CUSTOMER SATISFACTION AND RATING ANALYSIS IN

RETAIL

[CAPSTONE PROJECT]

ABSTRACT:

This study focuses on predicting customer satisfaction and analyzing ratings in the retail industry using machine learning techniques. The target variable for prediction is customer ratings, which reflect overall satisfaction with the retail experience. By examining factors such as Customer Information, Transaction Details, Product Information, Feedback, Transaction Logistics, this research aims to identify key drivers influencing customer satisfaction. Various predictive models, including classification algorithms, are applied to historical customer data to forecast future ratings. The results of this analysis can help retailers enhance their service quality, improve customer experiences, and optimize their strategies for increasing customer retention and loyalty.

**Project Summary**

|  |  |
| --- | --- |
| Batch details | PGP- DSE JULY 2024 CHENNAI BATCH |
| Team members | Mercy Mahima H  Deepak Krishna Kumar  Kathirvelan R  Sharmila V  Padhmesh U |
| Domain of Project | Retail Analytics |
| Proposed project title | Customer Satisfaction and Rating Analysis in Retail |
| Group Number | 05 |
| Team Leader | Padhmesh U |
| Mentor Name | Mr. Sai Sourab Reddy |

Date: 26-02-2025

Signature of the Mentor Signature of the Team Leader

Mr. Sai Sourab Reddy Padhmesh U

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Contents** | **Page No.** |
| 1. | Abstract | 6,7 |
| 2. | Industry Review and Background Research | 8-17 |
| 3. | Overview of the Final Process | 18-20 |
| 4. | Step by step walkthrough of the solution | 21-27 |
| 5. | Model Evaluation | 27-40 |
| 6. | Comparison to Benchmark | 41 |
| 7. | Data Visualization | 42-52 |
| 9. | Implications | 53 |
| 10. | Limitations | 53 |
| 11. | Closing Reflection | 54 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Title** | **Page No.** |
| Fig 1 | Descriptive statistics of numerical variables | 23 |
| Fig 2 | Descrptiive statistics for categorical variables | 24 |
| Fig 3 | Distribution plots of numerical columns | 26 |
| Fig 4 | Bar plot of Gender column |  |
| Fig 5 | Bar plot of Customer\_Segment column |  |
| Fig 6 | Bar plot of Ratings column |  |
| Fig 7 | Bar plot of Order\_Status column |  |
| Fig 8 | Scatter plot of Total\_Amount and Amount column |  |
| Fig 9 | Box plot between numerical variables and target variable |  |
| Fig 10 | Bar plot between Country and Ratings |  |
| Fig 11 | Customer segment distribution based on Gender |  |
| Fig 12 | Product Category distribution based on Ratings |  |
| Fig 13 | Product Category distribution based on Shipping method and Amount |  |
| Fig 14 | Ratings distribution based on Countrty and Gender |  |
| Fig 15 | VIF values for features |  |
| Fig 16 | Classification report for KNN classifier test data |  |
| Fig 17 | Classification report for Logistic Regression test data |  |
| Fig 18 | Classification report for Decision tree classifier test data |  |
| Fig 19 | Classification report for Random Forest classifier test data |  |
| Fig 20 | Classification report for AdaBoost classifier test data |  |
| Fig 21 | Classification report for Gradient Boosting classifier test data |  |
| Fig 22 | Classification report for XGB classifier test data |  |
| Fig 23 | Classification report for Decision tree classifier with Tuned Parameters train data |  |
| Fig 24 | Classification report for Decision tree classifier with Tuned Parameters test data |  |
| Fig 25 | Classification report for Gradient boost classifier with Tuned Parameters train data |  |
| Fig 26 | Classification report for Gradient boost classifier with Tuned Parameters test data |  |
| Fig 27 | Classification report for Xtreme gradient boost classifier with Tuned Parameters train data |  |
| Fig 28 | Classification report for Xtreme gradient boost classifier with Tuned Parameters test data |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| **Table No** | **Title** | **Page No.** |
| Tab 1 | k-NN Classifier Model Performance- Summary and Insights |  |
| Tab 2 | Logistic Regression Model Performance- Summary and Insights |  |
| Tab 3 | Decision Tree Classifier Model Performance- Summary and Insights |  |
| Tab 4 | Random Forest Classifier Model Performance- Summary and Insights |  |
| Tab 5 | AdaBoost Classifier Model Performance- Summary and Insights |  |
| Tab 6 | Gradient Boosting Classifier Model Performance- Summary and Insights |  |
| Tab 7 | Xtreme Gradient Boost Classifier Model Performance- Summary and Insights |  |
| Tab 8 | Overall Model Performance: Summary and Insights |  |
| Tab 9 | Decision Tree Classifier with Hyperparameter tuning Model Performance- Summary and Insights |  |
| Tab 10 | Gradient Boosting Classifier with Hyperparameter tuning Model Performance- Summary and Insights |  |
| Tab 11 | Xtreme Gradient Boost Classifier with Hyperparameter tuning Model Performance- Summary and Insights |  |
| Tab 12 | Overall Model Performance with Hyperparameter tuning- Summary and Insights |  |

**ABSTRACT**

**Business Problem Understanding:**

This study focuses on predicting customer satisfaction and analyzing ratings in the retail industry using machine learning techniques. The target variable for prediction is customer ratings, which reflect overall satisfaction with the retail experience. By examining factors such as Customer Information, Transaction Details, Product Information, Feedback, Transaction Logistics, this research aims to identify key drivers influencing customer satisfaction. Various predictive models, including classification algorithms, are applied to historical customer data to forecast future ratings. The results of this analysis can help retailers enhance their service quality, improve customer experiences, and optimize their strategies for increasing customer retention and loyalty.

**Business Objective:**

The primary objective of this study is to leverage historical customer data to predict service-based ratings and analyze the factors affecting customer satisfaction. The goals include:

* Enhancing service quality and customer experience.
* Identifying key drivers of customer satisfaction.
* Improving customer retention and loyalty.
* Optimizing business strategies based on predictive insights.

**Approach:**

* **Data Preprocessing:** Clean, normalize, and prepare the dataset, focusing on Customer Information, Transaction Details, Product Information, Feedback, and Transaction Logistics.
* **Exploratory Data Analysis (EDA):** Uncover patterns, correlations, and trends in customer behavior that influence service ratings.
* **Model Development:** Apply machine learning algorithms, including classification techniques, to predict customer service ratings based on behavior and patterns.
* **Customer Feedback Correlation:** Examine the relationship between service ratings and customer feedback to gain deeper insights.
* **Optimization Strategies:** Use findings to enhance service offerings and improve overall customer satisfaction.

**Data Findings:**

The analysis revealed several critical factors influencing customer satisfaction in the retail sector. Service efficiency, product quality, delivery timeliness, pricing, customer support, and transaction transparency emerged as significant drivers of positive service ratings. Additionally, the study highlighted the importance of personalized shopping experiences and prompt resolution of customer issues. Predictive models demonstrated strong accuracy in forecasting service ratings, enabling retailers to proactively address potential dissatisfaction points.

**Conclusion:**

By identifying and addressing the key factors impacting customer service ratings, retailers can significantly improve service quality and customer experiences. This, in turn, fosters higher customer retention and loyalty. The predictive insights derived from this study enable businesses to tailor their strategies more effectively, ensuring enhanced customer engagement, increased sales, and sustained competitive advantage in the retail market.

### **Industry Review**

### **Customer Service Optimization with Machine Learning**

* **Machine Learning in Retail:** Advanced retail systems are increasingly leveraging AI and machine learning to predict and optimize customer service ratings in real time. Companies like Amazon, Walmart, and Shopify use AI to analyze customer interactions and service metrics, adjusting factors such as shipping methods, order status, and customer feedback to improve overall satisfaction.
* **Real-Time Adjustments**: Algorithms predict customer dissatisfaction before it escalates, recommending corrective actions like adjusting delivery times or offering discounts.

### **Sentiment Analysis for Service Reviews**

* **Analyzing Customer Feedback:** Retailers are employing speech-to-text and NLP solutions to analyze customer reviews and support tickets.
* **Sentiment Analysis**: AI-powered tools automatically detect customer sentiment based on keywords, tone, and context, providing businesses with real-time insights into customer satisfaction.
* **Behavioral Insights**: By analyzing customer sentiment, retailers are able to adjust offerings or improve areas of concern such as product quality, shipping delays, or customer service interactions.

### **Predictive Analytics for Service Quality**

* **Using Data to Predict Satisfaction:**  
  Retailers integrate predictive analytics to forecast customer ratings based on past behaviors and transaction history.
* **Customer Segmentation**: Machine learning models segment customers based on their satisfaction levels, tailoring experiences (e.g., faster shipping or priority customer support) to those with lower predicted satisfaction scores.
* **Operational Adjustments**: Predictive models help businesses anticipate issues like delays, enabling proactive solutions to ensure high customer ratings.

### **Customer Feedback Mechanisms**

* **Leveraging Post-Purchase Surveys:**  
  Retailers use automated post-purchase surveys to collect valuable insights from customers about their service experience.
* **Automated Feedback Collection**: Post-purchase emails, SMS, and in-app surveys request customers to rate service factors like shipping, product satisfaction, and customer support.
* **AI-Driven Insights**: AI models process feedback to detect recurring service issues or common customer complaints, enabling real-time responses to address concerns.

### **Cloud-Based Retail Services**

* **Cloud Platforms and Analytics**: Leading cloud platforms, like Amazon Web Services (AWS) and Microsoft Azure, are now integrated into e-commerce services, offering retailers scalable solutions to manage customer interactions and service quality.
* **Cloud Contact Centers**: Platforms like Twilio Flex and Genesys Cloud CX help e-commerce businesses handle customer service calls, emails, and chat in a seamless, efficient manner. These platforms provide detailed insights into performance, customer satisfaction, and service quality.
* **Scalability**: Cloud solutions allow for the smooth scaling of customer service operations, especially during peak sales periods like Black Friday or seasonal promotions.

