Predicting User Churn for an E-commerce Platform

https://github.com/shoryasethia/e-commerce-churn

Shorya Sethia

India shorya-sethia shoryasethia shoryasethia4may@gmail.com

Dataset

Dataset: events.csv

Columns:

1. **event_time:** Datetime of the event (e.g., 2020-12-01 09:00:00)
2. **event_type:** Type of the event, one of view, cart, or purchase
3. **product_id:** Identifier of the product
4. **category_id:** Identifier of the product's category
5. **category_code:** Human-readable category code (if available)
6. **brand:** Brand of the product
7. **price:** Price of the product (numerical)
8. **user_id:** Identifier of the user who performed the event 9. user_session: Identifier of the user's session

Using events.csv I prepared following:

- **events_cleaned.csv:** Handeled anomalies, outliers, NaN, etc
- **events_features.csv:** Contains user based features
- **events_features_extended.csv:** Contains user based and several other misc features
- **user_churn_scores.csv:** Based on extracted features, calculated each user's churn score
- **user_churn_analysis.csv:** Extracted churn related features based on user's features
- **user_churn_risk_analysis.csv:** Annotated data of all users as per certain threshold and manipulations on churn scores and related features.
- **events_with_churn_score.csv:** Contains columns from events_features_extended.csv and user_churn_risk_analysis.csv

Exploratory Data Analysis (EDA)

Refer this notebook, which contains all insights regarding data, events_eda.ipynb

Feature Engineering

The dataset after adding features and cleaning comprises **13,353 rows** and **49 columns**, with each row representing a user. Explicit churn annotations were not provided; instead, churn signals were inferred from user behaviors captured in the event stream. The primary goal was to create user-level features that effectively capture churn signals for accurate prediction.

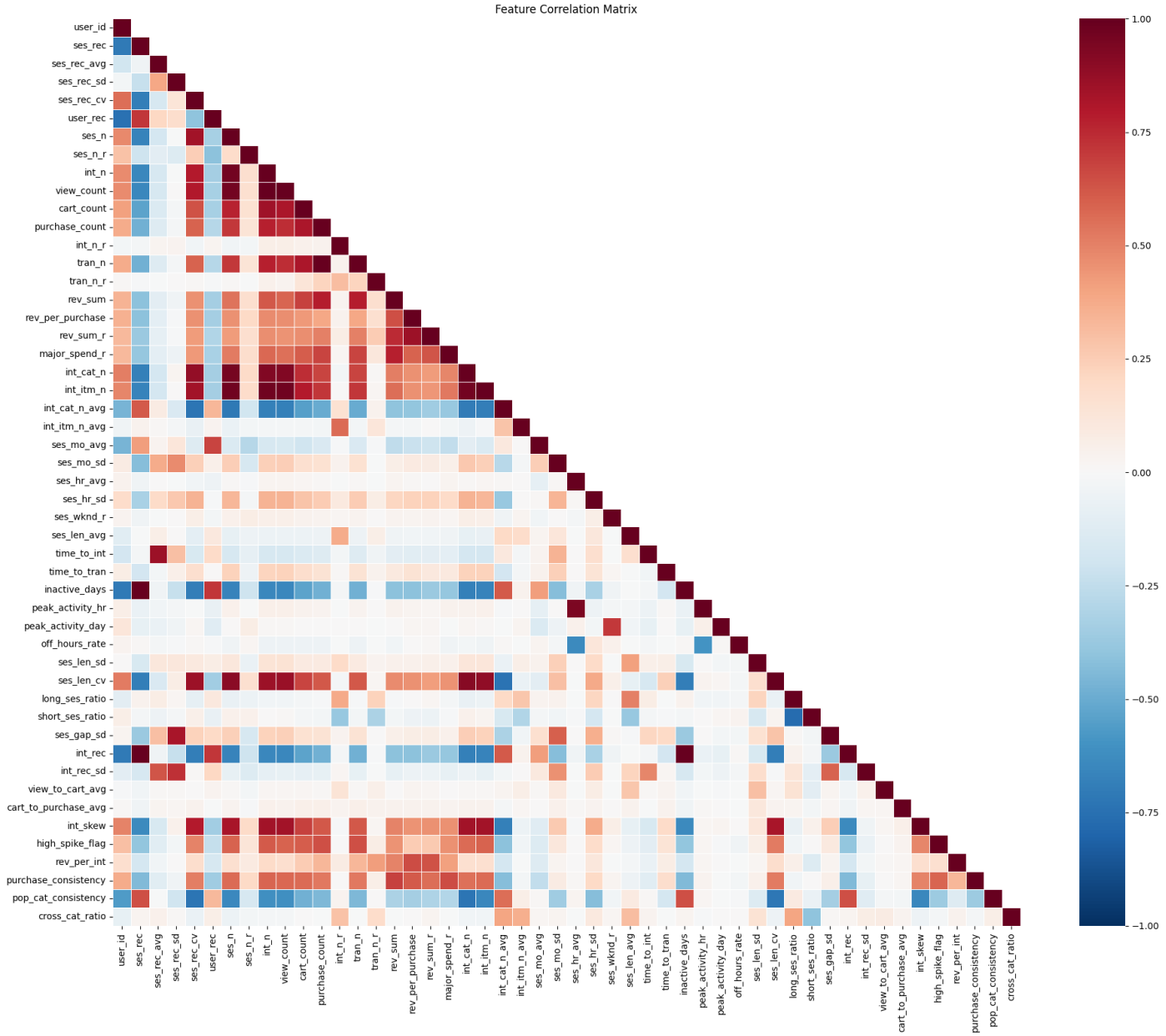
Feature Dictionary

Feature Name	Category	Abbreviated Name	Definition
Session Recency	Recency	ses_rec	Days between last user session and reference date
Average Session Recency	Recency	ses_rec_avg	Average days between consecutive sessions
Session Recency Standard Deviation	Recency	ses_rec_sd	Standard deviation of time between sessions (days)
Session Recency Coefficient of Variation	Recency	ses_rec_cv	Ratio of std dev to mean time between sessions (%)
User Maturity	Recency	user_rec	Days since user's first session
Session Count	Frequency	ses_n	Total number of sessions
Relative Session Frequency	Frequency	ses_n_r	Sessions per day since first activity
Interaction Count	Frequency	int_n	Total number of events (view/cart/purchase)
View Count	Frequency	view_count	Total number of view events
Cart Count	Frequency	cart_count	Total number of cart events
Purchase Count	Frequency	purchase_count	Total number of purchase events
Interaction Rate	Frequency	int_n_r	Average interactions per session
Transaction Count	Frequency	tran_n	Total number of transactions (purchases)
Transaction Rate	Frequency	tran_n_r	Average transactions per session
Total Revenue	Monetary	rev_sum	Total spending by user (USD)
Average Purchase Value	Monetary	rev_per_purchase	Average amount spent per purchase
Revenue per Session	Monetary	rev_sum_r	Average revenue per session
High Spender Flag	Monetary	major_spend_r	Whether user's spending is above average (1/0)

Feature Name	Category	Abbreviated Name	Definition
Category Interaction Count	Category & Item	int_cat_n	Number of unique categories interacted with
Product Interaction Count	Category & Item	int_itm_n	Number of unique products interacted with
Category Diversity	Category & Item	int_cat_n_avg	Average unique categories per session
Product Diversity	Category & Item	int_itm_n_avg	Average unique products per session
Average Month	Date & Time	ses_mo_avg	Average month of user sessions (1-12)
Month Standard Deviation	Date & Time	ses_mo_sd	Variation in session months
Average Hour	Date & Time	ses_hr_avg	Average hour of day for sessions (0-23)
Hour Standard Deviation	Date & Time	ses_hr_sd	Variation in session hours
Weekend Ratio	Date & Time	ses_wknd_r	Proportion of sessions on weekends
Average Session Length	Other	ses_len_avg	Average session duration (minutes)
Interaction Time Gap	Other	time_to_int	Average time between interactions (minutes)
Transaction Time Gap	Other	time_to_tran	Average time between purchases (days)
Inactivity Period	Engagement	inactive_days	Days since the user's last session relative to the reference date
Peak Activity Hour	Engagement	peak_activity_hr	Hour of day (0-23) with the highest number of user interactions
Peak Activity Day	Engagement	peak_activity_day	Day of the week (0-6, where 0=Monday) with the highest number of user interactions
Off-Hours Activity Rate	Engagement	off_hours_rate	Proportion of interactions occurring between 12 a.m. and 6 a.m.
Session Length Standard Deviation	Session Patterns	ses_len_sd	Standard deviation of session durations

