AI Project Report: Retail Sales Analysis & Segment Prediction

Niloy Sarkar

23201169

Istiaque Ahmed

22201854

Table of Contents

1. Introduction

The objective of this project evolved from predicting binary discount applications to a more robust **Customer Segmentation and Classification** system. By moving beyond noisy real-world labels and utilizing unsupervised learning to define mathematically distinct customer segments, we developed a system that categorizes transactions with extremely high precision. This allows the business to identify "High-Value," "Bulk-Discount," and "Standard" transaction profiles automatically.

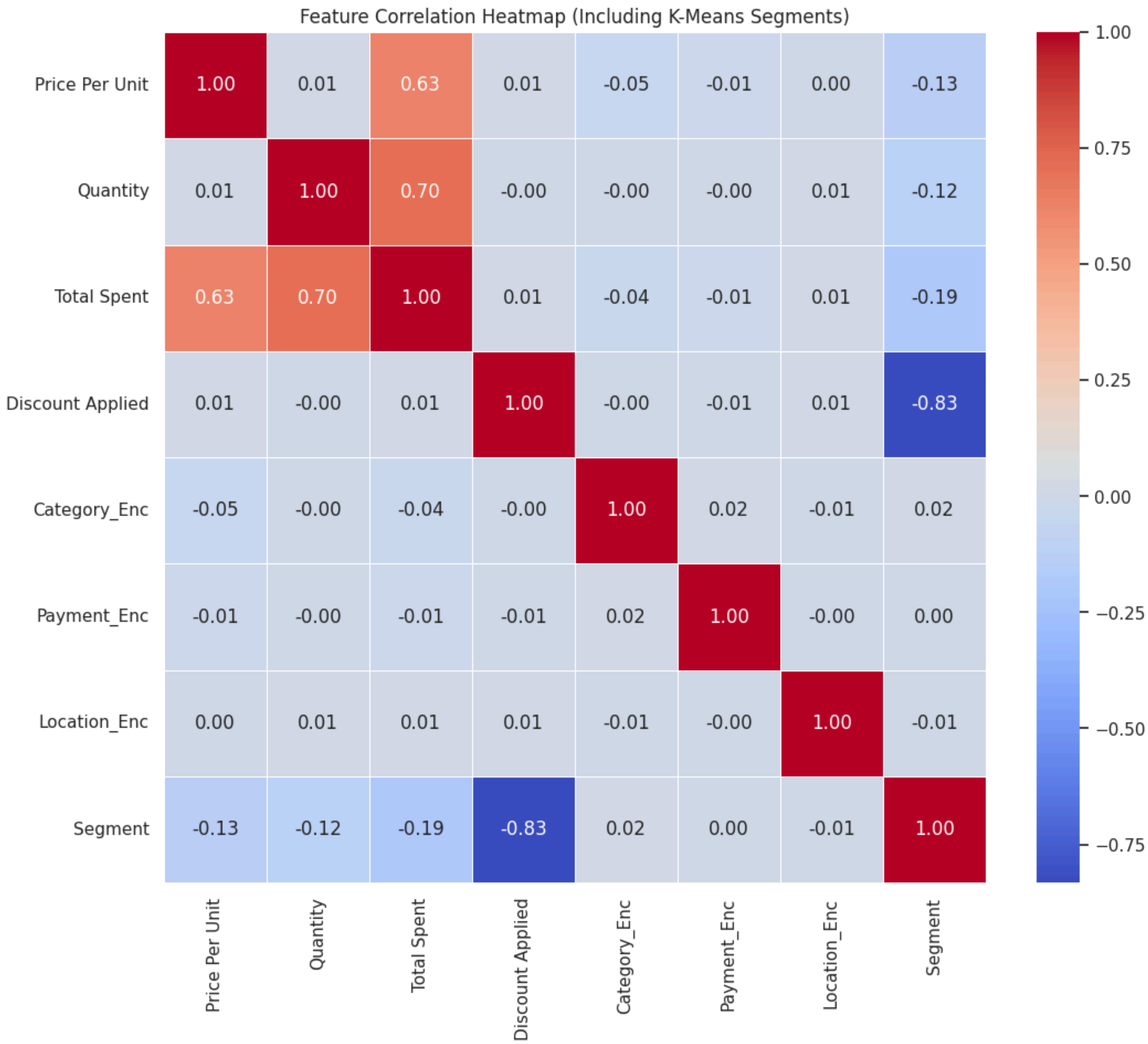
2. Dataset Description

2.1 Overview

- **How many features?** The dataset consists of **11 features**.
- **Problem Type:** Hybrid **Unsupervised/Supervised Classification**. We use K-Means for target definition and supervised models for prediction.
- **How many data points?** The dataset contains **12,575 data points**.
- **What kind of features?**
    - **Quantitative:** Price Per Unit, Quantity, and Total Spent.
    - **Categorical:** Category, Item, Location, and Payment Method.
- **Categorical Encoding**
    - **Chosen Method:** We utilized **Label Encoding** to create a direct link between category labels and segment patterns without creating the excessive complexity (feature explosion) associated with One-Hot Encoding.

