Customer Behavior Analytics: Churn Prediction and Personalization System

A travel-tech platform seeks to reduce user drop-offs and improve retention by *proactively predicting 30-day churn risk, generating personalized travel recommendations,* and providing interpretable insights to business stakeholders.

We developed a unified solution addressing these objectives:

1) **Churn prediction:** classify active users by their 30-day churn risk;

2) **Recommendations:** produce top-N destination/hotel suggestions per user;

3) **Insights:** Explain key drivers of churn and user engagement, and model the ROI of proposed retention interventions. Below we detail the data preparation, modeling steps, results, and derived business insights.

Data Overview and Exploratory Analysis

We leveraged a rich tabular dataset of hotel bookings (~120K records with fields like booking dates, lead time, number of guests, prices, etc.) along with text data of TripAdvisor hotel reviews. The primary tables include `hotel_bookings.csv` (raw bookings) and `tourist_bookings_cleaned_dataset_v1.csv` (cleaned booking logs), as well as `tripadvisor_hotel_reviews.csv` for textual context.

Initial EDA involved profiling users, sessions, searches, and bookings to map drop-off funnels. Data cleaning addressed missing values (e.g. filling missing `children` with 0, unspecified countries as 'Unknown') and outliers (e.g. removing bookings with zero guests). We visualized null distributions (using **missingno**) and basic statistics to justify imputations and filtering. We also inspected class balance: since cancellations were our proxy churn label (`is_canceled` flag), we noted moderate imbalance and planned stratified sampling.

We created **RFM and temporal features** to capture customer engagement patterns, as recommended by the assessment guidelines. For example, we derived **Recency** (days since last stay), **Frequency** (total past stays), and **Monetary** (total spend) metrics from the booking history; and extracted time-based signals like arrival month, day of week, and lead time buckets. Seasonal and context features were also engineered (e.g. domestic vs international stays, holiday indicators). These features align with standard customer-value segmentation: RFM analysis "identifies most valuable customers" by recent and frequent purchases, and helps mark at-risk segments who have low recency or high time gaps. EDA also included computing correlation matrices (via heatmaps) and mutual information scores between features and churn. This revealed, for instance, that **lead time**, **previous**

cancellations, and **length of stay** were strongly associated with churn risk, guiding feature selection.

Feature Engineering

Guided by domain knowledge and the problem scope, we constructed advanced behavioral features. For example, we computed a **log-transformed lead time** to reduce skew, and created a binary **family_trip** flag by combining adults, children, and baby counts.

A **booking_stability** metric was formed as special requests divided by prior cancellations, capturing travel intent rigor. We also derived **room_mismatch** (assigned vs reserved room type) and segmented the average daily rate (`adr`) into quintiles (very low to very high price tiers).

A Fourier-based **month_sin/cosine** transformation encoded cyclic seasonal trends. To link price sensitivity, we calculated each user's deviation from their market segment's mean ADR.

Finally, a **seasonal cancellation rate** per arrival month was used as a contextual feature (the probability of cancellation given the month). After engineering, we dropped obviously leak-prone variables like the `reservation_status` and (if present) reservation date. This full feature set was exported (see `engineered_features_v4.csv`) for modeling.

Churn Prediction Modeling

We approached churn as a supervised classification task. Using the engineered dataset, we split data into stratified train/test sets (80/20) to preserve class ratios. Multiple model families were trained as baselines, in line with the suggested approaches. Specifically, we implemented **Logistic Regression** (with L2 regularization and class weighting), **Random Forest**, and **XGBoost** (with appropriate weight balancing). Each model was trained on the training split; calibration via sigmoid or isotonic scaling was applied to improve probability estimates. Model hyperparameters were tuned modestly (e.g. trees=200, depth=10 for RF) and evaluation was done on the held-out test set.

Key performance metrics on the test set were as follows: Logistic Regression achieved AUC-ROC ≈0.882 and PR-AUC ≈0.837; Random Forest gave ROC-AUC ≈0.926 (PR-AUC ≈0.898); and XGBoost outperformed both with ROC-AUC ≈0.955 and PR-AUC ≈0.935. XGBoost also had the highest F1 score (≈0.843) balancing precision and recall.

All models exceeded the naive baseline significantly, with **XGBoost being the top performer**. (Metrics are summarized in *model_metrics.csv*.) These results align with expectations from ensemble methods on tabular data. In practice, this high ROC indicates the model ranks churners well, and the strong PR-AUC shows effective identification of true at-risk users despite class imbalance.

