**Sales forecasting for e-commerce: Application of Hybrid ARIMA and Auto-ARIMA**

Sahar Faraht, Bahçesehir University

**Abstract**

Sales forecasting is the technique of anticipating future revenue by calculating how much merchandise a volume of sales will offer soon. Although tremendous improvement has been achieved in establishing sales forecasting approaches over the last decades, the problem is so broad and multi-dimensional that high accuracy forecasts can only be reached in a few circumstances. In this work, we test a hybrid model that is suitable for modeling linear and non-linear sales trends by combining an ARIMA with a the most predictive clustering model and then testing the Auto-Arima model to forecast sells for near future.

**Keywords**: Sales Forecasting, e-Commerce, ARIMA, XGboost

**Introduction**

Sales predictions aid in the planning of resources to transport items, pay for marketing, hire workers, and so on. Accurate sales forecasting results in a well-oiled machine that fulfills consumer demand now and tomorrow. And, in sales teams, sales money that comes on schedule keeps leaders and collaborators satisfied, much like a package that arrives on time.

A sales forecast is an estimate and evaluation of how to manage future cash flow (regarding how money is going to come in and out). In other words, it is about anticipating when money will enter and exit the firm in the future so that you may identify business opportunities, prevent risks, and foresee problems. Sales forecasting is a fundamental business management topic that is often overlooked by online retailers but may provide competitive benefits when implemented correctly.

Business analytics is critical in optimizing the management of product marketing strategies. Sales forecasting is a prominent analytical method in business analytics. Businesses must conduct sales forecasting to optimize marketing management in the form of product availability predictions, capital sufficiency predictions, customer interest predictions, and product pricing governance (Bakri et al., 2019). However, a common issue in forecasting is the wide range of forecasting methods accessible, finding it challenging for businesses to select the optimal forecasting approach.

As a result, big data analytics algorithms and technologies are widely used to forecast sales of e-commerce commodities. There have been several research conducted in the field of sales forecasting. The sales forecasting methodologies used in this research may be broadly classified as time series models (TSMs) and machine learning algorithms (MLAs) (Ji et al., 2019).

TSMs span from exponential smoothing to the ARIMA families, which have been widely employed to extrapolate future trends based on previous observation data (Box et al., 2015). TSMs have been shown to be beneficial for sales forecasting; nevertheless, their forecasting performance is restricted by their assumption of linear behavior.

MLAs have also played an essential role in predicting. Existing MLAs have been heavily influenced by cutting-edge forecasting techniques such as artificial neural network (ANN), radial basis function (RBF), long short-term memory network (LSTM), support vector regression (SVR) and logistic regression (LR) among others.

First step we gained real data from the Brazilian e-commerce website “Olist”, which was available on Kaggle. Then we respectively used the ARIMA model and tested with Logistic Regression, XGboost and Random Forest clustering model to be able to predict the future values. The second step was to adopt the model with best RMSE to predict the sales of the few upcoming weeks. Third step was to apply the Auto-ARIMA model as an option for goal defining for next weeks.

**Related work**

Based on the XGBoost model, in their paper Ji et al. (2019) proposed a C-A-XGBoost forecasting model that takes commodity sales characteristics and data series trend into consideration. The C-XGBoost model is initially built to predict for each cluster of the generated clusters using a two-step clustering approach, with sales characteristics incorporated into the C-XGBoost model as forecasting influencing variables. Second, they applied an A-XGBoost model using the ARIMA model for the linear component and the XGBoost model for the nonlinear part is employed to anticipate the tendency. The results are computed by adding weights to the C-XGBoost and A-XGBoost predicting findings. When compared against the ARIMA, XGBoost, C-XGBoost, and A-XGBoost models using data from the Jollychic cross-border e-commerce platform, the C-A-XGBoost outperforms the other four models.

In their paper Shih and Lin, (2019) developed a model for forecasting short-term products demand in the context of E-commerce. They trained LSTM model to estimate future value based on a time series of sales and a sentiment grading of comments. The findings showed that the suggested LSTM technique performed well in terms of sales forecasting for items with short-term demand.