Feature Name	Category	Abbreviated Name	Definition
Session Length Coefficient of Variation	Session Patterns	ses_len_cv	Ratio of std dev to mean session duration
Long Session Ratio	Session Patterns	long_ses_ratio	Proportion of sessions lasting more than 30 minutes
Short Session Ratio	Session Patterns	short_ses_ratio	Proportion of sessions lasting less than 5 minutes
Session Gaps Standard Deviation	Session Patterns	ses_gap_sd	Standard deviation of time gaps between sessions (days)
Interaction Recency	Interaction Behavior	int_rec	Days between the last interaction and the reference date
Interaction Recency SD	Interaction Behavior	int_rec_sd	Standard deviation of time gaps between interactions (minutes)
Average View-to-Cart Time	Interaction Behavior	view_to_cart_avg	Average time taken to move from view to cart (minutes)
Average Cart-to-Purchase Time	Interaction Behavior	cart_to_purchase_avg	Average time taken to move from cart to purchase (minutes)
Interaction Skewness	Interaction Behavior	int_skew	Skewness of interaction counts per session (measure of bias towards high/low activity)
High Spending Spike Flag	Revenue Patterns	high_spike_flag	1 if there is a session with spending >2x average spending, else 0
Revenue per Interaction	Revenue Patterns	rev_per_int	Total revenue divided by total number of interactions
Purchase Consistency	Revenue Patterns	purchase_consistency	Standard deviation of spending per transaction
Popular Category Consistency	Category & Item	pop_cat_consistency	Fraction of sessions involving the most frequently interacted category
Cross-Category Interaction Ratio	Category & Item	cross_cat_ratio	Ratio of sessions with interactions in more than one category
Churn Score	Other	churn_score	Weighted score combining recency, frequency, monetary, and other behavioral metrics
Risk Category	Other	risk_category	Categorization of user risk level (e.g., low, medium, high)

Feature Name	Category	Abbreviated Name	Definition
Likely to Churn Flag	Other	is_likely_churn	1 if user is predicted to churn, else 0
Risk Velocity	Other	risk_velocity	Rate at which the user's risk of churn is increasing over time
Primary Risk Drivers	Other	primary_risk_drivers	Main factors influencing the user's risk of churn
Estimated Days to Churn	Other	estimated_days_to_churn	Estimated number of days until the user is likely to churn



Total features extracted = 37

The engineered features comprehensively capture user engagement, monetary trends, and behavioral patterns. These metrics form the foundation for accurate churn prediction as depicted in predictive modelling and actionable business insights.

Broad Feature Categories

1. RFM Metrics

- Captures Recency, Frequency, and Monetary signals to identify user engagement and loyalty levels.
- Examples:
  - ses_rec, ses_n, rev_sum

2. Session-Based Metrics

- Includes insights into user session behavior, such as session count, average session duration, and time gaps.
- Examples:
  - ses_len_avg, ses_gap_sd, long_ses_ratio

3. Product/Brand Preferences

- Tracks category and product-level interactions, providing insights into user preferences.
- Examples:
  - int_cat_n, int_itm_n, pop_cat_consistency

4. Behavioral Patterns

- Focuses on user behavior during the funnel process, e.g., moving from view to cart and from cart to purchase.
- Examples:
  - view_to_cart_avg, cart_to_purchase_avg, cross_cat_ratio

5. Churn Indicators

- Designed to predict churn using composite metrics like the churn score and behavioral trends.
- Examples:
  - churn_score, is_likely_churn, risk_velocity

Churn Definition & Reasoning

Definition of Churn

In my e-commerce events data exploration and extrapolation, a user is considered to be churned when they meet the following criteria:

1. Primary Threshold: Their composite churn score exceeds **68.43** (high-risk threshold), calculated using a sophisticated multi-component scoring system that incorporates:

- RFM (Recency, Frequency, Monetary) metrics
- Engagement decay patterns
- Behavioral patterns
- Purchase patterns
- Temporal rhythm analysis

2. Supporting Indicators:

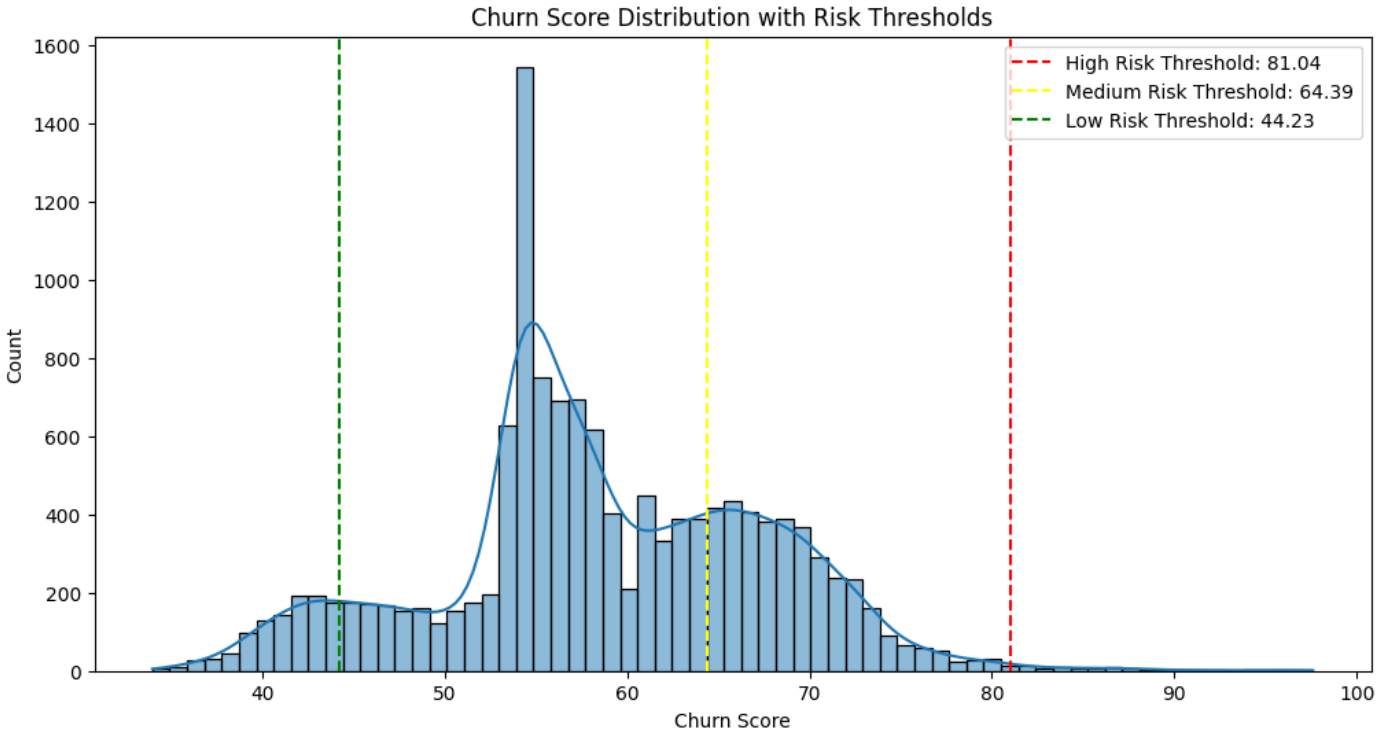
- Inactivity Period: Extended period of no activity relative to their historical engagement patterns
- Engagement Decay: Significant decline in interaction frequency and depth
- Purchase Pattern Disruption: Deviation from established purchasing rhythms
- Behavioral Anomalies: Significant changes in browsing and interaction patterns

	user_id	churn_score	confidence_level	rfm_component	engagement_component	behavioral_component	purchase_component	temporal_component
count	1.335300e+04	13353.000000	13353.0	13353.000000	1.335300e+04	13353.000000	1.335300e+04	1.335300e+04
mean	1.515916e+18	58.608127	100.0	0.251966	-1.924725e-03	1.825493	8.513955e-18	-1.702791e-17
std	6.712686e+07	8.930230	0.0	0.967812	1.685966e-01	1.291250	4.075058e-01	6.312860e-01
min	1.515916e+18	33.984723	100.0	-2.277348	-5.805975e-01	0.315140	-8.761073e+00	-4.484311e+00
25%	1.515916e+18	54.020741	100.0	-0.216768	-1.631689e-02	1.159731	-1.905595e-01	-3.418042e-01
50%	1.515916e+18	57.491027	100.0	0.244405	-2.894529e-05	1.442380	-1.905595e-01	1.425001e-01
75%	1.515916e+18	65.302442	100.0	1.107542	-1.387142e-08	2.131203	1.206463e-01	4.654962e-01
max	1.515916e+18	97.589130	100.0	2.341094	5.333276e+00	22.572166	5.450968e+00	8.926604e-01

