IS597MLC-SP24: Final Project Proposal

NetID: sshah023
Student Name: Shrey Shah

Title

"FastTrack Forecast": Predicting Driving Speed Violations with DeepAR

Motivation & Objective

In today's world, as technology has advanced a lot, the production of automobiles and the number of people owning vehicles have increased. In 2023, it was estimated that there would be 1.47 billion vehicles worldwide. Due to the increase in the vehicle count, there is always a risk of increased cases of accidents, road rage, and traffic speed violations. There is a need to cultivate a safe environment for drivers and pedestrians. The simple aim of the project is to forecast any speeding violations based on the data collected over time.

Throughout my journey of learning with respect to machine learning algorithms and practices, I have explored different domains through academic and self-learning projects ranging from classification, clustering, association rule mining, computer vision, and natural language processing. I have never worked with time-series-related data. Hence, my main motivation behind the project was to explore time-series analysis and work with some algorithms. The data related to speed violations interested me, as this is a concerning matter in today's world, and the data collected can reveal a lot of useful patterns that can be used to make smarter decisions for traffic and city planning.

There are some specific analytical questions that I would like to find answers to:
1) The initial task is to forecast the speed violations; can we predict the likelihood of speed camera violations at specific intersections based on historical violation data?
2) Are there spatial patterns or clusters of speed camera violations across different intersections?
3) How do speed camera violation rates vary over time, such as daily, weekly, or seasonal patterns?

Related Articles

This paper by (Amiruzzaman, 2018) describes how data mining techniques are used to predict traffic violations and when a traffic violation is most likely to occur. It also discusses the contributing factors that could result in damages. Furthermore, the majority of accidents occurred at particular times and days. The methodology evaluated different algorithms like Naive Bayes, SVMs, J48 decision-tree, decision table and almost all had very good metric values for a classification task. A few specific times are likely for traffic violations, according to the analysis of the national database for traffic violations.

The study by (Mozaffari et al., 2015) emphasizes the impact of driver behaviour and outside variables on road conditions while attempting to predict vehicle speed using numerical tools. By using driving data from urban roads in San Francisco, the authors propose an evolutionary least learning machine (E-LLM) for forecasting speed sequences. Using metrics like mean square error (MSE) and root mean square error (RMSE), they compare E-LLM against well-established techniques like auto-regressive (AR) and neural networks through the use of sliding window time series (SWTS) analysis. Results show E-LLM's superior predictive capabilities, potentially aiding automotive engineers in designing efficient predictive powertrain controllers.

With speeding violations being the most common traffic offense, the paper by (Shawky et al., 2017) investigates the criteria for choosing the best locations for speed cameras on rural highways. Data from 76 speed camera locations was analyzed, and it was discovered that over 8 million tickets were issued for speeding between 2008 and 2015. Traffic volume, average speed, and posted speed limit are among the fifteen significant variables that a predictive model utilizing negative binomial regression identified as influencing the frequency of speeding violations. The analysis was done using Excel and SPSS software for developing the predictive model. The results showed that traffic-related variables and type of day (weekday or weekend) significantly influence the occurrence of speeding violations at a confidence level of 95%.

The DeepAR approach—which trains an auto-regressive recurrent network model on multiple related time series—is presented in the paper by (Salinas et al., 2020). The goal of this approach is to generate precise, probabilistic forecasts. It emphasizes the value of probabilistic forecasting in streamlining business operations. DeepAR seeks to overcome the shortcomings of conventional forecasting methods by utilizing deep learning methods. Comprehensive empirical tests on real-world datasets show 15% or more accuracy gains over modern techniques. In general, DeepAR presents great progress in forecasting, which could have consequences for different sectors looking to improve their predictive capacities.

The paper by (Wang et al., 2022) addresses the area of vehicle speed prediction in traffic management. It proposes a method based on adaptive Kalman filtering within the Autoregressive Moving Average (ARMA) framework to predict high-speed moving vehicle speeds. By utilizing it to model speed time series prediction, the method assigns weights to each coefficient to account for their varying contribution rates. Furthermore, the approach involves multisource traffic data fusion and interval speed prediction, tailored to different traffic states. Experiments show high accuracy in speed prediction, highlighting the potential effectiveness of the proposed algorithm in enhancing traffic management and vehicle safety.