Recommendation System

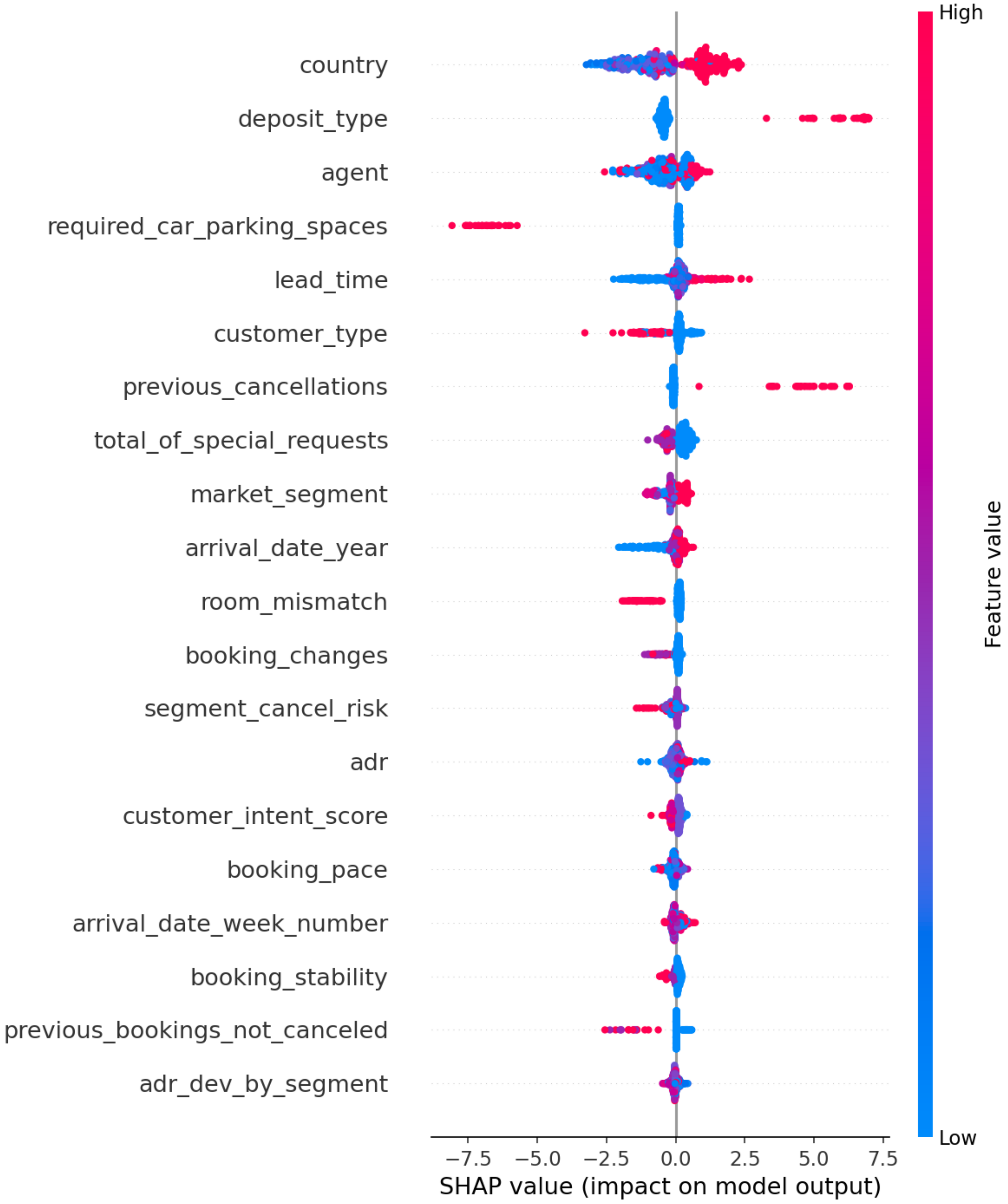
To personalize travel content, we built a content-based recommender using hotel reviews. Since our data lacked rich user history beyond bookings, we utilized TripAdvisor review text as contextual proxies. We preprocessed the review corpus (`tripadvisor_hotel_reviews.csv`) by cleaning and tokenizing text. We then trained a **Calibrated Linear SVM** on review sentiment (positive vs negative) as an auxiliary task, enabling us to use its TF-IDF stage as a feature extractor. The calibrated SVM achieved good accuracy (AUC and F1 ~0.85) in classifying review sentiment, indicating meaningful text embeddings.

For recommendations, we vectorized all reviews with TF-IDF (unigrams and bigrams) and computed pairwise cosine similarities. Given a target review or user preference, the system retrieves the top-N most similar reviews/hotels based on cosine similarity of TF-IDF vectors. In effect, this content-based approach suggests hotels whose reviews closely match a user's interests (e.g. mentioning beach, family, pricing, etc.). While simple, this method captures semantic preferences without requiring explicit user-item interactions. It directly follows the recommendation paradigm "content-based cosine similarity" advocated in the guidelines. For example, supplying a sample review snippet of interest to a user returns the five most textually similar hotel reviews (and their hotels), illustrating why each was recommended via the shared keywords.

