Hochschule Hof
University of Applied Sciences

Hof Large eight (2023), 1–10

ARTICLE

Sales Analysis and Prediction

Foning Keubou Loique Steve[1]

[1]Applied Research in Computer Science, Hof, 95028, Bavaria, Germany
[2]Email, loique.keubou.foning@hof-university.de
[3]Supervisor, Prof. Dr. Michael Spangenberg

Abstract

In the world, there is constant growth in shopping abilities which makes sales management a very exigent work in all shopping stands. Managers or store owners face really high challenges in order to ensure the satisfaction of each customer; hence, customer requirements have to be updated monthly in order to stay in the stand of needs.

In this work, we propose a data analytic-base approach with proposed Machine learning models to predict the sales performance based on 2019 data in which we'll get a prediction on the next quantity of products to be ordered and the sales to be done. This analysis will perform a series of tasks such as:

Identifying the months with the best sales, Identifying the months with the highest earns, Identifying the grouped sales for each purchase, Identifying the store with the best sale, and identifying the time which will be good for advertisement. Answering these questions, our work will be essential in a business in order to increase profit and facilitate decision-making.

Keywords: Online store, Data mining, Machine learning, sales, forecast, prediction

Introduction

The development of computer networks and more specifically the Internet has revolutionized habits and ways of doing things in several sectors of activities, including commerce.

Today, the reliability and success of electronic commerce or e-commerce are no longer at stake prove and thanks to the multiple advantages that this system presents (increased visibility, the disappearance of borders, reduced costs, automation of the calculation of the return on investment, possibility to offer many more products, etc.), it turns out to be an essential option for all traders wishing to make his business as profitable as possible. However, classic e-commerce has some limitations.

For example, in the Study of Selling Behavior of Salesperson, we could observe that individual sales productivity is too low as the employers are unable to know how much product may be needed for the sales of a specific month. Sometimes the team members experience sales anxiety as they are unable to set clear goals in order to target customers, which leads to an absence of ads at the right time in order to ensure customers get to know the new product or discount that may exist for a specific item they cherish. Also, the team members are unable to identify which product was mostly sold in case the company has enlarged the market to multiple stores all over the world, hence incapable of identifying which month, which city, which store has the highest sales of a specific product or to know which of them needs an increase of quantity produced. This turns to have little or too much productivity for a specific sale or item, wastage of resources, and inefficiency of the team members.

This situation does not generally occur in the context of physical commerce because we have salespeople who closely follow the actions of the customer and can then make proposals and recommendations to them.

At a time when Machine learning and artificial intelligence pre-feel many applications and encouraging results, it would be interesting to analyze their usefulness in correcting this deficiency. We are interested in this work on the application of automatic learning techniques by the computer (Data mining/ Analysis and Machine Learning) in the design of a system for detecting user future purchase intention on platforms and mostly online sales.

This system will allow e-merchants to better detect potential customer needs and thus have the possibility of offering them personalized promotions, with correctly estimated stocks which will lead to a higher purchase rate and better returns on investment.

Working Plan/Methodology of Work

The research methodology we used was made up of three steps:

1. `Existing Works`
2. `Data exploration and Analysis`

3. `Drawing out a specification book`
4. `Proposed Model for Prediction`
5. `Analysis Architecture`

This permitted us to identify and understand the aim of the work, identify the population concerned by this, and also permitted us to come across the related works that were done in relation to sales predictions. Some of which are:

Existing Works:

- An Approach of Sales Prediction of Customers Using Data Analytic Techniques.

  This work proposed a data analytic-based approach for predicting the sales details based on the last 12 months' data and it will generate a report. This system performs various tasks like finding the best months for sales, calculating monthly earned money from different products, which city sold the maximum products, and when will be the best time for an advertisement to maximum customers to buy the products also generates reports, which are items that are frequently bought by the customer to keep stock for upcoming months and, increase the profit of the shopping mall business.

- Online Ad-Sales Analysis and Understanding Customer Behavior Towards Online Ads.

  In this work, they considered how important ads have become for revenue increment in different companies. Hence, using the Analysis of Variance (ANOVA) method, they analyse online ads sales revenue generated by Meta and Google. And also analyze how different groups of customers respond to online ads in order to draw significant inferences and conclusions.

Data exploration and Analysis :

Thanks to the Kaggle platform, we could get a sample data set in the form of CSV files. With this collected data, we will start with the exploration of the data, getting to know how the data looks like, what the needed fields for analysis, getting rid of useless rows and columns, and identifying if we need to create additional columns/rows before jumping into the data analysis.

The Specification book :

In order to facilitate our analysis process, we wrote down a specification book and thanks to this, inspired by research work of (Improving Sales Analysis written by Ivan Kononov in 2021Best Selling Product and Category Prediction Using Sales Analysis written by Ms. Archana NikoseTejal, MungaleMinal and Shelke in 2022, and the book Electric Vehicle Sales Analysis Model Based on User Purchase Intention Analysis written by Ying Zhang, Yibing Chen, and Peisen Huang in 2022)this books permitted us to identify some pertinent questions in our research such as:

- **What was the best month for sales and how much was earned that month?**
- **What city sold the most products?**
- **What time should we display advertisements to maximize the like-hood of customer's buying product** ?
- **What products are most often sold together?**
- **What Product Sold the most? Why do you think it sold the most?**

Maybe adding these questions will allow us to answer our main question which is: **how can managers improve their sales?**. Also, these series of questions permitted us to elaborate and test the hypothesis:

- **Hypothesis:** Most of our sold items are cheap items!

