Customer Churn Analysis with Logistic and Tree-Based Models October 2025

Chiawei Wang
Data & Product Analyst
chiawei.w@outlook.com

Executive Summary

Challenge

The persistent failure of customers to rebook within six months is driving **significant customer attrition** and directly **eroding long-term key value**. This analysis was commissioned to definitively **diagnose the root drivers of churn** and deploy a **predictive, proactive intervention strategy.**

Objective

- **Diagnose and quantify** the key behavioural factors driving customer attrition.
- **Develop a robust, high-precision predictive engine** to flag **high-risk customers** in real-time.
- Design and deploy **targeted, automated interventions** to **reduce churn rate and lift customer lifetime value**.

Solution

We evaluated multiple techniques, confirming the **XGBoost model** as the optimal predictive engine (**ROC AUC = 0.874**). This analysis identified **loyalty, recency, channel, and platform** as primary churn drivers. The predictive engine is now **ready for deployment into the CRM system** to enable **real-time risk scoring** and launch automated, targeted retention campaigns.

Key Insights

- **Recency threshold:** Churn risk triples after 90 days without engagement, defining the critical intervention window.
- **Loyalty status risk:** Non-members and new customers are the primary source of churn volatility.
- **Marketing channel risk:** Direct traffic is highly brand-loyal (low churn), but acquisition channels introduce a measurable risk profile (volatility).
- **Platform value:** App users are 15% more secure (lower churn rate) than web users.

Strategic Recommendations

- **Recency:** Deploy an automated, risk-scored 90 day intervention campaign to pre-empt the churn acceleration.
- **Loyalty:** Implement mandatory lottery enrollment immediately post-booking to stabilize the high-risk new customer transition..
- **Channel:** Score and strategically divert high-risk acquisition traffic into high-priority loyalty enrollment funnel to protect ROI.
- **Platform:** Use targeted incentives to migrate high-risk web users to the app platform.

Next Steps

We must establish the **predictive engine** to accurately **flag high-risk customers** (failing to rebook within 6 months), and the next step is securing **approval for the execution roadmap and engine deployment**.

