EXPERIMENT REPORT

Student Name	Sahil Kotak
Project Name	AT3 - Data Product with Machine Learning
Date	10/11/2023
Deliverables	kotak_sahil_24707592_catboost.ipynb CatBoostRegressor

---

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 overarching goal of this project is to provide users in the USA with a reliable and user-friendly data product that accurately estimates local airfares based on trip details. By leveraging advanced machine learning models, we aim to generate predictions that help travelers budget more effectively, enable travel agencies to offer more competitive pricing, and assist airlines in adjusting their fare structures based on demand and market conditions. The anticipated impacts include: • **For Travelers:** Improved financial planning with more accurate fare estimates. • **For Travel Agencies:** Enhanced service offerings through precise fare predictions, potentially increasing customer satisfaction and loyalty. • **For Airlines:** Data-driven insights into fare structures, which could inform strategic pricing adjustments. Accurate results will foster trust in the tool, increase usage, and establish the application as a go-to resource for travel planning. Conversely, inaccurate predictions could result in mistrust, financial misestimations, and potential loss of market competitiveness.
1.b. Hypothesis	The central hypothesis is that machine learning models can predict flight fares with a high degree of accuracy using historical data and real-time inputs. We believe that among the various factors influencing airfares, distance traveled, airport location, cabin class, and time-related variables (day of the week, month) are critical predictors. The experiment will test the validity of this hypothesis by evaluating the model's performance in real-world scenarios. This hypothesis is rooted in the understanding that airfare pricing is complex and multifaceted, but certain patterns and factors consistently influence pricing strategies. By identifying and effectively harnessing these factors, we expect to make accurate

	fare predictions.
1.c. Experiment Objective	The expected outcome of the experiment is a set of robust models that accurately predict airfares within an acceptable margin of error, as indicated by RMSE, MAE, and R2 metrics. We aim to achieve an RMSE of under $100 and an R2 of 0.8 or above, which would signify high predictive accuracy and reliability. Possible scenarios resulting from this experiment include: ● **Successful Prediction:** The models meet or exceed the performance goals, and the app is deployed for public use, receiving positive feedback and high adoption rates. ● **Partial Success:** The models perform adequately but highlight areas for improvement, leading to further iterations and refinements before public release. ● **Underperformance:** The models fail to meet the performance expectations, necessitating a reassessment of the feature selection, model choice, and potentially the hypothesis itself. In all scenarios, the experiment will yield valuable insights into fare prediction dynamics and guide future development and optimization efforts.

---

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	The initial data preparation involved cleaning the dataset by handling missing values, particularly in the 'totalDistance' column, which was crucial for fare estimation. Given that 'totalDistance' is a significant predictor, we tried to calculate it by adding the distance in each leg of the journey. This allowed us access to totalDistance without potentially introducing bias. This step was essential to maintain the integrity and reliability of the model predictions. We also converted categorical variables into a format suitable for machine learning models through one-hot encoding. This step was necessary to enable the models to process non-numeric data. We decided against bucketing continuous variables, as preliminary analysis did not suggest significant non-linearity or outliers that would require such transformation. For future experiments, it might be worthwhile to explore adding in more data for distance and duration using external sources.

2.b. Feature Engineering	Feature engineering was a critical step in the modeling process. • We created a hierarchy of cabin classes to capture the relative luxury level of each class, which is a significant factor in pricing. This involved mapping each cabin code to a numerical scale and then determining the highest cabin class for each flight segment. • Additionally, we processed the 'segmentsDistance' data to calculate the total distance for multi-leg flights, ensuring that our distance variable accurately reflected the total travel distance. • We decided to remove features that were too granular or had a high cardinality, such as segment data. The engineered features capturing the hierarchy of cabin classes and total travel distance are likely to be important for future experiments due to their strong influence on fare prices.
2.c. Modelling	We trained a CatBoost model for this experiment, chosen for its robust handling of categorical features and its ability to model complex relationships without extensive hyperparameter tuning. The key hyperparameters adjusted included the number of iterations, learning rate, and tree depth. We tested various values to balance model complexity and prevent overfitting. We decided not to train a simple linear regression model, as preliminary analysis suggested that the relationship between the features and the target variable was too complex for such a model to capture effectively. For future experiments, exploring other ensemble methods like Random Forest or advanced neural network architectures could provide comparative insights. Additionally, fine-tuning hyperparameters such as the model's learning rate and depth could potentially yield improvements in predictive performance.

