict GDP per capita, we had the following results:

Test	**Results**
Mean Absolute Error (MAE)	872.08
Root Mean Squared Error (RMSE)	1153.98
Mean Absolute Percentage Error (MAPE)	5.83
Accuracy	94.17%

Given we have a highly accurate model at 94.17%, we can now look at the feature importance and try to understand what features are most important and why:



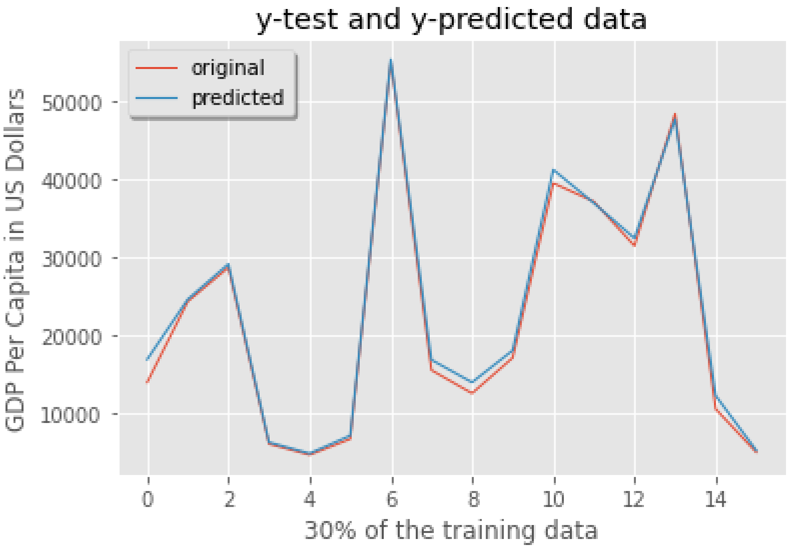
Here we can see there are six distinct features that are far above the rest in terms of feature importance:

Feature	Description	Feature Importance Score
Mark-Based Price Index (PCE)	Measure of prices that people in the US pay for goods and services	0.21
Services Price Index (PCE)	Measure of costs of services in the US	0.14
Exports of Durable Goods (QI)	Quantity index on exporting goods that do not need to be purchased often (lasts 3+ years)	0.12
Real GDP	Inflation adjusted value of goods and services produced by labor and property	0.12
Imports of Nondurable Goods (QI)	Quantity index on importing goods that are used up in a short period	0.11
Non Durable Goods Price Index (PCE)	Price index on goods that are used up in a short period	0.11

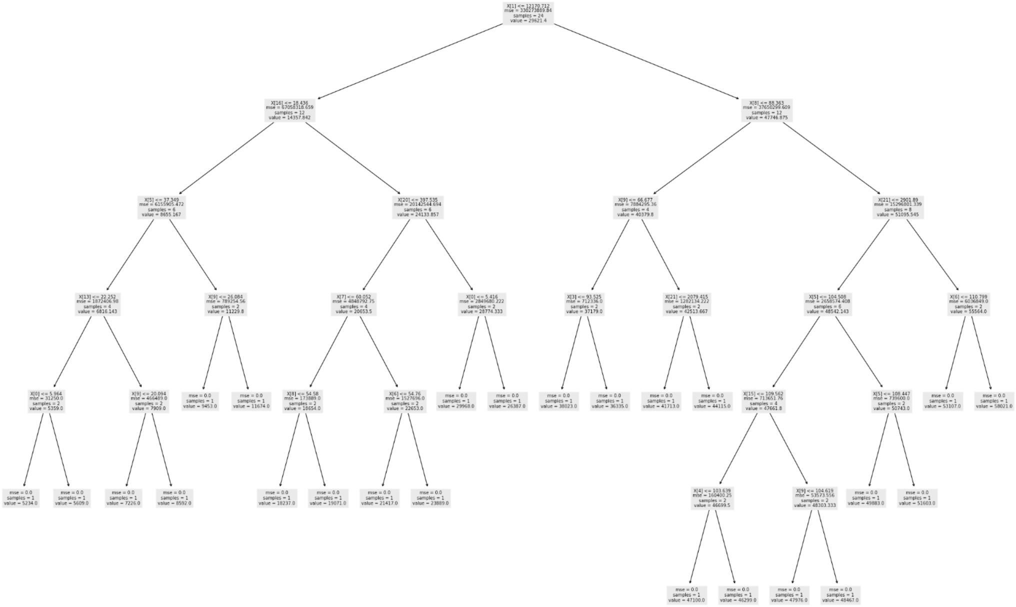
NOTE: PCE is the personal consumption price index (a measure of US inflation) and QI is a quantity index

Market-Based price index is highly indicative of GDP per capita because employers adjust the wages in the US based on the cost of living and pricing of goods. Just like the Market-Based Price Index, as services become more expensive (Services Price Index), GDP per capita must go up with the services otherwise people won't be able to afford the services. The next four features are very close in feature importance scores. The exports on durable goods and imports on non durable goods would help the US economy making GDP per capita go up. As people make more goods and services in the US (real GDP), the GDP per capita would go up. The Non Durable Goods Price Index would mean that if prices are high, GDP per capita would have to go up to be able to purchase those non durable goods.

