Predicting Traffic Patterns In New York City

Katie Park and Jonatan Peguero
Emory University CS 470

April 23, 2025

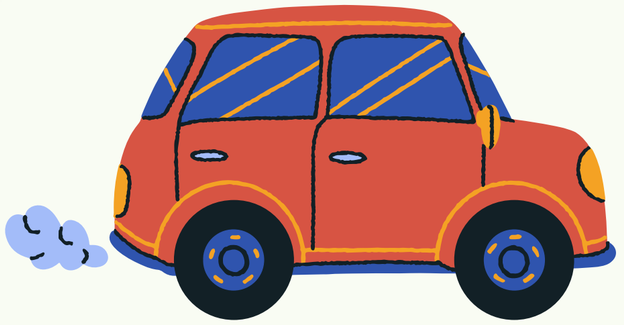
Introduction

- **Problem:** Traffic congestion in cities like NYC causes delays, pollution, and higher costs.

- **Traffic Leads to**
  - Longer Commutes
  - Increased Stressed
  - Loss of Time and Productivity
  - Increased greenhouse gas emissions

Research Question: How to avoid traffic in New York City during the Summer?

Summer is a busy season, as tourism increases and students are out of school, which often leads to more potential vehicles on the road.



Related Works

Lv, Duan, Kang, Li and Wang (2015)

- Used deep learning on large-scale traffic data.
- Outperformed traditional methods like ARIMA and SVR.
- Showed deep learning’s strength in automatically learning traffic features.

Google Maps + DeepMind (2021)

- Uses AI + real-time mobile data for traffic prediction.
- Improved ETA accuracy using Graph Neural Networks (GNNs).
- Highlights the power of AI in real-world traffic forecasting systems.

Gomes, Coelho & Aidos (2021)

Surveyed traffic prediction techniques:

- Deep Learning (RNNs, CNNs)
- Parametric Models (AR, Regression)
- Genetic Programming

Our Dataset

Automated Traffic Volume Counts Transportation

New York City Department of Transportation (NYC DOT) uses Automated Traffic Recorders (ATR) to collect traffic sample volume counts at bridge crossings and roadways.These counts do not cover the entire year, and the number of days counted per location may vary from year to year.

Title : NYC Open Data Automated Traffic Volume Counts

What It Contains:

- Hourly vehicle counts at various traffic sensors
- Location-based data for different street segments
- Time-based attributes: year, month, day, hour, and minute
- Traffic direction and segment ID

Our Focus (What we filtered):

- **Summer months:** June–August
- **Years:** 2016, 2018, 2021, 2022
- **Borough:** Brooklyn
- **Key features used:**
  - Month
  - Day
  - Hour
  - Minute
  - Traffic Direction

Methodology

Step 1: Data Cleaning & Filtering

- **Filtered**
  - Months: June–August
  - Years: 2016, 2018, 2021, 2022
  - Borough: Brooklyn

Step 2: Feature Engineering

- **Extracted features:**
  - Month
  - Day
  - Hour
  - Minute
  - Traffic Direction

Step 3: Model Building

- **Models Trained and Tested:**
  - 80% train, 20% test split
  - Random Forest Regression
  - LSTM (Long short-term memory)
- **Evaluated using:**
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - R^2 Score
  - K-Fold CV to test generalization

(Data → Features → Model → Prediction → Evaluation)

Results

Mean Absolute Error (MAE): On average, how far off prediction values are compared to actual data values.

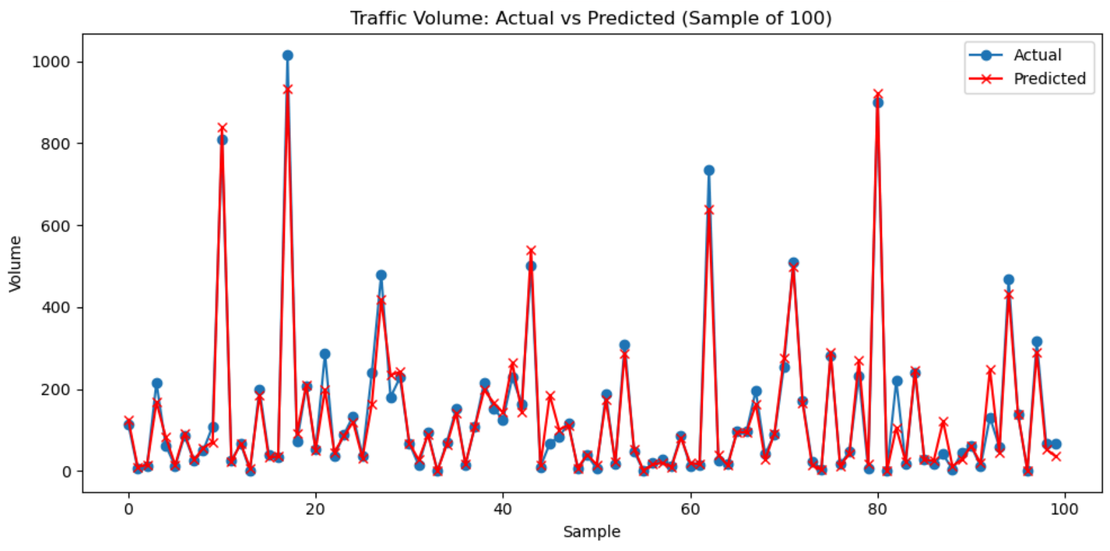
Root Mean Squared Error (RMSE): On average, how far off predictions values are with more weight to larger deviations

