ProfitAero

Jaykumar A Kakkad
jayk@gatech.edu

Priyank Sharma
psharma349@gatech.edu

Sanjay P Josh
sjosh3@gatech.edu

Aswin Nagarajan
anagarajan38@gatech.edu

1 Problem Definition

Our objective is to create an intuitive and interactive ML based web application for Small and Medium Business(SMB) retailers. Since SMBs are unable to collect large data, we propose to build this web tool based solely on transaction data. The descriptive and prescriptive analytics capability that we provide will help the SMB gain insights into customers and product market base.

URL: 'ProfitAero' Webtool

2 Background

There are enterprise tools from companies like SAP which take care of Business management and data insights for retail companies. Firms like Nielsen and IRI also provide product and customer level analytical solutions on syndicated sales data for portfolio optimization and price pack architectures. Typical cost structure for such tools ranges from $400,000 to $1,500,000. This is a hefty solution in terms of investment, requires significant amount of data and is beyond the reach of a small enterprise company.
As per UN [17], SMBs account for 98-99% of total enterprises and 44-80% of employment. With this web tool, we plan to introduce SMBs to the power of analytics and help them address a few low hanging opportunities related to customers and products.
Our discussion with few SMBs suggest that they are hesitant to invest in analytics due to upfront costs and lack of awareness. With this low cost solution, we expect SMBs to overcome their inertia and appreciate analytics. Also, SMBs will be able to take data driven decisions related to customers and products.

3 Literature Review

Companies in the online retail ecommerce space use several techniques to derive consumer insights. Non-parametric models like Random forest[1] are used to predict the purchase propensity in specific customer segments[16] and later aggregated across possible outcomes to give Customer Lifetime Value(CLTV)[3]. Use of Recency, Frequency and Monetary[3],[16] values as features for prediction is also common. The algorithms used in all above approaches vary from K-means algorithm on RFM features [4] to hierarchical models[18] to complex ANN architectures [5] which calculate the customer churn[3] at any specific time range. Some algorithms also use distribution of the customer survival or hazard function [2].
However, an important drawback with some of the approaches mentioned is their reliance of a diverse set of features[1],[2],[18] which includes behavioural metrics like clickstream, contact history and reason, etc. Most SMBs lack the resources to collect and maintain this diverse set of data at scale. SMBs also hesitate adopting solutions that are not user-friendly and has a steep learning curve.
There are three sections to our project:

3.1 Market Overview

The features in this section will help users to get a overview of key business indicators in various Geographies. Visualization on sales trends are provided in [13] which we can use in our project too. However, this visualization ignores macro trends that we plan to show to set a context.

3.2 Customer Analytics

With bipartite graphs[6][8] with two types nodes - product and customer; we run a risk of overcrowding the UI which can hinder the usability of the application. Louvian algorithm[7] and dimensionality rediction techniques like tSNE[19] could be used to detect communities within this graph to show clusters of similar customers. We can use the extracted features[5] for optimizing the algorithm[7]. There are many tools[8] to visualize relationship between the nodes in the graph[6]. Our approach here, is to rely on access-able data for a business i.e. invoice and sales history to generate

relevant features out of them [5] and utilize feature selection using chi square [2]. Existing research shows use of naive models [4] may lead to underfitting while predicting customer survival. Hence, we plan to experiment on a range of algorithms to model customer churn by also utilizing macro features of the economy.
Designing a system for large datasets involves usage of dedicated graph databases[8]. One of the works proposes a design for calculating customer churn that also involves feedback pipeline and continuous learning [1] on incoming data. Due to the complexity of this design [1], we also aim to make the system more efficient by refreshing only the data at specific intervals but keeping the algorithm constant for longer duration.

3.3 Product Analytics

Two features in product analytics are:
Cross selling opportunities: There are mainly two approaches to finding the association rules that can help cross selling. One approach is based on Minimum Spanning tree [12] that creates graph of all products. Even though the approach is visually appealing, the user may find it tedious to find rules and hence is not be useful for our project. Paper from [9], [10] describe an Apriori algorithm based on support, Confidence and Lift that is used today to find association rules. This is useful because the algorithm is efficient. However, the algorithm is not visually appealing. To make this algorithm interactive and intuitive, we can combine it with visualization techniques proposed in [11].
Sales Driver : A sales driver model helps find out relative importance of the key factors impacting sales. Some approaches to find features are found in [14] and [15]. These approaches use algorithms like KNN regression and may not work for our data which has less attributes. High uncertainty of prediction and computational complexity are key limitations of these approaches. Hence, multiple regression seems a more reliable choice.

