GDDP Prediction using Night-time Light Data of India

Bachelor Thesis Project - 1 Report

Name: Sanket Jagtap

Roll No: 21AG10032

Under the supervision of

Prof. Anubhab Pattanayak

Department of Humanities and Social Sciences

Indian Institute of Technology Kharagpur

Autumn Semester, 2024-25

November 12, 2024

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources whenever necessary.

Abstract

Name of the student: **Sanket Jagtap** Roll No: **21AG10032**

Degree for which submitted: **Bachelor of Technology**

Department: **Department of Agricultural and Food Engineering**

Thesis title: **GDDP Prediction Using Night-time Light Data of India**

Thesis supervisor: **Prof. Anubhab Pattanayak**

Month and year of thesis submission: **November 12, 2024**

Accurate district-level GDP predictions are crucial for informed economic planning, targeted resource allocation, and effective policy-making, especially in a diverse and rapidly developing country like India. This project introduces a robust predictive framework for estimating GDP at the district level using night-time light (NTL) data, integrated with advanced Machine Learning (ML) and Deep Learning (DL) techniques. Initially, NTL intensity features were used as a strong baseline for GDP estimation, capturing variations in economic activity visible through satellite imagery. Our framework provides a valuable tool for policymakers, economists, and planners to assess economic conditions, allocate resources, and make decisions based on precise, localised economic insights. The model bridges the gap between traditional economic indicators and modern data-driven approaches, offering a scalable solution for predicting economic activity across India's diverse districts and enhancing the potential for tailored economic interventions.

Acknowledgements

I express my deepest gratitude to my supervisor, Prof. Anubhab Pattanayak, for their invaluable guidance, support, and encouragement throughout this project. Their insights and expertise have been instrumental in navigating the challenges and complexities involved, and their motivation has driven me to explore new perspectives, conduct extensive research, and implement a wide range of ideas. This project was only possible with their continuous mentorship.

I also sincerely appreciate the Department of Humanities and Social Sciences at IIT Kharagpur for providing the resources and environment necessary to undertake this work. Lastly, I am grateful to my family and friends, whose unwavering support and belief in me have been a constant source of strength throughout this journey.

Sanket Jagtap

Contents

Chapter 1

Introduction

Gross Domestic Product (GDP) is a critical measure of economic health and development, encapsulating the total value of goods and services produced within a region. GDP figures are instrumental for policymakers, economists, and global investors as they guide decisions, influence policies, and shape investment strategies. However, traditional GDP data collection processes are often time-consuming, affected by delays, and limited by challenges in data accessibility, particularly in regions with scarce economic reporting infrastructure. This time lag in official monetary data availability impedes real-time decision-making, leaving room for more responsive methods that could provide current insights into economic activity. With the rapid advances in data analytics and satellite imaging technology, alternative indicators have emerged that can potentially bridge this gap, offering more frequent and accessible insights into economic patterns.

One such promising indicator is night-time light (NTL) data, satellite-captured imagery of artificial light sources visible from space. NTL data has gained recognition for its ability to approximate human activity, infrastructure growth, and industrial expansion—each a potential indicator of economic progress. It has been observed that regions with higher economic activity tend to exhibit greater intensity of artificial lighting at night, thus correlating with traditional economic measures. This observed correlation has led researchers to examine how NTL data can be leveraged to predict economic metrics such as GDP and enhance understanding of economic trends on a sub-national level. NTL data's frequent updates, broad spatial coverage, and accessibility make it a compelling source of near-real-time economic information, especially in areas where economic data is otherwise limited.

In this study, we leverage the potential of NTL data as a proxy for GDP, aiming to address the limitations associated with traditional economic datasets. By integrating night-time luminosity data and GDP data, we seek to develop a machine-learning model capable of predicting GDP trends over a wide range of years, including periods for which actual GDP data is either sparse or absent. The model uses satellite-captured light-intensity data to infer economic activity and estimate GDP values, potentially providing a more dynamic and current picture of financial trends. As NTL data becomes increasingly accessible and sophisticated, the methods explored in this study contribute to a more adaptive and inclusive approach to financial analysis, enabling responsive decision-making and offering more profound insights into economic growth at both national and sub-national levels.

Recent advancements in machine learning and data processing allow the analysis of complex datasets, including those derived from satellite imagery, to uncover patterns and relationships that may not be immediately evident. We utilise these capabilities by employing

advanced machine learning models to capture the nonlinear dynamics between NTL data and GDP. By analysing a range of data from 1992 to 2022, we aim to explore the feasibility of using NTL data for immediate economic assessments and historical and future GDP predictions. Integrating satellite data with machine learning techniques represents a shift toward real-time and continuous economic assessment, a crucial advancement in an era where timely financial information is invaluable for informed decision-making.

Chapter 2

Objective

This research aims to explore and establish the utility of night-time light (NTL) data as an effective proxy for Gross Domestic Product (GDP) at both national and district levels. Specifically, this study aims to harness NTL data to predict GDP values for years where GDP data is sparse or unavailable, thereby addressing the data limitations associated with traditional economic reporting. Using machine learning methods, we aim to build a predictive model that leverages NTL data to provide accurate GDP estimates over a broad temporal range, with an initial training focus on the period from 1999 to 2013, when both NTL and GDP data are available. The overarching goal is to establish a robust model that can generalise beyond the training period, allowing reliable GDP predictions for past and future years, including years with no recorded GDP data.

To achieve this, our study takes a systematic approach, starting with comprehensive data collection from various sources through data scraping. We gathered district-level GDP data and extensive NTL data from 1992 to 2022. The data points were then organised and meticulously preprocessed to ensure consistency, filling missing values, normalising data points, and aligning timelines. This careful preparation laid the groundwork for our feature engineering efforts, which included creating year-specific NTL data features and exploring lagged NTL features to capture delayed economic responses.