Table of Contents

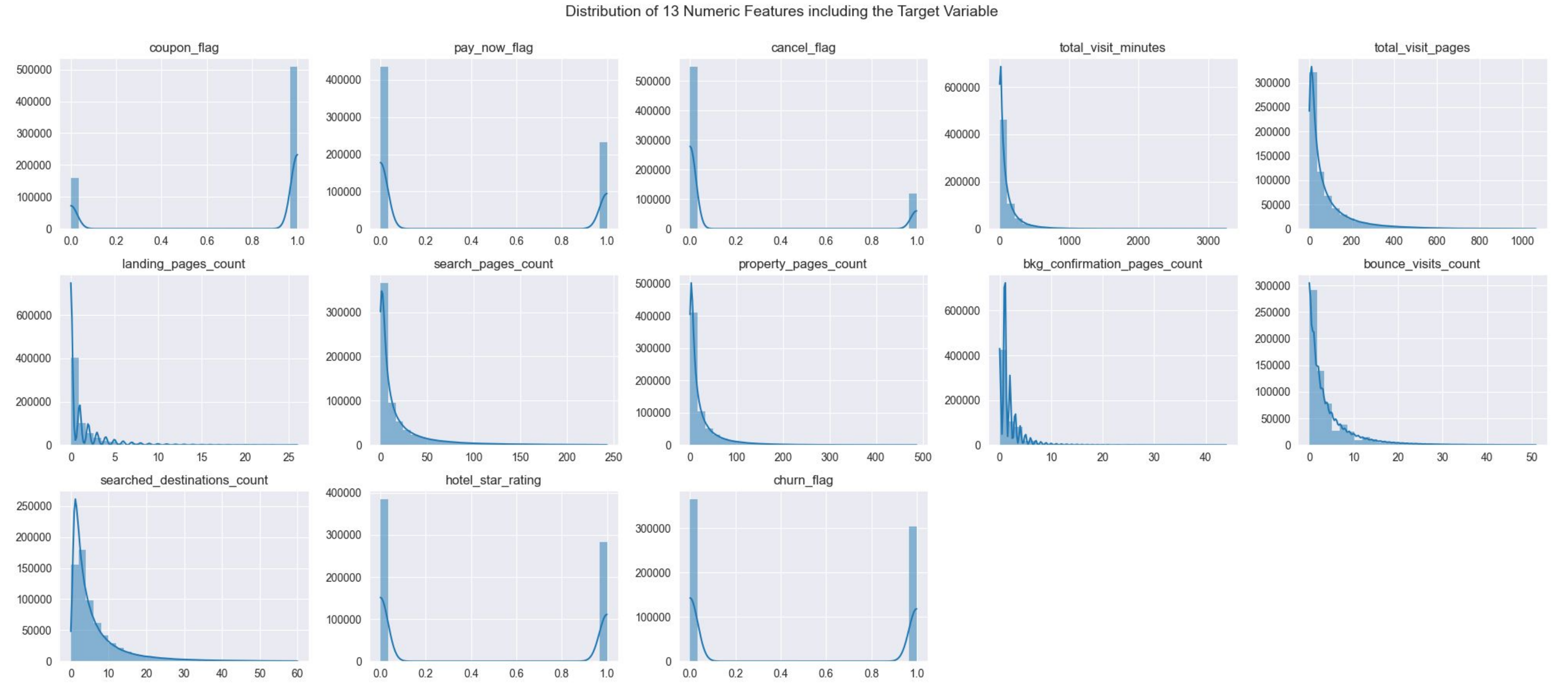
Key sections for this predictive churn analytics report

