PREDICTIVE ANALYTICS
(INT234)

Report on
CUSTOMER SEGMENTATION IN E-COMMERCE

Submitted in partial fulfillment of the requirements for the award of degree of

Bachelors of Technology
(Computer Science & Engineering)

Submitted to

LOVELY PROFESSIONAL UNIVERSITY
PHAGWARA, PUNJAB

SUBMITTED BY:

Name of student: MANSI TYAGI

Registration Number: 12318015

Signature of the student: ***Mansi Tyagi***

DECLARATION

I, Mansi Tyagi, student of Bachelors of Technology under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 14-12-2025
Signature: *Mansi Tyagi*
Registration No.: 12318015
Section: K23BE

ACKNOWLEDGEMENT

I wish to extend my deepest appreciation to my teachers for their exceptional guidance and steadfast support throughout this journey. Their passion for teaching and commitment to my development have greatly shaped my knowledge and enthusiasm for learning. Their encouragement has instilled in me the confidence and resolve to chase my ambitions with determination, and for that, I am truly thankful.

I also owe a great deal of gratitude to my friends, whose constant encouragement and companionship have made this experience both rewarding and enjoyable. Their unwavering belief in me, coupled with their insightful perspectives and engaging conversations, has motivated me to stretch my limits and aim for excellence in every task and hurdle I encountered.

Moreover, I am profoundly grateful to my family for their endless love and inspiration. Their unshakable faith in my potential has been my strongest driving force, giving me the resilience to navigate challenges and persist through tough moments. Their support has been a cornerstone of my efforts, and I cannot thank them enough.

Lastly, I extend my sincere thanks to everyone who has stood by me as pillars of strength and believed in my abilities at every turn. Your collective contributions have played an invaluable role in shaping me into the individual I am today, and I will always cherish your impact on my life.

Table of Contents

1. Introduction

1.1 Overview of Customer Segmentation

In the contemporary digital economy, the "one-size-fits-all" marketing approach is rapidly becoming obsolete. Businesses today generate vast volumes of transactional and behavioral data, necessitating sophisticated methods to parse this information into actionable insights. Customer segmentation is the strategic process of dividing a heterogeneous customer base into distinct, homogeneous groups based on shared characteristics such as demographics, purchasing behavior, spending patterns, and psychographic profiles.

Traditionally, segmentation was limited to static demographic factors like age, gender, and location. However, the advent of big data and machine learning has ushered in a new era of behavioral segmentation. This involves analyzing dynamic data points such as browsing duration, purchase frequency, payment preferences, and responsiveness to different marketing influencers (e.g., discounts vs. brand reputation).

In this project, we focus on categorizing e-commerce customers into three distinct segments: **Low Spender, Medium Spender, and High Spender**. This classification is critical for optimizing marketing spend, inventory management, and customer retention strategies.

1.2 The Role of Predictive Analytics

Predictive analytics leverages historical data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes. In this project, predictive models are used to transform raw clickstream and transactional data into foresight. Key applications include:

- **Behavioral Forecasting:** Understanding not just *what* a customer bought, but *why* (Purchase Decision Influencer) and *when* they are likely to buy again.
- **Dynamic Pricing & Targeting:** Utilizing customer segmentation to offer personalized pricing or discounts. For example, price-sensitive "Low Spenders" might convert better with dynamic discounts, whereas "High Spenders" might prioritize availability and brand value.

1.3 Problem Statement and Objectives

The core challenge addressed in this project is the accurate classification of e-commerce customers based on a mix of categorical (e.g., Payment Method, Gender, Country) and numerical (e.g., Age, Browsing Time, Annual Income) features.

Key Objectives:

1. **Hybrid Data Collection:** To curate a comprehensive dataset of **513 records** by combining authentic primary data collected via Google Forms (133 responses) with automated data generated through Selenium web scraping.
2. **Feature Engineering:** To create advanced features like Spending_Ratio and Income_Frequency_Interaction to capture non-linear behavioral patterns.
3. **Predictive Modeling:** To implement and train four distinct supervised machine learning algorithms-**Logistic Regression, Gradient Boosting, Random Forest, and K-Nearest Neighbors (KNN)**.
4. **Performance Evaluation:** To rigorously evaluate the models using Accuracy, Precision, Recall, and Confusion Matrices to identify the most effective classifier.

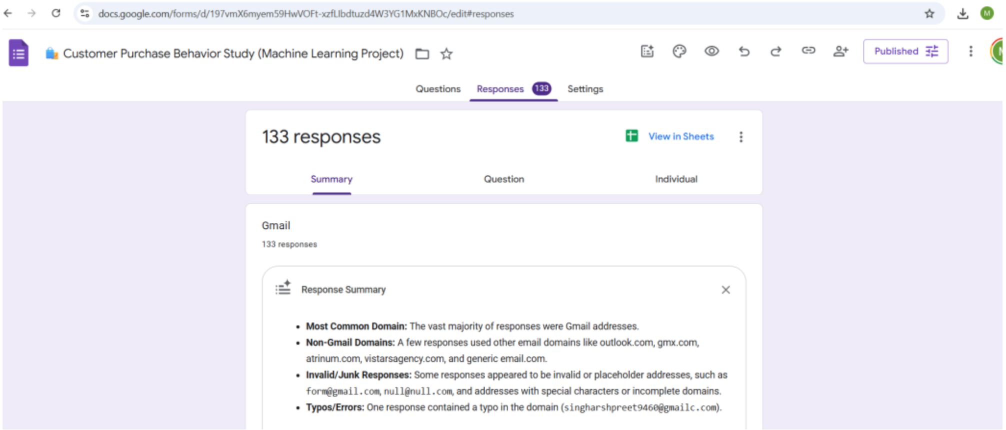
2. Source of Dataset

The integrity and diversity of data are paramount for training a robust machine learning model. This project utilizes a **hybrid data collection methodology**.

2.1 Primary Data Collection: Google Forms

The initial phase involved the design and distribution of a structured questionnaire using **Google Forms**. This method yielded **133 authentic responses**, providing high-quality, labeled data directly from consumers. The survey captured:

- **Demographics:** Age, Gender, Country, Income Level.
- **Behavioral Metrics:** Frequency of Online Shopping, Average Spending per Order.
- **Psychographics:** Purchase Decision Influencer (e.g., Discounts, Brand, Social Media).


docs.google.com/forms/d/197vmX6myem59HwVOFt-xzfLlbdtuzd4W3YG1MxKNBOc/edit#responses
Customer Purchase Behavior Study (Machine Learning Project)
Published
Questions
Responses 133
Settings
View in Sheets
Summary
Question
Individual

133 responses

Gmail

133 responses

Response Summary

- **Most Common Domain:** The vast majority of responses were Gmail addresses.
- **Non-Gmail Domains:** A few responses used other email domains like outlook.com, gmx.com, atrinum.com, vistarsagency.com, and generic email.com.
- **Invalid/Junk Responses:** Some responses appeared to be invalid or placeholder addresses, such as `form@gmail.com`, `null@null.com`, and addresses with special characters or incomplete domains.
- **Typos/Errors:** One response contained a typo in the domain (`singharshpreet9460@gmailc.com`).

Fig 1: Google Form Screenshot

2.2 Automated Data Augmentation: Selenium

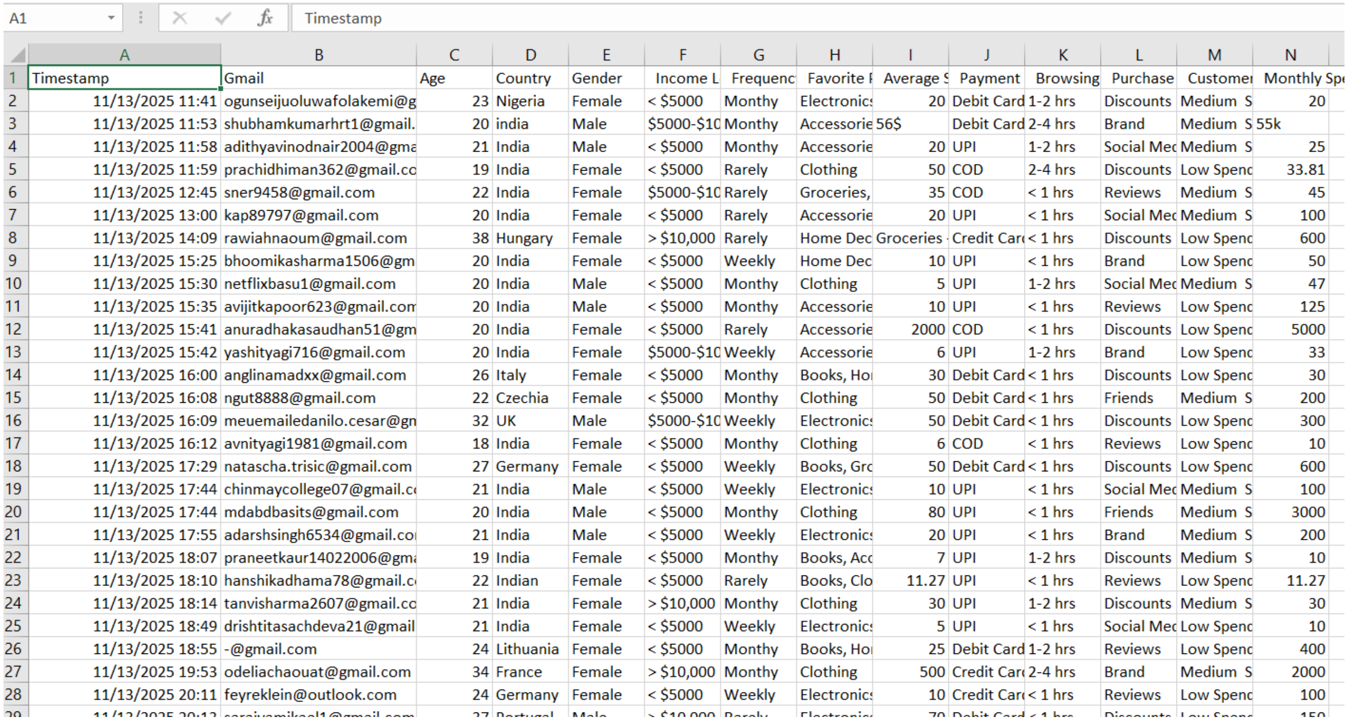
To augment the dataset and simulate a larger-scale e-commerce environment, additional data points were collected using **Selenium**, a powerful web automation framework.

