Machine Learning Project

Predicting Trip Retention: A Data Science Approach to Travel Agency Churn Rate

Francesco Tumminello
Riccardo Perego
Giovanni Piva

Problem Recap: The Return at the Yeti

Context

- Travel agency specialized in **school trips**, observed a **drop in sales**, especially among **regular clients**.
- Company relies heavily on **loyalty** due to the **niche nature** of their service.

Objective

- **Predict** customer churn using historical client data.
- Use the model to **identify key features** influencing loyalty.
- Ultimately, design **targeted strategies** to retain at-risk clients.

Steps

- Data Preparation
- Feature Engineering
- Train the Models
- Identify a good Strategy

Data Preparation

- Variable type formatting
- Cat. variables encoding
- Missing values (& Filling Strategies)

Missing Values

Table 1: Missing Values (NaN Counts) by Feature

Feature	NaN Count
From_Grade	226
To_Grade	269
Early_RPL	1167
Latest_RPL	36
Initial_System_Date	16
FPP_to_School_enrollment	159
FirstMeeting	596
LastMeeting	596
DifferenceTraveltoFirstMeeting	596
DifferenceTraveltoLastMeeting	596
SchoolSizeIndicator	159

Table 2: Count of Zero Values by Feature

Feature	Zero Count
FRP_Active	221
FRP_Cancelled	822
Num_of_Non_FPP_PAX	27
Cancelled_Pax	417
Total_Discount_Pax	27
Total_School_Enrollment	159
NumberOfMeetingswithParents	596

Missing Values

- 4000+ observations in 50+ columns
- Some columns lack almost 25% of values

- Filling missing values (Examples):
    - ‘From_Grade’ and ‘To_Grade’
    - ‘LastMeeting’

Missing Values: From_Grade + To_Grade

Both From_Grade and To_Grade are missing

- For each **Group_State**, compute the most frequent **(From_Grade, To_Grade) pair**.
- If both are missing → impute using this pair.

Conditional Imputation (One Value Missing)

- If only To_Grade is missing → Use the most frequent **To_Grade** for the given **From_Grade** in the same Group_State.

Missing Values: LastMeeting

Reference-Based Median Shifts

1. **Departure_Date:** reference point, (no missing entries) + forward-looking event in time.
2. For each incomplete date column:
   - Compute **median** time delta (days) between **column** and **Departure_Date**, using rows where both dates are available.
3. Fill missing values by **subtracting** the median delta from Departure_Date.

This method **preserves** realistic **temporal relationships** between events

New Columns

Table 1: Engineered Time-Based Features

Feature Name	Construction Formula
Days_from_Initial_System_to_Departure	Departure_Date - Initial_System_Date
Days_from_Latest_RPL_to_Departure	Departure_Date - Latest_RPL
Days_from_FirstMeeting_to_Departure	Departure_Date - FirstMeeting
Days_from_LastMeeting_to_Departure	Departure_Date - LastMeeting
Days_from_Deposit_to_Departure	Departure_Date - Deposit_Date
Days_from_Departure_to_Return	Return_Date - Departure_Date
Months_to_Departure	(Departure_Date - Deposit_Date) * 12 + (Departure_Date - Deposit_Date)
Days_between_Meetings	LastMeeting - FirstMeeting
Days_between_System_and_FirstMeeting	FirstMeeting - Initial_System_Date
Days_between_System_and_LastMeeting	LastMeeting - Initial_System_Date

Table 2: Engineered Business Logic Features

Feature Name	Construction Formula
Cancellation_Rate	Cancelled_Pax / (Total_Pax + Cancelled_Pax)
Revenue_per_PAX	SPR_Group_Revenue / Total_Pax
Deposit_Ratio	Tuition / Total_Pax
Insurance_Cancellation_Rate	FRP_Cancelled / FRP_Active
Past_Bookings	Grouped rank by [Program_Code, Group_State] on Initial_System_Date
Initial_Year	Initial_System_Date.year
Initial_Season	Initial_System_Date.month % 12 // 3
Grade_Span	To_Grade - From_Grade
Departure_Season	Departure_Date.quarter

