EXPERIMENT REPORT

Student Name	Shi Wu
Project Name	Experiment1——Lightgbm
Date	23/09/2023
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	Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results? The goal of this project is to develop an application that leverages machine learning models to predict flight fares in real-time, aiding users in budget planning and fare comparison. This service would primarily benefit users looking to travel, whether for business or leisure, by providing them with a tool to estimate and manage their travel expenses more effectively. Accurate predictions will significantly benefit the business by fostering customer trust and loyalty, as travelers will likely return to the app for future travel planning. Additionally, the predictive data can inform the company's marketing strategies and provide insights into consumer behavior and preferences. On the other hand, inaccurate predictions could result in a loss of user trust, potentially leading to a decline in engagement and damage to the company's reputation.

1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, The hypothesis we are exploring is whether machine learning models can accurately predict flight fares in real time. We're looking to answer if such predictions can indeed help travelers make better decisions about their travel plans. This is important because airfares frequently change due to various factors like seasonality, demand, and airline strategies. Accurate predictions could lead to cost savings for travelers and provide a personalized experience, which is highly valued in today's market. This endeavor is worthwhile because technological advancements have now made it possible to process large amounts of data and apply sophisticated algorithms to forecast prices. If successful, the predictive tool could not only save money for consumers but also give companies in the travel industry a way to stand out from the competition by offering a valuable service. The insights from this project could also guide airlines and travel agencies in their pricing and marketing efforts, making the entire exercise a strategic step for businesses looking to leverage data for better decision-making.

1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The expected outcome of the airfare prediction experiment is to create a model that can predict flight prices with reasonable accuracy. We hope to achieve a level of precision where the predictions are consistently useful for budget planning by travelers. If the model is successful, it will mean travelers can rely on the tool to estimate their flight costs, potentially leading to cost savings and better trip planning. For the business, it would mean providing a valuable service that could attract and retain customers. If the model is moderately successful, it might still offer benefits by giving travelers a general idea of expected costs, and it could serve as a foundation for further refinement. Should the model not predict accurately, it would signal a need to revisit the data and model approach, perhaps by incorporating more variables or using more advanced algorithms. In any case, the outcome will provide insights: a successful model validates the approach, while less accurate predictions indicate areas for improvement. Each result will guide further development and refinement of the tool.

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

Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments

Initially we combined the multiple raw data files into a single data

1. Separates Information: It splits the departure and arrival information into separate columns based on the first and last segments of the journey.
2. Timestamp Handling: It converts timestamps into a usable format, extracting the date, time, and hour for both departure and arrival times.
3. Numeric Transformations: It changes strings that represent numbers into actual numeric values, which are needed for calculations.
4. Sum Calculations: It adds up numeric values in certain columns, like the total flight duration or distance, to give a single sum for each row.
5. Cleans Data: It removes unnecessary columns that are not useful after the initial extraction and parsing of the data.

2.b. Feature Engineering

Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments

For Numerical Data:

- Filling in any missing values with the average of the rest of the values in the column.
- Standardizing the values so they're centered around zero and have a standard deviation of one.

For Categorical Data:

- Replacing missing values with the word ‘missing'.
- Changing categories into numbers so that the machine learning model can understand them, with a special number for any new categories that weren't seen when the model was being made.

2.c. Modelling

Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments

LightGBM is a gradient boosting framework that is designed for speed and efficiency. It's particularly suitable for large datasets and can handle categorical features directly without the need for preprocessing. The choice of the Poisson objective makes it well-suited for count-based predictions, such as sales forecasting.

The hyperparameters tuned and the values tested in the RandomizedSearchCV for the LightGBM model are:

1. model__learning_rate: [0.005, 0.01, 0.05, 0.1]：This determines how quickly the model learns. Lower rates can lead to better long-term accuracy but require more trees (n_estimators) to train.
2. model__n_estimators: [50, 100, 500]：This is the number of trees to be built in the boosting process. More trees can capture more complex patterns but also risk overfitting.
3. model__num_leaves: [20, 31, 40]：This is the maximum number of leaves for one tree. Having more leaves makes the model more complex and can fit the training data better, but also can overfit.
4. model__boosting_type: [‘gbdt’]：This specifies the boosting algorithm type. 'gbdt' is chosen because it is standard for regression and generally performs well.
5. model__objective: [‘regression’]：This defines the task the model is used for, which is regression in this case.
6. model__metric: [‘l2’]：The 'l2' metric, also known as Mean Squared Error, is a common choice for regression problems.
7. model__subsample: [0.1]：This is the subsample ratio of the training instance, which is a technique to speed up training and help with overfitting.

Models Not Trained: No specific models are mentioned in the code snippet that have been decided against training. However, in practice, one might decide not to train certain models like deep neural networks or support vector machines due to:

	1. Complexity: These models can be too complex and computationally expensive for the given task. 2. Data Size: They may require large amounts of data to perform well, which might not be available. 3. Interpretability: Simple models like linear regression are often preferred for their interpretability, especially in industries like finance or healthcare where understanding the model's decisions is crucial.

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

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

An RMSE (Root Mean Squared Error) of 99 can be interpreted as underperforming or acceptable depending on the context of the problem and the scale of the target variable. If the target flight prices are generally in the range of thousands, an RMSE of 99 might be quite good. However, if flight prices are in the range of hundreds, then an RMSE of 99 would indicate that the model's predictions are off by a significant margin on average.