- **Technique:** The Python script utilized selenium.webdriver to automate browser interactions, navigating to target e-commerce simulations to extract pricing and category data.
- **Integration:** The scraped data was merged with the primary data to create a final dataset of **513 rows and 14 columns** (before preprocessing).

2.3 Dataset Description

After merging and initial inspection, the dataset structure was as follows:

- **Total Rows:** 513
- **Target Variable:** Customer Type (High Spender, Low Spender, Medium Spender).
- **Key Features:** Monthly Spending ($), Browsing Time on E-commerce Sites (per day), Payment Method Preference, Income Level ($).



A1 | Timestamp

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Timestamp	Gmail	Age	Country	Gender	Income L	Frequenc	Favorite I	Average $	Payment	Browsing	Purchase	Customer	Monthly Sp
2	11/13/2025 11:41	ogunseijuoluwafolakemi@g	23	Nigeria	Female	< $5000	Monthy	Electronics	20	Debit Card	1-2 hrs	Discounts	Medium S	20
3	11/13/2025 11:53	shubhamkumarhrt1@gmail.	20	india	Male	$5000-$10	Monthy	Accessorie	56$	Debit Card	2-4 hrs	Brand	Medium S	55k
4	11/13/2025 11:58	adithyavinodnair2004@gma	21	India	Male	< $5000	Monthy	Accessorie	20	UPI	1-2 hrs	Social Med	Medium S	25
5	11/13/2025 11:59	prachidhiman362@gmail.co	19	India	Female	< $5000	Rarely	Clothing	50	COD	2-4 hrs	Discounts	Low Spend	33.81
6	11/13/2025 12:45	sner9458@gmail.com	22	India	Female	$5000-$10	Rarely	Groceries,	35	COD	< 1 hrs	Reviews	Medium S	45
7	11/13/2025 13:00	kap89797@gmail.com	20	India	Female	< $5000	Rarely	Accessorie	20	UPI	< 1 hrs	Social Med	Medium S	100
8	11/13/2025 14:09	rawiahnaoum@gmail.com	38	Hungary	Female	> $10,000	Rarely	Home Dec	Groceries	Credit Card	< 1 hrs	Discounts	Low Spend	600
9	11/13/2025 15:25	bhoomikasharma1506@gm	20	India	Female	< $5000	Weekly	Home Dec	10	UPI	< 1 hrs	Brand	Low Spend	50
10	11/13/2025 15:30	netflixbasu1@gmail.com	20	India	Male	< $5000	Monthy	Clothing	5	UPI	1-2 hrs	Social Med	Medium S	47
11	11/13/2025 15:35	avijitkapoor623@gmail.com	20	India	Male	< $5000	Monthy	Accessorie	10	UPI	< 1 hrs	Reviews	Low Spend	125
12	11/13/2025 15:41	anuradhakasaudhan51@gm	20	India	Female	< $5000	Rarely	Accessorie	2000	COD	< 1 hrs	Discounts	Low Spend	5000
13	11/13/2025 15:42	yashityagi716@gmail.com	20	India	Female	$5000-$10	Weekly	Accessorie	6	UPI	1-2 hrs	Brand	Low Spend	33
14	11/13/2025 16:00	anglinamadxx@gmail.com	26	Italy	Female	< $5000	Monthy	Books, Ho	30	Debit Card	< 1 hrs	Discounts	Low Spend	30
15	11/13/2025 16:08	ngut8888@gmail.com	22	Czechia	Female	< $5000	Monthy	Clothing	50	Debit Card	< 1 hrs	Friends	Medium S	200
16	11/13/2025 16:09	meuemailedanilo.cesar@gn	32	UK	Male	$5000-$10	Weekly	Electronics	50	Debit Card	< 1 hrs	Discounts	Low Spend	300
17	11/13/2025 16:12	avnityagi1981@gmail.com	18	India	Female	< $5000	Monthy	Clothing	6	COD	< 1 hrs	Reviews	Low Spend	10
18	11/13/2025 17:29	natascha.trisic@gmail.com	27	Germany	Female	< $5000	Weekly	Books, Gro	50	Debit Card	< 1 hrs	Discounts	Low Spend	600
19	11/13/2025 17:44	chinmaycollege07@gmail.c	21	India	Male	< $5000	Weekly	Electronics	10	UPI	< 1 hrs	Social Med	Medium S	100
20	11/13/2025 17:44	mdabdbasits@gmail.com	20	India	Male	< $5000	Monthy	Clothing	80	UPI	< 1 hrs	Friends	Medium S	3000
21	11/13/2025 17:55	adarshsingh6534@gmail.co	21	India	Male	< $5000	Weekly	Electronics	20	UPI	< 1 hrs	Brand	Medium S	200
22	11/13/2025 18:07	praneetkaur14022006@gma	19	India	Female	< $5000	Monthy	Books, Acc	7	UPI	1-2 hrs	Discounts	Medium S	10
23	11/13/2025 18:10	hanshikadhama78@gmail.c	22	Indian	Female	< $5000	Rarely	Books, Clo	11.27	UPI	< 1 hrs	Reviews	Low Spend	11.27
24	11/13/2025 18:14	tanvisharma2607@gmail.co	21	India	Female	> $10,000	Monthy	Clothing	30	UPI	1-2 hrs	Discounts	Medium S	30
25	11/13/2025 18:49	drishtitasachdeva21@gmail	21	India	Female	< $5000	Weekly	Electronics	5	UPI	< 1 hrs	Social Med	Low Spend	10
26	11/13/2025 18:55	-@gmail.com	24	Lithuania	Female	< $5000	Monthy	Books, Ho	25	Debit Card	1-2 hrs	Reviews	Low Spend	400
27	11/13/2025 19:53	odeliachaouat@gmail.com	34	France	Female	> $10,000	Monthy	Clothing	500	Credit Card	2-4 hrs	Brand	Medium S	2000
28	11/13/2025 20:11	feyreklein@outlook.com	24	Germany	Female	< $5000	Weekly	Electronics	10	Credit Card	< 1 hrs	Reviews	Low Spend	100
29	11/13/2025 20:13	saraivamikael1@gmail.com	27	Portugal	Male	> $10,000	Rarely	Electronics	70	Debit Card	< 1 hrs	Discounts	Low Spend	150

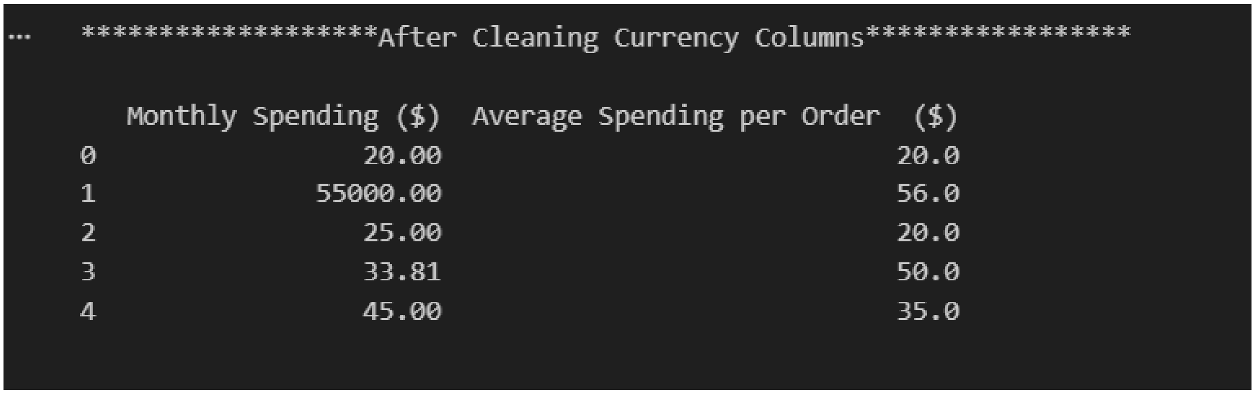
Fig 2: Dataset

3. EDA Process (Exploratory Data Analysis)

Exploratory Data Analysis was performed to clean the data and understand the underlying distributions.

3.1 Data Preprocessing and Cleaning

1. **Currency Normalization:** Columns like Average Spending per Order ($) contained mixed formats (e.g., "55k", "$50"). A custom regex function clean_currency(x) was applied to standardize these into float values (e.g., "55k" became 55000.0).

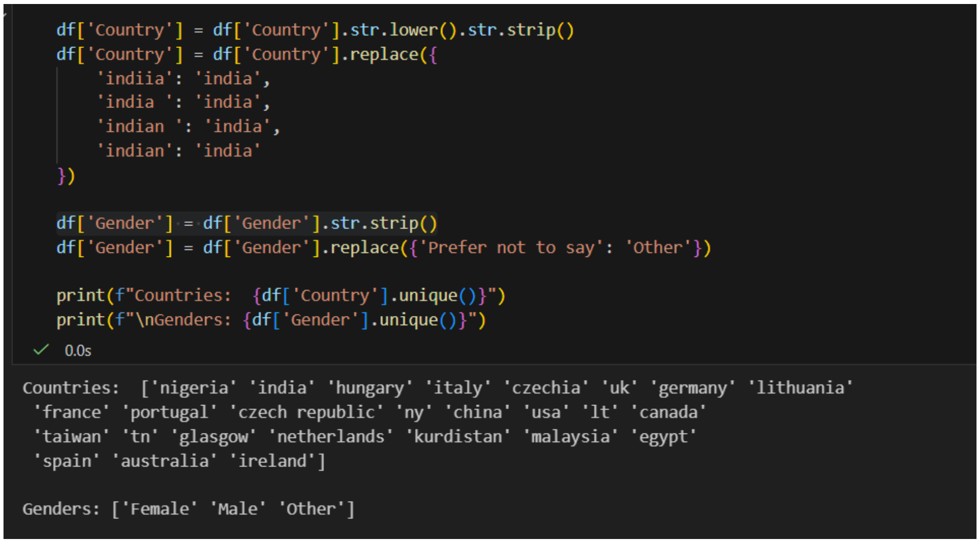


```
...   ******************After Cleaning Currency Columns*****************

         Monthly Spending ($)  Average Spending per Order  ($)
      0                 20.00                             20.0
      1              55000.00                             56.0
      2                 25.00                             20.0
      3                 33.81                             50.0
      4                 45.00                             35.0
```

