Order demand forecasting

BRIEFLY INTRODUCTION

Demand and supply are two fundamental concepts of sellers and customers, predicting demand accurately is critical for organizations in order to be able to make plans. In this paperwork, I propose a new approach for demand prediction on an E commerce website which means multiple warehouses for a variety of products. There are multiple ways of demand forecasting, with linear models being the most popular choices tested for data that exhibits a linear trend and or when the residuals have a normal distribution. The working data-set is mostly skewed and lacks a normal distribution thus one needs to look at alternate models for building accurate forecasts. In other theoretical words, demand forecasting is the concept of predicting the quantity of a product that consumers will purchase during a specific time period. Predicting the right demand for a product is an important sight in terms of space, time, and money for the sellers/providers. Most sellers may have limited time and they need to sell their products as soon as possible due to the money and product storage restrictions. Therefore the demand for a product depends on many facts such as price, popularity, time, space occupies in the deposit. Forecasting demand is being hard when the number of factors increases. Demand prediction is also closely related to seller revenue, if sellers store much more products than the demand then this may lead to surplus. On the other hand storing less product in order to save warehouse costs when the product has high demand, in most cases, will cause less revenue. Because of these and plus many more reasons, demand forecasting has become an interesting and important topic for researchers in distinct areas such as water demand prediction, data center app, energy demand prediction etc. Demand forecasting can be grouped into three categories:

- hybrid methods;
- AI methods;
- statistical methods.

Let's talk about the last category first: statistical methods. Here we meet linear regression, moving average/estimated moving average (MA/E MA), regression tree, Bayesian analysis, weighted average, which are just some of the statistical methods used for demand forecasting. For simplicity and interpret-ability, regression trees are preferred they also have a weak point for often relatively inaccurate and sometimes unstable.
I applied demand prediction on data-set found on kaggle.com which contain encoded data from privacy reasons made available only for research, learning and analysis purposes. The time-series data-set consists of several fields (Product_Code, Warehouse, Product_Category, Date (when the customer needs the product), Order_Demand (single order quantity). Each field has multiple items, such as warehouse a, b, c; product 1, 2, 3; category 1, 2, 3; dates between 2012 and 2016. Our second category, AI methods are commonly found in the literature for demand forecasting due to their primary advantage of being efficient and accurate used Artificial Neural Networks to predict sales of women's clothes and ANN clearly outperformed two statistical-based models [1]. In a performance comparison of several prediction methods that use artificial intelligence, the development of a hydro-logical forecasting model, based on past records is critical to productive hydro-power tank management and scheduling. Primarily, time series analysis and modeling are used in the building of mathematical models to generate hydro-logic records including water resources. Artificial intelligence or simply AI, as a fork of computer science, is capable of analyzing

long series and large-scale hydro-logical data. For a couple of years, is one of the front issues to apply this technology concept to hydro-logical forecasting modeling [2]. ARMA (auto-regressive moving average) models, ANNs (artificial neural networks), ANFIS (adaptive neural-based fuzzy inference system) techniques, GP (genetic programming) models, and SVM (support vector machine) method are tested using the long term observations of monthly river flow dispense. Four quantitative typically statistical performance- evaluation measures namely: RMSE (root mean squared error), MAPE (mean absolute percentage error), R (coefficient of correlation), E (efficiency coefficient), are waged to evaluate the performances of various models [2]. The hybrid methods is another approach to forecasting sales or demands. These methods exploit more than one method and use the robustness of these methods. Zhang utilized ARIMA and ANNs hybrid methodology in time series forecasting and suggest a method that can reach more accuracy than the methods when they were used separately [3].

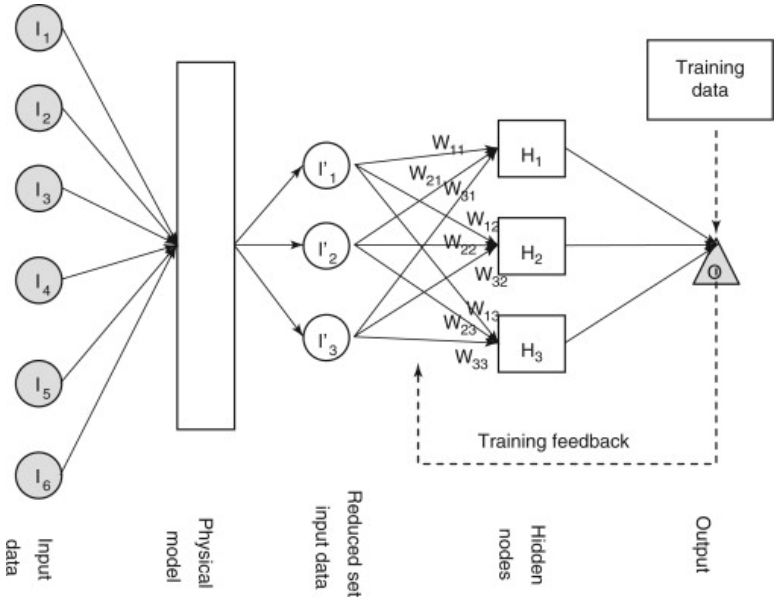


Figure 1: Artificial Neural Networks Model Overview [5]

