Forecasting transport flows using big data and machine learning technology

Komoliddin Tashmetov, Mukhamadaziz Rasulmukhamedov, and Timur Tashmetov

1. Tashkent State Transport University, Tashkent, Uzbekistan

Abstract. Efficient management of urban transportation is crucial for addressing the growing challenges posed by increasing traffic and population. In this context, the utilization of big data and intelligent systems has become paramount. This paper introduces a comprehensive approach to traffic flow management at a key intersection in Tashkent, capitalizing on the integration of big data analytics and predictive model.

Keywords: Big Data, machine learning, CRISP-DM, ITS, traffic flow, gradient boosting regressor, intersection, correlation, regression.

1. Introduction

In the dynamically developing capital of Tashkent, the issues of managing transport flows are becoming increasingly relevant and critical. Every day, hundreds of thousands of vehicles enter the roads of Tashkent, with a total number reaching an impressive 900 thousand. This remarkable volume not only exceeds the designed capacity of the infrastructure but also raises serious questions about the sustainability and efficiency of urban traffic.

With a population of five million people, steadily increasing year by year, and a registered annual growth of 76 thousand vehicles, the challenges in ensuring the efficiency of urban transport become an integral part of the lives of Tashkent residents. Unsatisfied mobility needs, prolonged traffic jams, economic losses, and adverse environmental impacts—all these aspects pose challenges that demand innovative and effective solutions.

The era of Big Data brings innovation and transformation to the world, encompassing various spheres of human activity. One of the key areas where Big Data has a significant impact is the transportation industry. In this overview, we will explore how the use of Big Data has become a catalyst for change in the transportation sector and what perspectives are opening up in other fields.

1. Optimization of Traffic Flows

One of the most crucial issues in urban infrastructure is the effective management of traffic flows. Big Data enables the collection and analysis of vast amounts of data on vehicle movements, passenger flows, and even route preferences. This allows cities to optimize their road systems, prevent traffic jams, and reduce pollution levels.

2. Improvement of Road Safety

* Corresponding author: tashmetov1993@gmail.com
Big Data plays a critical role in road safety. Analyzing data on traffic incidents, weather conditions, and even driver behavior helps create warning and prediction systems for accidents. This not only contributes to improving safety but also reduces costs associated with emergency medical assistance and road infrastructure restoration.

3. Efficient Management of Transport Resources

Smart transport systems supported by Big Data enable more efficient management of transport resources. Tracking the movement of goods, monitoring the condition of vehicles, and forecasting delivery times become more accurate and automated. This contributes to the optimization of logistics and a reduction in operational costs.

4. Impact on Global Environmental Issues

The use of Big Data allows for effective responses to environmental challenges related to transportation. Analyzing emissions and identifying ways to reduce the carbon footprint enables the creation of more environmentally sustainable transport systems.

The unlimited capabilities of Big Data influence not only the transportation industry but also extend to sectors such as healthcare, finance, marketing, and various business domains. In these areas, Big Data serves as a means to make more informed strategic decisions, predict trends, and identify new opportunities.

This article aims to examine contemporary challenges in urban transportation and analyze the possibilities that Big Data provides for optimizing transportation infrastructure and predicting traffic flows.

2. Literature review

Upon analyzing materials in this direction [1-5], the first article introduced a novel approach to traffic flow prediction using deep learning and the Stacked Autoencoder (SAE) model. The main distinction of this method lies in its ability to successfully uncover hidden spatial and temporal correlations in transportation flow data, including nonlinear dependencies. Researchers apply a greedy unsupervised learning method for pre-training a deep neural network with Big Data, followed by fine-tuning aimed at improving model parameters and enhancing its performance. The performance of the proposed method is evaluated on PMS (Performance Measurement System) data and compared with similar competing methods. The analysis results indicate the superiority of the proposed method over competing approaches in the task of traffic flow prediction [1].

In the second article [2], the role and importance of using Big Data in Intelligent Transportation Systems (ITS) are discussed. Given the vast amount of data available in ITS, the necessity of employing data-driven methods is emphasized. Big Data algorithms are applied to enhance the intelligence of transportation applications. The article covers a broad spectrum of applications of Big Data algorithms in ITS, including signal recognition, object detection, traffic flow prediction, travel time and route planning, as well as ensuring vehicle and road safety. The article provides a deep analysis of the application of Big Data algorithms in ITS, highlighting various areas of application and integrating models and applications.

The third article [3], [10] is dedicated to the development of a modern Intelligent Transportation Information Management System using cloud technologies and Big Data. The authors identify a series of challenges, such as complex types of objects, a large volume of collected data, high demands on transmission and computation, as well as a weak real-time capability in scheduling and management. The research is based on cloud management system theory and focuses on constructing the physical architecture of an intelligent transportation information management system. The article applies a predictive algorithm based on deep learning and ELM (Extreme Learning Machine) for accurate traffic flow forecasting in a comprehensive transportation network with a large number of traffic movement detection nodes and pre-forecasting.
3. Methods

Based on the reviewed articles [1-7], it can be stated that some methods in this field are challenging to implement in practice due to computational difficulties and configuration complexities. Practical experience has shown that intricately designed systems often do not yield very good results. In practice, simple and straightforward systems typically demonstrate good performance.

