**U.S GDP Analysis: Predicting the impact of Major Sectors using Linear Regression**

Sachin Muralidharan, Dhwanil Nikunj Mehta

Introduction:

For our project, we wanted to build a machine learning model which could make predictions about the GDP of the USA, and could predict what the GDP would be in a given year, depending on certain features. Gross Domestic Product is a metric that is dependent on several sectors and factors. GDP is a measure of the total goods and services produced within a country’s borders in a given year and includes a summation of a country’s total household consumption, business investment, and government spending, among others. GDP is considered to be the clearest metric to assess the health of a country’s economy, thus making it incredibly important for policymakers and leaders to have a concrete understanding of the statistic. We believe that a strong machine learning model that can be applied to GDP will provide a better understanding of the U.S. economy’s strengths and weaknesses, and this was the baseline for our project idea.

Objectives and Design:

We designed our project around the idea that GDP is reliant on a set of sectors within a healthy economy. In building this model, it would be unrealistic to attempt to find and interpret data for every single sector within the U.S economy, so we selected specific sectors and industries which would carry greater weight in a GDP prediction. We ultimately chose 3 sectors, which were manufacturing, energy, and agriculture. These 3 sectors not only had easily accessible data available to work with, but in the year 2021, manufacturing, energy, and agriculture combined for 12.0, 4.4, and 5.4 percent respectively, combining for nearly 20 percent of the total U.S GDP. Therefore, we defined our learning task to be as follows: develop individual linear regression models for each of these sectors, and discover a way to combine these into 1 large predictive model to assess GDP.