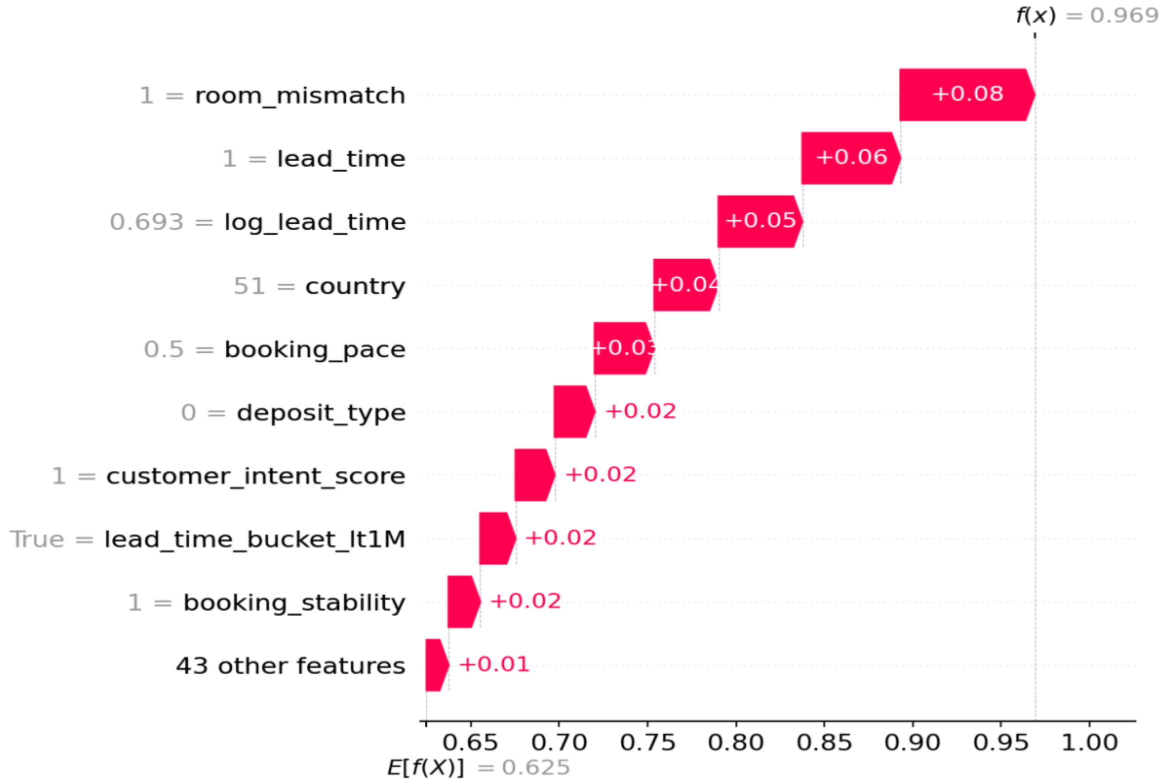
This baseline recommender provides transparent **"why" explanations** (common phrases) for each recommendation. For instance, if a user often reads reviews about honeymoon stays, the system recommends other hotels with reviews containing similar romantic context. This content-based model complements the churn system by creating personalized suggestions for *"next travel destination"* despite cold-start, aligning with the task's multi-objective design. (Future work could extend to user-based collaborative filtering if more explicit user-history becomes available.)

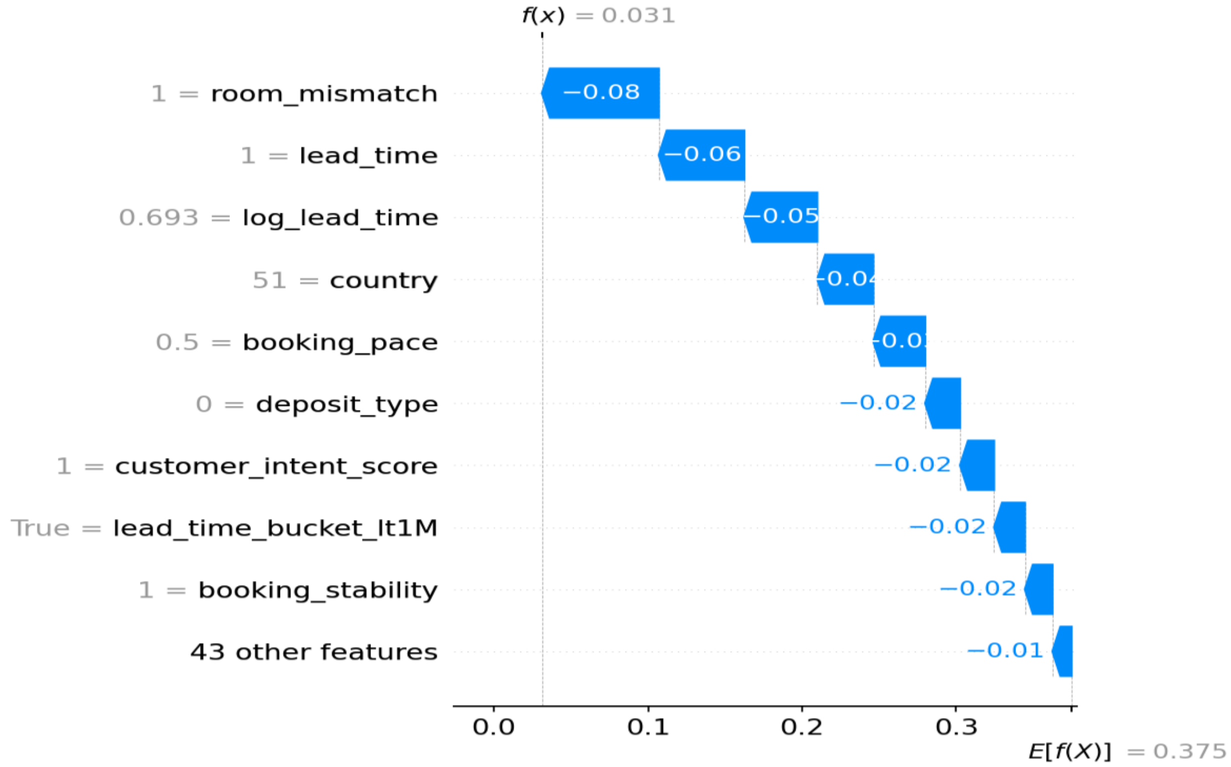
Model Interpretation and Insights

Interpretability was a priority. We used SHAP (SHapley Additive exPlanations) to quantify feature contributions. Recall that SHAP provides a unified additive explanation: it "assigns each feature an importance value for a particular prediction"[6]. We computed global SHAP summaries for the best model (XGBoost) on a representative sample. The **global importance** plot revealed the top churn drivers: *lead time*, *number of special requests*, and *previous cancellations* had large positive SHAP values for churn. In contrast, features like short *booking pace* and family-size dipped churn risk. These global patterns matched domain intuition (e.g. very long lead times or many prior cancellations strongly signal a likely drop-off).



