IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predictive Analysis of Time Series Data of Online Sales of Products

Dr. M. Arathi, Nazifullah

Associate Professor of Computer Science and Engineering, MTech Student
School of Information Technology,
Jawaharlal Nehru Technological University, Hyderabad, Indian

Abstract: According to Satista (a German company specializing in market and consumer data) the amount of data that is created in 2020 is 64.2 zettabytes, and it is going to increase rapidly, in 2025, the data produced, copied, capture and consumed is projected to reach 181 zettabytes.

"Sales data is a term that includes a large array of metrics" but, broadly speaking, if we can measure something in relation to the sales process, it's viable sales data. Modern software solutions can help us collect this data, but it's important to learn how to read this data to understand what it means for the business and where we can improve.

In this project, we will perform the Explanatory Data Analysis (EDA) for sales data, the data will be visualized to find patterns and performance of the business in the past. Then we will move on to predictive analysis, at first. Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Averages with exogenous regressors (SARIMAX) model will be created to project future sales. Facebook prophet which is new open-source model provided by Facebook will be also used prediction and at last deep learning methods LSTM will be applied to data.

Index Terms - Data analysis, time Series analysis, sales data analysis, predictive analysis, future sales prediction, ARIMA, SARIMAX, Facebook Prophet, LSTM, explanatory data analysis

1. INTRODUCTION

Data is "Factual information (like measurements or statistics) used as a basis for reasoning, discussion, or calculation,"
In the 1996 Webster's II New Dictionary Revised Edition. Data is defined as "factual information (like measurements or statistics) used as a basis for reasoning, discussion, or calculation," according to the Merriam Webster Online Dictionary.

Using the given definitions as a reference, data can be defined as numbers, characters, photos, or any other type of recording in some kind of a form that can be processed to make a decision or determination concerning a certain activity. We can uncover patterns in data to perceive information, and afterwards information can be used to improve knowledge by examining it attentively. Quantitative and qualitative analysis can indeed be accomplished in a number of approaches. Different methods enable data analysts to build a "logical sequence" for the usage of various procedures, giving them a more ordered approach to dealing with data. following, four examples of quantitative analysis methodologies are provided that might be examined as you work with and build our data analysis skills, as well as reasons why we might want to use them. When looking at online sales data in, several of these strategies will be applied in our project.

1- Visualization: Visualization of the Data Encompasses Creating a visual "picture" or graphic representation of the data.
2- Explanatory Data Analysis: EDA: Looking at data to determine or explain "what's going on" is what exploratory analysis implies. – it establishing a basic framework (baseline) for additional investigation.
   Every analysis includes EDA, it helps us understand the data and uncover basic information, it can provide us information easily that may not need complex computation which can have huge significance in business.
3- Trend Analysis: analysing at data at different periods of time. Trend analysis can be really help to visualize the performance of the business.
4- Predictive Analysis or Estimation: Predictive Analysis includes predicting a future value by using the actual data or historical data. Predictive analysis is used to plan for future, develop strategies accordingly and make our policies that best fit our business.

In our project we will start form explanatory data analysis and visualize our finding and then we will progress to find and uncover trends and seasonality in our data and visualize that as will and at last we will create machine learning models to predict the future and visualize at.

2. LITERATURE REVIEWS

Time series creates relation between cause and effect where one variable is the time and another variable is data. Where time is independent variable and data is dependent variable.

Time series classification in neural net is associated with LSTM. Time series data is ubiquitous from IOT devises to self-driving car it affects our life beyond we can imagine. Building a predictive model for time series is a challenging task. Machine learning has been a field of research for a few last decades for a lot of data scientists. We have different methodologies for analysis using machine learning,

dynamic time warping (DTW): dynamic time warping is a similarity methodology approach which apply dynamic time warping to actually align the time series and apply supervised classification on top of it to do time series classification because this technique uses dynamic time warping the time complexity is really high which causes to become incapable to work in real time. Key issues in DTW are: the time complexity in DTW is high, it requires huge computation power. And it is hard to implement in real time.

Auto-Regressive Integrated Moving Average (ARIMA): From the statistical part we have the ARIMA modelling which actually uses the auto regressive nature of the time series and based on that it tries to predict the sequence but assumption is basically that there is linearity in our current data which causes the model to predict the future accurately. But when the data is noisy which in many real-world scenarios the data is much noisier than then we may think. Thus, ARIMA modelling will not work in such scenarios. Key issues in ARIMA: if the data is heavily noisy and seasonal the ARIMA model may not work. It requires a lot work and skills to make the data ready for this model.

