Which customers are worth Retaining?

A Machine Learning and Optimization Framework for Targeted Customer Retention

Joshua Thompson

Date: 03/28/2025

Table of Contents

1. Business Context

The Telco-Customer-Churn dataset contains detailed records for over 7,000 customers of a telecommunications company. Each row represents a single customer, capturing both demographic and service-related attributes. Key fields include:

- Demographics: Basic information such as gender, senior citizen status, partner, and dependents.
- Subscription Details: Contract type (month-to-month, one-year, two-year), payment method, and paperless billing preferences.
- Service Usage: Internet service type (DSL, Fiber optic, None), phone and multiple line usage, and optional add-on services such as online security, online backup, device protection, streaming TV, and streaming movies.
- Financials: MonthlyCharges indicating how much a customer pays per month, and TotalCharges representing the cumulative amount paid throughout their tenure.
- Churn Indicator: A binary flag showing whether the customer has discontinued service.

In the telecommunications industry, customer churn is a major business challenge, especially for subscription-based models. Reducing churn directly impacts profitability by improving customer lifetime value (LTV), increasing marketing efficiency, and reinforcing brand loyalty. Even small improvements in churn rates can lead to significant financial gains.

This analysis is set in the context of a highly competitive market with constrained retention resources such as loyalty incentives and discounts. As such, the business requires a data-driven approach to optimize who to target for retention, given a limited budget.

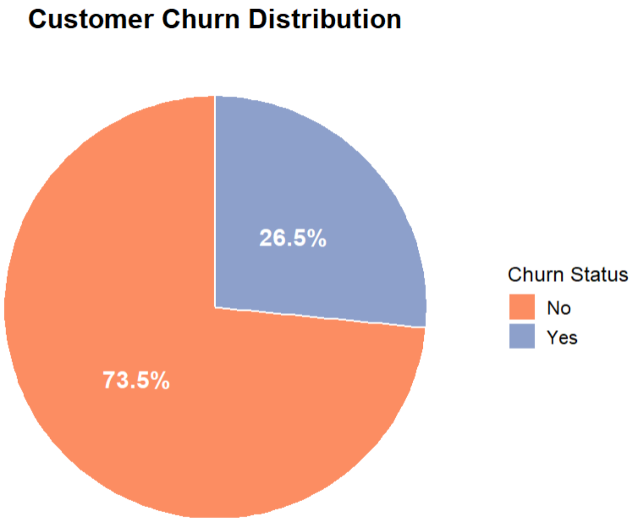
Business Objectives:

1. Predict which customers are most likely to churn.
2. Segment customer groups based on behavior and churn risk.
3. Prioritize customers for retention under budget constraints.
4. Maximize the expected net profit from retained customers.

To address these challenges, I conducted a comprehensive churn analysis and designed an optimized retention strategy. By combining machine learning with budget-constrained optimization, this solution not only predicts churn but also determines which customers to retain to maximize profitability. This framework ensures that retention efforts are focused on high-value, high-risk customers, preserving revenue while avoiding unnecessary spending.

