PREDICTION OF DAILY DEMAND FOR GOODS USING WEKA SOFTWARE TOOL

Milica Mitrović [a,*], Slađana Janković [a], Snežana Mladenović [b]

[a] University of Belgrade, Faculty of Transport and Traffic Engineering, Serbia

Abstract: *Predicting the demand for goods is an important task that directly affects procurement planning, inventory management and sales. Companies need to make an accurate forecast of demand for goods in order to successfully respond to customer requests and have an adequate range and quantity of stock goods. Based on the collected historical data and the application of predictive analysis methods in data mining software tools, companies can predict the demand for goods. In this paper, the prediction of the daily demand of five products was performed using supervised machine learning in the Weka software tool. The following features were selected as independent attributes that affect the quantity of goods demanded: product, day, month and day of the week. The datasets on which machine learning models were trained and tested covers the period from January 2018 to September 2020. Research shows that supervised machine learning is an effective method for predicting the demand for goods. The best results in predicting the demand for selected products were shown by the Decision Table and Random Forest algorithms.*

Keywords*: logistics, machine learning, prediction, demand.*

1. INTRODUCTION

Predicting demand for goods is an important part of the supply chain management process (Qasem, 2019). Forecasts affect decision making throughout the supply chain from the operational to the strategic level, such as the procurement of raw materials, goods, assets, transport organization, deployment and employment of workers, etc. The prediction of daily demand for goods is important in organization, planning and scheduling of activities in logistics processes because it enables the right time to meet customer requirements, reduce costs and improve the efficiency of realization of activities in the supply chain. Machine learning is a technique that can be used for different types of predictions, one of which is the prediction of demand for goods. Specialized tools for

* milica.mitrovic@sf.bg.ac.rs

creating and using prediction models can facilitate operational and strategic planning for logistics companies.

In literature, there are numerous papers in which researchers used machine learning algorithms for different types of predictions in the field of logistics. Qasem (2019) has implemented a PART strategy to predict the daily demand of the Brazilian logistics company's products. The results showed that PART classifiers with high accuracy gave a prediction of daily product demand. Alnahhal et al. (2021) dealt with the dynamic prediction of delivery times in logistics companies to optimize the consolidation of goods. Consolidation of goods reduces costs, but can increase delivery time. Kilimci et al. (2019) have integrated 11 different models that include multiple time series algorithms, the support vector method for regression analysis, and the deep learning method to predict demand. The authors of this study concluded that by inclusion of several different training algorithms, other than time series models, they improve the performance of models for predicting demand. Janković et al. (2020) conducted a descriptive and predictive analysis in the case study of foreign trade in food products for the Republic of Serbia. The authors of this research used supervised machine learning to predict the volume and structure of imports and exports in the food industry by using Weka software tool. They concluded that descriptive analysis and visualization of dependent and independent variable ratios on the initial dataset makes predictive analysis more efficient.

The aim of this paper is to obtain a prediction of the daily demand of five products by applying the methods of predictive analysis using Weka software tool. The data was obtained from the company Milšped, which is one of the leading logistics providers operating in Serbia. The basic services provided by Milšped are warehousing and transport services. The data is divided into two parts, so that about 2/3 of the available instances are a dataset for model training, and about 1/3 of the instances is a dataset for model testing. The paper presents the results of algorithms with the best performance, namely Decision Table and Random Forest. After training the model, model testing was conducted by predicting on the test dataset, and the projected demand values were compared to the actual demand values.

The paper is organized in three sections, in addition to the introduction and conclusion. The second section of the paper describes the machine learning method and algorithms that produced the best results in this study. The third section contains a brief description of the case study that is a part of this research. In the fourth section, the results of the best performance algorithms are presented and analyzed. The last section contains the conclusions of this case study and the directions of future research.

2. MACHINE LEARNING METHOD

Machine learning is a branch of artificial intelligence and a method of data analysis based on the idea that a computer system can automate learning from historical data and thus improve its work. It is applied to detect invisible patterns, market trends, hidden correlations, and other useful information for a particular aspect of data usage. Machine learning uses algorithms that allow a predictive model to learn based on historical data and predict the future values of the observed variable. Today, this method is most commonly used on large data sets. Machine learning consists of three phases: model training phase, model testing phase and prediction phase. Types of machine learning are supervised learning, unsupervised learning and reinforcement learning. Supervised

learning applies over label data for prediction the target variable. Unsupervised learning applies to a set of unlabeled data to detect hidden patterns in the data. Reinforcement learning uses both labeled and unlabeled data for training, i.e. model learns by method of attempts and errors, while each action carries a reward or punishment. The prediction phase in machine learning involves applying of the best trained prediction model on a new unlabeled dataset and obtaining new data that represent the prediction of a specific target variable (Salkuti, 2020).