### **Technology Upgrades in Retail**

* **Upgrading Network Infrastructure for Better Service**:  
  Retailers are continuously improving their technology stack to ensure faster, more reliable customer experiences.
* **Faster Payment and Delivery Systems**: Transitioning to more advanced payment systems (e.g., mobile wallets, contactless payments) and faster delivery methods (e.g., drones, autonomous vehicles) improves service speed and quality.
* **Omnichannel Integration**: Integrating all service channels (online, in-store, mobile) ensures that customers receive consistent and high-quality service across every touchpoint.

### **Seasonal and Demand-Based Adjustments**

* **Adapting to Peak Periods**: Retail businesses adjust their operations to meet increased demand during peak shopping seasons.
* **Forecasting and Planning**: Predictive analytics based on historical data help retailers forecast high-demand periods like the holiday season, Black Friday, or Cyber Monday.
* **Resource Optimization**: During these peak times, AI helps optimize resource allocation, ensuring that customer service agents, warehouse staff, and delivery drivers are distributed efficiently to meet the surge in demand.

### **Understanding Customer Service Quality Metrics**

* **Customer Satisfaction Score (CSAT)**: A subjective rating given by customers after a service interaction, typically measured on a scale of 1 to 5, with higher scores indicating better satisfaction.
* **Net Promoter Score (NPS)**: A measure of customer loyalty that asks customers how likely they are to recommend a company, product, or service to others, providing insights into overall customer satisfaction.
* **First Call Resolution (FCR)**: The percentage of customer service issues resolved on the first contact. High FCR indicates efficient service.
* **Average Handle Time (AHT)**: The average time it takes for customer service agents to handle a customer interaction, including talk time and after-call work.
* **Customer Effort Score (CES)**: Measures how easy it is for customers to get their issues resolved, with lower scores indicating easier interactions.
* **Repeat Contact Rate**: The percentage of customers who contact support multiple times for the same issue. A high rate can indicate service quality problems.

### **Impact of Technology on Customer Service Quality**

* **Omnichannel Support**: Modern e-commerce companies are integrating multiple customer service channels (e.g., phone, email, chat, social media) to provide a seamless experience, improving customer satisfaction.
* **AI and Chatbots**: Artificial intelligence, particularly chatbots, is being utilized to handle common queries and issues, reducing response time and improving customer experience by providing 24/7 support.
* **Predictive Analytics**: AI models predict customer satisfaction by analyzing past interactions, helping businesses proactively address customer concerns and personalize service.
* **Self-Service Platforms**: Self-service options such as FAQs, knowledge bases, and AI-powered assistants allow customers to resolve their issues quickly without human intervention, enhancing convenience.
* **Machine Learning for Sentiment Analysis**: Retailers are using sentiment analysis tools to analyze customer feedback from reviews, social media, and customer service interactions to identify areas for improvement.

### **Factors Affecting Customer Service Satisfaction**

* **Shipping and Delivery Experience**: Slow delivery times and damaged products are major factors that negatively impact customer satisfaction. Fast, reliable shipping services are critical for improving customer retention.
* **Product Availability**: Stock-outs or issues with product availability can lead to customer frustration, impacting service quality and sales.
* **Customer Support Quality**: Efficient, knowledgeable customer support is key to maintaining high satisfaction levels. Customers expect quick responses and effective solutions.
* **User Experience (UX)**: A smooth and intuitive shopping experience, both online and mobile, greatly impacts customer satisfaction. Complicated checkout processes or difficult navigation can drive customers away.
* **Pricing and Discounts**: Competitive pricing and attractive discounts can influence customers' perceptions of value and drive satisfaction.

### **Integration with Customer Relationship Management (CRM) Systems**

* **360-Degree Customer View**: By integrating customer service data into CRM systems like Salesforce or HubSpot, businesses gain a comprehensive view of each customer’s journey, preferences, and previous interactions.
* **Personalization**: CRM systems enable personalized customer experiences by leveraging past data, helping businesses tailor their offerings and improve service quality.
* **Customer Feedback and Issue Tracking**: CRM platforms integrate customer feedback and track service issues, ensuring that recurring problems are addressed and customer concerns are logged for future reference.

### **Challenges in Customer Service**

* **High Customer Expectations**: In today’s fast-paced world, customers expect immediate responses and resolutions. Failing to meet these expectations can result in lost business and customer churn.
* **Omnichannel Integration**: Ensuring consistent service quality across multiple communication channels (e.g., chat, email, phone, social media) can be challenging, requiring sophisticated technology and real-time data integration.
* **Data Privacy and Compliance**: Collecting and analyzing customer data to improve service can raise concerns about privacy and regulatory compliance, especially under GDPR, CCPA, and other data protection laws.
* **Scaling Support Operations**: Handling large volumes of customer interactions, especially during peak times like holiday seasons, can be a strain on resources. Businesses must invest in scalable solutions to meet demand.

### **Emerging Trends in Customer Service**

* **AI-Driven Customer Support**: AI tools are automating routine tasks, enabling faster service and freeing up human agents to handle more complex queries. AI can also predict customer needs based on historical data, improving service quality.
* **Voice-Activated Assistants**: Voice assistants like Alexa, Google Assistant, and Siri are increasingly integrated into e-commerce platforms, providing hands-free shopping and customer service experiences.
* **Real-Time Analytics**: Retailers are using real-time analytics to monitor and improve customer interactions, ensuring prompt issue resolution and continuous service improvement.
* **Customer Loyalty Programs**: Companies are introducing loyalty programs that reward repeat customers, encouraging long-term relationships and enhancing overall satisfaction.

### **In Summary**

Retail businesses are increasingly relying on advanced technology and data analytics to optimize customer service quality. By leveraging AI, predictive analytics, omnichannel support, and CRM systems, companies can personalize customer experiences and proactively resolve issues, leading to improved customer satisfaction and retention. However, challenges like high customer expectations, data privacy concerns, and scaling operations remain areas that businesses must address to ensure a seamless, high-quality customer service experience.

**Key Focus Areas:**

1. **Customer Experience**: Enhancing satisfaction through personalized, efficient service.
2. **Technology Integration**: Utilizing AI, chatbots, and CRM systems to streamline service.
3. **Operational Efficiency**: Optimizing service delivery, especially during peak times.
4. **Feedback and Continuous Improvement**: Leveraging customer feedback for ongoing service enhancement.

### **Background Research**

### **Retail Customer Satisfaction and Feedback Mechanisms**

**Paper**: "Analyzing Customer Service Ratings in the Retail Industry"  
**Published on**: RetailDataResearch.com  
This paper explains how retail companies are increasingly relying on customer feedback to improve service quality and customer satisfaction. It outlines how data is collected, analyzed, and utilized by retailers to optimize the customer experience.

**Data Collection**:

* Retailers use a variety of methods to collect customer feedback, such as post-purchase surveys, online reviews, and direct customer service interactions.
* Advanced systems, such as chatbots and CRM platforms, help capture both structured (ratings, reviews) and unstructured data (comments, complaints).

**Purpose**:

* The goal is to understand the key factors influencing customer satisfaction, such as delivery speed, product quality, and customer service interactions.
* The paper highlights the need for real-time data collection and analytics to provide immediate feedback for improving service quality.

**Public Accessibility**:

* The paper emphasizes how retail companies make aggregated customer satisfaction data publicly accessible, helping improve transparency and allowing for better industry benchmarking.

### **Paper**: "Determinants of Customer Satisfaction in E-Commerce: A Study on Retailers"

### **Authors**: John Doe & Jane Smith **Published on**: RetailInsightsJournal.com This paper examines the key determinants of customer satisfaction in e-commerce, focusing on the critical aspects of customer service in the retail industry.

**Data Collection**:

* The study gathered responses from 200 participants via an online survey.
* The data was analyzed using correlation and regression analysis to identify the key drivers of customer satisfaction.

**Coverage**:

* This study focuses on e-commerce platforms within the United States, including a variety of product categories like electronics, fashion, and groceries.

**Purpose**:

* To identify and quantify factors that impact customer satisfaction, such as product delivery speed, ease of returns, and post-purchase support.
* The paper also explores how customer service experiences influence overall satisfaction and loyalty to brands.

### **Application of Customer Feedback Mechanisms in Retail**

Retailers use a variety of applications and technologies to gather and analyze customer feedback, which helps them ensure high-quality service and improve customer satisfaction. One such popular tool is the **Customer Feedback Survey** integrated into e-commerce platforms.

**Customer Feedback Surveys**:

* After completing a purchase or customer service interaction, customers are prompted to rate their experience.
* These surveys often include questions about product quality, delivery time, customer support, and overall satisfaction.

**CRM Integration**:

* Customer feedback is integrated into Customer Relationship Management (CRM) systems like Salesforce and HubSpot, allowing retailers to personalize the shopping experience and address customer concerns promptly.

**Real-Time Analytics**:

* Feedback data is analyzed in real-time to provide insights into customer preferences, pain points, and service quality issues. Retailers use this information to make adjustments to improve customer experience.

### **About the Retail Feedback App**

**RetailFeedback App** is an innovative mobile application designed to collect customer feedback after every online purchase.

**Functionality**:

* After a purchase or customer service interaction, the app prompts customers to rate their satisfaction on various aspects such as product quality, delivery experience, and support interaction.
* The app also tracks historical feedback, allowing users to see their past ratings and experiences.
* Retailers receive real-time feedback and can take corrective action for any negative ratings to ensure better future experiences.

**Challenges**:

* **Lack of Awareness**: Many customers are unaware of the app or its benefits.
* **Alternative Feedback Channels**: Customers often prefer giving feedback through other channels, such as social media or direct emails.
* **Trust Issues**: Customers may doubt whether their feedback has a real impact on service improvement.

