EXPERIMENT REPORT

Student Name	Brandon Ji
Project Name	AT3
Date	9-11-2023
Deliverables	

1. EXPERIMENT BACKGROUND	
Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.	
1.a. Business Objective	The goal of this project is to create a model that can accurately predict flight fare amounts based on certain variables. By being able to predict flight fare amounts, users can plan budgets based on this prediction and also compare the effects that different variables would have on flight fares.
1.b. Hypothesis	The hypothesis that will be tested is XGBoost will create a strong model due to its ensemble method and gradient boosting properties.
1.c. Experiment Objective	I am expecting that the initial model will have decent results, with the model being quite a bit better after tuning hyperparameters.

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation	Rows containing null values were removed from the dataset as they may negatively affect the predictions of the model. Certain columns were also removed as they were deemed irrelevant or redundant such as ‘LegID’, ‘segmentsDepartureTimeRaw’ , ‘segmentsArrivalTimeRaw’ etc.
2.b. Feature Engineering	The ‘‘segmentsArrivalTimeRaw’ contained the flight arrival times for each leg of the flight separated by ‘||’. As there was such a varied number of legs for each flight, it was not viable to keep the data for each leg of the flight. Due to this, only the last segment was kept to represent the arrival of the final flight. For the same reasons, only the first departure time segment was kept from the ‘‘segmentsDepartureTimeRaw’ column. Only the first airline code was also kept from the ‘segmentsAirlineCode’ column and stored in a new column called ‘airline_code’. To compensate for the removal of the middle, a new feature called ‘stopnumber’ was created. This counted the number of stops made in each trip. For ‘segmentsCabinCode’, a new column was created for each unique cabin type. These columns were then filled with ‘1’ if the corresponding flight had used that cabin type and 0 if not. These new columns were labeled ‘iscoach’, ‘ispremiumcoach’, ‘isfirst’ and ‘isbusiness’. The ‘DepartureTime’ and ‘ArrivalTime’ columns were also split into month, day, hour and minute columns as the model cannot process datetime objects. The columns ‘startingAirport’, ‘destinationAirport’ and ‘airline_code’ were all label encoded to allow the model to process the categorical variables.

2.c. Modelling

XGBoost is a machine learning algorithm that is based on a decision tree ensemble. It utilizes extreme gradient boosting to enhance the predictions made by the decision tree ensemble. It can be used in both classification and regression tasks. The hyper parameters tested were

- Colsample_bytree
    - Percentage of columns used for each tree. A lower value can result in less overfitting.
- max _depth
    - Maximum depth of each tree. A higher value can capture more complex patterns but may lead to overfitting.
- Learning_rate
    - Step size shrinkage. As this value gets higher, computation becomes faster but the performance of the model may suffer. The inverse is true for when the value gets lower.
- N_estimators
    - Number of trees in ensemble. Increasing this value can improve model performance at the cost of computation time.

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance	The model with default hyper parameters had an MSE of 10473 and 102.34 RMSE. The optimised model had an MSE of 5595.04 and RMSE of 74.8.
3.b. Business Impact	The current results have a decent result however they can only provide an estimate. As the purpose of the model is to provide accurate results, the current model may not be suitable.
3.c. Encountered Issues	Due to the many segments in the segment based columns, a lot of data was lost during preprocessing.

4. FUTURE EXPERIMENT

Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.

4.a. Key Learning	The current model is quite strong, however there is room for improvement. Changes in the preprocessing could likely lead to better results.

4.b. Suggestions / Recommendations	Instead of removing middle segments, perhaps these segments could be split into separate columns. This will lead to a much larger dataset however it could improve model performance.