Project Report for CS661: BIG DATA VISUAL ANALYTICS by GROUP-04 2022-2023 Semester II

Project Title: AmazoLens: Analyzing Amazon E-commerce

Aryan Agarwal 241110012
Harsh Baid 241110026
Himalaya Kaushik 241110029
Jatin Jangir 241110031
Pritindra Das 241110054
Sahil Basia 241110061
Tanuj Agarwal 241110076
Yuvraj Raghuvanshi 241110084
IIT Kanpur

1 Introduction

We are presenting a solution in **domain of E-commerce Big Data Visualization**, and we have performed corresponding domain's Analysis techniques to solve the **problem of-Making better Strategy** as an E-Commerce Company to **boost company revenue, understand key strengths and weakness of product, identify underlying buying patterns, customer cohorts and User Sentiments from Natural Language Based reviews collected.**

We present ourselves as Big Data Consultants to an E-commerce Company, Amazon, and present our analysis done using Exploratory Data analysis, Machine Learning Models and Visualization Techniques to arrive at strategies, suggestions and improvements to boost their overall sales and enhance user experience. We had their transactional data (available in open source dataset on Kaggle link to which is given in project proposal), and customer reviews data along with product details. We build a Dashboard for presenting our analysis and to serve as a tool for Product, Data Science and Engineering Team of Amazon to help them understand and strategise better, while providing an enhanced visualization experience of the entire data.

From technical perspective, we performed exploratory data analysis, data cleaning and building a backend with this data hosted on a database to serve into our APIs written. With these we used React and D3 library to build engaging plots, along with certain plots made during Machine learning modeling part using Python. Code link is present in the code section for

same.

Key Tasks and Techniques Used: Exploratory Data Analysis and Visualization for given demographics, product and sales data,**Sales Forecasting** Region-wise distribution of several key metrics filtered by date ranges and product categories, **Market Basket Analysis** using Metrics like Confidence, Lift, **Association Rule Mining, Apriori Algorithm, FP-growth Tree, Topic Mining using Latent Dirichlet Allocation and Hierarchical Density Based Clustering (HDBSCAN)** on Transformer Based Embeddings (BERT) of User Reviews, **UMAP dimensionality reduction, Natural Language Processing techniques, Sentiment Analysis Product Recommendation system using collaborative Filtering** and Item-Item Similarity and Item-User Similarity approaches to build Sankey Plots, Product Affinity Networks, **Customer Segmentation using RFM** (Recency Frequency and Monetary) analysis, **Cohort analysis** to identify patterns and issues in marketing strategy and reporting overall key performing metrics for Sales Transactional Data in an engaging way.

2 Details of Tasks in Dashboard

For different aspects of data, we have divided the dashboard in multiple segments. Each segments provides an analysis and insight. Here we have described each segment, charts explanation, **reasoning for why that chart and analysis method is used**, benefits yielded and **insights driven out for ultimate problem solving of Visual Analytics Problem we solved** using this project. **Details of ML Based models is also explained in this.**

2.1 Customer Segmentation

Customer segmentation is technique of dividing customer data into groups that reflect similarity among customers in each respective group. The goal of this process is to decide how to relate to customers in each segment in order to maximize the value of each customer to business.

Its very important to understand our customer base. Platforms like Amazon who cater to millions of customers with diverse preferences, purchasing patterns, and loyalty levels, need to treat different group of customers differently. This is where customer segmentation is needed now, which we implemented below.

2.1.1 Customer Segment Analysis

1. **Technique Used**

   For this project, we applied the RFM (Recency, Frequency, Monetary) analysis technique to segment Amazon's customers based on their purchases :

   - **Recency (R):** How recently a customer has made a purchase (in days).

- **Frequency (F):** How often the customer makes a purchase(frequency of purchase).
- **Monetary (M):** How much money the customer spends on purchases.

The RFM values were calculated using transaction sales data. Customers were scored on each metric, and a combined RFM score was calculated. Based on these scores, we segmented the customers into distinct groups such as *Champions*, *Loyal*, *Promising*,*At Risk*, and others. The customers with high frequency and low recency scores were labeled as *Champions*. Those with high recency and low frequency were identified as *At Risk*. And others in the middle as appropriate.

2. **Analysis and Insights** We plotted them on a bubble chart to visualize the different customer segment and analyzed which customers are more valuable to the company. Figure 1 shows customer segments by recency, frequency, and monetary value. Bubble size represents the number of customers of each segment. Colors show different segment categories. Some insights we get from the chart are as follows:

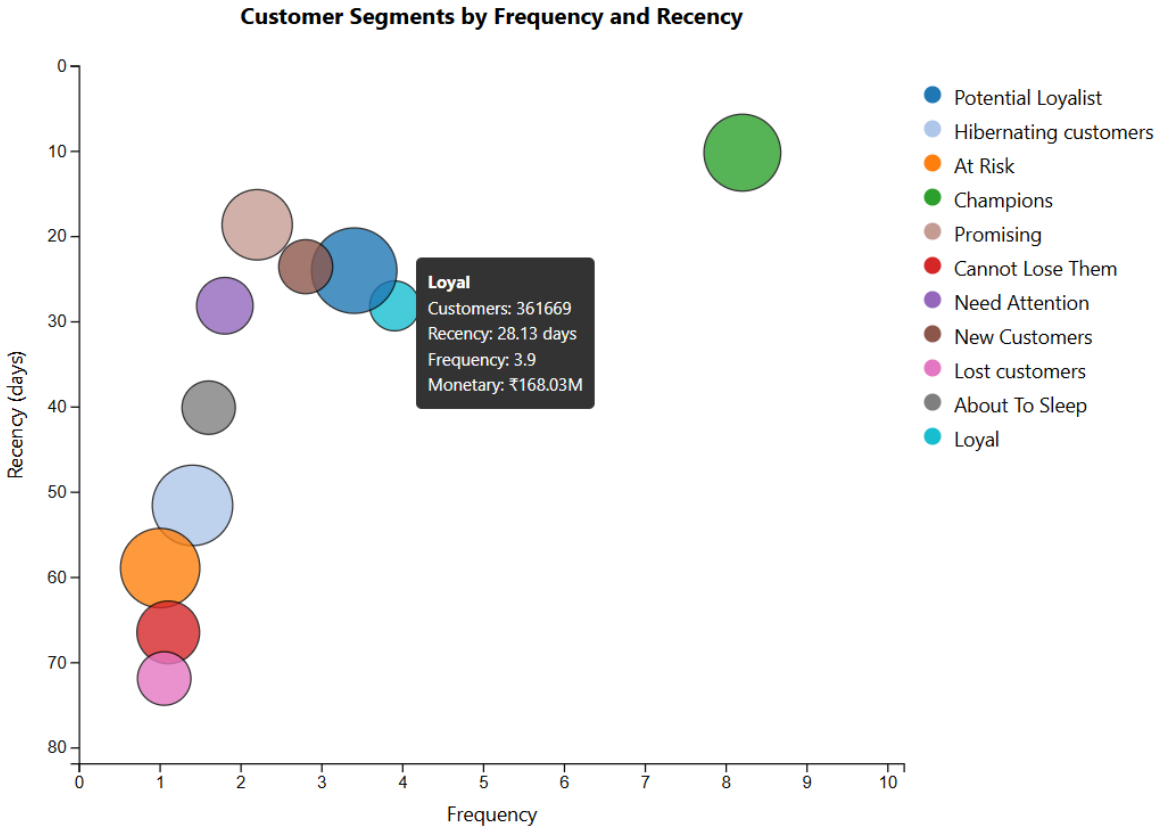

Figure 1: RFM Bubble Chart

- **Champions:** The most valuable customer segment with over 1.1 million customers. They purchase frequently (8.2 times on average) and have the lowest

recency (10.15 days). They also contribute the highest monetary value (₹357 crore+). these customers are ideal for new plans and benefit.

- **Loyal Customers:** With 3.9 purchases on average and over ₹168 crore in spending, this segment of 360K+ users represents a stable base. They may not buy as frequently as Champions but consistently return.

Overall, this segmentation allows Amazon to personalize marketing, and target the right customers with the right messages.

2.1.2 Cohort Retention Analysis

Cohort retention analysis is process of tracking user engagement with the platform over time. For this analysis, the customers were grouped based on their purchases. We tracked the weekly purchase history for each category of the products. The no. of active customers were recorded to plot the heatmap showing the customer retention.

Analysis Figure2 shows the cohort retention analysis for the *Cold Drinks & Juices* category. This reveals to us that customer retention during the late March showed around 80% in the first few weeks but it starts dropping which shows the challenges in maintaining customer loyalty.

2.2 Regional Sales

1. We made a regional sales distribution map , that automatically showcases dense region of key metrics - Total GMV (Gross Merchandise Value), Total Customers, Total Products Sold, Total Orders Made, Average Order Value for a selected time range.

2. Using this map a user can zoom in and out to see what region is performing well or not, and can make better Marketing and Sales Strategies and Policies to enhance these metrics based on patterns visible. They can analysis if certain region has any affinity for particular product type or is responding better based on certain Marketing Campaign done during selected time, and overall impact at any granular level, thus offering deeper insights.

3. We made this plot using folium Library in Python, using custom Tile Layers and Aggregation function for Marker Clusters. Figure 3 shows an example of it.

4. We also plotted top 10 states and corresponding revenue, to filter down options into a particular state as well, and use the map to zoom in and see which all regions are performing well or not using both charts simultaneously.

2.3 Market Basket Analysis

Market Basket Analysis (MBA) is technique that discovers frequent patterns from past purchases and helps us make upselling and cross selling strategies. It extract customers purchasing habits, that allow companies to make decisions as to how product positioning

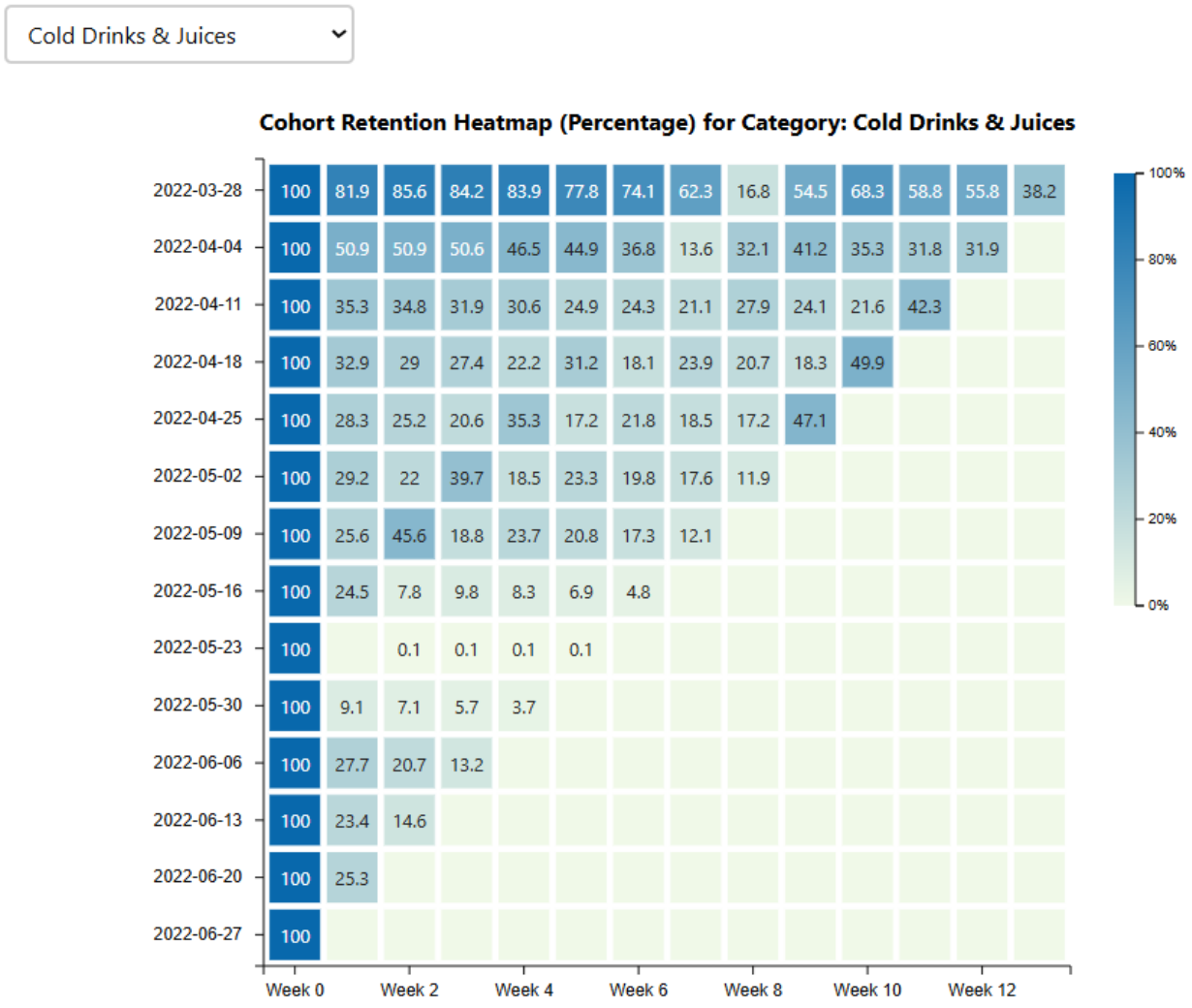

Figure 2: Cohort-Analysis Heatmap for a specific category of product

can improve customer experience, cross-selling, and marketing strategies and campaigns that leads to better sales and revenue.