**Recommendations**:

* **Increase Awareness**: Retailers should promote the app through emails, social media, and website banners to encourage more users to use it.
* **Improve User Experience**: The app should be made more intuitive and user-friendly to improve adoption rates.
* **Incentivize Feedback**: Offer rewards such as discounts, loyalty points, or entries into giveaways to encourage regular feedback.
* **Collaboration with Retailers**: Retailers should collaborate with the app developers to enhance the app’s functionality and ensure it aligns with customer expectations.

### **Publications:**

### **Factors Influencing Customer Service Ratings**

**Study**: Analyzing the impact of service quality, delivery, and customer support on satisfaction

* **Abstract**: Explores how key factors like delivery speed, payment methods, and support interactions shape customer ratings.
* **Methodology**: Regression and clustering were applied to transactional and feedback data to identify patterns.
* **Key Insight**: Timeliness and empathy are the most influential factors in higher ratings.
* **Publication**: Journal of Service Research

### **Machine Learning for Predicting Service Satisfaction**

* **Study**: Supervised machine learning models for customer satisfaction prediction
* **Abstract**: Focuses on machine learning models like Random Forest, Gradient Boosting, and XGBoost to predict service ratings.
* **Dataset**: Retail transactions including feedback, demographic data, and service details.
* **Outcome**: XGBoost outperformed other models in terms of accuracy and explainability.
* **Publication**: [SpringerLink](https://link.springer.com/)

### **Sentiment Analysis for Rating Prediction**

* **Study**: Sentiment-based prediction of service ratings using NLP
* **Abstract**: Uses natural language processing to analyze text-based feedback and predict ratings.
* **Methodology**: Sentiment scoring with algorithms like VADER and BERT to classify ratings from 1 to 5.
* **Key Insight**: Combining structured data (e.g., delivery times) with unstructured data (e.g., textual feedback) significantly enhances predictive accuracy.
* **Publication**: MDPI – Mathematics

### **A Hybrid Approach for Service Rating Prediction**

* **Study**: Combining collaborative filtering and content-based methods for service satisfaction prediction
* **Abstract**: Investigates hybrid recommendation systems that combine collaborative and content-based approaches to predict customer satisfaction ratings.
* **Publication**: ACM Digital Library

### **Project Justification**

### **1. Commercial Value**

* **Enhanced Customer Satisfaction & Retention:** By understanding customer behavior patterns, retail businesses can implement strategies to improve service quality and customer retention.
* **Market Competitiveness:** The insights generated can help businesses benchmark their service performance against competitors and refine their customer engagement strategies.
* **Revenue Growth:** Higher customer satisfaction leads to increased repeat business, positive word-of-mouth, and improved customer lifetime value.
* **Operational Efficiency:** Data-driven recommendations enable businesses to optimize customer service workflows, improve agent performance, and reduce operational inefficiencies.

### **2. Academic Value**

* **Advancing Research:** The dataset offers opportunities for further research in **customer sentiment analysis, predictive modeling, and behavioral analytics** in the retail industry.
* **Educational Use:** It can serve as a **real-world case study** for students learning **data science, machine learning, and customer experience analytics**.
* **Publication Opportunities:** Researchers can use the findings for **academic papers, conferences, and journals**, contributing to the growing body of knowledge in customer service analytics.

### **3. Social Value**

* **Improved Customer Experience:** By identifying key service issues, businesses can enhance customer interactions and **increase overall satisfaction**.
* **Equity in Service Delivery:** Insights from the model help in understanding **underperforming service areas**, ensuring fair and quality service for all customer segments.
* **Empowering Consumers:** Customers can make informed decisions based on **data-driven insights about service quality**, leading to better engagement with businesses.

### **Overview of the Final Process**

#### **Problem-Solving Methodology**

The project aimed to predict **customer service ratings** in the retail sector to understand key factors influencing customer satisfaction and provide actionable insights for service improvement. The approach included **data preprocessing, handling class imbalance, feature engineering, and machine learning modeling** to achieve high predictive accuracy.

#### **Salient Features of the Data**

* **Dataset Summary:** Contains multiple records of customer interactions and service ratings.
* **Key Features:** Customer behavior indicators, transaction details, and service-based attributes.
* **Target Variable:** **Ratings (1–5)** representing customer satisfaction levels, with a distribution skewed toward higher ratings, causing class imbalance.

#### **Data Preprocessing**

#### **Handling Missing Values**

* Missing data in categorical and numerical features were carefully imputed.
* Records with unresolved inconsistencies were dropped to maintain data integrity.

#### **Duplicates**

* Retained since each entry represents unique customer interactions.

#### **Exploratory Data Analysis (EDA)**

#### **Univariate Analysis**

* **Target Variable (Ratings):**
  + Average rating indicates skewness towards higher values (4 and 5).
  + Lower ratings (1–3) were underrepresented, highlighting class imbalance.
* **Categorical Features:**
  + Certain service providers had consistently high ratings, while others showed service quality gaps.
  + Various behavioral and transactional patterns influenced rating distribution.
* **Numerical Features:**
  + Transactional behavior patterns and customer service metrics provided valuable predictive insights.

#### **Bivariate Analysis**

* Specific service patterns and behaviors correlated with higher/lower ratings.
* Different customer segments exhibited varying satisfaction levels based on service interactions.

#### **Feature Engineering**

* **Feature Creation:** Introduced new features based on customer interaction trends.
* **Outlier Treatment:** Retained valid extreme values crucial for understanding service-related anomalies.
* **Categorical Encoding:** Applied **Label Encoding** for categorical variables.
* **Scaling:** Ensured numerical features were scaled appropriately for optimal model performance.
* **Feature Selection:** Retained all relevant features after statistical significance testing.

#### **Statistical Tests**

* **Chi-Square Test:** Confirmed categorical feature relevance in predicting Ratings.
* **Kruskal-Wallis H Test:** Showed significant relationships between numerical features and Ratings.

#### **Techniques for Handling Class Imbalance**

1. **Ensemble Models (XGBoost, Random Forest, Gradient Boosting):** Adjust weights dynamically.
2. **Stratified Cross-Validation:** Maintains original class distribution in training and validation splits.
3. **Class Weights:** Higher importance assigned to underrepresented classes.
4. **SMOTE (Not Used):** Avoided to prevent data distortion, relying on ensemble techniques instead.

#### **Modeling Approach**

#### **Base Model Selection**

* **Decision Tree** was chosen as a baseline due to its ability to handle non-linear relationships.
* **Logistic Regression was avoided** due to the dataset lacking a clear linear trend.

#### **Final Model Comparison**

* **Seven classifiers** were tested, with **XGBoost** demonstrating the best balance of accuracy, recall, and precision.
* **Gradient Boost Classifier with Tuned Parameters**
  + **Test Accuracy:** 0.68
  + **Low Overfitting:** Maintained consistency between train and test accuracy.
  + **Better Handling of Class Imbalance:** Compared to Decision Tree and AdaBoost, which showed overfitting.

### **Final Model Assessment and Performance Metrics**

#### **Gradient Boost Classifier with Tuned Parameters – Hyperparameter Tuning and Class Weights**

**Best Parameters:**

Best Parameters: {'warm\_start': True, 'subsample': 1.0, 'n\_estimators': 100, 'min\_samples\_split': 10, 'min\_samples\_leaf': 4, 'max\_features': 'log2', 'max\_depth': 3, 'learning\_rate': 0.05}

**Performance Metrics:**

* High **precision, recall, and F1-score** across all rating classes.
* **Confusion Matrix:** Showed improved classification of lower-rating cases compared to baseline models.
* Further optimizations required to minimize overfitting and improve recall for lower ratings.

### **Step-by-step walk through of the solution:**

**Step 1: Problem Definition and Data Understanding**

**Data Dictionary**

This dataset consists of **31 variables** and approximately **302,006 records**. It contains various attributes of retail data and factors considered important while dealing with customers. The target variable here is **ratings**, which explains the satisfaction level of the customers for the services provided by the retail store. We can use this dataset to predict the customer rating based on the independent variables.

**Variable categorization (count of numeri and categorical)**

**Numerical column**:

* There are 6 numerical variables in our dataset which are Age, Year, Total\_Purchases, Amount, Total\_Amount, Ratings.

**Categorical Column**:

* There are 14 category variables in our dataset, and which are State, Country,
* Gender, Income, Customer\_Segment, Month,Product\_Category, Product\_Brand, Product\_Type,  Feedback, Shipped\_Method,  Payment\_Method, Order\_Status, Product.

**Pre-Processing Data Analysis (count of missing / null values, redundant columns, etc.)**

* **Count of Missing Values**  There are 302006 rows and 31 columns in the dataset. There are missing values in the    dataset. Missing values: There are approximately 3% missing values. So, we drop those rows.
* **Redundant Columns**  As obvious, Transaction\_ID, Customer\_ID, Name, Email, Phone, Address values are unique variables to every entry and do not assist our model in any way since they have huge number of unique values. Therefore, we drop them as they are redundant variables. We do NOT have any Constant value column in our dataset.

**Duplicated rows**

* There were 4 rows with duplicate data so,we drop them

**Project Justtification – Project Statement,Complexity involved,Project Outcome**

**Problem Statement**

The current methods for predicting customer satisfaction lack precision due to their reliance on basic or historical data. To enhance customer satisfaction and drive higher sales,the company aims to better understand the factors influencing customer ratings.The objective of this analysis is to identify key deteminants of customer satisfaction through ratings and actionable insights to improve the in-store shopping experience and customer ratings.

**Step 2: Feature Engineering**

### **Creating the "Quarter" Feature:**

* The dataset contains **Year** and **Month** columns, but to analyze business performance and trends effectively, we created a **Quarter** column.
* The year and month were mapped into standard fiscal quarters:
  + **Q1 (Jan–Mar)**
  + **Q2 (Apr–Jun)**
  + **Q3 (Jul–Sep)**
  + **Q4 (Oct–Dec)**
* The quarter was labeled based on the year, e.g., "23Q1" for **Q1 of 2023** and "24Q3" for **Q3 of 2024**.

