FLYR Data Science Challenge

FLYR is the world's premier machine learning and data company in the flight and travel pricing space. Better than anybody else, we understand where people want to travel, how much they want to pay and how the prices currently available will change over time. Still, there is a lot more that we can improve upon and a lot more data that we have yet to tap into. The success of the products that we can offer to consumers, online travel companies and airlines is largely driven by our ability to measure and predict the "market value" of a flight being offered to potential travelers and the travelers' willingness to accept the offered airfare.

The data and prediction challenges confronted in travel and pricing are both complex and nebulous. Most of the time there are multiple steps to the problem and multiple considerations that must be balanced at the same time while working toward a goal that is not fully defined. Below is a simplified, although certainly not simple, example of the challenges we face daily.

In this challenge there are two goals.
One, **use the attached data sets to build a model that predicts an offered airfare**. The decisions on how which data points you use to validate your model and which metrics you choose to optimize and use to report your findings are left to you, but please clearly communicate both the results, methodology, assumptions and reasons for these decisions in your final report.

Two, **use the attached data sets to predict which airfares are ultimately booked**. Again, you are free to choose the simplifications, methodologies and metrics that you feel are appropriate and are most comfortable with.

Once you are have completed your models and validated your results, please **create a simple report documenting your methods, assumptions and results**. This report should be both concise and clear and highlight the aspects of that you believe are most important to your work. The report should take no more than 10 minutes for another data scientist to read and understand.

In addition to the above, please provide a section in the report explaining the features used in your methodologies and which features you found to be most impactful. The report must contain at least one data visualization that supports your results and a section on proposed next steps (only so much can be done in such a short time period).

A few hints:
Before beginning, spend some time on various flight search website refamiliarize yourself with the flight purchase experience. Try different websites and think about what flight parameters would most matter to you.
The data sets have a decent amount of missing information. These are real data sets that have been only slightly modified for this challenge.

The Data

Search Data

- **search_id** - Unique id for the search.
- **search_time** - UTC timestamp of search.
- **currency** - 3-letter currency code of the airfare.
- **destination** - 3-letter airport code of the destination airport.
- **search_user_id** - FLYR's unique user ID.
- **language** - Language in which the search was done.
- **origin** - 3-letter airport code of the origin airport.
- **partner_id** - FLYR's anonymized ID for a partnering travel website.
- **passengers** - Number of passengers in the flight search.
- **pos** - Point of Sale, country to which the airfare is being offered.
- **session_id** - Unique identifier of a search session.
- **user_agent** - Web browser user agent.
- **num_requests** - The number of flight itineraries in the search results as seen by FLYR.
- **fare** - The airfare offered.
- **supplier** - FLYR's anonymized ID for a airfare supplier to the travel website.
- **cabin_class** - cabin class of the flight search.
- **carrier_1** - The two letter airline code of the outbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **carrier_2** - The two letter airline code of the inbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **flight_num_1** - The flight numbers of the outbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **flight_num_2** - The flight numbers of the inbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **departure_datetime_1** - Departure datetime for each flight segment of the outbound flight itinerary localized to departure airport.
- **departure_datetime_2** - Departure datetime for each flight segment of the inbound flight itinerary localized to departure airport.
- **arrival_datetime_1** - Arrival datetime for each flight segment of the outbound flight itinerary localized to arrival airport.
- **arrival_datetime_2** - Arrival datetime for each flight segment of the inbound flight itinerary localized to arrival airport.
- **itinerary_id** - Unique ID for an offered flight itinerary.

Booking Data

- **booking_id** - Unique ID for the search
- **booking_time** - UTC timestamp of booking.
- **currency** - 3-letter currency code of the airfare.
- **destination** - 3-letter airport code of the destination airport.
- **search_user_id** - FLYR's unique user ID.

- **language** - Language in which the search was done.
- **origin** - 3-letter airport code of the origin airport.
- **partner_id** - FLYR's anonymized ID for a partnering travel website.
- **passengers** - Number of passengers in the flight search.
- **pos** - Point of Sale, country to which the airfare is being offered.
- **session_id** - Unique identifier of a search session.
- **user_agent** - Web browser user agent.
- **num_requests** - The number of flight itineraries in the search results as seen by FLYR.
- **fare** - The airfare offered.
- **supplier** - FLYR's anonymized ID for a airfare supplier to the travel website.
- **cabin_class** - cabin class of the flight search.
- **carrier_1** - The two letter airline code of the outbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **carrier_2** - The two letter airline code of the inbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **flight_num_1** - The flight numbers of the outbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **flight_num_2** - The flight numbers of the inbound flight itinerary. If there are multiple flight segments then the carrier codes are separated by semicolons.
- **departure_datetime_1** - Departure datetime for each flight segment of the outbound flight itinerary localized to departure airport.
- **departure_datetime_2** - Departure datetime for each flight segment of the inbound flight itinerary localized to departure airport.
- **arrival_datetime_1** - Arrival datetime for each flight segment of the outbound flight itinerary localized to arrival airport.
- **arrival_datetime_2** - Arrival datetime for each flight segment of the inbound flight itinerary localized to arrival airport.
- **itinerary_id** - Unique ID for an offered flight itinerary.

Airport Data

- **iata_code** - 3-letter IATA airport code.
- **city** - The name of the city served by the airport, if known.
- **country** - The name of the country in which the airport is located.
- **latitude** - Latitude of the airport, if known.
- **longitude** - Longitude of the airport, if known.
- **altitude** - Altitude of the airport, if known.
- **timezone** - Local timezone for the airport.
- **dst** - Daylight saving time rules for the airport.
- **aggregate_code** - 1 if the code is for multiple airports in a metropolitan area. 0 if the code is for a specific airport.