FACTORS AFFECTING DEMAND FORECAST ACCURACY

In principle, there are two characteristics of forecasting expected degree of accuracy and demand. In retail industry, the foresee demand can be a function of different kinds of structural variations such as seasonality or trend. The proper demand patterns of customers can be affected by price fluctuations and outside weather. The demand peaks be due to promotions and holidays. The elements which affects the expected demand can be classified in this way [4]:

- Trend and periodic demand patterns (seasonal)
- Price fluctuations and discounts (Changes in market may lead to price fluctuations which in turn affect the customer patterns. The price discounts may be planned or unplanned, encourage the customers to purchase more, and make the demand more volatile. If the price discounts are not planned well, they lead to unnecessary warehousing.)
- Holidays (Changes in demand because of holidays and festivals depend on the location, and cultural habits of customers. During the festival season, the retail stores located nearby touristic places and country boundaries may have higher demand due to tourist visitors.)

- Weather (Extreme weather conditions such as rainfall, snowfall, very hot or cold temperatures disturbs the customer purchasing behavior. If it keeps the customers in-home or forces them to visit their nearby stores.)

From the above-mentioned bullet-points, the price discounts of a product and/or other coupled products and changes in weather patterns may produce a short-term shocks in variation in seasonality with or without leading effect. However, in distribution changes and introduction of new products may lead to long-term fluctuations (shift in existing products demand). The accuracy (of forecasting) is eventually limited by the nature of the time series being forecast, it depends on:

- Data availability (The availability of complete and longer historical data is a major key to identifying and understand the external factors which affect the sales.)

- Data Quality (Model highly depends on the quality of input data (training, validation and testing data.)

- Forecast Horizon (Long forecast horizon may increase variability leading to high inaccuracy in the forecast.)

Generally, the sales forecast imprecision results in two types of problems: under-stocking and over-stocking. The under-stocking lead to stock-outs along with lower customer confidence and fall of the market image which is difficult to quantify. The over-stocking leads to insufficient space. Many factors like short shelf-life and bad product quality usually intensify the amount of waste [4].

SARIMAX MODEL

Acronym for Seasonal Auto Regressive Integrated Moving Average with eXogenous regressors derived from non-seasonal ARIMA (p,d,q) model represents a time-series with p: auto-regressive terms, q: moving-average terms and d: non-seasonal differences:

$$\phi_p(B)(1-B)^d Z_t = c + \theta_q(B)\varepsilon_t \quad (1)$$

where, B - Delay or lag operator, time series observation lag k period is symbolized $B^k X_t = X_{t-k}$

$\phi_p(B)$ - Autoregressive operator of p-order

$(1-\phi_1(B)-\phi_2(B^2)-\ldots-\phi_p(B^p))$

$\theta_q(B)$ - Moving average operator of q-order $(1-\theta_1(B)-\theta_2(B^2)-\ldots-\theta_q(B^q))$

$(1-B)^d$ - Differencing operator of order d to remove non-seasonal stationarity

Z_t - Sales of a product at time t

ε_t - Residual error in SARIMA model

c - Constant [4]

The SARIMA model can be represented as:

$$\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D Z_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (2)$$

where, $\Phi_p(B)$ - Seasonal autoregressive operator with p-order

$\Theta_q(B)$ - Seasonal moving average operator with q-order

$(1-B)^D$ - Seasonal differencing operator of order D

$(1-B)^d$ - Differencing operator of order d [4]

S - Seasonal length (e.g. in quarterly data s=4 and in monthly data s=12)

The unique advantage of the SARIMA approach is its capability to handle stationary and non-stationary time-series with seasonal elements. The generation of time-series forecasts using SARIMA is better if no outlying data occur. Based on the behavior of the time-series, outliers could have a potential impact on the estimates of the model parameters. The outlying data in a time-series may often point out important events or exceptions and provide useful information for management [4]. The SARIMAX model is a SARIMA model with external variables, called SARIMAX (p,d,q) $(P,D,Q)_S$ (X), where X is the vector of external variables. The external variables can be modeled by multi-linear regression equation, expressed as:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \omega_t \quad (3)$$

where, $X_{1t}\, X_{2t}\, X_{kt}$ are observations of k number of external variables corresponding to dependent variable Y_t; $\beta_0, \beta 1, \ldots, \beta_k$ are regression coefficients of the external variables; ω_t is a stochastic residual, The residual series ω_t can be represented in the form of ARIMA model as follows:

