RIDWAN AFOLABI
MUHAMMAD (AIU23102185)

ALBUKHARY INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTING AND INFORMATICS

SEMESTER 3 2024/2025
CCS2313_ Data Mining & Analytics
LECTURER:
DR. UMMI KULTHOM

TIME-SERIES ANALYSIS & FORECASTING

A STUDY ON LEVERAGING MACHINE LEARNING AND DEEP LEARNING FOR ACCURATE WEB TRAFFIC PREDICTION

PROJECT BACKGROUND/PROBLEM STATEMENT

Website traffic data is complex, non-linear, and non-stationary. This project tackles the challenge of capturing these patterns to produce reliable forecasts, enabling better resource management and strategic planning.

OBJECTIVES

- To implement, train, and evaluate traditional machine learning models, the Random Forest Regressor and the XGBoost Regressor, for forecasting daily website visitors.
- To implement, train, and evaluate a deep learning model, the Long Short-Term Memory (LSTM) neural network, and to conduct a comparative analysis of its performance against the other two models.

DATA AND ANALYSIS

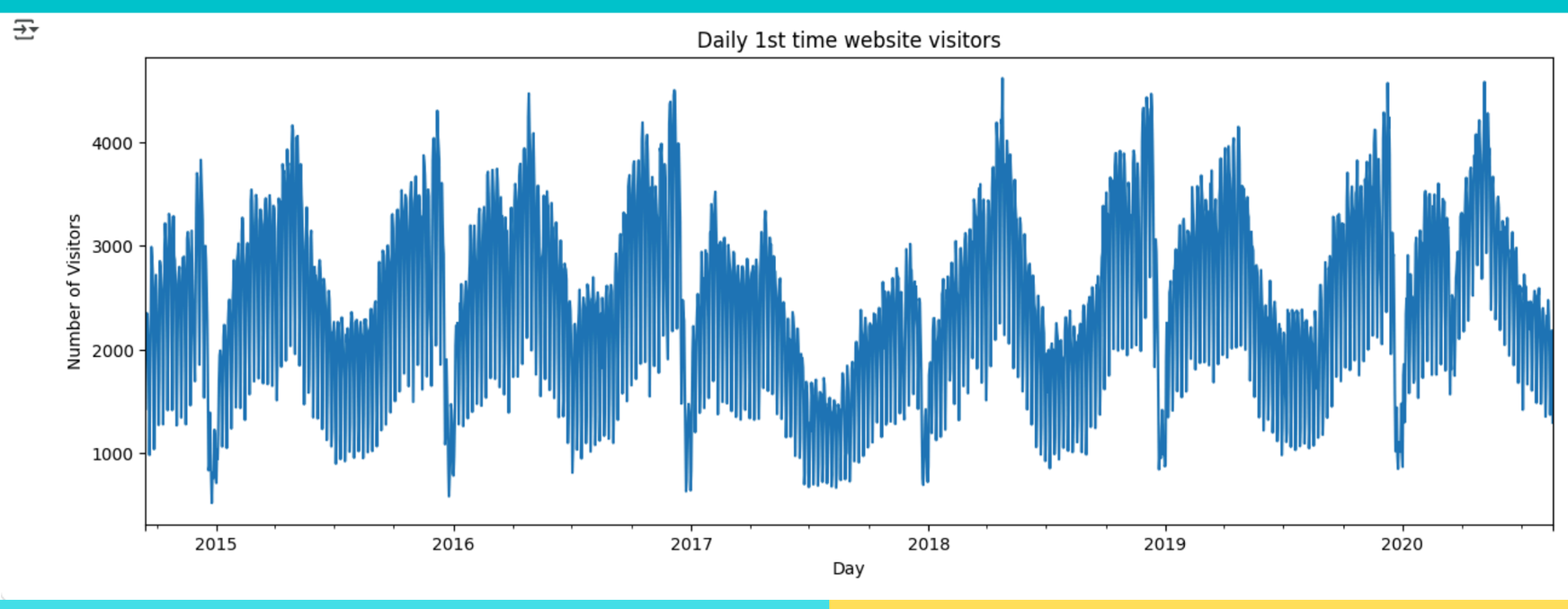
Data Source

Used a publicly available Kaggle dataset containing daily website visitor information from September 2014 to June 2020. The dataset was preprocessed to handle missing values and convert columns to the appropriate data types.