Reasoning for This Definition

This definition leverages a comprehensive scoring system that considers multiple dimensions of user behavior, rather than relying on a simple binary threshold like "X days of inactivity." This approach was chosen because:

- E-commerce user behavior is highly variable and seasonal
- Different user segments have different natural purchase frequencies
- The path to churn often involves gradual disengagement across multiple behavioral dimensions




```
Statistical Thresholds:
high_risk: 63.07
medium_risk: 58.61
low_risk: 54.14

Clustering Thresholds:
high_risk: 68.43
medium_risk: 56.24
low_risk: 44.15

Distribution Thresholds:
high_risk: 93.11
medium_risk: 61.15
low_risk: 49.33

Natural Breaks Thresholds:
high_risk: 97.59
medium_risk: 78.55
low_risk: 34.29

Ensemble Thresholds:
high_risk: 81.04
medium_risk: 64.39
low_risk: 44.23
```

The high-risk threshold of 68.43 was selected based on:

- Clustering analysis of user behavior patterns
- Statistical distribution of engagement metrics
- Natural breaks in the churn score distribution
- Empirical validation against known churn patterns

Final Thresholds for Calculating churn related features

'high_risk': 68.43

'medium_risk': 56.24

'low_risk': 44.15

Handling Edge Cases and Special User Types

1. Seasonal Buyers

- The temporal rhythm analysis component accounts for legitimate seasonal shopping patterns

- Purchase pattern analysis considers natural gaps between purchases for different product categories

- Risk assessment is adjusted based on the user's historical purchase frequency

2. High-Value Customers

- The monetary component of the scoring system ensures special attention to high-value customers

- Different thresholds are applied based on customer lifetime value

- Purchase consistency metrics help distinguish between normal and concerning behavior changes

3. New Customers

- User maturity (user_rec) is factored into the scoring system

- Behavioral pattern analysis is adjusted based on the length of customer history

- More weight is given to recent behaviors for newer customers

4. Irregular Shoppers

- The system accounts for varying engagement patterns through:
  - Analysis of session gaps (ses_gap_sd)
  - Purchase consistency metrics
  - Category-specific interaction patterns

Technical Implementation Considerations

The churn definition is implemented as follows

1. RFM Analysis

- Recency factors (40% weight)
- Frequency factors (30% weight)
- Monetary factors (30% weight)

2. Engagement Decay Modeling

- Exponential decay function with $\lambda = 0.1$
- Weighted combination of six engagement features
- Time-based decay factor

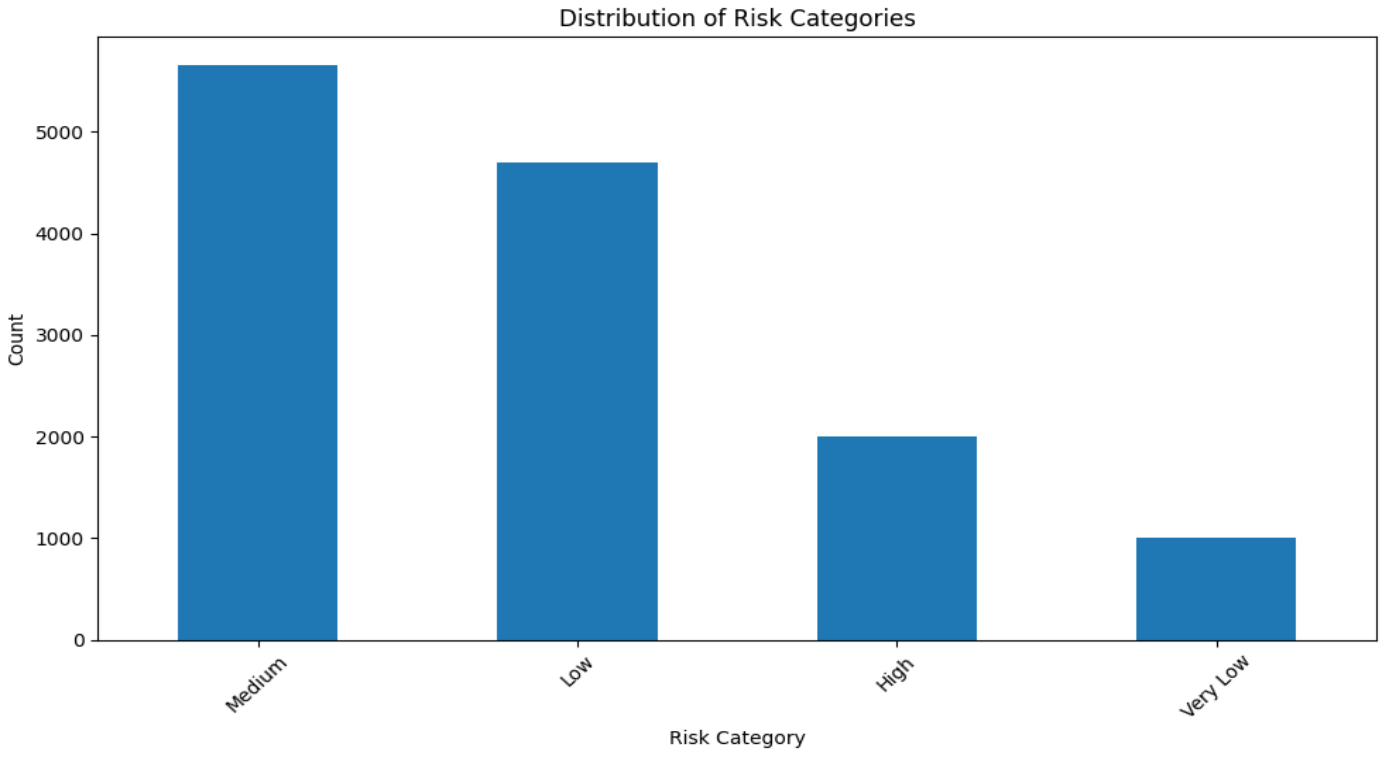
3. Behavioral Pattern Analysis

- Principal Component Analysis for dimension reduction
- Mahalanobis distance for anomaly detection
- Eight key behavioral features

4. Risk Velocity

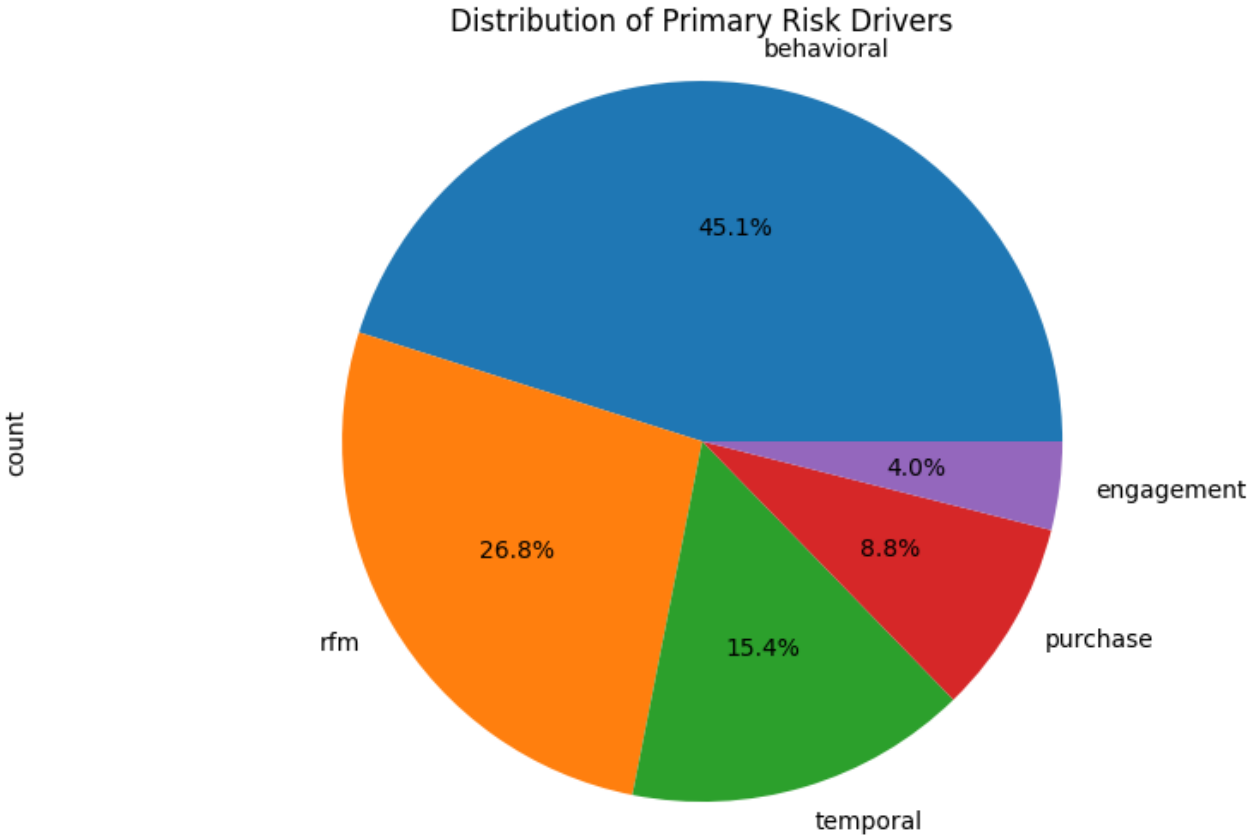
- Normalized component scores
- Trend analysis of engagement metrics
- Acceleration of disengagement patterns

