Shopper Pulse Estimation using Outliers in Retail Shopping Data

Srikar Chilakamarri[1], Ankur Agarwal[2]

[1]*Director Technology. Hyderabad,India*
[2]*Data and Analytics Lead. Reading,UK*

---***---

Abstract - *Data science in the retail world has provided an understanding of customer preferences and choices that have helped in cross selling, upselling and improving customer service. General applications of demand forecasting, price optimization, sales prediction, marketing strategies, customer behavior pattern, customer segmentation and many more are integral part of many businesses. The advancement of technologies has provided an ability to provide real time intelligence to make quicker and faster decisions. However, during the data cleansing process, variations in buying patterns usually end up being outliers and are eliminated from the decision making analytics. We propose analysis of these outliers along with other buying patterns to estimate the pulse of the customer at the time of shopping. This paper defines a method along with guidelines of implementation to determine the same.*

Key Words: Outlier Analysis, Machine learning, Retail, Artificial intelligence, Consumer Pulse

1. INTRODUCTION

The shopping/ purchase data is used to derive patterns to understand the trends and sufficiently predict the future patterns. Understanding the shopper's behavior helps in determining his inclination towards the products during the visit.

A large volume of data is parsed, filtered and enhanced to obtain right data quality to support the analysis. Outliers are patterns in data that do not conform to a well-defined notion of normal behavior [1]. Science today is mature enough to determine the outliers in the given set of data using clustering methods. Outliers have extensive use in a wide variety of applications such as military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems [1]. In this paper, we define a method to estimate a shopper pulse by analyzing retail shopper data that potentially drives his buying behavior and helps business with real time cross sell and upsell opportunities.

1.1 Shopper Pulse

We define Shopper pulse as the estimation of human behavior based on the past behavior patterns in specific conditions.

In this method, we peruse the outliers, usually referred to as noise, to define a methodology using the retail shopping data of individual customers and review it with sales & product data to estimate the behavior.

1.2 Methodology

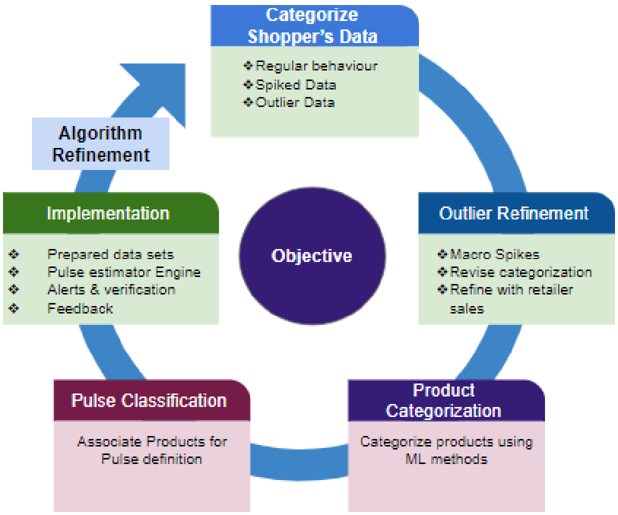


Fig -1: Shopper Pulse estimation methodology

The methodology is based on the foundational methodology of CRISP-DM model [2]. The below is the mapping of the phases prescribed

Table -1: Methodology comparison

Crisp-DM	**Shopper Pulse Estimation**
Business Understanding	Objective
Data Understanding	Categorize Shopper's data
Data Preparation	Outlier refinement
Modelling	Pulse Classification

Evaluation	Implementation
Deployment	Implementation

Data sample from repository [3] has been used to perform a sample analysis to describe the methodology. In this set logical grouping of products to obtain a right sample. Generalization methods for products, categories & pulse is used for representation and ease of understanding.

