IST 652

Final Project Report

COMPREHENSIVE MACROECONOMIC AND MARKET DATA ANALYSIS FOR S&P 500 PREDICTION

GROUP 5

Prashant Gautam

Suyog Surana

Sourabh Pandya

Md Jami Bin Mosharf Navid

Natasha Lobo

Table of Contents

Overview	03
Introduction	05
Data Explorations	06
Data Wrangling	08
Program Overview	09
Model development	10
Insights	16
Ways to improve	18
Conclusion	19
Team Contribution	21

Overview

The S&P 500, one of the most recognized stock market indices, tracks the performance of 500 of the largest publicly traded companies in the United States. Often considered a barometer for the U.S. economy, the S&P 500 reflects trends in corporate performance, investor sentiment, and macroeconomic conditions. Accurately predicting its movements has immense value for investors, policymakers, and businesses as it enables data-driven decision-making.

In recent years, financial markets have become increasingly interconnected with macroeconomic factors and alternative assets. Assets like Bitcoin, often referred to as "digital gold," have emerged as key indicators of market sentiment. Simultaneously, traditional hedging instruments like physical gold and Crude Oil continue to influence investor behavior and economic activity. The Federal Funds Rate, a critical monetary policy tool, impacts liquidity, borrowing costs, and

ultimately, market stability. Understanding how these variables interact and contribute to market trends forms the foundation of this project.

Financial Context

1. **S&P 500**: This index serves as a "team score" for the stock market, rising and falling based on the performance of its constituent companies. Factors such as corporate earnings, economic growth, interest rates, and geopolitical events significantly influence its movements. For investors, the S&P 500 provides a benchmark for portfolio performance and a measure of market sentiment.
2. **Bitcoin**: As a decentralized digital currency, Bitcoin operates independently of central banks and governments. Its value often correlates with investor sentiment, particularly during periods of inflation or economic uncertainty. Bitcoin's volatility and emerging role as a "store of value" make it a key variable in understanding alternative asset behavior.
3. **Gold**: Historically regarded as a safe haven asset, gold retains value during economic instability and inflationary periods. Its price trends provide insights into market risk aversion and broader economic stability.
4. **Crude Oil**: This commodity underpins global industrial activity, with its price driven by supply-demand dynamics and geopolitical factors. Oil prices serve

as a leading indicator of economic activity, particularly in energy-dependent industries.

5. **Federal Funds Rate**: Managed by the Federal Reserve, this rate influences liquidity, borrowing costs, and overall economic growth. Low rates stimulate economic activity, while high rates control inflation. Movements in this rate ripple through financial markets, affecting equity and bond performance.

Introduction

The primary aim of this project is to model complex relationships between macroeconomic and market variables to gain actionable insights. By leveraging statistical analysis and machine learning techniques, the project addresses the following key questions:

1. What patterns and relationships exist among macroeconomic and market data, and how do they influence the S&P 500?
2. Can predictive models accurately forecast S&P 500 trends, and what are the most influential variables?
3. How well do these models perform in capturing both long-term trends and short-term market fluctuations?

4. What insights can be derived to support investment strategies and economic policy?

The project adopts a multi-disciplinary approach, integrating financial, statistical, and machine learning methodologies to analyze structured time-series data and derive meaningful predictions.

Data Explorations

Out of several economic and market factors, a few primary indicators were used for the sake of complexity. As mentioned above, macroeconomic and market factors such as Federal Interest Rates, Gold Prices, Bitcoin Prices and Crude Oil Prices were utilized in the feature engineering process.

Data for S&P 500 was sourced using the Yahoo Finance library.

In this project, we are gathering and organizing data from multiple financial and economic sources to analyze the relationships between stock markets, interest rates, cryptocurrency, and commodity prices. The data includes several key components.

First, we collect historical data for the S&P 500 index, which tracks the performance of 500 major publicly traded companies in the U.S. and serves as a benchmark for the overall health of the U.S. stock market. This data is sourced from Yahoo Finance, covering the period from January 1, 2010, to December 1, 2024, to analyze long-term trends and their relationships with other variables.

Next, we retrieve the Federal Funds Rate, which represents the interest rate set by the Federal Reserve and significantly impacts borrowing, lending, and overall economic activity. This data is obtained using the FRED API (Federal Reserve Economic Data) and is provided in JSON format, containing historical interest rate values. This allows us to examine how changes in monetary policy affect financial markets, commodities, and cryptocurrencies.

