Sales Prediction for E-commerce Product Categories to Increase the Revenue and Profit.

Implementation Report

Abstract— **The emergence of online shopping has changed the way the retail industry operates. Over the past decade, the industry has grown exponentially year by year. Seasonal and Buyer specifics trends, need of expansion, competency among businesses, customer satisfaction made use of Predictive analysis in the field of e-commerce business essential. In this research, we will be focusing on predicting one of the important aspects of the business which is Sales. The investigation of past research is performed to select and understand the applicable technique and gain domain knowledge. The analysis is performed on 3 years of historical data of the Ecommerce business. The key business features, trends and seasonality are extracted using exploratory data analysis. For achieving accurate prediction of sales for 12 months ahead we have implemented Seasonal Autoregressive Moving Average. The implementation of SARIMA is done by adapting the CRISP-DM methodology. The model predicts with the Root Mean Square Error, range from 2655 to 8962 across various product categories. Different Plots, KPI's, Bar charts, Line charts and Area charts are used to represent the predicted vs actual sales values.**

Keywords- *E-commerce, Sales, SARIMA, RMSE, ME, AIC, CRSIP-DM, Profit*

I. Introduction

A. *Background of the Domain*

Over the past few decades, the modes of shopping have changed significantly, the advances in the field of internet computers and mobile devices made this change possible. This allowed the consumer to buy anything in the world from anywhere in the world using a single click of the device. E-commerce accumulates over 26 Trillion revenue over the past two decades. Currently, there are 2 Billion users of e-commerce platforms worldwide. This is 25 % of the world's total population. Predictions have been made that these numbers will grow exponentially every year. 95 % of the retail sales is expected to be done via e-commerce platforms by 2040. By the year 2021, there will be 2.20 billion e-commerce platform users [1].

Despite the exponential increases in the number of users, not all e-commerce businesses were successful. The main reason for this was increasing competition among businesses, lack of market trend analysis. If an e-commerce business is completely aware of consumer behavior, what are the products which they are likely to buy? What are the payment methods used? The company would not only outperform competitors but it will improve its performance as well. This can be achieved using Market Trend and Demand analysis [2].

This paper attempts to predict the sales for the e-commerce company for the coming month, quarter and year. This would help the company accommodate the consumer demand which will lead to better customer satisfaction, increased profit and an edge over competitors. The dataset chosen for analysis is taken from [1] Kaggle. The data is published by the [2]Data Co organization. The Seasonal Auto-Regressive Moving Average (SARIMA) is the time series model considered for the implementation for prediction.

The research is organized in the following manner. The research and investigation of the applicable techniques are done in Section II. The implementation of the selected technique is done in section III. The qualitative and quantitative interpretation of the finding is done in Section IV.

II. Research and Investigation Of Applicable Technique

This research [3] has successfully predicted the sales for an e-commerce fashion store, for which multiple machine learning algorithms were considered. Decision Tree, Generalized Linear model and Gradient Boosted Trees were compared using Accuracy Rate (%) Error Rate Precision Recall as evaluation metrics. The dataset was consisting 83,000 rows of 2015-2017 data. Out of which 75 % was used for training and 25 % used for testing purposes. The methodology chosen for model implementation is CRISP-DM. Among implemented, Gradient Boosted Trees performed significantly better than the other two models with an accuracy of 98%.

Research conducted [4] in this paper proposed an eXtreme Gradient Boosting (XGBoost) with feature engineering for sales forecasting. Publicly available Walmart retail dataset which contains 1913 days of historical data is used for model building. Before applying the XGBoost algorithm multiple features are extracted which are time, price, lag, rolling based statistical features. Model has achieved 16 % ,15% lesser RMSE value in comparison to Linear and Ridge regression respectively. The model can achieve more significant result as there is scope to explore more features which will increase prediction accuracy and reduce the error.

The sales of computer peripherals are in increased demand since the technology advances. The research [5] presented the models which predict the sales of Notebooks, LCD Monitors, Motherboards. The models implemented are Support Vector regression, multivariate adaptive regression

[1] https://www.kaggle.com/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis

[2] http://dataco-global.com/

splines and a combination of both. Here MARS is used to extract the important predictors and Support Vector Regression is used to predict the sales for the next quarter. For all three product categories, the hybrid model achieved the best results but results were not constant across product categories. The model is not able to capture non-linear relationships among the predictors.