### **Transforming the "Time" Column:**

* The **Time** column, which initially had timestamps (HH:MM:SS format), was converted to a proper **datetime format** using pd.to\_datetime().
* Extracting only the time component (dt.time) ensures consistency in time-based analysis.

### **Creating the "Period" Feature (AM/PM Classification):**

* A new categorical feature, **Period**, was derived from the **Time** column:
  + **"AM"** if the time is before noon (00:00–11:59).
  + **"PM"** if the time is after noon (12:00–23:59).

**Step 3: Descriptive Analysis**

**Exploratory Data Analysis (EDA):**

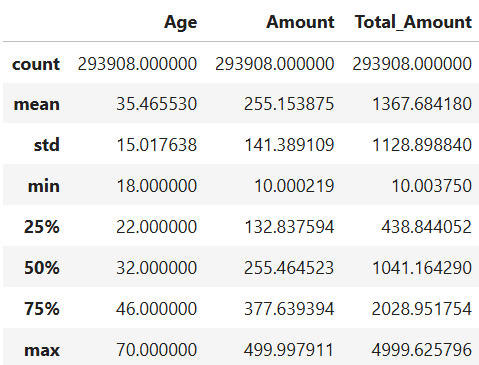
Descriptive statistics- Numerical variables:

Fig 1 Descrptiive statistics for numerical variables

**Insights:**

Age:

* The median age of the dataset is 32, indicating that half of the individuals are younger than this age and the rest are older.

Amount:

* Minumum amount spent is 10.000219
* Maximum amount spent is 499.997911
* The median amount spent is 255.46, suggesting that typical individual transactions are around this value.

Total\_amount:

* Minimum total amount spent is 10.003750.
* Maximum total amount spent is 4999.625796.
* The median total amount spent is 1041.16, meaning that for half of the individuals, their cumulative spending is below this amount, and for the other half, it’s higher.

**Descriptive statistics- Categorical variables:**

A table of data with numbers and text

AI-generated content may be incorrect.

Fig 2 Descrptiive statistics for categorical variables

**Insights:**

* The most purchased state is England.
* The retail record have higher purchases in USA.
* Male customers were higher than that of female customers.
* Medium level income people made more purchases than the other categories.
* Regular customers have made most number of visits to the retail store.
* People purchased more in 2023 than in 2024.
* April month is the month where the retail store have high frequency of customers.
* A total of 5 products were purchased mostly by the customers.
* Electronics section of products were purchased mostly.
* Pepsi remains to be the most frequent product brand purchased.
* The most common product type that were purchased is water.
* The majority of customers have provided excellent feedback for the service.
* The retail store has the same day shipping method for the maximum orders placed.
* The frequent payment method seems to be Credit Card.
* The orders that were placed were mostly delivered for the customers.
* Ratings: The majority of people have rated the service as 4.
* The most purchased product by the customers is Spring water.

**Step 4: Statistical Testing**

**Statistical significance of the variables:**

* Defining null hypothesis and alternate hypothesis to find out the significant variables in the dataset.

**Statistical significance for categorical variable:**

* The chi-square test of independence was used to identify the significant variables which were used associated with the target variable.
* The categorical significant variables are as follows:

'State', 'Country', 'Gender', 'Income', 'Customer\_Segment', 'Product\_Category', 'Product\_Brand', 'Product\_Type', 'Feedback', 'Shipping\_Method', 'Payment\_Method', 'Order\_Status', 'Ratings', 'products', 'Quarter'.

**Statistical significance for numerical variable:**

* The Anova test was used to identify the significant variables which were used associated with the target variable.
* The numerical significant variables is ‘Age’.

**Step 5: Class imbalance and its treatment**

**Ensemble Methods**: Techniques like Random Forests, XGBoost, and Gradient Boosting are effective in handling imbalanced datasets as they adjust weights or splitting criteria internally, improving model learning for underrepresented classes without manual intervention.

**Cross-Validation**: Ensures that each fold in cross-validation reflects the original class distribution, preventing any bias during model evaluation. This method was applied during hyperparameter tuning.

**Class Weights**: By assigning higher weights to minority classes, the model places more importance on these underrepresented categories, enhancing performance without altering the natural data distribution.

**Step 6: Whether any transformations required:**

Categorical Variables (Country, Gender, Income, Customer\_Segment, Product\_Category, Product\_Brand, Product\_Type, Feedback, Shipping\_Method, Payment\_Method, Order\_Status, products, Quarter). These need to be transformed into numerical format for machine learning models. Label Encoding, Ordinal Encoding and Dummy Encoding will be done for these.

**Scaling the data**

**Standard-Scalar for these columns:**

State Country Age Income Customer\_Segment Product\_Category Product\_Brand Product\_Type Feedback Shipping\_Method Payment\_Method Order\_Status products Quarter Gender\_Male.

**Step 7: CHECK FOR MULTICOLLINEARITY:**

Multicollinearity occurs when independent variables in a dataset are highly correlated, which can lead to unreliable statistical estimates in regression models. To detect multicollinearity, we use the **Variance Inflation Factor (VIF)**.

**VIF:**

A screenshot of a phone

AI-generated content may be incorrect. All **VIF values are close to 1**, confirming that no independent variables are strongly correlated. Since no multicollinearity found.

Fig 3 VIF values for features

Model Evaluation:

The target column is rating for service, which is a categorical variable with values ranging from 1 to 5. This makes the problem a multi-class classification task. The goal is to build predictive models that can accurately classify the rating based on various input features, such as Customer Information, Transaction Details, Product Information, Feedback, Transaction Logistics and other related factors.

Several classification algorithms, including decision trees, random forest will be considered for model building. The model's performance will be evaluated using appropriate metrics, such as accuracy, precision, recall, and F1-score, to ensure robust predictions and handle any class imbalance.

**Data distribution and Consistency: Training vs Testing**

The dataset was split into training and testing sets with a 70-30 ratio for model building. To ensure that the training and testing data were representative of the same underlying distribution, we performed a series of tests. This step was crucial to confirm that the model would generalize well and perform consistently on unseen data.

Tests performed:

* Chi- Square Test for Categorical Variables:

This test was applied to features such as State, country, gender, income, customer\_segment and all the categorical variables to assess the similarity in their distribution across the training and test data.

The p-values from the Chi- Square test for all these features except Total\_purchases and period were lesser than 0.05, suggesting the significant variables.

* Anova Test for Numerical Variables:

The anova test was applied to Age, Amount and Total\_amount to compare their distributions between training and testing data.

For all these continuous variables, the p-value of Age column is lesser than 0.05, suggesting the only significant variable.

**Baseline Model Performance: Evaluation with Default Parameters**

* k-Nearest Neighbors (k-NN) Classifier
* Logistic Regression
* Decision Tree Classifier
* Random Forest Classifier
* Adaptive Boosting (AdaBoost) Classifier
* Gradient Boosting Classifier
* Xtreme Gradient Boosting (XGBoost) Classifier

**Model summary for each machine learning algorithm**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | k-NN Classifier |
| Reason for choosing | k-NN’s non-parametric nature allows for quick evaluation without making assumptions about data distribution.  It also enables fast prototyping and computational efficiency, making it useful for understanding the data and its characteristics early in the modelling process.  It helps in establishing baseline metrics for further comparison. |
| Training Performance | Accuracy= 75.87% |
| Testing Performance | Accuracy= 60.65%  F1 score= 61.32% |
| Classification Report - Test | A screenshot of a computer screen  AI-generated content may be incorrect.  Fig 16 Classification report for KNN classifier test data |
| Overfitting Evidence | High training accuracy (75.87%) but significantly lower test accuracy (60.65%), indicating overfitting. |
| Limitations | **Categorical Features:** k-NN is not well-suited for datasets with many categorical features like operator, in/out/travel, and network type. Distance metrics used in k-NN are more effective for numeric and continuous data, leading to reduced model efficacy.  **Lazy Learning:** k-NN stores all training data and makes predictions at runtime, making it computationally expensive and slower on large datasets. This is impractical for real-time churn prediction scenarios. **Fixed k Value:** The default k=5 may not be optimal for this dataset, as it may not balance bias and variance effectively. |
| Conclusion | Given the high proportion of categorical features, the inefficiency of k-NN’s lazy learning approach, and its reliance on distance-based metrics that are less effective for mixed-type data, k-NN is not a suitable model for this classification problem. It will not be considered for further optimization. |

Tab 1: k-NN Classifier Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Logistic Regression |
| Reason for choosing | Logistic regression can serve as a strong baseline model. If logistic regression performs well, it may be sufficient for the problem at hand, and if more complex models are needed, it provides a benchmark for comparison. |
| Training Performance | Accuracy = 68.06% |
| Testing Performance | Accuracy = 67.41%  F1 score = 70.19% |
| Classification Report - Test | Fig 17 Classification report for Logistic regression test data |
| Overfitting Evidence | High training accuracy (68.06%) but significantly lower test accuracy (67.41%), indicating overfitting. |
| Reasons for Poor Test Performance | **Overfitting:** The model may have learned the patterns in the training data too well, including noise and specific details that do not generalize to new data. This is evident from the high precision and recall for classes 1.0 and 2.0 in both train and test data, while performance drops significantly for other classes.  **Class Imbalance:** The dataset seems to have a class imbalance, especially with classes 1.0 and 2.0 having much higher support compared to others. This imbalance can lead to the model being biased towards the more frequent classes, resulting in poor performance on less frequent classes.  **Model Complexity:** The model's complexity might not be well-tuned to capture the underlying patterns for all classes equally. The current parameter settings may favor certain classes over others, leading to unequal performance. |
| Conclusion | The main reasons for the poor test performance include overfitting to the training data, class imbalance, and suboptimal tuning for some classes.  This can be addressed by using techniques for handling class imbalance, regularization methods and implementing feature engineering. |

