H-STORE ANALYSIS

Which customer are likely to return?

Prepared By: Senit Ghebregziabher (20050)
BAN 501: Quantitative Methods for Business
Prof. Angie Lu

H Store Case Study

Company and Data

Company: H Store

Dataset Properties

```
[5 rows x 47 columns]
```

```
Dataset shape:
(20000, 47)
```

Objectives

1. **Customer return prediction**
2. **Insight Generation**
3. **Interpretation**

Business Insights

Identify high-value customers for targeted campaigns.

Refine marketing efforts based on traffic sources and engagement behavior.

Prioritize regions with high revenue potential.

My strategy

Data Processing

Data Understanding and Preparation:

Addressed missing values:

- Numerical columns: Imputed with median.
- Categorical columns: Used placeholders like "Missing" or mode

Feature Engineering

1. **General Features**
2. **Revenue Engineered Features**
3. **Visualization**

Modeling Approach

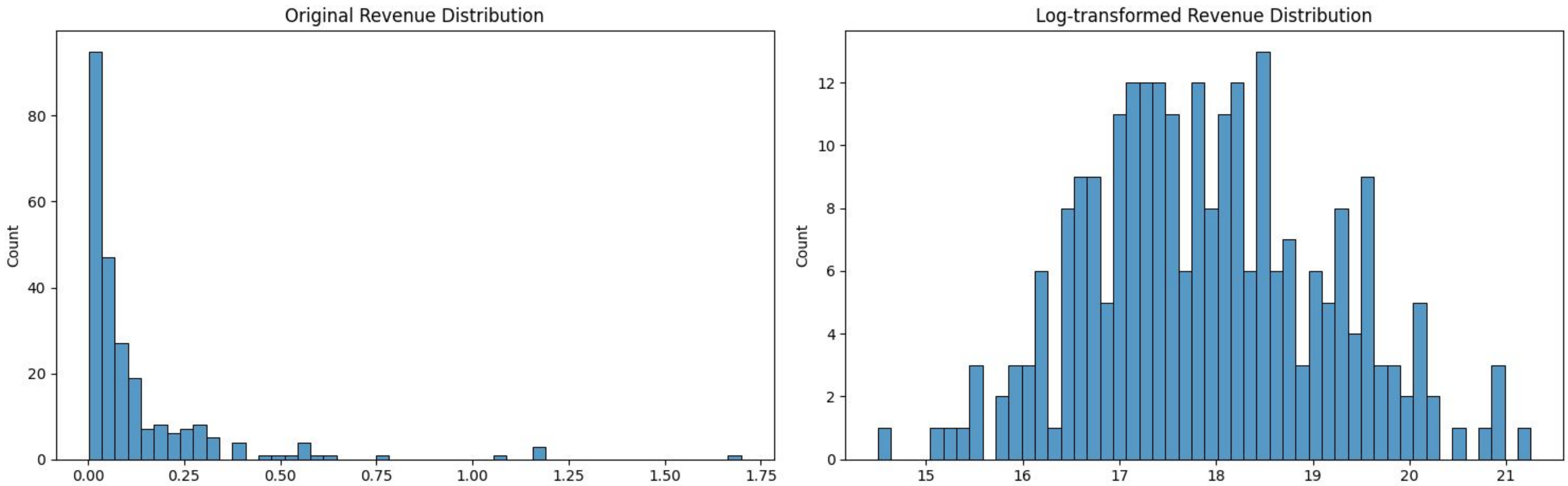
Classification: Predicted whether a customer would generate revenue.

Regression: Estimated revenue amount for returning customers using log-transformed revenue.

- Models used:

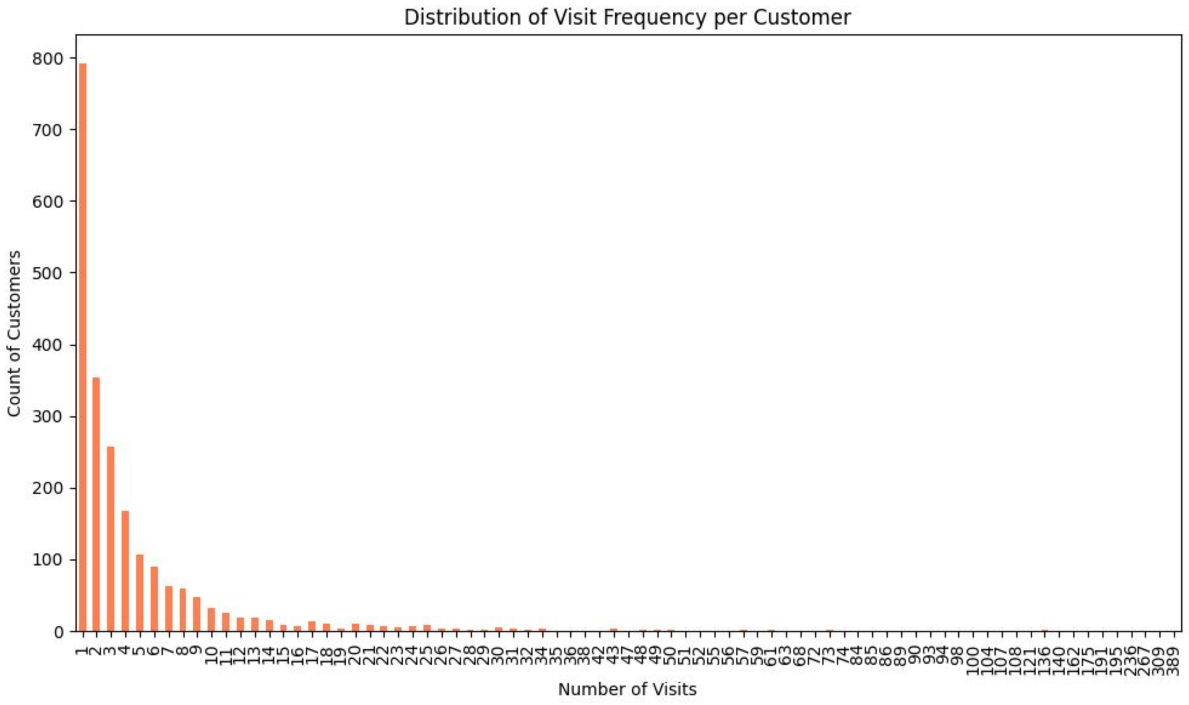
HistGradientBoostingClassifier and **Random Forest**

