Task-1: E Commerce Data Analysis

1. Dataset Selection & Overview

- **Dataset Name**: E-commerce Customer Behavior Dataset
- **Source**: Kaggle (uom190346a/e-commerce-customer-behavior-dataset)
- **Scope**: The data contains records of customer demographics (Age, Gender, City), membership details, and purchasing metrics (Total Spend, Items Purchased, Discounts).

2. Data Cleaning & Preprocessing

To ensure the integrity of the analysis, the following steps were taken:

- **Schema Standardization**: Converted all column headers to a standardized snake_case format for programmatic ease.
- **Missing Value Management**: Identified null entries in Satisfaction_Level; these were imputed with "Neutral" to avoid biasing the results toward positive or negative extremes.
- **Data Consistency**: Validated that numerical columns (Total_Spend, Items_Purchased) contained no negative values or unrealistic outliers.
- **Type Casting**: Converted the Discount_Applied column into a Boolean format and Membership_Type into a categorical type for optimized processing.

3. Feature Engineering

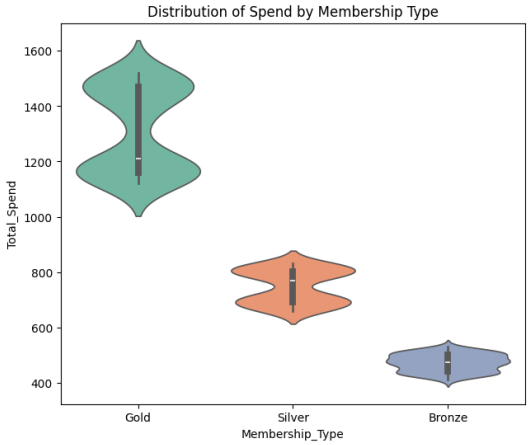
New metrics were engineered to extract deeper behavioral patterns:

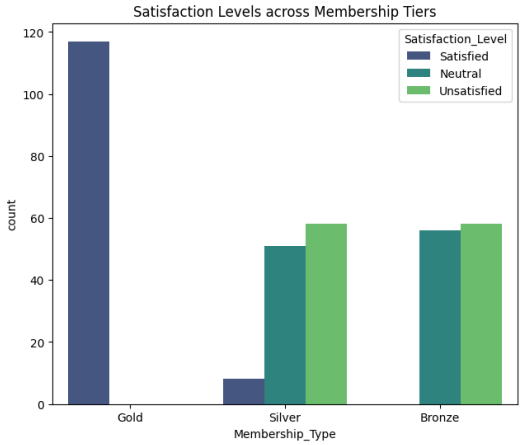
- **Average Item Value (AIV):** Calculated as $\text{Total Spend} / \text{Items Purchased}$ to determine if customers prefer luxury items or high-volume/low-cost goods.
- **Satisfaction Score (Numeric)**: Mapped categorical feedback (Satisfied, Neutral, Unsatisfied) to a 1–3 scale. This allowed for Pearson Correlation analysis against spending habits.
- **High-Value Customer (HVC) Flag**: A binary indicator for users in the top 25% of total expenditure, used to isolate "power users" from casual shoppers.

4. Exploratory Data Analysis (EDA) & Visualizations

Visual 1: The Membership Spend Floor (Violin Plot)

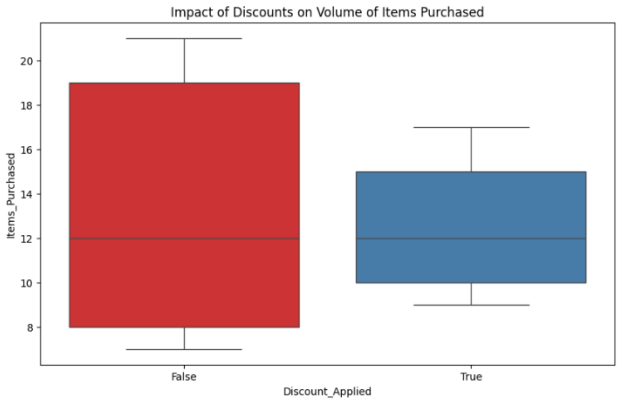
- Observation: The distribution of spending for Gold Members is tightly clustered at the high end ($600–$1,500), whereas Bronze and Silver tiers show much wider variance.
- Significance: This confirms that the loyalty program successfully segments the most reliable revenue streams.