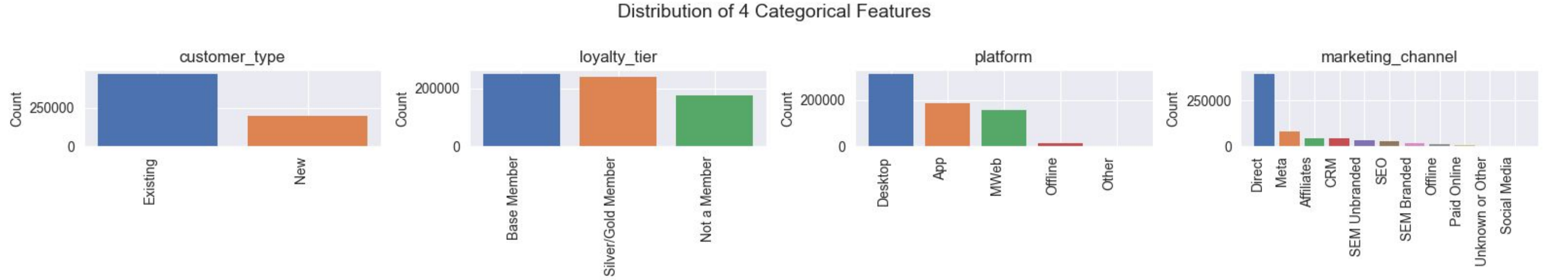
1. Overview of Data

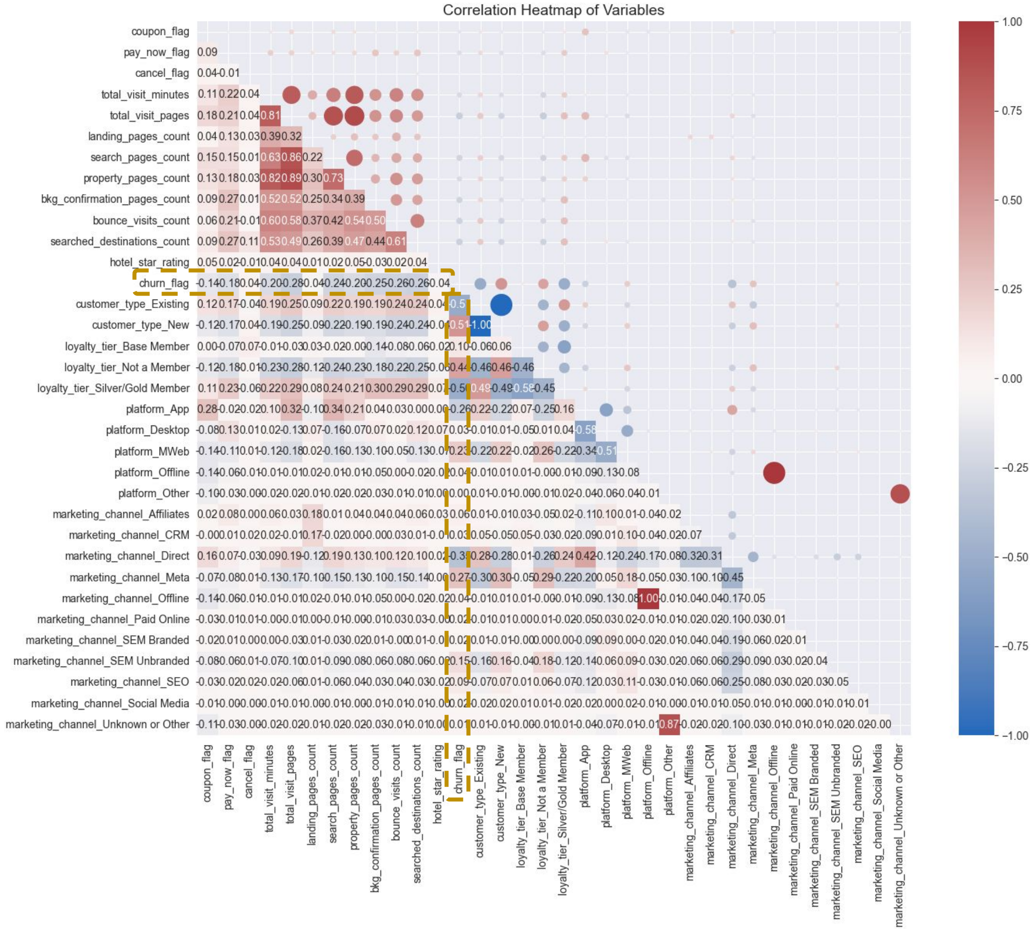
The dataset contains 689,742 entries and 21 columns. The target variable is 'churn_flag', which indicates whether a customer has churned (1) or not (0). The dataset includes various features related to customer behavior, booking details, and website interaction metrics.

