Rahnema College
Machine Learning Bootcamp

Demand Prediction

Shirjeh Team

Mentor:
Amirhossein Shahdaei

August 2023

Content

Project Statement

- Introduction
- Main objective

Data Exploration

- Data structure
- Visualization

Approach

- Demand
- Time series
- Evaluation metrics

Phases

- Daily prediction
- 3-hour interval prediction
- Pair locations prediction

Conclusion

- Achievements

Project Statement

Problem introduction

- Marketplace forecasting (supply and demand)
- Directing drivers to high-demand areas
- Increasing the number of trips and earnings

Main objective

- Predicting demand
- Analyzing various factors influence demand prediction

Data Exploration

Data structure

- First four month of 2023
- Data frame shape :

  (12672629, 19)
- Important columns :

  pick-up and drop-off locations

  pick-up and drop-off date/time

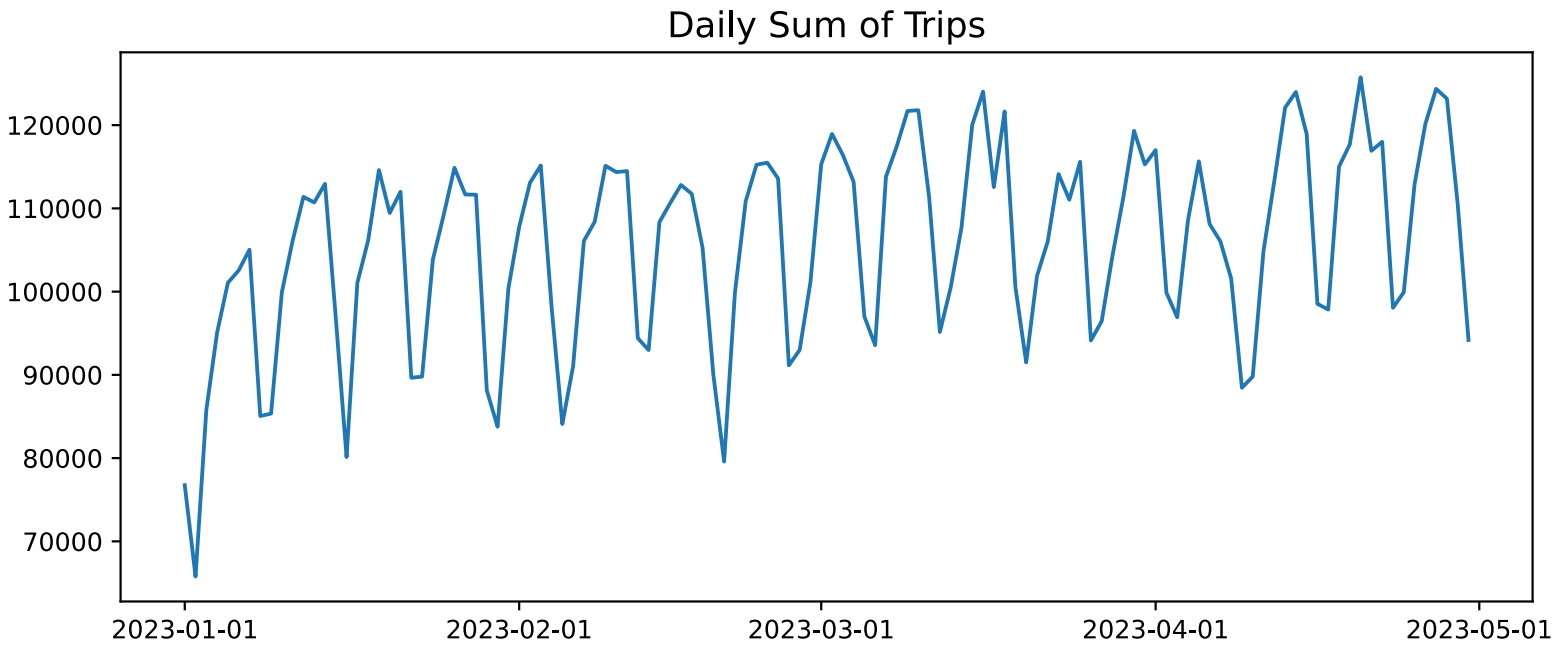
	tpep_pickup_datetime	PULocationID	DOLocationID	PU_date
0	2023-01-01 00:32:10	161	141	2023-01-01
1	2023-01-01 00:55:08	43	237	2023-01-01
2	2023-01-01 00:25:04	48	238	2023-01-01
3	2023-01-01 00:03:48	138	7	2023-01-01
4	2023-01-01 00:10:29	107	79	2023-01-01

