Reduce Customer Churn for Trusource

PREDICTIVE ANALYTICS CASE COMPETITION

Saren Chatham, Fran Jimenez, Shamsa Khoja, Willow Noltemeyer, Arnold Songa

10 February 2026

THE BUSINESS PROBLEM

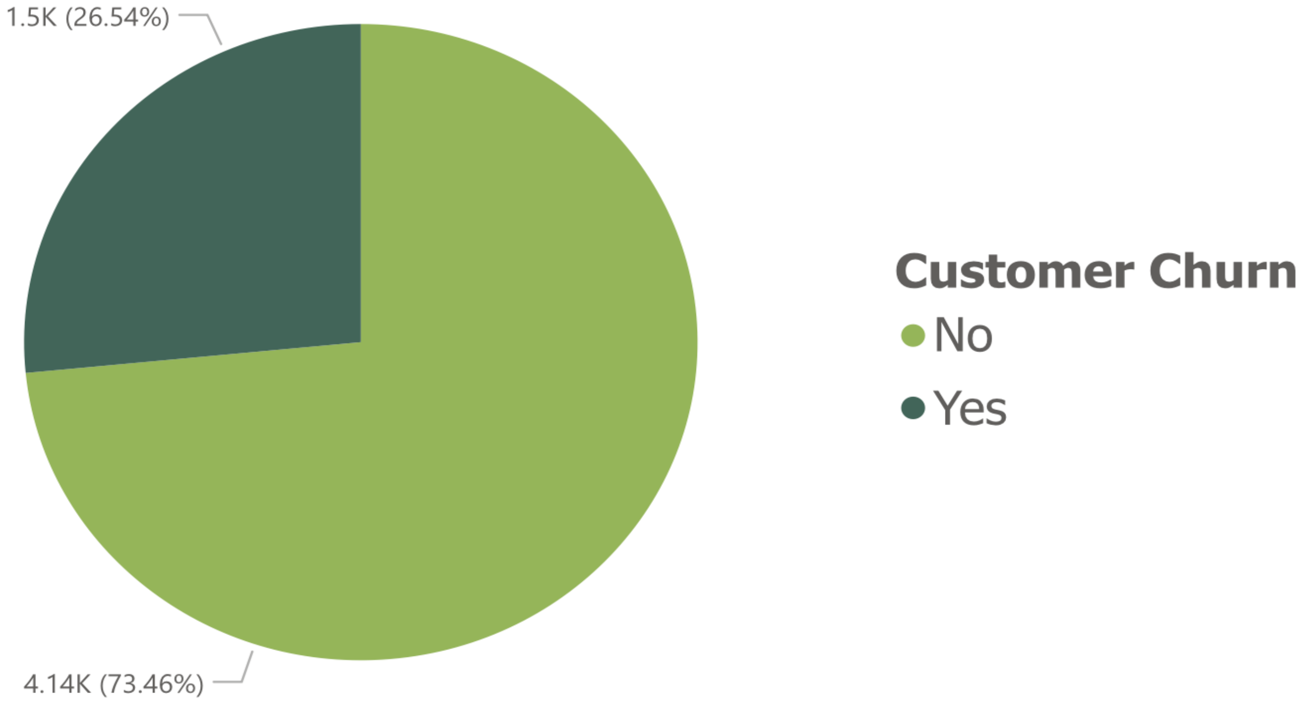
BUSINESS PROBLEM

THE CORE BUSINESS PROBLEM IS PREDICTING AND PREVENTING CUSTOMER CHURN

OBJECTIVE IS
TO IDENTIFY EARLY WARNING SIGNALS THAT INDICATE WHICH CUSTOMERS ARE MOST LIKELY TO CHURN IN THE FUTURE.

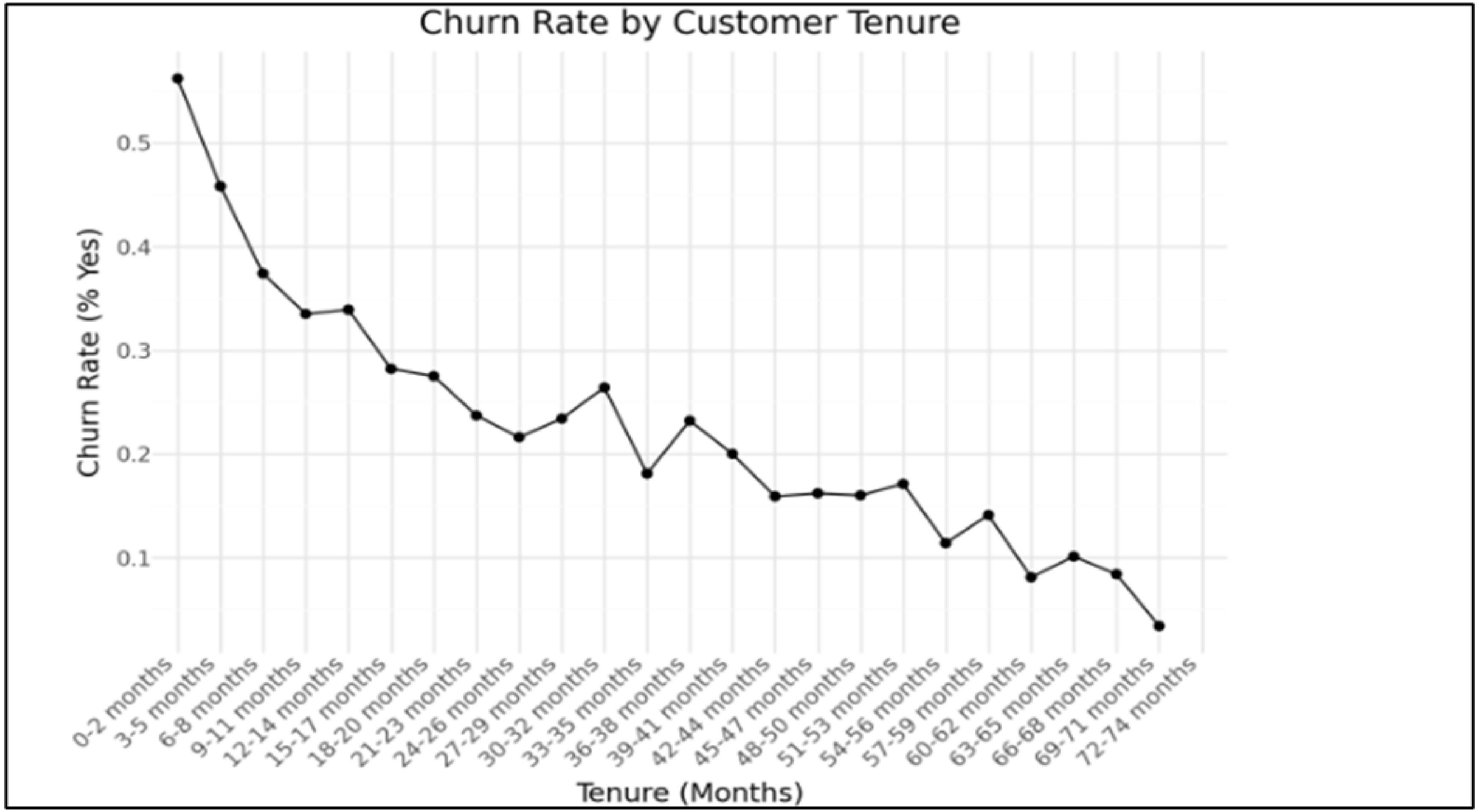
CHURNS REDUCES REVENUE AND INCREASES ACQUISITION COSTS.

