Journal of Physics: Conference Series

Market segmentation for profit maximization using machine learning algorithms

To cite this article: Sruthi Janardhanan and Raja Muthalagu 2020 *J. Phys.: Conf. Ser.* **1706** 012160

Market segmentation for profit maximization using machine learning algorithms

Sruthi Janardhanan and Raja Muthalagu

Department of Computer Science, Birla Institute of Technology and Science Pilani, Dubai Campus, Dubai, UAE

Abstract. In this present era with growing population, markets play an important role in providing the required and desired utilities to the people. For this it is important to enhance the Customer Relationship Management which can be achieved by segmenting the market utilities by various factors like weekly sales, demand and supply. In this research paper we discover the valuable information that is weekly sales and develop an efficient business strategy model to increase the profitability of the market through supply as well meet the demands of the customers. Our objectives are i) To find the top profitable products of the market across all the branches as well as their correlation with other products, ii) Understand the customer behavior according to the market flow and iii) Forecast the sales using ARIMA (Auto Regressive Moving Average). In this project, as we lack proper information about the customer identity, we use k means clustering which is an unsupervised learning model to cluster the customers according to the weekly sales behavior. Since we concentrate on the Weekly Sales to understand the market behavior, the best method to be applied on the dataset is the moving average technique. Here, we use ARIMA model to get the best results for forecasting as our time series data is considered to be stationary.

1. Introduction

In today's growing world, it is important for the market to implement various business strategies to profitably earn as well to satisfy the customer for a long run. This doesn't happen with just recording the customer's data or the sales data. A very detailed analysis is required to understand the behavior of their customers as well the demand of their products in order to gain profit. Currently we have a lot of visualizing tools that includes various libraries in python and business intelligent tools like Power BI to analyze the market with the historical data after which we apply various predictive algorithms to forecast the future of the market. Data analysis techniques have become a vital tool in the market industry to obtain patterns from the available data and predict the future of the market and make profitable decisions accordingly.

In this design project, I have worked on market segmentation using historical data that contains various features like Weekly sales, unemployment, fuel price, holiday and customer price index (CPI). On the first part of my design project I have worked on analyzing sales data of 45 stores eachcontaining 81 products labeled with 1 to 99. It is important for markets or retail stores that have large number of products to understand the fact that picking up profitable products in terms of demand, correlation with other products and net worth is something very vital.

Since the dataset lacks customer identity attribute, I have used the weekly sales of profitable products and performed clustering using unsupervised k-means algorithm to find the group of customers with similarity in their purchasing behavior. By doing this the market can concentrate on every segment of products for customer satisfaction and in turn flourish the market flow.

Forecasting and predicting the future of the market is another important step for any kind of business in order to understand the nature of demand and flow of the products in terms of sales. Here, I have used ARIMA model for forecasting the future sales.

2. Literature Review

The Instacart grocery dataset to perform customer segmentation using various algorithms like k means, hierarchical and density based algorithms is present in [1]. It is concluded that K means clustering provide the highest accuracy in order to segment the customer. The customer segmentation based on feature with datasets that contains 200 tuples with 2 features is proposed in [2]. The k means clustering, agglomerative clustering and mean shift algorithm are used. Also, they have considered two internal clustering measure namely, silhouette score and Calinski-Harabasz index.

The consumer behavior using data mining techniques are proposed [3]. They have analyzed the importance of various data mining techniques to recognize and derive the pattern of the market and its consumer. Various marketing strategies have been developed with market basket analysis through association rule mining. Minimum support and minimum confidence have been calculated to find association rules.

The clustering technique for customer segmentation is proposed in [4]. It showed that various clustering algorithms results in varying cluster outputs and thus they have compared their performance. For a good clustering algorithm, within the cluster, customers should behave more alike when compared to customers in the other clusters, thus performance assessment criteria were selected accordingly. It also shown that the developed clustering algorithm produces good results when compared to single link, k-means and complete link algorithms.