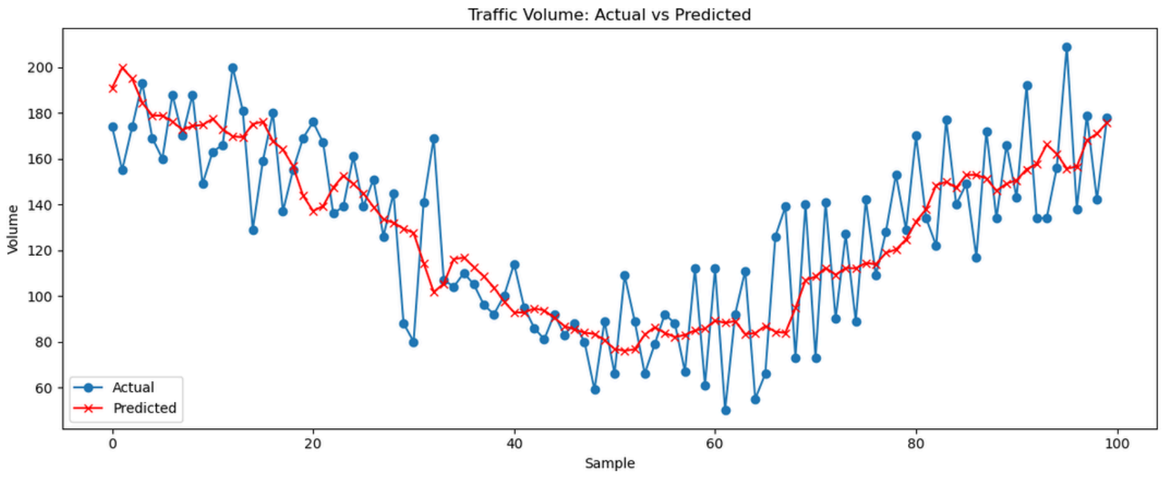
R^2 Score (0 - 1): How well the model matches the actual data

Model	MAE	RMSE	R^2 Score
Random Forest	**15.45**	29.42	**0.97**
LSTM	16.84	**21.13**	0.86

