Integrating Open Data for GDP Prediction Leveraging ALITE and Machine Learning

Team Information

- Madhu Mausam Thapa, u1463636@utah.edu, u1463636
- Sanjay Luitel, u1559592@utah.edu, u1559592
- Sushil Rijal, u1323213@utah.edu, u1323213

Abstract

This project aims to develop a machine learning pipeline for Gross Domestic Product (GDP) prediction by integrating multiple open economic datasets leveraging ALITE, a tool for integrating Data Lake Tables. The focus is on aligning and integrating Capital Stock data from the International Monetary Fund (IMF) with other economic datasets from World Bank and Penn World Table to build a more comprehensive forecasting model. By applying various machine learning models, we seek to evaluate how well capital stock and other factor of productions predicts GDP. This project highlights the advantages of automated data integration for macroeconomic forecasting while assessing different ML approaches in defining a production function.

Introduction

Motivation

Economic forecasting is critical for policymakers, businesses, and financial institutions. GDP and its growth serve as fundamental indicators of economic health, but accurately predicting is complex, which requires integrating multiple data sources. Traditional GDP estimation models rely on limited datasets and static assumptions, whereas machine learning models identify underlying complex relationships between GDP and factor of productions. Understanding and analyzing the relationship between economic indicators and GDP is crucial for assessing economic conditions and supporting data-driven planning. By identifying this relationship with confidence, we can estimate the required capital stock, which guides prudent investment decisions and contributes to informed decision-making.

Why is it interesting?

Bridges data management and ML:

- Demonstrates the power of integrating heterogeneous datasets for ML applications.
- Uses ALITE for automated data integration, reducing preprocessing overhead.

Real-world impact:

- Understanding relationship between Factor of productions and GDP as well as enhancing prediction accuracy benefits economic planning and decision-making.

Why is it needed?

- Manual data integration is prone to errors and scalability issues.
- Existing GDP models often suffer from the approximate relationship defined between the economic indicators and GDP.
- A unified approach to integrating open datasets and leveraging ML for forecasting can improve predictive accuracy.

Hypothesis

By integrating Capital Stock data with additional factors of productions, we hypothesize that a machine learning model can predict GDP and define its interrelationship more accurately than simple models. Furthermore, data integration using ALITE will improve data completeness and lead to better model performance.

Methodology

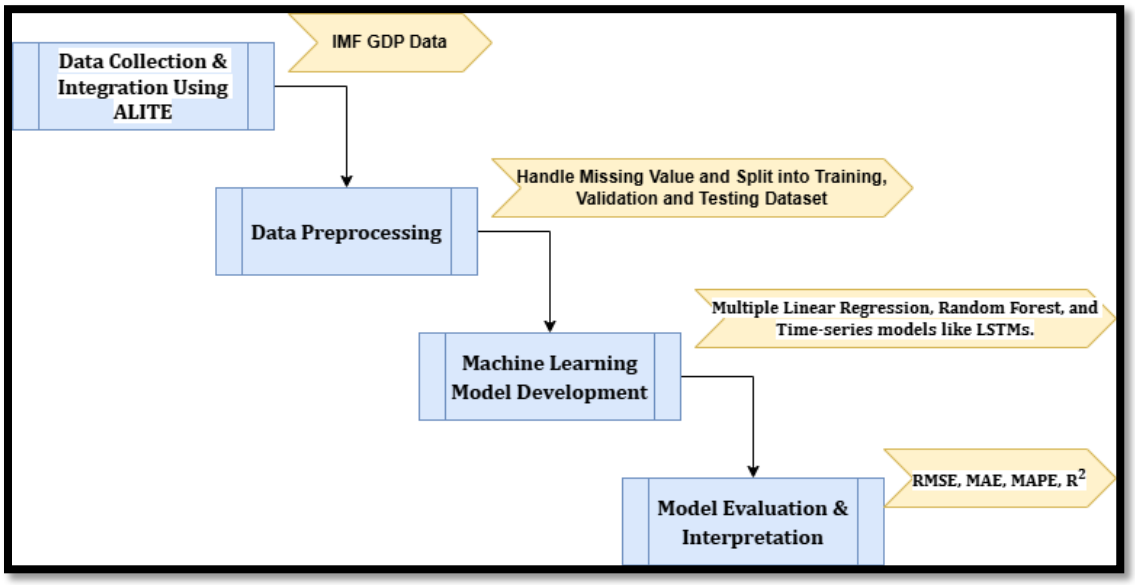

Fig: Implementation Flowchart

1. Data Collection & Integration Using ALITE

- Extract Capital stock data of available country from IMF.
- Integrate additional datasets using ALITE such as:
  - ✓ Historical IMF GDP data on Purchasing Power Parity (PPP) for facilitating economic comparison
  - ✓ World Bank economic indicators (Working population, unemployment)
  - ✓ Productivity data from Penn World table

2. Data Preprocessing

- Handle missing values using imputation techniques.
- Perform feature engineering.
- Split data into training, validation, and test sets.

3. Machine Learning Model Development

- Train multiple ML models including:
    - Multiple Linear Regression
    - Random Forest
    - Time-series models like LSTMs.
- Apply K cross-validation.
- Perform hyperparameter tuning.

4. **Model Evaluation & Interpretation**

- Assess Models using performance metrics:
    - Root Mean Squared Error (RMSE)
    - Mean Absolute Error (MAE)
    - R^2 (Coefficient of Determination)

-Conduct Feature importance analysis to assess the contribution of Capital stocks and other factors on GDP prediction.

Evaluation

1. Data Completeness: Evaluating the Added Value of Data Integration using ALITE
- Measure the percentage of missing values in manually-integrated data vs ALITE-integrated data
- Compute Pearson correlation for manually-integrated data vs ALITE –integrated data

2. Scalability: Assessing ALITE's Automation in Multi-Source Data Preparation
- Measure integration time manually vs. using ALITE.
- Compare manual operations required in traditional vs. alite integration.

Data Sources

1. IMF Capital Stock Dataset (primary feature)
2. IMF GDP Data (target variable)
3. World Bank Economic Indicators (e.g., Working Population, Unemployment)
4. Productivity and other macro-economic indicators (Penn World Table)

Task List & Responsibilities

Task	Madhu	Sanjay	Sushil
Literature review on ALITE and GDP prediction	✅		✅
Download economic datasets		✅	✅
Data integration with ALITE	✅	✅	
Data preprocessing & feature engineering	✅	✅	✅
ML model development (Baseline + Advanced)	✅	✅	✅
Model evaluation & interpretation	✅	✅	✅
Report writing & presentation	✅	✅	✅

Timeline & Deliverables

Week 1: Define scope & download datasets - Initial project plan
Week 2: Perform ALITE integration - Integrated dataset
Week 3 & 4: Preprocess data, create ML pipeline - Preprocessed dataset, Baseline model
Week 5 & 6: Train ML models, tune parameters - Trained models, results
Week 7: Analyze results, prepare report - Final results, Interpretability analysis
Week 8: Presentation & submission - Report, Code, Presentation slides