Hourly Traffic Volume Forecasting

Springboard Data Science Career Track – Capstone Three

Forecasting hourly traffic using time-series models

Problem Statement

- Urban traffic congestion creates operational, economic, and safety challenges.
- Traffic management teams need reliable short-term forecasts to allocate resources.
- The goal is to forecast hourly traffic volume up to 7 days ahead.
- Success metric: accurate out-of-sample prediction using RMSE.

Dataset Overview

- Metro Interstate Traffic Volume dataset.
- Hourly traffic counts collected over multiple years.
- Includes weather variables and holiday indicators.
- Target variable: traffic_volume.

Data Wrangling – Issues

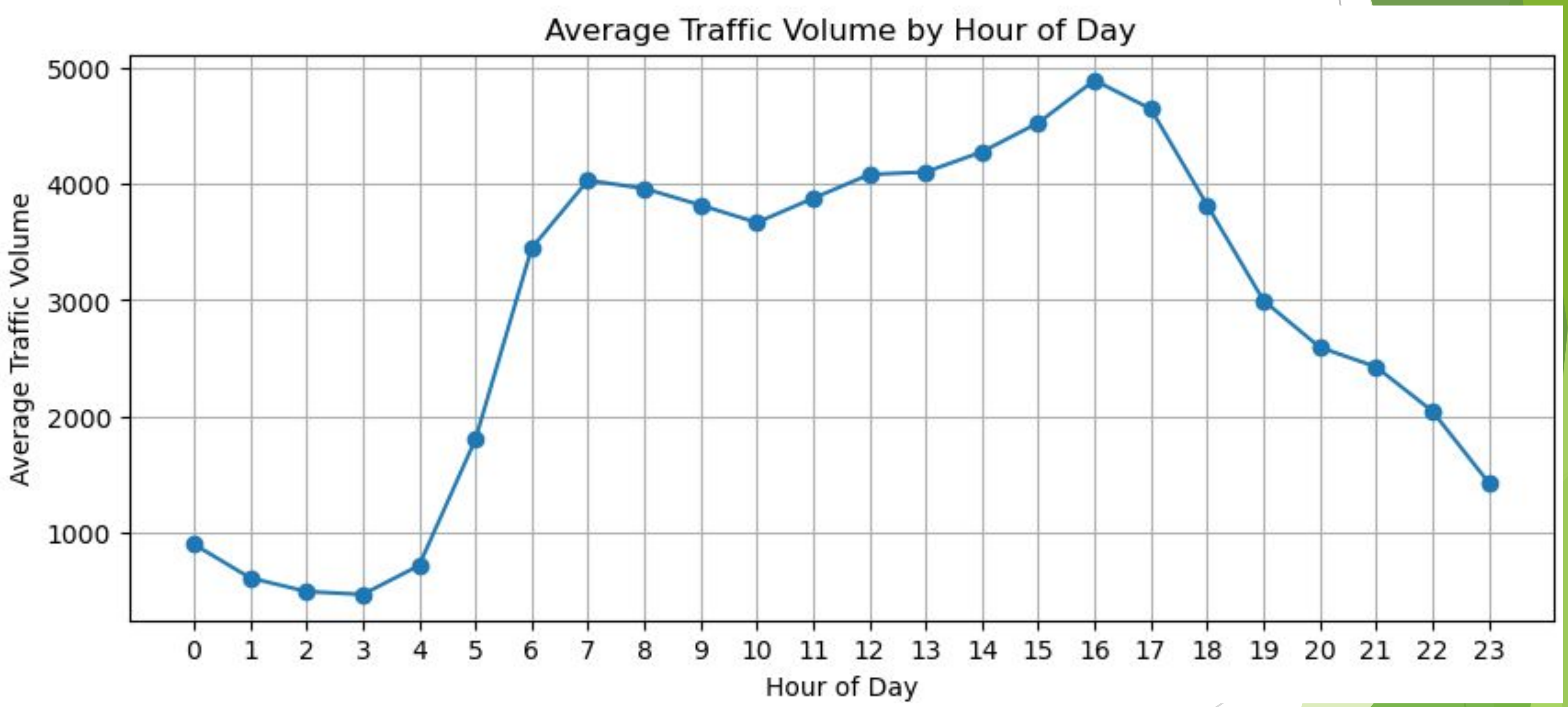
- Duplicate hourly timestamps.
- Missing hours.
- Mixed numeric and categorical features.