2. OBJECTIVE

The objective of the analysis is to review the outliers in the retailer shopping data to identify patterns which shall help understand the Shopper pulse

2.1 Categorizing Shopper Data

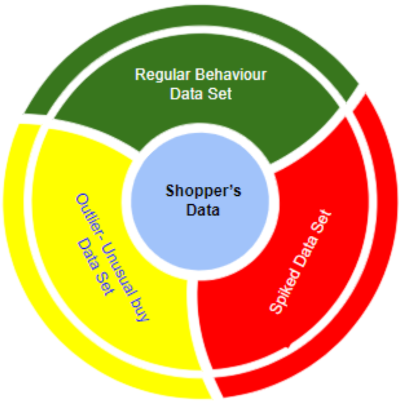


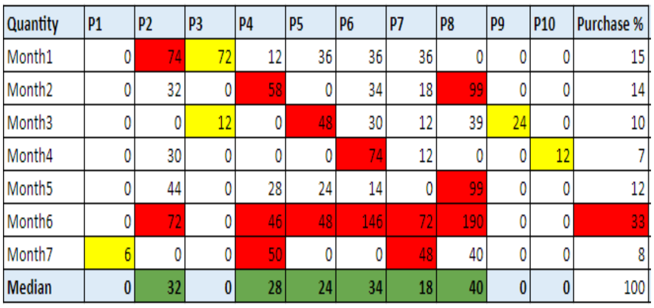
Fig -2: Shopper Data Categories

Shopper data is classified into 3 sets for the purpose of this paper

1. *Regular Behavior Data set:* The data set which shows a regular buying pattern of the customer of a product / set of products in the historical data analysis.
2. *Spiked Data set:* The data set where an increase in the buying pattern is noticed. This, in one example, can be derived as the distance of the data point from the median of total
3. *Outlier- Unusual Buy Data set*- The data set which does not conform to the regular buying data set. It may however have a sporadic trend over a period of time in the historical data set

Considering a sample data set of a shopper's data across months and segregating into the above categorization. P1, P2, P3 represent different products.

Table -2: Example- Shopper data categories



Quantity	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Purchase %
Month1	0	74	72	12	36	36	36	0	0	0	15
Month2	0	32	0	58	0	34	18	99	0	0	14
Month3	0	0	12	0	48	30	12	39	24	0	10
Month4	0	30	0	0	0	74	12	0	0	12	7
Month5	0	44	0	28	24	14	0	99	0	0	12
Month6	0	72	0	46	48	146	72	190	0	0	33
Month7	6	0	0	50	0	0	48	40	0	0	8
Median	0	32	0	28	24	34	18	40	0	0	100

The notations used are:

1. Regular behavior data set- Green (consistent purchase over a period of 7 months. Ignoring intermittent no purchases)
2. Spiked data set - Red (Rule 1- Farthest point from the median considered as greater than 10% of the total sales quantity)
3. Outlier data set- Yellow (Rule 2- with median value tending to zero, indicating unusual buy pattern)

2.2 Outlier Refinement

Step 1: *Eliminate macro spike patterns*

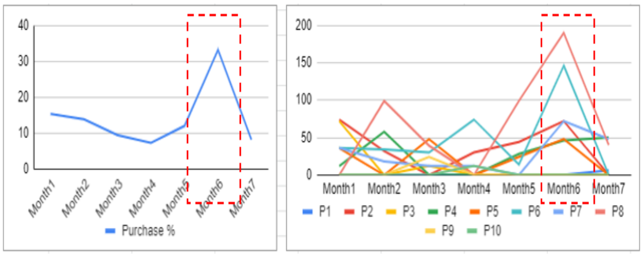
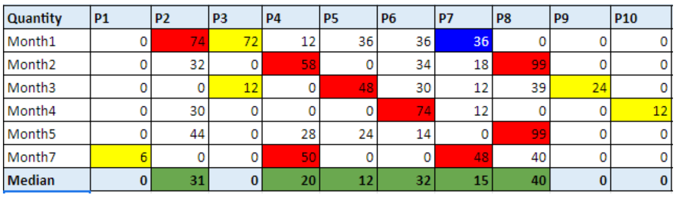


Chart -1: Macro Spike Detection

Further analysis of the shopper's trend of total purchases, eliminate the data sets which have witnessed a generic spike. The observation infers that {Month 6} has an overall spike in the purchase pattern. Further, there are no outliers identified in this month (Ref- Table-2). Hence, this data set is discarded from further analysis. Inclusion of this data shall result in increase in processing time and also skew of results.