In this study, supervised machine learning was applied on the available data set with the aim of obtaining a prediction of the target variable. First, the model was trained on the labeled training dataset, and then using the trained model, a prediction of the target variable over the testing dataset has been made. After the model testing, a comparative analysis of the values of actual and projected demand quantities of five products was performed, according to the two algorithms that gave the best performance of the predictive models. The observed performances are (Folorunsho, 2013):

- Mean absolute error - represents the average of the absolute values of the differences between the actual and projected value of the target variable, i.e. it is the average prediction error.

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

- Square root of mean square error - represents the square root of the mean square of the difference between each projected value and the corresponding actual value of the target variable.

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

- Relative absolute error - represents the ratio of the total absolute error of the projected values in relation to the actual values of the target variable and the total absolute deviation of the actual values from their average value. It is expressed as a percentage.

$$Relative - absolute\ error = \frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|} \quad (3)$$

- Root relative square error - represents the square root of the ratio of total square error of projected values in relation to the actual values of the target variable and the sum of squares of deviations of real values from their average value. It is expressed as a percentage.

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

- Correlation coefficient - represents the degree of difference between actual and projected values of the target variable, i.e. the strength of their mutual connection. The higher the value of the correlation coefficient, the stronger the relationship.

$$Correlation\ coefficient = \frac{S_{PA}}{\sqrt{S_P S_A}} \quad (5)$$

$$S_{PA} = \frac{\sum_{i=1}^{n}(p_i-\bar{p})(a_i-\bar{a})}{n-1},\ S_P = \frac{\sum_{i=1}^{n}(p_i-\bar{p})^2}{n-1},\ S_A = \frac{\sum_{i=1}^{n}(a_i-\bar{a})^2}{n-1} \quad (6)$$

Where:

- n - the total number of instances for testing,
- p_1,..., p_n - the projected values on test instances,
- a_1,..., a_n - actual values,
- $\bar{p}$ - mean projected values,
- $\bar{a}$ - mean actual values.

This study uses the specialized software tool Weka (Waikato Environment for Knowledge Analysis), which is open source and developed at the University of Waikato in New Zealand. Weka is a simple tool for use and provides the ability to perform the following data mining tasks: data preparation, classification, regression analysis, clustering, association rule learning, selection of relevant attributes, and data visualization (Witten and Frank, 2005). The files that can be loaded in software tool Weka are text files format ARFF (Attribute-Relation File Format) and Excel files format CSV (Comma-Separated Values). In the case study, all algorithms available in the Weka software tool were applied in the model training phase, but the paper presents two algorithms that had the best results in model testing. These algorithms are Decision Table and Random Forest, and they are described below.

The Decision Table is a machine learning algorithm that is arranged as a set of If-Then rules. This algorithm is used to analyze a dataset by using a decision table that contains the same number of attributes as the original dataset. He evaluates subsets of attributes using the "first best" search based on the method of the nearest neighbor and applying cross-validation (Kohavi, 1995; Kalmegh, 2018). The Decision Table algorithm is in the Rules category in the Weka software tool.

The Random Forest is a machine learning algorithm used to solve regression and classification problems. It combines multiple different algorithms to create a prediction. The Random Forest consists of many decision making trees based on whose predictions determine the outcome i.e. the final outcome is determined based on a majority vote. The algorithm gives a prediction based on averages from different trees. The structure of the decision making tree can be described as a tree made up of nodes connected by branches. The Random Forest reduces overfitting of datasets and increases accuracy (Breiman, 2001). This algorithm is in the Trees category in the Weka software tool.