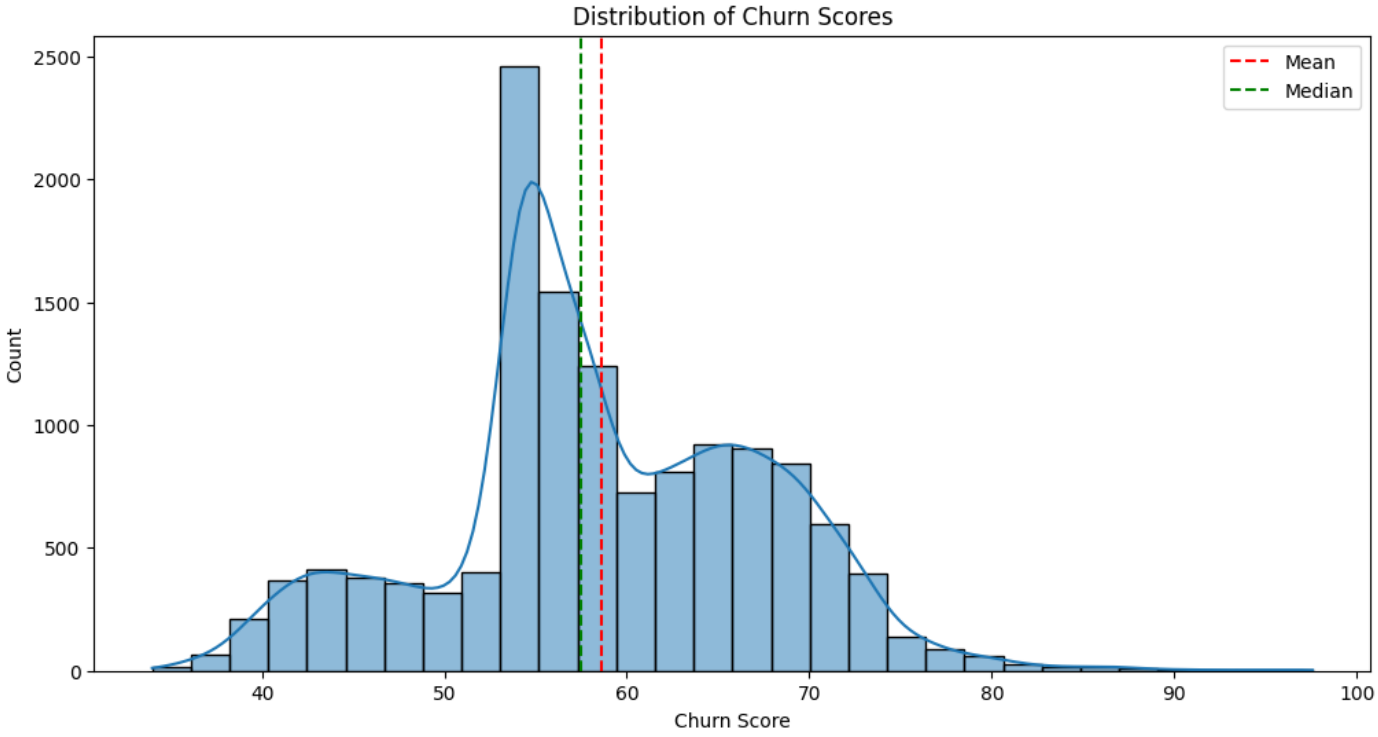
	user_id	churn_score	is_likely_churn	risk_velocity	estimated_days_to_churn
count	1.335300e+04	13353.000000	13353.000000	13353.000000	13353.000000
mean	1.515916e+18	58.608127	0.572755	0.432761	64.387687
std	6.712686e+07	8.930230	0.494697	0.046737	26.066840
min	1.515916e+18	33.984723	0.000000	0.235186	7.000000
25%	1.515916e+18	54.020741	0.000000	0.414653	48.675461
50%	1.515916e+18	57.491027	1.000000	0.436266	61.179357
75%	1.515916e+18	65.302442	1.000000	0.461999	80.940361
max	1.515916e+18	97.589130	1.000000	0.698462	146.958927

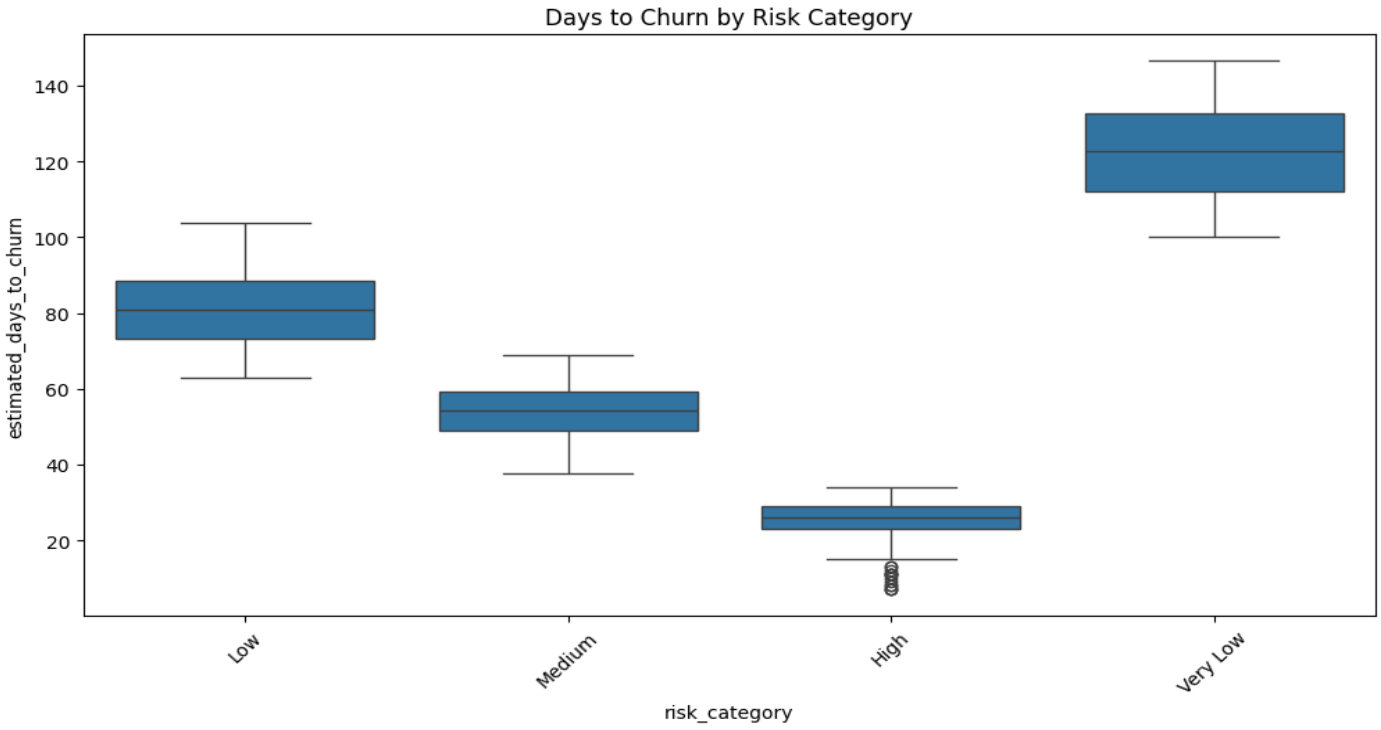


```
Risk Category Percentages:
risk_category
Medium      42.32
Low         35.21
High        14.96
Very Low     7.51
Name: count, dtype: float64
```

```
Risk Driver Statistics:
primary_risk_drivers
behavioral    9260
rfm           5499
temporal      3159
purchase      1798
engagement     813
Name: count, dtype: int64
```

