EXPERIMENT REPORT

Student Name	Shivatmak Sharma
Project Name	ML Data Product
Date	11/11/2023
Deliverables	train_model_Shivatmak.py best_model_Shivatmak.py main_Shivatmak.ipynb

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 primary goal of this project is to develop a reliable and efficient predictive model for forecasting flight fares. Accurate predictions would enable travelers and businesses to make informed decisions, optimize pricing strategies, and enhance budget planning. Accurate results would strengthen customer trust and decision-making, while incorrect results could lead to misguided estimations, impacting customer satisfaction and business strategies.
1.b. Hypothesis	The hypothesis is that Long Short-Term Memory (LSTM) networks can effectively model and predict flight fares by capturing temporal dependencies and patterns in historical data. Given the sequential nature of flight data (time series), it is anticipated that an LSTM model would perform well in this context.
1.c. Experiment Objective	The expected outcome is to develop a LSTM model that provides accurate and reliable flight fare predictions. The performance goal is set in terms of minimizing prediction errors, aiming for high accuracy (low RMSE) and strong predictive power. Possible outcomes include a successful model meeting performance criterion, or identification of the need for further refinement.

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	A comprehensive data pre-processing approach was implemented. Central to this was the DataPreprocessor class, designed to merge datasets and expand multi-entry columns, thus accurately reflecting the complexities of travel routes. Key to our process was parsing time and date features, like hours and weekdays, from crucial fields, allowing the model to effectively identify temporal patterns. Categorical data, such as airport codes, were meticulously converted into numerical formats compatible with the LSTM's embedding layers, with special attention to managing unknown categories. To ensure data integrity, we rigorously handled missing values and duplicates, guaranteeing dataset cleanliness and model efficacy. Addressing outliers, we normalized fare values to their mode fares for each group, thus stabilizing the prediction target against extreme variations. Our pipeline also included type optimization for efficiency, down casting where feasible to boost performance during training. Additionally, imputation, scaling, and encoding strategies were applied to maintain the dataset's quality. We enriched the dataset with contextual benchmarks like median travel distances and durations, equipping the model with essential referential statistics for more accurate fare estimations. Finally, we meticulously saved the pre-processed data, ensuring the predictive system remained updated and accurate, demonstrating our commitment to a robust and reliable flight fare prediction model.
2.b. Feature Engineering	In the LSTM model's feature engineering process, significant attention was given to parsing time and date features, such as hours and weekdays, from key data fields. This enabled the model to capture and utilize temporal patterns effectively. Categorical data, like airport codes, were transformed into numerical formats compatible with the LSTM's embedding layers, with special handling for unknown categories to enhance model resilience. This approach effectively consolidated varied datasets into a unified source, vital for the LSTM's performance. Through the segmentation and expansion of columns, the model adeptly managed complex data, enabling the extraction of a rich set of features. This thorough preprocessing and feature engineering were key to boosting the LSTM model's performance, enhancing its predictive accuracy and precision.

2.c. Modelling	The LSTM model was constructed with a combination of categorical and numerical inputs. The model incorporated an LSTM layer with 50 neurons, utilizing ReLU activation to capture intricate data patterns and a dropout layer with a 0.2 rate for regularization. The Adam optimizer was used for training, with Mean Squared Error as the loss function. The model was trained over 50 epochs with batch sizes of 26024. To address the challenge of imbalanced data, the model adjusted fare data to mitigate skewness and included dropout layers to promote balanced learning.

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 LSTM model achieved an RMSE of 160.97 and an MAE of 109.32. This suggests a good performance but still has room for improvement. The model's ability to process time-series data was a significant factor in capturing temporal trends, which is reflected in these metrics.
3.b. Business Impact	From a business standpoint, the LSTM model's performance in processing time-series data is particularly valuable for capturing seasonal fare adjustments. This capability is essential for airlines and travel agencies in optimizing pricing strategies and anticipating market trends. The results indicate that further refinement, such as hyperparameter tuning, could enhance the model's utility in a competitive business environment.
3.c. Encountered Issues	Computational difficulties with such a large dataset and performing deep learning on it without any cloud platform. Future iteration on LSTM model could focus on refining hyperparameters, incorporating more diverse temporal features, and exploring advanced techniques like bidirectional LSTMs or attention mechanisms to enhance its ability to model complex time-related dependencies in fare 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	The experiment demonstrated the viability of using LSTM models for flight fare prediction. The insights gained in handling time series data and feature importance are crucial for refining the model further.
4.b. Suggestions / Recommendations	Future steps include experimenting with more complex LSTM architectures, incorporating additional external data (like economic indicators), and exploring ensemble methods. If the model meets business requirements, steps towards production deployment, like integration with live data streams and user interface optimization, are recommended.