Name: Akhila Mediboyina

Bid : B00115521

Sales prediction in the tourism industry

Question 1: Elevator Pitch

Our project focuses on sales prediction in the tourism industry, leveraging historical booking and sales data to forecast future trends. By employing advanced machine learning models, we aim to empower stakeholders with actionable insights to optimize revenue, enhance customer experiences, and plan resource allocation effectively.

Question 2: Dataset Details

1) Collector(s):

Kaggle, World Tourism Organization (UNWTO), or a specific government tourism board.

2) Year:

The dataset spans a range of years, such as 2010–2024, depending on the chosen source.

3) Title of Dataset:

"Tourism Sales Data Across Regions"

4) Version Number (if any):

If applicable, use version numbers provided by the dataset publisher, e.g., v1.2.

5) Publisher:

Kaggle, UNWTO, or a government tourism department.

6) DOI or URL:

- Example Kaggle Dataset: https://www.kaggle.com/
- UNWTO Data: https://www.unwto.org/statistics

7) Study/Paper/Reason:

The dataset was collected to analyze trends in tourism sales, assess seasonal effects, and improve marketing strategies for the industry.

Question 3: Language and Libraries

Language:

- Python 3.13.1

Libraries:

- **Data Preprocessing:** Pandas, NumPy
- **Visualization:** Matplotlib, Seaborn, Plotly
- **Model Development:** Scikit-learn, TensorFlow/Keras, PyTorch
- **Evaluation:** Scikit-learn
- **APIs/Data Access:** Requests for API data or SQLite for local datasets

Question 4: Code will Write Our Own

- **Data Preprocessing:** Cleaning, encoding, and normalizing sales and seasonal data.
- **Feature Engineering:** Creating features like holiday periods, weather, promotions, and regional factors.
- **Model Development:** Implementing regression models or time-series models (e.g., ARIMA, LSTM).
- **Hyperparameter Tuning:** Writing scripts to test different hyperparameter combinations.
- **Performance Evaluation:** Creating custom functions to calculate and display error metrics.
- **Visualization:** Developing dashboards or plots for sales trends and predictions.

Question 5: Best Choice of Model

Model Choice:

- **Time-series Models:** LSTM, GRU, or ARIMA.
- **Regression Models:** XGBoost, Random Forest, or Gradient Boosted Trees.
- **Why:** Time-series models are ideal for sequential data, while regression models handle categorical and numerical predictors effectively, making them versatile for tourism sales data.

Question 6: Hyperparameters and Optimization

Key Hyperparameters:

1. Learning Rate
2. Batch Size
3. Number of Layers (for deep models)
4. Number of Units (LSTM/GRU)
5. Sequence Length (for time-series models)
6. Seasonal Order (for ARIMA)

Optimization Strategy:

- Begin with grid search or random search for basic tuning.
- Use AutoML libraries or Bayesian optimization for refined tuning.
- Validate hyperparameter choices using k-fold cross-validation.

Question 7: Performance Evaluation

Metrics:

1. **Mean Absolute Percentage Error (MAPE):** To measure prediction accuracy relative to sales volume.
2. **Root Mean Squared Error (RMSE):** To assess overall prediction error.
3. **R-squared (R^2):** To determine model fit quality.

Techniques:

- Evaluate predictions on seasonal and regional subsets of the data.
- Visualize actual vs. predicted sales trends to identify anomalies.
- Test the model on unseen data to ensure generalization capability.