4 Methodology-Intuition

a) Unique value proposition for SMBs: We have created a capability that

- Provides Interactive and intuitive Visualization even for non-tech SMBs
- Uses minimalistic data i.e. only transactions
- Uses proven ML techniques
- Provides data driven recommendations for a profitable business

b) One of a kind platform: Our research shows no tool in the market that uses just transnational data to provide customer and product level insights over a common platform. We are combining multiple algorithms in the space to generate insightful Consumer and Product level insights in the form of highly intuitive visuals. (details provide in approach section)

c) Direct Actionable Insights: Only by using the transaction level data the tool is able to provide impactful actionable insights in customer and product analytics space. For example, the user will see flash cards under each section that provide the monetary impact of insights like:

d) Innovative algorithmic approach to Customer analytics: For calculating CLTV and Customer Churn, we have used Non-parametric techniques like Boosting and Bagging instead of logistics regression which is widely used. We created a churn probability score to classify customers at high risk of churn. Using this and average customer value, we calculated CLTV. This method led to a better performing predictive model compared to traditional approaches.

5 Description of Approach

5.1 Data Preparation

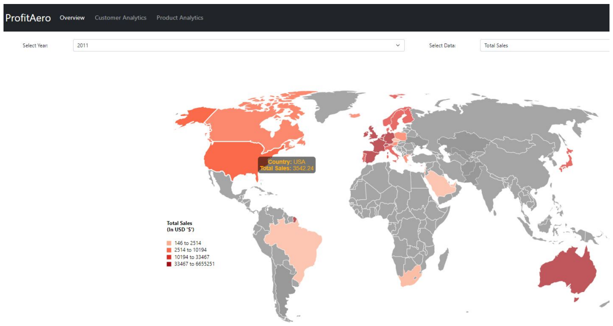
Our base data set is 261MB after combining UCI retail data(1.04mn rows) with relevant Macro data (totally 16 columns). We have created a master dataset from raw after following pre-processing steps:

- Missing values: Missing values under product description have been imputed by similar values from elsewhere to avoid reduction in data size
- Outlier Treatment: Values with sales outlier below zero have been capped (to still retain the customer purchase behavior), unusual products codes have been removed from the data
- Product Key generation: Set of EDA codes were developed to create a unique key across the complete dataset using data mining principles for product description clustering under a common code. This is to get a unique identifier for each product regardless of variations in description.

5.2 Market Overview

This tab acts as a bird's eye view to visually explore and summarize different markets. Before serving this visualization, we are combining invoice data with the

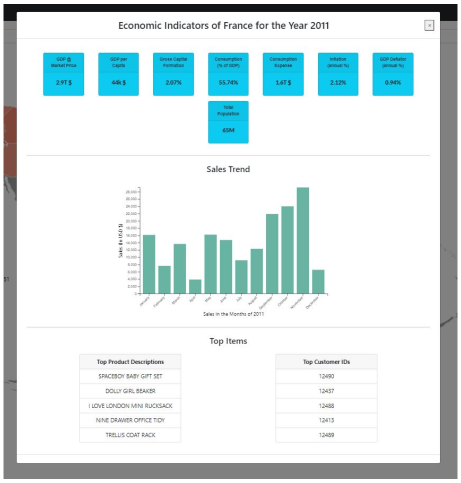
Economic Indicators of geographies. The Interactive Interface is a choropleth of the world map, with the drop down for the time periods and various metrics. When clicked on a specific Geography, a pop up shows various micro and macro details about specific market



Select Year: 2011

Select Data: Total Sales

Figure 1: Snapshot of Market Overview



Economic Indicators of France for the Year 2011

GDP @ Market Price: 2.9T $

GDP per Capita: 44k $

Gross Capital Formation: 2.07%

Consumption (% of GDP): 55.74%

Consumption Expense: 1.6T $

Inflation (annual %): 2.12%

GDP Deflator (annual %): 0.94%

Total Population: 65M

Top Product Descriptions
SPACEBOY BABY GIFT SET
DOLLY GIRL BEAKER
I LOVE LONDON MINI RUCKSACK
NINE DRAWER OFFICE TIDY
TRELLIS COAT RACK

