Building a model to predict Uber trip requests

Natalia Shcheglova, Oct 19, 2024

Why Would We Want To Predict The Demand?

Higher revenues

- Demand-based pricing.
- Increased capacity.

Improved operational efficiency

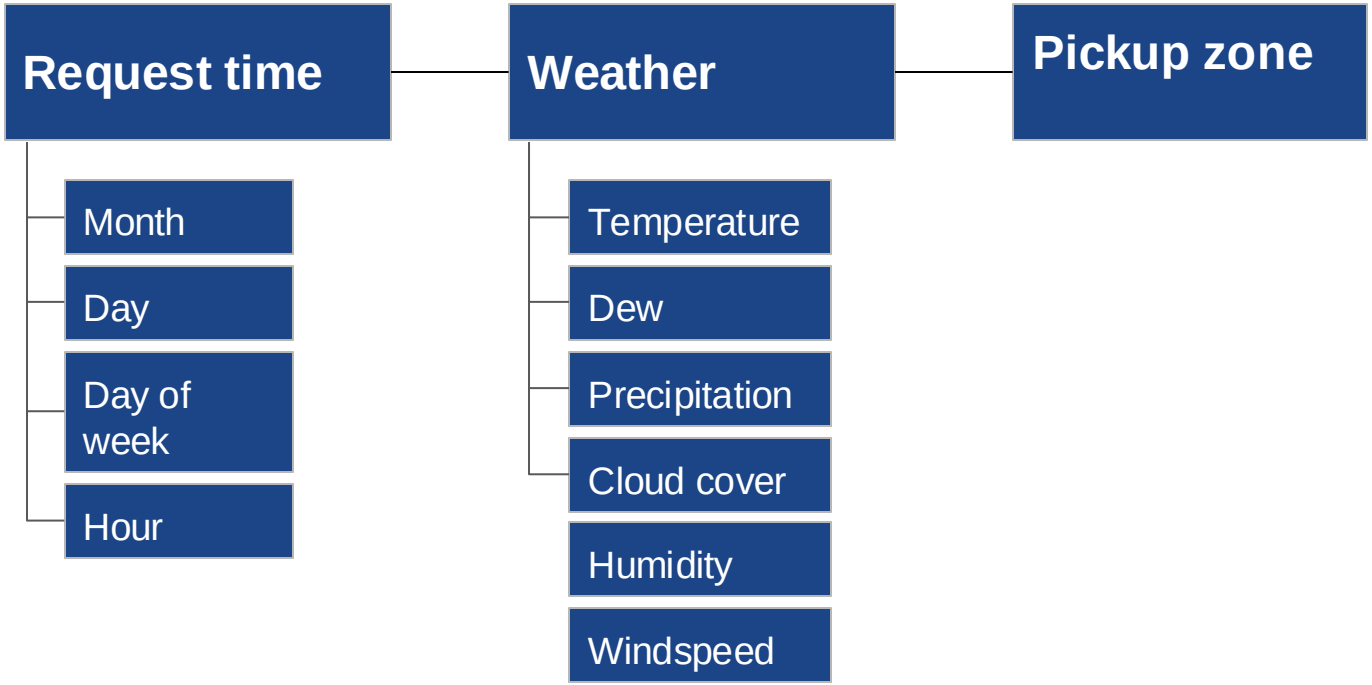
- measures to meet the surge in demand by incentivizing more drivers to work during peak hours