In a more explicit way, the data obtained from Kaggle came in the form of multiple CSV files of the year 2019 in monthly files. The first task was to combine these files into a single file from which we could draw out one data frame. From this data frame, we performed a series of operations in order to get a general overview of how the data looks like, operations in order to clean the data or add required files for the analysis, and also in order to prepare the data through the answering of the analysis questions elaborated in our specification, then finally we used the obtained results in these tasks to test for the feasibility of the elaborated hypothesis. The next step was to find the right algorithm that could draw out the models that would predict the sales and the quantities to be ordered for the years 2020 and 2021.

Figure 1 below will clearly demonstrate the methodology we used in a detailed way

For the model used, we had :

1. `Linear Regression Model`
2. `Polynomial Linear Regression Model`

Linear Regression Model:
Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatter plot can be a helpful tool in determining the strength of the relationship between two variables.

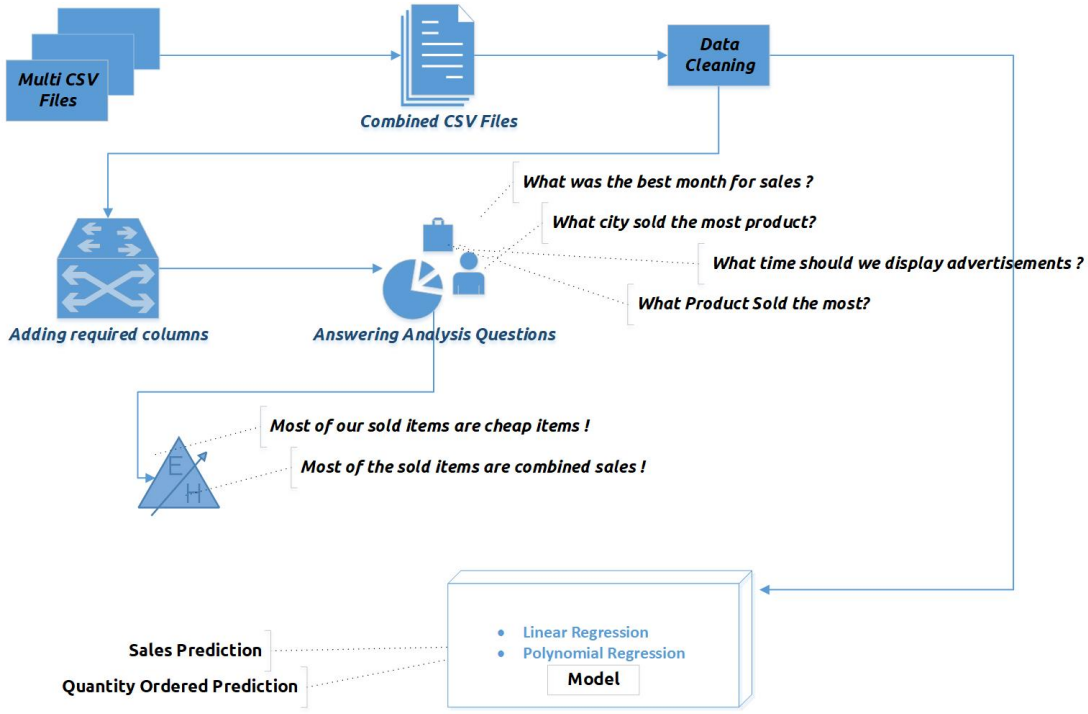

Figure 1. Working Schema / Methodology used

If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

A linear regression line has an equation of the form

Y = a + bX,

where **X** is the explanatory variable and **Y** is the dependent variable. The slope of the line is **b**, and a is the intercept (the value of **y** when **x** = 0).

Polynomial Linear Regression Model:

Polynomial regression is often considered a special multiple linear regression. It is a statistical method of determining the relationship between an independent variable (x) and a dependent variable (y) and modeling their relationship as the nth-degree polynomial.

The relationship between the independent and dependent variable on a graph turns out as a curvy-linear relationship with the help of a polynomial equation. Polynomial regression is used when there is no linear correlation between the variables. Hence, it explains why it looks more like a non-linear function.

There exist three forms of polynomial equations which all work out with some mathematical assumptions:

- **Assumption 1**
  The behavior of a dependent variable is explained by a linear, or curvy-linear, the additive relationship between the dependent variable and a set of k independent variables **(xi, i=1 to k)**
- **Assumption 2**
  The relationship between the dependent variable and any independent variable is linear or curvy-linear.
- **Assumption 3**
  The independent variables do not depend on each other too.
- **Assumption 4**
  The errors are independent and normally distributed with mean zero and constant variance.

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$Y = b_0 + b_1 x_1 + b_2 x_1^2 + b_2 x_1^3 + ...b_n x_1^n$$

There exist three types of polynomials:

Linear

$$ax + b = 0$$

Quadratic

$$ax^2 + bx + c = 0$$

Cubic

$$ax^3 + bx^2 + cx + d = 0$$

As you can see, the linear polynomial has a degree of 1, the quadratic polynomial has a degree of 2 and the cubic polynomial has a degree of 3. As the degree of the polynomial equations goes up, the curve better fits the data set.