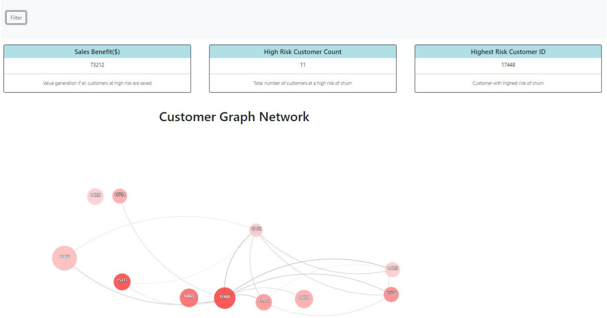
Top Customer IDs
12490
12437
12488
12413
12489

Figure 2: Popup when clicked on a Geography

5.3 Customer analytics

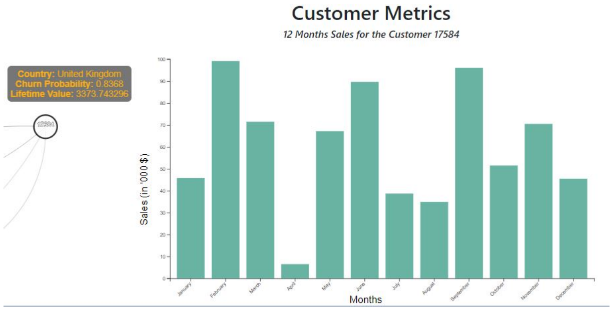
a) **Customer Churn Model**: With only the invoice data, we first focused on deriving features that could be useful determinant of customer churn. We derived various features including Recency, frequency, monetary, top products bought, countries, etc. The two year data was then divided into two parts - first 1.5yr for training data and last 6months for testing. Customers who did not buy anything in the last 6months were considered churned. Based on the experimentation, we decided to use the Gradient boosting to calculate churn probability of each customer. We used the churn probability customer revenues to calculate the CLTV

b) **Graph Network Model**: We created an adjacency matrix across all customers to visualize similar groups and purchase patterns using a weighted undirected graph network. Using the Pareto principle, we restricted the number of unique customers to only 1500 to visualize the graph. The graph weights were calculated as a similarity score of common products bought by customers. Given the size of data and over 2Million combinations to analyze, this section was coded on AWS. Using these weights, an adjacency matrix is deduced to feed as input to the graph UI. We used TSNE based dimensionality reduction to generate the X-Y coordinates to plot a customer graph. This co-ordinate system also helps in identifying communities within a graph.



Sales Benefit($)	High Risk Customer Count	Highest Risk Customer ID
73212	11	17448
Value generation if all customers at high risk are saved	Total number of customers at a high risk of churn	Customer with highest risk of churn

Figure 3: Interactive Graph and Insights



Country: United Kingdom
Churn Probability: 0.8368
Lifetime Value: 3373.743296

Figure 4: Bar chart when a customer is clicked

c) **User Interface**: Using Customer Churn, Lifetime Value and Purchase Similarity generated by the models, we created a graph network where each node represents a customer. Customers with similar buying patterns are positioned near each other to form clusters. The edges between the nodes represent the strength of the similarity on buying pattern. The size of nodes shows how valuable that customer is for the business (CLTV). The node colors are based on the customer churn probability i.e the three classes – Low, Medium and High represented by Green, Yellow and Red colors. The strength of the colors show how likely the customers can churn. The customer nodes can be clicked and hovered to reveal more details at the customer level. When a node/customer is clicked, a bar graph is displayed that shows the buying pattern of that customer for the last 12 months. For every filter selection, four actionable insights appear at the bottom of the graph.