3. CASE STUDY

Employees in the procurement sector monitor the demand, sales and changes that are happening on the market then analyze them to make a decision on ordering an adequate range and quantity of goods. This decision represents a particular challenge for employees because it is necessary to order the optimal quantity of goods that will not lead to situations where the company runs out of stock or has an excessive level of stock. Both situations create certain costs. Lack of stock creates the cost of missed opportunities, that

is, the inability to respond to customer requests. On the other hand, an excessive level of inventory creates higher costs of storing goods. The goal is to avoid both situations and not create unnecessary costs, and this is achieved by ordering an adequate quantity of goods. Modern technologies can facilitate and accelerate the process of ordering goods, in relation to the traditional way of ordering. The application of machine learning models has great potential in the supply chain, among other things for prediction demand and optimizing inventory.

In this case study, the daily demand of five products was predicted based on data obtained from Milšped. Milšped is a logistics provider that offers warehousing and transportation services, so it cooperates with clients and customers. Clients are companies that store their goods in the logistics provider's warehouse, and customers are retail stores that distribute goods. According to customer requirements, the dynamics of the output of goods from the warehouse can be monitored, on the basis of which the provider plans future demand. The obtained data is encrypted for security and confidentiality. As independent attributes affecting commodity demand, they are selected: product, day, month, day of the week (Monday to Friday), while the target variable is the quantity of goods demanded.

Data cover the period from January 2018 to September 2019 were used to train the model, and data cover the period from October 2019 to September 2020 were used to test the model. The initial available dataset contains 352.740 instances with information on the ordering date, purchase order number, product, customer and quantity of ordered goods. As the aim of the case study is to predict the demand for five products, certain data were neglected and only those related to the date and quantity of demand for five products were observed. In the data preparation phase, the data was "cleaned" by generating and executing SQL (Structured Query Language) queries on the entire available dataset. The training dataset consists of 1911 instances and the test dataset of 1262 instances.

4. RESULTS AND ANALYSIS OF RESULTS

In the data mining software tool Weka conducted the process of machine learning and prediction of daily demand for five products. Table 1 shows the performance of the two predictive algorithms measured on the training dataset, obtained using 10-fold cross-validation. Cross-validation is a statistical method for obtaining a reliable model performance assessment using only training data (Witten and Frank, 2005). Most of the algorithms tested gave excellent results, but two algorithms with the best results were presented: Decision Table and Random Forest. The excellent result of the model implies that the correlation coefficient is greater than 0.9, indicating that further analysis of the data may continued.

Table 1. Performance of Prediction Models obtained using Training Dataset

Algorithm	Correlation coefficient	Mean absolute error	Square root of mean square error	Relative absolute error (%)	Root relative square error (%)
Decision Table	0.965	219.184	367.071	19.368	26.205
Random Forest	0.961	232.368	383.471	20.534	27.376

Table 2 shows the performance for the best algorithms measured on the test dataset. A comparative analysis of the obtained performance of the algorithms measured on the dataset for testing and training shows that the results of the training algorithms are slightly better. Precisely such results indicate that there is no problem of overfitting and that algorithms can be used for prediction (Janković and Mladenović, 2020).

Table 2. Performance of Prediction Models obtained using Test Dataset

Algorithm	Correlation coefficient	Mean absolute error	Square root of mean square error	Relative absolute error (%)	Root relative square error (%)
Decision Table	0.946	330.688	487.550	28.970	36.865
Random Forest	0.941	321.890	512.617	28.199	38.760

Figure 1 presents five diagrams, which show the relationship between actual and projected demand, obtained on the test dataset, separately for each product in the third quarter of 2020 year. The blue line represents the actual quantity of demand on the test dataset, the black line represents the projected quantity obtained by the Random Forest algorithm, while the dashed orange line represents the projected quantity obtained by the Decision Table algorithm. From the five diagrams shown, it can be observed that the prediction of daily demand for products, obtained by the Random Forest algorithm, better follows the actual values of quantities than the projected values obtained by the Decision Table algorithm. Algorithms gave the best prediction for products 2 and 5, then for product 4, while products 1 and 3 observed a greater deviation between the values of actual and projected quantities. The reason for the greater deviation between the values of actual and projected quantities for product 1 may be the uneven daily demand of quantities that are not common, i.e. there is no specific trend. Such deviations in the dataset have a bad effect on prediction. For product 3, the reason for deviations is lack of data by date on the training dataset. One of the advantages of the Random Forest algorithm is to function well on a large dataset and on a set with a large number of missing data, which is an explanation for why this algorithm gave a better prediction for product 3 compared to the Decision Table algorithm. The forecast of daily demand for goods is shown for all products with the intention of showing how the forecast result is affected by the lack of certain values in the datasets and the becoming of deviation values. The

