Predictive Analytics Lab Project: Real-Time Traffic Congestion System

Traffic Volume Prediction Report

Submitted to- Achalya Shakaya

Submitted by- Anshika Nagar

-500110785

Vyom Verma

-500105694

Introduction

Traffic congestion poses a significant challenge in urban areas, impacting travel efficiency and contributing to poor air quality. Effective prediction of traffic volume can empower city planners and traffic management systems to make data-driven decisions aimed at alleviating congestion. This report details an analysis of historical traffic data using a linear regression model to predict vehicle volume based on time-related features.

Objective:- The primary goal of this project is to predict and manage traffic congestion by leveraging historical and real-time data, optimizing traffic flow, and minimizing delays.

Data Overview

The analysis utilized a dataset obtained from `traffic.csv`, which comprises historical traffic volume records. Initial data examination focused on identifying missing values, which were addressed using forward filling to maintain continuity in the time series. Subsequently, the `DateTime` column was transformed into two essential features: the hour of the day and the day of the week. These features are crucial, as they significantly influence traffic patterns.

Exploratory Data Analysis

To gain insights into the dataset, a time series plot was generated to illustrate traffic volume trends over time. This visualization facilitated the identification of peak traffic hours and variations across different days. Additionally, a correlation matrix was created using Seaborn, highlighting relationships between the number of vehicles, hour, and day of the week. The analysis indicated that both the hour and day have a substantial impact on traffic volume, emphasizing the importance of these features in our predictive model.

Role of Data Structures

1. Priority Queues: Manage dynamic routing and traffic signal control effectively.

2. Graphs: Model road networks, where nodes represent intersections and edges represent roads, facilitating shortest path and congestion calculations.

3.Hash Maps: Enable efficient storage and retrieval of traffic data based on geolocation.

4. Queues: Model waiting times at intersections and toll gates to optimize traffic flow.

Python Libraries

1. Pandas: Used for data manipulation and preprocessing, efficiently handling large datasets like traffic volume and sensor data.

2. NumPy:Facilitates numerical operations and the handling of arrays of traffic data.

3. Scikit-learn: A robust machine learning library for building and training predictive models, including linear regression.

4. Matplotlib: Provides visualization capabilities for illustrating traffic flow and congestion prediction results.

5. GeoPandas: Handles geospatial data, such as road networks and sensor locations, enhancing spatial analysis.

Methodology

1. Data Collection: Historical and real-time traffic data were collected from sensors and cameras.

2. Data Preprocessing: Data was cleaned and normalized using Pandas and NumPy to prepare for analysis.

3. Model Building:Regression models for traffic prediction were developed using Scikit-learn.

4. Real-Time Prediction: Models were deployed to process real-time data, enabling dynamic adjustments to traffic flow.

Use Cases

1. Smart Traffic Lights: Traffic light timings are adjusted based on predicted congestion levels to enhance flow.

2. Navigation Systems: Provide real-time alternate routes to drivers, helping them avoid congested areas.

Benefits

1. Reduced Congestion:Optimized traffic flow leads to fewer delays and reduced bottlenecks.

2. Environmental Impact: Decreased fuel consumption and emissions due to smoother traffic conditions.

3. Improved Safety: A more consistent traffic flow reduces the likelihood of accidents.

Challenges

1. Data Accuracy Managing noisy and incomplete real-time traffic data poses significant challenges.

2. Scalability: Efficiently processing large datasets in real-time requires robust infrastructure and methodologies.

Model Development

A linear regression model was chosen for this analysis due to its simplicity and effectiveness in predicting continuous outcomes. The dataset was split into training (70%) and testing (30%) subsets using the `train_test_split` function from Scikit-learn. This division allows for training the model on one portion of the data while validating its predictive accuracy on another.

The model was trained using the training set, where the independent variables comprised scaled features (hour and day of the week), while the dependent variable was the number of vehicles. After training, predictions were made on the testing set, enabling evaluation of the model's performance.

Real-Time Prediction

To assess the model's applicability in real-world scenarios, a sample input was constructed representing a typical weekday at 2 PM. This input was processed through the same scaling method employed during training, leading to a traffic volume prediction. The predicted volume was then compared to a predefined traffic threshold of 0.8. If the predicted volume exceeded this threshold, the model recommended adjustments to traffic light timings to mitigate congestion; otherwise, the current traffic flow was deemed normal.

Conclusion

This report demonstrates the successful application of linear regression for predicting traffic volume based on time-related features. Insights from the exploratory analysis, combined with the model's predictive capabilities, provide valuable information for traffic management systems. Future enhancements could focus on improving prediction accuracy by exploring more complex models, integrating additional features such as weather conditions, and utilizing real-time data feeds. These advancements would contribute to more effective traffic management strategies, ultimately leading to improved urban mobility and reduced congestion.