Additionally, we collect historical Bitcoin price data from Yahoo Finance to analyze this emerging asset class. Bitcoin represents a new and volatile investment type, and its data, spanning the same period, will help us explore its interaction with traditional financial markets and economic indicators.

We also incorporate commodity price data for gold and crude oil, sourced from Yahoo Finance. Gold is considered a "safe-haven" asset during times of economic uncertainty, while crude oil serves as a key driver of the global economy and reflects energy market trends. By including these commodities, we aim to

understand their price dynamics in comparison to other variables, such as stocks, Bitcoin, and interest rates.

Data Wrangling

Data Consolidation: The S&P 500, Fed Funds Rate, Bitcoin, Gold, and Crude Oil datasets are then merged into a single DataFrame. The data is aligned by date, and rows with missing values are dropped to ensure a complete dataset for analysis.

Preprocessing Data: Data Preprocessing refine the data, making it more robust for predictive modeling and helping to identify underlying trends in market behavior

1. **Creating Lagged Variables**: We introduce lagged variables (shifted by one day) for key economic indicators such as the Fed Funds Rate, Bitcoin, Gold, and Crude Oil prices. This accounts for the delayed impact these factors may have on market movements.
2. **Handling Missing Data**: After creating lagged variables, we remove rows with NaN values that arise due to the shifting operation, ensuring the dataset is clean and complete.

3. **Rolling Window Calculations**: We calculate the 5-day rolling mean and standard deviation for each of the economic indicators (Fed Funds Rate, Bitcoin, Gold, and Crude Oil). This helps smooth out short-term volatility and highlights longer-term trends and variability.
4. **Dropping NaN Values**: After applying the rolling mean and standard deviation, we again drop rows with NaN values caused by the rolling calculations, ensuring the dataset is ready for analysis.

Program Overview

1. Data Collection: Gathering historical data from Yahoo Finance (S&P 500, Bitcoin, Gold, Crude Oil) and the FRED API (Federal Funds Rate).

2. Data Preprocessing: Aligning datasets by date, creating lagged variables, and engineering rolling features for enhanced prediction.

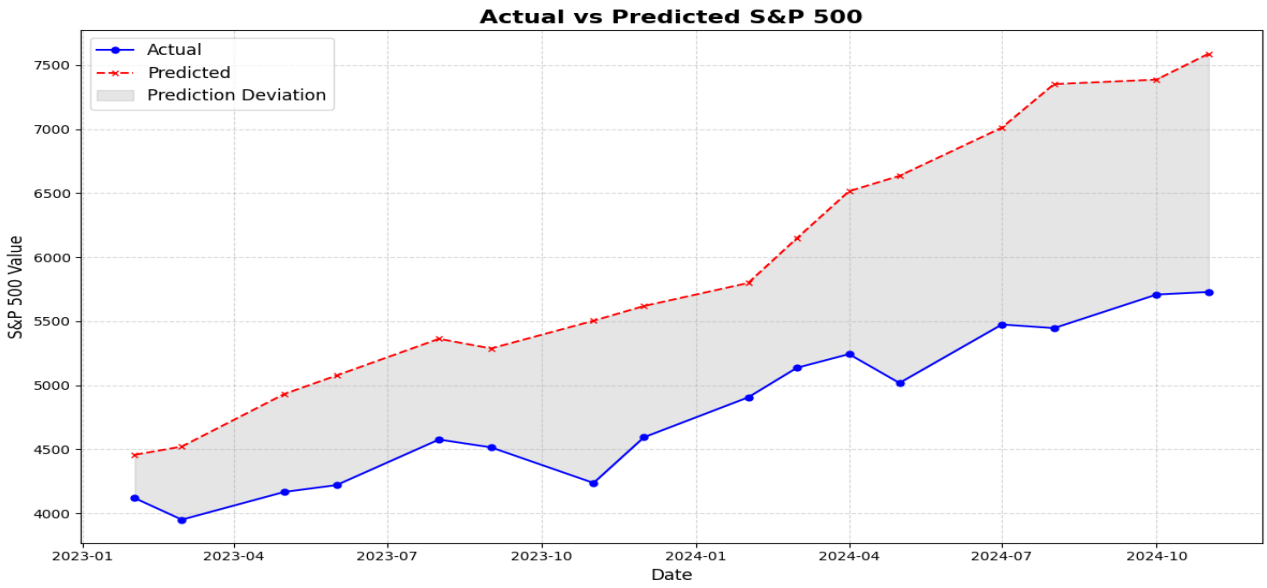
3. Exploratory Data Analysis (EDA): Using time-series trends and correlation heatmaps to understand relationships among variables.