average error in demand for goods obtained by the Random Forest algorithm per product expressed in units piece, from July to September is as follows: product 1 around 11 (in minus), product 2 around 47, product 3 around 116, product 4 around 290 and product 5 around 327. Also, the average error in demand for goods obtained by the Decision Table algorithm is as follows: product 1 around 13 (in minus), product 2 around 60, product 3 around 458, product 4 around 514 and product 5 around 569. The correlation coefficient is higher with the Decision Table algorithm, but the mean absolute error, i.e. the average prediction error is smaller with the Random Forest algorithm, as shown in Table 2 and observed in the diagrams.

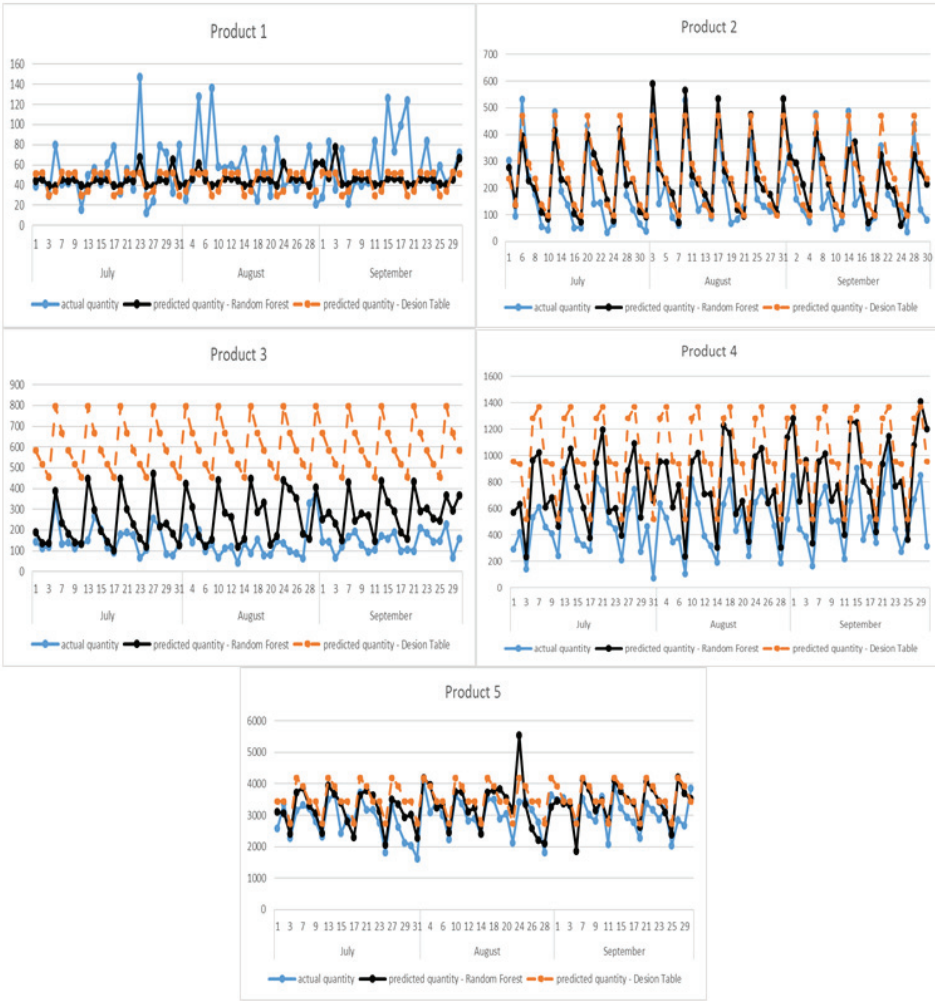

Figure 1. Actual and projected quantity of demand for five products in the third quarter of 2020

Figure 2 shows diagrams for five products showing the ratio of actual and projected demand for goods to the days of the week in June 2020. The perceived trend for products 2, 3, 4 and 5 is higher demand for goods at the beginning of the workweek, followed by a

drop in demand. The day of the week when the highest demand for product 2 is Monday, and in the most common cases it is for products 3 and 4, while for product 5 the highest demand is mostly on Tuesdays. Product 1 has not seen a specific recurrence trend and products of this type should always be in inventory because their demand is unpredictable. The results show that by applying machine learning algorithms, good predictions of the target variable can be obtained if the dataset contains quality data for as long as possible. Also, results cannot be predicted until the machine learning process is carried out on a specific dataset. The results show that the predictive models developed in this research cannot be used to predict the demand of all five types of products, but they can be used to predict the demand of product types 2 and 5.

