Machine learning models for prediction of mobile network user throughput in the area of trunk road and motorway sections

Zoran Ćurguz[1], Milorad Banjanin[2], Mirko Stojčić[1]

[1] *Faculty of Transport and Traffic Engineering, University of East Sarajevo, Vojvode Mišića 52, 74000 Doboj, Bosnia and Herzegovina (zoran.curguz@sf.ues.rs.ba, mirko.stojcic@sf.ues.rs.ba)*

[2] *Faculty of Philosophy Pale-Department for computer science & systems, University of East Sarajevo, Alekse Šantića 1, 71420 Pale, Bosnia and Herzegovina (email: milorad.banjanin@ff.ues.rs.ba)*

Abstract

User throughput in the telecommunications network stands out as one of the key performance indicators. Today, telecommunications service providers have the task of providing a reliable and secure connection for users in all geographical locations and at all times, and adequate network throughput to meet the growing need for streaming services. These requirements primarily apply to areas around important roads, such as motorways and trunk roads. The main goal of the research is to create models based on machine learning techniques for predicting the average user throughput in the M:tel network, in a geo-area that includes the section of Motorway "9th January" (M9J), Banja Luka-Doboj, between the Johovac node and the town of Prnjavor, and the area of the M17 trunk road section, between the Johovac node and the town of Doboj. Predictive models were created on the IBM SPSS Modeler software platform, and a comparative method was used to compare and select the models that show the highest prediction accuracy. The results have shown that k-Nearest Neighbors (k-NN)-based models have the highest prediction accuracy for both sections, with the model created for the trunk road section having significantly better performance.

Keywords*: Average user throughput, Predictive models, Machine learning techniques, k-NN*

1 Introduction

The development of wireless network technologies, the enormous increase in the number of mobile applications and the expansion of the range of telecommunications services have conditioned the need for constantly better network performance. Today, telecommunications service providers and network applications have the task of providing a reliable and secure connection for users in all geo-areas and at all times, and appropriate throughput in the network to meet the growing need for streaming services. These requirements primarily apply to areas around important roads, such as motorways and trunk roads. Special attention is paid to the throughput in downlink traffic which makes up the largest part of generated network traffic. To meet customer needs and improve the quality of user experience (QoE), telecommunications providers must use the prediction of key performance indicators, such as throughput, which determine the direction of development and expansion of network capacity and resources in the future. This task is solved by predictive modeling, where prediction is defined as a prediction modeling method from the present to the future based on data obtained in the past.

For the case study in this research, it was selected the geo-area of the road zone of Motorway "9th January" (M9J), Banja Luka-Doboj section, between the town of Prnjavor and the Johovac node and the area of the M17 trunk road section between the Johovac node and the town of Doboj. The observed sections are of great importance in the road system, i.e. the road network of the Republic of Srpska and Bosnia and Herzegovina. The M9J Banja Luka-Doboj is a key road connecting the western and eastern part of the Republic of Srpska, and the M17 trunk road is one of the busiest roads in BiH. The geo-area of the research is covered by the 4G – Long Term Evolution (LTE) telecommunications network managed by the M:tel BL provider.

Throughput prediction in the cellular network at locations related to roads is the subject of numerous studies. In the previous period, a large number of scientific papers referring to this topic were published. According to [1], methods for predicting user throughput can be divided into three groups: methods based on formulas, methods based on historical data and methods based on machine learning techniques. In [2], the authors presented the Random Forest (RF) model for throughput prediction in the LTE network, emphasizing its application in maintaining a reliable connection of autonomous vehicles with infrastructure. The same machine learning model was proposed in [3] for LTE network throughput prediction, for different mobility scenarios. Additionally, the RF model was created in [4] to predict Video Streaming throughput. As a result of research in [5], it has been created models based on different Deep Neural Network (DNN) approaches, which enable throughput prediction in areas where there is no previous data on mobile network performance. When using a Live Streaming service, especially at high vehicle speeds, frequent fluctuations of Uplink connection throughput occur, which causes service delays. As a possible solution to this problem, the authors in [6] suggest PERCEIVE, a bandwidth prediction framework based on the Long Short-Term Memory (LSTM) model. Throughput prediction in data transmission between vehicles in future 6G networks is the subject of research in [7]. For this purpose, the authors created several models: Artificial Neural Network (ANN), RF and Support Vector Machine (SVM). Ur Rehman et al. in [8] modeled downlink throughput in the LTE network based on several independent variables related to the conditions of radio networks (traffic) using multilayer neural networks.