Response Variable Distribution



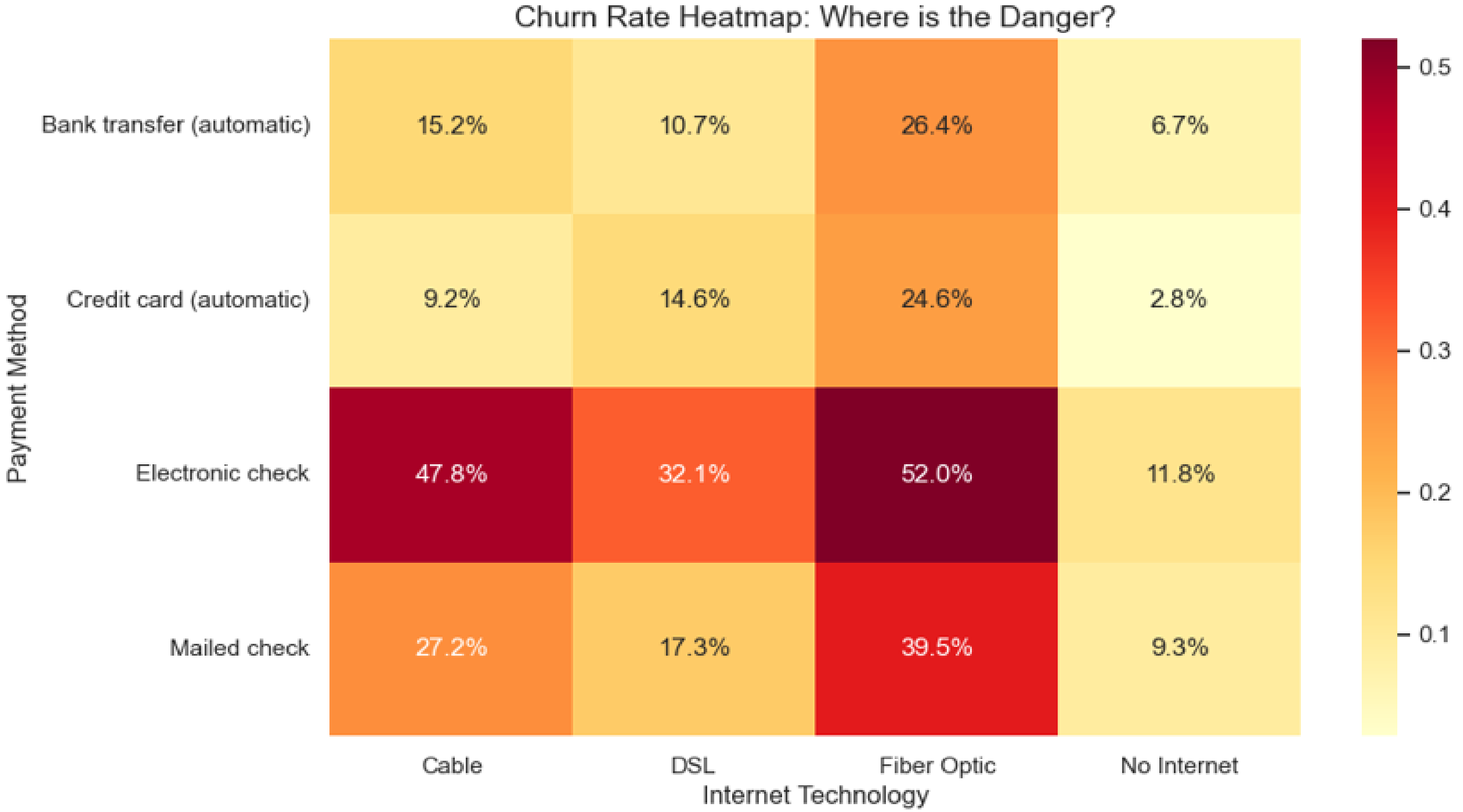
Response Variable	***left_flag*** indicates customer churn
Number of Observations	5636
Number of Predictor Variables	35

Dataset Overview



How customer churn rates change over time

Churn is higher for the early life cycle



Automatic payment methods like credit card and bank transfers records lower churn rates as it has the component of "locking the customers in"

DATA CLEANING AND PROCESSING

Missing Records	***internet_tech***, ***recent_offer*** Impute "Unknown", "No Offer"
Identifiers	***cus_ref***, ***account_ref*** Remove identifier variables
Imputed	***total_billed*** Missing values due to recent onboarding: customers yet to make a payment
Zero variance	***fiscal_qtr*** Variable dropped due zero variance
One-Hot Encoding	***payment method***, ***internet type***, ***recent offer*** Categorical variables (payment method, internet type and recent offer) to retained to preserve flexibility.

FEATURE ENGINEERING	
Add-on Services	***num_add_ons*** Combined numerous add-on features into a count of all add-ons
Streaming Services	***stream_tv***, ***stream_movies***, ***stream_music*** Aggregated into a binary feature indicating whether customer subscribes to any media streaming services
Binary Indicators	***extra_data_fees***, ***refunds_total*** Convert to binary indicator

The Model

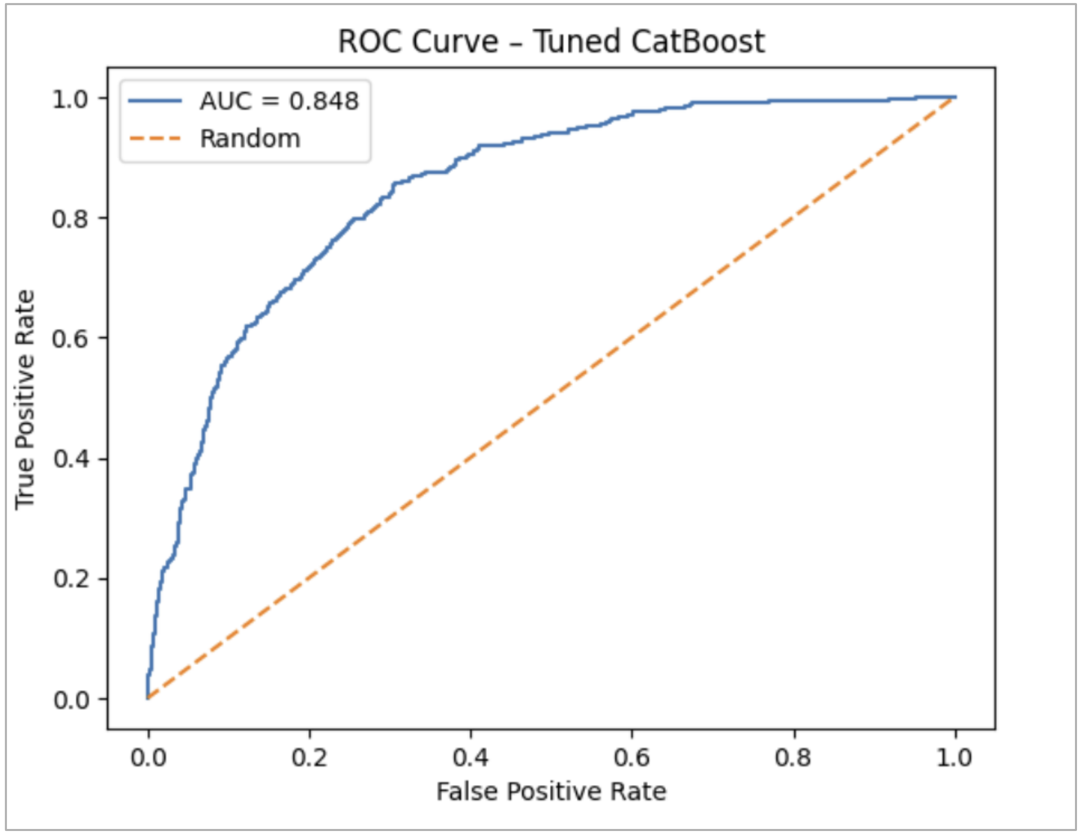
Multiple models evaluated to determine optimal predictive performance