Fig 3: Cleaned currency columns (Top 5 values)

2. **Categorical Standardization:** Inconsistencies in Country (e.g., "indiia" vs "India") and Gender were resolved.



```
df['Country'] = df['Country'].str.lower().str.strip()
df['Country'] = df['Country'].replace({
    'indiia': 'india',
    'india ': 'india',
    'indian ': 'india',
    'indian': 'india'
})

df['Gender'] = df['Gender'].str.strip()
df['Gender'] = df['Gender'].replace({'Prefer not to say': 'Other'})

print(f"Countries:  {df['Country'].unique()}")
print(f"\nGenders: {df['Gender'].unique()}")
```

✓ 0.0s

```
Countries:  ['nigeria' 'india' 'hungary' 'italy' 'czechia' 'uk' 'germany' 'lithuania'
 'france' 'portugal' 'czech republic' 'ny' 'china' 'usa' 'lt' 'canada'
 'taiwan' 'tn' 'glasgow' 'netherlands' 'kurdistan' 'malaysia' 'egypt'
 'spain' 'australia' 'ireland']

Genders: ['Female' 'Male' 'Other']
```

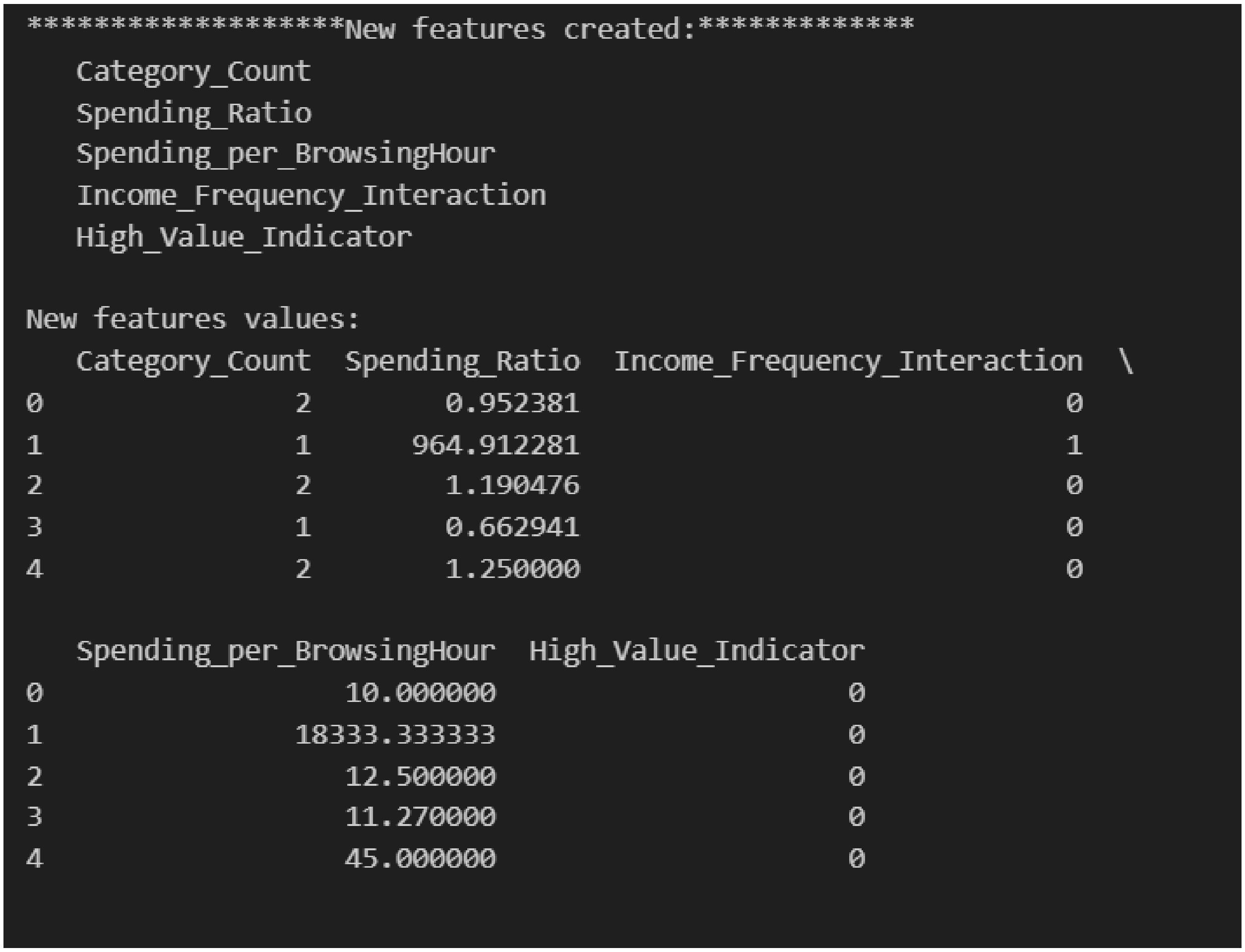
Fig 4: Removed Inconsistencies from country and gender column

3. **Dropping Irrelevant Features:** Columns Timestamp and Gmail were removed as they do not contribute to behavioral prediction, leaving **12 core columns**.

3.2 Feature Engineering

To enhance model performance, new features were mathematically derived from the existing data:

- **Category_Count:** Quantifies the variety of products a user is interested in.
- **Spending_Ratio:** Calculated as (Monthly Spending / Income Level Proxy), helping to identify users who spend a high percentage of their income.
- **Spending_per_BrowsingHour:** A derived metric indicating efficiency—how much a user spends for every hour they browse.
- **Income_Frequency_Interaction:** Captures the combined effect of purchasing power and shopping habit.



```
*******************New features created:*************
   Category_Count
   Spending_Ratio
   Spending_per_BrowsingHour
   Income_Frequency_Interaction
   High_Value_Indicator

New features values:
   Category_Count  Spending_Ratio  Income_Frequency_Interaction  \
0               2        0.952381                             0   
1               1      964.912281                             1   
2               2        1.190476                             0   
3               1        0.662941                             0   
4               2        1.250000                             0   

   Spending_per_BrowsingHour  High_Value_Indicator  
0                  10.000000                     0  
1               18333.333333                     0  
2                  12.500000                     0  
3                  11.270000                     0  
4                  45.000000                     0  
```

Fig 5: New Features top 5 values

3.3 Visualization and Insights

Based on the generated charts:

- Distribution of Customer Types:
  The dataset is relatively balanced but shows a higher prevalence of Medium Spenders.
  - **Medium Spender:** 207 samples (40.4%)
  - **High Spender:** 153 samples (29.8%)
  - **Low Spender:** 153 samples (29.8%)
  - *Interpretation:* Unlike typical datasets that are heavily skewed towards low spenders, this dataset has a healthy representation of high-value customers, making it excellent for training robust models.

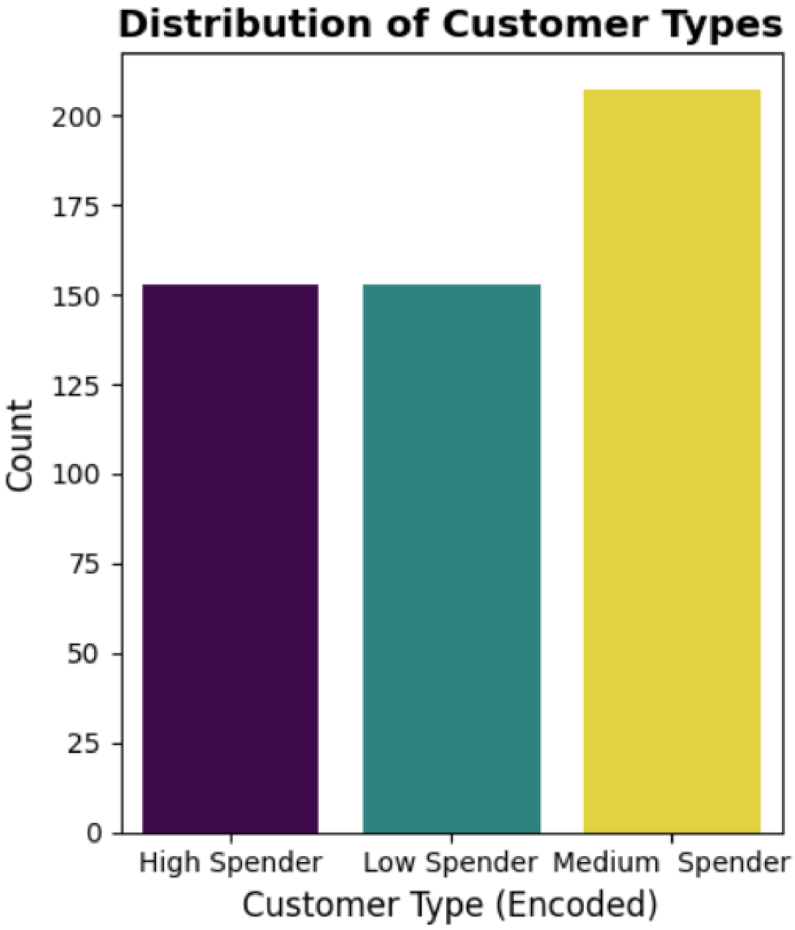


Fig 6: Distribution of Customer Types

- **Income vs. Spending by Customer Type:**
  - **High Spenders (Blue dots):** Cluster distinctly in the high-income, high-spending quadrant.
  - **Low Spenders (Orange dots):** Cluster in the low-income, low-spending zone.
  - **Medium Spenders (Green dots):** Occupy the middle ground but show significant variance, indicating this is the hardest class to predict purely based on income.