When we have a data set and we plot it on a graph, there has to be a straight line where the scatter plots lie. But what if we have a data set that gives us no straight but a curve? This is when polynomial regression comes in.

The difference between linear regression and polynomial regression is that the line of best fit is a curve in polynomial regression. The scatter plots are scanned for a pattern and the line is drawn (curve) following that pattern of the points. Another difference is, polynomial regression does not make it compulsory for the data to have a linear relationship between them.

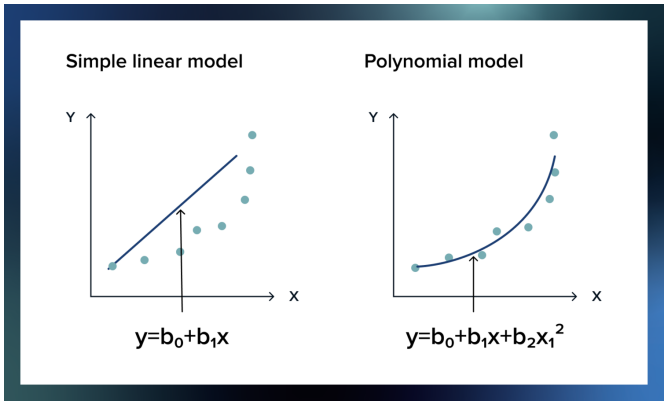

Figure 2. Linear Regression vs Polynomial Regression

So, as observed in fig2. when linear regression fails to determine a linear relationship between variables, polynomial regression does it for us.

Analysis and Results

1. Cleaning the Data and Adding Additional Columns

The data set Fig.3 contains fields like Order ID, Product, Quantity Ordered, Price Each, Order Date, and Purchase Address. The combined data contains hundreds of thousands of electronics store purchases broken down by month, product type, cost, purchase address, etc.

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
0	176558	USB-C Charging Cable	2	11.95	04/19/19 08:46	917 1st St, Dallas, TX 75001
1	NaN	NaN	NaN	NaN	NaN	NaN
2	176559	Bose SoundSport Headphones	1	99.99	04/07/19 22:30	682 Chestnut St, Boston, MA 02215
3	176560	Google Phone	1	600	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001
4	176560	Wired Headphones	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90001

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month
519	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
1149	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
1155	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
2878	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or
2893	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Or

Figure 3. Sorting and ignoring unwanted records

As the first step, the combined data set was filtered in order to identify the NAN values and also to identify error entries that contained OR in the recorded fields. This was done because all the missing values always reduce the accuracy of the final expected results. These errors were all filtered and ignored in the analysis using the code in Fig.4:

```
nan_df = all_data[all_data.isna().any(axis = 1)]
nan_df.head()

all_data = all_data.dropna(how = 'all')
all_data.head()
```

Figure 4. code used to ignore unwanted records

The second step is type conversion, which converts the data either categorical to numerical or numerical to categorical using astype method (). For the expected prediction calculation we need to understand, which are columns having major impacts on the predictions. So both concerned columns should have the right data types in order to perform an operation or apply a formula.

The third step was to augment data with an additional field.
The first data added was the month column where we had to extract the months from the Date-Time format of the order date column. This was done with the .str() method and then we created a column called a month in order to store each extracted date with respect to the order ID.

This was done the same for the hour and minutes found in this Date-Time format. Both the minutes and the hours were extracted and stored in respective columns in order to facilitate the analysis. A short view of the code is seen below:

The second field we added was Sales. In order to come out with this column, we had to determine the revenue from sales (gross revenue for a manufacturing unit) using the following steps:

- Firstly, we had to determine the number of units sold

```
#As the Order date column is in string, we'll first convert that to dateime formate
all_data['Order Date'] = pd.to_datetime(all_data['Order Date'])

#Creating columns of the hours, minutes
all_data['Hour'] = all_data['Order Date'].dt.hour
all_data['Minutes'] = all_data['Order Date'].dt.minute
all_data['Count'] = 1
all_data.head(30)
```

Figure 5. Columns added in the data frame

during a specific period, say for each order.

- Now, since the number of units produced is driven by demand, which forms the basis of the function for the price, we had to assess the average sales price per unit.
- Finally, the revenue is calculated by multiplying the number of units sold (step 1) and the average sales price per unit (step 2).

Hence, in fig.5 we obtained this derived formula:

Net Sale = Number of Units X Unit Price

The third field we added was the City column. In the Purchase Address column, we use the split() method and took the first member of the list that is in between the (,). The first item to the left of the first (,) is a space. This value is then stored in the newly created column (City). Fig. 5 shows an overview of these added columns in our data frame.

