CUSTOMER CHURN FORECASTING

From Data Analysis to Business Strategies for Customer Retention

Francesca Gaeta | Machine Learning Project

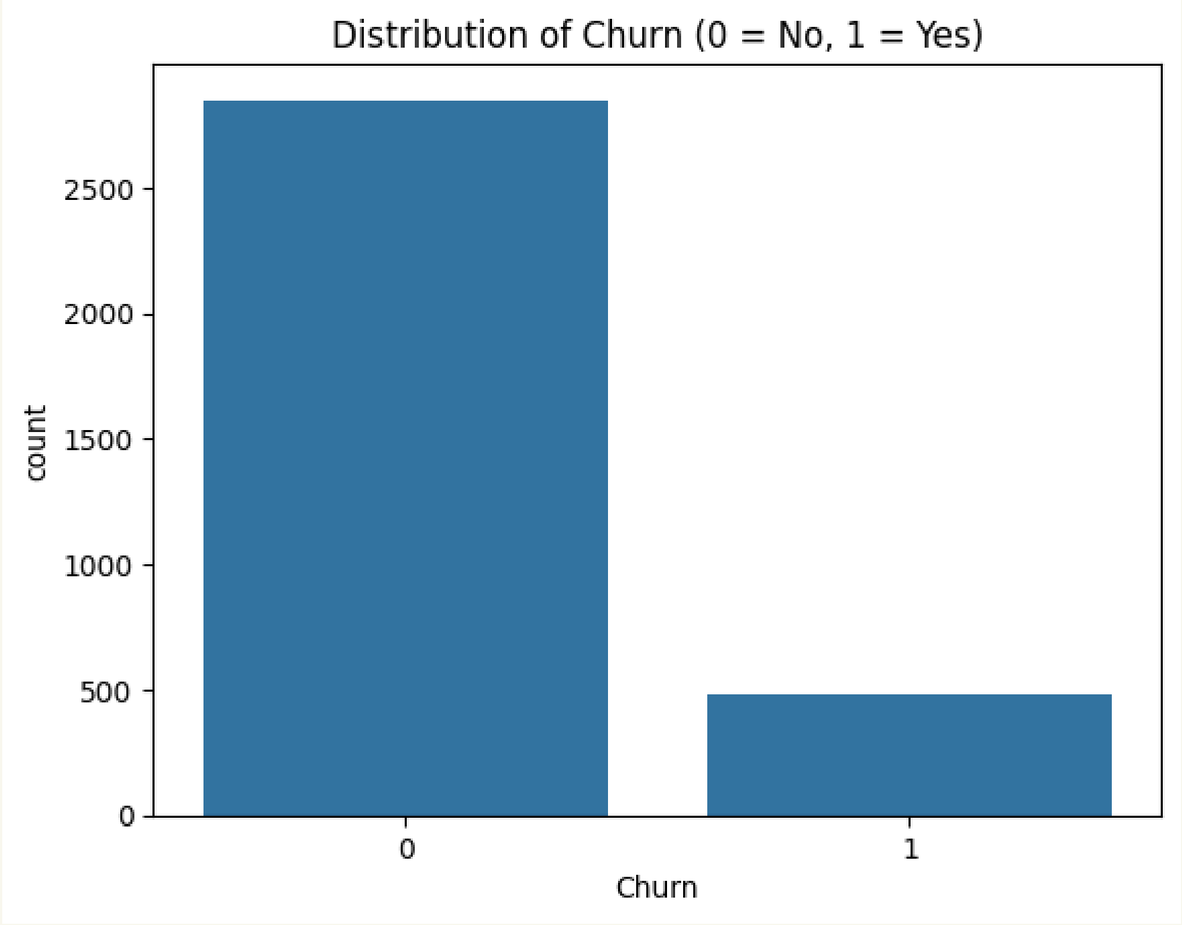
The Data Landscape & the Imbalance Challenge

Dataset Profile: "Telecom-churn" dataset containing customer-level minutes, charges, data usage, tenure and service calls.

The Imbalance Problem: The dataset is heavily skewed with 85.5% Loyal customers versus 14.5% Churners.