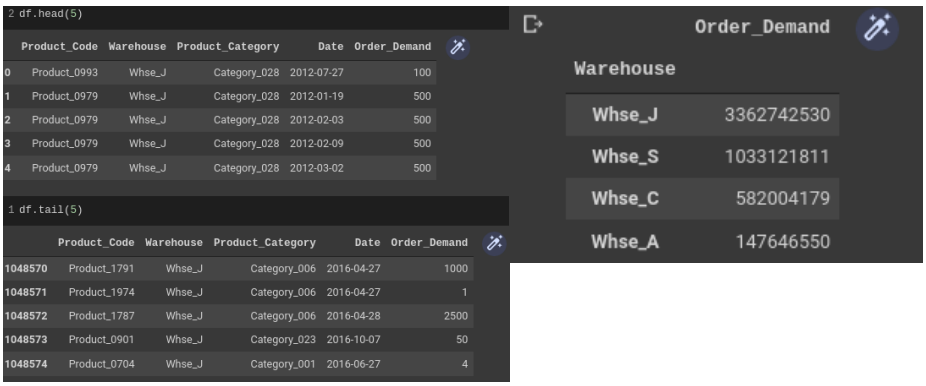
$$\omega_t = \frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t \quad (4)$$

The general SARIMAX model equation can be obtained by substituting Equation 4 in Equation 3, in this case, regression coefficient is interpreted in an usual and easier way [4]:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_k X_{k,t} + \left(\frac{\theta_q(B)\Theta_Q(B^S)}{\phi_p(B)\Phi_P(B^S)(1-B)^d(1-B^S)^D}\varepsilon_t\right) \quad (5)$$

DATA COLLECTION VIEW AND ANALYSIS

In this section I will present a general peek in the working encoded dataset, extracted from a database, the entire dataframe is shown in figures below and the univariate analysis for amounts of orders shipped by each warehouse in 01/01/2012 – 31/12/2016 period and bivariate for each product category (from 1 to 33), category 19 having the most orders.



```
2 df.head(5)
```

	Product_Code	Warehouse	Product_Category	Date	Order_Demand
0	Product_0993	Whse_J	Category_028	2012-07-27	100
1	Product_0979	Whse_J	Category_028	2012-01-19	500
2	Product_0979	Whse_J	Category_028	2012-02-03	500
3	Product_0979	Whse_J	Category_028	2012-02-09	500
4	Product_0979	Whse_J	Category_028	2012-03-02	500

```
1 df.tail(5)
```

	Product_Code	Warehouse	Product_Category	Date	Order_Demand
1048570	Product_1791	Whse_J	Category_006	2016-04-27	1000
1048571	Product_1974	Whse_J	Category_006	2016-04-27	1
1048572	Product_1787	Whse_J	Category_006	2016-04-28	2500
1048573	Product_0901	Whse_J	Category_023	2016-10-07	50
1048574	Product_0704	Whse_J	Category_001	2016-06-27	4

Warehouse	Order_Demand
Whse_J	3362742530
Whse_S	1033121811
Whse_C	582004179
Whse_A	147646550

Figure 2: Head and tail of dataframe lookup(left), orders shipped amount(right)