	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address	Month	Sales	City	Hour	Minutes
0	176558	USB-C Charging Cable	2	11.95	2019-04-19 08:46:00	917 1st St, Dallas, TX 75001	4	23.90	Dallas (TX)	8	46
2	176559	Bose SoundSport Headphones	1	99.99	2019-04-07 22:30:00	682 Chestnut St, Boston, MA 02215	4	99.99	Boston (MA)	22	30
3	176560	Google Phone	1	600.00	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	600.00	Los Angeles (CA)	14	38
4	176560	Wired Headphones	1	11.99	2019-04-12 14:38:00	669 Spruce St, Los Angeles, CA 90001	4	11.99	Los Angeles (CA)	14	38
5	176561	Wired Headphones	1	11.99	2019-04-30 09:27:00	333 8th St, Los Angeles, CA 90001	4	11.99	Los Angeles (CA)	9	27
6	176562	USB-C Charging Cable	1	11.95	2019-04-29 13:03:00	381 Wilson St, San Francisco, CA 94016	4	11.95	San Francisco (CA)	13	3
7	176563	Bose SoundSport Headphones	1	99.99	2019-04-02 07:46:00	668 Center St, Seattle, WA 98101	4	99.99	Seattle (WA)	7	46
8	176564	USB-C Charging Cable	1	11.95	2019-04-12 10:58:00	790 Ridge St, Atlanta, GA 30301	4	11.95	Atlanta (GA)	10	58

Figure 6. Columns added in the data frame

2. Answering Questions in the Specification Book

As explained in the methodology section, we laid down a specification book made up of questions and hypothesis which was to be answered and checked throughout the process of analysis. The first question we tagged in our analysis was:

1. `What was the best month of sales ?`

This was in coherence to identify how much was earned each month. Fig. 7 below shows the extracted data of monthly sales of products from our combined data set.

To perform this plotting, the original data set having the order date attribute from each sale per month is been extracted using string functions. This same string data is converted to an integer datatype with the astype() function. Using the data pre-processing method we eliminated all missing values from our data set. Finally, we represented this as a bar plot where the Y-axis is the Sales of the products in CAD and the X-axis shows the months.

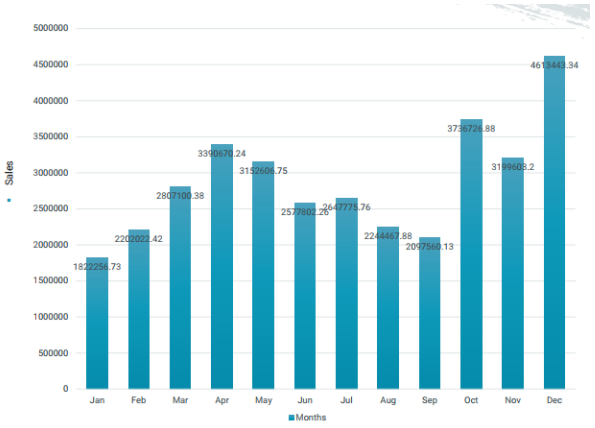

Figure 7. Best month of Sales

As the graphical representation shows above, every year begins with very few sales which gradually increase in May and get highest in the month of December. This is probably because, at the beginning of the year, people turn to saving and reducing expenses and the previous end of the year had lots of spending. During the month of December and May, people have a high tendency to spend more because of the Easter holidays around March to May and the Christmas holidays.

2. `What City Sold the Most ?`

Fig. 8 below shows the sales of products per city where the Y-axis represents the product sold in USD and the X-axis shows the city name. On this city, the name has appended the state where the city is found as they are similar city names in each state. We should note that the city column was not available in the initial data set and was created as explained in the sections above. In the first step the column 'Quantity ordered' and 'Price Each' is converted into integers using the "to numeric" function in python, then sales are calculated based on our generated formula obtained in the sections above.

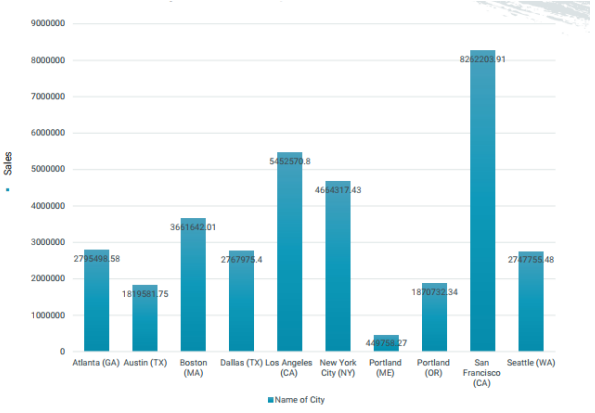

Figure 8. City with highest Sales

We observe extremely high sales for the cities of San Francisco (CA), Los Angeles (CA), and New York City (NY). This is probably because these are very big cities in the USA and people have little time to go shopping on-site and prefer to order online and also because they have enough money due to high living standards in big cities.

3. `What time should we display advertisements?`

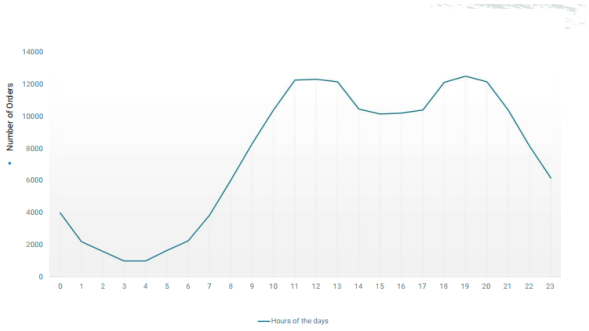

Figure 9. City with highest Sales

Fig.9 Illustrates the peak time when staff members should base themselves in order to prepare their advertisements and tag the maximum number of customers based on the sales sets of data. The X-axis represents the hours at which parches are made and the Y-axis shows the number of orders made. Both data are extractions of the Order Date Column as explained in the section of adding date field.