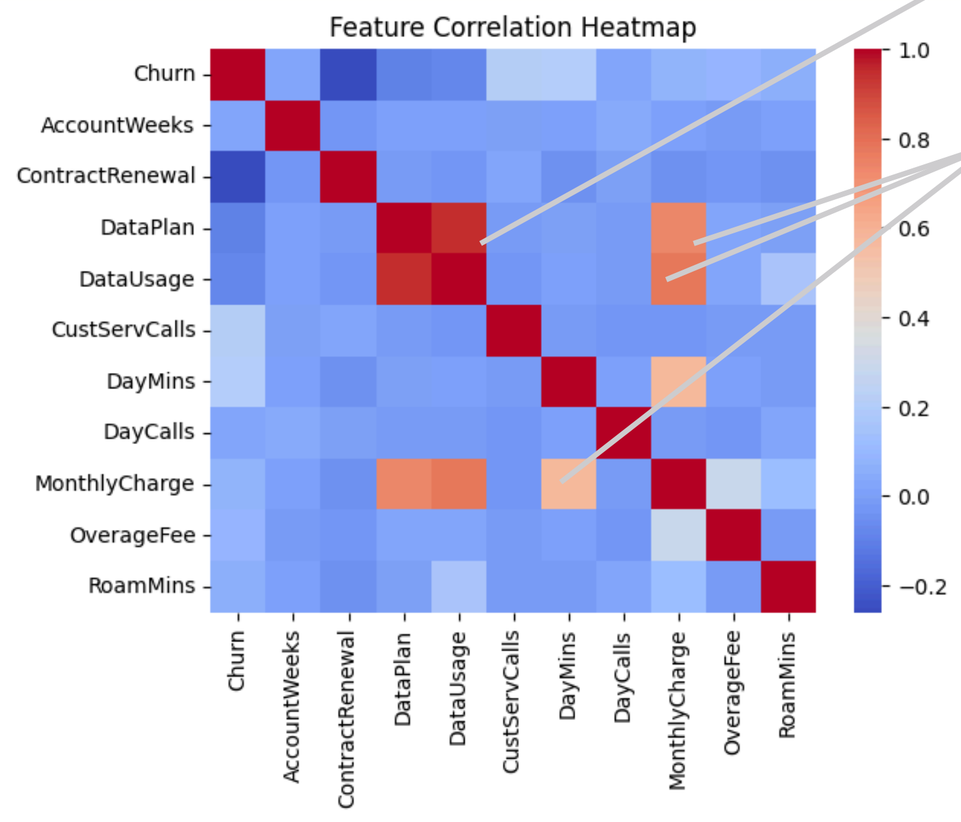
Analytical Consequence: A naive model could predict 'No Churn' for everyone and achieve 85% accuracy while failing the business objective.

Strategy: We prioritize Recall over simple Accuracy (we prefer false positives).



Correlation & The Heavy User Paradox

Core revelation: Contrary to expectations, our churners are highly active users. High engagement (DayMins) correlates with Churn, likely driven by 'Bill Shock' from overage fees.



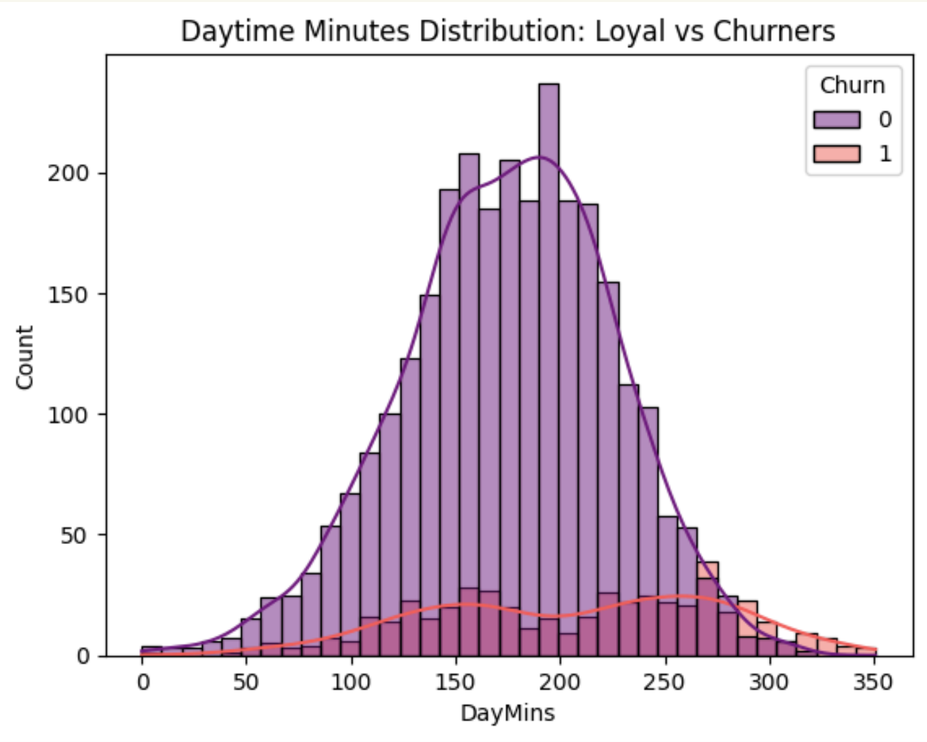
Strong Positive Links: Trivial almost perfect correlation (dark red).