Increase customer satisfaction

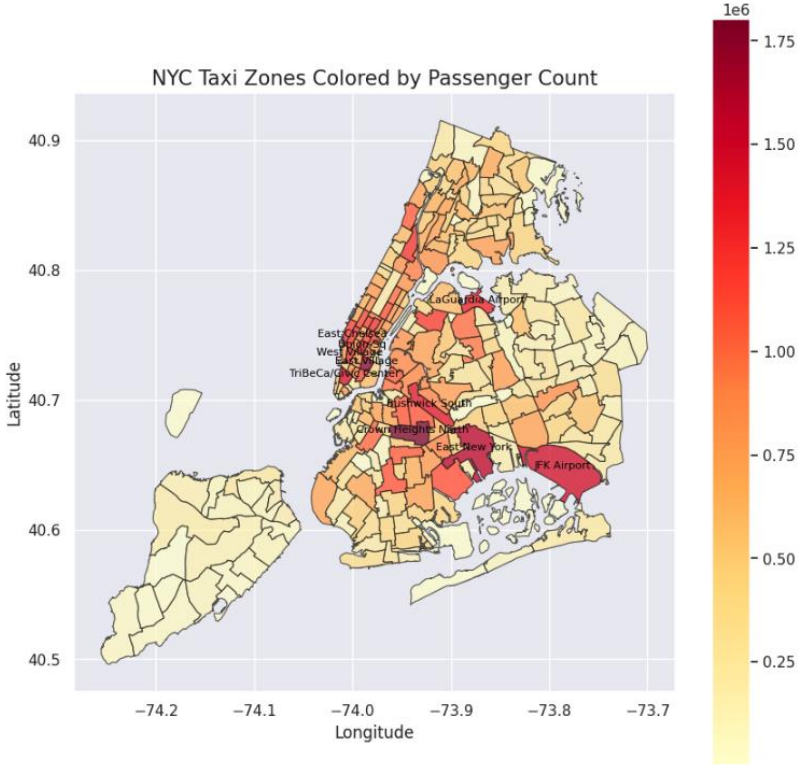
- Reduced wait times.

The Data

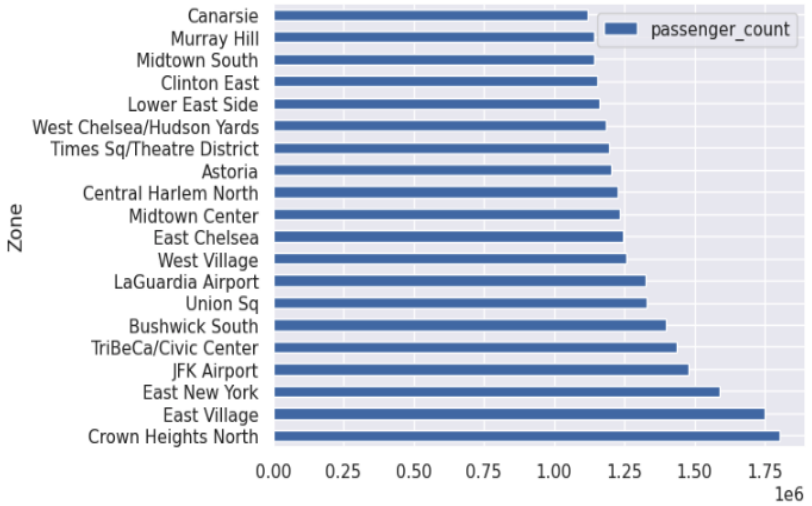
New York for-hire vehicles trip data for 2021 from Taxi and Limousine Commission