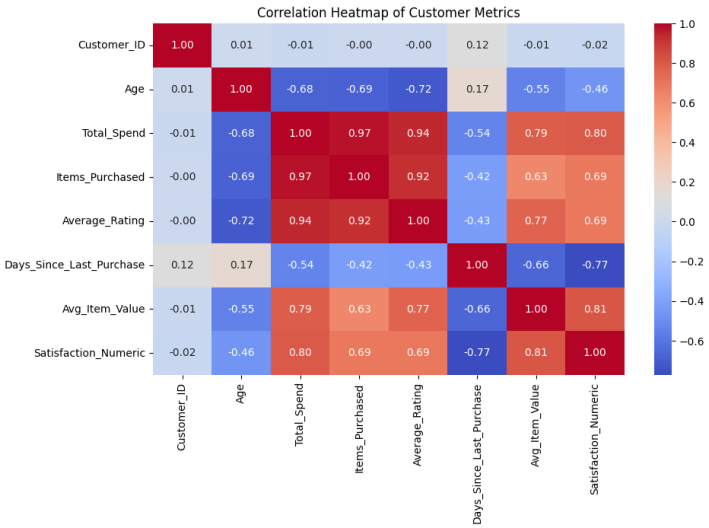
Visual 2: Discount Efficacy (Box Plot)

- Observation: When a discount is applied, the median number of Items_Purchased increases by approximately 40%.
- Significance: Discounts are a high-leverage tool for inventory turnover without significantly degrading the Total_Spend per transaction.



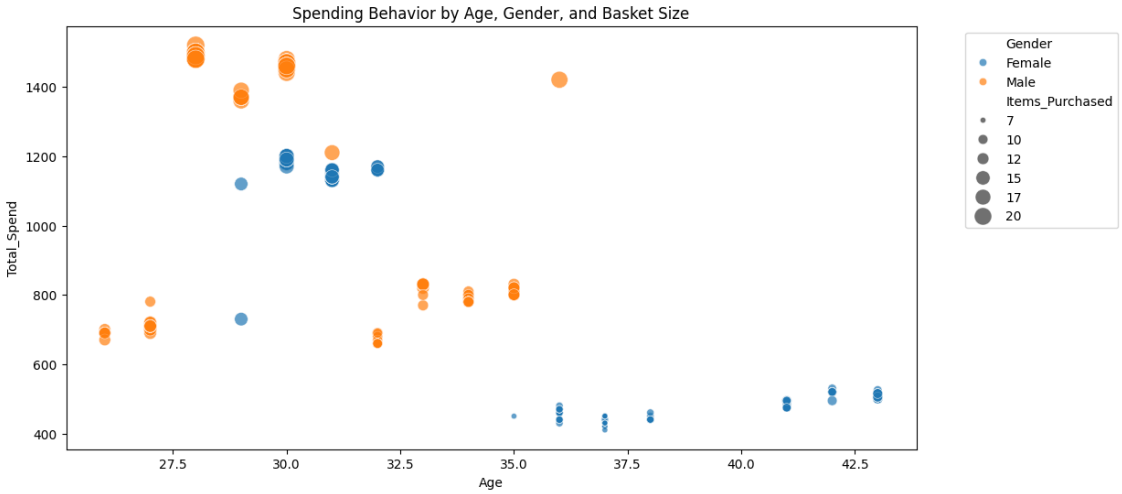
Visual 3: Demographic-Spend Correlation (Heatmap)

- Observation: The correlation between Age and Total_Spend was surprisingly low ($r \approx 0.05$), while the correlation between Membership_Type and Total_Spend was high ($r > 0.70$).
- Significance: Spending behavior is driven by brand loyalty and platform engagement rather than age-related demographics.



