DA 231o: **Data Engineering at scale**

Presentation Slides

Team: Matric Miners

Rahul Rai, CISCO-IISC, rahulrai@iisc.ac.in
Mukunda Saiteja Annam, CISCO-IISC, mukundaannam@iisc.ac.in
Manoj Kumar Yelamarthi, CISCO-IISC, manojyk@iisc.ac.in
T Senthilkumar, CISCO-IISC, 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.

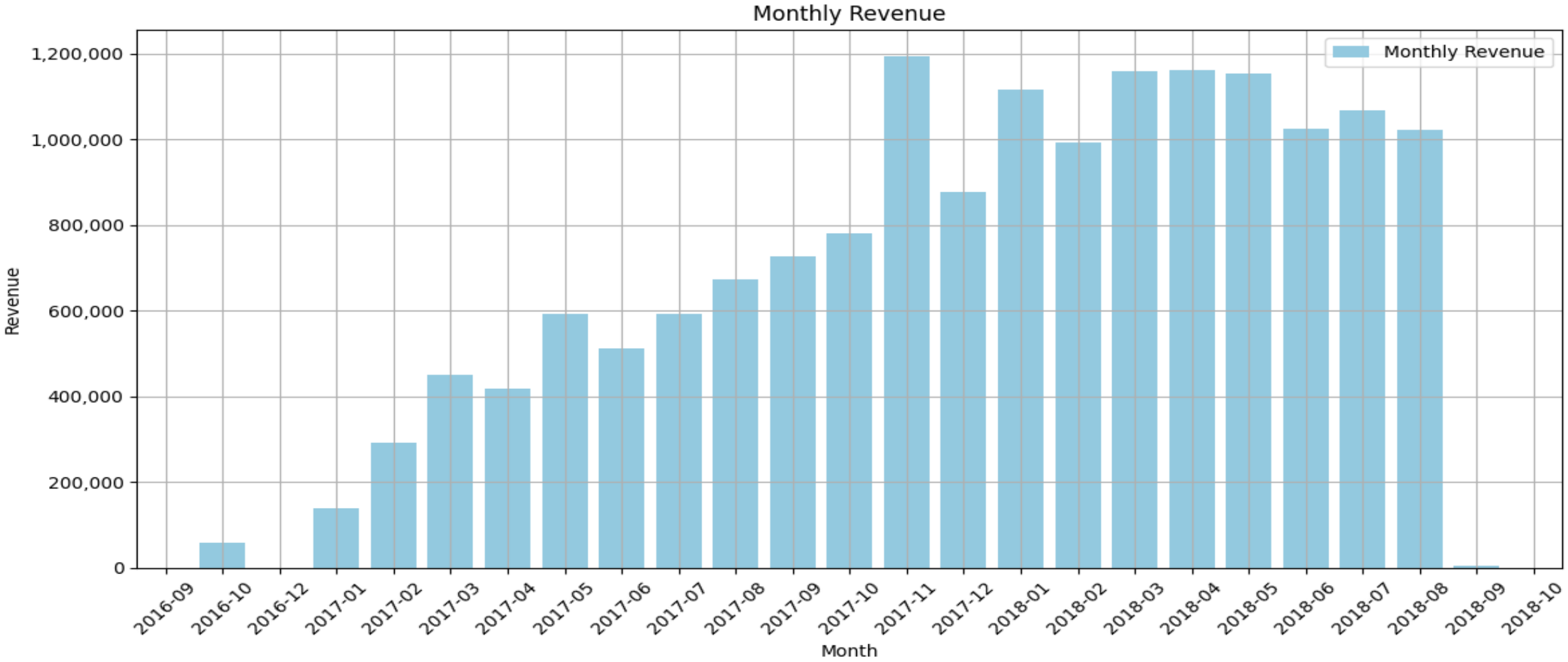
Data Collection and Preparation

- **Data Sources and Data Models:**
- **Primary Data:** **Brazilian E-Commerce Public Dataset by Olist**
- https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/
- **Data Model:** olist_customers_dataset, product_category_name_translation,olist_sellers_dataset,olist_products_dataset,olist_orders_dataset,olist_order_reviews_dataset,olist_order_payments_dataset,olist_order_items_dataset,olist_geolocation_dataset, olist_customers_dataset
- Records- Around 95K in each dataset
- Years – 2 years

