The Impact of Covid-19 on Air Traffic:

Spatiotemporal/Time Series Forecasting and Benchmarking

Bo Yan, Yuan Hu, & Adelle Driker
(Group 6)

Table of Contents

Team Introductions

Professors Rose Yu & Ilkay Altintas de Callafon

- Advisors/Mentors

Bo Yan

- Record Keeper
- Software/ML/DL Engineer

Yuan Hu

- Budget Manager
- Data Engineer/Solution Architect

Adelle Driker

- Project Coordinator/Manager
- Data/Business Analyst

Table of Contents

1. Team
2. Problem & Data Definition
3. Accuracy/Significance of Results
4. Main Insights
5. Experimental Trials
6. Next steps

Recap - Problem Definition

- Many global industries have been affected by the COVID 19 pandemic, the airline industry being one of the most heavily hit
    - E.g. London’s Heathrow Airport reported a 97% decrease in passenger numbers between May 2019 and May 2020
- Creates uncertainty for both passengers and airline companies, especially due to the multiple waves of virus mutations
    - How should airlines plan future flights? When should passengers schedule their travels?

In other words, given a country’s COVID situation, how should an airline/passengers plan ahead?

Recap - Data Definition

OpenSky Flight Data (Jan 2019 - Present)

- As-is, new files released on a monthly basis, multiple entries with missing data
- CallSign*, Number, ICAO24, Reg, TypeCode, Origin, Dest, First/Last Seen DT, Lat/Long/Alt of Origin & Dest

Johns Hopkins COVID19 Data (Jan 2020 - Present)

- Updates for historical inaccuracies, new files released daily
- Province/State, Country/Region**, Lat, Long, Dates

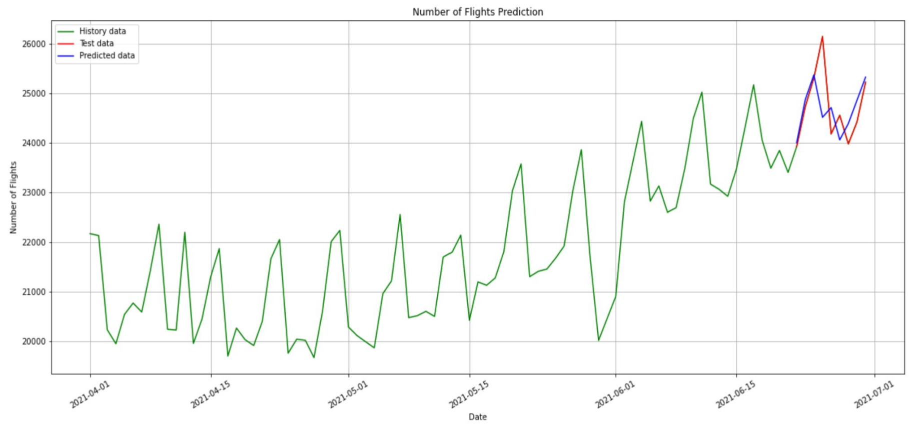
Airline Code and Country Mapping

- Sourced from IATA and ICAO, mostly complete
- Airline Name, IATA Designator, 3-Digit Code, ICAO Designator*, Country**
- Will be used to link together the OpenSky Flight and COVID19 datasets