2.3.1 Why Market Basket Analysis

- We wish to help amazon arrange items in particular order and combinations in stores or online stores, making it easier for customers to find what they need and improve seeling of more products.
- Companies can create more effective promotions, like offering discounts on complementary items, ultimately making customers to buy more.
- Keep popular product combinations in stock, and reducing risk of missed sales or overstock.

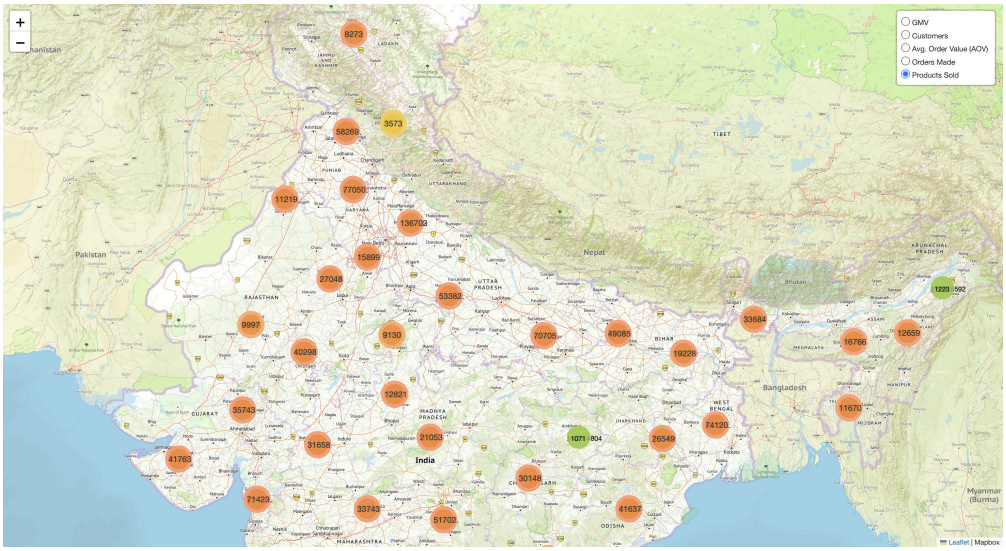

Figure 3: Regional Metric Distribution Map- With different metrics as option, and feature to zoom in and out extreme as per need for analysis

2.3.2 Key MBA Metrics

MBA uses **association rule** mining with the following fundamental metrics:

- **Support:** Measures how frequently an itemset appears in transactions

$$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$$

- **Confidence:** Indicates the likelihood of item Y being purchased when item X is purchased

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

- **Lift:** Measures how much more likely Y is purchased when X is purchased, compared to its general purchase probability

$$\text{Lift}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

- We will be making use Apriori and FP-Growth Tree Algorithms to get this data.

2.3.3 Market Basket Analysis Graphs and Their Significance

1. **Frequently Bought Together List (Fig. 4):**
    - It shows the top product combinations customers likes to buy together, with statistics like **confidence** (how often they're bought together), **support** (how common the combination is), and **lift** (how much more likely they are to be bought together than separately).

2. **Product Associations Network (Fig. 5):**

   - A graph of products as node. Here edge shows how often these two nodes(product) are bought together.
   - This makes it easy to see how products are connected and make better decisions in placements of these in counters/shelfs or online banners.

3. **Product Co-Occurrence Heatmap (Fig. 6):**

   - Here shades of colors show how product pairs are popular together. Unlike the Frequently Bought Together List, this one works on pairs only.

4. **Sankey** Purchase Flows Diagram (Fig. 7):

   - This flowchart keeps a track of flow of items base don how customer selects and buys items. It resembles a story line mapping of products and how they are bought. It also reveals how customers make decisions. We can use this info for arranging products in particular order or suggesting items during online checkout.

5. **FP-Tree Visualization (Fig. 8):**

   - A compact tree for shopping patterns. The FP-Tree (Frequent Pattern Tree) compress a lot of transaction data into a neat structure that highlights common item combinations without loss of useful informations.
   - It's useful for handling huge datasets. Instead of storing and analysing through every transaction, this tree quickly shows recurring patterns.

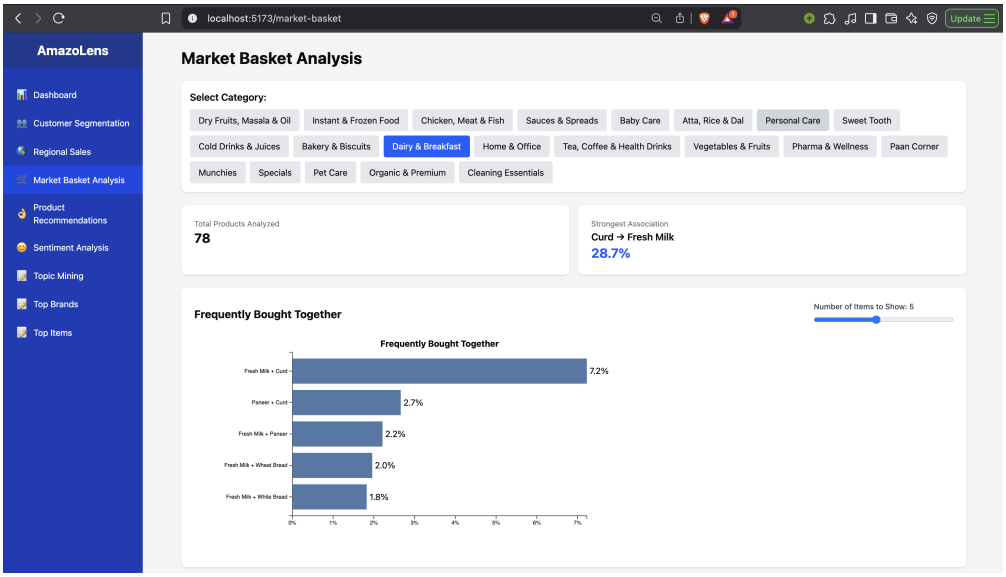

Figure 4: Frequently Bought Together analysis showing top product associations and their strength percentages