Data Exploration

Visualization

The daily sum of trips for the first four month

- Positive trend
- Weekly seasonality

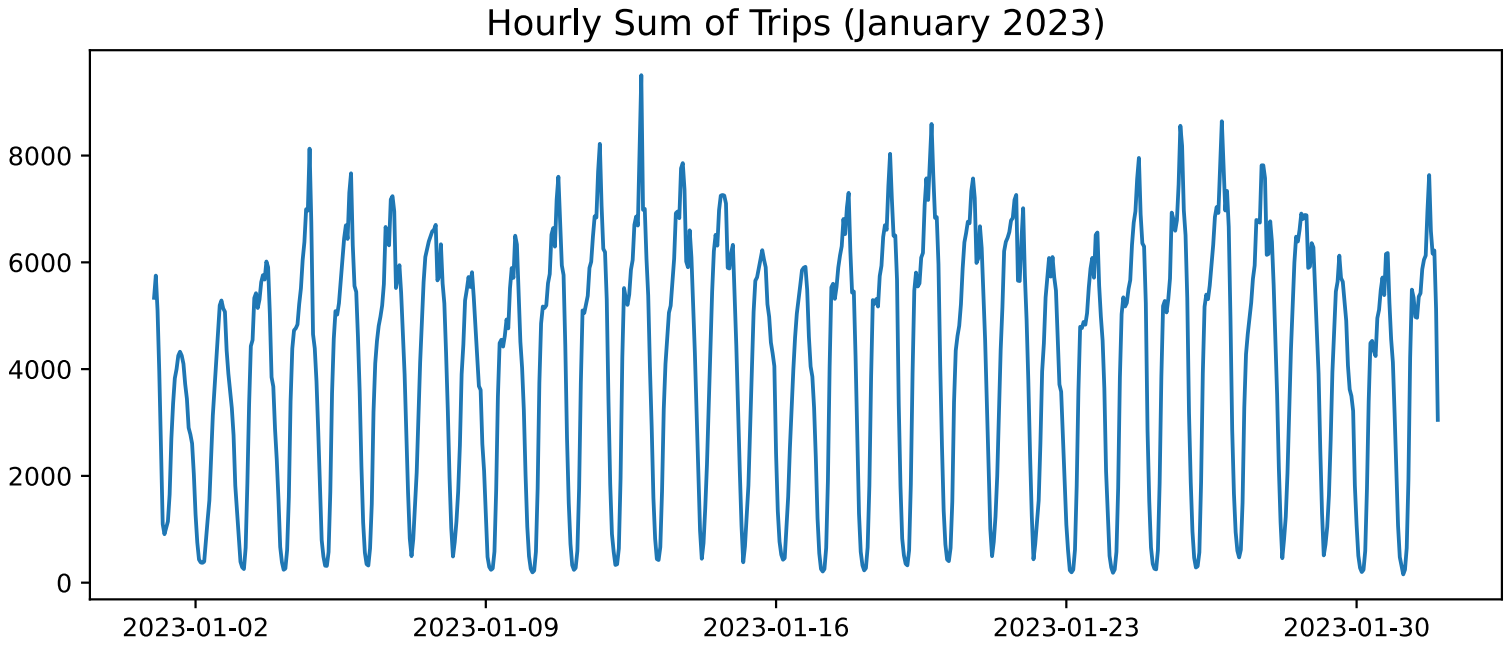


Data Exploration

Visualization

The hourly sum of trips in a given month (January)