Step 2: *Revise shopper's data categorization*

Table -3: Example- Revised Shopper data categories



Quantity	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Month1	0	74	72	12	36	36	36	0	0	0
Month2	0	32	0	58	0	34	18	99	0	0
Month3	0	0	12	0	48	30	12	39	24	0
Month4	0	30	0	0	0	74	12	0	0	12
Month5	0	44	0	28	24	14	0	99	0	0
Month7	6	0	0	50	0	0	48	40	0	0
Median	0	31	0	20	12	32	15	40	0	0

After Step 1, a new spike point {Month 1, P7) is identified due to change in the median points. This is marked in blue for notation purpose in Table-3

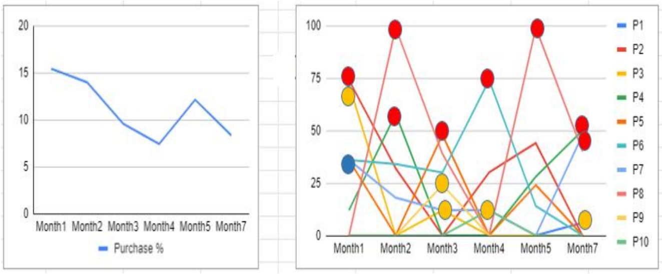


Chart -2: Spikes & Outlier identification

Chart -2 represents the product sales trend of products for one customer over a period of time. The Red & Blue dots represent the spike in the regular items purchased by the customer, while the yellow dots represent the outliers for the shopper during his lifetime with the retailer.

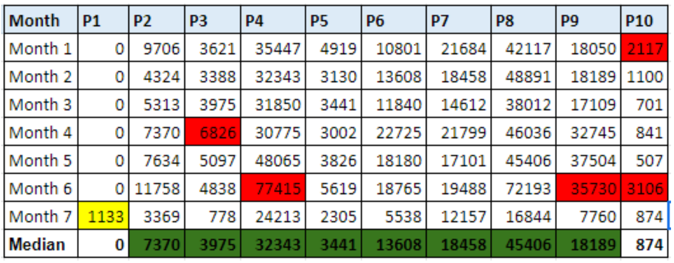
The outlier data set for the customer to be considered for further analysis is
{P1,Month7}, {P2,Month1}, {P3,Month1}, {P3,Month3}, {P4,Month2}, {P4,Month7}, {P5,Month3}, {P6,Month 4}, {P7,Month1}, {P7,Month7}, {P8,Month2}, {P8,Month5}, {P9,Month3}, {P10,Month4}

Step 3: *Refine outliers with the retailer sales trend*

The data set is further analyzed with the total sales data for the retailer to further refine the set of outliers. If a spike &/or outlier data of the customer correlates to the overall spike/outlier in the retailer's data, this data can be eliminated from the outlier set. Any intersection of data sets with similar data categories (Normal, spiked, outlier) between retailer & individual customer is eliminated.

The total retailer sales data is also categorized based on the same rules-Rule 1 & 2, mentioned in Section 2.1, as applied to customer data. The notations of Green for normal behavior, red for spikes and yellow for outlier is used for consistency.