We observe peek time at around 10 am to 1 pm which is due to the fact that people have the tendency to buy on the site during their break time, which is also a good time to send out ads that could help increase sales. This purchase peek time is also observed around 6 pm to 9 pm, so we assume that at that time people are comfortably resting home and relaxing while making some purchases online, offering recommendation ads may be the right time to attract more customers and also increase sales.

4. `What products are most often sold together?`

Fig.10 Illustrates the grouped items which can be used to recommend products after a customer puts a selected product in his online basket. This is also in coherence with the advertisement recommendations mentioned in the previous question as the team member can also tag the right time to perform these ads.

The Y-axis represents the quantity ordered and the X-axis represents the Items. These items are actually grouped products that were purchased with the same order ID. This was then arranged using the "most common()" method and the sorted order IDs were then counted and classified in order to have the number of items with the same IDs. A formalized data frame was then created having the columns 'Items',' Quantities Ordered', and' Products'. We then split this into small chunks of 50 and obtained 200 chunks of grouped items. On fig.10 you can observe the first 50 chunks.

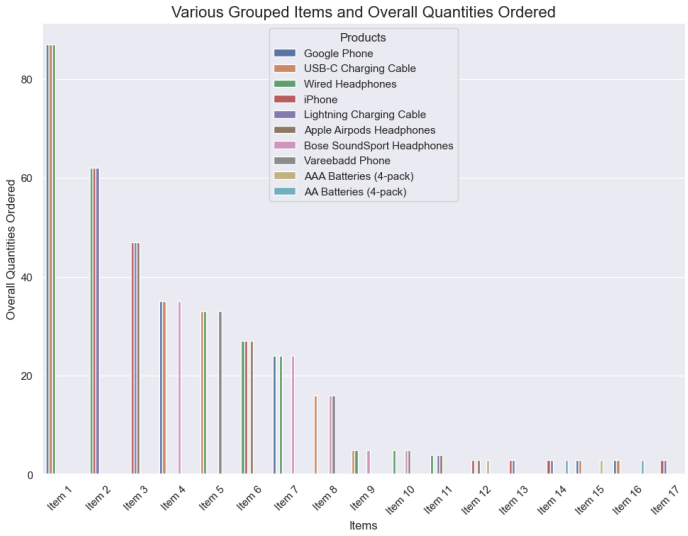

Figure 10. First 50 grouped orders

From here we observe that the first three grouped items are Google Phone, USB-C Charging Cable, and Wired Headphones. So, we assume that if a customer puts in his basket a Google Phone, the recommended ads just below his basket list may be USB-C Charging cables and Wired Headphones in order to increase sales.

5. `What Product Sold the most?`

Fig.11 illustrates the respective count of individual products in the set of data. The X-axis shows the name of each product and the Y-axis shows the customer's ordered quantity. This is to understand why certain products are sold more than others in order to find strategies for solving the problem and push customers to have interest in less sold products.
From here, we could Identify that the most sold items are AA Batteries (4 packs), and AAA Batteries (4 packs). So, we assume that these are batteries for children's games which means more children use the platform for purchase than adolescents who are concerned with the other highly sold products like Lightening Charging Cable and Wired headphones.

From these presumptions, we could Elaborate and test our hypothesis:

- **Hypothesis 1**

  `Most of the sold Items are cheap Items`

This can be proven by overlaying the graph above with the actual average price of the items in order to check for

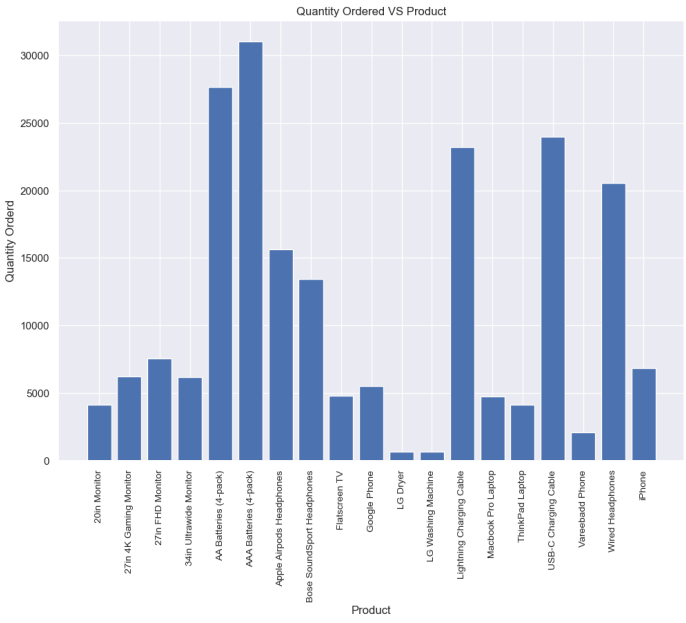

Figure 11. First 50 grouped orders

correlation.