Recurrent Neural Nets (RNN): With the advent of deep learning, we have recurrent neural networks. In the past few years there have been an influx of papers working on time series classification using RNN or recurrent neural nets. RNN in applied mostly in time series its major applications are music generation, Natural Language Processing (NLP) or speech recognition, machine translation, application of auto encoders and etc. RNN are model where we actually exploit the sequential nature of the time series we have our hidden vector which is actually propagated through time and the intuitive appeal of recurrent of a recurrent neural net is essentially the hidden vector activation is actually the representation of the entire time series. Each output is the function of the previous outputs and learn pattern via parameter sharing. RNN is able to glean context by parameter sharing and unbounded nature of the RNNs.

Key issues in RNN: when the data is missing or it noisy it will not give us accurate output, it has time complexity and takes longer time to train the nets and trade over it, it's due to complexity dependencies, it is slow to train. It requires huge amount of data to provide reasonable results.

3. METHODOLOGY

In this project at first the explanatory data analysis is performed. then the time series analysis to understand the trends and seasonality in data. The models are produced to predict the future of the business using time series analysis. We will perform the predictive analysis to understand the future sales of the company. We will use the ARIMA, SARIMAX, Facebook Prophet, and LSTM Neural Network to develop our model and visualize the result.

Time series data:

Time series is a tricky thing it is very complex and controversial. They are everywhere in our life, from personal planning to business planning and up to the government making economical strategies and future planning its all-time series, it is just sometime we don't recognize it, global warming is a time series, global temperate, natural disasters like earthquakes or the hurricanes heavy rainfall and droughts they all occurs in time if look closer at it we can find the time interval or continuity in time. stock prediction is a time series, when will the next peak occur? or is the trend going up or down? is there any kind of seasonality in data? It is all about time.

Predictive analysis or Estimation:

Predictive analysis is an insight to future data based on current or historical data by implementing statistical methods, predictive modeling and machine learning techniques. Predictive analysis can be used to optimize resources, make future plans and strategies, detect fraud, optimize business marketing, improve business operations, and reduce future risk and confront challenges.

3.1. Dataset:

The data set that is used in this analysis is form online product sales form Brazil called Olist, the data is analyzed previously but nobody has implemented Time Series analysis or Predictive analysis over it, at least up to my knowledge. The data set has 9 table and 48 total columns, with hundreds of thousands or record.

3.2. Data Preprocessing:

After completion of EDA which we were able to understand our data, identified outlier and data entry or any other problems next step is preprocess our data and make it ready for our machine learning model to get desired output.

Any data that we receive is in raw format and is not ready for analysis or at least we should make sure that it is ready for analysis, the process of transforming or converting the raw data into a more meaningful understandable format and most importantly to make it ready for the machine learning model is data preprocessing.

The primary aim of data preprocessing is to make the data ready for model and increase the accuracy and the speed of the model.

3.2.1 Data Cleaning: after collection of any data and initiating data analysis the first step is to clean the data. Noisy data is the common problem for any kind of data analysis. Missing data is a problem that the data analyst may have to deal with, especially in our case when we are doing time series analysis the data needs to be filled or we may discard the big chunk the data. Some of the models such as ARIMA and SARIMAX needs the data to be consistent based on some type of regular time interval.

3.2.2 Dealing with Missing Values:

There are two ways to deal with missing values. During the feature engineering we can remove those columns that has maximum number of records missing, it depends on the dataset and our objectives of data analysis. Removing the column is less desirable than removing the records, but in many situations when we even can't remove the column the next option is fill the columns with means. Fill with mean can be a little tricky depending on the data set specially when data is seasonal. In such cases we must be careful what kind of mean we have to choose from.

3.2.3 Dealing with Noise:

Noise in our data can be a great problem and dealing with it is not always easy, Noise in features mean that there are some fields or columns that doesn't relate to the target. Noise in items mean that there are outliers that exist in our data, we can visualize a specific column and identify the noise and remove it or we can provide ceiling to our column to specify how high the values can go. Noise in the records simply means that some records don't go along with the rest of the data. Those might be fixed somehow or removed from the data.

Dealing with noise there are many methods proposed in different algorithm can be used. The is binding methods that sort and partitions the data to find mean, median and boundaries. There is clustering methods that divide the data with multiple cluster that has similar values, and there is the regression method to find the best fit line.

Filtering methods can be used to remove the noise from the data such as Linear Discriminant Analysis (LDA) to find the characteristics of the features, or Pearson Correlation to find the relationship among the features and Chi-Square to find the relationship between various variables. And then there is a method called Wrapper in which we select a set of features and run it through model and see the result, we change the features and re-examine it until we find the best result. And finally, there are unsupervised Anomaly Detection methods such as density-based anomaly detection like KNN and LOF based methods, clustering based anomaly detection.

3.3. Feature Engineering:

At this stage of our project, we select the futures that is recommended for our system. In feature engineering the analyst is required to find the feature and the label, feature are the independent fields that are required to estimate the output or target variable and the target or output is the field or column that is used to be estimated or predicted or the case of future sales prediction. We are going to predict the sales so sale is our dependent variable and independent variable could be one in the case of univariate Times Series analysis or it can be multiple in the case of multi-variate time series analysis.

