Traffic Forecasting for 2024: SARIMA, LSTM, and Random Forest Approaches

June 25, 2026

1 Traffic Forecast modeling

1.0.1 Activity-Based Model (ABM) and Its Significance

An Activity-Based Model (ABM) is an advanced simulation tool used in transportation planning. Unlike traditional four-step travel demand models, ABMs simulate individual travel behavior based on the daily activities of households and persons. The core principle of an ABM is to replicate real-world travel patterns by modeling the sequence of activities performed by individuals, considering factors such as:

- Socio-economic characteristics (e.g., age, employment, household size).
- Travel preferences and constraints (e.g., mode of transportation, trip purpose).
- Time and spatial interactions (e.g., work, school, recreational activities).

By capturing these detailed interactions, ABMs provide a more realistic representation of travel behavior and traffic dynamics.

1.0.2 The Role of the Atlanta Regional Commission (ARC)

The Atlanta Regional Commission (ARC) is a leader in transportation planning and modeling in the United States. The ARC's ABM is widely regarded as one of the best implementations of this model due to:

- **High-Quality Data Inputs:** The ARC integrates comprehensive data sources, including household travel surveys, land use data, and economic forecasts, ensuring a robust model foundation.
- **Granular Resolution:** The ABM for Atlanta captures fine-grained details at the Traffic Analysis Zone (TAZ) level, offering high spatial accuracy.
- **Dynamic Temporal Analysis:** It models travel behavior across different time periods, enabling detailed insights into peak and off-peak traffic patterns.
- **Scenario Testing:** The ARC uses its ABM to evaluate future scenarios, such as changes in land use, economic growth, and transportation investments, making it a valuable tool for policy and infrastructure planning.

These strengths make the ARC's ABM a benchmark for transportation modeling in the U.S., providing reliable and actionable insights for traffic forecasting.

1.1 The Importance of Traffic Volume Forecasting

Accurate traffic volume forecasting is critical for:

- **Infrastructure Planning:** Forecasting helps transportation planners design road networks and transit systems that can handle future demand, reducing congestion and improving efficiency.
- **Safety Improvements:** Predicting traffic patterns allows for proactive identification of high-risk areas, guiding investments in safety measures.
- **Environmental Impact Assessment:** Understanding future traffic volumes supports the evaluation of vehicle emissions and the development of sustainable transportation policies.
- **Economic Growth:** Effective traffic management fosters smoother transportation of goods and services, promoting regional economic development.

1.2 Forecasting Traffic Volume for 2024 Using ABM Outputs

This project aims to forecast traffic volumes for 2024 by leveraging the outputs of the ARC's ABM for 2020, 2022, and 2023. Key aspects of this approach include:

- **Historical Trends:** The ABM outputs for 2020, 2022, and 2023 provide insights into evolving traffic patterns, influenced by factors such as population growth, employment changes, and infrastructure developments.
- **Model Calibration:** By training predictive models (SARIMA, Random Forest, and LSTM) on the historical ABM data, we aim to capture both temporal and spatial dependencies in traffic patterns.
- **Predictive Modeling:** The trained models are used to forecast 2024 traffic volumes, enabling planners to anticipate future transportation needs and address potential challenges.

1.2.1 Why Use ABM Outputs for Forecasting?

The ARC's ABM outputs offer a detailed and reliable dataset for forecasting due to their:

- **Granular Resolution:** The inclusion of time-of-day and TAZ-level details ensures the forecast captures localized traffic dynamics.
- **Scenario Validity:** The model accounts for socio-economic and land use changes, making its outputs robust for future predictions.

This project demonstrates the power of integrating advanced modeling techniques (ABM) with predictive analytics to achieve accurate traffic forecasts, contributing to effective transportation planning and policy-making.

1.3 Data Resources

The data for this project is sourced from the Atlanta Regional Commission, covering the Atlanta Metropolitan area. To align with the project's focus, the dataset is aggregated to include only Fulton County. The dataset can be found in the following links:

- Download the model or other associated files: http://abmfiles.atlantaregional.com/.
- Download the model shapefile: https://opendata.atlantaregional.com/datasets/arc-model-transportation-analysis-zones-2020.
- Review the model documentation: https://atlregional.github.io/ARC_Model/index.html.

1.4 Data Preparation Using ArcGIS

To ensure the dataset was well-structured and spatially accurate, data preparation was conducted using ArcGIS. This process involved selecting, spatially joining, and exporting variables of interest for predictive modeling. The detailed steps are as follows:

2 Steps in Data Preparation

1. **Selection of Traffic Analysis Zones (TAZs):**
   - All TAZs relevant to the study area (Fulton County, GA) were selected by location in ArcMap.
   - This ensured that the spatial extent of the data matched the geographic region of interest.
2. **Spatial Join of Volume Variables:**
   - Traffic volume variables (e.g., *Sum_V_EA, Sum_V_AM, Sum_V_PM*) were spatially joined to the selected TAZs.
   - The **sum function** was applied during the spatial join to aggregate traffic volume data for each TAZ.
   - This step ensured that the total traffic volume for each TAZ was accurately calculated.
3. **Spatial Join of Speed Variables:**
   - Traffic speed variables (e.g., *Avg_SPD_EA, Avg_SPD_AM, Avg_SPD_PM*) were spatially joined to the selected TAZs.
   - The **average function** was applied during the spatial join to calculate the mean traffic speed for each TAZ.
   - This provided an aggregated measure of traffic speed, capturing the overall efficiency of traffic flow within each zone.
4. **Spatial Join of Socio-Economic Data:**
   - Node data containing socio-economic variables (e.g., *Sum_POP, Sum_EMP, Sum_HSHLD*) were spatially joined to the TAZ shapefiles.

- This step ensured that socio-economic information was accurately linked to each TAZ, enabling a comprehensive analysis of demographic and economic factors influencing traffic.

5. **Exporting Data:**
   - All spatially joined variables were exported into Excel files for further analysis and feature engineering.
   - This step facilitated integration with predictive modeling tools and ensured a seamless transition between spatial and statistical analyses.

2.0.1 Significance of the Data Preparation Process

- **Spatial Accuracy:** The use of spatial joins ensures that all traffic and socio-economic data are accurately aligned with their corresponding TAZs, providing high geographic fidelity.
- **Data Aggregation:** Aggregating volume data by sum and speed data by average ensures that the dataset captures key characteristics of each TAZ while reducing noise from finer-scale variations.
- **Comprehensive Dataset:** By combining traffic volume, speed, and socio-economic data, the final dataset supports robust feature engineering and predictive modeling.
- **Seamless Integration:** Exporting the data into Excel facilitates its use in machine learning workflows, enabling efficient analysis and forecasting.

2.0.2 Outcome

The data preparation process in ArcGIS provided a geographically accurate and well-structured dataset that integrates traffic and socio-economic variables at the TAZ level. This forms the foundation for the predictive modeling scenarios described in subsequent sections.

3 Data Description

The following variables have been used in this project, as represented in the correlation chart:

3.0.1 Traffic Volume Variables

- **Sum_V_EA (Early AM Traffic Volume):** Total traffic volume during the early morning period. Used to understand traffic flow patterns during pre-peak hours.
- **Sum_V_AM (AM Traffic Volume):** Total traffic volume during the morning peak period. Reflects commuter traffic patterns and congestion during peak hours.
- **Sum_V_MD (Mid-Day Traffic Volume):** Total traffic volume during mid-day hours. Useful for understanding traffic flow in non-peak daytime periods.
- **Sum_V_PM (PM Traffic Volume):** Total traffic volume during the evening peak period. Captures traffic patterns as people return home from work.

- **Sum_V_EV (Evening Traffic Volume):** Total traffic volume during the evening hours after the peak. Reflects late-night traffic patterns.
- **Sum_V_TOTD (Total Daily Traffic Volume):** Aggregate traffic volume across all time periods in a day. Serves as a comprehensive measure of daily traffic flow.

3.0.2 Traffic Speed Variables

- **Avg_SPD_EA (Early AM Average Speed):** Average traffic speed during early morning hours. Helps evaluate road conditions and flow efficiency in pre-peak hours.
- **Avg_SPD_AM (AM Average Speed):** Average traffic speed during the morning peak period. Indicates congestion levels during peak commuting hours.
- **Avg_SPD_MD (Mid-Day Average Speed):** Average traffic speed during mid-day hours. Reflects traffic flow efficiency during non-peak daytime periods.
- **Avg_SPD_PM (PM Average Speed):** Average traffic speed during the evening peak period. Highlights congestion levels during evening commutes.
- **Avg_SPD_EV (Evening Average Speed):** Average traffic speed during evening hours. Shows road usage efficiency in late-night periods.
- **Avg_SPEED (Overall Average Speed):** Aggregate average speed across all time periods in a day. A general indicator of traffic conditions and road performance.

3.0.3 Socio-Demographic Variables

- **Sum_POP (Population):** Total population in a given Traffic Analysis Zone (TAZ). Represents the potential source of traffic generation within the zone.
- **Pop_Acres (Population Density):** Population per acre within the TAZ. Indicates the concentration of residents in the area, impacting traffic patterns.
- **Sum_HSHLD (Households):** Total number of households within the TAZ. Acts as a proxy for the residential base contributing to local traffic.
- **Sum_EMP (Employment):** Total employment in the TAZ. Represents the job density in the area, influencing commuting patterns and traffic volume.