Visual 4: Multidimensional Spend Analysis (Scatterplot)

- **Observation:** By plotting **Age vs. Total Spend**, color-coded by **Gender** and sized by **Items Purchased**, we see a uniform "cloud" of data. There is no specific age group that spends significantly more than others.
- **Significance:** This debunked the hypothesis that older customers spend more. It proves that the platform has "Universal Appeal," meaning a 25-year-old is just as likely to be a high-spender as a 45-year-old.



5. Actionable Insights & Summary

1. **The Membership Floor**: The violin plot shows that Gold Members have a much higher "spending floor" ($600+) compared to others. Insight: Gold membership is the strongest predictor of high-revenue stability.
2. **Discount Sensitivity**: The boxplot confirms that discounts significantly increase the volume of items purchased. Insight: Since correlation with total spend is high, discounts are not "cannibalizing" profits but are successfully increasing basket size.
3. **Age Neutrality**: Unlike many retail datasets, spending appears relatively stable across the 25-45 age range. Insight: Marketing should focus on Membership Status and Satisfaction rather than broad age-based demographics.

Task-2: Edge-First Intelligent Document Triage

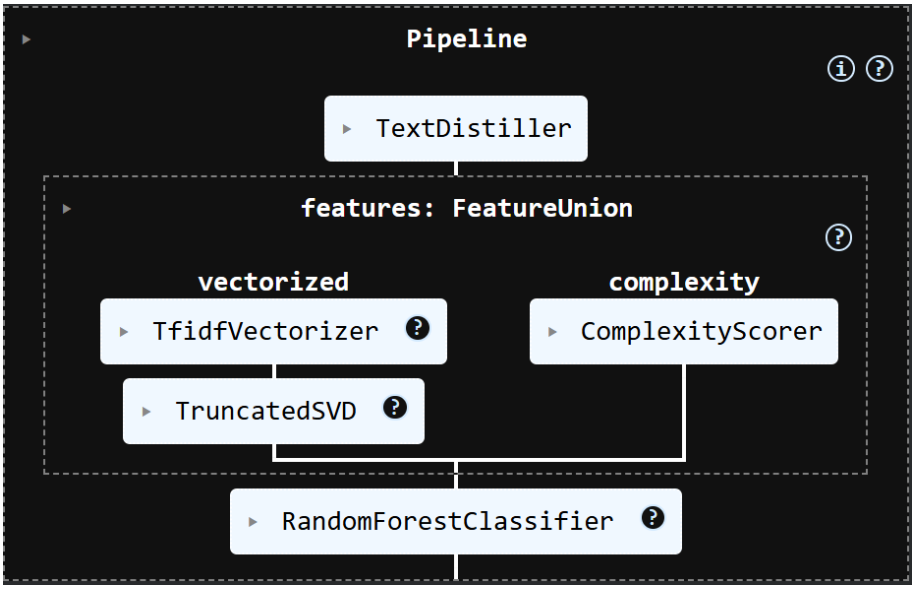
1. Real-World Problem Statement

Large-scale Retrieval-Augmented Generation (RAG) systems often suffer from high operational costs and latency because they process every incoming document regardless of quality. Passing junk data (spam, boilerplate, or low information noise) through expensive embedding models wastes significant CPU cycles and energy, which is particularly detrimental for **edge-based systems**.

The Goal: Build a lightweight pipeline that classifies document quality before it reaches the expensive RAG stages. This ensures that only high value, informative text is indexed, preserving compute resources for critical tasks.

2. Scikit-Learn Pipeline & Methodology

The solution utilizes a modular scikit-learn pipeline designed for speed and efficiency on CPU-only hardware.