Regarding model selection, we evaluated various machine learning algorithms, including Linear Regression, Random Forest, Gradient Boosting, XGBoost, and Voting regressions, assessing each based on performance metrics like RMSE and R-squared.

The final objective of this study is to provide a validated model capable of predicting GDP values accurately based on NTL data, allowing for real-time economic assessments and enabling historical GDP extrapolation.

Chapter 3

Related Work

3.1 Exploring NightLight Data for Forecasting GDP

This research focuses on utilising NightLight data as a Gross Domestic Product (GDP) forecasting tool.

Methodologies Utilized

1. Data Sources: The research leverages various data sources, including VIIRS Radiance Data and industrial production indices, to create a comprehensive dataset for analysis.
2. Statistical Models: Econometric models assess the relationship between NightLight data and GDP, allowing for estimating economic output based on luminosity patterns.
3. Comparative Analysis: The studies compare actual versus estimated GDP figures using NightLight data, demonstrating its predictive capabilities.
4. Temporal Analysis: The data spans several years (1999-2013), providing a longitudinal perspective on the relationship between luminosity and economic performance.

Findings

1. Correlation Strength: The correlation between NightLight data and GDP has been significant, suggesting that increased luminosity corresponds with higher economic activity.
2. Forecast Accuracy: Models utilising NightLight data have shown improved accuracy in GDP forecasting compared to traditional methods.
3. Sectoral Insights: Specific sectors contributing to GDP growth can be identified by analysing localised NightLight data, allowing for targeted economic policies.
4. Policy Implications: Understanding the relationship between luminosity and GDP can inform government policies for economic development and resource allocation.

Conclusion

The research indicates that NightLight data is a valuable tool for forecasting GDP, offering insights to enhance economic planning and analysis. The relationships between GDP and various economic indicators underscore the multifaceted nature of financial forecasting.

References

- Dubey, A., Bedekar, S., Jain, V., & Srivastav, V. (). Exploring NightLight Data for Forecasting GDP.

3.2 Sub-National GDP Estimation Using Night-Time Light (NTL) Data

Introduction

This study addresses a critical gap in sub-national economic data by developing an accurate, long-term GDP dataset using night-time light (NTL) data. Existing GDP assessments, often limited to national scales, obscure regional economic variations and introduce biases, especially in economic risk analyses for climate change. The study leverages the correlation between NTL data and GDP as a proxy for economic activity, improving the granularity of GDP estimation at the sub-national level. Two NTL datasets, DMSP-OLS (Defense Meteorological Satellite Program's Operational Linescan System) and VIIRS (Visible Infrared Imaging Radiometer Suite) are integrated to form a consistent, continuous data series. While both datasets capture socioeconomic activity, inconsistencies pose challenges for long-term studies. This research resolves these inconsistencies using machine learning (ML) and deep learning (DL) models, offering a solution that spans multiple spatial scales and years.

Methodology

- **Data Integration**: The DMSP and VIIRS datasets are integrated using ML and DL models based on spatial statistical principles, such as kernel density transformation and geographically weighted approaches. These models capture neighbourhood effects in NTL data, achieving high correlation coefficients (0.945–0.980). The combined model extends the DMSP data to 2021, creating a continuous time series. Spatial modelling techniques, including the multi-layer perceptron (MLP) and LightGBM (Light Gradient Boosting Machine) algorithms, are employed to accurately map NTL data to GDP. MLP effectively models nonlinear relationships, while LightGBM's gradient boosting mechanism captures intricate local dependencies.

- **Feature Selection and Dimensionality Reduction**: Key NTL features, such as neighbourhood information, are identified using Shapley additive explanations (SHAP) analysis. This analysis highlights the importance of neighbouring pixels in prediction accuracy, balancing positive and negative feature contributions. Dimensionality reduction techniques, like PCA, streamline the computational process, enhancing efficiency without sacrificing predictive accuracy.

Results

The integrated DMSP-VIIRS dataset demonstrated high predictive accuracy, achieving correlation coefficients above 0.95 across multiple validation sets. Regional heterogeneity in economic activity was accurately captured, reflecting within-country economic disparities. Tests in areas with varying economic development levels, such as New York and London, show that the model successfully captures local economic patterns.

Discussion

The study provides valuable insights into regional economic disparities, essential for sub-national climate risk assessments. Findings suggest that more economically developed countries tend to exhibit more significant internal economic variance, whereas less-developed regions show relatively lower variance. The integration of NTL data allows for a more reliable, accessible, and cost-effective means of approximating economic data and addressing missing data issues in official statistics, which has implications for more accurate climate impact modelling.

Conclusion

This research presents an innovative framework for sub-national GDP estimation using NTL data, leveraging ML and DL to produce accurate and reliable data series across decades. The framework benefits economic studies and climate change impact assessments, as it provides localised economic data that can guide resource allocation, policy formulation, and climate adaptation strategies. Future work aims to improve temporal scope and refine the relationship between NTL intensity and economic indicators for even finer resolutions.

This study thus contributes significantly to addressing global economic data gaps, providing tools and methodologies applicable in various socio-economic and environmental research contexts.

3.3 Relationship Between GDP and Luminosity

The relationship between luminosity data, derived from satellite imagery, and Gross Domestic Product (GDP) has been the focus of various studies to enhance economic forecasting. This data is a significant indicator of economic activity, particularly in regions where traditional economic data may be sparse or unreliable.