The prediction of sales in the food sector is another challenging task. The research [6] has proposed a methodology that has implemented Support Vector Regression and Random Forest Regression for predicting the sales for different food products. Out selected 20 features only 5 were significant for model prediction. The Random Forest Regression model achieved better accuracy of 87.7 % which is 3 % greater than the support vector regression model. The model seems to be overfitting the data as Training accuracy is higher than the testing accuracies.

Another research [7] used a neural network approach for predicting the short-term sales for the supermarket. A three-layer feed-forward perceptron model was designed which takes time series with three independent variables as input and predicted values for future as output. The model achieved significant results but data fed to the network was small. The model could not achieve similar results for a large dataset. The model is built using an excessive number of input neurons; therefore, the model is computationally exhaustive. There is scope to improve the feature vector using modeling techniques to reduce the training error.

Trend, Seasonality and Randomness of historical sales have also proven to be effective in accurately predicting the sales for the future. Research [8] implemented Winter Exponential Smoothing is used to predict sales. The model begins with decomposing the time series into Stationary, Trend and Seasonality pattern. The model takes input parameters such as Alpha, Beta, Gamma, number of periods to forecast. The model has achieved MAPE values between 20-30% between different product categories. The dataset contains day-wise data of one month therefore there is concern whether the model will reproduce similar results for time series with large historical data.

Research [9] implemented a combination of time series and neural network model to predict the e-commerce sales for a cargo company. The models used are ARIMA-BP. Both the models independently were not exhibiting accurate results but the combination of two of them was found to be successful in getting accurate results and reduced error. The model has achieved significant results but the model has high complexity thus it is less interpretable.

Amazon is the leading e-commerce business which is a typical business to consumers. Research [10] implemented Holt's Winter Exponential Smoothing, Neural Network and SARIMA to predict future sales using historic data of the past 4 years. The SARIMA model achieved the highest result across implemented models. As the MAPE obtained is significantly less this model can be adapted for other domains as well. As the model predicts data quarterly it captures significant seasonality and trend. MAPE is the only evaluation metric used, models could be tested using other metrics as well.

Research [11] used modified Knowledge Discovery in databases methodology to predict the sales for the retail chain. The research considered several regression algorithms which focus on feature extraction and accurate prediction using those feature extractors. Multiple Regression, ElasticNet, Lasso Regression, Polynomial Regression, Ridge Regression were used to predict the daily sales. Polynomial Regression achieved the lowest RMSE value among the implemented models. Data Exploration, Data Pre-Processing, Feature Engineering and Model Building were the steps involved.

The research [12] considered a combination of qualitative and quantitative forecasting methods. The qualitative method is based on the different variables which are responsible for accurate prediction whereas the quantitative method focuses on the record of sales. The model uses ARIMA for the Quantitative method and RNN for the Qualitative method. The implementation of these two methods forms a complex structure with 15 different steps which are computationally exhaustive. The prediction results of individual models were not acceptable but the combined model achieved better results.

After reviewing the past work, their merits and demerits, the decision of implementing the SARIMA model is made.

III. Methodology

The methodology used to implement the proposed machine learning model is CRISP-DM. CRISM-DM has its advantages when it comes to building Data Mining Models for industries as Business Understanding is an important step that help analyze the business goal and constraints. The remaining steps are universal as available in other methodologies as well [13].

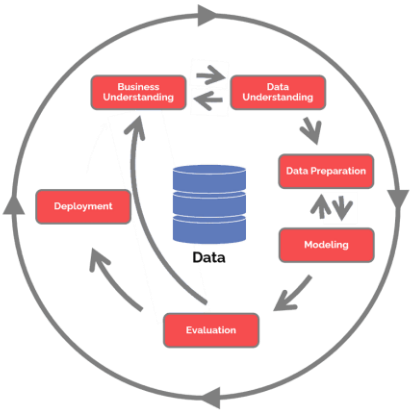


Figure 1: CRISP-DM Methodology

A. *Business Understanding*

The e-commerce business is always thrived to grow exponentially but increased competition, fluctuating trends have found to be concerning points for businesses.

Ecommerce businesses are always evolving and adapting to the market trends but to accommodate those growth plans there needs to be sufficient planning and funds available for future expansion. One of the key aspects of the E-commerce industry is their pricing of the products which induces the competition of aggressive pricing. Therefore, to keep the advantage business pre-fill the stock at a low price point when there is no demand.

This study focuses on predicting the sales for e-commerce business so that business can formulate accurate inventory plan which would help them avoid excess of stock available that is subject to wear and tear. Also, adequate stock can avoid a shortage of stocks when in demand. Profits can be increased as purchases can be made in advance at the low price point.

