RESEARCHERID
THOMSON REUTERS

[ICRTCET-2018]

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

DATA MODELING WITH RETAIL ANALYTICS- LOCATION AND MARKET PENETRATION

Sivakumar D[*1], Krithika D[2], Nandakumar T[3] & Anirudh M[4]

[*1,2,3&4]Computer Science and Engineering, Rajarajeswari College of engineering Bangalore, India

ABSTRACT

In this paper, it addresses at building a system to analyse market penetration based on geo locations for retail stores. This may include identifying various patterns of events & activities happening at retails to identify and improvise the business process. The models created give a means to evaluate alternative methods and the effects of many promotions & activities done at a retailer. Sales of any product depends on its demand, location, purpose and several other factors. Existing systems do not provide a complete -insights into how the sales of the product happens, hence retail analytics comes into picture which gathers the data, extricating the significance from this grouped information. In a supply chain, a product is sold through the distributors via the retailers. Here the company can market and promote its sales by doing loyalty program. Data is collected is from various departments (Sales, Marketing, Distribution). The data is diverse and sometimes unstructured. To address this problem. It has to combine all the data which includes data from sales, marketing, distribution, manufacturing, location and loyalty program and prepare multiple data models to feed our analytics and build insights on top of them using pandas.

Keywords: *Market Penetration, Pandas, Numpy, Scikit, Matplot..*

I. INTRODUCTION

Sales of any item relies upon its request, area, reason, supply, demand and a few different variables. Most of the applications in existence are only used to connect and store data. Multiple departments within an organisation maintain their own applications & databases [1]. There are no ways by which the organisations/businesses are able to leverage upon the existing data. Almost all the applications are used for data collection but only few do reporting and the reporting are all basic ones. Businesses have now started to collect data from multiple sources. The data might be unstructured, unverified and information are dynamic in nature. Organisations don't have an understanding about how, when and why a client purchases an item. Retail examination is a technique of gathering, getting ready and extricating the significance from this group information [2]. It basically prepares the data, and identify the required features by feature engineering, analyze the search & engagement patterns of customers towards the products/services. It proposes a solution to create data model by combining information from multiple data sources, and provide analytical dashboards for businesses in making informed decisions. Several loyalty programs like discounts, price cut downs, gift vouchers, lottery prize are offered by the wholesaler to the retailer in order to hike the sales and shares of the organization.

Loyalty programs attract the retailers and boost their sales. But these are just an addition to develop the income, but the real task lies in predicting the location and market penetration for the particular product [3]. The system which is addressed in this paper says that the input from various sources can be collected - Retail POS systems, user purchase data, inventory management system, order inventory system. Programming language like python is used in this system. The collected data has been linked, integrated using Apache Ni-fi, a software that supports data integration by following various techniques. The customers are categorized by identifying various features, and this can be done with the help of Pandas and Matplot. Scikit learn can be employed to identify leads by. This produces a dashboard that provides a user or retailer the complete decision as to whether his sales and business in the particular geographic location and market will be success or loss. Many startup companies and developing businesses will face problems in deciding where they have start it. This retail analysis will play a vital major role for business startups and developing business to flourish in the particular field. Business losses financially or ethically can be absolutely avoided by using this model. This model helps the retailer to make an informed

ISSN 2348 – 8034
Impact Factor- 5.070

decision, whether his proposal in the particular geographic region or location can be successful or not preparing multiple data models for the retailer to show the insight of product.

II. LITERATURE SURVEY

Transactions on Visualization and Computer Graphics [5] existing desire procedures generally speaking require a tremendous measure of tests and long planning estimate precision is poor for programs that experience a high apex or sharp lessening in universality. This paper shows our upgraded gauge approach in light of example recognizable proof The genuine doubts made in this model excludes the group and it's totally a desire course of action.

SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations [6], in this examination, we attempt to use visual examination that joins the best in class mining and observation techniques to deal with this issue using broad scale GPS bearing data.