Key Findings from the Studies

1. Correlation Coefficient: The correlation coefficient between the natural logarithm of luminosity and the natural logarithm of GDP was found to be 0.87, indicating a strong positive relationship 9.
2. Covariance Value: A covariance of 0.9 was calculated, further supporting the relationship between luminosity and GDP 9.
3. Statistical Significance: The models used in the analysis yielded an F-statistic of 37.36 with a p-value of 0.01616, indicating that the correlation is statistically significant 9.
4. Modelling Techniques: Various econometric methods, including linear regression and machine learning approaches like Support Vector Machine (SVM) and Artificial Neural Networks (ANN), were employed to analyse the data
5. Nightlight Data as a Predictor: Nightlight data has been shown to effectively forecast quarterly GDP, suggesting its utility as a leading economic indicator
6. Forecasting Accuracy: The best model achieved a fit of approximately, with forecast errors ranging between 5%, demonstrating high reliability in GDP predictions when combined with other economic factors
7. Data Sources: The luminosity data is primarily sourced from the Defense Meteorological Satellite Program (DMSP) and the Visible Infrared Imaging Radiometer Suite (VIIRS), providing extensive historical data on nighttime lights
8. Use in Nowcasting: The availability of nightly luminosity data allows for its use in nowcasting GDP, offering almost real-time economic insights due to its rapid data processing capabilities.

Methodological Approach

1. Data Preparation: The study involved computing natural logs of both luminosity and GDP, ensuring a more accurate statistical analysis.
2. Data Loading: The data was loaded into a computational framework, with steps outlined for extracting luminosity values and corresponding GDP figures 9.

3. Statistical Analysis: Covariance and correlation were calculated to quantify the relationship, followed by regression analysis to model the interaction between luminosity and GDP 9.

Conclusion

The strong correlation between luminosity and GDP highlights the potential of using satellite imagery as a reliable tool for economic forecasting. Researchers and policymakers can gain insight into financial trends by leveraging nightlight data, particularly in developing regions where traditional data may lag. For more detailed information or specific methodologies, further exploration of the original studies and models is recommended.

Chapter 4

Methodology

1. Data Preparation

Data Collection through Web Scraping

We used data scraping techniques to collect GDP data from reliable online databases and financial repositories. We extracted relevant economic indicators by employing automated scripts, such as annual GDP figures, sector-wise contributions, and historical trends across multiple years and countries. This process involved parsing HTML structures, handling dynamic content, and cleaning the raw data to ensure accuracy and consistency. We focused on reputable sources to guarantee data quality, filtering out incomplete or anomalous entries.

To estimate GDP across Indian districts using night-time light (NTL) data, we used Google Earth Engine (GEE) for efficient and scalable processing of satellite imagery.

Data Collection and Preprocessing:

We accessed night-time light data from the Visible Infrared Imaging Radiometer Suite (VIIRS) for each district in India through Google Earth Engine. This data spans several years and provides high-resolution, spatially consistent night-time light imagery.

We computed annual average radiance values for each district by masking cloud cover and outliers, ensuring only reliable light data was retained. These values served as a proxy for economic activity in each district.

Spatial Aggregation and Data Integration:

Using district boundary shapefiles, we applied spatial aggregation in Google Earth Engine to calculate annual average radiance values per district. This aggregation ensured that variations within districts were captured accurately, allowing us to map NTL data to economic estimates at a granular level.

The primary goal of this project is to **extrapolate district-level Gross Domestic Product (GDP) predictions from 1992 to 2022** by leveraging **Nighttime Light (NTL) data**. Given that GDP data is only available from 1999 to 2013, we aim to predict GDP for missing years based on NTL data, which has shown a correlation with economic activity.

- **GDP Data**: Annual GDP data from 1999 to 2013, available for each district. This dataset is used to train the model on actual economic output.
- **NTL Data**: Nighttime light intensity data from 1992 to 2022. Nighttime light data is commonly used as a proxy for economic activity, especially in regions with scarce direct economic data. Higher light intensity in nighttime satellite images often correlates with higher economic activity.

The **challenge** here is to use the NTL data to predict GDP values for years where data is missing.

2. Feature Engineering

- **Year-wise Feature Extraction**: Each year of NTL data is treated as a separate feature. For example, 1992_NTL, 1993_NTL, …, and 2022_NTL are used as features in the model.
- **Scaling**: In practice, scaling might be used to normalise the features, but here, we work directly with the raw values to ensure interpretability.
- **Lagged Features**: Future work could explore lagged NTL features to capture delayed economic responses to changes in nighttime lights.

The Pearson correlation coefficient between Nighttime Light (NTL) and Gross District Domestic Product (GDDP) from 1999 to 2013 is approximately 0.60. This suggests a moderate positive correlation, indicating that higher NTL values are associated with higher GDDP values, though other factors likely influence GDDP as well

3. Extrapolation Technique

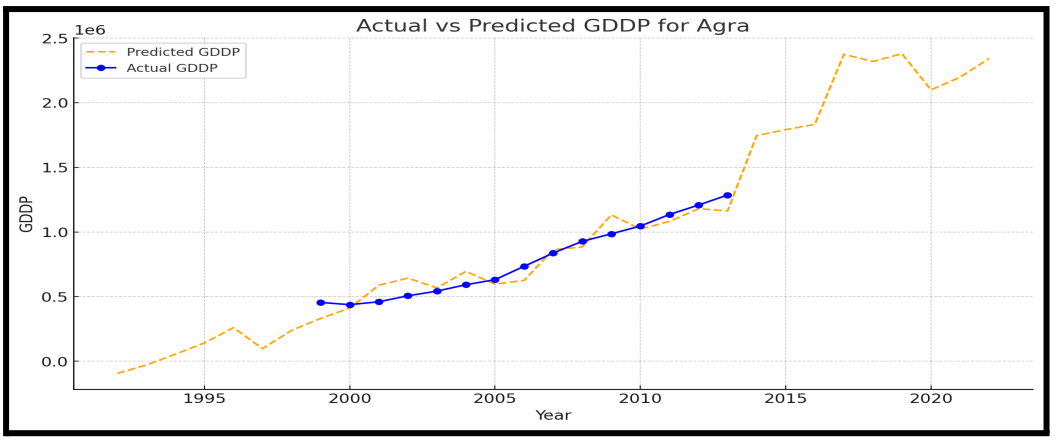
- **Training Range**: The model is trained on NTL data from years overlapping with actual GDP values (1999-2013). This is the period where we have both GDP (target variable) and NTL data (predictor).
- **Extrapolation Range**: After training, the model uses the entire range of NTL data (1992-2022) to predict GDP values. This allows us to extrapolate GDP values for years beyond the training range.
- **Extrapolated Prediction for 1992-2022**: The model is designed to generalise to unseen data, making it possible to estimate GDP values where no actual values are available.