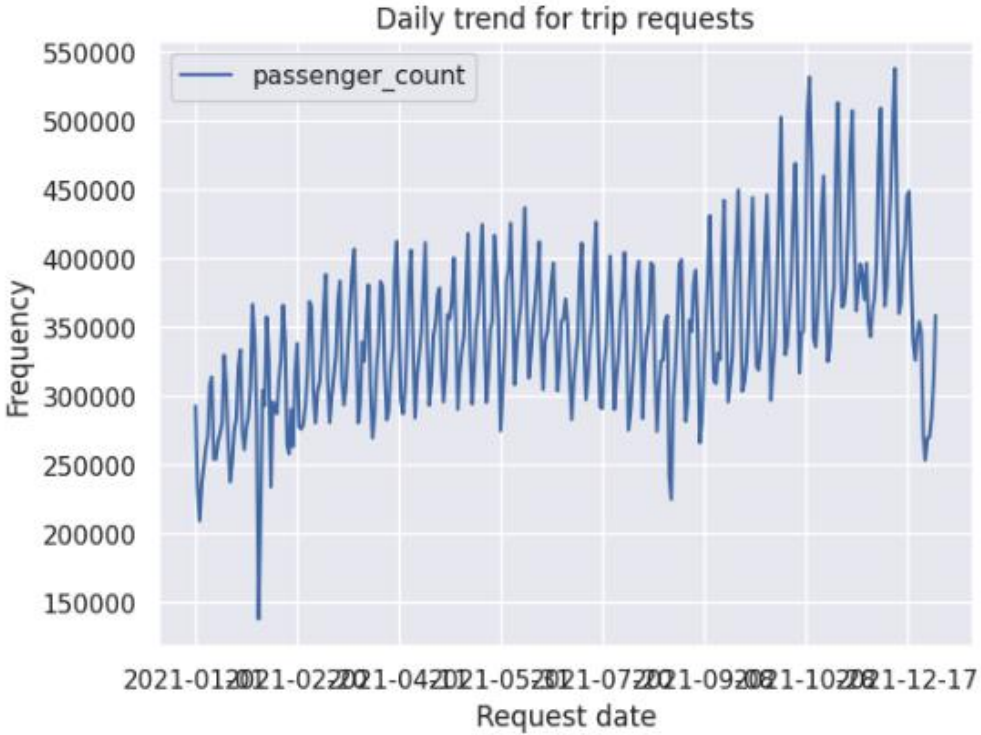
Data Exploration



Top 20 zones by number of requests

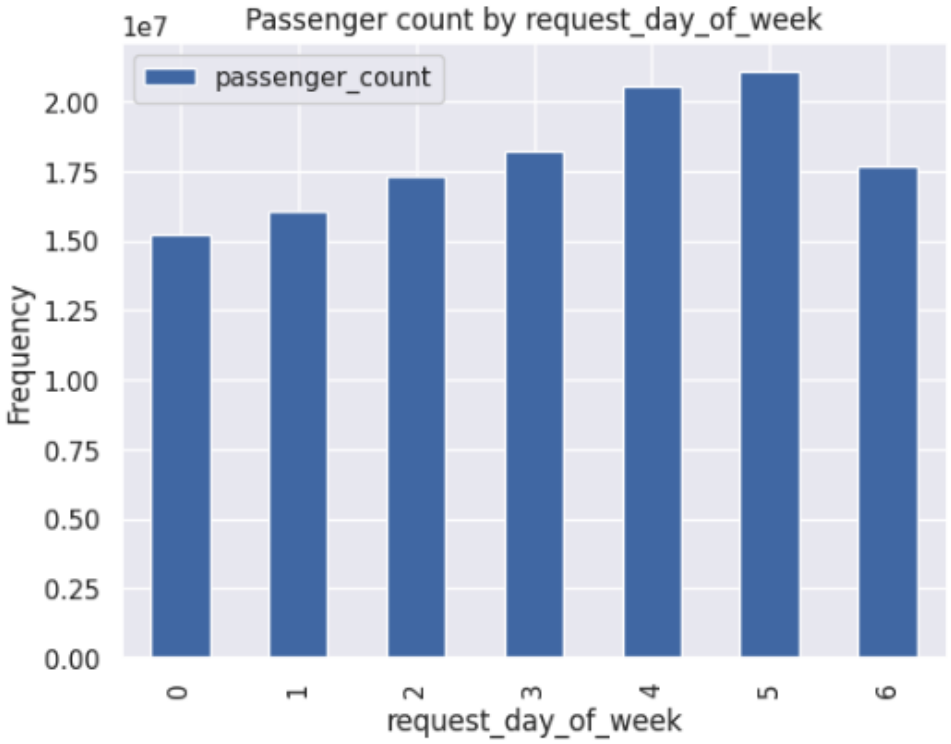


Annual trend



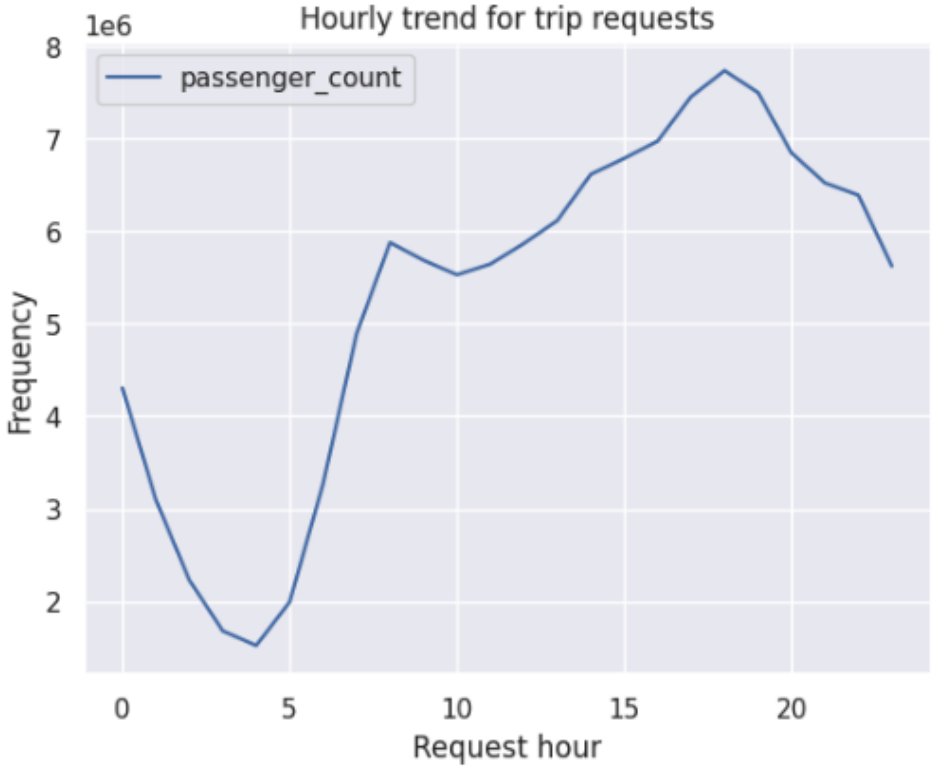
The number of requests is the lowest in the