Table of Contents

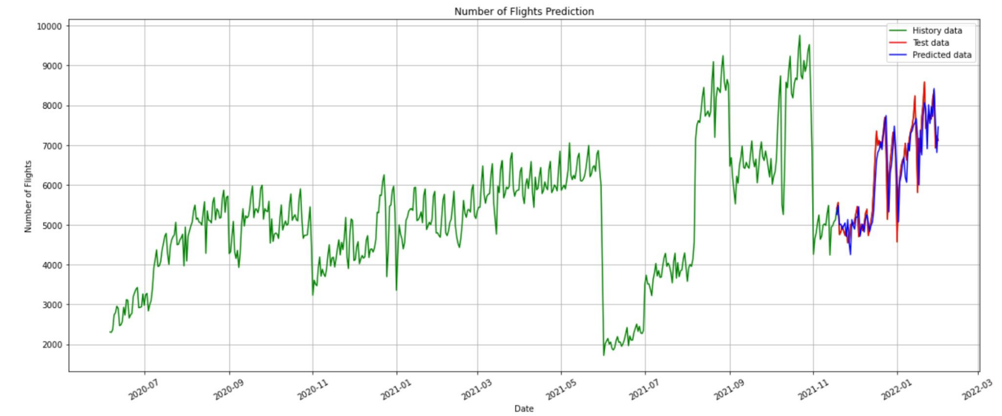
Accuracy/Significance of Results

—— Before vs. After: AR Model



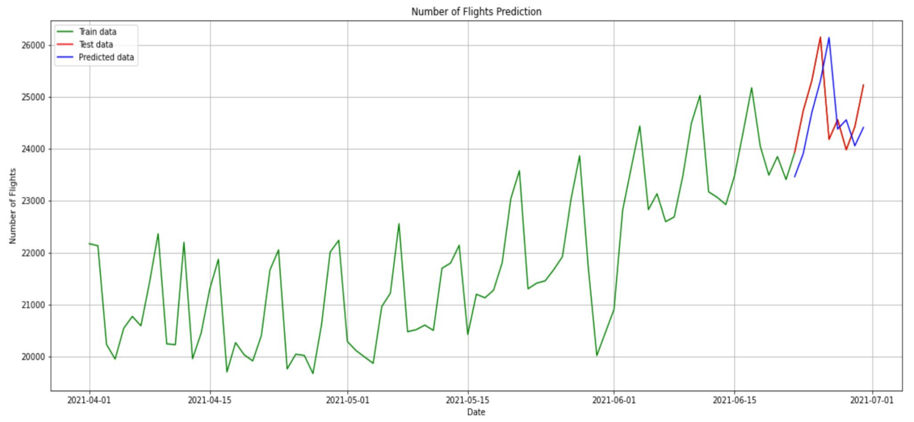
	Before	After
r^2	0.15	0.81
RMSE	633.40	509.27
MAE	433.98	381.05

- r^2: 81% of the data fit the AR model(better fit)
- Strong effect size, 81% of the variance of the dependent variable can be explained by the variance of the independent variable.
- The lower value of MAE, and RMSE implies higher accuracy of AR model.