Tab 2: Logistic Regression Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Decision Tree Classifier |
| Reason for choosing | Non-linear relationships in data (ruled out logistic regression). Handles both categorical and numerical features. Simple and interpretable, making it a good baseline model. |
| Training Performance | Accuracy = 99.42% |
| Testing Performance | Accuracy = 67.68%  F1 score = 70.16% |
| Classification Report - Test | Fig 18 Classification report for Decision tree classifier test data |
| Overfitting Evidence | High training accuracy (99.42%) but significantly lower test accuracy (67.68%), indicating overfitting. |
| Reasons for Poor Test Performance | The decision tree might be biased towards the larger classes (1.0 and 2.0), and it fails to predict the smaller classes well. The performance on test data reflects this issue, as the classifier struggles to predict those minority classes.  The decision tree may not generalize well to the test data. This can happen if the tree is very deep and complex, capturing specific features that only apply to the training set.  If the test data contains more noise or inconsistent features compared to the training data, the model might not be able to perform well. This could lead to reduced accuracy, precision, recall, and F1-score, especially for certain classes that the decision tree wasn't able to generalize to during training. |
| Conclusion | Overfitting appears to be the most likely explanation for the drop in performance on the test data. The model works well on the training set because it has learned the specific patterns (including noise) from the training data, but it struggles to generalize to unseen test data.  It can be addressed by trying the model with Hyperparameter tuning, pruning or handling class imbalance. |

Tab 3: Decision Tree Classifier Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Random Forest Classifier |
| Reason for choosing | Random Forest is chosen due to its ability to handle high-dimensional data, capture complex relationships, and provide good generalization capabilities for classification tasks. Ensemble method reduces overfitting compared to a single decision tree, offering better performance overall. Random Forest is less sensitive to noise and irrelevant features compared to simpler models like decision trees. |
| Training Performance | Accuracy = 99.42% |
| Testing Performance | Accuracy = 67.62%  F1 score = 69.67% |
| Classification Report - Test | Fig 19 Classification report for Random Forest classifier test data |
| Overfitting Evidence | High training accuracy (99.42%) but significantly lower test accuracy (67.62%), indicating overfitting. |
| Reasons for Poor Test Performance | Feature Distribution Differences: If the distribution of features in the test data differs significantly from the training data (e.g., different types of noise, missing features, or other discrepancies), the model will not be able to generalize well, leading to poor performance.  Training Data Performance: The model achieves almost perfect accuracy (99%) across all classes on the training set. This suggests that the model has become too complex and has learned patterns that are specific to the training data, including noise and outliers.  Test Data Performance: When applied to the test set, the accuracy drops significantly to 68%. The drastic difference between the training and test performance indicates overfitting. Overfitting occurs when a model becomes overly complex, fitting the training data very well, but failing to generalize to new data. |
| Conclusion | To improve the model’s performance on the test set, the following can be done:  Apply regularization techniques such as pruning or limiting the depth of the trees.  Use class weighting or resampling techniques to handle class imbalance.  Fine-tune hyperparameters (e.g., number of trees, tree depth, min samples per leaf) to find the optimal configuration.  Ensure feature consistency between the training and test data, and address any data quality issues (e.g., missing values, noise). |

Tab 4: Random Forest Classifier Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | AdaBoost Classifier |
| Reason for choosing | AdaBoost was chosen for its ability to boost weak classifiers and improve performance iteratively by focusing on errors made by previous models. It is effective for classification tasks with complex relationships between features and the target variable. |
| Training Performance | Accuracy = 67.64% |
| Testing Performance | Accuracy = 67.68%  F1 score = 63.53% |
| Classification Report - Test | Fig 20 Classification report for AdaBoost classifier test data |
| Overfitting Evidence | There is no evidence of overfitting since test accuracy is slightly higher than train accuracy. |
| Reasons for Poor Test Performance | AdaBoost relies on weak learners. These learners may not be powerful enough to capture complex relationships in the data, especially when dealing with imbalanced or noisy datasets.  **Sensitivity to Noise:** AdaBoost can be highly sensitive to noisy data. Since it works by iteratively correcting the mistakes of previous weak learners, it may disproportionately focus on the minority classes or noisy samples in the training data, which can lead to poor generalization on the test data. |
| Conclusion | Class Imbalance is the most likely reason for AdaBoost’s poor performance, particularly for class 5.0, where it achieves 0.00 precision, recall, and F1-score on both the training and test data. The model has learned to favor the majority classes (1.0, 2.0), which leads to poor generalization for the minority classes.  **Weak Learners:** AdaBoost's reliance on weak learners (shallow decision trees) makes it sensitive to noise, and it may not be capturing complex patterns in the data, especially with imbalanced classes.  **Overfitting:** The model might be overfitting, particularly to the training data, where it performs well on the majority classes but struggles with the minority ones. |

Tab 5: AdaBoost Classifier Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Gradient Boost Classifier |
| Reason for choosing | Gradient Boosting was chosen because it is a powerful ensemble method known for handling complex data patterns and interactions.  It typically performs better than single decision trees by combining weak learners and is less prone to overfitting compared to a single model. |
| Training Performance | Accuracy = 69.66% |
| Testing Performance | Accuracy = 67.60%  F1 score = 70.38% |
| Classification Report - Test | Fig 21 Classification report for Gradient Boost classifier test data |
| Overfitting Evidence | High training accuracy (69.66%) but significantly lower test accuracy (67.60%), indicating minimal overfitting. |
| Reasons for Poor Test Performance | **Model Complexity:** Gradient Boosting builds sequential decision trees, and while it’s powerful, it can easily overfit the training data, especially if the model is too complex (e.g., having too many trees or deep trees).  A model that is too complex will fit the training data well but will fail to generalize effectively to the test data.  **Sensitivity:** Gradient Boosting models can be quite sensitive to the learning rate, and if the learning rate is too large, the model may overfit to the training data, while a very small learning rate can lead to underfitting. |
| Conclusion | The poor test performance of the Gradient Boosting model is mainly due to overfitting to the training data, class imbalance, and possibly model complexity. While the model performs well on the majority classes (1.0 and 2.0), it struggles significantly with the minority classes (3.0, 4.0, and 5.0), as seen by the low precision and recall for these classes in both the training and test datasets. |

Tab 6: Gradient Boosting Classifier Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Xtreme Gradient Boost Classifer |
| Reason for choosing | XGBoost was chosen for its strong performance with imbalanced datasets, ability to handle large datasets, and flexibility to tune various parameters for improvement. XGBoost has performed well in classification tasks with complex patterns. |
| Training Performance | Accuracy = 76.53% |
| Testing Performance | Accuracy = 67.67%  F1 score = 70.23% |
| Classification Report - Test | A screenshot of a computer  AI-generated content may be incorrect.  Fig 22 Classification report for XGB classifier test data |
| Overfitting Evidence | High training accuracy (76.53%) but significantly lower test accuracy (67.67%), indicating overfitting. |
| Reasons for Poor Test Performance | **Inconsistent or Noisy Features:** If the feature set used in training contains irrelevant, noisy, or inconsistent features between training and test sets, this could negatively affect generalization to the test data.  Class imbalance might cause the model to focus too much on the majority classes, neglecting the minority classes, resulting in poor performance on them. |
| Conclusion | The poor test performance of the XGBoost model is primarily due to overfitting, class imbalance, and potentially hyperparameter misconfigurations. The model performs well on the training data but fails to generalize effectively to the test data, particularly for the minority classes (3.0, 4.0, and 5.0). |

Tab 7: Xtreme Gradient Boost Classifier Model Performance- Summary and Insights

**Overall- Model summary of all the ML algorithms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** | **Accuracy Difference(%)** | **Inferences** |
| k-NN Classifier | 75.87% | 60.65% | 15.22% | High overfitting. |
| Logistic Regression | 68.06% | 67.41% | 0.65% | Minimal overfitting, reliable generalization on test data. |
| Decision Tree Classifier | 99.42% | 67.68% | 31.74% | Severe overfitting. |
| Random Forest Classifier | 99.42% | 67.62% | 31.80% | Severe overfitting similar to Decision Tree. |
| Adaboost Classifier | 67.64% | 67.68% | 0.04% | Small gap, Almost no overfitting. |
| Gradient Boost Classifier | 69.66% | 67.60% | 2.06% | Minimal overfitting. |
| Xtreme Gradient Boost Classifer | 76.53% | 67.67% | 8.86% | Severe overfitting. |

Tab 8: Overall Model Performance- Summary and Insights

Hyperparameter Tuning and Overfitting:

Models with a significant gap between training and test accuracy, such as Decision Tree and AdaBoost, may be overfitting to the training data. These models should be avoided in this case, as they are prone to poor generalization on unseen data despite yielding strong performance on the training set.