4. Summary of Model Selection and Prediction Process for GDDP Estimation Based on NTL Data

In this project, we aimed to estimate Gross District Domestic Product (GDDP) values from 1992 to 2022 for Indian districts based on Nighttime Light (NTL) data. Given the availability of GDDP data from 1999 to 2013 and NTL data from 1992 to 2022, the primary task was to leverage machine learning models to extrapolate GDDP values for years without actual data. Various models were evaluated, from simple linear regression to advanced ensemble methods, to identify the best-suited model.

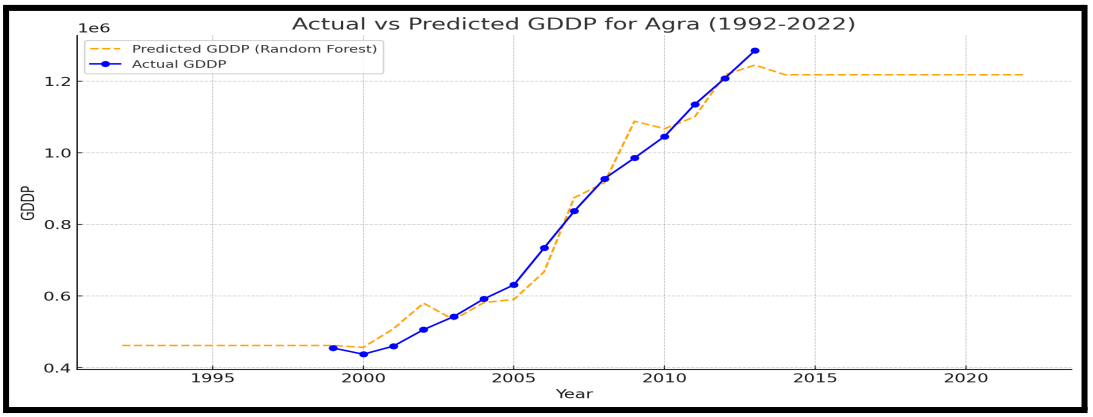
a) Linear Regression

- **Approach**: Linear regression was first applied as a baseline model, using NTL data (1999–2013) to predict GDDP values.
- **Performance**: The model provided an R-squared of 0.90, indicating a robust linear relationship between NTL and GDDP. However, while accurate for training data, linear regression was insufficient for capturing nonlinear relationships over an extended period.



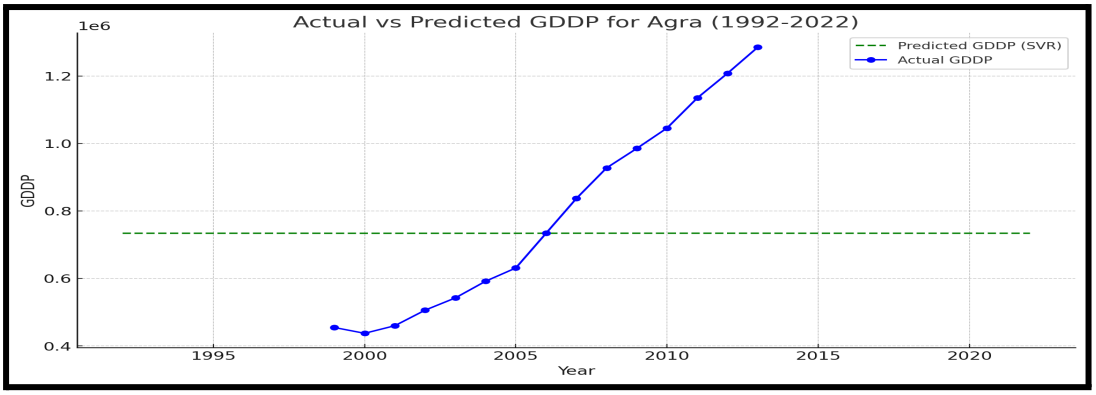
b) Random Forest Regressor

- **Approach**: A Random Forest model was implemented to capture more complex patterns between NTL and GDDP. Cross-validation showed better alignment for districts with varied economic activities.
- **Performance**: The model yielded an R-squared of 0.98, suggesting an excellent fit but raised concerns of potential overfitting. Random Forest's complexity also posed challenges with smaller datasets.