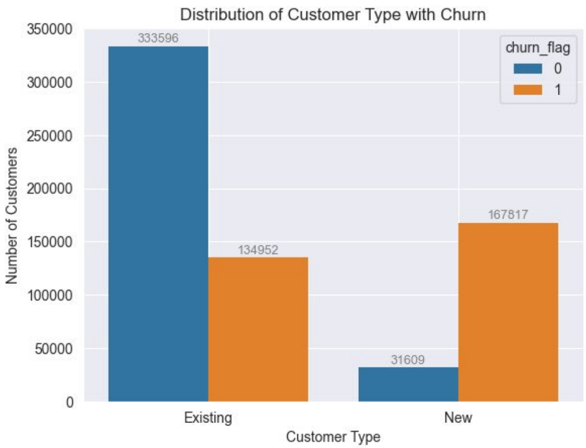
Description of Variables

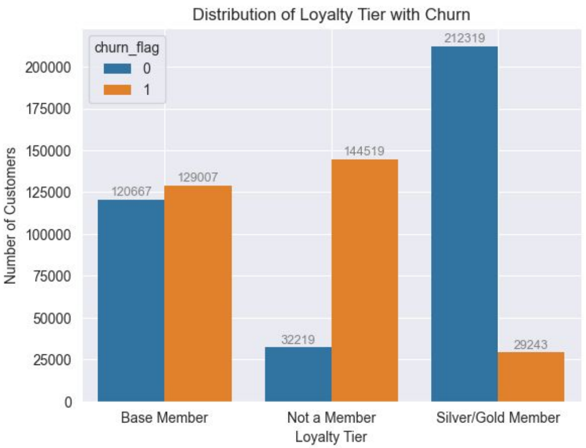
Variable	Type	Description
email_address	object	Unique id for each customer
booking_id	object	Unique id for each booking
bk_date	datetime64	Booking date, format YYYY-MM-DD
coupon_flag	int64	1 = used coupon when booking
pay_now_flag	int64	1 = payment choice 'pay now', 0 = 'pay later'
cancel_flag	int64	1 = canceled, 0 = not canceled, may contain missing values
cancel_date	datetime64	Cancel date, format YYYY-MM-DD, only present if booking was canceled
customer_type	object	Indicates customer type (e.g., new or returning customer)
loyalty_tier	object	Customer's loyalty level: 0 = not a member, 1 = base member, 2 = silver/gold member
platform	object	Platform used for booking (e.g., desktop, mobile website)
marketing_channel	object	How the visit came to the website (e.g., Google, TripAdvisor), may contain missing values
total_visit_minutes	int64	Total minutes of visits to the website, may contain missing values
total_visit_pages	int64	Total number of pages visited on the website
landing_pages_count	int64	Number of landing pages visited on the website
search_pages_count	int64	Number of search pages visited on the website
property_pages_count	int64	Number of property pages visited on the website
bkg_confirmation_pages_count	int64	Number of booking confirmation pages visited on the website
bounce_visits_count	int64	Total number of single page visits to the website
searched_destinations_count	int64	Total number of different destinations searched on the website
hotel_star_rating	int64	Star rating score of the booked hotel
churn_flag	int64	1 = No repeat booking 6 months after check out (churned), 0 = repeat booking (not churned)

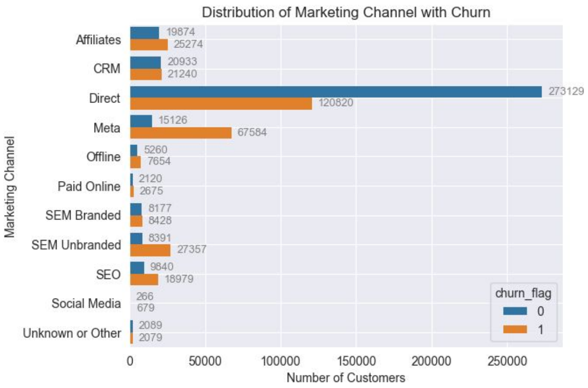

Distribution of 13 Numeric Features including the Target Variable
coupon_flag
pay_now_flag
cancel_flag
total_visit_minutes
total_visit_pages
landing_pages_count
search_pages_count
property_pages_count
bkg_confirmation_pages_count
bounce_visits_count
searched_destinations_count
hotel_star_rating
churn_flag