1.Data Preprocessing: Log Revenue



1. Better handling of skewed revenue distributions
2. More interpretable visualizations
3. More stable statistical relationships
4. Better feature engineering based on revenue patterns

Features: General



1.Visit Frequency Analysis:

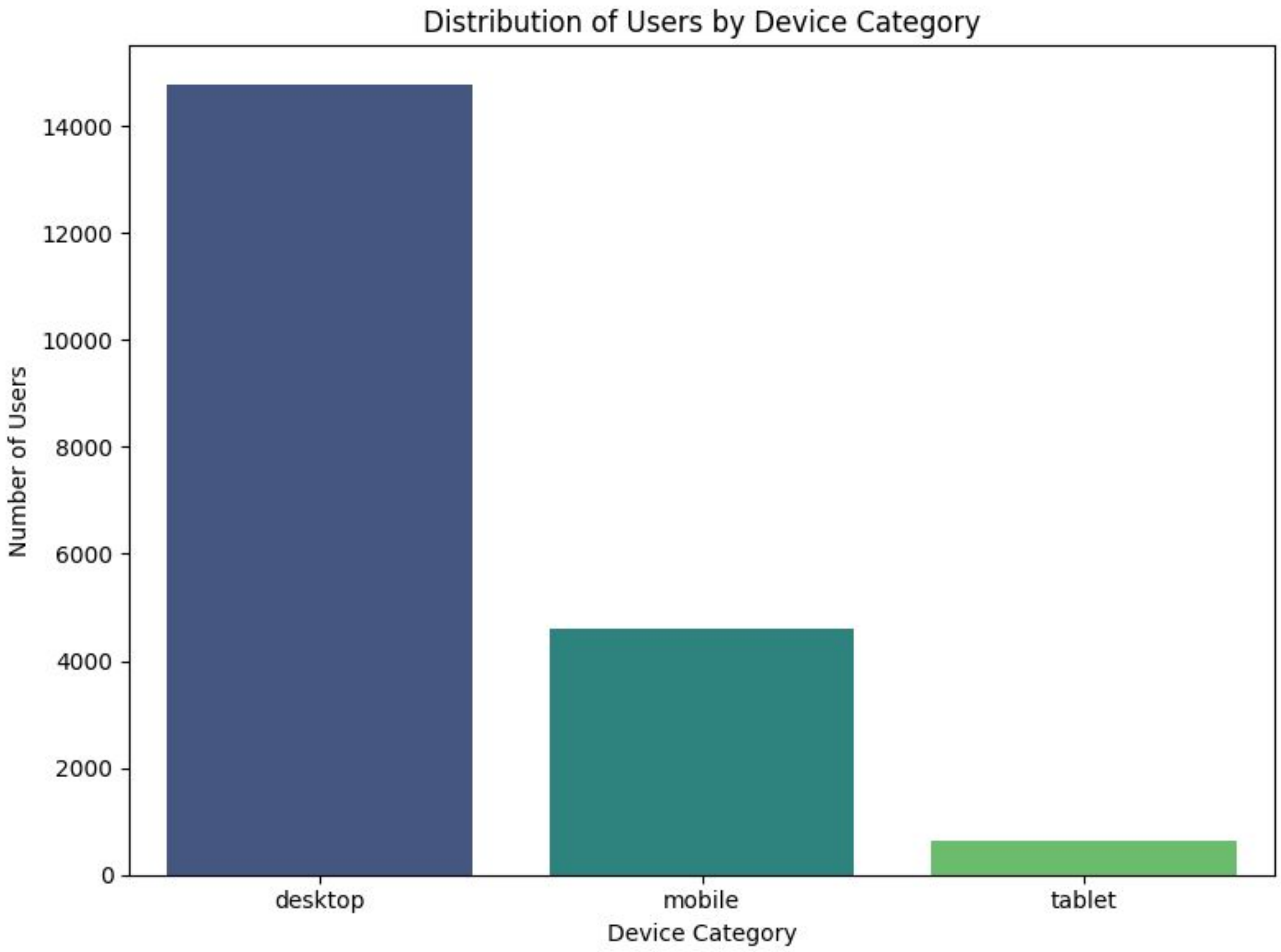
Displays customer loyalty patterns

Insights:

- Skewed Distribution
- High Customer Drop-off
- Long Tail of Frequent Visitors

Recommendations

- Enhance Onboarding Experience,
- Monitor Churn Causes
- personalized follow-up for Low Frequency Visitors



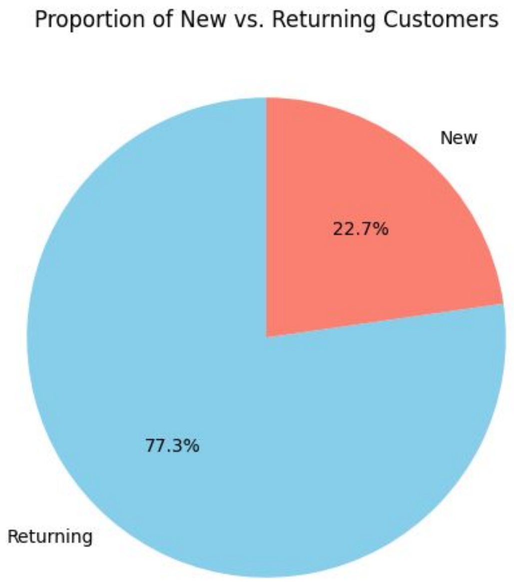
2. Users Device Category

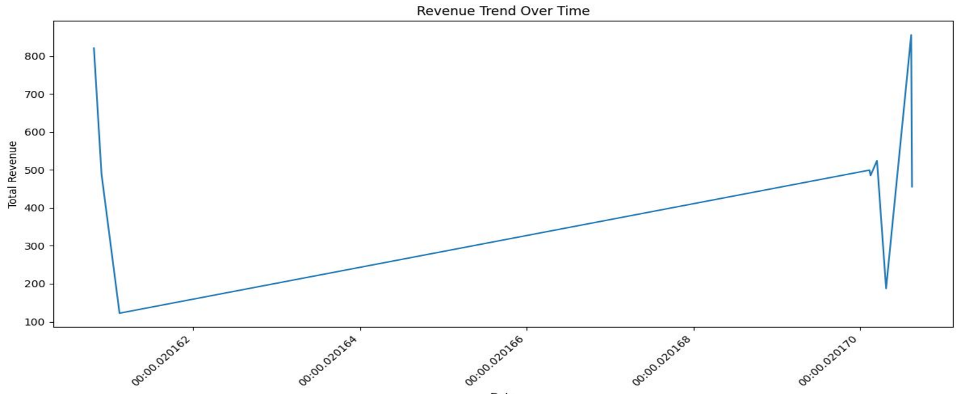
Insights:

- Dominance of Desktop Users
- Minimal Tablet Traffic

Recommendations

- Optimize for Desktop Experience
- Improve mobile optimization (responsive design, faster loading)
- Device-specific marketing strategies.
- A/B Testing Across Devices