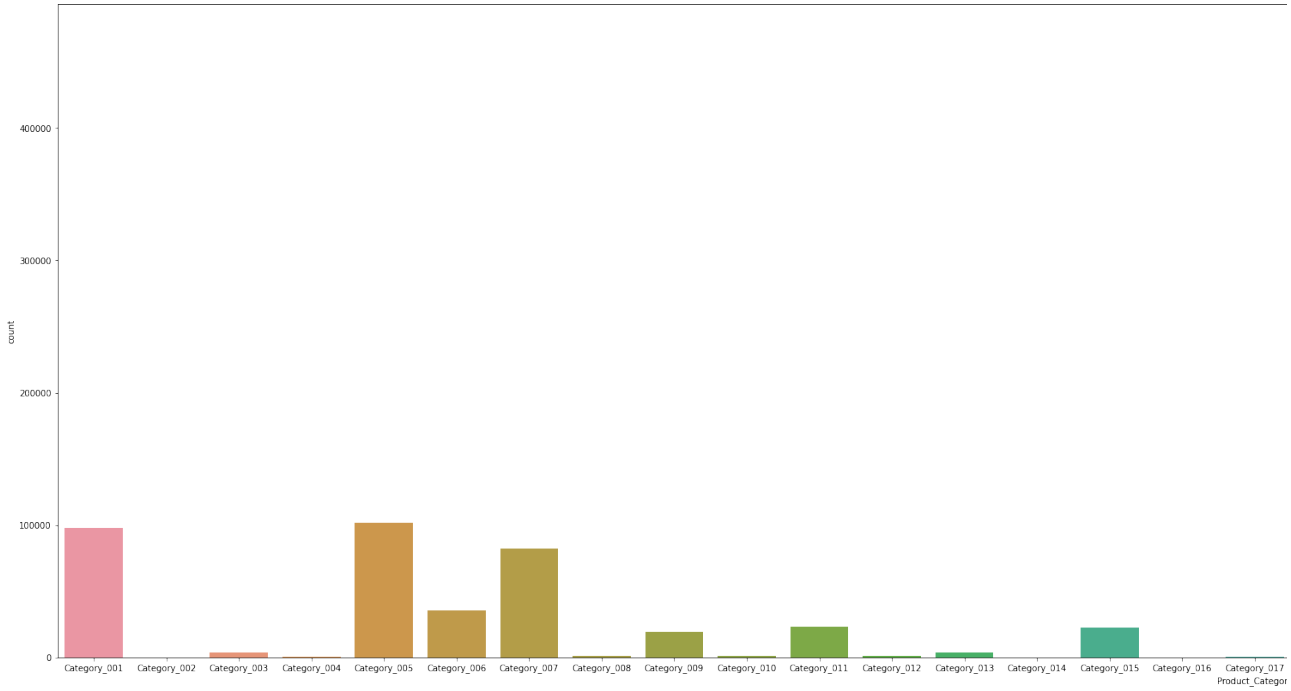


Figure 3: Bivariate Analisys - Product Category with target variable (1-17)

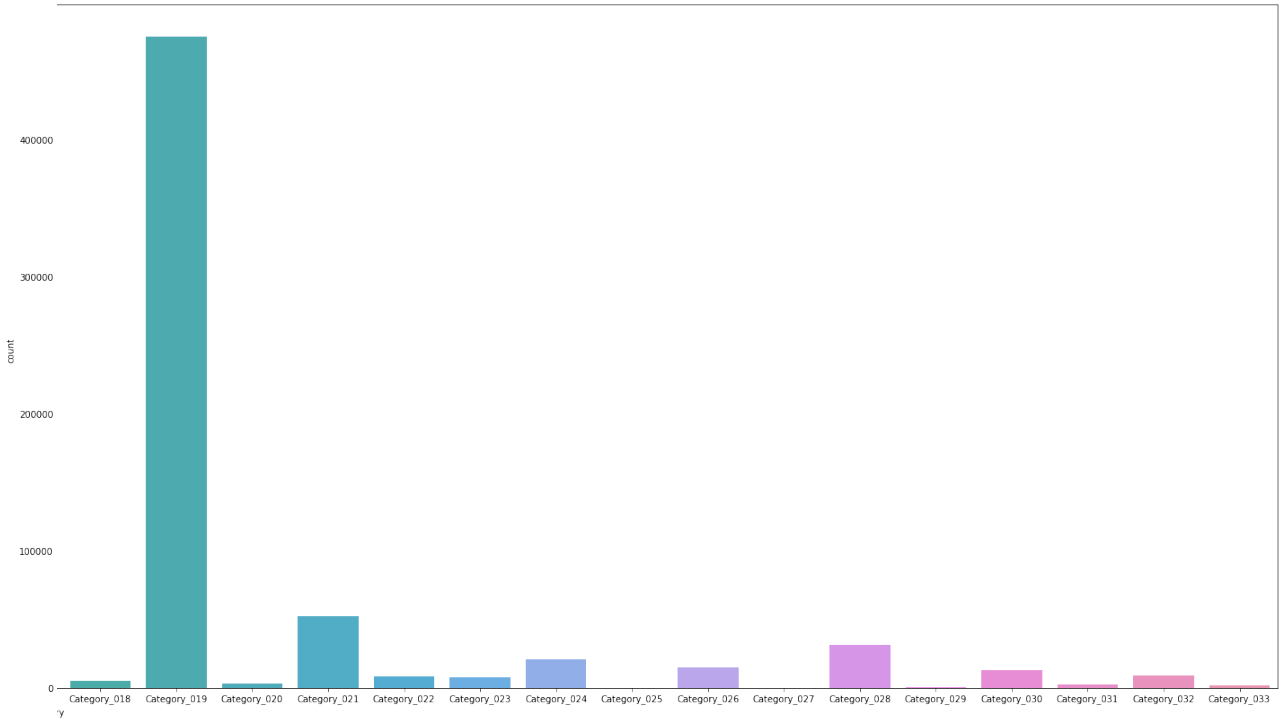


Figure 4: Bivariate Analisys - Product Category with target variable (18-33)

