17-634: Dataset Selection Report

Xuchao Zhou

Dataset Description

The dataset I selected is the US Airlines Routes and Fares 1993-2024 (Jikadara, 2024) dataset from Kaggle. This dataset includes information about the United States aviation industry between years 1993 and 2024, and this information enables analysis on the trend of air ticket price, passenger flow, carrier competition, and operational efficiency of different airports. To be specific, this dataset includes the following variables:

- **tbl**: Table identifier
- **Year**: Year of the data record
- **quarter**: Quarter of the year (1-4)
- **citymarketid_1**: Origin city market ID
- **citymarketid_2**: Destination city market ID
- **city1**: Origin city name
- **city2**: Destination city name
- **airportid_1**: Origin airport ID
- **airportid_2**: Destination airport ID
- **airport_1**: Origin airport code
- **airport_2**: Destination airport code
- **nsmiles**: Distance between airports in miles
- **passengers**: Number of passengers
- **fare**: Average fare
- **carrier_lg**: Code for the largest carrier by passengers
- **large_ms**: Market share of the largest carrier
- **fare_lg**: Average fare of the largest carrier
- **carrier_low**: Code for the lowest fare carrier
- **lf_ms**: Market share of the lowest fare carrier
- **fare_low**: Lowest fare
- **Geocoded_City1**: Geocoded coordinates for the origin city
- **Geocoded_City2**: Geocoded coordinates for the destination city
- **tbl1apk**: Unique identifier for the route

There is an analysis on the US Airlines routes and fare published on Github, using this dataset (BestEver2000). This research provides insights on the air ticket fare in different time periods and in various geographical locations.

Problem Statement

17-634: Dataset Selection Report

Xuchao Zhou

Inspired by the BestEver2000 analysis, the problem I am interested in is what the average air ticket fare for a given route at a given time in the future would be. To accomplish this goal, I plan to use this dataset to train a model that can estimate the ticket price given the time, route, and carrier; if the carrier is not selected, the model should give the estimated lowest fare among all carriers. This model will enable its end users to plan their trip effectively, that they do not have to browse through all options, and it allows them to keep their travel cost low. On the other hand, such a model would also bring value to the carriers as this allows them to make projections on the market in the future and set prices that can maximize their profit.

Potential Challenges

Given the dataset and the problem statement, I have identified several key challenges in the project. The first one is that the dataset contains missing and inconsistent values, which should be dealt with before putting into model training. I need to delete rows with missing values or to use imputation techniques to fill them in (e.g. mean and median of other values within the same category). The second one is the complexity of routes - there are a gigantic number of routes for air travel within the US, and some cities even have multiple airports (which price may differ significantly). To handle this issue, I need to be very careful selecting the method or algorithm used in training the model and monitor closely to make sure such complexity does not raise problems. In addition, overfitting is a common issue in all ML projects and it also applies to this one. I should test rigorously with data outside the training dataset to see if the model does the right thing and pay attention to the performance statistics to look for anything that is suspicious. Last but not least, the COVID pandemic disturbance to the aviation industry may have a negative impact on the accuracy of the model because air travel during this period is largely different from its normal status. I need to carefully evaluate the impact of the pandemic on the market and decide if including these data in the training process of the model would be helpful in predicting the future, or it is better to wipe them out and only keep those data which the market conditions when they are generated are consistent with our expectations for the future.

References

Jikadara, B. (2024, August 4). *US airline flight routes and fares 1993-2024*. Kaggle. https://www.kaggle.com/datasets/bhavikjikadara/us-airline-flight-routes-and-fares-1993-2024/data

BestEver2000. (n.d.). *BESTEVER2000/US-airline-route-and-fare-analysis*. GitHub. https://github.com/BestEver2000/US-Airline-Route-and-Fare-Analysis?utm_source=chatgpt.com