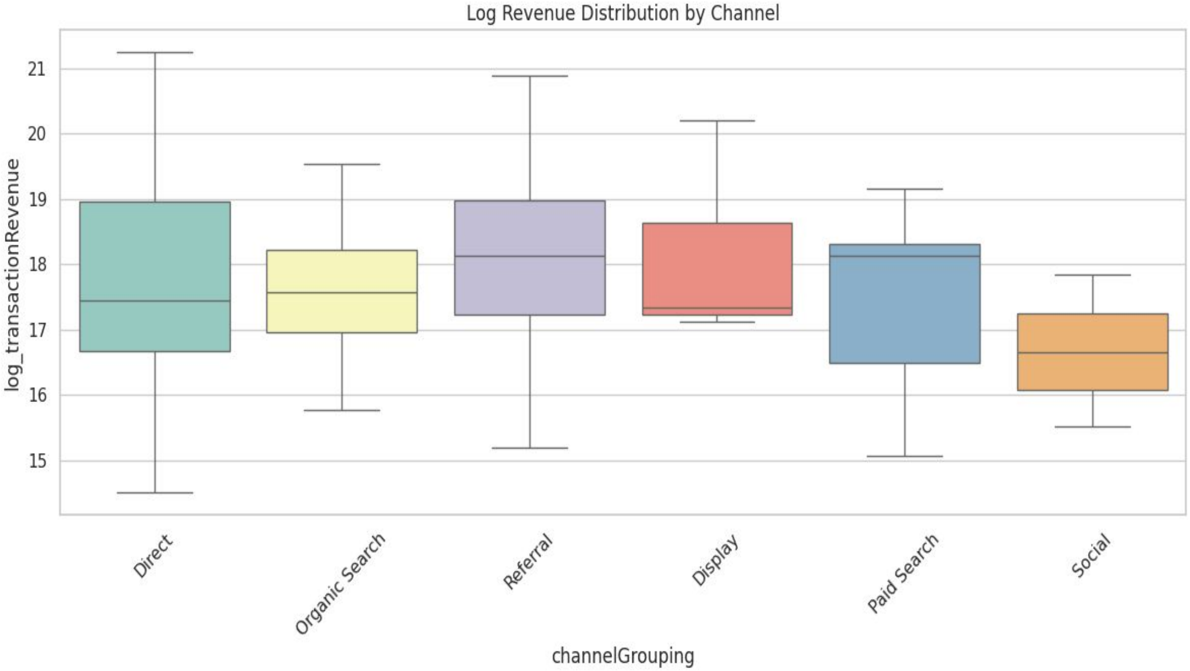


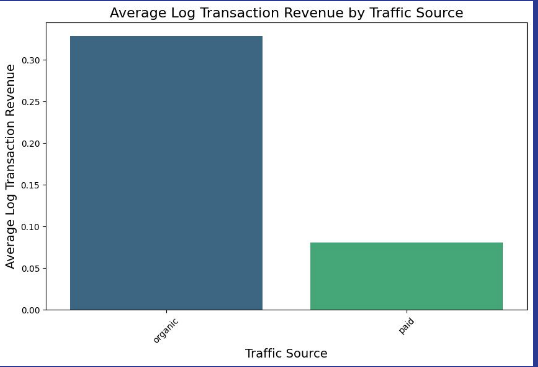
Returning vs. Non-Returning Customers

- Customer Retention: Is the proportion of returning customers high? This indicates customer loyalty.

Revenue trend over time

2. Features: Revenue Indexed

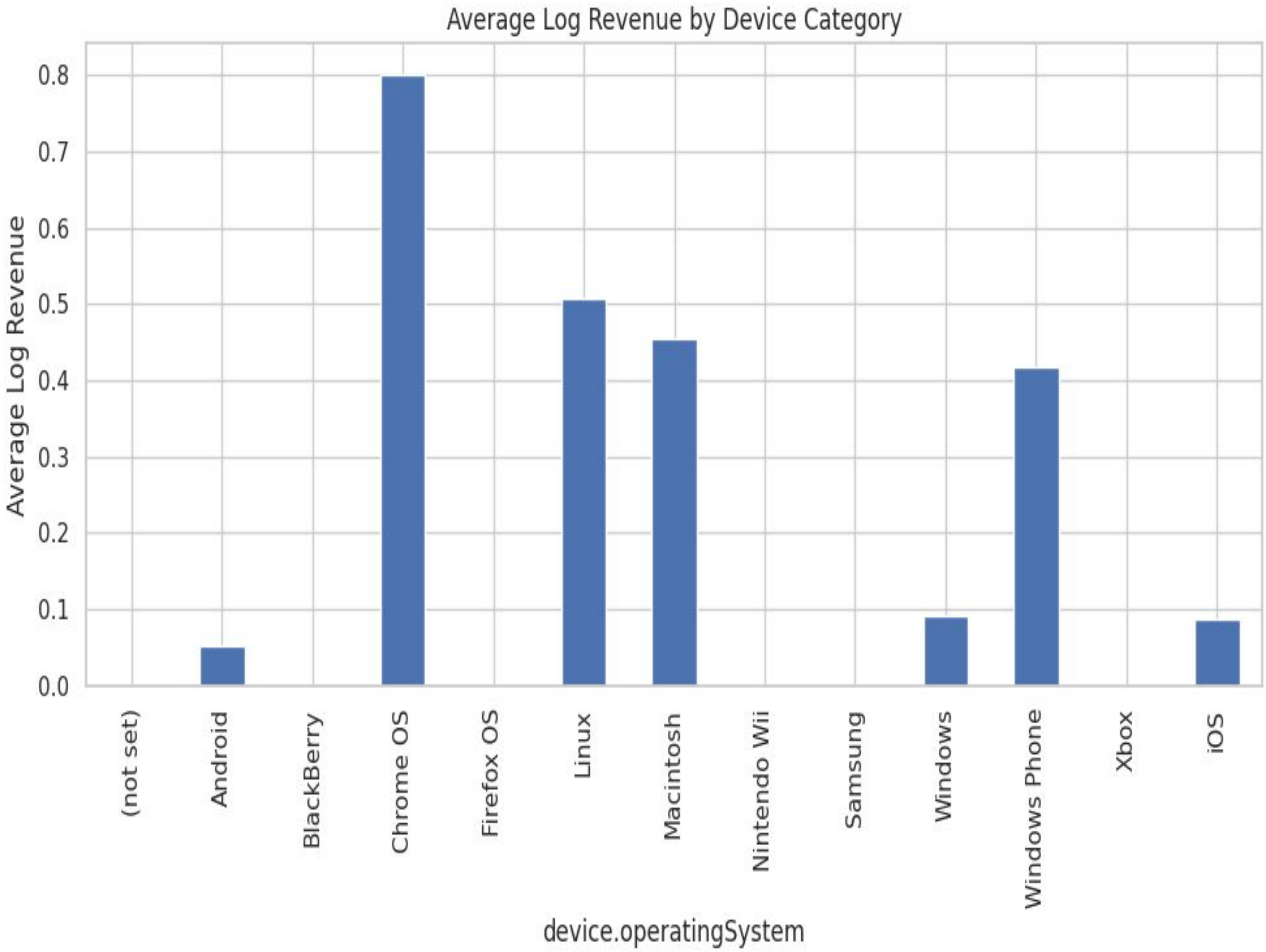




A. Traffic Source: how different marketing channels contribute to revenue

Recommendations

- Focus on Direct and Paid Search
- Explore optimization strategies for Social and Organic Search



