Technical Report:

Customer Churn Prediction and Retention Drivers

By: SOLOMON BIBINU

1. Problem Definition and Business Context

Customer churn represents a direct and recurring revenue loss. In this specific dataset, I identified a **high-risk cohort (Month-to-Month Fiber users) representing over $70,000 in monthly recurring revenue exposure.**

When a customer leaves, the business not only loses future cash flow but often incurs additional costs for new customer acquisition. As a result, reducing churn is typically far more cost-effective than driving new customer acquisition.

Churn is defined as a binary outcome indicating whether a customer discontinued the service during the observed period. The objective thereby is to predict which active customers are at the highest risk of churning and to identify behavioral and contractual factors that meaningfully influence that risk.

From a business perspective, failing to identify a churner is more costly than incorrectly flagging a non-churner. Therefore, my objective while modeling prioritized **high recall on churners**, while maintaining a sufficient precision which allows targeted retention efforts to remain operationally feasible.

1. Problem Definition and Business Context

The intended use of this model is **decision making support**: enabling retention teams to proactively engage a subset of high-risk customers with targeted interventions. Consequently, model interpretability and stability were emphasized over marginal gains in predictive performance.

2. Dataset Overview

The dataset contains 7,043 customers with 20 features, including demographics, service usage patterns, contract details, and billing information.

The target variable, Churn, is Boolean, indicating whether a customer discontinued the service (Yes) or remained active (No) during the observed period.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No

TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	No	No	One year	No	Mailed check	56.95	1889.5	No
No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

2. Dataset Overview

Features were grouped into four main categories:

- **Demographics** (e.g., tenure, age, gender),
- **Service usage** (e.g., Internet Service type, Tech Support, Online Security),
- **Contract and Billing (e.g.** Contract type, monthly charges, total charges), and
- **Derived metrics such as ARPU-related features.**

Redundant and highly correlated columns (e.g. **Customer ID**) were removed to avoid inflating model complexity, including derived totals and simple bins. Several engineered features were tested but excluded due to minimal predictive gain, such as **Stickiness Score** and **Charges_by_Tenure**, demonstrating a deliberate and disciplined feature selection process.

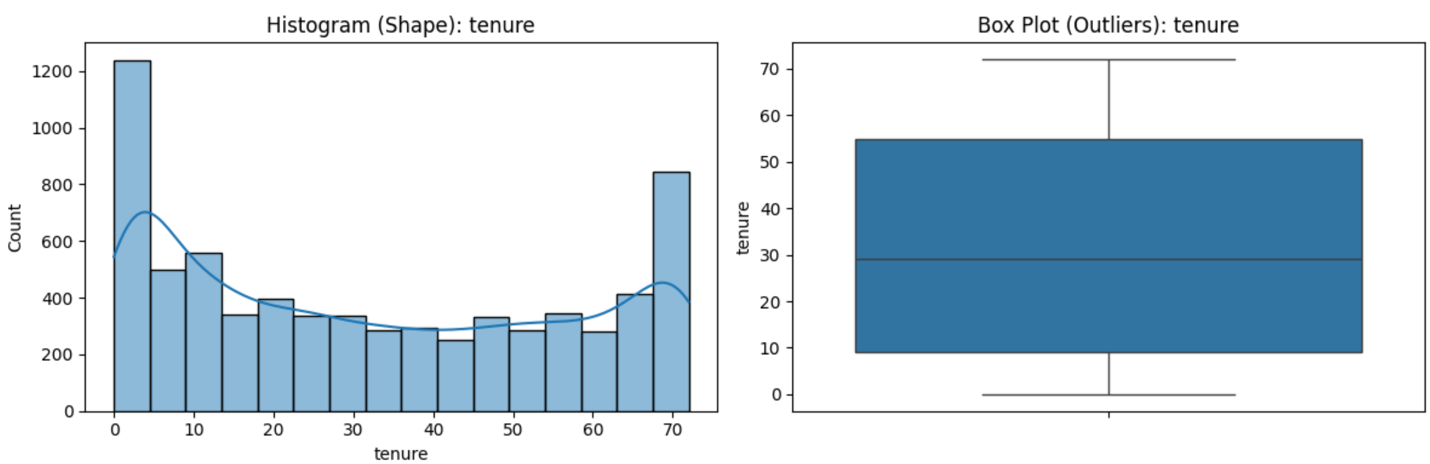
The dataset exhibits a churn rate of approximately 26%, creating a moderately imbalanced classification problem that informed subsequent modeling choices.

