Flight Delay Prediction

陈嘉懿，陈琪颖，金相恒，张天成

Abstract

Civil aviation is a vital component of the transportation system. However, various factors have caused prolonged delays in civil flights, causing significant inconvenience to passengers. To address this issue, we utilize ARIMA and TFT to predict flight delays based on flight data, aiming to minimize disruptions caused by unforeseen delays.

Keywords

Civil Aviation, Flight Delays, TFT

Contents

1. Introduction

Civil aviation, encompassing all non-military and non-state aviation, undeniably plays a crucial role in the transportation system. However, various factors, including but not limited to weather conditions, military control, major events, and emergencies, contribute to a significant likelihood of prolonged delays in civil flights. These delays can persist for several hours or even longer, causing significant inconvenience to people's lives. Recognizing this issue, we have decided to employ ARIMA and TFT methods to predict flight delays based on flight data, aiming to minimize the disruption caused by unforeseen flight delays.

1.1 Problem Description

To analyze the problem more effectively, we break it down into the following components:

- Predict flight delays based on all available pre-landing information (wheel on time).
  - Serving as the benchmark, this task is expected to be relatively straightforward since the majority of emergencies occur before the aircraft touches down.
  - The primary focus of this task is accuracy, ensuring the subsequent precise and reliable predictions.
- Predict flight delays based on all information before planes leave the runway (wheel off time).
  - Considering the limited time between landing and actual arrival, providing pre-

dictions after the aircraft has landed may not hold much significance.

  - Therefore, the second task aims to predict the duration of aircraft delays based on the available information before take-off.
  - This includes crucial details known to the tower at the time of departure, such as military control and major events.
  - By utilizing this information, we can offer insightful predictions regarding potential delays, enhancing our ability to manage and respond to such situations effectively.

- Apply an interpretable time series modeling to gain insights from segmentation of different airports/departure flights, etc.
  - Utilize an interpretable time series model to forecast data from distinct airports and flights independently, and scrutinize the disparities in the significance of various features within each group.
  - This analysis will uncover valuable insights after conducting the necessary additional analyses.

1.2 Data Sources

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers.

1.3 Methods & Innovation

We plan to base our method on the paper Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting, in which an interpretable transformer structure temporal model is introduced. This model has built-in interpretability and state-of-art prediction accuracy.

2. Main Idea

Our extensive flight data can be primarily divided into two distinct categories: scheduled and circuit flights. Specifically, scheduled flights are characterized by a continuous operation, with some routes operating a single flight per day, while others may have multiple flights within the same day. In the latter case, the destination of one flight leg often serves as the origin for the subsequent leg. These two categories constitute the vast majority of all data entries, rendering other minor categories insignificant for our analysis. By focusing on these predominant flight types, we can address the central issues with precision. Subsequently, we will select four most representative flights from each of the scheduled and circuit flight datasets. We will then conduct in-depth analysis on these flights using both ARIMA (AutoRegressive Integrated Moving Average) and TFT methods.

3. Progress

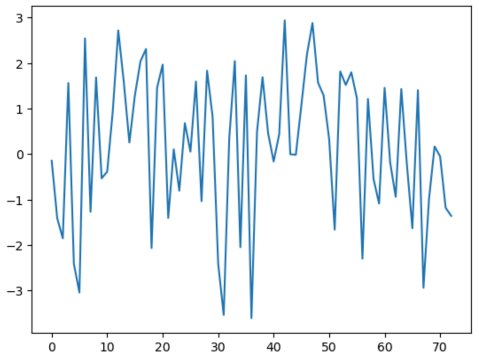
To date, the following work has been completed:

3.1 Continuous data

- **Data Preprocessing:** Utilizing continuous flight data, we addressed outliers and segmented the data into manageable slices. We developed two functions, `outl(data, column)` for outlier processing and `split(data)` for data segmentation, to facilitate subsequent operations.

- **Main Loop:** Within the main loop, we leveraged the `pmdarima` library, to be specific the `auto_arima` function. A suite of parameters was defined, where `y_train` was employed to train the model and optimize it over the training dataset. The parameters `start_p` and `start_q` pertain to the key components of the ARIMA model, with `m` representing the seasonal differencing (a challenge we faced and will discuss in detail in the subsequent subsection). We conducted a grid search to iterate through the main loop.

- **Residual Plots:** After forecasting the continuous data, we generated ten residual plots, one of which is presented below:

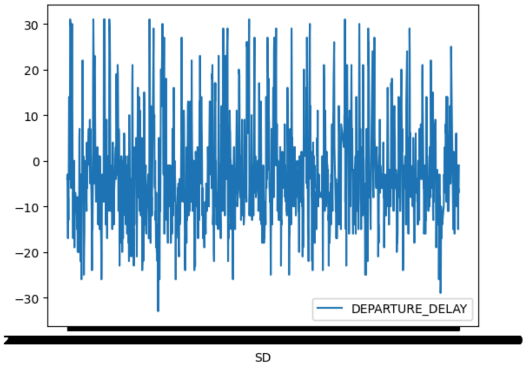


3.2 Circuit Data

- **Delay Pattern Analysis** We focus on flight AS64, which has the highest number of occurrences in the circuit dataset, to investigate the presence of a delay pattern. However, upon examination of the first forty data points for this flight, even after the removal of outliers, no discernible pattern has yet emerged. Further exploration of the subsequent data is required to draw any conclusions.