2. Exploratory Data Analysis

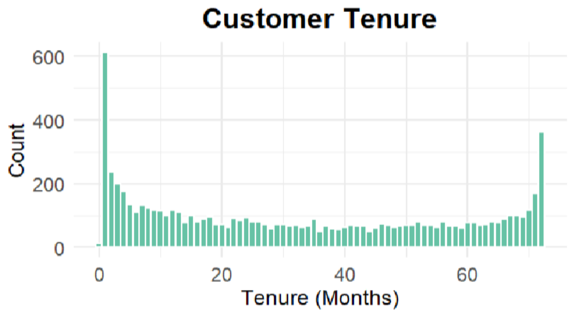
2.1 Summary Statistics

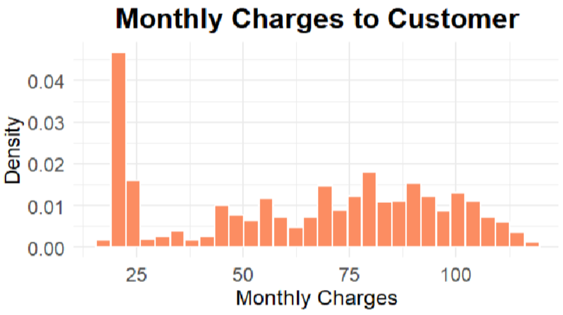


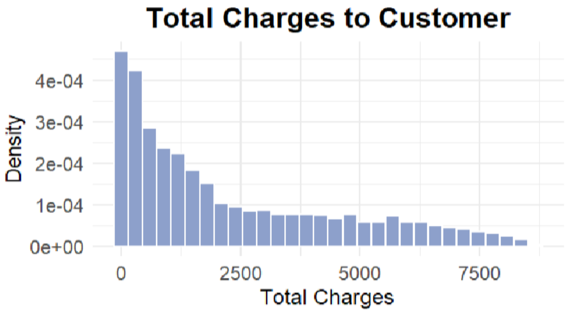
I began with looking at the distribution of churn status among customers in the dataset. Approximately 26.5% of customers have churned while 73.5% have remained loyal.

The churn rate is relatively high at 26.5%, signaling a significant portion of customers who have stopped using the service for various reasons.

This level of churn could have substantial implications on revenue and customer lifetime value (CLV). Thus, highlighting the importance of implementing targeted retention strategies.







Customer Tenure:
Indicating two main groups — short-term churners and long-term loyal customers, with fewer medium-tenure customers.

Monthly Charges:
Right-skewed, with most customers paying between $20–$50, but a significant portion paying higher amounts, reflecting diverse pricing tiers.

Total Charges:
Highly skewed due to the interaction between tenure and monthly charges. Higher total charges are associated with long-tenured, high-value customers.

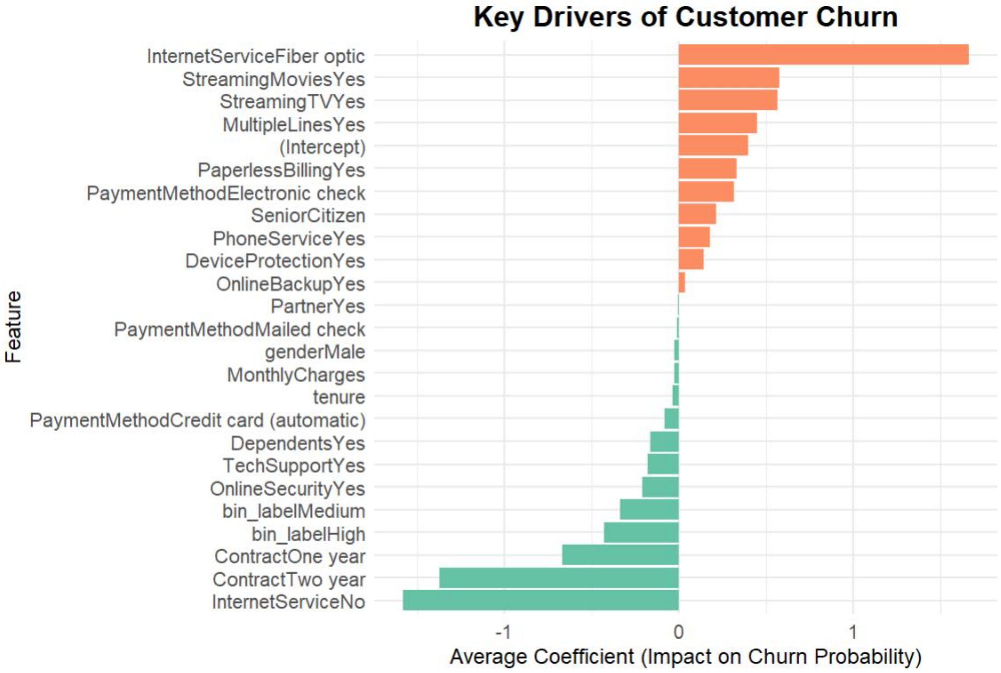
After exploring the key patterns in customer behavior, I wanted to dive deeper into understanding the drivers of churn. This led me to develop a predictive churn model and, more importantly, design an optimized retention strategy to help the business effectively prioritize which customers are worth retaining.