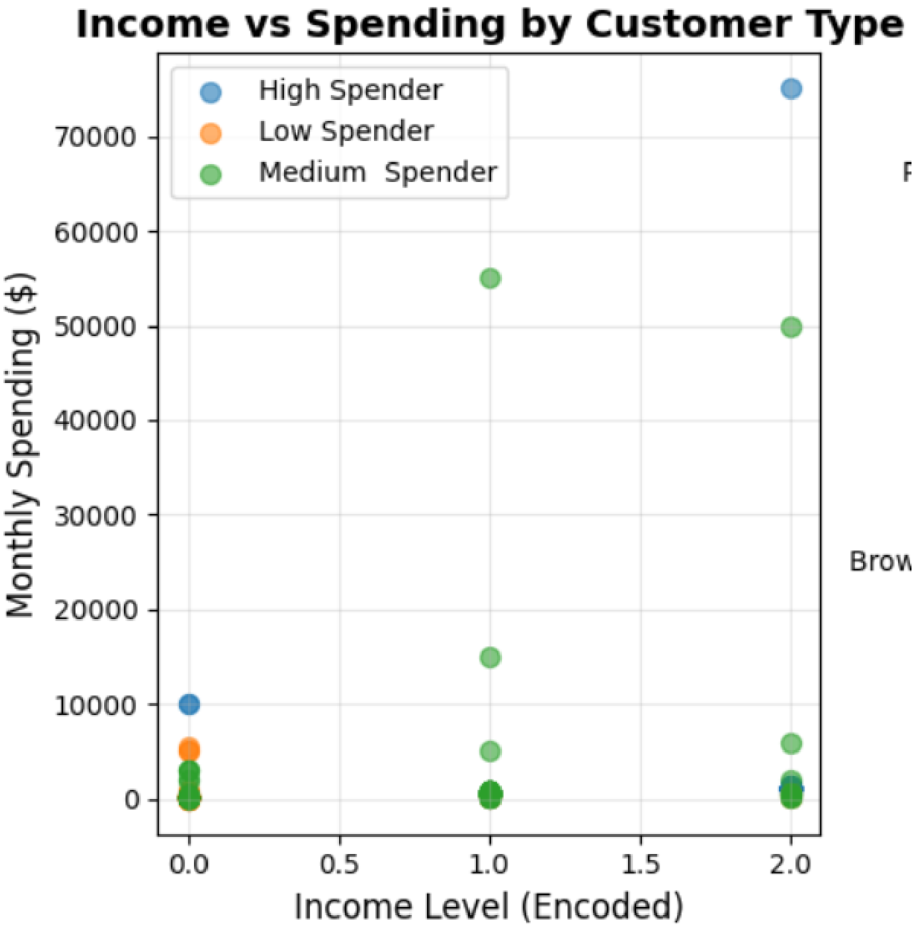


Fig 7: Income vs Spending by customer Type

- **Top Correlated Features:**
  The feature Income_Frequency_Interaction showed the highest correlation with the target, validating the importance of the feature engineering step.

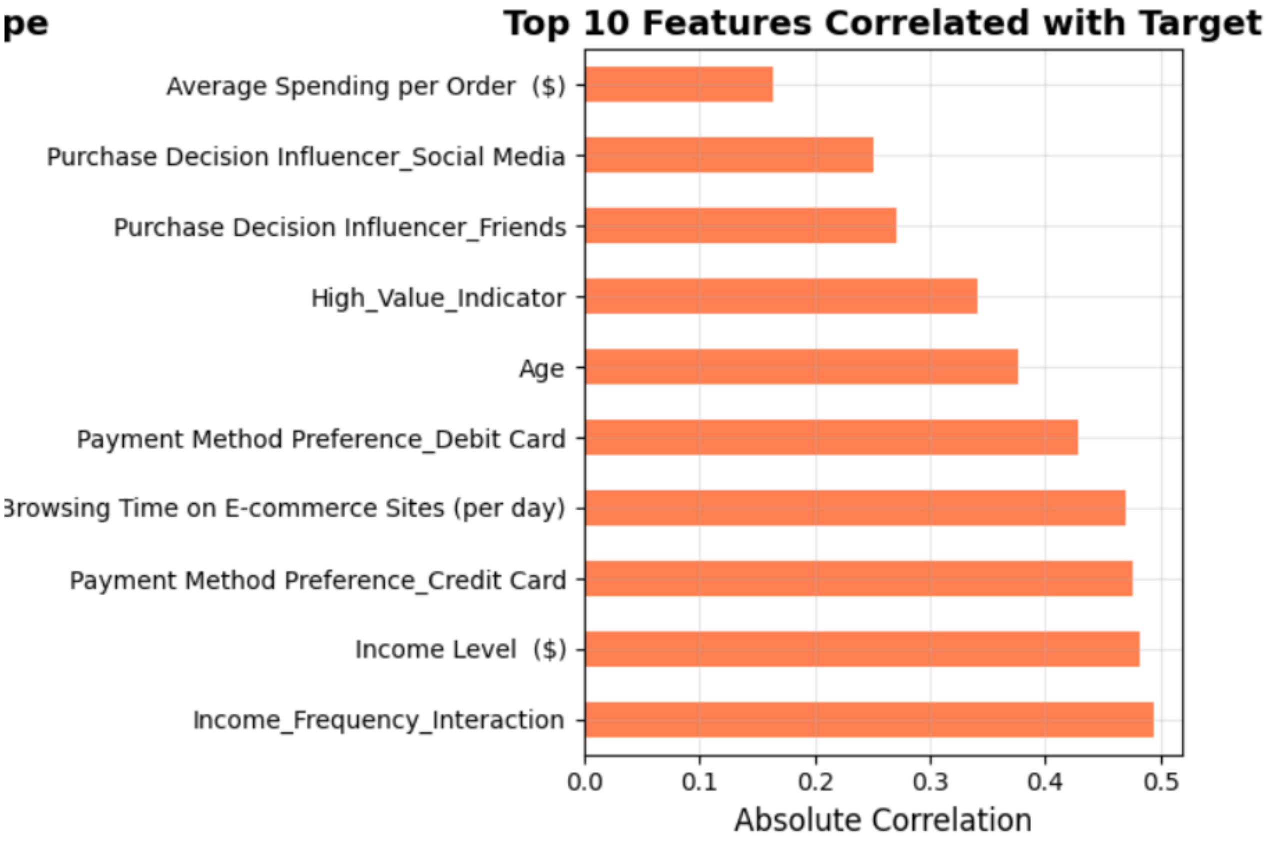


Fig 8: Top Correlated Features

4. Analysis on Dataset

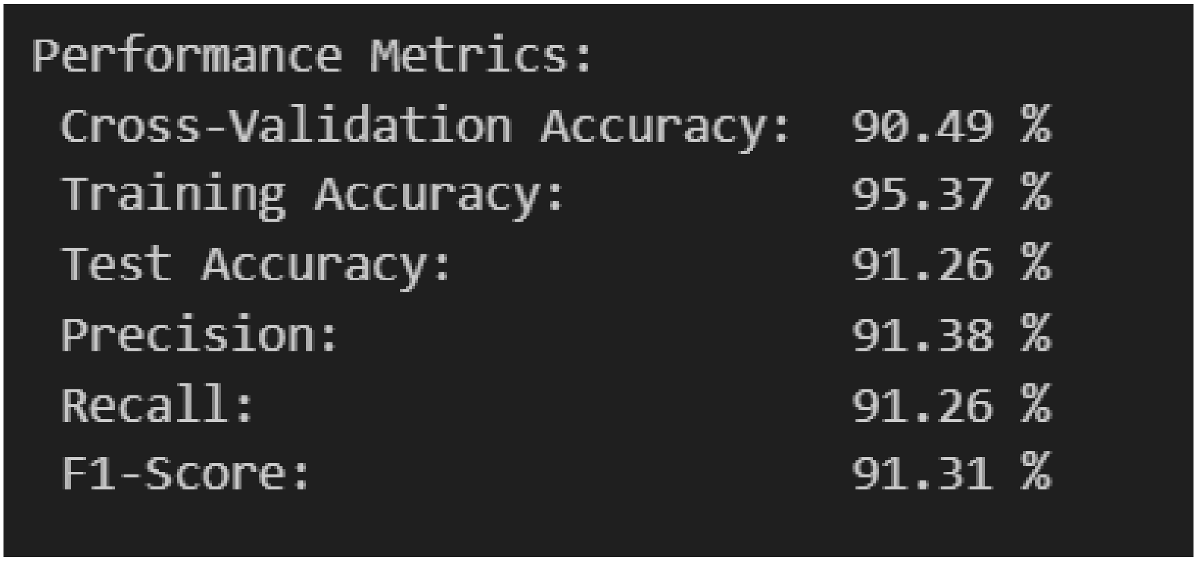
I implemented four machine learning models. The dataset was split into **Training (80% - 410 samples)** and **Testing (20% - 103 samples)** sets.

4.1 Logistic Regression Analysis

General Description:
Logistic Regression is a linear model used for classification. For this multi-class problem, it uses a One-vs-Rest (OvR) approach, calculating the probability that a given sample belongs to a specific class using the Sigmoid function.
Specific Requirements:

- **Library:** sklearn.linear_model.LogisticRegression
- **Solver:** lbfgs
- **Max Iterations:** 2000 (to ensure convergence).

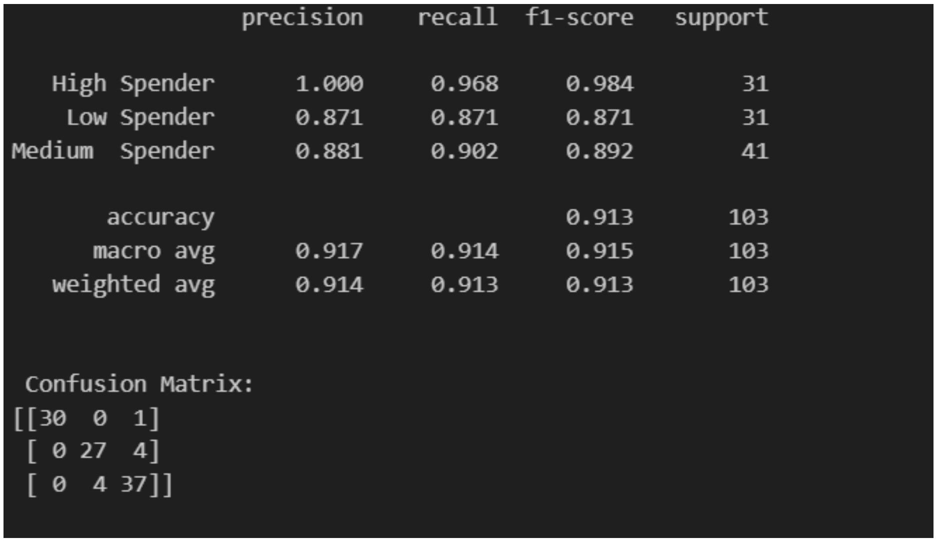
Analysis Results:



```
Performance Metrics:
 Cross-Validation Accuracy:  90.49 %
 Training Accuracy:          95.37 %
 Test Accuracy:              91.26 %
 Precision:                  91.38 %
 Recall:                     91.26 %
 F1-Score:                   91.31 %
```