B. *Data Understanding*

Data understanding is an essential building block before beginning the model building and analysis. This step provides more knowledge, insights, interconnections, dependencies available within the data.

The dataset contains a total of 3 Years of daily sales data of each order placed on the platform. There around 50 product categories but for this research, the top 8 categories sales-wise were chosen to perform the analysis. The same approach can be later expanded for the remaining product categories. The platform accepts 4 types of payment modes. There are four different types of shipping classes namely Standard, Same Day, First and Second class.

Below figure 2, provides the view of the Shipping Status of an order. The majority of the orders have been delivered late. The reason for that is stock is made available after the order is placed on the portal thus the majority of the time is taken for order preparation and transit. Time taken for delivery is one of the key aspects of customer satisfaction. Late delivery Significantly affects customer satisfaction and brand reputation.

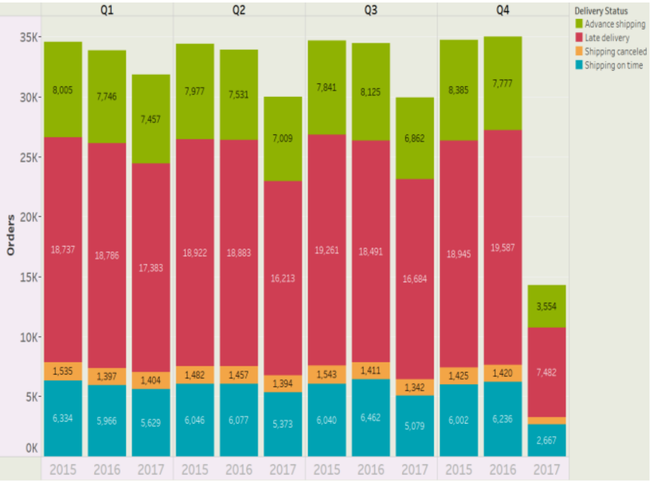


Figure 2: Order Count as per Delivery Status

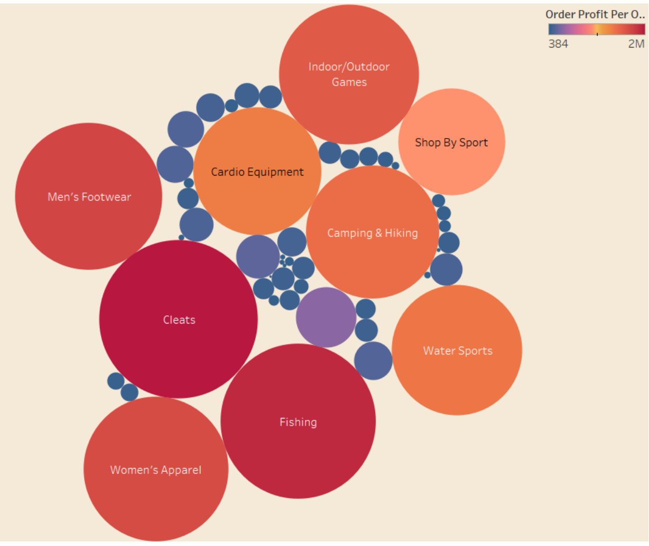


Figure 3: Product Category Wise Sales Distribution

From Figure 3, we can interpret that sales are not uniform across product categories rather it is concentrated with 8 Categories. Bigger circles indicate that the sales for that category are more compare to others and vice versa. Performing prediction on 50 different categories won't be feasible as we have to create separate time series for each category, thus we will be focusing on only 8 Categories and later on the model can be replicated for other categories.

C. *Data Preparation*

Data preparation is an important step which makes dataset ready for modeling task. It contains different subtasks.

Data Selection: As specified above prediction will be done on 8 product categories which are having the highest sales. Thus, data of these categories are extracted from the original dataset. 8 different time series have been constructed as per the product category.

Data Cleaning: The order Date and Sales column was having missing values which have been removed as date and sales would affect the trend and seasonality which is already existing in the time series.

Construct Data: As we have daily data of each order data has been aggregated month and year-wise. Therefore, each time series will have data of 36 months.

Format Data: The date column contains data in two formats dd-mm-yyyy and mm-dd-yyyy thus converted the latter one to the correct format so that it will easy for the model to interpret the period.

D. *Modeling*

1) *Decomposition of Time Series:*

The modeling stage begins with converting the Data frame into a time-series variable. After that next step is associated with the decomposition of Time Series into Trend

Seasonality and Randomness. This gives us insight into whether time series contains any of the above features and to what extent these features are present.