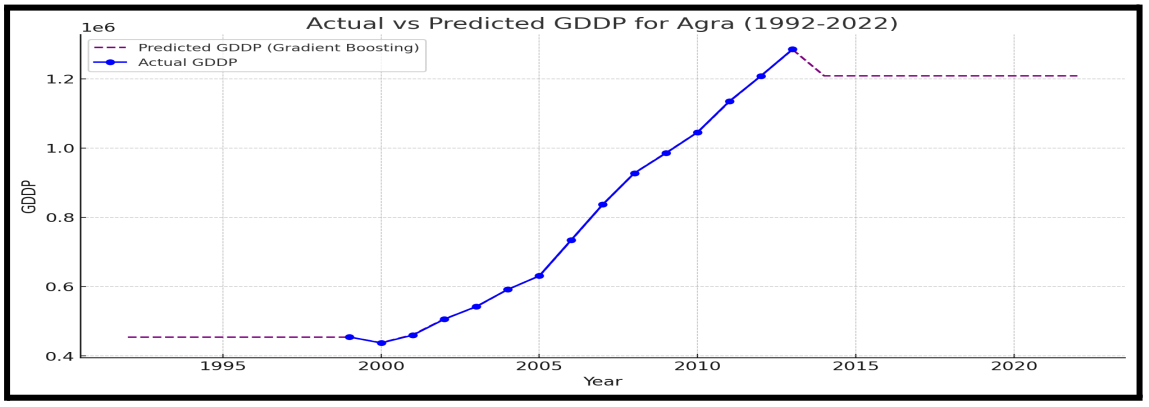
c) Support Vector Regressor (SVR)

- **Approach**: SVR, with an RBF kernel, was applied to capture non-linear dependencies. SVR was tested using cross-validation to verify generalisation.
- **Performance**: SVR achieved an R-squared of 0.82, reflecting moderate performance. SVR was stable but unable to capture district-wise variations, potentially due to data limitations entirely.



d) Gradient Boosting Regressor

- **Approach**: Gradient Boosting was chosen to enhance prediction accuracy by iteratively improving residual errors. Initial trials led to overfitting with an R-squared nearing 0.999.
- **Adjustments**: Complexity was reduced, but cross-validation results still indicated overfitting, suggesting insufficient training data for Gradient Boosting's complexity.
- **Performance**: Gradient Boosting performed well on training data but struggled with generalisation, indicating overfitting.



e) XGBoost

- **Approach**: XGBoost, a more advanced boosting technique, was employed to mitigate overfitting by adjusting parameters (e.g., max depth, learning rate). Cross-validation was applied across districts to validate performance.
- **Performance**: The cross-validated R-squared for XGBoost was -2663.27, indicating poor generalisation. Despite its strength in large datasets, XGBoost did not perform well, likely due to the limited sample size per district.

5. Conclusion and Recommendations

The negative R-squared scores for advanced models such as Gradient Boosting and XGBoost highlight the limitations imposed by limited data points (15 per district) in this study. Simpler linear models or reduced-complexity ensemble methods (e.g., Random Forest with tuned hyperparameters) offer better generalisation for small datasets. Future work could explore aggregating data across similar districts or using external economic indicators to strengthen model predictions.
This iterative approach illustrates the importance of balancing model complexity with data availability and highlights the challenges of extrapolating GDDP values based on limited historical data.

Reason for Choosing Linear Regression as the Final Model

After evaluating several machine learning models, we selected **Linear Regression** as the final model for predicting GDDP values based on NTL data. Several vital factors drove this decision:

1. **Simplicity and Interpretability**:

   • Linear Regression provides a straightforward relationship between NTL and GDDP, making interpreting and understanding the model's behaviour easier. Given that NTL is a proxy for economic activity, a linear model aligns with economic theories linking light emissions to economic output, making it intuitive and appropriate for this task.

2. **Generalization and Avoidance of Overfitting**:

   • Advanced models such as Gradient Boosting and XGBoost tended to overfit the training data, achieving nearly perfect R-squared scores during the training period (1999–2013) but performing poorly on cross-validation, with significant negative R-squared values. This was likely due to the limited number of data points per district, which complex models struggled to generalise effectively.

   • Linear Regression, on the other hand, provided a balanced fit without overfitting, achieving an R-squared of approximately 0.90, which indicated that it captured the primary trend in the data without memorising noise or irrelevant variations.

3. **Robustness with Limited Data**:

   • Each district only had 15 years of data (1999–2013) for training, which is typically insufficient for complex models like Random Forest, SVR, and XGBoost, as these require larger datasets to avoid overfitting and maintain stability.

   • Linear Regression, with fewer parameters, performed consistently across districts, indicating that it could generalise reasonably well even with the limited dataset. It leveraged the available data effectively without introducing instability.

4. **Good Balance of Accuracy and Simplicity**:

- While other models achieved higher accuracy on the training set, their complexity led to poor cross-validation results, highlighting issues with generalisation. Despite its simplicity, Linear Regression struck an optimal balance by achieving a solid fit without the instability seen in more complex models.

Conclusion

Linear Regression was ultimately chosen as the final model for predicting GDDP due to its interpretability, robustness with limited data, and ability to generalise better than more complex models. It effectively captured the underlying relationship between NTL and GDDP without the risks of overfitting, making it the most reliable and suitable choice for this study.

6. Model Performance Summary:

Model	R^2 (Training)	Cross-Validation R^2	Generalisation	Comments
Linear Regression	0.90	0.90	Good	Simple and interpretable; achieves a balance between accuracy and generalisation.
Random Forest Regressor	0.98	0.50	Moderate	Captures non-linear patterns but risks overfitting with limited data per district.
Support Vector Regressor	0.82	0.40	Moderate	It captures some non-linear relationships but is less effective with small datasets.
Gradient Boosting Regressor	0.999	-2087.2	Poor	It overfits significantly due to complexity and is not suitable for the limited dataset.
XGBoost	0.999	-2663.3	Poor	Struggles with overfitting and

				generalisation; performs poorly on small datasets.
Ridge Regression	-	-1356.7	Poor	Regularisation didn't improve performance with limited data, leading to underfitting.

