DOI 10.20544/TTS2021.1.1.21.p46
UDC••656.11.02:004.85(497.11)
Original scientific paper

Traffic Flow Prediction Using Machine Learning

Slađana Janković[1], Dušan Mladenović[2], Snežana Mladenović[3] and Stefan Zdravković[4]

Abstract –**The main objective of this research was to define and verify the methodology of predicting the volume and structure of traffic flows, based on the building and application of a supervised machine learning models. The proposed methodology was applied in the case study of the prediction of traffic flows on selected routes in the Republic of Serbia.**

Keywords – **Machine learning, Big data analytics, Traffic flow.**

I. INTRODUCTION

Accurate and timely traffic flow information is currently strongly needed for individual travelers, business sectors, and government agencies [1]. It has the potential to help road users make better travel decisions, alleviate traffic congestion, reduce carbon emissions, and improve traffic operation efficiency.

The monitoring of traffic flows are followed by the permanent generation of large amounts of data. Datasets that have Big Data features provide the ability to apply modern data mining techniques. A significant class of these techniques is predictive analytics based on the application of supervised machine learning. The aim of predictive analytics is to predict what will be happening or is likely to happen in the future by exploring data. It attempts to accurately predict the future events and discover the reasons [2]. Traffic flow prediction is regarded as a critical element for the successful deployment of intelligent transportation systems, particularly advanced traveler information systems, advanced traffic management systems, advanced public transportation systems, and commercial vehicle operations [3].

In [4], Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed". We can say that machine learning is generalization of a knowledge based on the previous experience (data related to phenomena that are our subject of learning). Today, in the Big Data era, machine learning is used as one of the leading techniques in predictive analytics [5]. The aim of predictive analytics in this research was to predict the volume and structure of traffic flow on selected routes in the Republic of Serbia.

In the Section 2 of the paper all stages of the machine learning process were described: building, evaluation, testing and application of machine learning models, i.e. prediction of dependent variables. The proposed methodology was applied in the case study of the prediction of traffic flows on five selected routes in the Republic of Serbia. The results of two examples of prediction from a case study are presented in the Section 3. The last section of the paper contains conclusions about machine learning algorithms that have shown the best results in predicting the volume and structure of traffic flow on the available datasets.

[1]Faculty of Transport and Traffic Engineering University of Belgrade, Vojvode Stepe 305, Belgrade, Serbia, s.jankovic@sf.bg.ac.rs
[2]Faculty of Transport and Traffic Engineering University of Belgrade, Vojvode Stepe 305, Belgrade, Serbia, d.mladenovic@sf.bg.ac.rs
[3]Faculty of Transport and Traffic Engineering University of Belgrade, Vojvode Stepe 305, Belgrade, Serbia, snezanam@sf.bg.ac.rs
[4]Faculty of Transport and Traffic Engineering University of Belgrade, Vojvode Stepe 305, Belgrade, Serbia, s.zdravkovic@sf.bg.ac.rs

II. METHODOLOGY

Since we had labeled dataset at our disposal, we have developed supervised machine learning models. Building each of the machine learning model consisted of the following phases:

1. defining the goal of the model;
2. choosing dependent variables (label, class), i.e. the dataset attribute which value we want to predict using the machine learning model;
3. selecting relevant attributes (features) of a dataset;
4. selecting supervised machine learning algorithm, according to the nature of labels and attributes [6];
5. datasets (training and test) preprocessing that fulfills requirements of the selected algorithm;
6. model tuning – setting hyperparameters that are specific for each type of the machine learning algorithm;
7. model training – application of the selected machine learning algorithm on the training dataset in order to obtain model parameters;
8. model evaluating using cross-validation. Cross-validation is a method for getting a reliable estimate of model performance using only training data. Witten et al. in [7] proposed several alternative measures that can be used to evaluate the success of numeric prediction: mean-squared error - Eq. (1), mean-absolute error - Eq. (2), root mean-squared error - Eq. (3), relative-squared error - Eq. (4), root relative-squared error - Eq. (5), relative-absolute error - Eq. (6) and correlation coefficient - Eq. (7). The total number of test instances is n; the predicted values on the test instances are $p_1, p_2, \ldots, p_n$; the actual values are $a_1, a_2, \ldots, a_n$; $\overline{p}$ and $\overline{a}$ are the average values of the predicted/actual values;