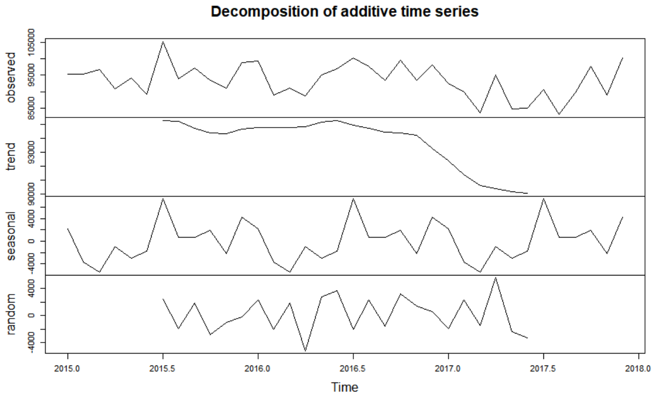


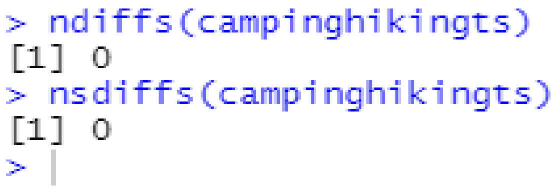
Figure 4: Decomposed Time Series for women’s Apparel

Here we have used Additive decomposition as sales are not increased in multiples within three years. From the above Figure 4, we can interpret that the time series is stationary but at the end, there is a downward slope. Time series exhibits significant seasonality as the same pattern is repeated three times. The Series also contains few random points as well. A similar analysis is done for the remaining 7 series as well.

2) NDIFF and NSDIFF :

Ndiff function performs unit root test to find out the number of differences need to make time-series stationary.

Nsdiff function performs seasonal unit root test get the number of seasonal differences to need to make time-series stationary.
If we get a value greater than 0, we perform the differentiation using the diff function. All the 8-time series are checked for ndiff and nsdiff and made differentiation if required.

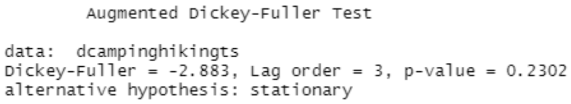


```
> ndiffs(campinghikingts)
[1] 0
> nsdiffs(campinghikingts)
[1] 0
> 
```

Figure 5: NDiff and NSdiff

3) ADF Test:

This is a hypothesis test used to determine whether the time series is stationary or not. A significant p-value indicates that the time series is stationary. The non-significant p-value is the indication of non-stationary time series.



```
        Augmented Dickey-Fuller Test

data:  dcampinghikingts
Dickey-Fuller = -2.883, Lag order = 3, p-value = 0.2302
alternative hypothesis: stationary
```

Figure 6: ADF Test for Women’s Apparel

4) ACF AND PACF

ACF and PACF plots are used to determine the value of P and Q in SARIMA. The presence of significant Autocorrelation and Partial Auto Correlation at initial lags and subsequent lags helps us identify the values of P and Q. After determining the value of P and Q we fit the model. In the next section, we will look at the results obtained from the model building.

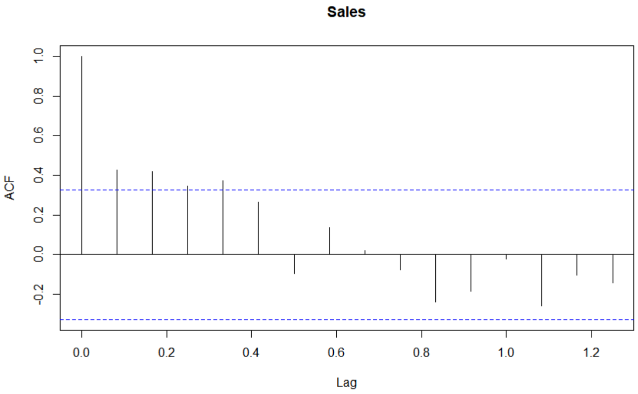


Figure 7: ACF Plot

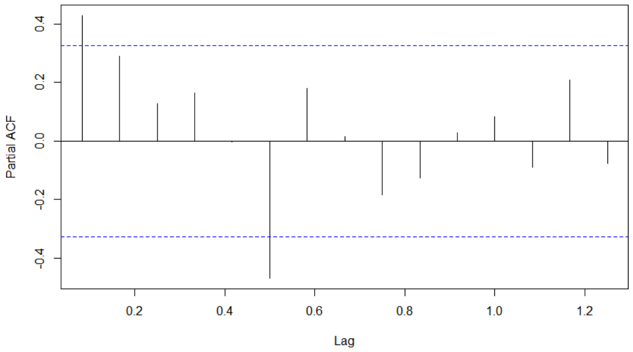


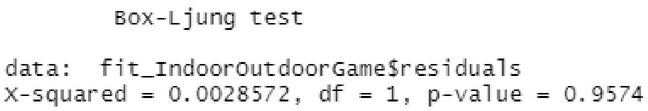
Figure 8: PACF Plot

5) Lung-Box Test:

This test is to verify whether the model has not left any pattern in residuals.

