Report on Data Science

Dataset Description

The dataset contains 1918 entries with the following columns:

1. **Date**: The date of passenger data (string format, needs conversion to datetime).
2. **Local Route**: Passenger journeys for local routes.
3. **Light Rail**: Passenger journeys on light rail services.
4. **Peak Service**: Passenger journeys during peak hours.
5. **Rapid Route**: Passenger journeys for rapid routes.
6. **School**: Passenger journeys for school services.
7. **Other**: Miscellaneous journeys (contains some missing values).

1. Key Insights from the Dataset

1. General Trends: Local and Rapid Routes Dominate Usage

Observation: The Local Route and Rapid Route service types consistently record the highest number of passenger journeys across the dataset. Their mean daily usage is approximately 9,891 and 12,597 journeys, respectively.

Inference: These services are likely the backbone of the transport system, covering essential and frequently used routes. The high numbers suggest a stable demand for daily commuting, possibly due to their coverage of urban and suburban areas.

---

2. School Services Show High Variability

Observation: Passenger journeys for School services range from 0 to a maximum of 7,255. The median value is only 568, indicating a skewed distribution with frequent low usage.

Inference: The large variability is indicative of a strong dependence on school term calendars and holidays. On days when schools are open, usage spikes, suggesting efficient targeting of this demographic. However, low median values imply that these services might be underutilized on non-school days, raising questions about cost-effectiveness during those times.

---

3. Peak Services: Small Contribution with Notable Spikes

Observation: The Peak Service category averages 179 passenger journeys per day, with occasional spikes to 1,029 journeys.

Inference: These spikes align with rush hours, suggesting that Peak Services effectively target high-demand periods. Despite this, their overall contribution to daily totals is small, indicating potential room for optimization to address under-utilization outside peak times.

---

4. Light Rail: Steady Usage with Seasonal Increases

Observation: Light Rail services maintain moderate and steady usage, with an average of 7,195 journeys per day. Usage shows periodic increases, possibly tied to seasonal or event-based demand.

Inference: The steadiness reflects its reliability and possibly its role as a feeder service. However, periodic increases hint at opportunities to align scheduling or capacity adjustments with peak seasons or special events.

---

5. Seasonal and Event-Driven Fluctuations

Observation: Passenger journeys for most service types show clear fluctuations corresponding to seasons, holidays, or specific dates.

Inference: These patterns emphasize the impact of external factors, such as festivals, school vacations, or weather conditions. By analyzing such trends further, the transport authority can better allocate resources and manage capacity.

Suggestions for Improvement:

Resource Optimization for School Services: Introduce dynamic scheduling to cater to school terms and holidays to reduce under-utilization.

Enhance Peak Services: Expand coverage during high-demand hours to capture a larger share of commuters and reduce congestion on other services.

Promote Light Rail Usage: Develop campaigns or offers to drive consistent usage, particularly during non-peak or off-season periods.

Leverage Data-Driven Decisions: Use advanced analytics to anticipate high-demand periods and allocate resources efficiently.

2. Forecasting

Model Overview: Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN) capable of learning long-term dependencies in sequential data. Traditional machine learning models, such as linear regression, often struggle with sequential data as they lack memory to retain information from earlier data points. LSTM, however, is designed to address this issue by using gates to control the flow of information over time, allowing the model to remember important past data and forget irrelevant data.

Steps Taken for Forecasting using LSTM

- **Data Preprocessing**

Data preprocessing is a critical step in ensuring the quality of input data for the model. Tasks performed include converting dates, sorting data, handling missing values, and outlier removal.

- **Outlier Removal**

Outliers can significantly affect the model's ability to learn and make accurate predictions. In time series data, outliers might result from incorrect data entries or rare events. We used the Interquartile Range (IQR) method for detecting and removing outliers.

The IQR is the range between the first quartile (25th percentile) and the third quartile (75th percentile). Any data point outside the range:

Lower Bound: Q1 - 1.5 × IQR

Upper Bound: Q3 + 1.5 × IQR

These points are considered outliers and are removed from the dataset. This process ensures that any extreme values are removed, helping the model focus on actual trends.

- **Data Normalization**

LSTM models are sensitive to the scale of input data, so it is crucial to normalize the dataset. MinMaxScaler is used to scale the data to a range of [0, 1], helping the model learn efficiently.