3. Exploratory Data Analysis (EDA)

3.1 Methodology:

To gain a clear understanding of the data and guide feature engineering, a structured EDA approach was employed:

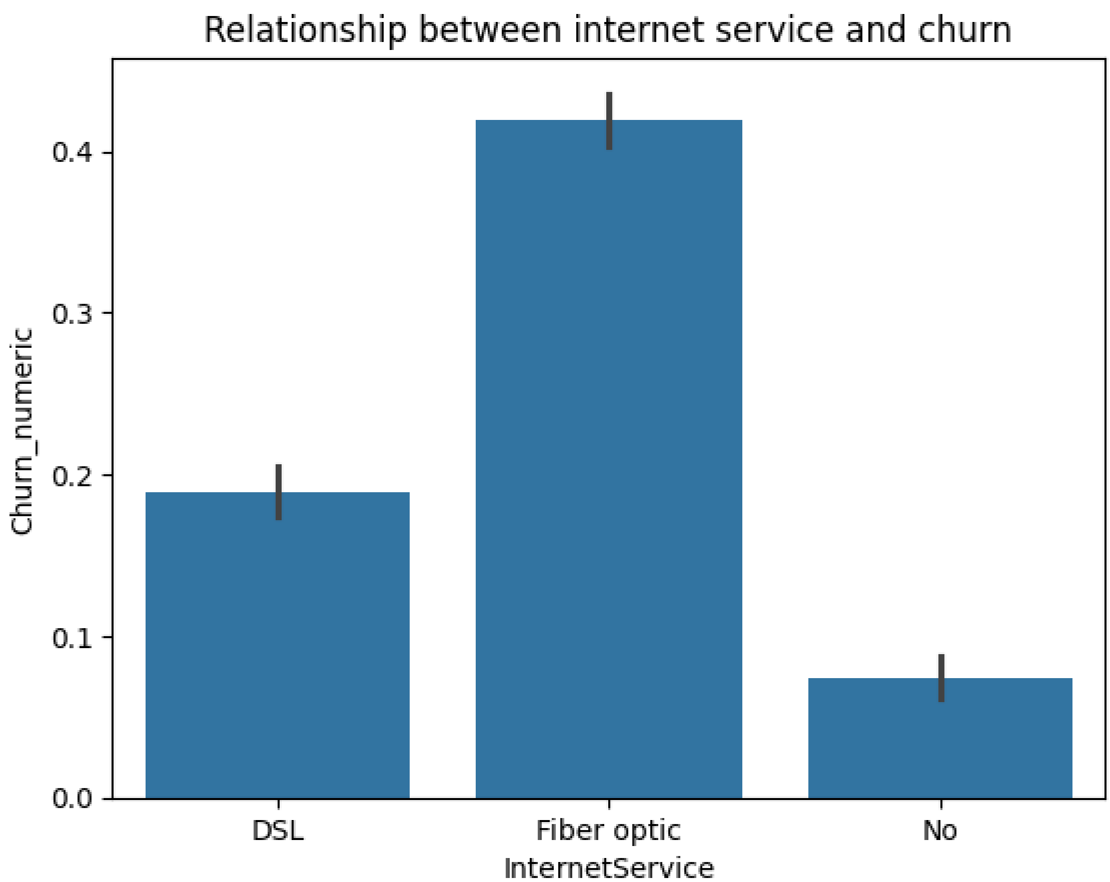
- ***Univariate Analysis***:
  Examined the distribution of each feature individually to detect anomalies, skew and missing values. For categorical features, frequency counts and proportions were analyzed. For continuous features, histograms, boxplots and descriptive statistics were used

