Modelling Buying Demand in the Tourism Industry based on Machine Training Methods

Viktoriya Tkach, Anatolii Pavlenchyk, Olena Sadchenko, Svetlana Nikola, Valeriia Drozdova, Iryna Davydenko

***Abstract*: The business processes of companies in the tourism industry lend themselves well to formalization and, consequently, computer automation. This study focuses on the process of creating a demand forecast model for a travel agent based on data mining algorithms. The program code was developed in the Anaconda development environment, which allows to process the initial data and to give the prediction results for two indicators of MAE and program accuracy. The program code is designed to improve the performance of the entire system by selecting the correct functions.**

***Index Terms*: ARIMA, Box-Jenkins model, buying demand, machine training, tourism industry.**

I. INTRODUCTION

Retail market conditions are a vital indicator of a country's economic development. As one of the most mobile areas of the economy, the tourism sector reflects the work of economic laws, both micro and macro. After the trials of the global financial crisis, the current structural and economic crisis, the widespread development of the Internet, the competitiveness of tourism industry enterprises has become primarily determined by the ability to optimize internal business processes and sales flows. This leads to the creation of new methods and approaches in the policy of organizing the tourism business, in particular, in the field of planning, forecasting and coordinating the service movement.Modern methods of applied statistics are widely used in many sectors of the economy due to the technological boom. The development of speed and efficiency of computational algorithms today allows processing large data arrays (Big Data) even on personal computers using open source software. Having the basis for the development of an effective platform for modelling consumer demand, the task is to increase the accuracy of forecasting. Based on the availability of methods developed in theory and practice - unique linear regression methods, decision trees and their compositions, the process of support vectors, neural networks, gradient boosting, etc. - the task of high-precision modelling needs is achievable.

Revised Manuscript Received on July 06, 2019.

ViktoriyaTkach, Department of International Tourism, Hotel & Restaurant Business and Language Training Alfred Nobel University Dnipro, Ukraine

AnatoliiPavlenchyk, Department of Economics and Management Lviv State University of Physical Culture Lviv, Ukraine

OlenaSadchenko, Department of Marketing and Business Administration Odessa I.I. Mechnikov National University Odessa, Ukraine

Svetlana Nikola, Department of Finance, Banking and Insurance Odessa I.I. Mechnikov National University Odessa, Ukraine

ValeriiaDrozdova, Department of Management and Logistics Odessa National Academy of Food Technologies Odessa, Ukraine

Iryna Davydenko, Department of Tourism, Hotel and Restaurant Business Odessa National Economic University, Odessa, Ukraine

To build a developed consumer demand forecasting system, there is an urgent need to implement a forecasting model based on the methods of the modern theory of econometrics, statistical and machine learning, as well as elements of the theory of the economics of commerce, while having a concrete form in the form of an implemented software and computer complex integrated into information systems the company. The presence of this need and due to the relevance of this study. To unambiguously define the concept of forecasting the demand for a product, it is necessary to reflect the general idea of forecasting in the modern scientific environment. For example, there is a marketing approach to demand forecasting, which can be reflected in the following definition of Kotler and Keller: "Forecasting is the art of foreseeing customer behaviour in certain circumstances based on an analysis of survey results" [1]. Lopatnikov gives the classic definition of demand forecasting: "Forecasting of demand is a study of the future (possible) demand for goods and services in order to substantiate relevant production plans better" [2]. This clarifies that for the implementation of the forecast statistics is used on the application of services, and, as a consequence, it is assumed the analysis of previous sales. Accordingly, it is clear from the definitions that the main essence of demand forecasting is the characteristic of the object of prediction - in this case, the demand for services - and the approach to forecasting. Within the article, the procedure is defined as forecasting based on statistics about previous sales and the conditions under which they were carried out. Here we should single out the space of classical methods for demand forecasting: time series analysis and regression-econometric analysis and advanced machine learning methods [3-5]. Considering the chosen techniques, the mathematical nature of demand forecasting is laid. At the same time, precise prediction is understood as a formalized type of forecasting, which is based on the construction of a mathematical model of a process that captures most of its laws. Before proceeding with the prediction procedure based on mathematical methods, it is necessary to analyze the data set (X, Y) on which the model will be built. As part of the task of forecasting demand, data usually consists of a set of characteristics of the situation in which a particular product was claimed by the buyer in a defined quantity [6]. These characteristics can be quite a lot: from the weather conditions near the tour operator to the cost of the tour, for which demand is predicted. In the terms of a large amount of data on possible features X a correlation analysis is carried out. To begin with, correlation analysis reveals the tightness of the signs X with the independent variable Y. This procedure identifies the most informative