The main goal of the research is to create models based on machine learning techniques for the prediction of the average user traffic throughput in the M:tel network in the observed geo-area. The sections are exposed to different conditions: different average speeds of users' vehicles, different number of handovers, different concentrations of users, different cell sizes, etc. Therefore, to increase prediction accuracy, it is necessary to create predictive models separately, for each section individually.

The syntactic structure of the paper consists of four sections. After the introduction, Section 1, Section 2 contains described materials and research methods. The main emphasis is on Section 3, which provides the most important research results and discussion. Concluding remarks are given in Section 4. At the end of the paper, there is an overview of references.

2 Materials and methods

Following the basic goal of the research, the paper uses the *Data-Driven Prediction* approach to create predictive models. The application of this approach has been in expansion in recent years along with the enormous increase in the availability of *Big Data*. As a result of the ability to learn from data, machine learning models establish connections between dependent (output) and independent (input) variables and, based on "learned" functions, generate output values for given inputs. The most well-known machine learning techniques are ANN, Decision Trees, SVM, k-Nearest Neighbors (k-NN), etc. Thus, the data-driven prediction approach implies the existence of a data set that is divided into two parts: a data set for training and a data set for model testing. Therefore, data collection is the first step of the methodological research procedure, which is algorithmically shown in Fig. 1.

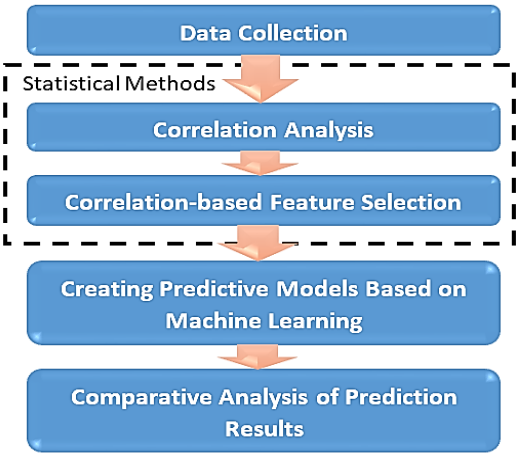

Fig. 1. Methodological research procedure

2.1 Data Collection

Fig. 2 shows the geo-area of the research. The length of the M9J Banja Luka-Doboj section from Prnjavor to the Johovac node is 35 km and in Fig. 2 is marked in blue. The section of the M17 trunk road, between the Johovac node and the town of Doboj in the length of 12 km, is marked in red in the same figure.

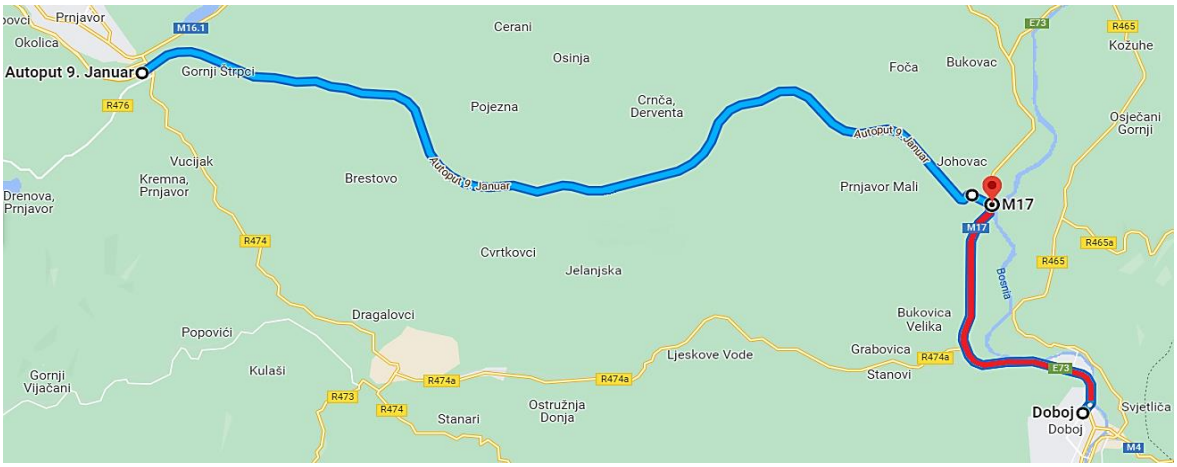

Fig. 2. Observed road sections: Prnjavor-Johovac (blue) and Johovac-Doboj (red)

Research data were obtained from the mobile operator M:tel based on a previously submitted Request specifying the necessary variables [9]. The obtained database for the LTE network contains data from a total of 71053 measurements. The values of the variables were registered in a period of 30 days (from 15 December 2020 to 15 January 2021), with a sampling frequency of one hour [9]. The data are structured into (input/output) vectors, where 17 independent variables, listed in Table 1, represent the input part, and the dependent variable –average user throughput (USER_DL_TR), the output part of the vector [9]. In addition to the names of variables, Table 1 also provides abbreviated labels (V…) used further in the paper.