The illustration in fig.12 below shows the sold product intersecting with the quantity ordered at respective prices. in this, we have the X-axis representing the Sold products with two Y-axis. Y1-axis is the Quantity ordered and Y2-axis represents the price of each product. On this plot, we can validate our Hypothesis that most of the sold products are cheap ones. This is observed with the first 3 highest sales which are respectively AA Batteries (4 packs), AAA Batteries (4 packs), and USB Charging cables.

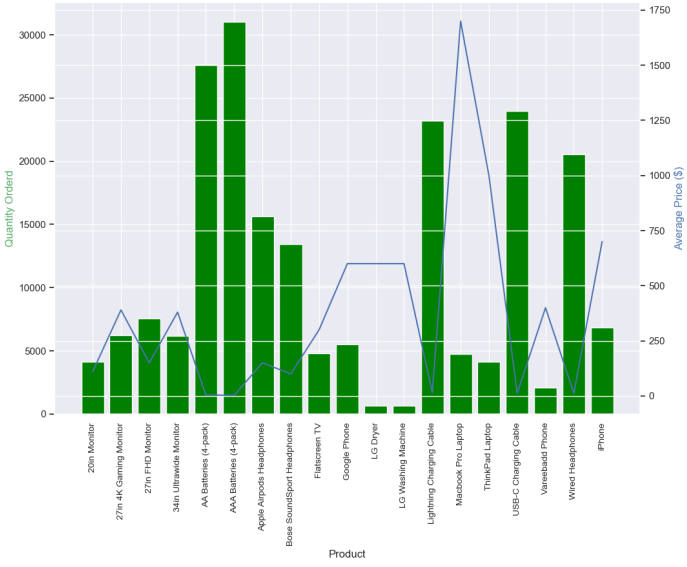

Figure 12. Sold Products Based on Quantity Ordered and Price

Here comes our Assumption: **That due to the interrelation that exists between the above measurable quantitative research study, there is a directly proportional relationship between the price of an item and the quantity ordered.**

Proposed Model

Linear Regression Model

In order to propose a model that could fit our analysis done above, observing the values of sales, month, and quantity ordered we observe continuous values which can be used to propose a linear regression. From these selected variables we could draw out a correlation table in order to find out which would be the right values to be used to come out with a model. We can observe a correlation table in fig.13.

	Month	Quantity Ordered	Sales
Month	1.000000	0.609091	0.589156
Quantity Ordered	0.609091	1.000000	0.998604
Sales	0.589156	0.998604	1.000000

Figure 13. Correlation table

From this correlation table, we can go on to use the high correlation that exists in order to make an estimation. **degree of efficiency of the model**; even though we had a little amount of data. The MSE(Mean Squared Error) of the model predicting Quantity Ordered for a given month is **13555931.214160837** and the prediction score is **0.3709918977726181** giving 37.0 Percent.

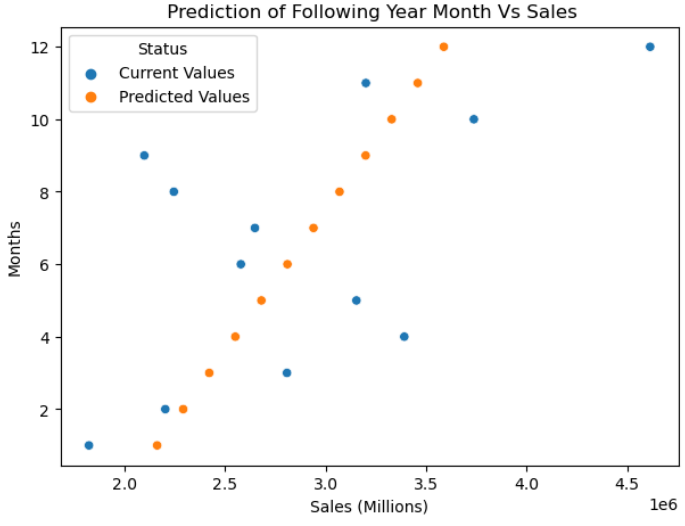

Figure 14. Above you talked about months and quantity ordered, here we are showing the plot of Months and Sales.

The Score of this prediction showed to be approximately 37 percent, which is not very reliable due to lack of data and also because the correlation between the months and Sales was approximately 58 percent. The X-axis presents the Sales

and Y-axis is the months expressed in numbers from 1 to 12. Also, as observed from our above prediction, the independent variable is the month as the sales will keep on fluctuating.

The mean squared error of actual value and predicted value for Quantity Ordered and Sales are respectively; **1608837104.96794**| **0.9972104396818211**. Fig.15 shows the prediction for the following year (2020) with the X-axis representing the Quantity ordered and the Y-axis representing the months with current values in blue and predicted values in yellow.

We obtained a correlation of approximately 60 percent, which is quite more specifiable compared to that obtained for the Months VS Sales. Here, the independent variable we used was also the month because of the fluctuation of the quantity ordered in relation to the items a customer may order.

