Economic Analysis and GDP Predictions for US Counties

This final project report submitted in fulfilment of the

requirement for the award of degree of

Master of Science

In

Data Science

By

SUSHRUTH VELDI

Under the Guidance of

Dr. Abdi Awl, D.Eng.

Professorial Lecturer

Columbian College of Arts & Sciences

GEORGE WASHINGTON UNIVERSITY

Table of Contents:

Glossary of Terms

- **Economic Indicators:** Variables that provide insights into the economic performance of a region, such as employment rates, income levels, and migration data.
- **Feature Engineering:** The process of creating new variables or transforming existing data to improve the performance of predictive models (e.g., creating "Establishments per Capita").
- **Log Transformation:** A data normalization technique that applies a logarithmic function to reduce the skewness in data distributions.
- **Correlation Matrix:** A table showing the relationship between variables, used to identify strong positive or negative relationships.
- **Random Forest:** A machine learning model that uses multiple decision trees to improve prediction accuracy and control overfitting.
- **Gradient Boosted Machines (GBM):** An ensemble machine learning technique that builds models sequentially to correct errors from previous iterations.
- **Support Vector Machines (SVM):** A model used for classification and regression that finds a hyperplane in high-dimensional space to separate data points.
- **Linear Regression:** A statistical method used to model the relationship between a dependent variable and one or more independent variables.
- **Root Mean Squared Error (RMSE):** A measure of the differences between predicted and observed values, used to evaluate model accuracy.
- **R-squared (R^2):** A statistical measure indicating the proportion of variance in the dependent variable explained by the independent variables in a model.
- **MAE (Mean Absolute Error):** A metric that measures the average absolute difference between predicted and actual values, indicating prediction accuracy.
- **Hyperparameter Tuning:** The process of optimizing model parameters to enhance predictive accuracy and generalization.

1. Introduction

1a. Introduction/Background

Risk of economies getting stuck in a vicious cycle is a pressing issue, particularly at the local level, where disparities in economic performance can lead to prolonged stagnation. Predicting county-level GDP is crucial to identifying regions at risk of economic stagnation and informing policymakers about where to focus investments and resources. Addressing this problem is essential for informing policy decisions and economic planning at a granular level, making it a valuable focus for this capstone project.

1b. Problem Statement

Economic disparities among counties can create significant challenges, particularly for regions at risk of long-term economic stagnation. Predicting county-level GDP is essential to identifying these vulnerable areas and guiding efforts to support local economic growth. This project aims to predict GDP at the county level by analyzing key economic indicators, such as employment rates, income levels, population growth, median age, migration data, poverty rates, Median household income ,labour statistics, number of business establishments etc. By developing and testing predictive models, the project will assist policymakers in making informed decisions and allocating resources effectively to prevent economic decline and stimulate development.

1c. Problem Elaboration

County-level GDP is a key measure of economic health and helps pinpoint areas in need of targeted support. This project aims to predict GDP at the county level by leveraging various economic indicators, such as employment rates, income levels, population trends, median age, migration patterns, poverty rates, labor data, median household income, and the number of business establishments. By building and testing predictive models, the project will uncover valuable insights to support data-driven policymaking. These findings will aid in resource allocation, identify at-risk areas, and inform strategies to counter economic decline and encourage growth. The ultimate goal is to provide actionable recommendations that reduce economic disparities and foster regional development.

1d. Motivation

This project is motivated by the pressing need to tackle economic inequalities among counties, which often lead to stagnation, inequality, and a decline in living standards. Many struggling regions lack the necessary tools and data to identify their challenges and create effective growth plans. By predicting county-level GDP and examining key economic factors, this project aims to provide policymakers with practical insights to allocate resources wisely, focus investments, and develop strategies that support fair economic development. Using advanced machine learning methods and detailed datasets, this work seeks to strengthen vulnerable regions, promote steady growth, and build resilience against economic challenges.

1e. Project Scope

This project aims to predict county-level GDP to identify regions at risk of stagnation and provide actionable insights for policymakers. Using data from 2017 to 2022 from sources like BEA, USCB, CBP, and USBLS, the analysis incorporates key economic indicators and engineered features like GDP per capita. Advanced machine learning models, including gradient boosting and random forests, will be optimized and evaluated using metrics like R-squared and RMSE. The findings will guide resource allocation, policy decisions, and strategies to promote equitable economic growth while addressing limitations and exploring future enhancements like integrating sentiment data or advanced modeling techniques.

2. Literature Review

This project draws upon established research in economic forecasting, machine learning applications, and socio-economic analysis, emphasizing the use of predictive models to address regional economic disparities. Key studies and methodologies informing this work are outlined below.