Data Wrangling – Solutions

- Duplicate timestamps were identified and aggregated to one row per hour.
- Missing hourly intervals were explicitly introduced to ensure continuity.
- Time index was restructured to a complete hourly sequence.
- Holiday missing values were treated as non-holiday periods.

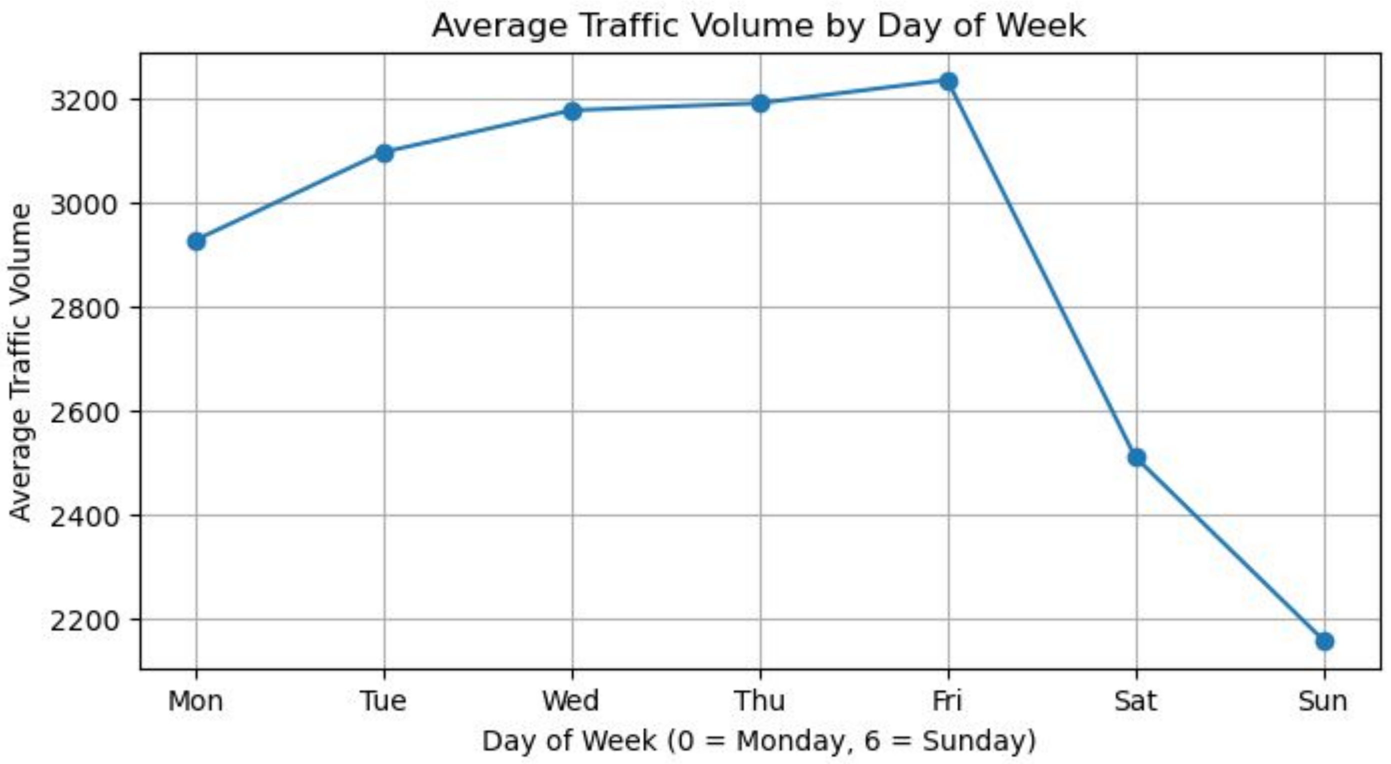
Exploratory Data Analysis

- Strong daily seasonality with clear morning and evening peaks.