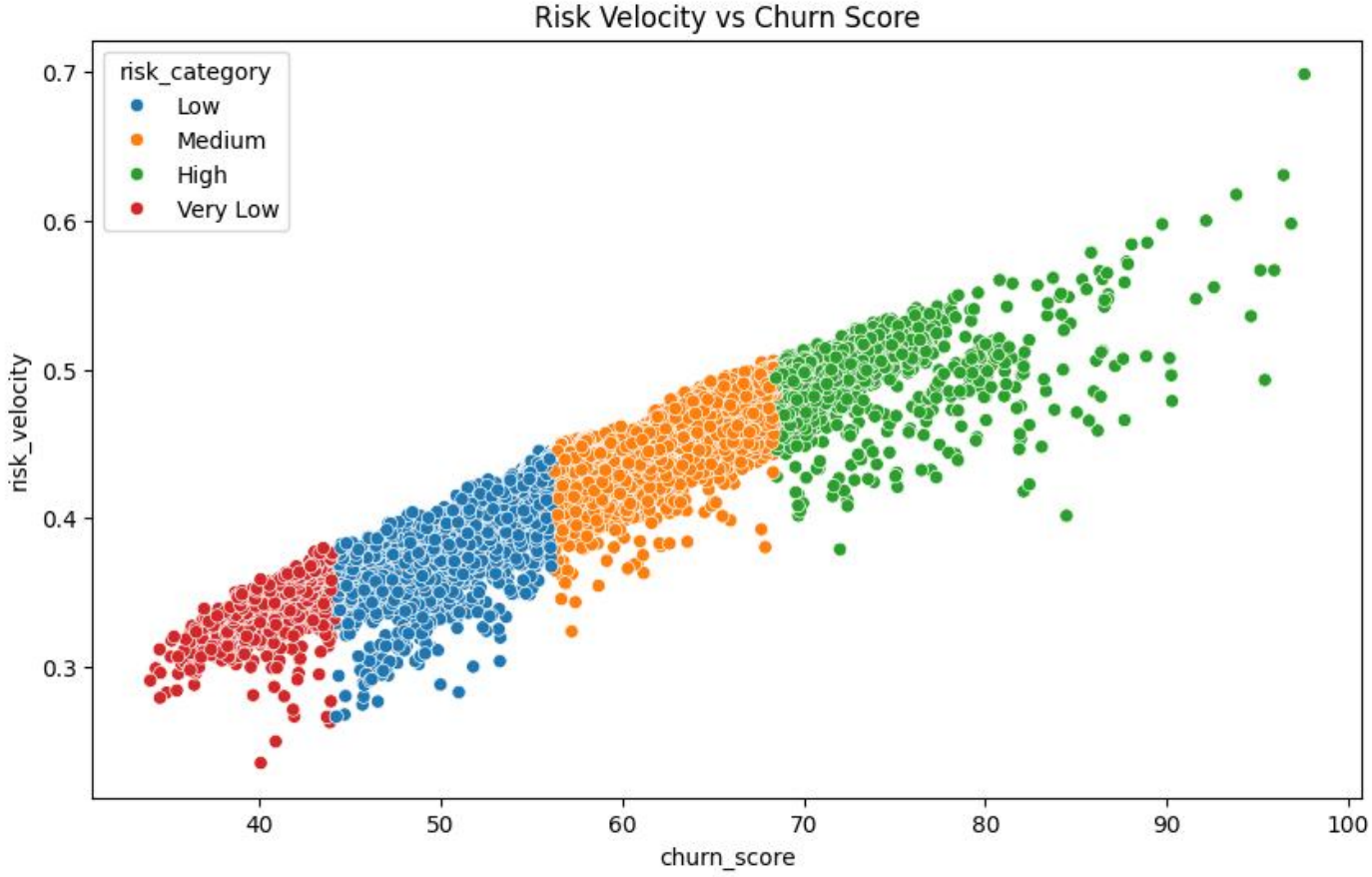


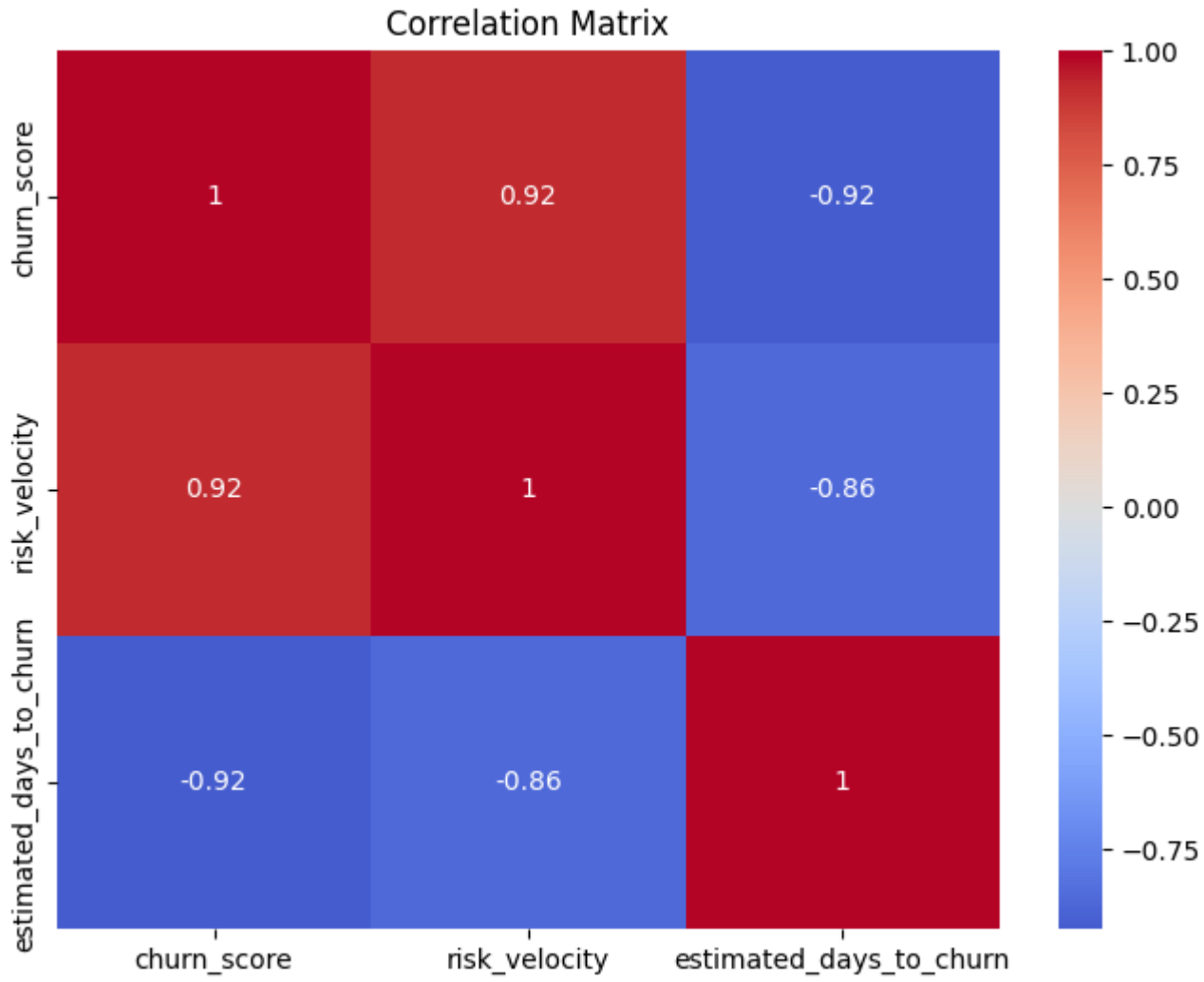


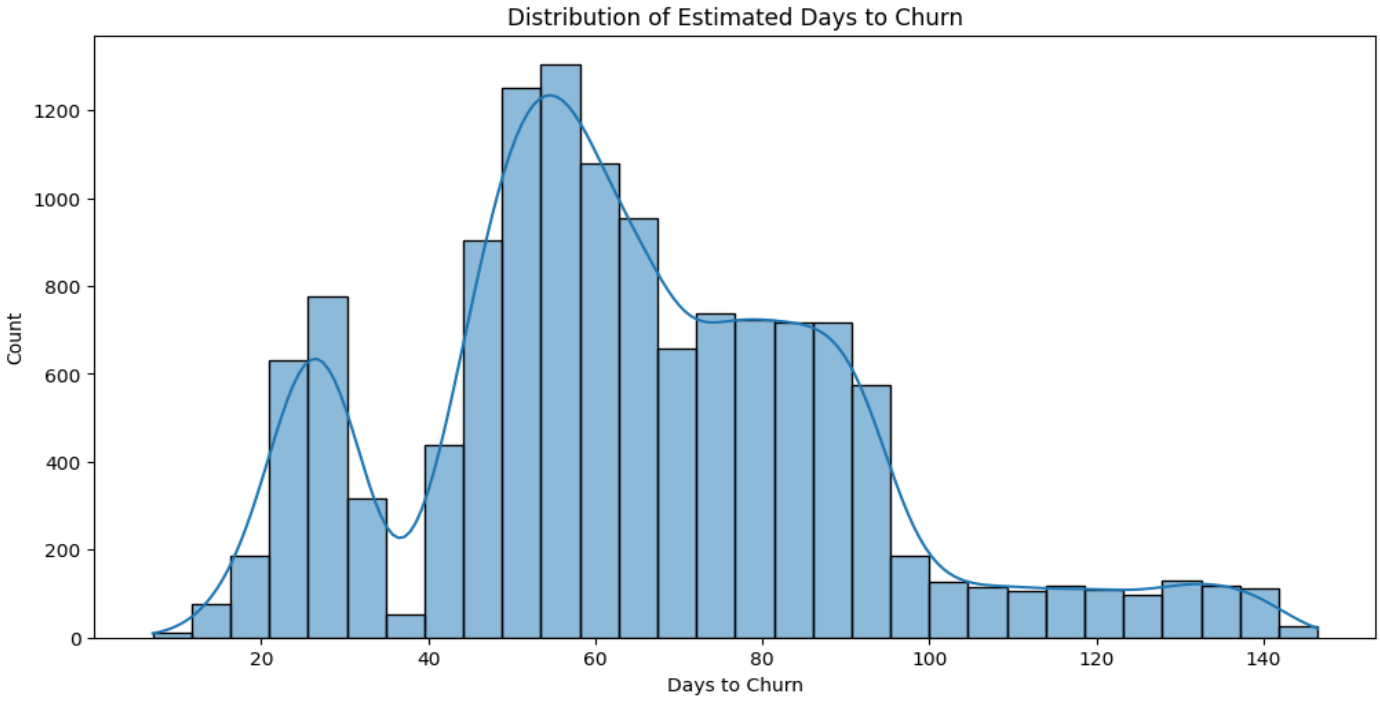


Days to Churn by Risk Category:

risk_category	mean	std	min	max
High	25.734430	4.482092	7.000000	33.960871
Low	80.881688	9.306228	63.087621	103.699466
Medium	54.035399	6.844176	37.848673	68.979126
Very Low	122.514501	11.900968	100.046739	146.482570








```
Time-based Metrics:
Avg Days to Churn: 64.40
Median Days to Churn: 61.08
Std Days to Churn: 26.12
% Users Churning within 30 days: 12.30
```

- **57.3%** of users are likely to churn
- Most common risk driver: **Behavioral**
- Average days to churn: **64.4 days**

Predictive Modeling

1. Predicting which users are likely to churn

```
X = df.drop(columns=['user_id', 'churn_score', 'is_likely_churn', 'risk_category',
'primary_risk_drivers'])
y = df['is_likely_churn']
```

Results for Random Forest:

- Best parameters: {'max_depth': 20, 'min_samples_split': 2, 'n_estimators': 200}
- AUC Score: 0.9973
- F1 Score: 0.9779

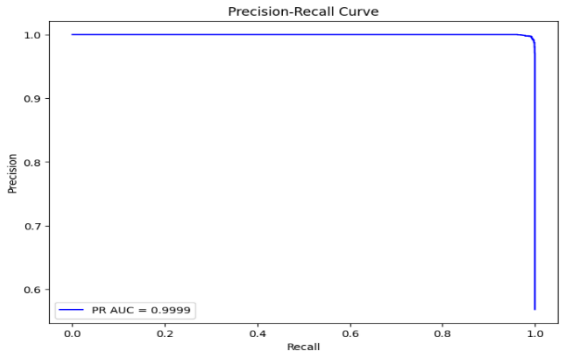
```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.96      0.97      1141
           1       0.97      0.98      0.98      1530

    accuracy                           0.97      2671