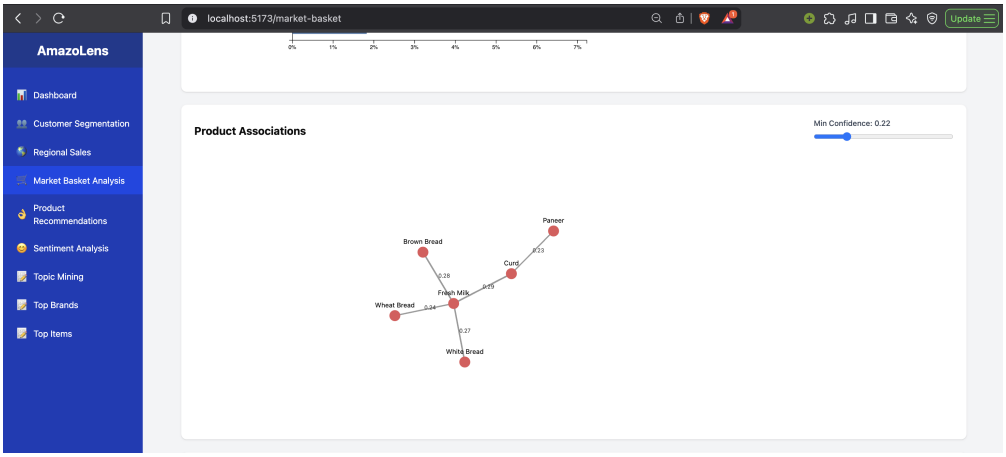

Figure 5: Product Association Network revealing connections between related products

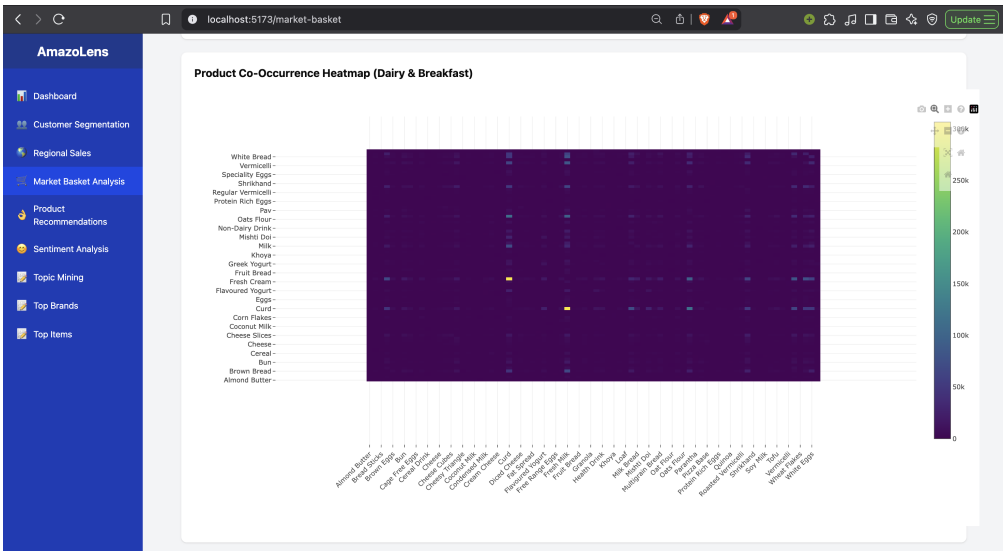

Figure 6: Heatmap analysis showing product co-Occurrence and their association

2.4 Product Recommendations

1. For each Category, we build a Product Recommendation system using Item-Item and Item-User Similarity collaborative Filtering approach. This was done based on different buying patters we observed during exploratory data analysis across categories. Our Objective to make this model is so that company can use this model to achieve better User Experience and Cross-Selling/Up-selling the product like Market Basket Analysis, and boost revenue and company growth.

2. We made the similarity models using cosine similarity on item-item combinations based on frequency patterns. For this we had build a sparse matrix for all encoded product types. To ensure model perform well, we had to aggregate at category level. We then defined a UDF to perform predictions for top N similar products to us on a selected category, and order them based on scores. This is shown in fig 9

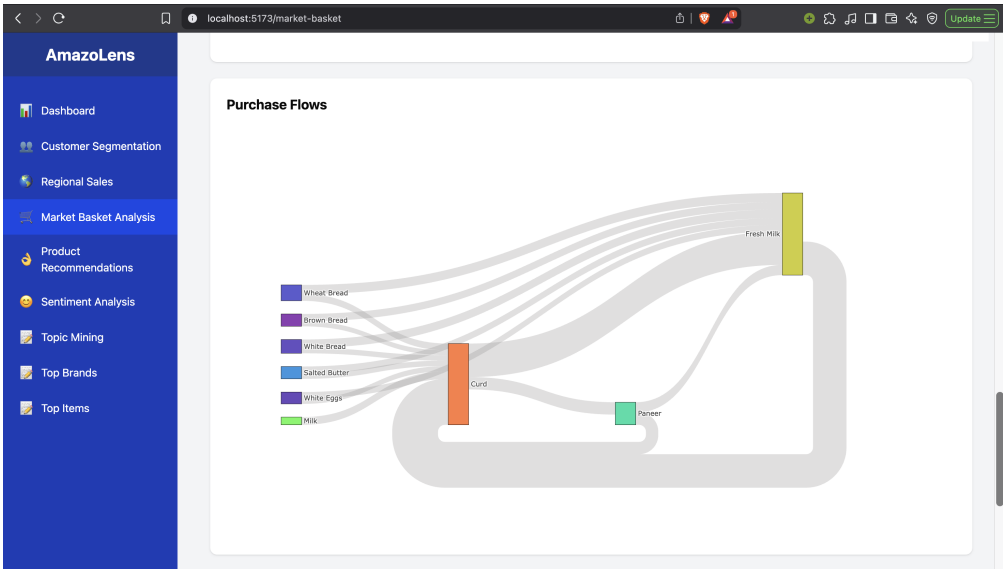

Figure 7: Sankey: Track common purchase sequences.