3.0.4 Key Insights

- The dataset includes both traffic and socio-demographic factors, allowing for a comprehensive analysis of how human activity influences traffic flow and congestion.
- Temporal granularity (e.g., early AM, AM, mid-day, PM, evening) enables detailed modeling of traffic patterns across different parts of the day.
- Socio-demographic factors, such as population, employment, and households, provide essential context for understanding the spatial distribution of traffic.

These variables form the basis for predictive modeling, enabling the development of models that capture both temporal and spatial traffic dynamics effectively.

4 Data Analysis and Insights

This section provides an in-depth analysis of traffic, population, and employment data to uncover patterns and relationships that influence traffic volume and speed. The goal is to leverage these insights for enhancing predictive modeling and traffic management strategies. The analysis is structured into the following key components:

- **Correlation Analysis**: Examines the relationships between traffic-related variables, such as speed, volume, population, and employment, using a correlation heatmap. This helps identify the most influential factors driving traffic patterns.
- **Temporal Analysis**: Explores trends in traffic speed, volume, and population over time, providing insights into how these variables have evolved from 2020 to 2023.
- **Periodic Traffic Trends**: Analyzes traffic dynamics across different times of the day (e.g., AM, PM, and evening periods) and across years, highlighting consistent patterns in traffic behavior.

Objective: This section aims to:

- Identify key variables and trends influencing traffic patterns.
- Provide actionable insights for improving traffic prediction models.
- Highlight areas of focus for urban planners and traffic management authorities.

By combining correlation, temporal, and periodic analyses, this section lays the groundwork for informed decision-making in traffic modeling and congestion mitigation. Each visualization is accompanied by detailed insights, offering a comprehensive view of the data.

4.1 Correlation Analysis

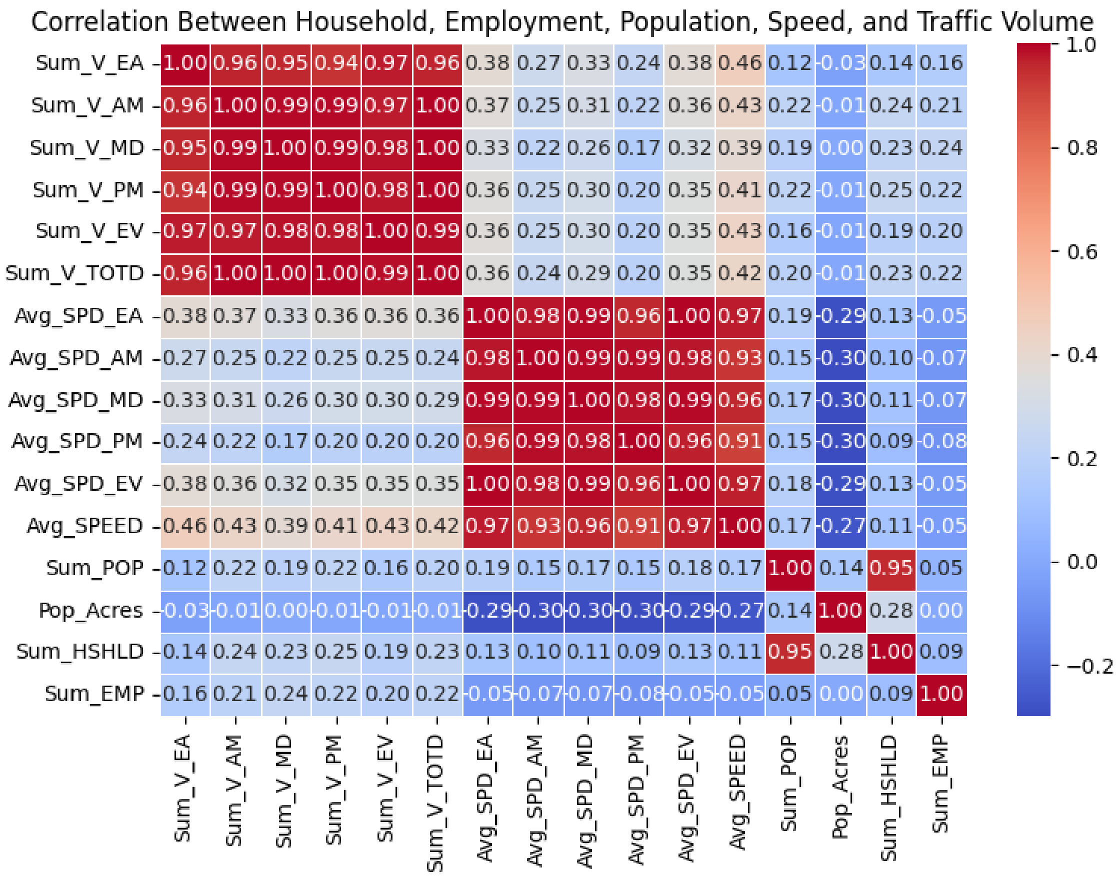

Figure 1: Correlation Between Household, Employment, Population, Speed, and Traffic Volume

The correlation heatmap highlights the relationships between traffic volume, speed, population, employment, and other variables. Key findings include:

- **Traffic Volume Correlations**: Strong positive correlations exist between total traffic volume (*Sum_V_TOTD*) and individual time-period volumes (e.g., *Sum_V_AM, Sum_V_PM*), indicating consistent traffic patterns across periods.

- **Speed Relationships**: Traffic speed shows weak negative correlations with population and employment, suggesting that denser areas tend to experience slower traffic speeds.

- **Population and Employment**: Population and employment densities are moderately correlated with traffic volume, underlining their role as key determinants in traffic forecasting.

This smaller correlation heatmap focuses on key variables, including total traffic volume (*Sum_V_TOTD*), average speed (*Avg_SPEED*), population (*Sum_POP*), population density (*Pop_Acres*), household counts (*Sum_HSHLD*), and employment (*Sum_EMP*). Key insights from this chart include:

- **Traffic Volume and Speed**: A moderate positive correlation ($r = 0.42$) exists between traffic volume and average speed, suggesting that higher speeds are often associated with greater traffic flow in certain conditions.

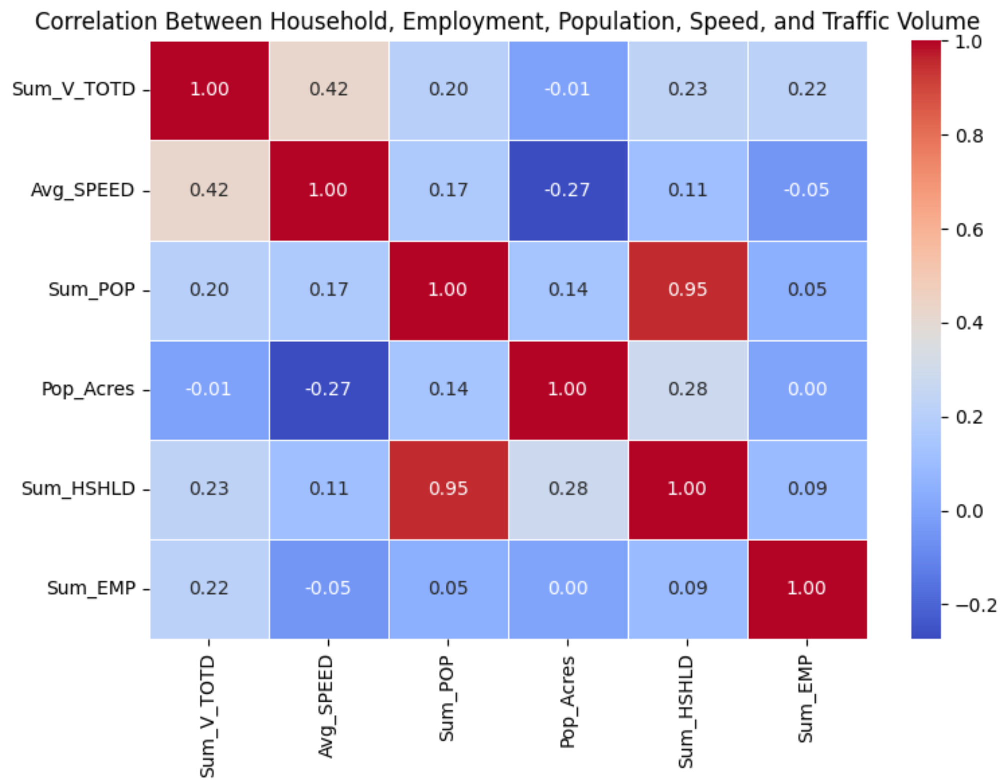

Figure 2: Correlation Between Household, Employment, Population, Speed, and Traffic Volume

- **Population and Household Counts**: There is a strong positive correlation between population (*Sum_POP*) and household counts (*Sum_HSHLD*), with $r = 0.95$. This indicates that areas with larger populations tend to have more households, which likely contribute to higher traffic volumes.
- **Employment and Traffic Volume**: Employment (*Sum_EMP*) is positively correlated with traffic volume ($r = 0.22$), reflecting the impact of job locations on daily commuting patterns.
- **Population Density and Speed**: Population density (*Pop_Acres*) exhibits a weak negative correlation with average speed ($r = -0.27$), suggesting that denser areas tend to experience slower traffic speeds, likely due to congestion or urban infrastructure constraints.
- **Traffic Volume and Households/Population**: Total traffic volume correlates positively with household counts ($r = 0.23$) and population ($r = 0.20$), emphasizing their role in traffic generation.