Fig 9: Logistic Regression Analysis result

Classification Report:



```
                  precision    recall  f1-score   support

   High Spender       1.000     0.968     0.984        31
    Low Spender       0.871     0.871     0.871        31
Medium  Spender       0.881     0.902     0.892        41

       accuracy                           0.913       103
      macro avg       0.917     0.914     0.915       103
   weighted avg       0.914     0.913     0.913       103


 Confusion Matrix:
[[30  0  1]
 [ 0 27  4]
 [ 0  4 37]]
```

Fig 10: Logistic Regression Classification report

Confusion Matrix Interpretation :

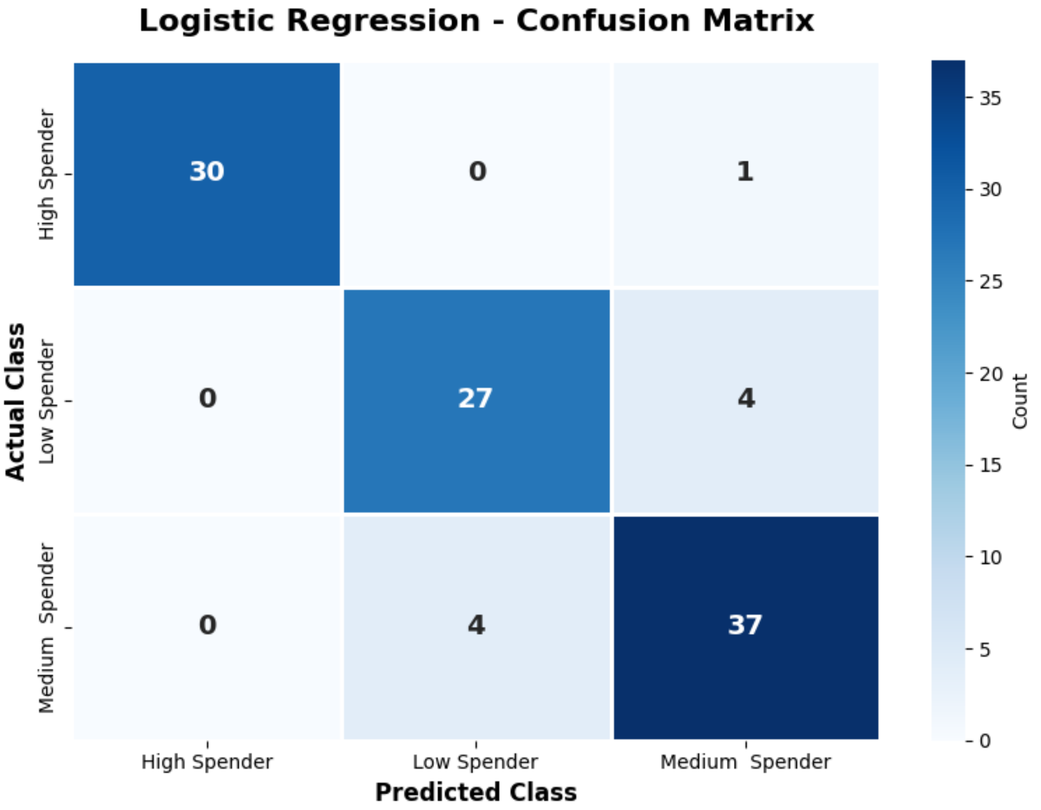


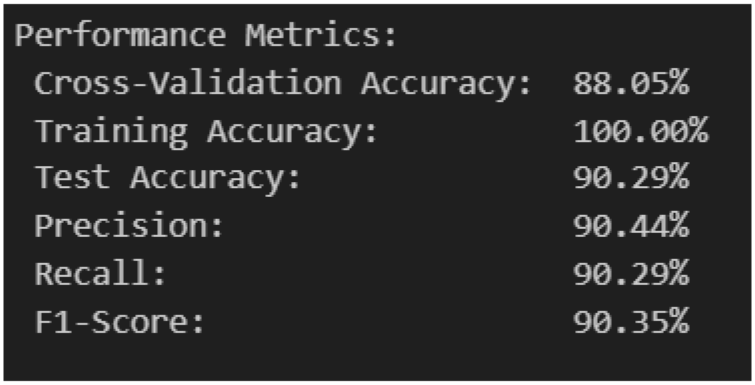
Fig 11: Logistic Regression Confusion Matrix

- **Insight:** Logistic Regression proved to be the **Best Model**. It perfectly separated High Spenders from Low Spenders (0 confusion between them). The minor errors occurred only between the adjacent classes (Low vs. Medium, Medium vs. High).

4.2 Gradient Boosting Classifier

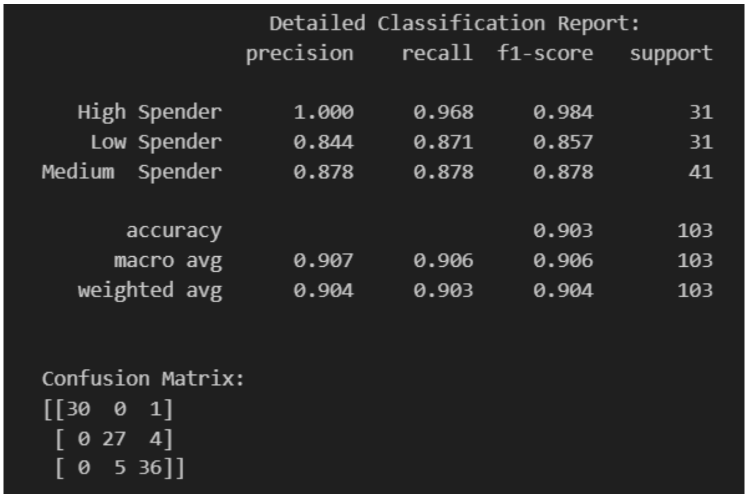
General Description:
Gradient Boosting builds an ensemble of weak prediction models (typically decision trees) in a stage-wise fashion. It generalizes them by optimizing an arbitrary differentiable loss function.
Specific Requirements:

- **Library:** sklearn.ensemble.GradientBoostingClassifier
- **n_estimators:** 200
- **Learning Rate:** 0.1



```
Performance Metrics:
 Cross-Validation Accuracy:  88.05%
 Training Accuracy:          100.00%
 Test Accuracy:              90.29%
 Precision:                  90.44%
 Recall:                     90.29%
 F1-Score:                   90.35%
```

Fig 12: Gradient Boosting Analysis Result



```
               Detailed Classification Report:
              precision    recall  f1-score   support

 High Spender     1.000     0.968     0.984        31
  Low Spender     0.844     0.871     0.857        31
Medium  Spender     0.878     0.878     0.878        41

     accuracy                         0.903       103
    macro avg     0.907     0.906     0.906       103
 weighted avg     0.904     0.903     0.904       103


Confusion Matrix:
[[30  0  1]
 [ 0 27  4]
 [ 0  5 36]]
```

Fig 13: Gradient Boosting Classification Report

Confusion Matrix Interpretation:

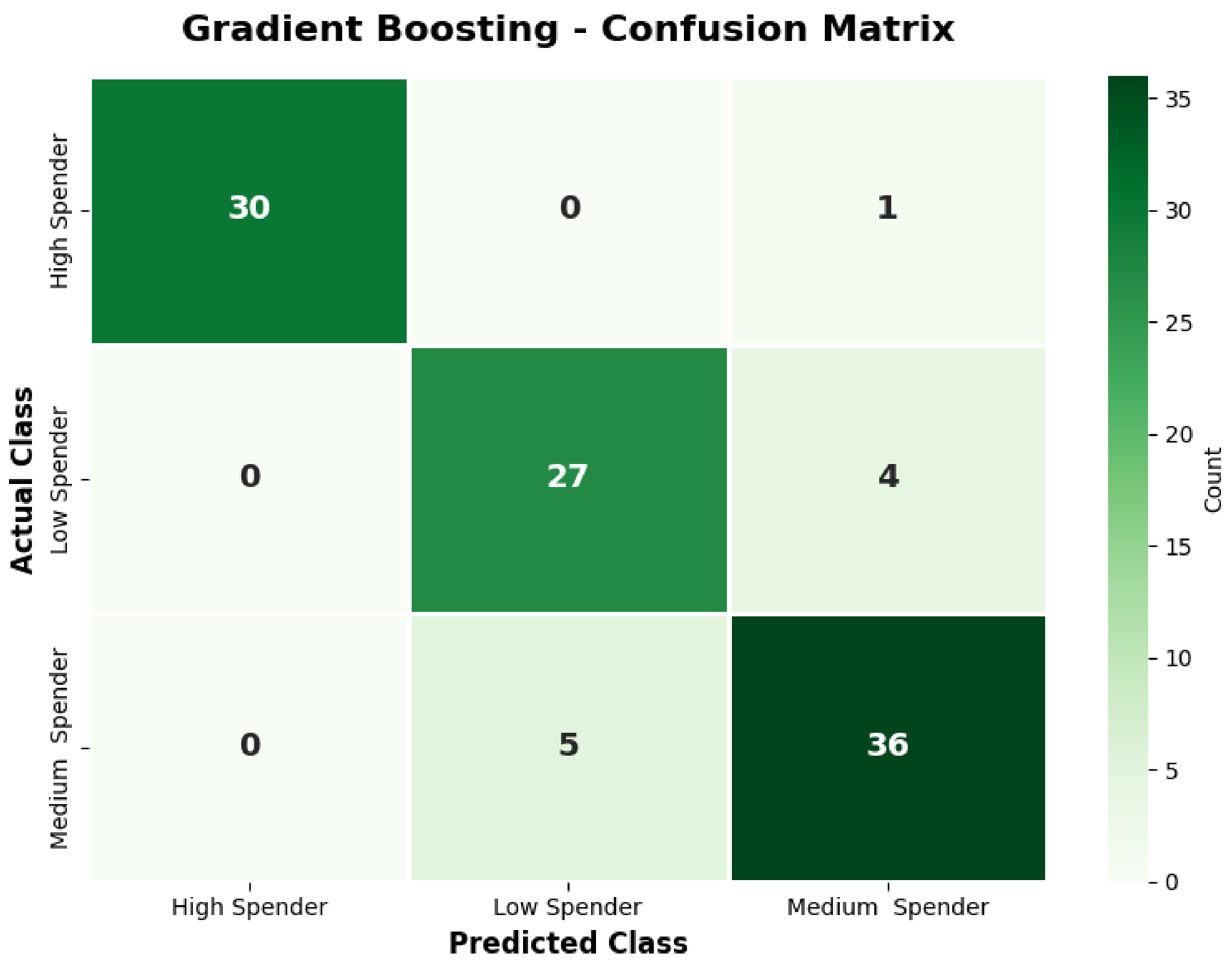


Fig 14: Gradient Boosting Confusion Matrix

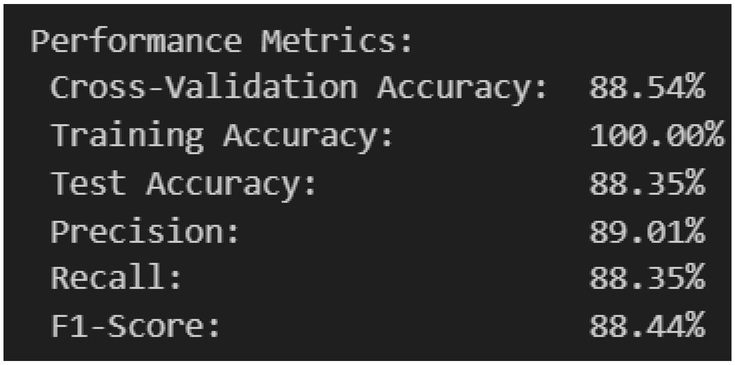
- **Insight:** While it achieved 100% on training data, its test accuracy dropped slightly below Logistic Regression. It misclassified 5 Medium Spenders as Low Spenders, slightly more than the Logistic model.

4.3 Random Forest Classifier

General Description:
Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.

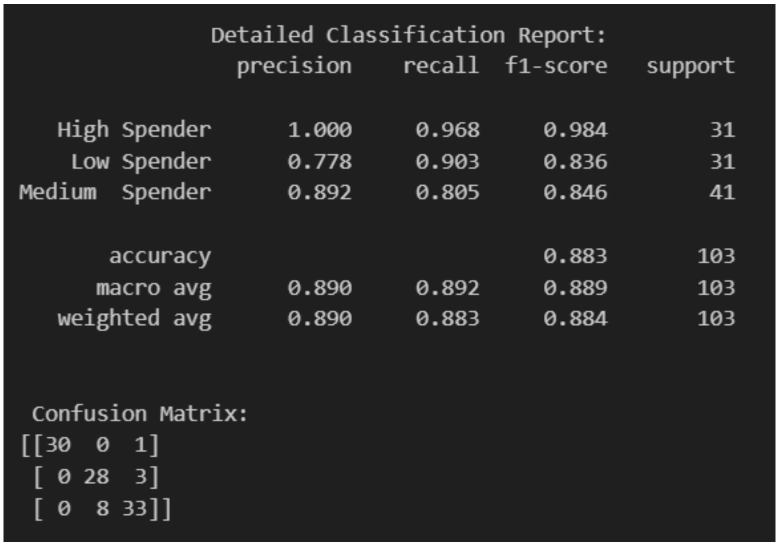
Specific Requirements:

- **Library:** sklearn.ensemble.RandomForestClassifier
- **n_estimators:** 300
- **Criterion:** Gini Impurity



```
Performance Metrics:
  Cross-Validation Accuracy:   88.54%
  Training Accuracy:           100.00%
  Test Accuracy:               88.35%
  Precision:                   89.01%
  Recall:                      88.35%
  F1-Score:                    88.44%
```

Fig 15: Random Forest Analysis Result



```
               Detailed Classification Report:
                precision    recall  f1-score   support

   High Spender      1.000     0.968     0.984        31
    Low Spender      0.778     0.903     0.836        31
 Medium  Spender     0.892     0.805     0.846        41

       accuracy                          0.883       103
      macro avg      0.890     0.892     0.889       103
   weighted avg      0.890     0.883     0.884       103


 Confusion Matrix:
[[30  0  1]
 [ 0 28  3]
 [ 0  8 33]]
```

Fig 16: Random Forest Classification Report

Confusion Matrix Interpretation :

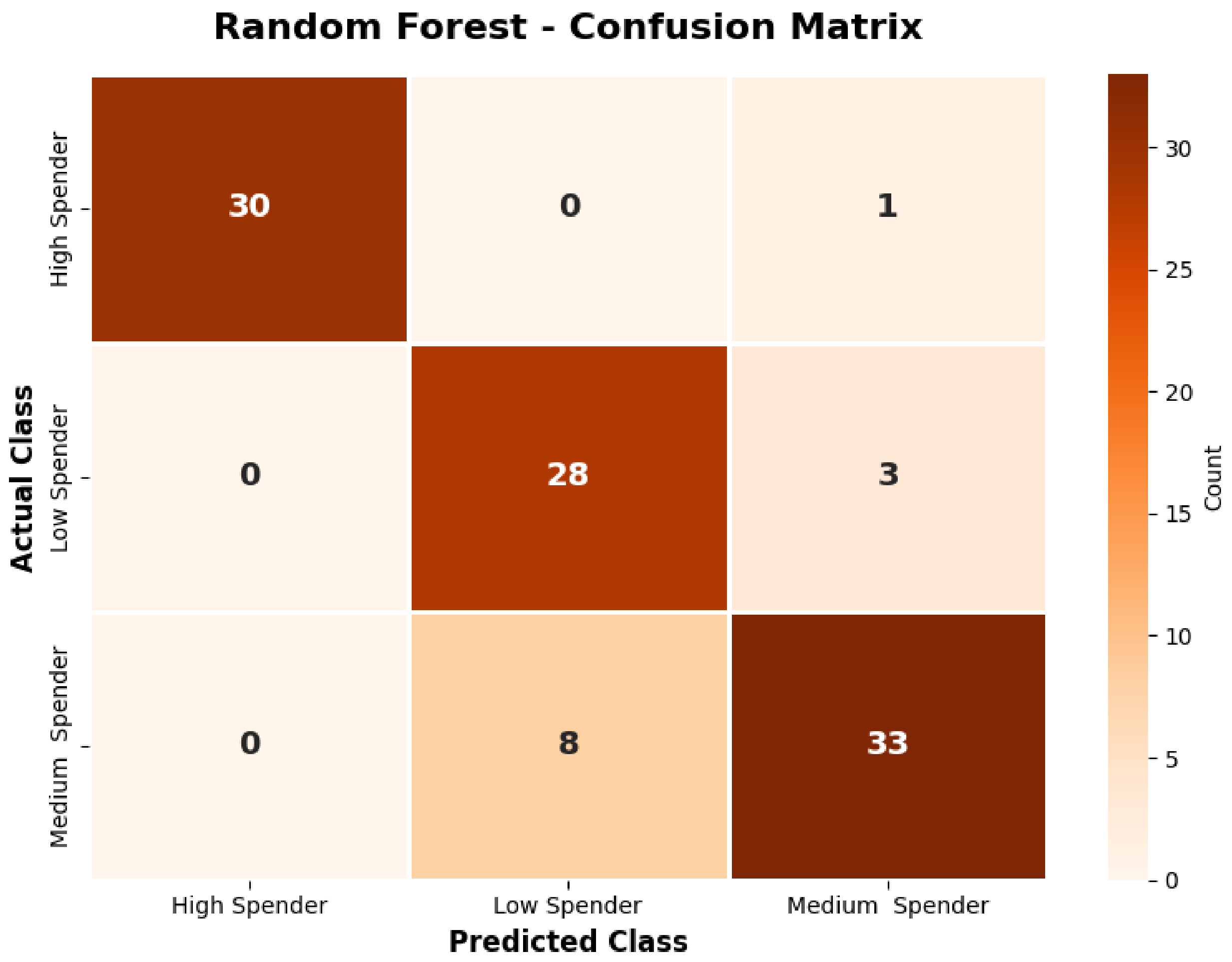


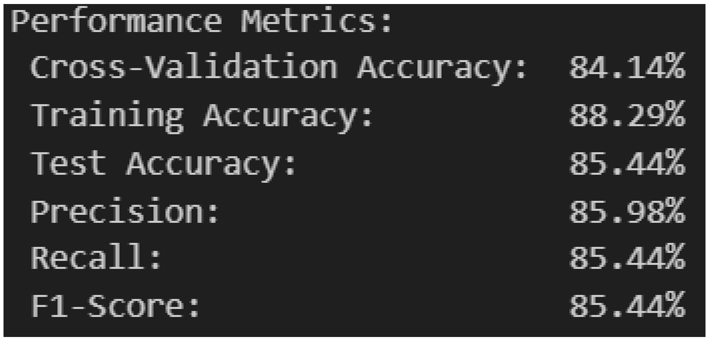
Fig 17: Random Forest Confusion Matrix

- **Insight:** Random Forest struggled the most with the "Medium Spender" class, misclassifying 8 of them as Low Spenders. However, it correctly identified 28 Low Spenders, one more than Logistic Regression.