Explanatory Data Analysis:

Explanatory Data Analysis answer the most basic and sometime the most important question about the data and the business. How much sale did we had this month? How much sales did we had the same month last year? What is the mean difference between those months? What is the trend that may exist in data.

3.4 Data Transformation:

Data transformation involve the process to transform or convert the raw data to machine learning model ready data, every machine learning model may require their data to be transformed into specific format for example the ARIMA model needs the data to be stationary and it also requires the data to be timely and consistent in terms of time. The data type of the input variable has to be timestamp etc. in data transformation we change the form of the data in term of values, structure or format. The generalization or normalization strategies can be used to transform the data.

3.5 Data Reduction:

It is also known as features selection. it's a process of selecting the most desired features or removing the feature that doesn't have relationship with output of data. There might be some variable in our data that is not related to the label or output. We must select the features that best fit our model's requirement and that the output is dependent on.

We can also the encoding techniques to reduce the size of the data in order to speed up our model training or we could represent our data in more summary format for the same purpose.

In case of having a very large set of data we may consider the subset of data to be used for training and testing of the machine learning model.

3.6. Model Training: the main purpose of using machine learning algorithm in data analysis is to build and train the model get the desired result. In our project we will user ARIMA, SARIMAX, Facebook prophet and LSTM Neural Network model and provide the result based on the trained data.

Visualizing the result: not everybody can interpret the data if they are provided with just numbers, in other cases the top management may not have the time to look at all the numbers provided. They need visual representant to get the broader picture.

1. Auto-Regressive Integrated Moving Average (ARIMA):

Stationarity:

Most of the statistical models requires the data to be stationary. If a time series follow a particular behavior such as trend, seasonality, and cycle it has a very high probability to follow the same pattern in the future. The formulas applied to the time series are more mature for the stationary data and they are also easy to be implemented. A data is tended to be stationary if it has a constant mean, constant variance and autocovariance that does not depend on time

Checking the stationarity: To check the stationary there are multiple test available that when applied can tell us if the time series is stationary or not. These tests include:

1. rolling statistic
2. ADCF test or Augmented Dicky Fuller Test
3. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

ARIMA model has 3 parameters it is; p = auto regressive lags, q = moving average lags, And Integration (I) part which is referred to as d = order of differentiation

to predict the value of 'p' we use Partial Auto Correlation Function (PACF) graph, to predict the value of 'q' we use the Auto Correlation Function (ACF) graph and the value of 'd' is which is equal to order of differentiation, to make our time series stationary we use some kind of differentiation, order of differentiation defines the value of 'd'.

ARIMA(p,d,q) if p = 1 and d =1 and q = 2

ARIMA (1,1,2)

In our data applying the ACF and PACF our values are (1,1,1).

2. SARIMAX:

ARIMA is the best and most used model for univariate time series prediction but it has one shortcoming that is it cannot handle the seasonality which in reality there is almost some kind seasonality present in any kind of data.

Seasonal Auto Regressive Integrated Moving Average eXogenous (SARIMAX) model is used when there is seasonal component on our data. It is the same as ARIMA except it handles seasonality. All the three parameter that is present in ARIMA, AR ('p' value), MA ('q' value) and Integration part ('d' value) also exist in SARIMAX. Additionally, it as four extra parameters represented by capital 'P', capital 'Q', capital 'D' and 'S' which represent Seasonality.

- P = Seasonal autoregressive order.
- Q = Seasonal difference order.
- Q: Seasonal moving average order.
- m: The number of time-steps for a solo regular or seasonal period of time.

SARIMAX (p, d, q) (P, D, Q) m

where after we placed the values, it will be like:

SARIMAX (1,1,1) (1,1,0)12

Significantly, the m parameter effects the P, D, and Q parameters. For example, an m of 12 for monthly data advocates a yearly seasonal cycle and if the m=1 it will show the weekly seasonality and if m = 3 that means the seasonality is there for each three months repeating.

A P=1 would make use of the first seasonally offset observation in the model, for example, t-(m*1) or t-12. A P=2, would use the last two seasonally offset observations t-(m * 1), t-(m * 2). Similarly, a D of 1 would calculate a first order seasonal difference and a Q=1 would use a first order errors in the model (for example, moving average).

3. Facebook Prophet:

Facebook Prophet is an open-source business forecasting tool that was introduced in 2018. As the Facebook says the tool is aimed at the people who don't want to lean machine learning models that are sometimes difficult and the primary goal of Prophet is to make the business prediction easy. As being open-source its available in PyPI and CRAN and can be downloaded and used in pip and Conda environment. It can be consumed by Python programming language.

