ESJ Natural/Life/Medical Sciences

Predicting Customer Behavior Using Prophet Algorithm In A Real Time Series Dataset

Ledion Liço
MS in Data Science, University of Rochester, New York
Indrit Enesi
Department of Electronics and Telecommunication, Polytechnic University of Tirana, Albania
Harshita Jaiswal
MS in Data Science, University of Rochester, New York

Doi:10.19044/esj.2021.v17n25p10

Submitted: 30 May 2021
Accepted: 16 June 2021
Published: 31 July 2021

Cite As:
Liço L., Enesi I. & Jaiswal H. (2021). *Predicting Customer Behavior Using Prophet Algorithm In A Real Time Series Dataset.* European Scientific Journal, ESJ, 17(25), 10. https://doi.org/10.19044/esj.2021.v17n25p10

Abstract

Customer Relationship Management is important in analyzing business performance. Predicting customer buying behavior enables the business to better address their customers and enhance service level and overall profit. This paper focuses on proposing a model that predicts future period sales in a real retail department store with low prediction error rate, and it also discovers the main sales trends over time. A model based on the Prophet algorithm is implemented and modified according to different parameters in order to lower the prediction error. The modifications consisted of the insertion of a new seasonality pattern, changes in the Fourier order of the existing and the new seasonality pattern, inclusion of the holiday data, and parameterizing its impact. The performance of the standard and modified model is evaluated in terms of the MAE (mean absolute error) and MAPE (mean absolute percentage error). The standard and the modified model were tested on a real dataset consisting of the sales between 2011-2019 in a department store of a shopping center in Albania. Implementation results show that the MAE in sales prediction for the modified model is reduced, while the MAPE in sales prediction for the modified model was measured for different prediction periods. The implementation results indicate a comparable or even better performance than the standard model.

Keywords: CRM, Time Series Analysis, Prophet Algorithm, Holiday Data, Accuracy In Prediction, Cross-Validation

Introduction

Choosing the best technique to forecast data is a problem that arises in any forecasting application. Decades of research have resulted into an enormous amount of forecasting methods that stem from statistics, econometrics, and machine learning (ML), which leads to a very difficult and elaborate choice to make in any forecasting exercise (Aditya Satrio et al., 2021). Prediction of sales in department stores is usually very hard as they are impacted by several factors such as seasonality, holiday seasons, stock quality, price changes and many other factors that are usually not easy to predict. This is because there are different types of articles in a department store varying from clothes to accessories, cosmetics, household appliances and food, as well as the respective seasonality changes. However, the way these seasonalities impact the sales is not the same.

In this paper, a Time Series Model is proposed based on a modified version of the Prophet algorithm (Taylor et al., 2017) in order to predict the sales in a department store located in Albania. A new bimonthly seasonality pattern is introduced to the base Prophet model. The Fourier order that governs the seasonality pattern is changed for the new bimonthly seasonality and the weekly one. Holiday Data in Albania are implemented to the model and its impact is parameterized. The prediction results are compared for both the base model and the modified model based on MAE (mean absolute error) and MAPE (mean absolute percentage error). The modified model is fitted on the real data and the MAPE error is evaluated for quarterly sales predictions.

A real time series dataset is considered in the simulations. It includes the daily total sales of the department store from 2011-2019. The dataset is taken from a department store in Tirana, Albania, which is part of an international chain that operates in Europe. The department store's main products are clothes, accessories, cosmetics, and household appliances. The prediction is done for the total sales of the shop, and it is not divided into categories to understand the general trend in customer purchases over time. It is expected that the model fits into all sales datasets taken from other department stores in Albania as seasonalities and holiday data patterns are the same.

Literature review is discussed in the second part of this paper, while time series is analyzed in the third part of the paper. Section four describes the Prophet algorithm and its three core components. Implementation and analysis compound the fifth section. This is followed by the conclusions and future works.

Literature Review

The usage of Machine learning techniques to forecast sales data has been studied intensively in the last years. The main techniques used for this purpose are Neural Networks, Time Series Models, and combined models. Artificial Neural Networks were used to predict sales on the biggest grocery retailers in Turkey (Penpece et al., 2014) and the authors report an accuracy of 90%. The methodology used for measuring the accuracy levels was not clearly explained in the study.

A combined model of ANN and ARIMA Time Series Model was proposed (Li et al., 2018) and used on Chinese E-Commerce Sales. The model was compared to traditional ARIMA and ANN models which produced lower Root Mean Square Error (RMSE).