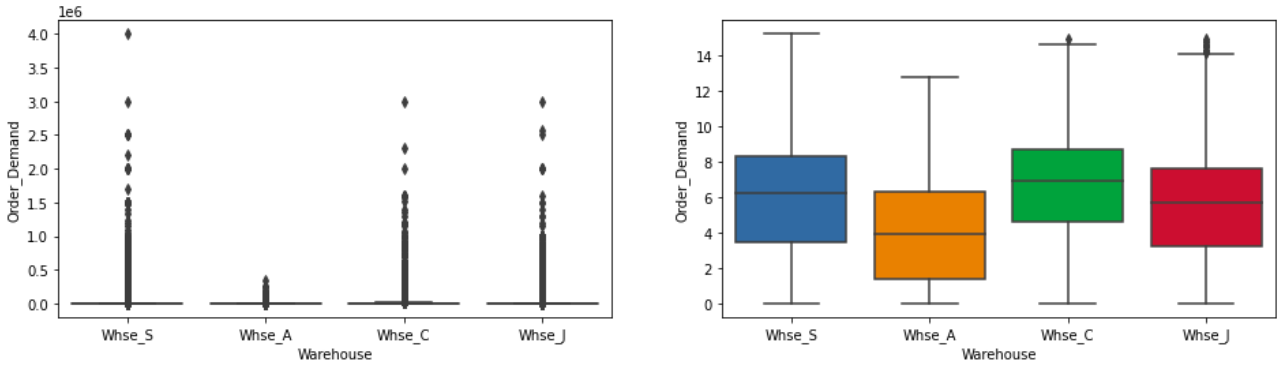


Figure 5: Bivariate analysis - Warehouses

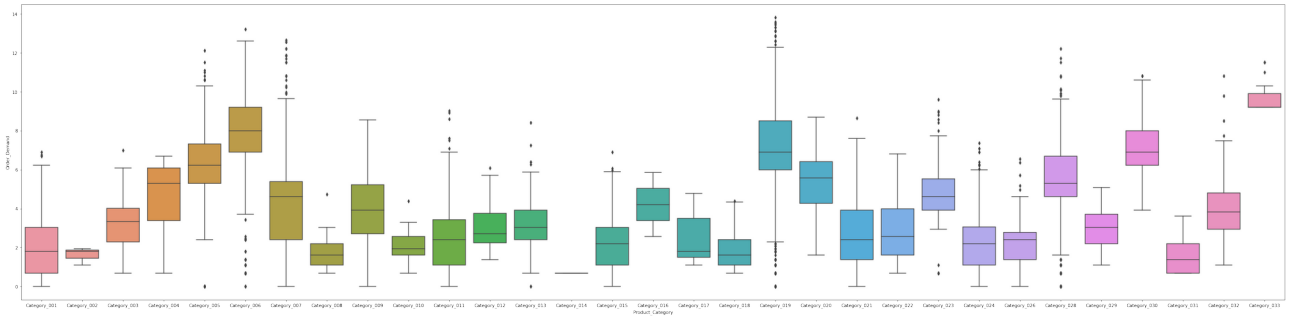


Figure 6: Checking the orders by product category with log transformation

The date-time index looks like: "['2012-01-01', '2012-01-02', '2012-01-03', '2012-01-04', '2012-01-05', '2012-01-06', '2012-01-08', '2012-01-09',….,'2016-12-20', '2016-12-21', '2016-12-22', '2016-12-23', '2016-12-25', '2016-12-26', '2016-12-27', '2016-12-28']".
Making a graphical representation of this time-frame in Figure 7, I noticed that the sales are always low for the beginning of the year and the highest peak in demand every year is in the last quarter. The observed trend shows that orders were higher during 2014-2016 then reducing.

