Measuring Business Data Volume in Serbia using Machine Learning Systems

Ana Uzelac, Sladana Jankovic, Snezana Mladenovic, Stefan Zdravkovic
University of Belgrade
Vojvode Stepe 305
11000 Belgrade, Serbia
{ana.uzelac@sf.bg.ac.rs}, {s.jankovic@sf.bg.ac.rs}, {snezanam@sf.bg.ac.rs}
{s.zdravkovic@sf.bg.ac.rs}

ABSTRACT: *There is an increasing tendency to process supply chain data to detect the patterns to improve the functions. We have generated a machine learning system which can be used to analyse the business data volume to judge the size of the trade activities.*

Keywords: Machine Learning, Prediction, Big Data Analytics

Received: 28 April 2020, Revised 15 July 2020, Accepted 16 August 2020

DOI:10.6025/jdp/2020/10/4/118-124

1. Introduction

Big data is a term used for massive data sets with complex structure. It refers to those datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze [1]. The four main characteristics defining Big Data are Volume, Velocity, Variety and Veracity [2]. As previously mentioned, Big Data exceeds the space of technical ability of storing, processing, managing, interpreting and visualizing of a traditional system [3].

Before the Big Data era, various data analytics technics were used to analyze data with the aim to find correlations between them. With the Big Data emergence, a great volume of data is generated every day creating a growing demand to investigate a greater amount of data in order to find useful patterns and correlations within. Big Data can be combined with analytics forming the Big Data Analytics (BDA). The term BDA can be defined as the application of advanced analytic techniques including data mining, statistical analysis, predictive analytics, etc. on big datasets as new business intelligence practice [4]. BDA has the ability to research massive amounts of data with the aim to reveal hidden patterns and secret correlations. Therefore, Big Data combined with analytics creates the ability to extract meaningful insights and turn data into information and intelligence [5]. BDA give firms competitive advantage by extracting significant value from massive amounts of data creating an imperative for business leaders in almost every industry sector: from healthcare to manufacturing.

One of the technics used in BDA is machine learning. Machine learning emerged many years before Big Data existed: in 1959, Arthur Samuel defined machine learning as a "Field of study that gives computers the ability to learn without being explicitly

programmed" [6]. Tom Mitchell provides a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [7]. In general, any machine learning problem can be assigned to one of two broad classifications: supervised and unsupervised learning. If instances are given with known labels (the corresponding correct outputs) then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled [8]. If we are trying to predict something that has a discrete value, that type of supervised learning is called classification. The other type of supervised learning is regression where the accuracy is measured by how close the estimate is to the actual value in the test dataset, rather than whether the predicted value is precisely right or not. Regression typically requires different algorithms than those used for classification. The aim of classification and regression is to predict something accurately. The part of the data is used for learning while the rest of the data is used for testing the result of the learning. In comparison, unsupervised learning is where we do not know the right answer ahead of time for any of the data—there is no prior basis to judge how good our result is. The goal of unsupervised learning is to find interesting and useful generalities within the data. A common form of unsupervised learning is clustering - given a collection of data, separate instances into two or more groups (clusters) based upon their similarities. Unsupervised learning has no objective measure of success, and therefore, all the data can be used as input to the algorithm. As previously seen, selecting the suitable machine learning technic depends mostly on the structure of the data and the goals of the study.

The aim of this paper is to explore possibilities to predict the volume of the foreign trade using supervised machine learning and to propose the development methodology and application of appropriate machine learning models. In the second section we have covered literature review. In the third section a methodology including development and the application of the machine learning model that predict the volume of import and export is presented. Some of the results obtained using proposed methodology on the dataset of foreign trade of food industry in the Republic of Serbia are shown in the fourth section. In the last section a conclusion about machine learning algorithms that have shown the best results in predicting the volume of the foreign trade on the available datasets is given.

2. Literature Review

Gartner estimates that by 2020 there will be around 26 billion devices in the supply chain. All of them generate a great amount of data every day. Therefore, there is a growing need to analyze huge amounts of data in Supply Chain Management (SCM). Scholars agree that BDA has the potential to transform the entire business process, by improving the various supply chain processes and logistics management [9]. Although the term "Big Data" is not new, there are not many applications of Big Data in the SCM field.