Positive Links: Day Minutes & Monthly Charges are highly correlated.
High usage equals high bills.

CustServCalls shows a 0.20 correlation. While numerically low, it is a critical signal.

Counterintuition: Daytime minutes suggests churn is not caused by a lack of engagement

Avg calls for loyal: **1.45**.
Avg calls for churners: **2.23**.
54% increase = warning sign.



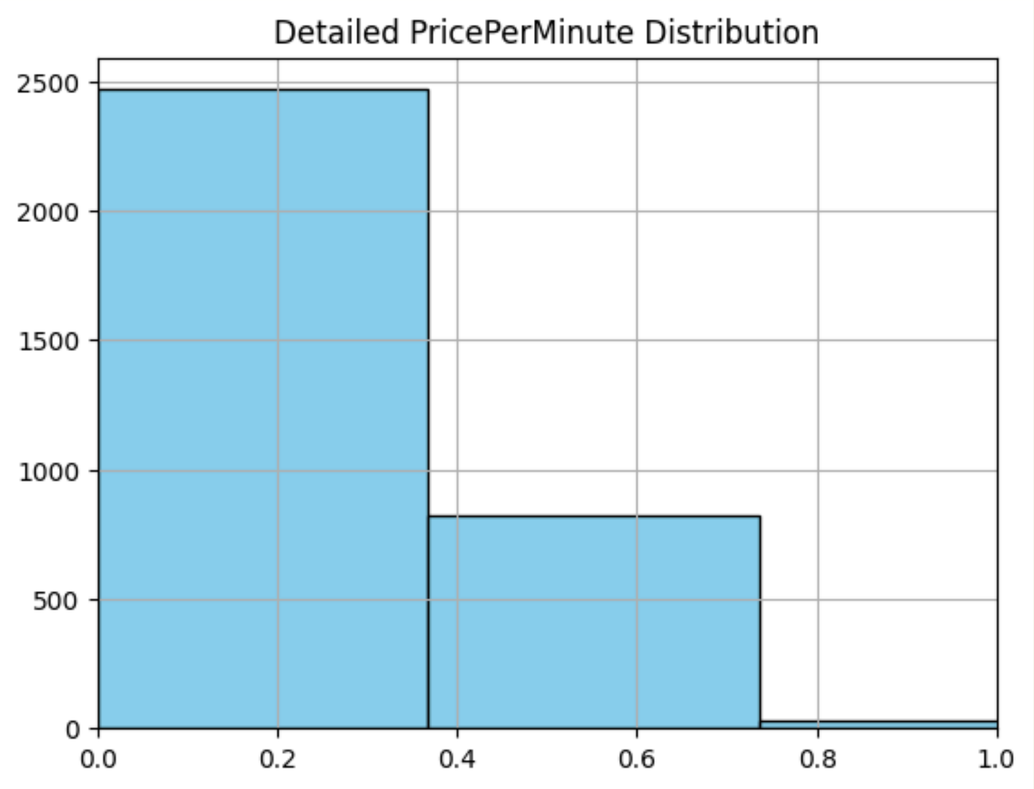
Dimensionality Reduction → Handle Redundancy
Feature Engineering → Define Value for Money

Dimensionality Reduction:

- drop **DataPlan** vs more informative DataUsage
- drop **MonthlyCharge** because correlated with usage variables

Feature Engineering:

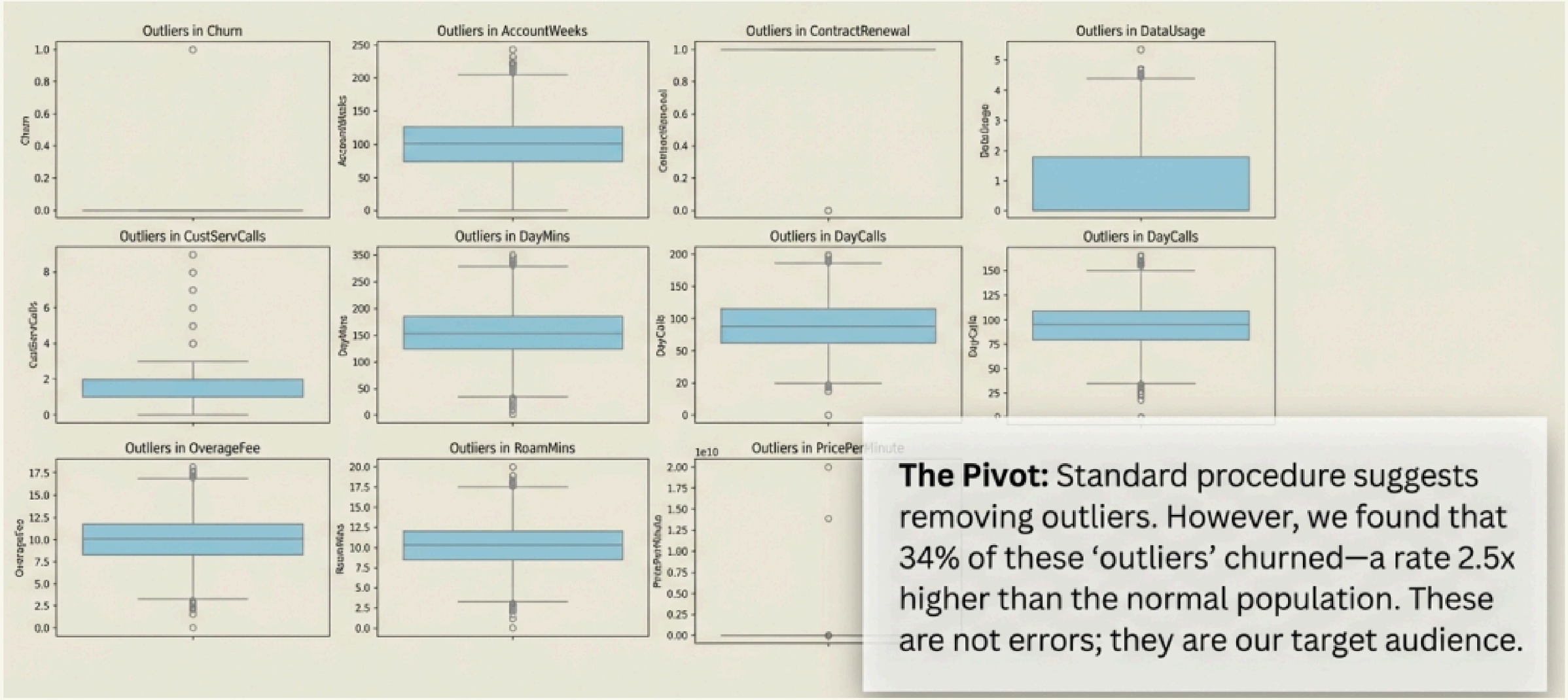
Before dropping MonthlyCharge, we divide it by DayMins to create **PricePerMinute** → preceived value for money