Table -4: Example- Outlier categorization Retailer sales data



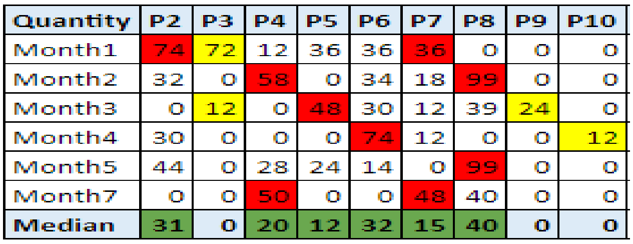
Month	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Month 1	0	9706	3621	35447	4919	10801	21684	42117	18050	2117
Month 2	0	4324	3388	32343	3130	13608	18458	48891	18189	1100
Month 3	0	5313	3975	31850	3441	11840	14612	38012	17109	701
Month 4	0	7370	6826	30775	3002	22725	21799	46036	32745	841
Month 5	0	7634	5097	48065	3826	18180	17101	45406	37504	507
Month 6	0	11758	4838	77415	5619	18765	19488	72193	35730	3106
Month 7	1133	3369	778	24213	2305	5538	12157	16844	7760	874
Median	0	7370	3975	32343	3441	13608	18458	45406	18189	874

On comparison of data for the customer in Table-3 and retailer data in Table-4, the data of {P1, Month7} is an intersection for the outlier data set from the customer. This is eliminated from the outlier data set.

The resultant outlier of this exercise is
{P2,Month1}, {P3,Month1}, {P3,Month3}, {P4,Month2}, {P4,Month7}, {P5,Month3}, {P6,Month4}, {P7,Month1}, {P7,Month7}, {P8,Month2}, {P8,Month5}, {P9,Month3}, {P10,Month4}

The refined data set post total sales data comparison is

Table -5: Final Outlier Data set



Quantity	P2	P3	P4	P5	P6	P7	P8	P9	P10
Month1	74	72	12	36	36	36	0	0	0
Month2	32	0	58	0	34	18	99	0	0
Month3	0	12	0	48	30	12	39	24	0
Month4	30	0	0	0	74	12	0	0	12
Month5	44	0	28	24	14	0	99	0	0
Month7	0	0	50	0	0	48	40	0	0
Median	31	0	20	12	32	15	40	0	0

For the method proposed here, the outlier data set {P3, Month1}, {P3, Month3},{P9,Month3}, {P10,Month4} is used. The other datasets support validation.

2.3 Product Categorization

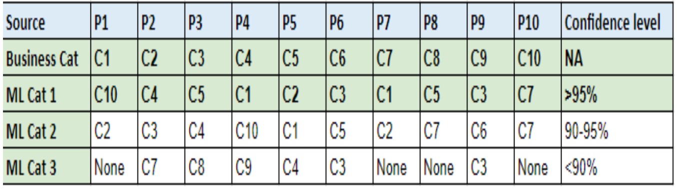
Products offer multiple features and are often difficult to categorize a product in a single hierarchy and category [4]. E.g. A portable battery charger could be categorized under any device accessory like electronics, accessories, and toys and also under travel categories. This is also dependent on the line of business of the retailer. To identify the possible categorizations, we propose to apply multiple classification [5] tags.

The changing customer needs and market demands, these tags are generated using the regular data collected by web scraping [6] the online retailers & manufacturer websites, and then applying multi-label classification [7] to pick multiple category tags associated with that product with confidence factor. Depending on the following additional information, weightings for the tags could be defined which feeds as features to ML based multi-label classification [7] [8] process:

i) Tags identified/derived from the manufacturer's website of the product - this source is assumed to provide most authentic categorization about the product and hence highest weighting is assigned.
ii) Frequency of the occurrence of the tags across multiple websites - if a same category tag is appearing from multiple sources then weighting is accordingly increased for that tag.
iii) Regular manual adjustment of the weighting by the data owner in the organization - Based on the core line of business, the data owner manually adjusts the weighting of specific tagging.

For illustration, we consider the existing categorization of the retailer (business cat) and the ML based classification, depicted in Table-6. Same category classifications are not considered with ML classification. The association can be done programmatically using the methods cited in [7] [8] and shall yield the affinity within products.

