

Time Series Data Forecasting: Alcohol Sales Prediction Based on Prophet

Yuhan Kuang *

Pickering College (High school), Newmarket, Canada

* Corresponding Author Email: kyhphoebe@gmail.com

Abstract. Sales forecasting plays an important role in a business's operation. It points out the outline of future development and customer targets for companies. Due to the fact that sales forecasting is mostly based on large amount of past data, machine learning has a successful implementation in this field. This paper employed Prophet, a package specialized in dealing time series, to forecast the alcohol sales amount across the USA. The data showed significant trends and holiday effects that fit with the algorithm behind Prophet. By adjusting the parameter functions, the model generated by Prophet demonstrated impressive ability on forecasting the third year's data based on the previous two. A variety of performance indicators, including Mean squared error (MSE), Mean absolute percent error (MAPE), and Median absolute percentage error (MDAP) and so on, were used to evaluate the predicted results. The result of this study implies that machine learning has a huge potential in the field of sales forecasting.

Keywords: Sales forecasting, time series data forecasting method, alcohol, Prophet, USA.

1. Introduction

Sales forecasting, which involves predicting revenue within a certain time period, has always been an important factor of a business' success. Forecasts help a business to attain revenue efficiency by offering insight into the likely behavior of valuable customers. It also enables the business to estimate their costs and revenues accurately and improve pricing and product development further on [1]. Due to the fact that future forecasts are mainly based on economics models and past datasets, machine learning has a significant application in the field. Machines can use algorithms to generalize trends and learn to deal with changes and outliers, returning developers with much more accurate and informative models than human representatives. Machine learning also allows business to create more complicated business models that take maximum factors into consideration with minimum human effort. The model generated by algorithms can be furtherly applied into business processes to promote the making of informed business decisions.

This paper focuses on the forecast of alcohol sales across the USA. The dataset applied is retrieved from the official website of U.S. Department of Agriculture. The series is recorded weekly based on proprietary scanner data from a nationally representative sample of retail food establishments collected by Information Resources, Inc. (IRI) since October 2019 [2]. Note that the dataset did not include data from Alaska, Hawaii and other 10 states because IRI state-level data is not available. The dataset contains sales data of numerous categories of food including beverage, dairy, vegetables etc. In order to narrow down research subject, this report picked alcohol sales as the only target because it has a relatively high sales value and experience more significant fluctuations compared to other food categories. To analyze the dataset, this report performed machine learning with Prophet, an open source forecasting software package specialized in dealing time series released by Facebook. We apply version 1.1.1 of Prophet in this research using the developing language of Python 3.10.8. In order to better process and visualize the data, Pandas 1.5.2 and Matplotlib 3.6.2 are also installed.

2. Method

This paper forecasts alcohol sales across the United States. The first is to preprocess the data so that it can better adapt to Prophet algorithm. Through the prediction of this algorithm, the results are

analyzed and discussed, and the process is shown in Figure 1. Prophet is an algorithm for forecasting time series data. It applies additive model to fit non-linear trends with yearly, weekly, and daily seasonality, plus holiday effects [3]. It is also designed to enable adjustments to parameters without knowing the entire underlying model.

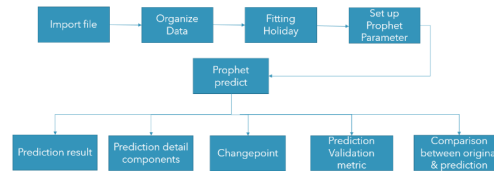


Figure 1. Alcohol prediction process based on Prophet algorithm.

There are three main components for the Prophet model: trend, seasonality and holidays. The components are combined in the following equation with ϵ_t representing the error term.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \quad (1)$$

Two models that cover most circumstances are implemented in the trend function [4]. One kind of market experience saturated growth where a nonlinear growth saturates at a carrying capacity. Such sort of growth is typically fitted with logistic growth models, which are extensions of the following basic form:

$$g(t) = \frac{c}{1 + \exp(-k(t-m))} \quad (2)$$

with C as the carrying capacity, k as the growth rate and m as an offset parameter. However, in real life circumstances, the formula can be much more complicated. For example, the carrying capacity may not remain constant. Instead, it may be represented by a function with respect to time, $C(t)$. Moreover, the growth rate does not stay constant either, so the package applied a much more complicated method to fit with historical data, essentially through defining changepoints in growth rates [5].