$$Mean - squared\ error = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n} \quad (1)$$

$$Mean - absolute\ error = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{n} \quad (2)$$

$$Root\ mean - squared\ error = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{n}} \quad (3)$$

$$Relative - squared\ error = \frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2} \quad (4)$$

$$Root\ relative-squared\ error = \sqrt{\frac{(p_1-a_1)^2+\cdots+(p_n-a_n)^2}{(a_1-\bar{a})^2+\cdots+(a_n-\bar{a})^2}} \quad (5)$$

$$Relative-absolute\ error = \frac{|p_1-a_1|+\cdots+|p_n-a_n|}{|a_1-\bar{a}|+\cdots+|a_n-\bar{a}|} \quad (6)$$

$$Correlation\ coefficient\ = \frac{S_{PA}}{\sqrt{S_P S_A}} \quad (7)$$

where:

$$S_{PA}\ = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(a_i-\bar{a})}{n-1} \quad (8)$$

$$S_P\ = \frac{\sum_{i=1}^{n}(p_i-\bar{p})^2}{n-1} \quad (9)$$

$$S_A\ = \frac{\sum_{i=1}^{n}(a_i-\bar{a})^2}{n-1} \quad (10)$$

9. model testing - to predict the performance of a model on a new dataset, we need to assess its performance measures on a dataset that played no part in the formation of the model. This independent dataset is called the test dataset. Comparing test vs. training performance allows us to avoid overfitting. If the model performs very well on the training data but poorly on the test data, then it is overfit;
10. selecting a winning model - model that has the best performance on the test dataset;
11. labels prediction using the winning model.

III. Case Study

For predictive analytics we used an open source data mining software called Weka 3.8.3. Weka is a collection of machine learning algorithms used in data mining. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. We used Weka for data preparation and regression.

According to the above methodology we built and implemented machine learning models to predict the volume and structure of traffic flows. For training and testing machine learning models we used data generated by selected automatic traffic counters. Since our dataset labels (*total number of vehicles* and *percentage participation of different categories vehicles in traffic flow*) are numeric, we built machine learning models based on the most popular regression algorithms: Linear Regression, Multilayer Perceptron (Neural Network), SMOreg (Support Vector Machine for Regression), IBk (k-Nearest Neighbors), M5P, Random Forest, Random Tree and REPTree.

Records related to the period 2011-2017 were used for creating training dataset, while records belonging to the year 2018 were used to make test dataset. The training and test datasets were generated using MS Access 2016.

A. Example 1

Training dataset: *Monthly traffic flow for 21 traffic counters on three selected routes in the Republic of Serbia, for the period 2011-2017*; Number of instances: 1740; Attributes: *counter*, *month*; Dependent variable: *total number of vehicles*. Test mode: 10-fold cross-validation. Traffic counters are located at the following routes: Preljina-Pojate, Preljina-Gostun (state border between Serbia and Montenegro), ring road Aleksinac-ring road Trupale (on the highway A1).

Using the eight machine learning algorithms listed above, eight machine learning models were created to predict traffic flow volume, depending on the month of the year and traffic counter. The performance of the four best machine learning models created by using four different algorithms on this training dataset are shown in Table I.

TABLE I
PERFORMANCE OF THE TOP FOUR PREDICTION MODELS MEASURED ON THE TRAINING DATASET USED IN EXAMPLE 1

Machine learning algorithm	IBk (k=1)	Random Forest	Random Tree	REP Tree
Correlation coefficient	0.9779	0.9777	0.9779	0.9611
Mean absolute error	17492.9	17633.3	17492.9	21874.9
Root mean squared error	26075.9	26184.7	26075.9	34471.6
Relative absolute error [%]	18.03	18.17	18.03	22.54
Root relative squared error [%]	20.92	21.01	20.92	27.66

The test dataset is comprised of 247 instances, which capture traffic volume data for all 12 months of 2018 and for each of the 21 automatic traffic counters. The performance of the three selected models, obtained on the test dataset, is shown in Table II. The models based on the IBk (k-Nearest Neighbors) and Random Tree algorithms were chosen as the best prediction models because theirs performance is improved on the test dataset over the training dataset and because they show the best performance on the test dataset (Table II). Prediction was done by applying the best machine learning models (IBk and Random Tree) to the test dataset. The prediction results using these two models are identical.