The implementation of three clustering algorithms that are k-means, advanced k-means and agglomerative clustering is proposed in [5]. From the experimental results, it can be observed that agglomerative clustering is not feasible because of its long execution time. The advanced k-means lessens the overall running time by 27.8% when compared with standard k-means and 97.8% when compared to agglomerative clustering. Also, in terms of clustering quality and speed it performs better than standard k-means. Thus, they have concluded that for customer segmentation with respect to RFM model, advanced k-means clustering is more efficient and feasible.

A model to segment the customers using data mining technique called as clustering is proposed in [6]. In clustering, they have used k-means clustering to cluster the customers based on the given data which includes 82,648 transactions. An RFM model was built based on the transactions, which resulted in 102 customers who were further clustered into 2 clusters. With this the market strategy could be improved by understanding the customers' behavior in each cluster. An innovative approach of combining RFM and ABC analysis, and data mining techniques for clustering the customers for segmentation is proposed in

[7]. An investment criterion is applied on each cluster of customers namely, Best, Good, Average and others. For best customers, to enhance the relationship we can invest more on this cluster which in turn will maximize the profit of the market. Thus, by segmenting the customers, the market can focus on each cluster for maximizing the profit of the market.

A forecast sale of the truck component is also proposed in [8]. They have studied various machine learning algorithms like Support Vector Regression, Random Forest, Ridge Regression and Gradient Boosting. It was found that Ridge Regression performed well because of regularization approach followed by SVM in order to forecast. Gradient Boosting produced bad results due to over fitting performance as well as weak tuning. Various machine learning models to predict the sales using Extra Tree, Random Forest, ARIMA, Lasso and Neural Network is proposed in [9]. It was found that extra tree performed well compared to the other in terms of validation error. Sales forecasting is said to be a regression problem rather than a time series problem.

A sales forecasting is being conceptually performed in this research paper using various intelligent machine learning models namely, Gradient Boosted Trees, Decision Trees and Generalized Linear Model [10]. A dataset consisting of 85,000 records was used to perform the forecasting analysis. It was observed that Gradient Boosted Trees performed better than the rest giving an accuracy rate of 98%.

3. Proposed Methodology

3.1. Data Collection:

The dataset used for this project consists of weekly sales data of 45 branches of a store each consisting of 80 to 90 products named as dept. The dataset also consists of attributes Date and Is holiday which specifies if that particular date is holiday or not. The dataset is very huge, making it useful to analyze the market at the same time they lack customer attribute like customer id. Thus, in this work we make sure to predict the behavior of customers with the given data.

3.2. Visualizing the dataset:

Before we analyze the dataset, it is important to visualize the dataset since to understand the relationship between various attributes which will be useful to analyze the data accordingly. I have used Power BI and python code to visualize the dataset. Figure 1 is a pie chart visualization using Power BI and Figure 2 is a time series plot using matplotlib library.

