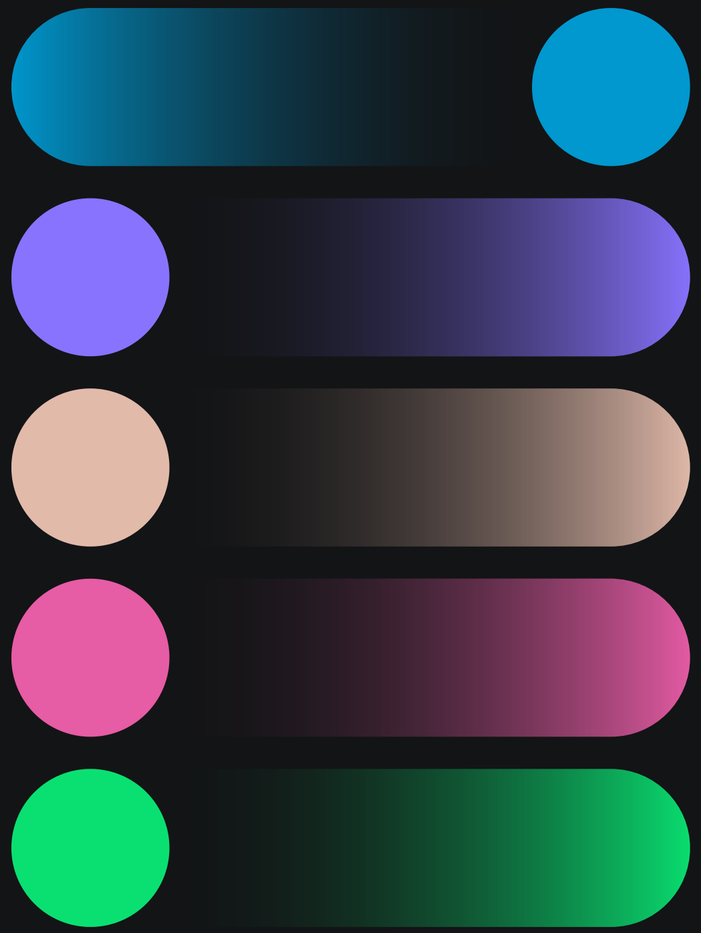
ENTERPRISE DATA SCIENCE BOOTCAMP - GROUP 20

One out of Four Customers Is Leaving
Can We Stop the Next One?

Can we reliably predict churn early enough to intervene, and how should the business act on it?

NOVA IMS
INFORMATION MANAGEMENT SCHOOL



Understanding the Problem

If we can predict churn early, we can intervene before customers leave

Customer churn prediction estimates the chance of a customer ending their relationship with a company. Tracking who leaves and who stays helps reveal behavior patterns and the factors that contribute to churn risk.

With predictive models, it becomes possible to spot risks ahead of time and act before customers decide to walk away.

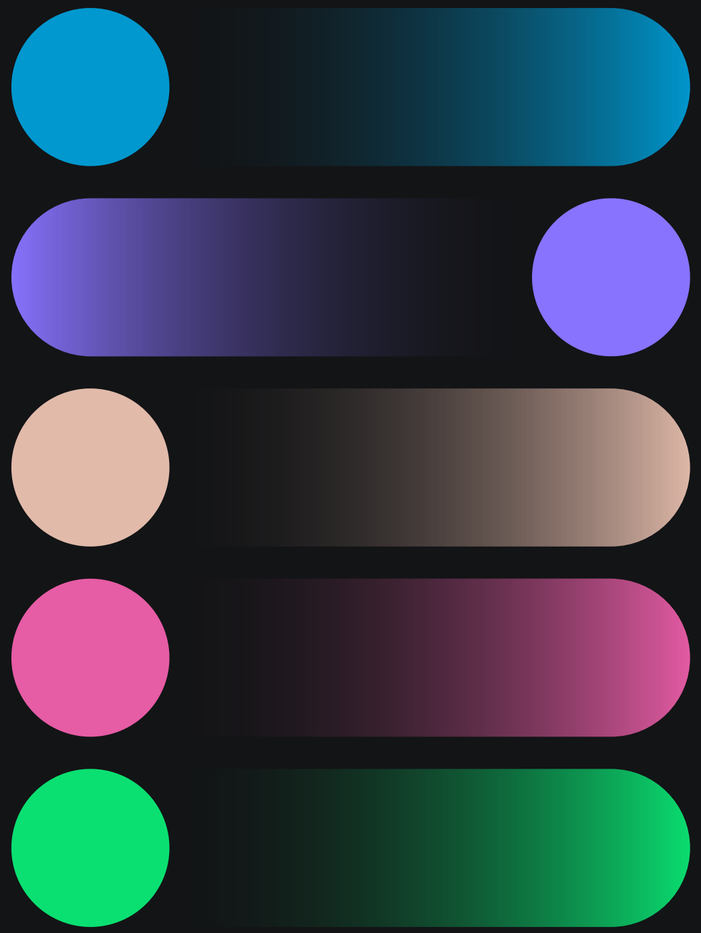
26.5% of customers churn

Losing customers impacts revenue, and acquiring new ones can cost 5 to 25x more than retaining existing users

Purpose

Predict which customers are most likely to churn, understand why they churn and enable targeted retention actions

Fonte: Growth-O-Nomics – Churn Metrics Guide



Understanding the Data

We’ve built a complete behavioral and financial profile of every customer

From 5 independent business domains that directly map to each business departments:

- **Demographics** → **MARKETING**
- **Location Context** → **STRATEGY**
- **Population Context** → **STRATEGY**
- **Services & subscriptions** → **SALES**
- **Customer Status & Churn History** → **OPS**

To a clean, future-safe dataset that supports reliable churn predictions.

- 7,043 customer records
- 31 final predictive variables
- 0 duplicate customers
- All records validated for consistency

Risk & Leakage Control

- All post-churn and outcome-based variables removed
- Model trained only on information available **before** churn

We’ve built a complete behavioral and financial profile of every customer

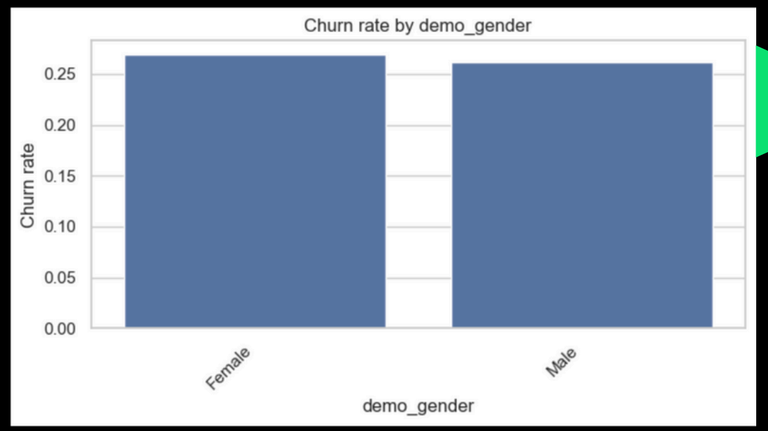
From 5 independent business domains that directly map to each business departments:

To a clean, future-safe dataset that supports reliable churn predictions.