4.4 K-Nearest Neighbors (KNN)

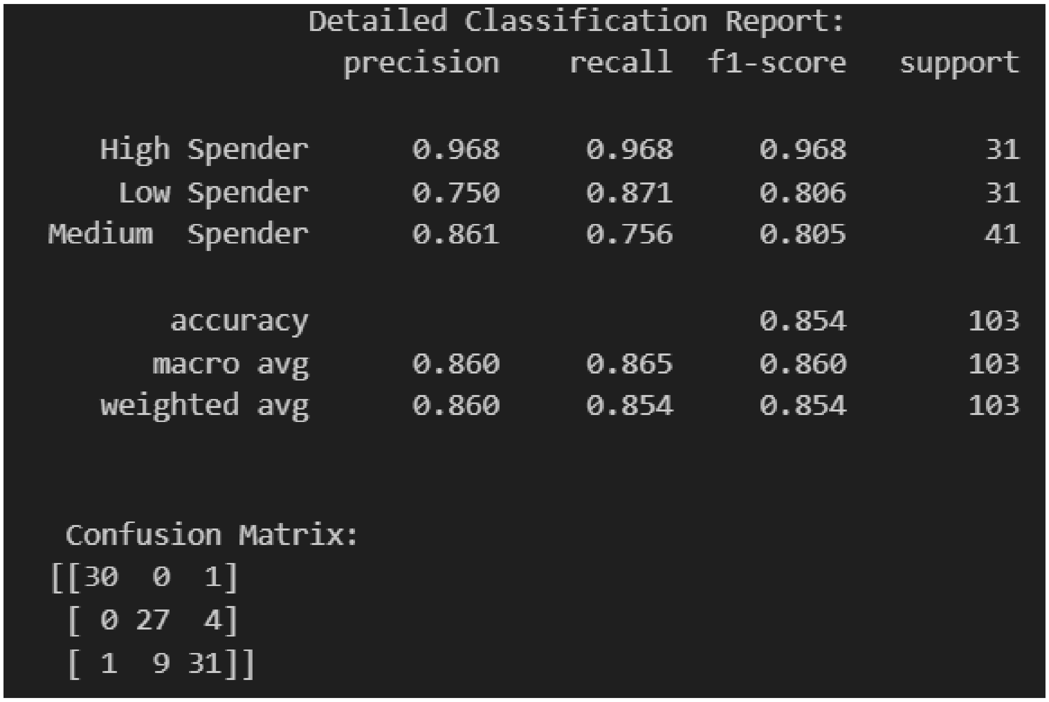
General Description:
KNN is a non-parametric method that classifies a sample based on the majority class of its 'k' nearest neighbors in the feature space.
Specific Requirements:

- **Library:** sklearn.neighbors.KNeighborsClassifier
- **n_neighbors:** 5
- **Metric:** Euclidean Distance



```
Performance Metrics:
 Cross-Validation Accuracy:  84.14%
 Training Accuracy:          88.29%
 Test Accuracy:              85.44%
 Precision:                  85.98%
 Recall:                     85.44%
 F1-Score:                   85.44%
```

Fig 18: KNN Analysis Result



```
                 Detailed Classification Report:
                  precision    recall  f1-score   support

    High Spender      0.968     0.968     0.968        31
     Low Spender      0.750     0.871     0.806        31
  Medium  Spender      0.861     0.756     0.805        41

        accuracy                          0.854       103
       macro avg      0.860     0.865     0.860       103
    weighted avg      0.860     0.854     0.854       103


 Confusion Matrix:
[[30  0  1]
 [ 0 27  4]
 [ 1  9 31]]
```

Fig 19: KNN Classification Report

Confusion Matrix Interpretation:

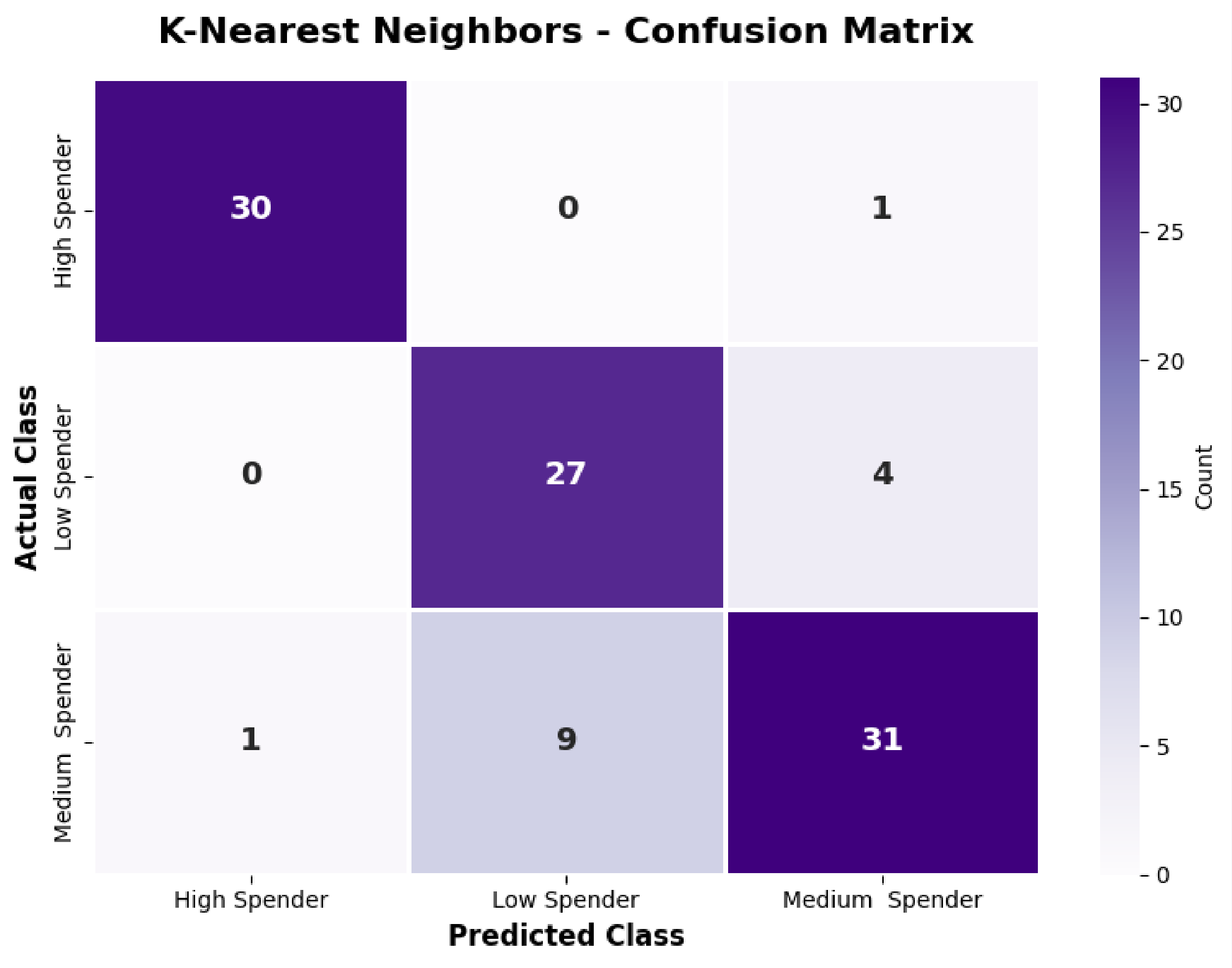


Fig 20: KNN Confusion Matrix

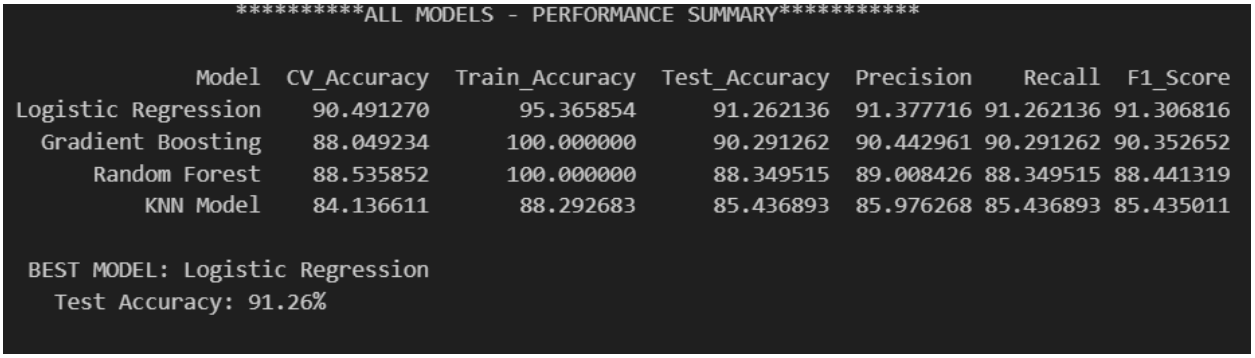
- **Insight:** KNN had the lowest accuracy. Notably, it was the only model to misclassify a Medium Spender as a High Spender (1 instance in bottom-left), indicating it struggles with the boundaries in high-dimensional space compared to linear or tree-based models.

5. Conclusion

This project successfully implemented a predictive analytics pipeline to segment e-commerce customers. By leveraging a hybrid dataset of 513 records (collected via Google Forms and Selenium), we achieved high-accuracy classification.

Key Findings:

1. **Best Model: Logistic Regression** achieved the highest test accuracy of **91.26%**. This suggests that the decision boundaries between Low, Medium, and High spenders in this specific dataset are largely linear and well-defined by features like Income and Spending_Ratio.



```
                 **********ALL MODELS - PERFORMANCE SUMMARY***********

                Model  CV_Accuracy  Train_Accuracy  Test_Accuracy  Precision    Recall  F1_Score
Logistic Regression    90.491270       95.365854      91.262136  91.377716 91.262136 91.306816
  Gradient Boosting    88.049234      100.000000      90.291262  90.442961 90.291262 90.352652
      Random Forest    88.535852      100.000000      88.349515  89.008426 88.349515 88.441319
          KNN Model    84.136611       88.292683      85.436893  85.976268 85.436893 85.435011

 BEST MODEL: Logistic Regression
   Test Accuracy: 91.26%
```