---

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 CatBoost model achieved a **Root Mean Squared Error (RMSE) of 103.73**, a **Mean Absolute Error (MAE) of 69.80**, and an **R-squared (R2) value of 0.75**. These metrics indicate that the model has a reasonable predictive performance, though there is room for improvement, especially in reducing the RMSE to enhance accuracy. Upon closer analysis, the underperforming cases were often associated with outlier fares or unusual routes that are not well-represented in the training data. This could be due to special events, pricing anomalies, or data entry errors. A potential root cause for these inaccuracies could be the lack of contextual data that might influence fare prices, such as holidays or airline-specific promotions.

3.b. Business Impact	From a business perspective, the model's ability to predict airfare with an R2 of 0.75 suggests that it can capture most of the variability in fare prices. This level of performance can be beneficial for users trying to estimate travel costs, aiding in budget planning and decision-making. However, the errors represented by the RMSE could lead to significant misestimations for certain trips, particularly those involving routes or cabin classes that deviate from the norm. Incorrect results could lead to a loss of trust in the application, impacting user retention and the perceived reliability of the service.
3.c. Encountered Issues	During the experiment, we faced several issues: ● Data Quality and Completeness: Missing values in key variables like 'totalDistance' led to the decision to drop certain records, potentially reducing the model's training data diversity. ● Model Complexity: Tuning CatBoost's parameters to avoid overfitting while still capturing complex relationships was challenging. ● Computational Resources: The amount of data and the complexity of the model required significant computational resources, which could be a limitation for continuous retraining or scaling up the service. ● Integration with External APIs: Fetching real-time data from RapidAPI introduced additional complexity, such as handling API errors or data mismatches. ● Solutions and workarounds included rigorous data cleaning, feature selection to reduce model complexity, and ensuring robust error handling when integrating external APIs. For future experiments, addressing issues like data imputation, feature encoding strategies, and computational optimizations will be crucial. Additionally, implementing a more systematic approach to hyperparameter tuning, possibly through automated methods like grid search or Bayesian optimization, could yield performance improvements.

---

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 experiment has yielded several key insights that are valuable for both technical model refinement and business strategy: ● Feature Impact: The significant influence of 'totalDistance' and airport features on fare prediction underscores the importance of geographical and logistical factors in fare determination. This insight can guide further feature engineering. ● Model Robustness: The CatBoost model's current performance highlights the need for a more diverse dataset that includes outlier scenarios and more instances of less common routes. ● Data Dependency: The reliance on quality external data, such as from AeroAPI, indicates that maintaining data pipelines and ensuring data accuracy is critical for the model's real-world application.

	These insights suggest that pursuing further experimentation with the current approach is warranted, with a focus on enhancing data quality, exploring additional features, and refining the model.
4.b. Suggestions / Recommendations	The following potential next steps and experiments could be considered: ● **Data Augmentation:** To improve model robustness, incorporating more data, especially for underrepresented scenarios, could be beneficial. ● **Advanced Feature Engineering:** Exploring non-linear transformations or interaction terms might capture complex patterns in the data. ● **Hyperparameter Optimization:** Employing a more systematic approach to tuning could lead to performance gains. ● Model Ensembling: Combining predictions from multiple models could improve accuracy and reliability. ● **Deployment Considerations:** Before deploying the model into production, ensuring infrastructure readiness, establishing continuous integration and delivery pipelines, and setting up monitoring for model drift are crucial. Each of these steps should be assessed for expected uplift or gains, with a prioritization based on the potential impact on model performance and business value. If the model meets business requirements, I recommend beginning the process to deploy it into production, alongside developing a user feedback mechanism to continuously refine the model based on real-world usage.