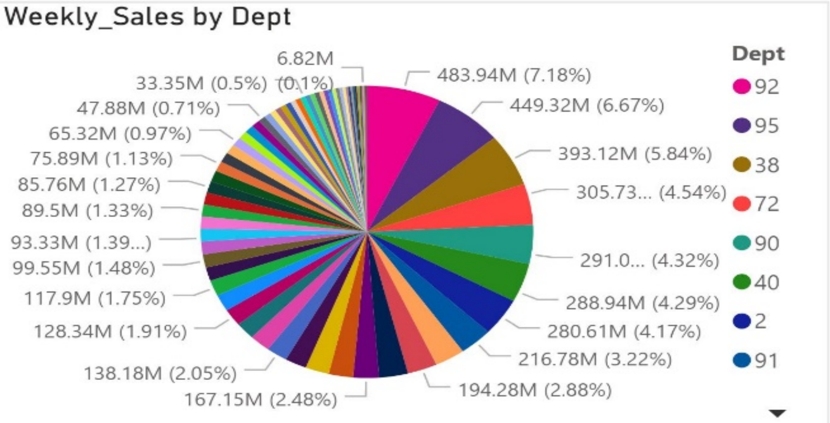

Figure 1. Visualization using Power BI

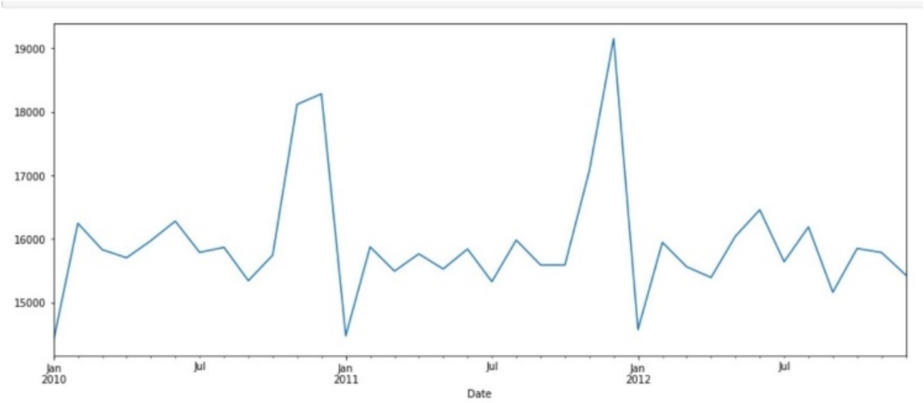

Figure 2. Time series plot

3.3 Obtain profitable products and most correlated products.

We obtain top profitable products according to the weekly sales. We obtain products whose weekly sales is greater than 30 million across the 45 branches. This shows that these are the most moving products in this market as well as it shows their demand accordingly. To understand the correlation between these products we apply correlation algorithm, which infers how correlated these top products are with one another in regards with the sales.

We filter top 11 products in regards with both the factors and the results can be seen in Table 1. We perform the elbow method in order to get the optimum number of clusters for performing k means algorithm. We then apply k means algorithm in order to cluster the products. From these clusters we can understand the customer behavior. Table 2 shows the number of products in every cluster. From Figure 3 which was obtained using Power BI model shows how much each product is purchased by customers belonging to each cluster. Each segment of the bar graph depicts the sales of a product which in turn shows the behavior of the customer with respect to the product.

For example, product 11 in Blue color is purchased more by customers belonging to cluster 4 rather than people in cluster 3. We can also infer that customers belonging to cluster 3 spend less than customers belonging to customer 4. We can also infer which product is having more value but bought very much less in number. For example, from figure 3 product 10 is bought less in number but the value of the product is high. Since the dataset has no records of the name of the product, we can assume it as some branded product which costs more.