As machine learning is able to discover patterns in supply chain data, it has been identified ten ways how machine learning can revolutionize supply chain management: 1) improving demand forecast accuracy, 2) reducing freight costs, 3) improving SCM performance, 4) opening up many potential applications in physical inspection and maintenance of physical assets across an entire supply chain network, 5) lowering inventory and operations costs and getting quicker response times to customers, 6) forecasting demand for new products, 7) extending the life of key supply chain assets including machinery, engines, transportation and warehouse equipment, 8) improving supplier quality management and compliance, 9) improving production planning and factory scheduling accuracy, and 10) providing end-to-end visibility across many supply chains for the first time [10]. Artificial intelligence with machine learning can help the logistics industry fundamentally shift its operating model from reactive actions and forecasting to proactive operations with predictive intelligence [11].Additionally, machine learning represents a new tool that can enable companies to better understand the impact of demand drivers such as media, promotions and new product introductions, and to then use that knowledge to significantly improve forecast quality and detail [12].

Artificial Intelligence is set to transform the foreign trade, for example; by reducing the cost of numerous processes throughout the trade lifecycle [13]. Furthermore, machine learning can help in reducing foreign exchange risk [14]. Moreover, machine learning can be used to build a metalearning model that is able to detect the error in the foreign trade transactions [15].

Although machine learning models are recognized to be a great tool to discover patterns in large amounts of data with the aim to improve different parts of SCM, currently there are not many studies that investigate their practical applications. As

foreign trade represents an important part in SCM, this investigation is of a great importance as it shows how machine learning models can be used in practice in order to predict the volume in the foreign trade.

3. Methodology

The research process consists of the following three phases: Data exploration, Data preprocessing and Predictive analytics.

During data exploration phase we examined some characteristics of the initial dataset, such as its volume, completeness, validity of data, potential relations between individual data elements, different ways raw data is organized and stored.

Data Preprocessing phase consists of standard ETL (Extract, Transform and Load) operations. During Data preprocessing we performed the following operations: data importing, data querying, data cleaning, data formatting end data exporting. As a result of data processing, a new dataset on which it can be applied different machine learning technics is generated. Additionally, as the final step we divided the original dataset into two different datasets: testing and training datasets.

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. The aim of predictive analytics in this research was to predict the volume and structure of import and export of food products in the Republic of Serbia. In order to get the answer, we used different machine learning techniques. Machine learning process consists of the following steps: 1) data preprocessing, 2) model building, 3) model evaluation, 4) model testing, and 5) model deployment. Machine learning is an iterative process which is repeated until a satisfying performance is achieved.

The first step was to build a model. As we had labeled dataset, we could build only supervised machine learning model. First, we defined the goal of our model, then selected dependent variables (labels) and relevant attributes, performed necessary preprocessing of the dataset in order to prepare it to fulfill requirements of the selected algorithm. The next step was model tuning where we set hyperparameters that are specific for each type of the machine learning algorithm. The next phase was model training where we applied selected machine learning algorithm on the training dataset in order to obtain model parameters.

Since our dataset labels (Net weight [kg] and Amount [EUR]) are numeric continuous, we have chosen machine learning models based on the most popular regression algorithms: Linear Regression, k-Nearest Neighbors, Decision Tree, Support Vector Machine for Regression and Neural Network. Machine learning model results from learning algorithm applied on a training dataset.

We performed model evaluation using 10-fold crossvalidation. 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 model formation. This independent dataset is called the test dataset. We assume that both the training data and the test dataset are representative samples of the underlying problem. 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. The success of numeric prediction was evaluated using various performance metrics, as they are: 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 $a_1, a_2, \ldots, a_n$; p and 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,

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

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

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

We selected the best model comparing different criteria, such as performance (examining if the model has the best performance on the test dataset), robustness (if the model performs well across various performance metrics), consistency (if the model has one of the best cross-validated scores from the training dataset) and win condition (if it solves the original business problem).

4. Results

Previously described methodology is applied on the dataset of foreign trade of food products in the Republic of Serbia for the period from 2015 till 2017. 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.