3. Customer Churn Analysis

3.1 Logistic Regression Model for Churn Classification

Model	Accuracy	Recall	Precision	F1-Score	AUC
Logistic Regression	0.739	0.805	0.505	0.620	0.843
Naïve Bayes	0.687	0.857	0.453	0.592	0.819
CART	0.686	0.836	0.458	0.587	0.797

During my churn analysis, I compared different machine learning model results and among the three models, logistic regression achieved the best overall performance, offering a strong balance of recall, precision, and AUC. While Naïve Bayes excelled in recall, it suffered from low precision. CART, despite being highly interpretable, underperformed in both classification and ranking power.

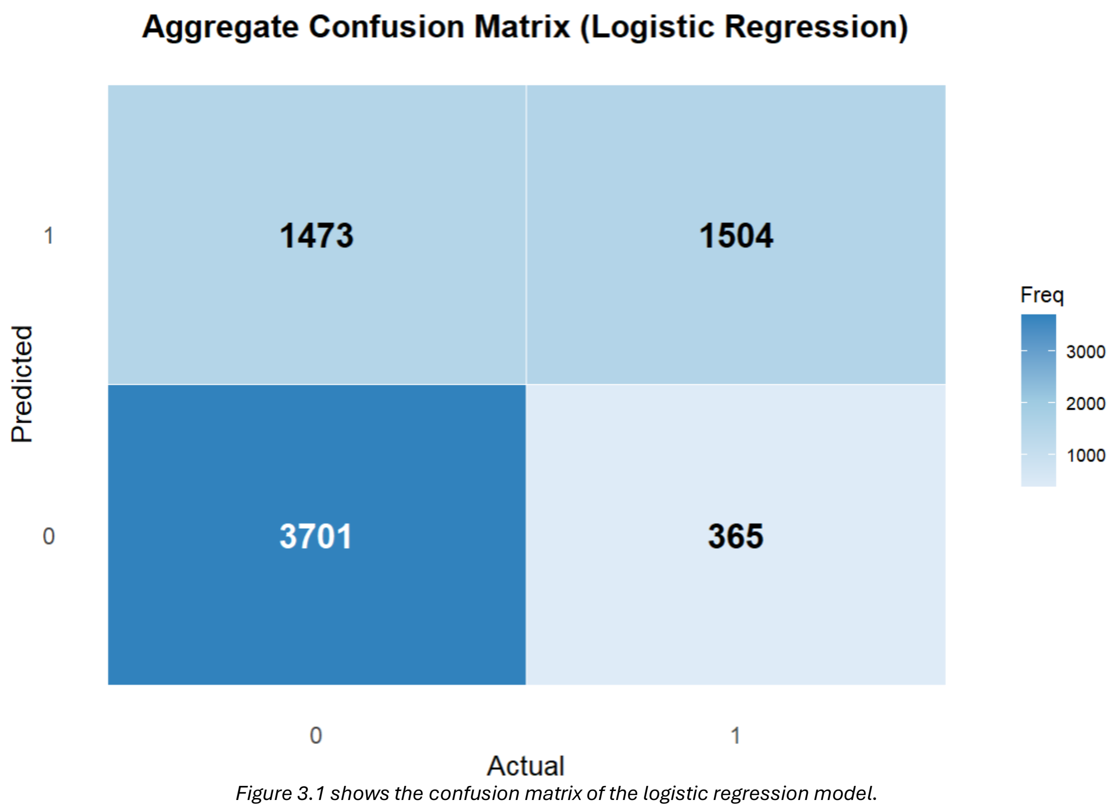


This plot visualizes the average effect of each feature on churn probability, as learned by the logistic regression model.

Insight	Description
Long-Term Contracts Reduce Churn	Customers with 1-year and especially 2-year contracts are significantly less likely to churn, confirming the effectiveness of contract-based retention.
Fiber Optic Internet Increases Churn	Customers subscribed to fiber optic internet show the highest churn risk. This group may require targeted interventions such as service quality improvements or loyalty incentives.
Engagement with Support Services Lowers Churn	Usage of technical support, online security, and phone services are associated with lower churn, suggesting that engaged customers who use multiple services tend to stay longer.
Streaming Services & Paperless Billing Increase Churn Slightly	Customers who use streaming services or have paperless billing show a modestly higher churn risk. These customers may be more digitally aware and willing to switch providers.
High Customer Lifetime Value (CLV) Segments are More Loyal	Customers classified into medium and high-value CLV bins show a lower likelihood of churn, making them prime candidates for retention efforts.