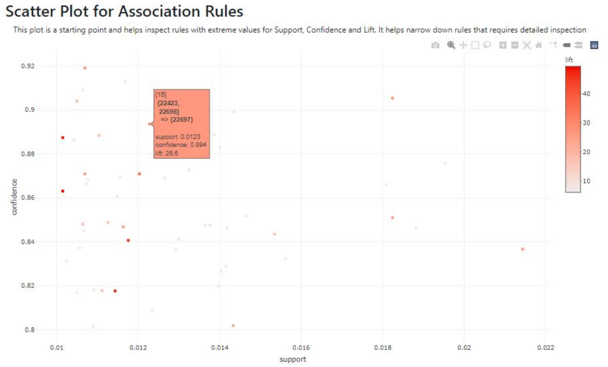
5.4 Product Analytics

a) **Association mining Model**: We implemented the Apriori algorithm to find association rules for products. We created an appropriate data structure to use the arules library in R to create rules.

b) **Sales Driver Model**: A regression-based attribution model is created to identify price elasticity of each product. Using Pareto principle, we reduced the number of unique product (1200) such that they constitute top 80% sales for the retailer. Our primary variable was price along with controlling variables for seasonal patterns, organic trends and customer density. The response variable was quantity sold on a given date

c) **Product Clustering Model**: We used item names to cluster similar products using semantic meaning of the names by treating this as an NLP problem. We used state of the art attention based transformer model RoBERTa (large) to generate sentence embeddings of the product names. This gives us a 1024 dimensional vector which we use for clustering. For EDA on the generated embeddings, we have visualized them on a 2D plane by using tSNE dimensionality reduction techniques to get the x and y coordinates. We then experimented with various types of clustering techniques: DBSCAN, K Means, Spatial Clustering, Agglomerative Clustering. For each product clusters, we assigned the most frequently occurring word in cluster as the cluster label.

d) **User Interface**: In this tab, we implemented UI for association mining. We used the arules library to create html widgets that offer interactive visualization. Two interactive visualizations i.e. scatter plot and graphs were created. The scatter plot provides visualization based on support, confidence and lift. Each bubble has a tool tip that displays the product basket. The graph helps visualize rules along with products. Each rule points to various products. When a rule or a product is clicked all associated items are displayed. The color of bubbles is based on the lift of each rule.



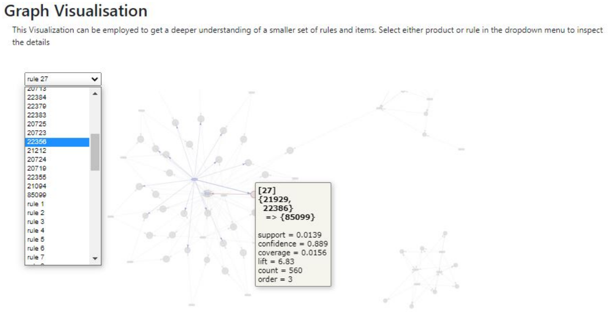
Scatter Plot for Association Rules

This plot is a starting point and helps inspect rules with extreme values for Support, Confidence and Lift. It helps narrow down rules that requires detailed inspection

[15]
{22423,
22698}
=> {22697}

support: 0.0123
confidence: 0.894
lift: 26.6

Figure 5: Scatter plot of Rules



Graph Visualisation

This Visualization can be employed to get a deeper understanding of a smaller set of rules and items. Select either product or rule in the dropdown menu to inspect the details

[27]
{21929,
22386}
=> {85099}

support = 0.0139
confidence = 0.889
coverage = 0.0156
lift = 6.83
count = 560
order = 3

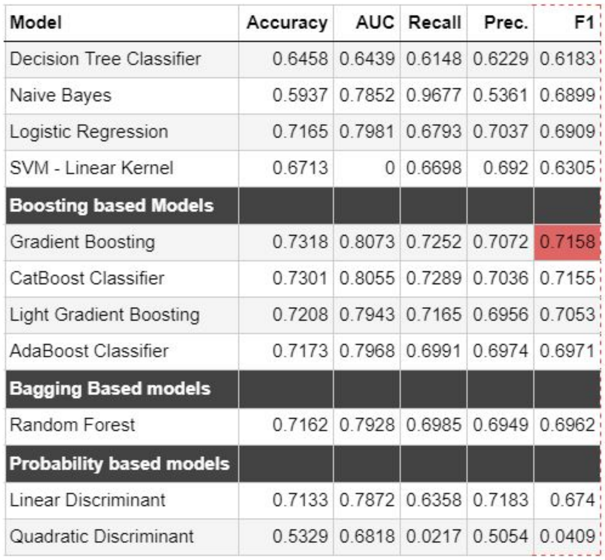
Figure 6: Interactive Graph of Rules with Products