- Daily and weekly seasonality



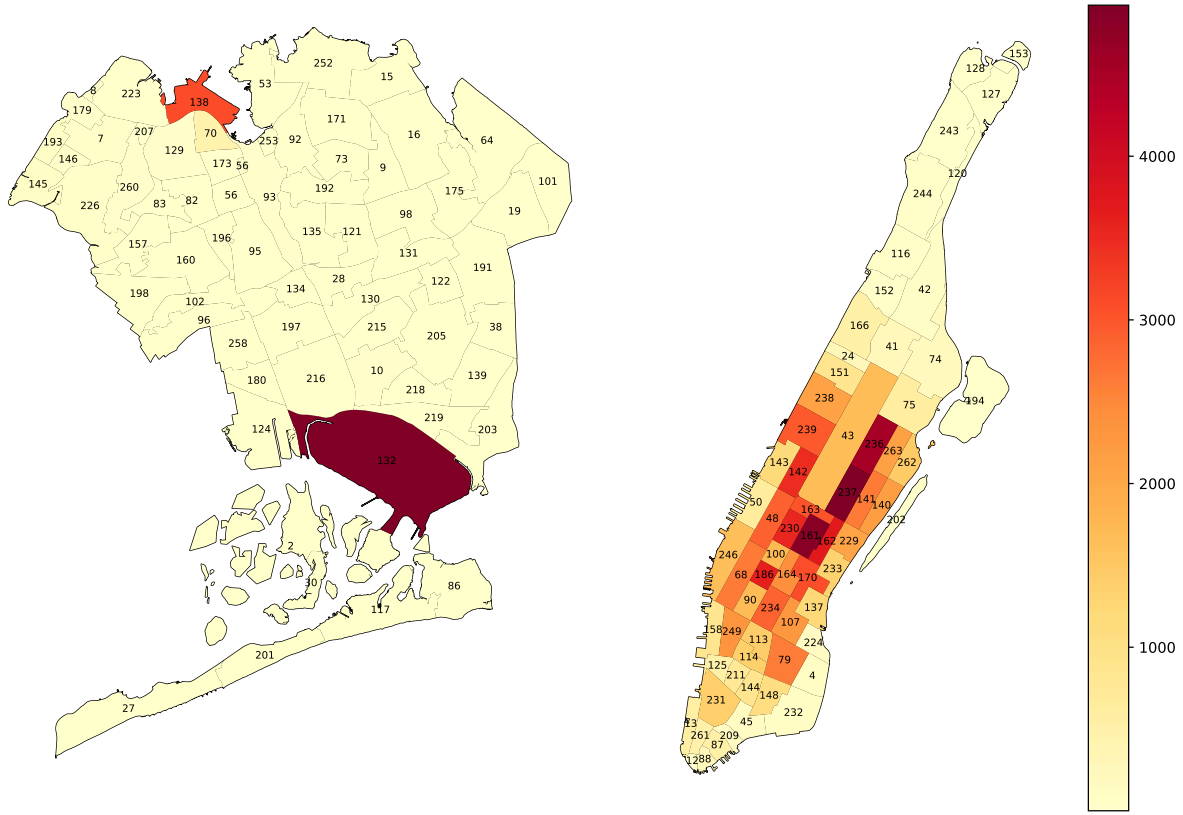
Approach

Demand

Number of rides in a specific time interval for each location

- The most high demanded boroughs

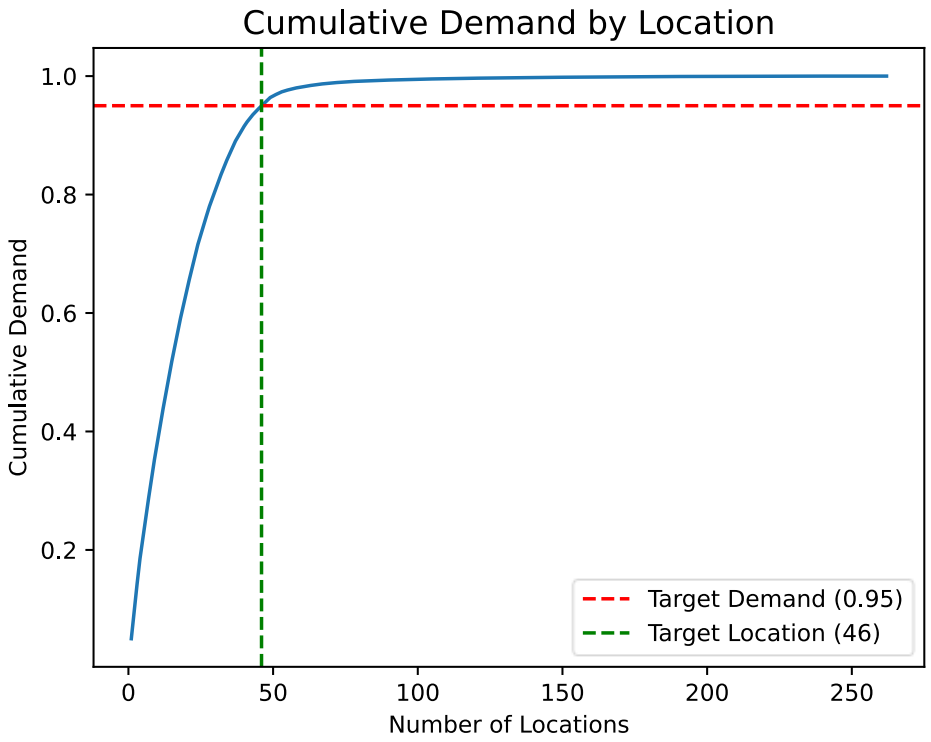
  (Manhattan and Queens heatmap)

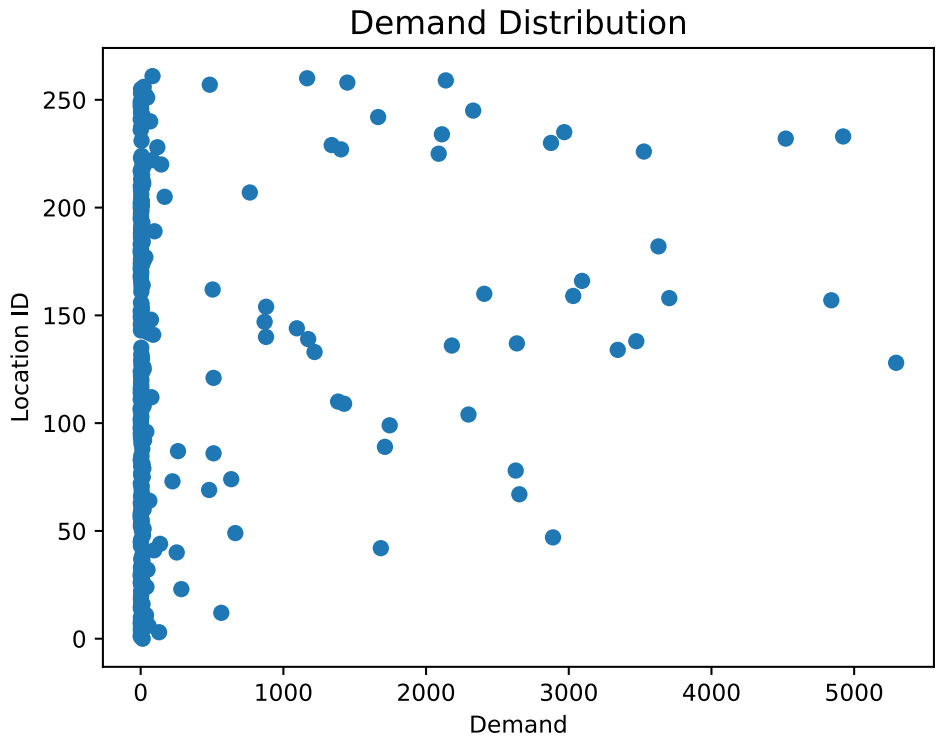


Approach

Demand

95% of demand is in just about 20% of locations





Approach

Time series features

- Feature engineering :

  There is no concept of input and output in time series problems

  Transforming time series problems into supervised learning problems

- Three classes of features :

  Date time features

  Lag features

  Window features (Summary features)

Approach

Evaluation metrics

- Mean Absolute Percentage Error (MAPE)

  For high demand locations

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_{exp,i} - y_{pred,i}}{y_{exp,i}}\right|$$

- Mean Absolute Error (MAE)

  For low demand locations

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{exp,i} - y_{pred,i}\right|$$

Phases

Phase 1

Daily Demand Prediction

Models

- Classical & Statistical
  - ARIMA
  - Prophet
- Machine Learning
  - Ridge regression
  - XGBoost
  - Random forest