4. Modeling: Implementing baseline (Linear Regression) and advanced models (TensorFlow) to capture non-linear relationships.

5. Evaluation: Comparing model performance using metrics like Mean Squared Error (MSE) and R^2 Score to assess predictive power.

Model development

Linear Regression Model



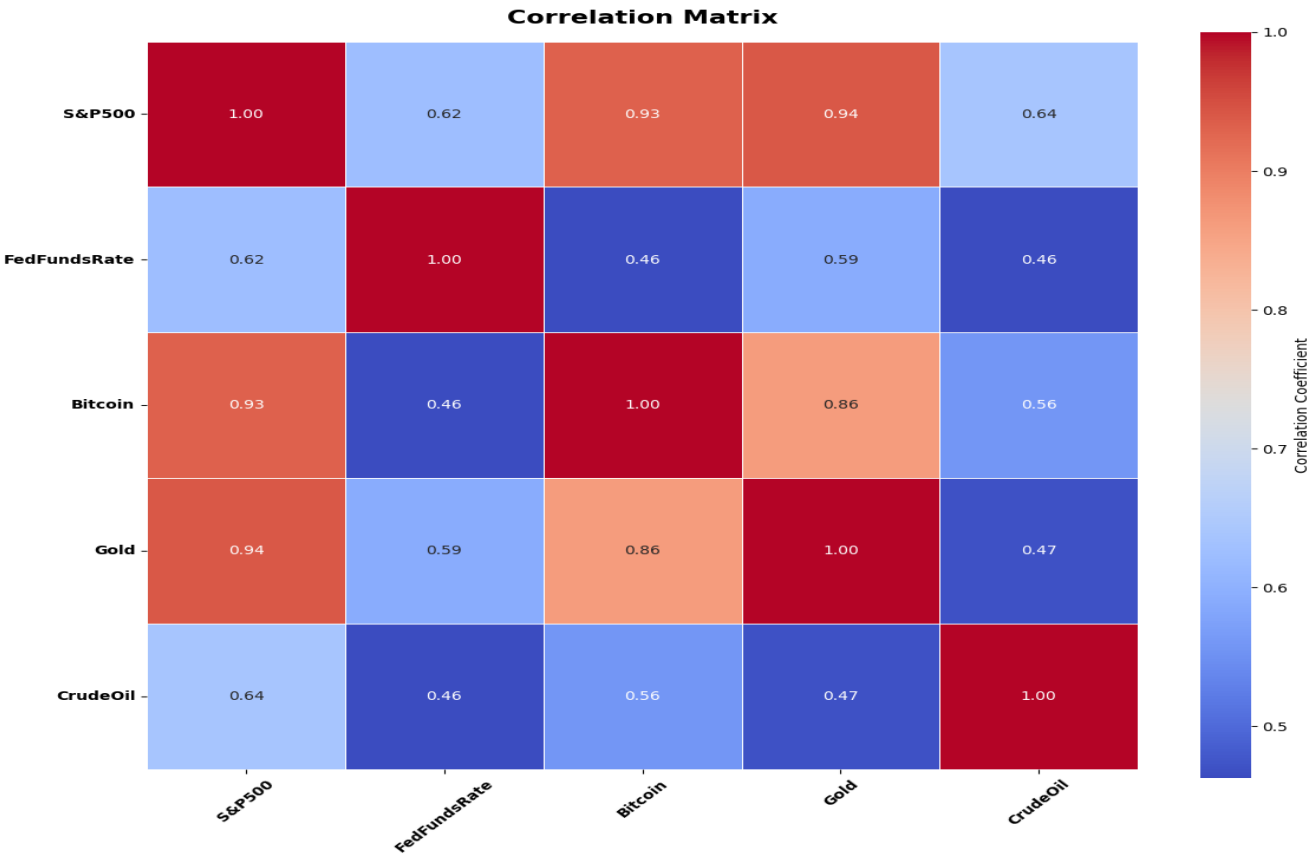
Why Linear Regression? Linear Regression was chosen as the baseline model due to its simplicity and interpretability. It provides a straightforward way to evaluate the relationships between the dependent variable (S&P 500) and the independent variables (e.g., Bitcoin, Gold, Crude Oil, and Fed Funds Rate). The model is widely used in finance to establish initial benchmarks and assess the significance of predictors.