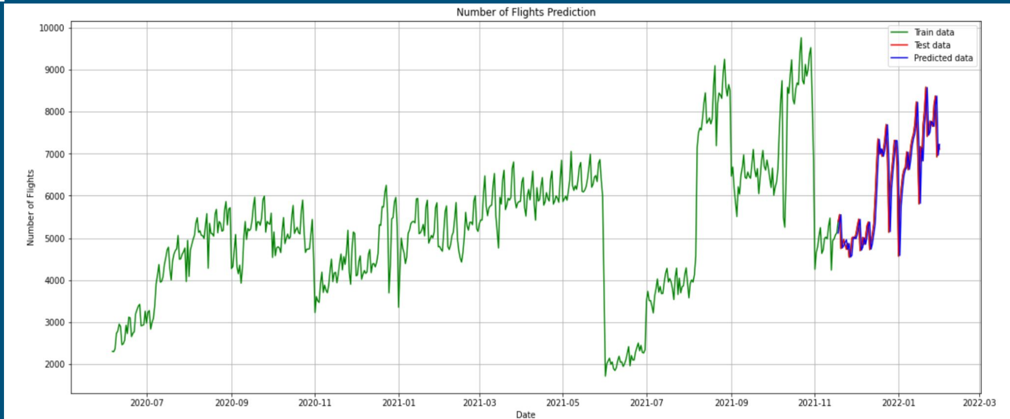
Accuracy/Significance of Results

—— Before vs. After: ARIMA Model



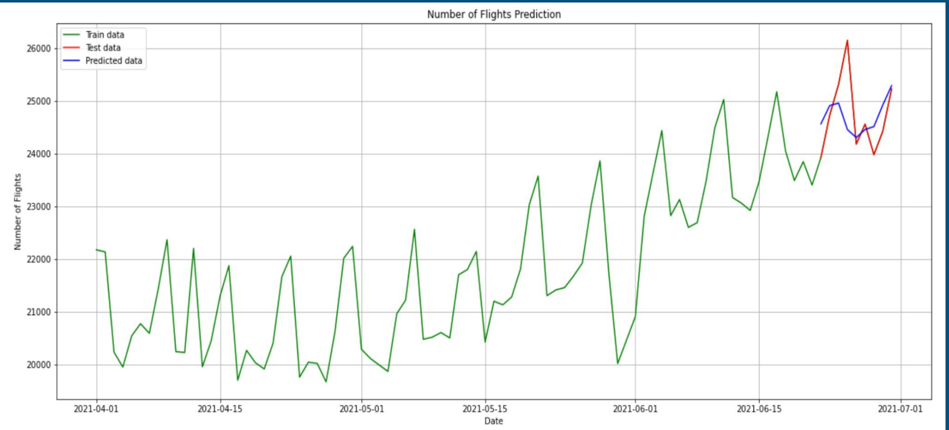
	Before	After
r^2	-0.65	0.73
RMSE	882.81	603.36
MAE	739.61	429.24

- r^2: 73% of the data fit the ARIMA model (better fit)
- Strong effect size, 73% of the variance of the dependent variable can be explained by the variance of the independent variable.
- The lower value of MAE, and RMSE implies higher accuracy of ARIMA model.



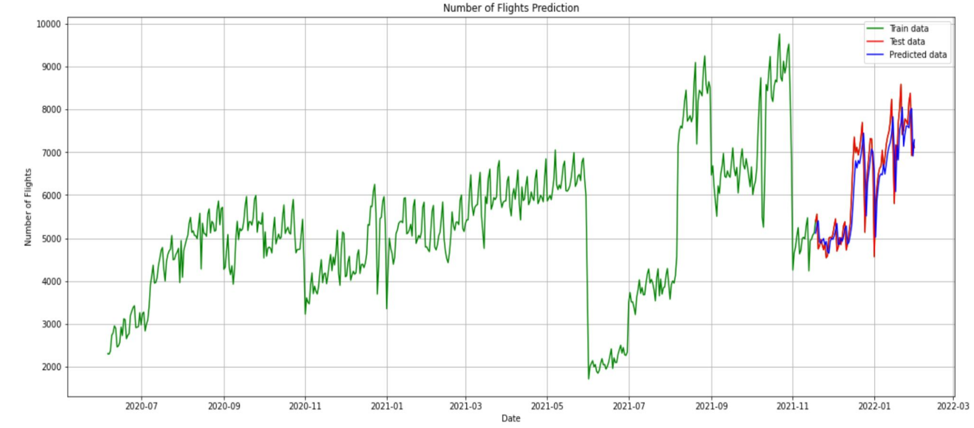
Accuracy/Significance of Results

—— Before vs. After: LSTM



	Before	After
r^2	0.06	0.76
RMSE	667.30	565.07
MAE	466.12	433.86

- r^2: 76% of the data fit the LSTM model (better fit)
- Strong effect size, 76% of the variance of the dependent variable can be explained by the variance of the independent variable.
- The lower value of MAE, and RMSE implies higher accuracy of LSTM model.