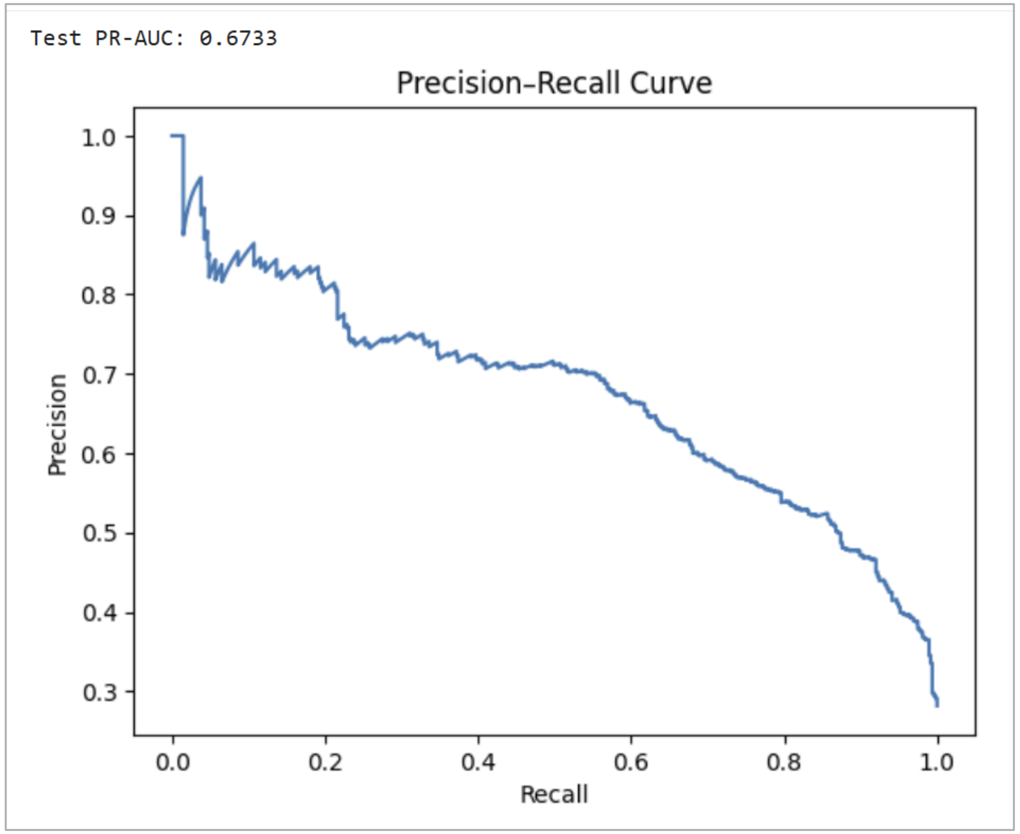
MODEL	TUNED HYPERPARAMETERS	TEST ROC-AUC
Logistic Regression (Lasso)	Max iterations = 2000 Solver = SAGA Penalty = L1 CV = 5-fold	0.853
Random Forest	Trees = 50–150 Depth = 3–5 Min leaf = 40–60 Max leaf nodes = 15–30	0.843
CatBoost	Depth = 4–8 Learning rate = 0.01–0.03 Iterations = 400–800 L2 regularization = 1–10	0.848
XGBoost	Trees = 400–800 Depth = 4–8 Learning rate = 0.05–0.1 Row sample = 0.8–1.0 Feature sample = 0.8–1.0 Min child weight = 1–3	0.829

MODEL EVALUATION METRICS:

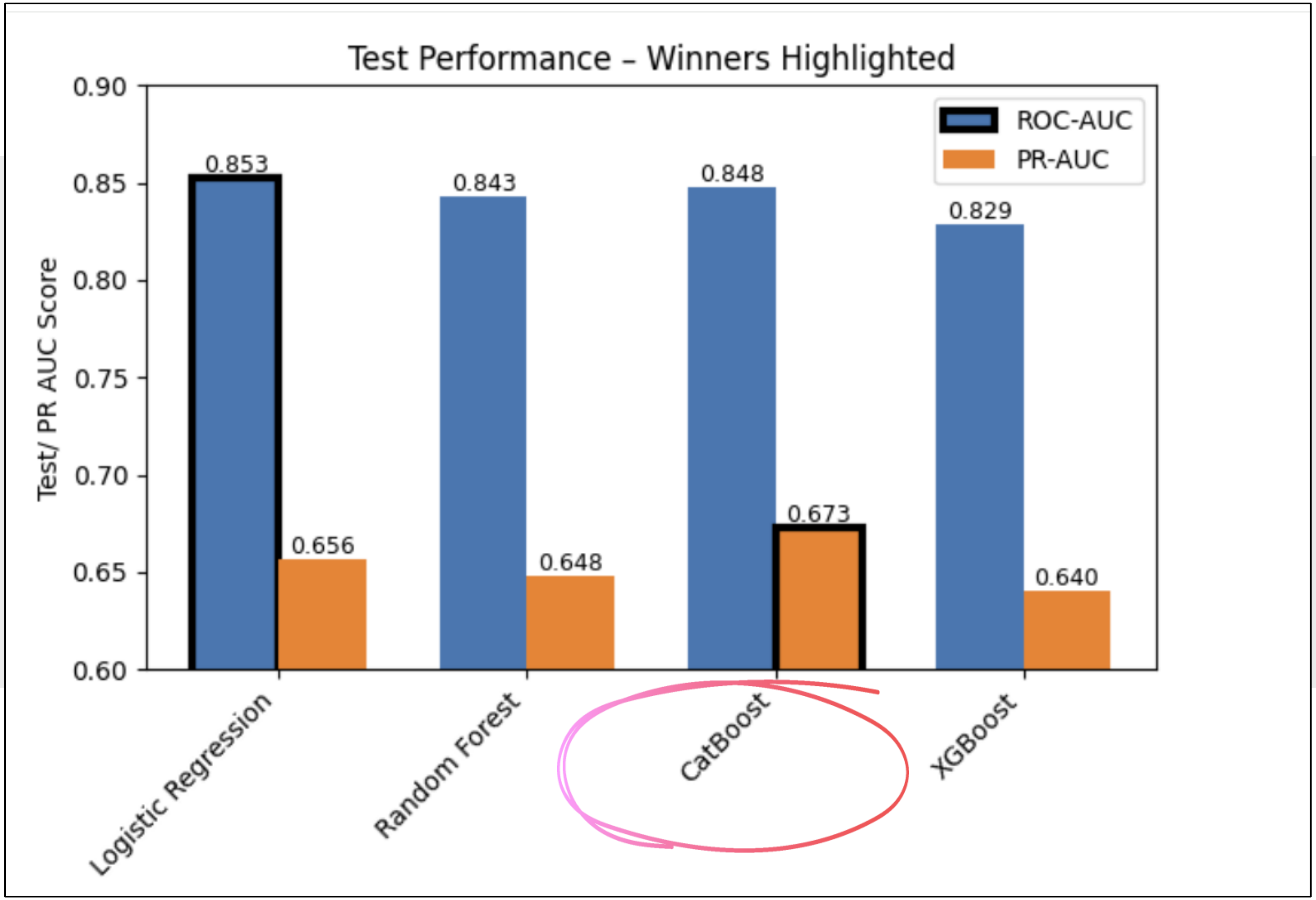
ROC – AUC → how well can the model distinguish between classes at any threshold

PR – AUC → when predicting positive, am I right or missing many.





HEAD-TO-HEAD BENCHMARKING ON UNSEEN DATA



Catboost Wins! Let’s see why?