Facebook Prophet is a generalized additive model combining the seasonality function, the trend function, the holiday facts and the error term which can be represented as:

y(t) = g(t) + s(t) + h(t) + e(t)

- g(t) is the trend (upward, downward or horizontal trend)
- s(t) refers to seasonality (daily, weekly, monthly, yearly)
- h(t) refers to effects of holidays to the forecast
- e(t) refers to the unconditional changes that is explicit to a business or a person or a circumstance. It is also called the error term.
- y(t) refers to forecast.

Facebook provides two model one is the piecewise linear model and second is logistics growth model. Be default the piecewise linear model is selected. If the times series data has linear properties and has a upwards or downward trend than piecewise model is a good choice and we can change the default model using the model property and make it logistics growth model, choosing model can be a little tricky as it depends on the features of the times series dataset such as size of the data, growth rate, business model and etc. if the time series data has a saturating and non-linear data which basically mean that the data grows non-linearly and when reaching the saturation point it shows no growth at all or it may show very little growth then choosing the logistics growth model is a best choice.

As first component piecewise trend(t) Facebook uses the L1 regularized trends shift, which can be represented as following equation:

Or in case of logistic growth model is fit using the following statistical equation,

$$y = \begin{cases} \beta_0 + \beta_1 x & x \le c \\ \beta_0 - \beta_2 c + (\beta_1 + \beta_2)\, x & x > c \end{cases}$$

$$g(t) = \frac{C}{1 + e^{-k(t-m)}}$$

And as a second component s(t) which is seasonality Facebook prophet uses the Fourier Series, and for the holiday effects h(t) it uses the dummy variable and the last component is to take care of noise or outliers in the time series data.

The trend in data is automatically detected by the Prophet in both piecewise linear and logistic growth curve trends. It detects the yearly seasonality and weekly seasonality by Fourier series and dummy variable and the user defined important parameter that the user can provide to the model as a list.

4. LSTM Neural Network:

Recurrent networks are artificial neural network that recognized the pattern of data such as text, gnome, handwriting and spoken language or numerical time series data that may generate from sensor or any other resources.

Recurrent neural networks, or RNNs, are explicitly designed to work, learn, and predict sequence data.

In RNN the output of the network from previous step is provided as an input to the succeeding time step. In this model the prediction occurs on both the current input and the result from the previous step and after these two step the result is feed to the next step as well as new input.

Here we have an input ‘t-1’ which is feed it to a neural network and we get the output ‘t-1’. In the second step ‘t’ is provided as an input as well the result or information from the last step ‘t-1’ to the neural network and we get the output of ‘t’, it continuous until the last step or last sequence of data, so there is loop going on the RNN neural network that happens again and again which gets input from both new data and the information from the previous neuron provide to it.

LSTM also has a chain like structure just like RNN, but the repeating module has different structure. Instead of having single neural network here there are four which interact in very special way. The upper line in the diagram is cell state we can consider it as a conveyer belt