beginning of the year, the highest in the last 3 months of the year, and drops around Christmas.

Weekly trend



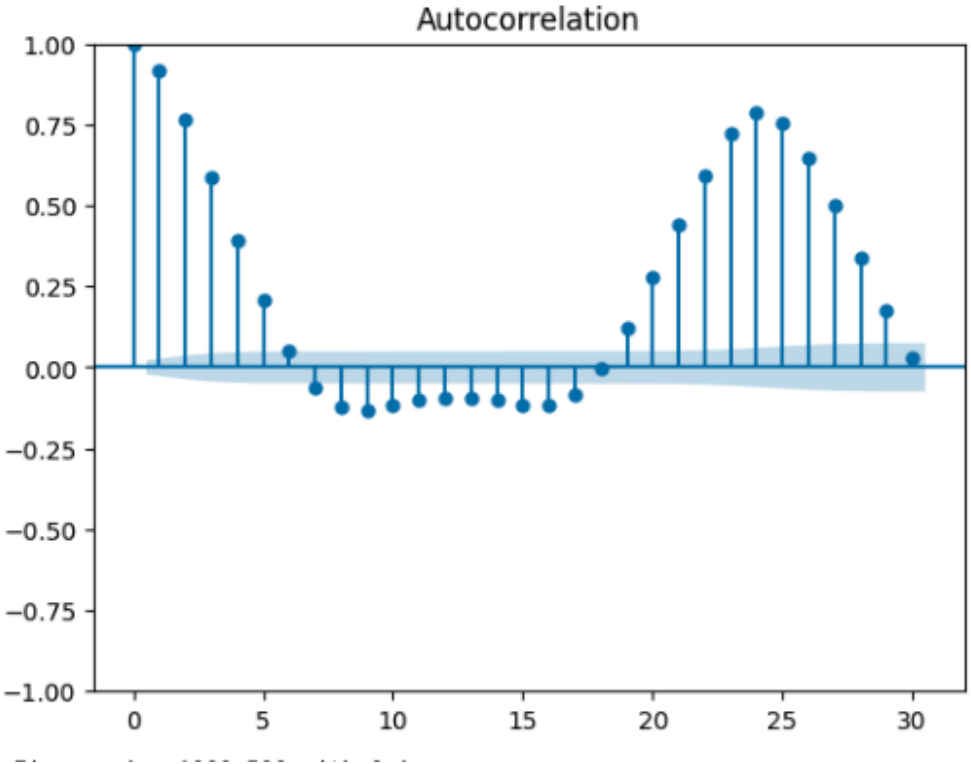
The number of requests is the lowest on Mondays, and the highest on Fridays and Saturdays.