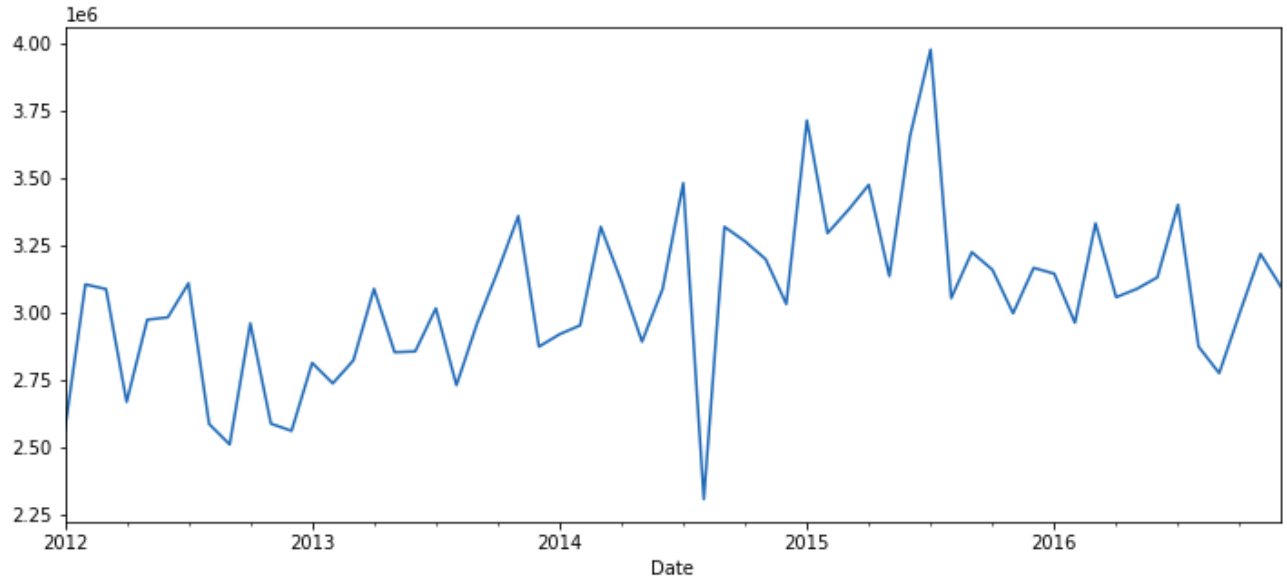


Figure 7: Time series analysis

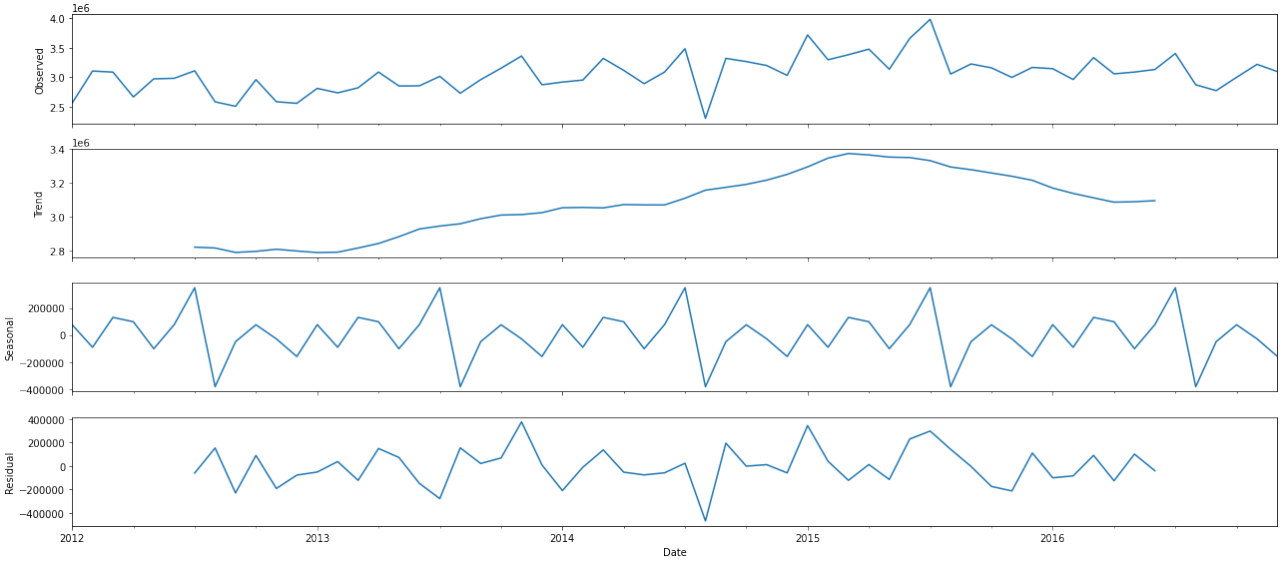


Figure 8: Time series decompositon (Observed, Seasonal, Trend, Residual)

I use the Akaike Information Criterion (shorter AIC) for measures how well a model fits the data while taking into account the overall complexity of the model. AIC is an estimator of the relative quality of statistical models for a given set of data. The lowest AIC value is, the best parameters for sarimax model we get (fits good with lesser features) as we can see in Figure 9. If the AIC value is large the model will fit good but will use more features and will become much slower. An ARIMA model is characterized by 3 parameters: p (seasonality), d (trend), q (data noise):

- “p” represents number of lags of Y to be used as predictors. For example: if it snowed for the last week, it is likely it will snow tomorrow;
- “q” is the order of the moving average term;
- “d” is the number of differences required to make the time-series stationary, if is already stationary then d=0.

When dealing with seasonality, it is best to incorporate it as 's' ARIMA(p,d,q)x(P,D,Q)s. Where 'P,D,Q' are nonseasonal parameters and 's' is the periodicity of the time-series, means: 4-quarter or 12-yearly.

```
SARIMA(1, 1, 1)x(1, 0, 0, 12)12 - AIC:1285.44377400188
SARIMA(1, 1, 1)x(1, 0, 1, 12)12 - AIC:1259.1781884157738
SARIMA(1, 1, 1)x(1, 1, 0, 12)12 - AIC:960.5164122018646
SARIMA(1, 1, 1)x(1, 1, 1, 12)12 - AIC:3106.0496679810417
```

Figure 9: Best combination of parameters found

Quickly generate a model diagnostics object (Figure 10) and when investigating an object like this for any unusual behavior, we need to make sure that:

- **Residuals should be normally distributed**, checking the top right figure: the KDE line (orange colored) should be closely matched with green colored N(0,1) line. This is the standard notation for normal distribution with mean 0 and sd 1. In bottom left the Q-Q plot shows the ordered distribution of residuals (blue dots) follows the linear trend of the samples taken from a standard normal distribution with N(0, 1);
- **Residuals are not correlated**, checking the top left figure we observe the standard residuals don’t display any obvious seasonality and appear to be white noise. The auto correlation plot on the bottom right, shows that the time-series residuals have low correlation with its own lagged versions.