Demographics	**MARKETING**	[gender, age, marital status, number of dependents]
Location Context	**STRATEGY**	[city, zip code, latitude, longitude]
Population Context		
Services & subscriptions	**SALES**	[tenure in months, offer, phone service, avg monthly long distance charges, multiple lines, internet service, internet type, avg monthly gb download, online security, online backup, device protection plan, premium tech support, streaming tv, streaming movies, streaming music, unlimited data, contract, paperless billing, payment method, monthly charge, referred a friend, number of referrals]
Customer Status & Churn History	**OPS**	[churn label (Yes, No)] → our **target**

Life-stage stability matters: single customers or with no dependents are more likely to leave

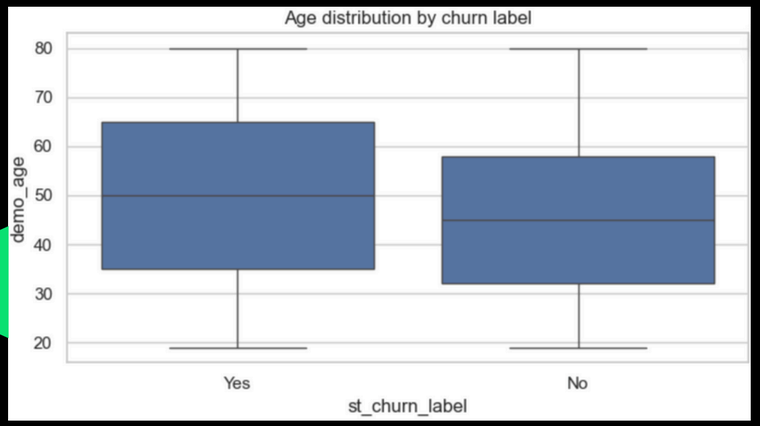
DEMOGRAPHICS VS CHURN

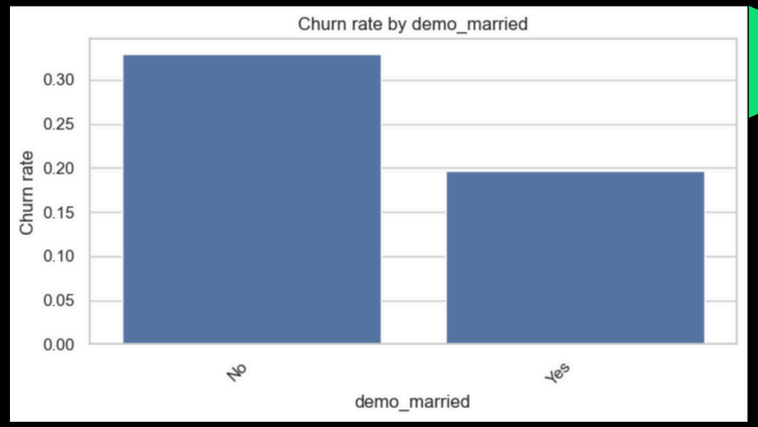


Key Findings

Gender: almost no effect (≈26–27% churn both Male/Female)

Age: churners are slightly older on average (≈50 vs 45), but effect is modest.



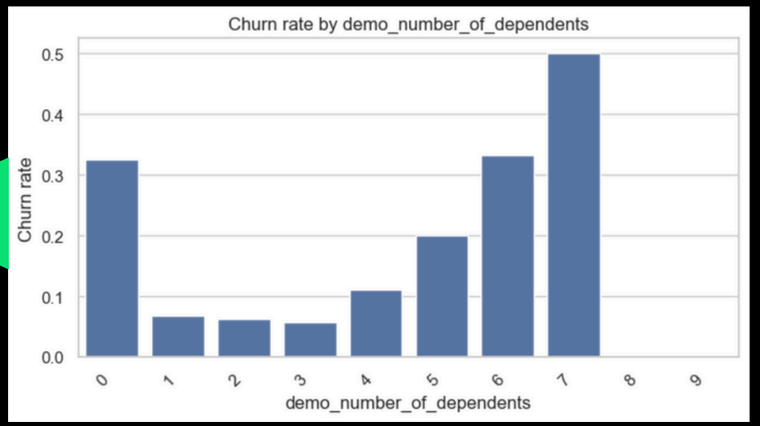


Marital status:

- Not married: churn ≈ 33%
- Married: churn ≈ 20%

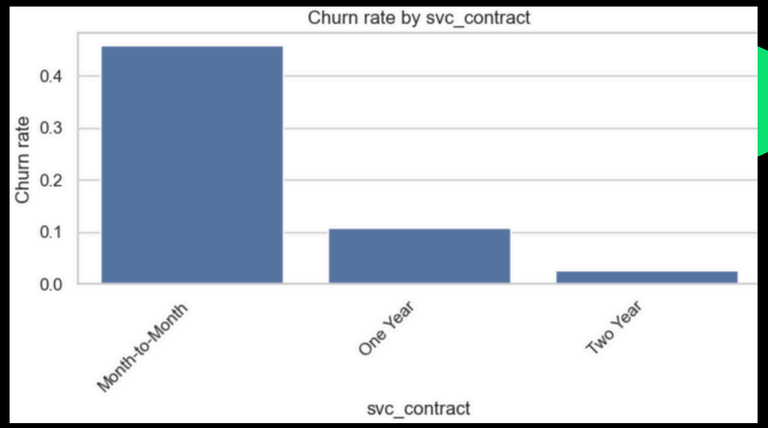
Dependents:

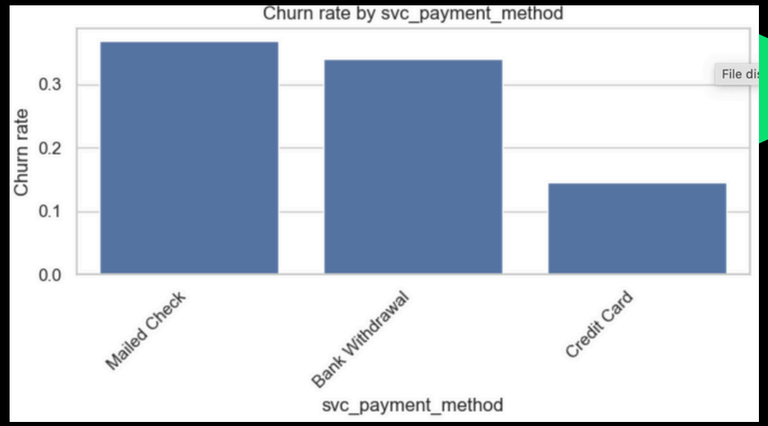
- 0: high churn (~32%)
- 1–3: churn < ~10%(very stable)
- 4+: churn increases again, but sample is tiny.



Understanding how customers are billed and the contract they have are key for churn

CONTRACT AND BILLING VS CHURN





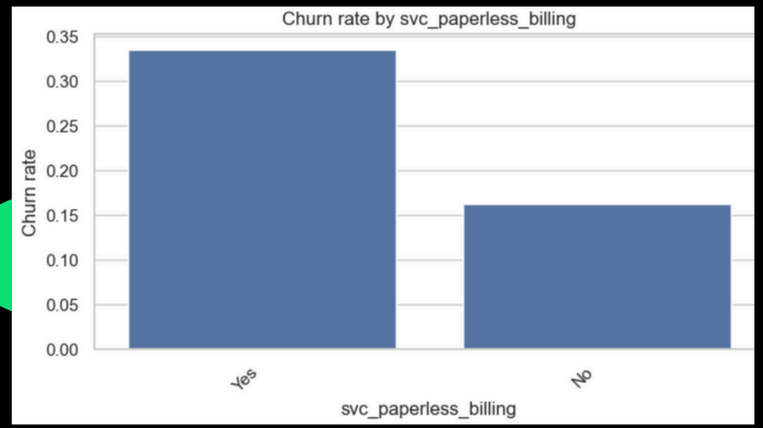
Key Findings

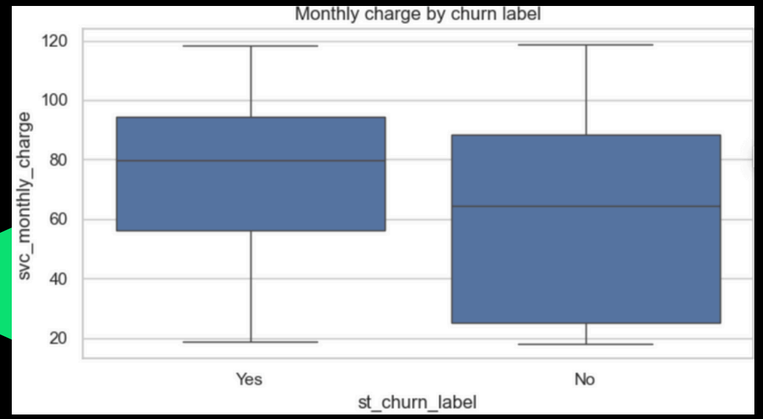
Contract: Month-to-Month customers are 18x more likely to churn than 2-year customers.

Paperless Billing: correlates with churn but is mostly a proxy for flexible online contracts, not a root cause.

Payment Method: Independent driver of churn (even within Month-to-Month contracts, churn increases from credit card < bank withdrawal < mailed check).