Phases

Phase 1 (daily prediction)

Features

- Date time features
  - Day of week
  - Day of month
- Lag features
  - Previous 1-14 day demand
- Window features
  - Max demand of previous 1-2 week (each location)
  - Max/min/mean demand of previous week (grouped locations)
- Model-driven input (for XGBoost and Random forest)
  - Ridge model predictions

Phases

Phase 1 (daily prediction)

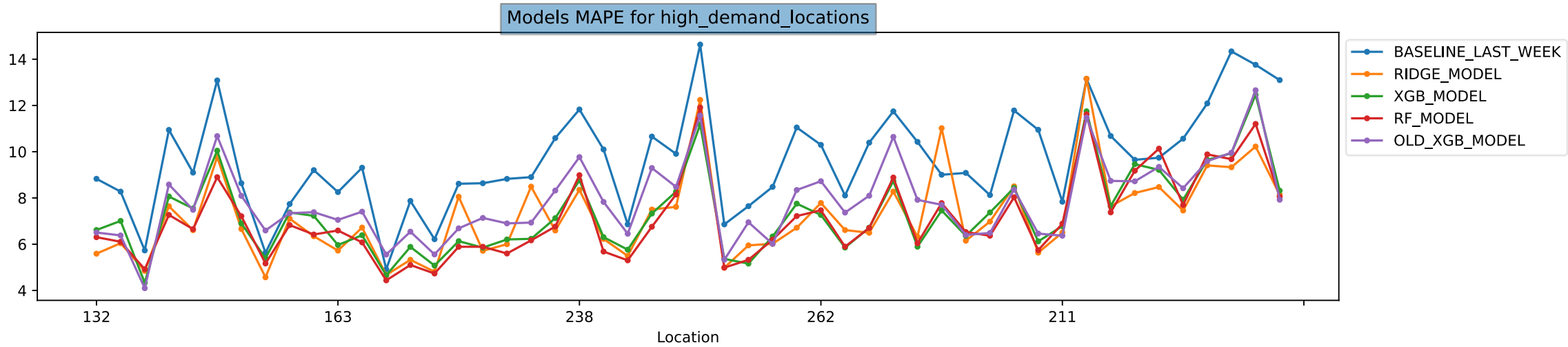
Table of results

	High Demand Locations	Low Demand Locations (Mean of Demand = 17.80)
Models	MAPE	MAE
Baseline (Last week demand)	9.64	4.59
Old XGBoost model	7.85	3.88
Ridge regression model	7.21	3.95
XGBoost model	7.31	3.78
Random forest model	7.09	3.58

Phases

Phase 1 (daily prediction)

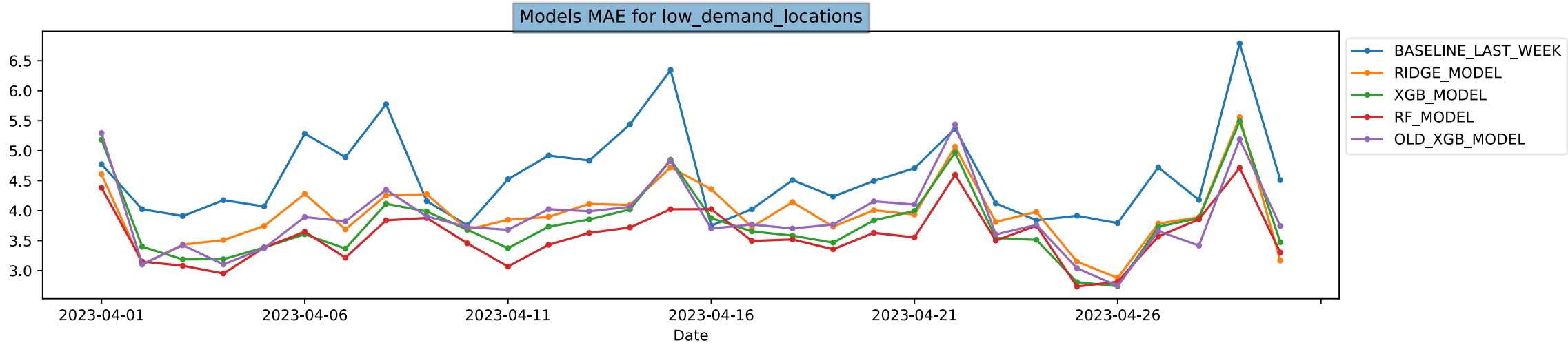
Results (MAPE high demand)



Phases

Phase 1 (daily prediction)

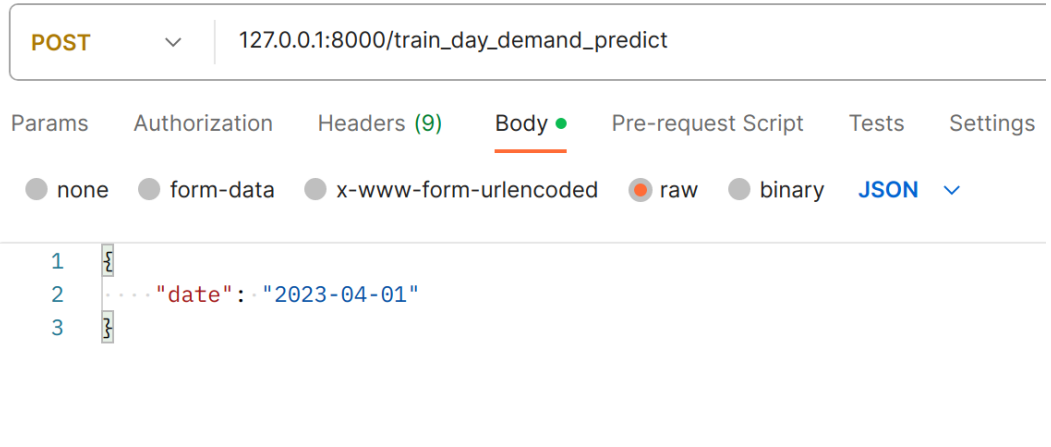
Results (MAE low demand)

