Machine Learning for Flight Prediction

CSE 6242 Poster - Team 8 (Super Fly)

Nicola Ciocchini, Yusheng Xia, Makram Harmoush, Suresh Sivakumaran, Rutvij Rana, Xiaojie Yu

INTRODUCTION

In an era where travel efficiency is paramount, passengers are often torn between choosing cost-effective flights and opting for routes with a higher on-time performance. Existing platforms, such as Google Flights, prioritize fare cost without accounting for reliability of flight routes. This project aims to revolutionize how passengers select flights by introducing a recommendation engine that integrates both price and reliability metrics. Utilizing a blend of historical and real-time data, our engine presents a dynamic solution that caters to the modern traveler's needs.

PROBLEM DEFINITION

Navigating the complexities of flight selection, passengers often grapple with the decision on whether to prioritize lower fares or more reliable routes. The lack of comprehensive platforms that consider both cost and flight reliability leaves a significant gap in the decision-making process. Our engine addresses this gap by providing a nuanced recommendation system, thereby enhancing the travel planning experience.

APPROACHES

We propose to develop a flight recommendation engine that will redefine the flight selection process, offering a user-centric platform that melds price comparison with reliability assessments.

Real-Time Data Library

By leveraging API data connections to Trip Advisor's online data repository, users of our platform will be able to obtain estimates of flight delays for their particular flight that are based on the latest available flight patterns along with flight prices.

Probabilistic Modeling

Through our access to historical data across all flights, we can provide estimates on delay likelihood and forecast delay for any US based airline flying from a particular departure city to arrival city. Connecting flights have their delay probabilities chained using the product rule so that users will be able to obtain a holistic estimate across all their flights, a view that does not currently exist today in real time. This algorithm allows users to understand the historical reliability of their airline at different airports. Criteria for price can also be layered on to these predictions such that users can ensure the flight recommendation does not exceed a certain price while ensuring maximum reliability.

ML Predictors on Estimated Delays

While simply understanding the probability of delays is one potential method of understanding airline performance, our platform aims on expanding that view to include machine learning predictions on what the estimated delay time would be based on numerous indicators provided by the user. A boosted decision tree regression model trained on historical flight data will be used to provide flight delay expectations for any one-way flight, with the model continuously updating as newer Trip Advisor data comes in. Flight delay estimates could also include confidence intervals so that delay estimates will also contain a range of values for users who need a bit more accuracy when planning their trip.

User Friendly Interface

At the forefront of our platform is a user-centric interface that simplifies the exploration of flights and enables travelers to input their preferences and constraints easily. At minimum the platform will generate two outputs: Estimated delay, and Up-to-date Price. It should be easy to use by individuals and allow for the ability to input complex flight patterns that include multi-stop flights across multiple airports and airlines.

INNOVATION

1. **Dual-Focused Flight Recommendations**: By considering both price and reliability, our engine provides a more comprehensive assessment of flight options that has not been achieved yet by any other platform.
2. **Dynamic, Real-Time Modeling**: Unlike traditional systems that rely on static data, our engine continuously updates its predictions and recommendations based on real-time data and changing conditions.
3. **Predictive Insights into Flight Reliability**: Our model goes beyond reporting historical delays only; it anticipates both the likelihood and potential duration of delays for specific flights, offering a predictive edge that allows users to plan with greater confidence.
4. **Interactive, User-Friendly Interface**: Our platform makes complex data accessible and engaging, allowing users to explore flight options with ease and gain clear insights into the trade-offs between cost and reliability.

DATA

The data that was used to train and test our various models were downloaded from Kaggle - Flight Delay and Cancellation Dataset (2019-2023). The dataset was 614.1MB. We opted not to incorporate flight data prior to 2019 because its antiquity rendered it insufficient for accurate prediction, considering the ongoing advancements in airline practices and schedules.

To enhance flight predictions for travelers, we integrated API data connections with TripAdvisor's expansive online database, spanning millions of data points. This access to a substantial volume of live data greatly enhances our ability to provide more accurate estimates of flight delays.

EXPERIMENTS

Probabilistic Modeling: Although the team has successfully implemented a probabilistic model, we decided not to include it in the final implementation of modeling, due to its overlapping with ML models and the complexity of incorporating edge-case scenarios into custom flight patterns.

ML Predictors on Estimated Delays: The team experimented with different ML models to understand the accuracy of predicted flight delays: **1)** Random forest ensemble (subsampled in python), **2)** Linear regression (subsampled in python), **3)** Linear and logistic regression models (full dataset in Azure), **4)** Boosted Decision Tree Regression (full dataset in Azure). The team evaluated model performance based on statistical metric evaluations such as Mean Absolute Error (MAE) and Mean Squared Error (RMSE) as listed in Table 1 and chose Boosted forest regression model (highlighted in red in Table 1).