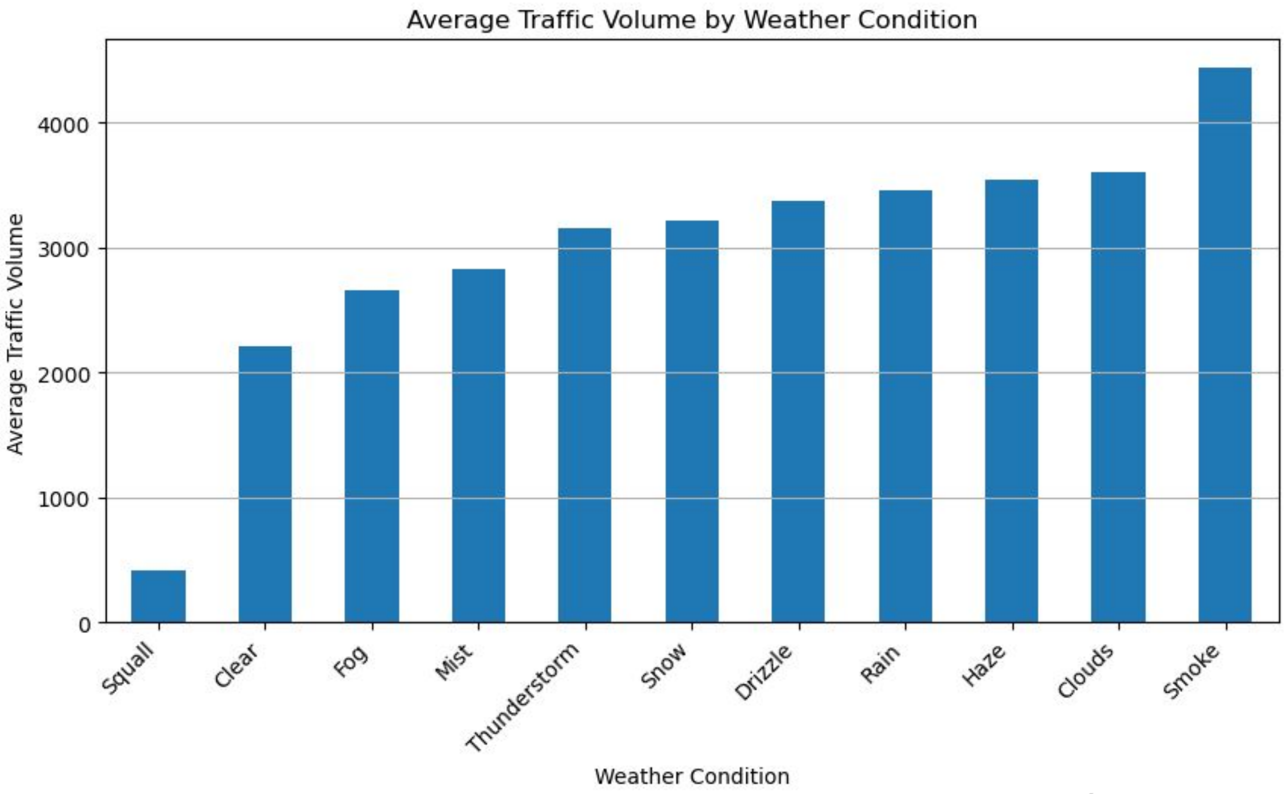
Exploratory Data Analysis

- Consistent weekly pattern: higher weekday traffic, lower weekends.



EDA – Weather & Holidays

- Weather and holidays affect traffic.
- Effects are secondary to seasonality.
- Do not override core patterns.



Preprocessing & Feature Engineering 1/2

- Ensured evenly spaced hourly time series required for time-series models.
- Created binary holiday indicators.
- Prepared light and full exogenous feature sets for SARIMAX models.

Preprocessing & Feature Engineering 2/2

- Engineered a binary holiday indicator.
- Selected compact weather features (temperature, rain, snow, cloud cover).
- Excluded high-cardinality weather descriptions to reduce noise.

Key takeaway:

preprocessing ensured a clean, continuous time series, while feature engineering captured external signals without overriding core temporal patterns.

Modeling Strategy

- Progressive approach: baseline → SARIMA → SARIMAX.
- Goal: test whether added complexity improves accuracy.
- Models evaluated using consistent time-based split.

Models Implemented

- Seasonal Naive Baseline (same hour, one week earlier).
- SARIMA without seasonality.
- SARIMA with daily seasonality.
- SARIMAX with full exogenous features (no seasonality).
- SARIMAX with daily seasonality and light exogenous features.

