Forecasting Product Demand using Statistical and Machine Learning Methods:

ARIMA, Neural Network and Support Vector Machine

By

REINALDO MADEIRA RIBEIRO ZEZELA

Department of Engineering and Technology
UNIVERSITY OF DERBY

A dissertation submitted to the University of Derby in accordance with the requirements of the degree of MASTER OF SCIENCE in Big Data Analytics.
2017/18

AUGUST 2018

Abstract

In the academic literature machine learning have been proposed as alternative to statistical approach for time series forecasting. It is due to his popularity gaining awards in many prediction applications. However, it is hard to find in the literature the intersection between time series forecasting and machine learning.

Hence, we decided to explore it applying monthly product demand forecasting from one of the competition problems on Kaggle platform. We used publicly available real-life data to develop and evaluate a model to forecast the product demand.

The aim of the present study is to evaluate if machine learning algorithms outperform the traditional one, in terms of forecasting accuracy. Since, typically the last one is more used for this kind of problems. To this end, we applied three algorithms widely used, Autoregressive Integrated Moving Average ($ARIMA$) for statistical approach and Neural Network (NN) and Support Vector Machine (SVM) for machine learning methods.

The major contribution of this study is first, to prove that machine learning model does not outperform the statistical approach toward time series forecasting. Second, to point the needs to work out the reason of this underperformance and find a way to reverse the situation. Therefore, we are disappointed with these results and we continue believing in the greatest potentiality of machine learning for forecasting applications.

Table of Contents

CHAPTER 1

INTRODUCTION

Demand forecasting is the art-science of predicting consumer demand for goods or services. This is one of the most challenging fields of predictive analytics. Any company that wants to compete in the market needs to understand consumer demand [11].

In supply chain management, demand forecast play a paramount role in that it allows the optimization of operations and stock. This will imply developing optimal strategies for procurement, production planning, stock management and storage cost reduction, and sometimes assessing the current capacity to meet future demand. However, the demand of certain products in future is affected by a several factors which makes complex the effective management of the supply chain.

Accurate predicting models are vital in order to meet the consumer's needs. Although, no forecasting model is flawless. By knowing what things shape demand, we can drive behaviours around our products better. The present study focusing on the process of predicting product demand using machine learning algorithms.

For many years the time series forecasting was dominated by the traditional method, the statistical approach such as exponential smoothing and Autoregressive Integrated Moving Average (ARIMA).

However, in the predictive modelling research in the last two decades, Machine Learning has becoming the most attractive area due the facto of developing algorithms capable to learn from data automatically. It became popular due to high performance of their prediction algorithms in several dominions such as speech recognition, natural language processing, spam detection, and so on. Yet, it is harder to find in the literature the intersection between machine learning and time series forecasting [3].

In the literature, three models are recommended to apply machine learning techniques in

time series problems, namely support vector machines (SVM), artificial neural network (ANN) and decision trees. The first two are due to their highly accuracy mainly in non-linear problems while the third one due to his interpretability [5] [6].

Therefore, the present study focus on SVM and ANN due to the characteristics of the data. We will use publicly available real-life data to develop and evaluate a model to forecast the product demand. The aim is to compare these two approaches machine learning and statistical one in terms of accuracy forecast for time series problems. To this end, we will apply autoregressive integrated moving average ($ARIMA$) model the most widely-used approaches to time series forecasting.

1.1 Motivation

After researching current and past literature, we reached at the conclusion that most of the forecasting problems are applied the traditional model, and it is hard to find a study that involves supervised learning algorithms, particularly in supply chain.

Machine Learning models has been attracting attention in the forecasting community in the last two decades, because of their competitiveness with the classic traditional model [5] [6]. These models are flexible to nonlinear and non-parametric characteristics and they use only historical data to learn and establish the relationship between the past and the future [5] [6].

Nowadays, these models are creating fascinate debates in the data science arena due to their extensive exploration in the forecasting competitions. Therefore, it was based on this spirit that we selected one of the prediction problems on the kaggle competition platform to apply machine learning algorithms and evaluate their accuracy compared to the traditional model.