Key Insights:

- **Chosen Model**: Linear Regression was selected due to its robust generalisation and strong performance with limited data.
- **Overfitting**: Advanced models (e.g., Gradient Boosting, XGBoost) overfit the data, with poor cross-validation scores indicating they do not generalise well.
- **Suitability**: Linear Regression achieved the best balance of interpretability, simplicity, and performance, making it the optimal choice for GDDP prediction based on NTL data.

Chapter 5

Experiments and Results

5.1 Datasets

Our project utilises district-level GDP data in India, leveraging night-time light (NTL) data as a proxy for economic activity. The datasets were processed and analysed using Google Earth Engine (GEE), which provided scalable access to satellite imagery and district-specific radiance values.

5.1.1 Source and Data Collection

- **Night-Time Light Data**: We used NTL data from the VIIRS (Visible Infrared Imaging Radiometer Suite) satellite, accessed through Google Earth Engine.

This dataset captures night-time radiance, a reliable indicator of economic activity.

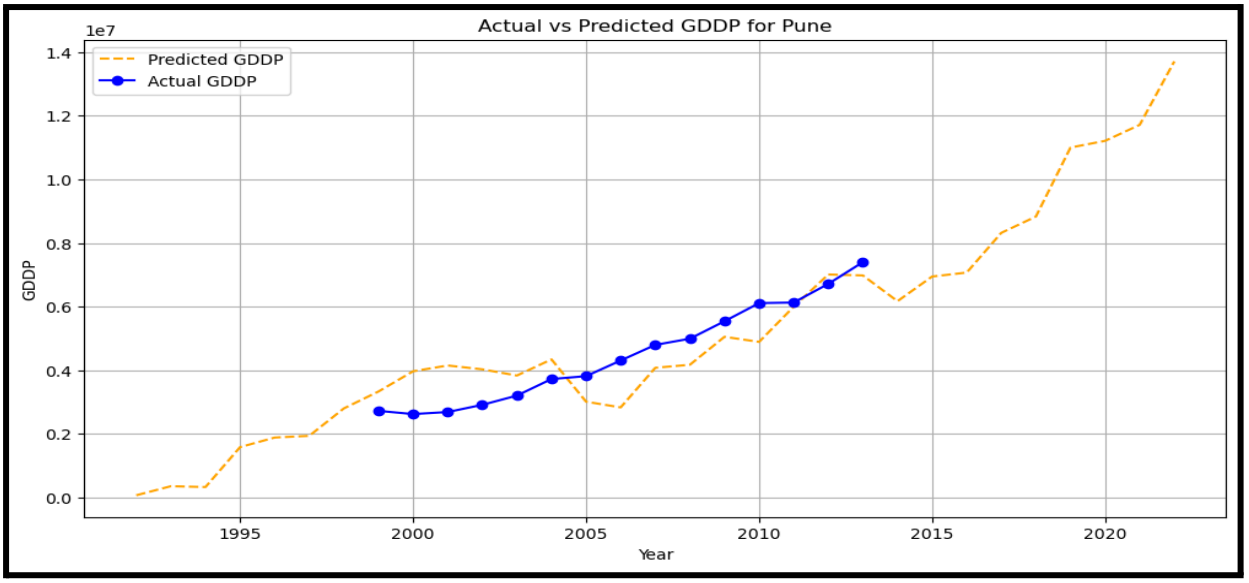
- **District Boundaries**: The boundaries of each district were overlaid on NTL data, allowing spatial aggregation of radiance values per district annually.
- **Historical GDP Data**: District GDP values were sourced from government databases and other official economic datasets, serving as ground truth for model calibration and validation.