Data Cleaning and Preprocessing

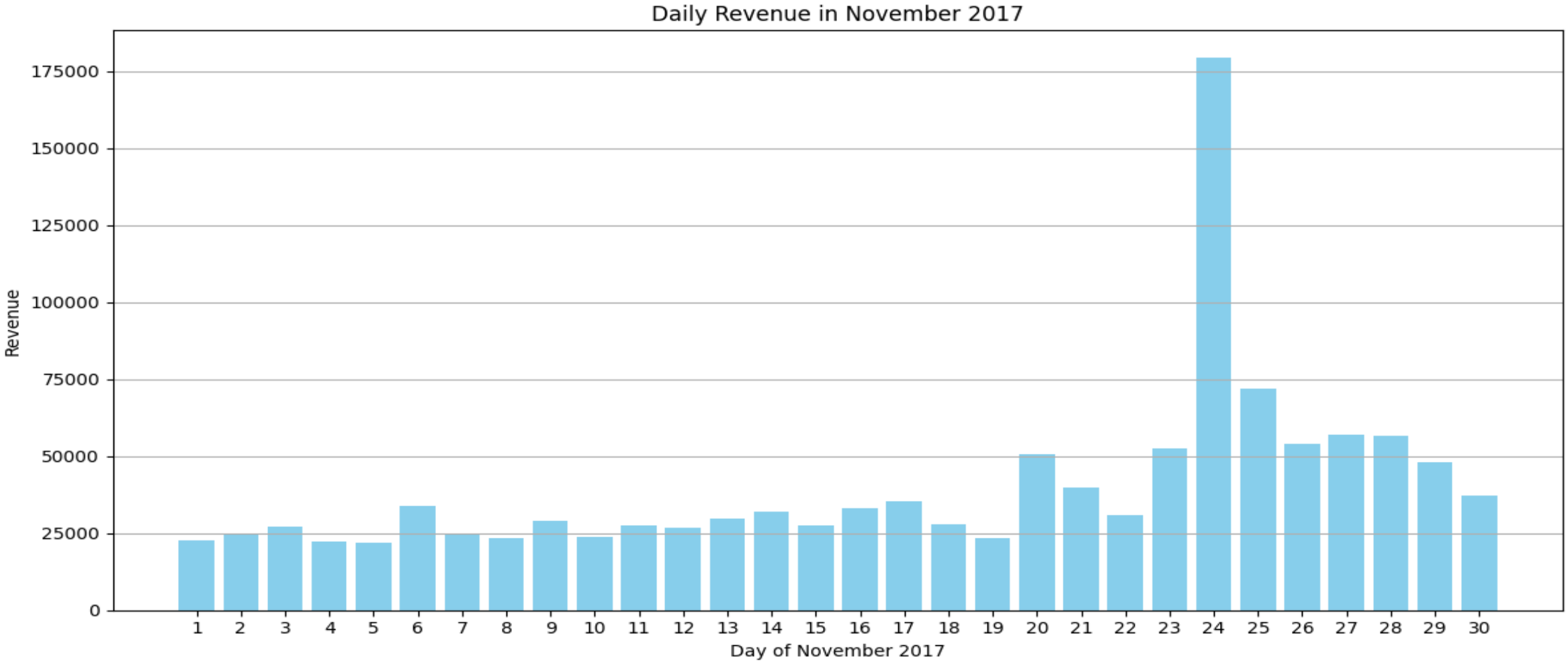
- **Removed Missing Values:** Used .dropna() to remove rows with missing values in the orders_df.
- **Removed Duplicates:** Used .dropDuplicates() on order_id in order_items_df to ensure each order ID has unique entries.
- **Filled Missing Values:** Replaced null values in the review_comment_message column with an empty string '' using .fillna().
- **Converted Date Columns:** Transformed order_purchase_timestamp, order_delivered_customer_date, order_delivered_carrier_date, and order_estimated_delivery_date columns to a uniform date format for consistency using to date.
- **Created Positive Review Indicator:** Added a column is_positive_review to identify reviews with a score of 5 and a non-empty review comment.
- **Encoded Customer Satisfaction:** Introduced a binary target column customer_satisfaction to label scores above 4 as satisfied (1), otherwise not satisfied (0).

Revenue Analysis



From the above metrics, we can observe that revenue generally increases over time, although there was some fluctuations. Notably, there was a sudden spike in revenue from October to November 2017, followed by a significant dip in the subsequent month.