3. We also made a visual for the team to understand the Product Affinity Graph that shows connection between two products shown as nodes based on similarity, and magnitude of that connection is displayed by a threshold similarity value(cosine similarity) as an edge in the graph. With this we can pick a threshold, and see which products are closely related in a 3D space for better visualization, instead of tiring long tables to read from. This is shown in fig 10

4. With this approach, we wish to offer Amazon a tool improve their recommendation systems (based on whatever data we have)

2.5 Topic Mining

1. We wanted to understand what are key issues and topics the Users are talking about, and hence we used their feedbacks and reviews data to perform an Unsupervised Learning - Topic Mining. We used two techniques to do this to arrive at good topic Distributions- **Latent Dirchlet allocation** which used probabilistic approach to model word distribution within each topic and probability distribution of topics within each document or review in our case, to identify topic clusters. We also, then tested our approach to make use of word embeddings and perform **Hierarchical Density based clustering of topics (HDBSCAN)** , and to our assumptions it performed well. It showed that topics do have similar traits, and hence have a hierarchical/nested nature and using a BERT based embeddings gave powerful representation for contextual and semantic information in clustering space. for model evaluation we got following results for a **TF-IDF Based Vectorised** data with 800 vocabulary tokens- Log Likelihood (Should be high) :-142570.5263 and Perplexity (Should be low): 267.5928

2. Once the topics were identified, we mapped them to a **Intertopic Distance Map (via multidimensional scaling)** to visualize the topic cluster distribution, and see how

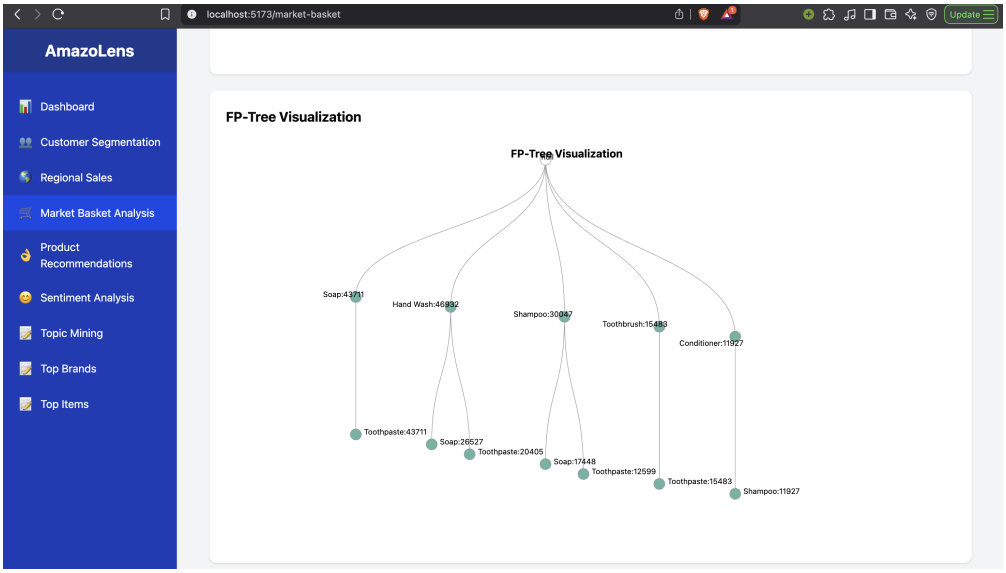

Figure 8: FP-Tree Visualization identify frequent itemsets

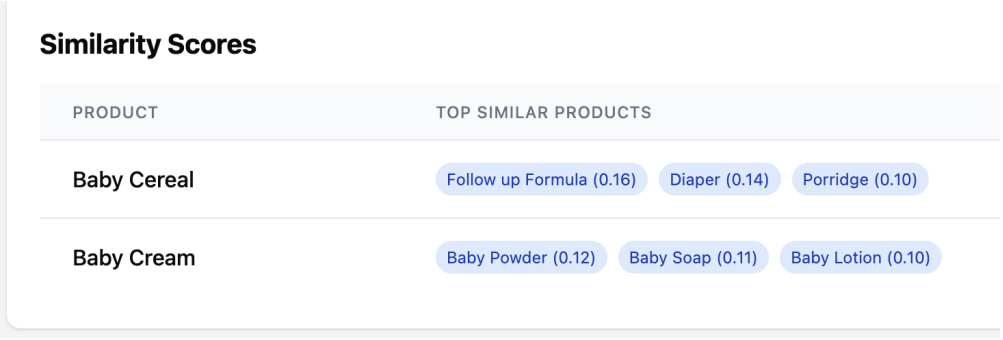

Figure 9: Product Recommendation System for selected Category showing top similar products

semantically similar they are based on the words/token distributions for each cluster. This is visualized as shown in figure 11.

3. We also plotted the reviews embedding (high dimensional) into a 3D space using **t-SNE approach**, and visualised it for seeing cluster forming and similarity patterns as seen in figure **??**. Users can also see how the **reviews are clustered in high dimensional space** with this, and hover and zoom as per need by filtering across categories. We used lower perplexity value of 5 for the model as per the number of training reviews, and 3 output dimensions to visualize a 3D plot.

4. With topics identified, we made plots to understand which product type has most issues or which products are performing well and are talked about over the time. These plots are displayed in fig 14. This was done for company employees to pick a category or all categories and see how that product is reviewed by customers over time across topics identified.

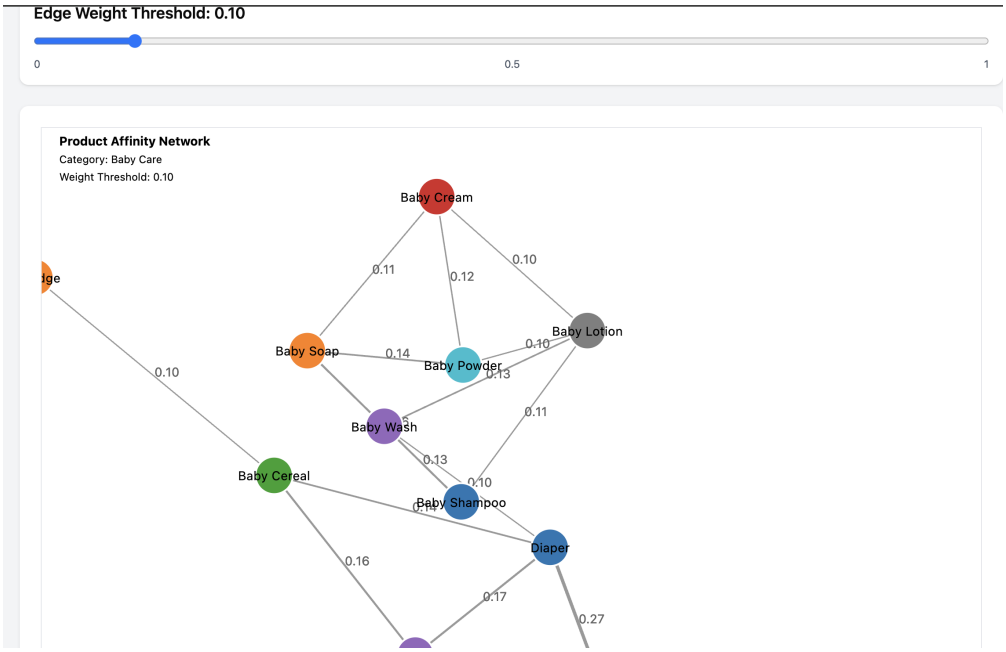

Figure 10: Product Affinity Graph Example for Selected Category and Threshold in Product Recommendation Systems

2.6 Sentiment Analysis

1. For understanding how customers react to the products, we create a **Sentiment Analysis Dashboard** , which can be really useful to see the trend of sentiment throughout the weeks and also overall sentiment for a specific category. We can use this dashboard to help business team to see early dissatisfaction in users and support marketing teams to increase a product sale.