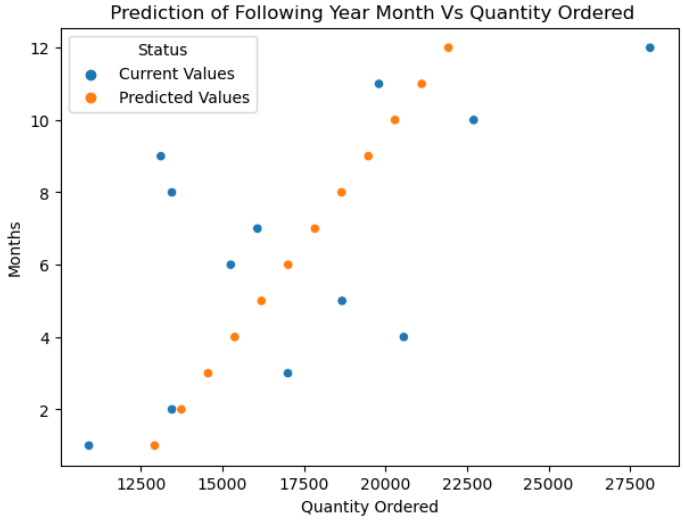

Figure 15. Prediction of months Vs Sales

on the other hand, in order to verify our assumption stated above, we plotted a scatter of the prediction of the following year (2020) of the quantity ordered and sales as observed below in fig.16

On the plot, we have on the Y-axis the number of sales made and on the X-axis the quantity ordered. We observed the mean squared error of actual value and predicted value for Quantity and Sales; **376547598167.19415**. This is coherent with the value of 99 percent obtained from our correlation table. This high correlation of values is also represented on the plot as some of the predicted values overlap some of the current values. This confirms our assumption that due to the interrelation that exists between the above measurable quantitative research study, there is a directly proportional relationship between the price of an item and the quantity ordered.

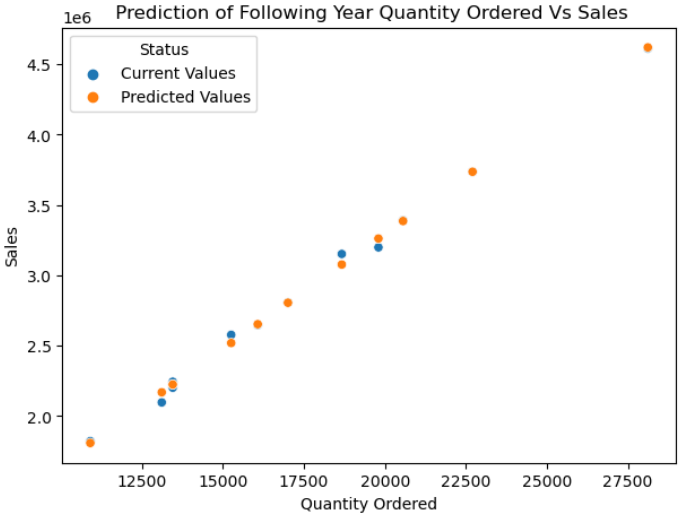

Figure 16. Prediction of months Vs Sales

Polynomial Regression Model

Because machine learning algorithms have their limitation and also due to the fact that a single algorithm may not always make the perfect prediction for a given data set, we decided to boost the overall accuracy of our predictions with the use of a Polynomial Linear Regression model.

We proposed a polynomial regression model of degree 3 of the sales vs months. We used 80 percent of the data as training data and the remaining 20 percent was used for prediction. This permitted us to get a better approximation as observed on fig.17

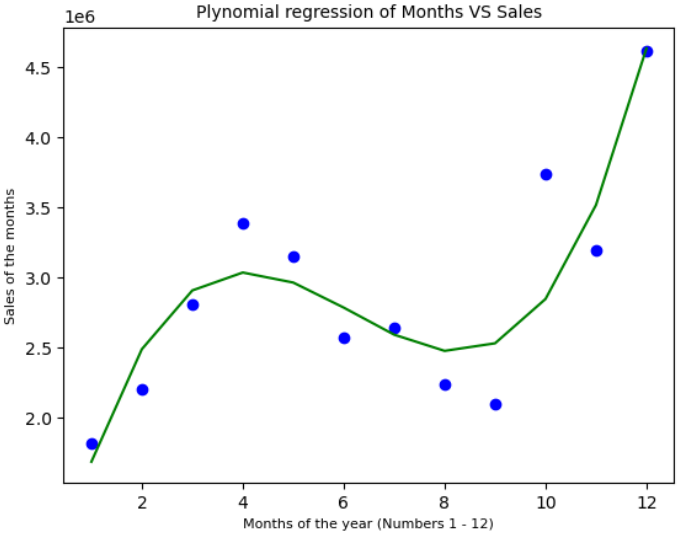

Figure 17. Prediction of months Vs Sales

This permitted us to come out with a prediction of Sales vs months where the Sales is on the Y-axis and the month on the X-axis. We plotted the prediction for the years 2021 and 2022 as observed in fig.18