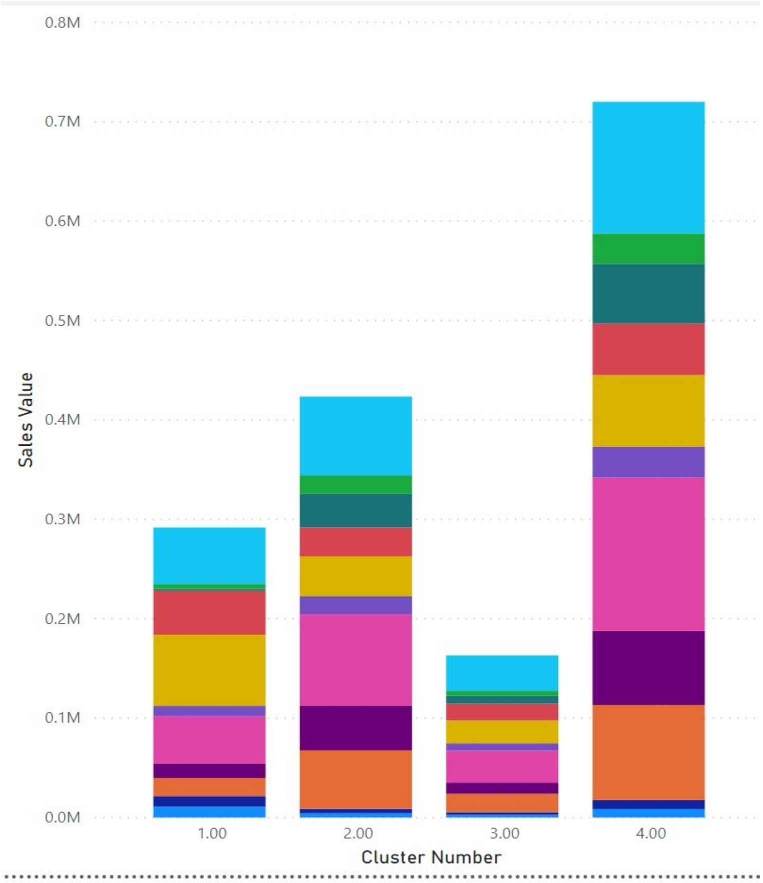

Figure 3. Bar Graph shows the behavior of customers that is depicted in terms of clusters.

3.4. Forecasting using ARIMA model

We forecast the future sales using ARIMA model in order to analyze the time series character of the data as well as to predict the future sales.

ARIMA is an autoregressive Integrated moving average which helps to perform the above. Auto Regressive is known for the lags of differenced series which is stationary and Moving Average is the average of lagged errors obtained from the data that is to be forecasted. I have used R programming to build the ARIMA model using Power BI tool. Figure 4 is the generated ARIMA model that is used to forecast the sales of next 52 weeks using 3 years weekly sales data which is considered to be stationary.

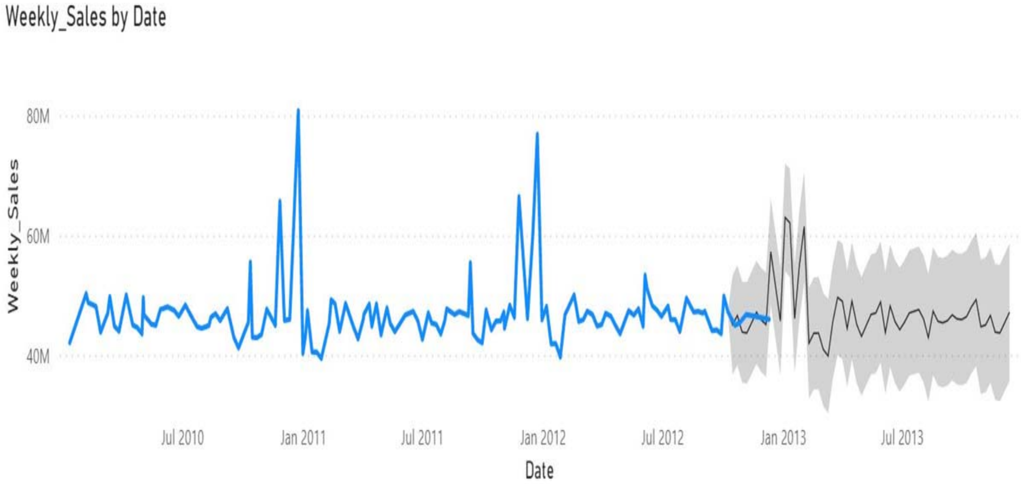

Figure 4. forecast of the sales for next 52 weeks by analyzing 3 years of data.

4. Results and Discussion

Applying the above methodology we have understood that market data is always very huge and analyzing it using various data mining mining algorithms for maximizing the profit is something very vital.

The dataset that we have used here consists of weekly sales and product number which is named as dept. 91 different products across 45 branches. In this research we make sure to predict the behavior of customers with the given data and from the above analysis we have understood that k means can be used for analyzing the customer behavior according to the market.