Initial Data Insights

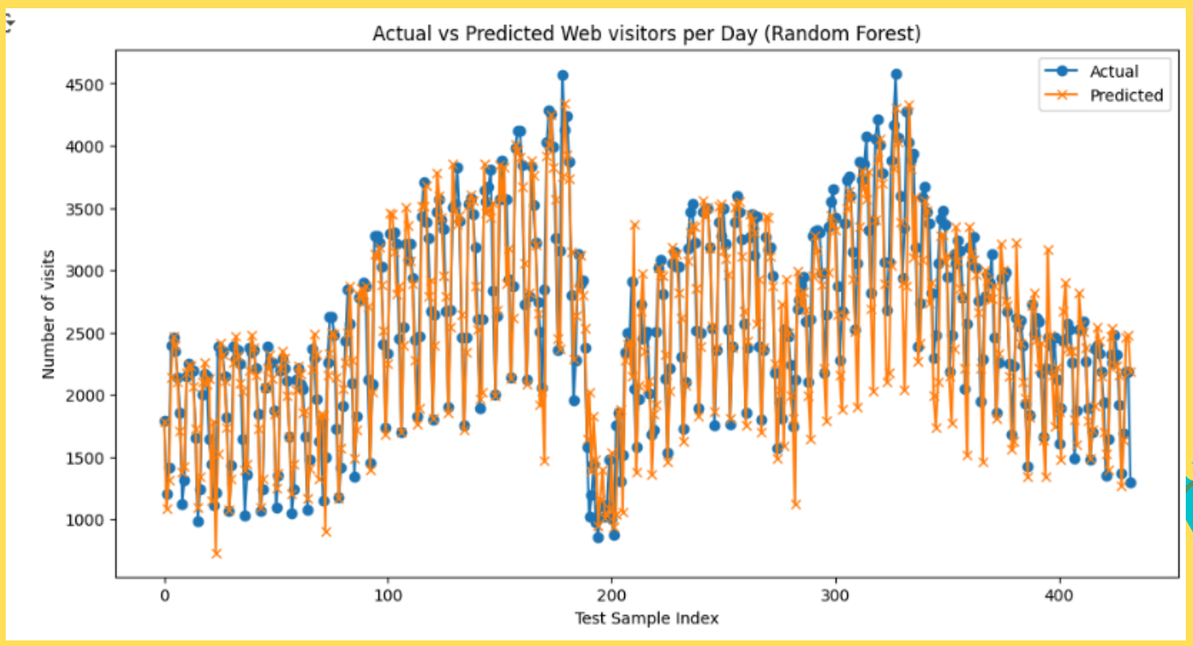
An initial analysis of the "First-Time Visits" time series revealed two key patterns:

- **Strong Seasonality:** The data exhibits a clear annual cycle with regular peaks and troughs, suggesting a yearly repeating pattern.
- **Upward Trend:** Over the six-year period, there is a slight, but consistent, increase in the number of daily visitors.