Prophet is a forecasting model released by Facebook's Core Data Science team (Taylor et al., 2017), which handles common features of business time series by adjusting parameters without prior knowledge of the underlying model details. It consists of a procedure developed for forecasting time series data based on an additive model where nonlinear trends are fit with yearly, weekly, and daily seasonality plus holiday effects. It was used in previous research to predict sales (Žunić et al., 2020) and the results were reported based on mean absolute percentage error (MAPE) level for sales prediction of different type of articles. They achieved a MAPE level of less than 30% for 70% of the products on a quarterly prediction. Default seasonality patterns were used in this study.

Time Series Analysis

A time series is a series of data points ordered in time which is used to predict the future of data. Time series may be stationary, seasonality or auto correlated (Athiyarath et al., 2020). They are represented using different data visualization techniques to uncover the hidden patterns in datasets. Sequential data points are mapped at a certain successive time duration. Various methods tend to understand the underlying concept of the data points in time series to make predictions. Some of them use significant models to forecast the future conclusions on the basis of known past outcomes (Suhermi et al., 2018). Another objective is to explore and understand patterns in changes over time, where these patterns signify the components of time series. When such components reside in a time series, the data model must be considered for these patterns to generate accurate forecasts. Time series models are Autoregressive, Integrated, Moving average or a combination of all. Machine learning is a powerful technique that is available for a huge, explicated dataset

(Suhermi et al., 2018). Nevertheless, problems based on time series do not usually have interpreted datasets. Time series analysis requires some sorting algorithms that can allow it to learn time-dependent patterns across multiple models. Various machine learning tools such as classification, clustering, forecasting, and anomaly detection depend on real-world business applications. Multiple models and methods are used as an approach for time series forecasting. Machine learning methods for time series forecasting are divided into two, namely: Univariate time-series Forecasting method, where one variable is time and the other is the field that requires forecasting; and Multivariate Time-series Forecasting method, where the forecasting problem contains multiple variables that keep time as fixed (Aditya Satrio et al., 2021). Machine learning models are classified in ARIMA model, which is a combination of three different models, i.e., AR, MA, and I (Aditya Satrio et al., 2021). It tries to fit the data into the model. ARIMA also depends on the accuracy over a broad width of time series.

Prophet Model Algorithm

Prophet is a forecasting model which handles common features of business time series by adjusting parameters without prior knowledge of the underlying model details (Taylor et al., 2017). Three model components are implemented, namely: trend, seasonality, and holidays as shown in equation 1:

$$\boldsymbol{y(t) = g(t) + s(t) + h(t) + \epsilon_t} \qquad \textbf{(1)}$$

where g(t) implements the trend function as a non-periodic change of time series values, s(t) implements periodic changes, h(t) implements the holidays effects, and the error term represents idiosyncratic changes as a normally distributed variable (Aditya Satrio et al., 2021). The forecast is considered as a framed curve-fitting problem based on the temporal dependence structure of the data. The system is flexible such that multiple periods of seasonality are easily embedded, measurements do not need to be regularly spaced, fitting is very fast, and users can interactively manipulate many model specifications and parameters or include new components (Sahay & Amudha, 2020).

The Trend Model

Two most used trend model implementations are saturating growth model and piecewise linear model (Sahay & Amudha, 2020). The saturating growth model is described by the basic equation (2):

$$\boldsymbol{g(t) = \frac{C}{1+\exp(-k(t-m))}} \qquad \textbf{(2)}$$

where c is the carrying capacity, k is the growth parameter, and m is the offset. Analyzing the equation 2, it can be highlighted that C and k are not usually constant parameters. The first one is replaced by C(t). For the second

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

parameter, changepoints are incorporated where growth rate is allowed to change. By denoting the change points with s_j where j= 1 … S, a vector of adjustments $\delta \epsilon R^S$ (the rate) at time t is calculated by the base parameter k and the adjustments up to that point as shown in equation (3):

$$\boldsymbol{t = k + a(t)^T \delta} \qquad \textbf{(3)}$$

Where $\boldsymbol{a_j(t) = \{ \begin{matrix} 1, \ if \ t \geq s_j \\ 0, otherwise \end{matrix}}$

The offset parameter m must also be adjusted according to parameter k in order to connect the endpoints of the segments (Žunić et al., 2020). The adjustments at endpoints j are shown in equation (4):

$$\boldsymbol{\gamma_j = \left(s_j - m - \textstyle\sum_{l<j} \gamma_l\right)\left(1 - \frac{k+ \sum_{l<j} \delta_l}{k+ \sum_{l\leq j} \delta_l}\right)} \qquad \textbf{(4)}$$

