©LUMEN Proceedings

Available online at: http://lumenpublishing.com/proceedings/published-volumes/lumen-proceedings/nashs2017/

3rd Central & Eastern European LUMEN International Conference
New Approaches in Social and Humanistic Sciences | NASHS 2017 |
Chisinau, Republic of Moldova | June 8-10, 2017

New Approaches in Social and Humanistic Sciences

Optimized Demand Forecasting by Cross-Validation

Răzvan Daniel ZOTA*, Yasser AL HADAD

https://doi.org/10.18662/lumproc.nashs2017.50

How to cite: Zota, R., D., & Al Hadad, Y. (2018). Optimized Demand Forecasting by Cross-Validation. In V. Manolachi, C.M. Rus, S. Rusnac (eds.), *New Approaches in Social and Humanistic Sciences* (pp. 563-574). Iasi, Romania: LUMEN Proceedings. https://doi.org/10.18662/lumproc.nashs2017.50

©LUMEN Proceedings

3rd Central & Eastern European LUMEN International Conference
New Approaches in Social and Humanistic Sciences |
NASHS 2017 | Chisinau, Republic of Moldova | June 8-10, 2017

Optimized Demand Forecasting by Cross-Validation

Răzvan Daniel ZOTA[1]*, Yasser AL HADAD[2]

Abstract

Sales forecasting plays an important role in business strategy. An appropriate demand forecasting model is necessary for reducing the cost of storage. At a company level, lowering the warehouse costs and optimizing the value chain is a prominent requirement for an optimum stock management. In this paper a demand forecasting model is built to support the stock management activity of medium enterprises by means of data mining algorithms. SQL server analysis service is used for implementing the demand forecasting model. The paper studies a list of available algorithms that are offered by SQL server analysis service and the performance of the aforementioned algorithms is tested using the cross-validation feature that is provided by SQL server analysis service to optimize the performance of the model. We also aim to explore in our research the ability of RMSE (Root mean Squared Error) to include time series algorithms in the cross-validation phase. The proposed model is tested based on a dataset of a timber export company and the output is used for analysing the performance of the proposed model. The paper reached a group of conclusion and one of most the importance conclusion is neural network algorithms performance was the better in adapting our tested dataset comparing with the other algorithms.

Keywords: *Demand forecasting, BI (Business intelligence), SAS (SQL analysis services), cross-validation, data analysis.*

[1] The Bucharest University of Economic Studies, Bucharest, Romania, zota@ase.ro.
[2] The Bucharest University of Economic Studies, Bucharest, Romania, dukeyasser10@yahoo.ro.

https://doi.org/10.18662/lumproc.nashs2017.50
Corresponding Author: Răzvan Daniel ZOTA
Selection and peer-review under responsibility of the Organizing Committee of the conference

1. Introduction

Demand forecasting has an increasing interest for economic processes. An appropriate demand forecasting model is necessary in order to adapt the operational activities. In this paper, we aim to build demand forecasting model that can be integrated within existent transactional system using business intelligence. The proposed model is tailored and generalized to meet the requirement of Romanian medium companies. The proposed model is analyzed and validated based on data set of Romanian Timber Export Company.

The model performance will be optimized by testing multiple algorithms and comparing their results against each other. SQL server Analysis service is used to build the forecasting model. The following sections document the available algorithms in SSAS (SQL server analysis services) that can be used for building a demand forecasting model. The sections that follow describe the process of designing the proposed model. The model will then be evaluated using the cross-validation feature of SQL server as well as RMSE (Root mean Squared Error) to include time series algorithms in validation the phase. Some different data aggregation scenarios are tested in order to optimize the results. The database of a Timber Export Company is used for validating the selected model.

2. Problem Statement

The research carried out for this paper is motivated by the small number of studies that address analysis systems related to Romanian timber export. In the context of the absence of such systems, the demand for a forecasting model is considerable. The most prominent parameters that can influence the demand forecast of timber are analysed and configured in the structure of the proposed demand forecasting model. A well-defined demand forecasting model helps in determining the minimal level of inventory that should be guaranteed. An efficient demand forecasting model includes usage of appropriate forecasting methods.