In this study, for forecasting based on Big Data, we employed the gradient boosting regressor machine learning algorithm [8, 9, 14]. This choice is motivated by the fact that this model has proven to be straightforward in real-world applications and has demonstrated excellent results.

Within the scope of this work, a street intersection located on Bogishamol Street (Fig. 1) was chosen as the research object.

Fig. 1. Tashkent ring road with the intersection of Bogishamol street. Uzbekistan.
To manage this intersection using an intelligent system, we need parameters of the traffic flow in these directions. These parameters include traffic intensity, traffic density, and average speed.

Fig. 2. Traffic flow schedule by day and hour

The data presented in (Fig. 2) is just a portion of the information collected in the year 2023. This dataset represents the number of vehicles moving from west to east.

After obtaining the data, we faced the question of which methodology to use, so we chose the CRISP-DM methodology. This methodology consists of several stages: understanding the processes, defining the analytical approach, data requirements, data collection, data analysis, data preparation, modeling, model evaluation, and project implementation. After acquiring the data, it is necessary to transform them into the required format. To achieve this, we need to clean the data from various erroneous values and, if necessary, fill them with data.

Table 1. Hourly collection of traffic flow data

<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Vehicle</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>00:00</td>
<td>1340</td>
</tr>
<tr>
<td>23</td>
<td>01:00</td>
<td>1320</td>
</tr>
<tr>
<td>24</td>
<td>02:00</td>
<td>1320</td>
</tr>
<tr>
<td>25</td>
<td>03:00</td>
<td>1280</td>
</tr>
<tr>
<td>26</td>
<td>04:00</td>
<td>1280</td>
</tr>
<tr>
<td>27</td>
<td>05:00</td>
<td>1360</td>
</tr>
<tr>
<td>28</td>
<td>06:00</td>
<td>1400</td>
</tr>
<tr>
<td>29</td>
<td>07:00</td>
<td>...</td>
</tr>
<tr>
<td>30</td>
<td>08:00</td>
<td>...</td>
</tr>
<tr>
<td>31</td>
<td>09:00</td>
<td>...</td>
</tr>
<tr>
<td>01</td>
<td>10:00</td>
<td>...</td>
</tr>
<tr>
<td>02</td>
<td>11:00</td>
<td>...</td>
</tr>
<tr>
<td>03</td>
<td>12:00</td>
<td>...</td>
</tr>
<tr>
<td>04</td>
<td>13:00</td>
<td>...</td>
</tr>
<tr>
<td>05</td>
<td>14:00</td>
<td>...</td>
</tr>
<tr>
<td>06</td>
<td>15:00</td>
<td>...</td>
</tr>
<tr>
<td>07</td>
<td>16:00</td>
<td>...</td>
</tr>
<tr>
<td>08</td>
<td>17:00</td>
<td>...</td>
</tr>
<tr>
<td>09</td>
<td>18:00</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>19:00</td>
<td>...</td>
</tr>
<tr>
<td>11</td>
<td>20:00</td>
<td>...</td>
</tr>
<tr>
<td>12</td>
<td>21:00</td>
<td>...</td>
</tr>
</tbody>
</table>

After all the data is formatted as needed (Table 1), to observe the relationship between the data, we can visualize them using a correlation plot (Fig. 3).
From this plot, we can see that the number of cars depends on both the day of the week and the time of day. Although the correlation is not very strong, it exists. After establishing these dependencies, the data is split into two parts, as commonly done in machine learning: training data (train_data) and testing data (test_data). Once again, it's important to consider data distribution. An 80/20 or 70/30 data split would be appropriate.

After obtaining the results, we can determine the mean squared error. If the error is acceptable, we can proceed with the work; otherwise, we may need to adjust the data or model parameters.

4. Result

We are satisfied with our result, by extracting a scatter plot with a regression line, we can analyze the result in more detail (Fig 4).
If we examine the forecasted result in (Fig. 5), the difference between the actual benchmark and the predicted data is very small. With this result, we can manage intersection systems, such as traffic lights, by forecasting the flow of cars in this direction, thereby preventing or predicting traffic congestion on the roads.

Fig 5. Graph comparing the actual result with the forecast.

5. Conclusion

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology was employed within the scope of the conducted research to analyze and forecast traffic flow at an intersection, based on data collected in 2023. The use of the gradient boosting regressor machine learning algorithm enabled successful modeling and prediction of the number of vehicles moving from west to east.

The obtained results not only allow visualizing relationships between various parameters of traffic flow but also facilitate efficient management of the traffic signal system at the intersection. The minimization of the gap between actual data and forecasts signals the high accuracy of the model, providing the ability to prevent or predict potential traffic jams. The utilization of Big Data and machine learning methods in the field of traffic management has the potential to enhance urban mobility efficiency, reduce congestion, and optimize traffic signal systems.

References