Relevant Research

1. **Regional Economic Disparities and Growth** Studies on regional economic disparities emphasize the role of granular data in identifying areas at risk of stagnation. Research by the International Monetary Fund (IMF, 2021) highlights the importance of migration patterns, GDP per capita, and employment rates in understanding the economic health of regions. These indicators are critical for developing targeted interventions to address local economic challenges.
2. **Predictive Modeling for GDP Estimation** Predictive models have demonstrated significant effectiveness in estimating GDP. Kuhn and Johnson (2013) highlighted

the value of feature selection and regularization in improving model accuracy and reliability for complex datasets. Additionally, research by Adewale and Ebembe (2024) showed that ensemble methods, such as Random Forest, achieved exceptional accuracy (R^2 = 0.96) by incorporating features like life expectancy, healthcare expenditure, and infrastructure access. These findings demonstrate the strength of advanced machine learning techniques in capturing intricate economic relationships.

3. **Socio-Economic Impacts of Inequality** Research indicates that economic inequality influences social cohesion, migration patterns, and employment dynamics. The IMF (2021) emphasized how poverty and unemployment rates contribute to disparities between urban and rural areas. Including socio-economic variables like poverty estimates and household income in predictive models provides a more comprehensive understanding of regional economic performance.

4. **Machine Learning Applications in Economic Forecasting** Advanced machine learning models, particularly Gradient Boosting Machines and Random Forests, are highly effective in handling non-linear relationships and complex datasets. These models provide detailed insights into feature importance, helping to identify critical drivers of economic performance. Despite their complexity, these methods are increasingly favored for their ability to offer robust predictions.

Methodological Integration

Building on these studies, this project incorporates findings and techniques such as:

- Utilizing robust machine learning models, including Gradient Boosting and Random Forest, optimized through techniques like hyperparameter tuning.
- Leveraging socio-economic variables like poverty rates, migration data, and labor statistics to capture regional disparities.

By leveraging these insights, the project aligns with established research while addressing specific challenges associated with county-level GDP prediction.

3.Methodology

3a. Dataset Description

The dataset for this project covers the years 2017–2022 and consists of 17,808 instances with 20 features. It is compiled from reliable sources, including the Bureau of Economic Analysis (BEA), U.S. Census Bureau (USCB), County Business Patterns (CBP), and U.S. Bureau of Labor Statistics (USBLS). Key features include economic indicators such as employment rates, poverty levels, migration data, median household income, and labor statistics. The data provides a comprehensive view of county-level economic conditions, enabling detailed analysis and predictive modeling. These features were selected to capture the diverse factors influencing GDP and to facilitate robust forecasting efforts.

3b. Dataset Collection

The dataset was collected from reliable sources, including the Bureau of Economic Analysis (BEA), U.S. Census Bureau (USCB), U.S. Bureau of Labor Statistics (USBLS), and County Business Patterns (CBP), covering the years 2017–2022. It includes county-level gross domestic product (GDP), population trends, migration data, employment statistics, and industry metrics such as the number of establishments and annual payroll. These sources provide accurate and granular data essential for analyzing and predicting economic trends across counties. The integration of these datasets ensures a strong foundation for addressing regional economic disparities.

3c. Data Preprocessing and Visualization

Data Preprocessing Steps

Dataset Integration:

Data from multiple sources, including BEA, USCB, CBP, and USBLS, were merged into a unified dataset. This ensured a comprehensive collection of economic indicators such as GDP, migration data, and employment statistics at the county level.

Missing Value Handling:

Variables with over 10% missing data were removed, while the remaining missing values were imputed using the median. This approach-maintained data integrity and reduced bias from skewed distributions.

Outlier Treatment and Log Transformations:

Outliers were identified and treated using capping methods, while highly skewed variables like GDP per capita were log-transformed to normalize distributions. These steps stabilized variances and improved model interpretability.

Feature Engineering:

New metrics such as GDP per capita, payroll per capita, and establishments per capita were created to provide more meaningful insights into economic performance at the county level.

Standardization and Scaling:

Features were standardized using z-scores to ensure uniformity across variables, preventing scale differences from impacting model training.