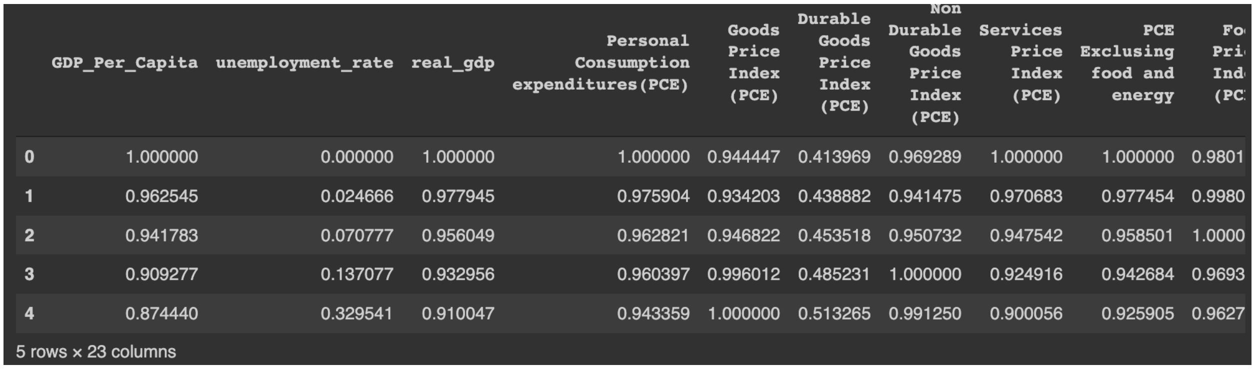
Next, we can visualize how the model did on the testing data to see how close the predicted was to the original data.



As you can see, the model did very well at predicting the testing data. Next, we can see what a single decision tree looks like within the random forest:



The decision trees are not too deep to prevent overfitting of the data. Finally, to finish the analysis, I will be normalizing the data and putting it on the same graph to see how the data compares:

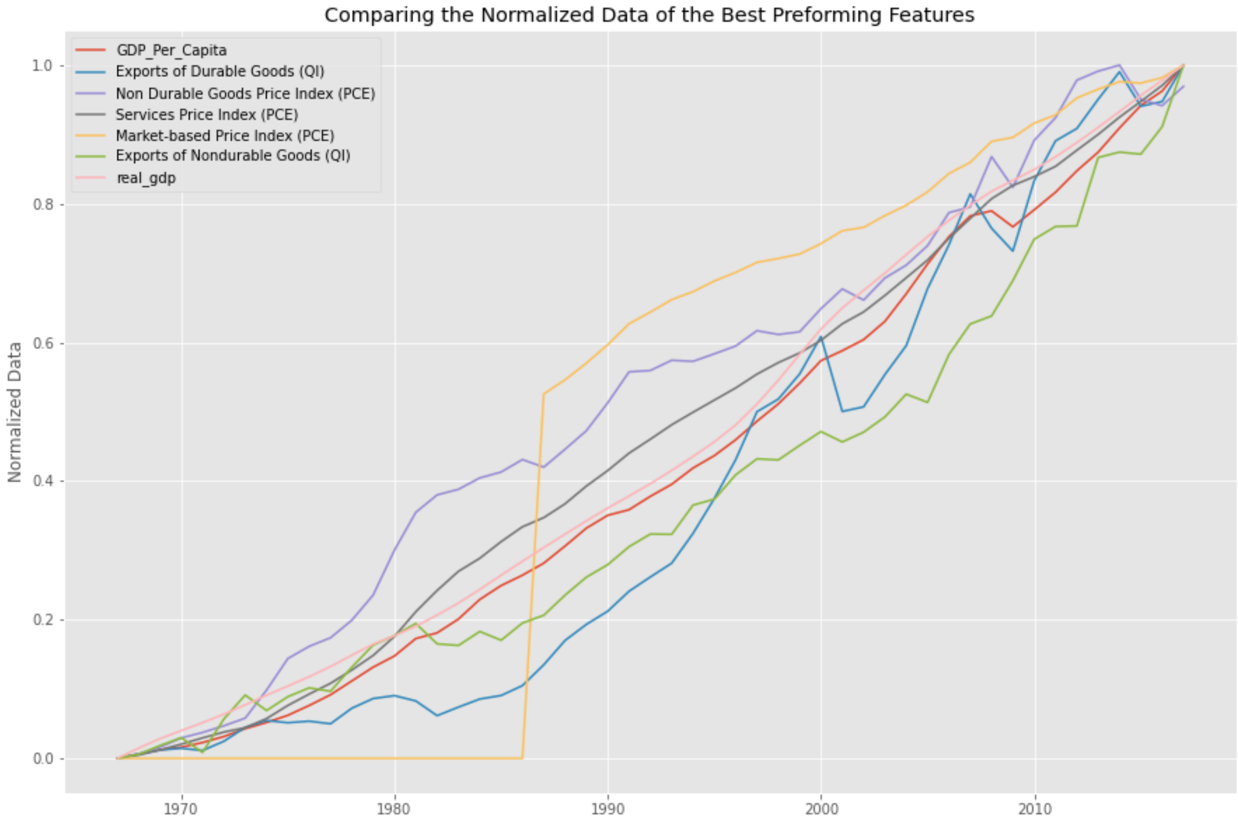


	GDP_Per_Capita	unemployment_rate	real_gdp	Personal Consumption expenditures(PCE)	Goods Price Index (PCE)	Durable Goods Price Index (PCE)	Non Durable Goods Price Index (PCE)	Services Price Index (PCE)	PCE Exclusing food and energy	Fo Pri Ind (PC
0	1.000000	0.000000	1.000000	1.000000	0.944447	0.413969	0.969289	1.000000	1.000000	0.9801
1	0.962545	0.024666	0.977945	0.975904	0.934203	0.438882	0.941475	0.970683	0.977454	0.9980
2	0.941783	0.070777	0.956049	0.962821	0.946822	0.453518	0.950732	0.947542	0.958501	1.0000
3	0.909277	0.137077	0.932956	0.960397	0.996012	0.485231	1.000000	0.924916	0.942684	0.9693
4	0.874440	0.329541	0.910047	0.943359	1.000000	0.513265	0.991250	0.900056	0.925905	0.9627

5 rows × 23 columns

NOTE: Index 0 is 2017

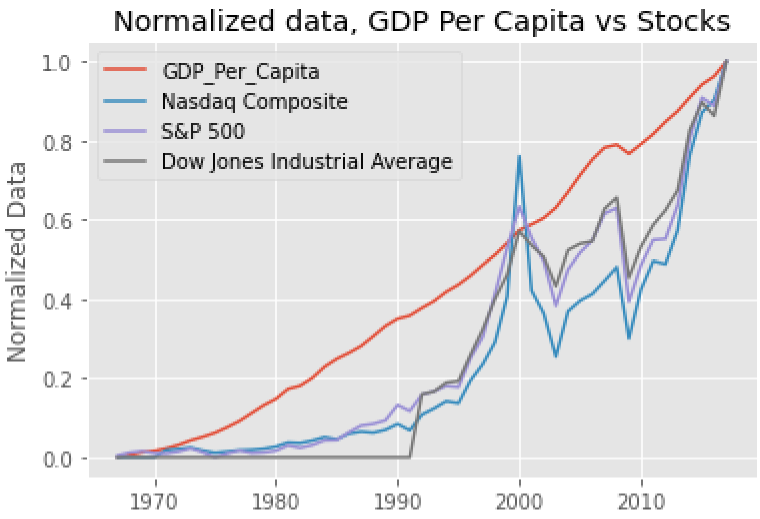
We will be looking at the normalized data for the six most predictive features of GDP per capita:



NOTE: Market-based price index did not have data before 1987

We can see the curves for four of the predictors are above the GDP per capita for most of recorded history and two are below. GDP per capita, Non Durable Goods Price Index, and Exports of Durable good show that they struggled during the 2008 recession, with all three going down during the time period.

We can then look how stocks (using the three largest stock indexes) influence GDP per capita:



What we find is that there is no easily identifiable correlation between stocks and GDP per capita. The curves are very different and do not show a relationship other than they have growth over time. Based on the model and the graph above, GDP per capita is not necessarily dependent on stocks. This goes against popular belief that if stocks are doing well, people across the US do well with the stocks. From my model and findings, people are much more impacted by the economy than the stock market.

Technical Challenges:

I found it difficult to find a model that fits the data and makes an accurate prediction. After trying a few models, I had found that the random forest regression mad the best prediction after changing the parameters (# of trees and max features), so that the model isn't overfitting.

Conclusion:

In this project, I was able to find 6 predictors of GDP per capita as well as find that stocks are not as big of a predictor of GDP per capita as popular opinion might think.

Datasets:

Yahoo Finance API: https://www.yahoofinanceapi.com/
GDP Per Capita Table: https://www.macrotrends.net/countries/USA/united-states/gdp-per-capita
NASDAQ API Datasets:

1. Unemployment: https://data.nasdaq.com/api/v3/datasets/FRED/NROUST.json?api_key=iZoWyhptYoFNawrYZYKd
2. Real Gross Domestic Product: https://data.nasdaq.com/api/v3/datasets/FRED/GDPPOT.json?api_key=iZoWyhptYoFNawrYZYKd
3. PCE: https://data.nasdaq.com/api/v3/datasets/BEA/T20804_M.json?api_key=iZoWyhptYoFNawrYZYKd
4. QI: https://data.nasdaq.com/api/v3/datasets/BEA/T40203_Q.json?api_key=iZoWyhptYoFNawrYZYKd