Models

Model	$F_{0.5}$ Score (Threshold = 0.5)	$F_{0.5}$ Score (Threshold = 0.6)
Random Forest	90.16%	91.09%
XGB Classifier	87.45%	87.67%
Logistic Regression	77.88%	78.92%
MLP	87.57%	88.02%
AdaBoost Classifier	77.40%	65.37%
GradientBoost Classifier	85.72%	86.68%

Models - Random Forest

- F(0.5) Score = **91.09%**
- Only **29** False Positives

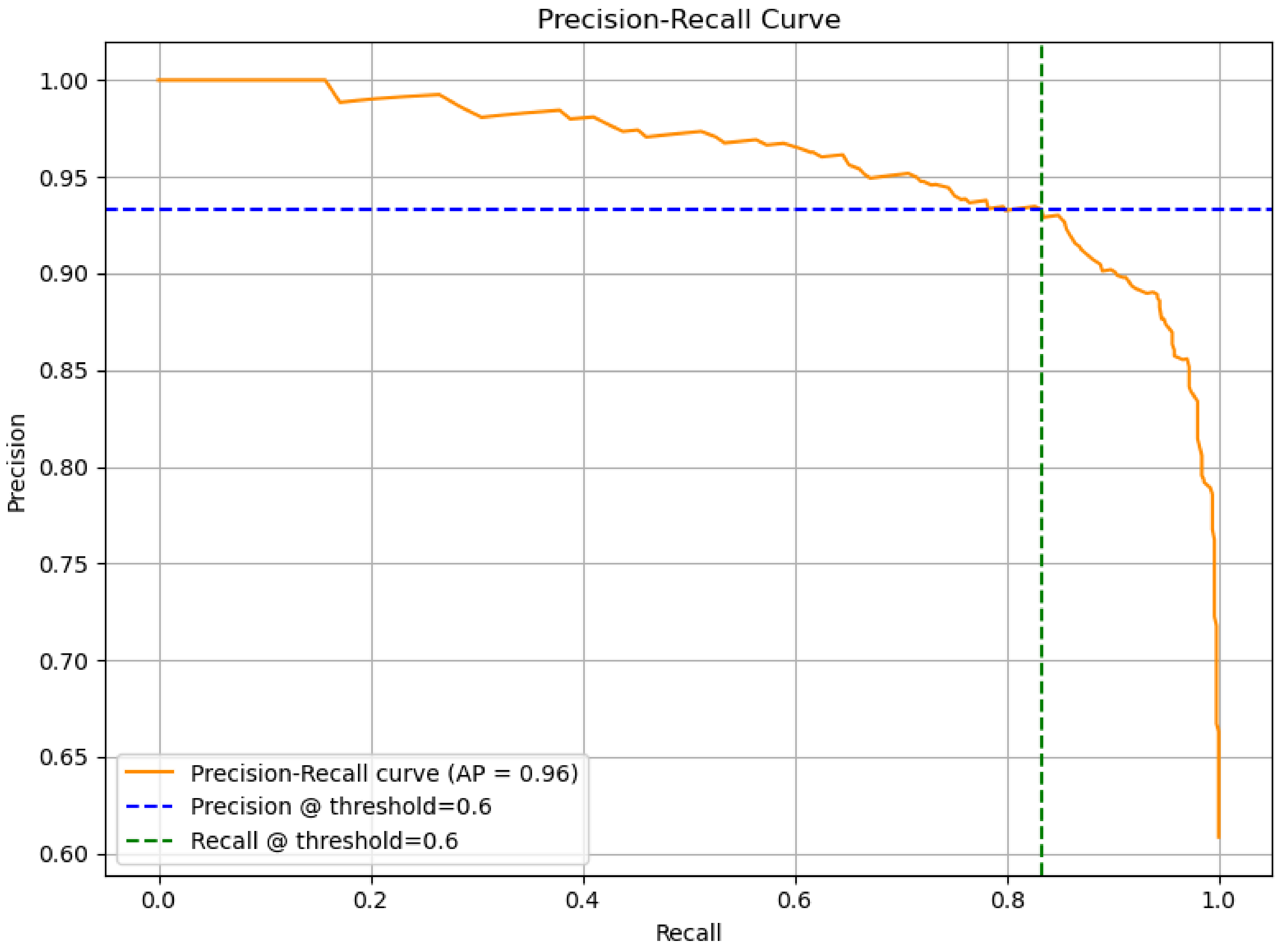
Table 1: Random Forest Performance (Threshold = 0.6)