Table 1. Names of independent variables with abbreviated labels

Name of Variable	Abbreviated Label
Downlink (DL) Physical Resource Block (PRB) usage rate [%]	V1
Average Channel Quality Indicator (CQI)	V2
Number of attempts by the User Equipment (UE) to establish a connection with an eNodeB	V3
Number of successfully completed connection setup procedures	V4
Average number of UEs in the connected state in the cell	V5
DL retransmission rate [%]	V6
Initial Block Error Rate (iBLER) [%]	V7
Total aggregated DL traffic in the cell [Gbit]	V8
Number of Transport Blocks (TB) with initial errors under 16QAM modulation	V9
Number of TB with initial errors under 64QAM modulation	V10
Number of TB with initial errors under QPSK modulation	V11
Number of retransmitted TB into DL shared transport channel under 16QAM modulation	V12
Number of retransmitted TB into DL shared transport channel under 64QAM modulation	V13
Number of retransmitted TB into DL shared transport channel under QPSK modulation	V14
Number of initially emitted TB into DL shared transport channel under 16QAM modulation	V15
Number of initially emitted TB into DL shared transport channel under 64QAM modulation	V16
Number of initially emitted TB into DL shared transport channel under QPSK modulation	V17

By extracting data from the obtained M:tel database, a set of a total of 9886 input/output vectors was formed for the Prnjavor-Johovac section, while 2301 input/output vectors were selected for the Johovac-Doboj section. This difference in the number of vectors for the two sets is due to the different lengths of the sections, and thus the different number of cells covering them.

2.2 Correlation Analysis

The quantitative expression of the measure of linear correlation between two variables is the Pearson correlation coefficient (r), which can be defined by the ratio of the covariance of two variables and the product of their standard deviations. The values of the Pearson correlation coefficient range from -1, which represents a

perfect negative linear correlation, and 1, which represents a perfect positive linear correlation. A value equal to zero means that there is no correlation between the variables. Any value from the specified interval can be interpreted according to the scale shown in Table 2.

Table 2. Pearson Correlation Scale [10]

Absolute correlation coefficient value	Qualitative assessment
$0< \|r\| \leq 0.19$	Very Low Correlation
$0{,}20 \leq \|r\| \leq 0.39$	Low Correlation
$0{,}40 \leq \|r\| \leq 0.59$	Moderate Correlation
$0{,}60 \leq \|r\| \leq 0.79$	High Correlation
$0{,}80 \leq \|r\| \leq 1.0$	Very High Correlation

2.3 Correlation-based Feature Selection (CFS) Method

The CFS method allows the number of input/independent variables to be reduced based on previously performed correlation analysis to simplify the machine learning model. Pearson correlation coefficients help identify independent variables that may have a stronger influence on dependent variables. Thus, a higher correlation coefficient means that the observed independent variable can be considered a strong predictor of the dependent variable [11]. According to [12], a set of variables is representative for the prediction model if, in addition to a strong correlation between independent and dependent variables, there is as low correlation as possible between independent variables. The mathematical function that defines this correlation is Merit – heuristic evaluation function:

$$M_s = \frac{k\, r_{cf}}{\sqrt{k+k(k+1)r_{ff}}} \qquad (1)$$

where M_S is the heuristic evaluation function of the subset S containing k variables; r_{cf} – arithmetic mean of correlation between independent and dependent variables; r_{ff} – arithmetic mean of correlation between independent variables. There are three heuristic strategies for finding the best subset (with the largest Merit): forward selection, backward elimination, and best first [11]. This paper uses the forward selection strategy, which is presented algorithmically in Fig. 3.