This chart provides a focused view of the relationships between essential traffic, demographic, and socio-economic variables, helping to identify the key drivers of traffic patterns. These insights can guide feature selection and model optimization in predictive modeling tasks.

4.2 Temporal Analysis of Key Variables

The temporal analysis reveals the following trends:

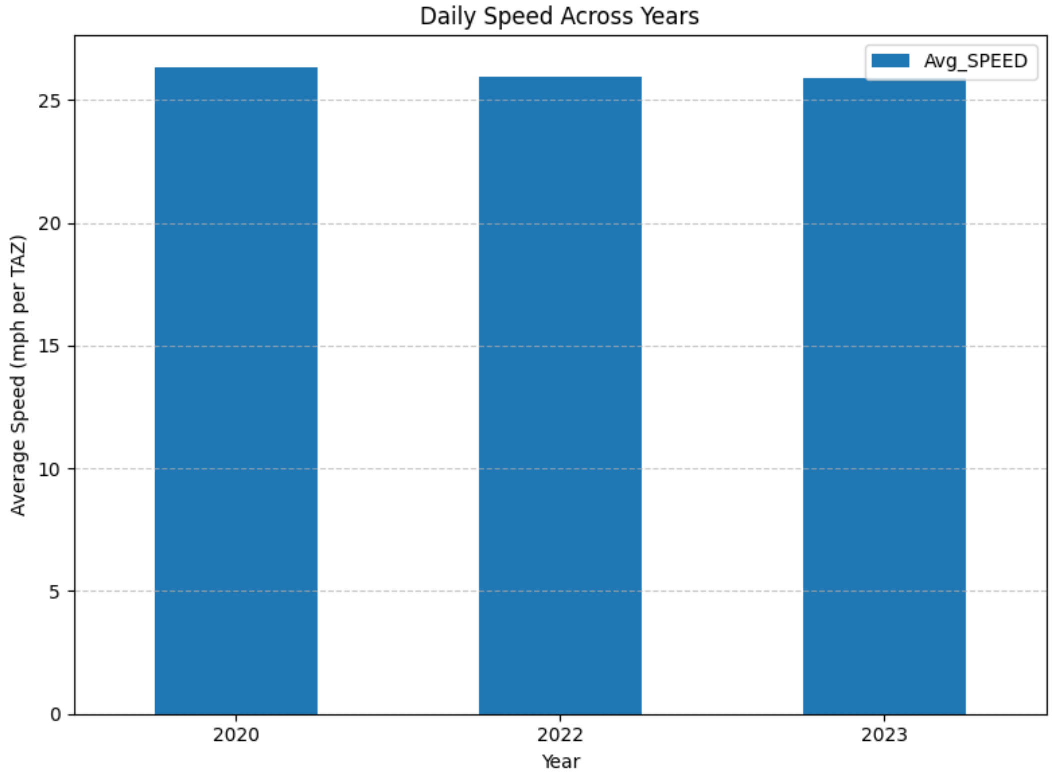

Figure 3: Daily Speed Across Years (2020, 2022, 2023)

- **Daily Traffic Speed**: Average daily speeds remain relatively stable over the years, with a slight decrease from 2020 to 2023, possibly reflecting increasing urbanization and traffic congestion.
- **Daily Traffic Volume**: Traffic volumes have seen minor increases over the years, reflecting gradual population growth and economic activity.
- **Population Trends**: Population density, measured per TAZ (Traffic Analysis Zone), shows consistent growth over the years, aligning with increasing traffic volumes.

4.3 Periodic Traffic Trends

Analysis of periodic traffic trends shows:

- **Traffic Speeds**: Speeds are highest during early mornings and evenings, with a significant dip during peak traffic hours (AM and PM). This pattern is consistent across all analyzed years (2020, 2022, 2023).
- **Traffic Volumes**: Volumes peak during the AM and PM periods, reflecting commuting patterns. These trends are consistent over time, emphasizing the importance of these periods in traffic management.

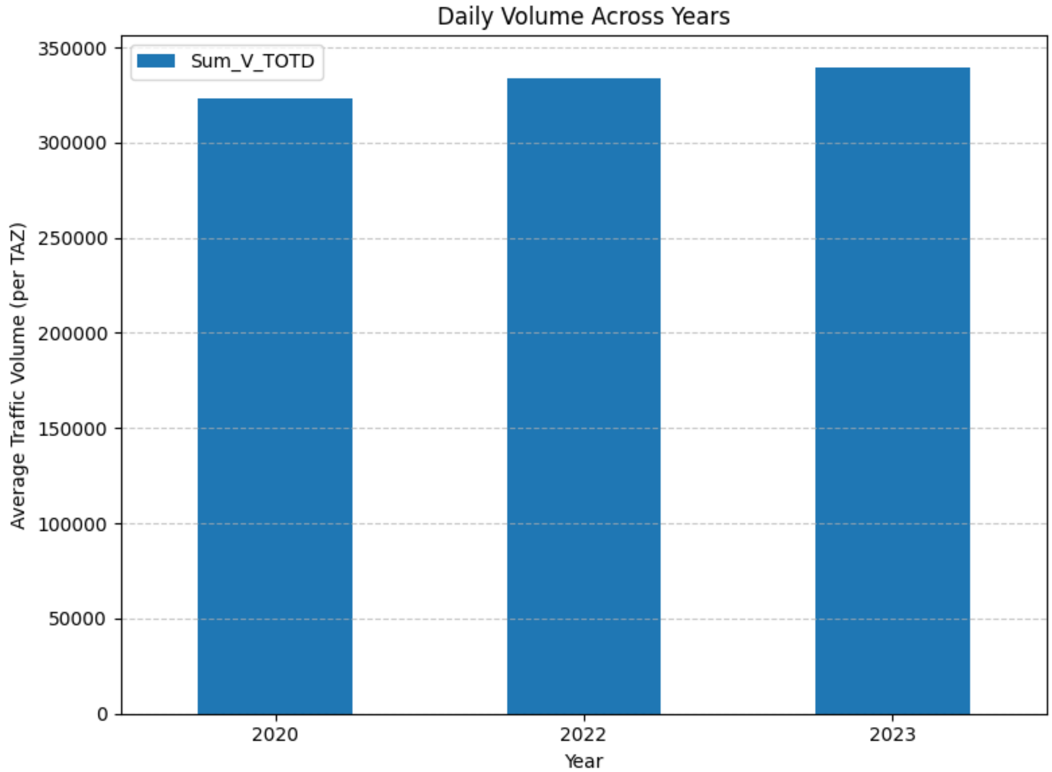

Figure 4: Daily Traffic Volume Across Years (2020, 2022, 2023)

4.4 Key Takeaways

- The data highlights the importance of time-period-specific variables, such as speeds and volumes, in predicting traffic dynamics.
- Increasing population and employment densities have a direct impact on traffic patterns, necessitating proactive traffic management strategies.
- Future predictive models should incorporate both temporal and spatial factors to capture the nuanced relationships between these variables.

5 Feature Engineering

Feature engineering is a critical step in preparing data for predictive modeling. For this project, the following steps were implemented to create a robust dataset using Activity-Based Model (ABM) outputs from the Atlanta Regional Commission (ARC) for the years 2020, 2022, and 2023. The key feature engineering steps and their importance are detailed below:

5.1 Steps in Feature Engineering

1. **Data Combination:** Data from three years (*2020, 2022, 2023*) were merged into a single dataset, ensuring consistent columns of interest such as traffic volume, speed, and socio-demographic variables (e.g., population, household density).

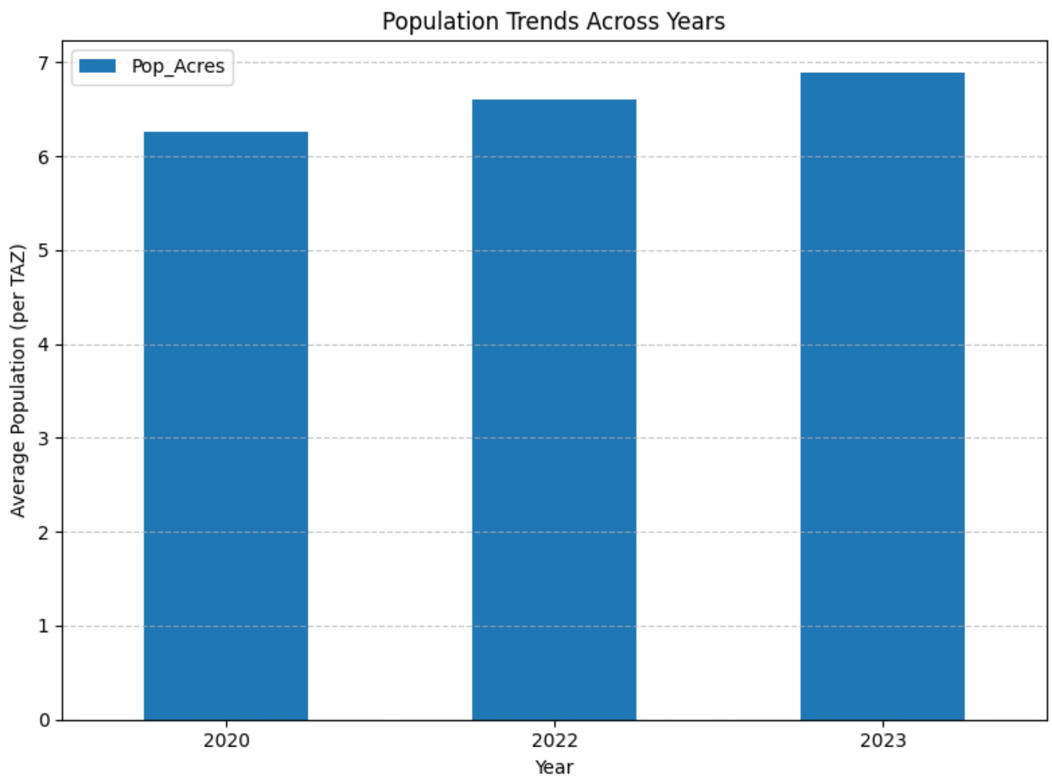

Figure 5: Population Trends Across Years (2020, 2022, 2023)

Figure 6: Traffic Speed Trends Across Periods and Years

2. **Handling Missing Data:** Missing values were filled with column means to ensure the dataset remained complete for modeling purposes, minimizing the risk of bias due to missing information.