N_0 : Autocorrelations are not Significant and the Model fits the data well.

N_1 : There is significant Autocorrelation and the model doesn’t fit the data well.



```
        Box-Ljung test

data:  fit_IndoorOutdoorGame$residuals
X-squared = 0.0028572, df = 1, p-value = 0.9574
```

Figure 9: Box Ljung Test

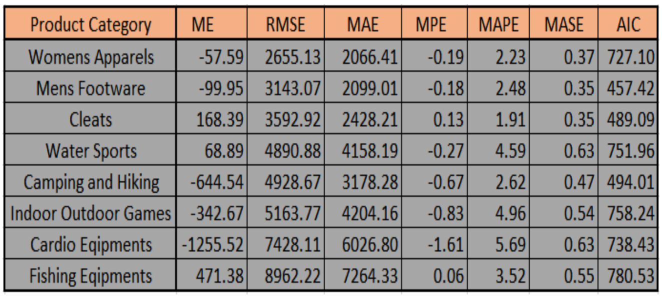
From Figure 9, we interpret that p-value 0.9574 is greater than 0.05 thus it confirms that autocorrelations don't differ from Zero and the model fits the data well.

IV. Evaluation

A. *Quantitative Analysis*

1) Evaluation Metrics:

Seasonal ARIMA models can be evaluated using multiple evaluation metrics such as RMSE, MAPE, ME, MAE, MPE MASE and AIC.



Product Category	ME	RMSE	MAE	MPE	MAPE	MASE	AIC
Womens Apparels	-57.59	2655.13	2066.41	-0.19	2.23	0.37	727.10
Mens Footware	-99.95	3143.07	2099.01	-0.18	2.48	0.35	457.42
Cleats	168.39	3592.92	2428.21	0.13	1.91	0.35	489.09
Water Sports	68.89	4890.88	4158.19	-0.27	4.59	0.63	751.96
Camping and Hiking	-644.54	4928.67	3178.28	-0.67	2.62	0.47	494.01
Indoor Outdoor Games	-342.67	5163.77	4204.16	-0.83	4.96	0.54	758.24
Cardio Eqipments	-1255.52	7428.11	6026.80	-1.61	5.69	0.63	738.43
Fishing Eqipments	471.38	8962.22	7264.33	0.06	3.52	0.55	780.53

Table 1: Evaluation Table for Product Categories

From Table 1, we can interpret that the lowest RMSE value obtained is for Women's Apparel. RMSE value ranges from 2655 to 8962. These are acceptable values for RMSE as monthly sales are in the 100 Thousand range. The ME is the mean error which is the average of all the error terms predicted for particular time series. Men's Footwear, Women's Apparel, Camping & Hiking, Indoor & Outdoor games and Cardio Equipment's are having a negative mean error which means that model has predicted lesser sales values more than actual sales. Mean Absolute Error (MAE) nullifies the effect of predictions that have sales lesser than actual sales. Thus, we get the absolute error of the model.

2) Predicted VS Actual Sales:

Now we will look at how the model has predicted the values for the sales of different product categories.

a) Cleats Apparel:

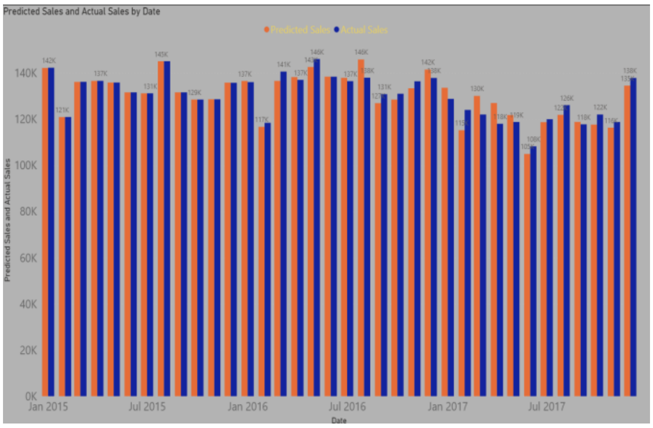


Figure 10: Predicted vs Actual Cleats Category

Figure 10, represents predicted sales from the ARIMA model and Actual sales in the form of Building Blocks. Looking at the building blocks we can interpret that model is accurately fitting the historical sales.