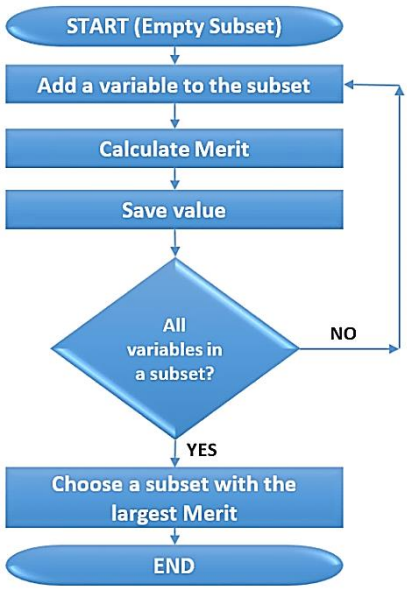

Fig. 3. Forward selection search algorithm

The algorithm shown in Fig. 3 starts with an empty subset of variables. In each subsequent step, based on a certain criterion (e.g. correlation), the following independent variable is added to the existing subset and Merit is calculated. When all the variables of the initial set are added to the subset, the subset with the highest value of the heuristic function is selected. The algorithm ends with that step.

2.4 Creating predictive models and comparative analysis of prediction results

Models for predicting average user throughput in the network were created in the SPSS Modeler software package. This software platform is one of the leading solutions in the field of Data Science, and especially machine learning. Supported techniques include Neural Networks, Classification and Regression Tree (C&R), Chi-square Automatic Interaction Detection (CHAID), linear regression, generalized linear regression, logistic regression, Bayesian Network, SVM, k-NN. A key role in this paper is played by the method of automatic modeling, which simultaneously examines several models of machine learning with different parameters according to a supervised learning paradigm. The SPSS Modeler automatically ranks offered solutions, which is possible based on correlation, relative error or the number of variables used. Comparative analysis of the offered solutions is given in the Results and discussion section based on relative

error, which represents the ratio of deviations of the observed values of the test set from those predicted by the model and deviations of the observed values from the arithmetic mean of the test set.

3 Results and discussion

This section presents the most important research results. According to the methodological steps in Fig. 1, first the results of correlation analysis are given, then the results of the application of the CFS method, and finally prediction results with comparative analysis of solutions are presented.

3.1 Results of correlation analysis

As an initial step in the statistical processing of data, a correlation analysis was performed to determine the measure of linear correlation of research variables. The matrix of Pearson correlation coefficients between the research variables for the M9J Prnjavor-Johovac section is shown in Fig. 4.

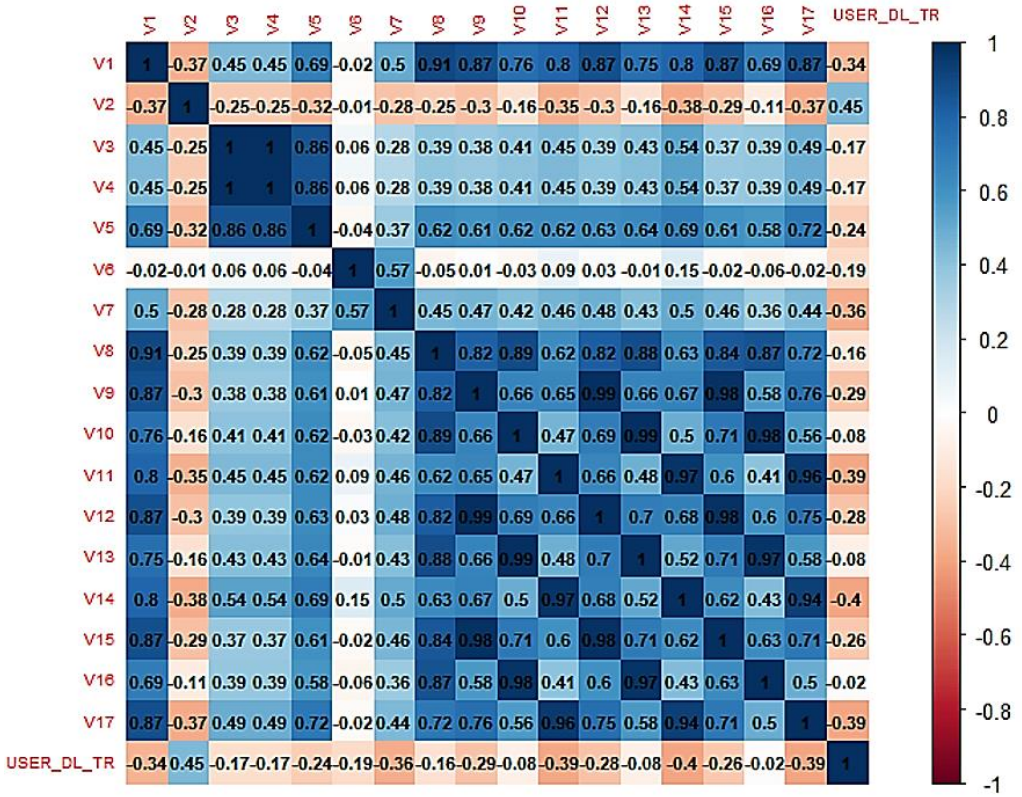

Fig. 4. Matrix of Pearson correlation coefficients (correlogram) of research variables for the M9J Prnjavor-Johovac section