An Integrated eVoucher Mechanism for Flexible Loads in Real-Time Retail Electricity Market [7]: This paper proposes an unique money related and constructing coupled structure to stimulate versatile loads or load aggregators, for instance, parking structures with high invasion of electric vehicles, to take a premium particularly in the persistent retail influence promote in perspective of a planned program. At whatever point executed, the eVoucher program grants typical versatile weights, for instance, electric vehicle stopping zones, to change their demand and usage direct as demonstrated by cash related stimuli from an EDC. An assignment structure director (DSO) fills in as an outcast to hustle exchanges between such stopping zones and EDCs, and furthermore the esteem clearing process. Definitely, both power retailers and power structure directors will be benefited by the dynamic collaboration of the versatile weights and imperativeness customers.

Learning Dynamic Prices in Multi Seller Electronic Retail Markets with Price Sensitive Customers, Stochastic Demands, and Inventory Replenishments Reinforcement learning is used. Machine learning based approach for dynamic esteeming in ordinary electronic retail publicizes [8]. Honest to goodness systems have not been associated before a bit of the assumptions made by us in the retail promote show ought to be easygoing: the nature of volume discounts; the nature of stock procedure; and the doubts as for clients and detainees. Besides, prosperity stock to be kept up by merchants can be exhibited as a decision variable and the model debilitated by them end up being generously all the more charming and complex.

A Generic Operations Framework [9] for Discos in Retail Electricity Market, headway of a careful operational device that a disco can expeditiously use when working in an open power promote condition; recommendation of a twoorchestrate logical model reasoning about disco's operational needs, goals and confinements both in its day-ahead and consistent errands. Since its only models being used, there are lot of assumptions made, hence accuracy might be lacking.

Putting Analytics on the Spot or How to Lower the Cost for Analytics [10] Cloud services and computing by using Hadoop like systems, it increases fault tolerance. Hadoop requires reading and writing a lot of data from and to disk, which makes the system slow.

Investing and Pricing with Supply Uncertainty [11] in Electricity Market: A General View Combining Wholesale and Retail Market decision problems are formulated and solved using the data model (market game model) Since this game model, makes use of assumptions, the accuracy in the market trends cannot be guaranteed.

ISSN 2348 – 8034
Impact Factor- 5.070

III. PROPOSED ARCHITECTURE

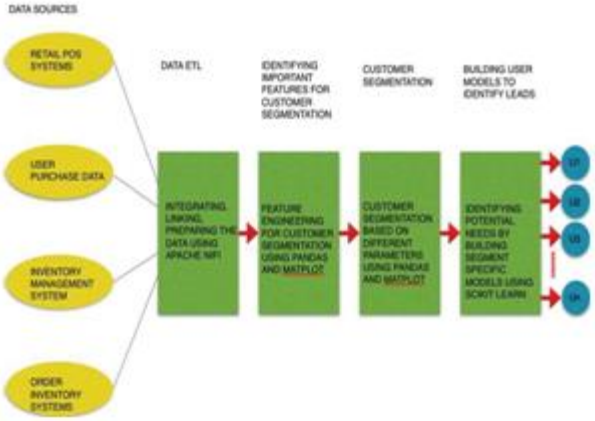

Fig. 1. Architecture diagram

The Fig.1 architecture shows the following models. DATA SOURCES- information or data source is the essential area from where information comes. The data source can be a database, a dataset, a spreadsheet or even hard-coded information. Here the data sources are from an assortment of frameworks that contain the client data. So as to accumulate the data we utilize different sources like POS Systems, E-Commerce, Customer information, Data Inventory and so on.

DATA ETL- stands for extracting transforming and loading the data. The connecting and stacking are refined by a program which is differently called the linker, or the loader, or the connecting loader. Despite the fact that connecting and stacking are theoretically particular, it is exceedingly regular that a solitary program join those capacities.