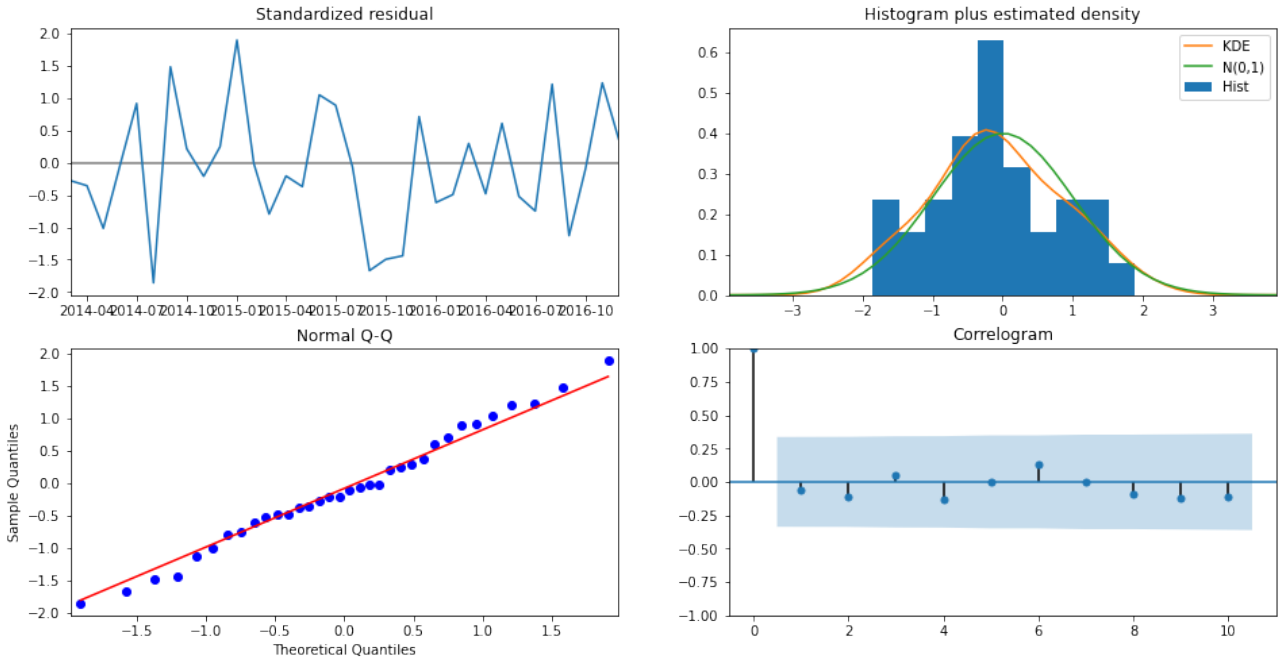


Figure 10: Diagnostics

The forecast will start from the 1st of Jan 2017, the forecast seems to be fitting well to the data. The blue/purple thicker plot shows the confidence level in the forecasts. Far away values are naturally more flat to variance, the gray area presents the confidence we have in the prediction.

