***Fin-A-Lytics***

Team: DATA CRUNCHERS

### Team Members:

1. Aniruddha Mukherjee
2. Ritika Banerjee
3. Diptarup Chakravorty (Team Leader)

# Model Description

We made use of the XGBoost algorithm, which is well-known for its efficacy and accuracy, in order to make projections regarding the "Booking Class Fare USD" for aircraft fares. The gradient boosting framework successfully manages enormous datasets and addresses the problem of overfitting by including various methods of regularisation. In order for the model to accurately capture the complex nature of the price structure for flight reservations, it required a thorough collection of 472 attributes that included both categorical and numerical variables. This was done in order to ensure that the model was as accurate as possible. The introduction of temporal data into the investigation was one of the most important aspects of the study. More specifically, the Day, Month, Year, Hour, and Minute variables were all included. Because of this, the model was able to take into account temporal fluctuations, which included time-based trends, seasonal swings, and other time-sensitive features. In addition, we have constructed a one-of-a-kind property that we refer to as "Day\_of\_Week." This attribute contains the consistent patterns that are seen across multiple days of the week and is named after the day of the week. This quality proved to be very acute in identifying the nuances that are typically noted in airline price, such as rises over the weekends or drops in the middle of the week, for example. The incorporation of these features inside the robust framework of XGBoost made it possible to construct a highly accurate predictive model, which shown an excellent level of precision when forecasting the "Booking Class Fare USD."



# Assumptions

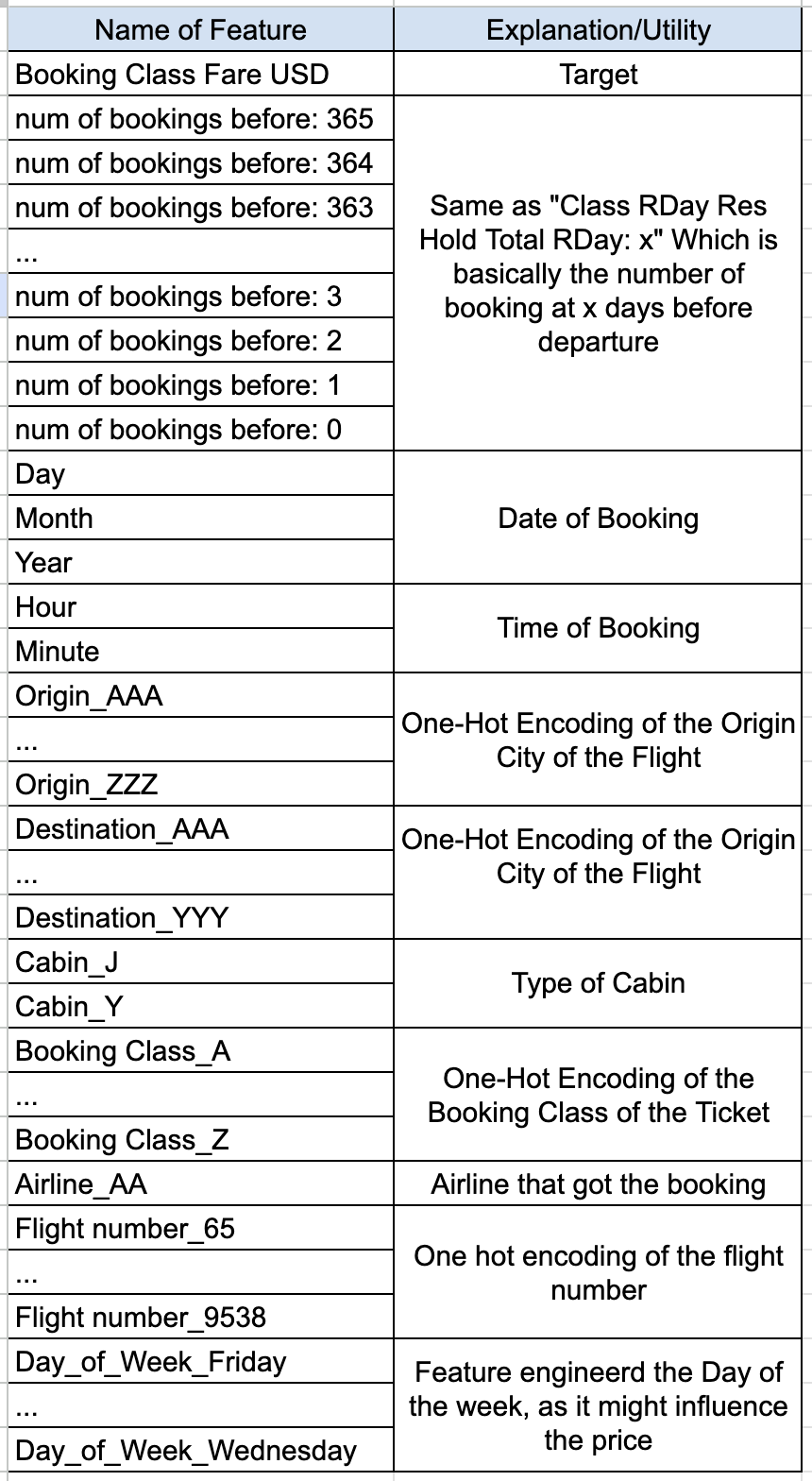
We assumed wherever there was “NULL” in the dataset on the columns that had “Class RDay Res Hold Total RDay: x”, that meant the booking hadn’t started and it was null.