The other model for forecasting problems that do not include saturating growth is a piece-wise constant rate growth, which is also used in this report. The trend model is defined by:

$$g(t) = (k + a(t)T\delta)t + (m + a(t)T\gamma) \quad (3)$$

Where k is the growth rate, δ represents the rate adjustments, and m is the offset parameter. An important factor of this model is the selection of changepoints, modified by adjusting the value of `changepoint_prior_scale` parameter. Higher values create models that are more sensitive to change points and vice versa [6-7].

The next component, seasonality, is considered based on the fact that business time series often have multi-period seasonality based on human behaviors. The Prophet seasonality model is a periodic function of t . In order to estimate the parameters in the function, Prophet constructs a matrix of seasonality vectors for each t and assign each element in the matrix a parameter b such that $b \sim \text{Normal}(0, \sigma^2)$ to smoothen seasonality changes.

The last major component of Prophet is the holiday function. It is the determining parameter of the model in this report due to the fact that alcohol sales are significantly impacted by holiday effects. Similar to seasonality, Prophet predicts holidays by adding an indicator function representing whether time t is during the holiday and assign each holiday with a parameter $k \sim N(0, 2)$, with the date of the holiday being the peak of the normal distribution and gradually decrease values as the date gets farther from the holiday [8]. Another additional feature for the holiday function is the lower and upper window. Adding this parameter directs the program to include surrounding days as the holiday itself, enhancing the flexibility of the model.

3. Result

3.1. Settings

The dataset retrieved from the official website is readily formatted in a csv file. However, unnecessary elements need to be removed, and alcohol related data is kept only. Because of the limited

amounts of data, the set needs to be split into training sets and validation sets for machine learning [9-10]. The first two years of data is categorized into original data for training. The third year of data is used to validate the model and ensure if the model is acceptably accurate. This report then uses the data for the first three years for the model to predict trends for the fourth year so that the model has more samples to learn from. Other preprocessing includes converting strings into numeral data and replacing extra characters. This report also converted the variable unit from dollar to million dollars so that the returned graph can be more detailed.

The dataset retrieved from the official website is readily formatted in a csv file. However, unnecessary elements need to be removed, and alcohol related data is kept only. Because of the limited amounts of data, the set needs to be split into training sets and validation sets for machine learning. The first two years of data is categorized into original data for training. The third year of data is used to validate the model and ensure if the model is acceptably accurate. This report then uses the data for the first three years for the model to predict trends for the fourth year so that the model has more samples to learn from. Other preprocessing includes converting strings into numeral data and replacing extra characters. This report also converted the variable unit from dollar to million dollars so that the returned graph can be more detailed.

The holiday function of this program included all federal holidays from October 2019 to September 2023. As alcohol usually has a preholiday effect, this report extended the holiday to the previous week, adding a lower window of -7 to the function.

For the changepoint scale parameter, several values are tested in order to increase accuracy and prevent overfit. We then compare the corresponding performance (measured by several accuracy indicators) and select the value with the best accuracy.

3.2. Performance Analysis

3.2.1. Accuracy

The following Table I-II evaluates the accuracy of the predicting model by the cross-validation conducted on the third-year data. By importing the performance metrics function in Prophet, the program returns several accuracy determinants.

Table 1. Meaning of performance indicator

Abbreviation	MSE	RMESE	MAE	MAPE	MDAPE	SMAPE
Accuracy Determinants	Mean squared error	Root mean squared error	Mean absolute error	Mean absolute percent error	Median absolute percentage error	Symmetric mean absolute percentage error

Table 2. Performance result

MSE	RMESE	MAE	MAPE	MDAPE	SMAPE
9472.1	82.8	68.0	0.05	0.04	0.7

The only variable that we can manually modified to improve accuracy is the changepoint scale. After several trials, 0.15 yields the lowest MAPE value and is recognized as the most suitable value. The resulting graph has four major change points in the general trend. In order to test the accuracy of the model, the cross-validation function is applied to the third-year data, and we also generated performance metrics for direct evaluations of the model (comparison of real data and the predicted data is shown in figure 2).