Data Splitting and Validation:

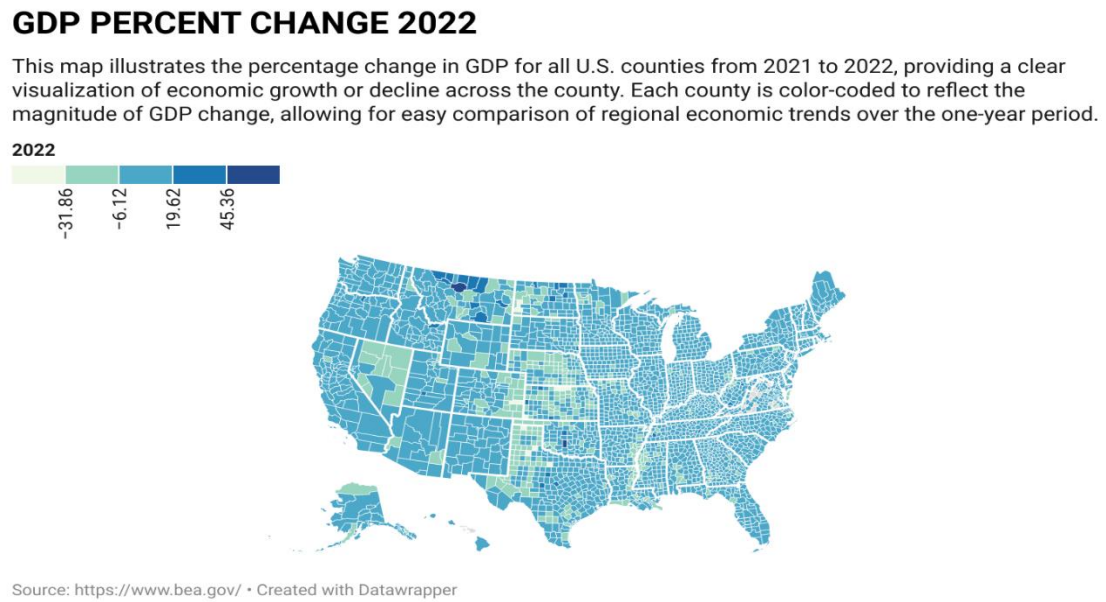
The dataset was split into training (2017–2020), validation (2021), and testing (2022) sets.

Data Visualization

Visualizations were instrumental in understanding the data and identifying key patterns.

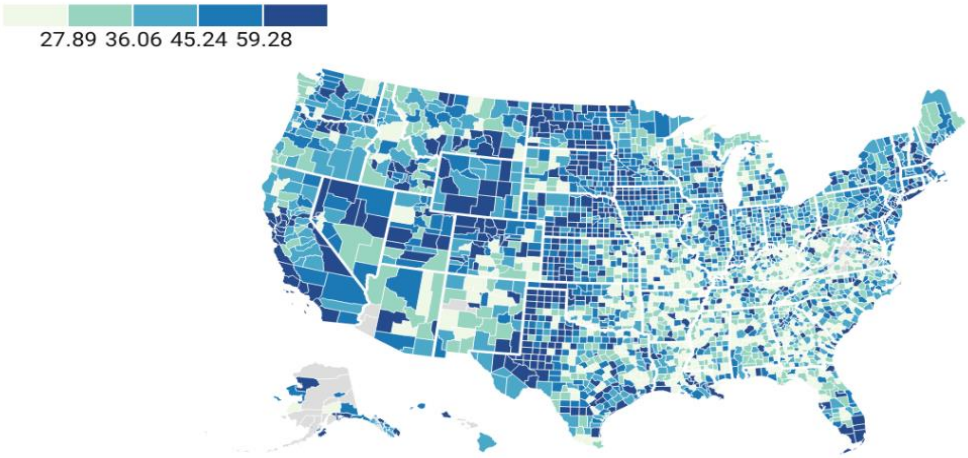
Geospatial Insights

Maps displaying GDP percent changes and GDP per capita across U.S. counties provided a geographical perspective on economic performance. These visualizations highlighted regions with substantial growth or decline, offering valuable insights into regional disparities