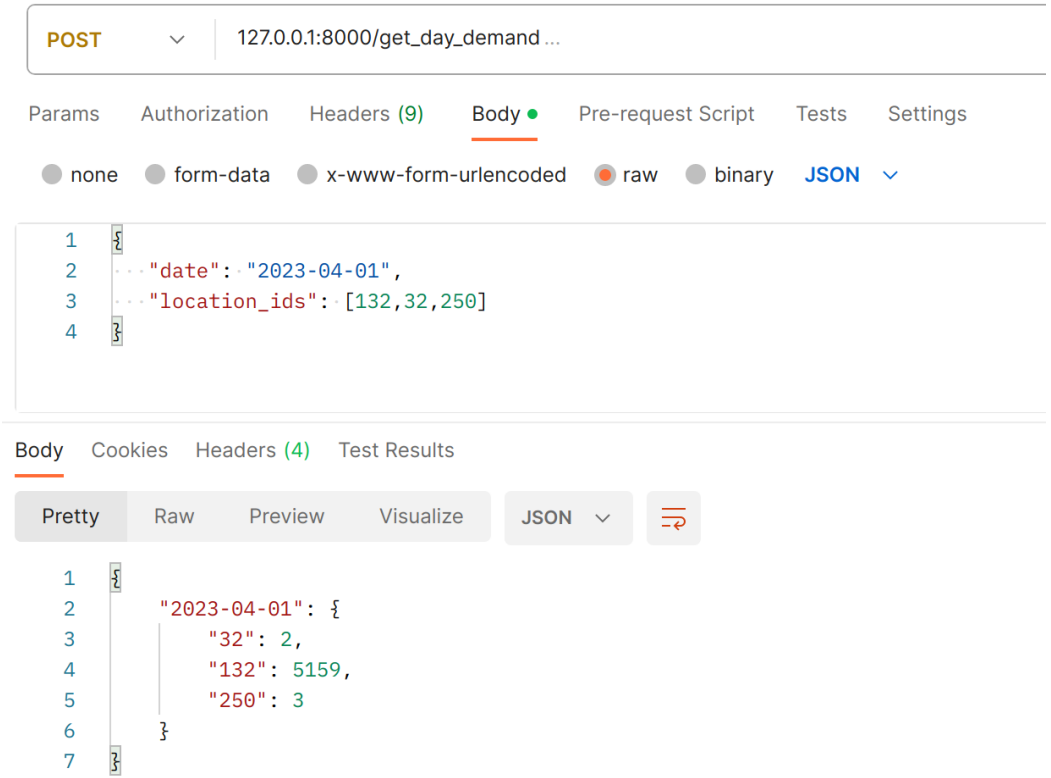


Phases

Phase 1 (daily prediction)