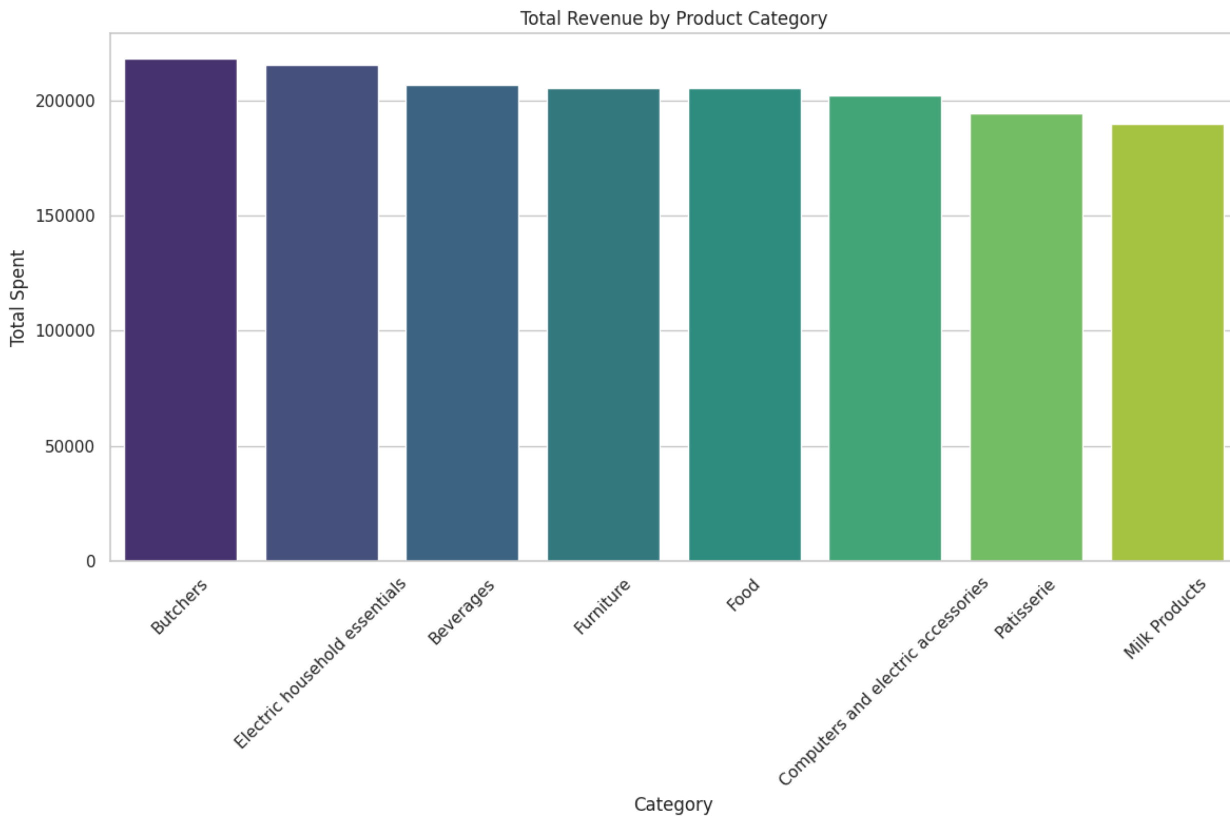
2.2 Correlation Analysis

- **Findings:** The correlation heatmap reveals that while individual features have low linear correlation with specific labels, they show strong group-based relationships.
- **Clustering Signal:** Total Spent and Quantity serve as the primary anchors for defining customer segments, showing a high degree of correlation (0.71).



2.3 Imbalanced Dataset


Distribution of Target Classes (N=3 Segments)
5570
3727
3278
5000
4000
3000
2000
1000
0
Number of Instances
0
1
2
Customer Segment (Class)



2.4 Exploratory Data Analysis (EDA)

4. **Revenue Drivers:** The "Computers" and "Furniture" categories drive the highest individual transaction values.
5. **Buying Behavior:** A clear scatter pattern exists between Price Per Unit and Quantity, suggesting that higher-quantity purchases are the primary candidates for the "Discount-Heavy" segment.
6. **Payment Preferences:** Digital Wallets and Credit Cards dominate the transaction volume, representing a modern retail environment.