GDP PER CAPITA 2022

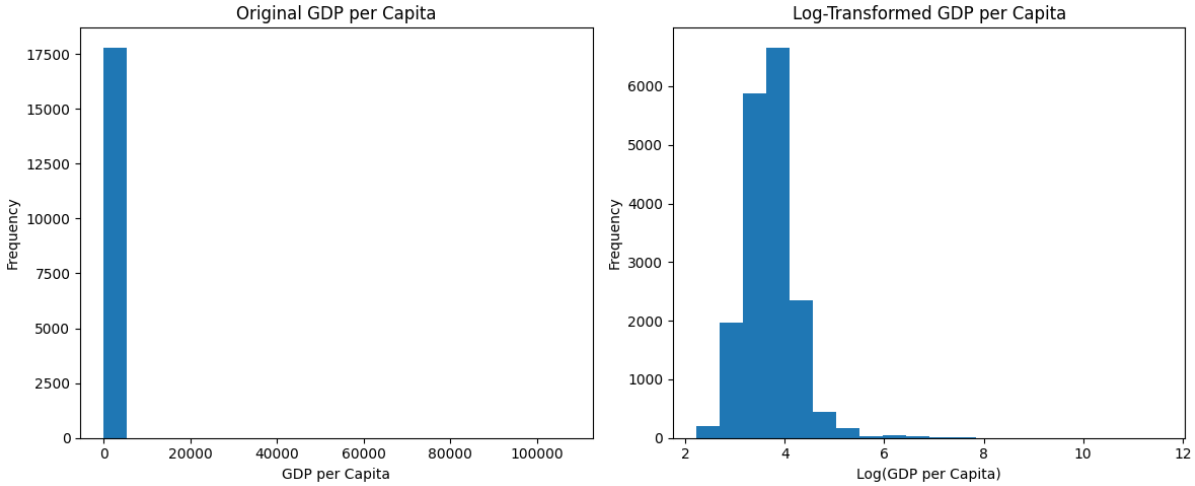
This map displays the GDP per capita for all U.S. counties in 2022, offering a visual representation of economic performance across the nation. Counties are color-coded based on the level of GDP change, enabling straightforward comparison of regional economic trends and highlighting areas of growth or decline over the past year.



Source: https://www.bea.gov/ • Created with Datawrapper

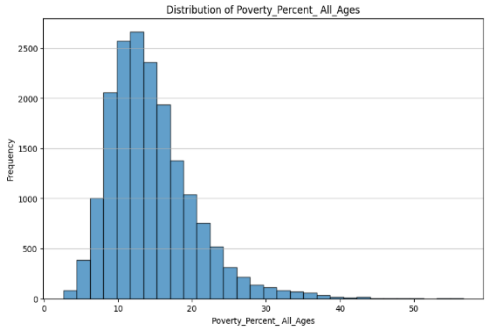
GDP Distribution:

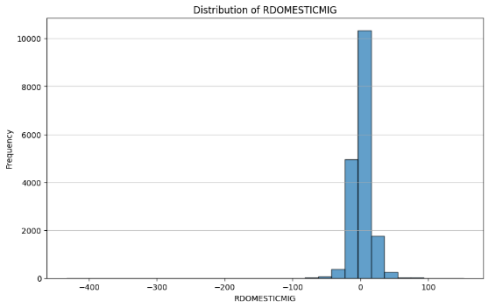
The distribution of GDP per capita before and after log transformation. The original distribution was heavily skewed, while the log-transformed version showed a more normalized pattern, making it suitable for machine learning models.

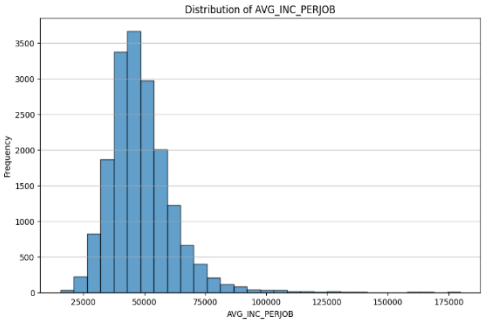


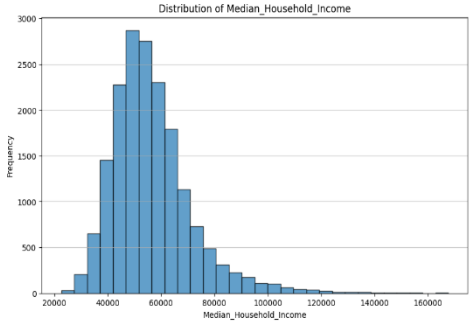
Feature Distributions:

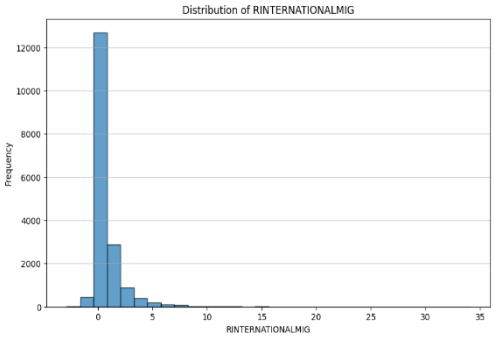
Distribution plots for variables such as poverty rates, median household income, unemployment rates, and migration data highlighted significant variability across counties. These visualizations provided insights into the underlying data structure and informed feature engineering.