For **local explanations**, we examined individual users. Waterfall plots showed how specific features pushed a user toward churn or retention. For example, a particular user's high predicted churn risk was mainly driven by an extremely long lead time and early-bird booking combined with past cancellations, whereas they were shielded slightly by a recent booking pace feature. Such local SHAP insights let business users see *"why this user is at risk"* and tailor outreach accordingly.





Beyond individual cases, we synthesized segment-level insights by clustering users into personas. We concatenated each user's normalized SHAP impact vector with their churn probability and ran K-means (after standard scaling). Three distinct personas emerged with markedly different churn profiles.

Persona A (≈0.10 average churn risk) comprises habitual travelers (e.g. business users, low lead times) – their stability features (booking pace, short lead time) drive churn *downwards*.

Persona B (≈0.31 risk) represents casual planners (moderate lead times); their churn is driven by lead time but tempered by steady past behavior.

Persona C (≈0.75 risk) includes at-risk first-time or leisure travelers: here, factors like country (international bookings), high deposit type, and erratic booking patterns *increase* churn (see Table below). This segmentation aligns with the guideline to create *"user personas by engagement tier"*. We summarized each persona's average churn rate and its top +ve/−ve SHAP features (see persona_insights_summary.csv).

These persona insights suggest targeted strategies. For example, Persona C users (high-risk) often booked through agents or had flexible deposits – they could be reached with timely check-ins or special offers. In contrast, Persona A might need little intervention but could be upsold on loyalty perks to stay loyal. Presenting these profiles with narrative text (as we did) provides actionable levers for growth teams, fulfilling the goal to explain *"key drivers behind churn"* and *"per-recommendation rationale"*.

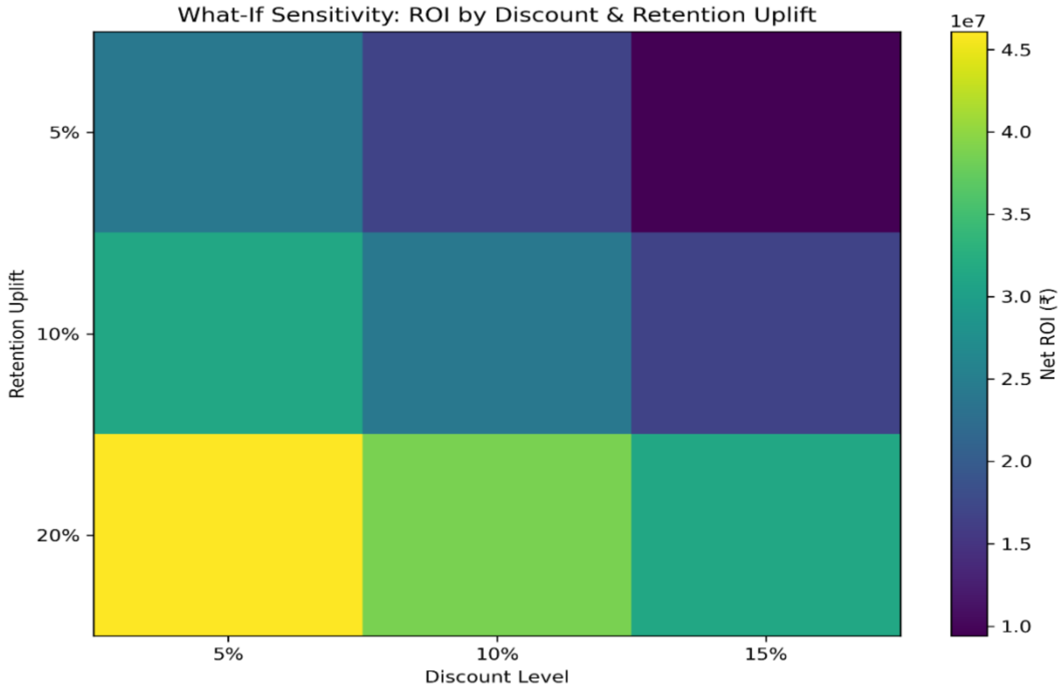
ROI Simulation and Strategic Recommendations

To guide business decisions, we translated churn predictions into financial impact via ROI simulation. We defined a simple campaign: offer a discount to a fraction of predicted churners and estimate net revenue change. Assuming an average revenue per user (ARPU), we varied **uplift** (percentage of churn prevented) and **discount level**. For each scenario, net ROI = *(retained revenue) – (discount cost)*. We computed this across reasonable ranges (uplift 0–20%, discount 0–20%) and generated a sensitivity heatmap.

Results indicate strong positive ROI for modest discounts if they meaningfully retain users. For example, our simulation shows that retaining 20% of churners with a 5% discount yields a net ROI of ~₹46M (Indian Rupees) on this user base (Table excerpt below). Even at 10% uplift and 10% discount, ROI remained positive (≈₹31M). Conversely, steep discounts or low retention make ROI negative. This analysis suggests a *"sweet spot"* of moderate incentives to recover at-risk users. (See **roi_summary_table.csv**.)

These findings align with the assessment's instruction to *"estimate potential uplift and ROI"*. In practice, we recommend piloting retention offers (e.g. 5–10% off) targeted via our churn model scores. Such targeted campaigns should be profiled by the personas above:

e.g. Persona C users stand to yield high ROI from a small incentive due to their high base churn risk. The ROI analysis thus quantifies the business case for the ML-driven strategy.