API



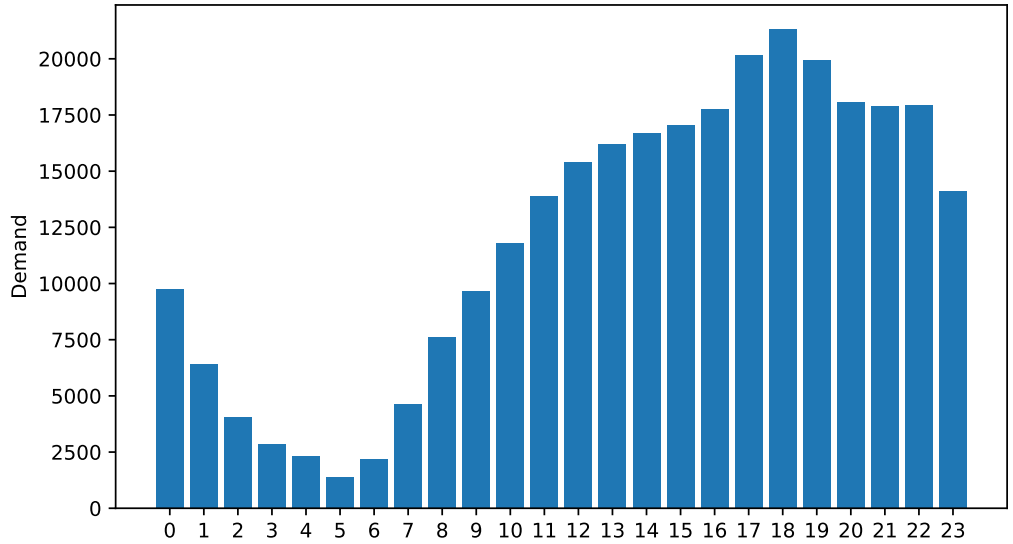


Phases

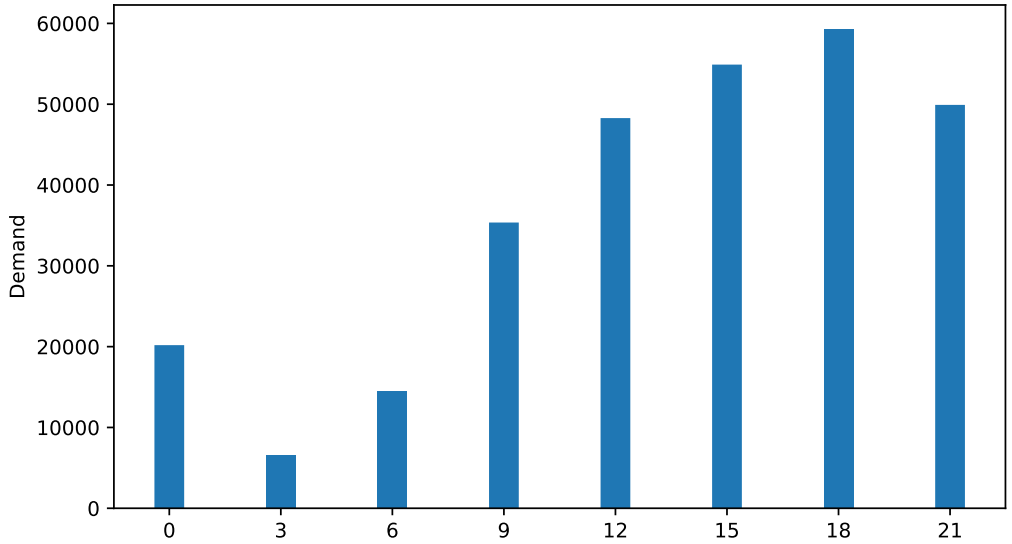
Phase 2

3-hour interval Demand Prediction

Demand for each 1-hour interval



Demand for each 3-hour interval



Phases

Phase 2 (3-hour interval prediction)

Features

- Date time features
  - Day of week
  - Day of month
- Window features
  - Max demand of previous week (each location)
  - Weighted moving average of previous day intervals
- Lag features
  - Previous 1-14 day demand
  - Previous 1-2 interval of previous day
- Model-driven input (for XGBoost and Random forest)
  - Ridge model predictions

Phases

Phase 2 (3-hour interval prediction)

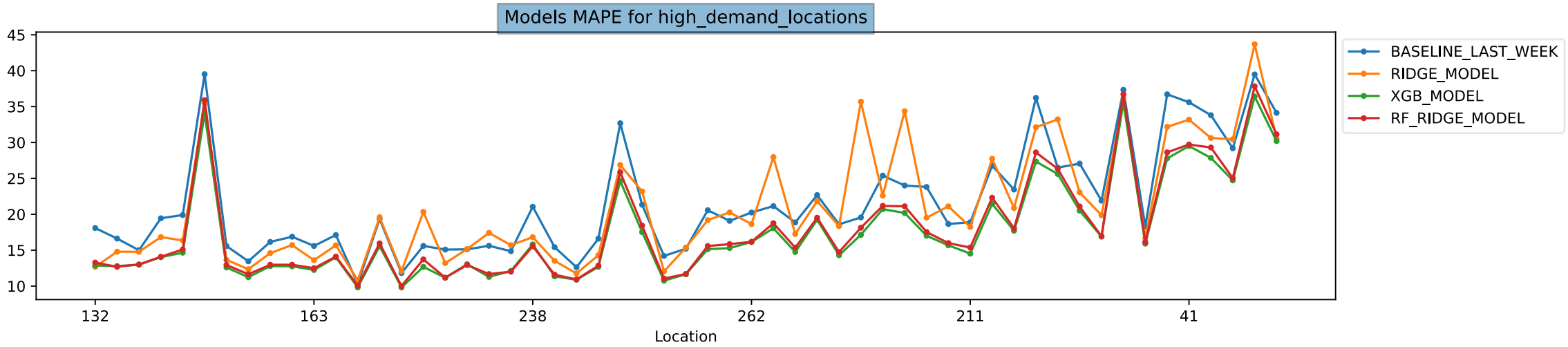
Table of results

	High Demand Locations	Medium Demand Locations (Mean of Demand = 17.80)
Models	MAPE	MAE
Baseline (Last week demand)	21.43	3.28
Ridge regression model	18.08	2.43
XGBoost model	17.18	2.56
Random forest model	17.65	2.48

Phases

Phase 2 (3-hour interval prediction)

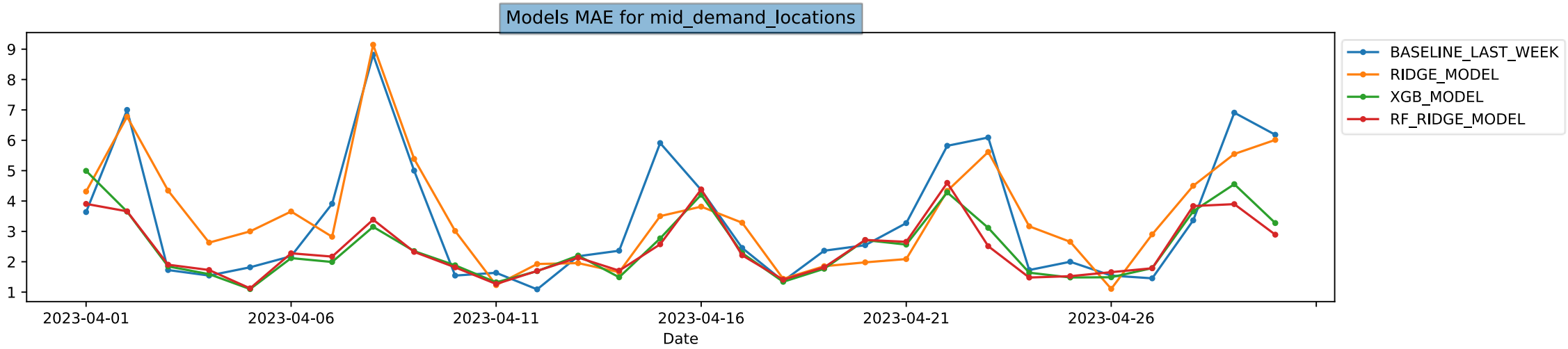
Results (MAPE high demand)