The available dataset consisted of 772517 instances, each having the following attributes: IE (Import/Export), ClearanceProcedure, RegistrationNumber, VATIN, CompanyName, CountryOfBuyer/Seller, CountryOfImport, CountryOfExport, CustomsTariff, CustomsTariffName, Year, Month, Quarter, TradeName, UnitOfMeasure, Quantity, NetWeightKG, and AmountEUR. Data preprocessing consisted of cleansing the dataset from the incomplete records, eliminating attributes that were uniquely dependent on the other attributes (such as VATIN, CompanyName, CustomsTariffName) and creating different SQL (Structured Query Language) on the dataset. Some of the queries performed only records grouping in different ways such as: Import/Export by Year and ClearanceProcedure, Import/Export by Year, Month and CountryOfImport/Export, Import/Export by Year, Month, RegistrationNumber and CountryOfImport/Export, Import/Export by Year, Month, RegistrationNumber, CountryOfImport/Export and CustomsTariff, etc. The second group of queries records were selected by different criteria, such as: Clearance Procedure, CountryOfImport/Export, Registration Number, Customs Tariff, while the third group of queries grouping and selecting of records were performed. The results of these three categories of queries represent different datasets on which different machine learning models were built. Each of the obtained datasets were divided on the training and test dataset. Records related to the years 2015 and 2016 were used for creating training dataset, while records belonging to the year 2017 were used to make test dataset. SQL queries were created and training and test datasets were generated using MS Access 2016.

The following attributes were selected as labels (attributes which values will be predicted): Net weight [kg] and Amount [EUR]. On different datasets different machine learning models were created based on the application of the following algorithms: Linear Regression, Multilayer Perceptron (Neural Network), SMOreg (Support Vector Machine for Regression), IBk (k-Nearest Neighbors), M5P, Random Forest, Random Tree and REPTree.

Example: Training dataset: "COCA-COLA HELLENIC BOTTLING COMPANY-SERBIA" Export of Water to Montenegro

by Year and Month and Custom Tariff - NetWeightT 2015-2016, for custom tariffs: 2201101100 and 2202100000; number of instances: 175; attributes: Month, CountryOfExport, CustomsTariff, SumOfNetWeightKG; Test mode: 10-fold cross-validation.

The performance of the first four different machine learning models created by using four different algorithms on this training dataset are shown in Table 1.

Machine learning algorithm	**Linear Regression**	**Multilayer Perceptron**	**SMOreg**	**IBk**
Correlation coefficient	0.7265	0.9477	0.6777	0.972
Mean absolute error	1050.97	404.73	950.78	231.16
Root mean squared error	1425.92	690.34	1609.56	489.84
Relative absolute error [%]	71.25	27.44	64.45	15.67
Root relative squared error [%]	68.56	33.19	77.39	23.55

Table 1. Performance Measures For The First Group of Prediction Models

The performance of the second four machine learning models created on the same training dataset are shown in Table 2.

According to the results of the models shown in Tables 1 and 2, the model that has the best performance was based on IBk (k-Nearest Neighbors) algorithm. Considering all performance measures this model shows the best performance. The second place share two models with very similar performances: the first is based on Random Forest and the second one on the Random Tree algorithm. The Fig. 1 shows relationships between actual values for the year of 2017 from test dataset, and the values predicted using the machine learning models that are selected as the best ones (IBk and Random Forest).

Algorithm	**M5P**	**Random Forest**	**Random Tree**	**REPTree**
Correlation coefficient	0.9265	0.9691	0.9691	0.9585
Mean absolute error	448.98	256.98	248.62	321.73
Root mean squared error	791.97	514.14	516.88	592.76
Relative absolute error [%]	30.44	17.42	16.85	21.81
Root relative squared error [%]	38.08	24.72	24.85	28.50

Table 2. Performance Measures for the Second Group of Prediction Models

5. Conclusion

The most important conclusion of this research is that food foreign trade dataset can be used to build supervised machine learning models that can perform satisfying results in predicting of volume and the structure of import and export of the food

