Project Group 17: *Flight Fare Forecasting–Machine Learning for Cost-Effective Travel*

Sami Eltawil	Indumini Jayakody	Andre Melo	Nikolai Pastore
seltawil@ucsd.edu	ijayakody@ucsd.edu	amelo@ucsd.edu	npastore@used.edu

Background

This project focuses on the aviation and travel industry, specifically analyzing airfare dynamics and flight logistics using purchasable ticket data from Expedia. By exploring patterns in ticket pricing, travel behaviors, and market trends across key U.S. airports, the goal is to identify factors influencing airfare and to develop predictive models for optimal pricing. This analysis has practical applications for travelers seeking cost-effective bookings, for airlines aiming to optimize pricing strategies, and for travel platforms looking to enhance user recommendations.

Problem Definition

We aim to develop a model that accurately predicts flight ticket prices. The model's inputs include an extensive variety of relevant features, such as search date, flight date, and starting and destination airports. The corresponding price prediction output is expressed both in terms of the base fare and the total fare (which includes taxes and fees). This model thus provides insights for consumers to optimize travel planning and for stakeholders to make data-driven pricing decisions.

Motivation

Machine learning is effective for this problem due to the complex and non-linear interactions among factors influencing flight ticket prices, such as demand fluctuations, seasonality, and airline pricing strategies. The vast Kaggle dataset that we have selected, which encompasses flights from April 16, 2022, to October 5, 2022, provides an extensive summary of ticket pricing in this timeframe. While it is somewhat limited temporally, the dataset offers sufficient depth and diversity to train a model capable of capturing key pricing trends and relationships. We therefore provide a foundation that could ideally be extended with more recent training data, thereby allowing for the prediction of future ticket prices and offering more practical insights for decision-making.

Literature Review

The research community has explored various approaches to predicting flight fares, often employing machine learning (ML) models to tackle this problem. For instance, Tziridis et al. (2017) used eight state-of-the-art ML models, including Bagging Regression Trees, Random Forests, and Support Vector Machines (SVMs). They achieved up to 87% accuracy for single-flight fare predictions by focusing on features like departure time, holiday indicators, and the number of stops. Biswas et al. (2022) incorporated Random Forest, Decision Tree, and XGBoost models, emphasizing the importance of feature engineering for attributes like airline,

route, and total stops. Another approach by Groves and Gini (2013) focused on purchase timing optimization using partial least squares regression.

However, much of the existing work is constrained by limited datasets, statically or manually selected features, and a narrow focus on specific routes or airline datasets. In contrast, our approach uses a more comprehensive dataset of one-way flights from Expedia, encompassing multiple airports; diverse fare types; and granular details about flight segments, schedules, and fares. This dataset offers spatial and categorical features that enable more nuanced modeling. Moreover, rather than relying on generic machine learning approaches, we plan to incorporate more specialized models like gradient-boosting algorithms (e.g., XGBoost or LightGBM). These methods will help us address the dataset's complex feature interactions and temporal dynamics, creating a scalable fare prediction model applicable to dynamic market scenarios.

Our Approach

We plan to implement a machine learning-based approach, exploring models such as XGBoost and LightGBM. These models have demonstrated effectiveness in handling structured data with complex feature interactions and appear to be well-suited for our dataset, which contains temporal, spatial, and categorical features. We also intend to explore deep learning architectures for tabular data. Such architectures dynamically learn feature importance and can capture complex relationships within the data. Our approach is inspired by prior work, such as Tziridis et al. (2017) and Biswas et al. (2022), which highlighted the potential of machine learning models for airfare prediction in more constrained settings. By applying these ideas to a more detailed dataset, we aim to test their effectiveness and adapt our methods as necessary to develop a robust and generalizable model.

For feature extraction, we preprocess categorical variables using encoding techniques (e.g., one-hot or target encoding) and standardize numerical features to ensure consistency across scales. We engineer additional features, such as time until departure and fare-per-mile, to enhance model performance. Regarding our choice of algorithm, we expect that gradient-boosting models, such as LightGBM or XGBoost, are suitable for their ability to handle tabular data with mixed feature types and complex interactions. These models have shown strong performance in similar contexts, and we hope to validate their efficacy in our analysis. Additionally, we are considering deep learning architectures to dynamically learn feature representations and potentially capture relationships not evident through traditional methods. Our models are evaluated under metrics that assess both predictive accuracy and model robustness. Primary metrics include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which provide statistical summaries of our model's predictive error. Additionally, we consider the R^2 metric to evaluate the proportion of variance explained by the model. For a more comprehensive assessment, we analyze feature importance and use cross-validation to ensure the model generalizes well to unseen data. Low error rates and high generalizability are the primary indicators of success we anticipate.