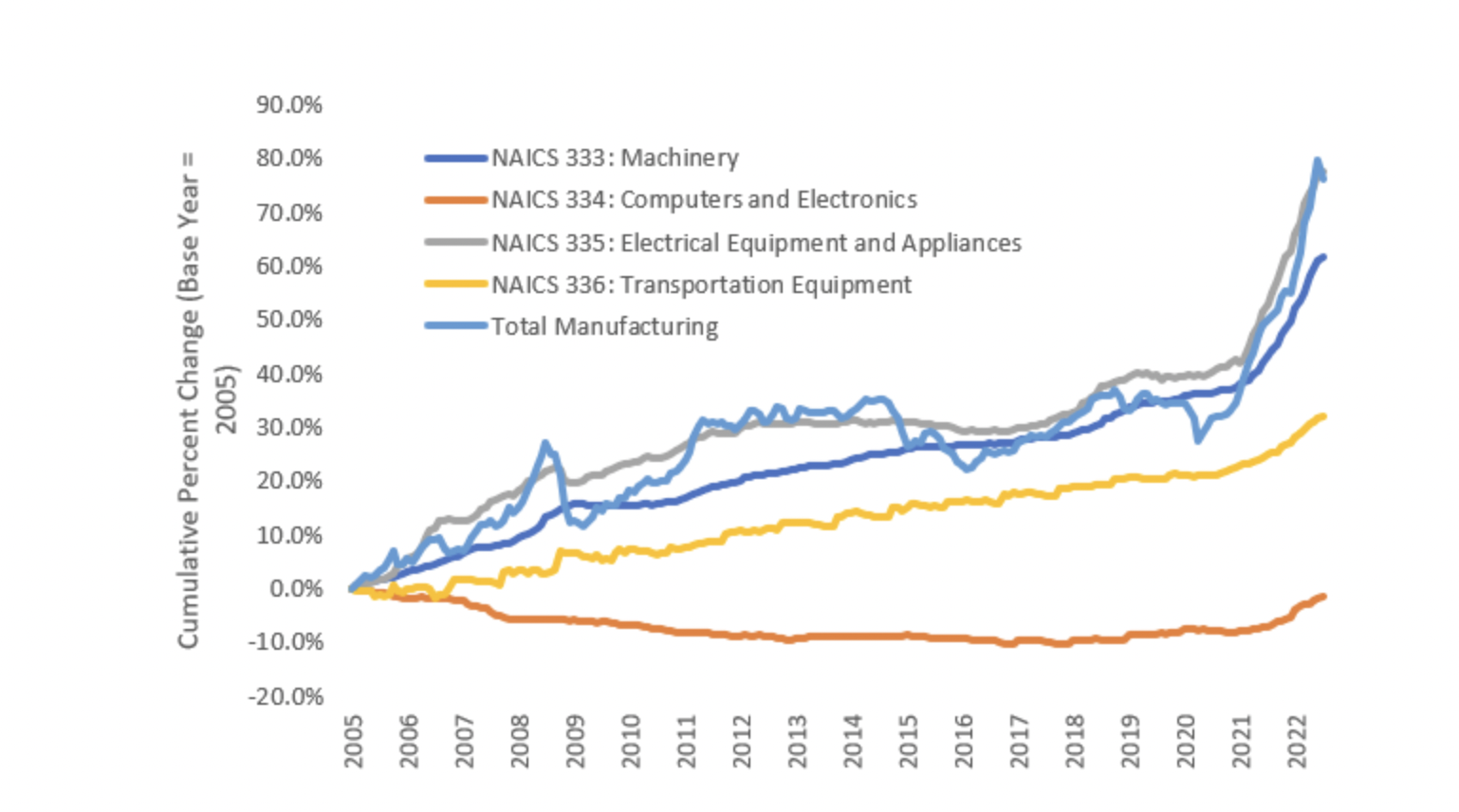


Figure 1: Manufacturing Trends

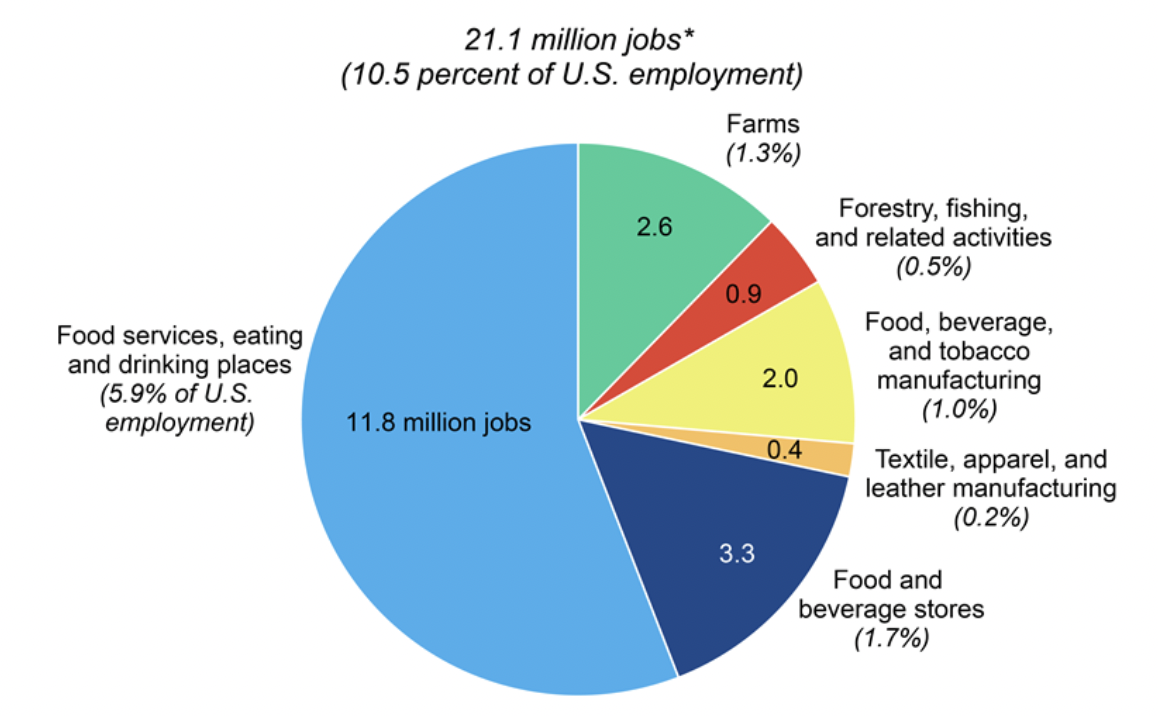


Figure 2: Agricultural Segmentation

Python Pandas:

For our project, we used the Python Pandas library as our primary method of storing and manipulating datasets. Pandas provided an easy way to conduct data preprocessing, such as handling NAN values and transforming our data into a format applicable to linear regression.

Scikit-Learn:

We used the Sklearn library to build our linear regression models for the project. Sklearn offered an easy way to implement our models, but it also provided us with quality documentation and instructions to implement a stacking regressor that combined our individual models together.