Based on the values of correlation coefficients shown in Fig. 4, it is concluded that all independent variables, except V2, have a negative linear correlation with USER_DL_TR. According to the scale shown in Table 2, the variables V3, V4, V6, V8, V10, V13 and V16 have a very low correlation with the dependent variable. Low correlation defines the linear correlation of variables V5, V7, V9, V11, V12, V15 and V17 with the average user throughput. Variables V2 and V14 have a Moderate correlation with the observed output variable, which, for this section of the motorway, is the largest measure of correlation of independent and dependent variables. Therefore, the correlation coefficient higher than 0.45 (V2) was not determined between the independent and dependent variables for the observed section. Fig. 5 shows Pearson correlation coefficients between the observed variables for the M17 section, Johovac-Doboj.

From Fig. 5, it can be concluded that there is a negative correlation of all independent variables with user throughput, except variable V2, as is the case with the M9J Prnjavor-Johovac section (Fig. 4). Also, it is evident that there is a Very Low correlation between the variable V6 and user throughput. Variables V3 and V4 have slightly higher coefficients (r=-0.35) and

according to the scale shown in Table 2, they have Low Correlation with the dependent variable. Moderate correlation defines the relationship between variables V2, V7 and V16 with USER_DL_TR. Most of the independent variables have a high correlation with the observed dependent variable, namely V5, V8, V9, V10, V11, V12, V13, V14, V15 and V17. Variable V1 with a value of Pearson coefficient r=-0.80 has a very high, negative linear correlation with the average user throughput.

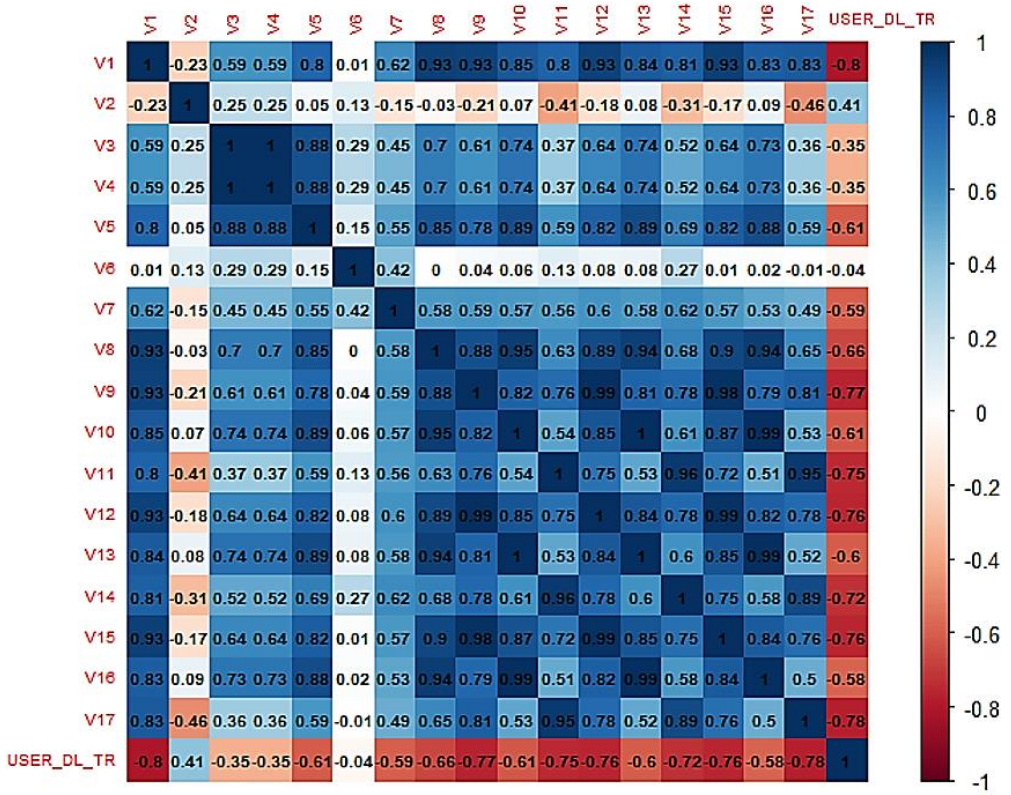

Fig. 5. Matrix of Pearson correlation coefficients (correlogram) of research variables for the M17 trunk road section, Johovac-Doboj