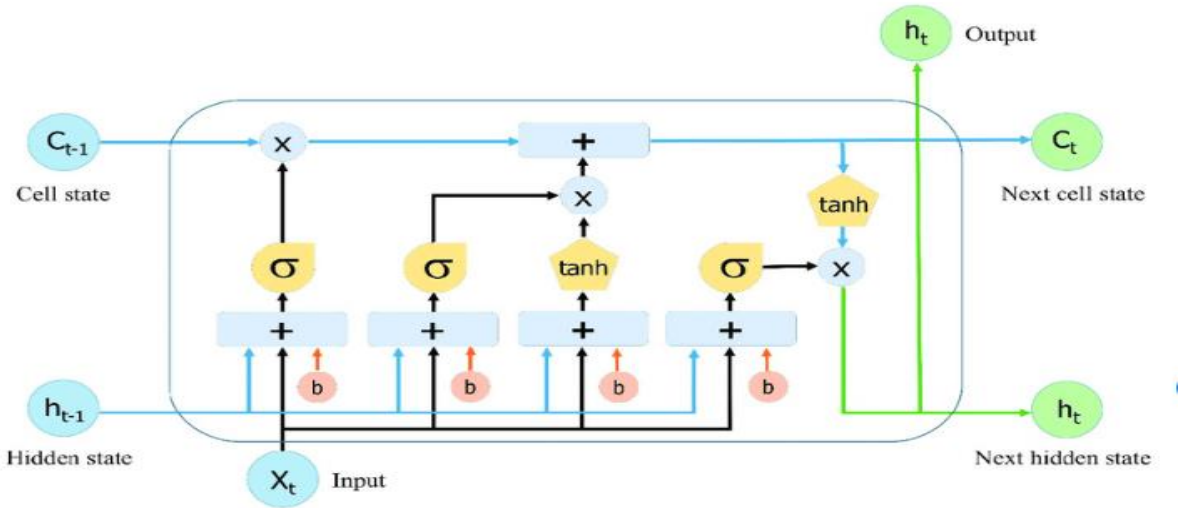

Figure 1 LSTM cell Structure

4. RESULTS

ARIMA Model: test and train

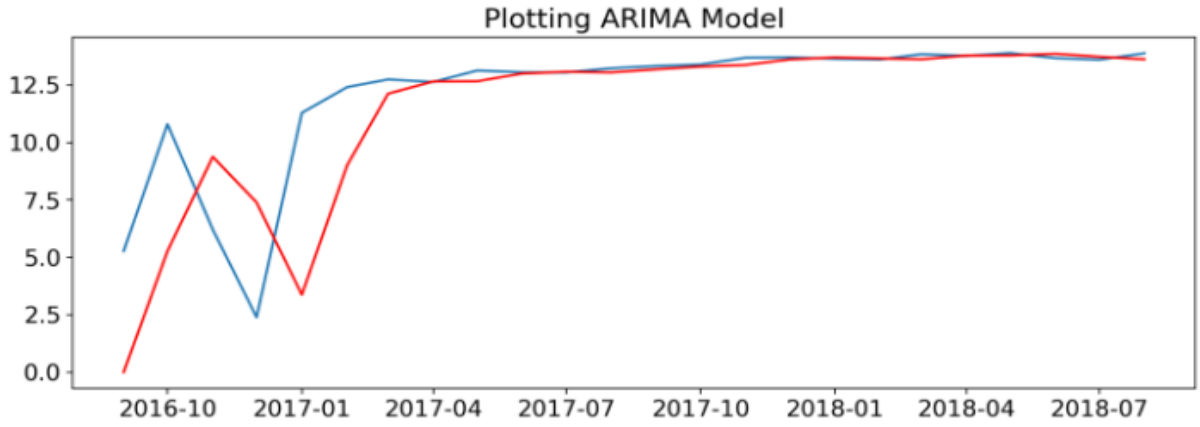

Figure 2: Training ARIMA

4.1 ARIMA future Sales Prediction:

Figure 3: ARIMA sales forecast

4.2 SARIMAX future Sales Predictions:

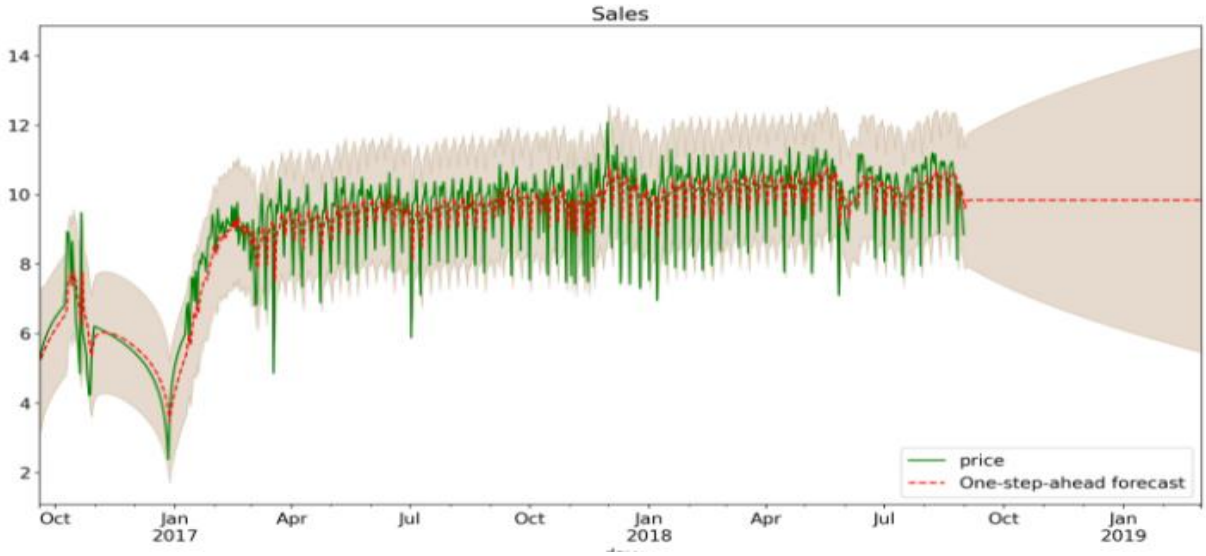

Figure 4: SARIMAX Sales forecast

4.3 Facebook Prophet:

Future sales prediction by Fb Prophet

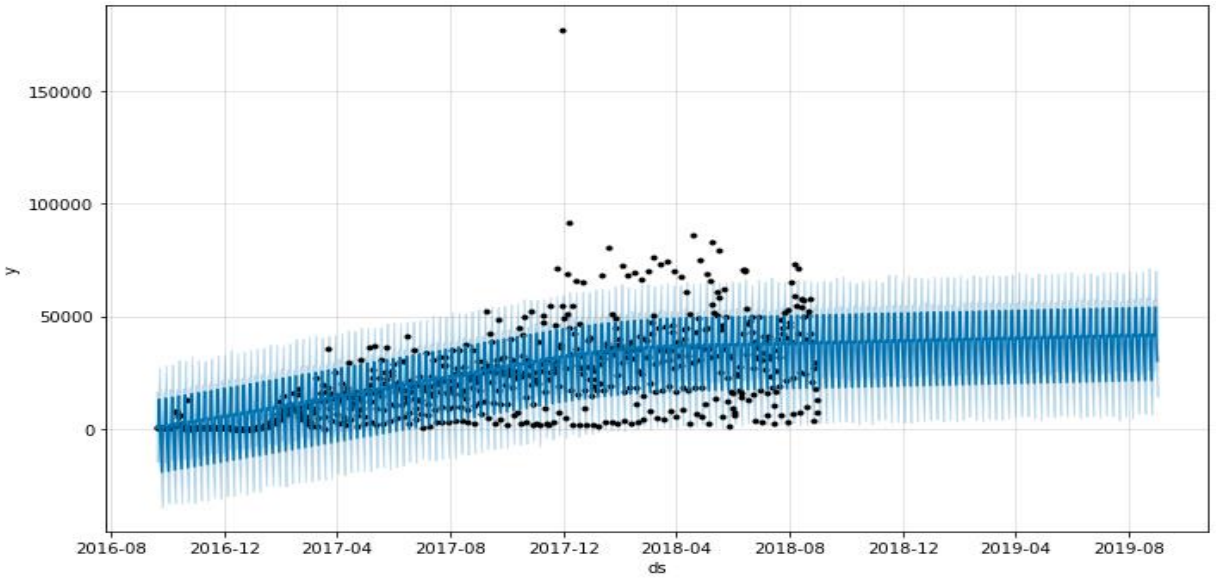

Figure 5: Prophet one year prediction

Fb Prophet performance Metrics:

	horizon	mse	rmse	mae	mape	mdape	smape	coverage
0	18 days	1.332582e+08	11543.751907	8226.897272	2.565716	0.363629	0.670732	0.583333