The saturating growth model is shown in equation (5):

$$\boldsymbol{g(t) = \frac{C(t)}{1+exp\left(-(k+ a(t)^T\delta)(t-(m+ a(t)^T\gamma)\ )\right)}} \qquad \textbf{(5)}$$

In a situation whereby the forecast does not satisfy the saturating growth model, the piecewise constant growth rate is used. Thus, the trend model is shown in equation (6):

$$\boldsymbol{g(t) = (k + a(t)^T\delta)t + (m + a(t)^T\gamma)} \qquad \textbf{(6)}$$

where: $\boldsymbol{\gamma_j = -s_j\delta_j.}$

Seasonality Model

Multi-periods of seasonality often appear at time series business. Seasonality models must be specified as periodic functions regarding t.
To provide flexibility for periodic model effects, Fourier series are used. The seasonal effect is shown in equation (7):

$$\boldsymbol{s(t) = \textstyle\sum_{n=1}^{N}\left(a_n cos\left(\frac{2\pi nt}{P}\right) + b_n sin\left(\frac{2\pi nt}{P}\right)\right)} \qquad \textbf{(7)}$$

Where P is the period of time series. 2N parameters β = [a1, b1, …aN, bN]^T are estimated for the seasonality. The seasonality matrix is constructed for each value of t in past and future data (Li et al., 2020). The seasonality component is shown in equation (8):

$$\mathbf{S(t) = X(t)}\ \boldsymbol{\beta} \qquad \textbf{(8)}$$

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

Holidays

Holidays and events are usually not periodic and provide large predictable shocks to business time series. Since they happen every year, it is important to incorporate them in the forecast (Toharudin et al., 2021). Assuming that the results of holidays effects are independent, the list of holidays is inserted into the system. Also, past and future days of the holiday are considered in the system. A function can be seen if the time *t* during a holiday *i* assigns the parameter k_i as the change point for the forecast. Therefore, a matrix of regressors is created:

$$\boldsymbol{Z(t) = [1(t \in D_1], \dots 1(t \in D_L]} \qquad \textbf{(9)}$$

and h(t) = Z(t) k. Additional parameters are inserted for the surrounding days of the holiday.

Methodology and Performance Metrics

The prophet model has produced high accuracy rate which is compared with existing tools as observed in the previous studies. Its prediction error rate depends on the Fourier order of the model. A modified version of the Prophet model is proposed in this paper for the purpose of sales forecasting in a department store. The sales data is queried and aggregated daily from the transactional database of the department store and used as an input to the model. A new bimonthly seasonality pattern is included to the base Prophet model and the Fourier Series Order described in the previous sections is changed to produce the best fit and lower prediction error rate. The holiday data in Albania is gathered from 2011-2019 and inputted to the model. The impact of the day before the holiday is inputted also as a parameter to the model.

The base model and the modified model are trained and tested on the dataset and their prediction performance is evaluated based on MAE (Medium Average Error). The medium average error is the average magnitude of the errors in a set of predictions. It is calculated as shown below where y is the real data and ȳ is the predicted value:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \bar{y}_j\right|} \qquad \textbf{(10)}$$

The performance of the modified model is cross validated over different periods of time during 2011-2019. It is trained within a period of 2 years and tested for the prediction of the following 6 months. The prediction performance is evaluated based on the mean absolute percentage error (MAPE). MAPE is the mean or average of the absolute percentage errors of forecast, and it is calculated as shown below where y is the real data and ȳ is the predicted value:

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{y_j - \bar{y}_j}{y_j}\right| x\ 100\%} \qquad \textbf{(11)}$$

Implementation and Analysis

Data Exploration and Preprocessing

For the Time Series Analysis task, a dataset of 3271 records of the daily sales in the department store was used. The sales covered the period from 2011-2019, and the model was based on the Prophet algorithm. The data was converted in the format required by the algorithm. Another small dataset of the official holidays during these years was created and was used later on in the model. The aim was to predict the sales for 2020.

Implementation of Holidays in Prophet Model

The Prophet base model and the modified model were tested. In the modified model, the holidays were imported in the model and the changes summarized in Table 3 were also indicated. A new bimonthly seasonality period was created and it was estimated that the department store is better described by this seasonality. Seasonality is estimated using partial Fourier Sum. The Fourier order determines how quickly the seasonality can change (default order for yearly seasonality is 10, while weekly seasonality order is 3). Increasing this Fourier order allows the seasonality to fit faster-changing cycles. Since there are sudden changes in the weekly sales of the data, the weekly Fourier order was changed to 20 and the bimonthly Fourier order was then set to 5. Furthermore, the holiday data were imported and the lower window was set to -1. This gave more weight to the day before the holiday, and it is estimated that the sales usually spike during this period. The holiday_prior_scale was set to 25 (default is 10) to strengthen the effect of holidays.

