Week 7 Milestone: Flight Price Predictor

Initial Modeling Attempts/Results:

For our initial modeling attempts to analyze historical flight price fluctuations, we averaged ticket prices for t-n days before departure across all airport pairs and sampled 10 flights, since the original dataset contains 82 million rows and is therefore too computationally expensive to use for model testing. After data processing, we applied a SARIMA model with m=7 seasonality, but the fit was suboptimal as seen by the inconsistencies in the residuals. As an alternate approach, we then implemented a Prophet model, again incorporating seasonality, which led us to an improved model, but still with poor residual performance. Having reached the limitations of univariate models, we plan to integrate ML techniques to improve predictions.

Next Steps:

In our initial experiments, ARIMA and SARIMA were mildly effective in the short term, but are missing many of the variables that are necessary to understand the complex relationships of demand and competition, so they fail to predict in the long term due to this inability. Given these limitations, we concluded that it makes the most sense to utilize a machine learning approach.

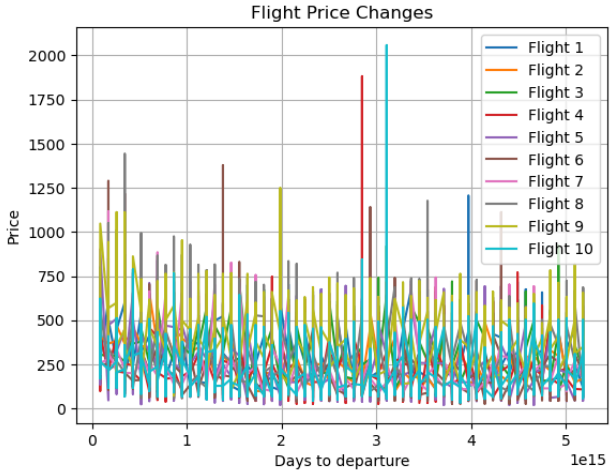
Specifically, we plan to begin by training some baseline models using bagging models like XGBoost and LightGBM due to their speed and high tabular data handling ability. This will also allow us to break from our hierarchical strategy mentioned above to create a unified global model. However, since we are concerned about these models' ability to capture the time-dependent long-term temporal patterns, we plan to run these against additional deep-learning models. This will likely include the use of a RNN or LSTM, both of which are able to regress with continuity and therefore better model sequential price movements.

We will then experiment with ensemble models, utilizing the traditional time series models' short-term forecasting and ML models' long-term price dynamics based on conditioning.

Extensions:

If time permits, we will also attempt more theoretical models that have emerged recently such as temporal fusion transformers and linear RNNs, which are proven to scale well against the large amount of data we have (82 million rows), and perform well when the external variables strongly influence price (which is the case here).

Appendix:




```
SARIMA Model for Flight 1:
                                SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 4330
Model:               SARIMAX(5, 1, 0)   Log Likelihood              -27825.546
Date:                Wed, 19 Feb 2025   AIC                        55663.092
Time:                        09:10:03   BIC                        55701.330
Sample:                             0   HQIC                       55676.592
                                 - 4330
```