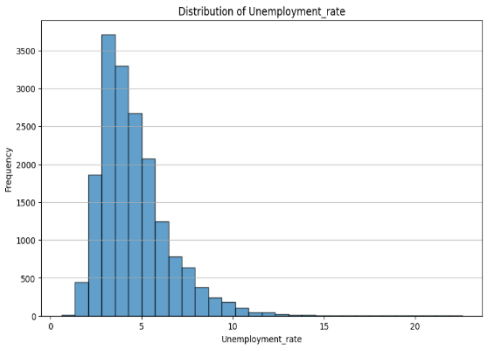






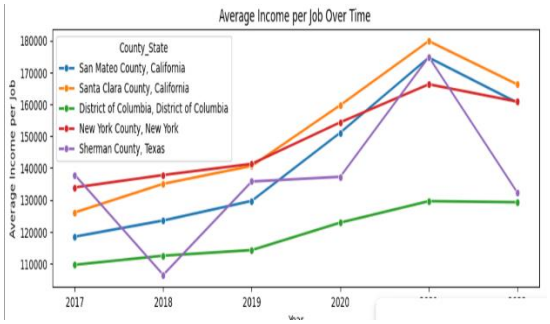


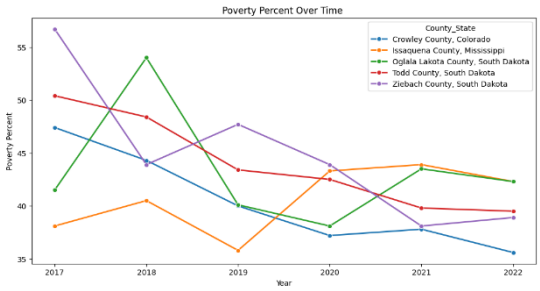


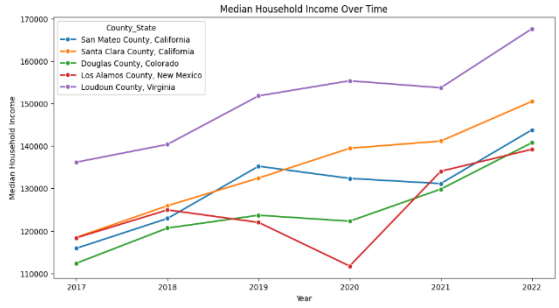


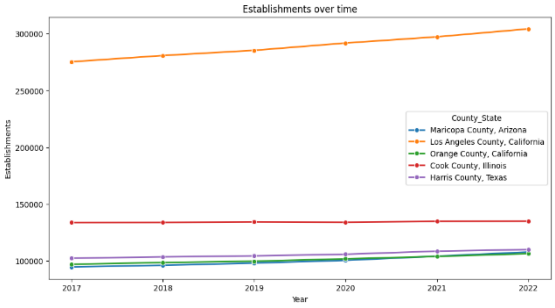
Temporal Trends:

Line charts illustrated trends over time for key metrics like poverty rates, median household income, and average income per job. These trends showed variations across selected counties, highlighting regional economic disparities and growth patterns from 2017 to 2022.



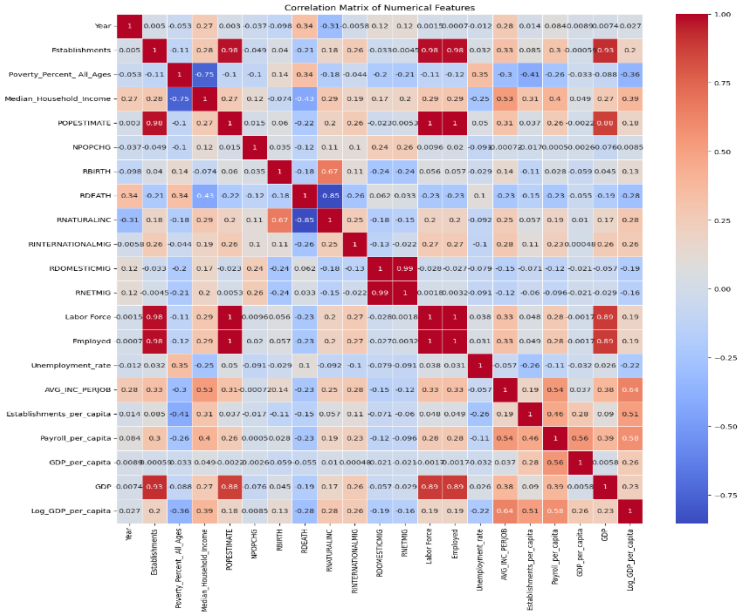






Correlation Matrix:

A heatmap of the correlation matrix revealed relationships between features and the target variable (GDP per capita). Strong positive correlations were observed with average income per job and payroll per capita, while unemployment and poverty rates showed significant negative correlations.