Daily trend



The number of requests is the lowest at nighttime, and the highest after work hours – between 5 pm and 7 pm.

Autocorrelation



Number of requests in previous hours have impact on the current number of requests.

Modelling

Time series forecasting with seasonal terms & weather as exogenous features

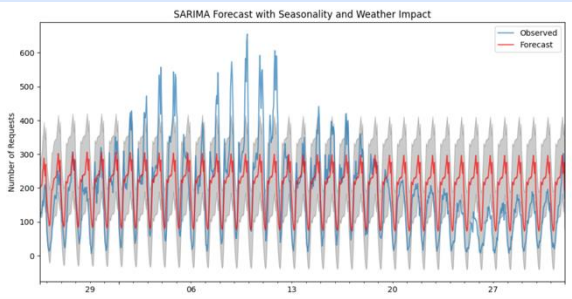
vs

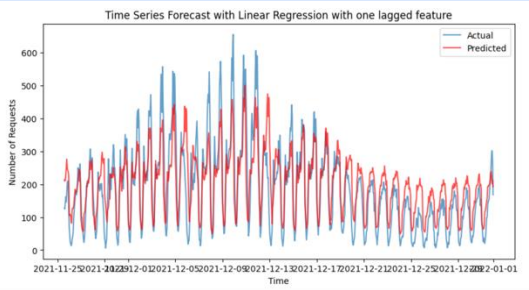
Regression models with seasonal and lagged features

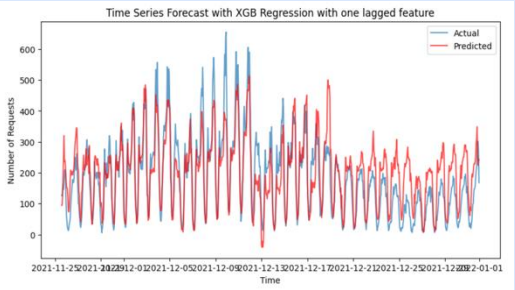
SARIMAX:
- Fourier terms (yearly, weekly and daily seasonality).
- Exogenous features (incl. weather)
- Splitting into train and test.

Linear Regression:
- Lagged features (trip requests at t -24 hours).
- Features scaling
- Splitting into train and test.

GXBoost Regression:
- Lagged features (trip requests at t -24 hours).
- Features scaling
- Splitting into train and test.