Metric	**Class 0**	**Class 1**	**Weighted Avg**
Precision	0.77	0.93	0.87
Recall	0.91	0.83	0.86
F1-score	0.84	0.88	0.86

Confusion Matrix

	Predicted 0	**Predicted 1**
Actual 0	294	29
Actual 1	87	415

Models - Random Forest



Models - Random Forest

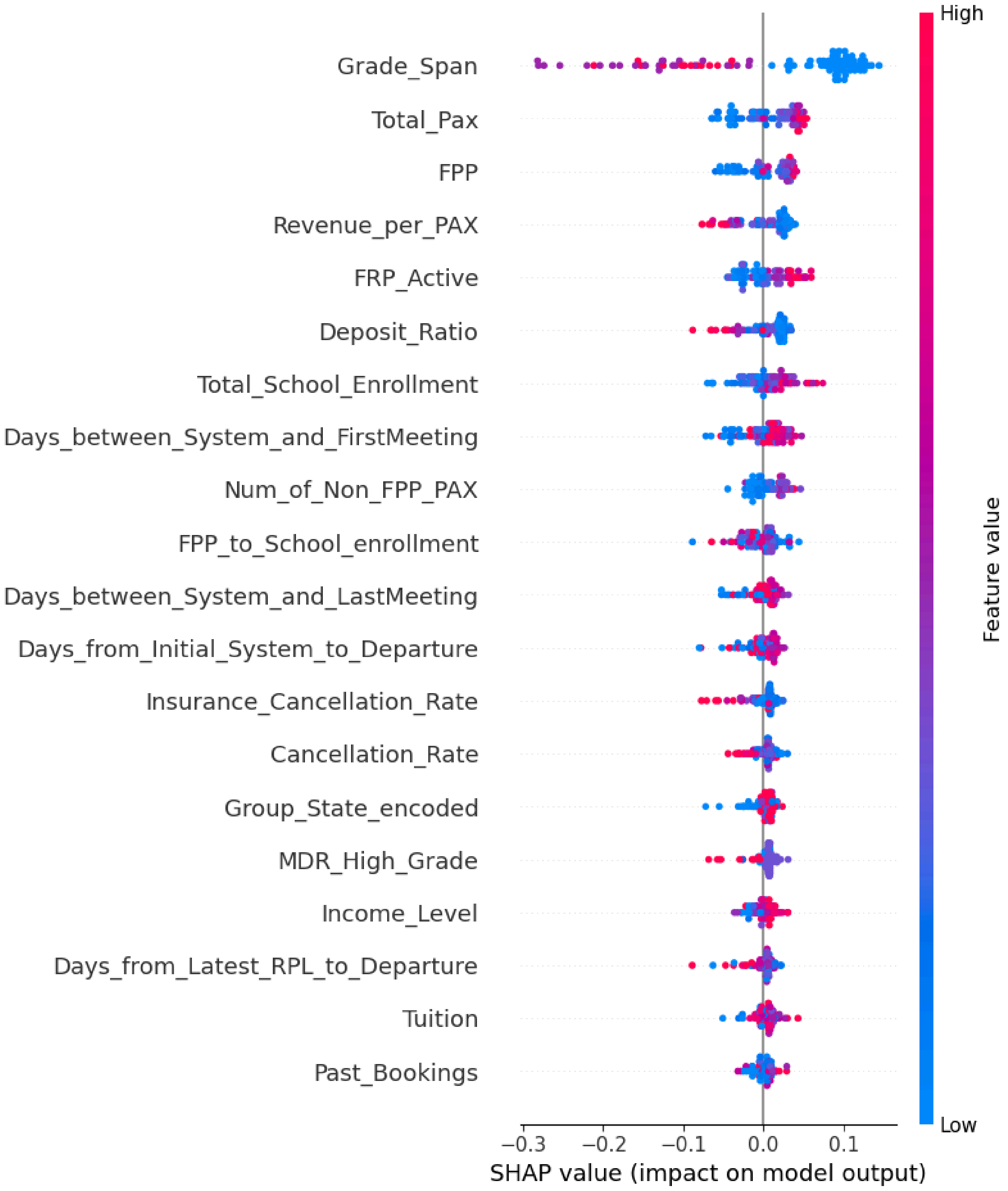
Table 2: Optimal Random Forest Hyperparameters