Phases

Phase 2 (3-hour interval prediction)

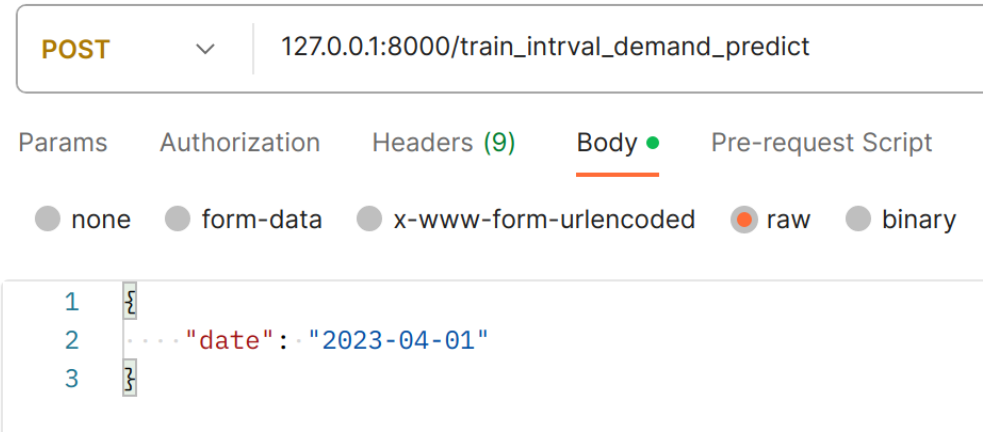
Results (MAE mid demand)

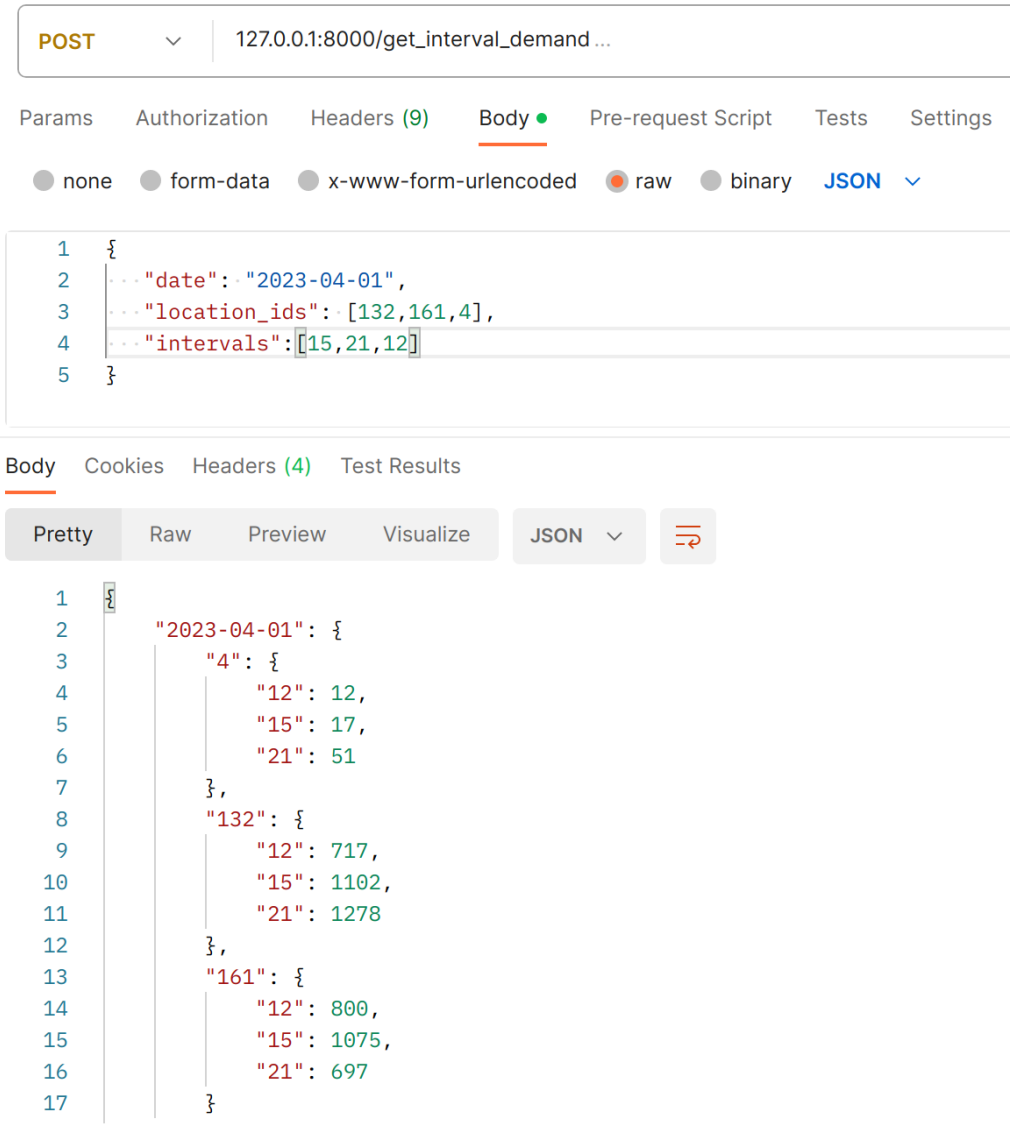


Phases

Phase 2 (3-hour interval prediction)