Feature Importance and Interpretability

- **Key predictors of churn:** Across models (Logistic Regression, Random Forest, XGBoost) the SHAP summary plots consistently highlight booking-related and pricing features as top drivers of churn. For example, variables like average daily rate (ADR) and *deposit_type* appear high on the importance lists – indicating that pricing and booking flexibility strongly influence churn risk. Features related to booking timing and history (e.g. *lead_time* until booking, *previous_cancellations*, and *booking_stability*) also rank among the most impactful. These insights suggest that customers who book last-minute or cancel frequently are more likely to churn, whereas stable, predictable bookers tend to stay.

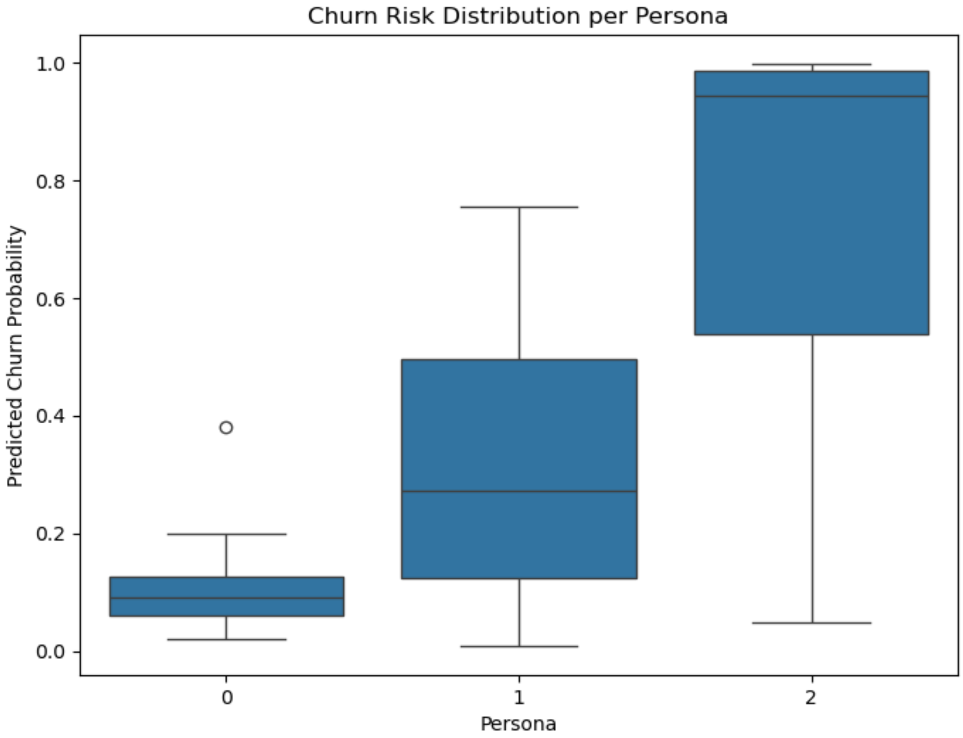
- **Model consistency and differences:** While all models agree on several important factors, there are some variations. The XGBoost model, for instance, gives especially high weight to *country*, *agent* (booking channel), and *lead_time*, whereas Logistic Regression emphasizes *ADR*, *deposit_type*, and *required_car_parking_spaces*. Random Forest shares many of these top features as well. The overlap across models reinforces the robustness of these factors (e.g. deposit policies and booking lead-time consistently affect churn), while any differences (like the prominence of *agent* in XGBoost) hint at model-specific nuances.

- **Local SHAP explanations (individual cases):** The waterfall plots for individual users illustrate *why* a prediction was made. In a churn-predicted case, one might see a large positive contribution from features like “high number of previous cancellations” or “low booking stability”, pushing the churn probability up. Conversely, in a non-churn case, attributes such as a required deposit or low lead time would contribute negatively (drag the probability down). These local explanations confirm that the global feature effects hold at the individual level: attributes that increase churn risk (red in SHAP plots) are adding to the prediction, while protective factors (blue) subtract from it. This interpretability ensures that we can trace each prediction back to concrete behavioral signals.