Datasets and Data Preprocessing:

We decided to combine various datasets because every dataset we used was specialized towards a single sector. For obtaining the necessary information, we used official data provided by government websites. For example, we used data from the API provided by the EIA to gather information about energy trends during the past fifty years. We had to restructure some datasets using the pivot function in pandas to predict the average output of that sector for years, as the data was not originally provided in a neat column format. Additionally, because sector outputs were dependent on many different features, we determined a need to use multivariate linear regression, as a sector like Agriculture is dependent on over 30 sub-features. Lastly, there were some complexities in obtaining the data due to certain limitations with JSON retrieval. For example, the EIA API only allows a user to retrieve 5000 rows at a time, and therefore we needed to manually add in data after a certain number of requests.

Agriculture Sector:

We selected agriculture as one of our sectors because it plays a major role in USA’s GDP and the dataset we found had several features leading to agricultural output like capital inputs and crop outputs. The dataset we obtained from the USDA consisted of over 30 features each having its own column and a total agricultural output variable having its own column as well. We had to pivot the dataset in a way where the agricultural output feature became the dependent variable and the other features were the independent variables for every single year.

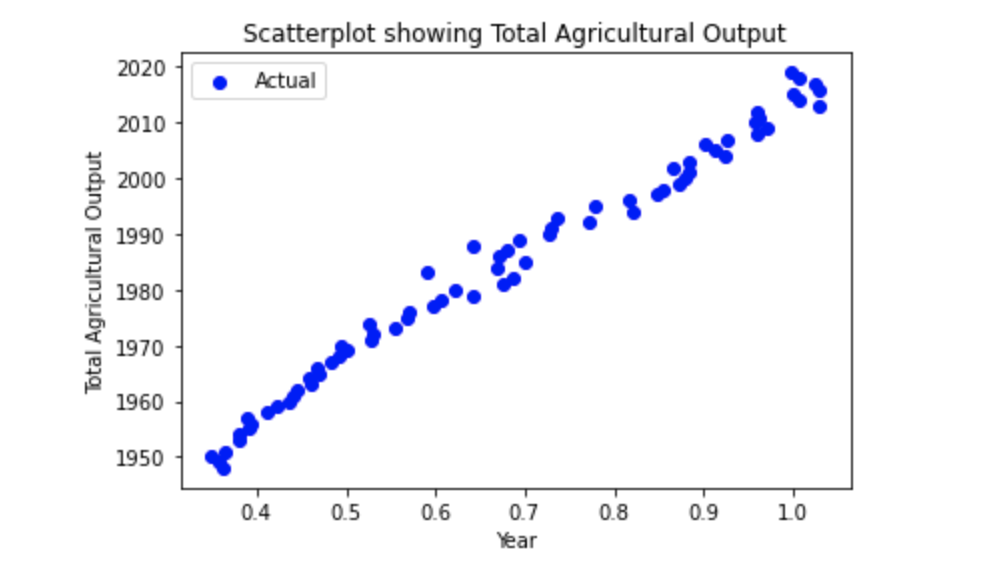


Figure 3: Total Agricultural Output Graph

Coal Sector:

We wanted to select three important sub-sectors from the energy sector, and determined that coal would be an ideal fit based on past trends. For this sector, we built our model with five key features, which were average employees, labor hours, number of mines, production, and productivity. We then inserted a column for GDP, based on the row’s year value, and this was our new dependent variable.