***Histogram**: about density; **bins** = 50 (18.38/50=0.37)*

Outlier Detection

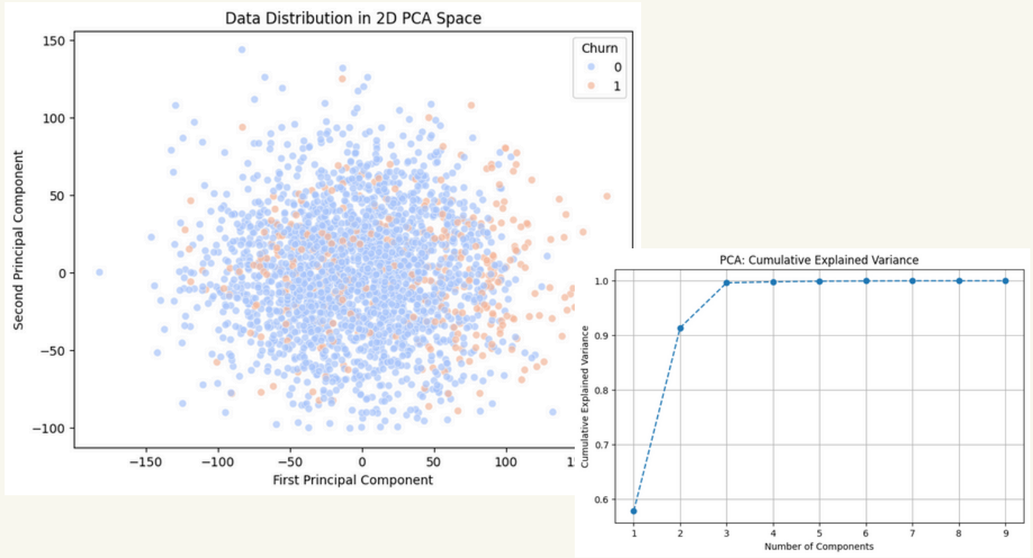
METHODS: IQR and Isolation Forest

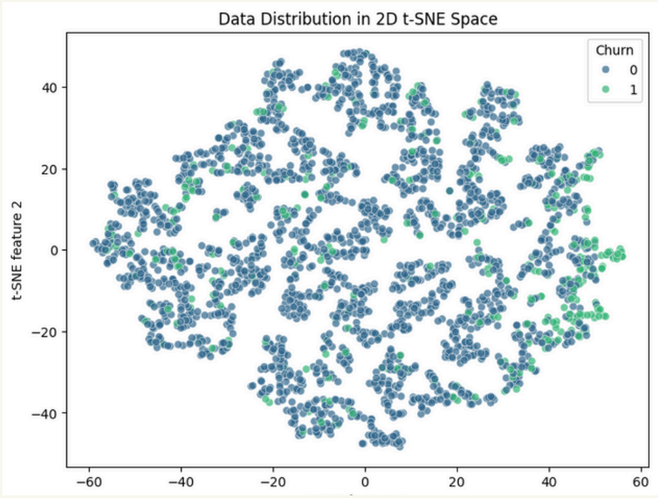


The Pivot: Standard procedure suggests removing outliers. However, we found that 34% of these 'outliers' churned—a rate 2.5x higher than the normal population. These are not errors; they are our target audience.

Complexity Analysis: Why Linear Models Fail

The Test: We utilized PCA (Linear) and t-SNE (Non-Linear) to test separability.





PCA Result: 90% information retention but poor separation.

t-SNE Result: A complex 'cloud' with no distinct clusters (Silhouette score 0.071)

Implication: Simple Logistic Regression achieved only 77% accuracy. The data is non-linear, requiring Tree-Based models (Random Forest / XGBoost) to find the 'hidden boundaries'.

Model Performance: ROC AUC

The Contenders:

Logistic Regression, ANN, Random Forest, XGBoost.

The Winner:

Tree-based ensembles broke the 85% null accuracy barrier.