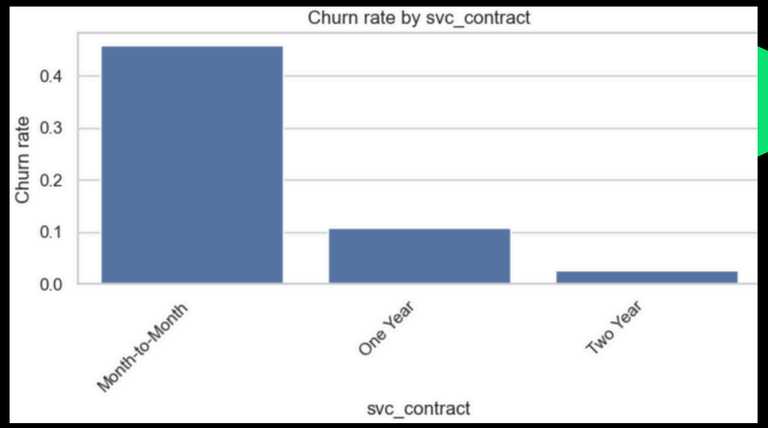
Monthly Charge & Offers: Some offers attract high-churn customers rather than retaining them.





Understanding how customers are billed and the contract they have are key for churn

CONTRACT AND BILLING VS CHURN



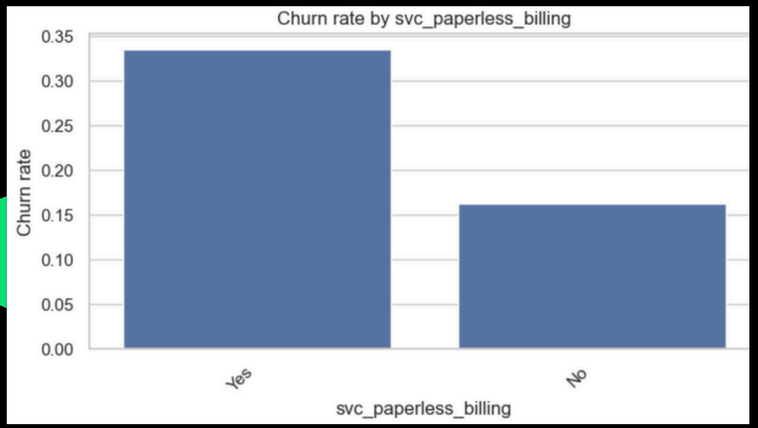
Key Findings

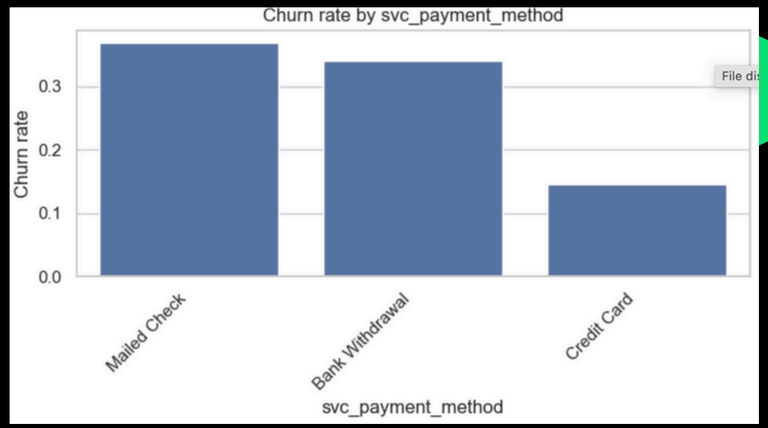
Contract: Month-to-Month customers are 18x more likely to churn than 2-year customers.

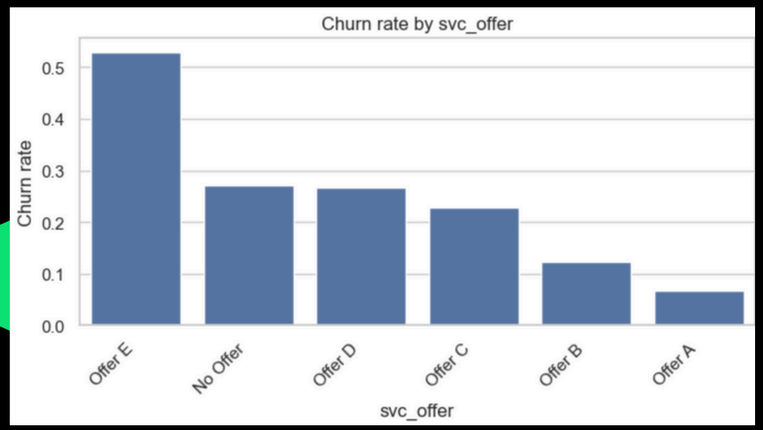
Paperless Billing: correlates with churn but is mostly a proxy for flexible online contracts, not a root cause.

Payment Method: Independent driver of churn (even within Month-to-Month contracts, churn increases from credit card < bank withdrawal < mailed check).

Monthly Charge & Offers: Some offers attract high-churn customers rather than retaining them.

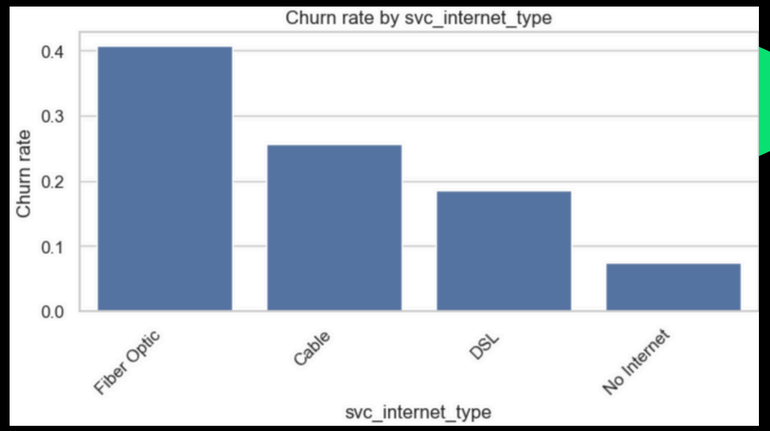






Add-ons & engagement protect against churn, while premium usage increases churn risk

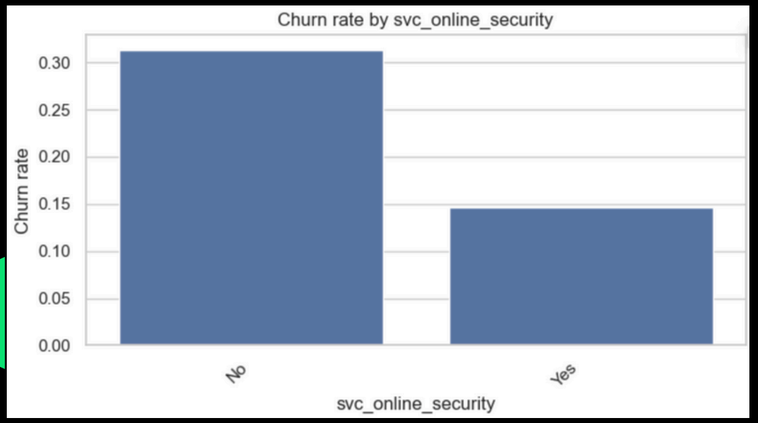
SERVICES & ADD-ONS VS CHURN

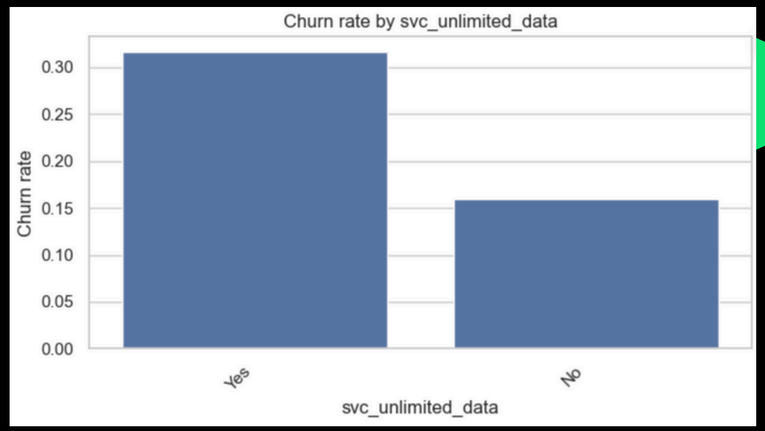


Key Findings

Internet Service & Type: Premium internet customers (especially fiber) are the most likely to leave - price-sensitive and more competition.