TABLE II
PERFORMANCE OF THE TOP THREE PREDICTION MODELS MEASURED ON THE TEST DATASET USED IN EXAMPLE 1

Machine learning algorithm	IBk (k=1)	Random Forest	Random Tree
Correlation coefficient	0.9768	0.9766	0.9768
Mean absolute error	39582.5	39594.2	39582.5
Root mean squared error	50469.5	50488.9	50469.5
Relative absolute error [%]	35.34	35.35	35.34
Root relative squared error [%]	33.88	33.90	33.88

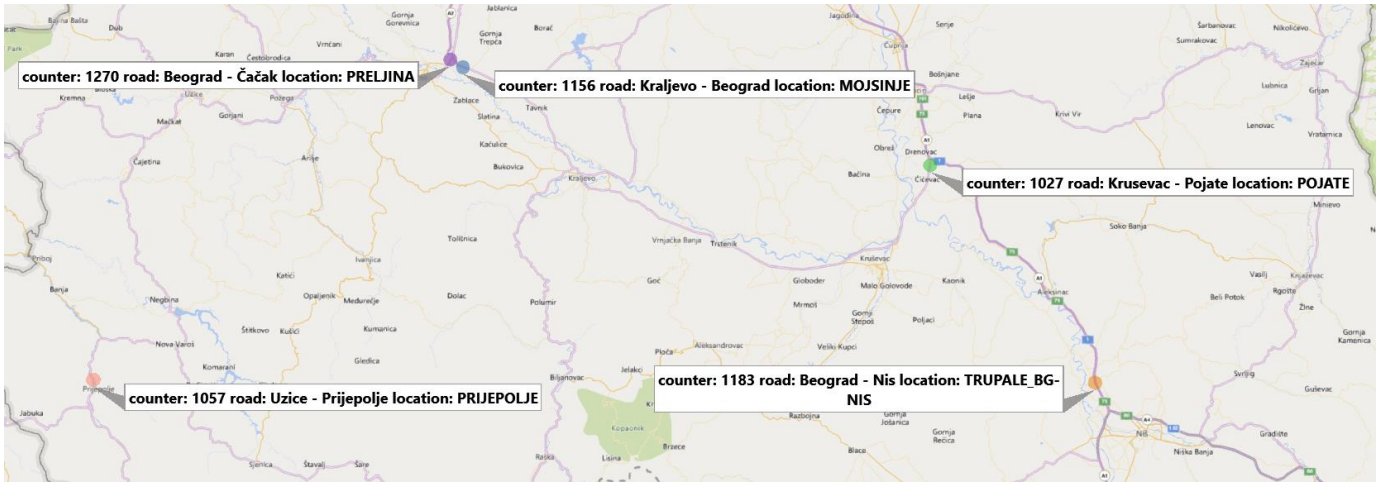

Fig. 1. Locations of the traffic counters selected for prediction analysis

Prediction results are analyzed for traffic counters whose labels are: 1027, 1057 and 1183. The locations of these counters are shown in Fig. 1. The Fig. 2 shows relationships between actual values of the monthly traffic flows for the year of 2018, and the values predicted using the machine learning models that are selected as the best ones (IBk and Random Tree), for three selected traffic counters.