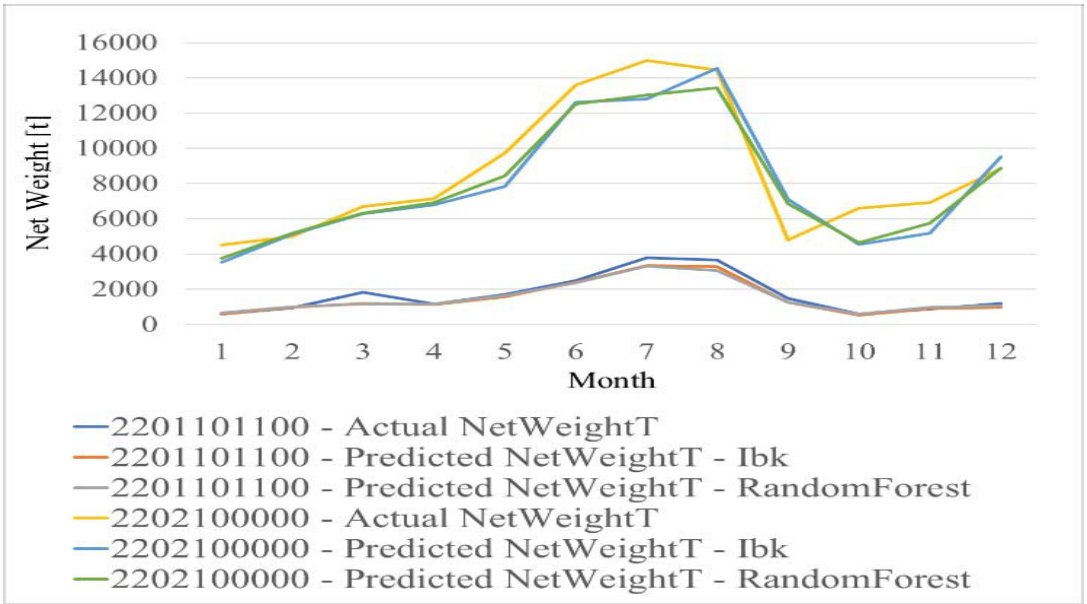

Figure 1. Actual and predicted "Coca-Cola Hellenic Bottling Company-Serbia" export of water to Montenegro by month and custom tariff – net weight [t]

products in the Republic of Serbia. Models that have shown the best performances were based on k-Nearest Neighbors, Random Forest 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. The Serbian Food Foreign Trade dataset has been provided by the company CUBE Team d.o.o Belgrade.

References

[1] Manyika, J., Chui, M., Brown, B., Bughin, J. (2011). Big Data: The next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011. Available at: https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf

[2] Dijcks, J. P. (2011). Big Data for the Enterprise, Oracle Corporation, 2011. Available at: http://resources.idgenterprise.com/original/AST- 0054994_DW_US_EN_WP_BigData.pdf

[3] Kaisler, S., Armour, F., Espinosa, J. A., Money, W. (2011). Big data: Issues and challenges moving forward. *46th Hawaii International Conference on System Sciences*, p 995–1004, Wailea, 2013.

[4] Russom, P. (2011). Big data analytics, TDWI Best Practices Report, Fourth Quarter, 2011.

[5] Sanders, N. R. (2016). How to Use Big Data to Drive Your Supply Chain, *California Management Review*, 58 (3), p 26–48, 2016.

[6] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers, *IBM Journal of research and development*, 3 (3), p 210-229, 1959.

[7] Mitchell, T. (1997). *Machine Learning*. McGraw Hill, 1997.

[8] Jain, A. K., Murty, M. N., Flynn, P. (1999). Data clustering: a review, *ACM Comput Surveys*, 31 (3), p 264–323, 1999.

[9] Hofmann, E. (2015). Big data and supply chain decisions: the impact of volume, variety and velocity properties on the bullwhip effect, *International Journal of Production Research*, p 5108-5126, 2015.

[10] Columbus, L. (2011). 10 Ways Machine Learning Is Revolutionizing Supply Chain Management, 2011. Available at: https://www.forbes.com/sites/louiscolumbus/2018/06/11/10-ways-machine-learning-is-revolutionizing-supply-chainmanagement/# aff94263e370

[11] Gesing, B., Peterson, S. J., Michelsen, D. (2018). Artificial Intelligence in Logistics - A collaborative report by DHL and IBM on implications and use cases for the logistics industry, *DHL Customer Solutions & Innovation*, 2018. Available at: https://www.logistics.dhl/content/dam/dhl/global/core/documents/pdf/glo-artificial-intelligence-in-logistics-trend-report.pdf

[12] Shamir, J. (2014). Machine learning: A new tool for better forecasting, CSCMP's Supply Chain, Quarter 4, 2014.

[13] Finextra. (2018). Intelligent Machines and FX Trading, 2018. Available at: https://www.finextra.com/blogposting/15405/intelligent-machines-and-fx-trading

[14] Editorial Team, How Machine Learning Can Help Reduce Foreign Exchange Risk, 2018. Available at:https://insidebigdata.com/2018/03/30/machine-learning-canhelp-reduce-foreign-exchange-risk/

[15] Zarmehri, M. N., Soares, C. (2015). Metalearning to choose the level of analysis in nested data: A case study on error detection in foreign trade statistics, *International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland 2015.