variables for building a model, which is very important in terms of limited resources for using data. Further, correlation analysis is used to identify strong dependencies within economic data on features *X*, to use them corrected for this dependence or reduce their dimension by eliminating elements that do not carry significant information. At the same time, there is an understanding of the presence of non-linear relationships within the framework of specific manifestations of the signs. This non-linearity can be taken into account by creating non-linear combinations between features, for example, the procedure of pairwise multiplication, polynomial, exponential, logarithmic transformation and many others. Usually seek to improve the predictive ability of the result, without losing the meaning of interpretation. An essential part of choosing the set of features *X* for predicting the quantity is the absence of logical contradictions. For example, when forecasting demand, it is incorrect to use the value of the residual at the time as a predictor, despite the strong relationship with the target variable. The reason is trivial - any fluctuations in the final commodity balance entirely depend on the actual demand for the product, and not vice versa. Otherwise, there is a risk of "looping" the model, making it inoperable. The same can be said about quantities about which there is no information at the time of the demand estimate of $t0$. Therefore, to use such features, it is necessary to develop prognostic models for such variables. If this is feasible, the basic prediction model can function normally or even better compared to more straightforward implementations [7-9].

II. ANALYSIS OF PECULIARITIES AND SELECTION OF THE MACHINE TRAINING METHOD FOR FORECASTING DEMAND

There are various methods for predicting time series, among which the following groups are particularly distinguished:

- Box-Jenkins techniques;
- regression methods;
- neural network methods.

In this paper, we will consider the Box-Jenkins methodology, better known under the name ARIMA (AutoRegressive Integrated Moving Average), which is a model of autoregression and an integrated moving average. The ARIMA model can include both models at once or each separately and is designated: *ARIMA (p, d, q)*, where *p* is the order of the moving average. This model is an extension of the *ARMA (p, q)* model (ARMA serves to predict stationary time series) and is used if the process is non-stationary and to take it to a stable form, it was necessary to take several differences [8]. The ARIMA model belongs to the class of statistical models for analyzing and forecasting time series. The model name is an acronym, the capital letters of which mean the following: AR: Autoregression is the use of a connection between current and some late observations.

I: Integrated is a process that uses the difference between the current and the previous observation to keep the time series unchanged. MA: Moving-average is a model where the relationship between observation and residual errors from the moving average model is used concerning delayed views.

Each of these components is explicitly listed as a parameter to the model.

The standard notation used in ARIMA (p, d, q), where parameters are replaced with numbers in order to quickly indicate the particular model used.

Model parameters mean:

p is the number of delayed observations contained in the model, also known as the lag order;

d is the number of the order of the time series difference;

q - the size of the moving average window (the request of the moving average).

The ARIMA model (p, d, q) for a non-stationary time series Yt:

$$\Delta^d Y_t = c + \sum_{i=1}^{p} a_i \Delta^d Y_{t-1} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t \qquad (1)$$

where c, a_i, $_{bj}$ - the model parameters;

Et - stationary time series;

Δd - difference operator of order d.

ARIMA's approach to time series is that the stationarity of the series is first evaluated. Various tests reveal the presence of unit roots and the order of integration of the time series (usually limited to the first or second order). Further, if necessary (if the order of integration is more significant than zero), the series is transformed by taking the difference of the corresponding request and already for the modified model some ARMA model is built, since it is assumed that the resulting process is stationary, unlike the original non-stationary method (differential-order or integrated process d).

III. PRACTICAL APPLICATION OF THE METHODOLOGY FOR FORECASTING THE BUYING DEMAND

A. *Means of implementation*

The following methods of implementation was used:

- RStudio is a free open source software development environment for the R programming language, which is intended for statistical data processing and graphics. RStudio is available in two versions: RStudio Desktop, in which the program runs on a local machine as a typical application; and RStudio Server, which provides access via a browser to RStudio installed on a remote Linux server. RStudio Desktop distributions are available for Linux, OS X and Windows. RStudio is written in the C ++ programming language and uses the Qt framework for the graphical user interface.
- XGBoost is a library of open source software that provides improved performance and adds new functionality for programming languages such as C ++, Java, Python, R, and Julia. It runs on Linux, Windows, and MacOS. In addition to working on the same machine, it also supports Apache Hadoop, Apache Spark, and Apache FLiNK distributed processing databases.
- Microsoft Excel is a spreadsheet program created by Microsoft for Microsoft Windows, Windows NT and Mac OS, as well as Android, iOS and Windows Phone. It provides economic statistics, graphical

tools and, except for Excel 2008 on Mac OS X, VBA (Visual Basic for Application). Microsoft Excel is part of Microsoft Office and today, Excel is one of the most popular applications in the world.