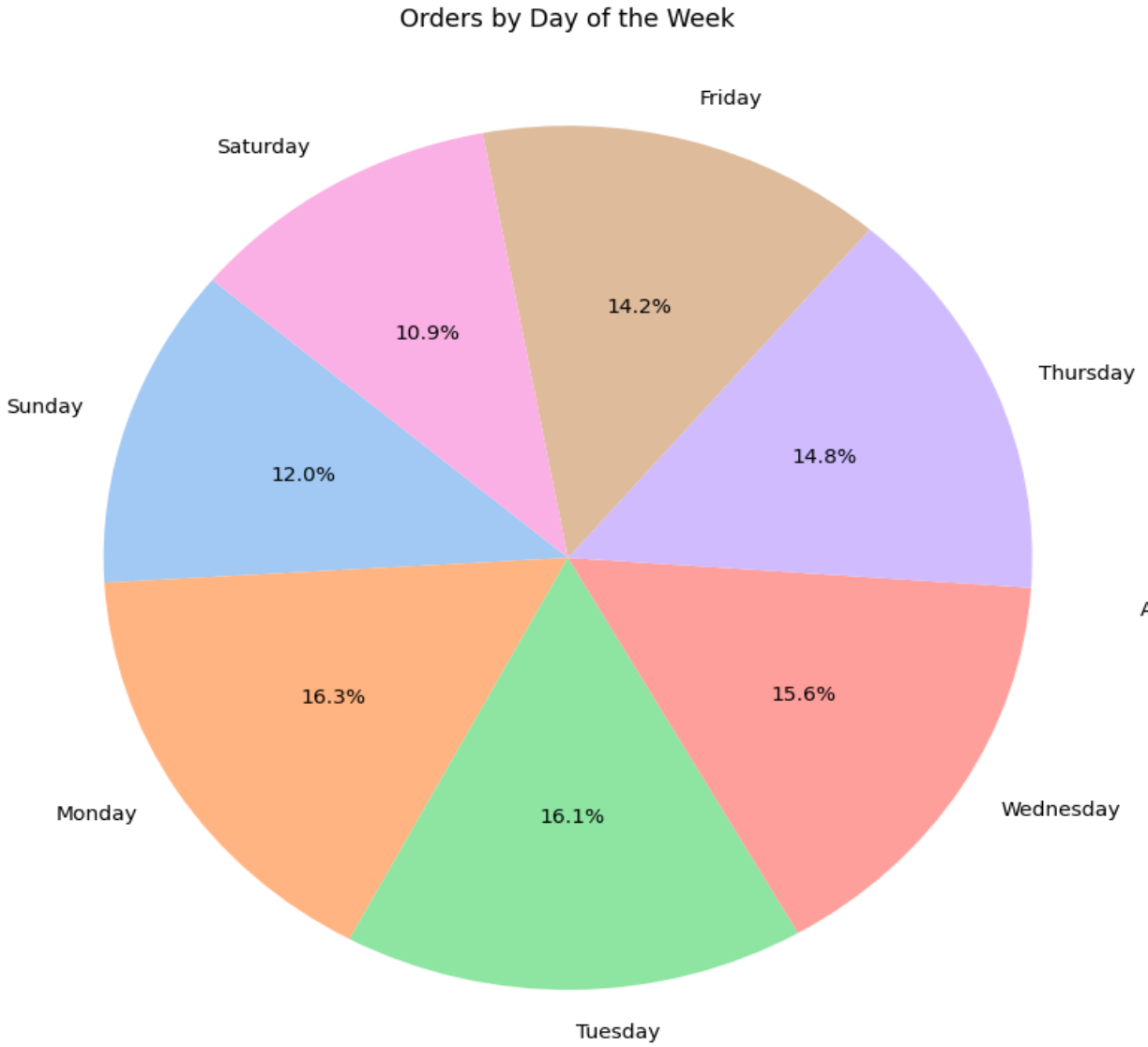
To gain deeper insights, it would be beneficial to further analyze the revenue for November 2017 on a day-by-day basis. This detailed analysis could help identify specific days or events that contributed to the spike in revenue