1.2 Problem statement

The dataset contains historical product demand for a manufacturing company with footprints globally. The company provides thousands of products within dozens of product categories. There are four central warehouses to ship products within the region it is responsible for. Since the products are manufactured in various locations all over the world, it normally takes more than one month to ship products via ocean to different central warehouses. If forecasts for each product in different central with reasonable accuracy for the monthly demand for month after next can be achieved, it would be beneficial to the company in multiple ways. These forecasts can be used by the company as input to optimizes operations, such as plan capacity in the longer term or delivery vehicle route. Available in https://www.kaggle.com/felixzhao/productdemandforecasting/home.

1.3 The purpose

The research focuses on the intersection between machine learning and the classic statistical model: in fact, we intend to explore the Predictive Analytics concept. In this perspective, the aim of this study is to evaluate if machine learning algorithm outperform the traditional one in terms of forecasting accuracy in time series problems. Since, the last one is widely used for this kind of analysis. In affirmative case, we will examine if the difference is statistically significative using ANOVA test.

To achieve this goal, we will apply the literature's most used and recommended algorithms for both models, namely ANN and SVM for Machine Learning and ARIMA for statistical approach. These models will be applied to a dataset that contains the historical demand for products in a manufacturing company as part of a competition problem on the Kaggle platform. Also, their accuracy prediction will be evaluated using more common metrics, that are Mean Squared Error (MSE) and Mean Absolute Percent Error (MAPE).

1.4 Research questions

Q1: To what extent machine learning algorithms can predict accurately time series data better than traditional method?

First, we focus on develop and evaluate prediction models for product demand using supervised learning algorithms (ANN and SVM) and statistical approach ($ARIMA$).

Second, to compare their results in terms of forecasting accuracy as well as to validate if these differences statistically significative using ANOVA test.

BIBLIOGRAPHY

[1] I. ALON, M. QI, AND R. J. SADOWSKI, *Forecasting aggregate retail sales: a comparison of arti"cial neural networks and traditional methods*, Journal of Retailing and Consumer Services, (2001), p. 10.

[2] N. BEHNIA AND F. REZAYAN, *Coupling wavelet transform with time series models to estimate groundwater level*, Arabian Journal of Geosciences, 8 (2015).

[3] S. BEN TAIEB, *Machine learning strategies for multi-step-ahead time series forecasting*, (2014).

[4] J. BERBIĆ, E. OCVIRK, D. CAREVIĆ, AND G. LONČAR, *Application of neural networks and support vector machine for significant wave height prediction*, Oceanologia, 59 (2017), pp. 331–349.

[5] G. BONTEMPI, S. BEN TAIEB, AND Y.-A. LE BORGNE, *Machine Learning Strategies for Time Series Forecasting*, in Lecture Notes in Business Information Processing, vol. 138, Jan. 2013.

[6] ——, *Machine Learning Strategies for Time Series Forecasting*, in Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures, M.-A. Aufaure and E. Zimányi, eds., Lecture Notes in Business Information Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 62–77.

[7] S. CANKURT AND A. SUBASI, *Developing tourism demand forecasting models using machine learning techniques with trend, seasonal, and cyclic components*, COMPUTER ENGINEERING, 3 (2015), p. 8.

[8] R. CARBONNEAU, R. VAHIDOV, AND K. LAFRAMBOISE, *Machine Learning-Based Demand Forecasting in Supply Chains*, IJIIT, 3 (2007), pp. 40–57.

[9] K.-Y. CHEN AND C.-H. WANG, *Support Vector Regression with Genetic Algorithms in Forecasting Tourism Demand*, Tourism Management, 28 (2007), pp. 215–226.

[10] J. D. CRYER AND K.-S. CHAN, *Time Series Analysis: With Applications in R*, Springer, New York, 2nd edition ed., Nov. 2010.

[11] X. DU, S. LEUNG, J. LONG ZHANG, AND K. K. LAI, *Demand forecasting of perishable farm products using support vector machine*, International Journal of Systems Science - IJSySc, 44 (2011), pp. 1–12.

[12] G. GEORGE, E. C. OSINGA, D. LAVIE, AND B. A. SCOTT, *Big Data and Data Science Methods for Management Research*, Academy of Management Journal, 59 (2016), pp. 1493–1507.

[13] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer-Verlag, New York, 2 ed., 2009.

[14] S. S. HAYKIN AND S. S. HAYKIN, *Neural networks and learning machines*, Prentice Hall, New York, 3rd ed ed., 2009.
OCLC: ocn237325326.