Figure 2. Actual and projected quantity of demand for five products in June 2020

5. CONCLUSION

In this conducted case study for predicting daily demand of five products, models based on the Decision Table and Random Forest algorithms had the best performance. These algorithms showed better performance than other algorithms available in the Weka software tool. Performance of models based on these two algorithms are approximate. The correlation coefficient with the Decision Table algorithm is higher than with the Random Forest algorithm. However, the mean absolute error is smaller with the Random Forest algorithm. On the diagrams shown, it was observed that the prediction by the Random Forest algorithm follows the actual values better than the prediction obtained by the other algorithm. The advantage of the Random Forest algorithm is that it functions well over a large dataset and over a set with a significant number of missing data, as confirmed by this study. The quantity and quality of datasets are of great importance in the application of machine learning models (Zhou et al., 2020). The results of the research indicate a data-driven characteristic of machine learning. The data-driven feature refers to the fact that one machine learning model can give excellent results on one dataset, but at the same time it can be unusable on another dataset, which describes the same phenomenon. The reasons that lead to this problem can be lack of data and diversity of data.

The results of this research prove that machine learning and the application of Weka software tools can be of significant benefit to logistics companies to predict demand for goods. Using data mining tools and machine learning algorithms for certain predictions could help companies better plan, organize, reduce costs, and increase efficiency.

In future research, more attributes related to the demand of goods can be observed to detect certain regularities and patterns in data using machine learning. Then, it would be interesting to predict the demand for goods using time series algorithms from the Forecast tab of the Weka software tool.

ACKNOWLEDGMENT

This research was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the project number 036012.

The authors thank the company Milšped from Belgrade, which is provided data for the research presented in the paper.

REFERENCES

[1] Alnahhal, M., Ahrens, D., Salah, B., (2021). Dynamic Lead-Time Forecasting Using Machine Learning in a Make-to-Order Supply Chain. MDPI, 11(21), 1-16.
[2] Breiman, L., (2001). Random Forests. Machine Learning, 45, 5-32.
[3] Folorunsho, O., (2013). Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database. International Journal of Advanced Research in Computer Science and Software Engineering, 3(3), 11-15.
[4] Janković, S., Kilibarda, M., Uzelac, A., (2020). Big Data Analytics in Logistics. In Quantitative Methods in Logistics. University of Belgrade – Faculty of Transport and Traffic Engineering, 197-214.

[5] Janković, S., Mladenović, D., (2020). Application of software tools in Big Data analysis of sensor data in traffic. XXXVIII Symposium on new technologies in the postal and telecommunications traffic – PosTel, 239-248.
[6] Kalmegh, S.R., (2018). Comparative Analysis of the WEKA Classifiers Rules Conjunctiverule & Decision Table on Indian News Dataset by Using Different Test Mode. International Journal of Engineering Science Invention, 7(2), 1-9.
[7] Kilimci, Z.H., Akyuz, A.O., Uysal, M., Akyokus, S., Uysal, M.O., Bulbul, B.A., Ekmis, M.A., (2019). An Improved Demand Forecasting Model Using Deep Learning Approach and Proposed Decision Integration Strategy for Supply Chain. Complexity, 1-15.
[8] Kohavi, R., (1995). The Power of Decision Tables. Computer Science Department. Stanford University, 912, 174-189.
[9] Qasem, S.N., (2019) A PART-Based Algorithm for Prediction of Daily Demand Orders. International Journal of Computer Science and Network Security, 19(9), 91-95.
[10] Salkuti, S.R., (2020). A survey of big data and machine learning. International Journal of Electrical and Computer Engineering, 10(1), 575-580.
[11] Witten, I.H., Frank, E. (2005). Data Mining Practical Machine Learning Tools and Techniques, Second Edition.Elsevier Inc.
[12] Zhou, Q., Lu, S., Wu, Y., Wang, J., (2020). Property-Oriented Material Design Based on Data-Driven Machine Learning Technique. The Journal of Physical Chemistry Letters, 3920-3927.