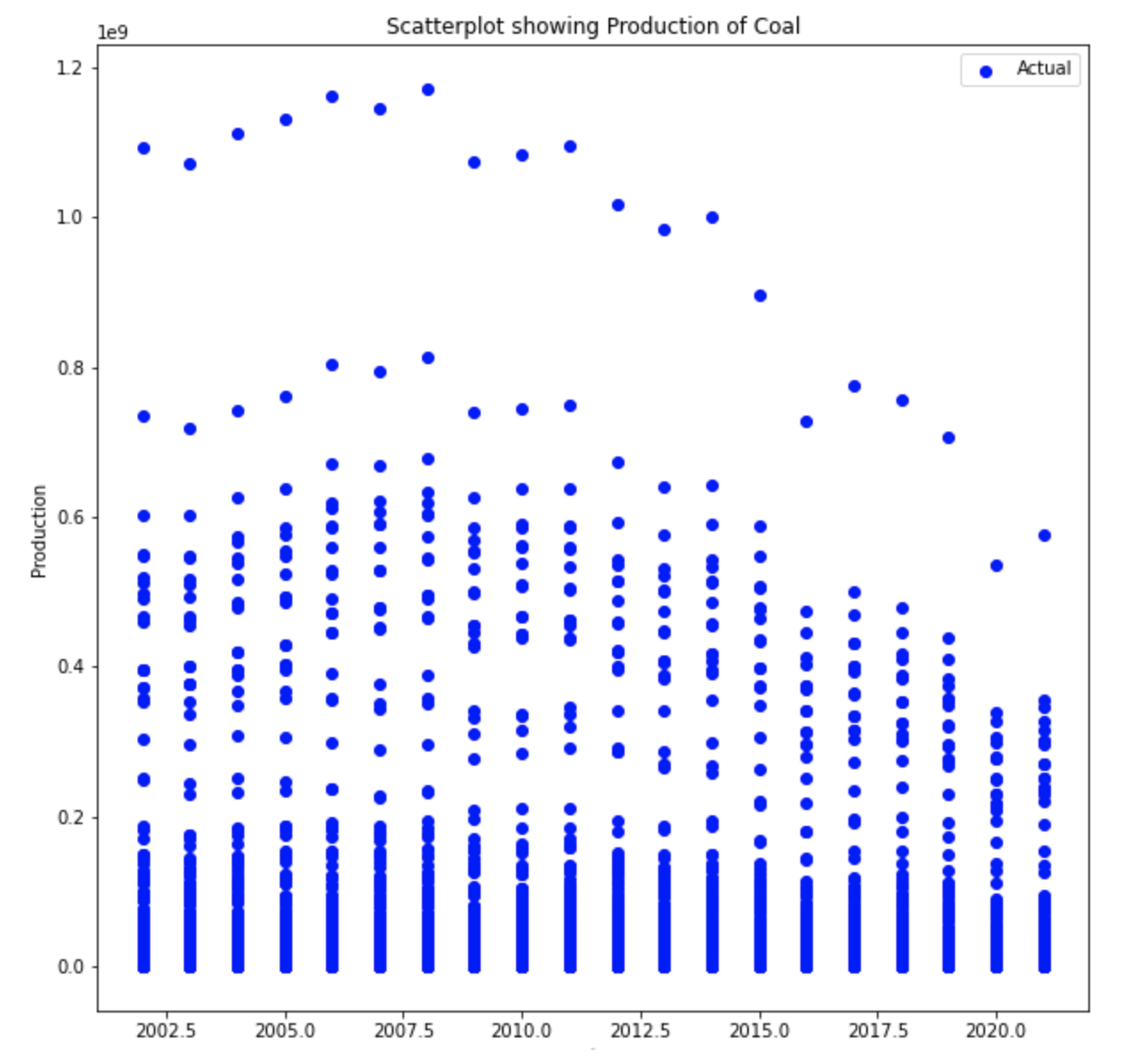


Figure 4: Coal Production Graph

Manufacturing Sector:

The manufacturing sector consists of several productions of goods that are exported as well as sold in the country. For this sector, in order to build a comprehensive dataset, we combined the growth of different individual manufacturing industries, such as plastics, computers and electronics, and industrial production.

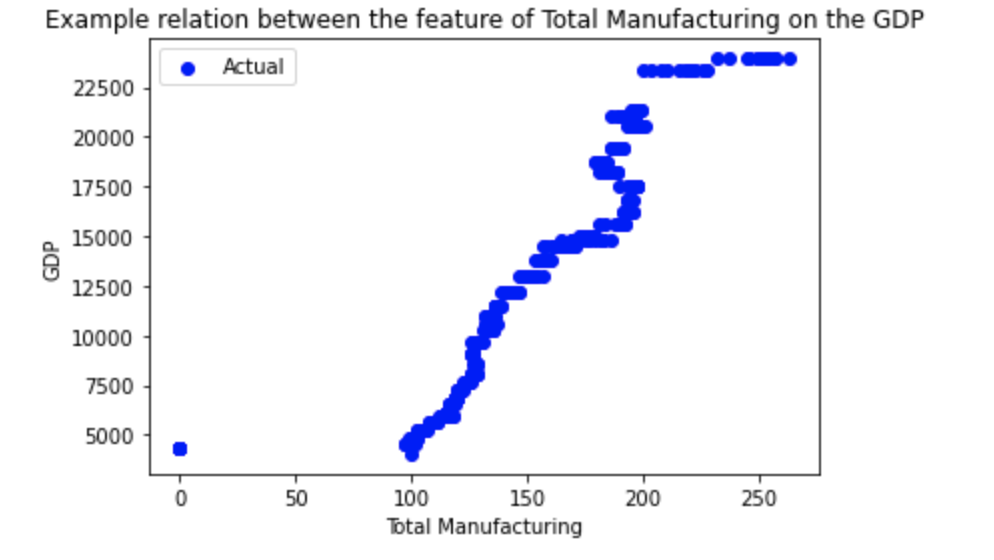


Figure 5: Manufacturing and GDP Graph

Electricity Sector:

The electricity sector focuses on the revenue generated for every state as electricity is provided throughout the country and taxes and costs need to be paid to utilize it. Thus, it directly impacts the GDP of the country. In the dataset, we combined statewide electricity production with a GDP column for each year to predict the impact of electricity revenue on the GDP.

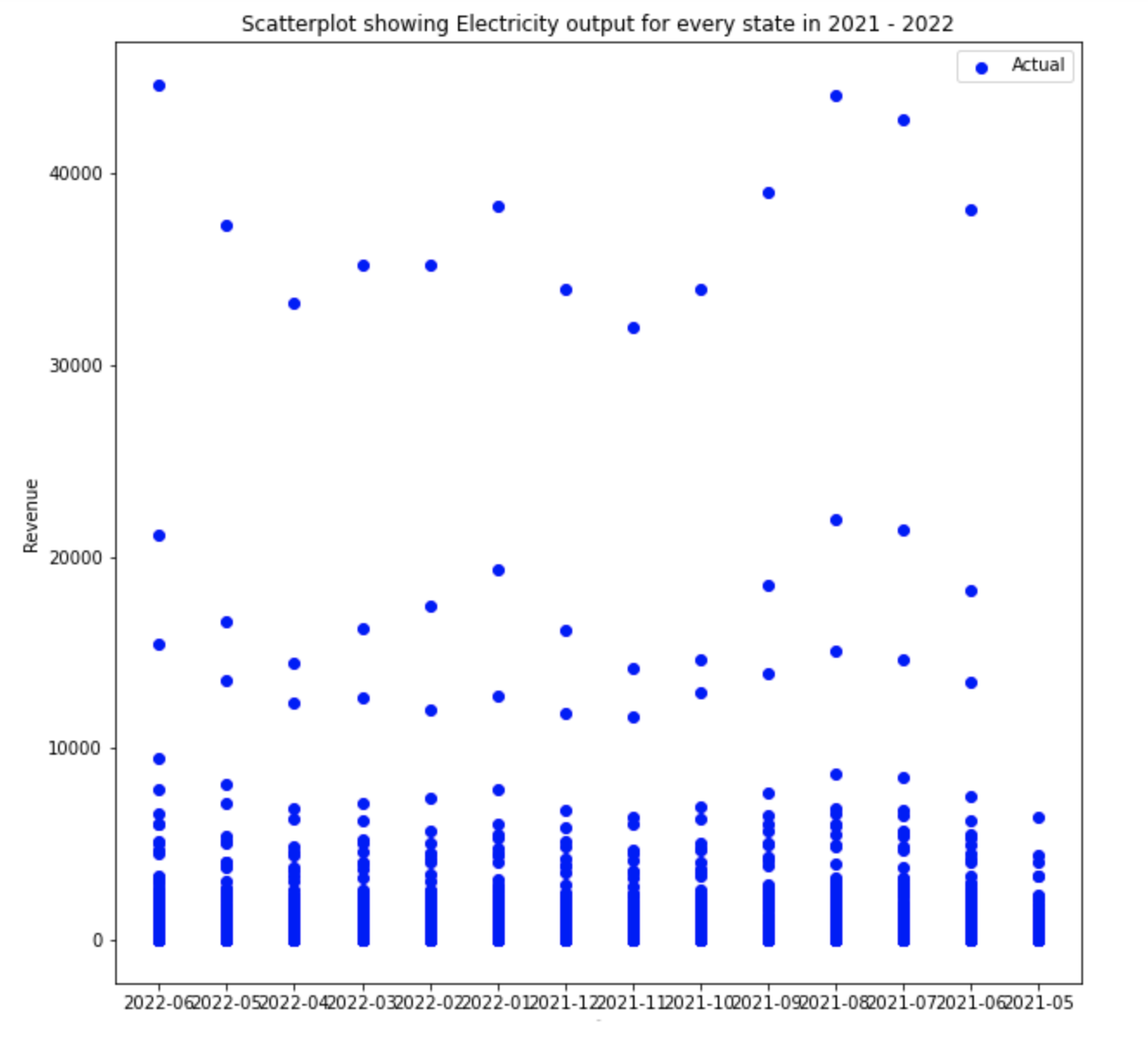


Figure 6: Electricity Production Graph

Natural Gas Sector:

The natural gas sector is one of the highest contributors to the GDP of the country due to the impact it has on exports as well as internal distribution. This was the last sub-sector we used for the energy sector, and this is arguably the one that has the most impact on America’s overall GDP. The dataset consists of the value of the output of Natural Gas for each year in the dataset, and like the previous sectors, we simply added a column for GDP to build our model.