b) Fishing

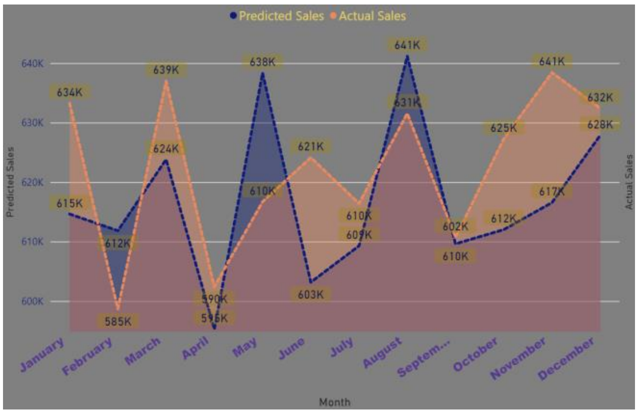


Figure 11: Actual vs Predicted Fishing Category

Figure 11, represents the sales aggregated for each month over 3 years plotted using Area Chart. Model is accurately predicting the month-wise sales values but when we combine the results for 3 years month-wise the accuracy is not matched.

c) Cardio Equipment:

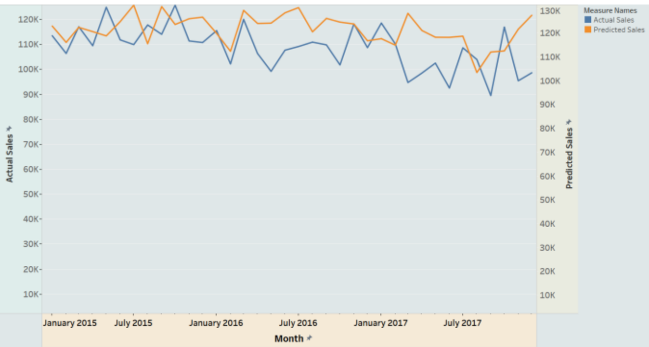


Figure 12: Actual vs Predicted Cardio Equipment Category

Figure 12, represents the Actual vs Predicted sales for Cardio Equipment's category using a Line chart. The model has able to match the trend and seasonality but has not achieved exact accuracy for this category.

d) Camping and Hiking

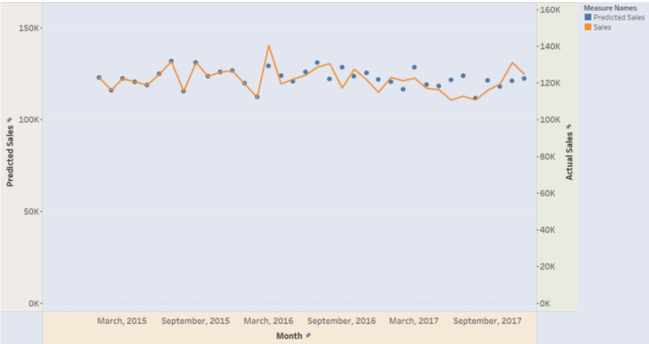


Figure 13: Actual vs Predicted Camping & Hiking Category

Figure 13, represents the actual vs predicted graph of the Camping and Hiking Product Category. Here we can see that initially for the around 1-year model has accurately predicted the sales value thus both the values are almost identical. Later the model exhibits an error due to which values are not matching on the graph.

e) Indoor Outdoor Games

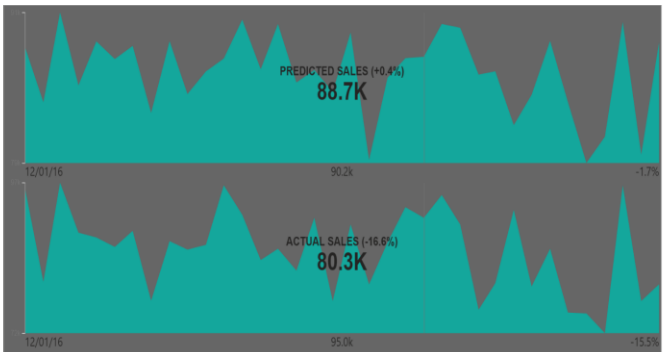


Figure 14: Actual Vs Predicted Indoor & Outdoor Games

Figure 14, represents two area charts with the KPI. As we move across the graph, we can see the predicted vs actual values along the x-axis, Month and Year at Left Corner. This chart also provides us percentage increase from the previous month at right corner.