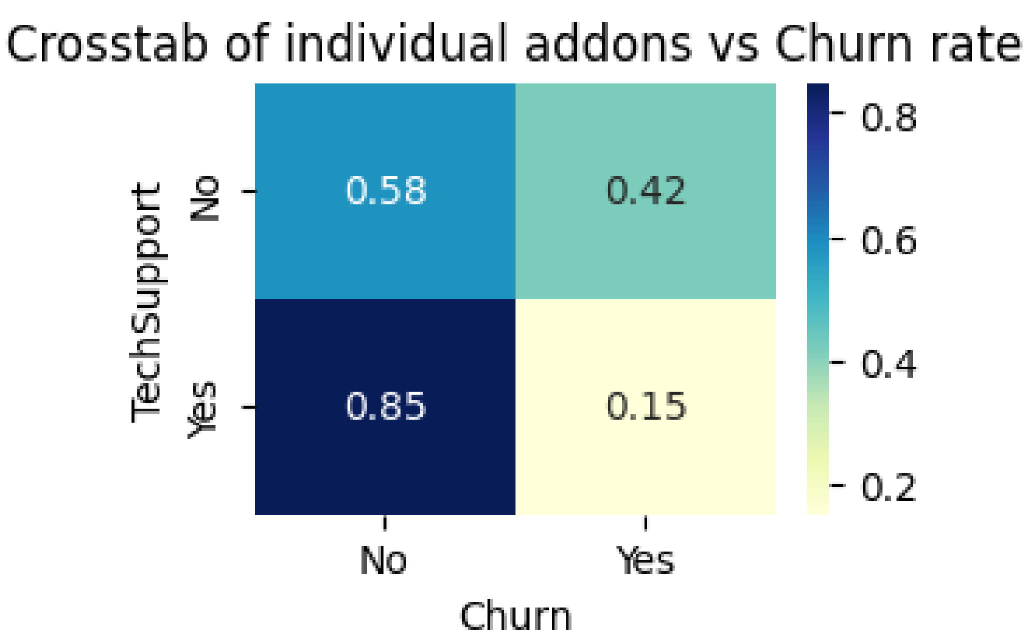


Tenure plots, this is a regular distribution type for this type of data

- ***Bivariate Analysis***:
  Explored relationships between each feature and the target variable, Churn. For categorical features, churn rates were compared across groups. While continuous features were visualized against churn to identify trends

3. Exploratory Data Analysis (EDA)





3. Exploratory Data Analysis (EDA)

- ***Multivariate Analysis:***
  Investigated interactions between multiple features and their combined effect on churn. The analysis was structured around key business questions:

  1. ***Payment method vs. churn**:*
     Examined which customer segments primarily use e-checks and whether this correlates with higher churn rates
  2. ***Customer Stickiness:***
     Identified the types of users exhibiting higher engagement or loyalty ("stickiness") and explored factors contributing to this behavior.
  3. ***High Spender Churn:***
     Analyzed why approximately 70% of high-ARPU customers leave within a year, seeking patterns in service usage, contracts, and support features.
  4. ***Contract Tenure vs. conversion:***
     Investigated why customers on month-to-month contracts for several years do not convert to longer-term contracts, identifying potential barriers to retention.
  5. ***Service Type Concerns:***
     Assessed whether specific service types, such as fiber optic internet, inherently contribute to elevated churn risk.

3. Exploratory Data Analysis (EDA)

This structured approach allowed for targeted insights that directly informed feature engineering and model design.