2. We first built a classifier that labels each review as *positive*, *neutral*, or *negative*. These predictions were aggregated weekly using ClickHouse queries. The dashboard allows filtering sentiment trends by category using the `split_2_category` field, so one can visualize how sentiment varies for each category separately. At default the user will be able to see the overall sentiment across all the product

3. We added visual components to help the team easily interpret review-based insights:

    - The **Gauge Chart** displays the percentage of *positive* sentiment for the selected category. If "All" is selected, the gauge graph shows the overall sentiment score across all category. This chart helps users quickly assess sentiment score and prioritize categories that are underperforming in terms of review sentiments. This is shown in Fig 15.
    - The **Weekly Trend Line Chart** plots sentiment trend over weeks, showing peaks and drops which often show deep insight in to key events like new product releases or customer service issues. This chart supports time-based storytelling and helps correlate user sentiment with business actions. This is shown in Fig 16.

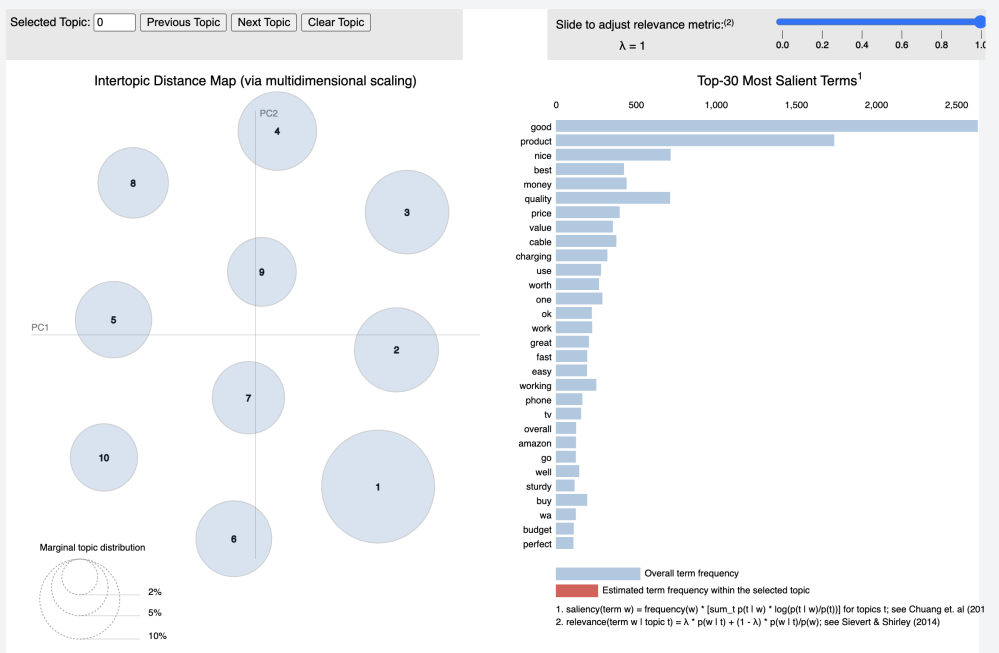

Figure 11: Intertopic Distance Map and Words Distribution Visualization for Topic Mining

4. The dashboard also displays top reviews (positive and negative) for each category. This allows business analysts or users form the company to read direct customer feedback, adding qualitative context to the sentiment scores.

2.7 Sales Forecasting

1. To assist the company decision makers, we build a Sales forecasting model as well. This helps visualize the sales trend in upcoming weeks which they can select based on the inputs they give. This helps in mitigating overstock/understock risks and enhancing supply chain planning as well. They can optimize their marketing campaigns as well with this information. Details of this model and usage is explained below.

2. **Model Details:**
   We used Facebook's Prophet time series model. It has many benefits like ease of use, robustness against missing data, seasonality and trend shifts:

   - our ClickHouse database storing weekly/daily sales quantity data along with product metadata.
   - Python backend that queries required data i.e top 50 selling products and applies Prophet model on that specific product.
   - The model outputs forecasted values with upper and lower confidence intervals.
   - The backend is given multiple filters options, allowing granularity (`daily` or `weekly`) and prediction horizon to be dynamically controlled by the user (best part).

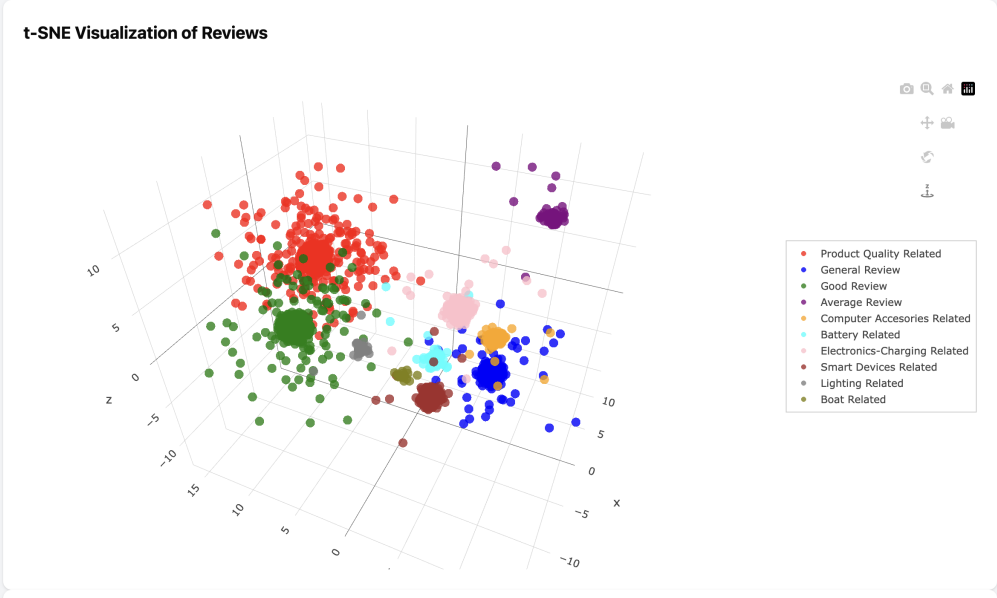

Figure 12: t-SNE Plot for the User Reviews data forming clusters of Topics in Topic Mining