One of the main methods that are widely used for generating a demand forecasting model is the time series regression method. Methods such as ARIMA and Holt-Winters were tested for demand forecasting in food retail [1], and it is stated that both models reached a compatible performance with minimal difference in favour of Holt-Winters due to the feature of seasonality. The Holt-Winters method is recommended for short term predictions whereas ARIMA can be still accurate with longer periods.

Both approaches are implemented in SSAS: ARTXP (Autoregressive Tree Models with Cross Prediction) that is optimized for short-term prediction algorithms, and ARIMA (autoregressive integrated moving average). Moreover, the seasonality feature is handled in both algorithms [2].

Other methods that can process and incorporate more assumptions are used in demand forecasting such as neural network that was used in demand forecasting to support the supply chain management application [3], and the Monte Carlo method that was used for forecasting or estimation of expected demand of automotive aftermarket market [4]. Both algorithms enable the feature of what-if analysis to test various scenarios that is required by managers to build an order scenario. A further comparison of both algorithms in case of nonlinear optimal estimation is performed. It points out that the calculation amount depends on the size of designed neural networks, with the possibility of reducing this size using the Monte-Carlo method. On the other hand, the potential adaptive properties of the neural network method reduce the burden of designing model [5]. The performance of the methods mentioned above is proven by many researches. However, their results are difficult to interpret. It is hard to understand the correlation between data and how the result is derived due to heavy calculations performed by every algorithm. In this context, a tree decision algorithm can also be used to model the forecast model due to its ability to predict continuous values [6]. The main advantage of the decision tree method is the graphical representation of its result that is easy to interpret.

3. Research Questions

This study aims to solve the problems of demand forecasting encountered by medium enterprises. The study includes the particular characteristics of the Romanian business environment. In this context, a dataset for a Romanian Timber Export Company is considered for designing and testing the proposed model. In order to build the structure of the proposed model, related aspects such as demand, environmental specificities and available information sources are taken into consideration. The most important research questions that are investigated by this paper can be summarized as follows:

- What are the characters and features of exporting problem?
- What is the best forecasting model that can adapt better per our data set?
- What is the optimum aggregation data type that can improve the forecasting task?

Choosing the best forecasting model is the central concern of the research. Based on the reviews put forth in the specialized literature, the best performing methods are implemented, and the performances of the selected algorithms for the demand forecasting are tested using RMSE (Root mean Squared Error) as a performance measure. The potential of SQL server and its features for implementing the proposed model is presented.

4. Research Methods

The objective of this section is to implement a model that can forecast the total sales amount based on existing historical data. SQL server Analysis service is used to build the forecasting model. All required tools as well as the building of the model, including the validation phase for implementing the model are presented in the following subsections:

4.1 SQL Server Analysis Services

SSAS is one of leaders in business logic analysis software. Its integrated platform and services is one of the largest frameworks that provide BI solution. SSAS solution offers a number of data mining algorithms. A data mining algorithm is a set of calculations and heuristics that analyses the data source, looking for patterns or trends and uses the results to determine the optimal parameters for building the mining model. These parameters are applied across the data set to extract detailed statistics and actionable pattern. SSAS offers some data mining algorithms. Three types of algorithms that have the ability to predict continuous parameters will be tested in this paper.

4.1.1 Microsoft Time Series Algorithm

It is one of algorithms that are provided by SQL server analysis services. It offers two algorithms for predicting continuous values: ARTXP for short-term prediction and ARIMA for long-term prediction [7]. By default, Microsoft time series algorithm trains both models separately. Then, both models are blended to get a better result. There are many applications where the Time series can be used such as forecasting sale over time. The most important assumptions of the time series algorithm are that the discovered historical pattern will not change during the forecast period. If the model assumptions are true, we obtain an excellent forecast. Otherwise

the results cannot be satisfactory. We use ARTXP algorithm as it is applied more successfully in comparable settings [1].