6 Experiments and Evaluation

Our experiments were designed for following:

- Model selection: Which approach or model has the best performance?
- Parameter tuning: What are best hyper-parameters of models in terms of performance and User friendliness
- Insights: Can we derive business insights from the model implementation?

a) **Customer Churn and CLTV**: We first divided data into training (3/4) and test sets (1/4). We then fitted training data on various models and tested their performance on the test data. We used traditional models (decision tree, logistics regression,etc) boosting techniques (adaboost, gradient boosing,etc), bagging techniques (Random forest) and probability based models (QDA, LDA, Naive Bayes). We decided to use model that near highest accuracy with the best F-score. This was to ensure that False classification is given equal importance. On these parameters, the Gradient boosting model performed the best as it has the best F1 score and accuracy. We also tuned the hyper-parameters were also tuned using 10 fold cross validation.



Model	Accuracy	AUC	Recall	Prec.	F1
Decision Tree Classifier	0.6458	0.6439	0.6148	0.6229	0.6183
Naive Bayes	0.5937	0.7852	0.9677	0.5361	0.6899
Logistic Regression	0.7165	0.7981	0.6793	0.7037	0.6909
SVM - Linear Kernel	0.6713	0	0.6698	0.692	0.6305
Boosting based Models					
Gradient Boosting	0.7318	0.8073	0.7252	0.7072	0.7158
CatBoost Classifier	0.7301	0.8055	0.7289	0.7036	0.7155
Light Gradient Boosting	0.7208	0.7943	0.7165	0.6956	0.7053
AdaBoost Classifier	0.7173	0.7968	0.6991	0.6974	0.6971
Bagging Based models					
Random Forest	0.7162	0.7928	0.6985	0.6949	0.6962
Probability based models					
Linear Discriminant	0.7133	0.7872	0.6358	0.7183	0.674
Quadratic Discriminant	0.5329	0.6818	0.0217	0.5054	0.0409

Figure 7: Models tested for churn probability

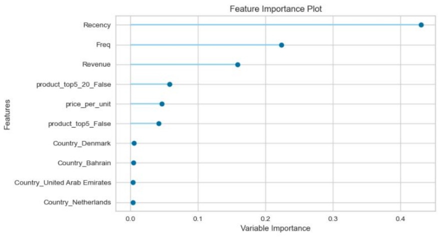


Figure 8: Feature importance for churn probability

b) **Graph Network**: The experiments involved-
1) Testing various Algorithms: TSNE dimensionality reduction and Louvain algorithm-based community detection. Experiments were focused towards generating an intuitive graph network along with reducing the run time given the size of customers pairs.
2) Comparing performance: We used a link-pair approach to identify the nodes that comprise a community and the degree of each node helps in locating valuable nodes. Run time is measured for a sample run(10,000 combinations) to compare and decide on the final algorithm before it is executed over the complete dataset(2 million combinations). TSNE was chosen as the final algorithm due to its intuitiveness to generate x-y coordinate system for graph communities and comparable run times with Louvain

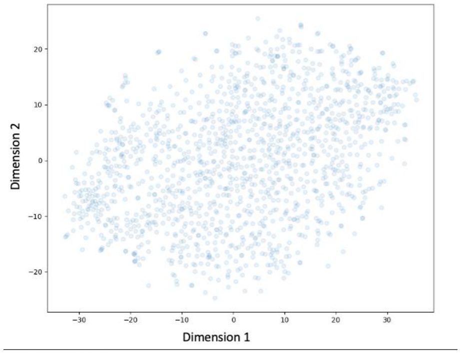
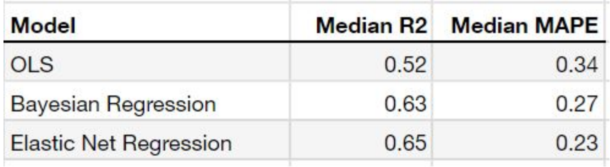


Figure 9: Customer similarity on two dimensions

c) **Sales Driver**: Experiments involved :
1) Algorithms: OLS, Bayesian Regression, Elastic Net Regression with bounded parameters were experimented upon to create the final model.
2) Comparing performance: Metrics on R2, MAPE were used to asses model performance in terms of accuracy. This was done over a cross-validated dataset to avoid bias. Bayesian Regression was chosen as a final model despite a superior performance from ENet due to the range of price elasticity. A bounded ENet had capped most of the coefficients towards lower and upper limits but the ability to provide initial set of priors in Bayesian Regression helped overcome this problem.