Table -6: Product Categorization



Source	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Confidence level
Business Cat	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	NA
ML Cat 1	C10	C4	C5	C1	C2	C3	C1	C5	C3	C7	>95%
ML Cat 2	C2	C3	C4	C10	C1	C5	C2	C7	C6	C7	90-95%
ML Cat 3	None	C7	C8	C9	C4	C3	None	None	C3	None	<90%

A sample inference can be drawn as below

- Product P1 can belong to category C1 or C10
- Category C1 can consist of P1, P4 & P7 (Based on ML categorization with >95% confidence level)

Hence with the above inference, the following products {P1, P4, P7} are associated closely but are viewed under different categories by the retailer. These products can be grouped into another category of 'Pulse' referring to 'Shopper pulse'.

The other methods of using Affinity algorithms [9] can also be augmented to this method.

The pulse categorization would result in the below grouping set

Table -7: Pulse Categorization

Pulse categorization	Products		
Pulse 1	P1	P4	P7
Pulse 2	P2	P5	
Pulse 3	P3	P6	P9
Pulse 4	P4	P2	
Pulse 5	P5	P3	P8
Pulse 6	P6		
Pulse 7	P7	P10	
Pulse 8	P8		
Pulse 9	P9		
Pulse 10	P10	P1	

The final behavioral outlier data set is examined with the product grouping/ categorization.To explain the use of this data in the algorithm definition- if P3 is found to be an outlier in a particular month, then Pulse 3 & Pulse 5 are the possibilities of the shopper's pulse. The study would need to consider the association of the outliers with the spikes/ outliers of following products P6, P9 (from Pulse3) & P5, P9 (from Pulse 5). This shall result in a classification of the shopper's pulse.

2.4 Pulse Classification

Rule 3- The Pulse classification would be derived based on the occurrences. The formula used is

n= number of products in pulse categorization

i= number of products for which match is found

*Confidence =(i/n) *100*

For e.g./- if P5 match is found in outlier data set in the same month and P8 is not, then Pulse is classified by 66% confidence

The review of the outlier data sets before and after refinement, can be found in Table-2 & Table-5. The outlier data points, as arrived in section 2.2, are {P3, Month 1}, {P3, Month3}, {P9, month3}, {P10, Month4}. The data with spikes will be used as reference only for this evaluation.

Table -8: Pulse Classification

Data Set	Condition	Result	Output
{P3, Month 1}	Pulse 3- Check P6 & P9	P6 & P9- do not show a spike / outlier	Pulse 3- Ruled out
{P3, Month 1}	Pulse 5 - Check P5 and P8	P5 & P8- do not show a spike / outlier	Pulse 5 - Ruled out
{P3, Month 3}	Pulse 3- Check P6 & P9	P6- No spike/Outlier P9- Outlier	Pulse 3- with 66% confidence
{P3, Month 3}	Pulse 5 - Check P5 and P8	P5 - Outlier P8- No spike/Outlier	Pulse 5 - with 66% confidence
{P9, month3}	Pulse 9- No other products	P9- Outlier	Pulse 9- 100% confidence
{P9, month3}	Pulse 3 - Check P3 and P6	P3 - Outlier P6 - No spike/Outlier	Pulse 3- with 66% confidence
{P10, Month4}	Pulse 7 - Check P7	P7- No Spike/Outlier	Pulse 7 - Ruled out
{P10,Month4}	Pulse 10- Check P1	P1- No Spike/Outlier	Pulse 10- Ruled out
Based on the analysis- Pulse 9, Pulse 3 and pulse 5 are the probable candidates. However Pulse 9 with 100% confidence shall be treated as the final Pulse of the customer			

Hence for this data set of the customer, Pulse 9, Pulse 3 & Pulse 5 can be considered. Based on the set of pulse classes identified for any shopper, top 'N' can be used during implementation.

3. IMPLEMENTATION

Implementation of this method requires:

1. Prepared data sets - Shopper's data consisting of historic regular items & buying patterns, Product categorization & Pulse categorization data sets
2. Engine hosting the algorithm which identifies outliers from real time feed and connects the prepared data sets to arrive at the classification
3. Alerts mechanism to provide outputs in a consumer application recognizable format

We propose following implementation steps for estimating the pulse of the customer.