- **STL Analysis** An analysis of the circuit dataset using STL revealed no apparent trends or seasonal characteristics.

- **Distribution of Delays** After processing all the data, it was observed that the delays exhibit a more pronounced normal distribution, as illustrated in the following figure:



3.3 Explorations in TFT

- **Time Index** Raw data dates are converted into a day-time format to facilitate temporal analysis.

- **TFT Data Preprocessing** Canceled flights and those with irregular departures, such as return flights post takeoff, are excluded from the dataset. During training, we also filter out airline and flight number, as these can be collectively represented by the unique Flight ID feature.

- **Time Series Format Conversion** To transform the data frame into a time series format, non-standard times are removed, allowing for panel data analysis.

- **Exclusion of Non-standard Times** For the encoder to convert the data frame into a time series format, non-standard times are excluded, enabling analysis in a panel data framework.

- **Tail Number** Sometimes delays may be caused by intrinsic factors of a specific aircraft. Therefore, we define group_ids as [”Flight_ID”, ”Tail_Number”]. Moreover, we will continue to pay attention to Tail Number in the subsequent model analysis.

- **Target Normalizer** We opted for the count method over softplus to circumvent potential errors.

4. Challenges

When analyzing the data and applying the TFT method, we have encountered a series of challenges.

4.1 Challenges in Data

- **Data Granularity:** The dataset for civil flights spans only one year, which does not align with either quarterly or monthly reporting periods. Consequently, the seasonal differencing parameter m in the main loop requires careful consideration. The parameter m typically denotes the number of periods in a seasonal cycle, but with our limited data duration, we must decide whether to treat the data as monthly ($m = 12$) or quarterly ($m = 4$).
- **Irregular Flight Patterns:** Within the `continuous_flights_data.csv` dataset, there are instances where flights operate for only a handful of days throughout the year. This sporadic activity makes it challenging to discern any monthly or quarterly patterns.
- **Irregular Data Spacing** The Time Frequency Transform (TFT) requires equidistant data points, which poses a challenge due to the irregular spacing in our dataset.
- **Lack of Pattern** Analysis of the first forty data points for flight AS64, while visually suggesting the presence of peaks and troughs, reveals no discernible pattern upon closer examination.
- **Outlier Treatment** In dealing with circuit flights data, we encountered the dilemma of whether to exclude outliers. It is difficult to ascertain whether certain flight data points are anomalies or carry significant information. We may need to create a training set and a testing set for comparative analysis. Alternatively, we could analyze the data by adjusting for scheduled delays and removing extreme delays.
- **TFT and Discrete Variables** Time Frequency Transform (TFT) can accommodate discrete variables but is not designed to process textual data. Consequently, it is necessary to further convert textual information into discrete variables for analysis.
- **Encoder Length Parameters** Setting different values for `min_encoder_length` and `max_encoder_length` results in time series datasets of varying lengths. To maintain consistency, we are currently constrained to set `min_encoder_length` equal to `max_encoder_length`.
- **Early Stopping Problem** In our analysis using Time Frequency Transform (TFT) on the validation dataset, we consistently face the issue of early stopping.
- **Training Cutoff Anomaly** A frequent issue arises where, for example, Flight A may have been consistently using Air-

craft B or C for the first 335 days, but then switches to Aircraft D in the last 30 days. This sudden change can disrupt our indexing process, preventing the retrieval of data. Notably, such aircraft changes have predominantly been observed to occur on a particular day in December. As an interim measure, we have temporarily excluded these data points from our analysis.

5. Next-Period Outlook

- **Data Focus**

  Considering the complexity of the dataset, we will concentrate our analysis on four flights that are most representative within the continuous dataset. Subsequently, we will generate graphical representations and proceed with data fitting.

- **Irregular Flight Operations**

  In the `continuous_flights_data.csv` dataset, we observe instances where certain flights operate for only a few days throughout the year, making it challenging to discern any monthly or quarterly patterns.

- **Innovation Highlight**

  We will incorporate a baseline method and conduct comparative analyses to underscore the innovative aspects of our approach.

- **Seasonality**

  In subsequent research, the question arises as to whether scheduled flights can be considered non-seasonal, while circuit flights exhibit seasonality.

- **TFT Delay Analysis** Initially, we planned to conduct both Departure Delay and Arrival Delay predictive analyses. However, we are now at a critical juncture where we must carefully deliberate on which of these two analyses to pursue.

- **Training Cutoff Anomaly** To address the Training Cutoff Anomaly mentioned earlier, we can explore modifications to our handling of the Flight_ID or Tail number. By doing so, we aim to resolve the issue without the need to exclude data points associated with aircraft changes. It is important to note that we must also consider the continuity of the data in this process.