3.2 Results of CFS application

The Correlation-based Feature Selection (CFS) method was applied to reduce the dimensionality of the space of independent variables, based on the correlation coefficients presented in the previously given analysis. The subset of independent variables in the initial step consists only of the variable that has the highest correlation with the dependent variable. In the next step, the subset is expanded with a variable that has the second largest correlation coefficient with the average user throughput (from Fig. 4 and Fig. 5). This step is repeated until all available independent variables are included in the subset. For each subset, according to Eq. (1), Merit is calculated. The final subset of independent variables, used to create machine learning models, is the one with the highest Merit value. Table 3 shows the calculated Merit values for each subset of variables on the M9J Prnjavor-Johovac section.

Table 3. Merit values for each subset of variables on the Prnjavor-Johovac motorway section

Subset of independent variables (variable labels without the initial letter V)	Merit
2	0.447
2;14	0.508
2;14;11	0.487
2;14;11;17	0.470
2;14;11;17;7	0.491
2;14;11;17;7;1	0.475
2;14;11;17;7;1;9	0.459
2;14;11;17;7;1;9;12	0.442
2;14;11;17;7;1;9;12;15	0.426
2;14;11;17;7;1;9;12;15;5	0.417
2;14;11;17;7;1;9;12;15;5;6	0.431
2;14;11;17;7;1;9;12;15;5;6;3	0.422
2;14;11;17;7;1;9;12;15;5;6;3;4	0.411
2;14;11;17;7;1;9;12;15;5;6;3;4;8	0.394
2;14;11;17;7;1;9;12;15;5;6;3;4;8;13	0.376
2;14;11;17;7;1;9;12;15;5;6;3;4;8;13;10	0.358
2;14;11;17;7;1;9;12;15;5;6;3;4;8;13;10;16	0.338

Based on the values given in Table 3, it is obvious that a subset consisting of variables V2 and V14 (0.508) has the largest Merit. With the expansion of the subset, the decrease of Merit is evident, to a final value of 0.338. Table 4 provides an overview of the calculated Merit values for the subsets of independent variables, for the M17 Johovac-Doboj section.

Table 4. Merit values for each subset of variables on the Johovac-Doboj trunk road section

Subset of independent variables (variable labels without the initial letter V)	Merit
1	0.804
1;17	0.825
1;17;9	0.822
1;17;9;12	0.815
1;17;9;12;15	0.808
1;17;9;12;15;11	0.816
1;17;9;12;15;11;14	0.818
1;17;9;12;15;11;14;8	0.809
1;17;9;12;15;11;14;8;10	0.801
1;17;9;12;15;11;14;8;10;5	0.794
1;17;9;12;15;11;14;8;10;5;13	0.785
1;17;9;12;15;11;14;8;10;5;13;7	0.791
1;17;9;12;15;11;14;8;10;5;13;7;16	0.782
1;17;9;12;15;11;14;8;10;5;13;7;16;2	0.800
1;17;9;12;15;11;14;8;10;5;13;7;16;2;3	0.782
1;17;9;12;15;11;14;8;10;5;13;7;16;2;3;4	0.789
1;17;9;12;15;11;14;8;10;5;13;7;16;2;3;4;6	0.756

A subset of variables V1 and V17 has the highest Merit value (0.825), according to the results given in Table 4. Also, as is the case with the variables on the Prnjavor-Johovac section, the expansion of the subset leads to a decrease in Merit values. The complete set, with all independent variables, has Merit equal to 0.756.

3.3 Prediction results and comparative analysis of the results

Given that the average user throughput is continuous, in the SPSS Modeler software environment, training and testing data are processed using the *Auto Numeric* option to automatically create different predictive models. In this way, in just one pass through the modeling process, *Auto Numeric* examines models based on different machine learning techniques, different combinations of parameters for each of these models, and ranks the solutions according to relative prediction error. Table 5 presents the three best models of machine learning for both observed road sections.