Accuracy/Significance of Results

—— Summary

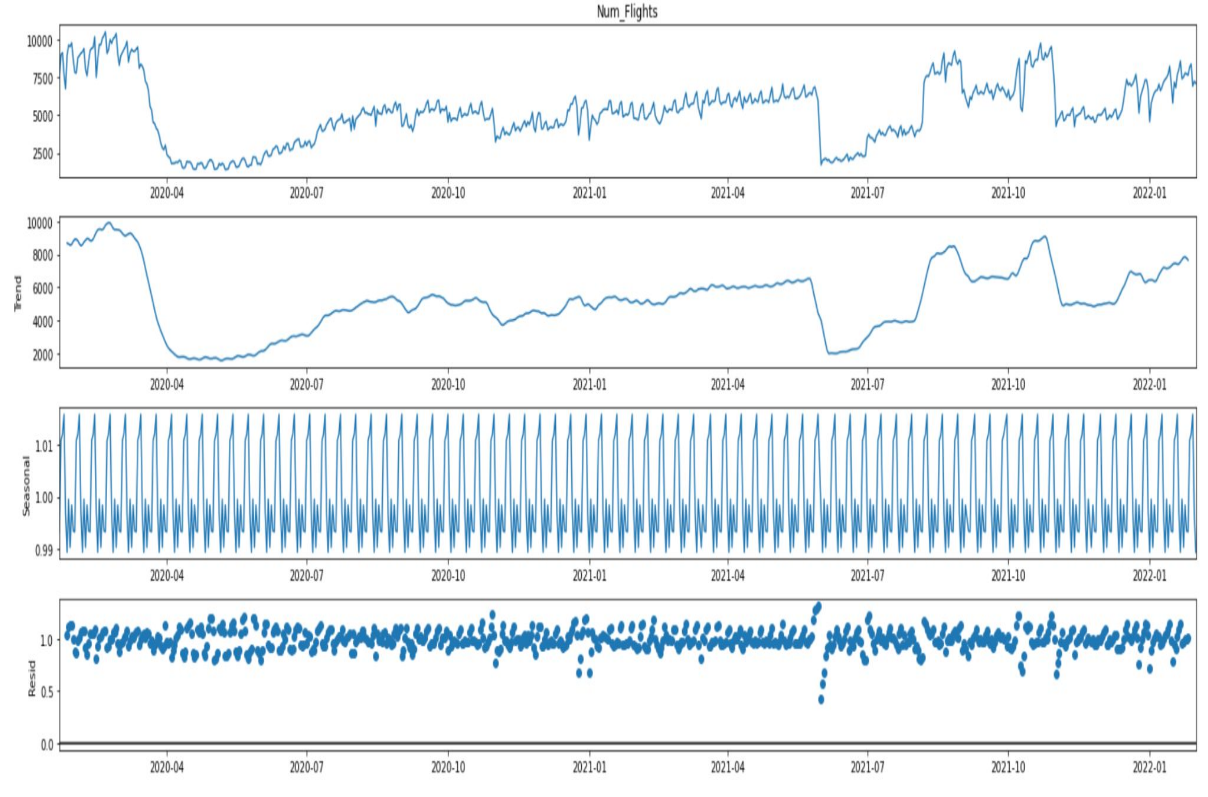
- **Trend**
  - Before: General trend of predictions matches overall trend of actuals
  - After: Much higher match between predictions and actuals
- **Accuracy**
  - Incorporation of full dataset helped improve accuracy in both Baseline and Deep Learning Models
  - Before: Smaller sample size resulted in large error
  - After: Larger sample size resulted in less error
- **Robustness**
  - Before: Although COVID is a transient event, must take historical seasonality into account
  - After: Although irregularities occurred around Jun, Aug, and Nov, the model still performed well

Table of Contents

Main Insights

——Decomposed Dataset

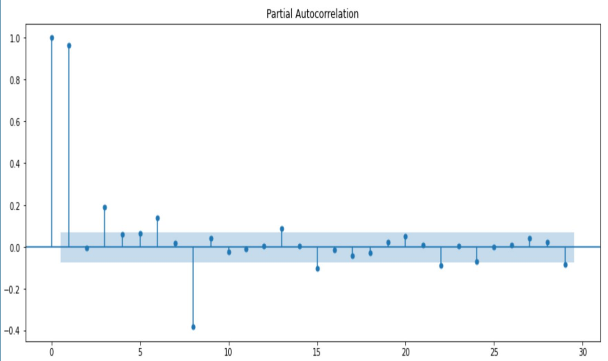
- Seasonality
  - 10 days cycle
- Residuals
  - high variability
- Trend
  - March 2020 to April 2020: Sharp drop (CDC : facemasks and social distance regulations on late Feb)
  - May 2021 to July 2021: Sharp drop (Vaccine come out at early 2021)
  - Nov. 2021 to Dec 2021: Sharp drop (Omicron breakout)