B. Revenue by Device OS:

HIghest Revenue by

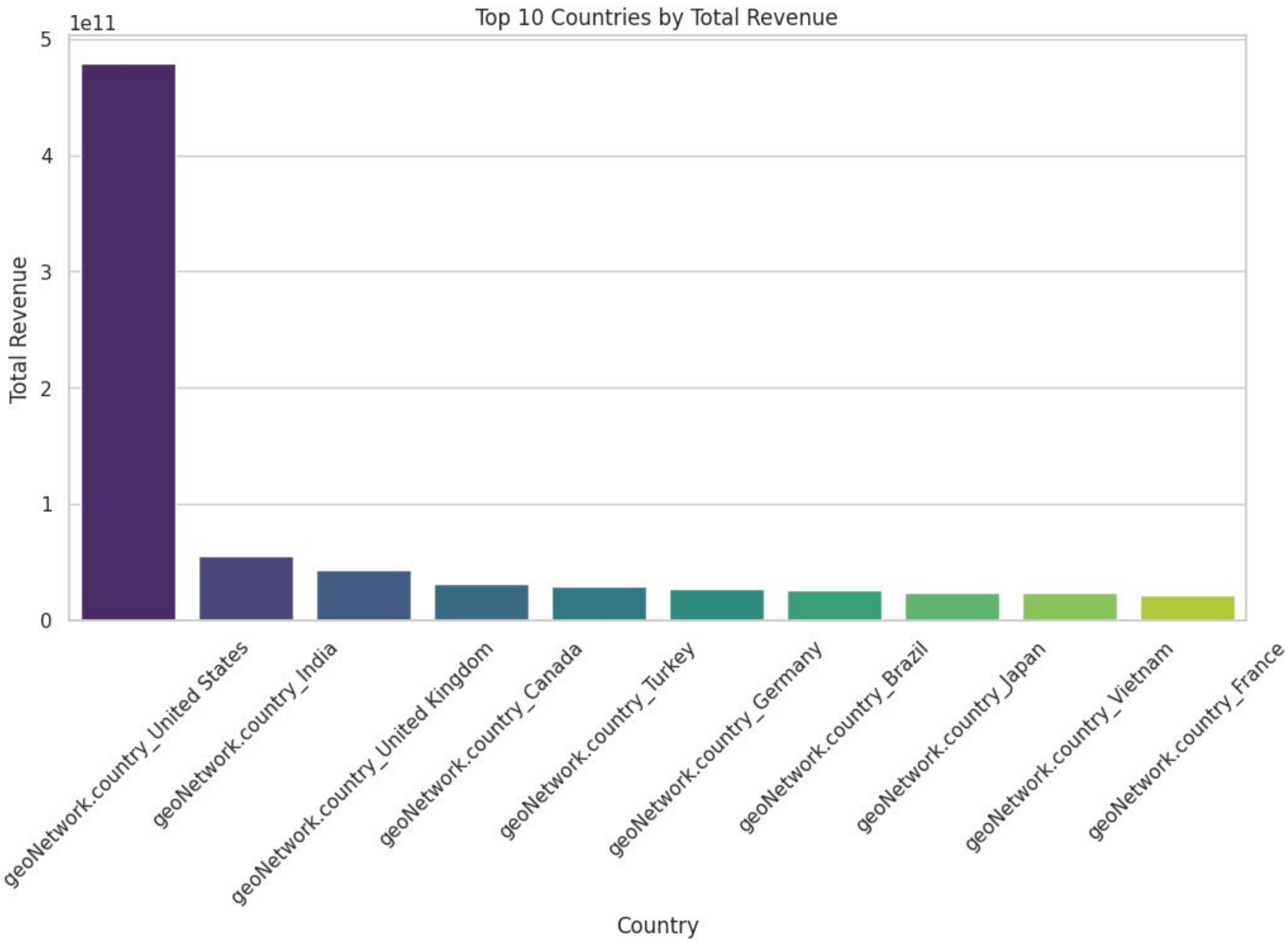
Chrome OS

Reccs:

- Focus marketing efforts on users of Chrome OS, Macintosh, and Linux

- Identify potential barriers (e.g., UI issues, limited functionality) and improve user experience