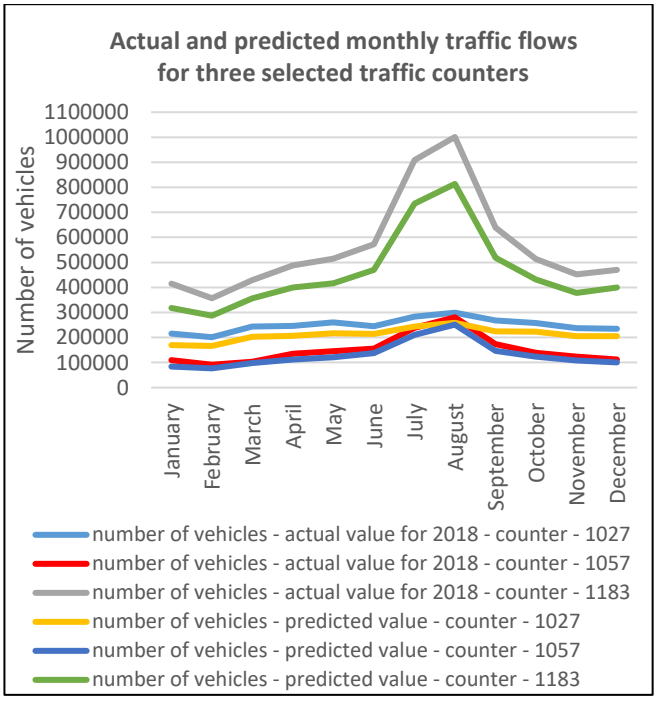

Fig. 2. Actual and predicted monthly traffic flows for three selected traffic counters

B. Example 2

Training dataset: *Monthly percentage traffic flow structure for 21 traffic counters on three selected routes in the Republic of Serbia, for the period 2011-2017*; Number of instances: 1740; Attributes: *counter*, *month*; Dependent variables: *A0%*, *A1%*, *A2%*, *B1%*, *B2%*, *B3%*, *B4%*, *B5%*, *C1%*, *C2%*, *X%*; Test mode: 10-fold cross-validation. Traffic counters are located at the routes Preljina-Pojate and Preljina-Gostun.

The performance of the first four different machine learning models created by using four different algorithms on this training dataset are shown in Table III. Dependent variable in these models is *A1% - percentage participation of A1 category vehicles in traffic flow*. Vehicles of category A1 are passenger cars and passenger cars with trailers.

TABLE III
PERFORMANCE OF THE TOP FOUR PREDICTION MODELS MEASURED ON THE TRAINING DATASET USED IN EXAMPLE 2 (A1 CATEGORY VEHICLES)

Machine learning algorithm	IBk (k=1)	M5P	Random Forest	Random Tree
Correlation coefficient	0.9701	0.9502	0.9699	0.9701
Mean absolute error	0.7323	1.1788	0.7357	0.7323
Root mean squared error	1.171	1.596	1.1733	1.171
Relative absolute error [%]	18.61	29.97	18.70	18.61
Root relative squared error [%]	24.29	33.10	24.33	24.29

TABLE IV
PERFORMANCE OF THE TOP THREE PREDICTION MODELS MEASURED ON THE TEST DATASET USED IN EXAMPLE 2 (A1 CATEGORY VEHICLES)

Machine learning algorithm	IBk (k=1)	Random Forest	Random Tree
Correlation coefficient	0.9817	0.9815	0.9817
Mean absolute error	0.6833	0.6836	0.6833
Root mean squared error	0.924	0.9291	0.924
Relative absolute error [%]	17.12	17.12	17.12
Root relative squared error [%]	19.06	19.17	19.06

Prediction results are analyzed for traffic counters whose labels are: 1057, 1156 and 1270. The locations of these counters are shown in Fig. 1. According to the results of the models shown in Table IV, the models that have the best performance were based on IBk (k-Nearest Neighbors) and Random Tree algorithm. Considering all performance measures these models show the best performance. The model based on Random Forest algorithm has very similar performance.

The Fig. 3 shows relationships between actual values for the year of 2018, and the values predicted using the machine learning models that are selected as the best ones (IBk and Random Tree), for three selected traffic counters.