Initial Models Comparison

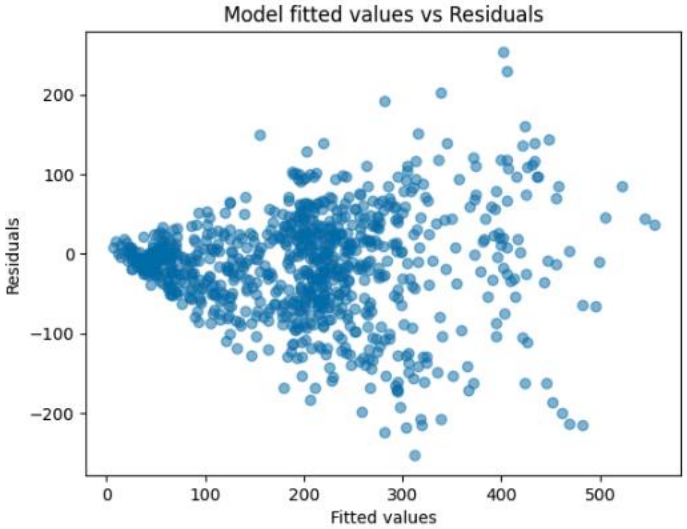
Model	MAE	RMSE	AIC	R-squared
SARIMAX	77.7	99.5	74,997	n/a
Linear Regression	55	72.7	10,007	0.67
XGBoost Regression	54	70	n/a	0.69

Hyperparameters optimization

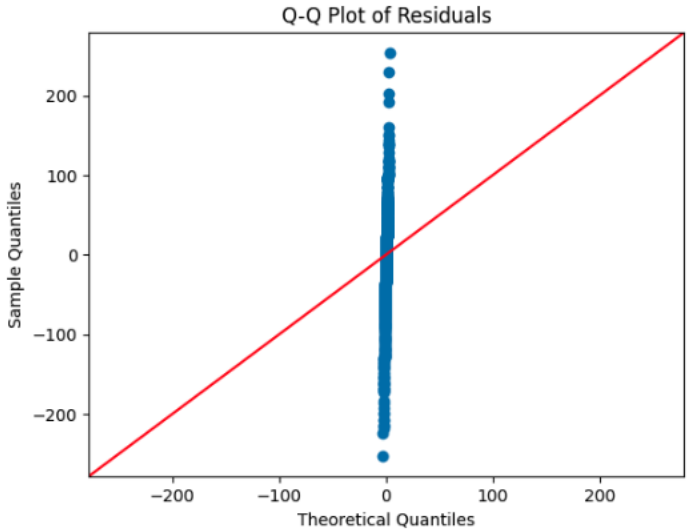
- Randomized Search with cross-fold validation.

- Optimal parameters:
  - colsample_bytree: 0.85,
  - gamma: 0.31,
  - learning_rate: 0.11,
  - max_depth: 9,
  - min_child_weight: 7,
  - n_estimators: 111,
  - reg_alpha: 0.33,
  - reg_lambda: 0.73,
  - subsample: 0.82

Model Diagnostics



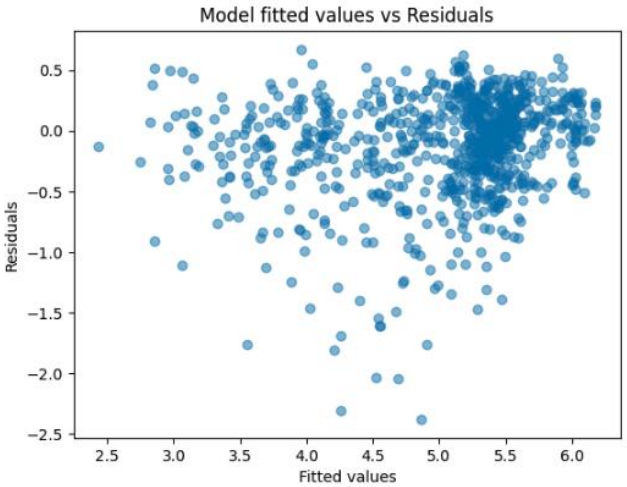
Residuals are not randomly scattered => heteroscedasticity problem.