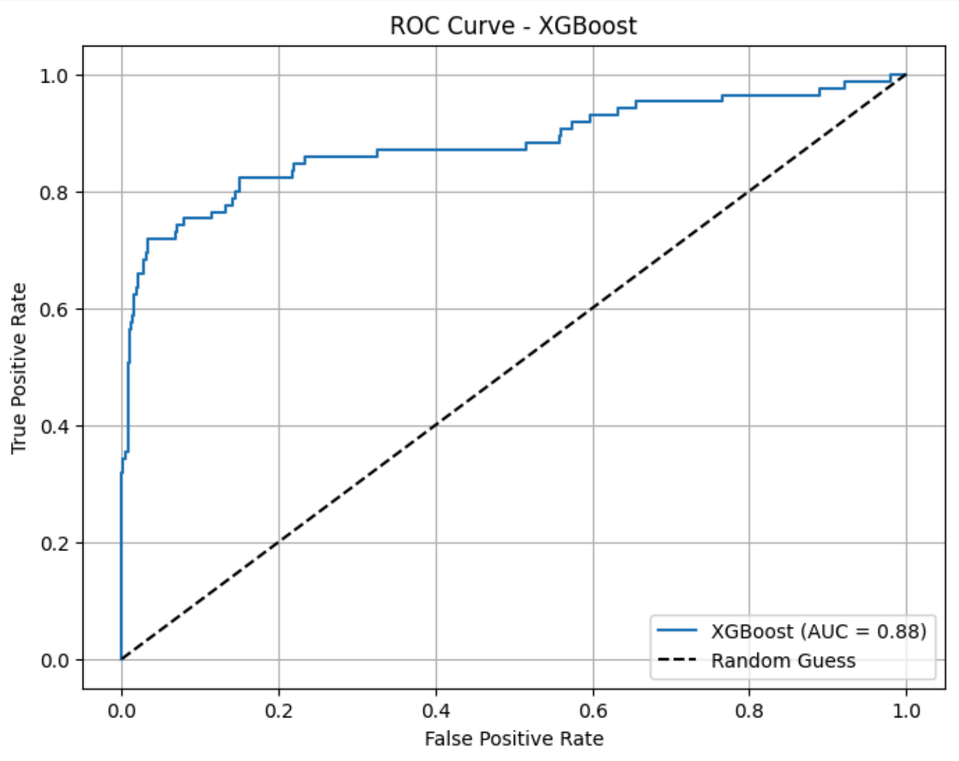
Random Forest:

Stable, high ROC-AUC **(0.90)**. Good for general separation.

XGBoost:

Selected as the Business Winner. Slightly lower AUC (**0.88**) but superior **Recall**.

Metric	Value	Business Implication
XGBoost Accuracy	92.9%	Significant lift over 85% baseline
Recall (Class 1 - Churn)	**72%**	Captures majority of at-risk users
Recall (Class 0 - Stay)	**96%**	Minimizes disturbance to loyal base

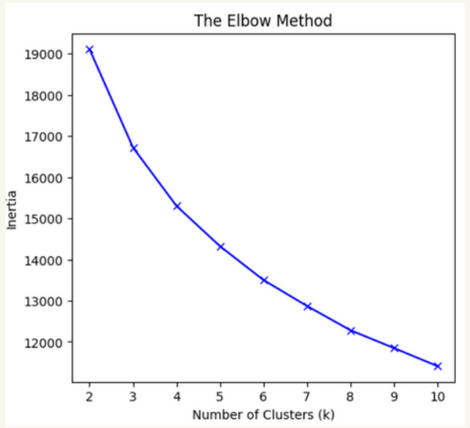


ACCURACY: percentage of total correct guesses.
PRECISION: the churners are correctly predicted (no false positives)
RECALL: out of all the people that actually churned how many could the model catch (no false negatives)

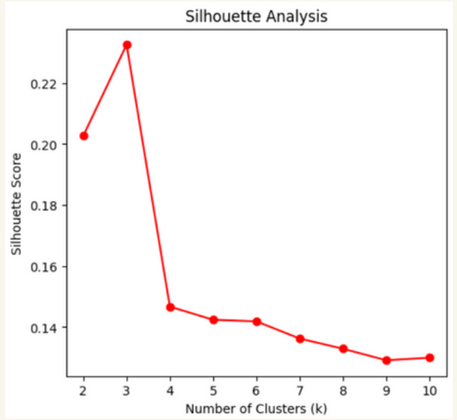
XGBoost is the selected model. In business, missing a churner (False Negative) is expensive.

XGBoost caught 10% more churners than Random Forest, achieving ~93% accuracy.

Customer Segmentation via K-Means: The 4 Personas



K=3 vs **K=4** for greater business interpretability



9,3% churning probability

Persona 1: Quiet Loyalists

Long tenure, balanced usage. They see value in the brand despite higher price-per-minute. Low churn.

4,7% churning probability

Persona 2: Premium Data Users

High data consumption. Tolerant of high prices. Minimal churn. The most profitable and stable group.

Persona 3: Price-Sensitive Light Users

Short tenure, very low spend. High churn risk. A 'trial-phase' relationship easily broken by confusion.