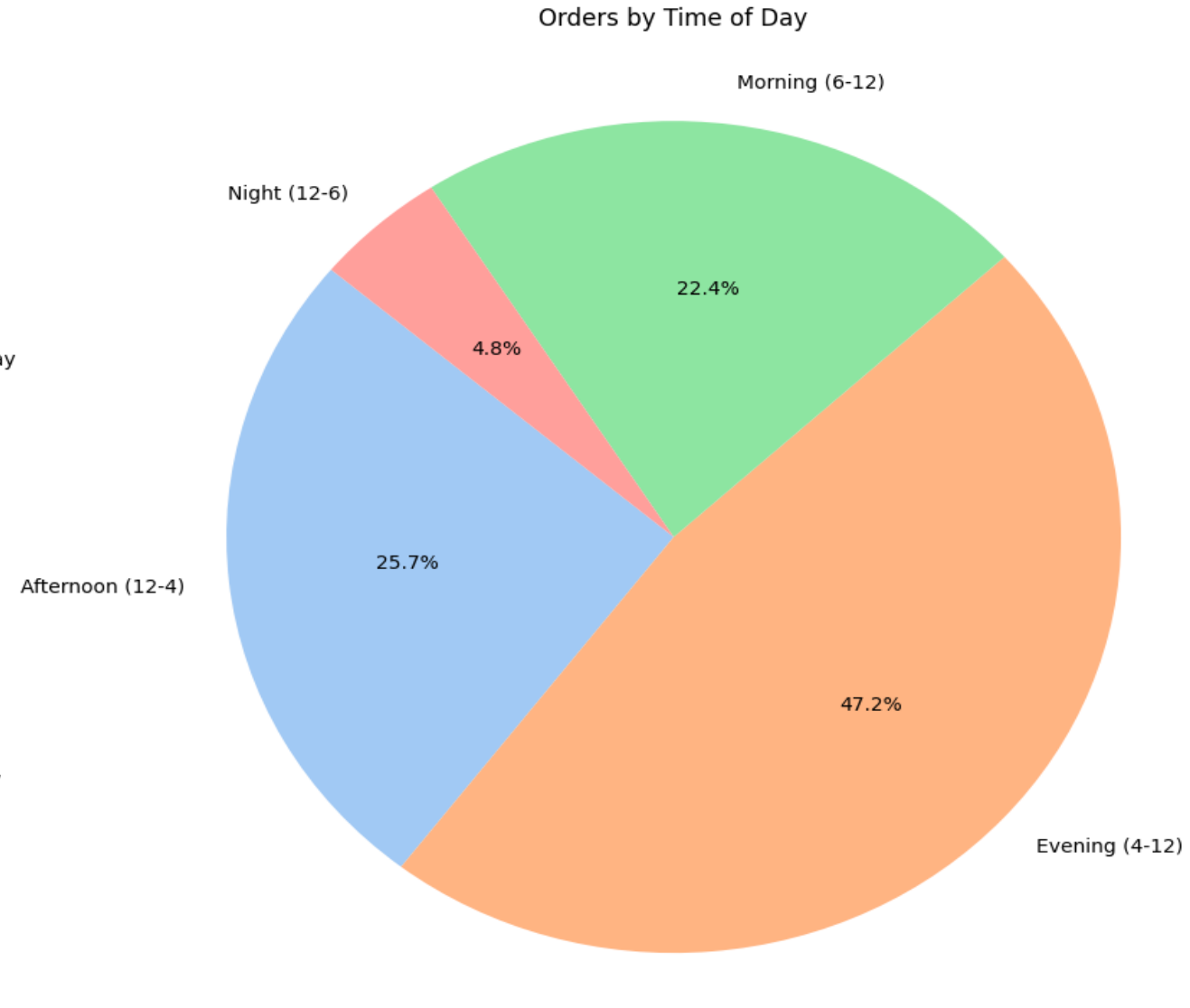


A notable spike in revenue was observed on November 24th, 2017, which aligns with Black Friday, a significant shopping event.This indicates a strong correlation between the holiday season and increased customer activity.

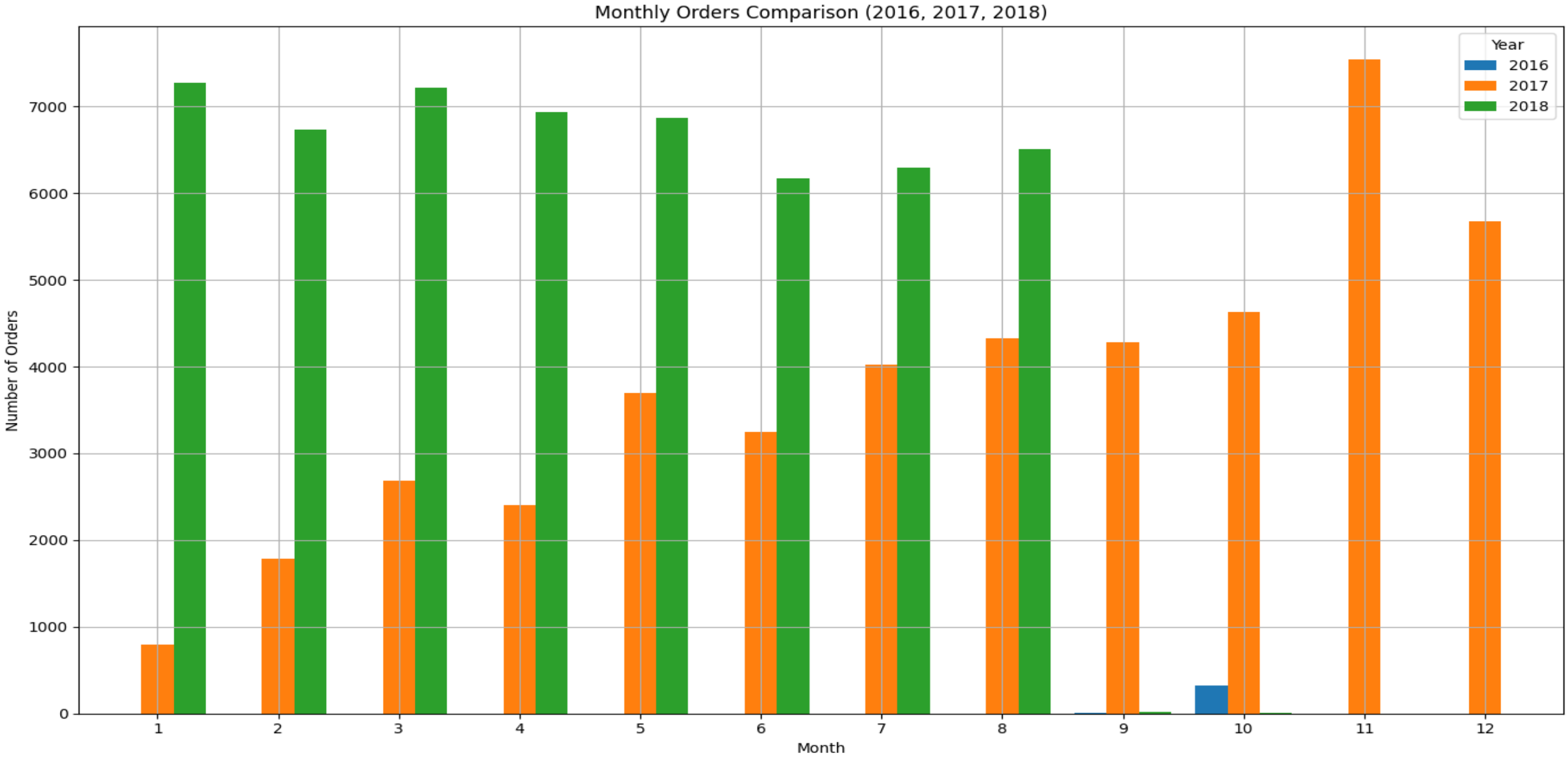
To optimize resource allocation for future peak seasons, a strategic approach can be implemented to dynamically scale computing resources, ensuring the system can handle increased traffic during high-demand periods like Black Friday and prevent performance degradation, thereby ensuring a seamless customer experience and maximizing revenue potential.