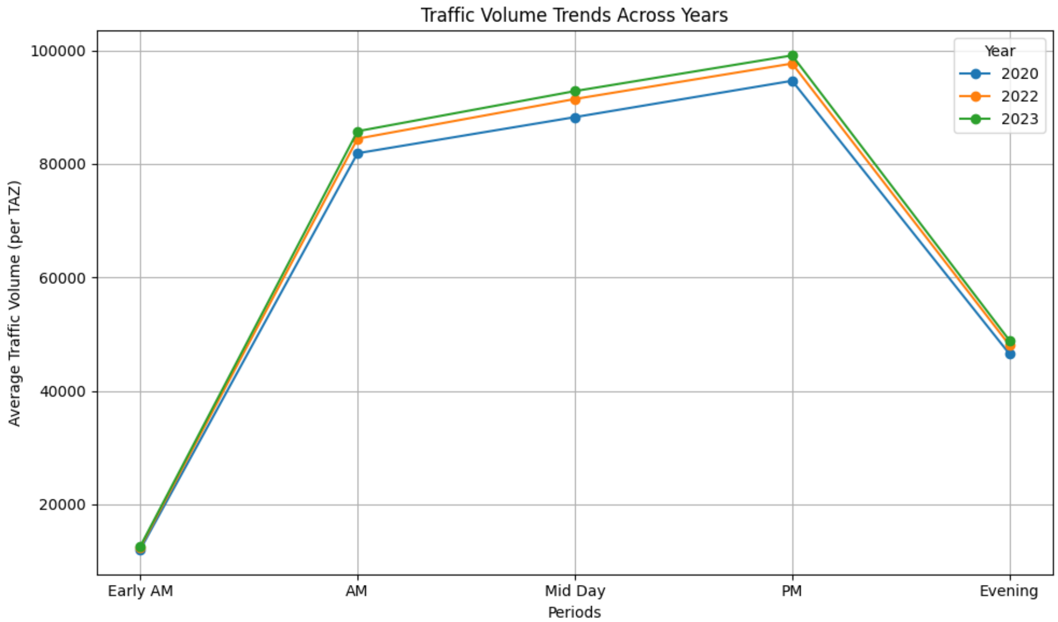

Figure 7: Traffic Volume Trends Across Periods and Years

3. **Creation of Lag Features:** Lagged features were generated for key variables, including traffic volume and speed. For example:
   - *Sum_V_EA_lag1* represents the traffic volume for the early morning of the previous time step.
   - Lag periods of 1, 2, and 3 time steps were applied to capture temporal dependencies and enhance model predictions.
4. **Adding Time-Based Variables:** New features such as *hour of the day, day of the week*, and *is_weekend* were created from datetime information. These variables provide contextual understanding of temporal patterns in traffic behavior.
5. **Normalization of Data:** All numerical variables were normalized using MinMaxScaler to ensure consistent scaling across features. This step prevents model bias towards variables with larger scales.

5.2 Importance of Feature Engineering in Scenarios

- **Scenario Relevance:** The feature engineering process enables the models to leverage lagged and time-based variables, which are critical in scenarios focusing on temporal patterns (e.g., *Scenario 2: Enhanced Feature Engineering* and *Scenario 3: Advanced Feature Engineering*).
- **Model Accuracy:** Lag features enhance the ability of models such as LSTM to capture sequential dependencies, while Random Forest benefits from the added explanatory variables for non-linear relationships.
- **Temporal and Spatial Insights:** Time-based features provide the models with additional context to understand variations across hours, days, and weekends, which is crucial for scenarios with dynamic traffic patterns.

- **Improved Forecasting:** By normalizing the data and incorporating comprehensive temporal features, the models are better equipped to provide accurate forecasts for 2024, leveraging trends and patterns from the historical ABM outputs.

5.3 Outcome

The final feature-engineered dataset integrates historical traffic and socio-demographic data with lagged and time-based variables, creating a robust foundation for predictive modeling. This ensures that all four scenarios leverage the full potential of the ABM outputs, enabling precise traffic volume forecasts.

6 Scenarios

This section explores the performance of three predictive models—SARIMA, Random Forest, and LSTM—across four distinct scenarios, each designed to assess the impact of feature engineering and dataset complexity on predictive accuracy. The scenarios are structured progressively, beginning with a basic feature set and advancing to a comprehensive set of lagged and explanatory variables. The goal is to understand the strengths and limitations of each model under varying levels of feature richness and temporal dependencies.

Model Characteristics:

- **SARIMA**: A traditional time-series forecasting model that relies on seasonality and trends. It is included to evaluate the performance of linear, statistical methods compared to more advanced machine learning techniques.
- **Random Forest**: A tree-based ensemble learning method known for its robustness and ability to capture non-linear relationships. It serves as a baseline for non-linear regression tasks.
- **LSTM (Long Short-Term Memory)**: A recurrent neural network designed for sequential data, capable of capturing both short-term and long-term temporal dependencies. Its performance is expected to improve as temporal features become more prominent.

Scenarios:

1. **Scenario 1: No Feature Lag Engineering**—Examines model performance using a basic feature set without lagged variables or temporal enhancements.
2. **Scenario 2: Enhanced Feature Engineering**—Introduces lagged variables to explore their impact on temporal modeling and predictive accuracy.
3. **Scenario 3: Advanced Feature Engineering**—Incorporates additional explanatory variables to evaluate the models' ability to capture complex relationships.
4. **Scenario 4: Comprehensive Feature Set**—Combines lagged variables and a wide range of features to maximize dataset richness and complexity.

Objective: By evaluating these scenarios, this study aims to:

- Identify the model best suited for traffic volume prediction under varying feature setups.
- Assess the role of feature engineering in enhancing predictive accuracy.

- Provide actionable insights for practical applications and future research in predictive modeling.

The results provide a comprehensive understanding of the trade-offs between traditional statistical methods and modern machine learning approaches in handling dynamic and high-dimensional data. Each scenario is discussed in detail, including performance metrics, insights for each model, and recommendations for future applications.

6.1 Scenario 1: No Feature Lag Engineering

6.1.1 Variables Used

In this scenario, the following variables were used:

- **Sum_EMP**: Total employment per TAZ
- **Sum_POP**: Total population per TAZ
- **Avg_SPEED**: Average speed per TAZ

In this scenario, no lagged features (historical values) were included. The models relied solely on the current values of these variables to predict the target variable (**Sum_V_TOTD**).

6.1.2 Performance Summary

Model	**RMSE**	**MAE**	**MAPE (%)**	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$
SARIMA	960,343.24	565,514.10	690.56	-0.0041	-0.0228
Random Forest	0.052	0.033	351.15	0.7787	0.7779
LSTM	0.108	0.080	187.76	0.1081	0.1045

Table 1: Model Performance Metrics

6.1.3 Analysis

SARIMA

- **Performance**: Poor, with an R^2 of -0.0041 and MAPE of 690.56%.
- **Limitation**: Struggles significantly in this context, likely due to an inability to leverage external predictors effectively without lagged features.
- **Use Case**: Not recommended for datasets without clear seasonality or trend-driven patterns.

Random Forest

- **Performance**: Strong, with an R^2 of 0.7787 and MAPE of 351.15%.
- **Strength**: Captures complex relationships between predictors and the target effectively.
- **Recommendation**: Well-suited for this dataset due to its robustness, though MAPE could be improved with additional feature engineering.

LSTM

- **Performance**: Moderate, with an R^2 of 0.1081 and MAPE of 187.76%.
- **Strength**: Captures sequential dependencies better than SARIMA but lags behind Random Forest in overall accuracy.
- **Limitation**: High error rates indicate potential overfitting or insufficient tuning.