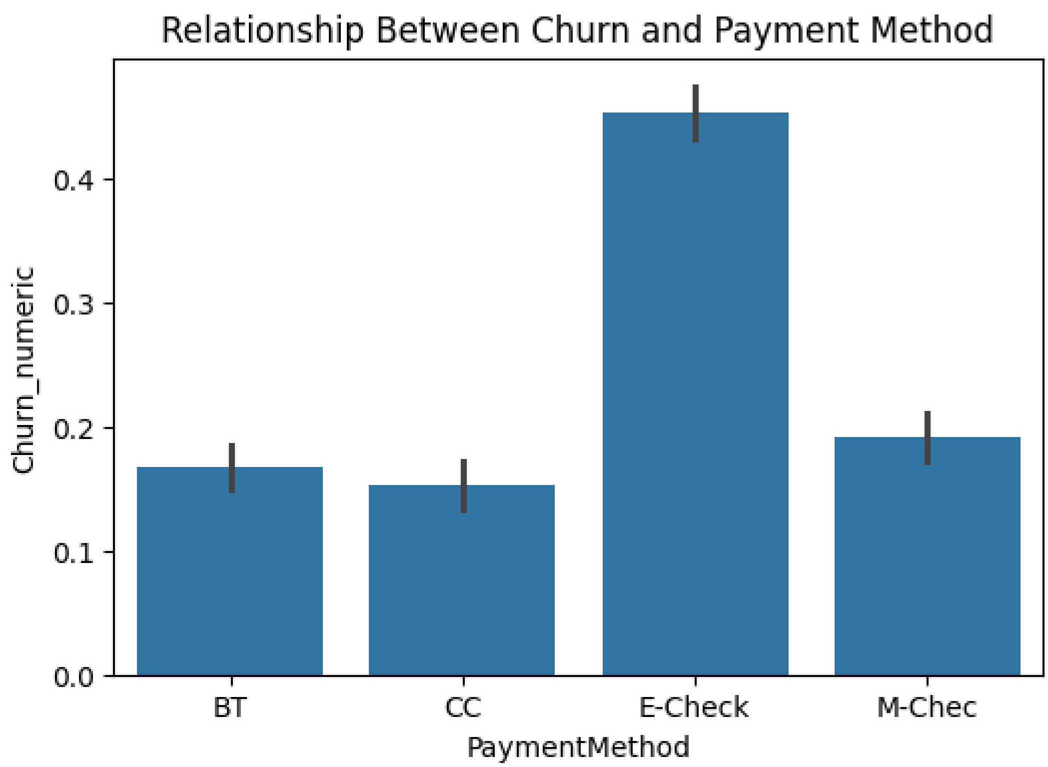
3. Exploratory Data Analysis (EDA)

3.2 Insights:

Following the multivariate analysis framework, the following critical patterns were identified:

- **Insight 1: The "E-Check" proxy effect:**

    - **Finding:** Electronic Check users exhibit a **churn rate of ~45%**, significantly higher than the average churn rate of ~15-18% for automatic payment methods (Credit Card/Bank Transfer), and more than double the churn rate of Mailed Checks ~20%



    - **Root-Cause Analysis:** Multivariate analysis revealed that this was a proxy for inherently risky customers, not a payment method issue.
        - **~67%** of E-Check users are subscribed to fiber-optic (our highest churning product)

3. Exploratory Data Analysis (EDA)

HighRisk_Fiber	False	True
PaymentMethod		
BT	0.581606	0.418394
CC	0.607753	0.392247
E-Check	0.325581	0.674419
M-Chec	0.839950	0.160050

Crosstab showing the distributions of payment methods and Subscription to fiber optic

- **~78%** of E-Check users are on month to month contracts

HighRisk_Contract	False	True
PaymentMethod		
BT	0.618523	0.381477
CC	0.643233	0.356767
E-Check	0.217759	0.782241
M-Chec	0.446030	0.553970

Crosstab showing the distributions of payment methods and contracts

- **~66%** of E-Check users also exhibit low stickiness (defined as <= 1 sticky service)

HighRisk_LowStickiness	False	True
PaymentMethod		
BT	0.510363	0.489637
CC	0.519054	0.480946
E-Check	0.335729	0.664271
M-Chec	0.254963	0.745037

Crosstab showing the distributions of payment methods and Stickiness.

- **Business Implication:** Simply changing payment options will yield limited ROI. Retention efforts must focus on migrating these users to longer-term contracts or bundling them with sticky services

- **Insight 2: The “Stickiness” Ecosystem and the Retention Wall:**
  - **Finding:** “Stickiness” (adoption of ancillary services) is not evenly distributed
    - **Demographic Driver:** Users with partners have a 50% higher stickiness score (mean of 1.53) than single users (mean of 1.01)

```
Partner
No     1.014831
Yes    1.534098
Name: StickinessScore, dtype: float64
```

Crosstab showing the distributions of stickiness and partner status

    - **The “Retention Wall”:** Stickiness does not grow linearly. It remains sluggish (0.55) -> (0.94) for the first 24 months and then **nearly doubles** (to 1.76) after the 2-year mark

```
...   tenure_group
      0 - 12 months     0.555402
      12 - 24 months    0.939453
      24+ months        1.755283
      Name: StickinessScore, dtype: float64
```