Another study is of Vavliakis et al. (2021), were they presented a new hybrid model for predicting linear and nonlinear sales trends by merging an ARIMA model with an LSTM neural network. They used a dataset from a real-world e-commerce site and compared the proposed solution to three competing models and showed that their method beats all three rival models which are hybrid ARIMA-ARNN, hybrid ARIMA-XGBoost, and hybrid ARIMA-SVM.

**Methodology**

This section describes various prediction models and their composite models designed by hybrid and decomposition technique. contains basics of forecasting and various machine learning models.

Since linear models are incapable of effectively capturing nonlinear patterns, their residue (which incorporates nonlinear patterns) is projected by a nonlinear model to enhance prediction results. To generate the consequent prediction of time series, this forecast result is added to the forecast acquired by the linear model. As a result, for improved outcomes, the hybrid approach considers both a linear and a nonlinear model.

***A. Forecasting Model***

ARIMA is an advanced econometric modeling approach is autoregressive integrated moving average. ARIMA examines historical time-series data and uses back-fitting optimization methods to account for historical autocorrelation (the connection of one value to another in time) as well as data stability to adjust for nonstationary features. By correcting its predicting errors, our predictive algorithm learns over time. To create excellent prediction models using this method, advanced econometrics knowledge is usually necessary.

The general steps to implement an ARIMA model are –

1. *Load the data:* The first step for model building is of course to load the dataset
2. *Preprocessing:* Depending on the dataset, the steps of preprocessing will be defined. This will include creating timestamps, converting the dtype of date/time column, making the series univariate, etc.
3. *Make series stationary:* To satisfy the assumption, it is necessary to make the series stationary. This would include checking the stationarity of the series and performing required transformations
4. *Determine d value:* For making the series stationary, the number of times the difference operation was performed will be taken as the d value
5. *Create ACF and PACF plots*: This is the most important step in ARIMA implementation. ACF PACF plots are used to determine the input parameters for our ARIMA model
6. *Determine the p and q values*: Read the values of p and q from the plots in the previous step
7. *Fit ARIMA model:* Using the processed data and parameter values we calculated from the previous steps, fit the ARIMA model
8. *Predict values on validation set:* Predict the future values
9. *Calculate RMSE:* To check the performance of the model, check the RMSE value using the predictions and actual values on the validation set

***B. Predictive models***

1.Logistic regression model is a Supervised machine learning algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable, and the outcome is binary or dichotomous in nature. That means Logistic regression is usually used for Binary classification problems (Field, 2009). It has the following function:

g(E(y)) = α + βx1 + γx2

Here, g() is the link function, E(y) is the expectation of target variable and α + βx1 + γx2 is the linear predictor ( α,β,γ to be predicted). The role of link function is to ‘link’ the expectation of y to linear predictor.

2. Xgboost is an abbreviation for eXtreme Gradient Boosting. It is built on the foundation of gradient boosting. Gradient boosting is a machine learning approach that may be used to solve classification, regression, and ranking issues. To control overfitting, Xgboost features regularized model formalization, which improves its performance. It produces decent results for most datasets involving linearity and nonlinearity. It's efficient since it allows for parallel computing on a single computer (Gurnani et al., 2017). The Xgboost main function is:

Here the is the real value (label) known from the training dataset and the can be seen as f(x+Δx) where .

3. Random Forest: A random forest is made up of many separate decision trees that work together to form an ensemble. Each tree in the random forest produces a class prediction, and the class performing the best becomes the prediction of the mode (Madani, and Alshraideh, 2021). When using a library the random forest uses for each decision tree, Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes (binary tree):

ni\_j=w\_j C\_j-w\_(left(j)) C\_(left(j))-w\_(right⁡(j)) C\_(right(j))

Here ni sub(j) is the importance of node j, w sub(j) is the weighted number of samples reaching node j, C sub(j) is the the impurity value of node j, left(j) is the child node from left split on node j and right(j) is the child node from right split on node j.