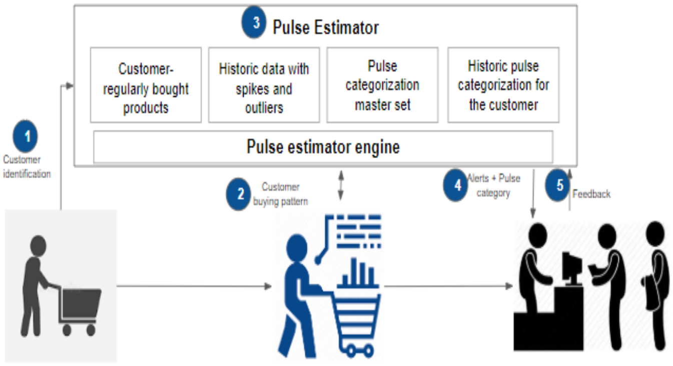
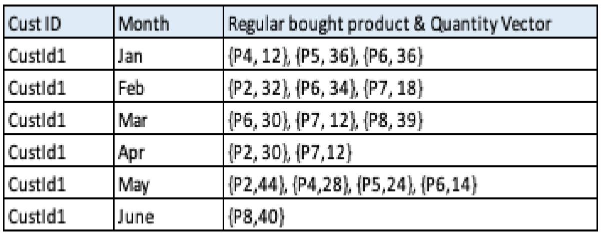

3 Pulse Estimator
Customer- regularly bought products
Historic data with spikes and outliers
Pulse categorization master set
Historic pulse categorization for the customer
Pulse estimator engine
1 Customer identification
2 Customer buying pattern
4 Alerts + Pulse category
5 Feedback

Fig -3: Implementation approach

1. **Customer Identification**: Customer submits his identity. This can be manual submission of his id, member card scanning or advanced IoT sensor based scanning [10]
2. **Real time customer behavior feeds**: The basket of the customer transmits the products added to the cart to the real time algorithm. This algorithm tracks his purchases and triggers the pulse estimator algorithm
3. **Pulse estimator** : has the following components:

a. **Master customer data set of regularly bought products and quantities** in an offline analytics process. We propose to create and regularly refresh the master list of regularly bought products and quantities for each customer using the historical sales transactions data. To cover the seasonal buying behavior, this list is to be maintained for each season (or quarter/ months) as the customers regular buying behavior changes during the year. This master data is going to help in the isolating unusual buying behavior of the customers in the further steps.

Table -9: Customer- Regular items



Cust ID	Month	Regular bought product & Quantity Vector
CustId1	Jan	{P4, 12}, {P5, 36}, {P6, 36}
CustId1	Feb	{P2, 32}, {P6, 34}, {P7, 18}
CustId1	Mar	{P6, 30}, {P7, 12}, {P8, 39}
CustId1	Apr	{P2, 30}, {P7,12}
CustId1	May	{P2,44}, {P4,28}, {P5,24}, {P6,14}
CustId1	June	{P8,40}

b. **Historic buying pattern data:** with identification of spikes in the past purchase & also outliers of the past.

c. **Pulse categorization master list:** A master list of all classified pulse with product affinity. Example Table 7.

d. **Historic Pulse Classification**: A list of pulse identified for the customer in the past.
e. **Pulse Estimator Engine:** Detects the unusual behavior and Alert the "Pulse Estimator" algorithm during the customer shopping journey.
In order to estimate the shopper's pulse during a shopping journey, we propose to analyze the customer basket as soon as the customer scans the product. A streaming data solution as referenced in [11] is proposed to analyze each transactions in the basket as they happen, and compare those with the historical buying pattern data and to detect any of following 2 conditions
a) if the customer scans an unusual quantity of a regular product. We define this data as Set A. Set A data is used to generate alerts and to modify the data set for future reference
b) if the customer scans a product which is not in the master list (Table 9). We define this data as Outlier (Set B). Set B data is used by the algorithm to identify the pulse of the customer
In other solutions, detection of this unusual behavior can be implemented using unsupervised real-time anomaly detection for streaming data [12] algorithms or using data transformation APIs available in the data solution.