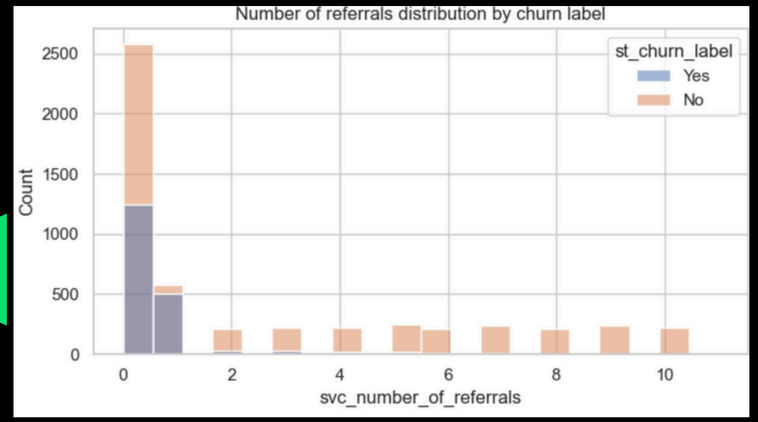
Protection & Support Services: Customers who buy security and support services are significantly more loyal.





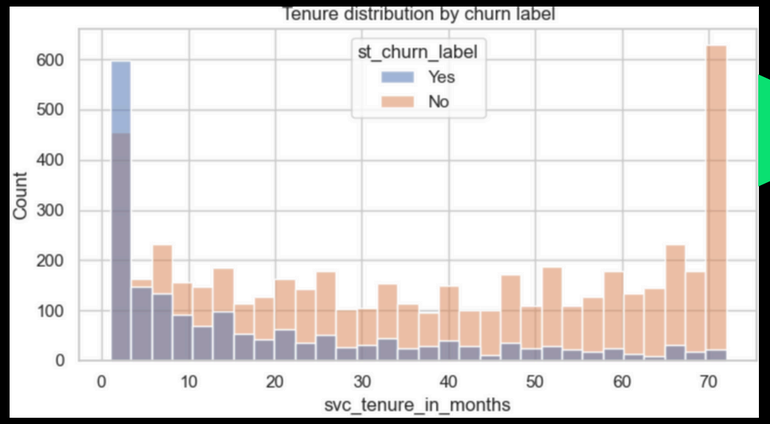
Streaming & Unlimited Data: Unlimited data signals high-usage, high-churn customers, while streaming has only a minor impact.

Referrals: Customers who refer others are much less likely to leave - social engagement strongly predicts loyalty.



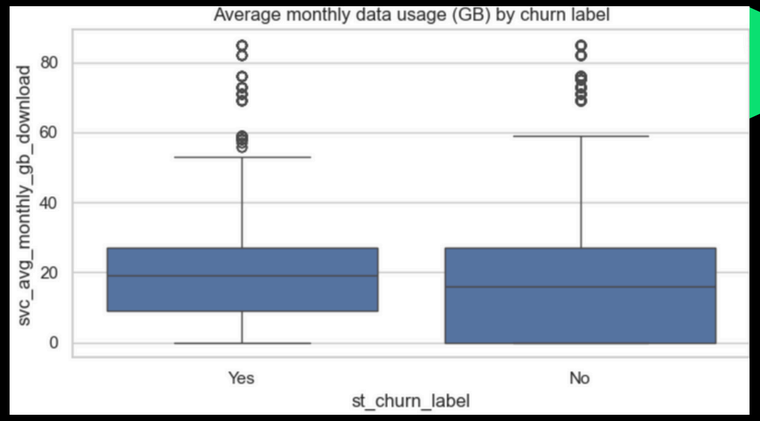
Customers who remain through the first year are significantly more likely to stay with us long-term

NUMERIC BEHAVIORAL DRIVERS VS CHURN



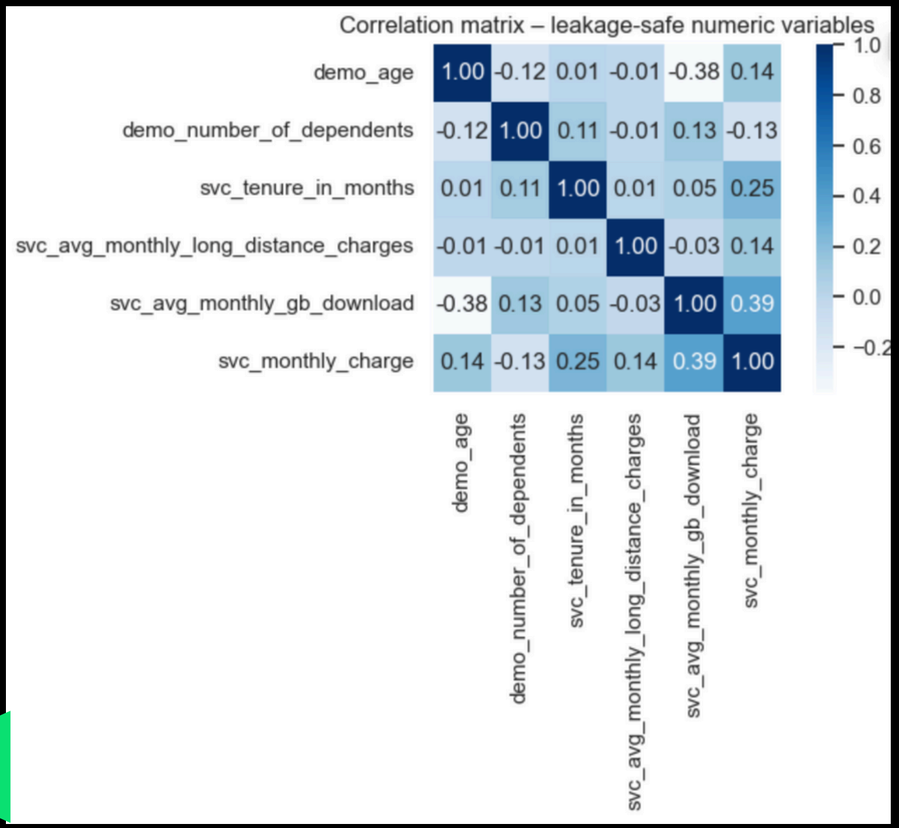
Key Findings

Tenure: Early customers are the most fragile - retention must start early in the lifecycle.



Usage & Charges: Heavy usage and higher extra charges slightly increase churn, but they are not the main drivers.

Correlation: No hidden found redundancies - each numeric variable adds unique information



Retention is influenced more by the design of the business than by how customers use the product.

WHAT ACTUALLY DRIVES CUSTOMER RETENTION?

Strongest Churn Drivers

- Flexible contracts
- Short customer history
- High monthly bills
- Billing/payment setup

Protective Factors

Customers who invest in protection services, refer others, and have stable family profiles are more likely to stay.

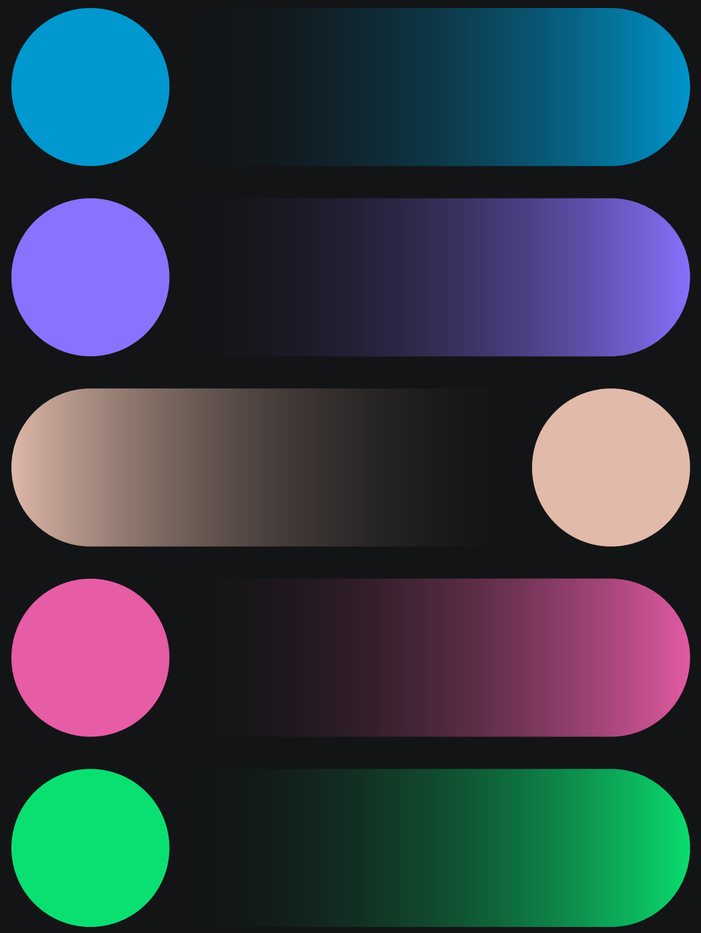
This signals:

- Product attachment
- High switching costs
- Stronger social connection

Risky Segments

High-risk profiles:

- Month-to-month with a high monthly price
- Fiber + unlimited data
- Early-tenure customers



Model

We've rewarded models that identified churners without triggering too many false alarms

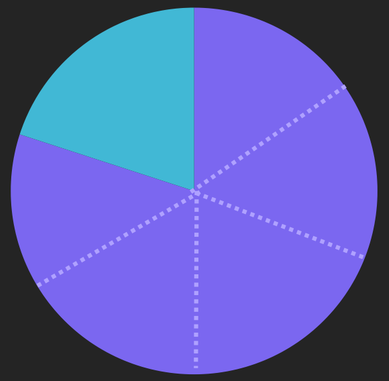
Evaluation Strategy

- **Final Hold-Out Test**

Final test is never seen during feature selection, never used in model tuning, and **used once only at the very end**

- **Internal Cross Validation**

Data split into 5 subsets, to ensure a balanced churn distribution.
Churn prediction evaluated using **F1-score** to balance catching churners early (recall) and avoiding wasted retention actions (precision)



Models Tested

- **Logistic Regression**

Simple, fast, and highly interpretable. Shows which factors drive churn and in what direction.

- **Random Forest**

Strong, stable performance. Captures non-linear patterns and interactions with high accuracy and low overfitting risk.

- **XGBoost**

Boosting model that delivers the best churn detection, especially on complex customer behavior.

Data Preparation & Governance

Customer information was standardized for model use.
Categorical data converted to numbers.

Final test set was kept isolated
Evaluation performed on untouched data to ensure reliability.

Irrelevant variables removed
to improve stability and business clarity.

To identify the most relevant drivers of churn, we've applied 3 feature selection methods

Statistical Filtering

Marital Status → **DROPPED**

- Overlap with number of dependents (multicollinearity).
- **Number of dependents shown as a more informative variable.**

Internet Service → **DROPPED**

- Redundant with monthly charge (internet always increases cost).
- **Monthly charge showed stronger predictive power in all models.**

Streaming services → **MERGED into 1** (TV, Movies, Music)

- Individually weak predictors but represent the same behavior.

Gender & Phone Service → **DROPPED**

- Low variance and minimal contribution to churn prediction.

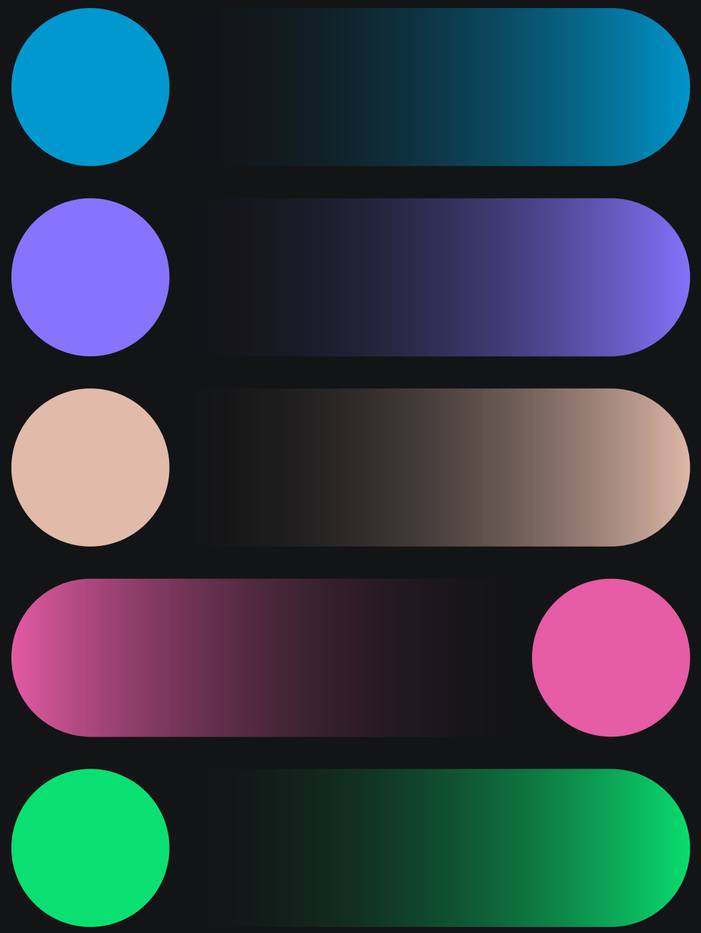
Recursive Feature Elimination (RFE)

- Applied Recursive Feature Elimination with Cross-Validation using:
  - XGBoost
  - Random Forest
  - Logistic Regression
- Each model selected the optimal number of features based on performance:
- XGBoost → 10–16 features
- Random Forest → up to 16 features
- Logistic Regression → 8–12 features

Consensus feature set:
Only the 11–14 variables that consistently improved prediction across multiple models were retained.

Model-Based Importance

- Variables were ranked based on their **contribution to prediction accuracy**
- **Key Result**: the analysis confirmed the core drivers of churn as:
  - **contract type**
  - **tenure**
  - **number of referrals**
  - **internet type**
  - **monthly charges**
- These variables consistently showed the strongest predictive impact across models.
- Strong presence of **service-related features** (security, backup, tech support) and **payment-method features** suggests **churn is strongly influenced by service bundle depth and that payment behavior matters.**



Results

Predicting Churn Before It Happens: Can We Keep the Next Customer?

Model	F1	Precision	Recall	ROC AUC	PR AUC
🏆 XGB	0.707447	0.703704	0.711230	0.901150	0.767324
RF	0.674487	0.746753	0.614973	0.897845	0.753698
LOG	0.689076	0.723529	0.657754	0.900814	0.762519

	Actually Stayed	Actually Churned
Predicted "Stay"	✅ Correctly ignored	🔴 Missed churners
Predicted "Churn"	🟠 Unnecessary offers	✅ Correct early saves

What Does This Mean in Practice?

Out of 100 customers who **WOULD** churn:

🟢 ~70 are correctly flagged in advance
🔴 ~30 are missed (leave anyway)

Business Impact:

✅ This means we can proactively intervene with the majority of customers who are truly at risk, instead of reacting after they've already left.

🟠🔴 YES, some churners will inevitably be missed, and a small number of stable customers may receive unnecessary offers - this is the natural cost–benefit trade-off of any predictive model."

We’ve moved from mass, untargeted discounting to precise, data-driven retention

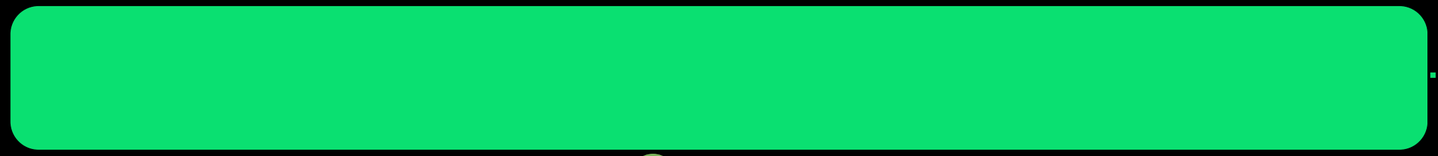
High Risk:
Month-to-Month / High monthly charge / Short tenure
Fiber / Unlimited Data

Business Actions

- Short-term loyalty bonus (3–6 months)
- Targeted bill optimization (“cheaper plan found!”)
- Usage-based plans
- Temporary discounts/bundled offers (internet + mobile).