f) Men's Footwear:

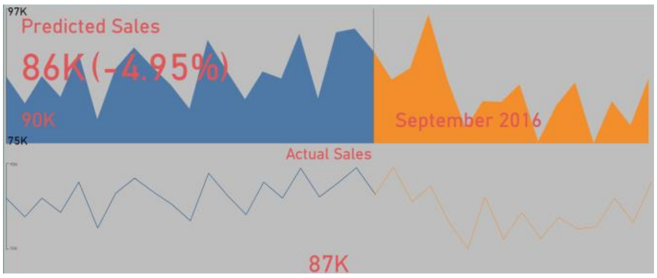


Figure 15: Actual Vs Predicted Men’s Foot ware

Figure 15, contains KPI, which represents the current value of the sales. As you move across the graph for a particular month it also provides percentage increase or decreases value between selected and current month. Along with this it also gives the comparison between predicted and actual sales.

g) Water Sports:

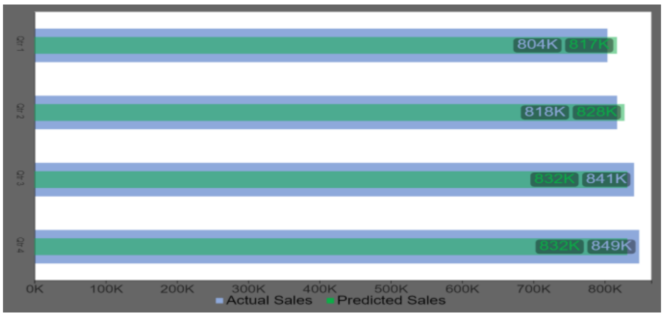


Figure 16: Actual vs Predicted Water Sports Category

Figure 16, is a stacked bar chart that represents sales values for 4 quarters of the water sports category. For the first quarter predicted sales are 13 k more than the actual sales. For second-quarter, there is a difference of 10 k between actual and predicted sales. For the last 2 quarters predicted sales are sales than actual sales.

B. *Qualitative Interpretation:*

1) Forecast for Camping & Hiking Products

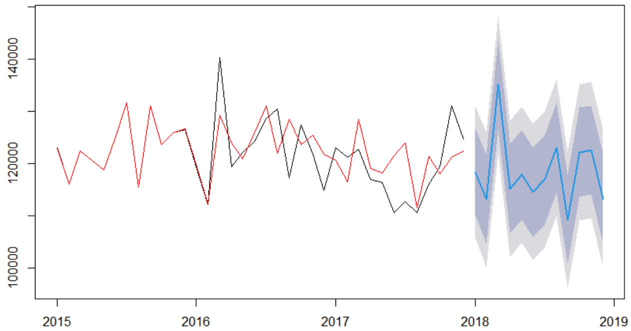


Figure 17: Forecast for Camping & Hiking Products

2) Forecast for Women’s Apparel:

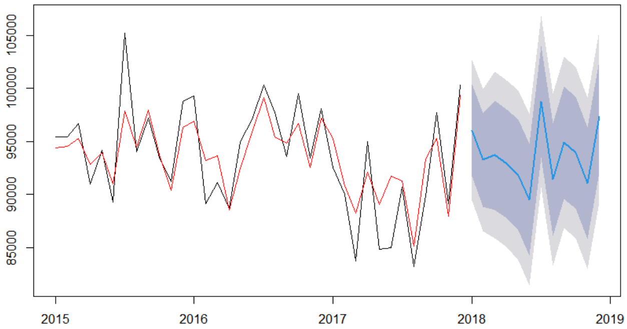


Figure 18: Forecast for Women Apparel Products

3) Forecast for Indoor Outdoor Games

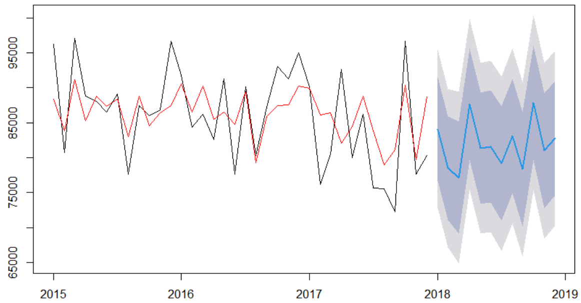


Figure 19: Forecast for Indoor Outdoor games

Figure 17,18,19 gives us a view of model forecasting for three different product categories. Even though these 3 categories are top sales accumulating their sales forecast is varied across 12 months period. For the first quarter Camping and Hiking, Products forecasted to be top-selling but later sales have been dipped down. The Indoor Outdoors games forecasted to have up and down sales across 12 months period. The Women’s Apparel category forecasted to gain average sales.