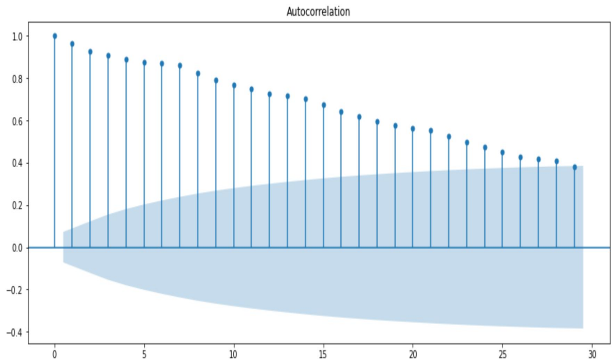


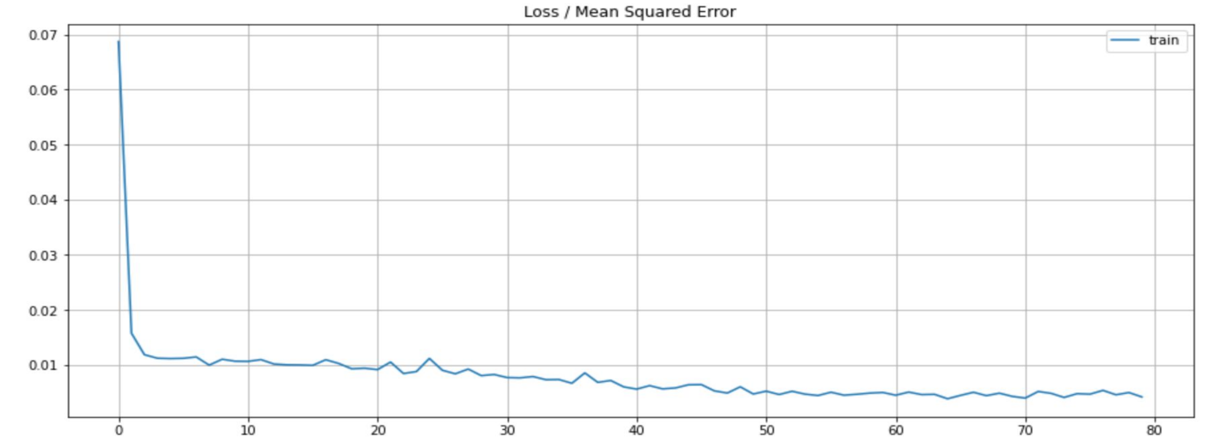
Main Insights

——Performance Tuning

- Optimal parameters
    - Autocorrelation: sharp cut-off
    - Partial Autocorrelation: Gradual decrease
    - Best epoch for LSTM: 60