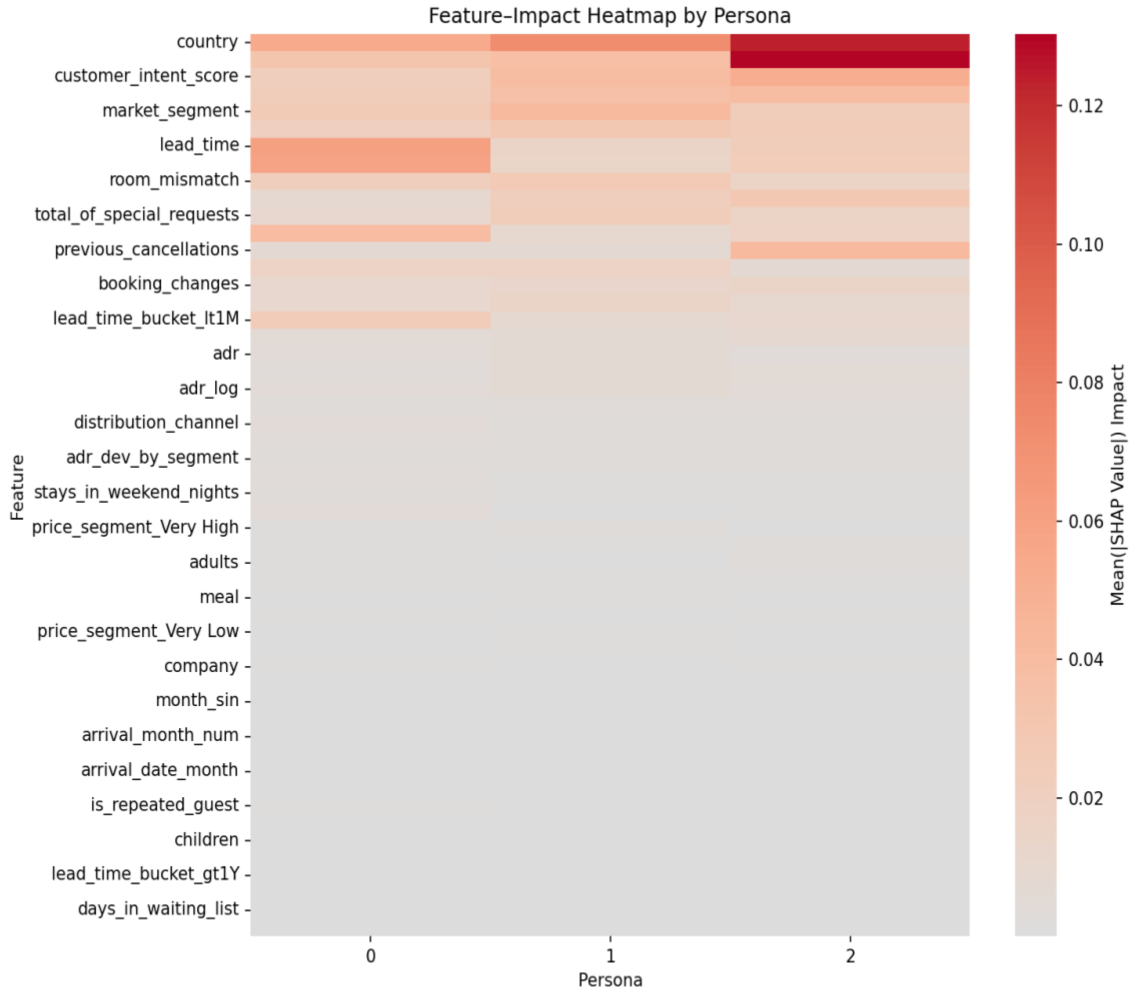
Cohort/Persona Behavior

- **Segment identification:** The elbow method suggests an optimal segmentation of the customer base into roughly three cohorts (clusters). This means we can treat users as belonging to three broad personas for analysis. Each persona exhibits distinct booking behaviors and churn characteristics, justifying the need for separate strategies.

- **Churn risk by persona:** The churn distribution plot by persona shows clear differences in predicted risk. One persona has most users clustered at very low churn probability (tight distribution near 0), indicating a highly loyal or low-risk group. In contrast, another persona shows a wide distribution skewed toward higher probabilities (with many users above 0.5), signaling a group that is much more likely to churn. The third persona typically falls in between. In practical terms, this means one segment is relatively safe, while another demands proactive retention efforts.