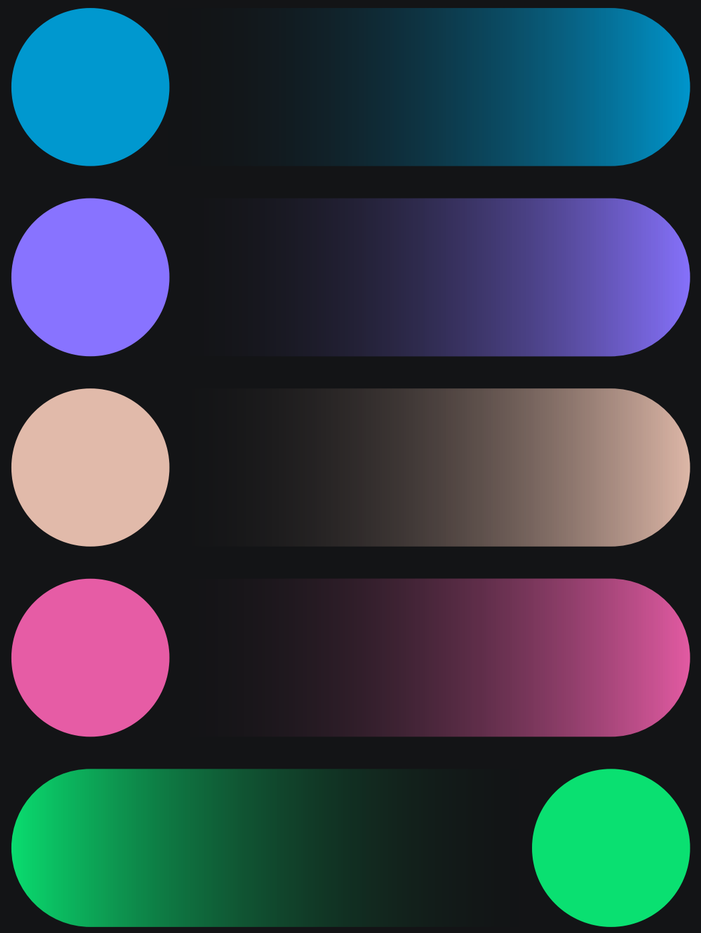
Medium Risk
Moderate price / Some add-ons / Mid-tenure

- Streaming / data vouchers
- Add-on bundles (security, backup)
- Contract renewal incentives



Low Risk
Long tenure / Multiple services / Referrers

- Loyalty rewards (extra data, streaming vouchers).
- Refer-a-friend credits
- Upselling premium services



Discussion

Auxiliary Section

A. Data Sources & Integration

B. Leakage Control & Governance

C. Data Cleaning Summary

D. Feature Engineering Summary

E. Model Configuration Details

F. Full Evaluation Metrics

APPENDIX A: DATA PIPELINE & SOURCES

```
===== demographics =====
Shape: 7043 rows × 9 columns

Data types:
Customer ID              object
Count                     int64
Gender                   object
Age                       int64
Under 30                 object
Senior Citizen           object
Married                  object
Dependents               object
Number of Dependents      int64
```

```
===== location =====
Shape: 7043 rows × 9 columns

Data types:
Customer ID     object
Count            int64
Country         object
State           object
City            object
Zip Code         int64
Lat Long        object
Latitude       float64
Longitude      float64
```

```
===== population =====
Shape: 1671 rows × 3 columns

Data types:
ID              int64
Zip Code        int64
Population      int64
```

```
===== status =====
Shape: 7043 rows × 11 columns

Data types:
Customer ID            object
Count                   int64
Quarter                object
Satisfaction Score      int64
Customer Status        object
Churn Label            object
Churn Value             int64
Churn Score             int64
CLTV                    int64
Churn Category         object
Churn Reason           object
```

```
===== services =====
Shape: 7043 rows × 30 columns

Data types:
Customer ID                        object
Count                               int64
Quarter                            object
Referred a Friend                  object
Number of Referrals                 int64
Tenure in Months                    int64
Offer                              object
Phone Service                      object
Avg Monthly Long Distance Charges  float64
Multiple Lines                     object
Internet Service                   object
Internet Type                      object
Avg Monthly GB Download             int64
Online Security                    object
Online Backup                      object
Device Protection Plan             object
Premium Tech Support               object
Streaming TV                       object
Streaming Movies                   object
Streaming Music                    object
Unlimited Data                     object
Contract                           object
Paperless Billing                  object
Payment Method                     object
Monthly Charge                    float64
Total Charges                     float64
Total Refunds                     float64
Total Extra Data Charges            int64
Total Long Distance Charges       float64
Total Revenue                     float64
```

Final dataset size confirmation:

7,043 customers

56 raw → 31 clean features

Merge tables

Four customer-level datasets (demographics, location, services, and status) can be merged one-to-one using Customer ID. The population dataset is auxiliary and used for enrichment via Zip Code.

APPENDIX B: LEAKAGE CONTROL & DATA GOVERNANCE

```python
target = "st_churn_label"

leakage_target_cols = [
    "st_count",
    "st_quarter",
    "st_satisfaction_score",
    "st_customer_status",
    "st_churn_value",
    "st_churn_score",
    "st_cltv",
    "st_churn_category",
    "st_churn_reason",
]

df_clean = df_clean.drop(columns=leakage_target_cols)

print("✅ Dropped target-related leakage columns:", leakage_target_cols)
print("✅ Remaining shape:", df_clean.shape)
print("✅ Target distribution:")
print(df[target].value_counts())
```

```
✅ Dropped target-related leakage columns: ['st_count', 'st_quarter', 'st_satisfaction_score', 'st_customer_statu
s', 'st_churn_value', 'st_churn_score', 'st_cltv', 'st_churn_category', 'st_churn_reason']
✅ Remaining shape: (7043, 32)
✅ Target distribution:
st_churn_label
No     5174
Yes    1869
```

APPENDIX C: DATA CLEANING SUMMARY

Feature	Action	Notes
demo_gende	Kept	
demo_age	Kept	
demo_under_30, demo_senior_citizen	Dropped	These variables are direct transformations of demo_age. To preserve the most informative representation of life stage, only demo_age was retained.
demo_married, demo_number_of_dependents	Kept	
demo_dependents	Dropped	This variable describes the same underlying concept as demo_number_of_dependents. Since the numeric feature provides richer information, it's the only one we decide to keep.
demo_count — Record/count indicator (non-behavioral	Dropped	Record/count indicator (non-behavioral)

APPENDIX C: DATA CLEANING SUMMARY

Feature	Action	Notes
loc_country, loc_state,	**Dropped**	Hierarchically redundant
loc_zip_code	**Kept**	
loc_latitude, loc_longitude	**Dropped**	concatenation of latitude & longitude - created multicollinearity.
loc_lat_long	**Kept**	

APPENDIX C: DATA CLEANING SUMMARY

Feature	Action	Notes
svc_tenure_in_months, svc_offer	Kept	
svc_referred_a_friend, svc_number_of_referrals, svc_contract	Kept	
svc_phone_service, svc_multiple_lines	Kept	
svc_internet_service, svc_internet_type, svc_unlimited_data, svc_avg_monthly_long_distance_charges, svc_avg_monthly_gb_download	Kept	
svc_streaming_tv, svc_streaming_movies, svc_streaming_music	Kept	
svc_online_security, svc_online_backup	Kept	
svc_paperless_billing, svc_payment_method, svc_monthly_charge	Kept	
svc_total_charges, svc_total_refunds, svc_total_extra_data_charges, svc_total_long_distance_charges, svc_total_revenue	Dropped	High leakage risk: likely to encode information after the churn event. (Low total revenue → customer must have churned early) (High total revenue → customer stayed longer)
svc_quarter, svc_count	Dropped	Metadata