Crosstab showing the distribution of stickiness and tenure

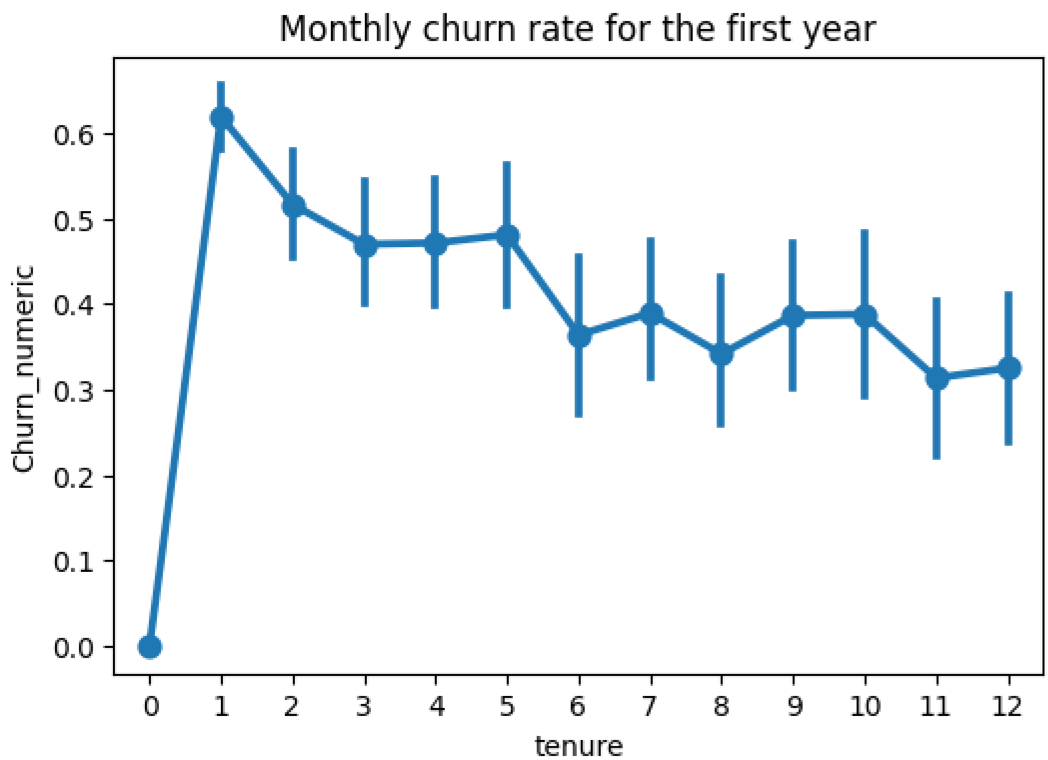
  - **The “Protection Gap”:** There is a critical misalignment in our product tiers
    - **DSL users (Lower ARPU):** Highly embedded with “Retention Anchors” (Tech support/security)

3. Exploratory Data Analysis (EDA)

    - **Fiber users (Highest ARPU): ~56%** have Streaming (Entertainment), but only **~27%** have Online Security. We are selling our most expensive product with the weakest "lock-in" features.
  - **Business Implication:**
    - **Immediate Action:** Launch a **"Year 2 Anniversary"** campaign to pull stickiness forward into the 12-18 month window, rather than waiting for organic adoption**.**
    - **Strategic Shift:** Stop selling Fiber as just "Fast Internet." Sell it as "Secure Internet" by defaulting new Fiber sign-ups into a "Security First" bundle (Tech Support + Online Security) for the first 3 months to reduce early-life churn.

- **Insight 3: The Month to Month paradox (High value, High risk)**
  - **Finding: Month-to-Month** users have the highest average revenue per user **(ARPU)** ($66 vs. $60 for 2-year-contracts). However, this segment is structurally unstable due to two distinct failure modes.
    - **Failure Mode A: The Month 1 “cliff” (Early Failure)**
      - **Data: ~60%** of M2M users churn in the very first month with a further decline of **~50%**, the next month however from here onwards, where users are statistically less likely to churn than to churn. However, at this point a **vast number of the cohort is gone already (from 100% down to 30%)** with a churn rate of **~45%** until month 6

3. Exploratory Data Analysis (EDA)



- **Root Cause:** This immediate exit suggests an onboarding failure (expectations mismatch or technical issues) rather than price sensitivity.