Table 5. The best machine learning models ranked by relative error for both road sections

Section	Selected input/independent variables	Created machine learning model	Relative error
Prnjavor-Johovac	V2 and V14	1. k-Nearest Neighbors	0.549
		2. C&R Tree	0.699
		3. Neural Network	0.703
Johovac-Doboj	V1 and V17	1. k-Nearest Neighbors	0.183
		2. C&R Tree	0.241
		3. Neural Network	0.247

Based on the results given in Table 5, it is obvious that the models created for the Johovac-Doboj trunk road section have significantly higher prediction accuracy. The best model for this section is based on the k-NN machine learning technique and is characterized by a relative error of 0.183.

The most accurate model for the M9J Prnjavor-Johovac section is based on the same technique, but its relative error is three times higher and is 0.549. Fig. 6 shows the Scatter plot of the prediction results of the k-Nearest Neighbors model for the M9J Prnjavor-Johovac section.

Based on Fig. 6, it is obvious that there is a large deviation in the data obtained by prediction from the actual values.

The coefficient of determination (R^2), as an indicator of the quality of the model, is 0.416, which can be considered a Moderate correlation [13]. Fig. 7 shows a scatter plot of the prediction results using the k-Nearest Neighbors model for the Johovac-Doboj trunk road section.

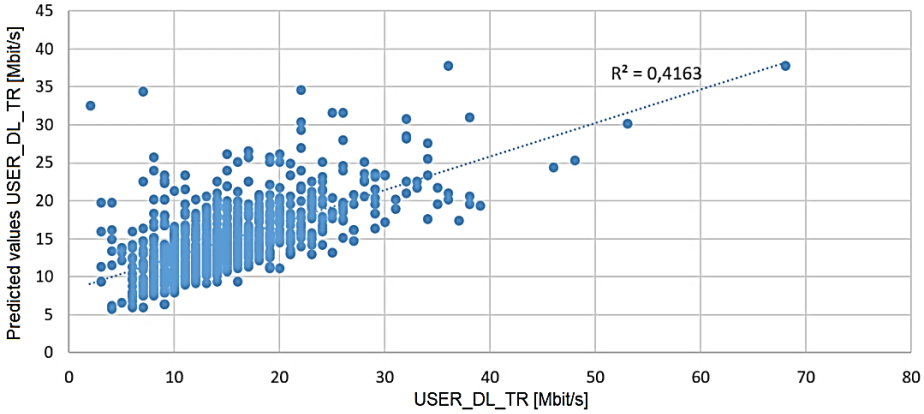

Fig. 6. Scatter plot of the prediction results for the M9J Prnjavor-Johovac section

In Fig. 7, it can be seen that the spots are largely concentrated in the vicinity of the line shown, which is indicated by the value of R^2 which is equal to 0.8005. This determination coefficient defines the High level of correlation of the data obtained by prediction with the real values of the dependent variable [13].