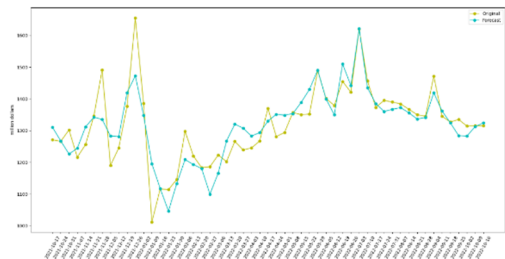


Figure 2. The comparison of real data and the predicted data

The model showed satisfying performance, ending up with a mean MAPE of 0.05. According to related research, a MAPE value less or around 5 % is considered as acceptably accurate for a machine learning model (Table II). Therefore, we can conclude that the prediction of the model trained in this case as trustworthy. The MAPE value graph indicates that at most time the error is under 0.1, but during the second, third and fourth months the model has a large bias in predicting sales values (Figure 3). The reason can be found in the last line diagram for the comparison of real and forecasted values. We notice that the sales amount of alcohol for winter 2021 rocketed up and approached an extremely high value. But, based on the data of the previous two years, the model failed to predict the extreme value that happened even though the time for the peak is correct. Nevertheless, apart from the mentioned error, the model predicts other parts of the data well and most rises and falls are fit perfectly.

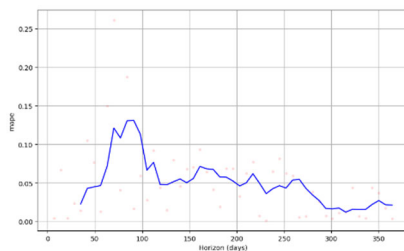


Figure 3. The change of MAPE value with respect to days in a specific year

3.2.2. Result analysis

The returning result is generally satisfying. The figure 4 shows the fitted model for the three years of data and the predicted line and confidence interval (blue area) for the fourth year. The figure 5-7 shown on the side is the key components of the model.

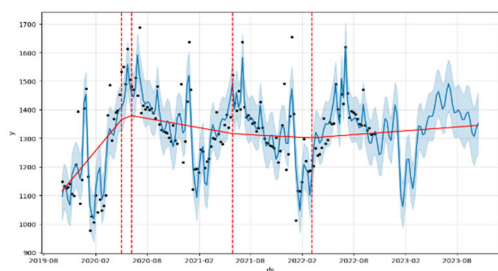


Figure 4. The overall prediction with changepoint,

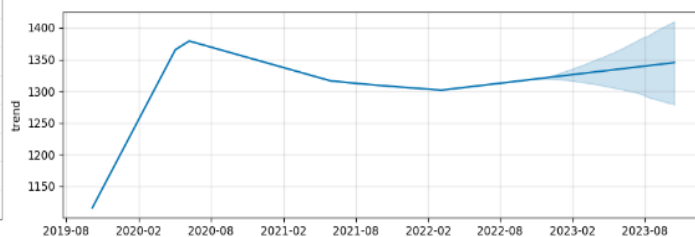


Figure 5. Component graph of trend

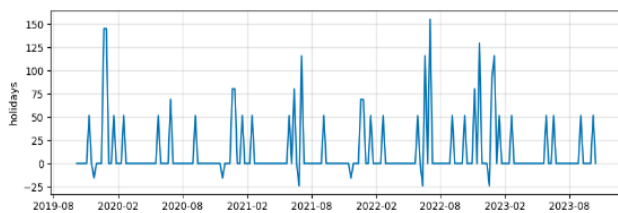


Figure 6. Component graph of holiday effects.

