EXPERIMENT REPORT

Student Name	Ya-Ping
Project Name	Liao
Date	09/11/2023
Deliverables	<dt_regression> <Decision Tree Regressor> <Random Search>

---

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 is to empower airfare prediction strategies, to empower customers to take control of their air ticket budget when planning their trips, and to give them the knowledge and resources necessary to decide wisely, select affordable solutions, and take pleasure in a flawless and economical travel experience.
1.b. Experiment Objective	The experiment aims to train a decision tree model using random search to enhance its performance. The target variable is 'totalFare,' representing the airline ticket price. The model's performance will be evaluated using MSE and RMSE. And establish a pipeline for preparing the model for deployment on a Streamlit app.

---

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	Data cleaning**:** use the dropna() to clean the missing value Define four functions to deal with raw data: 1. convert_duration: change the data from string to become total duration minute like the “PT7H52M” will transfer into integer minute 2. split_column_into_two: Define a function to split a column's values in the dataset, where transfer information is separated by "||," into two separate columns.*(‘segmentsDepartureTimeEpochSeconds’, ‘segmentsDepartureTimeRaw’, ‘segmentsArrivalTimeEpochSeconds’,etc.)* 3. drop_columns: define a function that removes columns that have already been processed and creates new columns, resulting in the removal of the original ones. 4. extract_transfer_airport_code: Define a function to keep the transfer airport.
2.b. Feature Engineering	Define five function to do feature engineering: 1. convert_columns_to_datetime_utc: Convert the input columns into a given datetime format and set the time zone to UTC for each converted datetime value. 2. calculate_transfer_waiting_time: Calculate the waiting time in minutes between two datetime columns within a DataFrame, and the result is stored in a new column specified by ‘new_column_name’. (Use preprocessing data to calculate the waiting time when arriving at the transfer airport, which is the time between `ArrivalTimeRaw_1` and `DepartureTimeRaw_2`) 3. calculate_time_duration: Calculate the number of days between the search and departure dates. 4. convert_object_columns_to_integer: Convert the object data type columns to integer data type. 5. drop_columns: Define a function that removes columns that have already been processed and creates new columns, resulting in the removal of the original ones.
2.c. Modelling	Model: Choose decision tree regressor handles both numerical and categorical features effectively, making it suitable for datasets with mixed data types without the need for extensive preprocessing. Hyperparameter tuning: use the random search to find the best model. The best hyperparameters are as follows: model__min_samples_split: 2 model__min_samples_leaf: 6 model__max_features: 0.3 model__max_depth: 14. Abstain from utilising linear regression due to the potential for a high MSE and RMSE stemming from the inclusion of a large number of features.

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	Following model training, the testing data was utilised to predict the target variable, 'totalFare.' Subsequently, the model was assessed using MSE and RMSE. The results revealed an MSE of 3278.4465 and an RMSE of 57.2577, respectively.
	`mse: 3278.4464852992737`
	`rmse: 57.25771987513364`

3.b. Business Impact	Using the predictive capabilities of the model for airfare enables customers to manage their budgets proactively, fostering an environment of financial control and strategic decision-making in the intricacies of travel planning.
3.c. Encountered Issues	Analyse the functionality and significance of the segment's columns within the dataset. Subsequently, implement strategies to handle and interpret these columns effectively, ensuring their integration into the overall data processing pipeline. Given the considerable size of the merged dataset, opt for a more manageable approach by utilising a 20% sample for the training phase. Furthermore, it confronts the prevalent issue of missing values within the dataset. Develop robust methodologies for imputation or, if necessary, consider the strategic removal of observations with missing data.

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	Learn how to separate a column delimited by "||" and handle UTC time format data, implement a function that employs appropriate string manipulation techniques and datetime conversions. Additionally, develop strategies within the function to handle string columns, converting relevant portions into numeric formats for subsequent target value calculation. Create a comprehensive feature engineering function that addresses the specific needs of the dataset, generating new columns to enhance model performance. Recognize that Labelencoder encounters issues when run through the pipeline; hence, consider utilising Ordinal Encoder for effective categorical encoding within the pipeline.
4.b. Suggestions / Recommendations	An attempt to utilise sample data for training a neural network model will be made. An efficient approach will be employed to handle the large dataset, and the model will be trained on the entire dataset using the decision tree regression model. More sophisticated strategies will be implemented to address the intricacies of the data.