6.1.4 Charts

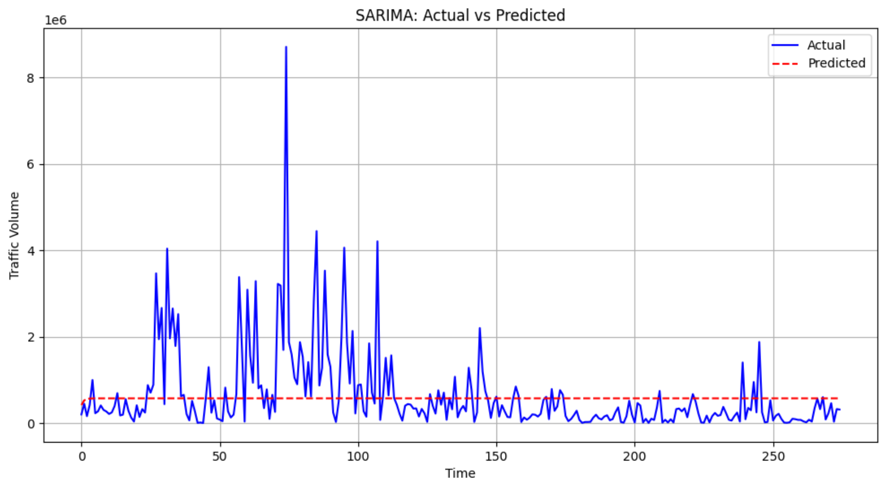

Figure 8: SARIMA Predictions vs Actual Values

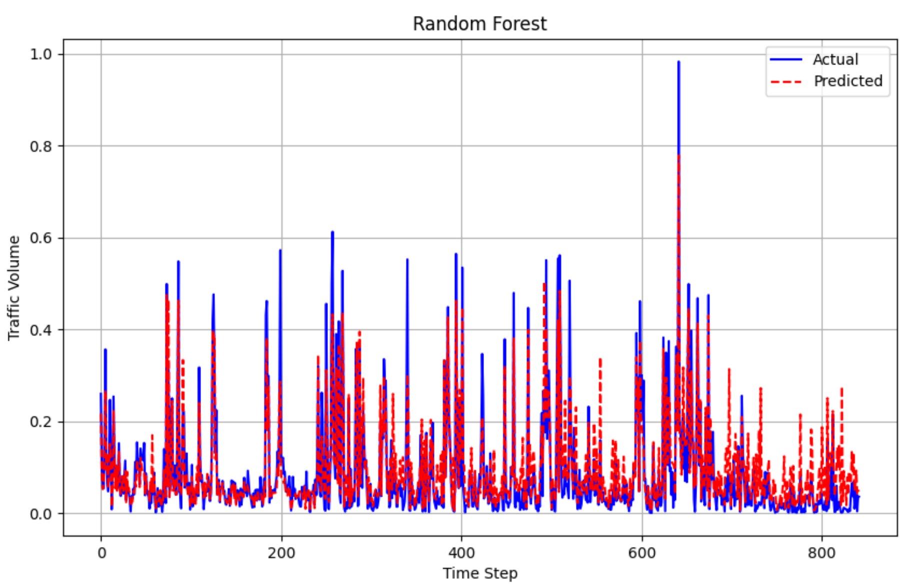

Figure 9: Random Forest Predictions vs Actual Values

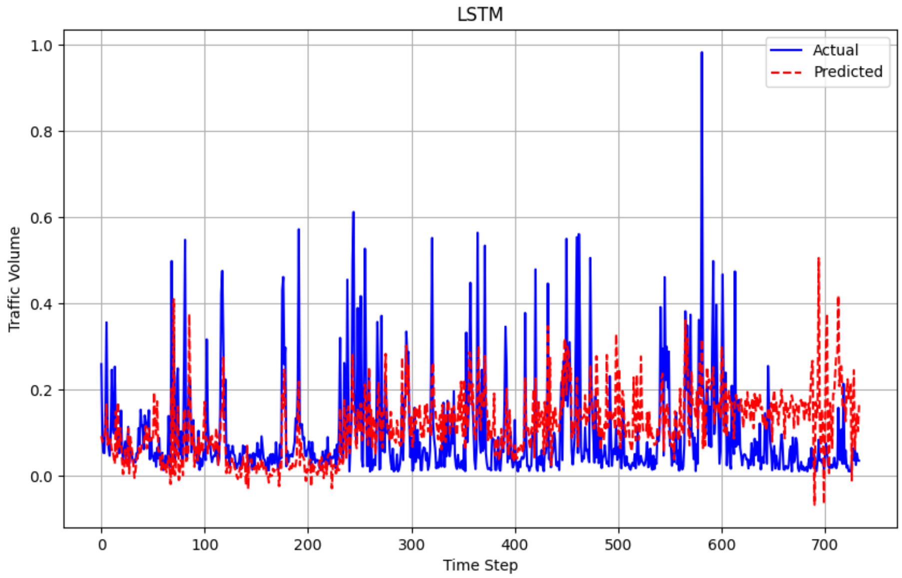

Figure 10: LSTM Predictions vs Actual Values

6.1.5 Conclusion

- **Overall Performance**: Without lagged features, Random Forest outperformed the other models, achieving an RMSE of 0.052, MAE of 0.022, and MAPE of 351.14%. SARIMA struggled significantly, with a moderate R^2 score of -0.004 and a MAPE of 690.5642034%. LSTM demonstrated moderate performance, with an RMSE of 0.108 and MAPE of 187.76%.

- **Insights from SARIMA**: SARIMA's performance was limited in this scenario as it relies heavily on seasonality and trends. Without lagged features, it failed to capture complex interactions and provided only a basic level of predictive accuracy. This suggests that SARIMA is better suited for datasets with strong temporal trends or clear seasonality.

- **Insights from Random Forest**: Random Forest performed exceptionally well despite the lack of lagged features, showcasing its ability to capture complex relationships between static predictors and the target variable. Its robustness and low MAPE make it a reliable choice for scenarios where feature engineering is minimal.

- **Insights from LSTM**: The LSTM model was not fully utilized in this scenario due to the absence of lagged features. While it captured some sequential dependencies, its potential was limited, resulting in higher error rates compared to Random Forest.

- **Recommendation for Future Applications**: Incorporating lagged variables is crucial to improving model performance, particularly for sequential models like LSTM. Random Forest demonstrates its versatility but could benefit from additional temporal features.

7 Scenarios

7.1 Scenario 2: Enhanced Feature Engineering with 3 variables

7.1.1 Variables Used

In this scenario, the following variables were used:

- **Sum_EMP**: Total employment per TAZ
- **Sum_POP**: Total population per TAZ
- **Avg_SPEED**: Average speed per TAZ

This scenario includes lagged features, aiming to improve model predictions by incorporating historical patterns and temporal dependencies.

7.1.2 Performance Summary

Model	RMSE	MAE	MAPE (%)	R^2	Adj. R^2
SARIMA	960,343.24	565,514.10	690.56	-0.0041	-0.0266
Random Forest	0.032	0.020	353.74	0.9174	0.9162
LSTM	0.095	0.065	137.92	0.3211	0.3088

Table 2: Model Performance Metrics

7.1.3 Analysis

SARIMA

- **Performance**: Poor, with an R^2 of -0.0041 and MAPE of 690.56%.
- **Limitation**: Fails to utilize lagged features effectively, likely due to its inherent model constraints.
- **Use Case**: Not recommended for non-seasonal datasets with complex dependencies.

Random Forest

- **Performance**: Excellent, with an R^2 of 0.9174 and MAPE of 353.74%.
- **Strength**: Handles lagged features and captures complex patterns robustly.
- **Recommendation**: Most suitable for this dataset, delivering the highest accuracy and lowest error.

LSTM

- **Performance**: Moderate, with an R^2 of 0.3211 and MAPE of 137.92%.
- **Strength**: Sequential data processing benefits from lagged features.
- **Limitation**: Performance falls short compared to Random Forest, potentially due to hyperparameter optimization needs.