K means clustering is useful since it can work well with distinct points as well as large dataset. It is applied on the top profitable as well as correlated products which helps us to maximize the market's profit. Table 1 and Table 2shows the results of the above analysis. Since our dataset is huge and lacks the customer information, k means helps us to analyze accordingly.

Table 1. Correlation results of profitable products

Product x	Product y	correlation
24	33	0.974936
90	92	0.971726
91	92	0.965310
81	92	0.959824
2	13	0.956589
93	97	0.953253
81	90	0.948514
92	95	0.943919
81	91	0.941699

Table 2. K Means clustering results

Cluster	**Count of Products**
1	2479
2	1909
3	1231
4	816

Further to forecast the future sales of the market we use ARIMA model. It is very vital for every retail market to forecast the future in order to meet the demand of the customers with which the supply of the product can be enhanced and the profit of the market can be maximized accordingly. In this research, the dataset consisted of 3 years of sales data which was sufficient to obtain accurate results with ARIMA. Figure 4 is the generated ARIMA model.

ARIMA model was chosen since the data showed a stationary behavior and univariate feature (i.e. Weekly sales) was being used for forecasting. It is seen from the research that ARIMA model works well on the data with single feature into consideration. The seasonality factor was considered to be TRUE since it takes into account the previous series of the sales.

5. Conclusion

With the growing population and demand, it is important for the markets to extract patterns of their products' sales and understand their customer's behavior to profitably run their store at the same time meet their demands. Data analysis techniques have become a vital tool in the market industry to obtain patterns from the available data and predict the future of the market and make profitable decisions accordingly. It is also important to forecast the future of the market with the obtained data in order to maximize their profit. I have used k means clustering model for clustering the dataset and obtaining various inference with reference to the sales of every profitable product. ARIMA model was used to forecast the sales of all the products as whole. This approach helps the market to maximize their profit.

References

[1] Kalyani Bhade, Vedanti Gulalkari, Nidhi Harwani, Sudhir N. Dhage *"A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization"* 9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru, India.

[2] Tushar Kansal, Suraj Bahuguna , Vishal Singh, Tanupriya Choudhury, *"Customer Segmentation using K-means Clustering"* University of Petroleum & Energy Studies (UPES),Dept. of Informatics, School of Computer Science, Dehradun, IEEE 2018.

[3] Abhijit Raorane, R.V.Kulkarni *"Data Mining Techniques" A source for Consumer Behavior Analysis."*International Journal of Database Management Systems Vol.3, No.3, August 2011.

[4] Prabha Dhandayudam, Dr. Ilango Krishnamurthi, *"An Improved Clustering Algorithm for Customer Segmentation"* International Journal of Engineering Science and Technology (IJEST) vol. 4 no.02 February 2012.

[5] Sabbir Hossain Shihab, Shyla Afroge, Sadia Zaman Mishu*, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering:A Comparative Study"*. International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019.

[6] Ina Maryani, Dwiza Riana, Rachmawati Darma Astuti *"Customer Segmentation based on RFM model and Clustering Techniques with K-Means Algorithm"*. STMIK Nusa Mandiri Jakarta.

[7] Jan Panuš, Hana Jonášová, Kateřina Kantorová, Martina Doležalová, Kateřina Horáčková, *“Customer segmentation utilization for differentiated approach”*. The International Conference on Information and Digital Technologies 2016

[8] Venishetty Sai Vineeth, *“Machine Learning Approach for Forecasting the Sales of Truck Components”*. Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

[9] Bohdan M. Pavlyshenko, *“Machine-Learning Models for Sales Time Series Forecasting”*. IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, 21–25 August 2018.

[10] Sunitha Cheriyan,*” Intelligent Sales Prediction Using Machine Learning Techniques”*. IT Department Higher College of Technology, ResearchGate 27 November 2019.