3d. Data Modeling

The data modeling process employed a combination of linear and non-linear machine learning techniques to predict county-level GDP with precision and reliability. Each model was chosen to balance interpretability, complexity, and predictive power.

Model Selection

A diverse set of models was employed to address the varying relationships within the dataset:

- Linear Regression: Served as a baseline due to its simplicity and ease of interpretation.
- Partial Least Squares (PLS): Addressed multicollinearity and reduced dimensionality while retaining predictive features.
- Support Vector Machines (SVM): Utilized for capturing non-linear relationships in high-dimensional data.
- Gradient Boosting Machines (GBM): Leveraged ensemble learning to boost prediction accuracy and rank feature importance.
- Random Forest: Offered robust performance by mitigating overfitting and identifying key predictors.

Feature Engineering

Key features were engineered and selected to optimize model performance:

- Critical Predictors: Variables such as Average Income Per Job, Payroll Per Capita, and Establishments Per Capita were identified as significant drivers of GDP.
- Transformations: Log transformation of GDP per capita normalized data distribution, enhancing the models' interpretability.
- Correlation Analysis: Multicollinearity was mitigated by identifying and managing redundant features.

Data Partitioning

The dataset, spanning 2017–2022, was divided to ensure robust training and evaluation:

- Training Set (2017–2020): Used to fit the models and establish foundational learning patterns.
- Validation Set (2021): Employed for hyperparameter tuning and model refinement.
- Testing Set (2022): Reserved for final evaluation to ensure model generalizability.

Hyperparameter Tuning

Hyperparameters were optimized using RandomizedSearchCV to enhance model accuracy and prevent overfitting:

- n_estimators: Number of trees in ensemble methods.
- max_depth: Controlled the complexity of the models.
- learning_rate: Adjusted step sizes for boosting iterations.

Evaluation Metrics

The following metrics guided model evaluation:

- R-squared (R^2): Quantified the proportion of variance explained by the model.
- Root Mean Squared Error (RMSE): Measured the average magnitude of prediction errors.
- Mean Absolute Error (MAE): Assessed average error in absolute terms.

The combination of these modeling strategies ensured the predictive models were both robust and interpretable, effectively capturing the economic dynamics at the county level.

4.Results & Analysis

This section highlights the performance of the predictive models and analyzes the findings to assess their effectiveness in predicting county-level GDP.

4.1 Model Performance

The table below summarizes the performance metrics of the five predictive models on the validation dataset. The models were evaluated using Root Mean Squared Error (RMSE), R-squared (R^2), and Mean Absolute Error (MAE).

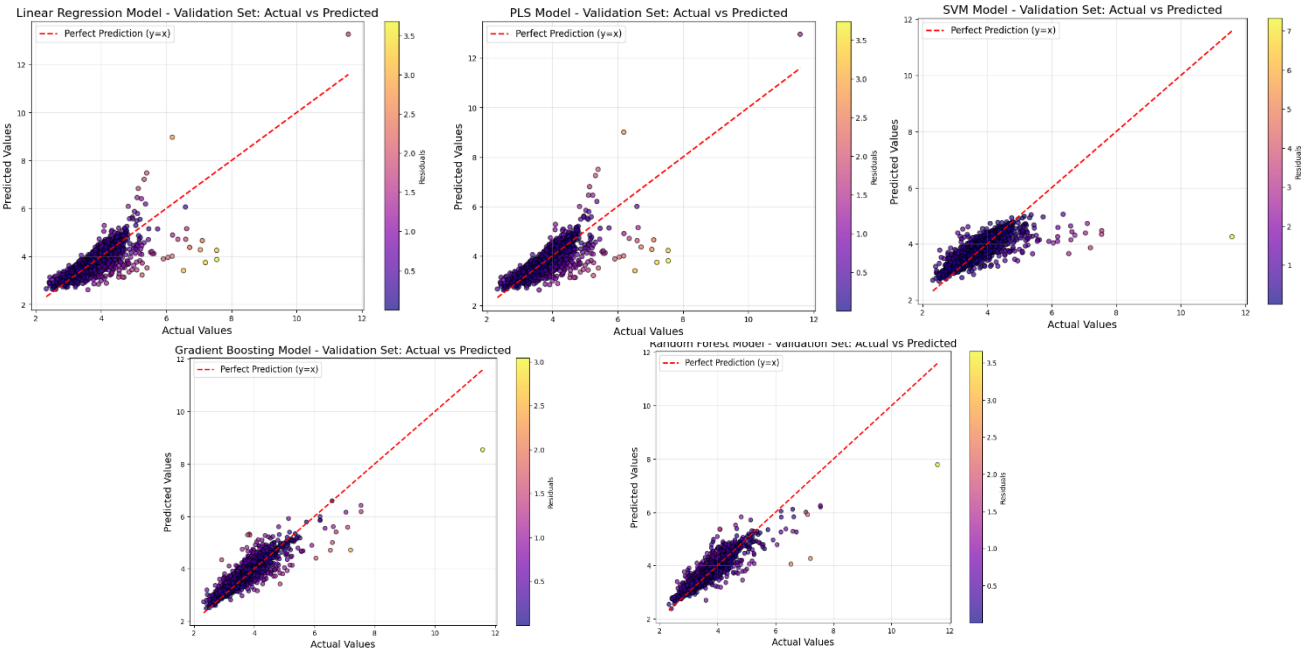
	Model	Validation RMSE	Validation R2	Validation MAE
0	Linear Regression	0.332183	0.612354	0.197497
1	Partial Least Squares (PLS)	0.332340	0.611988	0.196599
2	Gradient Boosting	0.256371	0.769103	0.168989
3	Support Vector Machines	0.353765	0.560346	0.213230
4	Random Forest	0.260688	0.761262	0.173094