B. *Description of the data source*

The data for solving the problem were obtained from the software package "Oli IS", developed for and used by the travel agent Olympia, which currently can analyze the number of orders and total revenue daily, but which does not offer solutions for tour optimization and demand forecasting. Therefore, we unloaded data for several years to train our neural network and draw up an algorithm for forecasting orders of tours. To do this, it was necessary to extract data and make an unusual sample, which can be used to make the most accurate and useful forecast. To begin with, the available data was cultivated for further beneficial use. Thus, the variable of the month, which captures seasonal changes, and the year assumed by macroeconomic factors, such as crises, inflation, etc., were obtained from the date variable as separate indications.

C. *Development of a demand forecasting model*

The data mining process presents the following steps in Fig. 1:

- creating a model using a specific algorithm;
- training the model using training data (in the training data, the initial attributes and attributes that will be predicted are known);
- providing the predictive data for the data mining model (predictable characteristics are unknown) followed by the determination of the values of hidden attributes.

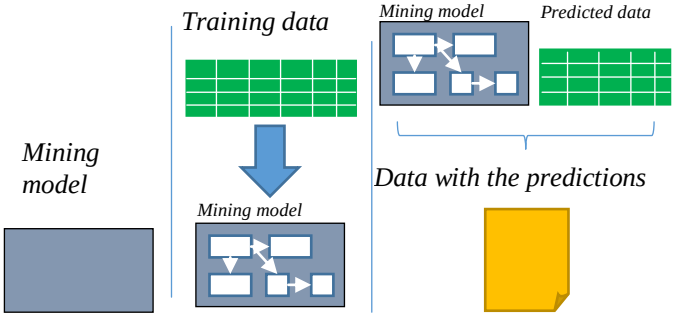

Fig. 1 The data mining process

The simulation was carried out according to the time series system (the future according to the past). Therefore, using various machine learning methods, different suitable models were obtained. The best technique turned out to be ensemble gradient boosting, which minimizes the loss function. In our case, nested in increasing models were trees, and the ensemble building procedure itself is iterative, which results in the presence of a set of adjustable parameters in the algorithm that need to be tuned.

The main functions used in the code to create a demand forecasting model:

- Disconnect on *train* and *test*
- Training on data
- Model training
- Prediction of results
- Calculation of MAE indicators and forecast accuracy

To understand the logic, we present a part of the program code for the prediction model (Fig. 2).

```
def trainModelTestTrainSplit(Data, Model):
    C =Data.Columns.differences(['Count'])
    X_train , X_test, y_train, y_test = sklearn.cross_validation.train_test_split(Data(C)), Data(Count), test_size = 0.3
    Model.Verbose =1
    Model.fit X_train.as_matrix(), y_train.as_matrix()
    pred =Model.predict(X_test)
    Print ("result: MAE is " = mae(pred, y_test))
    Print (Str(mape_vectorized(pred, y_test)) + "% is accurancy.")
    return model
```

Fig. 2 Code with essential functions for the prediction model

D. *The results of the prediction model*

The most crucial step in conducting research in the field of forecasting is the assessment of the quality of the results of the prediction models. To describe with what degree of reliability the created model explains the retrospective of the object under study, the characteristics of the information suitability of the model are used. In this case, the reliability of the model is usually evaluated by comparing real and predicted values. Various indicators are used to assess the quality and degree of reliability of the forecast values obtained as a result of the work of forecasting models. There is a set of purest quality indicators, based on which several forecasts can be compared. An essential feature of these indicators is the fact that they do not depend on the method of forecasting. These indicators include:

1. MAE (Mean Absolute Error); A metric that evaluates the absolute average error between the predicted and real assessment of the quality of translations on a test set. The smaller the error, the better the classifier.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - y_t^*| \qquad (2)$$

where *yt* is the real value at the moment of time t;

y^* is the predicted value at time t obtained as a result of the work of the prediction model;

n is the number of historical observations.

The modules in the formula still allow you to get rid of the signs and get some estimate of the distance from the actual to the calculated values, which will then need to be minimized. The undoubted advantage of MAE is that the modules do not increase by several times the deviations considered as outliers. Therefore, this estimate is more robust than MSE and actually corresponds to the median (in 50% of cases for a given value of x (t), the dependent value y (t) will not be less than the y ^ ty ^ t obtained).