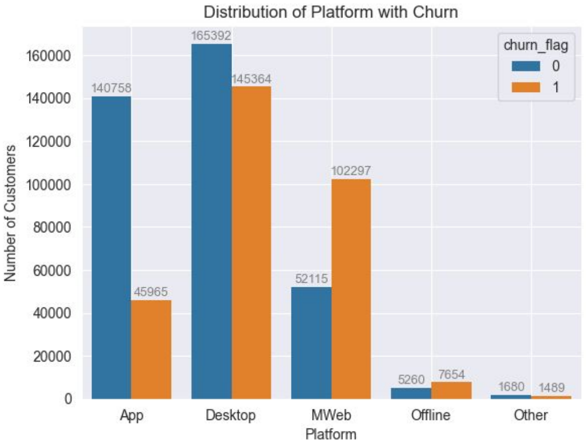

Distribution of 4 Categorical Features
customer_type
Count
Existing
New
loyalty_tier
Base Member
Silver/Gold Member
Not a Member
platform
Desktop
App
MWeb
Offline
Other
marketing_channel
Direct
Meta
Affiliates
CRM
SEM Unbranded
SEO
SEM Branded
Offline
Paid Online
Unknown or Other
Social Media


Correlation Heatmap of Variables
coupon_flag
pay_now_flag
cancel_flag
total_visit_minutes
total_visit_pages
landing_pages_count
search_pages_count
property_pages_count
bkg_confirmation_pages_count
bounce_visits_count
searched_destinations_count
hotel_star_rating
churn_flag
customer_type_Existing
customer_type_New
loyalty_tier_Base Member
loyalty_tier_Not a Member
loyalty_tier_Silver/Gold Member
platform_App
platform_Desktop
platform_MWeb
platform_Offline
platform_Other
marketing_channel_Affiliates
marketing_channel_CRM
marketing_channel_Direct
marketing_channel_Meta
marketing_channel_Offline
marketing_channel_Paid Online
marketing_channel_SEM Branded
marketing_channel_SEM Unbranded
marketing_channel_SEO
marketing_channel_Social Media
marketing_channel_Unknown or Other
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Distribution of Customer Type with Churn
churn_flag
0
1
Number of Customers
350000
300000
250000
200000
150000
100000
50000
0
333596
134952
31609
167817
Existing
New
Customer Type


Distribution of Loyalty Tier with Churn
churn_flag
0
1
Number of Customers
200000
175000
150000
125000
100000
75000
50000
25000
0
120667
129007
32219
144519
212319
29243
Base Member
Not a Member
Silver/Gold Member
Loyalty Tier


Distribution of Marketing Channel with Churn
Marketing Channel
Affiliates
19874
25274
CRM
20933
21240
Direct
273129
120820
Meta
15126
67584
Offline
5260
7654
Paid Online
2120
2675
SEM Branded
8177
8428
SEM Unbranded
8391
27357
SEO
9840
18979
Social Media
266
679
Unknown or Other
2089
2079
churn_flag
0
1
0
50000
100000
150000
200000
250000
Number of Customers


Distribution of Platform with Churn
churn_flag
0
1
Number of Customers
160000
140000
120000
100000
80000
60000
40000
20000
0
140758
45965
165392
145364
52115
102297
5260
7654
1680
1489
App
Desktop
MWeb
Offline
Other
Platform

2. Model Foundation & Feature Importance

Given the binary nature of the target variable (churned versus not churned), 4 classification algorithms were considered for modeling customer churn. The models evaluated include Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each model was assessed based on its performance metrics, interpretability, and computational efficiency to determine the most suitable approach for predicting customer churn.

Model Performance

Following the evaluation of four different models, the **XGBoost model emerged as the most effective predictive engine**, achieving a strong cross-validated **ROC AUC of 0.874**.

This score confirms **excellent discrimination** power. The model correctly ranks a churner above a non-churner 87.4% of the time. Crucially for targeted action, we can **correctly identify 8 out of every 10 customers who are actually going to churn**, ensuring a low false-negative rate for retention campaigns.

The model is statistically reliable, validated for its predictive accuracy, and is **ready for immediate deployment** to score the customer base daily.

* Note that terminology is explained in the notes throughout the document.