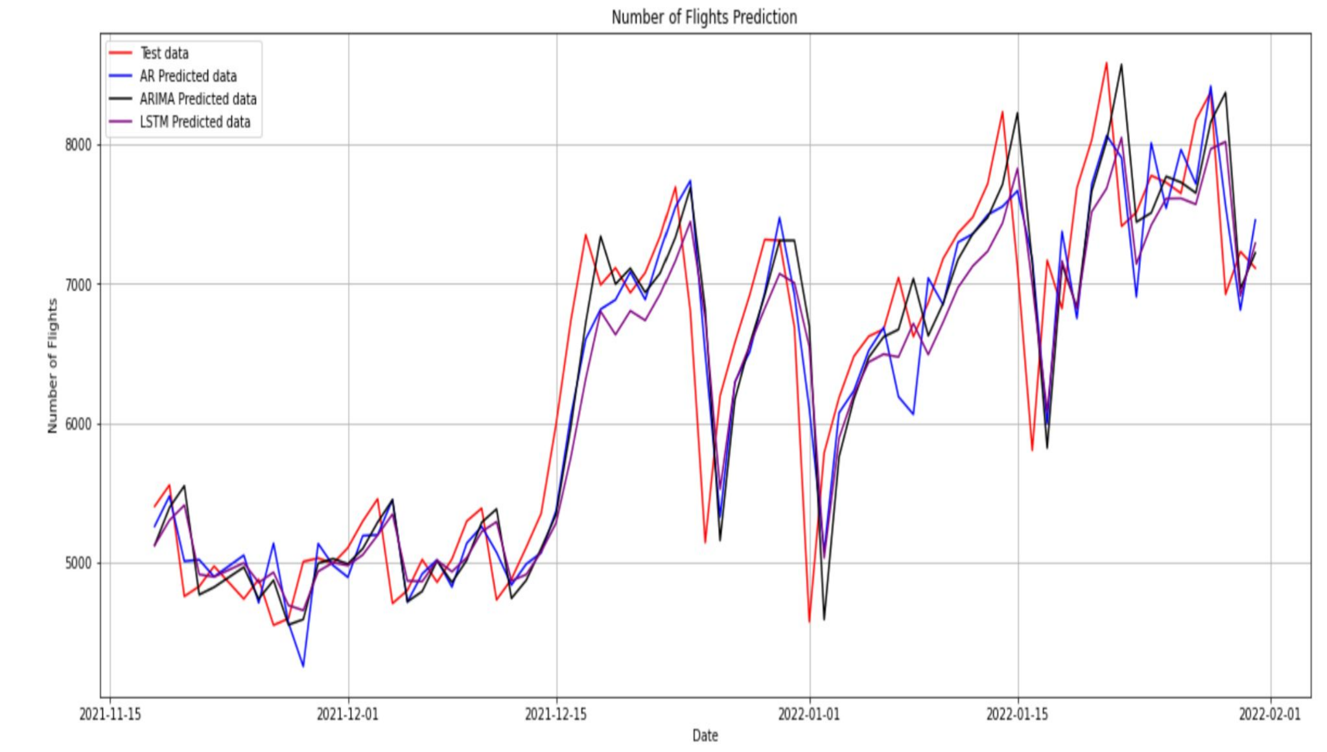






Main Insights
——Deep Learning vs. Baseline Models

- All models show major deviation throughout December 2021 and January 2022, which signifies the peak of the Omicron wave



Main Insights
——Summary

- External factors have side effects on the performance of our models
    - CDC regulations: wear masks, keep social distance
    - Coverage rates of vaccines
    - Omicron
- The Outbreak of Omicron doesn't significantly affect the number of departing flights
- Performance tuning can significantly improve the performance

Table of Contents

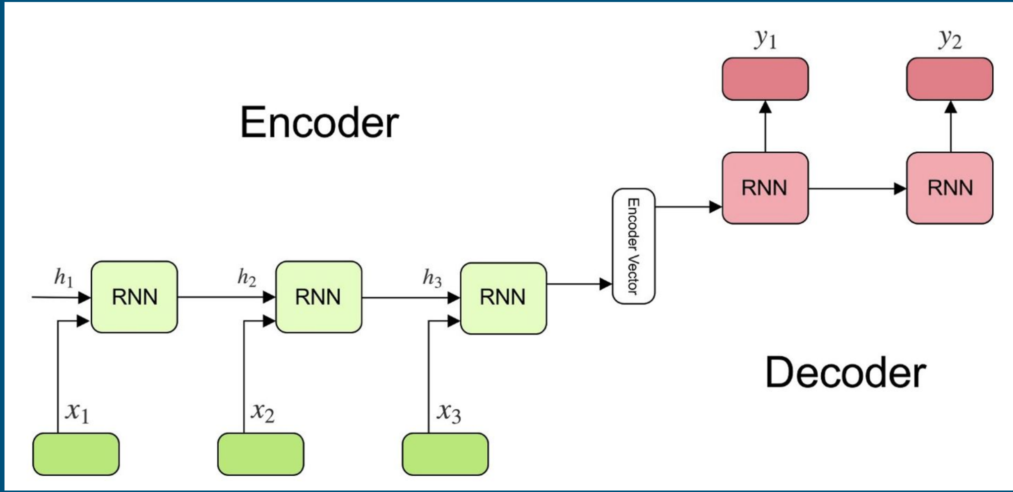
Multi-step time-series forecasting - Seq2seq with LSTM