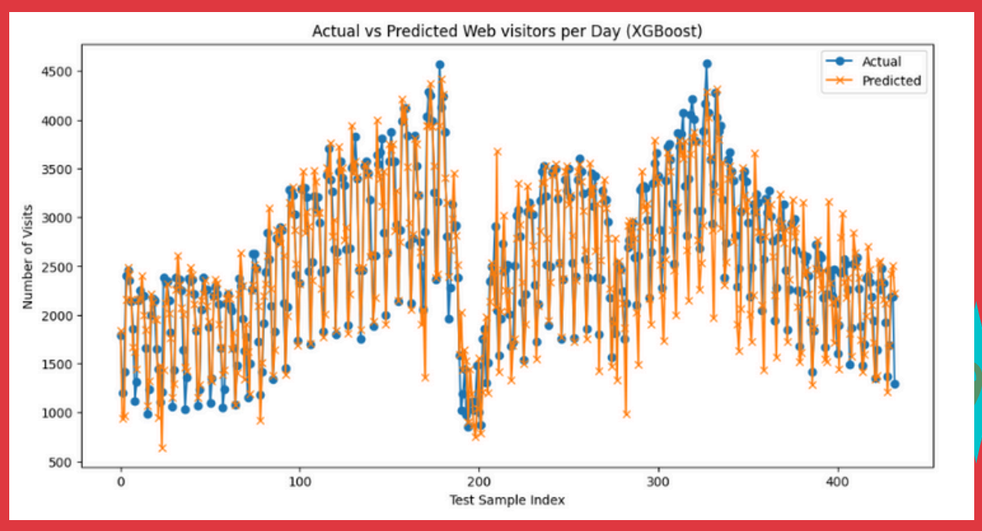
METHODOLOGY AND EXPERIMENTS

1. **Random Forest Regressor:** It was used as a baseline to evaluate its performance on this non-linear dataset.
   - **Approach:** data was transformed into a supervised learning problem by creating new features known as lags.
   - **Results**: Good.

