Predicting Customer Satisfaction for E-Commerce Dataset

Team Name: Matrix Miners

Authors:

- Manoj Kumar Yelamarthi, Rahul Rai, Mukunda Saiteja Annam and T Senthilkumar

- Emails: manojyk@iisc.ac.in, rahulrai@iisc.ac.in , mukundaannam@iisc.ac.in, senthilkuma1@iisc.ac.in

Problem Definition

The objective of this project is to analyze various aspects of an e-commerce dataset to predict customer satisfaction and derive actionable insights.

Motivation

Understanding customer satisfaction is crucial for e-commerce businesses to enhance customer experience, improve retention rates, and drive sales. By leveraging data analytics and machine learning, we aim to identify key factors influencing customer satisfaction and provide recommendations for business optimization.

Design Goals

- Accurately predict customer satisfaction based on historical data.
- Identify trends and patterns in revenue, orders, and product reviews.
- Provide actionable insights to improve business operations and customer experience.
- Ensure scalability and performance of the data analytics and modeling processes.

Features Required

- Data cleaning and preprocessing
- Feature engineering and transformation
- Machine learning model building and evaluation
- Visualization of key metrics and insights

Scalability/Performance Goals

- Efficient handling of large datasets using distributed computing frameworks.
- Real-time data ingestion and processing.
- Scalable machine learning models to accommodate growing data volumes.

Approach and Methods

High-Level Design

The project involves several stages, including data cleaning, data analysis, feature engineering, model building, and evaluation. The data is sourced from various CSV files and processed using Apache Spark for distributed computing.

Architecture/Data Model

- **Data Sources**: Multiple CSV files containing customer, order, product, and review data
- Data **Visualization: Using python libraries like seaborn, matplotlib**
- **Data Processing**: Apache Spark for distributed data processing
- **Machine Learning**: Predicting customer satisfaction using different models from SparkML

Big Data Platforms Used

- **Apache Spark**: For distributed data processing and machine learning.

ML Methods Used

- **Random Forest Classifier**: Used to predict customer satisfaction based on engineered features.
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1 Score.

Evaluation

Experiment Design

- **Data Cleaning and Preprocessing**: Removing missing values, duplicates, and transforming date columns.
- **Feature Engineering**: Calculating delivery times, on-time delivery indicators, and joining relevant datasets.
- **Model Building**: Training a Random Forest Classifier on the processed data.
- **Model Evaluation**: Using accuracy, precision, recall, and F1 score to evaluate model performance.

Scalability/Performance Metrics

- **Data Processing Time**: Time taken to process and clean the data.
- **Model Training Time**: Time taken to train the machine learning model.
- **Prediction Time**: Time taken to make predictions on the test dataset.

Feature Metrics

- **Delivery Time**: Average time taken for order delivery.
- **On-Time Delivery**: Percentage of orders delivered on or before the estimated delivery date.
- **Customer Satisfaction**: Average review scores and satisfaction rates.

Plots and Analysis

Order Analysis

The analysis indicates that most orders are placed during evening hours (4 PM to 12 AM), suggesting a preference for late-night shopping. While there is a noticeable difference in order volume between weekdays and weekends, no definitive pattern is evident. Comparing 2017 and 2018, there is a significant increase in orders, demonstrating the company's year-over-year growth. Certain states exhibit a higher percentage of delayed orders, necessitating a deeper investigation into the underlying causes. Addressing these delays is crucial to improving customer satisfaction, as there is a strong correlation between delayed orders and reduced customer satisfaction scores.

Revenue Analysis

Revenue generally increases over time, with some fluctuations. A notable spike in revenue from October to November 2017, followed by a significant dip, suggests a potential seasonal or promotional influence on sales. The spike on 24 November 2017 aligns with Black Friday, indicating a strong correlation between the holiday season and increased customer activity. To optimize resource allocation for future peak seasons, a strategic approach can be implemented to dynamically scale computing resources during high-demand periods, ensuring a seamless customer experience and maximizing revenue potential.

Product Reviews

The analysis provides valuable insights into product categories by identifying the top five best-selling products and the top five products with the highest review scores within a given region. This enables the company to provide targeted product suggestions to customers, enhancing engagement by aligning recommendations with local buying trends. Additionally, the analysis highlights products with lower average review scores, allowing the company to focus

on understanding and addressing the underlying causes of dissatisfaction. By improving product quality, the company can enhance customer satisfaction and strengthen its brand reputation.

Summary

Based on our predictive modeling, we have identified customers who are likely to discontinue their engagement with our company. To mitigate this risk and enhance customer retention, we propose the Following strategic initiatives:

1. ***Targeted Promotions:*** **Focus on providing personalized promotions to at-risk customers. By offering tailored incentives, we can improve customer loyalty and reduce churn.**

2. ***Product Recommendations:*** **Utilize data analytics to identify trending products with high review scores in the customer's state or region. These products should be highlighted and recommended to customers, leveraging their proven popularity and satisfaction metrics.**

3. ***Optimized Communication:*** **Based on previous analyses, the evening is the most effective time for customer engagement. Therefore, we recommend scheduling notifications and promotions through our app to reach customers during this peak period, maximizing the likelihood of interaction and purchase.**

4. ***Enhanced Delivery Strategy:*** **Develop a comprehensive strategy to ensure timely delivery of products to these customers. By focusing on improving delivery times, we can significantly boost customer satisfaction, leading to increased revenue and customer retention.**

Implementing these strategies will not only help retain customers but also strengthen our market position by building trust and loyalty among our customer bases.

Ability to Achieve Design and Performance Goals and Future Extensions:

As our dataset spans approximately 20 months and continues to grow, it is crucial to ensure our data analytics and modeling processes remain efficient and scalable. By leveraging Apache Spark, we can maintain robust data analysis capabilities with minimal impact on processing time, even as the dataset expands over several years. Spark's distributed computing framework allows us to employ distributed modeling techniques, significantly enhancing model performance and scalability.

Furthermore, all raw data is securely stored in a data lake. This architecture is inherently scalable, accommodating increasing volumes of data seamlessly. The data lake's flexible storage capacity ensures that as our data grows, it can be efficiently managed and accessed for analytics.

Data ingestion is facilitated by Apache Kafka, a scalable messaging system designed to handle high-throughput data streams. Kafka's ability to efficiently manage large volumes of real-time data ensures that our data pipeline remains robust and responsive, supporting our analytics infrastructure.

By integrating these technologies, we ensure that our data analytics and modeling processes are well-equipped to scale alongside our expanding dataset, maintaining performance and accuracy.