IDENTIFYING IMPORTANT FEATURES FOR CUSTOMER SEGMENTATION- "connecting" and "stacking" mean a similar thing, when talking casually, despite the fact that they are separate activities. This process of linking is the most important aspect of our system as it helps us analyse the customer behaviour and their pattern of searching for their item of interest, which in turn helps us come up with an effective recommendation model for the users. This process is achieved using Apachenifi. This representation is a productive method to see the best clients for example and their relative weight in your general business. This information is sectioned in light of the client's decision, inclination and needs. This portioned information causes us distinguish potential clients for any business, so the information is acquired by breaking down the client's example of obtaining merchandise. This arrangement of sectioned information that has diverse data for various client causes us fabricate a special proposal display for each client. This is done using pandas and matplot.

CUSTOMER SEGMENTATION- Division depends on valuable customer, one time customer, loyal customer basically look out for customers with shopping patterns like people who buy frequently, people who spend high, people who would spend high under certain conditions (say some spend huge in buying organic products). In large retails, it becomes impossible to keep loyal customers. So, the business from a repeating customer is higher than the business from one-time customer. Our solution will enable retailers create more loyal customers from one-timers. Feature engineering is the way toward utilizing space information which makes machine learning calculations work. Feature engineering is key to the utilization of machine learning, and is both troublesome and costly. The requirement for manual component building can be obviated by automated feature learning. An element is a trait or property shared by the greater part of the autonomous units on which investigation or expectation is to be finished. Any attribute could be a component, as long as it is helpful to the model. The reason for an element, other than being a characteristic, would be significantly simpler to context with regards to an issue.

ISSN 2348 – 8034
Impact Factor- 5.070

UILDING USER MODELS TO IDENTIFY LEADS- By recognizing the leads in the client division we will become more acquainted with the scope of clients, genuine clients, important clients and how frequently they buy a similar item, based on hit, regularly reordered and by investigating client propensity the client models for the specific client or client is worked by the client necessity and division. Better customer segmentation can create easier and more adaptable models, and they frequently yield better outcomes. The calculation we utilized are exceptionally standard for kagglers. This recommendation Models helps the retailers as well as the organisation whether the product will be a hit or not in a particular geographical location or any region.

SCIKIT- Is a free programming machine learning library for the Python programming language. It highlights distinctive portrayal, backslide and gathering computations including reinforce vector machines, k-means and DBSCAN, and is proposed to interoperate with the Python numerical and intelligent libraries NumPy and SciPy. Scikit gives an extent of directed and unsupervised learning calculations by means of a steady interface in Python. The library relies upon the SciPy (Scientific Python) that must be introduced before you can utilize the software scikit-learn. This bundle incorporates Numpy, Matplot and so forth. Anaconda is a free and open source dissemination of the Python and R programming dialects for huge scale information preparing, prescient examination, and logical figuring, that expects to improve bundle administration and organization. Bundle variants are overseen by the bundle administration framework anaconda. Customer Segmentation- this is based on different parameters using pandas and matplot. Building user models to identify leads- Identifying potential leads by building segment specific models using scikit learn. Since python language is being used, Anaconda and Jupyter notebooks are used to simulate the model. The resulting output is a dashboard that tells a retailer/user whether his business would be successful or not. These models would accurately predict whether their product will be a hit. Apache NiFi underpins intense and adaptable coordinated diagrams of information directing, change, and framework intercession rationale. Pandas is a product library composed for the Python programming dialect for information control and investigation. Specifically, it offers information structures and tasks for controlling numerical tables and time arrangement.

IV. RESULTS AND DISCUSSION

The Jupyter Notebook is an open -source web application that empowers you to make and offer reports that contain live code, conditions, perceptions and story content. Utilizations include: information cleaning and change, numerical reenactment, factual displaying, information perception machine learning and much more. Our job is to accurately predict which item will be reordered on the next order.