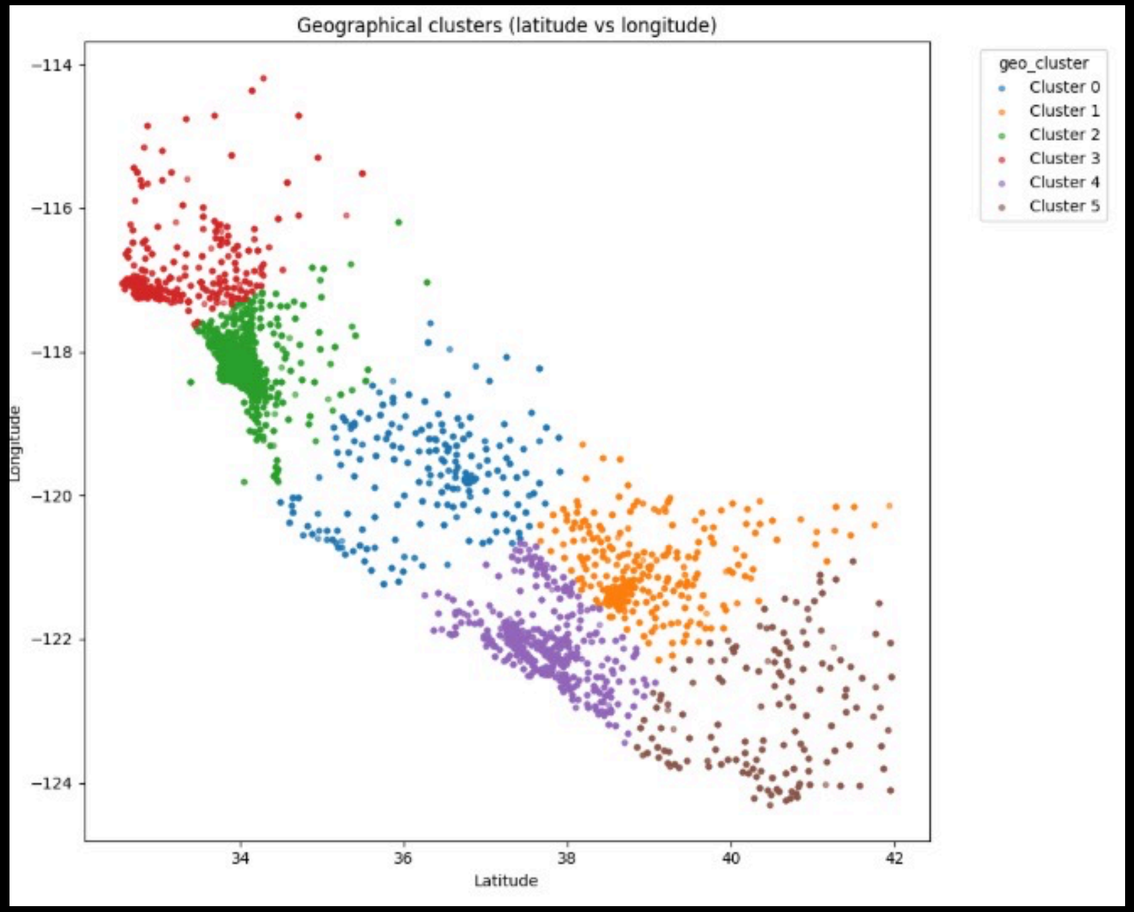
APPENDIX C: DATA CLEANING SUMMARY

Feature	Action	Notes
st_churn_label	Kept	Target
st_customer_status, st_churn_value, st_churn_score, st_cltv, st_churn_category, st_churn_reason, st_satisfaction_score	Dropped	Using these variables in training will cause data leakage and unrealistically high model performance.
st_count, st_quarte	Dropped	Metadata

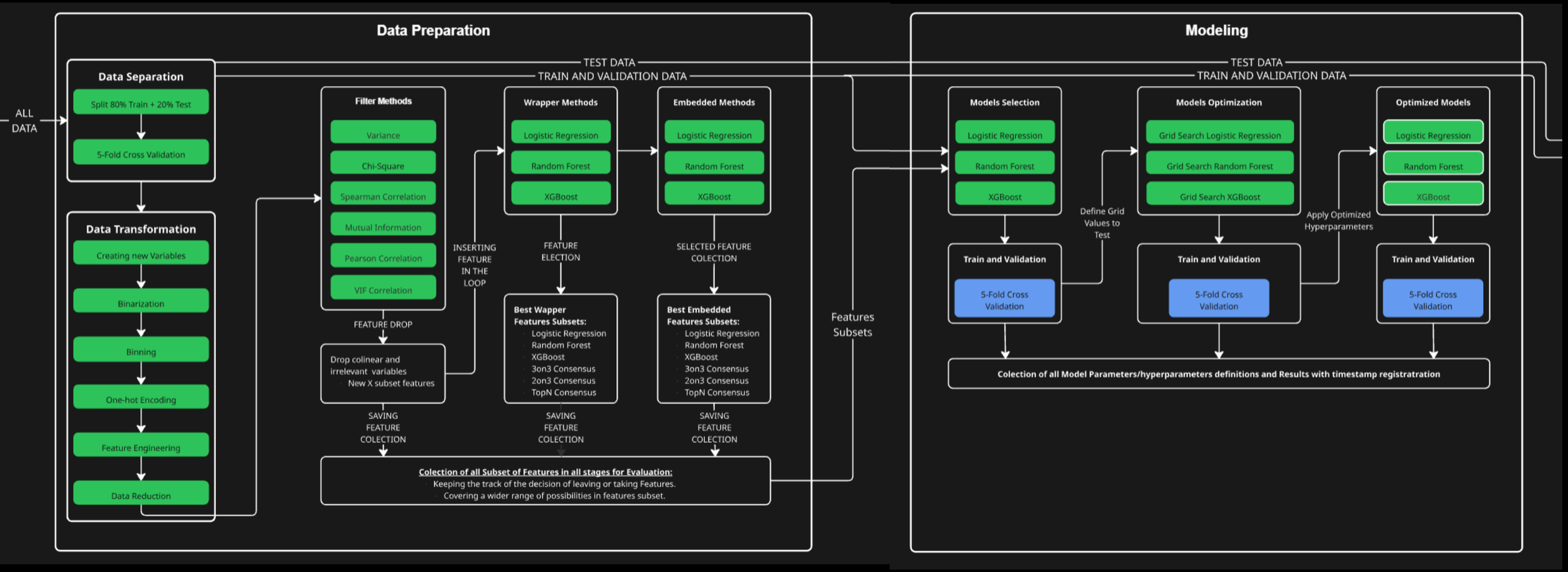
APPENDIX D: FEATURE ENGINEERING SUMMARY

Feature	Action	Notes
is_woman, has_dependents	Created	{"Male": 0, "Female": 1} {"No": 0, "Yes": 1}
demo_married, svc_referred_a_friend, svc_phone_service, svc_multiple_lines, svc_internet_service, svc_online_security, svc_online_backup, svc_device_protection_plan, svc_premium_tech_support, svc_streaming_tv, svc_streaming_movies, svc_streaming_music, svc_unlimited_data, svc_paperless_billing	Binarization	
svc_number_of_referrals svc_avg_monthly_gb_download	Binning	Referral_bin: {0: "0", 1: "1", 2: "2-6", 3: "7+"} GB_download_bin: {0: "0",1: "1–10",2: "11–30",3: "31–60",4: "60+"}
svc_internet_type, svc_contract, payment method, svc_offer	One-Hot Encoded	
streaming_any	Created	Merged ["svc_streaming_tv", "svc_streaming_movies", "svc_streaming_music"]
geo_cluster	Created	grouped loc_longitude & loc_latitude in clusters via K-Means
demo_gender, loc_city, svc_offer, demo_number_of_dependents, svc_number_of_referrals, loc_longitude, loc_latitude, loc_zip_code, svc_avg_monthly_gb_download	Eliminated	Redundant variables