To analyze the underperforming cases and identify potential root causes, we can take the following steps:

1. Error Analysis:
    - Examine the distribution of errors. If the errors are normally distributed around zero, the model might just need slight adjustments. If the errors are skewed, there may be systematic bias in the model.
    - Look at the residuals (the differences between predicted and actual values) across different values of predictors. This can show whether the model is consistently over- or under-predicting for certain ranges of variables.
2. Data Quality:
    - Check for any data entry errors, outliers, or anomalies in the dataset that could be affecting model performance.
    - Review if there's sufficient representation for all categories in your dataset. A lack of diversity in training data can lead the model to perform poorly on underrepresented categories.
3. Feature Engineering:
    - Consider whether the model might be missing important features that could explain variances in flight prices, such as holiday periods, special events, or airline promotions.
    - Revisit preprocessing steps to ensure that you're not losing valuable information. For instance, binning continuous variables too broadly or encoding categorical variables incorrectly might reduce the model's ability to capture nuances.
4. Model Complexity:
    - A model that's too simple may not capture all the complexities of the data

(underfitting), whereas a model that's too complex may capture noise in the training data rather than the actual signal (overfitting).

5. Hyperparameter Tuning:
   - The current hyperparameters may not be optimal. You can try more extensive searches or different ranges of hyperparameters.
   - Consider using other model evaluation metrics during hyperparameter tuning that might be more appropriate for your business objectives.
6. External Factors:
   - There might be external factors not captured in the data that are influencing flight prices, such as economic changes, fuel price fluctuations, or changes in competition.
7. Temporal Dynamics:
   - Flight prices are dynamic and can change based on the time of booking and the time until the flight departure. Ensure that the model is accounting for time-sensitive patterns.

3.b. Business Impact	Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others) Interpreting the RMSE of 99 in relation to the business objectives, the accuracy of flight fare predictions has direct implications for user satisfaction and the company's credibility. If the model's predictions are off by a large margin, users might misjudge their budget for flights, leading to potential overpayment or missed opportunities for cheaper travel. This can cause frustration and diminish their trust in the tool, making them less likely to use it in the future. From the business perspective, inaccurate fare predictions could affect not just the primary service but also ancillary product sales linked to flight bookings. Users may start looking elsewhere for more reliable information, impacting the company's market position and profitability. The degree of impact depends on how central fare prediction is to the business. If it's a core service, the effects of inaccuracy will be more pronounced compared to if it's just one of many features offered. To reduce these negative outcomes, the company will need to enhance the predictive model to provide users with more accurate, reliable flight fare estimates.
3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be deal with in future experiments. Issues Faced: • Data Quality: Problems with missing values and inconsistencies. We Imputed missing values and sanitized data and delete some observations with na value. • Computational Limitations: Due to out of memory and extended training times with large dataset. We just used efficient algorithms like LightGBM. • Hyperparameter Tuning: Automated methods like grid search is not applicable due to large dataset so we just applied Random Search. Future Experiment Concerns: • Model Diversity: Need to try diverse models beyond LightGBM. • External Factors: Sales influenced by factors not in the 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	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. The outcome of the experiment with an RMSE of 99 tells us that while the model has room for improvement, it's a good starting point. We've learned that predicting flight fares is complex and requires careful consideration of many factors like timing and airline pricing strategies. The experiment showed the importance of having the right features and the need for thorough hyperparameter tuning. Continuing experimentation seems worthwhile because there's potential to improve the model by exploring more data, refining features, and trying out different machine learning techniques. The current approach isn't a dead end; it's just the first step in understanding what works and what doesn't for flight fare predictions. With adjustments and further testing, the model can become more accurate and useful for budget planning and fare comparisons.

4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. Based on the results achieved, here are potential next steps and experiments, along with their expected uplift and ranked by potential impact: 1. Advanced Feature Engineering: Creating more sophisticated features, especially those capturing temporal dynamics and external factors like holidays and events, could lead to significant improvements in prediction accuracy. 2. Hyperparameter Optimization: Fine-tuning hyperparameters with methods like Bayesian optimization could result in a better-fitting model and improved RMSE. 3. Ensemble Learning: Combining the predictions of multiple models could capture a broader range of patterns in the data and reduce the likelihood of overfitting. 4. Expand Data Collection: Collecting additional data, such as more recent or diverse datasets, could help the model generalize better to current market conditions. 5. Model Complexity Evaluation: Investigating whether the model complexity is appropriate (neither too simple nor too complex) may provide gains in accuracy. 6. Alternative Algorithms: Testing out other regression algorithms might offer improvements or provide a benchmark to understand the current model's performance. 7. Data Quality Improvement: Further cleaning and preprocessing might yield some performance benefits, especially if there are still issues with noise in the data. If the experiment already achieved the required outcome for the business, the steps to deploy this solution into production would include: 1. Model Finalization: Finalize the model by training it on the full dataset or a representative subset. 2. Performance Monitoring: Establish monitoring for the model's performance over time to catch any degradation in accuracy. 3. Deployment Infrastructure: Set up the infrastructure to serve the model, which could be through a cloud-based platform or on-premises servers. 4. Integration: Integrate the model into the existing product or service, ensuring it can handle user requests and provide predictions in real-time. 5. User Documentation: Prepare documentation and support materials for users to understand how to interpret and use the predictions. 6. Continuous Learning: Implement a system for the model to update and retrain

	periodically with new data to stay accurate. 7. Fail-safes: Have backup systems or processes in place in case the model fails or provides unexpected results.