4.1.2 Microsoft Neural Network Algorithm

The Microsoft neural network algorithm is used in building data mining models. It has various desirable characteristics. It enables analysing complex patterns included in the input data. It is useful for deriving rules that other algorithms cannot perform properly. It is capable of exploiting different input information that might have some correlation with output and affect the results. Microsoft neural network algorithm combines each possible state of the input attributes with each possible state of the predictable attribute. The algorithm is influenced by the values of the parameters. For instance, if price is changed or promoted the demand will be affected, and it can distort the time series as it uses only the historical demand information for forecasting. Neural network can be used to analyse complex relationships between multiple inputs and multiple outputs. It is flexible and can analyse any combination of inputs and outputs. So it can have multiple predictable columns, but processing time can increase exponentially for a large data set. The resulted neural network can be presented by Microsoft neural network viewer.

4.1.3 Microsoft Decision Tree Algorithm

Decision Trees algorithm is a supervised classification algorithm that is implemented by Microsoft decision tree algorithm. The resulted decision tree can be used for regression of both continuous and discrete attributes [6]. Decision tree model tells about factors that influence the output value. In case of discrete attributes, the algorithm builds the decision tree based on the relations between input and output attributes in a dataset. The algorithm uses the values that are known as states, of those input attributes to predict the states of output attribute. The algorithm identifies the input attributes values that are correlated with the output attributes values. The algorithm infers the values that are a good predictor of output value. In case of continuous attributes, decision tree algorithm builds the decision tree using linear regression. Linear regression is used to determine how a decision tree should be split. A separate decision tree is built for every predictable attribute. Decision tree algorithm is used at this paper to implement demand forecast demand.

The algorithm builds a decision tree model by defining a series of splits in the tree. These splits are described in nodes. The algorithm adds a new node in case of discovering an input attribute that is correlated with the predictable attribute. The method of determining a split depends on whether the output attribute is a discrete column or a continuous column. Feature selection is used by Microsoft Decision Trees algorithm. It helps the data mining model to improve the quality and performance of analysis. It is a guide for the algorithm to select the most useful attributes for splitting the decision tree. The resulted decision tree can be presented by Microsoft tree viewer.

4.2 Cross -Validation

Cross-validation is an important feature in analytics. It is a standard tool that helps in fine-tuning data mining models. The processes of building and training data mining models are followed by the cross-validation phase in order to ascertain the validity of the created model. Cross-validation is used for validating the robustness of a particular mining model or identifying the best model by comparing multiple models. Cross-validation can be used by data mining designer or by running stored procedures. The Data Mining Designer tool enables configuring in a single dialog box both the training data and the accuracy results parameters. This makes measuring the accuracy easier for mining models that are included in a single mining structure. Cross-Validation is part of Mining Accuracy Chart view. It is available in either SQL Server Development Studio or SQL Server Management Studio. Cross-validation is available in the form of parameterized system stored procedures for advanced configuration. The stored procedures offer advantages such as editing the process and added customizations. For every type of mining model, cross-validation is performed by stored procedures in two separate phases: generating metrics for the entire data set or partitioning data and then generating metrics for partitions. During cross-validation, the data in a mining structure is divided by analysis services into multiple cross-sections. Then, cross sections are tested iteratively. This analysis outputs a set of accuracy measures for each model. The reliability of the model can be assessed by comparing the metrics.

4.3 Sample study of timber export company data set

The tested data set of Romanian timber Export Company is used to analysing and validation the proposed model. At timber export business

field, Romania is considered as one of the main timber exporters for a large number of countries. The activity of timber export companies is influenced by many factors. The prices of raw material for timber in Romania have witnessed a phenomenal growth in the last ten years so exporters have to adapt their activity to counteract the consequences of growing prices. Timber prices and production are directly determined by the demand for forest products. They are dependent on the capability of producers to harvest trees and transform them into products [8].