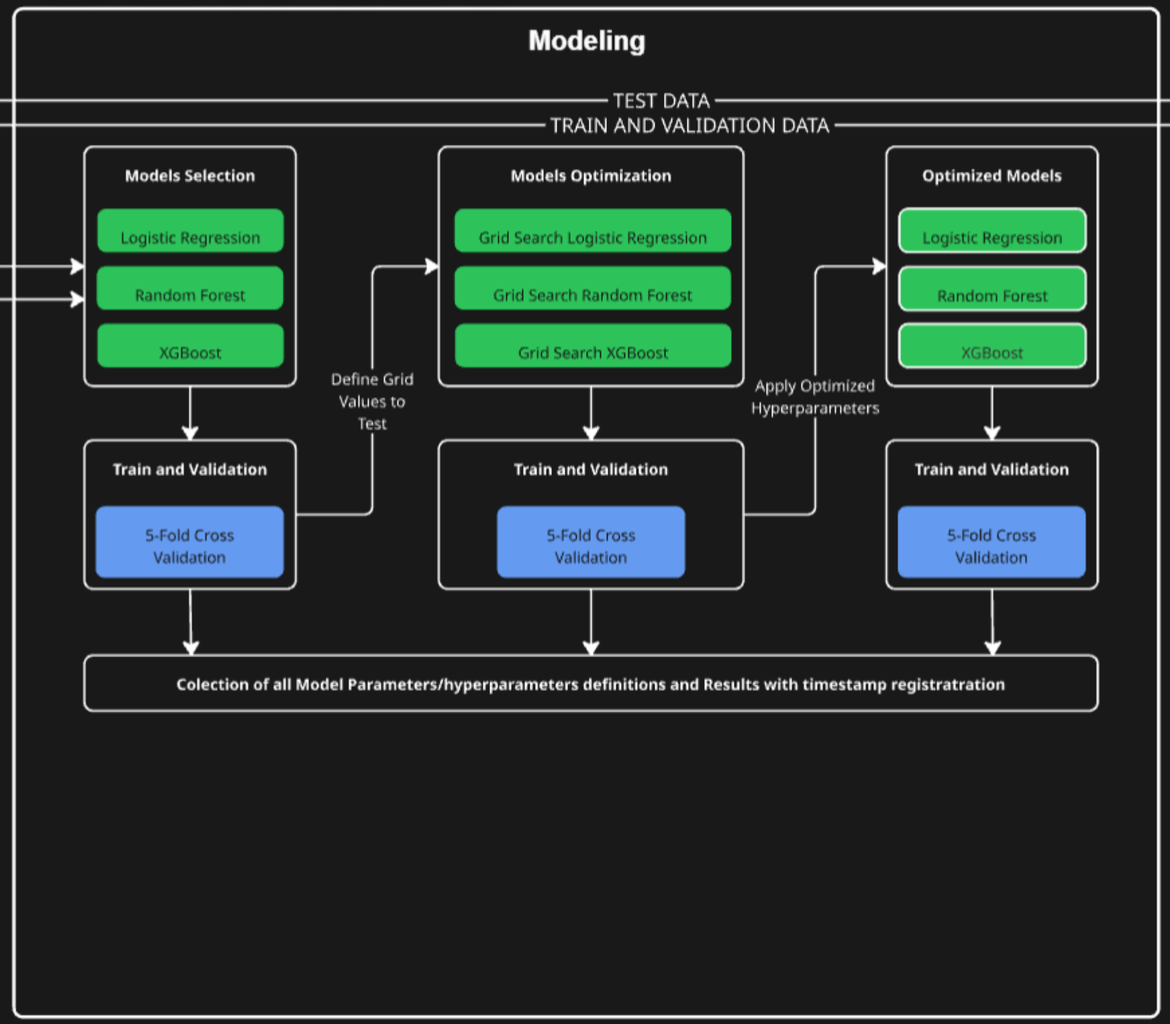
APPENDIX D: FEATURE ENGINEERING SUMMARY

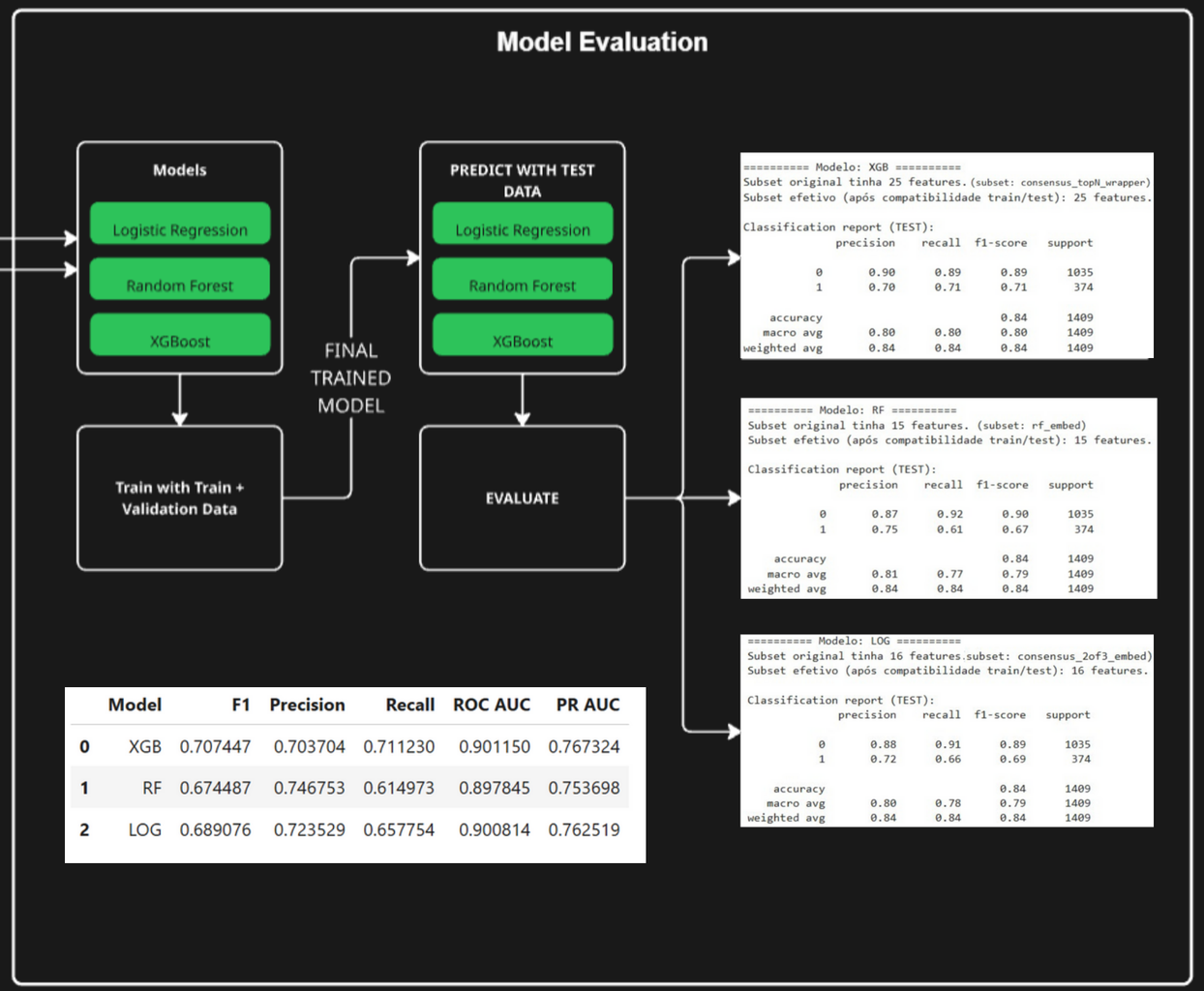


APPENDIX E: MODELING CONFIGURATION DETAILS



APPENDIX E: MODELING CONFIGURATION DETAILS






```
========== Modelo: XGB ==========
Subset original tinha 25 features. (subset: consensus_topN_wrapper)
Subset efetivo (após compatibilidade train/test): 25 features.

Classification report (TEST):
              precision    recall  f1-score   support

           0       0.90      0.89      0.89      1035
           1       0.70      0.71      0.71       374

    accuracy                           0.84      1409
   macro avg       0.80      0.80      0.80      1409
weighted avg       0.84      0.84      0.84      1409
```

```
========== Modelo: RF ==========
Subset original tinha 15 features. (subset: rf_embed)
Subset efetivo (após compatibilidade train/test): 15 features.

Classification report (TEST):
              precision    recall  f1-score   support

           0       0.87      0.92      0.90      1035
           1       0.75      0.61      0.67       374

    accuracy                           0.84      1409
   macro avg       0.81      0.77      0.79      1409
weighted avg       0.84      0.84      0.84      1409
```

```
========== Modelo: LOG ==========
Subset original tinha 16 features.subset: consensus_2of3_embed)
Subset efetivo (após compatibilidade train/test): 16 features.

Classification report (TEST):
              precision    recall  f1-score   support

           0       0.88      0.91      0.89      1035
           1       0.72      0.66      0.69       374

    accuracy                           0.84      1409
   macro avg       0.80      0.78      0.79      1409
weighted avg       0.84      0.84      0.84      1409
```

	Model	F1	Precision	Recall	ROC AUC	PR AUC
0	XGB	0.707447	0.703704	0.711230	0.901150	0.767324
1	RF	0.674487	0.746753	0.614973	0.897845	0.753698
2	LOG	0.689076	0.723529	0.657754	0.900814	0.762519

APPENDIX E: MODELING CONFIGURATION DETAILS

Accuracy

Overall % of correct predictions

(can be misleading in churn because most customers do not churn - not used)

Recall

Of all real churners, how many did we correctly detect?

(high recall = fewer missed churners).

Precision

Of all customers flagged as churners, how many truly churn?

(high precision = less wasted money on unnecessary retention offers)

F1 Score

Single score that balances Recall and Precision.

PR AUC

How reliable churn predictions are under class imbalance.

APPENDIX E: FULL EVALUATION METRICS