Context:
There are a lot of customers on the InvestorAI platform. Whenever they login in the app and view anything, the pings are gotten from their mobile phone indicating that they are using the app.
I have been provided with 3 weeks of training data and 1 week of test data.
The dataset has three files. The data in the first two films can be used for training the model and the third file contains test data.

- customers.csv: This file contains customers profile data
- pings.csv: This file contains: This file contains the customer pings
- test.csv: This file has the test data

Problem statement formation:

I want to predict how many hours the customer will be online / using the app on a given day. So the test data contains customer id, and date (during the test data period). The test data also contains the actual online hours, which is what the model should predict.

Also it is possible to do customer segmentation to predict using the app by a certain type of customer.

Scope of solution space:

So we are going to handle time related data for each customer.This could be a multivariate time series problem with multiple time series. So modeling strategy and model building need to be done to capture patterns.

Criteria for success:

We will be looking at Root Mean Squared Error or RMSE to see how good the model is. It's value should be lower than the baseline score. The baseline score was set by assuming the predictions for t(n) date are t(n-1) (i.e,today predictions are yesterday values).

Data source:

Kaggle dataset "Users active time prediction"
https://www.kaggle.com/datasets/bhuvanchennoju/mobile-usage-time-prediction?select=test.csv

Stakeholders:

Don't have any.

Constraints:

Which model should I choose? Should I train models with simple ARIMA or LSTM models?