Model	Median R2	Median MAPE
OLS	0.52	0.34
Bayesian Regression	0.63	0.27
Elastic Net Regression	0.65	0.23

Figure 10: Performance of Sales driver models

d) **Association mining**: The aim of experiment is to tune the support and confidence metrics such that we limit the rules that are important. Hence, we found number of rules at various support and confidence levels. Finally, we decided to use confidence of 0.8 and support of 0.01 so as to display less than 100 rules that are important.

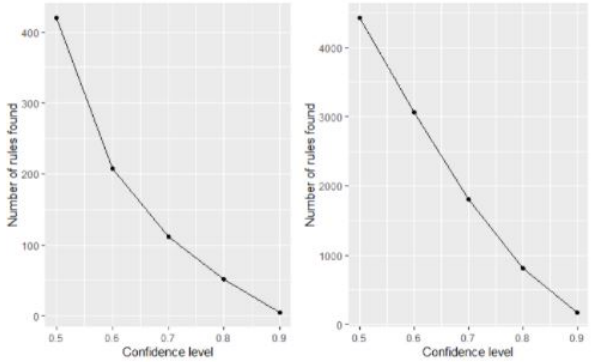
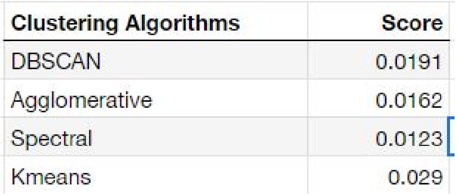


Figure 11: No of rules at Support of 0.005 and 0.01



Clustering Algorithms	Score
DBSCAN	0.0191
Agglomerative	0.0162
Spectral	0.0123
Kmeans	0.029

Figure 12: Chart of Product clustering experiment

e) **Product clustering**: To measure the performance of the various clustering algorithms used, we have used silhouette score of the results. This score ranges from -1 to 1 where score close to 1 shows that the distinction between clusters is more pronounced and -1 shows that the labels are wrongly assigned. A score near 0 shows that the clusters are overlapping.

7 Work distribution

Planned and implemented work of each Individual:

a) Jaykumar Kakkad:

- Planned: Algorithm for Customer Churn, CLTV and Association Mining.
- Implemented: Algorithm for Customer Churn, CLTV and Association Mining. Product Analytics UI and customer metric UI

b) Priyank Sharma:

- Planned: Algorithm, approaches and experiments for Sales Driver & Graph Network.
- Implemented: Algorithm, approaches and experiments for Sales Driver & Graph Network; Insights in Customer analytics tab.

c) Aswin Nagarajan

- Planned: UI for Customer Analytics; Algorithm and UI for Product Clustering
- Implemented: UI for Customer Analytics, Algorithm for Product Clustering and Final code deployment

d)Sanjay Josh

- Planned: Final code Deployment; UI for Market Overview and Product Analytics
- Implemented:Market Overview UI, Design and changes in Customer analytics UI

8 Conclusion and Discussion

We present below our conclusions based on experiments and analysis:
a) In calculating customer churn probability, we found that boosting based models yield better accuracy and F1 score compared to bagging based models and probability based models. Clearly, reducing bias (boosting) is more effective than reducing variance (bagging) in lowering the Mean squared error (MSE) in our case.
b) Using Apriori algorithm we computed association mining rules. These rules represent cross selling opportunities to the SMBs. The high confidence 50 rules we computed would have resulted into cross-selling revenue opportunity of approximately Euros 18,000 in the last two years
c) From the output of churn probability, we could identify a potential revenue of 73k Euro if the business prevents churn of high risk customers.
d) Our Sales driver algorithm helped identify products that have low elasticity to change in price. By taking price hikes in such products, we estimate additional revenue opportunity of 157K euros.

9 Future work and Improvements

We set an ambitious target even though two members of our team dropped out. Due to time constraints, we were unable to add product clustering and Sales driver features to our UI even though our algorithms were ready. Adding these two features could further enhance our web tool. Moreover, we could provide users options to tune algorithms to visualize different outputs based parameters like no. of clusters, no. of rules, etc.