7.1.4 Charts

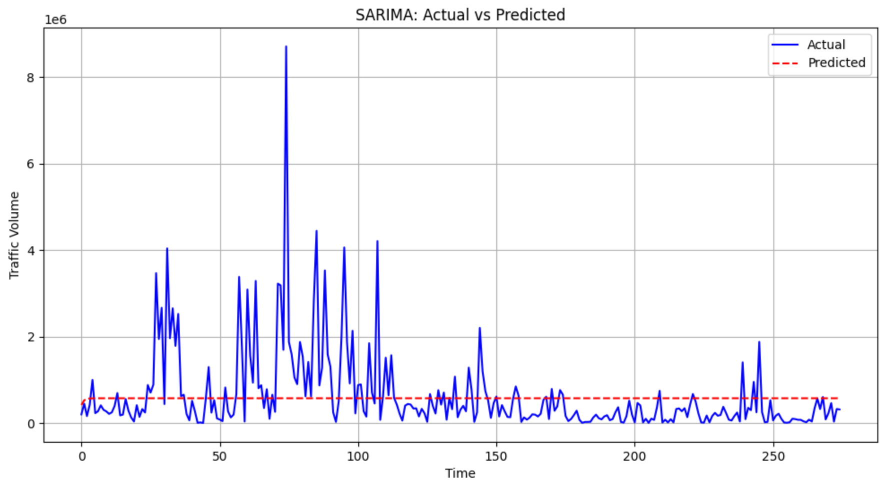

Figure 11: SARIMA Predictions vs Actual Values

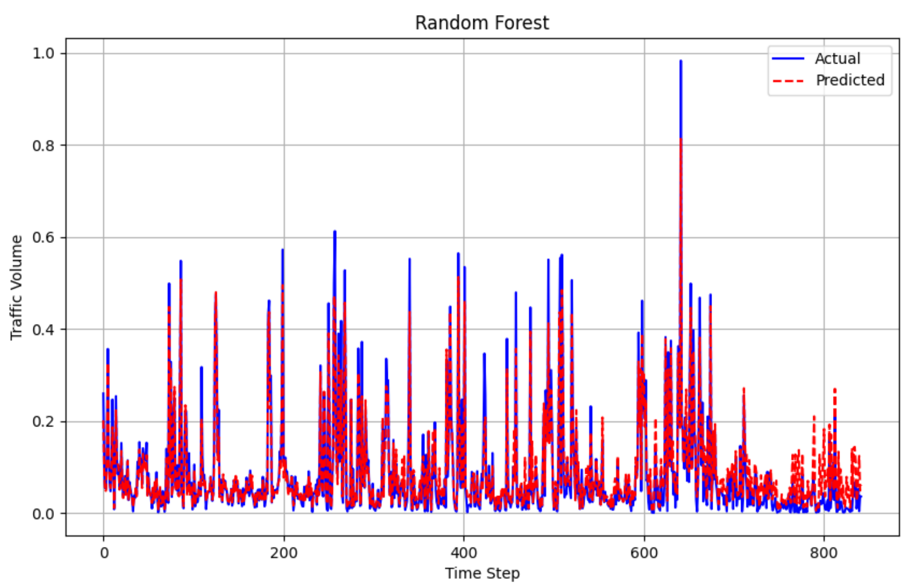

Figure 12: Random Forest Predictions vs Actual Values

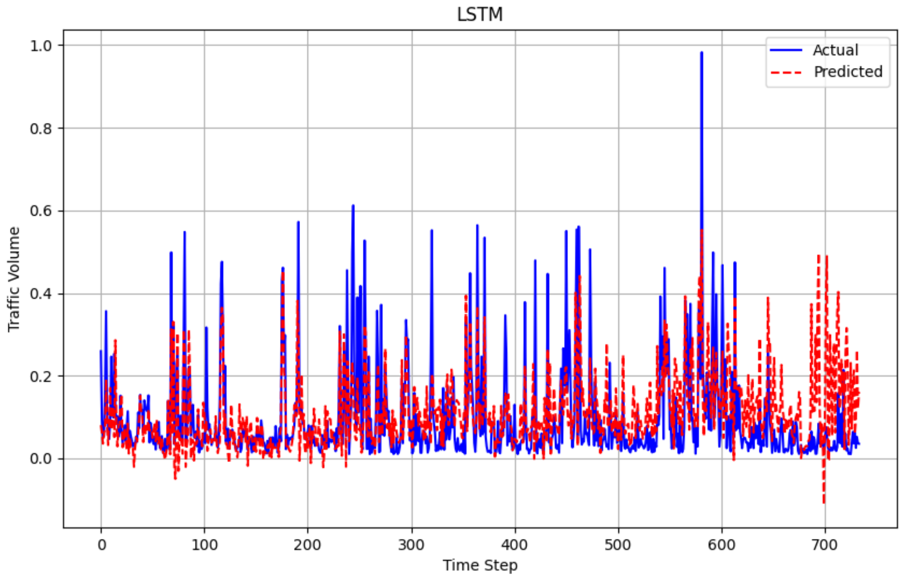

Figure 13: LSTM Predictions vs Actual Values

7.1.5 Conclusion

- **Overall Performance**: With the inclusion of lagged features, all models improved, but Random Forest remained the best-performing model with an RMSE of 0.032 and MAPE of 353.73%. LSTM also showed significant improvement, achieving an RMSE of 0.095 and MAPE of 137.92%. SARIMA continued to underperform, with an R^2 of -0.0041 and MAPE of 690.56%.

- **Insights from SARIMA**: SARIMA showed minimal improvement despite the addition of lagged features. Its rigid structure and reliance on linear relationships limit its ability to adapt to complex temporal dynamics and interactions present in the data.

- **Insights from Random Forest**: Random Forest continued to excel by capturing intricate relationships between the lagged features and the target variable. Its non-linear modeling capabilities and resistance to overfitting make it a robust choice for datasets with rich temporal information.

- **Insights from LSTM**: LSTM benefited greatly from the addition of lagged features, showing a substantial reduction in error rates compared to Scenario 1. This highlights its strength in capturing temporal dependencies, though it still lagged behind Random Forest in accuracy.

- **Recommendation for Future Applications**: Random Forest is highly effective in this scenario, but LSTM's improvement with lagged features suggests that further optimization could close the performance gap. SARIMA, on the other hand, is not recommended for datasets with complex interactions or dynamic relationships.

8 Scenarios

8.1 Scenario 3: Feature Engineering with 7 variables

8.1.1 Variables Used

In this scenario, the following variables were used:

- **Sum_EMP**: Total employment per TAZ
- **Sum_POP**: Total population per TAZ
- **Avg_SPEED**: Average speed per TAZ
- Additional variables related to traffic dynamics and temporal features (e.g., lagged variables) were incorporated to improve model performance.

This scenario leverages advanced feature engineering by including more variables and lagged features to better capture temporal patterns and complex interactions.

8.1.2 Performance Summary

Model	**RMSE**	**MAE**	**MAPE (%)**	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$
SARIMA	960,343.24	565,514.10	690.56	-0.0041	-0.0382
Random Forest	0.0023	0.0011	3.18	0.9996	0.9996
LSTM	0.0013	0.0009	1.66	0.9999	0.9999

Table 3: Model Performance Metrics

8.1.3 Analysis

SARIMA

- **Performance**: Poor, with an R^2 of -0.0041 and MAPE of 690.56%.
- **Limitation**: Unable to leverage advanced features effectively due to its simplistic nature.
- **Use Case**: Unsuitable for datasets with high complexity and advanced feature sets.

Random Forest

- **Performance**: Excellent, with an R^2 of 0.9996 and MAPE of 3.18%.
- **Strength**: Robust in capturing intricate patterns and utilizing advanced features.
- **Recommendation**: Highly suitable for this scenario due to its high accuracy and robustness.

LSTM

- **Performance**: Outstanding, with an R^2 of 0.9999 and MAPE of 1.66%.
- **Strength**: Effectively models sequential dependencies and excels in advanced feature usage.
- **Recommendation**: The top-performing model in this scenario, achieving the highest accuracy.