4.2 Analysis of Model Performance

- **Gradient Boosting**:
    - Demonstrated the best overall performance, with the lowest RMSE (0.256371) and highest R^2 (0.769103). It effectively captured complex relationships in the data, making it the most suitable model for GDP prediction.
- **Random Forest**:
    - Achieved comparable performance to Gradient Boosting, with an RMSE of 0.260688 and an R^2 of 0.761262. Its interpretability and robustness make it a strong contender.
- **Linear Regression and PLS**:
    - Both linear models showed moderate performance, with similar RMSE and R^2 values (~0.332 and ~0.61, respectively). These models provide baseline insights but lacked the predictive power of ensemble methods.
- **Support Vector Machines (SVM)**:
    - Performed the worst among all models, with the highest RMSE (0.353765) and the lowest R^2 (0.560346). The complexity of the dataset may have limited its effectiveness.

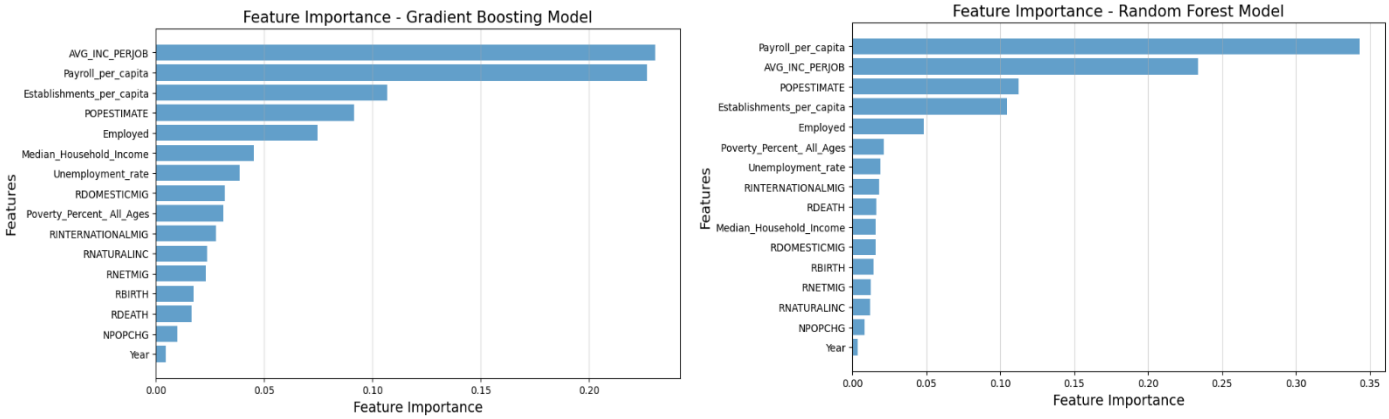
4.3 Actual vs. Predicted Plots

The **actual vs. predicted plots** visually assess the models accuracy by comparing the true GDP values against the predictions. Key insights include:



The actual vs. predicted plots show Gradient Boosting and Random Forest as top performers, closely aligning with the perfect prediction line and exhibiting minimal residuals. Linear Regression and PLS provide moderate accuracy, while SVM shows significant errors. The color gradient highlights residuals, with darker colors indicating smaller errors. Gradient Boosting performs the best overall.