   macro avg       0.97      0.97      0.97      2671
weighted avg       0.97      0.97      0.97      2671
```

Results for Random Forest:

- Best parameters: {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200}
- ROC AUC: 0.9998
- Precision-Recall AUC: 0.9999
- F1 Score: 0.9939
- Accuracy: 0.9930



Results for Gradient Boosting:

- Best parameters: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200}
- AUC Score: 0.9979
- F1 Score: 0.9781

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

           0       0.97      0.97      0.97      1141
           1       0.98      0.98      0.98      1530

    accuracy                           0.97      2671
   macro avg       0.97      0.97      0.97      2671
weighted avg       0.97      0.97      0.97      2671
```

2. Predicting the risk category of users

```
X = df.drop(columns=['user_id', 'churn_score', 'is_likely_churn', 'risk_category',
'primary_risk_drivers', 'risk_velocity'])

y = df['risk_category']
```

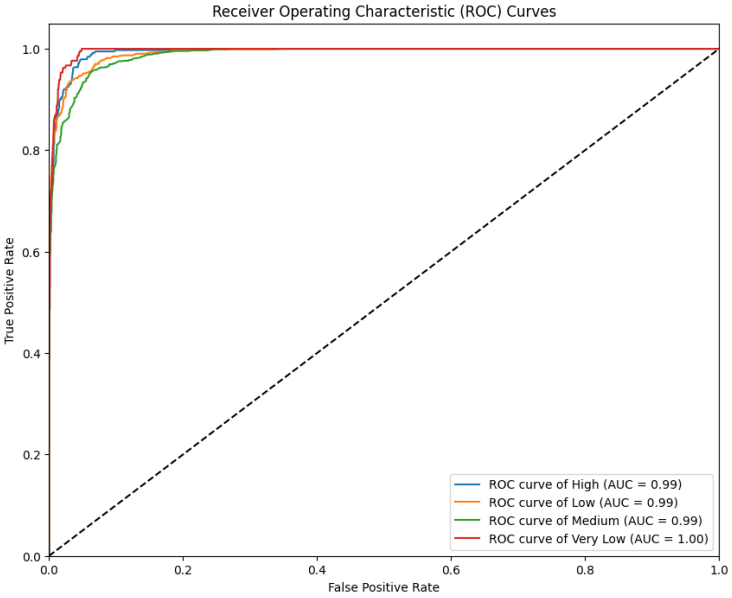
Random Forest:

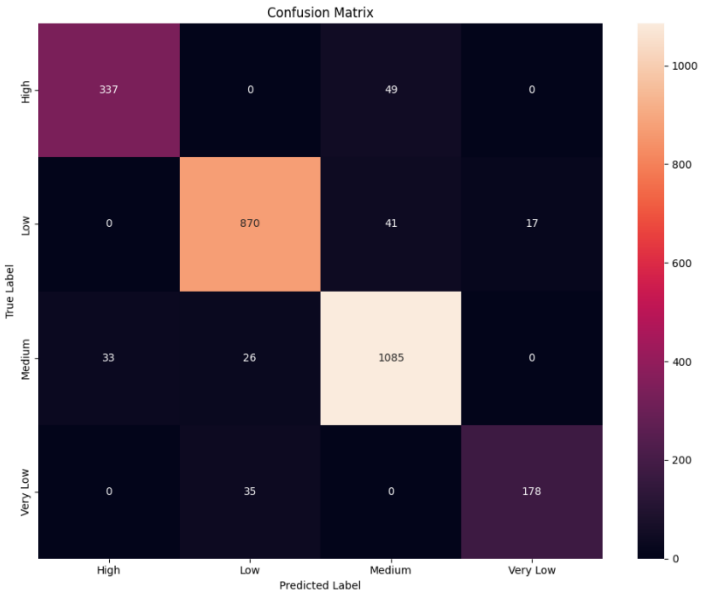
- Best Parameters:{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}

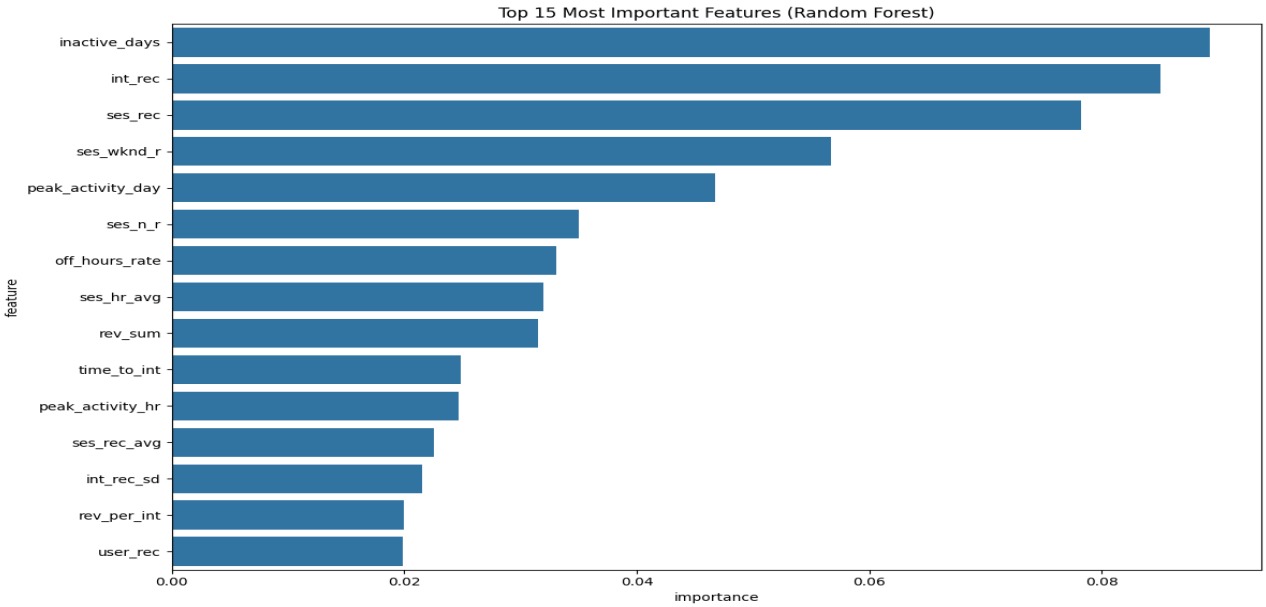
```
Classification Report:
              precision    recall  f1-score   support