Performance Metrics	Independent variables are all columns; arrival delay as dependent		Independent variables are subset of columns; arrival delay as dependent	
	Linear Regression	Boosted Forest Regression	Linear Regression	Boosted Forest Regression
MAE (Mean Absolute Error)	0.025136	2.067141	18.016774	17.439848
RMSE (Root Mean Squared Error)	0.294626	4.763171	44.697694	44.387381
RSE (Relative Squared Error)	0.000033	0.008614	0.75858	0.748084
RAE (Relative Absolute Error)	0.001047	0.08611	0.750519	0.726486
R2 (Coefficient of Determination)	0.999967	0.991386	0.24142	0.251916

Independent variables are subset of columns; arrival delay as dependent	
Performance Metrics	Multiclass Logistic Regression
Overall accuracy	0.925469
Micro_Precision	0.925469
Macro_Precision	0.929998
Micro_Recall	0.925469
Macro_Recall	0.906356

Performance Metrics	Independent variables are subset of columns; arrival delay as dependent	
	Two Class Logistics Regression (0.5 Threshold)	Two Class Decision Forest (0.5 Threshold)
Accuracy	0.927	0.947
Precision	0.953	1
Recall	0.836	0.851
F1 Score	0.891	0.919
AUC	0.982	1

Table 1. Performance of various models.

RESULTS

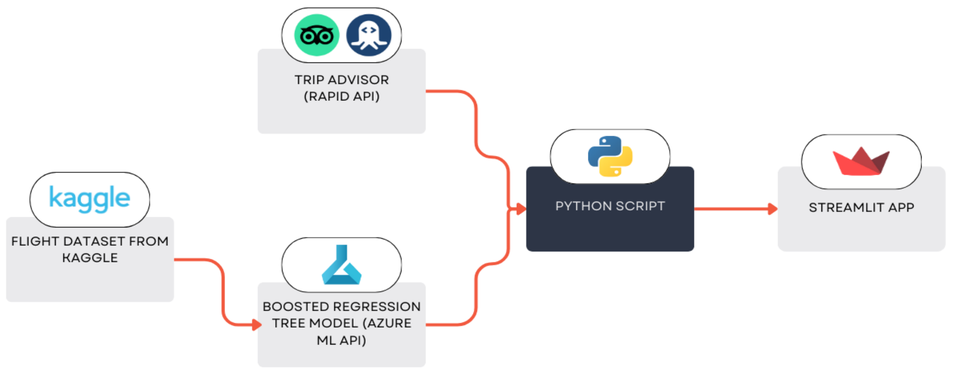

Figure 1. Data pipeline for Super Fly app

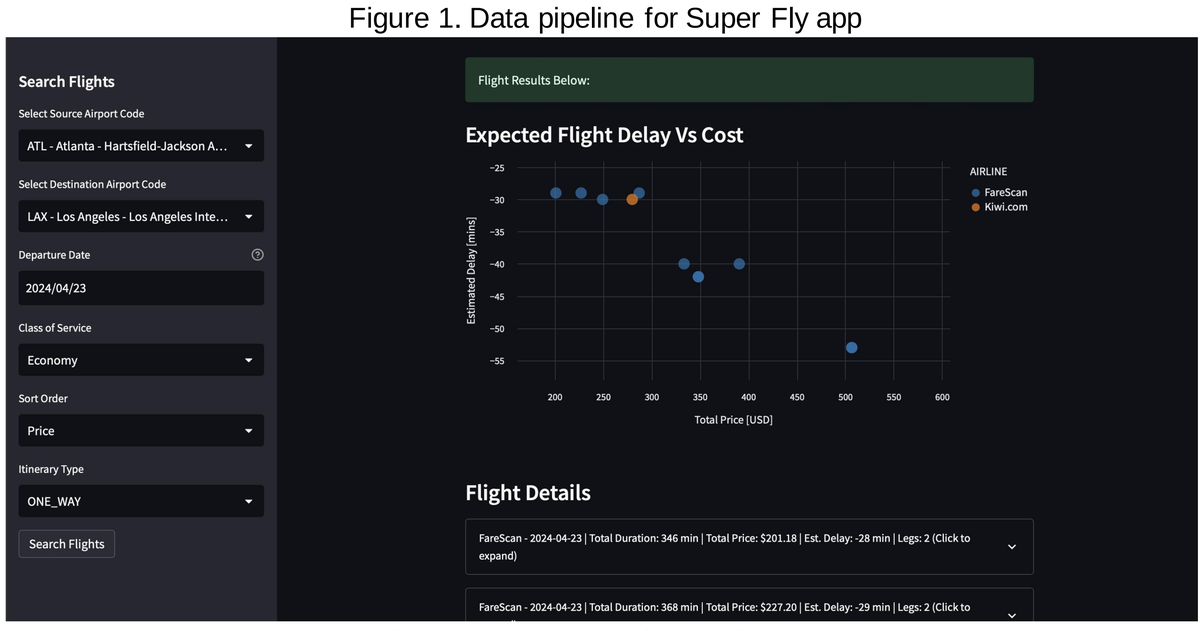

Figure 2: Interface of Superfly App

As shown in the workflow listed in Figure 1, we deploy the Boosted Decision Tree Regression model using Azure ML and developed a User-Interface 'Superfly' app using Streamlit. Upon selecting a specific origin, destination and date, users are presented with a comprehensive tool designed to aid in making informed decisions regarding their travel plans. This output will not only include a table showing the flight options with price and expected delays (data now shown), but also include an Expected Delay vs. Price graph (Figure 2), which visually highlights any measurable differences within different carriers.

CONCLUSION

The goal of this project is to build a strong proof of concept that could then be expanded to provide actual values to end users. We leverage historical flight data as well as live data to predict flight delays and correlate it with price. Our results can help many users understand the factors involved when deciding the airlines and flights they want to take, providing valuable insights for informed decision-making.