By producing actionable churn probabilities, this model enables the business to proactively reduce churn and improve Customer Lifetime Value (CLV) by targeting high-risk customers with tailored retention strategies.

3.2 Model Performance Evaluation



	0	1
1	1473	1504
0	3701	365

Figure 3.1 shows the confusion matrix of the logistic regression model.

Confusion Matrix

This plot represents the **aggregate confusion matrix** resulting from the **Logistic Regression** classifier evaluated across all folds of cross-validation.

Key Insights:

- The model correctly identified **1,504** customers who actually churned (**True Positives**).
- It also correctly identified **3,701** customers who did not churn (**True Negatives**).
- However, there were **1,473 false positives** (non-churners predicted as churners).
- There were **365 false negatives** (actual churners who were missed by the model).

Business Implication:

- While the model successfully captures a significant number of actual churners, there is a noticeable number of **false positives**. This means some retention resources may be allocated to customers who were unlikely to churn.
- The relatively low number of **false negatives** is encouraging, as fewer actual churners are being missed.

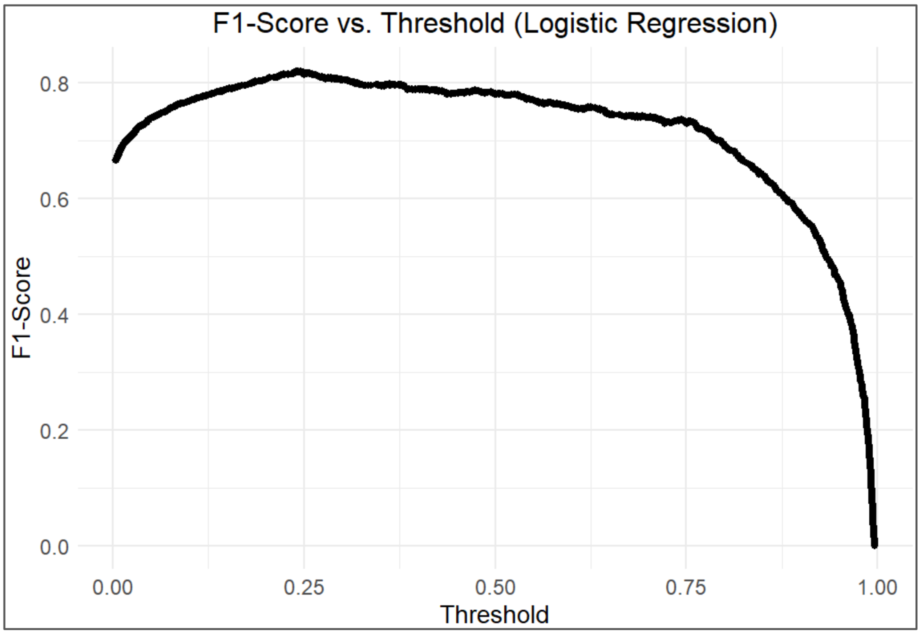
In customer retention contexts, this trade-off may be acceptable if the cost of offering retention incentives is lower than the cost of losing customers.

However, a single confusion matrix reflects only one threshold setting. To optimize retention strategies, we need to see how performance changes as we adjust that threshold.