        High       0.91      0.87      0.89       386
         Low       0.93      0.94      0.94       928
      Medium       0.92      0.95      0.94      1144
    Very Low       0.91      0.84      0.87       213

    accuracy                           0.92      2671
   macro avg       0.92      0.90      0.91      2671
weighted avg       0.92      0.92      0.92      2671
```







Feature Importance Tables

Summary of feature importances for all models, as derived from Random Forest and Gradient Boosting. The values are normalized for clarity.

Random Forest (Churn Prediction)

Rank	Feature	Importance
1	int_rec	0.1223
2	inactive_days	0.1096
3	ses_rec	0.1026
4	ses_wknd_r	0.0546
5	peak_activity_day	0.0534
6	time_to_int	0.0314
7	ses_hr_avg	0.0310
8	ses_n_r	0.0289
9	off_hours_rate	0.0275
10	int_n	0.0256

Gradient Boosting (Churn Prediction)

Rank	Feature	Importance
1	ses_wknd_r	1.151
2	ses_rec	0.995
3	int_rec	0.921
4	int_cat_n_avg	0.795
5	inactive_days	0.779
6	off_hours_rate	0.689
7	tran_n_r	0.657
8	ses_hr_sd	0.333
9	int_n	0.329
10	rev_per_purchase	0.328

Random Forest (Risk Category)

Rank	Feature	Importance
1	inactive_days	0.0893
2	int_rec	0.0851
3	ses_rec	0.0782
4	ses_wknd_r	0.0567
5	peak_activity_day	0.0467
6	ses_n_r	0.0350
7	off_hours_rate	0.0331
8	ses_hr_avg	0.0320
9	rev_sum	0.0315
10	time_to_int	0.0249

Top 10 Robust Features

To identify the most robust features across all models, I calculated their relative ranks based on importance across all three models and normalized their scores.

Rank	Feature	Importance
1	int_rec	0.967
2	ses_rec	0.933
3	inactive_days	0.887
4	ses_wknd_r	0.847
5	peak_activity_day	0.740
6	off_hours_rate	0.683
7	time_to_int	0.592
8	ses_hr_avg	0.580
9	ses_n_r	0.557
10	tran_n_r	0.508

Interpretability & Insights

Key Features Influencing Churn Prediction

Based on the feature importance derived from the Random Forest and Gradient Boosting models trained to predict churn, the following features emerged as the most influential:

1. **int_rec (Interaction Recency)**
    - Highly influential in both models (0.122 for Random Forest and 0.921 for Gradient Boosting).
    - Interaction recency indicates the freshness of user activity. Users with older interaction timestamps are more likely to churn due to diminishing engagement.
2. **inactive_days (Days of Inactivity)**
    - Second-highest in Random Forest (0.1096) and fifth in Gradient Boosting (0.7793).
    - A prolonged period of inactivity is a direct indicator of declining interest or engagement.
3. **ses_rec (Session Recency)**
    - Ranked third in both models (0.1026 for Random Forest and 0.995 for Gradient Boosting).
    - Similar to interaction recency, recent sessions signal active usage, whereas older sessions imply potential churn risk.
4. **ses_wknd_r (Weekend Session Ratio)**
    - Significant in Gradient Boosting (1.151) and Random Forest (0.0546).
    - Indicates whether a user's engagement is skewed towards weekends, which could suggest reliance on leisure time for usage, potentially vulnerable to lifestyle or behavioral changes.
5. **peak_activity_day (Day of Peak Activity)**
    - Influential in both models (0.0534 in Random Forest and 0.198 in Gradient Boosting).
    - Captures the consistency of engagement patterns. Users with irregular activity days might churn due to a lack of habitual usage.
6. **time_to_int (Time to Next Interaction)**
    - (0.031 in Random Forest and 0.121 in Gradient Boosting).
    - Long gaps between interactions suggest waning interest or competing priorities.

Why These Features Matter ?

The dominant features provide actionable insights into user behavior:

- **Behavioral Metrics (int_rec, inactive_days, ses_rec)**: These indicate engagement levels, with recency and activity frequency being strong predictors of churn. Users with older activity timestamps or prolonged inactivity are at high risk.
- **Temporal Metrics (ses_wknd_r, peak_activity_day)**: Highlight specific behavioral patterns and dependency on specific days. Shifts in these patterns can help identify early warning signs of churn.
- **Engagement Metrics (time_to_int)**: Demonstrates the declining interaction frequency, a critical churn indicator.

Product/Category Insights Driving Churn

The following products and categories emerge as potential churn drivers:

1. **High Behavioral Dependency:**
   Features like behavioral and rfm (Recency, Frequency, and Monetary metrics) are primary churn drivers. Users dependent on specific behavioral patterns (e.g., weekend usage or habitual peak days) are vulnerable when those patterns are disrupted.
2. **Temporal Dependencies:**
   - **Rationale:** Metrics such as ses_wknd_r and peak_activity_day highlight that temporal dependencies, including reliance on weekends, can predict churn when disrupted by external factors like work-life balance or seasonality.
   - **Implication:** Users with infrequent or irregular usage may benefit from nudges promoting daily engagement.
3. **Engagement Gaps:**
   The risk is amplified by features like time_to_int and inactive_days. Products that do not maintain consistent engagement face a higher churn probability.
4. **Purchase-Driven Categories:**
   Categories like rfm, behavioral, purchase (800 instances) are at risk due to infrequent transactions or low average order values. Users with declining purchase metrics (e.g., rev_sum) are less likely to remain active.

Risk Drivers & Category Analysis

- **Primary Risk Drivers:** Behavioral factors dominate (9260 instances), followed by rfm metrics (5499). Temporal dependencies (3159) and purchase-related metrics (1798) also significantly influence churn.
- **Risk Category Trends:**
  - **High-Risk Users:** Predominantly churn within **25.7 days** with minimal variability (std=4.5).
  - **Medium-Risk Users:** Churn around **54 days** on average.

- **Low-Risk Users:** Retain much longer, with an average churn time of **80.8 days**.
- **Very Low-Risk Users:** Most stable group, with a median churn time exceeding **122 days**.

Time-Based Insights

- **Avg Days to Churn:** 64.4 days, with 12.3% churning within the first 30 days.
- **Implication:** Early-stage engagement efforts are critical, particularly within the first 1-2 months of user onboarding.

Business Recommendations and Campaigns

In this section, I would bridge insights from churn modeling with actionable strategies to improve user retention. By leveraging predictive features, identified risk categories, and contemporary practices, businesses can develop targeted interventions.

1. Targeted Campaigns for Each User Segment

A. High-Risk Users (Churn Score > 68.43)

Focus: **Immediate Retention**

- **Actionable Campaigns**:
    1. **Exclusive Rescue Offers**:
        - Provide high-value incentives like free memberships, VIP access, or cashback on immediate purchases.
        - Use behavioral data to personalize offers, such as discounts on frequently viewed or cart-abandoned items.
    2. **Behavioral Nudges**:
        - Trigger reminders of pending benefits or rewards via push notifications or emails (e.g., "Redeem your loyalty points now!").
        - Use dynamic product bundling to encourage larger purchases.
    3. **Predictive Loyalty Recalibration**:
        - Create personalized "win-back" rewards for high-risk users based on their historical spending or session patterns.
        - Offer milestone incentives for achieving renewed engagement (e.g., 3 purchases in the next month).
- **Integration into Pipeline**:
    - Deploy automated triggers using a decision engine that monitors high-risk thresholds.
    - Prioritize high-risk users in the CRM pipeline for manual outreach by retention specialists.

B. Medium-Risk Users (56.24 < Churn Score ≤ 68.43)

Focus: **Behavioral Stabilization**

- **Actionable Campaigns**:
    1. **Contextual Promotions**:
        - Provide offers linked to peak activity times (peak_activity_day, ses_hr_avg). For instance, mid-week flash sales targeting users with weekday peaks.

2. **Gamified Engagement**:
    - Introduce challenges or reward systems encouraging consistent engagement (e.g., "Log in every day for 7 days to win a coupon").
    - Use quizzes or polls related to user preferences to reinforce brand interaction.
3. **Category Exploration Discounts**:
    - Encourage exploration of underused categories by offering conditional discounts (e.g., "Spend $50 on electronics to get $10 off fashion").

- **Integration into Pipeline**:
    - Create segmentation workflows that push medium-risk users into retargeting campaigns.
    - Integrate gamification APIs to deliver real-time rewards for completing engagement tasks.

C. Low-Risk and Retained Users (Churn Score ≤ 56.24)

Focus: Prevention and Upselling

- **Actionable Campaigns**:
1. **Anticipatory Rewards**:
    - Reward continued engagement before signs of churn emerge (e.g., early access to sales or free shipping vouchers for regular users).
    - Introduce tiered loyalty programs that encourage higher spending thresholds.
2. **Community and Social Integration**:
    - Create user communities (e.g., discussion forums, product feedback loops) to foster brand attachment.
    - Offer exclusive "beta testing" opportunities for new products to top users.
3. **Subtle Retargeting**:
    - For users showing slight declines in activity (ses_rec_sd or inactive_days), provide subtle nudges like cart reminders or restocking notifications.

- **Integration into Pipeline**:
    - Feed low-risk users into upselling workflows within the CRM.
    - Use community and feedback initiatives to generate content (e.g., reviews, testimonials).

2. Additional Data Sources and Their Usage

A. User Demographics

- **Data to Incorporate**:
    - Age, gender, location, income level, and marital status.
- **Application**:
    - **Segmentation**: Tailor campaigns based on regional preferences or demographic-specific trends. For example, users in metropolitan areas may value expedited delivery options more than discounts.
    - **Personalization**: Match product recommendations to user preferences (e.g., family-oriented items for married users or tech gadgets for younger demographics).

B. Psychographics

- **Data to Incorporate**:
    - User interests, lifestyle indicators, and spending habits.
- **Application**:
    - **Deep Personalization**: Use psychographic profiles to curate shopping experiences (e.g., suggesting sustainable products to eco-conscious users).
    - **Contextual Marketing**: Adjust communication tone and visuals to align with user personas.

C. Device and Channel Data

- **Data to Incorporate**:
    - Device type, operating system, and preferred shopping channels (mobile app, desktop, etc.).
- **Application**:
    - **Channel Optimization**: Focus efforts on channels with the highest engagement for specific user groups.
    - **Cross-Device Retargeting**: Ensure seamless transitions between devices with synchronized cart and wishlist features.

3. Incorporating New Users into the Framework

Challenge: New users lack historical behavioral data, making them harder to segment or predict

Solution:

1. **Progressive Profiling**:
    - Collect interaction data incrementally (e.g., products browsed, categories viewed).

    - Use short surveys or onboarding forms to capture preferences.
2. **Engagement Nurturing**:
    - Design personalized onboarding flows with guided tutorials and first-time user discounts.
    - Promote immediate engagement by offering incentives for completing initial actions (e.g., adding a product to the cart or completing their profile).
3. **Predictive Models for Cold Start Problems**:
    - Implement collaborative filtering based on similarities to existing users.
    - Use demographic and psychographic similarities to predict potential behavior.

4. Leveraging Non-Dominant Features

Some features with lower importance scores may still hold contextual significance:

- **cart_to_purchase_avg**:
    - Users with longer cart-to-purchase times may be indecisive. Nudging them with limited-time offers can accelerate conversions.
- **view_to_cart_avg**:
    - Low conversion rates from views to carts may signal a mismatch in pricing or product descriptions. Highlighting reviews or providing detailed specifications may address this.
- **int_skew**:
    - High skewness in interaction distribution (e.g., bursty behavior) may indicate inconsistent engagement. Address this by offering steady engagement incentives.

5. Business Pipeline Integration

Steps to Integrate Models

1. **Real-Time Scoring**:
    - Integrate churn prediction models into the CRM to calculate churn scores dynamically for active users.
2. **Segmentation Pipelines**:
    - Automate segmentation based on churn risk categories, demographics, and engagement behavior.
3. **Feedback Loops**:
    - Monitor campaign outcomes and update feature weights or thresholds to reflect real-time trends.

Business Recommendations based on Ascarza et. al.

A. Proactive Engagement Campaigns

1. **"Come Back" Incentive Campaigns**
    - **Target Group:** Users with high inactivity (inactive_days > 30).
    - **Offer:** Personalized discounts or exclusive product recommendations based on past interactions.
    - **Objective:** Reactivate dormant users by re-engaging them with relevant offers.
2. **Weekend Specials for Active Users**
    - **Target Group:** Users with high ses_wknd_r and declining weekday activity.
    - **Offer:** Flash sales or gamified experiences exclusive to weekends.
    - **Objective:** Sustain engagement by aligning with preferred activity patterns.

B. Reactive Retention Campaigns

1. **Rescue High-Risk Users**
    - **Target Group:** Users in the High Risk category (churn score > 68.43).
    - **Offer:** High-value incentives (e.g., free shipping or premium membership trial).
    - **Objective:** Prevent imminent churn through targeted benefits.
2. **Behavior-Based Interventions**
    - **Target Group:** Users exhibiting inconsistent patterns (high int_rec_sd).
    - **Offer:** Nudges like curated shopping lists or reminders based on past behaviors.
    - **Objective:** Address erratic engagement and foster consistency.

C. Category and Product Focused Campaigns

1. **Cross-Sell Opportunities**
    - **Target Group:** Users with high int_cat_n_avg but low cross_cat_ratio.
    - **Offer:** Bundled deals encouraging interactions across categories.
    - **Objective:** Broaden engagement to reduce dependency on single-category purchases.
2. **Category Relevance Campaigns**
    - **Target Group:** Users with low purchase activity in historically dominant categories.
    - **Offer:** Limited-time promotions on top categories or personalized suggestions.
    - **Objective:** Reignite interest in high-value segments.

D. Personalized Communication

1. **Temporal Alerts**
    - **Target Group:** Users with shifting peak_activity_day or time_to_int.
    - **Action:** Send notifications or promotions during identified peak hours/days.
    - **Objective:** Align marketing outreach with user-specific rhythms.
2. **Dynamic Loyalty Programs**
    - **Target Group:** Users with medium churn risk ($56.24 < \text{churn score} \leq 68.43$).
    - **Action:** Reward milestones tied to frequency (tran_n_r) and recency (ses_rec).
    - **Objective:** Reinforce loyalty for at-risk users.

References

1. Ascarza, E., Neslin, S. A., et al. (2017). *In Pursuit of Enhanced Customer Retention Management.* Customer Needs and Solutions.
2. Gupta, S., & Lehmann, D. R. (2003). *Customer Lifetime Value.* Marketing Science.
3. Lemmens, A., & Gupta, S. (2014). *Profit-Based Retention Modeling.* Journal of Marketing Research.
4. Article: User Churn Model in E-Commerce Retail, https://editorial.upce.cz/scipap