-Tailor promotions specifically for Chrome OS and Mac users



C. Revenue by Country:

HIghest Revenue by

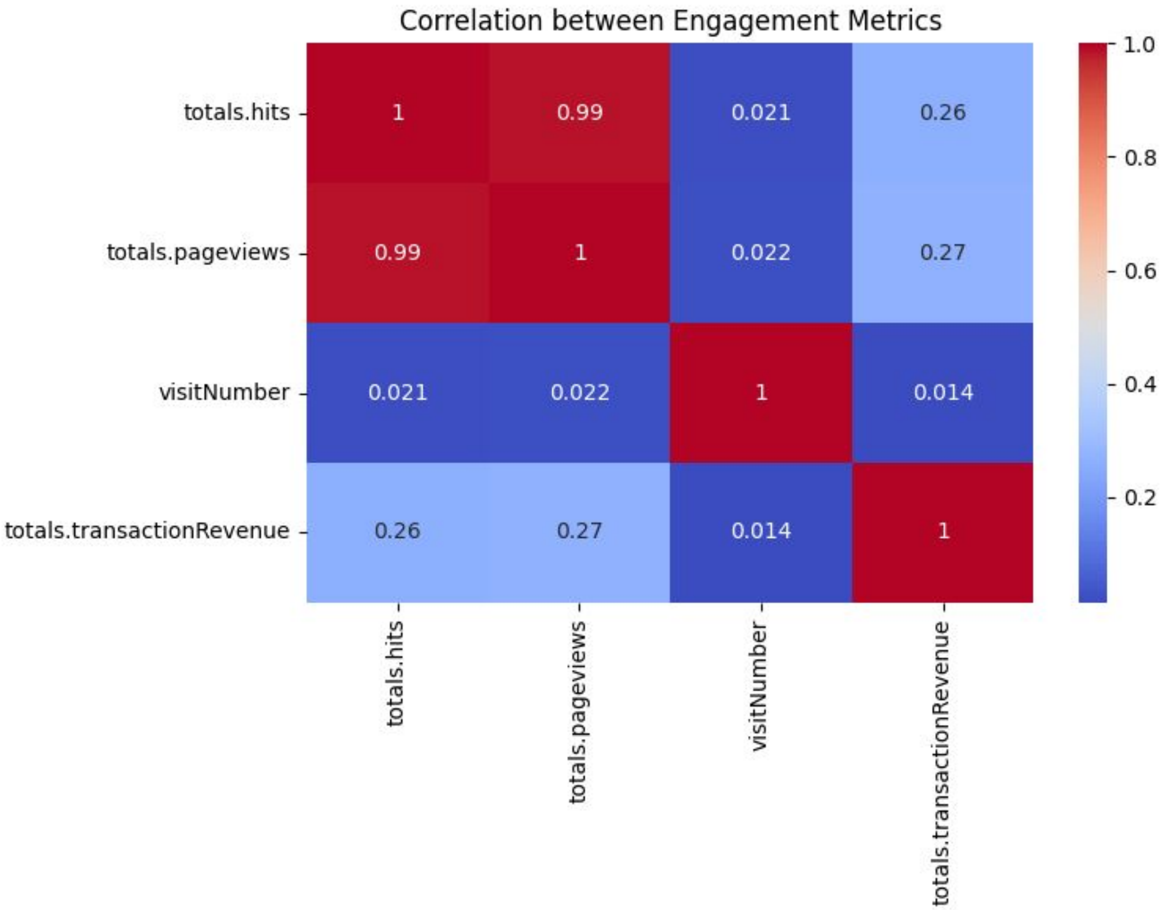
United States

Reccs:

-Focus on the U.S. Market: Maximize efforts in the highest revenue-generating region.

- Target India, UK, and Canada for growth opportunities.

-Localize strategies for Turkey and Germany..



D. Engagement Metrics

- Total Hits
- Total PageViews
- Visit Number
- Total Transaction Revenue

How customer engagement correlates with revenue?

Key Insights on Website Engagement and Revenue

- **Correlation** between *totals.transactionRevenue* and both t*otals.hits (0.26) and totals.pageviews* (0.27), indicates engagement (hits and pageviews) positively influences revenue
- **Moderate Revenue Impact:** emphasizing the need for engaging content and user experience.
- **Hits & Pageviews**: strong alignment. (**0.99**), users with more hits tend to view more pages
- **Visit Frequency:** Low Correlation (0.02), limited impact on revenue

Optimization Potential: Improving engagement strategies can unlock revenue growth opportunities. Focus on content. Use A/B Testing.

Regression Model

Model Training:

both **Random Forest and Gradient Boosting:** chosen due to their robustness

- Uses cross-validation for robust evaluation
- Calculates various performance metrics

Random Forest Results:

Classification Report:

	precision	recall	f1-score	support
0	0.12	0.02	0.04	436
1	0.89	0.98	0.93	3564
accuracy			0.88	4000
macro avg	0.51	0.50	0.48	4000
weighted avg	0.81	0.88	0.84	4000

Mean CV Score: 0.6151245802533782
CV Score Std: 0.014902796199868175

ROC AUC Score: 0.6053472415284342

Top 5 Important Features:

	feature	importance
5	geoNetwork.country	0.297051
6	totals.hits	0.146348
7	totals.pageviews	0.123949
0	channelGrouping	0.115672
3	device.browser	0.089911

Random Forest: Key Insights

Accuracy: 88%, driven by strong performance on Class 1 (returning customers).

Recall: High for Class 1 (98%), but extremely low for Class 0 (2%).

ROC AUC: 0.605, indicating limited ability to differentiate between classes.

Feature Importance:

- **Top Feature:** geoNetwork.country (29.7%) highlights geographic variations.
- Other Key Features: Engagement metrics like totals.hits (14.6%) and totals.pageviews (12.4%).

Limitations

Poor detection of non-returning customers (Class 0) due to class imbalance

Gradient Boosting Results:

Classification Report:

	precision	recall	f1-score	support
0	0.33	0.00	0.00	436
1	0.89	1.00	0.94	3564
accuracy			0.89	4000
macro avg	0.61	0.50	0.47	4000
weighted avg	0.83	0.89	0.84	4000

Mean CV Score: 0.6220771659521557
CV Score Std: 0.016420224124088983

ROC AUC Score: 0.6387711209958916

Top 5 Important Features:

	feature	importance
10	visitNumber	0.668849
5	geoNetwork.country	0.087987
6	totals.hits	0.052426
7	totals.pageviews	0.051914
3	device.browser	0.046852

Random Forest: Key Insights

Accuracy: 89%, slightly better than Random Forest.