3.3 Threshold and Trade-Off Analysis

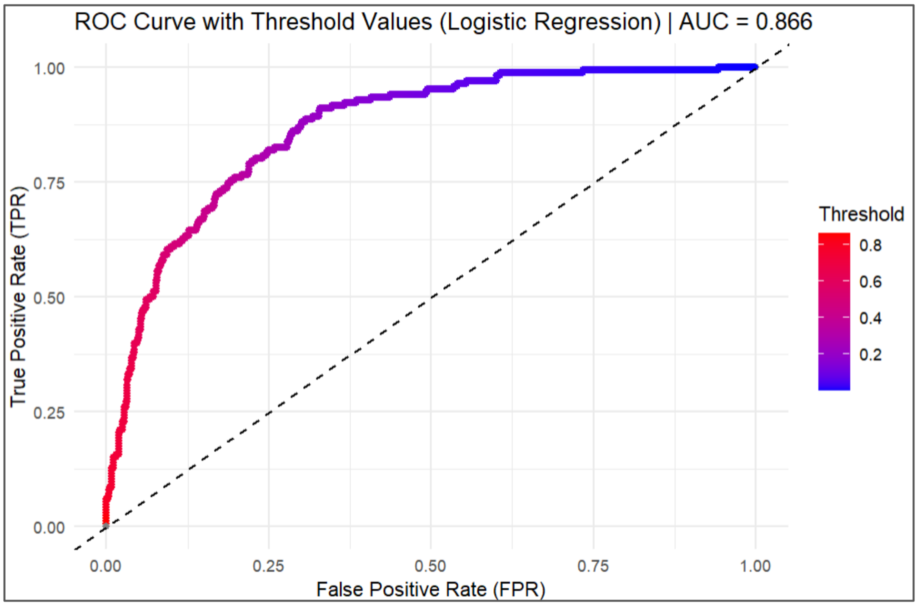
Here are some results from the tradeoffs considered when selecting a reasonable classification threshold value.

F1-Score vs. Threshold Analysis



This chart shows how well our churn prediction model performs as we adjust the threshold — the point where we classify a customer as an actual churner.
The key takeaway here is how to set the threshold to best identify churners without generating too many false alarms:

This pattern suggests that the model tends to assign low probabilities even to customers who are actually at risk of churning. Therefore, to maximize effectiveness, we need to adopt a lower threshold to ensure we capture most potential churners early.



Balancing Precision and Recall

This chart illustrates how our model balances the True Positive Rate (TPR, or recall — how many churners we successfully detect) and the False Positive Rate (FPR — how often we incorrectly predict churn for non-churners) as we adjust the classification threshold.

The trade-off visualized here is critical for churn prediction:

A higher TPR means we successfully identify more actual churners. A lower FPR ensures we minimize false alarms and focus retention efforts on truly at-risk customers.

The goal is to select a threshold that balances these two objectives — catching as many churners as possible while avoiding excessive false positives.

Choosing the right balance improves the efficiency of retention campaigns, allowing us to prioritize customers who are genuinely at risk. Based on this trade-off, I proceeded to select an optimal threshold that best aligns with business objectives.

Final Threshold Selection and Model Performance

After analyzing the trade-offs, I selected a threshold of 0.25 as the optimal cut-off for identifying churners. This choice balances business priorities with model performance, while being informed by the ROC Curve which achieved an AUC of 0.866 — indicating the model has strong discriminatory power.

Why this threshold?

- High Recall (0.805)
  Our main objective is to capture as many potential churners as possible. Since retention interventions are typically less costly than the revenue lost from churn, it's worth identifying a large portion of at-risk customers — even if it means including some who might not churn.
- Acceptable Precision (0.505):
  While not perfect, this precision level is acceptable given business needs. The retention team can manage a moderate number of false positives, especially since proactive engagement with these customers can still strengthen loyalty or even open up-selling opportunities. In many cases, the cost of engaging non-churners is outweighed by the value of preventing churn or increasing CLV.
- Balanced F1-Score (0.620)
  This F1-Score shows that the model strikes a practical balance — identifying enough churners (high recall) without overwhelming retention teams with too many false alarms.