4.4 Feature Importance



Gradient Boosting Insights: Key drivers of GDP prediction include Average Income Per Job, Payroll Per Capita, and Establishments Per Capita, highlighting the economic significance of workforce income and business density.

Random Forest Insights: Payroll Per Capita and Average Income Per Job emerged as the most critical predictors, consistent with Gradient Boosting results, while Poverty Percent and Unemployment Rate also showed notable importance in capturing socio-economic disparities.

5a. Conclusion

This project successfully developed predictive models to estimate county-level GDP, addressing economic disparities and informing policy decisions. Gradient Boosting and Random Forest emerged as the most effective models, demonstrating superior accuracy and identifying critical predictors like Average Income Per Job, Payroll Per Capita, and Establishments Per Capita. These insights provide actionable recommendations for resource allocation and targeted economic interventions. While the models were robust, limitations such as data availability and the interpretability of non-linear models highlight areas for improvement. Future research can incorporate additional data sources, such as public sentiment, and explore advanced techniques like deep learning to enhance predictive accuracy and practical utility.

5b. Project Limitations

While the project achieved its primary objectives, several limitations were identified that could influence the results and broader applicability. First, the dataset was limited to the years 2017–2022, which may restrict the generalizability of the models to long-term trends or future economic shifts. Additionally, the lack of granular or real-time data, such as industry-specific GDP metrics or precinct-level indicators, constrained the depth of the analysis. Non-linear models like Gradient Boosting and Random Forest, though highly accurate, are less interpretable compared to simpler models, which can make it challenging to communicate findings to policymakers. Lastly, factors such as external shocks (e.g., pandemics or policy changes) were not accounted for, potentially limiting the models' ability to adapt to dynamic economic conditions.

5c. Future Research

Future research can build on this foundation by integrating more granular datasets, such as real-time economic indicators, public sentiment data, or regional investment patterns, to enhance prediction accuracy and context. Advanced modeling techniques, including deep learning architectures, could be explored to capture complex patterns in larger datasets. Incorporating external factors, such as global economic trends or policy impacts, can improve the robustness of predictions.

6.References

https://www.bea.gov

https://www.census.gov

https://www.bls.gov

https://www.usda.gov

https://www.elibrary.imf.org/view/journals/001/2021/038/article-A001-en.xml

https://link.springer.com/book/10.1007/978-1-4614-6849-3

https://link.springer.com/article/10.1007/s43762-024-00116-2

https://www.sciencedirect.com/science/article/pii/S2772941924000619

7.Appendix

Technical Tools Used

1. Programming Language: Python
2. Libraries:
    - Scikit-learn (Model Building)
    - Pandas and NumPy (Data Processing)
    - Matplotlib and Seaborn (Visualization)

3.DataWrapper for Geospatial Analysis

	Year	Poverty_Percent_All_Ages	Median_Household_Income	POPESTIMATE	NPOPCHG	RBIRTH	RDEATH	RNATURALINC	RINTERNATIONALMIG	RDOMESTICMIG
count	17808.000000	17808.000000	17808.000000	1.780800e+04	17808.000000	17808.000000	17808.000000	17808.000000	17808.000000	17808.000000
mean	2019.500000	14.547125	56401.873035	1.041721e+05	411.273304	10.951839	11.771889	-0.820050	0.751994	1.540836
std	1.707873	5.752872	14841.529521	3.384258e+05	3604.549468	2.388327	3.403377	4.490913	1.544457	13.871106
min	2017.000000	2.600000	22679.000000	5.100000e+01	-180394.000000	0.000000	0.000000	-24.000000	-2.847077	-432.890317
25%	2018.000000	10.500000	46518.500000	1.069575e+04	-76.000000	9.508825	9.519099	-3.580677	0.000000	-5.406950
50%	2019.500000	13.500000	54086.500000	2.541250e+04	7.000000	10.828889	11.558582	-0.834796	0.293455	0.746584
75%	2021.000000	17.400000	63351.250000	6.678050e+04	220.000000	12.144239	13.798963	1.828399	0.888931	8.073240
max	2022.000000	56.700000	167605.000000	1.010371e+07	83011.000000	30.103142	33.579769	25.002861	34.093286	152.000000

The table summarizes key features in the dataset, including demographics, income, and migration metrics, for 17,808 records spanning 2017–2022. The average poverty rate is 14.55%, with median household income at $56,401 and population estimates ranging widely from small counties to over 10 million residents. Metrics like population change and migration show significant variability, reflecting diverse county-level dynamics. Birth and death rates are nearly balanced on average, while migration data highlights both domestic and international trends. These descriptive statistics underline the dataset's complexity and the importance of preprocessing for effective modeling.