Recall: Perfect for Class 1 (100%) but fails entirely on Class 0 (0%).

ROC AUC: 0.639, moderately better than Random Forest.

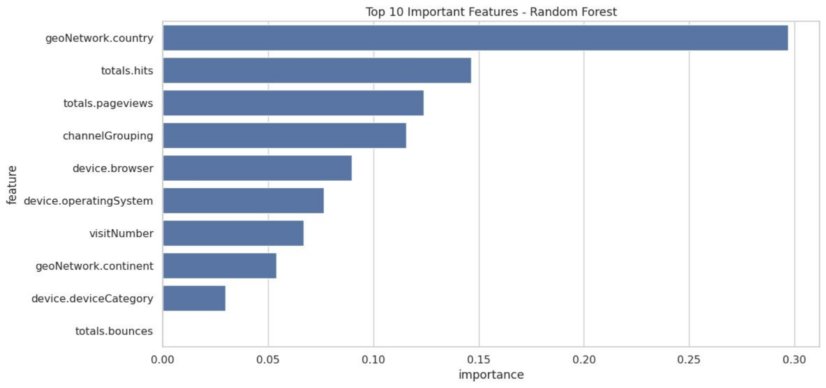
Feature Importance:

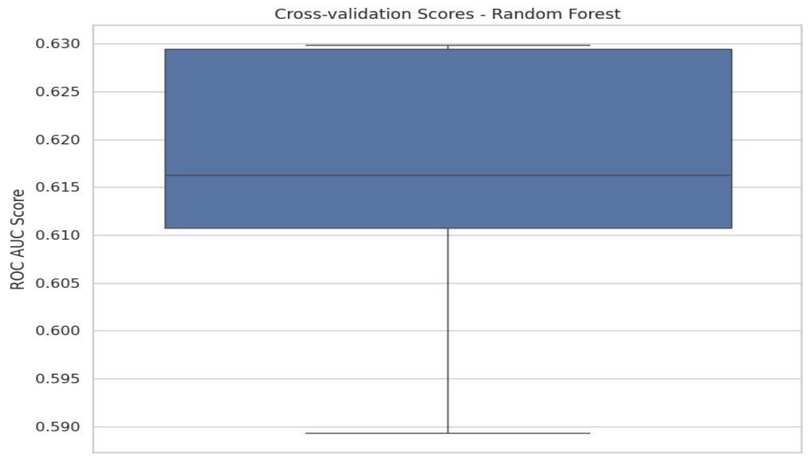
- Top Feature: visitNumber (66.8%) indicates customer visit frequency as the strongest predictor.
- Other Key Features: geoNetwork.country (8.8%), totals.hits, and totals.pageviews.

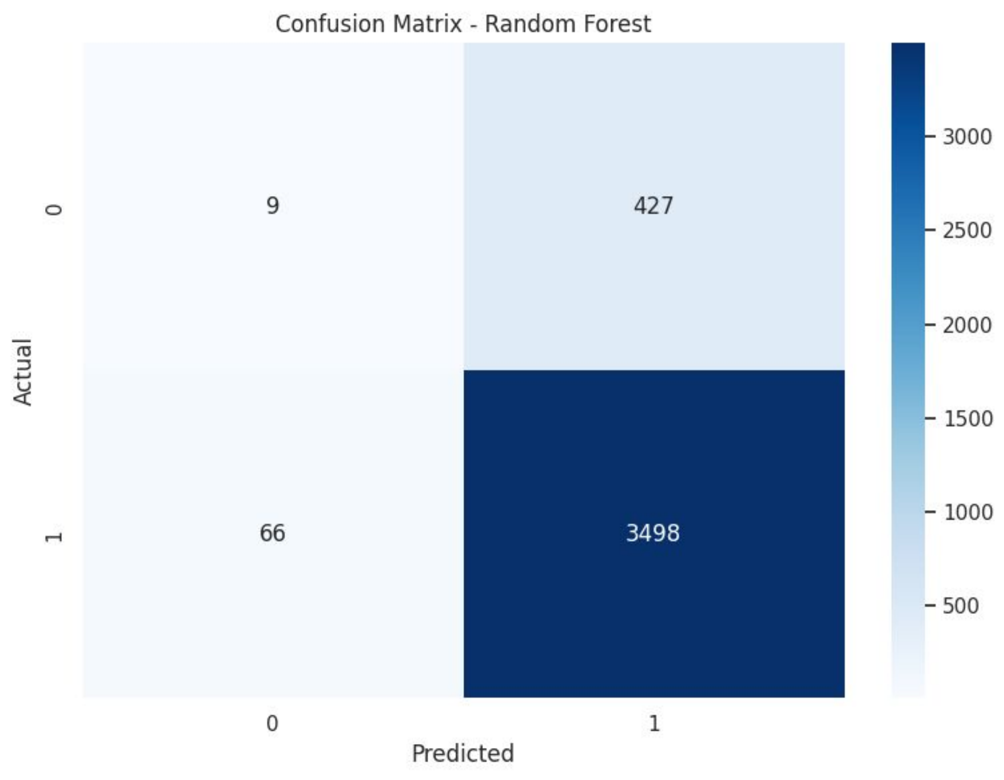
Limitations:

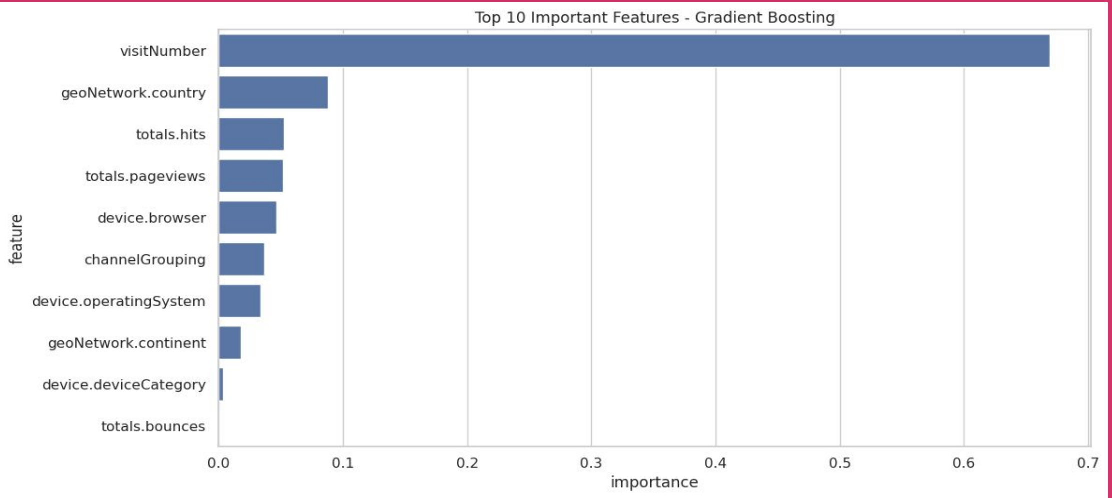
- Completely neglects Class 0, making it ineffective for predicting non-returning customers.
- Requires class-balancing strategies for better real-world application.

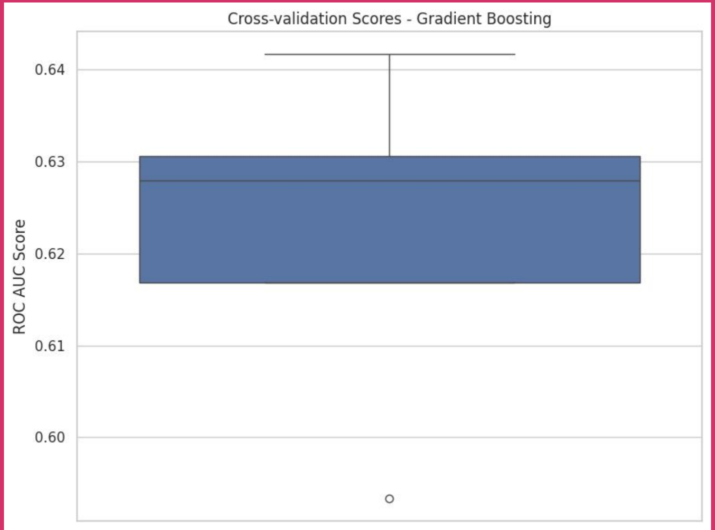
Random Forest Results

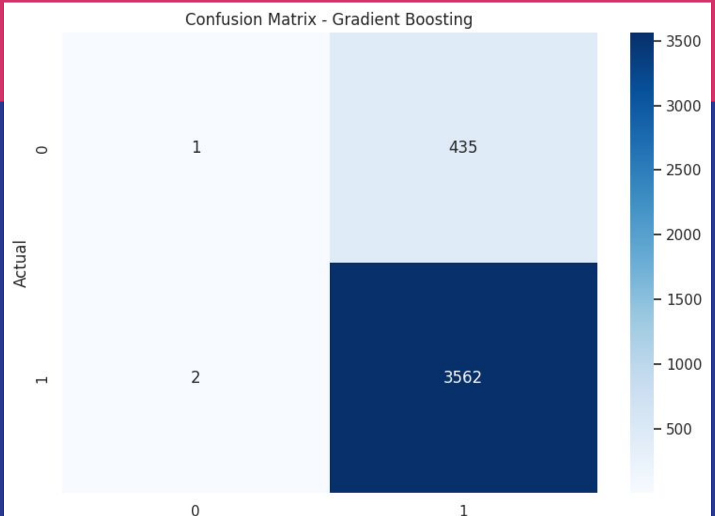












Gradient Boosting Results

Model Comparison

Random Forest vs. Gradient Boosting

Metric	Random Forest	Gradient Boosting
Accuracy	88%	89%
Recall (Class 0)	2%	0%
Recall (Class 1)	98%	100%
ROC AUC	0.605	0.639
Top Feature	geoNetwork.country	visitNumber

Class 0: Non Returning Customers

Class 1: Returning Customers

Weaknesses: Both models heavily favor Class 1, with poor performance on Class 0. Require resampling or cost-sensitive learning to address class imbalance.

Recommendation: Gradient Boosting is preferred for its slightly better general performance, but both models require improvement for practical customer churn prediction.

Final Thoughts

- Engagement metrics (hits and pageviews) are critical for driving revenue.
- A data-driven approach focusing on user quality over quantity will maximize profitability.
- Continuous testing and user behavior analysis are essential to refine the engagement-to-revenue pipeline

Thank you for your attention