Order Analysis

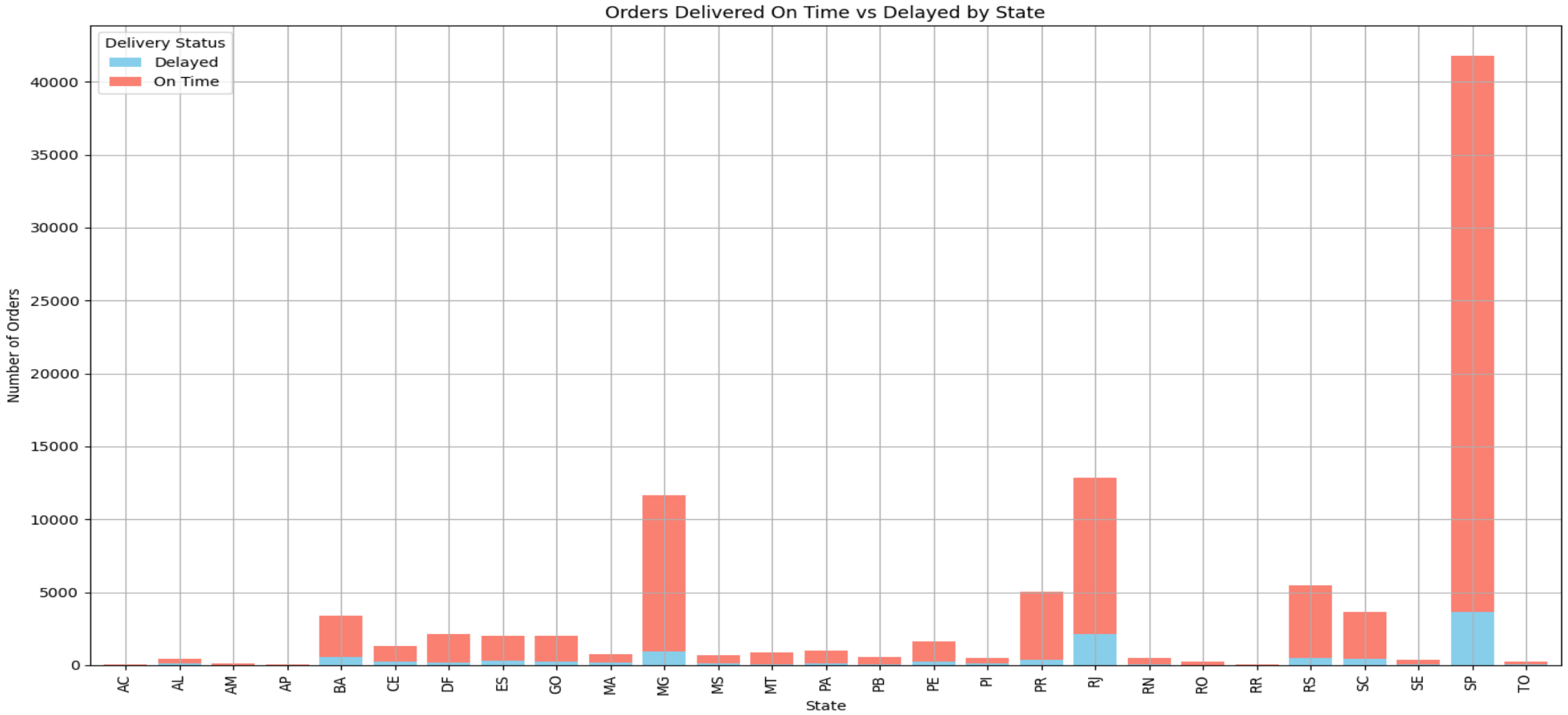




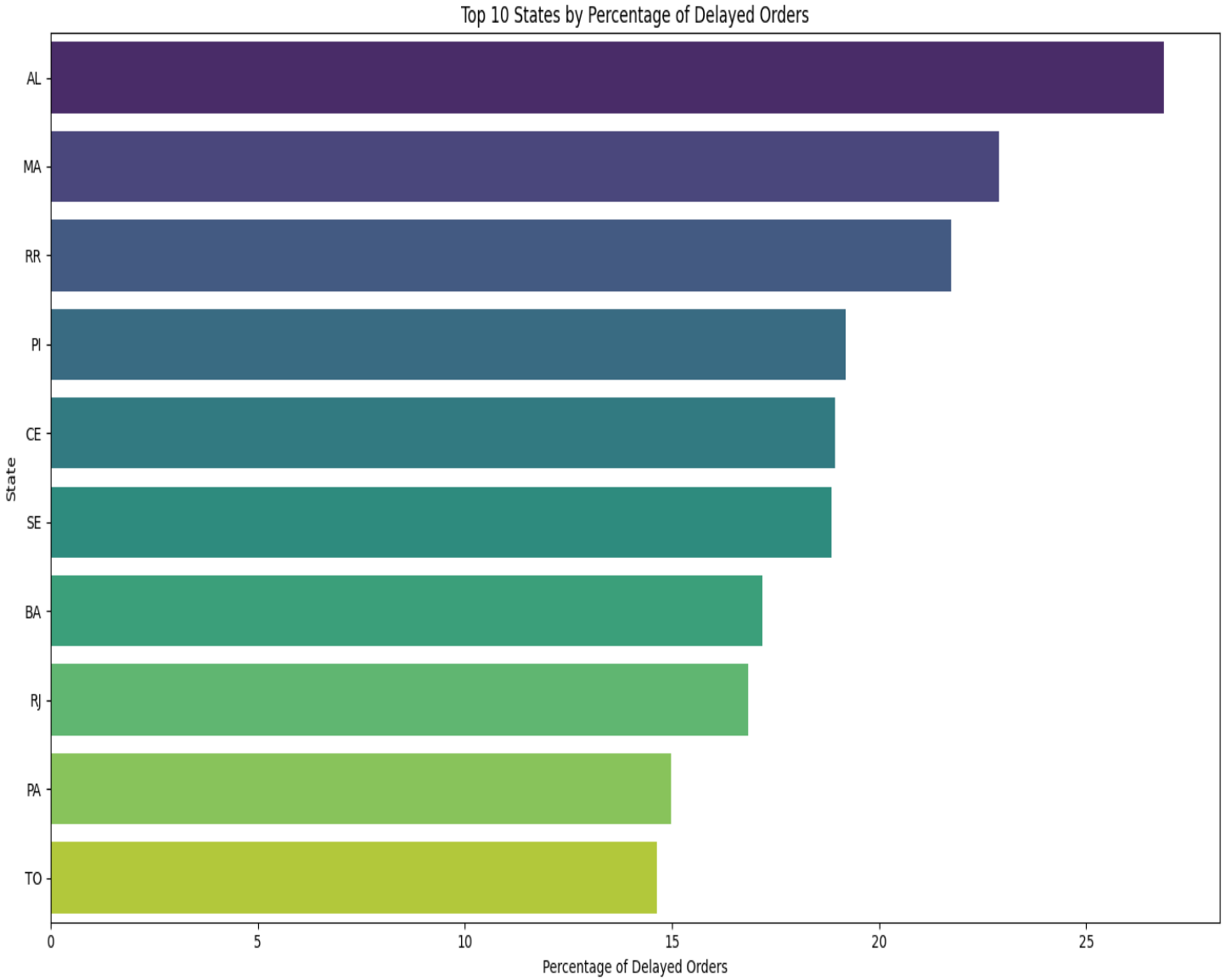
The analysis suggests that the majority of orders are placed during the evening hours (4 PM to 12 AM), indicating a preference for late-night shopping. While there is a noticeable difference in order volume between weekdays and weekends, a definitive pattern is not evident.