▪ **Failure Mode B: The “Vulnerable Veteran” (Late Failure)**
  - **Data:** A sub-segment of users with >24 months tenure remain on M2M contracts. They pay a premium **(~$79 ARPU) monthly**, yet **91%** of them are “unanchored” (missing tech support/online security).
  - **The Anchor Gap:** While 79% of these users purchase *some* add-ons (mostly Streaming/Entertainment), **91% are "Unanchored"** (lacking Tech Support or Online Security)
  - **Risk:** While these users are loyal, with them having – 0 stickiness score only 21% of the time, their churn

3. Exploratory Data Analysis (EDA)

        (45%) risk remains **2.5x higher** than anchored users **(18%),** despite their loyalty

  - **Business Implication:**
    - **Fix the Onboarding:** Audit onboarding flows for possible leaks and shift retention resources to days 0-60 as this has proven to be our most volatile period of the customer journey. Mandatory “Setup support” calls for new M2M users may mitigate the month 1 cliff.
    - **Anchor the Veterans:** Do not force the vulnerable veterans into long contracts (which adds friction). Instead, incentivize them to add **one anchor service** (e.g., "Free Tech Support Upgrade") to lock them in via product utility.

- **Insight 4: The “Unprotected” High spenders**
  - **Finding:** High-spending new customers (>$75/mo.) represent a **"High Value / High Risk"** paradox, churning at a rate of **~70% within the first year**.
    - **Spend vs. Churn:** The high churn correlates linearly with increased spend, indicating this is not a budget issue (lack of funds) but a value realization issue (lack of perceived utility.)
  - **The “Stickiness” void:**
    - **Data:** A staggering **82%** of these high-value users fall into the “Low-Stickiness” zone **(~55% chance of churn)** with 0-1 sticky services.
    - **Risk Multiplier:**
      - Users with **0 stickiness** churn at **65%**

3. Exploratory Data Analysis (EDA)

            - Adding just **one** service drops churn to **46%**
        - **Conclusion:** We are successfully upselling them on price, but failing to cross-sell them on **Retention Anchors.** We have successfully captured their wallet, but not their loyalty
    - **Service Spread Analysis (The Adoption Mismatch):**
        - **Finding:** First year high spenders are spending money on the ***wrong*** ancillary services. There is a massive misalignment between what they **buy** and what actually **retains** them.
            - **The "Placebo" Products:**
                - **High Adoption:** Online Backup (~24%) and Device Protection (~29%) are the most popular add-ons.
                - **Low Impact:** These services are statistically ineffective at preventing churn for this cohort (Churn remains extremely high at **~67-68%** even with them.)
            - **The "Real" Anchors:**
                - **Low Adoption:** Only **~13%** subscribe to Online Security and **~14%** to Tech Support.
                - **High Impact:** These services provide a statistically significant "Retention Shield", dropping churn risk to **~54-57%**
        - **Conclusion:** We have a **15 point adoption gap.** Out highest value customers are opting for passive "insurance" (backup/device protection) rather than active "utility"

3. Exploratory Data Analysis (EDA)

(security/support). This suggests our onboarding flow pushes the wrong upsells

- **Business Implication:**
    - **Re-Rank Upsells:** Update the sales script and website UI to prioritize **Online Security** and **Tech Support** as the default add-ons for high-tier plans.
    - **Bundle Logic:** Create a new “Customer Safety Pack” that bundles tech support with device protection. Use the popularity of the device protection to “Trojan Horse” the Tech support service that actually makes them stay

4. Feature Engineering

Driven by the strategic insights from the EDA, the feature engineering process focused on capturing **behavioral risk** rather than just raw metrics. The goal was to transform business logic (e.g. “The Protection Gap”) into model-ready inputs.

4.1. Retained Engineered Features:

- **Is_Anchored:** A Boolean flag, identifying users subscribed to both ***Tech Support*** and ***Online Security***.
    - *Rationale:* EDA showed that while “Stickiness” (count of services) was generally good, these two services specifically acted as a distinct “Retention Shield”, whereas other services (like streaming) do not.