What the Model Shows: The Linear Regression model revealed significant deviations between the actual and predicted values. Key findings include:

1. **Mean Squared Error (MSE)**: 1,496,494.83. This high value indicates substantial differences between the model's predictions and actual S&P 500 values, suggesting the model struggles to capture the complexity of the data.
2. **R^2 Score**: -3.487. The negative R^2 suggests that the model performed worse than a simple mean-based prediction, failing to explain any variance in the target variable.

Analysis: The model likely suffers from underfitting due to its inability to capture non-linear relationships and complex interactions between variables. While it provides a starting point, the poor performance underscores the need for more advanced methods to account for non-linearities and dynamic dependencies.

Correlation Heatmap



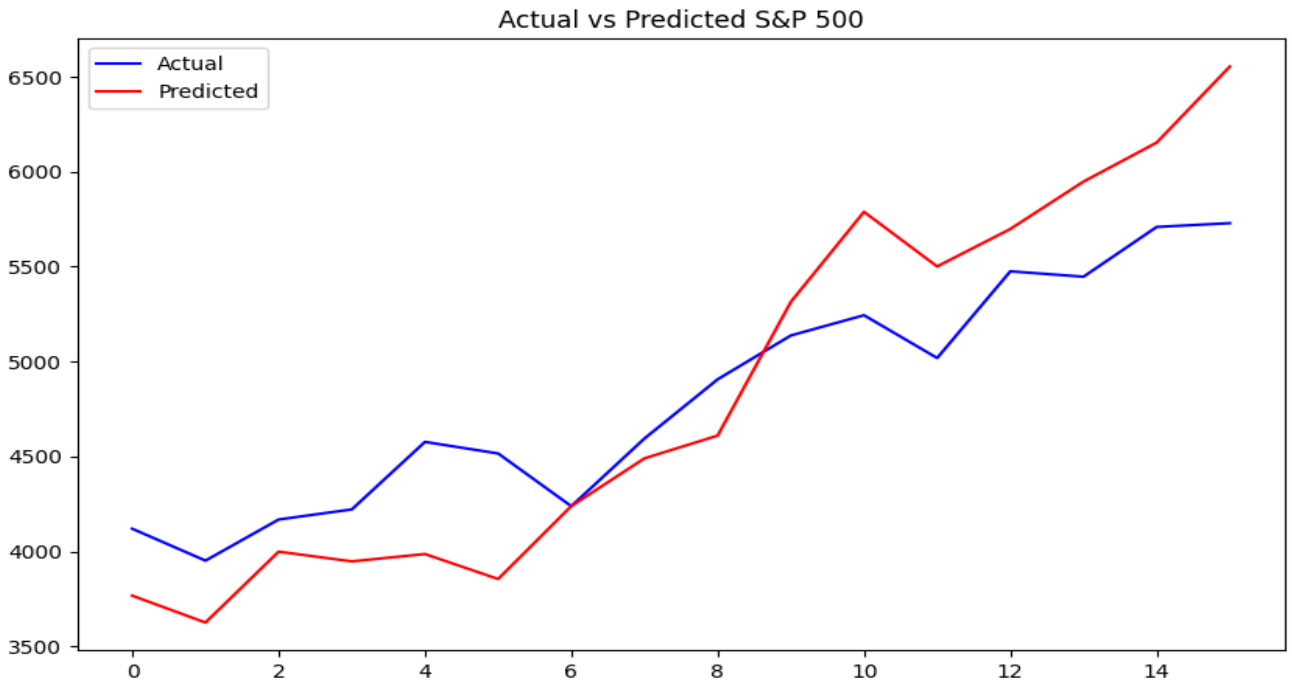
Purpose and Insights: The correlation heatmap is a critical tool for identifying relationships between variables. It provides a visual representation of how different features are associated, which helps in feature selection and engineering.