Logistic Regression

- ROC AUC: 0.858
- Accuracy: 0.78
- Precision: 0.76
- Recall: 0.78
- F1: 0.77

Decision Tree

- ROC AUC: 0.867
- Accuracy: 0.79
- Precision: 0.8
- Recall: 0.74
- F1: 0.77

Random Forest

- ROC AUC: 0.872
- Accuracy: 0.79
- Precision: 0.8
- Recall: 0.75
- F1: 0.77

XGBoost

- **ROC AUC: 0.874**
- Accuracy: 0.79
- Precision: 0.8
- Recall: 0.75
- F1: 0.77

Influential Factors by Coefficient & Gini Importance

Top 10 most influential factors by coefficient magnitude for Logistic Regression and by Gini importance for Decision Tree, Random Forest, and XGBoost models

Logistic Regression

	Coefficient
loyalty_tier_Not a Member_sum	1.015602
loyalty_tier_Silver/Gold Member_sum	-0.920002
customer_type_New_sum	0.817832
marketing_channel_Direct_sum	-0.680318
search_intensity	-0.598496
customer_type_Existing_sum	-0.516183
marketing_channel_Affiliates_sum	0.438632
marketing_channel_Meta_sum	0.381890
num_bookings	0.301649
platform_MWeb_sum	0.228698

Decision Tree

	gini_importance
customer_type_Existing_sum	0.593722
loyalty_tier_Silver/Gold Member_sum	0.190320
marketing_channel_Direct_sum	0.085609
loyalty_tier_Base Member_sum	0.026154
loyalty_tier_Not a Member_sum	0.020642
bkg_confirmation_pages_count_sum	0.018339
total_visit_pages_sum	0.011637
searched_destinations_count_max	0.009422
avg_page_time	0.007863
total_visit_pages_mean	0.005643

Random Forest

	gini_importance
customer_type_Existing_sum	0.282381
loyalty_tier_Silver/Gold Member_sum	0.200166
customer_type_New_sum	0.121633
loyalty_tier_Not a Member_sum	0.086373
marketing_channel_Direct_sum	0.083085
total_visit_pages_sum	0.028937
loyalty_tier_Base Member_sum	0.024421
avg_page_time	0.017842
total_visit_pages_mean	0.017625
searched_destinations_count_max	0.015363

XGBoost

	gini_importance
customer_type_Existing_sum	0.374207
loyalty_tier_Silver/Gold Member_sum	0.263442
customer_type_New_sum	0.186008
loyalty_tier_Not a Member_sum	0.049755
marketing_channel_Direct_sum	0.044610
loyalty_tier_Base Member_sum	0.029842
platform_App_sum	0.006548
marketing_channel_Meta_sum	0.005500
marketing_channel_Affiliates_sum	0.004874
bkg_confirmation_pages_count_sum	0.003226

Top 10 Drivers of Customer Churn

The XGBoost model's heavy reliance on the top features confirms **loyalty status and customer lifecycle are the primary drivers for intervention**, mandating an immediate focus on **loyalty enrollment** and stabilizing the **high-risk new customer transition**.

Rank	Factors	Gini Importance	Strategic Insight
1	customer_type_Existing_sum	0.374207	**Retention anchor:** The existing customer base is the core foundation of low churn. Protecting this segment is paramount.
2	loyalty_tier_Silver/Gold Member_sum	0.263442	**Maximum loyalty lift:** High-tier status provides the greatest reduction in churn probability. Loyalty works.
3	customer_type_New_sum	0.186008	**High-risk transition:** The period between the first and second booking is the most volatile and high-risk.
4	loyalty_tier_Not a Member_sum	0.049755	**Critical enrollment failure:** Not enrolled in the loyalty program is a churn indicator. Enrollment is a necessary retention step.
5	marketing_channel_Direct_sum	0.044610	**Brand affinity:** Customers who arrive directly (typing the URL) are inherently more secure and brand-loyal.
6	loyalty_tier_Base Member_sum	0.029842	**Mid-tier stability:** Base members are a secure segment, but engagement campaigns are needed to drive tier upgrades.
7	platform_App_sum	0.006548	**Platform stickiness:** Usage of the App platform contributes positively to long-term retention.
8	marketing_channel_Meta_sum	0.005500	**Acquisition risk:** Traffic acquired via Meta channels introduces measurable volatility that requires proactive management.
9	marketing_channel_Affiliates_sum	0.004874	**Acquisition risk:** Affiliate traffic also carries a measurable risk profile requiring strategic funneling.
10	bkg_confirmation_pages_count_sum	0.003226	**Conversion signal:** The total count of successful bookings is a clear signal of customer engagement and value.