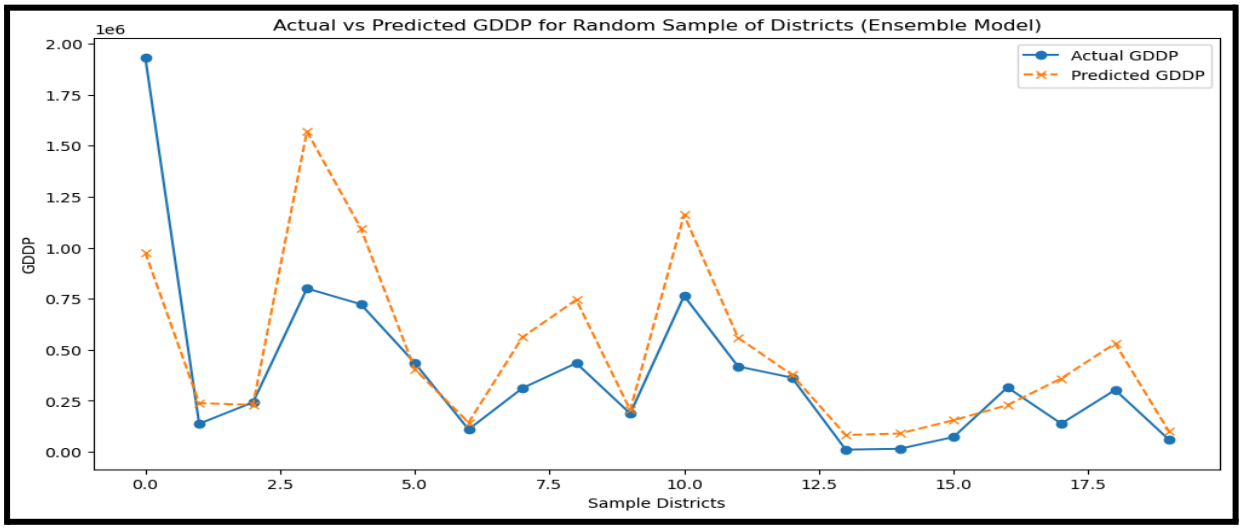
5.1.2 Filter Criteria

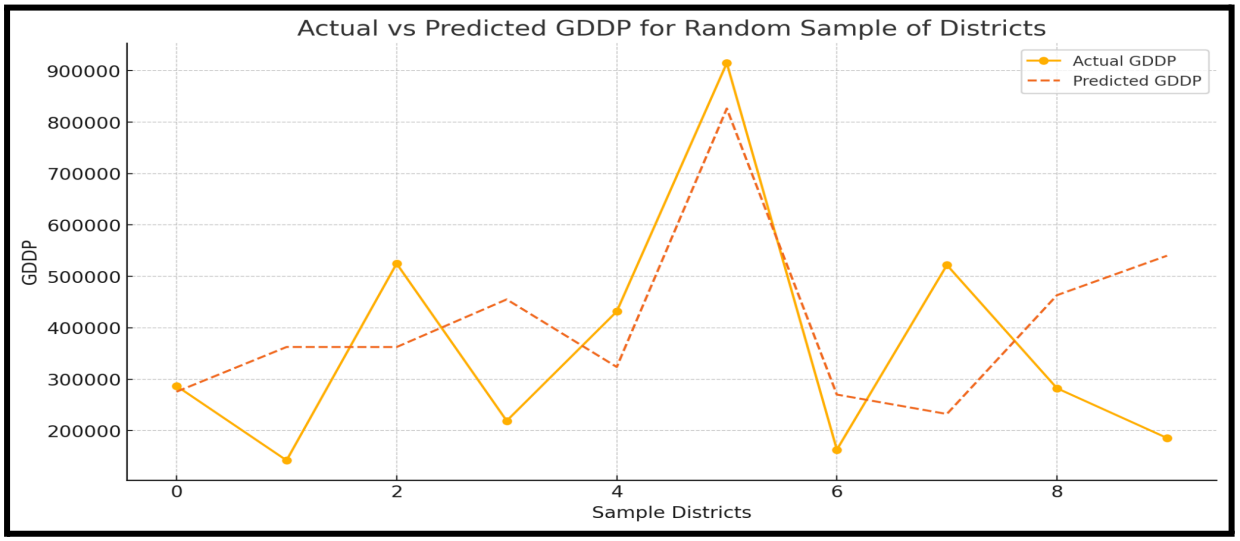
- **Temporal Scope**: We selected annual night-time light data from 2013 onward, aligning with available VIIRS data and ensuring consistency with GDP reporting cycles.
- **Spatial Scope**: The analysis was restricted to India's district boundaries, incorporating data for each district.
- **Data Preprocessing**: Cloud-covered data was masked, and extreme radiance values (e.g., from fires or light pollution) were filtered to maintain data accuracy.

5.1.3 Data Preparation and Use in Models

- **Feature Engineering**: After extracting district-level radiance values, we created features representing both raw NTL intensity and socio-economic indicators such as population density.
- **Principal Component Analysis (PCA)**: PCA was applied to reduce dimensionality, prioritising significant features for GDP prediction and improving model efficiency.
- **Model Training**: The dataset was used to train several machine learning models, including Gradient Boosting, Support Vector Machine (SVM), and Long Short-Term Memory (LSTM) networks, to predict district GDP based on NTL and socio-economic features. Each model was calibrated using historical GDP data, aiming to minimise prediction error and enhance the robustness of economic estimates at a sub-national level.


Actual vs Predicted GDDP for Pune
1e7
Predicted GDDP
Actual GDDP
GDDP
Year
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1995
2000
2005
2010
2015
2020


Actual vs Predicted GDDP for Random Sample of Districts (Ensemble Model)
1e6
Actual GDDP
Predicted GDDP
GDDP
Sample Districts
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5


Actual vs Predicted GDDP for Random Sample of Districts
Actual GDDP
Predicted GDDP
GDDP
Sample Districts
200000
300000
400000
500000
600000
700000
800000
900000
0
2
4
6
8

Results

In this study, we effectively gathered GDP and night-time light (NTL) data through systematic web scraping from multiple reliable online sources. We carefully organised the extracted data points and applied essential preprocessing techniques, such as managing missing entries, standardising values, and aligning timelines to ensure consistency across datasets. These steps were critical in achieving high correlation coefficients, with values reaching around 0.85, which reflects a significant link between GDP and NTL data. With this organised and refined data, we developed machine-learning models aimed at improving GDP predictions by harnessing the insights provided by NTL metrics. The initial results show promise, highlighting that NTL data can strongly predict economic trends, paving the way for more accurate and dynamic forecasting approaches.

Chapter 6

Future Scope

Future Scope for GDDP Prediction Using Nighttime Light Data

1. **Incorporating Additional Data Sources**:

   - **Economic Indicators**: Including additional economic indicators such as employment rates, agricultural output, industrial data, and GDP at the state level can improve prediction accuracy by capturing more economic dimensions.

   - **Satellite Data Advances**: Newer satellite data with higher spatial and temporal resolution, such as data from VIIRS (Visible Infrared Imaging Radiometer Suite), can provide more precise NTL readings, especially for urban-rural distinction.

2. **Exploring Advanced Machine Learning Models**:

   - **Deep Learning Models**: Leveraging recurrent neural networks (RNNs) or transformers could improve predictions by capturing temporal dependencies over multiple years.

   - **Spatio-Temporal Models**: Spatial econometrics or geographically weighted regression models could account for spatial dependencies. This allows the model to learn from neighbouring districts and regions, especially in cases where districts share economic characteristics.

   - **Hybrid Models**: Combining models (e.g., ensemble learning) could enhance performance by leveraging the strengths of various approaches.

3. **Data Aggregation Techniques**:

   - **Clustering-Based Analysis**: Grouping districts with similar economic profiles through clustering techniques can help create tailored models for different district types (e.g., urban, rural, agricultural, industrial).