WHY CATBOOST WINS ON CHURN DATA

Preserves **behavioral** information

Natively handles categorial variables

Preserves signal loss in one-hot coding due to sparse data

Critical for our data set with many important cat-features

Best performance on **accuracy**

Second highest Test ROC- AUC

More robust (ensemble approach)

Captures non-linear relationships

More prediction power

Superior **churn detection**

Highest Precision Recall – AUC

Identifies more true churners

Higher impact per campaign

Improves marketing ROI

Selecting the best parameters for model tuning

Hyperparameter	Option 1	Option 2	Option 3
Tree Depth	4	6	8
Learning Rate	0.01	0.03	—
Iterations	400	800	—
Regularization	1	3	10

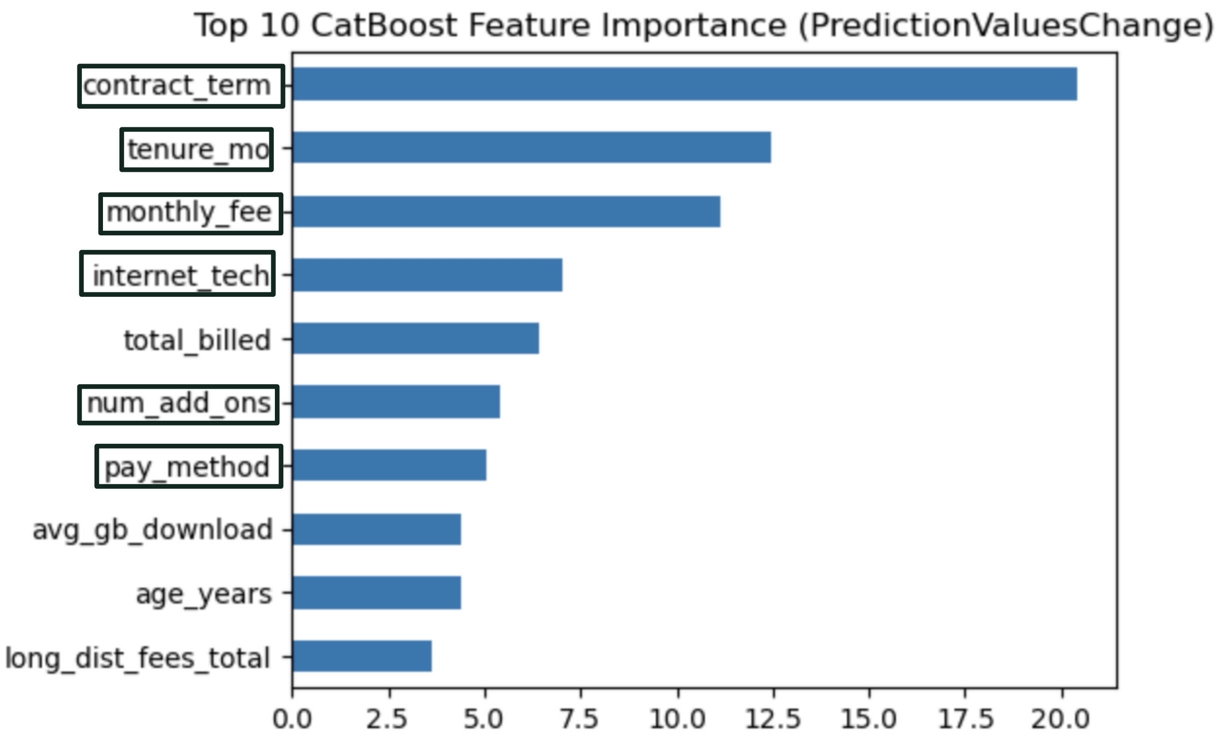
3 x 2 x 2 x 3= 36 models

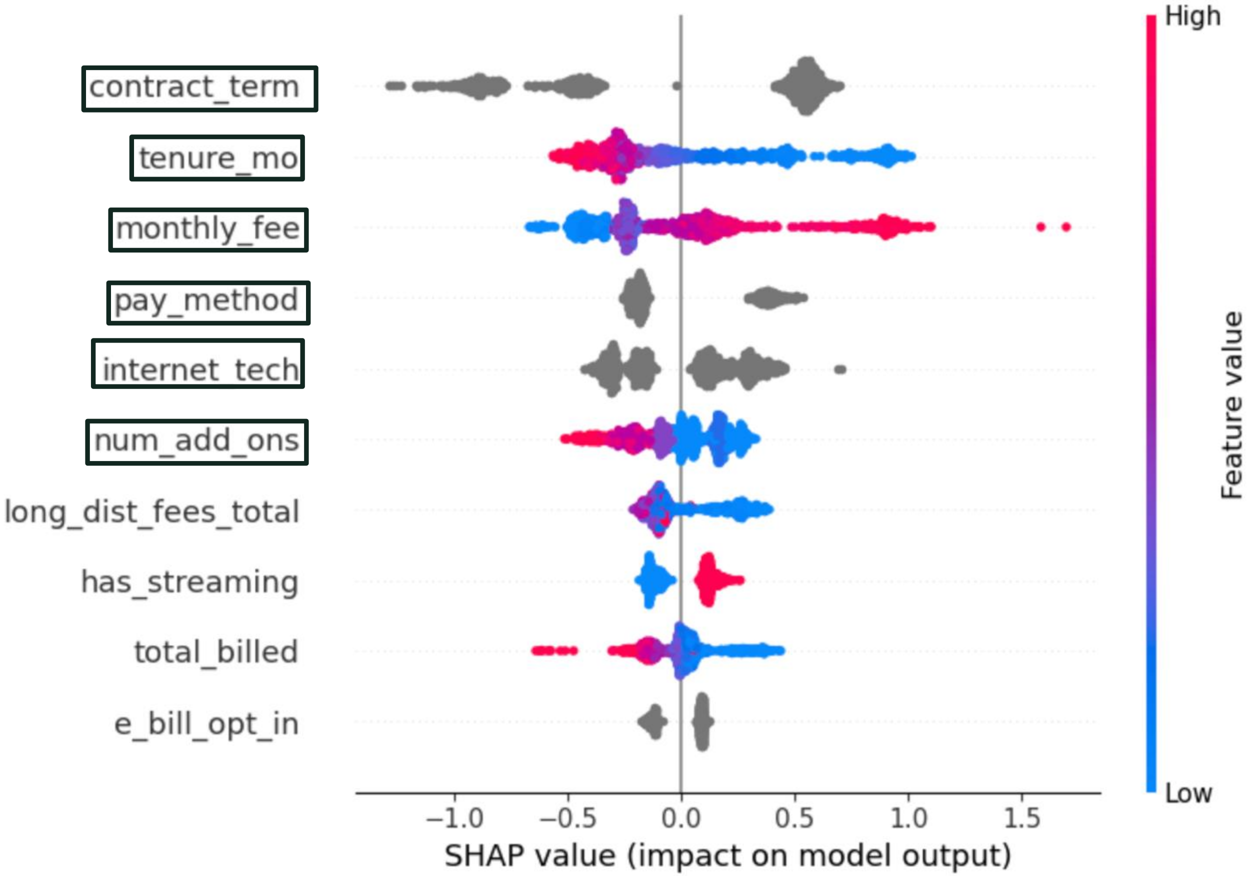
Using GridSearchCV we systematically tried **every combination**:

Fold 1 → AUC = 0.82
Fold 2 → AUC = 0.79
Fold 3 → AUC = 0.81
Fold 4 → AUC = 0.80
Fold 5 → AUC = 0.83
Mean AUC = 0.81 ← used for comparison