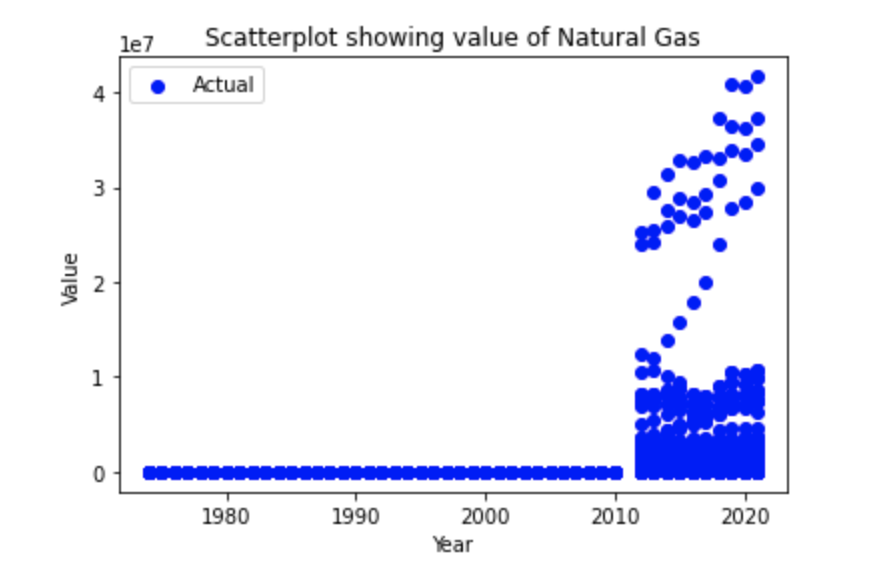


Figure 7: Natural Gas Graph Production(Data accumulated post-2011)

Meta-Analysis:

One of the key concepts we discovered in this project was that of Meta-Analysis, which is a way to analyze several different results, and draw one conclusion from them. This method considers the merging of several independent results to calculate the overall effect of a set of predictions. In our case, we predicted the individual outputs for every sector and combined the results to see the impact they have on GDP.

Meta-Analysis is based on 5 steps:

* Developing a hypothesis and understanding the individual datasets based on it
* Conduct a systematic review of the individual datasets to make sure that they are directly impacting the final output
* Data extraction from every independent prediction
* Decide appropriate summary measures from every prediction of every dataset
* Decide an appropriate model to summarize all the studies

Ultimately, our research about Meta-Analysis led us to discover the Stacking Regression method.

Stacking Regressor Implementation:

Once we had developed our five linear regression models, the main problem was how to combine each of them. The models were each trained with the same target variable, GDP, but the datasets and the features were completely different. Therefore, we needed to find a way to combine models which are data-independent, and the stacking regressor was the solution. A stacking regressor receives input from different machine learning models, and uses the model's coefficients and intercepts to create a “meta-model”. A stacking regressor falls under the classification of an “ensemble method”, which combines base estimators to improve overall prediction results. Many data scientists use the model stacking method to assess the results of different algorithms which are applied to datasets. Stacking does not only apply to regression, and the more frequent usage comes with the Stacking Classifier, which can receive different models such as KNN and SVM models, and combine them into a single prediction model. It’s important to note that data scientists typically use stacking within the same dataset, where you split the dataset into different portions, train different models, and then combine them using model stacking. However, in our project, we combined models which came from different datasets, which is a valid approach but is still unconventional. One drawback of the stacking regressor is the computational complexity involved with it. The process of model stacking involves the development of multiple machine learning models, which is inherently costly, however for our project, this was a necessity, because we needed some way to combine our already existing individual models. We also explored different approaches to combining our models, and these methods included bagging(bootstrap aggregation) and boosting, which are also ensemble methods. We eliminated the boosting approach, because this is a data-dependent approach, and also involves decision trees, which we felt was out of scope for this project. We also did not select bagging because of the decision tree components involved with it, however, this method has been revealed as successful in eliminating noises and overfitting, so to take this project further, a different approach could be to use bagging to increase our accuracy. Overall, the stacking regressor provided us with a straightforward way to blend our models together, but because we did not use the method in a typical way by splitting a main dataset into individual models, it's possible we could achieve better results by choosing a different ensemble method.