*Predictive Evaluation Method.*

We use root mean square error (RMSE) to evaluate the accuracy of fitting and prediction of each model.

***C. Hybrid model***

Time series can be represented as Yt = Lt + Nt where Yt is time series, Lt is Linear component of time series and Nt is its nonlinear component. The Working of hybrid model can be described below:

1. Initially, linear model like ARIMA is applied to time series(Y) which yields forecast. This forecast is considered as linear forecast because ARIMA captures Linear component very well as compared to nonlinear patterns. It is represented as F¹Lt+hº
2. Since, ARIMA doesn’t capture nonlinear patterns properly, hence its residue contains nonlinear component which is given by Et = Yt F¹Yt º where Et is residue, Yt is Actual Values of Training Set and F¹Yt º is Forecast Values of Training Set
3. Residue obtained by ARIMA is applied to nonlinear models like logistic regression, XGBoost and random forest to obtain forecast of nonlinear patterns missed by ARIMA. This forecast is represented as F¹Nt+hº
4. At the end, resultant forecast is obtained by adding forecast of both linear model and a nonlinear model. It is given below F¹Yt+hº = F¹Lt+hº + F¹Nt+hº where F¹Yt+hº is resultant forecast of time series in Test set, F¹Lt+hº is forecast of linear component of test set and F¹Nt+hº is forecast of nonlinear component of test set.

***D. Auto-ARIMA***

The Auto-ARIMA module automates portions of the standard ARIMA modeling by evaluating many permutations of model requirements and delivering the model with the best fit. The Auto-ARIMA module works in the same way as ordinary ARIMA forecasts do. The P, D, and Q inputs are no longer necessary, and alternative combinations of these inputs are automatically performed and compared.

A proper model in Time Series Analysis should have the highest Log-Likelihood and the lowest AIC. A model's AIC and BIC are dependent on log-likelihood; to be more specific, AIC and BIC employ Log-Likelihood in their formula. This would suggest that the log-likelihood is large, thus this technique cycles through the model with different ordering and parameters and returns the one with the lowest by adopting the following formulas:

Formula of Log-Likelihood

Formula of AIC

Formula of BIC

**Case study**

We applied the previously described method and algorithms to the Brazilian e-commerce website “Olist”. The dataset was obtained from Kaggle. It has information of 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. It is an anonymous real commercial data which a perfect case that we can use for our sales forecasting goal.

***Data preprocessing***

To get the data ready for the ARIMA we first executed some preparation and cleaning to make it adaptable for the forecasting. First, the data was divided into small csv files each is about a particular information such as product, seller, orders and so on, therefore, we had to create a master dataframe with all the necessary features. Second, we must consider that the data were normalized, with relationships between them. So, a single purchase has a unique 'order\_id' but may appear multiple times in the dataset as it may have more than one product, more than one payment method, more than one seller, etc. It is not a Primary Key. Thus, we must calculate the averages always considering the value of unique sales, not the frequency. Third, we checked for missing values, and we found that there is 0.68% or the “price” feature missing. We used the KNNimputation method to fill the messing data to be able to use the price as our main feature for prediction. The fourth step was to change the date feature type to datetime.

***Hybrid model application***

We started decomposing the time series which entails seeing it as a collection of level, trend, seasonality, and noise components. Decomposition is a useful abstract paradigm for thinking about time series in general, as well as for better comprehending challenges in time series analysis and forecasting.

To be able to apply the ARIMA we had to check if the time series is stationary. To explain more, A stationary time series' observations are not time dependent. If a time series has no trend or seasonal impacts, it is said to be stationary. Summary statistics based on time series, such as the mean or variance of the data, are consistent throughout time. It is easier to model a time series when it is stationary. To be effective, statistical modeling approaches presuppose or demand that the time series remain stationary. In this case we used the Augmented Dickey-Fuller (ADF) test to check for stationarity. A Dickey-Fuller test is a unit root test that examines the null hypothesis that =1 in the model equation below. The coefficient of the first lag on Y is denoted by alpha. Fundamentally, it has a similar null hypothesis as the unit root test. That is, the coefficient of Y(t-1) is 1, implying the presence of a unit root. If not rejected, the series is taken to be non-stationary.