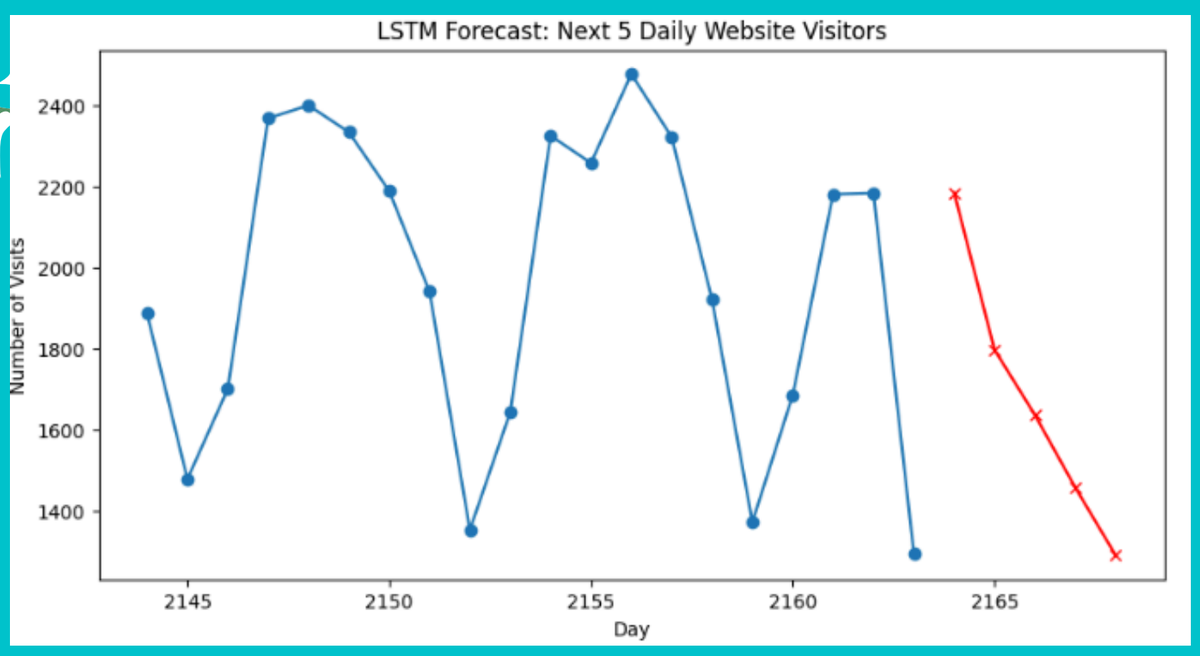


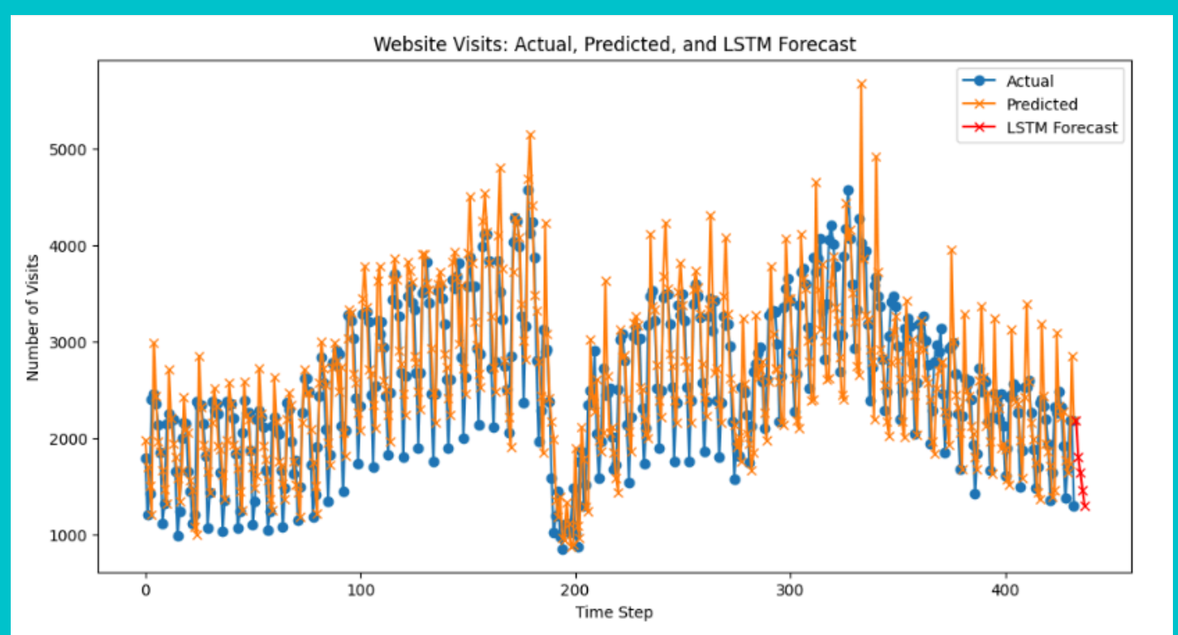
METHODOLOGY AND EXPERIMENTS

2. **XGBoost Regressor:** An advanced, highly accurate gradient boosting model.
   - Approach: Similar to Random Forest, it used the same engineered features to train on the data.
   - Results: Good performance.



3. **Long Short-Term Memory (LSTM) Network:** A type of Recurrent Neural Network (RNN) specifically designed to learn long-term dependencies in sequential data.
   - Approach: The LSTM model did not require extensive feature engineering. The data was scaled and reshaped to be used as a sequence.
   - Results: The LSTM model's predictions closely matched the actual data, successfully capturing both the overall trend and subtle daily fluctuations. When used for forecasting, the model produced a five-day forecast that followed the recent downward trend, demonstrating its ability to project future values.





FINDINGS AND CONCLUSION

- **Models Compared:** The project successfully implemented and compared three forecasting models: Random Forest, XGBoost, and LSTM. This covered both traditional machine learning and a modern deep learning approach.
- **Project Lifecycle:** The entire forecasting process was executed, from initial data collection and preparation to model implementation, evaluation, and a final comparative analysis of the results.
- **Feature Engineering:** Random Forest and XGBoost models were effective at capturing trends and seasonal patterns by relying on carefully engineered features.
- **LSTM's Strength:** The LSTM network proved to be more powerful by learning complex patterns directly from the data sequences even without the need for manual feature creation, which is a significant advantage for complex time series data.
- **Performance:** While tree-based models offer valuable and efficient solutions, the results align with current research, confirming that advanced methods like XGBoost and LSTM provide more accurate and robust forecasts for dynamic and non-linear data like website traffic.
- **Future Directions:** Further work could explore creating hybrid models that combine the strengths of both tree-based and deep learning approaches, incorporating additional variables (like holidays or marketing campaigns), or testing other models such as Facebook's Prophet.