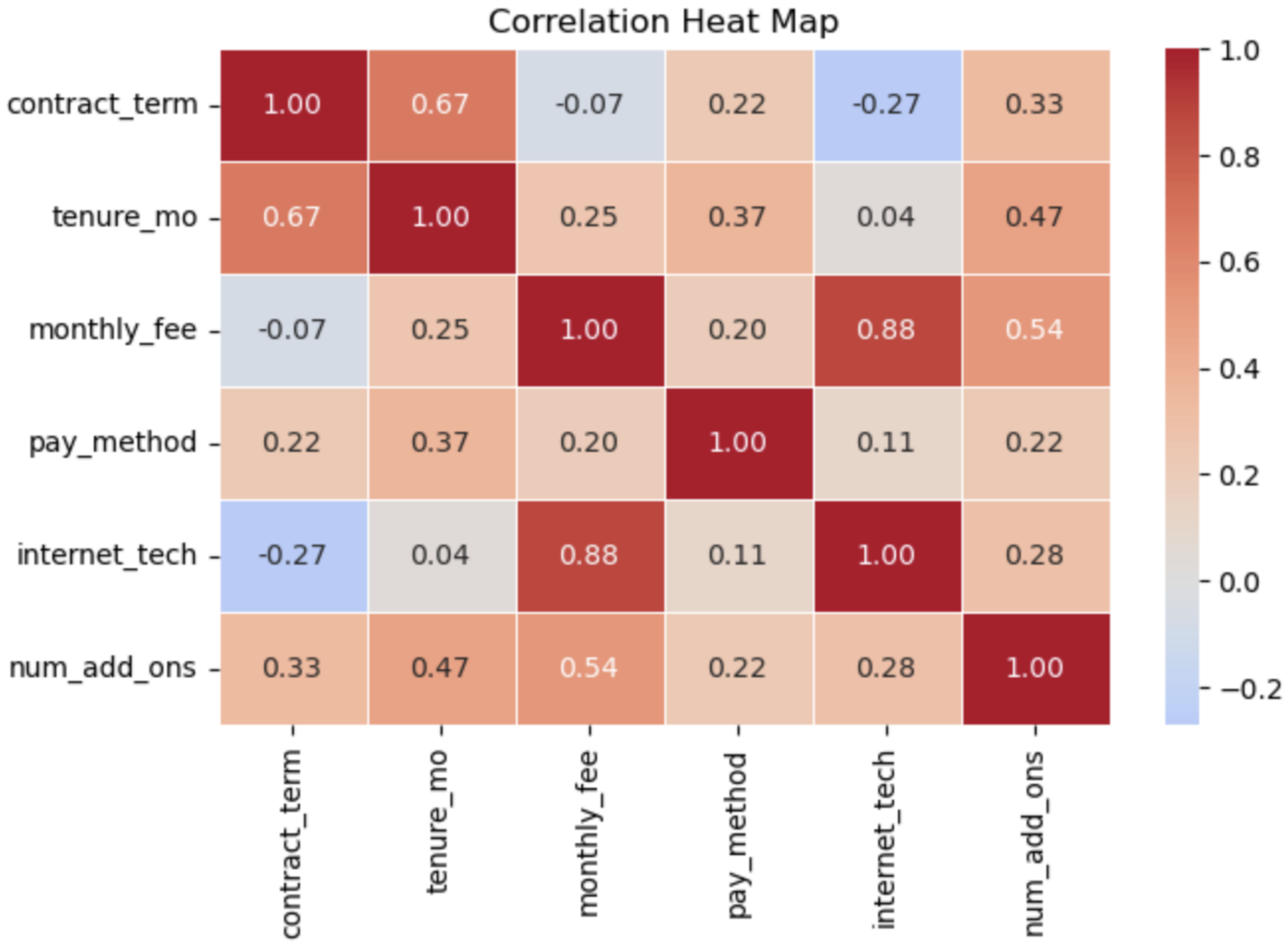
We used 4 parameters to train 36 different Catboost models across 5 CV folds.
(36 x 5 → 180 fits)

TOP 10 FEATURE IMPORTANCE AND SHAP VALUES





FEATURE CORRELATIONS



Correlations between contract term/tenure months, internet tech/monthly fee, and number of add ons/monthly fee are expected

Unexpected correlations:

Tenure months/number of add ons

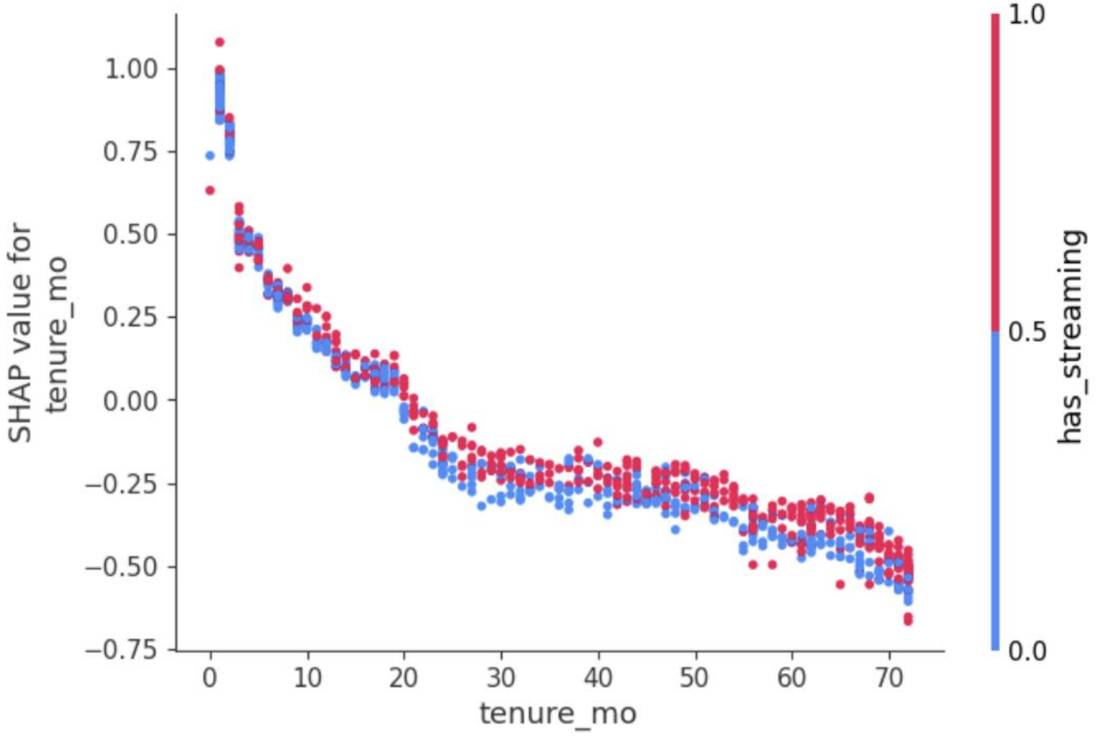
Contract term/internet tech

Contract term/monthly fee

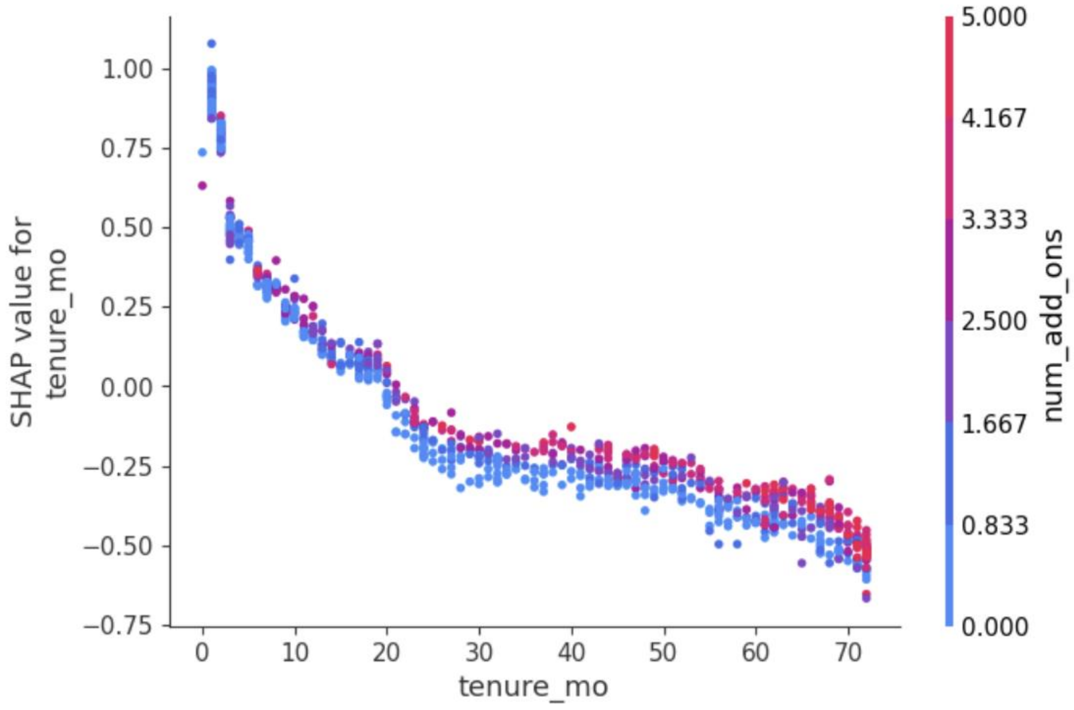
Key Driver: Engagement

Users without streaming services or add-ons in their first 10 months have dramatically higher churn risk → Pushing engagement in early on-boarding is important.

Streaming:



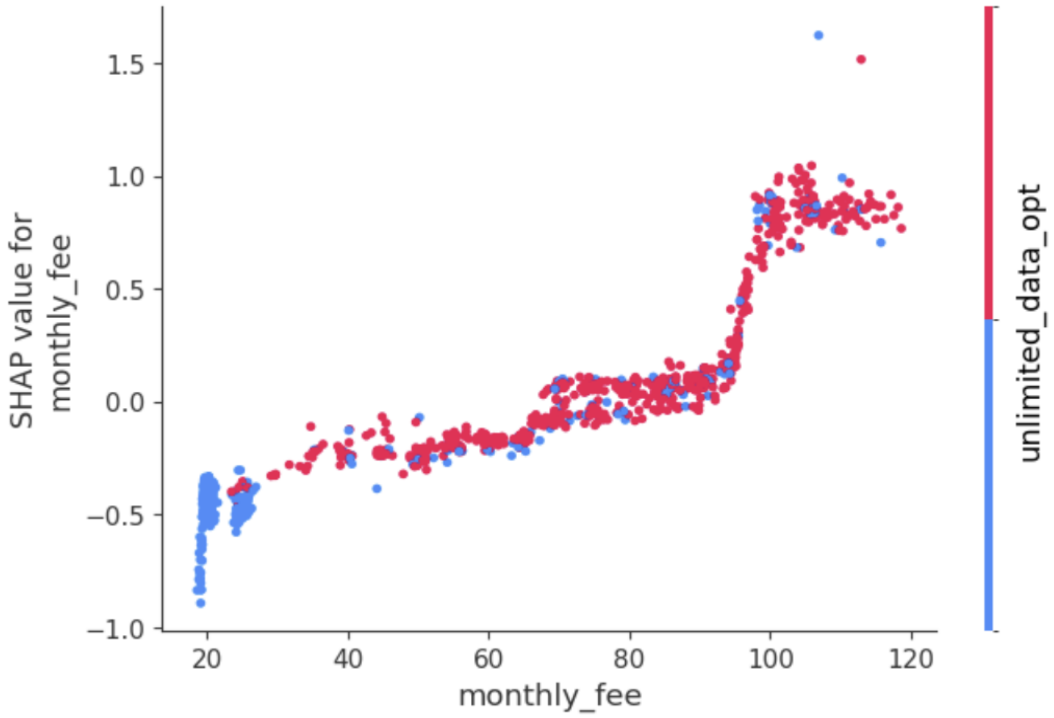
Add-ons:



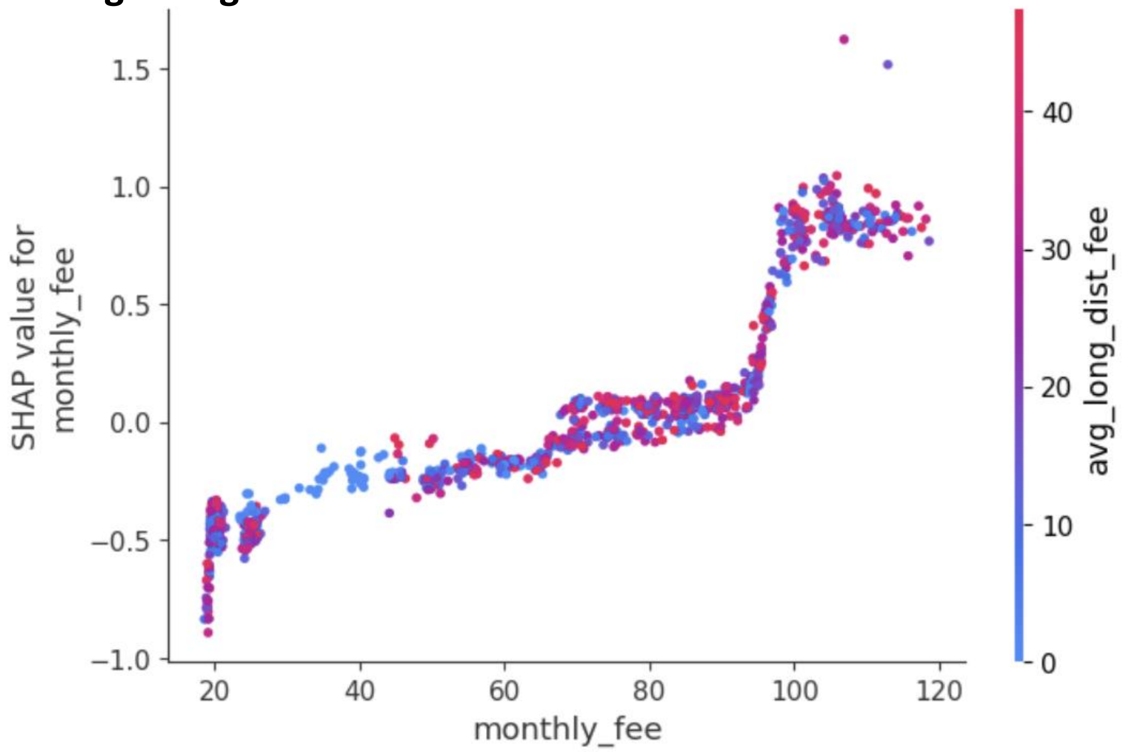
Key Driver: Paying above monthly fee

Users paying $85+ for extra data/ long distance charges are at churn risk.

Unlimited Data Opt-In:



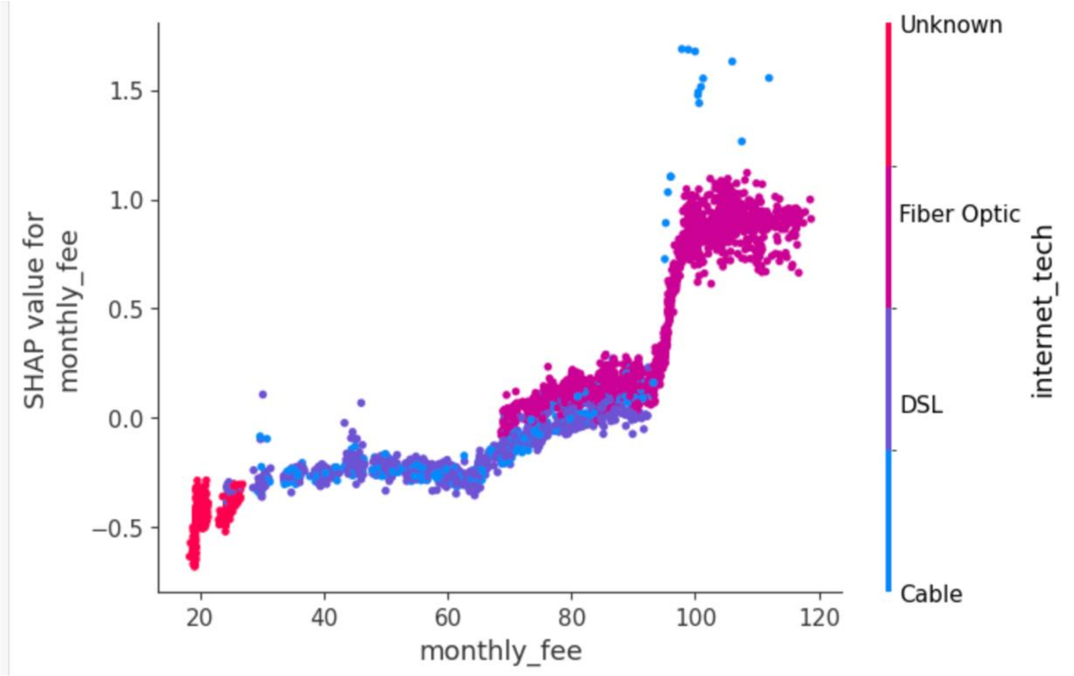
Average Long-Distance Fee:



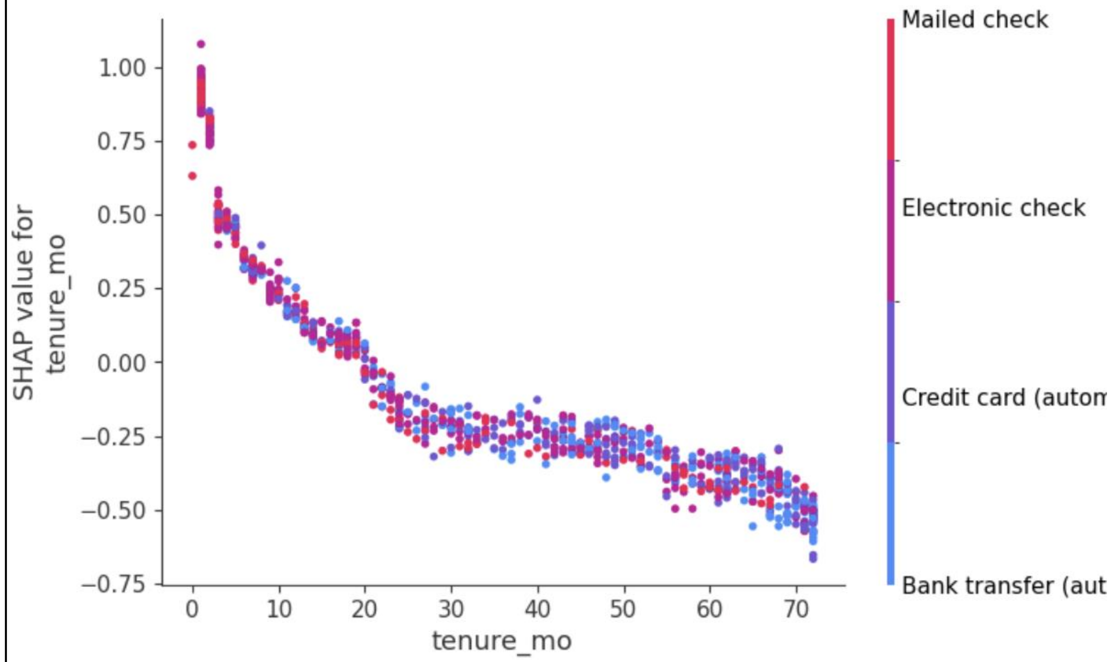
Key Driver: Tech-forwardness

Expensive internet plans have high churn risk when monthly fees exceed $100 and using checks have high churn risk in early life cycle particularly first 2 years.

Internet Tech:



Pay-method:



K-Means & Segmentation

WHAT IS K-MEANS?

WHAT

- CLUSTERS data into significant groups
- Exposes potentially HIDDEN insights

HOW

- Choose amount of clusters
- Calculates centroids for each cluster based on the mean
- Minimizes sum of squares within each cluster cluster

K-MEANS PROCESS

ASSESS SELECTED FEATURES

TEST VIA K-MEANS ALGORITHM

EVALUATE RESULTS

ASSESS

Initial Feature Experimentation

1. Original dataset
2. Select features from the original dataset
   - Significance in model
   - Minimize highly correlated features

FINALIZED FEATURE SELECTION

Tenure (months)
Monthly Fee
Number of Add-ons
Long Distance Fee per Month
Unlimited Data User
Electronic Check Payment Method
Fiber Optic User
Month-to-Month Contract
Age
Auto Payment
E-bill
Monthly Data Use (GB)

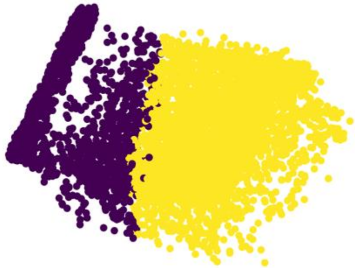
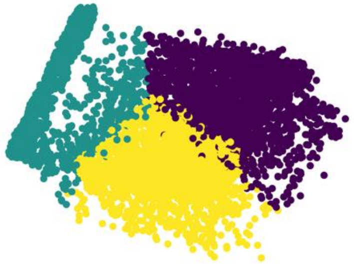
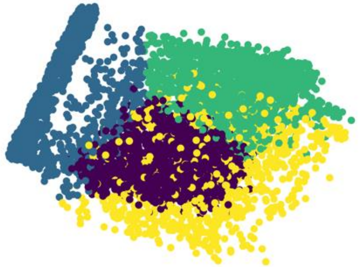
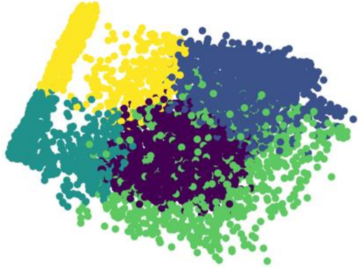
TEST