4.2. Tested but Dropped Features:

- **StickinessScore:** Dropped. The model performed better with the specific binary **is_anchored** flag than a raw count of services, further proving that the quality of service matters more than the quantity of services.
- **High_ARPU_Newbie:** Dropped. The model successfully inferred this risk profile from the interaction of **MonthlyCharges** and **Tenure,** without needing a hard-coded flag. Testing this feature also slightly dropped both recall and precision by 0.1, hence it was binned
- **Charges_by_Tenure:** Dropped due to introduced noise

4.3 Features for Analysis only (EDA)

Certain features were created strictly to aid visual analysis and business segmentation but were excluded from the training pipeline to avoid data redundancy (multicollinearity)

4. Feature Engineering

- **Binning Features:** *tenure_group, MonthlyCharges_group*
- **Risk Flags:** *HighRisk_Fiber, HighRisk_Contract, HighRisk_LowStickiness*
- **Rationale:** These concepts were already captured by the model through the raw features (*tenure, InternetService, Contract)* and their coefficients. Feeding the bins alongside the raw data would cause information duplication

5. Data Preprocessing

To ensure model generizability and prevent data leakage, a rigorous preprocessing pipeline was implemented prior to training.

- **5.1 Manual Binary Encoding:**

  Variables with a simple Yes/No structure (e.g., *Partner, PhoneService, is_anchored),* were manually mapped to binary format (Yes=1, No=0).

  This ensured consistent handling of custom Boolean flags created during feature engineering
- **5.2 One-Hot Encoding(with Multicollinearity Prevention):**
  - Nominal variables(e.g., *PaymentMethod, InternetService)* were converted using pd.get_dummies
  - The parameter drop_first = True was applied. This removes the first category of each variable.
  - **Rationale:** This prevents **Perfect Multicollinearity (**The Dummy Variable Trap), which is critical for linear models like Logistic Regression to function correctly.
- **5.3 Train-Test Split (Stratified):**
  - The dataset was split into **training (80%)** and **testing (20%)** sets.
  - Stratification was used to ensure the churn distribution (26%) remained consistent across both sets, preventing a non-representative test set
- **5.4 Feature Scaling:**
  - **StandardScaler** was applied to the continuous features (*tenure, MonthlyCharges)* to normalize their range. This ensures numerical stability and allows coefficient magnitudes to be meaningfully compared.

5. Data Preprocessing

- **Leakage Prevention:** The scaler was **fitted only on the training set** and then applied to the test set. This ensures that statistical information from the test set (mean/variance) did not leak into the model training process

6. Modeling Strategy

The modeling phase prioritized **performance on the minority class** (churners) and interpretability

- **6.1 Algorithm Selection: Logistic Regression**
  - **Choice:** Logistic regression was selected as the final model.
  - **Performance Lead:** Logistic Regression **consistently outperformed** the challenger model (Random Forest). During validation, **Random Forest exhibited material degradation in minority-class performance,** particularly in recall, despite the increased model complexity
  - **Rationale:** The dataset contains strong linear risk factors (e.g., Price increase correlates to higher churn risk). The Random Forest model likely struggled to generalize these relationships given the class imbalance, whereas the weighted logistic regression successfully captured the signal.
  - **Interpretability:** Beyond performance, Logistic regression offered transparent coefficients (e.g., quantifying exactly how much “Fiber Optic” increases the odds of churn), which is critical for explaining the model to stakeholders.

- **6.2 Handling Class Imbalance:**
  - **Method:** The model was configured with **(*class_weight = ‘balanced’*).**
  - **Mechanism:** This hyper parameter automatically adjusts weights inversely proportional to class frequencies. It penalizes the model more heavily for missing a “Churner” than for flagging a “Non-Churner”.

6. Modeling Strategy

- **Business Logic:** This aligns with the financial reality that the cost of missing a churner ($74/month ARPU + LTV) is significantly higher than the cost of a retention intervention.

7. Model Evaluation

The model performance was evaluated based on the **financial trade-off** between precision and recall, rather than raw Accuracy.

- **7.1 Key Metrics:**
    - **Recall (~72%):** The model correctly identifies **72 out of every 100 churners.** This was the primary optimization target to minimize revenue leakage.
    - **Precision (~54%):** When the model predicts churn, it is correct 54% of the time. While statistically moderate, this represents a **>2x Lift** relative to the baseline churn rate **(26%)**.
    - **ROC-AUC (0.84):** This indicates strong separability between the classes, confirming the model can effectively rank customers by risk.
- **7.2 The strategic Trade-Off:**
    - A lower precision (more “False Alarms”) was accepted to maximize recall as the cost of missing a churner exceeds that of retaining a non-churner
    - **Business Justification:** Intervening on a “False positive” remains cost effective, because retention efforts are cheaper than acquiring a new customer.