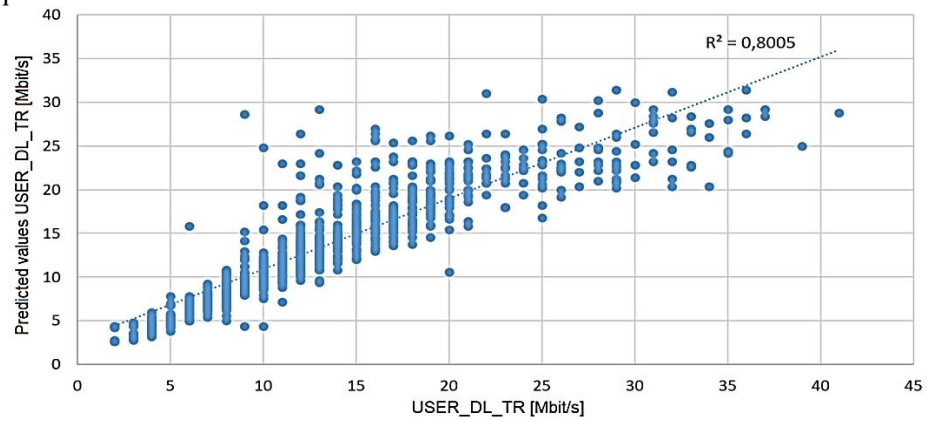

Fig. 7. Scatter plot of the prediction results for the M17 Johovac-Doboj section

4 Conclusion

In the paper, it is created several machine learning models for average user throughput prediction in the mobile network in the observed geo-area of the research. Based on the criterion of relative prediction error, the best solutions were ranked and one model was selected for each of the sections. The results showed much higher prediction accuracy for the selected k-NN model on the trunk road section, between the Johovac node and the town of Doboj. The reason for the determined difference in accuracy lies in the fact that there are not such large oscillations in the measured throughput on the M17 Johovac-Doboj road section, as is the case with the M9J section. Some of the main reasons are lower vehicle speeds on the trunk road, fewer handovers and fewer cells covering the 12 km long section. The research results and developed models have innovative theoretical and considerable practical significance, especially in terms of the needs of telecommunications service providers in the geo-area of the network that covers the observed roads. User throughput prediction in the network enables more precise planning and allocation of network resources in the future to meet user requirements. In relation to the previously published studies, which are listed in the introductory section, this paper is characterized by the following novelties: an original methodological approach to the application of machine learning methods in combination with modern statistical methods has been created; a combined geo-area in the zone of roads with different conditions for predicting the of telecommunication traffic performance has been selected; a representative set of correlative research variables in bimodal traffic has been identified. The orientation of future research may be to find other models for average user

throughput prediction on the M9J Prnjavor-Johovac section to determine a smaller relative error of prediction.

References

[1] L. Yan, J. YB Lee, "An empirical study of throughput prediction in mobile data networks", In *2015 IEEE global communications conference (GLOBECOM)*. IEEE, 2015, pp. 1-6

[2] F. Jomrich, A. Herzberger, T. Meuser, B. Richerzhagen, R. Steinmetz & C. Wille, "Cellular Bandwidth Prediction for Highly Automated Driving", *In Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2018)*, 2018, pp. 121-132, doi: 10.5220/0006692501210132

[3] C. Yue, R. Jin, K. Suh, Y. Qin, B. Wang & W. Wei, "LinkForecast: Cellular link bandwidth prediction in LTE networks", *IEEE Transactions on Mobile Computing*, *17*(7), 1582-1594, 2017, doi: 10.1109/TMC.2017.2756937

[4] D. Raca, A. H. Zahran, C. J. Sreenan, R. K. Sinha, E. Halepovic, R. Jana, ... & M. Varvello, "Empowering video players in cellular: Throughput prediction from radio network measurements", In *Proceedings of the 10th ACM Multimedia Systems Conference*, Jun 2019, pp. 201-212.

[5] J. Schmid, M. Schneider, A. HöB & B. Schuller, "A deep learning approach for location independent throughput prediction", In *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*, November 2019, pp. 1-5, IEEE.

[6] J. Lee et al., "PERCEIVE: deep learning-based cellular uplink prediction using real-time scheduling patterns", In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*, Jun 2020, pp. 377-390.

[7] B. Sliwa, R. Falkenberg & C. Wietfeld, "Towards cooperative data rate prediction for future mobile and vehicular 6G networks" In *2020 2nd 6G Wireless Summit (6G SUMMIT)*, March 2020, pp. 1-5, IEEE.

[8] T. ur Rehman, M. A. I. Baig, & A. Ahmad, "LTE downlink throughput modeling using neural networks", In *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, October 2017, pp. 265-270, IEEE.

[9] M. K. Banjanin, M. Stojčić, D. Drajić, Z. Ćurguz, Z. Milanović, & A. Stjepanović, A., "Adaptive Modeling of Prediction of Telecommunications Network Throughput Performances in the Domain of Motorway Coverage", *Applied Sciences*, *11*(8), 3559, 2021. https://doi.org/10.3390/app11083559

[10] M. Selvanathan, N. Jayabalan, G. K. Saini, M. Supramaniam, & N. Hussin, "Employee Productivity in Malaysian Private Higher Educational Institutions", *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(3), 66-79, 2020, doi: 10.48080/jae.v17i3.50

[11] M. A. Hall, "Correlation-based Feature Selection for Machine Learning", Department of Computer Science, Hamilton, New Zealand, Doctoral thesis, 1999.(https://www.cs.waikato.ac.nz/~mhall/thesis.pdf)

[12] A. Ranjan, V. P. Singh, R. B. Mishra, A. K. Thakur & A. K. Singh, "Sentence polarity detection using stepwise greedy correlation based feature selection and random forests: An fMRI study", *Journal of Neurolinguistics*, *59*, 100985, 2021, https://doi.org/10.1016/j.jneuroling.2021.100985

[13] I. Sobolev, S. Babichenko, "Application of the wavelet transform for feature extraction in the analysis of hyperspectral laser-induced fluorescence data", *International Journal of Remote Sensing*, *34*(20), 7218-7235, 2013, https://doi.org/10.1080/01431161.2013.817714