References

[1] Ali Vanderveld, Addhyan Pandey, Angela Han, and Rajesh Parekh. 2016. An Engagement-Based Customer Lifetime Value System for E-commerce. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 293–302.

[2] Lu, Junxiang, Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Techniques, Proceedings of the 27th Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc., 2002.

[3] Zhang, B. and Wang, L., 2017. Application of Survival Analysis for Predicting Customer Churn with Recency, Frequency, and Monetary. [online] Support.sas.com. Available at: <https://support.sas.com/resources/papers/proceedings17/1131-2017.pdf>

[4] Nikita Bagul , Prerana Berad , Chirag Khachane , Priya Surana, 2021, Retail Customer Churn Analysis using RFM Model and K-Means Clustering, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 03 (March 2021),

[5] Smeureanu, I., Ruxanda, G., & Badea, L. M. (2013). Customer segmentation in private banking sector using machine learning techniques. Journal of Business Economics and Management, 14(5), 923-939.

[6] Zan Huang. 2005. Graph-based analysis for e-commerce recommendation. Ph.D. Dissertation. University of Arizona, USA. Advisor(s) Hsinchun Chen and Daniel D. Zeng. Order Number: AAI3168598.

[7] Lili Zhang, Jennifer Priestley, Joseph DeMaio, Sherry Ni, and Xiaoguang Tian.Big Data.Apr 2021.132-143.http://doi.org/10.1089/big.2020.0044

[8] Jo, Sunhwa, Beomjun Park, Suan Lee, and Jinho Kim. 2021. "OLGAVis: On-Line Graph Analysis and Visualization for Bibliographic Information Network" Applied Sciences 11, no. 9: 3862. https://doi.org/10.3390/app11093862

[9] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.

[10] Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining quantitative association rules in large relational tables. In Proceedings of the 1996 ACM SIGMOD international conference on Management of data (SIGMOD '96). Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/233269.233311

[11] Hahsler, Michael. (2017). arulesViz: Interactive Visualization of Association Rules with R. R Journal. 9. 163-175. 10.32614/RJ-2017-047.

[12] Valle, Mauricio & Ruz, Gonzalo & Morrás, Rodrigo. (2017). Market basket analysis: Complementing association rules with minimum spanning trees. Expert Systems with Applications. 97. 10.1016/j.eswa.2017.12.028.

[13] Van Boeckel, Thomas P., Sumanth Gandra, Ashvin Ashok, Quentin Caudron, Bryan T. Grenfell, Simon A. Levin, and Ramanan Laxminarayan. "Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data." The Lancet infectious diseases 14, no. 8 (2014): 742-750.

[14] Ali, Özden. (2013). Driver Moderator Method for Retail Sales Prediction. International Journal of Information Technology and Decision Making. 12. 1-26. 10.1142/S0219622013500363.

[15] Panay, Belisario, Nelson Baloian, José A. Pino, Sergio Peñafiel, Jonathan Frez, Cristóbal Fuenzalida, Horacio Sanson, and Gustavo Zurita. 2021. "Forecasting Key Retail Performance Indicators Using Interpretable Regression" Sensors 21, no. 5: 1874. https://doi.org/10.3390/s21051874

[16] Griva, Anastasia & Bardaki, Cleopatra & Pramatari, Katerina & Papakyriakopoulos, Dimitris. (2018). Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data. Expert Systems with Applications. 100. 10.1016/j.eswa.2018.01.029.

[17] Secretariat, U. N. C. T. A. D. (n.d.). Improving the competitiveness of smes through ... - UNCTAD. Retrieved October 15, 2021, from https://unctad.org/system/files/official-document/iteteb20051_en.pdf.

[18] C. Chen, D. Agrawal and S. Kumara, "Retail Analytics: Market Segmentation through transaction data," IIE Annual Conference.Proceedings, pp. 1034-1042, 2015. Available at:https://search.proquest.com/scholarly-journals/retail-analytics-market-segmentation-through/docview/1791990378/se-2?accountid=11107.

[19] Arora, Sanjeev, Wei Hu, and Pravesh K. Kothari. "An analysis of the t-sne algorithm for data visualization." Conference On Learning Theory. PMLR, 2018.