Parameter	Optimal Value
`n_estimators`	200 trees
`criterion`	Gini impurity
`max_depth`	Unlimited (None)
`max_features`	$\sqrt{\text{total features}}$
`min_samples_split`	2 samples
`min_samples_leaf`	1 sample
`bootstrap`	False (uses full dataset)

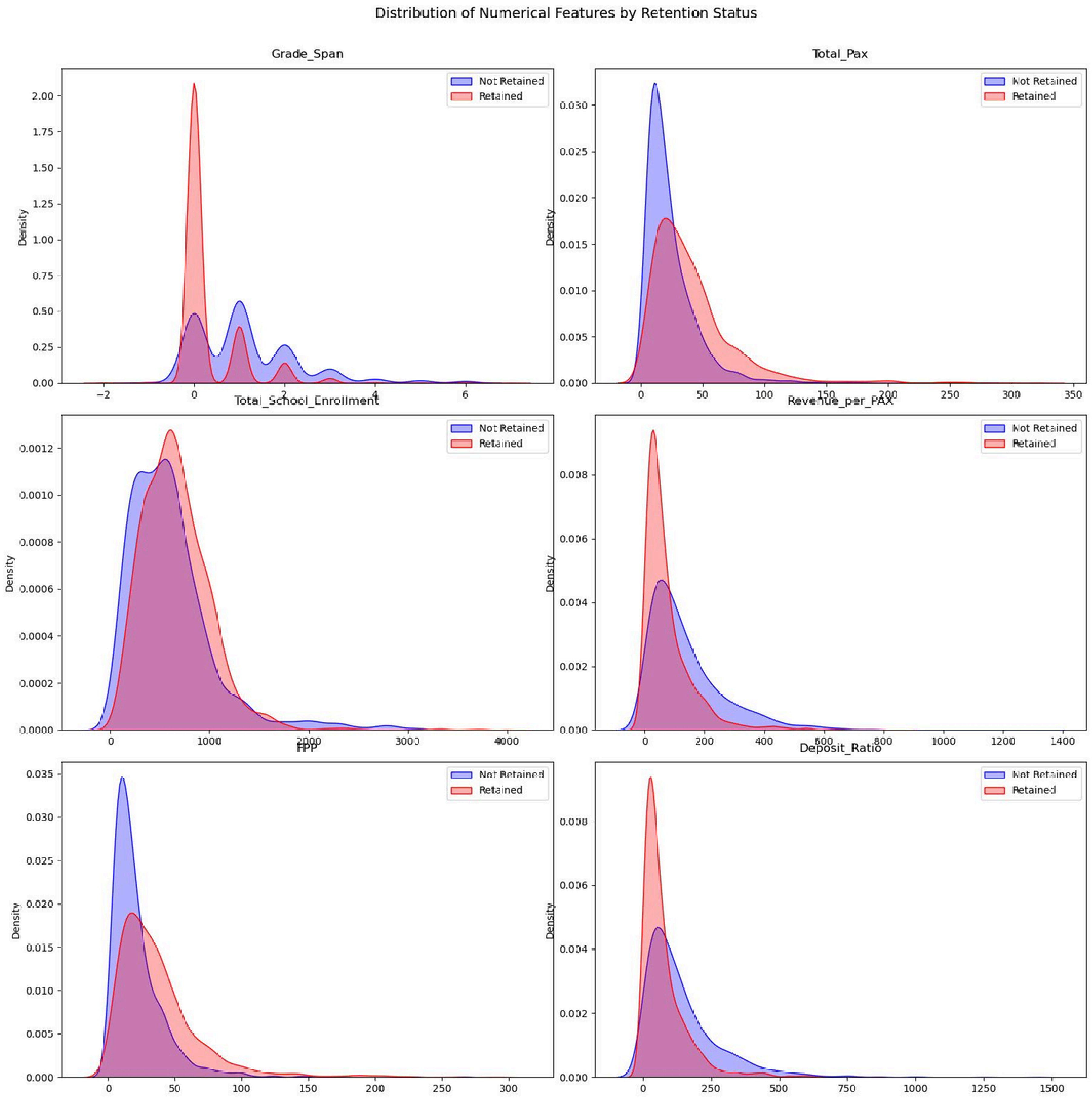
Feature Importance - the Shapley Value

- Shapley values are a game theory concept. They quantify the **contribution** of each feature to a model's output for a specific prediction, fairly distributing the "payout" (prediction difference from the baseline) among all input features.
- Example: if a trip's 'Total_Pax' has a Shapley value of +0.2, it means this feature increases the retention probability by 20% compared to the baseline.

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \left[f(S \cup \{i\}) - f(S) \right]$$



Grade_Span
Total_Pax
FPP
Revenue_per_PAX
FRP_Active
Deposit_Ratio
Total_School_Enrollment
Days_between_System_and_FirstMeeting
Num_of_Non_FPP_PAX
FPP_to_School_enrollment
Days_between_System_and_LastMeeting
Days_from_Initial_System_to_Departure
Insurance_Cancellation_Rate
Cancellation_Rate
Group_State_encoded
MDR_High_Grade
Income_Level
Days_from_Latest_RPL_to_Departure
Tuition
Past_Bookings
High
Feature value
Low
SHAP value (impact on model output)



Distribution of Numerical Features by Retention Status
Grade_Span
Total_Pax
Total_School_Enrollment
Revenue_per_PAX
FPP
Deposit_Ratio
Not Retained
Retained
Density

Business Strategy - Increase Retention

- Goal: Implement strategies to keep current clients.
  - Reward groups with a larger grade span with volume discounts.