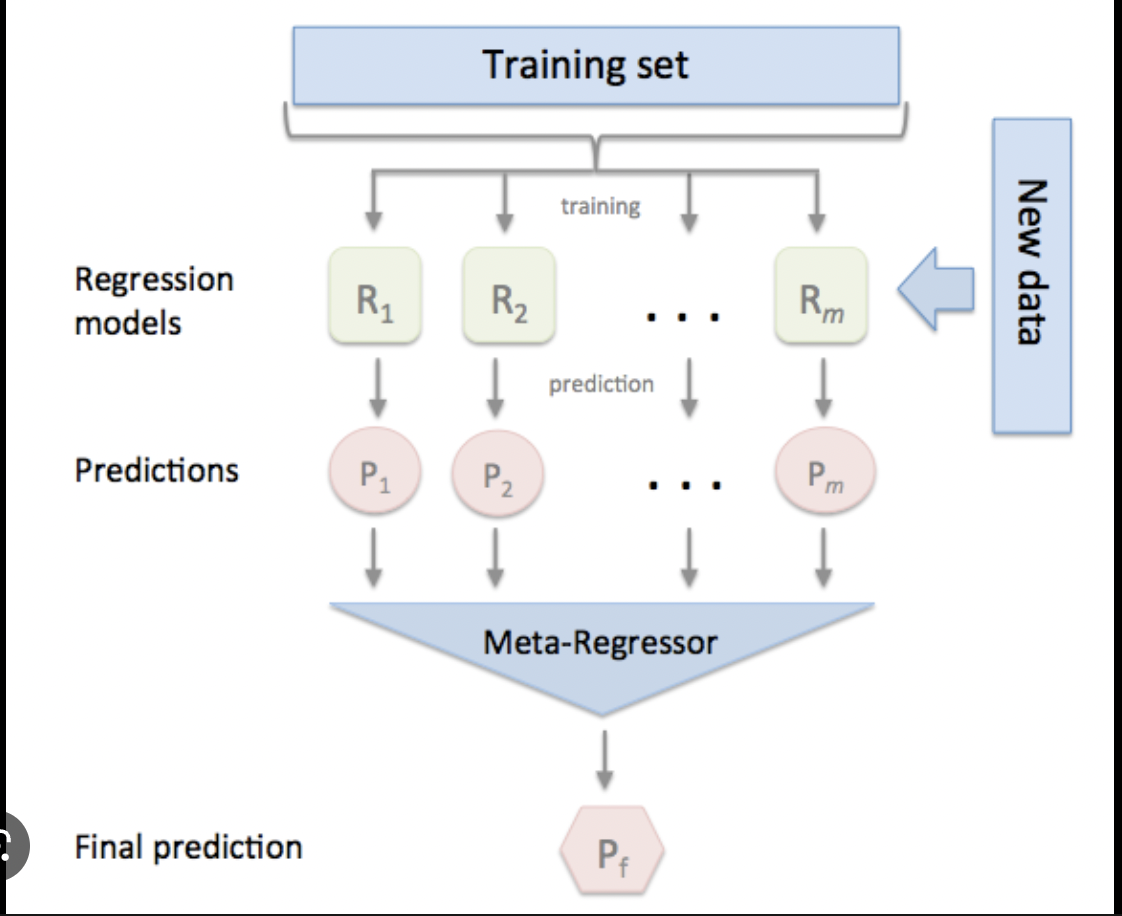


Figure 8: Stacking Regressor Representation

Final Results:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Percentage Error | Mean Absolute Error | R-squared Score |
| Manufacturing | 1.088 | 152.734 | 0.999 |
| Electricity | 4.193 | 484.145 | 0.979 |
| Agriculture | 2.381 | 402.053 | 0.994 |
| Natural Gas | 2.646 | 338.199 | 0.99 |
| Coal | 3.826 | 399.085 | 0.981 |
| Summary | 10.672 | 2192.913 | 0.857 |

Interpretations and Conclusion:

From the table’s results, we have a few key takeaways. Firstly, the individual models for each sector appeared to perform fairly well. We chose the mean absolute error, r-squared score, and percentage error as our three metrics for evaluation. Mean absolute error is one of the more easily interpreted metrics. MAE measures the average difference between the predicted values and actual values and averages them out for the whole dataset. A lower MAE indicates a superior model, while a high MAE indicates that the prediction data is drastically different from the training data. Next, we selected the R-Squared score. The R-Squared score is also known as the “coefficient of determination”, and is often used specifically for linear regression models. In technical terms, the R-Squared score measures the “proportion of the variance for a target variable that can be interpreted from independent variables in a model”. Variance describes how data points differ from the mean of a dataset, therefore, R-Squared essentially describes the model’s success in “fitting” to the data points. Higher R-squared values indicate a better fit to the dataset, while lower values indicate the opposite. One of the main drawbacks of the R-Squared score is that it doesn’t appropriately define the “reliability of a model”, and therefore a poor model can have a high R-Squared value, despite it being flawed. Moreover, the R-Squared score cannot detect whether a model suffers from overfitting. Finally, we used an unconventional method of evaluation, which was the percentage error. We calculated this value by taking the square root of the mean squared error, and dividing it by the difference between the maximum and minimums of the test dataset, multiplying the result by 100. This is not typically used to evaluate machine learning models, however, it's another metric that can give a broad estimation of a model's performance. Overall, the mean absolute error is clearly the most consistent way to interpret regression models. For our individual sector models, we received fairly low MAE values, indicating good performance. However, the final stacking regression model did not perform as well as the individual ones, which could have multiple explanations. Firstly, despite compiling sufficient amounts of data, it's possible that this was still not enough to predict a variable as complex as a country’s GDP. The more likely reason is that U.S GDP simply cannot be predicted by simply combining these sectors together. There are a variety of other factors which go into GDP accumulation, such, as investments, imports and exports, consumer spending, and inflation. At any given time, a certain event or policy could heavily impact GDP and play a larger role in the calculation, and these fluctuations can be difficult to predict. There are many small details which cannot be captured with the analysis we have conducted in this project, and therefore, a more comprehensive analysis of every single sector in the U.S economy, along with tracking federal activity with regard to economic policies would be needed to improve the accuracy of this GDP model. However, we can conclude from our analysis that the sectors we have selected do play a large role in GDP calculations, and this baseline conclusion can be very impactful in terms of informing economic policies for the future.

References:

*Wang, Jian. “How to Use Model Stacking to Improve Machine Learning Predictions.” Medium, 20 Sept. 2021, medium.com/geekculture/how-to-use-model-stacking-to-improve-machine-learning-predictions-d113278612d4.*

*Carter, Michael J. “Ensemble Methods in Machine Learning: What Are They and Why Use Them?” Toptal, 26 Feb. 2021,* [*www.toptal.com/machine-learning/ensemble-methods-machine-learning#:~:text=Ensemble%20methods%20are%20techniques%20that,winning%20solutions%20used%20ensemble%20methods*](http://www.toptal.com/machine-learning/ensemble-methods-machine-learning#:~:text=Ensemble%20methods%20are%20techniques%20that,winning%20solutions%20used%20ensemble%20methods)*.*

*"R-Squared." Investopedia, Investopedia, 14 Sept. 2021,* [*www.investopedia.com/terms/r/r-squared.asp*](http://www.investopedia.com/terms/r/r-squared.asp)*.*

*Homepage - U.S. Energy Information Administration (EIA), https://www.eia.gov/.*

*“Manufacturing Industry Statistics.” NIST, 4 Jan. 2023, https://www.nist.gov/el/applied-economics-office/manufacturing/total-us-manufacturing#:~:text=In%202021%2C%20Manufacturing%20contributed%20%242.3,an%20estimated%2024%20%25%20of%20GDP.*

*Maccarrone, Giovanni, et al. “GDP Forecasting: Machine Learning, Linear or Autoregression?” Frontiers, Frontiers, 20 Sept. 2021, https://www.frontiersin.org/articles/10.3389/frai.2021.757864/full.*

*“GDP by Industry.” GDP by Industry | U.S. Bureau of Economic Analysis (BEA), https://www.bea.gov/data/gdp/gdp-industry.*

*Https://Www.ers.usda.gov/Data-Products/Agricultural-Productivity-in-the-u-s/Agricultural-Productivity-in-the-u-s/#State-Level%20Tables,%20Relative%20Level%20Indices%20and%20Growth,%201960-2004-Inputs*.