Key Findings:

1. **Positive Correlations**: S&P 500, Bitcoin(0.93), and Gold(0.94) showed strong positive correlations, indicating that they tend to move together. This suggests that Bitcoin and Gold prices could be useful predictors of the S&P 500.
2. **Crude Oil** (0.64)demonstrated a moderate positive correlation with the S&P 500, reflecting its role as an economic driver.
3. **Fed Funds Rate**: The Fed Funds Rate(0.62) exhibited moderate positive correlations with other assets but requires careful interpretation, as changes in interest rates can have indirect effects on market variables.

Correlation does not imply causation. While relationships are observed, deeper statistical and causal analyses are necessary to confirm predictive utility and avoid spurious associations.

TensorFlow Model



Why TensorFlow? TensorFlow was selected to implement a neural network model capable of capturing non-linear relationships and interactions between features. Unlike Linear Regression, neural networks can model complex dynamics, making them suitable for financial data, where trends and dependencies are rarely linear.

What the Model Shows:

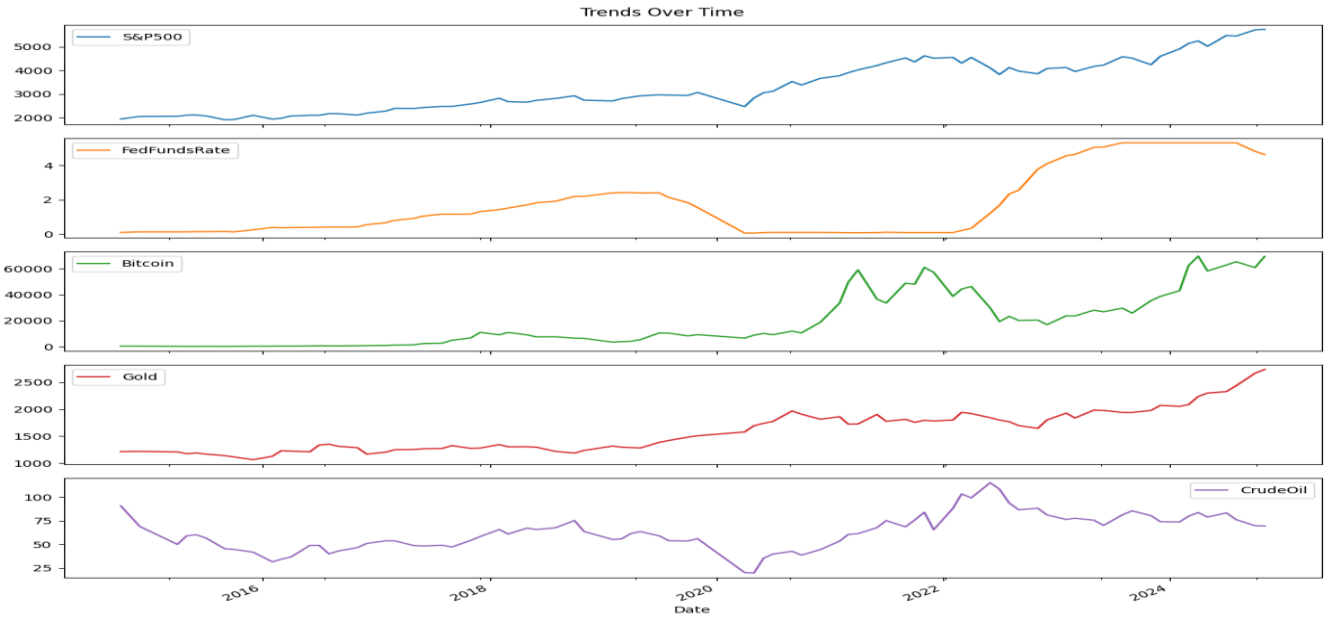
1. **Mean Squared Error (MSE)**: 84283.97. This lower error compared to Linear Regression indicates that the neural network better approximates the true relationships in the data.
2. **R^2 Score**: 0.7473 (~74.73%). While this is not a perfect score, it demonstrates that the model explains a significant portion of the variance in S&P 500 movements.

Analysis: The TensorFlow model outperformed Linear Regression by capturing non-linear patterns and dependencies between variables. However, it showed significant deviations in certain periods, such as overestimations and underestimations. These deviations highlight areas for improvement, including hyperparameter tuning, feature engineering, and potentially incorporating additional macroeconomic indicators.