Using the Pulse classification process defined in section 2.4, feature list specific to customer visit is created by joining set B with Pulse categorization master list. The Pulse classification is arrived at based on Rule 3 mentioned in section 2.4. In other methods, it is fed into a clustering algorithm [13] [14] [15] which in effect identifies similar transactions and aggregates transactions per customer [13] to segment the customer into a certain pulse category with a confidence factor. Stream based clustering allows the customer's pulse category or associated confidence factor to change as the customer continues to buy more products.
This algorithm is regularly trained using the updated historical buying patterns data (b) and real time feedback loop which is mentioned in step 5 below.
4. **Alerts + Pulse Class**, in one of the implementations, is fed to customer service representatives and other marketing algorithms, for immediate actions like real time campaigning [16], present customized offers and discounts [17] and a number of other use cases e.g. Optimize shelf stock planning [18].
5. **Feedback** is the inputs provided by the customer during the visit and is used to train the ML algorithm to improve its accuracy [19]. This could be acceptance, rejection or ignoring of the offers presented to the customer related to the pulse identified for his visit. If the customer accepts an offer which is related to his estimated pulse then that acts as a training data set for Pulse estimator model and improves its accuracy over time. On the other side if the customer rejects or ignores the offer that could be inferred as Pulse estimator could be calibrated further.

4. CHALLENGES & FUTURE WORK

This research addresses the approach to estimate the Shopper pulse from a data analytics perspective. However presents a few challenges from a technology solution & data governance perspective which can be addressed in the future research work. Some of the challenges, to cite a few:

1. **Capex Investment:** Real time monitoring & Pulse classification would require technology infrastructure at the store. IoT enablement of merchant stores to allow real time data capture, streaming and receiving the recommendation offer events from campaign management solutions based on the Pulse Estimator. A case study of digital signage-based online stores [11] provides the technical solution for data streaming and analysis for online stores and [10] proposes a scan & go solution for physical stores to capture the data and receive the recommendation offer events. Use of scan & go devices could eliminate the need of major investment in store infrastructure and at the same time will allow identification of customers and potentially allow joining up of customer online and store journeys history.
2. **Technology maturity** of the business: Pulse Estimator engine can be implemented in a hub and spoke model using AI at edge solution [20] to minimize data transfers cost and reduce latency in the pulse estimation. Once pulse is defined on the edge, real time offers can be pulled from central campaign management solutions hosted in the hub.
3. **RoI (Return on Investment)**: In order to improve the accuracy of the pulse of the customer, greater emphasis on manual product categorization is required to allow better correlation/affinity of products which in turn will improve the accuracy of the pulse of the customer.
4. **Determining accuracy**: As the estimation of the pulse of the customer is temporary and is supposedly valid for a specific customer visit only, insight derived out of this research should be separately and carefully studied prior to leveraging it with the organization's CRM & Retail strategy [21].

4.1 Future work

1. The alerts generated either due to detection of increase in volume buy and/or pulse generation can be further used to feed systems like Consumer Buying Agents [22]
2. Pulse estimator algorithm proposed in this paper currently considers customer segmentation into a single pulse category however, technically it will be further

refined to segment the customer into multiple pulse categories or combination of them to allow better micro segmentation. The algorithm has a scope to further refine by adding micro customer segmentation, market basket analysis for product affinity [9]. This shall increase the accuracy of the algorithm.

5. CONCLUSION

A shopper has a different mood, every time that he enters a shop. The outliers, when correlated with historic data patterns at both micro and macro level, yield interesting insights. This unique method correlates the real time feeds and converts them to meaningful insights. It shall help businesses increase customer centricity and also in enhancing their targeted real time marketing strategies. The outputs of this method can be further refined for high quality by bringing together other analytical solutions and data mining methods. The useful customer insights of this method can be leveraged to enhance other analytical studies within the businesses. The technology options of cloud & open source pave way to a wide spectrum of opportunities to explore the implementation of this method.