In the area of timber export, there are many constraints and criteria that should be taken into consideration. Supplying price and the possible risk of delivery interruptions can vary among producers. Production sites are located in areas that are exposed to bad weather conditions, consequently the production plans can be interrupted. The availability and pricing of raw materials for timber production vary largely based on the regions from which the material is sourced. The delivery of the supplies can be delayed in some locations having an infrastructure that is vulnerable to bad climate conditions. Long-term market stability is also affected by policies such as the introduction of the offset payment law for carbon emissions. As a result, these factors affect the supplying schedule and cause a fluctuation in the sales demand. Sales forecasts or demand forecasts depend on many factors such as the consumables price, market demand, promotion strategy used, quality of service, and so on. The sample study of timber export company data set that are used at this paper, includes sales data about 500 products that are extended on period of 41 months.

4.4 Preparing the Model

The objective of this section is to implement a model that can forecast the total sales amount based on existing historical data. SQL server Analysis service is used to build the forecasting model. It includes several algorithms that can deal with continuous parameters types such as Microsoft time series, neural networks and decision tree. Microsoft time series uses an auto regressive method to predict the output parameter based on a previous output parameter trend without requiring any other input parameter. Auto regressive algorithm is useful in case of the inexistence of any unknown parameters that can influence the output. The possibility of testing the input parameters that can influence the final output can be realized by using neural network algorithm. Microsoft neural network implements neural network algorithms. It permits adding input variables beside the output parameter to the structure of data mining model. Microsoft neural network viewer offers

an interactive user interface that shows the relation between parameters and their values. It also offers if-else feature that enables users to perform different experiments on the data to explore their trends and patterns. Neural network consists of input layer, output layer and hidden layer. The relation between input and output layers, and the way that hidden layer is configured is difficult to understand for any user. Therefore, Microsoft decision tree was used for testing their potential in implementing forecast demand model. It uses input parameters to classify the output values. The result of the decision tree algorithm is presented by decision tree viewer. The advantage of decision tree is that its diagram can be understood easily, so the user can realize how input and output values are related.

The wrong configuration of data mining model can affect the performance of the model. It is not recommended to use multiple columns that are correlated strongly with each other or derived from the same data. The model includes purchase price field and profit percentage field but not sales price as it can be derived from these fields. Also the payment type is an essential factor that can affect the sale but is not included in the model as it can be derived from total investment field and existing stock field that are included in the model. Also, sales date field is used to capture any possible seasonal pattern that can influence the output value. Finally total sales field is included in the model and it is used as predicted field.

4.5 Algorithms Validation

In the previous section, the demand forecasting structure has been analysed and modelled, and then several data mining models with different data mining algorithms have been added to the demand forecasting structure. The objective of this section is to test the performance of the implemented mining models and to compare their performance. A single data structure can have multiple data mining models, and every data mining model can have a different data mining algorithm and a different data filter. Data source cases are divided into training cases and validation cases. 70% of cases were used for training data mining models and 30% of cases were reserved for validation to prevent any overtraining. The algorithms of Microsoft Time series cannot be evaluated by the mining accuracy chart of SSAS as it can predict only the period that is followed by data set training. Lift chart was used for comparing neural network algorithm and decision tree algorithm. The monthly aggregated training data set was used for evaluation. Figure 1 shows that the score of the neural network algorithm is higher than that of decision tree algorithm.