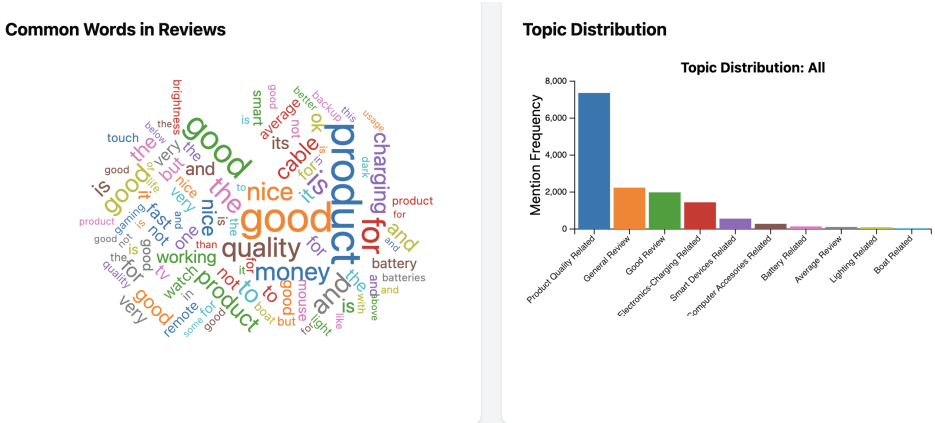

Figure 13: Charts for Analysing Topics Across Time and Products-1

3. **User Interface and Visualization:**
   Final dashboard consists of:

   - Selecting a product from top 50 selling products list.
   - Then users chooses granularity (daily or weekly).
   - they set number of days/weeks into future they want to predict.

   The prediction result is visualized using D3.js line chart (shown in Fig. 17): **Black line** shows historical sales. **Green line** shows forecasted values. **light green shaded area** represents 95% confidence interval. **Dashed green lines** mark upper and lower forecast bound.

4. **Insights and Interpretation:** This helps us understand that products with strong upward trends are going to be in more demand, so we can start planning better storage

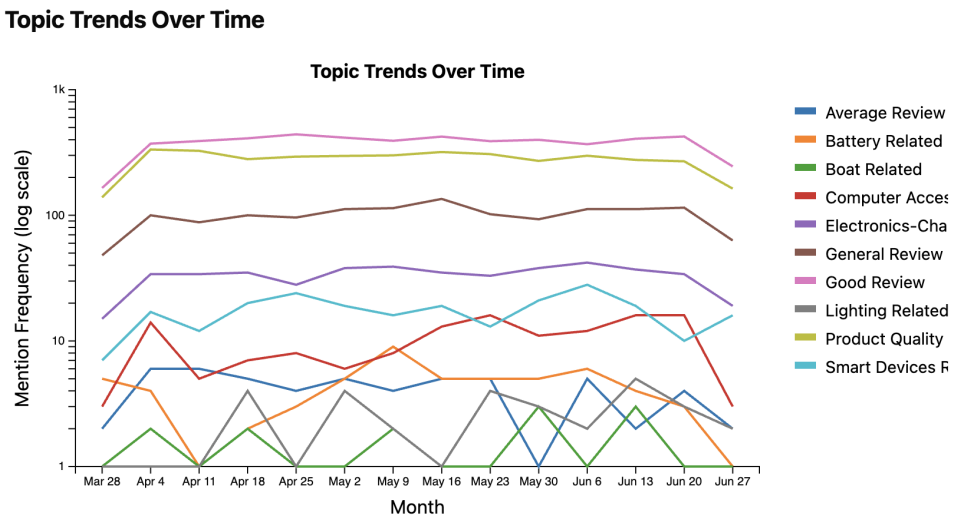

Figure 14: Charts for Analysing Topics Across Time and Products-2

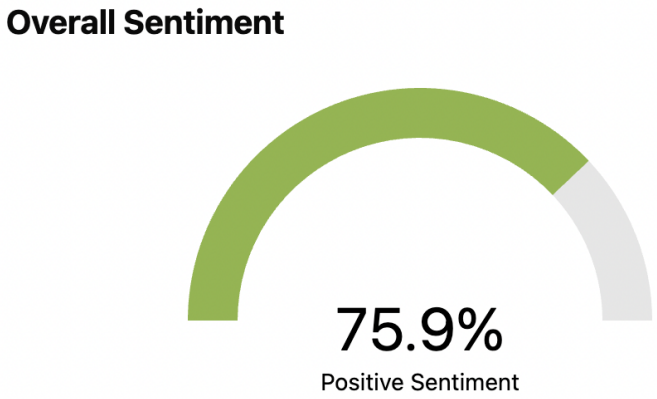

Figure 15: Gauge showing category-wise positive sentiment percentage.

for them in stock. on other hand, products with downward trend can be removed from stocks to save up costs.

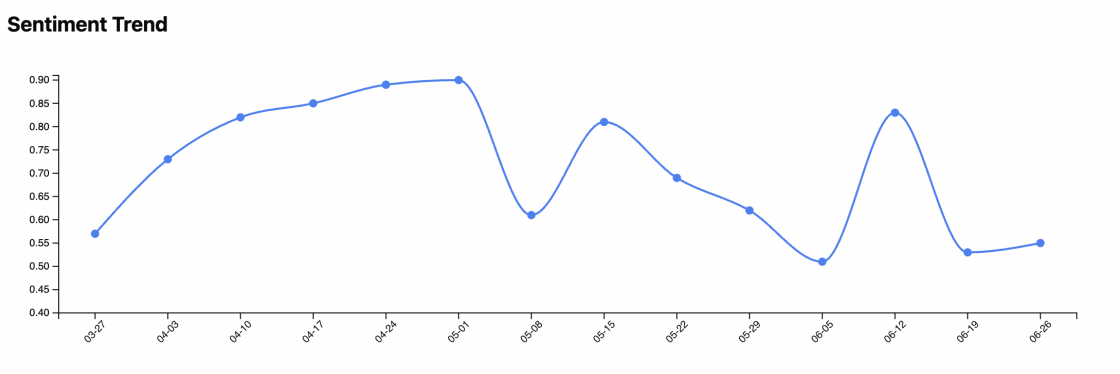

Figure 16: Line chart of weekly sentiment trends across time.