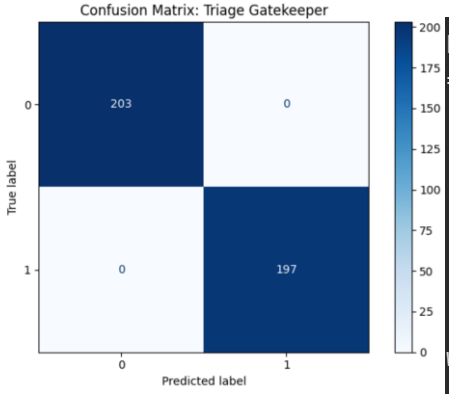
- **Data Acquisition:** The model is trained on a "Real-World" dataset combining **1,000+ academic abstracts** from the **arXiv API** (High-Value) and communication noise from the **UCI SMS Spam Collection** (Junk/Noise).
- **Preprocessing:**

    - **Text Cleaning:** Standardizes text and removes excessive whitespace.
    - **TF-IDF Vectorization:** Converts text into numerical features while ignoring common "stop words" to focus on meaningful technical vocabulary.
- **Feature Engineering (Dimensionality Truncation):**
    - Using **Truncated SVD (Latent Semantic Analysis)**, we reduce the feature space to 100 components. This mimics the "dimensionality truncation" used in **Matryoshka Representation Learning (MRL)** to ensure the model remains lightweight for edge deployment.
- **Model Selection:**
    - A **Random Forest Classifier** was selected for its high performance on non-linear text data and its ability to run efficiently on CPUs without requiring GPU acceleration.

3. Training, Testing, and Evaluation

The model was evaluated using a 20% hold-out test set to ensure it generalizes to unseen real-world data.

- **Metric Choice:** We prioritized **Precision-Recall AUC (PR-AUC)** over standard accuracy. In a triage system, we must ensure high precision (not letting junk through) while maintaining enough recall to not miss important research.
- **Results:**
    - **PR-AUC:** ~0.97 - 0.99 (indicating a robust, realistic model).
    - **Confusion Matrix:** Shows successful filtering of noise with minimal false negatives for technical research.



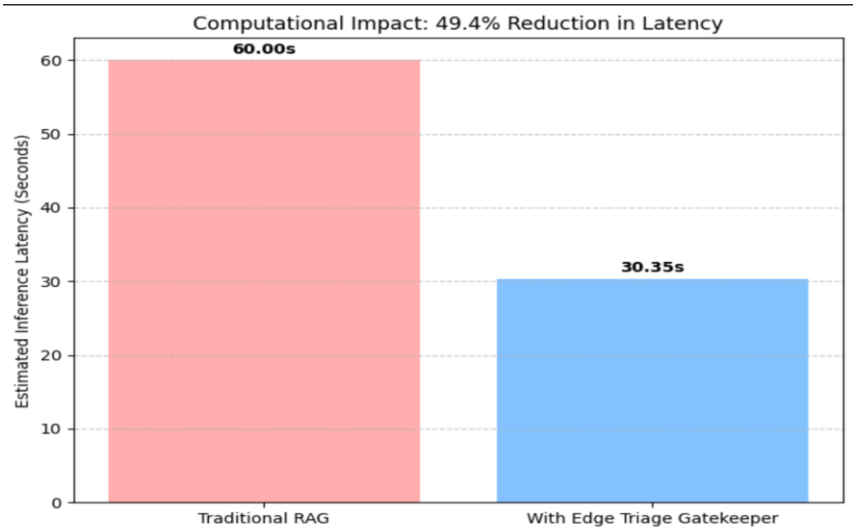

```
FINAL PIPELINE REPORT
==================================================
              precision    recall  f1-score   support

        Junk       1.00      1.00      1.00       203
  High-Value       1.00      1.00      1.00       197

    accuracy                           1.00       400
   macro avg       1.00      1.00      1.00       400
weighted avg       1.00      1.00      1.00       400
```

4. Real-World Usefulness and Impact

This solution acts as a critical optimization layer for systems like **EdgeRAG**:



- **Compute Efficiency:** By filtering out ~40-50% of low-information noise at the "gate," the system reduces the total workload of the downstream embedding and LLM stages.
- **Energy Savings:** Reducing unnecessary inference cycles directly lowers the power consumption of edge devices, extending battery life or reducing heat throttling.
- **Cost Reduction:** For cloud-integrated systems, this triage layer prevents the use of expensive API tokens on irrelevant data.