Metro Traffic Volume Forecasting

-

A Machine Learning Project

Noémie MAZEPA

Aymeric MARTIN

Lorrain MORLET

Auriane MARCELINO

DIA5 ESILV

Outline

1/ Business challenge and state-of-the-art

Nowadays, cities are increasingly relying on public transportation to manage the flow of people and reduce congestion. One of the key challenges for urban planners and transportation authorities is predicting the volume of traffic at different times of the day, especially on major interstate routes. By analyzing a dataset on Metro Interstate Traffic Volume. We aim to predict the number of people using public transport based on various factors such as the time of day, weather conditions, and the date. This information is crucial for optimizing public transport schedules, ensuring efficient resource allocation, and improving the overall commuter experience by anticipating periods of high affluence.

The state-of-the-art in transport demand prediction combines advanced machine learning techniques with time series forecasting models to provide accurate and scalable solutions. Traditional methods like ARIMA and its seasonal variant, SARIMA, are used for time series forecasting but may struggle with non-linear relationships. Machine learning models such as Random Forests, Gradient Boosting (e.g., XGBoost), and Support Vector Machines (SVM) are effective for handling complex relationships and large datasets, although SVMs may not scale well for time series problems. Deep learning models like LSTM and GRU excel at capturing long-term dependencies in time series data, while CNN-LSTM hybrids are powerful for complex spatial-temporal dependencies.

Many state-of-the-art models also incorporate external data sources like weather information and event schedules to improve prediction accuracy. Modern approaches use real-time data streams for dynamic, real-time predictions, employing adaptive learning algorithms that continuously update based on new information. However, challenges remain in scalability, data availability and quality, and model interpretability.

Addressing these challenges is crucial for fully realizing the potential of state-of-the-art techniques in operational environments, leading to better service planning, resource allocation, and customer satisfaction. Some work has already been done in this area. For example, five years ago, Ramya H R (now a project coordinator at Rakuten) created a notebook to analyze the dataset and implemented a simple machine learning model (XGBoost) to predict traffic volume. She achieved a Root Mean Square Error (RMSE) of around 1015, which is not bad. Additionally, she conducted extensive analysis of the dataset, providing valuable insights that can be built upon.

2/ Data description

The Metro Interstate Traffic Volume dataset consists of hourly traffic volume data for westbound I-94 in the Minneapolis-St. Paul, Minnesota area. It was gathered between the years 2012 and 2018 and has a total dimensionality of 48,204 instances with eight features and one target variable. This is both a time-series and multivariate problem whose principal goals will be to understand the tendencies of traffic in terms of weather and holiday influences. Some of the key features are the temperature in Kelvin, precipitation, rain and snow in millimetres, cloud coverage, and categorical weather conditions. It also contains additional information on US National holidays and regional events such as the Minnesota State Fair. The target variable will be the traffic volume, explained as the hourly traffic count on the given highway section.

These data were collected from the MN DoT using ATR station number 301. There is also a variation of weather conditions to capture the effects of weather on the flow of traffic; the variables range from categorical descriptions of the weather conditions down to precise measurements of rain and snow. This broad dataset is a perfect source for regression tasks, focused on the pattern of traffic volume to assess the roles of environmental and temporal factors that affect traffic flow, with no values missing.

3/ Business objectives and the scope

This dataset aims to meet the needs of public transport demand forecasting, particularly for metro and bus services, based on several influential factors such as time of day, holidays, special events, and weather conditions. The main objective is to enable transport agencies to anticipate user demand and adjust schedules and service frequency in real-time.

By leveraging this data, transport agencies can optimize operations by adjusting schedules to match peak usage periods. This not only enhances service efficiency but also ensures greater user satisfaction by reducing wait times and preventing vehicle overcrowding during peak hours.

The dataset includes various temporal aspects (time of day, days of the week, holidays, local events) and environmental factors (weather conditions like rain, snow, or temperature) that affect transport demand. By integrating historical traffic volume data with real-time data on events and weather, this forecasting model enables better resource planning, such as the number of vehicles and personnel needed. It is a valuable tool for public decision-makers, transport agencies, and urban planners, aiding in well-informed decisions for daily service adjustments as well as long-term infrastructure investments and development.

In summary, this dataset allows public transport services to adapt to user needs in real-time, making the system more responsive, efficient, and pleasant for users while optimizing resource allocation for transport operators.

4/ Workplan

Sprint 1: Implementation of Standard Solutions From 01/11 to 10/11 (*Deadline: 20/11)*

- Clean the dataset by handling missing values and outliers. (Noémie and Auriane)
- Perform an Exploratory Data Analysis (EDA) and create new features. (Noémie and Auriane)
- Implement simple models seen in class. (Aymeric and Lorrain)
- Fine-tune hyperparameters and use evaluation metrics to improve model performance. (Aymeric and Lorrain)

Sprint 2: Improving the Standard Solution From 11/11 to 20/11 (*Deadline: 20/11)*

- Implement new and more advanced models which can improve prediction accuracy. (Noémie and Aymeric)
- Analyze the performance of the model and compare it to the models from Sprint 1. (Auriane)
- Combine several algorithms to produce more accurate predictions. (Lorrain and Auriane)

Sprint 3: More Improvements From 21/11 to 11/12 *(Deadline: 11/12)*

- Choose an algorithm outside of the course and justify our choice with scientific papers. (Noémie and Auriane)
- Implement the algorithm and compare it to the previous models. (Lorrain)

- Compare results with already existing solutions on Kaggle and discuss how our model improves on the previous results or not. (Aymeric)

Conclusion

In this document, we have written down the basic information to lead a qualitative project and to develop a solid metro traffic volume forecasting model. Next, our work will focus on analyzing our dataset and implementing several models, testing their performance, and seeing how they could be applied in real-world traffic management. Our goal is to develop a tool that helps transport authorities improve efficiency and manage resources better. Finally, we want to help create a practical solution for more sustainable urban transportation.

References

Metro Interstate Traffic Volume Dataset:

https://archive.ics.uci.edu/dataset/492/metro+interstate+traffic+volume

Notebook on this dataset:

https://www.kaggle.com/code/ramyahr/metro-interstate-traffic-volume

ARIMA and SARIMA:

https://openclassrooms.com/fr/courses/4525371-analysez-et-modelisez-des-series-temporelles/5001226-les-processus-non-stationnaires-arima-et-sarima

https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average