REFERENCES

1. Outlier Detection: Applications and Techniques Karanjit Singh and Dr. Shuchita Upadhyaya. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012
2. Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000)
3. UCI Machine Learning Repository: Online Retail Data Set
4. What is Product Categorization
5. Ml-knn: A Lazy Learning Approach to Multi-Label Learning, Min-Ling Zhang, Zhi-Hua Zhou
6. Applied Webscraping in Market Research, Herrmann, Markusa and Hoyden, Laurab, Universitat Polit`ecnica de Val`encia, Val`encia, (2016)
7. Multi-Label Classification: An Overview, Grigorios Tsoumakas, Ioannis Katakis Dept. of Informatics, Aristotle University of Thessaloniki, 54124, Greece
8. Using Confidence Values in Multi-label Classification Problems with Semi-Supervised Learning, Fillipe M Rodrigues and Araken de M Santos and Anne M P Canuto (2013)
9. Consumer Buying Pattern Analysis using Apriori Association Rule by Dr.V. Srinivasa Kumar,Dr.R.Renganathan,Dr.C.VijayaBanu,Iyer Ramya (International Journal of Pure and Applied Mathematics Volume 119 No. 7 2018, 2341-2349)
10. The rise of Scan and Go technology and how it works by Julian wallis (2017)
11. Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores Alfian, Ganjar and Ijaz, Fazal (2019)
12. Unsupervised real-time anomaly detection for streaming data, Subutai Ahmada Alexander Lavina Scott Purdya Zuha Aghaab (2017)
13. Customer Segmentation Based on Transactional Data Using Stream Clustering, Matthias Carnein and Heike Trautmann (2019)
14. An Evaluation of Data Stream Clustering Algorithms Stratos Mansalis, Eirini Ntoutsi, Nikos Pelekis and Yannis Theodoridis (2018)
15. Time Series Clustering: A Superior Alternative for Market Basket Analysis, Swee Chuan Tan And Jess Pei San Lau (2013)
16. The influence of real-time marketing campaigns of retailers on consumer purchase behavior, Safura M. Kallier (2017)
17. Putting One-to-one Marketing to Work: Personalization, Customization, and Choice, Neeraj Arora and Xavier Drèze and Anindya Ghose and James D. Hess And Raghuram Iyengar (2008)
18. Hübner, A., Schaal, K. Effect of replenishment and backroom on retail shelf-space planning. Bus Res 10, 123–156 (2017). https://doi.org/10.1007/s40685-016-0043-6
19. Design of a recommendation system based on collaborative filtering and machine learning considering personal needs of the user, V. Lytvyn, V. Vysotska, V. Shatskykh, I. Kohut, O. Petruchenko, L. Dzyubyk, V. Bobrivetc, V. Panasyuk, S. Sachenko, M. Komar (2019)
20. Why the Retail Industry Needs to Utilize the Power of Edge Computing By Nick Shaw (2020)
21. Style and Statistics: The Art of Retail Analytics By Brittany Bullard (2017)
22. When Algorithms Go Shopping: Analyzing BusinessModels for Highly Autonomous Consumer Buying Agents Michael Weber1, Marek Kowalkiewicz, Jörg Weking1, Markus Böhm1, and Helmut Krcmar (2020)

BIOGRAPHIES

Srikar Chilakamarri is an advisor and practitioner of business analytics for over 2 decades. He specializes in digital transformation, data analytics & organization management. He has led many successful data management programs and is an inventor with patents in this space.

Ankur is a seasoned Data & AI leader who advises clients in data transformation & modernisation, AI and data monetisation initiates. He has led the presales and solution functions and managed delivery of large scale transformation programmes. He is an inventor with a patent in this space.