In this Fig.2 the dataset contains departments, ordered products, aisle. These file descriptions as an entity with unique identity, using this data we perform many operations using panda functions. The dataset gives the list of all orders we have in the dataset. Let’s have a look when people buy groceries online. We perform operations based on when do people order hour of a day, day of week, when do they reorder again, how many prior order there, how many items do people buy and which is the best seller amongst them and how often people are at the same item again, most often reordered which item do people put into the cart first.

ISSN 2348 – 8034
Impact Factor- 5.070

order_id	user_id	eval_set	order_number	order_dow	order_hour	days_since_prior_order
2539329	1	prior	1	2	8	
2398795	1	prior	2	3	7	15
473747	1	prior	3	3	12	21
2254736	1	prior	4	4	7	29
431534	1	prior	5	4	15	28
3367565	1	prior	6	2	7	19
550135	1	prior	7	1	9	20
3108588	1	prior	8	1	14	14
2295261	1	prior	9	1	16	0
2550362	1	prior	10	4	8	30
1187899	1	train	11	4	8	14
2168274	2	prior	1	2	11	
1501582	2	prior	2	5	10	10

Fig. 2. Input Dataset

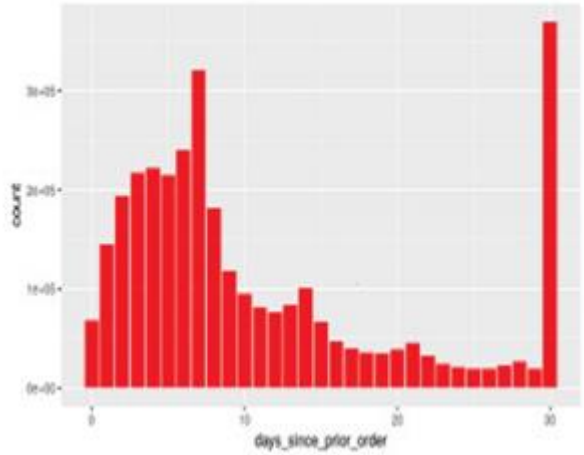

Fig.3 Histogram Plot

Fig.3 describes the basic information about the dataset and creates picture of data distribution, here the graph shows the number of customers on x -axis and number of prior orders on y-axis. Which shows how many people seems to order more often after exactly one week.

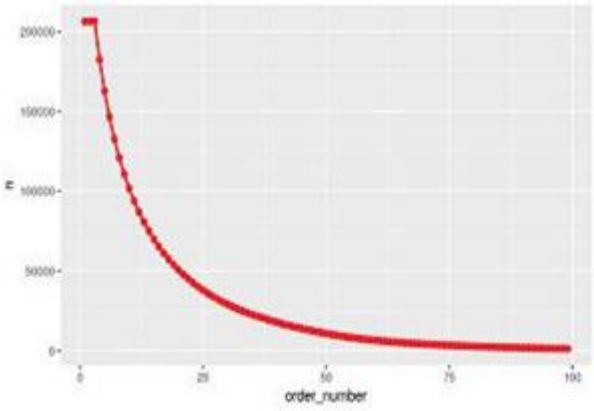

Fig.4 Kernel Density Estimation plot

Fig.4 describes the estimation based on the notified data of an unobservable underlying probability density function. Which s hows the number of customers on y-axis and order numbers on x-axis. It shows how many items are in the orders and compar es between the train and prior order set.

Fig.5 scatter plot display values for typically two variables for a set of data. Which shows the positive relationship between number of customers and proportion reordered so products with higher number of orders are more likely to be reorder.