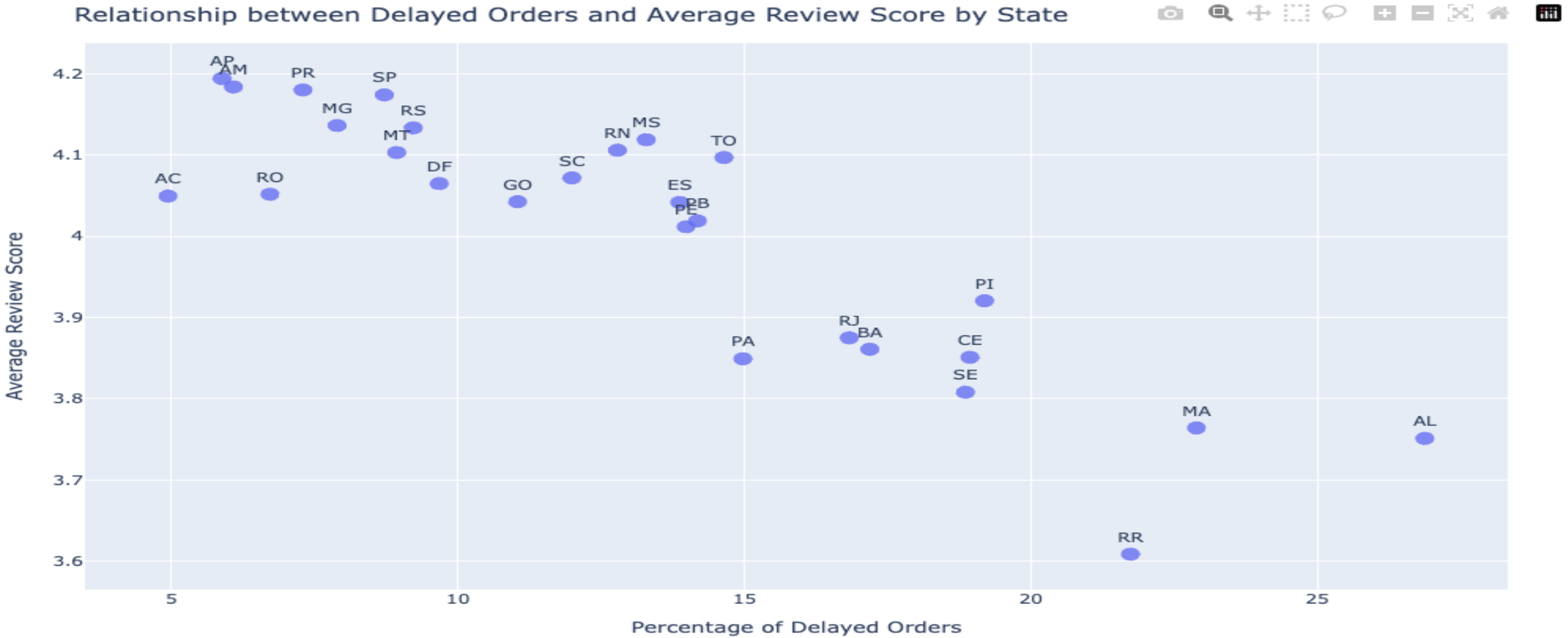
A comparison of 2017 and 2018 reveals a significant increase in orders, demonstrating the company's year-over-year growth.



Analysis of the recent data indicates that certain states exhibit a higher percentage of delayed orders compared to others. This discrepancy necessitates a deeper investigation into the underlying causes of delays in these specific regions