- **Creating Time Series Sequences**

LSTM models require the data to be formatted as sequences, where the input data is a series of previous observations used to predict the next value. A time step (e.g., 10) is defined to specify how many previous time points are used for each prediction.

- **Splitting the Data**

The dataset is split into training and testing sets, with the training set comprising 80% of the data and the testing set the remaining 20%. This split allows the model to be trained on past data and evaluated on unseen data.

- **Building the LSTM Model**

The model consists of:

LSTM Layer: This layer has 50 units and is set to return a single output value.

Dense Layer: This layer produces the output (predicted passenger journey value).

- **Training the Model**

The model is trained using the Adam optimizer and mean squared error as the loss function. We train for 20 epochs with a batch size of 32.

- **Making Predictions**

After training the model, we use it to predict future values. The model is given the last time_step number of observations to predict the next 7 days.

- **Inverse Transformation**

Since the data was scaled, the predicted values are transformed back to their original scale using the inverse transformation.

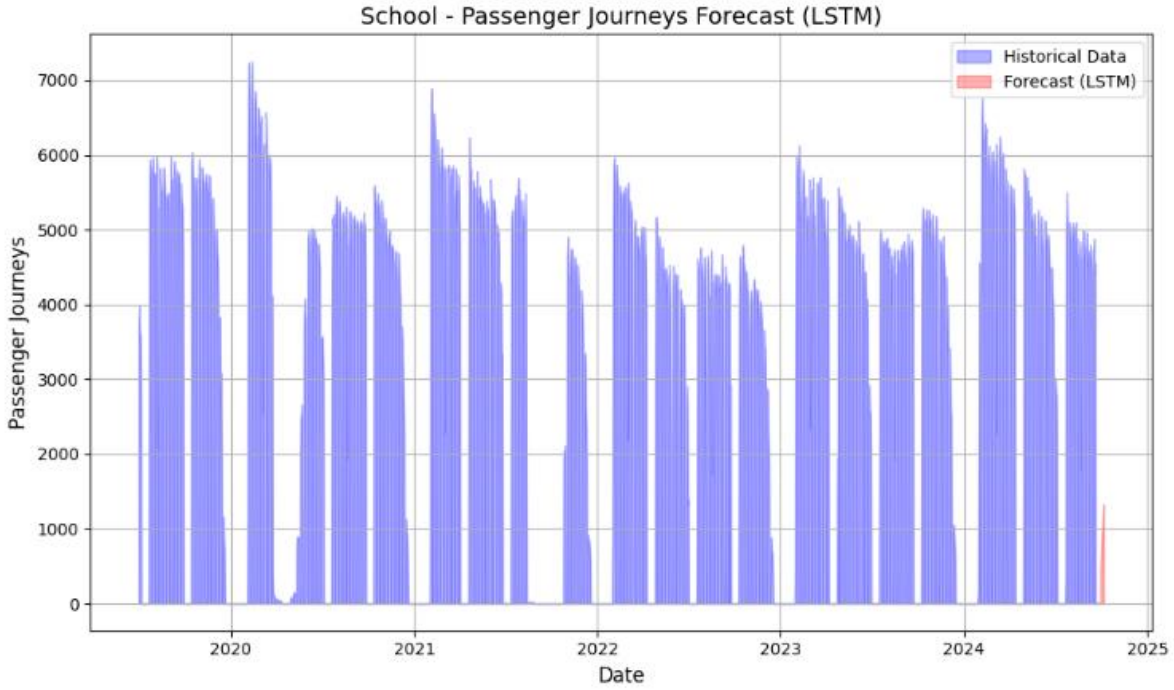
- **Visualization of Predictions**

Finally, the model's predictions are compared with the historical data in a graph. This provides a visual representation of how well the model has learned the patterns and made predictions.

Output Explanation

The model's predictions are compared with historical data using graphs. The blue line represents the historical data, while the red dashed line shows the forecasted values.

The table below displays the predicted values for the next 7 days. Note: Negative values might indicate parameter issues, which can be resolved through hyperparameter tuning.



Date	Predicted Passenger Journeys
2024-09-30	468
2024-10-01	642
2024-10-02	799
2024-10-03	948
2024-10-04	1088
2024-10-05	1217
2024-10-06	1330