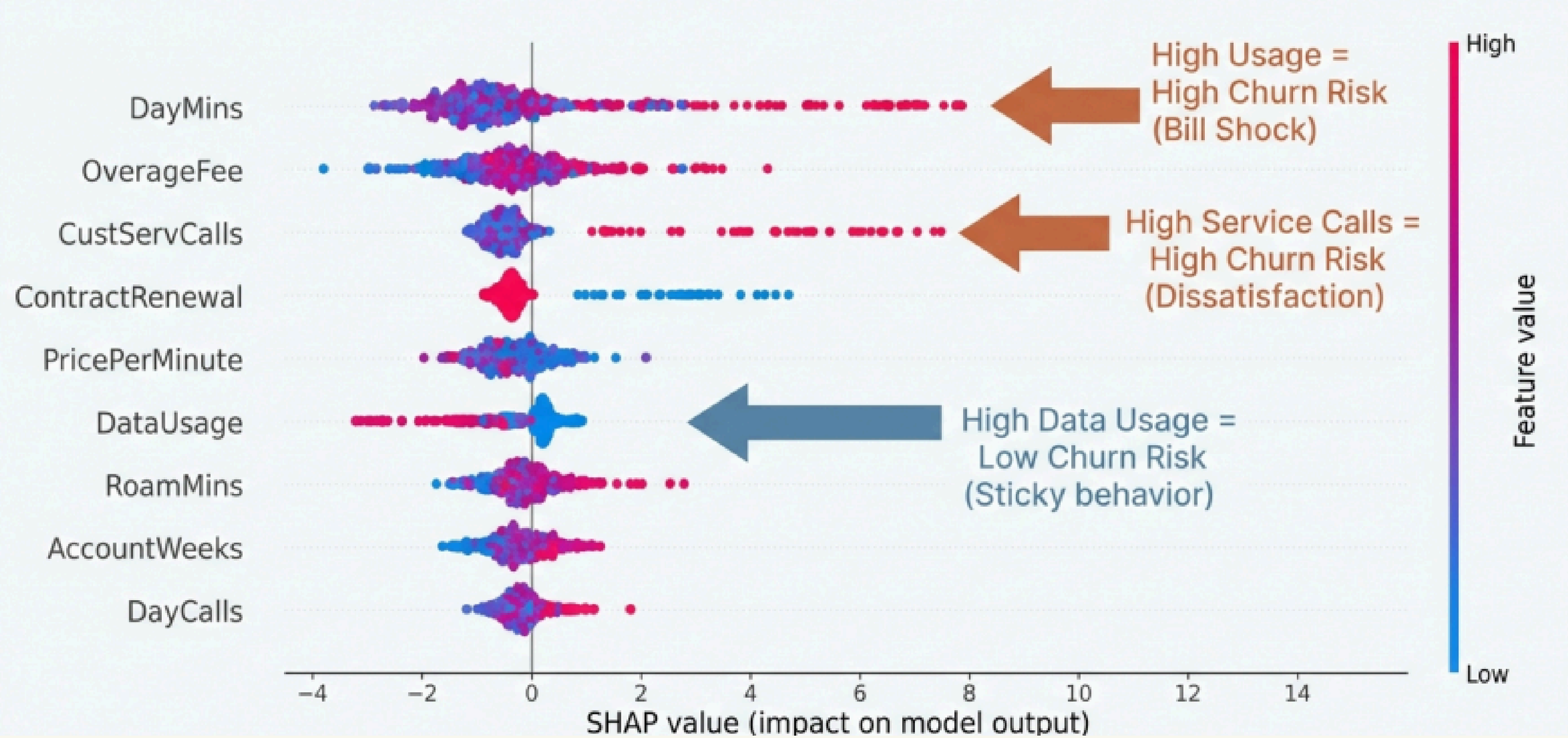
17,4% churning probability

Persona 0: Chronic Churners

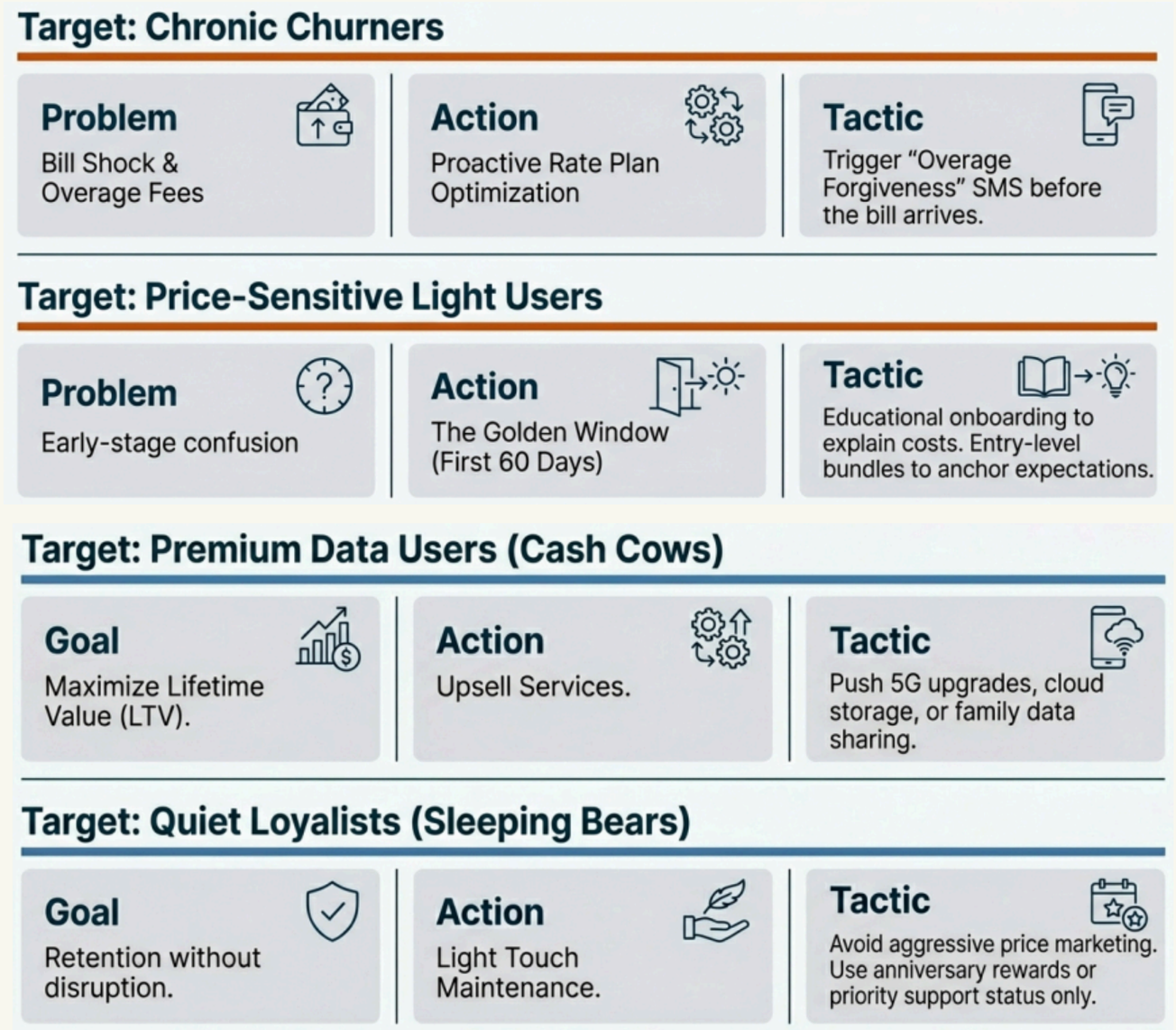
The 'Heavy User Paradox'. High voice usage, frequent overage fees, extreme churn rates. High value, high risk.

39,5% churning probability

Drivers of the Decision to Churn: SHAP Explainability



Business Strategies



RETENTION MARKETING

PERSONA 0:

- voice add-ons instead of full plan changes
- overage forgiveness

PERSONA 3:

- entry-level bundles
- contract incentives (discount after three months)
- on-boarding techniques

LOYALTY MARKETING

PERSONA 2:

- upgrade to premium data plans
- add-on services (cloud storage)

PERSONA 1:

- anniversary rewards
- loyalty perks
- no useless discounts if they are already loyal

Executive Summary & Conclusion

THE CONTEXT → **THE DISCOVERY** → **THE SOLUTION**

Recognized the high cost of acquisition vs retention in a saturated market.

Identified the 'Heavy User Paradox' (the most active users are the most prone to leave). Validated outliers as key target.

Deployable XGBoost (non-linear) for 93% Accuracy and superior Recall.

Final Takeaway: by segmenting customers into 4 different user personas, we have transformed raw data into targeted revenue protection. We prevent the loss of our highest-value, high-usage customers through proactive intervention.