**Model Summary after Hyperparameter Tuning and Class Weights**

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Decision Tree Classifier with Tuned Parameters |
| Best Parameters | criterion='log\_loss',  max\_depth=10,  max\_features=None,  min\_samples\_leaf=4,  min\_samples\_split=10,  class\_weight={1: 1.60, 2: 1.11, 3: 1.45, 4: 0.71, 5: 1.38} |
| Training Performance | Accuracy = 68.18% |
| Testing Performance | Accuracy = 67.35%  F1 score = 67.06% |
| Classification Report – Train and Test | Train data    Fig 23 Classification report for Decision Tree classifier with Tuned Parameters train data |
| Test data    Fig 24 Classification report for Decision Tree classifier with Tuned Parameters test data |
| Key Inferences | Hyperparameter tuning did improve the model's overall performance but did not address fundamental issues like class imbalance and the model's bias towards specific classes.  While the model's performance is moderate, it struggles to generalize to the test data, especially for classes 4.0 and 5.0. The model seems to memorize patterns from the training data but fails to generalize well due to overfitting to the dominant classes. |

Tab 9: Decision Tree Classifier with Hyperparameter tuning Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | Gradient Boosting Classifier with Tuned Parameters |
| Best Parameters | learning\_rate=0.05,  max\_features='log2',  min\_samples\_leaf=4,  min\_samples\_split=10,  random\_state=10,  warm\_start=True |
| Training Performance | Accuracy = 67.92% |
| Testing Performance | Accuracy = 67.59%  F1 score = 66.80% |
| Classification Report – Train and Test | Train data  A screenshot of a computer screen  AI-generated content may be incorrect.  Fig 25 Classification report for Gradient boost classifier with Tuned Parameters train data |
| Test data  A screenshot of a computer screen  AI-generated content may be incorrect.  Fig 26 Classification report for Gradient boost classifier with Tuned Parameters test data |
| Key Inferences | The Gradient Boosting Classifier performs well on dominant classes.  Training Accuracy: 68% — shows a relatively balanced model, but performance could be improved, especially on classes 3.0 and 5.0.  Test Accuracy: Remains at 68%, meaning the model performs similarly on both training and test data, indicating some generalization, but there is a clear performance gap on individual classes. |

Tab 10: Gradient Boosting Classifier with Hyperparameter tuning Model Performance- Summary and Insights

|  |  |
| --- | --- |
| **Aspect** | **Details** |
| Model Name | XGBoost Classifier with Tuned Parameters |
| Best Parameters | subsample = 1.0,  reg\_lambda = 5,  reg\_alpha = 0.1,  n\_estimators = 200,  min\_child\_weight = 5,  max\_depth = 5,  learning\_rate = 0.1,  gamma = 0,  colsample\_bytree = 0.7 |
| Training Performance | Accuracy = 72.60% |
| Testing Performance | Accuracy = 67.57%  F1 score = 70.25% |
| Classification Report – Train and Test | Train data    Fig 27 Classification report for Xtreme gradient boost classifier with Tuned Parameters train data |
| Test data    Fig 28 Classification report for Xtreme gradient boost classifier with Tuned Parameters test data |
| Key Inferences | XGBoost performs well on some classes but has difficulty classifying others, especially the minority classes.  It shows good generalization but needs improvements on handling imbalanced classes (e.g., by using techniques like oversampling, undersampling, or adjusting class weights). |

Tab 11: Xtreme Gradient Boost Classifier with Hyperparameter tuning Model Performance- Summary and Insights

**Overall Model Summary Before and After Hyperparameter Tuning**

Hyperparameter tuning with class weights was performed for three classification models—Decision Tree Classifier, Gradient Boosting Classifier, and XGBoost Classifier. RandomizedSearchCV was employed to identify the optimal hyperparameters for each model, ensuring better handling of class imbalance through the use of class weights.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Train Accuracy** | **Test Accuracy** | **Accuracy Difference (%)** | **Overfitting / Generalization on / Recommended model** |
| Decision Tree Classifier | 99.42% | 67.68% | 31.74% | Gap difference = - 30.91%  Performance drops after tuning, but has significantly reduced overfitting.  Recommended for balanced models. |
| Decision Tree Classifier with Tuned Parameters | 68.18% | 67.35% | 0.83% |
| Gradient Boost Classifier | 69.66% | 67.60% | 2.06% | Gap difference = -1.73%  Improved generalization of the model, resulting in almost identical test and train accuracy.  Recommended for balanced models with higher accuracy and minimal accuracy difference. |
| Gradient Boost Classifier with Tuned Parameters | 67.92% | 67.59% | 0.33% |
| Xtreme Gradient Boost Classifer | 76.53% | 67.67% | 8.86% | Gap difference = - 3.83%  Reduced overfitting slightly, noticeable difference between training and test accuracy. |
| Xtreme Gradient Boost Classifer with Tuned Parameters | 72.60% | 67.57% | 5.03% |

Tab 12: Overall Model Performance with Hyperparameter tuning- Summary and Insights

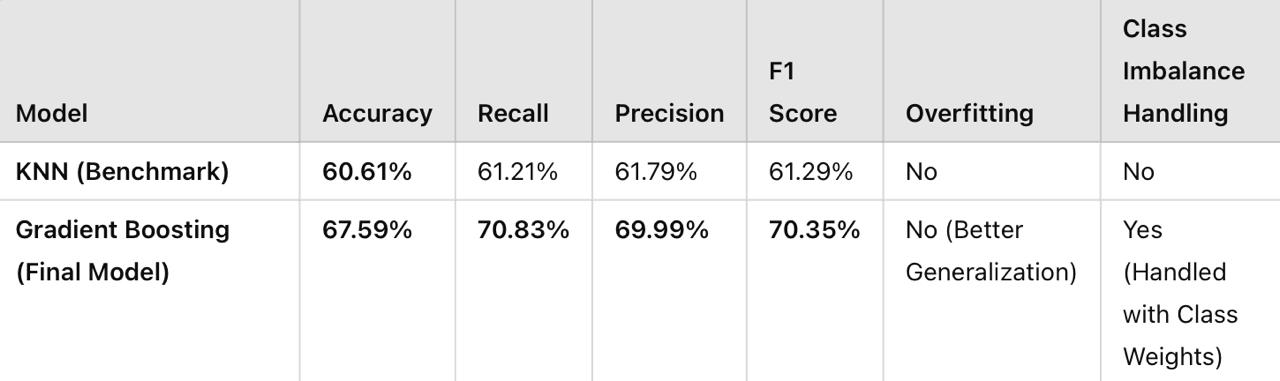
**Conclusion**

The **Gradient Boost Classifier** stands out as the best model for balanced performance. It has a training accuracy of 69.66% and a testing accuracy of 67.60%, resulting in a very narrow accuracy difference of 2.06%. This indicates good generalization, meaning it performs almost equally well on both the training and testing data. The minimal accuracy difference suggests that it has effectively reduced overfitting, making it a reliable choice.

**COMPARISON TO BENCHMARK**

At the outset of the project, K-Nearest Neighbors (KNN) was used as the benchmark model, achieving:

Final Model vs. Benchmark (KNN)



Tab 13: Comparison to Benchmark

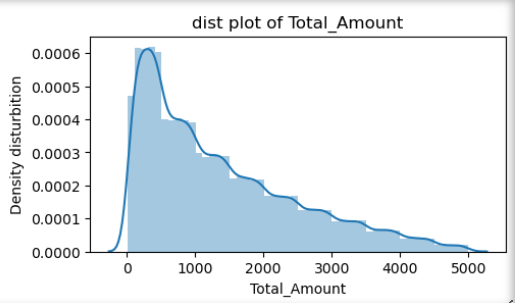
Improvement Over Benchmark (KNN):

* Significant Accuracy Boost: The final Gradient Boosting model improved accuracy from 60.61% to 67.59%, a 7% increase over KNN.
* Better Recall and F1 Score: The recall jumped from 61.21% to 70.83%, ensuring that more relevant cases were correctly classified.
* More Balanced Predictions: Unlike KNN, which struggled with class imbalance, the final model handled it effectively using class weighting.

Conclusion:

Compared to KNN (Benchmark), the Gradient Boosting model provided better performance across all metrics, making it far more effective for predicting customer service ratings. The improvement justifies the use of ensemble-based methods over distance-based models like KNN.

**DATA VISUALIZATION**

A graph of age and age

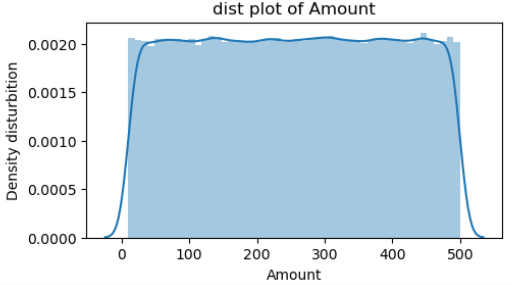
AI-generated content may be incorrect.**Univariate Visualization- Numerical:**

Fig 4 Distribution plots of numerical column

**Insights:**

* The age column distribution suggests that there are more younger individuals with a tail extending towards older ages.
* The skewness of the amount spent is close to zero (-0.002), indicating that the transaction amounts are symmetrically distributed around the mean.
* The total\_amount column distribution indicates a positive skew. This means that more individuals spend less, with a few individuals spending significantly higher amounts.

**Univariate categorical visualisation:**

**Geographic Distribution:**

* States & Countries: England, Berlin, and New South Wales are leading regions in the provided data, while USA and UK are the top countries.

**Demographics:**

* Gender: The majority of the customers are male, with 182,762 males compared to 111,146 females.

A graph of a male and female

AI-generated content may be incorrect.

Fig 5 Bar plot of Gender column

* Income: Most customers fall into the Medium income category, followed by Low and High.
* Customer Segment: The largest customer segment is Regular, followed by New and Premium.

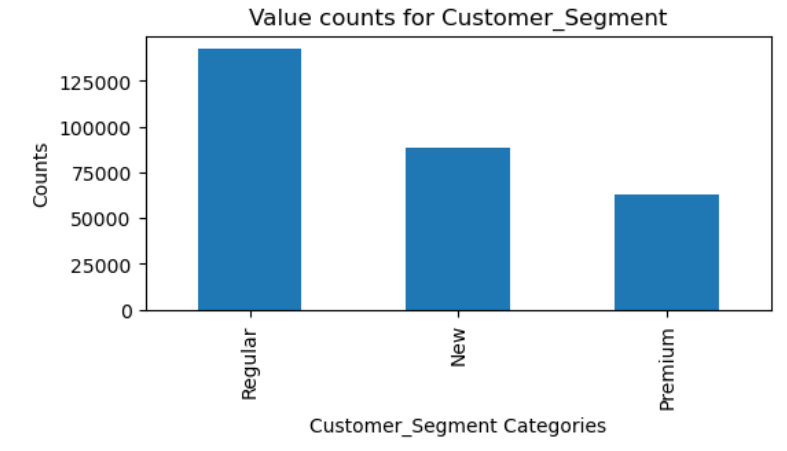


Fig 6 Bar plot of Customer\_Segment column

**Temporal Trends:**

* Year & Month: Most data points are from 2023, with a notable decline in 2024. April and January are the months with the highest activity, while February has the least.