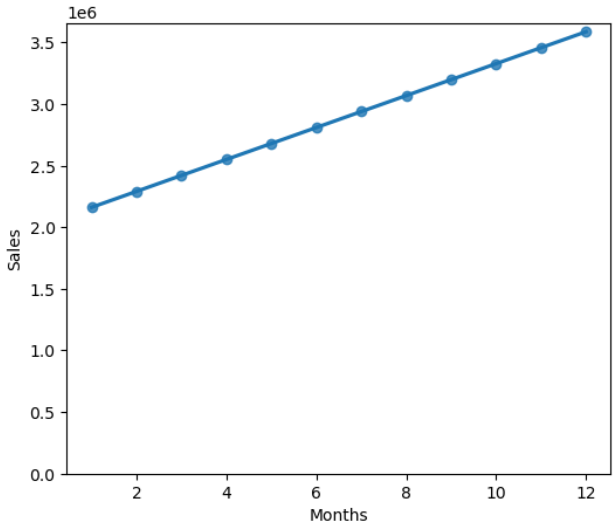

Figure 18. Prediction of months Vs Sales

The Same result was obtained due to the absence of enough data and also because the same training values were used to make a prediction of the two years(We have only a one-year data set). So, we normally have to expect the same prediction no matter the year for which we need to make a prediction.

Analysis Architecture

This concept is an umbrella term of a variety of technical layers that allowed our work organization to collect, organize, and parse the multiple data streams that we used more. Fig .19 below shows an analytic architecture reference to our system, protocols, and technology used to collect, store and analyze the data.

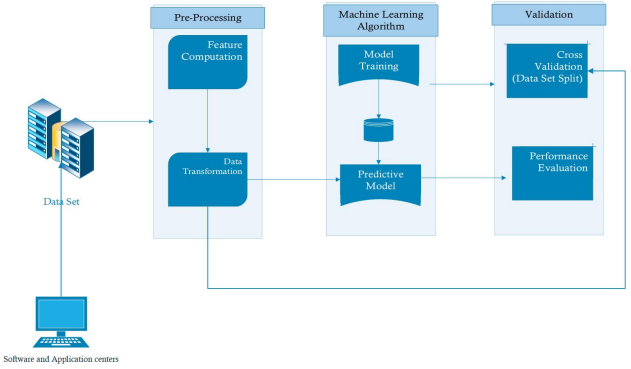

Figure 19. Prediction of months Vs Sales

After obtaining our data from Kaggle, we started with the first phase named Pre-processing, which consisted of feature computation and data processing. Here, we learned what the data looked like, identify the necessary fields to tag our analysis, got rid of unwanted fields in the data, and then add the required columns to facilitate the analysis. The Second phase was the Machine Learning algorithm which consisted of finding which model could fit based on the analysis previously made and training the model. Lastly was the Validation consisting of cross-validation and performance evaluation. Here, we split the data and observed the performance of the different models that could be drawn from the previous step.

Analysis Architecture

Conclusion

E-commerce with the many advantages it presents becomes the choice for every trader wishing to make his business as profitable as possible. However, classic e-commerce has some limitations when it comes to customer management. In order to remedy this, we propose a system based on the Analysis using python and Machine Learning algorithms which is able to predict the purchase intention of users of e-commerce platforms based on the data collected from the Kaggle platform. To do this, we used the Online Shoppers Purchasing Data-set and applied a strict methodology in order to get the most out of this data set as well as the most recent techniques and algorithms used in classification problems. In order to do this we answered a series of questions that could be useful for the prediction. We elaborated and test some hypotheses along with some assumptions that were equally verified.

Therefore in our study, we worked with different machine learning algorithms, some being classical machine learning algorithms. We trained different prediction models with these algorithms, then we evaluated them using metrics such as MSE and prediction score. This methodology allowed us to obtain interesting results with a model more efficient than the top models.

It should also be noted that our best result could not be approximated since the two models used perform the same. So, both models can be used for prediction. We can then highlight the fact that classical machine learning models can perform well in our analysis. The system we offer allows online merchants to receive proposed items when they select a specific product into their basket, identify the expected quantity to be sold for a specific month, and identify the time at which advertisement can be done. The system can then be used to offer personalized solutions to these potential customers or to be able to contact them if they do not end up making a purchase and this ultimately means increased revenue and better customer satisfaction.

Perspectives

We would like to find more in other techniques that can be used to clean and explore the data as this may be useful in upgrading our analytic skills.

In order to optimize this analysis, we should work out possible questions to be tagged in the analysis. Also, hoping that we obtained additional data we would like to explore other machine learning algorithms in order to test other models to ensure more accuracy in our prediction.

References

- Online Ads Sales | written by Anyaa Mishra | 2023
- Kaggle Platform, place we got our data-set
- Course on Machine Learning
- Voxco book we used for understanding the Polynomial Regression model
- An Approach of Sales Prediction System of Customers Using Data Analytics Techniques | 2020 paper
- The Economics of the Online Advertising Industry | Written by David S. Evans 2008
- The Online Advertising Industry: Economics, Evolution, and Privacy | Written by David S. Evans 2009
- Data visualization Library
- Data visualization 2 Library
- Data visualization Tutorial
- Parking Analysis Tutorial
- Predicting Crypto Prices in Python
- Study of Selling Behavior of Salesperson
- Specification Book
- Electric Vehicle Sales Analysis Model Based on User Purchase Intention Analysis written by Ying Zhang, Yibing Chen and Peisen Huang in 2022
- Best Selling Product and Category Prediction Using Sales Analysis written by Ms. Archana NikoseTejal, MungaleMinal and Shelke in 2022
- Improving Sales Analysis witten by Ivan Kononov in 2021