Results

Actual vs. Predicted values for each model





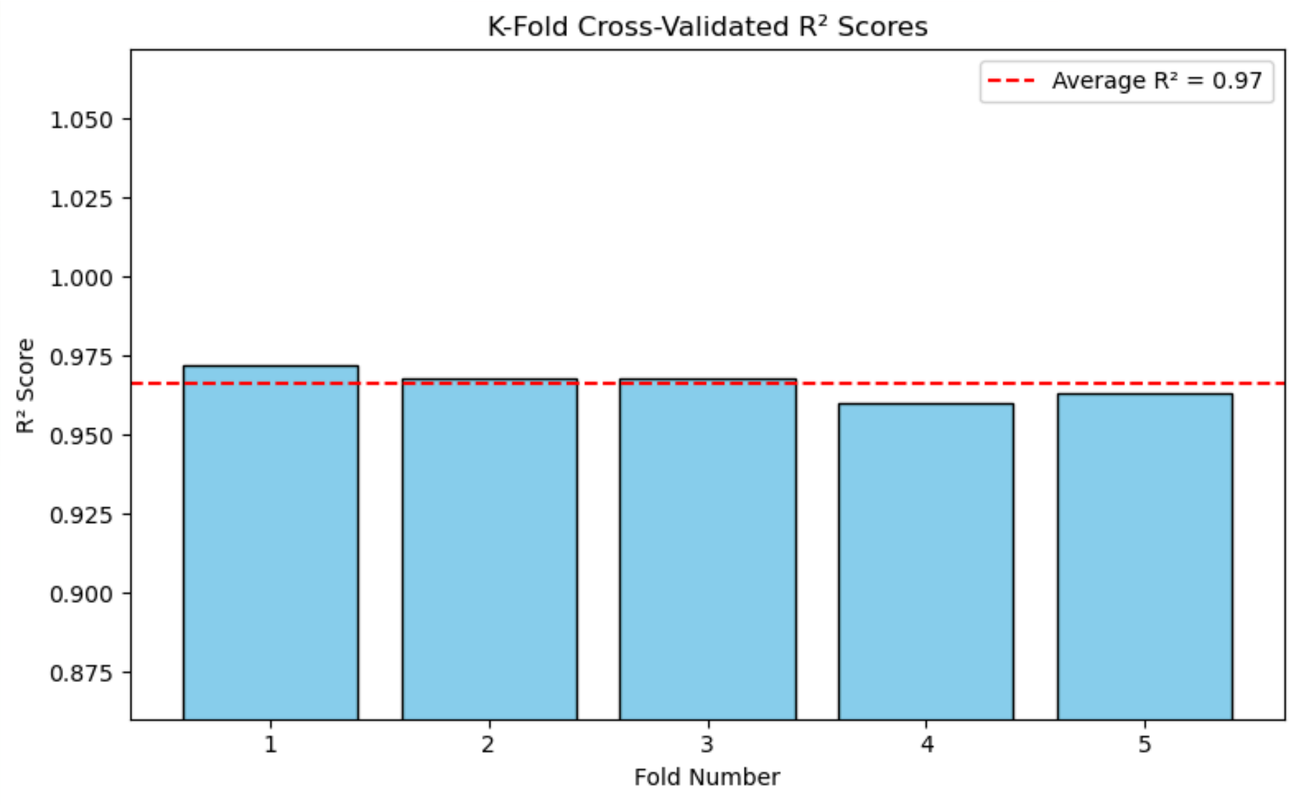
Random Forest Results

- Captured and Predicted traffic volume well with high accuracy
- Learned detailed patterns such as higher traffic spikes

LSTM Results

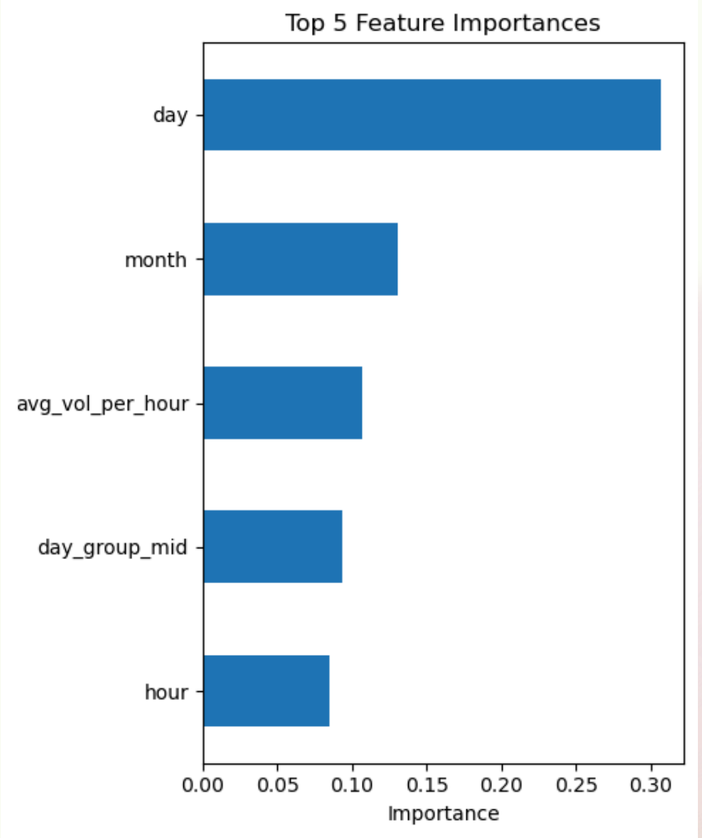
- Missed exact volume counts, but tracked general patterns over time
- Predicts smoother stable traffic trends

K-Fold Cross Validation and Most Important Features (Random Forest)



K-Fold Cross Validation

- The model was tested on 5 different subsets of the data
- R² scores were consistently high across all folds
- Model is reliable and not just memorizing the training data.

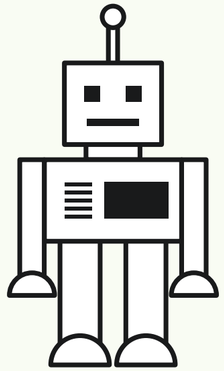


Most Important Features

- Day of the month was the most important predictor of traffic
- Month, avg_vol_per_hour, and hour of day helped make better predictons

Interpretation and Conclusion

VS



Random Forest
Predicted traffic volume more accurately

Long Short Term Memory (LSTM)
Captured overall trends, but missed exact volume counts

Why These Results?

- **LSTM was trained on one street segment**
  - Limited changes made it hard for the model to learn strong patterns
- **Small dataset**
  - LSTM typically requires larger datasets perform well
  - Random Forest works better on smaller datasets
- **Easy splitting features (Hour, minute)**
  - Random Forest could easily split and make decisions based on these values

Real World Application

- **Helps drivers plan the best time to travel**
  - Specifically in areas of Brooklyn, New York
- **Improves traffic management and planning**
  - Shows when and where traffic is likely to occur, helping with scheduling and road usage decisions
- **Reduces congestion and emissions**
  - Helps drivers avoid busy roads leads to lower emissions and improved air quality

Thank You!

Questions?