1	19 days	1.327662e+08	11522.422659	8235.669177	2.651072	0.373964	0.680525	0.583333
2	20 days	1.323641e+08	11504.958770	8231.957036	2.722103	0.381520	0.681887	0.583333
3	21 days	1.144084e+08	10696.186753	7926.894848	2.912155	0.395518	0.695077	0.574074
4	22 days	1.065951e+08	10324.488254	7623.170717	2.937187	0.395518	0.699053	0.601852

Figure 6: Prophet Performance Matrix

FB Prophet Cross Validation Graph:

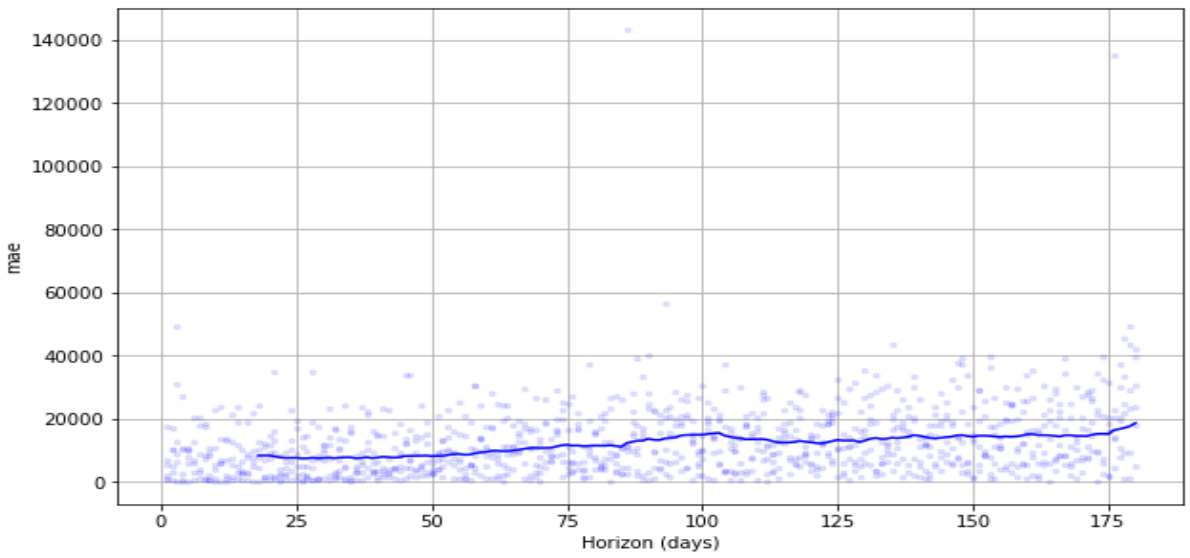

Figure 7: Fb Cross Validation Graph

4.5 RNN LSTM: preparing data for LSTM

Each sequence is going to contain 10 data points from the history:

`(630, 10, 7)(630,)`

- Let's start with a simple model and see how it goes.
- we will now train the model

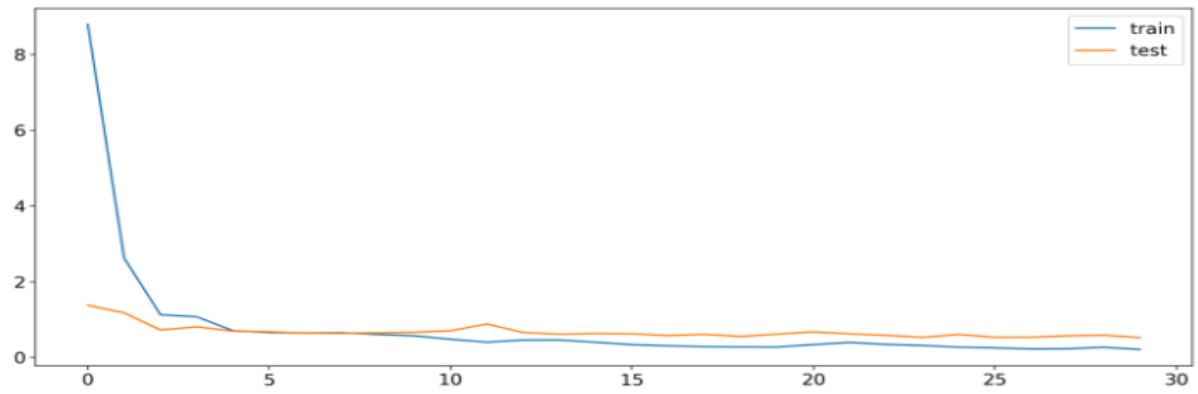

Figure 8: Training LSTM

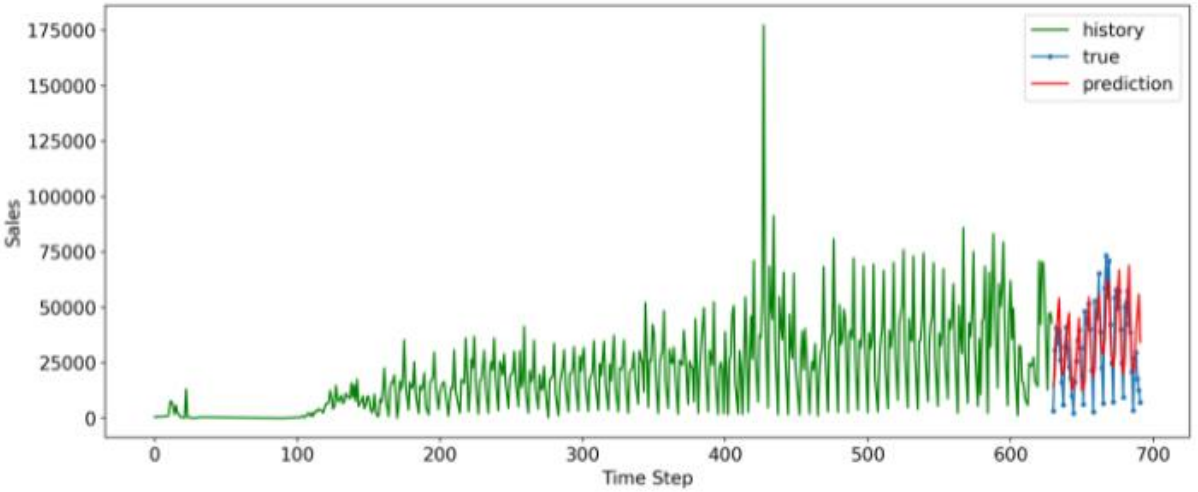

Figure 9: LSTM prediction

LSTM Validation:

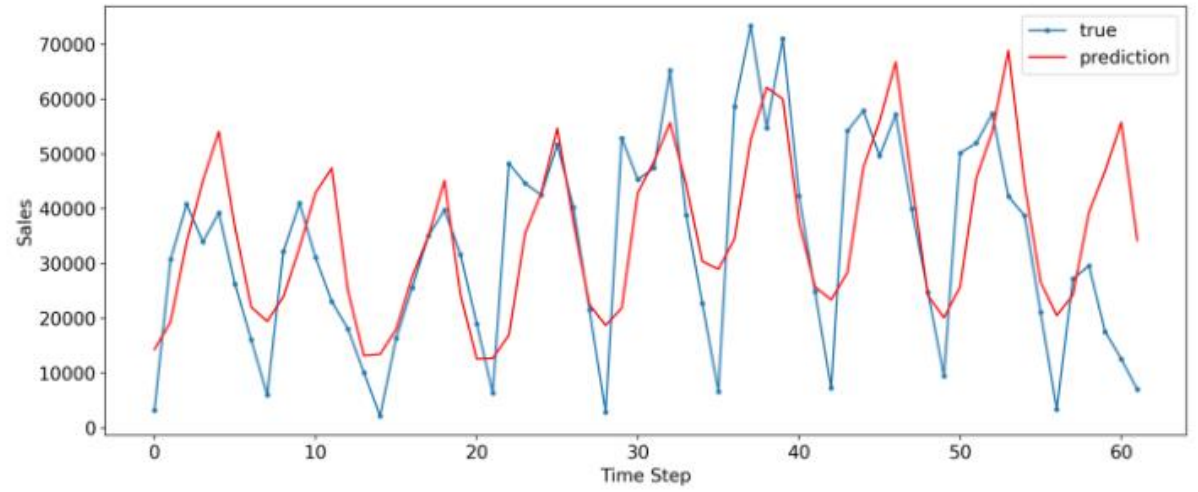

Figure 10: LSTM Validation

5.CONCLUSION AND FUTURE WORK

we did some basic study of our data and performed explanatory data analysis which gave a comprehensive information about our dataset, on the business side we provides concise information and graph for decision makers such as list of best sold item and best seller and etc. We also found some weak point in the business like most of the customers are one time buyers and mostly the orders included only one item which can be addressed as providing bundle offers, better recommendation system and to increase the items sold per order the business need to offer discount and encourage the customers to buy more. We used ARIMA model to draw a picture of the future sales, although ARIMA is a good model it is complex, it needs mastering to understand every aspect of it and the math behind it as well as the parameters that it uses needs to be selected carefully or multiple values maybe used and check the outcome based on each change to find the desired outcome. The preparation of data for ARIMA is also complicated, to remove trend and seasonality we must try different methods until we reach the goal. on the other hand, SARIMAZ provide the same result but it doesn't need the dataset to be stationary it can handle the seasonality and we may define it using parameter. Facebook prophet is a lot easier to use, it is easier to understand and learn, it has a lot of functionality, as being open source, the tools provide for validation and testing and graphing is best, but the result may not be much different from the ARIMA and SARIMAX although better graphic and presentation is provided. The last model used was deep learning model LSTM RNN which uses neural nets the is obviously the best feature for predicting the future of the business, it is more accurate it provides a large range of parameters, it can be trained to learn from features to use or throw the feature that is not needed if large dataset is provided to it. Even though it is prone to overfitting, we have to know the math behind it to better understand and use it properly. It requires more resources than any other model use in our project, because of the computation complexity in the neural cell.

The future work is to provide the recommendation system for the business and the use of CNN model, grid LSTMS and other changes the rapidly adopted for LSTM and time series.

6. REFERENCES

1. Ioannis E. Livieris, Emmanuel Pintelas & Panagiotis Pintelas, A CNN–LSTM model for gold price time-series forecasting, Springer, 13 April 2020
2. Taoying Li; Miao Hua; Xu Wu, A Hybrid CNN-LSTM Model for Forecasting Particulate Matter (PM2.5), 2020 16th International Conference on Computational Intelligence and Security (CIS), IEEE, 2020
3. Sean J Taylor, Benjamin Letham, Forecasting at scale, PeerJ September 27, 2017
4. Wenjie Lu, Jiazheng Li, Yifan Li, Aijun Sun, and Jingyang Wang, A CNN-LSTM-Based Model to Forecast Stock Prices, hindawi Article ID 6622927, 2020
5. William W.S. Wei, Time Series Analysis, Oxford Publications, Mar 2013
6. Daniel Peña, George C. Tiao, Ruey S. Tsay, A Course in Time Series Analysis, New statistical methods and future directions of research in time series, A Wiley-Interscience Publlications, 2001
7. Yatao Li; Fen Ying, Multivariate Time Series Analysis in Corporate Decision-Making Application, 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, IEEE 29 December 2011
8. A. V. Seledkova; L. A. Mylnikov; Krause Bernd, Forecasting characteristics of time series to support managerial decision making process in production-and-economic systems, 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM), 24-26 May 2017
9. Sai Swaroop Ratakonda; Sreela Sasi, Seasonal Trend Analysis on Multi-Variate Time Series Data, 2018 International Conference on Data Science and Engineering (ICDSE), IEEE, 12 November 2018.
10. Ildar Batyrshin, Constructing Time Series Shape Association Measures: Minkowski Distance and Data Standardization, 2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence, 17 July 2014.
11. Zhijie Wang, TCL stock price prediction model based on LSTM RNN, 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), IEEE, 02 August 2021.
12. Norshakirah Aziz; Mohd Hafizul Afifi Abdullah; Ahmad Naqib Zaidi, Predictive Analytics for Crude Oil Price Using RNN-LSTM Neural Network, 2020 International Conference on Computational Intelligence (ICCI), 09 November 2020.