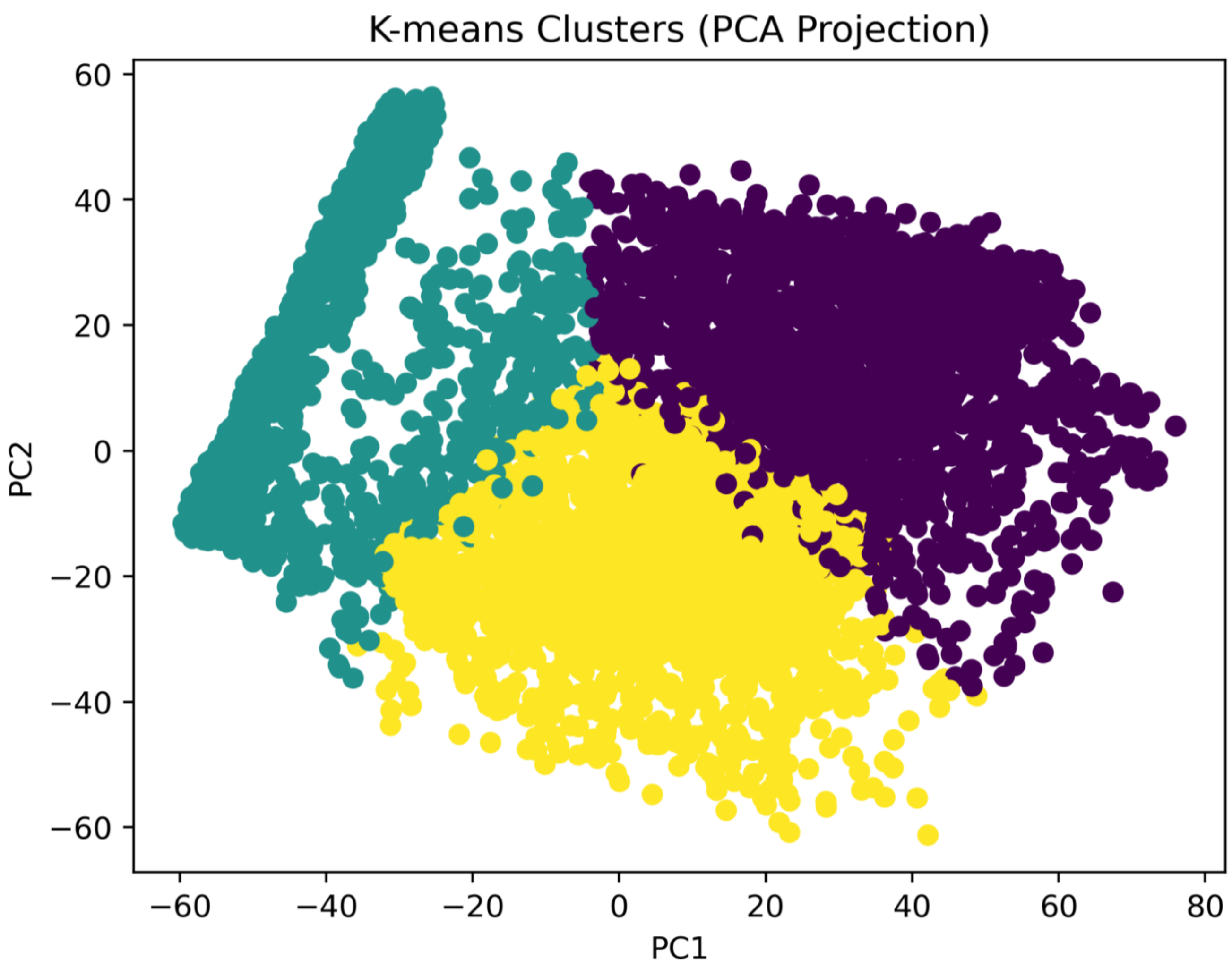
1. Run K-Means Algorithm
2. Test each dataset with various numbers of clusters
   - 2 – 5 clusters per set of features

EVALUATE

Metrics for Evaluation:

1. PCA Projection Plot
2. Within-Cluster Sum-of-Squares (WCSS) A.K.A. Inertia
3. Calinski-Harabasz Score A.K.A Variance Ratio Criterion
4. Silhouette Score
5. Davies-Bouldin Score

Metrics for Evaluation	2 Segments	3 Segments	4 Segments	5 Segments
PCA Projection Plot				
Within-Cluster Sum-of-Squares (WCSS) A.K.A. Inertia	9.3M	7.4M	6.1M	5.3M
Calinski-Harabasz Score A.K.A.Variance Ratio Criterion	2700	2400	2300	2300
Silhouette Score	0.30	0.27	0.29	0.28
Davies-Bouldin Score	1.33	1.40	1.20	1.23



PCA PROJECTION PLOT

3 SEGMENTS

FINAL SEGMENTATION

	Segment 1	Segment 2	Segment 3
Tenure (months)	57	30	13
Monthly Fee	$ 89.83	$ 26.05	$ 73.20
Number of Add-ons	3	0	1
Long Distance Fee per Month	$ 24.20	$ 19.82	$ 23.97
Unlimited Data User	86%	23%	87%
Electronic Check Payment Method	31%	15%	50%
Fiber Optic User	65%	0%	58%
Month-to-Month Contract	29%	43%	87%
Age	48	44	48
Auto Payment	62%	0.42	29%
E-bill	70%	37%	68%
Monthly Data Use (GB)	27	5	27
Percent	**33%**	**29%**	**38%**

Contextual **recommendations** for the three distinct groups:

3 Segments

1 — **(33%) Loyal Premium Bundled Users**

2 — **(29%) Split Wallet Customers**

3 — **(38%) Early-Life Month-to-Month Majority**

We chose 3 segments because it produces **clearer, more distinct customer personas** and **avoids** splitting customers into **overlapping** groups that lead to the same retention action.

	Segment 1	Segment 2	Segment 3
Tenure (months)	57	30	13
Monthly Fee	$ 89.83	$ 26.05	$ 73.20
Number of Add-ons	3	0	1
Long Distance Fee per Month	$ 24.20	$ 19.82	$ 23.97
Unlimited Data User	86%	23%	87%
Electronic Check Payment Method	31%	15%	50%
Fiber Optic User	65%	0%	58%
Month-to-Month Contract	29%	43%	87%
Age	48	44	48
Auto Payment	62%	42%	29%
E-bill	70%	37%	68%
Monthly Data Use (GB)	27	5	27
Percent	**33%**	**29%**	**38%**

SEGMENT 1 (33%) LOYAL PREMIUM BUNDLED USERS
PROTECT HIGH-CLV CUSTOMERS: VIP RETENTION SHIELD

Tenure	57 months (longest)
Monthly Fee	$89.83 (highest)
Add-ons	3 (highest)
Average Long Distance Fee	$24.20
Unlimited Data	86%
Month-to-month	29%

Recommendation:

- Create a **VIP retention experience**
- Careful upsell (Long-distance add-on, international calling plans,)
- Handle over-paying

KPI to Track:

- Avg. long-distance fee per customer, before vs. After
- Average support resolution time

This group is already loyal — churn is most likely triggered by **service failures or unexpected price changes**, not contract structure.

SEGMENT 2 (29%) SPLIT WALLET CUSTOMERS
DIAGNOSTIC + PREDICTABLE BUDGET BUNDLE (NOT DISCOUNTS)

Tenure	30 months (mid-term)
Monthly Fee	$26.05 (lowest)
Monthly Data Use (GB)	5 (lowest)
Avg. Long distance fee	$19.82
Unlimited Data	23% (lowest)
Month-to-month	43% (moderate risk)

Recommendation:

- Offer an upgrade trial (unlimited data / fiber)
- Split wallet diagnostic
- Reduce long-distance fee exposure

KPI to Track:

- Monitor the increase in average gb dowload after trial
- Conversion into primary provider
- Avg. Long distance fee, before vs. after

Our strategy focuses on winning wallet share through targeted data trials, understanding split-service behavior, and preventing long-distance bill shock.

SEGMENT 3 (38%) EARLY-LIFE MONTH-TO-MONTH MAJORITY

POTENTIAL HIGH-VALUE USERS... BUT EASIEST TO LOSE

Tenure	13 months (lowest)
Monthly Fee	$73.20
Add-ons	1
Electronic Check Payments	50%
Unlimited Data	87%
Month-to-month	88% (extremely high risk)

Recommendation:

- Run an offer or contract before they leave
- Curate fiber/data optimization upgrade
- Drive stickiness using add-on bundling
- Reduce Payment Friction – push auto-pay adoption

KPI to Track:

- Contract Lock-In Success
- Engagement/Stickiness Growth
- Payment Stability
- Revenue + CLV impact

This is our largest and most valuable churn risk: high-paying customers on month-to-month contracts. The strategy is to lock them in early, reduce bill shocks, increase engagement through add-ons, and stabilizing payment behavior.

TRADE-OFFS TO BE MINDFUL OF:

Segment 1:

- Premium support, priority service requires dedicated resources for already-profitable segment
- Over-investing in retention when these customers might stay anyway (diminishing returns)

Segment 2:

- Win-back data business: High acquisition cost to compete on data pricing vs. Low current ARPU ($22) - may not be profitable even if successful
- Investing heavily in low-value customers who may never become high-ARPU users

Segment 3:

- Expensive offers to convert month-to-month erodes margins on already high-ARPU customers
- Discount dependency risk- customers may stay for as long as the discounts exists. In the long-run, churn spikes after the discounts end.