  - Reduce deposit requirements, especially for repeating customers.
  - Avoid being pushy about the meeting dates, especially the first.
  - Offer discounts or other benefits to schools who are canceling passengers and insurance.

Business Strategy - Lead Generation

- Goal: Target marketing activities to find new clients.
  - Target schools in high income area.
  - Prioritize marketing towards larger schools - more stable budget.
  - Focus on schools with a narrow grade span, such as those that offer only high school classes.
  - Prioritize schools with a high tuition.

What if...?

- What if reducing FP wasn't actually the best strategy?
- During the analysis we have assumed (like it is common) that FPs cost significantly more than FNs: it costs much more to loose a client rather than implementing strategies to try and keep him.
- But in some particular condition, this does not hold.
- For example, suppose that the travel agency has a very low profit margin and it requires a lot of volume to be profitable. In that case offering discount promotions to retain customers can be quite relatively expensive, leading us to want to prioritize FN minimization rather than FP
- In this cases, we should prioritize **Recall** ⟶ F2 Score.

Random Forest - prioritizing Recall

- F(2) Score = **94.27%**

Table 1: Random Forest Performance (Threshold = 0.4)

Metric	Class 0	Class 1	Weighted Avg
Precision	0.95	0.83	0.88
Recall	0.69	0.98	0.86
F1-score	0.80	0.90	0.86

Confusion Matrix		
	Predicted 0	**Predicted 1**
Actual 0	222	101
Actual 1	12	490

Thank you