API



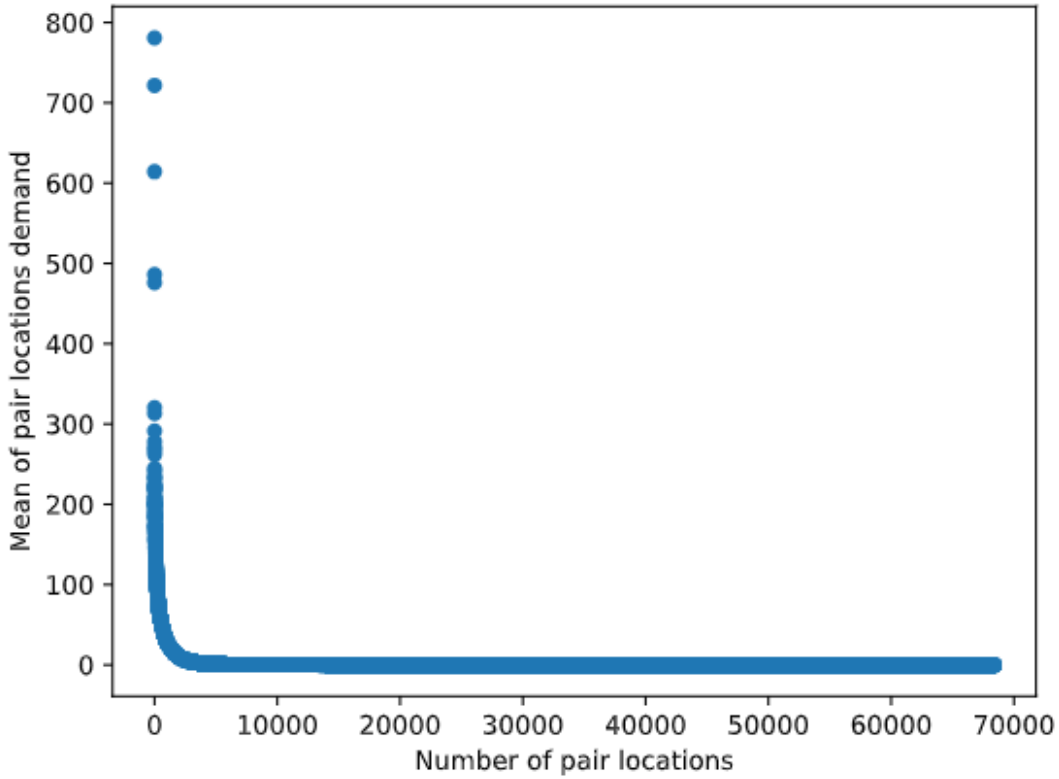


Phases

Phase 3

Pair locations prediction (pick-up and drop-off locations)

- Daily prediction
- Features (same as phase 1)



Phases

Phase 3 (pair locations prediction)

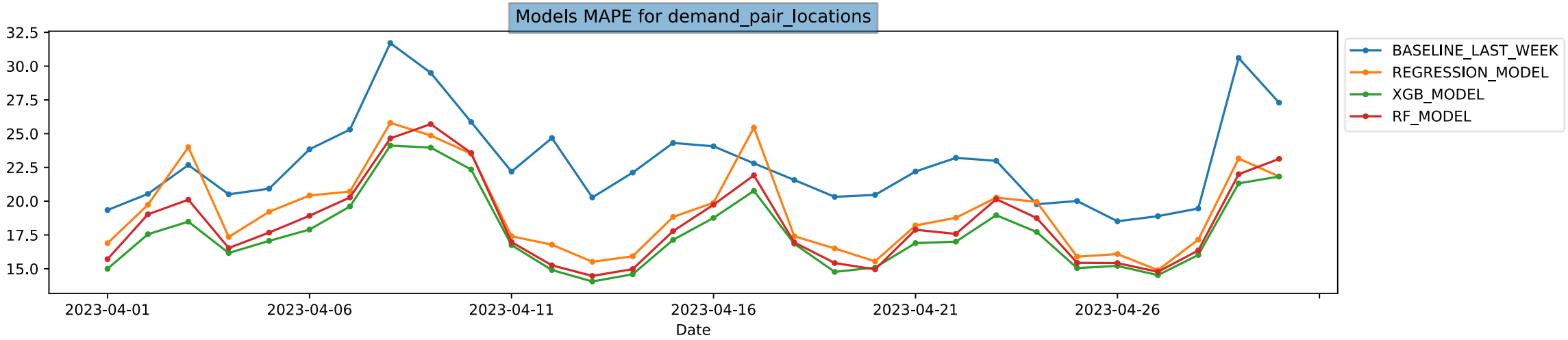
Table of results

	Demand for Selected Pair Locations
Models	MAPE
Baseline (Last week demand)	22.86
Ridge regression model	19.26
XGBoost model	17.68
Random forest model	18.40

Phases

Phase 3 (pair locations prediction)

Results (MAPE)



Conclusion

Achievements

Phase 1 (Daily prediction)	
Best Model	Random forest
MAPE (high)	7.09 (Improvement = 26%)
MAE (low)	3.58 (Improvement = 22%)

Phase 2 (3-hour interval prediction)	
Best Model	XGBoost
MAPE (high)	17.18 (Improvement = 20%)
MAE (mid)	2.56 (Improvement = 22%)

Phase 3 (Pair locations prediction)	
Best Model	XGBoost
MAPE (high)	17.68 (Improvement = 23%)

Thanks for your attention :)