EXPERIMENT REPORT

Student Name	Varun Singh Chhetri
Project Name	Airfare Prediction using Gradient Boost
Date	10th November 2023
Deliverables	<chhetri_varun_24711703_Gradient_Boost > <Gradient Boost>

---

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 project's primary goal is to build a data product (streamlit app) that can accurately predict the local travel airfare in the USA with a high degree of accuracy. This will help the customers to quickly check the airfare prices with a click on their devices. The accurate predictions also help to gain customer trust and suggest that businesses can make data-driven decisions. On the other hand, airlines and travel agencies can use this model to understand their customers' nature better. This can help them create customised pricing strategies, refine pricing according to the competitors or provide better pricing schemes to attract customers during off-seasons. Incorrect results can directly hamper the customer and business. Inaccurate price predictions will make the customers unhappy, and they could probably explore other forms of travel if they do not fit into their budget. For businesses, it could mean they have wrongly understood their customer nature and pattern, meaning the chances of customised strategies won't work according to the expectations. Hence, the predictions must be accurate.
1.b. Hypothesis	The project aims to accurately predict the airfare prices, using the provided input parameters which relate the input features to the dataset. Firstly, from the customer perspective, the ability to accurately predict fares would enable predictions based on real-time data, leading to more cost-effective travel fares. This could also benefit a broader base of travellers. Secondly, for businesses, this could help in identifying loopholes in their businesses and better manage their resources, which would in turn, lead to profitability in the business. In summary, the project aims to signify the use of a data-driven approach for businesses, which can not only benefit the business but also the customers.

1.c. Experiment Objective	The expected outcome from the given experiment is to correctly construct a predictive model than and accurately predict the airfares and at the same time its usability in a user-friendly streamlit application. It should be scalable to handle high volume of data and its adaptability to given accurate predictions on unseen or new data, using historical learning. If the model provides accurate predictions and smoothly integrates with the streamlit app, then the model can be used for business, but only after testing the model in real time data. If the model accuracy is not high, it could mean that the model is not understanding the data very well, meaning either additional data is needed to be provided or changes are needed to be made on the models trained and how the data is being prepared.

---

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	During the data preparation phase, common data was prepared, which was then further used for individual modelling. Only selective columns were selected for the modelling phase during the individual preparation phase. Columns: starting and destination airport, total travel distance, total travel duration, cabin type and departure time were used. Firstly, these columns were correctly formatted to their types by analysing the tables. Secondly, only these columns were selected because there was a limitation on user input. Since the user would only input five parameters, these were the maximum number of columns that could be extracted based on the user input. Specific additional columns were also added in the feature engineering part. During the initial ingestion, the missing values were replaced by zero, and since the data had nearly 13.5 million rows, I dropped columns where the total distance and duration were 0. Some other rows were also dropped, where specific columns have values of 0. The final raw count before data preparation was nearly 12.1 million rows. Also, the model was needed to be deployed and hence a pipeline was created so that when the model runs for predictions it does some pre-processing and then modelling. This would help in much smoother integration of model across different platforms.
2.b. Feature Engineering	Using the panda's datetime feature, the departure time column was converted to datetime. The reason for doing this was because of this column; five new features could be extracted. The extracted features were: ‘flightmonth’, ‘flightday’, ‘flightdayoftheweek’, ‘flightweekofyear’, ‘flightisweekend’. Since the user would be inputing the departure time, adding these features would help the model train better. These features help the model to understand seasonal trends and price variations during peak and off-peak seasons. Also, using the flight departure column, a new column for departure time was created where the time was subcategorised into 4 slots: morning, afternoon, evening, and night. This would help the model understand if there is price variation during different times of the day and if pricing changes according to the time of the day. It was also observed that there is a price fluctuation when total distance and total duration are higher, indicating a direct relation with the target variable.

2.c. Modelling	The gradient boost model is an ensemble machine learning model that builds a predictive model stage-wise. The core principle behind GBM is constructing new base learners that are maximally correlated with the negative gradient of loss function associated with the whole ensemble. Secondly, since the data is quite large, we need to train a mode that is computationally efficient and, at the same time, provides good results. The hyperparameters tuned for Gradient boost regressor are: 1. 'Loss': Optimises the loss function. Since we were doing a regression problem, it was set to 'squared_error'. 2. 'Learning_rate': It is the contribution of each tree, set to 0.1. 3. 'N_estimators': The number of boosting stages was set to 100. 4. 'Criterion': To measure the quality of the split, it was set to friedmen_mse for calculating the mean squared error. 5. 'Min_sample_split' and 'min_sample_leaf: They were set to their default values of 2 and 1 respectively. 6. 'Max_depth': The maximum depth of the individual regression estimator was set to 3. 7. 'Sub_sample': To use all the samples, it was set to 1. All the categorical columns were encoded using one-hot encoding to provide equal importance to each value. It is important to note that the model was trained using their standard values of hyperparameters as the size of the data was affecting the runtime of the model. Techniques like Gridsearchcv and model stacking can be applied which would increase the model scores, but at the same time it would be more computationally resource intensive.