Data

A. Data Collection

The dataset for this has been obtained from the Chicago Data Portal, and the number of infractions that have happened in Children's Safety Zones every day for each camera is reflected in this dataset. The dataset gets updated on a daily basis and can be fetched using the API or exported in CSV format. The dataset has data from July 2014 to present and consists of 376,419 rows currently. There are 9 attributes in the dataset containing details about the location of speed violations, the camera that captured them, the geographical coordinates, the date of the incident, and the count of the violations. The table below gives a detailed explanation about the attributes.

Attribute Name	Description
Address	Address of the location of the speed enforcement camera(s). There may be more than one camera at each address.
Camera ID	A unique ID associated with the physical camera at each location. There may be more than one camera at a physical address.
Violation Date	The date of when the violations occurred. NOTE: The citation may be issued on a different date.
Violations	Number of violations for each camera on a particular day.
X Coordinate	The X Coordinate, measured in feet, of the location of the camera. Geocoded using Illinois State Plane East

Y Coordinate	The Y Coordinate, measured in feet, of the location of the camera. Geocoded using Illinois State Plane East
Latitude	The latitude of the physical location of the camera(s) based on the ADDRESS column. Geocoded using the WGS84.
Longitude	The longitude of the physical location of the camera(s) based on the ADDRESS column. Geocoded using the WGS84.
Location	The coordinates of the camera(s) based on the LATITUDE and LONGITUDE columns. Geocoded using the WGS84.

B. Data Pre-processing

Whenever we start with any machine learning project, having good-quality data is critical for training the model and getting good performance over unseen data. Hence, the flow for this project will be similar. To perform data cleaning, I will be using the standard python libraries like numpy, pandas, matplotlib (for EDA) and scikit-learn.

Upon a first look at the data, there are some rows that contain details regarding the camera, violation date, and counts but lack geographical coordinates. These rows of data contain useful information, and hence I shall not discard them. To make the processing easier and follow a consistent format, I will convert the date from a string format to a datetime format. There are some rows where the camera location and violation counts are missing. I will use 0 for the violation count when data for a given camera on a given date is not available. In the event that I encounter any duplicate instances, I will remove all of them but retain the first instance.

Since we are working on a forecasting problem, there is no specific label or target column available in the dataset.

Analysis & Methodology

To achieve the goal of predicting speed camera violations, I plan to follow a structured approach that involves several steps. First, I will preprocess the dataset and extract relevant features such as time of day and day of the week. For spatial analysis, I will utilize the latitude and longitude coordinates to explore spatial patterns of violations across different locations.

I will use the DeepAR algorithm from AWS SageMaker, especially for time series forecasting tasks, for model training. Because DeepAR can capture intricate temporal patterns and dependencies in the data—a skill for predicting violations over time—it is a good fit for this task. I might also investigate alternative algorithms like LSTM networks or conventional time series models like ARIMA.

To evaluate the performance of the models, I will consider metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) to calculate the accuracy of predictions. Furthermore, I plan to use the cross-validation technique to assess the robustness of the models and avoid overfitting and generalizing them.

References

Amiruzzaman, M. (2018). Prediction of traffic-violation using data mining techniques. *Proceedings of the Future Technologies Conference (FTC) 2018*, 283-297. https://doi.org/10.1007/978-3-030-02686-8_23

Mozaffari, L., Mozaffari, A., & Azad, N. L. (2015). Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on san francisco urban roads. *Engineering Science and Technology, an International Journal*, *18*(2), 150-162. https://doi.org/10.1016/j.jestch.2014.11.002

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181â€“1191. https://doi.org/10.1016/j.ijforecast.2019.07.001

Shawky, M., Sahnoon, I., & Al-Zaidy, A. (2017). Predicting speed-related traffic violations on Rural Highways. *Proceedings of the 2nd World Congress on Civil, Structural, and Environmental Engineering*. https://doi.org/10.11159/icte17.117

Wang, Y., Yu, C., Hou, J., Chu, S., Zhang, Y., & Zhu, Y. (2022). Arima model and few-shot learning for vehicle speed time series analysis and prediction. *Computational Intelligence and Neuroscience*, *2022*, 1â€“9. https://doi.org/10.1155/2022/2526821