The above forecast is a key indicator for business to plan their inventory. The business can focus on buying inventories that are forecasted to be the highest selling in the first quarter. Also, they could buy inventory for the second quarter in advance so that they can buy that for a lower price point. Thus, making more profit than before. There is a scope to perform a comparative study that if the model has forecasted to gain fewer sales than the previous same period then analysis need to be done to find out the reasons for that.

V. Conclusion

This paper performed a prediction of sales for e-commerce product categories. The Seasonal ARIMA algorithm was found to be providing significant accuracy in prediction and forecasting sales for 1 year ahead. The data were first preprocessed to remove missing values after that data was split into a respective category. Data aggregation was performed to get the monthly sales data from daily orders. The Trend, Seasonality and Randomness were analyzed. Multiple tests were conducted to check whether Time Series is stationary, depending on the test result series was made stationary. Model prediction on historical sales value was visualized to check whether the model has accurately identified trends and seasonality. Model is evaluated using multiple evaluation metrics. The model achieved RMSE values within the range of 2655 to 8962. The model can be implemented for other product categories and other business domains as well.

References

[1] UNCTAD, "Global e-Commerce hits $25.6 trillion – latest UNCTAD estimates," UNITED NATIONS CONFERENCE ON TRADE AND DEVELOPMENT, 27 04 2020. [Online]. Available:

[2] https://unctad.org/press-material/global-e-commerce-hits-256-trillion-latest-unctad-estimates. [Accessed 25 03 2021].

[3] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics & Communications *Engineering (iCCECE)*, 2018, pp. 53-58, doi: 10.1109/iCCECOME.2018.8659115

[4] X. dairu and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), 2021, pp. 480-483, doi: 10.1109/ICCECE51280.2021.9342304.

[5] C. Lu, T. Lee and C. Lian, "Sales Forecasting of IT Products Using a Hybrid MARS and SVR Model," 2010 IEEE International Conference on Data Mining Workshops, 2010, pp. 593-599, doi: 10.1109/ICDMW.2010.11.

[6] S. N. Boyapati and R. Mummidi, “Predicting sales using Machine Learning Techniques,” p. 52.

[7] F. M. Thiesing, U. Middelberg and O. Vornberger, "Short term prediction of sales in supermarkets," Proceedings of ICNN'95 - International Conference on Neural Networks, 1995, pp. 1028-1031 vol.2, doi: 10.1109/ICNN.1995.487562.

[8] F. Adnan, P. Damayanti, G. W. Fajarianto and A. Cahya Prihandoko, "Winter Exponential Smoothing: Sales Forecasting on Purnama Jati Souvenirs Center," 2018 5th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2018, pp. 680-683, doi: 10.1109/EECSI.2018.8752767.

[9] T. Bowen, Z. Zhe and Z. Yulin, "Forecasting method of e-commerce cargo sales based on ARIMA-BP model*," 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2020, pp. 133-136, doi: 10.1109/ICAICA50127.2020.9181926.

[10] B. Singh, P. Kumar, N. Sharma and K. P. Sharma, "Sales Forecast for Amazon Sales with Time Series Modeling," 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T), 2020, pp. 38-43, doi: 10.1109/ICPC2T48082.2020.9071463.

[11] A. Krishna, A. V, A. Aich and C. Hegde, "Sales-forecasting of Retail Stores using Machine Learning Techniques," 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), 2018, pp. 160-166, doi: 10.1109/CSITSS.2018.8768765.

[12] R. G. Hiranya Pemathilake, S. P. Karunathilake, J. L. Achira Jeewaka Shamal and G. U. Ganegoda, "Sales Forecasting Based on AutoRegressive Integrated Moving Average and Recurrent Neural Network Hybrid Model," 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2018, pp. 27-33, doi: 10.1109/FSKD.2018.8686936.

[13] Z. A. Andriawan et al., "Prediction of Hotel Booking Cancellation using CRISP-DM," 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), 2020, pp. 1-6, doi: 10.1109/ICICoS51170.2020.9299011.