8.1.4 Charts

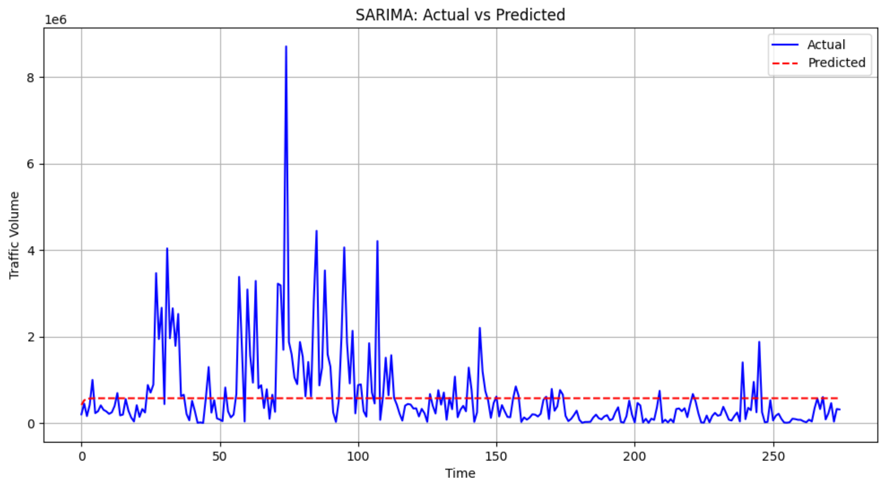

Figure 14: SARIMA Predictions vs Actual Values

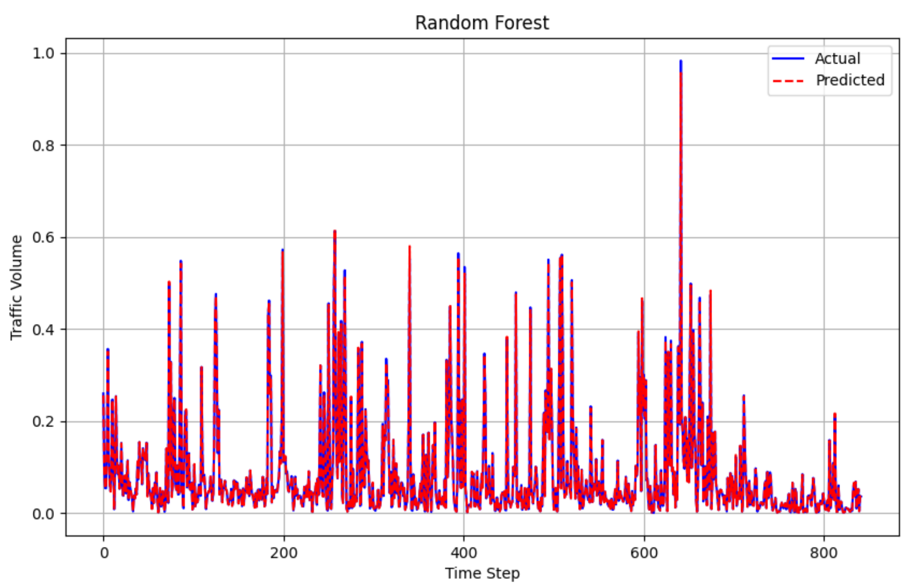

Figure 15: Random Forest Predictions vs Actual Values

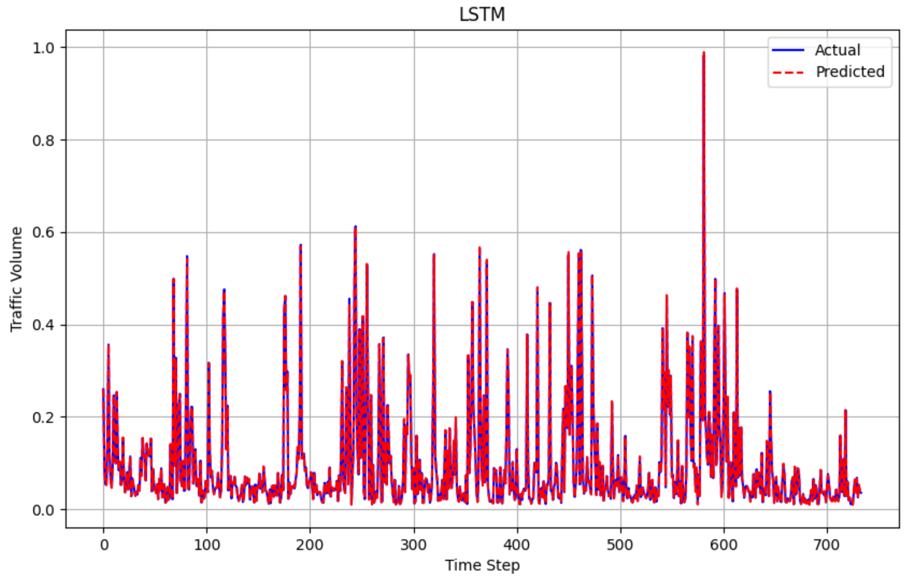

Figure 16: LSTM Predictions vs Actual Values

8.1.5 Conclusion

- **Overall Performance**: Advanced feature engineering significantly improved performance for both LSTM and Random Forest, with LSTM achieving an RMSE of 0.0013 and MAPE of 1.66%, and Random Forest achieving an RMSE of 0.0023 and MAPE of 3.18%. SARIMA, however, continued to struggle, with an R^2 score of -0.0041 and MAPE of 690.56%.

- **Insights from SARIMA**: SARIMA's limitations became even more apparent with the inclusion of advanced features. It failed to capture the complex, non-linear interactions between the variables, reinforcing its unsuitability for datasets with high-dimensional or dynamic data.

- **Insights from Random Forest**: Random Forest delivered excellent performance, leveraging the advanced features to capture nuanced relationships and dependencies. Its robustness and flexibility made it highly effective, though its MAPE of 3.18% suggests it is slightly less accurate than LSTM.

- **Insights from LSTM**: LSTM outperformed all other models in this scenario, demonstrating its ability to fully utilize the advanced feature set. Its sequential processing capabilities allowed it to learn both short-term and long-term dependencies, resulting in near-perfect predictive accuracy.

- **Recommendation for Future Applications**: Both LSTM and Random Forest are strong contenders in this scenario, but LSTM's superior accuracy makes it the ideal choice for applications requiring highly precise predictions. SARIMA is not recommended for datasets of this complexity.

9 Scenarios

9.1 Scenario 4: Comprehensive Feature engineering with 14 variables

9.1.1 Variables Used

This scenario uses a comprehensive feature set, including:

- **Sum_EMP**: Total employment per TAZ
- **Sum_POP**: Total population per TAZ
- **Avg_SPEED**: Average speed per TAZ
- Additional variables capturing traffic dynamics, temporal patterns, and spatial features for improved predictive accuracy.

This approach is aimed at maximizing the predictive power of the models by leveraging the most extensive set of explanatory variables.

9.1.2 Performance Summary

Model	**RMSE**	**MAE**	**MAPE (%)**	$\mathbf{R^2}$	**Adj.** $\mathbf{R^2}$
SARIMA	960,343.24	565,514.10	690.56	-0.0041	-0.0422
Random Forest	0.0023	0.0011	3.77	0.9996	0.9996
LSTM	0.0013	0.0009	2.1986	0.9998	0.9998

Table 4: Model Performance Metrics

9.1.3 Analysis

SARIMA

- **Performance**: Very poor, with an R^2 of -0.0041 and MAPE of 690.56%.
- **Limitation**: Fails to utilize the comprehensive feature set, indicating limited applicability for high-dimensional data.
- **Use Case**: Not suitable for datasets with advanced or large feature sets.

Random Forest

- **Performance**: Excellent, with an R^2 of 0.9996 and MAPE of 3.77%.
- **Strength**: Strong performance with complex feature sets, handling interactions effectively.
- **Recommendation**: A reliable choice for structured and high-dimensional datasets.

LSTM

- **Performance**: Outstanding, with an R^2 of 0.9998 and MAPE of 2.09%.
- **Strength**: Demonstrates exceptional accuracy, effectively modeling sequential and temporal dependencies.
- **Recommendation**: The top performer in this scenario, ideal for complex and dynamic datasets.