   - **Monthly/Seasonal Analysis**: Future studies could also investigate the variations in NTL data, as certain economic activities (such as agriculture) vary across seasons. Monthly or quarterly aggregation of NTL and financial data could reveal more granular insights.

4. **Real-Time Economic Monitoring and Forecasting**:

   - **Real-Time NTL Data for Rapid Assessment**: Using real-time NTL data, this model could become a tool for rapid assessment of economic activities and help in early detection of economic shifts (e.g., sudden changes in industrial activity or responses to natural disasters).

   - **Forecasting Future GDDP**: Extending the model to forecast future GDDP values based on projected NTL data and economic growth trends could provide actionable insights for policy and investment.

5. **Policy Analysis and Decision-Making**:

   - **Economic Policy Impact Evaluation**: By analysing changes in GDDP over time in response to specific policies (e.g., subsidies, economic zones), this model could help policymakers evaluate the financial impact of their initiatives.

   - **Monitoring Sustainable Development Goals (SDGs)**: This model could support progress monitoring of SDGs related to economic growth, urbanisation, and infrastructure by providing data-driven insights into regional economic development.

6. **Building a User-Friendly Platform**:

   - **Interactive Dashboard**: Creating an interactive platform or dashboard to visualise predicted GDDP values across regions and time would make the model accessible for stakeholders, including policymakers, researchers, and businesses.

   - **APIs for Integration**: Developing APIs to integrate the model's predictions with other governmental and research applications could expand its reach and utility.

Conclusion

Expanding this project along these dimensions would allow for more accurate, versatile, and actionable insights into economic conditions across districts, enhancing the model's applicability to various users and helping drive data-informed decisions at regional and national levels.

Chapter 7

Conclusion

This project presents a novel framework for district-level GDP estimation leveraging Night-Time Light (NTL) data, demonstrating the effectiveness of satellite imagery as a proxy for economic activity. The model achieves high accuracy in GDP prediction at the district level through a detailed methodological approach that includes data preprocessing, feature engineering, and advanced machine learning techniques. The Gradient Boosting Regressor (GBR), in particular, proved the optimal model, offering the lowest error rates and capturing complex, non-linear relationships in the data.

The results underscore a strong correlation between NTL intensity and economic output, highlighting the value of this approach for policy-making, resource allocation, and financial planning. This model bridges data gaps and provides policymakers and researchers with an adaptable, scalable tool to assess regional economic trends and disparities. The insights drawn from this study pave the way for data-driven, localised economic interventions, which are essential in addressing the diverse economic landscape of India.

The potential for real-time data integration and extended temporal analysis could further enhance this framework. Future research might incorporate additional socio-economic variables and explore advanced deep-learning models to improve predictive accuracy. This study's approach contributes significantly to the growing body of knowledge on economic forecasting and provides a foundation for refined, reliable, and efficient economic assessments in developing regions.

Bibliography

https://www.sciencedirect.com/science/article/pii/S1569843224004400?pes=vor#f0020

https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=GDP+nighttime+light&sid=7304f28c480bfb66e2c95ee83184faea&sot=b&sdt=b&sl=34&s=TITLE-ABS-KEY%28GDP+nighttime+light%29&origin=searchbasic&editSaveSearch=&yearFrom=Before+1960&yearTo=Present&sessionSearchId=7304f28c480bfb66e2c95ee83184faea&limit=10

https://www.scopus.com/record/display.uri?eid=2-s2.0-85201276605&origin=resultslist&sort=plf-f&src=s&sid=7304f28c480bfb66e2c95ee83184faea&sot=b&sdt=b&s=TITLE-ABS-KEY%28GDP+nighttime+light%29&sl=34&sessionSearchId=7304f28c480bfb66e2c95ee83184faea&relpos=5

https://www.pnas.org/doi/pdf/10.1073/pnas.0509842103

https://www.pnas.org/doi/full/10.1073/pnas.1919913118

https://xkdr.org/paper/but-clouds-got-in-my-way-bias-and-bias-correction-of-viirs-nighttime-lights-data-in-the-presence-of-clouds

http://164.100.161.239/plans/stateplan/ssphd.php?state=ssphdbody.htm

Sanket Jagtap - 21AG10032

ORIGINALITY REPORT

9%	8%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian Institute of Technology, Kharagpure Student Paper	2%
2	www.researchgate.net Internet Source	1%
3	d-nb.info Internet Source	<1%
4	joshuaebenezer.github.io Internet Source	<1%
5	Submitted to National Forensic Sciences University Student Paper	<1%
6	qspace.library.queensu.ca Internet Source	<1%
7	Submitted to University of Hong Kong Student Paper	<1%
8	www.mdpi.com Internet Source	<1%
9	Submitted to Liberty University	

	Student Paper	<1%
10	Submitted to University of Nottingham Student Paper	<1%
11	www.cfrpc.org Internet Source	<1%
12	Submitted to University of North Texas Student Paper	<1%
13	backend.orbit.dtu.dk Internet Source	<1%
14	discovery.researcher.life Internet Source	<1%
15	www.ijraset.com Internet Source	<1%
16	dokumen.pub Internet Source	<1%
17	kar.kent.ac.uk Internet Source	<1%
18	res.mdpi.com Internet Source	<1%
19	siddhanthaldar.github.io Internet Source	<1%
20	www.amazon.science Internet Source	<1%

21	www.geeksforgeeks.org Internet Source	<1%
22	www.science.gov Internet Source	<1%
23	fenix.iseg.ulisboa.pt Internet Source	<1%
24	staging-medinform.jmir.org Internet Source	<1%
25	www.qeh.ox.ac.uk Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		