These metrics indicate that the model effectively prioritizes high-risk customers for proactive retention, supporting targeted interventions.

8. Business Impact and ROI Simulation

To translate model performance into business value, a profitability simulation was conducted on the test set. This simulation estimates the operational impact of deploying the model for a targeted retention campaign.

8.1 Simulation Assumptions

- **Revenue at Risk (ARPU):** $74 (Average monthly spend of predicted churners).
- **Intervention Cost:** $30 per customer (Estimated cost of a discount, incentive or agent time
- **Retention Success Rate:** 30% (Conservative benchmark, assumes every 1 in 3 contacted churners is saved)

```python
#Simulation Assumptions
arpu = 74
intervention_cost = 30
success_rate = 0.30

true_positives = ((y_test == 1) & (y_pred_custom_threshold == 1)).sum()
total_flagged = y_pred_custom_threshold.sum()

campaign_reach = total_flagged
operational_cost = intervention_cost * campaign_reach
customers_saved = true_positives * success_rate
revenue_rescued = customers_saved * arpu
net_profit = revenue_rescued - operational_cost

print(operational_cost)
print(revenue_rescued)
```

[164] ✓ 0.2s

```
14970
5949.599999999999
```

8.2 The Calculation

Using the model's precision of ~54% and recall of ~72%,

8. Business Impact and ROI Simulation

1. **Campaign Reach:** The model flags ~480 **high-risk** customers in the test set.
2. **Operational Cost:** Contacting these 480 customers costs **$14,400 (**$30 * 480)
3. **True Churners Found:** Within this group, ~**260** are actual churners (True Positives).
4. **Customers Saved:** With a 30% success rate, we retain ~78 customers (260 * 0.3)
5. **Monthly Revenue Rescued:** These 78 customers represent **$5,772** in monthly recurring revenue **($74 * 78)**

8.3 ROI Conclusion

- **Net Monthly Benefit (Month 1): -$8,628 (**Initial investment).
- **Payback Period: ~2.5 Months.** The recurring revenue saved **($5,772/mo.)** fully covers the one time intervention cost **($14,400)** midway through the third month.
- **12-Month LTV Impact:**
  - If the saved customers stay for 1 year, the total revenue rescued is **~$69,264.**
  - This yields a **~380% ROI** on the **$14,400** spend over a 12-month horizon
- **Strategic Verdict:** While this campaign requires upfront working capital, it is highly profitable on a **Lifetime Value (LTV)** basis. It trades a one-time expense for long-term recurring cash flow.

9. Limitations, Recommendations and Next Steps

While the current model delivers immediate ROI, the analysis acknowledges specific constraints and outlines a roadmap for future optimization.

9.1 Strategic Recommendations:

Based on the model's findings and counterfactual simulations, two actions are recommended:

- **The Anchor Campaign:**
    - **Action:** Target "Vulnerable Veterans" (High tenure, Month-to-Month) with a bundle offer for **Online Security** or **Tech Support.**
    - **Rationale: EDA** confirmed these services act as a retention shield, significantly lowering churn probability for fiber users.
- **The Contract Migration Strategy (Validated via Simulation):**
    - **Action:** Target high risk Month to Month users with a "12th Month Free" offer to incentivize upgrading to a 1 year contract.
    - **Model Validation:** I ran a counterfactual simulation on the test set, artificially migrating all high-risk users to a 1 year contract.
    - **Result:** The model predicts this single action would reduce churn volume by **~34%** (499 -> 328 churners).
    - **Financial Impact:** This represents a potential **$12,600 monthly revenue retention** upside on the test population alone

9. Limitations, Recommendations and Next Steps

9.2 Technical Limitations

- **Snapshot View vs. Time-Series:** The dataset treats customer behavior as a static snapshot. There is no capture of temporal trends (e.g. “Usage dropped by 50% last month”), which Is often a leading indicator of churn.
- **Correlation vs. Causation:** While the model successfully identifies *who* is at risk, it is heavily reliant on correlation. High pricing correlates with churn, but reducing pricing isn’t guaranteed to fix it if the root cause is poor service quality.
- **Intervention Assumptions:** The ROI analysis assumed a flat 30% success rate for retention offers. This is a heuristic; actual pilot data will be required to calibrate the true lift of specific interventions