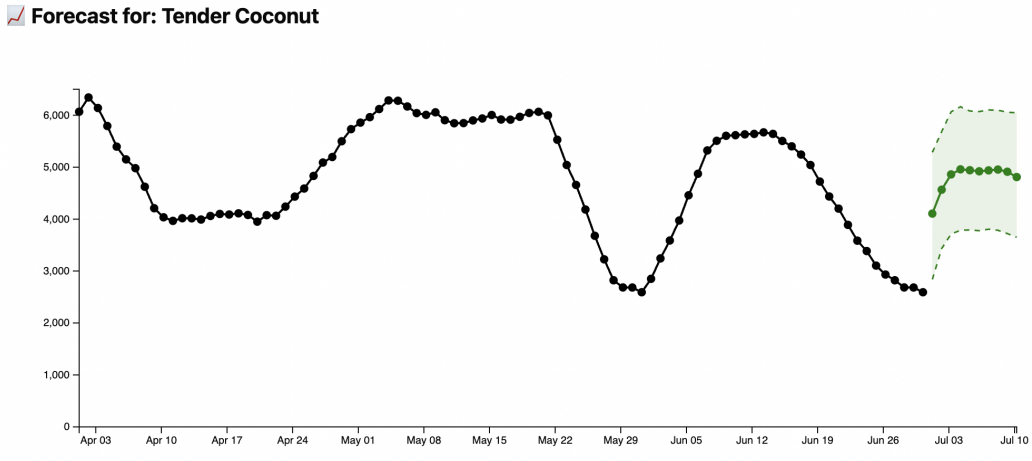

Figure 17: Interactive sales forecasting graph with dynamic controls and predicted time series using Facebook Prophet. Historical sales (black), forecast (green), and 95% confidence bounds (dashed).

3 Summary and Consultation Suggestions as final Results

Respective Tasks' results are explained in each section. With above analysis and visualization aid, we also wish to present the company following suggestion and consultations as conclusion to enhance their revenue and user experience based on our analysis of their big data:-

1. We need to focus more on increasing Average Order Value in densely marketed area by suggesting to make more marketing campaigns based on our association rules for Personal Care Category.

2. They need to target Sales in Delhi-NCR region as visible drop in this is seen from aid of the Regional chart

3. we suggest to run a coupon based sales strategy based on association rules to increase upselling and crosselling of products as listed in Market Basket Analysis and Product

Recommendation systems section

4. We see that State wise we need to perform better in multiple states like Maharashtra, Karnataka (as they have more potential for higher end products), and these top 50 selling products as pe Sales Forecasting method might be the first way to start with.

5. Topic Mining suggests these Electronics-charging Related issues are key issues in most products and we need to improve our Sellers group for this products category to work on these issues to keep our customer retention better

6. Our topic Mining Model can further be used in their Customer Queries and conflict resolutions Customer Care Automation process as well

7. cohort analysis shows that in the week of 2022-05-23 the users had a bad experience, and hence should inquire more to resolve this user group.

8. We should make use of the T-SNE plot for better visualization of upcoming reviews, and if the clusters keeps on getting dense or similar, thus reflecting certain patterns. If they cluster with Good Reviews cluster slowly, that is good indication of product improvements.

9. We Should make use of our Sales forecasting method to drive better sales and marketing campaigns.

10. **As explain in each section their use cases and enhancements, if we implement these suggestions, we can see significant improvement in Overall Sales and Customer Experience, thus helping a company grow!**

4 Link to Source Code Link and Instructions

The source code for this project is divided into two separate GitHub repositories (Machine Learning Code repository is linked in Backend Repo) :

- **Frontend Repository:** `github.com/harshbaid-13/AmazoLens`
- **Backend Repository:** `github.com/harshbaid-13/Backend_AmazoLens`

Prerequisites

- Node.js (v18 or higher) and npm
- Python 3.10+ and pip

Backend Setup

1. Clone the repository:
   ```
   git clone https://github.com/harshbaid-13/Backend_AmazoLens
   cd Backend_AmazoLens
   ```

2. Create and activate a virtual environment:
   ```
   python -m venv .venv
   source .venv/bin/activate  # For Linux/macOS
   .venv\Scripts\activate     # For Windows
   ```

3. Install dependencies:
   ```
   pip install -r requirements.txt
   ```

4. Run the backend:
   ```
   uvicorn app.main:app --reload
   ```

5. Info:

   ```
   Your application will be running at http://127.0.0.1:8000
   (if available, check logs otherwise).

   Endpoints
   GET / - Check if the server is running.
   GET /docs - API documentation (Auto-generated by FastAPI).
   ```

Frontend Setup

1. Clone the repository:
   ```
   git clone https://github.com/harshbaid-13/AmazoLens
   cd AmazoLens
   ```

2. Install dependencies:

   ```
   npm install --legacy-peer-deps
   ```

3. Start the frontend:
   ```
   npm run dev
   ```

4. Info:
   ```
   Open your browser and navigate to http://localhost:5173
   ```

Notes

Ensure the backend server is running before starting the frontend to make sure, everything runs perfectly.

5 Tech Stack Details

Machine Learning and Backend Development: We used a Python-based backend, along with Scikit-learn, Transformers, mlextend, Folium, TensorFlow-based ML libraries to facilitate our ML-driven tasks and approaches.
Frontend Development: We used NextJS and React framework with D3 and Folium Library for visualizations. We made use of different libraries, corresponding documentation, domain knowledge and material on visualization for data science.[1] [2] [3] [4] [5] [6] [7]

References

[1] M. Grootendorst. *BERTopic: Leveraging BERT and c-TF-IDF for Topic Modeling*, 2020. GitHub repository, https://github.com/MaartenGr/BERTopic.

[2] Joel Grus. *Data Science from Scratch: First Principles with Python.* O'Reilly Media, 2nd edition, 2019.

[3] Scott Murray. *Interactive Data Visualization for the Web.* O'Reilly Media, 2nd edition, 2017.

[4] S. Bird, E. Klein, and E. Loper. *NLTK: The Natural Language Toolkit*, 2009. O'Reilly Media.

[5] C. Sievert and K. Shirley. *pyLDAvis: Python Library for Interactive Topic Model Visualization*, 2014. Journal of Open Source Software.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *Scikit-learn: Machine Learning in Python*, 2011. Journal of Machine Learning Research, 12, 2825–2830.

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. *Transformers: State-of-the-Art Natural Language Processing for Pytorch and TensorFlow 2.0*, 2020. GitHub repository, https://github.com/huggingface/transformers.