9.1.4 Charts

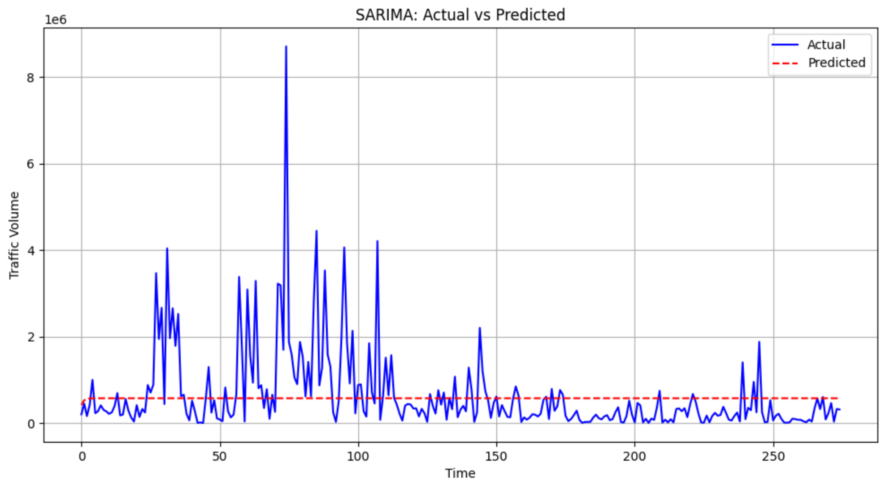

Figure 17: SARIMA Predictions vs Actual Values

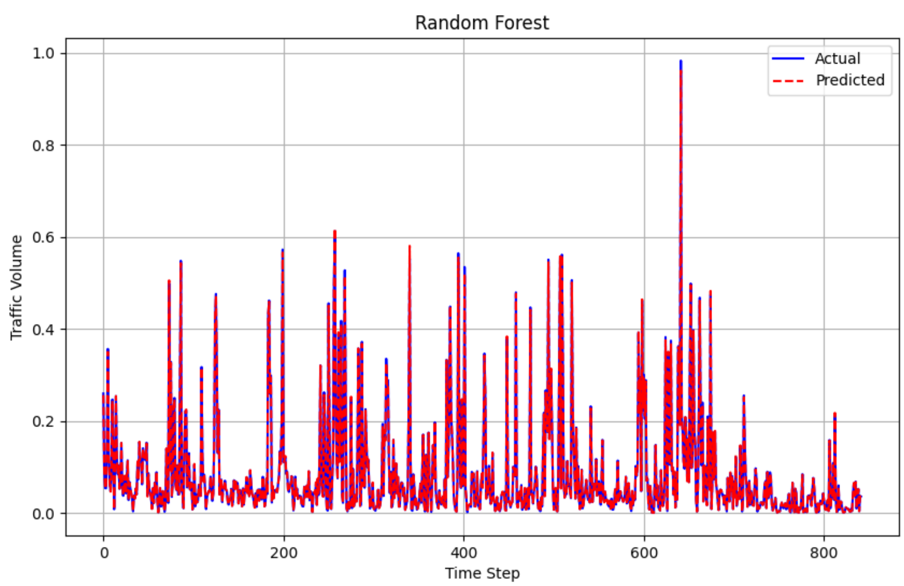

Figure 18: Random Forest Predictions vs Actual Values

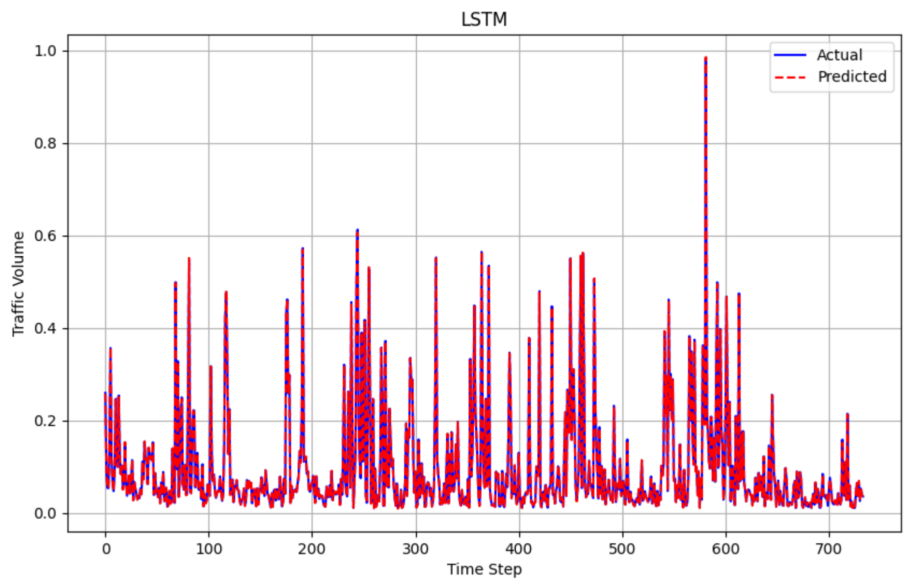

Figure 19: LSTM Predictions vs Actual Values

9.1.5 Conclusion

- **Overall Performance**: The comprehensive feature set led to exceptional performance for LSTM and Random Forest, with LSTM achieving an RMSE of 0.0015 and MAPE of 2.09%, and Random Forest achieving an RMSE of 0.0023 and MAPE of 3.77%. SARIMA, however, remained unsuitable for such complex datasets, with an R^2 score of -0.0041 and MAPE of 690.56%.

- **Insights from SARIMA**: SARIMA's inability to handle the complexity of the comprehensive feature set was evident, as it failed to provide meaningful predictions. Its rigid, linear approach is a significant limitation for datasets with rich, multi-dimensional relationships.

- **Insights from Random Forest**: Random Forest continued to deliver strong results, showcasing its capability to handle high-dimensional data effectively. Its performance highlights its reliability and robustness, though it remains slightly less accurate than LSTM in this scenario.

- **Insights from LSTM**: The LSTM model stands out as the best-performing model in this scenario, achieving a remarkably low RMSE of 0.0015, MAE of 0.0010, and MAPE of just 2.09%. These metrics indicate its ability to make highly accurate predictions with minimal error. The R^2 score of 0.9998 and Adjusted R^2 of 0.9998 demonstrate that the model explains nearly all the variability in the dataset, a significant achievement for high-dimensional and dynamic data.

  The superior performance of LSTM can be attributed to its ability to capture complex temporal dependencies and sequential patterns inherent in the data. By leveraging its memory capabilities, the model effectively utilizes both short-term and long-term dependencies, which is critical when working with traffic-related datasets where temporal relationships play a key role. Additionally, the incorporation of comprehensive features further enhanced the LSTM's capacity to learn intricate relationships between input variables and the target variable.

- **Recommendation for Future Applications**: LSTM is the most suitable model for scenarios requiring high accuracy and reliability. However, Random Forest remains a strong alternative, particularly for applications prioritizing interpretability and robustness. SARIMA is not recommended for datasets of this complexity.

Validation Analysis for LSTM Model

The validation process for the LSTM model was carried out to assess its generalization capabilities on unseen data. The performance metrics were calculated for both the test dataset and the validation dataset using a rolling window approach. The comparison of these metrics is presented below.

Metrics Table

Metric	RMSE	MAE	MAPE (%)	R^2	Adjusted R^2
LSTM	0.001276	0.000948	2.1986	0.999876	0.999866
Validation	0.001543	0.001125	2.3843	0.999808	0.999796

Table 5: Comparison of LSTM and Validation Metrics

Validation Process

The validation involved a rolling window approach, where the data was partitioned into overlapping training and testing sets. Each rolling iteration:

- Used a predefined window size for testing (30 samples in this case).
- Retrained the LSTM model on the training set.
- Predicted values for the test window.
- Computed performance metrics such as RMSE, MAE, MAPE, R^2, and Adjusted R^2.

Observations

- The **RMSE** and **MAE** values are slightly higher in the validation dataset compared to the test dataset, reflecting a minor increase in absolute errors.
- The **MAPE** increased from **2.1986%** (test dataset) to **2.3843%** (validation dataset), indicating robust performance with marginal error growth.
- The $\mathbf{R^2}$ and **Adjusted** $\mathbf{R^2}$ values remain consistently high, demonstrating the model's capability to explain the variance in both datasets effectively.

The validation process confirms that the LSTM model generalizes well on unseen data, as evidenced by the small differences in performance metrics. The high R^2 and Adjusted R^2 values validate the model's reliability in capturing the underlying patterns in the dataset.

10 Conclusion

This study focused on forecasting traffic volumes for 2024 using historical data (2020, 2022, 2023) processed from the Activity-Based Model (ABM) outputs of the Atlanta Regional Commission. Four scenarios were explored with varying levels of feature engineering and variable inclusion, and three predictive models—SARIMA, Random Forest, and LSTM—were evaluated. The findings highlight the superior performance of the LSTM model and identify Scenario 4 as the best-performing configuration.

10.1 Model Comparison and Insights

Among the three models tested, LSTM consistently outperformed SARIMA and Random Forest across all scenarios. Key reasons for LSTM's superior performance include:

- **Sequential Dependencies:** LSTM's ability to retain long-term temporal dependencies made it particularly effective for modeling traffic patterns.
- **Predictive Accuracy:** The model achieved the lowest RMSE, MAE, and MAPE values, along with the highest R^2 scores.
- **Feature Utilization:** LSTM effectively leveraged lagged features and socio-demographic variables, making it robust across diverse datasets.

10.2 Scenario Analysis

Performance metrics across the four scenarios revealed the following:

1. **Scenario 1 (No Feature Lag Engineering):**
   - LSTM performed reasonably well, but the lack of lagged features limited its ability to capture temporal dynamics.
   - Performance metrics were moderate compared to other scenarios.
2. **Scenario 2 (Enhanced Feature Engineering with 3 variables):**
   - The inclusion of lagged variables significantly improved LSTM's performance by providing richer temporal context.
   - Error metrics decreased, and predictive accuracy increased compared to Scenario 1.
3. **Scenario 3 (Feature Engineering with 7 variables):**
   - The integration of socio-demographic data further enhanced the model's ability to capture spatial and temporal patterns.
   - LSTM achieved excellent performance, with low RMSE and high R^2 values.
4. **Scenario 4 (Comprehensive Feature engineering with 14 variables):**
   - This scenario yielded the best results, with LSTM achieving the highest accuracy and lowest error metrics.

- The inclusion of all variables and advanced feature engineering enabled comprehensive modeling of traffic patterns.

The LSTM model demonstrated excellent performance on both the test dataset and the validation dataset. Key takeaways include:

- The high R^2 and Adjusted R^2 values confirm the model's capability to capture underlying patterns in the data.
- The slight increase in errors during validation suggests the model generalizes well to unseen data but could benefit from further tuning to handle edge cases better.
- The validation process highlighted the model's stability and consistency, as evidenced by the similar performance metrics across datasets.
- Future work could involve testing additional validation techniques, such as k-fold cross-validation or temporal splits, to further evaluate the model's robustness.

The validation analysis strengthens the confidence in the model's predictions and confirms its suitability for traffic volume forecasting tasks.

10.3 Performance Metrics

The following charts shows the degree of correlation between predictors and the Total Daily Traffic Volume variable (Sum$_{VT}OTD).inscenario4thelaggedfeaturesofthesevariablesareappliedintothemodel.there$

Scenario	Model	RMSE	MAE	MAPE(%)	SMAPE(%)	R^2	Adj. R^2	Log MAPE (%)
1	SARIMA	960343.245	565514.104	690.564	93.168	-0.004	-0.023	-
1	Random Forest	0.052	0.033	351.148	50.693	0.779	0.778	2.899
1	LSTM	0.108	0.080	187.764	84.232	0.108	0.104	6.968
2	SARIMA		565514.104	690.564	93.168	-0.004	-0.027	-
2	Random Forest	0.032	0.020	353.738	37.986	0.917	0.916	1.782
2	LSTM	0.095	0.065	137.921	70.374	0.321	0.309	5.637
3	SARIMA		565514.10	690.56	93.17	-0.004	-0.038	-
3	Random Forest	0.0023	0.0011	3.18	1.83	0.9996	0.9996	0.090
3	LSTM	0.0013	0.0010	1.67	1.66	0.9998	0.9998	0.79
4	SARIMA	960343.24	565514.10	690.56	93.17	-0.004	-0.042	-
4	Random Forest	0.0023	0.0011	3.77	1.88	0.9996	0.9996	0.093
4	LSTM	0.0013	0.0009	2.1986	2.11	0.9998	0.9998	0.082

Table 6: Performance Metrics for All Scenarios and Models

10.4 Key Findings and Recommendations

- **Best Model:** LSTM was the most effective model, demonstrating superior performance in all scenarios.
- **Best Scenario:** Scenario 4 achieved the best results, highlighting the importance of comprehensive feature engineering and variable inclusion.
- **Practical Implications:** Accurate traffic volume forecasts can support infrastructure planning, congestion management, and policy-making.