Fig 21: All models Performance summary

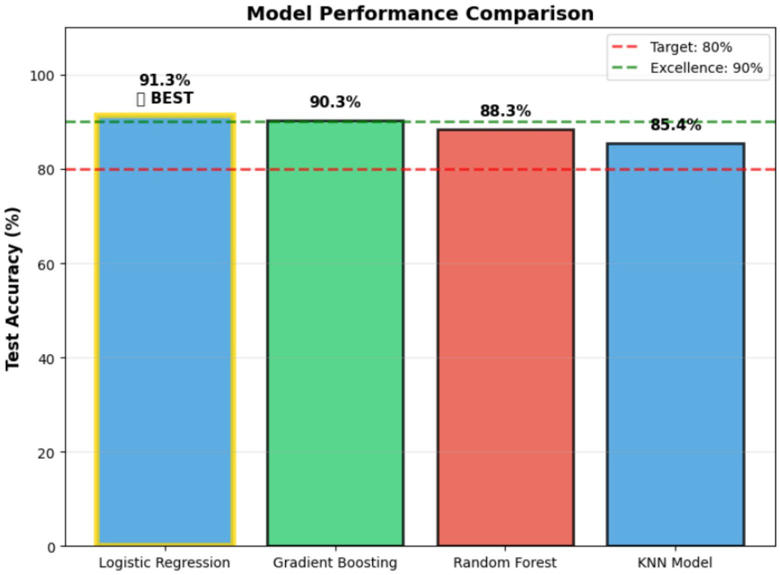


Fig 22: Model Performance Comparison

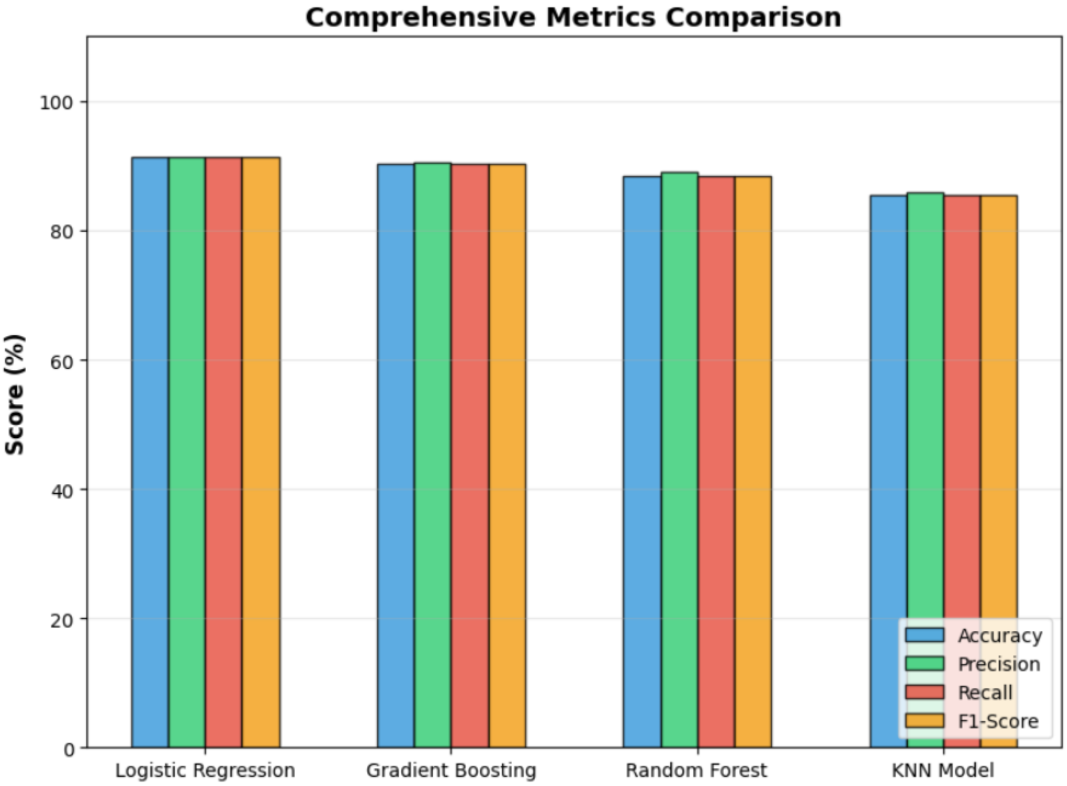


Fig 23: Metrices Comparison of all models

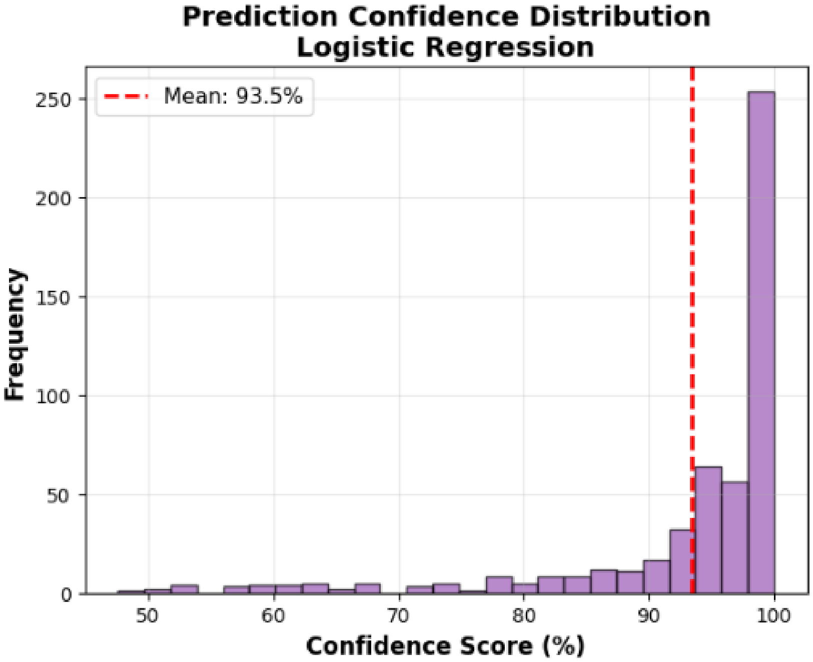


Fig 24: Logistic Regression Confidence Distribution

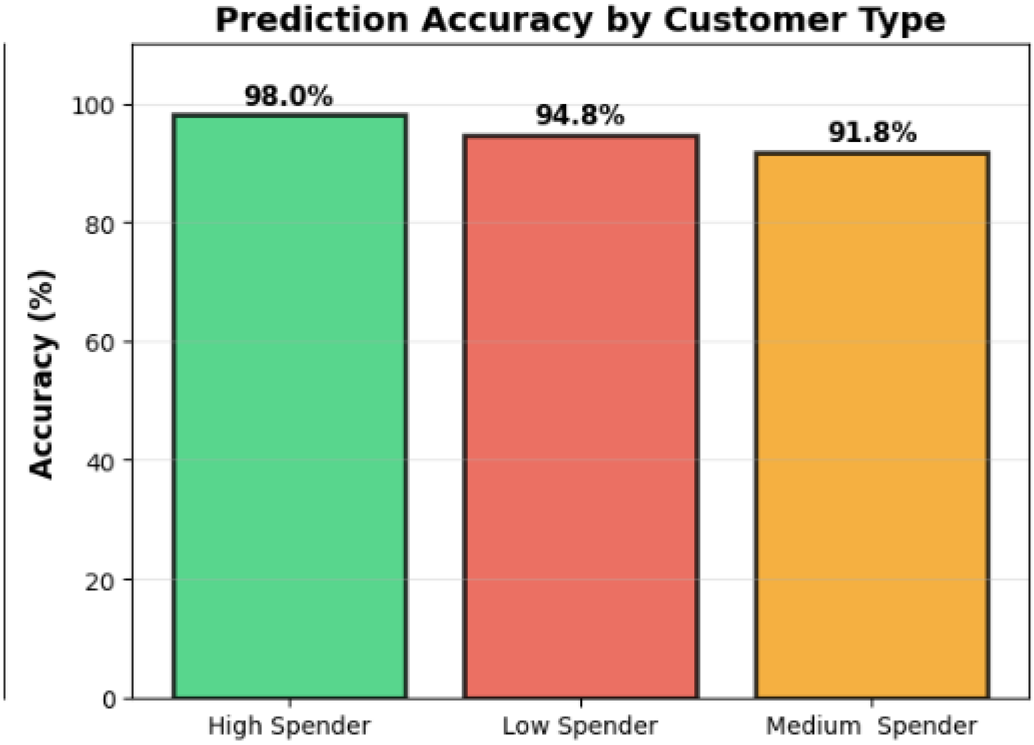


Fig 25: Prediction Accuracy by Customer Type

2. **Data Quality:** The rigorous preprocessing (currency cleaning, standardizing categorical variables) and feature engineering (creating interactions like Income_Frequency_Interaction) were crucial. The feature importance analysis showed that spending behavior is a better predictor than simple demographics.
3. **Class Distinction:** All models performed exceptionally well on the **High Spender** class (nearly 97-100% precision), likely because high-income/high-spending behaviors are very distinct. The main challenge for all models lay in distinguishing the "borderline" cases between Low and Medium spenders.

The project demonstrates that with clean data and intelligent feature engineering, simple models like Logistic Regression can outperform complex ensembles like Gradient Boosting in specific retail contexts.

6. Future Scope

This project lays the groundwork for a sophisticated **Real-Time Recommendation Engine**. The current model is static, but the future scope involves deploying this into a live e-commerce environment.

6.1 Website Integration (React + Flask)

The ultimate goal is to integrate the trained Logistic Regression model into a web application to personalize the user experience in real-time.

- **Frontend (React.js):** The website will track user actions (clicks, time spent on page, cart additions) in real-time.
- **Backend (Flask):** The Python backend will host the pre-trained model (saved as model.pkl).
- **Workflow:**
    1. A user browses the website. React captures their Browsing Time and Category_Count.
    2. This data is sent via API to the Flask backend.
    3. The model predicts the Customer Type (e.g., "High Spender") instantly.
    4. **Dynamic Customization:**
        - If **High Spender**: The website dynamically changes the homepage to show premium, high-margin products and offers "VIP Priority Shipping."
        - If **Low Spender**: The website automatically highlights "Deal of the Day," coupons, and discounts to encourage conversion.

This integration transforms the project from a theoretical analysis into a revenue-generating business tool.

7. References

1. J. Joseph et al., "Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms," *International Journal of Engineering Trends and Technology*, vol. 73, no. 2, pp. 92-106, 2025.
2. S. Agarwal, "Harnessing Machine Learning for Effective Customer Segmentation," *2024 International Conference on Signal Processing and Advance Research in Computing (SPARC)*, 2024.
3. M. Chavva, "Machine Learning Models for Enhancing Customer Segmentation and Targeting in Business Intelligence," *IEEE Conference Publication*, 2024.
4. Scikit-learn Documentation, "Supervised learning - Logistic Regression," available at scikit-learn.org.
5. Selenium Documentation, "Web Driver Automation," available at selenium.dev.

Google Drive link: https://drive.google.com/drive/folders/1b25leisWRufE2krxp5L-MnXIQ0ReWLhq?usp=sharing

Linkedin Link: https://www.linkedin.com/feed/update/urn:li:activity:7404395091416768512/

Github Link: https://github.com/mansityagi01/Predictive-Project