Table 3. Changes to the Prophet base Model

Base Model	**Modified Model**
Yearly/Weekly Seasonality	Seasonality
No holiday data	1.Annual
Linear Growth	2.Weekly fourier_order = 20
	3.Bi-Monthly fourier_order = 5
	holiday_season, holiday_prior_scale = 25,'lower_window': -1
	Linear Growth

The results of the prediction are plotted in Figure 1 and it can be seen that the results fit into the real data. The blue line represents the prediction and the black dots represent the real data. A prediction for 2020 was also produced.

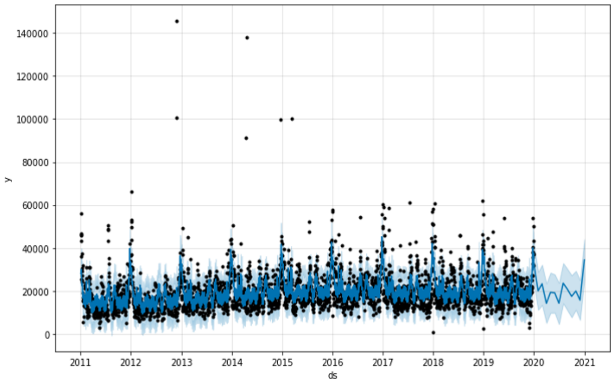

Figure 1. Modified Model Fit

The base model and the modified model are compared in terms of the mean absolute error (MAE). The error is measured based on the real data and predictions for the second half of 2019.

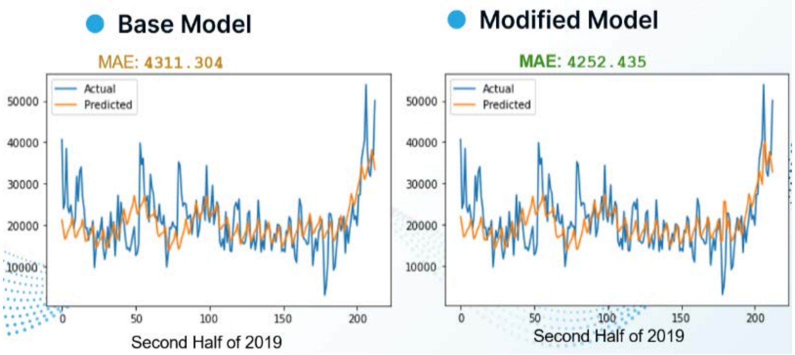

Figure 2. Comparison of model based on MAE

It can be observed that the modified model performs better in prediction and an improvement of 2% was obtained in prediction. The cross validation was also performed on the modified model in order to validate it. The algorithm was trained for 2000 days and a prediction was performed for a horizon of 180 days. In total, 7 forecasts were done and the results of MAPE (mean absolute percentage error) are plotted in Figure 3. From the results, it can be deduced that the model gave a very good prediction for about 90 days (<25%) and then the accuracy began to deteriorate within the interval of 90-140 days (max 40%). After 140 days, the error dropped again. This can be attributed to changes in sales seasons patterns for the period of 2018/2019.

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

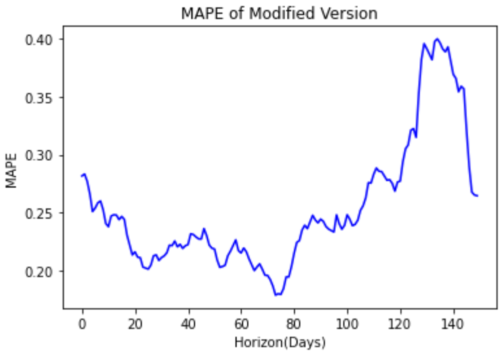

Figure 3. Cross Validation MAPE

In addition, the trends in the sales are shown in Figure 4 since they are valuable information that can be used by the department store.

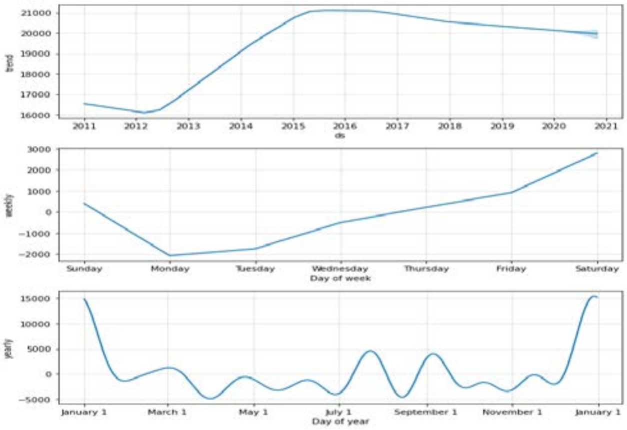

Figure 4. Trends in the Sales Data

Conclusion

Predicting customer buying behavior is very important in business performance. A modified Prophet model was proposed and evaluated. The modifications consisted of the implementation of a new seasonality, changing the Fourier order of the existing and new seasonalities, insertion of holiday data, and parameterizing their impact.