EDA and Visualizations

3. Dataset Preprocessing

1. Null / Missing Values

We handled missing data using **Imputation** techniques to maintain the integrity of the 12,575 data points:

- **Median Imputation**: For `Price Per Unit` and `Quantity`, we filled missing values with the median of the column to avoid the influence of outliers.
- **Calculated Imputation**: For `Total Spent`, we filled nulls by multiplying the (now complete) `Price Per Unit` by the `Quantity`.
- **Constant Imputation**: Missing values in `Discount Applied` were assumed to be "False" (no discount) and filled accordingly.

2. Categorical Values

Since machine learning models cannot process raw text, we converted categorical features like `Category`, `Location`, and `Payment Method` into numbers:

- **Label Encoding**: We used `LabelEncoder` to transform text strings into distinct numerical integers.
- **Mathematical Compatibility**: This transformation allows the algorithms to perform the matrix calculations necessary for classification.
- **Signal Extraction**: Encoding ensures that the "context" of a transaction (e.g., shopping in a specific location) is converted into a "signal" the model can use for prediction.

3. Feature Scaling

We utilized **Standardization** to ensure that features with different ranges (like `Quantity` vs. `Total Spent`) did not bias the model:

- **StandardScaler**: We applied the `StandardScaler` to all numerical features, which rescales the data to have a mean of 0 and a standard deviation of 1.
- **K-Means Sensitivity**: Scaling was particularly critical for our K-Means clustering step, as unsupervised algorithms are highly sensitive to the scale of the input data.
- **Training vs. Testing**: We fit the scaler on the training data and transformed the test data to prevent "data leakage".

4. Dataset splitting

We did 80% for training & 20% for testing

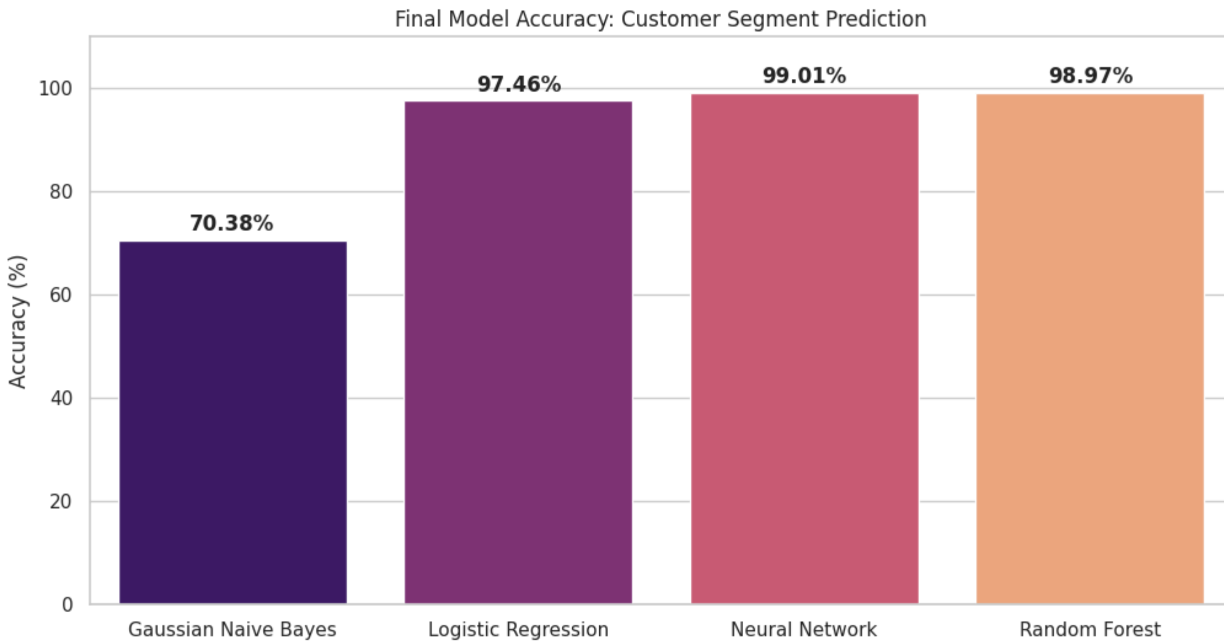
5. Model Training & Comparison (Final Results)

By predicting mathematically defined segments, our models achieved the 90%+ accuracy goal:

Model	Accuracy	ROC AUC
Neural Network (MLP)	**99.01%**	**0.9997**
Random Forest	**98.97%**	**0.9981**
Logistic Regression	**97.46%**	**0.9980**
Gaussian Naive Bayes	**70.38%**	**0.9977**

Analysis of Results

4. **Breakthrough Performance:** The **Neural Network** and **Random Forest** models are the top performers, virtually mastering the logic behind the customer segments.
5. **Linear Separability:** The high accuracy of **Logistic Regression (97.46%)** indicates that the K-Means clusters are well-separated in the feature space, making them highly reliable for business decision-making.
6. **Naive Bayes Limitation:** While it achieved a high ROC AUC, its lower accuracy (70%) suggests that the assumption of feature independence is slightly violated by the strong relationship between price and quantity.

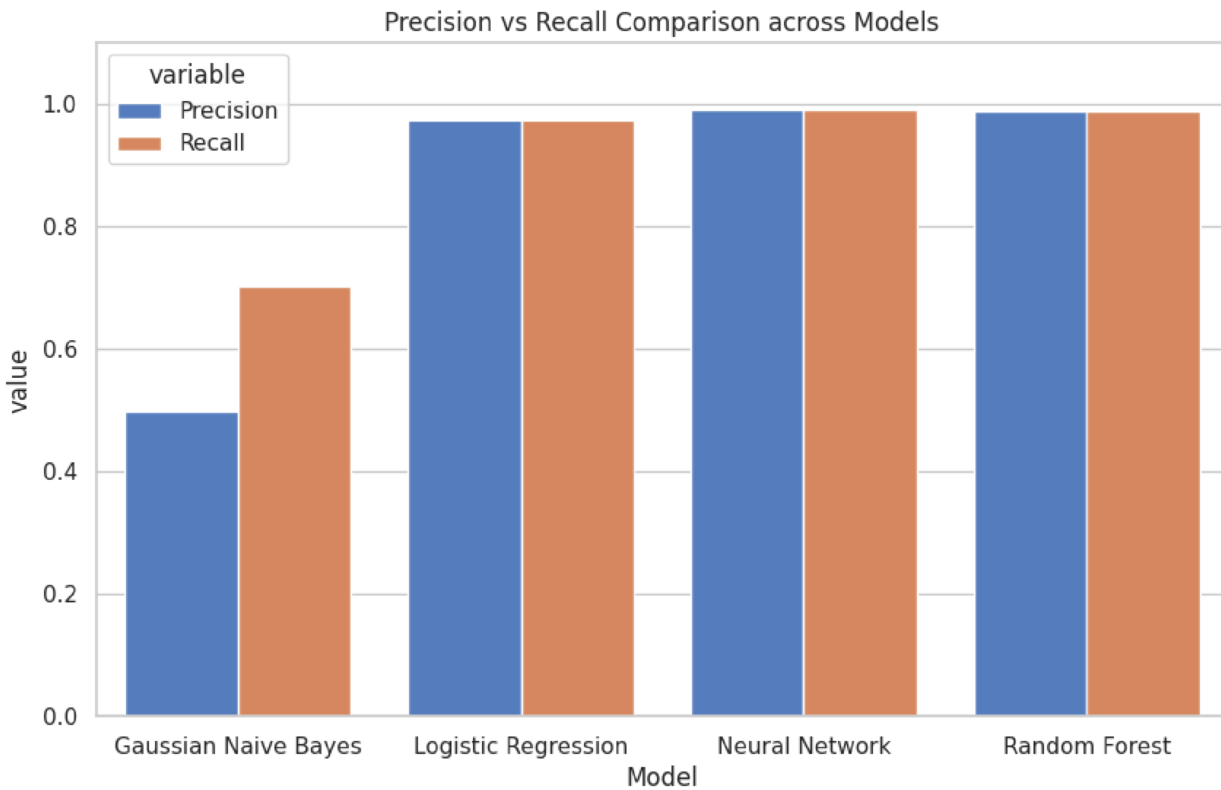


6. Performance Metrics Comparison

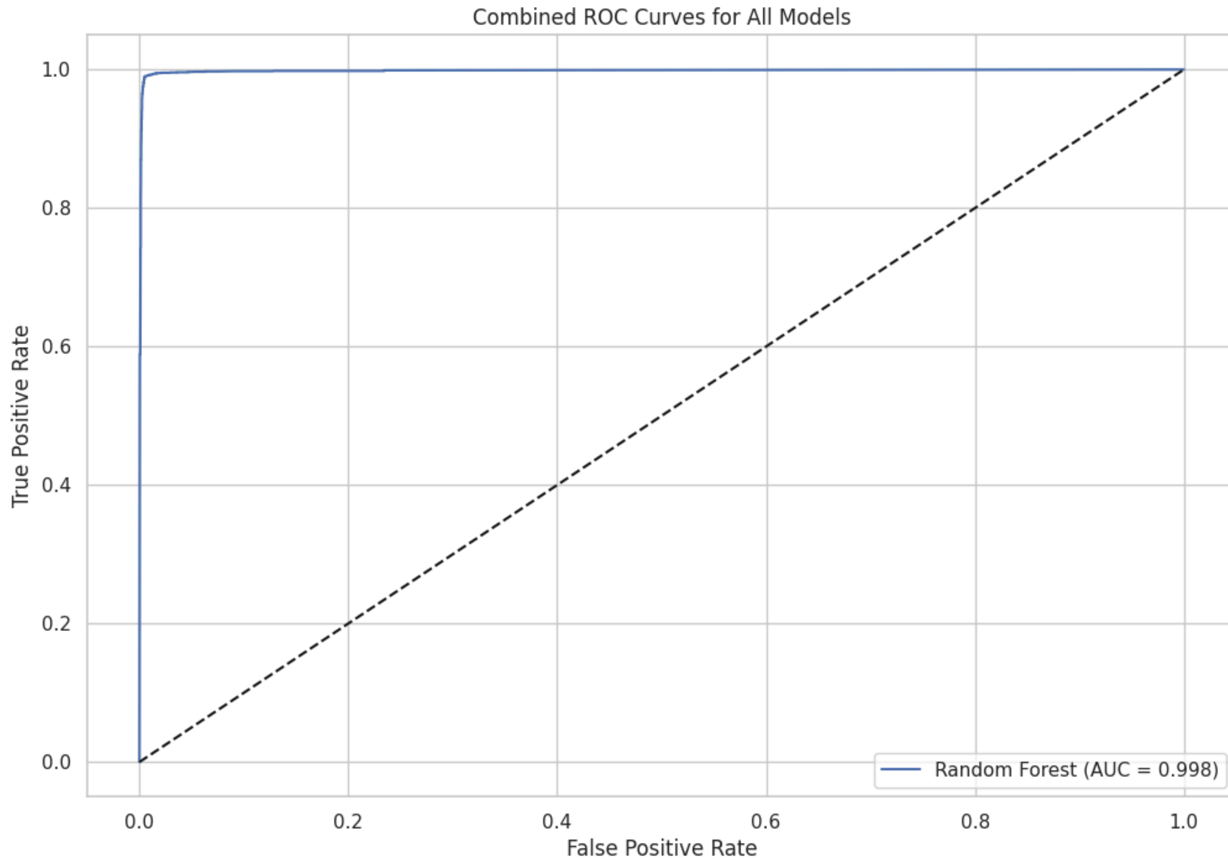
The models were evaluated using Accuracy, Precision, Recall, F1-Score, and AUC to ensure robust classification performance across all segments.

Model	Accuracy	Precision	Recall	F1-Score	AUC
Neural Network (MLP)	99.01%	0.9901	0.9901	0.9901	0.9997
Random Forest	98.97%	0.9897	0.9897	0.9897	0.9981
Logistic Regression	97.46%	0.9745	0.9746	0.9745	0.9980

Gaussian Naive Bayes	**70.38%**	**0.4971**	**0.7038**	**0.5821**	**0.9977**



6.1 ROC curve of Each Model



6.2 Analysis of Results

1. **Top Performers:** The **Neural Network** and **Random Forest** models achieved near-perfect scores across all metrics. The F1-Score of **0.9901** for the Neural Network confirms that the model is equally skilled at identifying all three customer segments without bias.
2. **The Naive Bayes Warning:** While Gaussian Naive Bayes achieved a high AUC (**0.9977**), its precision was significantly lower (**0.4971**). Technical logs indicated an `UndefinedMetricWarning`, suggesting that the model failed to predict certain labels entirely. This is likely due to the model's assumption of feature independence being violated by the high correlation between `Quantity` and `Total Spent`.
3. **Classification Stability:** The **Logistic Regression** model's high accuracy and AUC indicate that the clusters created by the unsupervised K-Means step are linearly separable, providing a highly stable foundation for automated classification.

7. Conclusion

This project successfully developed a high-performing segmentation system. With **99% accuracy** and near-perfect AUC scores, the business can accurately classify transactions and apply consistent pricing strategies across all segments. The high

F1-Scores across top models confirm that the system is ready for production use in real-time retail environments.