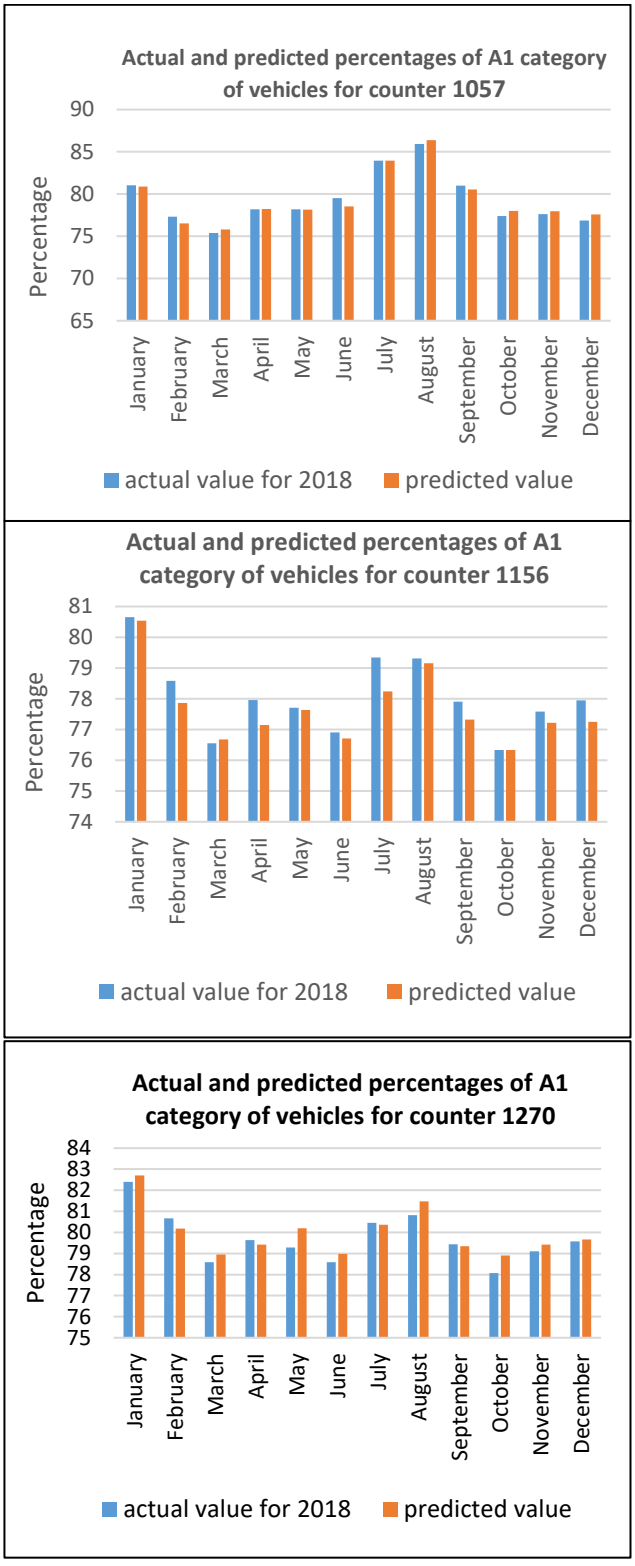

Fig. 3. Actual and predicted percentages of A1 category of vehicles for three selected traffic counters

IV. Conclusion

The research has shown that Big Data analytics based on machine learning technics can be successfully applied to predict volume and structure of traffic flows. A great number of machine learning models based on the application of the most popular regression algorithms were built, verified and tested: k-Nearest Neighbors (IBk), M5P, Random Forest, Random Tree and REPTree. Some of built machine learning models have shown satisfying performance, thus verifying the proposed prediction methodology. The best results were received by models based on k-Nearest Neighbors and Random Tree algorithms. This means that the independence of the attributes of the observed dataset is better described by nonlinear machine learning algorithms and ensemble machine learning algorithms than by linear machine learning algorithms.

Acknowledgement

This paper has been partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia project under No. 36012.

References

[1] N. Zhang, F.-Y. Wang, F. Zhu, D. Zhao, S. Tang, "DynaCAS: Computational experiments and decision support for ITS", IEEE Intell. Syst., vol. 23, no. 6, pp. 19-23, 2008.

[2] A. Uzelac, S. Janković, S. Mladenović, S. Zdravković, "Development of Machine Learning Models for Foreign Trade Volume Prediction", ICEST 2019, Conference Proceedings, pp. 228-231, Ohrid, North Macedonia, 2019.

[3] Y. Lv, Y. Duan, W. Kang, Z. Li and F. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865-873, 2015.

[4] A. L. Samuel, "Some studies in machine learning using the game of checkers", IBM Journal of research and development, vol. 3, no. 3, pp. 210-229, 1959.

[5] S. Linford, B. Bogdanovic, K. M. Chao, S. Janković, V. Maraš, M. Bugarinović & I. Trochidis, "Data Analysis on Big Data Applications with Small Samples and Incomplete Information". IEEE CSCWD 2019, Conference Proceedings, pp. 146-151, Porto, Portugal, 2019.

[6] A. K., Jain, M. N. Murty, P. Flynn, "Data clustering: a review", ACM Comput Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[7] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, Burlington, USA, Morgan Kaufmann, 2017.