Model Performance (RMSE)

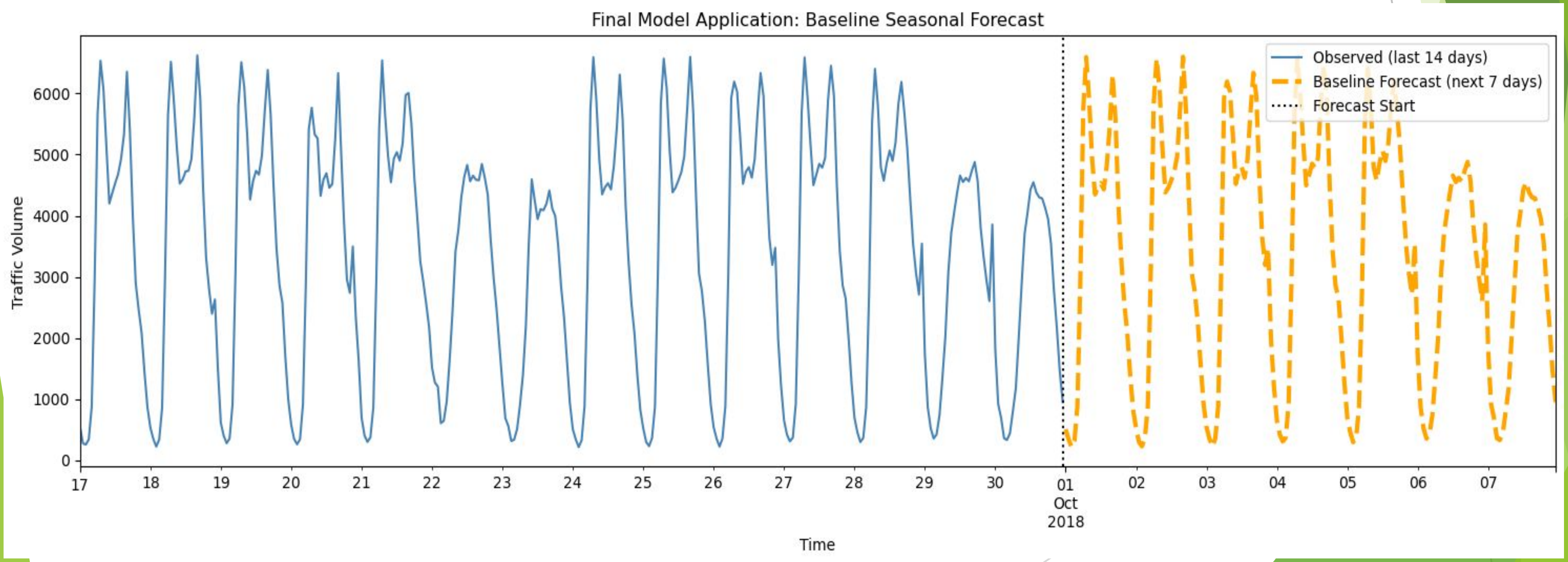
- Seasonal Naive Baseline: RMSE ≈ 646 (best).
- SARIMA (No Seasonality): RMSE ≈ 3858 (worst).
- SARIMA (Daily Seasonality): RMSE ≈ 983.
- SARIMAX (Full Exogenous, No Seasonality): RMSE ≈ 2015.
- SARIMAX (Daily Seasonality + Light Exogenous): RMSE ≈ 958.

Final Model Selection

- The seasonal naive baseline achieved the lowest RMSE.
- More complex models did not improve accuracy.
- Seasonality alone explains most traffic behavior.
- The baseline model was selected as the final model.

Final Model Application

- Generated a real 7-day hourly forecast beyond the dataset.
- Forecast preserves daily and weekly patterns.
- No artificial trend or instability introduced.
- Results are interpretable and operationally reliable.



Business Recommendations

- Use the seasonal baseline model for short-term traffic planning.
- Schedule staffing and congestion mitigation around predictable peak hours.
- Use weather and holidays as monitoring signals, not core forecast drivers.

Limitations & Future Work

- Baseline model assumes stable seasonal patterns.
- Does not adapt to structural changes or rare events.
- Future work could explore event-based modeling.
- Integration with real-time data could improve responsiveness.