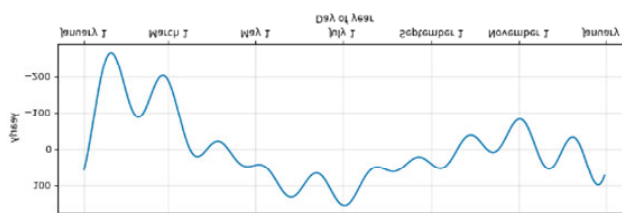


Figure 7. Component graph of seasonality

The figure 5, also presented as the red lines in the main graph, represents the general trend from 2019 to 2023. The mean sales amount of alcohol experienced a sharp increase at the beginning of 2020 and slightly dropped during the next two years. The model predicts a smooth rise in the fourth year with the confidence interval being the blue triangular area shown in the first subplot. The figure 6 demonstrates all the holiday effects added to the model. Every federal holiday sharply increases the sales amount of alcohol across the USA, with a few of them followed by a small downfall. This result is reasonable since celebrating events often happen on holidays, resulting in the rise of demand of alcohol products. The figure 7 represents the trend during a single period, one year, in this case. The line fitted shows that the sales amount of alcohol is relatively lower in the first season and is higher during the second and third season, during which it reaches the peak in July, marking the same period as National Day. We can conclude that consumers tend to buy more alcohol products during spring and summer, possibly due to friendlier weather and longer vacations.

4. Conclusion

This paper focuses on the forecast of alcohol sales across the USA. The prediction of alcoholic sales with Prophet has obtained a good accuracy, but there still presents some limits for this research. The most fundamental problem is the size of the dataset. If a dataset of 5 or 6 years or even longer period can be retrieved, the model will definitely perform even better as it has more data to learn from. The model can moreover be improved by considering more factors that affect the sales, including non-federal holiday events like Superbowl. Meanwhile, due to Covid-19 lockdowns since 2020 and the accompanying economic issues and political issues it brought, the sales amount of alcohol during the last three years may be very different from the pre-covid period. Factors including shutdowns of retail shops and significantly reduced social activities can have a great impact on the data. Though we will probably remain living in the aftermath brought by the pandemic in a few years, the model apparently needs to be adjusted when people completely return to normal life and the economy fixes itself back to normal.

Despite the limitations, the forecast of sales by machines have a far outlook and is anticipated to have more impact on businesses. Apart from alcohol, similar models could be used to predict the sales of other products or services with strong seasonal period and holiday effects. For instance, other beverages, fresh fruit, deliveries, online-shopping etc. The model can be built furtherly into applications to allow easy access to predictions. Businesses and companies can use the model to plan ahead of time and set revenue goals and better allocate resources. With advanced methods, the model can be capable of predicting downturns or sharp changes in the market, companies applying the model can therefore prepare for impacts beforehand. In conclusion, this report only provides a relatively simple experiment for applying machine learning in forecasting. The technology is already showing its potential in assisting companies in sales planning and gaining advantages in the market.

References

- [1] Mahalingam, K. "Importance of Sales Forecasting & Six factors to consider for Accurate Forecasting". Chargebee Compass. <https://www.chargebee.com/blog/importance-of-sales-forecasting/>.
- [2] Loureiro, Ana LD, Vera L. Miguéis, and Lucas FM da Silva. "Exploring the use of deep neural networks for sales forecasting in fashion retail." *Decision Support Systems* 2018, 114 (18), pp. 81 - 93.
- [3] Ma, Shaohui, and Robert Fildes. "Retail sales forecasting with meta-learning." *European Journal of Operational Research* 2021, 288.1, pp. 111 - 128.
- [4] Chen, Zheng, et al. "An overview of Prophet." *Algorithms and Architectures for Parallel Processing: 9th International Conference, ICA3PP 2009, Taipei, Taiwan, June 8-11, 2009. Proceedings 9*. Springer Berlin Heidelberg, 2009, pp. 1 - 8.
- [5] Ning, Yanrui, Hossein Kazemi, and Pejman Tahmasebi. "A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet." *Computers & Geosciences* 2022, 164: 105126.
- [6] Dalrymple, Douglas J. "Sales forecasting practices: Results from a United States survey." *International journal of Forecasting* 1987, 3.3 - 4, pp.379 - 391.
- [7] Lawrence, Michael, Marcus O'Connor, and Bob Edmundson. "A field study of sales forecasting accuracy and processes." *European Journal of Operational Research* 2020, 122.1, pp.151 - 160.
- [8] Sun, Zhan-Li, et al. "Sales forecasting using extreme learning machine with applications in fashion retailing." *Decision Support Systems* 2008, 46.1, pp. 411 - 419.
- [9] USDA ERS. (n.d.). Weekly Retail Food Sales. <https://www.ers.usda.gov/data-products/weekly-retail-food-sales>.
- [10] Taylor, S., J., & Letham, B. Forecasting at Scale. *PeerJ Preprints*. <https://doi.org/10.7287/peerj.preprints.3190v2>.