The t-test value= 0.12126 > Critical Value and p-value > 0.05 – accepted the null hypothesis i.e., time series have a unit root, meaning it is not stationary. It has a time-dependent structure. Therefore, we needed to make dome modification to make the data stationary in order to apply the ARIMA. The figure 1 represents the order purchase distribution by time stamp.

***Figure 1: Order purchase timestamp***

Chart, line chart

Description automatically generated

The next step was to get a monthly distribution of the time data to understand the seasonal trends, the results are represented in figure 2.

***Figure 2: order purchase timestamp: yearly distribution***

Chart, waterfall chart

Description automatically generated

By decomposing the data, we noticed a strong downward trend in values according to seasonality, at the end of each year, in December, probably after an up caused by the Black Friday. Then we investigated the autocorrelation factor to understand better what parameters we should use. We then noticed a strong downward trend in values according to seasonality, at the end of each year.

***Figure 3: Autocorrelation Figure 4: Partial autocorrelation***

Chart

Description automatically generatedChart, histogram

Description automatically generated

As observed on Partial Autocorrelation plot, a shift of 1 will be sufficient for a regression model. So, we will create this new feature to apply a regression model to forecast next week's price.

***Regression models***

As explained in the methodology section, we tested three different regression model to see which one is more accurate to predict the price with association with the ARIMA time series. By examining the RMSE metric we find that the best model is the Xgboost regressor. The results are represented in table 1.

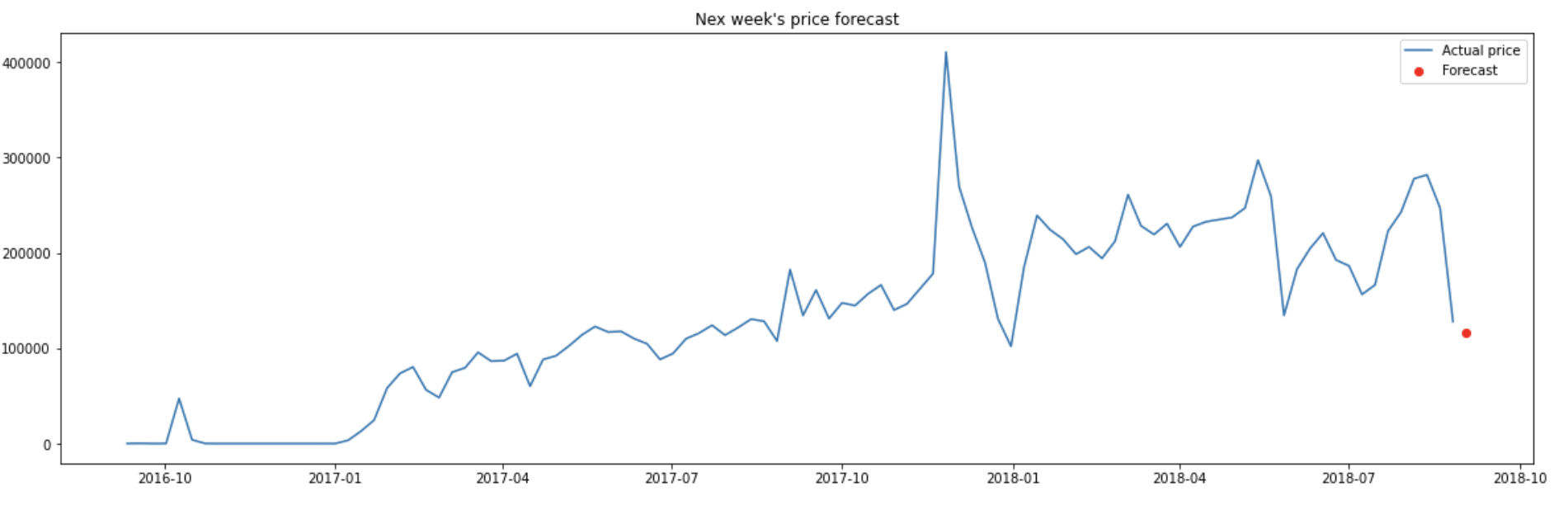
***Table 1: regression models RMSE results***