[15] R. J. HYNDMAN AND G. ATHANASOPOULOS, *Forecasting: principles and practice*, OTexts.com, Heathmont?, Victoria, print edition ed., 2014.

[16] G. JAMES, D. WITTEN, T. HASTIE, AND R. TIBSHIRANI, *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, Springer-Verlag, New York, 2013.

[17] K. KANDANANOND, *Forecasting Electricity Demand in Thailand with an Artificial Neural Network Approach*, Energies, 4 (2011), pp. 1246–1257.

[18] ——, *A Comparison of Various Forecasting Methods for Autocorrelated Time Series*, International Journal of Engineering Business Management, 4 (2012), p. 4.

[19] N. KHALIL ZADEH, M. M. SEPEHRI, AND H. FARVARESH, *Intelligent Sales Prediction for Pharmaceutical Distribution Companies: A Data Mining Based Approach*, 2014.

[20] M. KHASHEI AND M. BIJARI, *An artificial neural network (p,d,q) model for timeseries forecasting*, Expert Systems with Applications, 37 (2010), pp. 479–489.

[21] B. LANTZ, *Machine learning with R: discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R*, Community experience distilled, Packt Publishing, Birmingham Mumbai, second edition ed., 2015.
OCLC: 947168887.

[22] S. MAKRIDAKIS, M. HIBON, AND C. MOSER, *Accuracy of Forecasting: An Empirical Investigation*, Journal of the Royal Statistical Society. Series A (General), 142 (1979), pp. 97–145.

[23] S. MAKRIDAKIS, E. SPILIOTIS, AND V. ASSIMAKOPOULOS, *The Accuracy of Machine Learning (ML) Forecasting Methods versus Statistical Ones: Extending the Results of the M3-Competition*, Oct. 2017.

[24] S. MAKRIDAKIS, E. SPILIOTIS, AND V. ASSIMAKOPOULOS, *Statistical and Machine Learning forecasting methods: Concerns and ways forward*, PLOS ONE, 13 (2018), p. e0194889.

[25] M. NELSON, T. HILL, W. REMUS, AND M. O'CONNOR, *Can neural networks applied to time series forecasting learn seasonal patterns: An empirical investigation*, in Proceedings of the Hawaii International Conference on System Sciences, vol. 3, Feb. 1994, pp. 649–655.

[26] P. M. ROBINSON AND M. ROSENBLATT, eds., *Athens Conference on Applied Probability and Time Series Analysis: Volume II: Time Series Analysis In Memory of E.J. Hannan*, Lecture Notes in Statistics, Athens Conference on Applied Probability, Springer-Verlag, New York, 1996.

[27] A. K. ROUT, P. K. DASH, R. DASH, AND R. BISOI, *Forecasting financial time series using a low complexity recurrent neural network and evolutionary learning approach*, Journal of King Saud University - Computer and Information Sciences, 29 (2017), pp. 536–552.

[28] P. SAMUI AND B. DIXON, *Application of support vector machine and relevance vector machine to determine evaporative losses in reservoirs*, Hydrological Processes, 26 (2012), pp. 1361–1369.

[29] M. SARHANI AND A. EL AFIA, *Intelligent system based support vector regression for supply chain demand forecasting*, in 2014 2nd World Conference on Complex Systems, WCCS 2014, Nov. 2014.

[30] D. SARKAR, R. BALI, AND T. SHARMA, *Practical Machine Learning with Python*, Apress, Berkeley, CA, 2018.

[31] Z. TANG AND P. A. FISHWICK, *Feedforward Neural Nets as Models for Time Series Forecasting*, ORSA Journal on Computing, 5 (1993), pp. 374–385.

[32] J. W. TAYLOR, *Short-term electricity demand forecasting using double seasonal exponential smoothing*, Journal of the Operational Research Society, 54 (2003), pp. 799–805.

[33] E. YILMAZ, *Forecasting tourist arrivals to Turkey*, 63, p. 12.

[34] G. P. ZHANG, *Neural Networks for Time-Series Forecasting*, in Handbook of Natural Computing, G. Rozenberg, T. Bäck, and J. N. Kok, eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 461–477.

[35] P. ZHANG AND M. QI, *Neural network forecasting for seasonal and trend time series*, European Journal of Operational Research, 160 (2005), pp. 501–514.

[36] ——, *Neural network forecasting for seasonal and trend time series*, European Journal of Operational Research, 160 (2005), pp. 501–514.