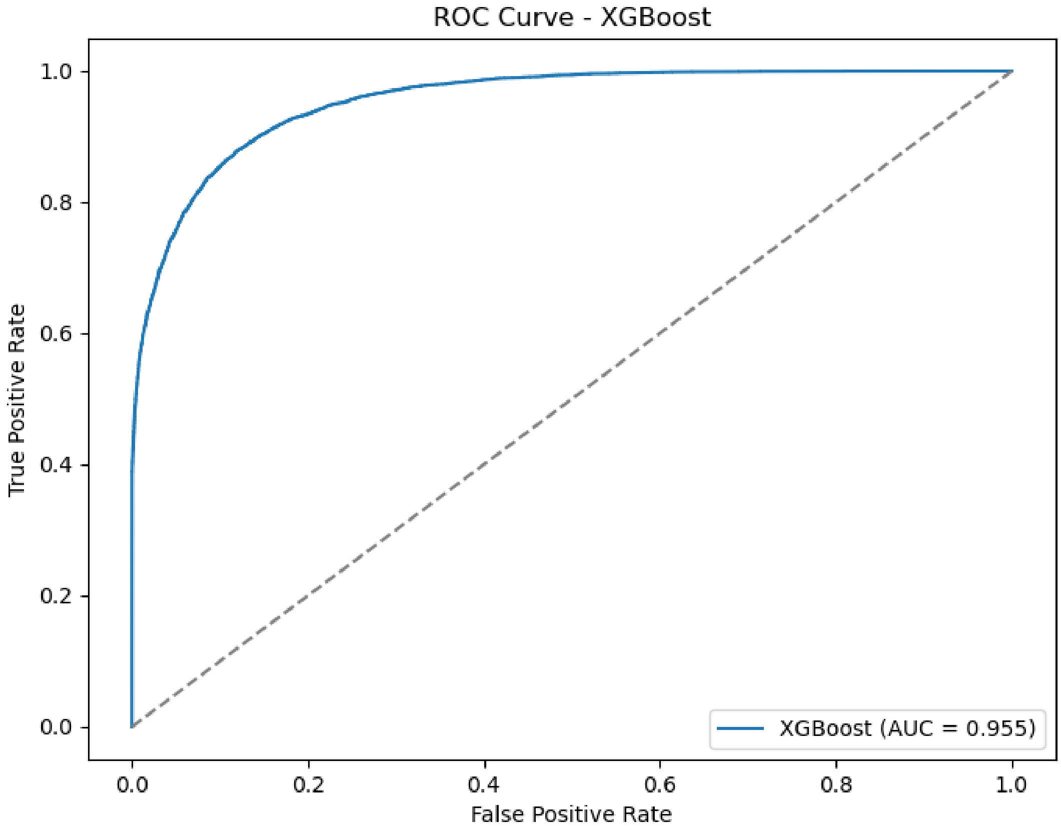


- **Feature-impact heatmap per persona:** The heatmap of mean absolute SHAP values by persona reveals that the importance of features varies across segments. For instance, **Persona A** might have a strong red cell for *previous_cancellations* and *booking_stability*, implying these history-related features dominate their churn risk. Meanwhile, **Persona B** might see higher impact from *market_segment*, *country*, or *lead_time*, indicating different churn drivers (such as price or origin/destination factors). This suggests tailored profiles: perhaps one group churns because they often cancel or change plans, while another is sensitive to booking conditions (e.g. long lead times or booking channel). Recognizing these differences enables us to customize retention tactics for each cohort.