2. Prediction accuracy

The approach to solving the problem as forecasting of time series showed an increase in forecast accuracy. ARIMA models[10], vector autoregression models, as well as the same boosting, were used, but the forecast was made only for future values, not mixed. Forecasts were made for the next few days, iterative estimates for the week and for each day (all years were used for training, and from the beginning of 2017 to October 2017 for forecasting). It is evident that at the beginning of the year, the model is not yet saturated with information and gives predictions that deviate more from real values, and then it

learns, and the error tends to decrease. In absolute terms, this is not the case due to the presence in the predicted data of outliers.

The percentage accuracy of the forecast is calculated using the following formula:

$$forecast\ accuracy = (1 - mape) * 100\% \quad (3)$$

The results of building the model were made in graphical dependence, and for clarity, two indicators were used to predict values and real values, which are presented in Fig.3.

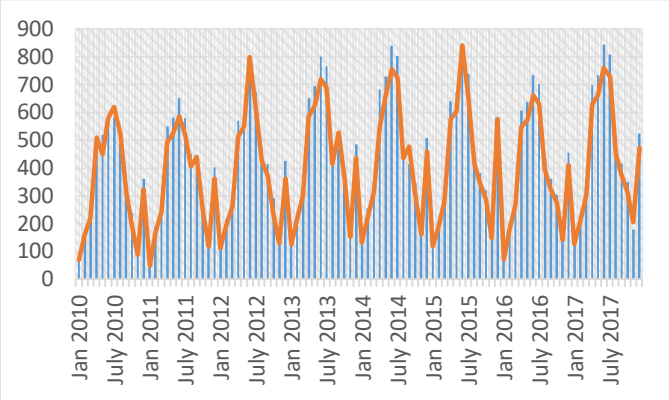

Fig. 3 Results of model prediction compared to real data

- On average, the model gives a forecast deviating from the real value by 45 tours;
- In percentage terms, the accuracy reaches an average of 89%, and in some cases is 94-98%;
- Model errors are typical, the size of errors only in rare instances reaches large quantities.

IV. CONCLUSION

The developed program code in the Anaconda development environment allows to process the source data and to give the prediction results for two indicators of MAE and program accuracy. The program code is designed to improve the performance of the entire system by selecting the correct functions. A further direction of research can be the development of a prototype interface that allows you to predict demand in real time without using other special programs. The interface should be visually and intuitively clear for the operator.

REFERENCES

1. Kotler, P.; Keller, K.L. *Marketing managem*ent,14th ed., Pearson; 812 p., 2011
2. Lopatnikov, L.I. *Economics and Mathematics Dictio*nary: Dictionary of Modern Economics, 5th ed., Moscow: Delo, 520 p., 2003.
3. Cichosz, P. *Data Mining Algorithms*: Explained Using R. Published: John Wiley & Sons, 716 p., 2015
4. Bashynska, I.O. Using the method of expert evaluation in economic calculations,*Actual Problems of Economics*, 7 (169), pp. 408-412, 2015
5. Domingos, P. A few useful things to know about machine learning, *Communications of the ACM*, Vol. 55, No 10, P. 78-87, 2012
6. Bashynska I.,Malanchuk M., Zhuravel O., Olinichenko K., Smart Solutions: Risk Management of Crypto-Assets and Blockchain Technology, *International Journal of Civil Engineering and Technology*, 10(2), pp. 1121–1131, 2019.
7. Athanasopoulos, G.; Hyndman, R. J.; Song, H.; Wu, D. The tourism forecasting competition,*International Journal of Forecasting*, 27(3), 2011
8. SvitlanaBondarenko, VolodymyrLagodienko, Iryna Sedikova and Olga Kalaman, Application of Project Analysis Software in Project Management in the Pre-Investment Phase, *Journal of Mechanical Engineering and Technology*, 9(13), 2018, pp. 676-684
9. Ponomarenko T., Zinchenko O., Khudoliei V., Prokopenko O., Pawliszczy D. Formation of the investment environment in Ukraine in the context of European integration: An example of Poland,*Investment Management and Financial Innovations*,Vol. 15, Issue № 1, 2018,pp. 361-373.
10. Wessa P., ARIMA Forecasting (v1.0.10) in Free Statistics Software (v1.2.1), *Office for Research Development and Education*, URL http://www.wessa.net/rwasp_arimaforecasting.wasp, 2