Data Quality: The model assumes that the input data is of high quality and has been preprocessed to handle missing values, outliers, and any inconsistencies.

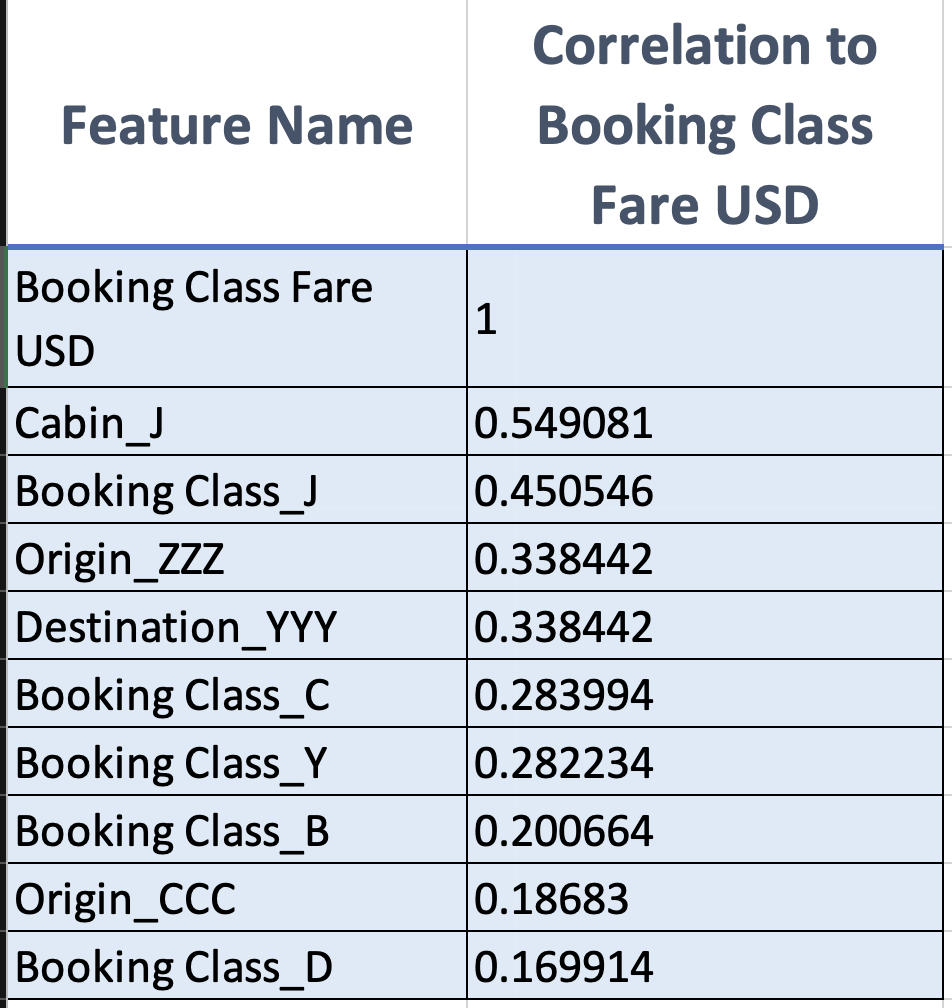
# Feature Explanation

Here is a table representing all the features we used to train our model.

We engineered a feature termed “Day\_of\_Week” to give the model information about weekends, weekdays etc.