Insights

The combination of Linear Regression and TensorFlow provides complementary insights. While Linear Regression establishes a baseline and highlights deficiencies in capturing complexity, TensorFlow offers a more nuanced approach to understanding non-linear relationships. Together with the correlation heatmap, these models and analyses form a cohesive framework for exploring and predicting S&P 500 trends.

Trends over time:



Post 2009 **S&P 500** trended upwards at a slower rate. In 2020, there was a big dip due to the Global Pandemic. Post pandemic when the market began to stabilize, the economy began functioning like normal, S&P 500 rose in value.

For the **Federal Interest Rate**, post 2020, interest rates were at a low point. This was to counteract the effects of pandemic. Lower interest rate allows more loans from banks, more disposable income for consumers hence more money in the economy.

Crude oil trends have fluctuated over the years with gradual increase and major decreases in 2020. As new technology comes into play, more energy is required for the servers and tools required. Crude oil plays a part as raw energy to power such technological advancements. Hence the trend for crude oil has been increasing in the last few years as shown in the graph.

BitCoin began booming post 2020 when it became a viral phenomenon through Social Media and News. It has an unpredictable trend due to being a decentralized currency. It generally follows people's sentiments and has risen in value over the past years as crypto and blockchain technology became more advanced.

Gold is considered as one of the most stable forms of Assets and is something that fiat and digital currencies are hedged against. It has a stable trend with increases

post 2020 due to uncertainty caused by global pandemic. It has been rising since at a stable rate.

Ways to improve

A few ways to improve our model and analysis would be to include more economic factors such as unemployment rate. GDP growth, inflation etc. Breaking down S&P 500 into sectors such as healthcare, technology etc industries can allow deeper insights into trends and drivers. We can also aim to include foreign global market indices such as FTSE 100 to assess how international markets influence S&P 500. Adding interaction variables can also allow us to identify combined effects of predictors. Deep learning models such as LSTM (Long-Short Term Memory) networks, boosting and Random Forest models are all beneficial for modelling and forecasting. We need to normalize the data to enhance our machine learning models. We should also conduct the VIF analysis to check for multicollinearity amongst the predicted variables.

We can also enhance predictions by incorporating sentiment analysis of financial news sources like Reuters, Financial news, Washington post, Blomberg etc. to capture public perception and emotional tone.

Conclusion

The ability to predict S&P 500 movements has significant implications across various sectors. For institutional investors, such as hedge funds, mutual funds, and pension funds, it provides valuable insights into market trends, helping them optimize portfolio allocations and manage risk more effectively. By anticipating market shifts, institutional investors can adjust their strategies in advance, ensuring better returns and reducing exposure to potential losses. For policymakers, forecasting S&P 500 movements serves as an essential tool for gauging the overall economic health. By understanding market signals, they can make informed decisions regarding fiscal and monetary policies, proactively responding to potential economic downturns or periods of growth. Individual investors also stand to benefit greatly, as accurate predictions enable them to time their investments more effectively, maximizing returns while minimizing risks. This allows for smarter asset allocation, improving investment outcomes.

However, creating a reliable predictive model for index forecasting requires considering multiple complex factors. The model must incorporate a diverse range of data inputs, including economic indicators (such as GDP growth, inflation, and

unemployment rates), financial market data, and sentiment analysis from news articles or social media. A larger and more varied data pool enhances the model's ability to recognize patterns and correlations. Feature engineering is also crucial for extracting the most relevant information from the data, improving the model's predictive accuracy. This involves selecting and transforming the right variables—such as sector-specific performance, global market indices, or interest rates—into meaningful features that drive better forecasting results. For optimal performance, the model should be robust, flexible, and adaptable to different market conditions, requiring continuous refinement and tuning.

Team Contribution

Prashant Gautam – Project Manager & Data Collection

Oversaw project progress and managed data collection from Yahoo Finance and FRED API.

Suyog Surana – Data Preprocessing & Feature Engineering

Handled data cleaning, created lag features, and implemented rolling averages.

Sourabh Pandya – Model Development & Evaluation

Developed and evaluated predictive models for S&P 500 using machine learning techniques.

Md Jami Bin Mosharf Navid – Data Analysis & Model Support

Conducted data analysis and assisted with model tuning and evaluation.

Natasha Lobo – Visualization & Reporting

Created visualizations and contributed to the final project report.