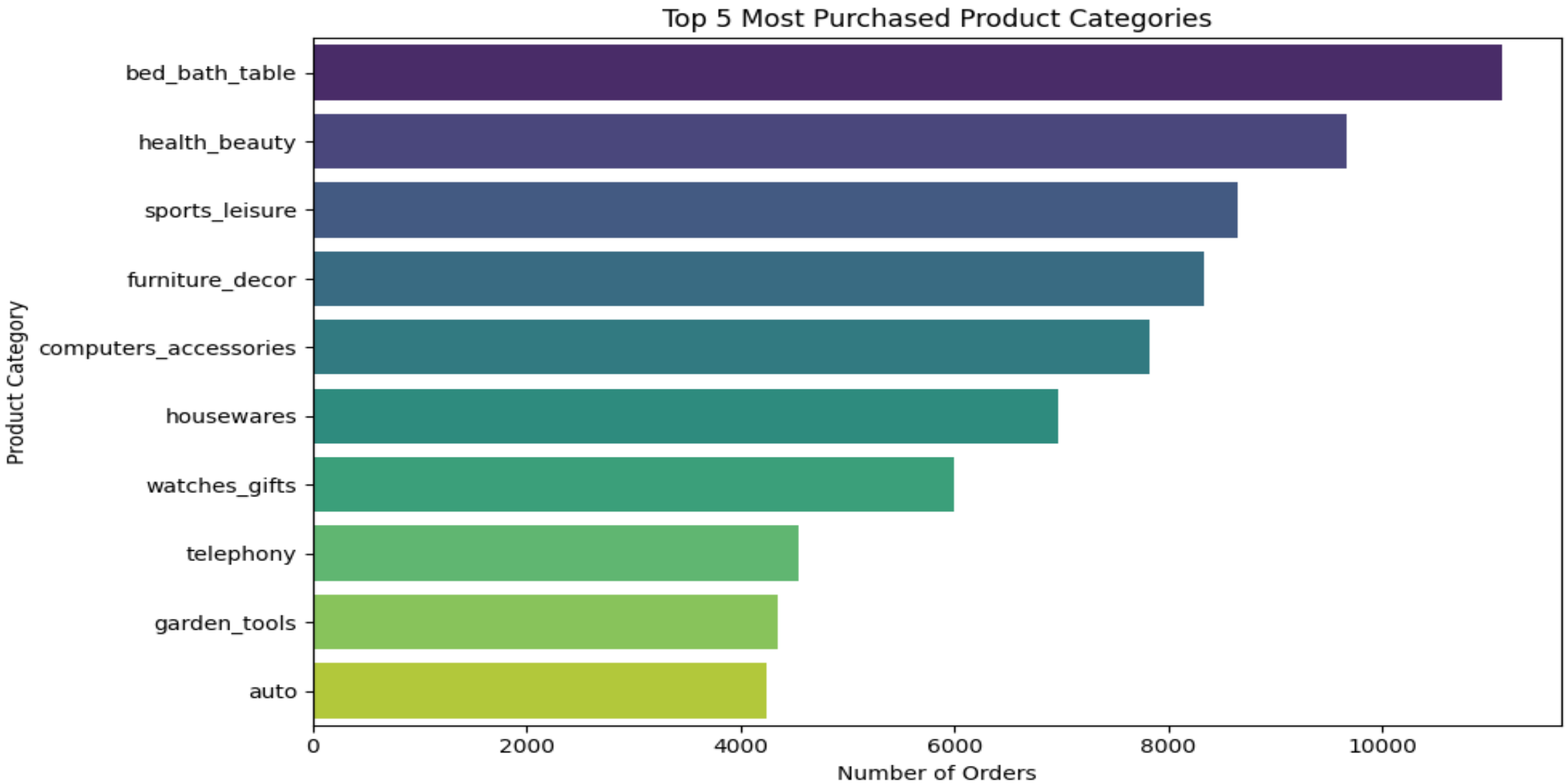


It is recommended that the company conducts a comprehensive review of its logistics and delivery operations in the states with significant delays. By identifying bottlenecks and inefficiencies, the company can optimize delivery times, ensuring timely fulfillment and enhancing operational efficiency in these regions.



The analysis has revealed a strong correlation between delayed orders and reduced customer satisfaction, as evidenced by lower satisfaction scores in customer reviews from these states. Addressing these delays is crucial to improving customer satisfaction and fostering positive customer relationships.

Product Analysis



We have created a utility(`get_top_products_by_state`) which enables the derivation of valuable insights into product categories by taking the state as an input parameter. This method effectively identifies the top five best-selling products and the top five products with the highest review scores within a given region.

Based on these insights, the company can provide targeted product suggestions to customers, highlighting products that are popular in their specific region. This approach enhances customer engagement by aligning recommendations with local buying trends.

Additionally, the analysis highlights products with lower average review scores. The company can focus on these products to understand the underlying causes of dissatisfaction. By investigating and addressing these issues, the company can enhance product quality and customer satisfaction, ultimately strengthening the brand's reputation.

Predicting Customer Satisfaction

Feature Engineering

- **Delivery Time Calculation:** Added a delivery_time column to compute the difference (in days) between order_delivered_customer_date and order_purchase_timestamp using datediff.
- **On-Time Delivery Indicator:** Created a binary feature delivered_on_time to indicate whether the delivery occurred on or before the order_estimated_delivery_date.
- **Shipped on Time:** Added a column shipped_on_time to check if the order_delivered_carrier_date was before or equal to the shipping_limit_date.
- **Joined DataFrames:** Combined order_reviews_df, orders_df, order_items_df, and order_payments_df using successive .join() operations on the order_id.

Feature Selection

Below are the features that are used for customer satisfaction

- shipped_on_time
- delivered_on_time
- delivery_time
- payment_value
- freight_value
- is_positive_review

Model building and Evaluation

This is a classification problem we used below models to predict customer satisfaction

- Logistic regression
- Decision tree classifier
- Random forest classifier
- Gradient boosted classifier

Out of this Random forest gave us better performance compared to other algorithms.

Metrics:

Accuracy: 0.67

Precision: 0.66

Recall: 0.67

F1 Score: 0.65

Customer Retention Strategy Based on Previous Analysis

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 base.

Scalability Considerations for Data Analytics and Modeling

- 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 can be 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.