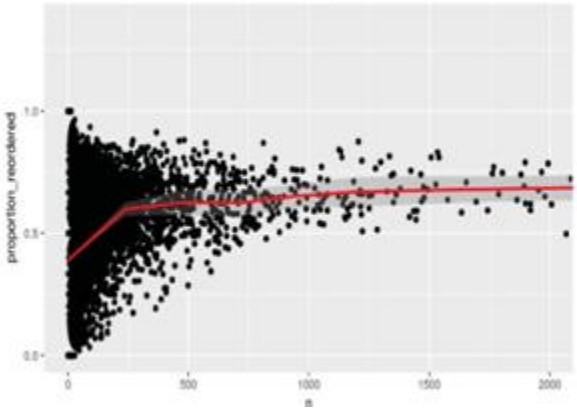

Fig.5 Scatter Plot

The above are the graphs obtained for the dataset, by plotting these graphs and can explore customer habits like the customer who just reorder the same product all the time so we can accurately predict the product which the customer reorder again in the next order and this model gives the retailer the detail description about the market for the product and the location where you can sell the products which helps in making informed decisions whether the business will be a success or not in a particular geographical area.

V. CONCLUSION

In this paper, the dataset is collected from various sources and it combines all the data which includes data from sales, marketing, distribution, manufacturing, location and prepare multiple data models for the retailers to make and informal decisions for their business to excel and the utilization of Jupyter notebook, pandas and matplot capacity to create the outcomes have turned out to be exceptionally proficient and by plotting the above graphs and comparing the values with ordered set and train set 95% accurate prediction of any product that he would buy can be made. The dashboard gives us the precise expectation as in whether the client or retailer can effectively maintain the business or not and it predicts the accurate result of the product in the particular geographic region or location and the resulting output is a dashboard that tells the retailer whether the product will be a hit or not.

REFERENCES

1. *Yilin Song, "Online Cost Efficient Customer Recognition System For Retail Analytics", Applications Of Computer Vision Workshops (Wacvw), 2017 Ieee Winter 27 April 2017.*
2. *Zhiqiang Ge, "Data Mining And Analytics In The Process Industry: The Role Of Machine Learning", Ieee Access 26 September 2017.*
3. *Chengang Zhu, "Big Data Analytics For Program Popularity Prediction In Broadcast Tv Industries", Ieee Access 27, Vol.27, No 5, October 2017.*
4. *A.R. Al-Ali, Imran A. Zualkernan, Mohammed Rashid, Ragini Gupta, Mazin Alikarar, "A Smart Home Energy Management System Using Iot And Big Data Analytics Approach", Ieee Transactions On Consumer Electronics, Vol. 63, No. 4, November 2017.*
5. *Chengang Zhu1, Guang Cheng1 And Kun Wang (2016), Ieee Transactions On Visualization And Computer Graphics Vol. 23, No. 1, January 2017.*

6. *Tao Chen, "An Integrated Evoucher Mechanism For Flexible Loads In Real-Time Retail Electricity Market", Ieee Access Volume 5 26 January 2017.*
7. *V. L. Raju Chinthalapati, Narahari Yadati, "Learning*
8. *Dynamic Prices In Multiseller Electronic Retail Markets With Price Sensitive Customers, Stochastic Demands, And Inventory Replenishments", Ieee Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 36, No. 1, January 2006.*
9. *Ayed A. S. Algarni, Student Member, Ieee, "A Generic Operations Framework For Discos In Retail Electricity Markets", Ieee Transactions On Power Systems, Vol. 24, No. 1, February 2009.*
10. *Elkhan Dadashov, Ugur Cetintemel, And Tim Kraska Brown University(2014), "PuttingAnalytics On The Spot", September/October 2014.*
11. *Li Xiaobo, Gao Li, Wang Gongpu, Gao Feifei, Wu Qi, "Investing And Pricing With Supply Uncertainty In Electricity Market: A General View Combining Wholesale And Retail Market", China Communications March 2015.*
12. *Y. Yao And F. Gao, ''A Survey On Multistage/Multiphase Statistical Modeling Methods For Batch Processes,'' Annu. Rev.*
13. *Control, Vol. 33, No. 2, Pp. 172–183, 2009.*
14. *N. F.Thornhill,S. L. Shah, B. Huang, And A. Vishnubhotla,*
15. *''Spectral Principal Component Analysis Of Dynamic Process Data,'' Control Eng. Pract., Vol. 10, No. 8, Pp. 833–846, 2002.*