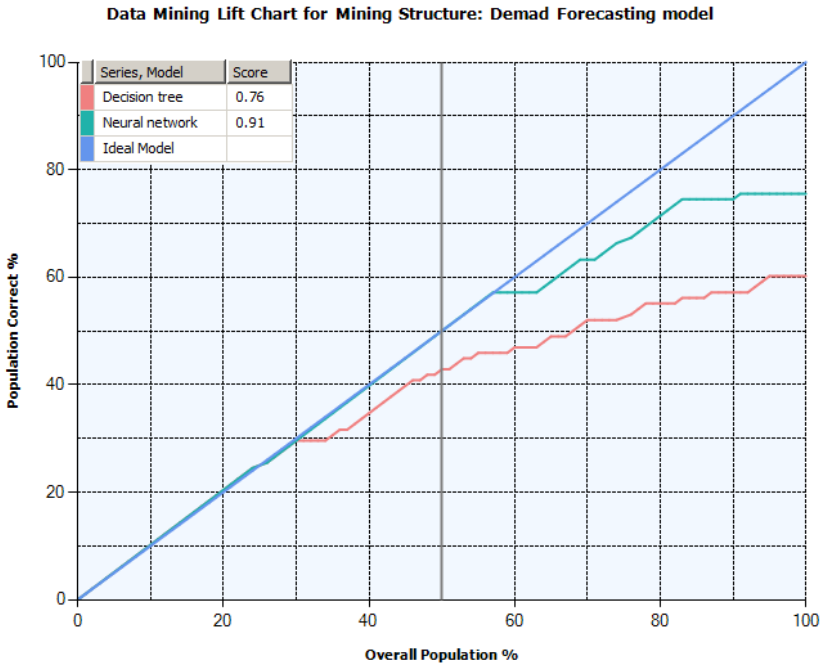

Figure 1. Data mining lift chart for demand forecasting model

The standard deviation of data mining models is calculated for validation. It is also called "Root mean Squared Error (RMSE)" or "Validation standard error of the Mean". It is used to measure the deviation of predicted data from real data. The dataset that was used for training data mining models were aggregated in three ways: monthly (30 days), bimonthly (60 days) and quarterly (90 days). The trend of sales amount is too fluctuated. We assume that data aggregation by longer period can decrease the fluctuation and enhance the performance of the data mining model. The standard deviation of data mining models is calculated for validation. It is used to measure the deviation of predicted data in comparison with the real data.

Another way to compare the models is by using testing data sets in addition to the training and validation data set. This helps in including the time series model into the validation step. Time series algorithm implements auto regressive methods. The auto regressive model can predict only the period that is followed by the training data set. In order to evaluate the auto regressive model, we have to process the model multiple times, every time processing with a portion of the training data, and then measuring the deviation of the resulting output against the actual testing data set. RMSE

was used for comparing the data mining models of the demand forecasting structure, see figure 2.

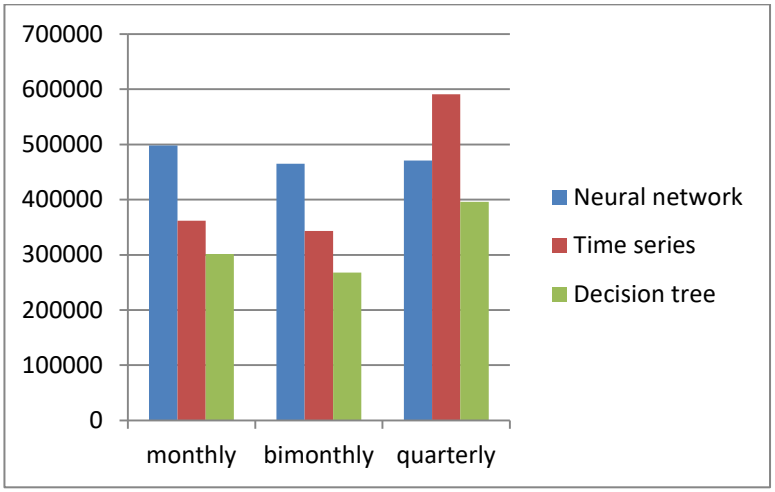

Figure 2. Testing results of demand forecasting structure using RMSE method

5. Findings

Figure 1 is generated using SSAS. It shows that the score of the neural network algorithm is higher than that of the decision tree algorithm. It is expected that the low size of training data set and the complex relation between the input and output parameters decrease the performance of decision tree. Figure 2 compares the performance of the data mining models using RMSE. The error rate of the quarterly aggregation was too high. The possible reason is that the size of data set wasn't enough to help any algorithm to build an accurate model. Decision tree algorithm failed to build a decision tree with many nodes. It just combines all data in one node without splitting them in a regular tree. The single formed node has constant output value that represents the mean output value of training data set. This single output value helped decision tree algorithm to perform better at bimonthly aggregated data set as the level of fluctuation is less than the monthly aggregated data. In the case of bimonthly aggregated data, the size of data set was not enough for neural network algorithm to perform optimally. Neural network had the best performance in forecasting based on monthly aggregated data. It also enables if-then analysis report whereas the decision tree algorithm failed to build a decision tree for if-then analysis test.