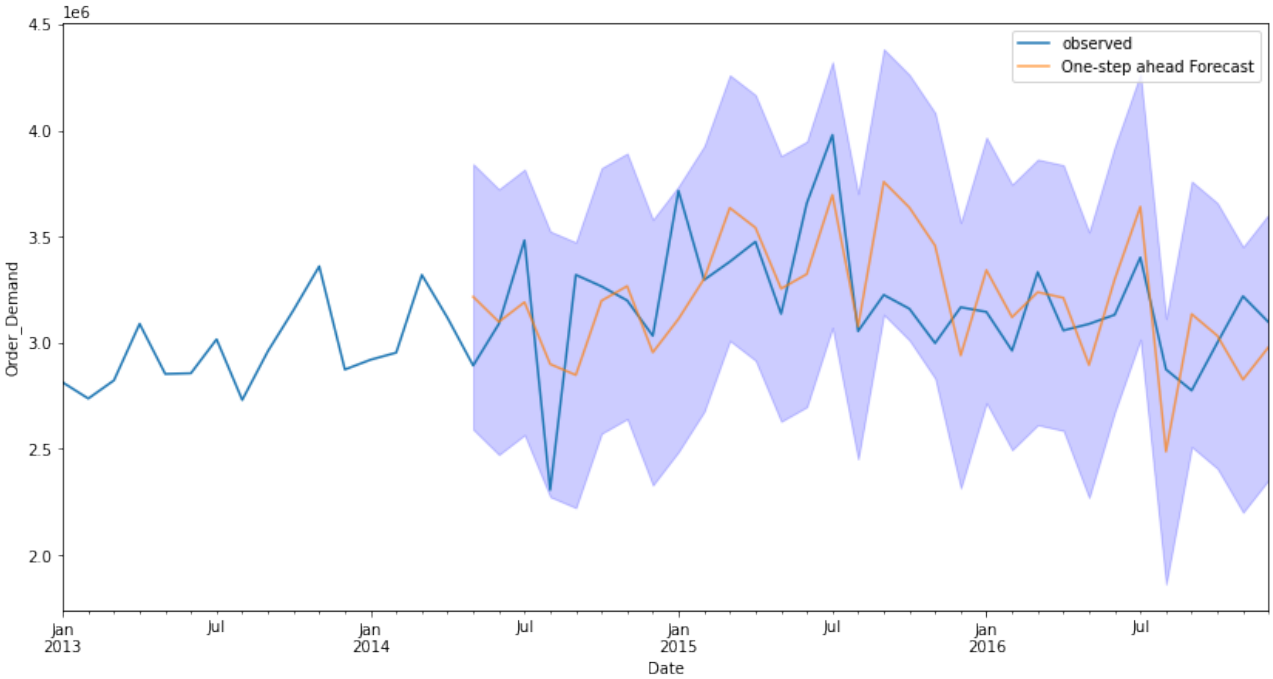


Figure 11: Plotting real and forecasted values

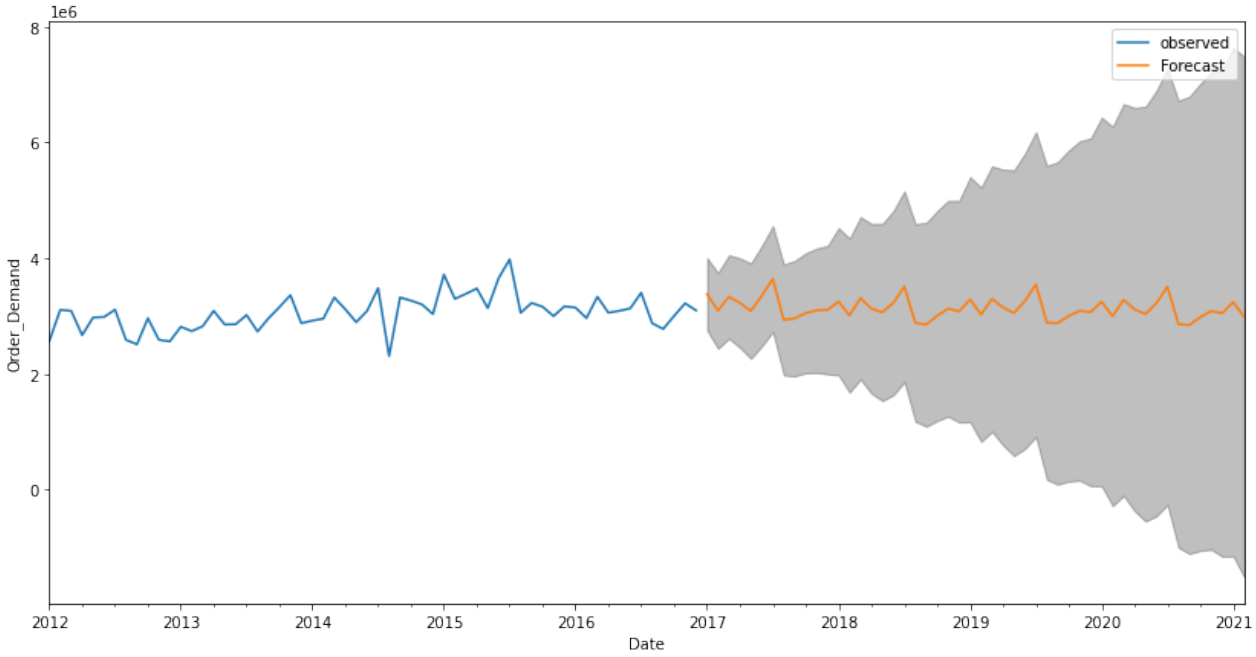


Figure 12: Forecast accuracy (mean squared error and root mean squared deviation(error))

REFERENCES

[1] Chang, P.-C. and Wang, Y.-W. (2006). Fuzzy and back-propagation model for sales forecasting in PCB industry. Expert systems with applications, 30(4):714–725;

[2] Kwok-Wing Chau, Chun-Tian Cheng: A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, Journal of Hydrology, Volume 374, Issues 3–4, 2009, Pages 294-306,ISSN 0022-1694;

[3] Zhang, G. P. (2003). Time series forecasting using a hybrid arima and neural network model. Neurocomputing, 50:159–175;

[4] Nari Sivanandam Arunra, Diane Ahrens, Michael Fernandes: Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry, Volume 7 Issue 2 April-June 2016: 3-4.

[5] https://www.sciencedirect.com/topics/engineering/artificial-neural-network-model