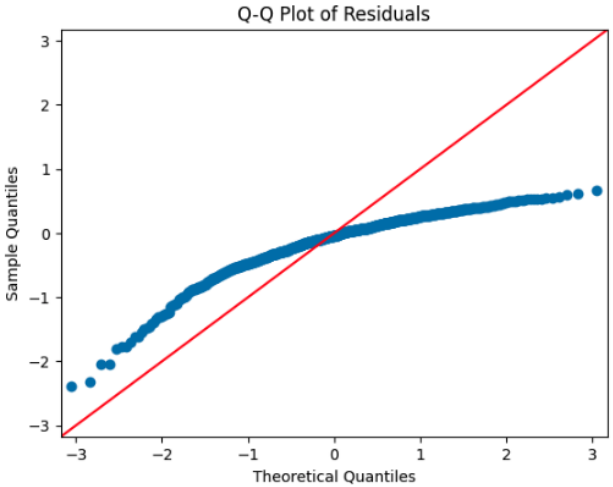
Residuals are not normally distributed.

Model Diagnostics

After log-transformation of a target feature:

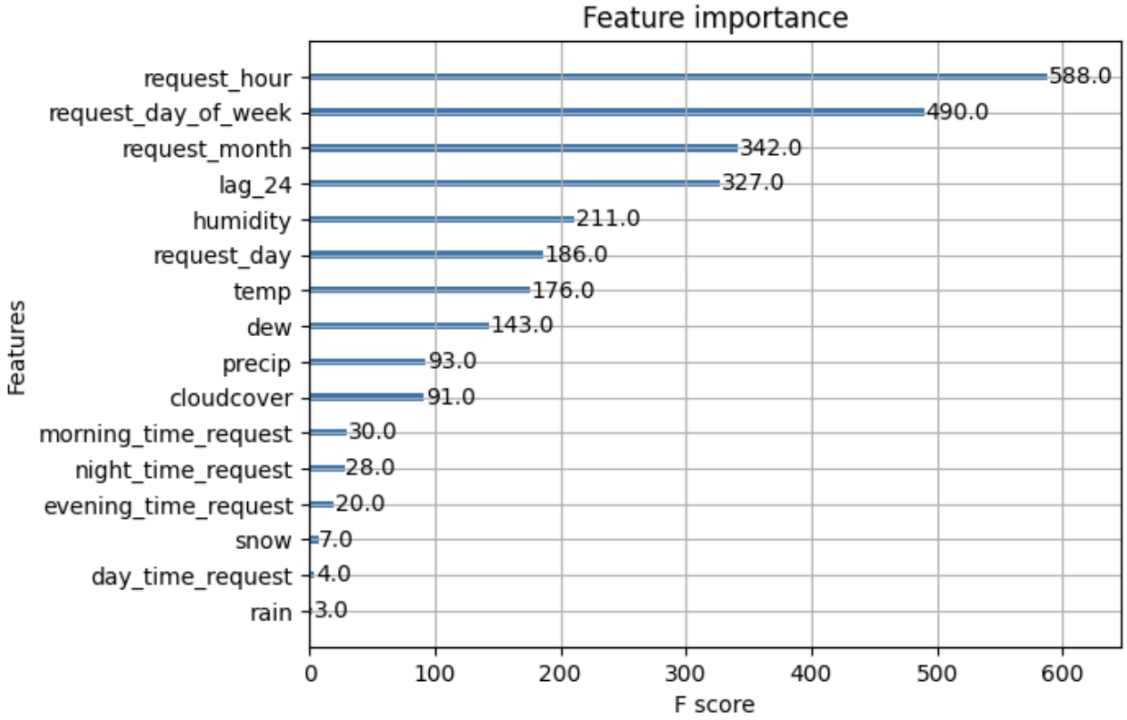


Residuals are scattered without a pattern.



Residuals are closer to being normally distributed.

Final Results



R-squared: 0.76
MAE: 0.3
RMSE: 0.44