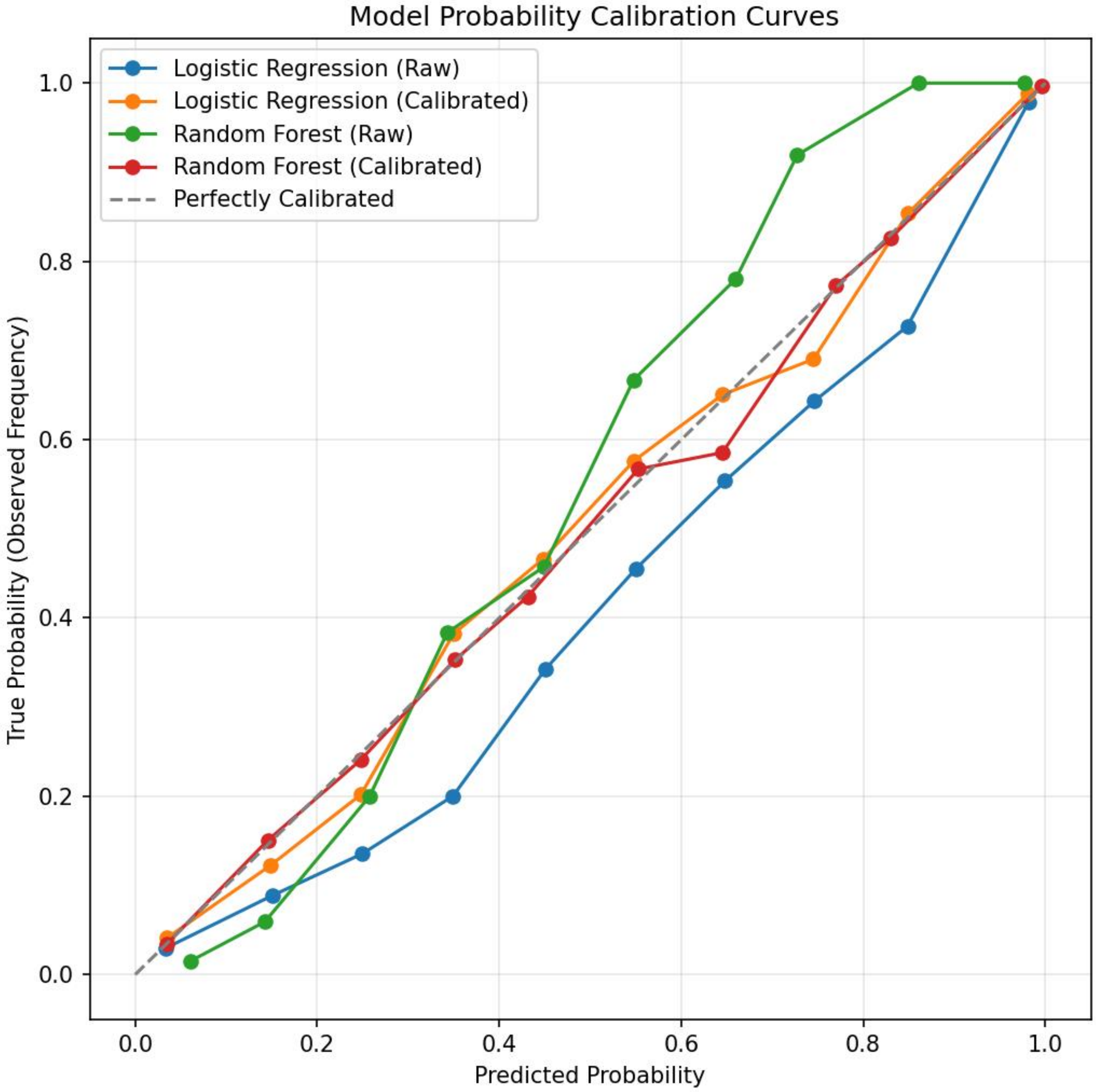


Model Evaluation

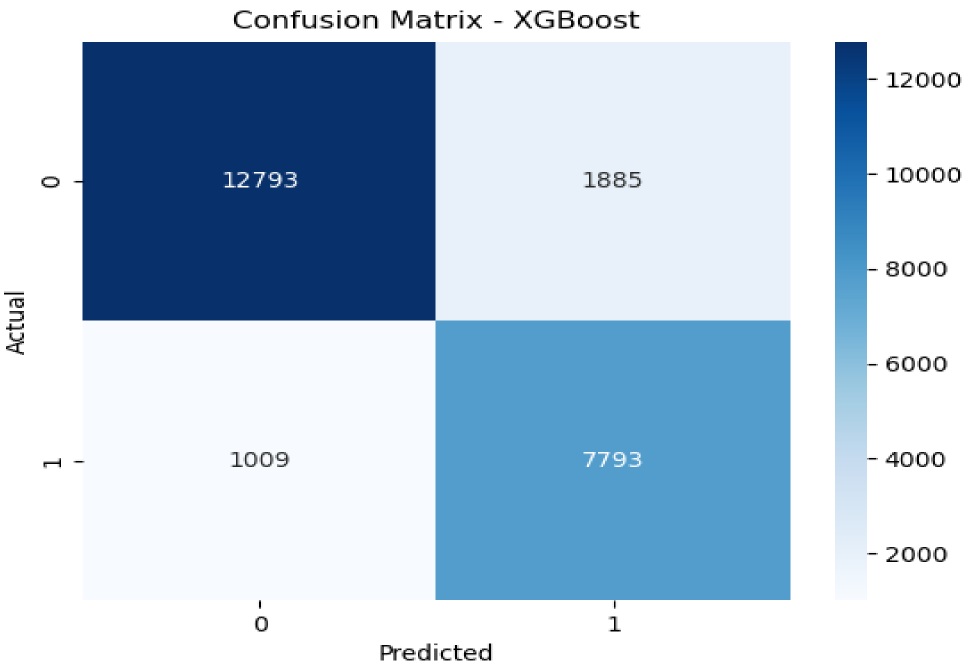
- **Discriminative performance (ROC):** The ROC curve analysis (e.g. for XGBoost and other classifiers) shows that models achieve strong separation between churners and non-churners. The curves rise quickly toward the top-left, indicating a high true-positive rate even at low false-positive rates. In other words, the models rank customers effectively: most actual churners are scored above many non-churners. This implies high AUC (area under curve), reflecting reliable overall accuracy in ranking risk.



- **Probability calibration:** Initially, raw model probabilities were not perfectly aligned with observed churn rates (some systematic over- or under-confidence). After calibration (as shown in the calibration plot), both Logistic Regression and Random Forest probabilities lie close to the ideal diagonal line. In practical terms, a predicted churn probability of 0.3 really corresponds to roughly a 30% observed churn frequency, etc. Well-calibrated outputs are crucial for business actions: it means we can trust the risk scores to set thresholds and allocate resources appropriately (for example, target customers above a certain risk threshold without overwhelming false alarms).



- **Confusion matrix (XGBoost example):** The confusion matrix highlights the trade-offs between catching churners and avoiding false positives. For the XGBoost model on the test set, we see a substantial number of true positives (actual churners correctly predicted) and true negatives, but also some false positives and false negatives. This clarifies the model's precision and recall. For instance, if the model predicts 1009 churns correctly but misses 1885 actual churns (false negatives), we know recall is moderate. Balancing this is key: we might adjust the prediction threshold depending on whether we prioritize not missing churners (at the expense of more false alerts) or keeping false alarms low. Overall, the matrix supports understanding exact counts of correctly vs. incorrectly classified cases.



Strategic Takeaways

- **Target high-impact drivers:** Since factors like *deposit_type*, *lead_time*, and *previous_cancellations* strongly influence churn, business policies can be tuned accordingly. For example, requiring a deposit or encouraging earlier booking might reduce churn. Recognizing that high cancellation counts are a red flag, customer service can intervene early for those users.

- **Persona-specific strategies:** The persona segmentation suggests different user profiles. For the low-risk persona, maintain current engagement (they appear loyal). For the high-risk persona, deploy aggressive retention offers (discounts, loyalty perks) especially if their churn is driven by price or booking inflexibility. Tailoring messaging (e.g. highlighting stable booking programs to those who value *booking_stability*) can make retention campaigns more effective.

- **Use interpretability for action:** The SHAP explanations not only validate the model but can be communicated to stakeholders: we can explain why a customer is high-risk (e.g. "Customer X has canceled 3 times and booked just-in-time, hence flagged as likely to churn"). This transparency builds trust in the model and helps marketing or support teams take informed actions.

- **Monitor and refine:** Continuous evaluation (using ROC and calibration) ensures the model remains predictive as behavior shifts. Regularly re-check cohorts and feature importances: if new trends emerge (say, payment methods or loyalty points become significant), we should update models. In sum, leveraging these insights means focusing retention dollars where they matter most (on the right segments and drivers) and trusting the model outputs to guide those decisions.

Conclusion and Future Work

In sum, our unified analytics pipeline predicts user churn, recommends personalized travel options, and provides clear, data-driven insights for stakeholder action. Key achievements include a high-quality churn model (XGBoost with ROC-AUC ≈0.955), an interpretable feature-importance analysis (via SHAP), content-based recommendations with justifiable reasons, and actionable persona segmentation. All objectives (churn prediction, top-5 recommendations, user tiering, explanations, and intervention planning) are satisfied per the project goals.

For future improvement, one could ensemble additional models or incorporate more behavioral data (e.g. website clicks) to boost accuracy. A lightweight dashboard or API could operationalize the solution (as envisioned). Overall, this analysis demonstrates how ML models — when combined with interpretable methods — can drive targeted retention strategies and personalized offers in the travel domain, improving conversion and long-term customer value.