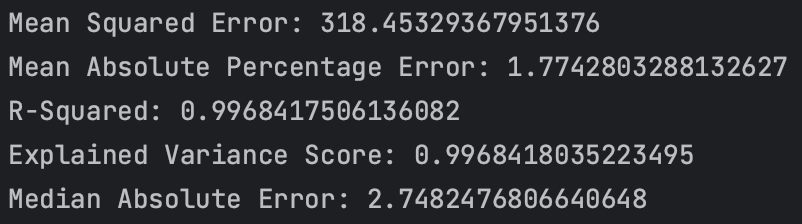
Here is a correlation matrix of the features that maximally influenced the XGBoost Algorithm



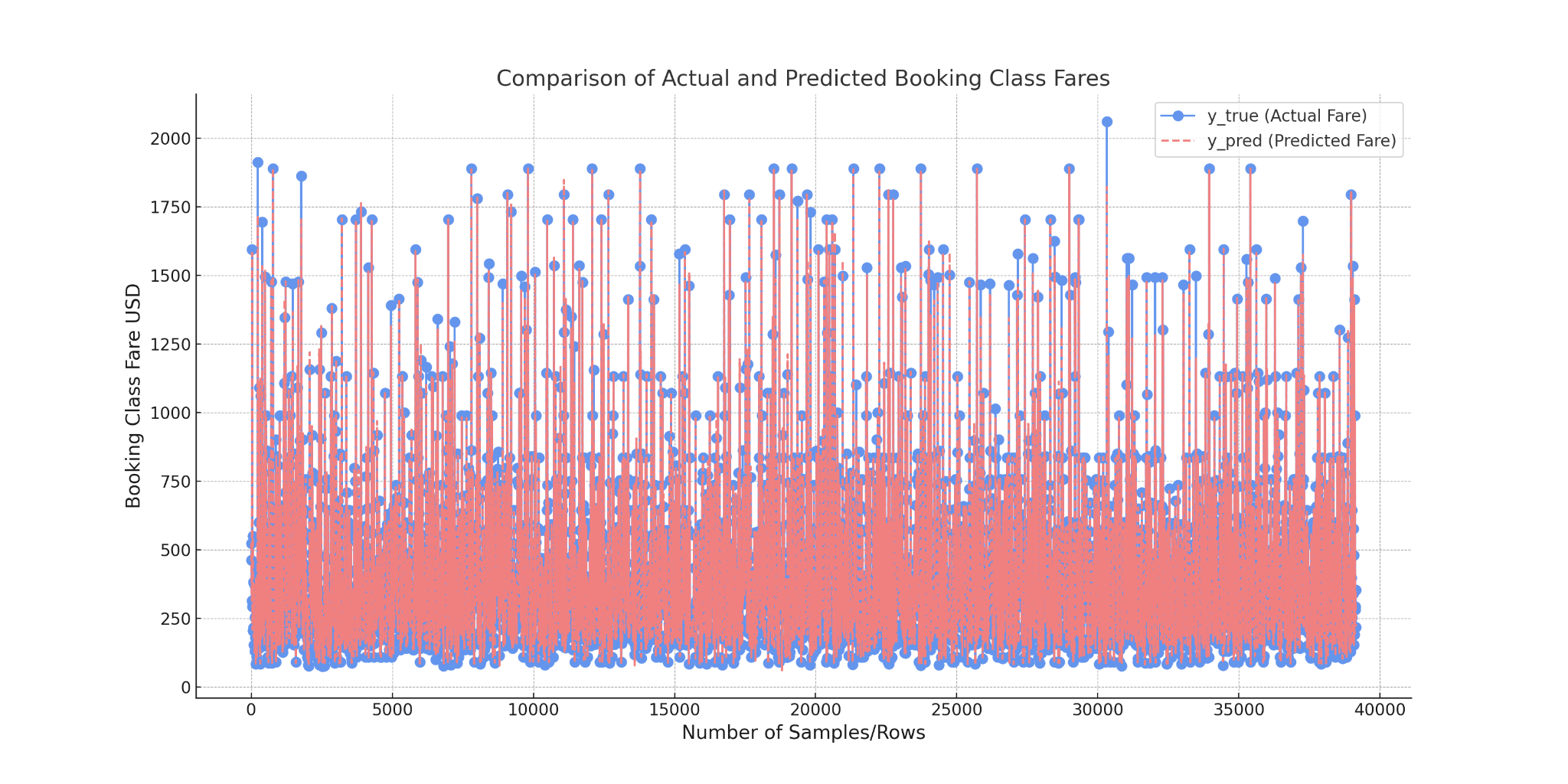
**We used a total of 472 features.**

**We had 156619 training samples and 39155 test samples**

# Prediction Output



Our model performed remarkably well on the test set with ~40,000 samples.



**Notebook Link:**

<https://drive.google.com/file/d/1n9gpBuQsGYiexB8FEZRKtlyBU6tEbL00/view?usp=sharing>

# Future extension or applicability of the model in continuous pricing/classless

We plan to incorporate a LSTM model to capture the time based anomalies in a better way. It has been built by us, but we couldn't train it in time.

# References

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. <https://doi.org/10.1145/2939672.2939785>

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Comput. 9, 8 (November 15, 1997), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>