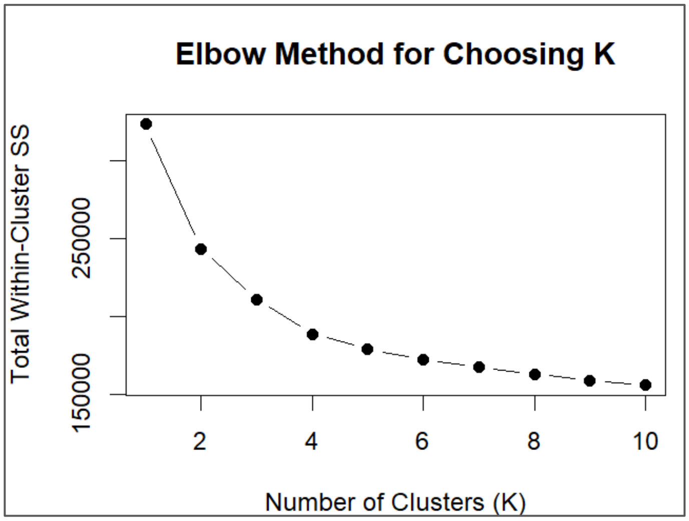
In business terms, using this threshold ensures we detect about **81% of real churners**, giving the company a strong opportunity to engage and retain them, even if it occasionally involves interacting with customers who might not have churned. Since missed churners cost far more than a few false alarms, we value recall most.

While the model provides valuable churn probabilities for individual customers, it is equally important to understand how different types of customers behave. To address this, I performed a customer segmentation analysis.

4. Customer Segmentation

The next step was to segment customers into distinct groups based on shared behavioral patterns, allowing the business to tailor retention strategies more effectively. To ensure the segmentation was objective and data driven, I used the Elbow Method to determine the optimal number of customer groups.

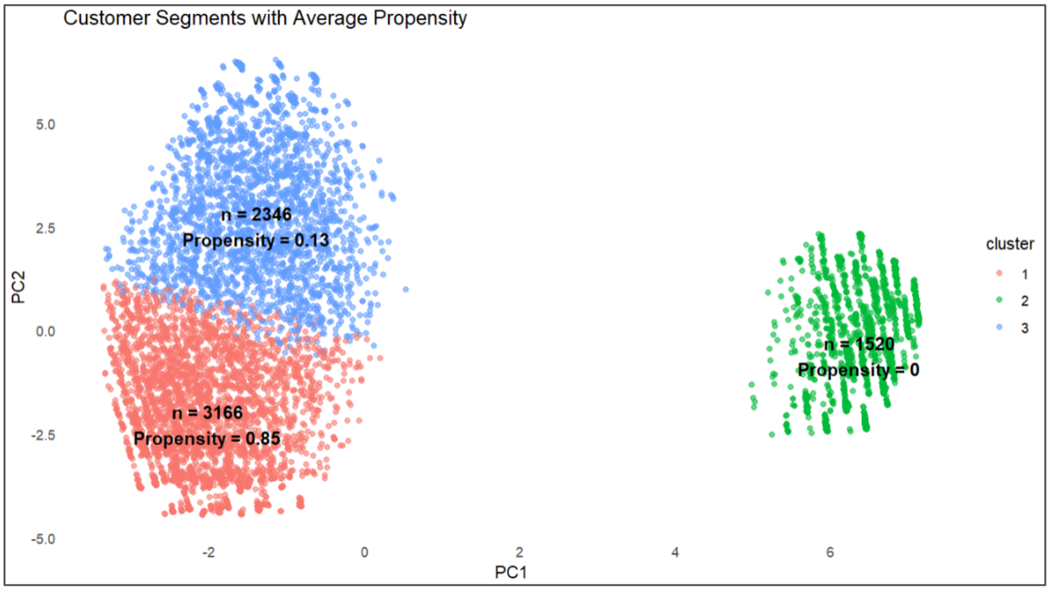
1.1 Determining Optimal Number of Clusters (Elbow Method)



Choosing the Optimal Number of Clusters (K)

To determine the ideal number of customer segments, I applied the **Elbow Method**. This technique shows a sharp bend at K = 3, suggesting that three clusters offer a good balance between compression and complexity. Therefore, I selected **K = 3** for the number of segmentations.

