REPORT

Title: Traffic Prediction Using Machine Learning Techniques

Introduction

The aim of this project is to forecast traffic volume(or the number of vehicles) at different junctions using a dataset from Kaggle. This dataset includes features like date, time, and junction-specific vehicle counts. By processing, visualizing, and analyzing these features, we aim to develop an effective model to predict future traffic volume.

Data Loading and Preparation

2. Date and Time Feature Engineering

- The date-time column is converted into multiple new features, such as year, month, day, hour, and day of the week. This transformation allows us to better capture temporal patterns and trends in traffic volume.

Data Visualization and Analysis

Distribution of Traffic Volume :

- Using a histogram, the distribution of vehicle counts is visualized. This step helps identify potential outliers, skewness, or irregularities, which are essential for preprocessing and normalizing data in later stages.

Time-Series Analysis at a Junction :

- Line plots are created to analyze traffic trends over time, split by junctions. This reveals any location-specific patterns in traffic volume and helps distinguish trends.

Hourly, Daily, and Monthly Patterns :

- Bar plots and line plots are used to assess traffic patterns by hour, day, and month, identifying peak traffic times and seasonal variations.
- Boxplots are also used to assess outliers in traffic volumes by junctions, which help highlight data points that might need further scrutiny.

Temporal Patterns and Rolling Averages :

- Using rolling means and standard deviations, the traffic patterns are smoothed, revealing underlying trends that are less influenced by short-term fluctuations.
- Separate plots for each junction over a limited timeframe allow for a focused analysis of traffic dynamics and seasonality.

Difference Analysis :

- A difference feature is created using a 7-day interval, revealing weekly patterns in traffic volume.
- This transformation highlights short-term changes and is useful for identifying sudden spikes or drops in traffic that could signal events or anomalies.

Data Preprocessing ➔

Missing Data and Descriptive Statistics :

- A quick check reveals no missing data, which simplifies the processing stage.
- Descriptive statistics offer insights into the range, mean, and distribution of each feature, guiding normalization and standardization steps.

Normalization and Feature Scaling :

- The Z-score normalization is applied to the `Vehicles` column. Although not strictly necessary, normalization ensures model training stability and allows predictions to be brought back to original scales if needed.

Correlation Analysis :

- A heatmap visually displays correlations between numerical features, helping to determine which factors have strong relationships with vehicle count. This informs feature selection in the modeling phase.

Modeling ➔

Data Splitting

- A train-test split of 70/30 is applied, with scaling applied to ensure that features fall within the same range, improving the performance of machine learning algorithms.

Linear Regression Model :

- A linear regression model is trained on the prepared data, and its performance is evaluated using Mean Squared Error (MSE). This provides a baseline prediction accuracy against which more complex models can be compared.

Linear Regression models are typically trained using a method which minimizes the **Mean Squared Error (MSE)** between the predicted and actual values. The MSE is calculated as:

where:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- yiy_iyi is the actual value of the target for the iii-th data point.
- y^i\hat{y}_iy^i is the predicted value from the model.
- NNN is the number of data points.

Multi-Layer Perceptron (MLP) Neural Network :

- A customized MLP model is defined to capture more complex, non-linear patterns in the data.
- The model architecture includes multiple layers with dropout for regularization.
- The model is compiled with Stochastic Gradient Descent and a mean squared error loss function.
- Early stopping is implemented to avoid overfitting, enhancing model generalization on unseen data.

MLP Model Training and Evaluation

Training Process :

- The MLP model is trained with an early-stopping criterion to monitor validation loss. The training and validation loss are stored for visual analysis.

Performance Evaluation :

- Mean Absolute Error (MAE) is used as the evaluation metric, reflecting the average error between predicted and actual traffic volumes.
- A plot of training and validation loss over epochs helps illustrate the model's learning curve and identify any overfitting or underfitting trends.

PLAY GROUND FOR ACTUALLY PREDICTING:

This is in the last part of code where you can actually use the trained both models to predict and also find the short time from a play toy example {A,B,C,D}

⇨ First it takes aan input of necessary features

Now

1)it calculates for LR trained model and predicts for the input YOU give:

2) it calculates for MLP original trained model and predicts for the input YOU give:

3) finally it uses a simple algorithm to find the shortest distance of the given easy path

Conclusion

This report walks through the traffic prediction pipeline, from data loading to feature engineering, visualization, and model training. The linear regression model offers a baseline prediction, while the MLP model leverages non-linear relationships to improve accuracy. Each step enhances understanding of traffic patterns, empowering data-driven insights for urban planning or traffic management solutions.