3. Key Insights & Strategic Recommendations

Based on the predictive model and factor importance, we should implement a four-pronged strategy focusing on recency, loyalty, channel and platform.

Key Insights

Timing, loyalty, channel, and platform are the dominant drivers of churn

Rencency Threshold

Insight

Churn risk **triples after 90 days** without engagement or booking activity after their last stay.

Implication

Timing is critical. Intervention must be highly automated and occur proactively **before the 90-day acceleration point** to stabilize the customer.

Loyalty Status

Insight

The stability of **Existing** and **Silver/Gold** members anchors retention. The **New Customer** and **Non-Member** segments create the highest churn volatility.

Implication

Enrollment is pivotal. We must convert new customers into loyal members immediately post-booking to stabilize the critical booking transition.

Marketing Channel

Insight

Direct traffic establishes security and brand loyalty. **Acquisition channels** (Meta/Affiliates) introduce a measurable risk profile.

Implication

Mitigate risk. Immediately score all acquisition traffic and prioritize its funnelling into the **loyalty enrollment** funnel to protect ROI.

Platform Value

Insight

App users are 15% more secure than customers using desktop or mobile web.

Implication

App stickiness is a retention lever. Driving platform migration for high-risk web users is a direct retention action.

Strategic Recommendations

Shift to proactive, risk-scored intervention across four key levers

Rencency

Deploy Automated 90-Day Intervention

- Intercept customers before the **90-day risk acceleration point**.
- Use the predictive score (e.g. churn risk score >0.5) to ensure **personalized offers** only hit customers truly at risk.

Loyalty

Launch Mandatory Loyalty Conversion Funnel

- Secure commitment by establishing **enrollment immediately after the first transaction**.
- **Target Not a Member** and **New Customer** segment to eliminate risk.
- Offer a tangible 'First Member Perk' (e.g. £20 credit) only accessible by completing **free loyalty sign-up** post-booking.

Channel

Manage and Funnel Acquisition Risk

- Immediately score and divert high-risk Meta/Affiliates traffic into the **loyalty enrollment funnel** to protect acquisition ROI.
- Recognize direct traffic loyalty with **exclusive member perks** (e.g. early access) to reinforce brand affinity.

Platform

Real-Time App Migration

- Flag **high-risk, non-app users** using the churn score.
- Integrate **real-time scoring** to drive platform migration incentives.
- Target them with a direct and **exclusive app-only incentive** (e.g. Get an extra 10% off your next stay when you download and book via the app).

4. Next Steps

We have successfully quantified the risk and identified the core levers for retention. The time for proactive intervention is now.

Implementation Plan & Approval

Technical Deployment

Embed the XGBoost Risk Score API into the CRM systems immediately. This is the foundational step required to generate the real-time, high-precision scores that will power all targeted campaigns.

Loyalty Foundation Launch

Launch the mandatory loyalty enrollment funnel immediately post-booking for all new customers. This is essential to eliminate the single highest risk factor: non-member status.

Pilot Campaign Initiation

Initiate the first-wave pilot campaigns: the automated 90-day intervention campaign (recency), the app migration campaign (platform), and channel risk funneling (channel), to validate ROI on the targeted strategy.

Call to action

- Protecting customer lifetime value is a near-term ROI opportunity.
- Approving these immediate steps allows us to reduce measurable customer attrition starting next quarter.

Any questions,
please reach out!