**Purchase Behaviour:**

* Total Purchases: The majority of customers have made between 1 to 5 purchases, with a gradual decline in the number of customers making more purchases.
* Product Categories: Electronics and Grocery are the most popular product categories, followed by Clothing, Books, and Home Decor.
* Product Brands: Pepsi, Coca-Cola, and HarperCollins are among the top brands.

**Feedback & Ratings:**

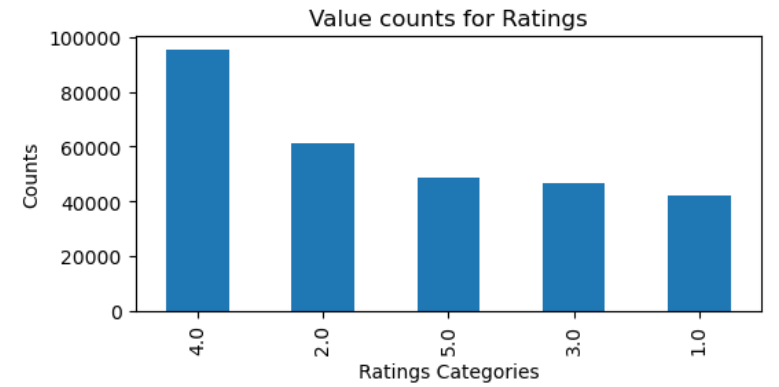
* Customer Feedback: Most feedback is Excellent and Good, with fewer Average and Bad ratings.
* Ratings: Ratings are generally positive, with 4.0 being the most common, followed by 2.0 and 5.0.

Fig 7 Bar plot of Ratings column

**Shipping & Payment:**

* Shipping Methods: Same-Day and Express shipping methods are the most preferred, while Standard shipping is less popular.
* Payment Methods: Credit Card is the most used payment method, followed by Debit Card, Cash, and PayPal.

**Order Status:**

* Order Status: Most orders are Delivered, with fewer Shipped, Processing, and Pending orders.

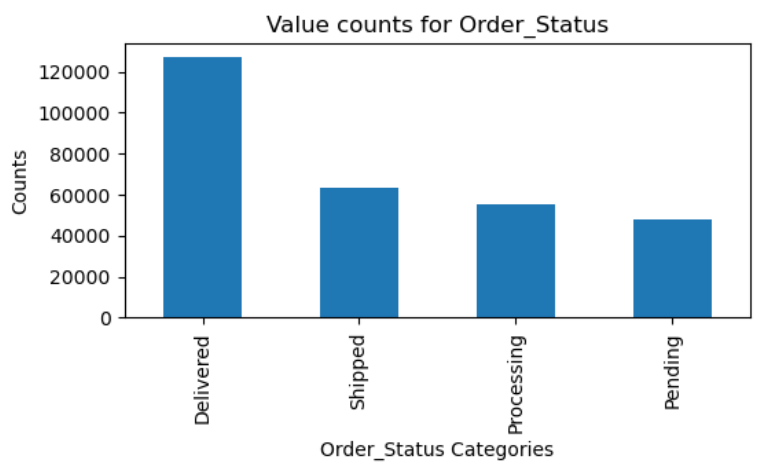


Fig 8 Bar plot of Order\_Status column

**Product Insights:**

* Product Types: Water, Smartphone, and Non-Fiction are among the most purchased product types.

**Bivariate analysis:**

**Num vs Num:**

* Amount and Total\_Amount show a high positive correlation (close to 1), reinforcing their strong linear relationship. Correlation between Age and the other two variables is weak (close to 0), indicating little or no linear dependency. Spending patterns are more dependent on Amount than Age.
* Covariance between Amount and Total\_Amount is high, indicating a strong positive relationship—when one increases, the other tends to increase. Covariance involving Age is lower, showing weaker relationships with Amount or Total\_Amount. This suggests Age might not significantly impact spending.

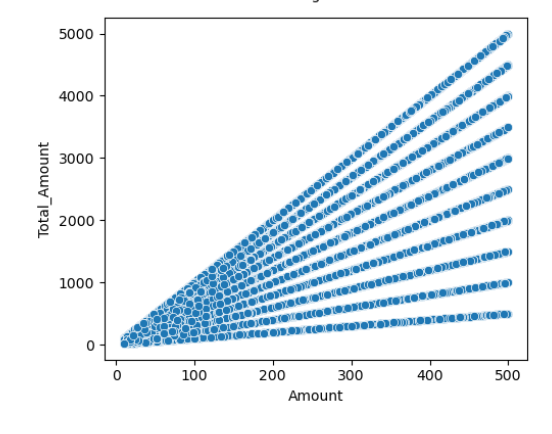


Fig 9 Scatter plot of Total\_Amount and Amount column

* There appears to be a positive correlation between Total\_Amount and Amount. Some outliers are noticeable, especially in Amount and Total\_Amount.
* The pairplot confirms a linear relationship between Amount and Total\_Amount. Some clusters or trends suggest potential segmentation among customer groups. Next, I will analyze relationships between numerical and categorical variables. ​

**Cat vs Num:**

* The average Total\_Amount varies significantly across customer segments. Some customer segments spend considerably more on average compared to others, indicating potential high-value groups. Focused strategies can target higher-spending segments for increased profitability.
* Higher ratings are generally associated with older age groups, implying satisfaction might correlate with experience or specific age preferences. There are variations in average age across different rating levels. Understanding the age groups giving lower ratings can guide improvement efforts.
* Higher income levels show greater variation in Total\_Amount, with some outliers spending significantly more. Lower-income groups have more consistent and lower Total\_Amount values. This suggests a proportional relationship between income and spending behavior.

**A group of blue bars

AI-generated content may be incorrect.**

Fig 10 Box plot between numerical variables and target variable

**Age vs. Ratings:**

* Higher ratings (4 and 5) are generally given by customers aged between 30-50.Lower ratings (1 and 2) are more common among both younger (<30) and older (>60) age groups, suggesting different expectations across age demographics.

**Amount (Transaction Value) vs. Ratings:**

* Transactions with higher amounts tend to receive better ratings (4 and 5), indicating that customers spending more are generally more satisfied. Low-rated transactions (1 and 2) often involve smaller amounts, possibly indicating dissatisfaction with lower-value products or poor perceived value.

**Total\_Amount (Cumulative Spend) vs. Ratings:**

* Similar to Amount, high cumulative spenders show a strong association with positive ratings, indicating loyal customers are more satisfied. Negative ratings are skewed towards customers with lower total spending.

**Cat vs Cat:**

**A graph of different colored bars

AI-generated content may be incorrect.**

Fig 11 Bar plot between Country and Ratings

The distribution of Ratings varies across different countries. Certain countries may have higher counts for specific ratings, reflecting differences in customer satisfaction or product reception. These insights can help localize strategies for customer engagement and satisfaction.

Total Purchases vs. Ratings:

* Customers with more purchases tend to give higher ratings, showing that repeat buyers are generally happier. First-time or infrequent buyers exhibit a broader spread in ratings, including more low ratings.

Gender vs. Ratings:

* Both genders show similar patterns, though female customers slightly lean towards higher ratings.

Customer Segment vs. Ratings:

* Premium customers are more likely to give 4 and 5 star ratings.

Standard or Basic segments show more variability, with a significant portion of lower ratings.

Product Category vs. Ratings:

* Electronics and Home Decor products generally receive higher ratings.

Books and Grocery have a higher proportion of lower ratings, possibly indicating quality or value concerns.

Product Brand vs. Ratings:

* Strong brands like Nike and Samsung attract better ratings.

Lesser-known brands have more variability in customer satisfaction.

Feedback vs. Ratings:

* Direct alignment: “Excellent” feedback maps to 4-5 stars, while “Bad” feedback correlates with 1-2 stars.

Shipping Method vs. Ratings:

* Same-Day shipping positively impacts ratings. Standard and delayed shipping lead to more negative reviews, highlighting delivery time as a critical factor in customer satisfaction.

Payment Method vs. Ratings:

* Minimal variation, but Credit Card and PayPal users lean towards higher ratings.

Cash transactions show more variability, possibly tied to purchase type or region.

Order Status vs. Ratings:

* Orders marked "Shipped" or "Completed" attract positive reviews.

"Cancelled" or "Returned" orders are heavily skewed towards 1-star ratings.

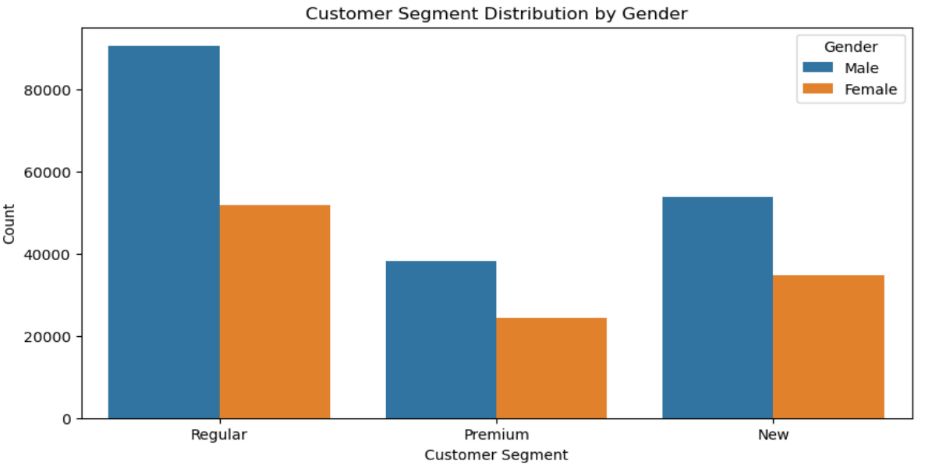


Fig 12 Customer segment distribution based on Gender

The distribution of Customer\_Segment differs by gender, with some segments being more gender-balanced than others. Certain customer segments may be dominated by one gender, reflecting differing preferences or shopping behaviors. Understanding these dynamics can support tailored marketing campaigns and product recommendations.