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	Baseline: 1. MSE: 44,166.24 2. RMSE: 210.15 3. MAE: 156.83 Gradient Boost: 4. MSE: 19159.73 5. RMSE: 138.41 6. MAE: 95.99 • Mean Squared Error (MSE): The average squared difference between the estimated values and the actual value. It gives a general idea of the error magnitude but is sensitive to outliers since the errors are squared. • Root Mean Squared Error (RMSE): The square root of the MSE. It's in the same units as the target variable, making it more interpretable. • Mean Absolute Error (MAE): The average absolute difference between the predicted values and the actual values. It provides a linear score, meaning all individual differences are weighted equally in the average.
3.b. Business Impact	• Both the MSE and RMSE are lower for the Gradient Boosting Model compared to the Baseline Model, which means that the Gradient Boosting Model has a better fit to the data and makes more accurate predictions. The MAE is also lower, indicating that, on average, the Gradient Boosting Model's predictions are closer to the actual values.

	• The Gradient Boosting Model has reduced the MSE to almost half of what the Baseline Model's MSE is, and similar improvement is seen in RMSE and MAE. This is a substantial improvement in predictive performance. • The RMSE of the Gradient Boosting Model is 138.42, which means that, on average, the model's predictions are off by this amount from the actual fare prices. This error may or may not be acceptable. For high-cost items (like expensive flights), this might be a reasonable error. For low-cost items, this might be too high. • Given that the Gradient Boosting Model significantly outperforms the Baseline Model in all three metrics, it is likely a better model for the given data and can be selected for further optimisation and testing. • In summary, the Gradient Boosting Model shows a significant improvement over the Baseline Model, indicating that it is learning from the features provided and is able to make more accurate predictions. Further model tuning, feature engineering, and possibly doing GridSearchCV or model stacking could lead to even better results. It is also important to note that the dataset is huge, and training models take a significant amount of time hence further tuning must be done where computational powers are high or the modelling is done in a small subset of data.
3.c. Encountered Issues	Firstly, the size of the dataset, consumed a lot of memory and hence the features were needed to be downsized. Secondly the formatting of rows was needed to be correctly encoded using a delimiter. Since the size of the data was, huge it could not be directly uploaded in the virtual environment, as while pushing, the data won't push. Hence the data was accessed from google drive. The data had a lot of rows, but not each row has been able to be utilized to get a higher accuracy score. Since the user input parameter was minimal, linking all the data with the parameters was not possible, thereby forcing to only use the data that can been linked or creating new features based on the input parameters. Training the model was utilizing a lot of computational resources and time. This could have been prevented using subset of data, where it could give an overview of how different models have been performing. Lastly, since my model had additional features like total distance and total travel time, the user was not providing this input, hence this information was linked with external API, to extract information using the starting and destination airport code. This issue was solved by working collaborately.

---

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 outcome of the experiment helped in gaining insights into different aspects of model performance and model deployment. • Model Performance: It provides a comprehensive view of how the model performs using the given hyperparameters and how the predicted values are fair with the actual values. Although the model performance is not highly accurate, it performs well if the parameters provided fall under the majority class, indicating that the model has trained well under the majority class but not the minority class. This could either mean the data cleaning techniques used must be tuned further to emphasise the minority classes or try different model and tune their hyperparameters. • Algorithm Performance: The scores from the model indeed indicate that tree-based mores are able to understand the complex relations of the data and work well under such circumstances.

	• Real-world Application: By successfully loading and running the model using the given input parameters, suggest that it can be integrated in real-world setting. The only downside is not all the data can be derived from input parameters and the model has to depend on external apis (AeroApi) for some values.
4.b. Suggestions / Recommendations	From the given experiment, it was observed that the steps taken during the entire process have been fruitful and in the right direction. It highlights the importance of primary features that are crucial to predicting the fares accurately. Since the data provided was not in a proper format, it is important to consider storing the raw data in correct shape and form. This will inturn help in better modelling as there can be proper pipeline setup which handles data pre-processing and data processing. The model scores also reflect the importance of size of the data, probably by adding more number of rows and columns could drastically help in improving the model performance scores. While the given model can be deployed, it is important to first examine the model in all the above-mentioned aspects and what are its effects on the business. Once the model has been deployed, it should be ensured that, it is been updated and trained with real time data to stay up to date and provide accurate predictions.