|  |  |
| --- | --- |
| Regressor | RMSE |
| Logistic regression | 38701.39 |
| Xgboost | 7057.48 |
| Random Forest | 16356.95 |

We tested if the predicted data using the XGB fits the price data plotting and we found out that the result are accurate. Next step we split the data into train and test data and then check if the model XGB Regressor fits on unseen data (test set). The results of the plotting showed that they are well fitted, and the model proved that it can predict the next weeks sales. The results are represented in figure 5.

***The auto-ARIMA model forecasting***

For the auto-ARIMA we started by splitting the data into test and training and we dropped the new feature that we have created to be able to apply the regression models. From the SARIMAX report we concluded that the term is statically significant with a p value= 0.051. All the results implied that the auto-ARIMA is a well fitted model that can forecast the sales accurately.



***Figure 5: next week’s prediction with XGB***

After, checking all the metrics of the auto-ARIMA we generated a prediction dataframe with exact sales prediction values for the next few weeks and plotted the results of the trained, tested and predicted data. The figure 6 shows the final plot of the model.

Chart, line chart

Description automatically generated

***Figure 6: Next week’s prediction with auto-ARIMA***

**Conclusion**

People use Artificial Intelligence to support themselves solving the tasks or problems in the future and increase the opportunities to make the personalized service real. In this study, we tested the ARIMA model to forecast the sales of an e-commerce platform in the short term. The additional value of this study is to test different regression models accuracy and adaptability to used in a hybrid system with the ARIMA. Our results aligned with Ji et al. (2019) who also used the XGboost as a regressor for the ARIMA. After we tested the Auto-ARIMA prediction model and we were able to see that it exceeded the previous model by giving accurate prediction for not only one week further but several weeks a head which makes it the best model to forecast the sales.

**References**

Bakri, R., Data, U., & Astuti, N.P. (2019). Aplikasi Auto Sales Forecasting Berbasis Computational Intelligence Website untuk Mengoptimalisasi Manajemen Strategi Pemasaran Produk. Jurnal Sistem Informasi Bisnis, 9, 244-251.

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019, December). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. In *International conference on neural information processing* (pp. 462-474). Springer, Cham.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Field, A. (2009). Logistic regression. Discovering statistics using SPSS, 264, 315.

Gurnani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., & Bhirud, S. (2017, February). Forecasting of sales by using fusion of machine learning techniques. In 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) (pp. 93-101). IEEE.

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage XGBoost-based model to sales forecasting of a cross-border e-commerce enterprise. *Mathematical Problems in Engineering*, *2019*.

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*.

Madani, B., & Alshraideh, H. (2021). Predicting Consumer Purchasing Decision in The Online Food Delivery Industry.

Riyadi, N., Mulki, M. F., & Susanto, R. (2019). Analysis of Customers Purchase Patterns of E-Commerce Transactions Using Apriori Algorithm and Sales Forecasting Analysis With Weighted Moving Average (WMA) Method. *vol. VII, no*, *7*, 45-58.

Shih, Y. S., & Lin, M. H. (2019). A LSTM approach for sales forecasting of goods with short-term demands in E-commerce. In *Asian Conference on Intelligent Information and Database Systems* (pp. 244-256). Springer, Cham.

Vavliakis, K. N., Siailis, A., & Symeonidis, A. L. (2021). Optimizing Sales Forecasting in e-Commerce with ARIMA and LSTM Models. 17th International Conference on Web Information Systems and Technologies

Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, *2021*.