6. Discussions

A demand forecasting model for the field of timber export was implemented using data mining algorithms. The model was trained using historically collected data. SQL server platform is used for implementing the model. The training data includes 41 month. The sale amount trend of the training data is too fluctuating. The assumption was that data aggregation over a longer period can decrease the fluctuation and enhance the performance of the model. So in order to evaluate this assumption, further training data were aggregated in three ways in order to test many scenarios: monthly (30 days), bimonthly (60 days) and quarterly (90 days).

Three algorithms were tested that have the ability to predict continuous parameters type. Microsoft time series uses the auto regressive method to predict the output parameter that is based on previous output parameter trends without requiring any other input parameter. The most important assumptions of the time series algorithm are that the historical discovered pattern will not change during the forecasting period. Microsoft neural network permits adding input variable to the structure of the data mining model. It offers an if-else feature that enables users to perform different experiments on the data in order to explore their trends and patterns, however the way that input and output data are related remains invisible for users. Microsoft decision tree was also used for testing its potential in implementing a demand forecasting model. The advantage of decision tree is that its diagram can be understood easily, so the user can identify with ease the correlation between input and output.

The proposed model structure takes into consideration previous related studies as well as our specific case of a Timber Export Company in order to identify some extra influencing parameters and particularities that characterize the Romanian business environment. For a more general overview of the issue, the study should be extended to a larger number of medium size companies. In addition, a larger dataset should be used in order to test the reaction of the selected algorithms.

7. Conclusions

A dataset provided by a Timber Export Company and its entire related business environment were analysed and structured in our proposed model. There are some characters and features that are captures in the phase of analyzing that can also be generalized to other business fields. Medium enterprises prefer to collaborate with medium enterprises and local

enterprises for supplying due to its advantageous prices. However, there are some consequences that should be taken in consideration such as low-quality workmanship and unstable production at the plant, supplying price fluctuation and raw material availability, and weather conditions and possible risk of delivering interruption.

The selected algorithms were tested based on the cross-validation feature and RMSE. Findings show that neural network has the ability to discover complex existing relationships between the input and output parameters. It has the best accuracy performance in processing monthly aggregated data, and more importantly, it enables if-else analysis query.

8. Acknowledgments

Part of this work is done under the auspices of the doctoral studies within the Doctoral School of Economic Informatics, Bucharest University of Economic Studies.

References

[1] Claudimar P. D., Cassia R. P., Anderson C., Ubirata T., & Wesley V. D. Demand forecasting in food retail: a comparison between the Holt-Winters and ARIMA models. WSEAS TRANSACTIONS on BUSINESS and ECONOMICS. 2014 (11). pp. 608-614

[2] Meek C., Chickering D. M., & Heckerman D. Autoregressive tree modes for time-series analysis. SIAM International Conference on Data Mining. 2002. (2). pp. 229-244

[3] Kochak A., & Sharma S. Demand forecasting using neural network for supply chain managemnt. IJMERR. 2015 Jan, 4(1). pp. 96-104

[4] Dobrican, O. Forecasting Demand for Automotive Aftermarket Inventories. Informatica Economică. 2013. 17(2). pp. 119-129

[5] Stepanov O. A, & Amosov O. S. The Comparison of the Monte-Carlo method and neural networks algorithms in nonlinear estimation problems. IFAC Proceedings. 2007. 9(1) pp. 392-397

[6] Pang-Ning T., Michael S., & Kumar V. Chapter 4: Classification: Basic Concepts, Decision Trees, and Model Evaluation. Introduction to Data Mining. Boston, MA, USA: Addison-Wesley Longman Publishing Co. 2005. p. 157

[7] ***, SQL Server 2012 Tutorials: Analysis Services - Data Mining [Internet]. SQL Server 2012 Books Online. 2012. (cited 2017 Jun 05). P. 55

[8] Buongiorno J. Global modelling to predict timber production and prices:the GFPM approach.. Forestry. 2014 Dec 4, 88. pp. 291–303. Available from: https://doi.org/10.1093/forestry/cpu047