- Maps input series to output series:

  (P is input history length, h is the forecasting horizon)

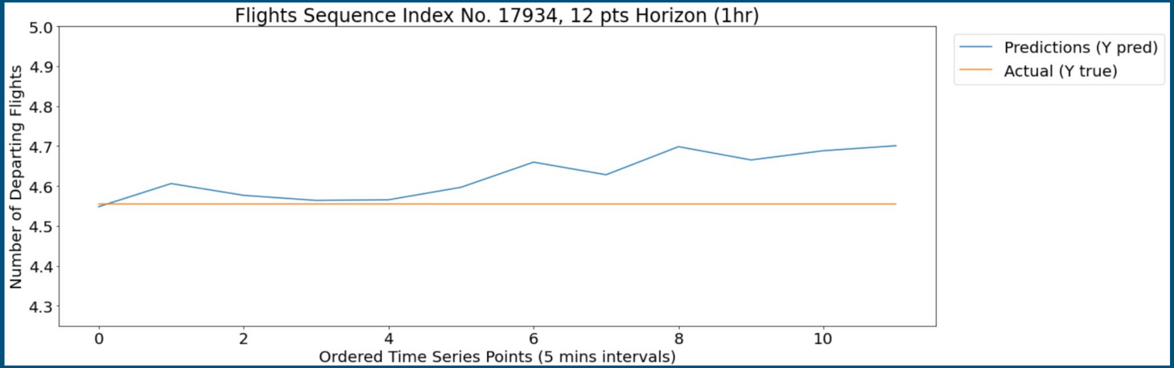
- Seq2Seq model consists of an encoder and a decoder
- Both encoder and decoder adapt LSTM components

$$x_{t-p}, x_{t-p+1}, \ldots, x_{t-1} \longrightarrow x_t, x_{t+1}, \ldots, x_{t+h-1}$$



Seq2seq with LSTM - Road blocks

- Steep learning curve for model and parameters
- Pytorch and GPU computing
- Model interpretation and evaluation
- Current results need either hyper-parameter tuning(ex: horizon, dimension) or map the origin data into a proper form of sequence



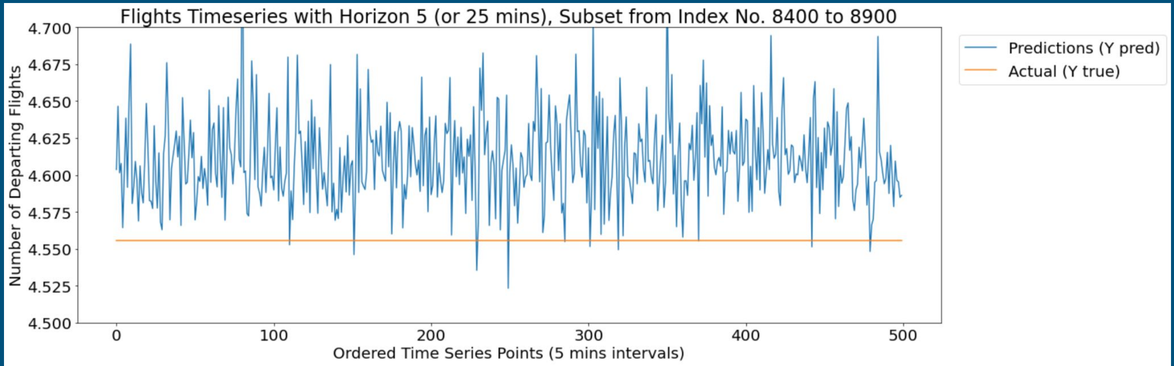


Table of Contents

Next Steps (Based on Progress)

- LSTM Model performed very well with flight data → ready to break down forecast by location and add COVID (and potentially Holiday) data as an additional feature
- Additional hyperparameter tuning will be required (LSTM, Seq2Seq)
- Look into DCRNN Model (aimed at spatiotemporal forecasting)
- Prepare for scalability → load data into database, setup SageMaker instance for deep learning notebooks
- Create benchmarks

Thank You

Q&A