The modified Prophet model was compared with the standard model. Implementation results show a reduction of 2% in the MAE level of the sales prediction from the standard model using the dataset. More so, the predicted curve fits the real sales data points even in the case of sudden changes. The model was cross validated over different periods of time and the sales predictions obtained correlate with the real data, which has a low MAPE under 25% for the quarterly predictions. This result is comparable and even better compared to previous studies discussed in literature review.

The error rise experienced within the interval of 90-140 days is attributed to changes in the sale's promotional seasons. The results obtained were supplied to the sales department for further analysis in order to exploit the detailed reasons behind the trend changes.

Future Work

This study will be carried out in larger datasets and will be expanded in different dimensions (shops, categories, brands) of the data. More experiments will be conducted with the Prophet model and other time series techniques in order to achieve better levels of MAPE.

References:

1. Aditya Satrio, C. B., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. Procedia Computer Science, 179, 524–532. https://doi.org/10.1016/j.procs.2021.01.036
2. Athiyarath, S., Paul, M., & Krishnaswamy, S. (2020). A Comparative Study and Analysis of Time Series Forecasting Techniques. SN Computer Science, 1(3). https://doi.org/10.1007/s42979-020-00180-5
3. Calster, V. T. (2020). Profit-oriented sales forecasting: a comparison of forecasting techniques from a business perspective. ArXiv.Org. https://arxiv.org/abs/2002.00949
4. Li, M., Ji, S., & Liu, G. (2018). Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model. Mathematical Problems in Engineering, 2018, 1–12. https://doi.org/10.1155/2018/6924960
5. Li, Y., Ma, Z., Pan, Z., Liu, N., & You, X. (2020). Prophet model and Gaussian process regression based user traffic prediction in wireless networks. Science China Information Sciences, 63(4). https://doi.org/10.1007/s11432-019-2695-6
6. Nakajima, K., Nakata, T., Matsuo, S., Doi, T., & Jacobson, A. (2019). 239Machine learning model for predicting sudden cardiac death and heart failure death using 123I-metaiodobenzylguanidine. European Heart Journal - Cardiovascular Imaging, 20(Supplement_3). https://doi.org/10.1093/ehjci/jez145.003
7. Penpece, D. & Elma, O. E. (2014). Predicting Sales Revenue by Using Artificial Neural Network in Grocery Retailing Industry: A Case Study in Turkey. International Journal of Trade, Economics and Finance, 5(5), 435–440. https://doi.org/10.7763/ijtef.2014.v5.411
8. Sahay, A. & Amudha, J. (2020). Integration of Prophet Model and Convolution Neural Network on Wikipedia Trend Data. Journal of

Computational and Theoretical Nanoscience, 17(1), 260–266. https://doi.org/10.1166/jctn.2020.8660

9. Suhermi, N., Suhartono, Prastyo, D. D., & Ali, B. (2018). Roll motion prediction using a hybrid deep learning and ARIMA model. Procedia Computer Science, 144, 251–258. https://doi.org/10.1016/j.procs.2018.10.526
10. Taylor, S. J. & Letham, B. (2017). Forecasting at scale. PeerJ Preprints. https://peerj.com/preprints/3190/
11. Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y., & Chen, R. C. (2021). Employing long short-term memory and Facebook prophet model in air temperature forecasting. Communications in Statistics - Simulation and Computation, 1–24. https://doi.org/10.1080/03610918.2020.1854302
12. Žunić, E., Korjenić, K., Delalić, S., & Šubara, Z. (2021). Comparison Analysis of Facebook's Prophet, Amazon's DeepAR+ and CNN-QR Algorithms for Successful Real-World Sales Forecasting. International Journal of Computer Science and Information Technology, 13(2), 67–84. https://doi.org/10.5121/ijcsit.2021.13205
13. Žunić, E., Korjenić, K., Hodžić, K., & Đonko, D. (2020). Application of Facebook's Prophet Algorithm for Successful Sales Forecasting Based on Real-world Data. International Journal of Computer Science and Information Technology, 12(2), 23–36. https://doi.org/10.5121/ijcsit.2020.12203