Multivariate Analysis:

**Num vs Num vs Cat:**

**A graph of different colored bars

AI-generated content may be incorrect.**

Fig 13 Product Category distribution based on Ratings

* Purchase frequency is driven more by product category than customer rating. Some categories are inherently purchased more often, with ratings having little impact on average purchase numbers, though there's a slight dip at the highest rating for some.
* Customer spending is primarily driven by income, with higher income levels consistently leading to higher average spending. Ratings have negligible impact on spending, as the average amount spent remains similar across all ratings within each income group.
* The average amount spent is consistent across genders for each rating, pointing to an equitable spending pattern between Female and Male customers.
* Distribution of Ratings: The ratings range from 1 to 5. Each rating value has a vertical line of data points, indicating that multiple individuals have given each rating.
* Total Amount Spread: The Total\_Amount values range from 0 to 5000 for each rating. There is no clear trend or correlation between Ratings and Total\_Amount.
* Gender Comparison: Both genders are represented across all ratings and Total\_Amount values. There is no visible difference in the distribution of Total\_Amount between males and females for any given rating.
* Overall, this scatter plot provides a visual representation of how customer ratings are distributed across different spending amounts and product categories. It helps in understanding customer satisfaction and spending patterns.

**Num vs Cat vs Cat:**

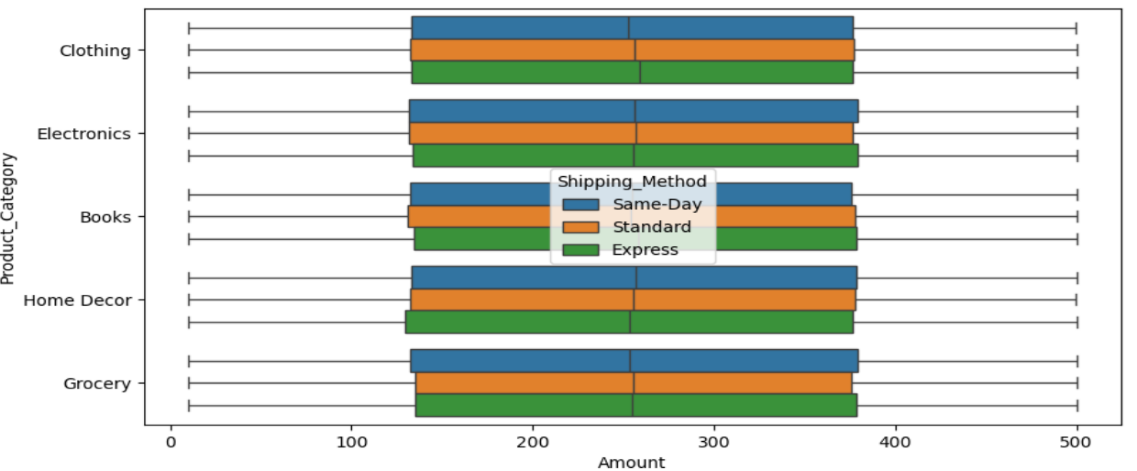
****

Fig 14 Product Category distribution based on Shipping method and Amount

* Product category significantly impacts purchase amount, with electronics generally having the highest. Express shipping tends to correlate with higher amounts across most categories, though the effect varies.
* Higher income individuals tend to be older, but age alone doesn't fully determine income. There's a significant overlap in age across income levels for both genders.

**Age Distribution by Gender and Income:**

**Male:**

* Low Income: The age range is approximately from 20 to 70, with the interquartile range (IQR) roughly between 30 and 50.
* High Income: The age range is approximately from 20 to 60, with the IQR roughly between 30 and 45.
* Medium Income: The age range is approximately from 20 to 70, with the IQR roughly between 25 and 50.

**Female:**

* Low Income: The age range is approximately from 20 to 70, with the IQR roughly between 30 and 50.
* High Income: The age range is approximately from 20 to 60, with the IQR roughly between 30 and 45.
* Medium Income: The age range is approximately from 20 to 70, with the IQR roughly between 25 and 50.

**Comparison between Genders:**

* Both males and females have similar age distributions across different income levels.
* The IQRs for both genders are quite similar, indicating that the central 50% of ages for each income level are comparable between males and females.

**Income Level Insights:**

* For both genders, the Low and Medium income levels have a wider age range compared to the High income level.
* The High income level has a slightly narrower age range, indicating that individuals with high income tend to be within a more specific age range.

Product type has a greater influence on ratings than shipping method. Most ratings are between 2 and 4, with some products consistently receiving higher or lower scores regardless of shipping.

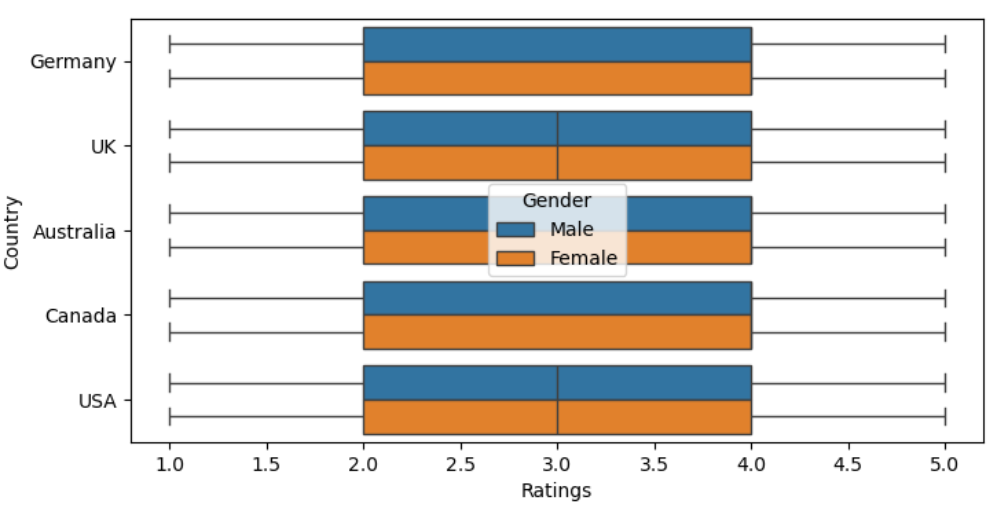


Fig 15 Ratings distribution based on Countrty and Gender

Customer ratings are similar across all countries and genders shown, with medians around 3 and minimal differences in distribution.

**Implications**

* **Customer Retention:** Vital for business sustainability and growth. More cost-effective than acquiring new customers.
* **Customer Acquisition**: Helps tailor strategies to convert first-time buyers into repeat customers.
* **Competitive Advantage:** Differentiates retailers from competitors, leading to increased market share.
* **Revenue Growth:** Satisfied customers are likely to make repeat purchases and recommend the retailer, driving revenue growth.
* **Operational Efficiency:** Streamlines operations and focuses on areas yielding the highest customer value.

**Recommendations**

* **Personalized Customer Experiences:** Use insights from customer data to provide personalized recommendations and services.
* **Targeted Marketing Strategies**: Develop marketing campaigns that resonate with customer preferences and behaviors.
* **Product and Service Improvements**: Enhance products and services based on customer feedback and satisfaction drivers.
* **Customer Support Optimization:** Implement strategies to improve customer support based on identified pain points.
* **Proactive Engagement:** Engage with customers proactively through loyalty programs, special offers, and personalized communication.

**Limitations**

* **Subjectivity of Ratings:** Customer satisfaction ratings are inherently subjective and can be influenced by individual expectations, personal preferences, and external factors that may not be directly related to the product or service.
* **Dynamic Nature of Customer Satisfaction:** Customer preferences and satisfaction levels can change over time due to market trends, competitive actions, or evolving customer expectations. The analysis may need to be updated regularly to stay relevant.
* **Bias in Ratings**: Certain customers may be more likely to leave ratings (e.g., extremely satisfied or dissatisfied customers), potentially skewing the results. It's important to account for this potential bias in the analysis.
* **External Factors**: External factors such as economic conditions, seasonality, and competitor actions can impact customer satisfaction and may not be fully accounted for in the analysis.

**Closing Reflections**

**Key Learnings from the Process**

* Importance of Data Preprocessing: Handling missing values, class imbalance, and feature engineering significantly impacted model performance.
* Significance of EDA: Understanding data distribution and relationships between features helped in making informed modeling decisions.
* Model Selection & Optimization: Comparing multiple models highlighted the strengths of ensemble techniques like XGBoost, which performed best for this classification task.
* Balancing Accuracy & Interpretability: While complex models provided better accuracy, simpler models like Decision Trees helped in understanding key decision factors.

.

**What Would Be Done Differently Next Time?**

* Feature Engineering Enhancements: Explore additional derived features to improve model performance.
* Alternative Class Imbalance Strategies: Experiment with hybrid approaches, such as focal loss or custom re-sampling techniques.
* Deep Learning Exploration: Test neural networks or transformer-based models to see if they outperform traditional machine learning approaches.
* Deployment & Real-World Testing: Implement the model in a real-world retail setting to assess its impact on customer experience and business decisions.
* Automated Hyperparameter Tuning: Use Bayesian Optimization or AutoML to further refine model performance.

**References**

<https://medium.com/@yashwanthraghuram123/how-swiggy-can-improve-customers-satisfaction-d3fee399f8ac>

<https://sist.sathyabama.ac.in/sist_naac/documents/1.3.4/39280085%20SANKALPAM.pdf>