BIG DATA VISUAL ANALYTICS (CS661) PROJECT PROPOSAL REPORT GROUP-04

AmazoLens: Analyzing Amazon E-commerce and AWS

1 Introduction

In today's digital age, businesses and consumers are increasingly dependent on eCommerce platforms like Amazon for products and services. To stay competitive, companies need to leverage analytics and insights from these platforms to optimize operations, improve customer experience, identify trends, and drive growth. Understanding data-driven insights allows businesses to make informed decisions, enhance performance, and stay ahead in a rapidly evolving market.

That is exactly what we aim to inculcate by performing Exploratory Data Analysis with Visualization to prepare a dashboard for Amazon E-Commerce and AWS data collected, to perform various analyses including Recommendation systems, Review Sentiment Analysis, Cohort analysis, and Vital metrics dashboards to visualize different aspects of data trends.

2 Data Sources

Various sources of raw data for our project:

- **E-Commerce Sales Dataset:** Kaggle Dataset that provides detailed insights into Amazon sales data, including SKU Code, Design Number, Stock, Category, Size, and Color, to help optimize product profitability. `https://www.kaggle.com/datasets/thedevastator/unlock-profits-with-e-commerce-sales-data`
- **Amazon AWS SaaS Sales Dataset:** Kaggle dataset that contains transaction data from a SaaS company selling sales and marketing software to other companies (B2B). `https://www.kaggle.com/datasets/nnthanh101/aws-saas-sales`
- **Amazon Sales Dataset:** Kaggle Dataset that provides user reviews for different Amazon products. `https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset`

Data Description: The datasets primarily represent the transactional data for Amazon E-commerce platform and AWS services respectively. It includes attributes like Order ID, UserID, shipping date, shipping location, user reviews and rating, profit from a particular order, revenue, etc. We aim to perform EDA on this data and derive useful features to channel them into actionable insights and strategies that can be clearly understood through visual representations.

3 Specific Tasks

For this project, we will perform the following main tasks:

- **Data Preprocessing:** Perform data cleaning, removing noise, feature engineering, removing redundant features, and data transformation for analysis.

- **Data Visualization:** Develop interactive visualizations to explore sales and customer analytics like cohort analysis, user segmentation, heatmaps/geospatial data, time series forecasting on revenue data, and key metrics of sales reporting on dashboards for efficient understanding of sales data.
- **Insights Generation:** Implement algorithms for deriving insights from the data, such as customer segmentation, recommendation systems, sales trends, user sentiment analysis, and topic mining on customer reviews to abet strategies as part of our consulting objective.
- **User Interaction:** Enable users to interact with the visualizations, customize views, and save insights.

Key Insights: Customer Segmentation and Cohort Analysis, Region-wise distribution of user base and corresponding Sales Trends across different customer/product segments to drive sales and marketing strategy. ,Product Recommendation Systems, User Review Sentiment Analysis and Topic Mining on User Reviews

4 Overall Solution

Since transactional data has a high velocity of data generation and volume, we aim to incorporate these aspects of Big Data into a Visual Analytics System. This system will handle aggregated or summarized data insights to drive a Sales Dashboard that can be used by the designated company's business teams. We aim to build a user-interactive dashboard that will facilitate users to customize views and execute the following 5 key tasks to derive insights:

1. **Customer Segmentation and Cohort Analysis:** Analyze customer data to segment users based on behavior, demographics, and purchase history. This insight helps tailor marketing strategies, personalize offers, and improve customer retention by addressing specific cohort needs.
2. **Region-wise Distribution and Sales Trends:** Examine user base distribution and sales patterns across regions and customer/product segments. It identifies growth opportunities, optimizes regional marketing campaigns, and informs product positioning to increase sales and market penetration.
3. **Product Recommendation Systems:** Build recommendation engines using customer behavior and product data. This increases cross-selling, upselling, and customer satisfaction by offering relevant product suggestions, ultimately boosting revenue and customer loyalty.
4. **User Review Sentiment Analysis:** Perform sentiment analysis on user reviews to gauge customer satisfaction and product reception. It enables real-time feedback for product improvements, marketing adjustments, and better customer support, increasing brand trust and product quality.
5. **Topic Mining on User Reviews:** Extract common topics and themes from user reviews to identify trends and issues. This helps Amazon improve product development, enhance customer service, and create more targeted marketing campaigns based on customer concerns and desires.

These tasks help optimize Amazon's sales, marketing, and product strategies by leveraging data insights and improving business outcomes. This project showcases how a subset of fast-growing transactional Big Data can be harnessed for demonstration purposes.

5 Tech Stack

Backend Development: We will use a Python-based backend, along with Scikit-learn, Transformers, TensorFlow-based ML libraries to facilitate our ML-driven tasks and approaches.
Frontend Development: We plan to use NextJS and React framework

6 Team Members and Responsibilities

While all team members plan to contribute to each task for collective learning, we will tentatively divide the workload as follows:

- **Machine Learning and Visualization Module:** Aryan Agarwal (241110012), Himalaya Kaushik (241110029), Tanuj Agarwal (241110076), Jatin Jangir (241110031)
- **Backend Development and Database Management:** Jatin Jangir (241110031), Aryan Agarwal (241110012), Yuvraj Raghuvanshi (241110084), Sahil Basia (241110061)
- **Frontend Development and Application Hosting:** Harsh Baid (241110026), Pritindra Das (241110054), Himalaya Kaushik (241110029), Sahil Basia (241110061)

Unlocking actionable insights from Amazon's data to drive smarter decisions, optimize strategies, and fuel business growth.