4.2 Customer Segmentation Visualization (PCA)

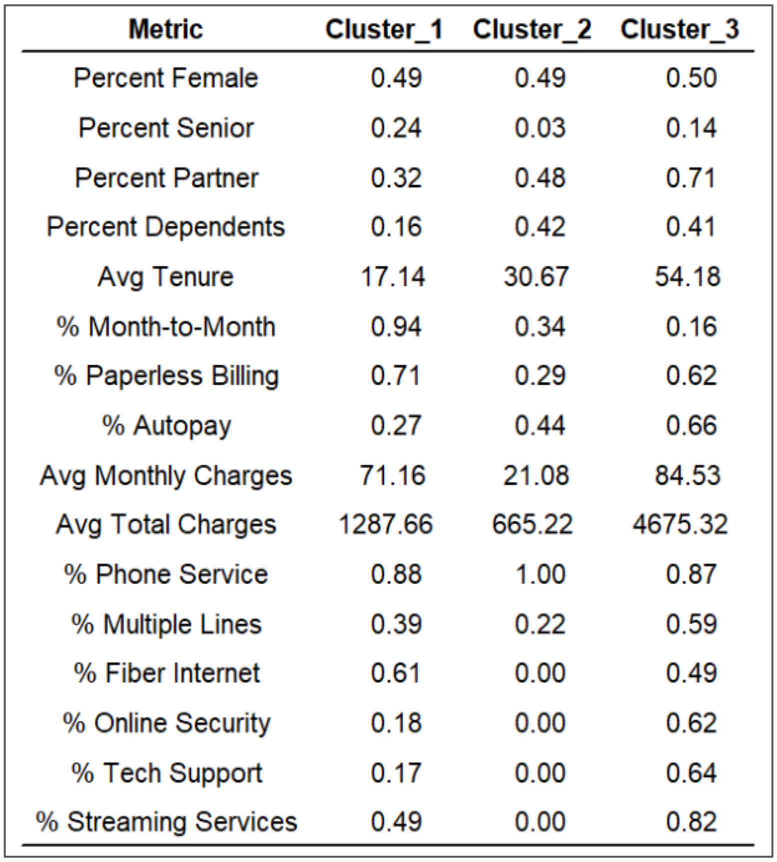


To uncover patterns in customer behavior, I performed K-means clustering using a variety of customer attributes such as contract type, tenure, monthly charges, and usage patterns. The objective was to group customers with similar characteristics into distinct, actionable segments.

Using Principal Component Analysis (PCA) for dimensionality reduction, I visualize the three customer segments. Each point in the scatter plot represents a customer, colored by their cluster assignment.

In the next step, I analyzed these clusters in detail to understand their defining characteristics and identify actionable customer personas.

4.3 Customer Segments & Personas



Metric	Cluster_1	Cluster_2	Cluster_3
Percent Female	0.49	0.49	0.50
Percent Senior	0.24	0.03	0.14
Percent Partner	0.32	0.48	0.71
Percent Dependents	0.16	0.42	0.41
Avg Tenure	17.14	30.67	54.18
% Month-to-Month	0.94	0.34	0.16
% Paperless Billing	0.71	0.29	0.62
% Autopay	0.27	0.44	0.66
Avg Monthly Charges	71.16	21.08	84.53
Avg Total Charges	1287.66	665.22	4675.32
% Phone Service	0.88	1.00	0.87
% Multiple Lines	0.39	0.22	0.59
% Fiber Internet	0.61	0.00	0.49
% Online Security	0.18	0.00	0.62
% Tech Support	0.17	0.00	0.64
% Streaming Services	0.49	0.00	0.82

A summary of customer behavioral & demographic Insights are used to better understand each segment:

Cluster 1 (High) – 85% likelihood of churn

- Mostly month-to-month contracts (94%)
- Low tenure and low autopay use
- Heavy use of fiber internet, limited support services

Cluster 2 (Loyal) – 0% likelihood of churn

- Long-tenured customers
- Low monthly charges and strong service adoption
- High percentage have phone service

Cluster 3 (Moderate) – 13% likelihood of churn

- Highest revenue customers
- Long tenure, high overall service use

Equipped with customer personas and churn probabilities, I next focused on designing a data-driven retention strategy. The following analysis compares two approaches to deciding which customers should be targeted under realistic budget constraints.

5. Customer Retention

This analysis evaluates two approaches for selecting customers for retention initiatives:

1. A Rule-Based Decision Model without budget constraints.
2. A Budget-Constrained Knapsack Optimization Model.

The goal is to maximize net profit from retention actions while considering practical budget limitations.

Assumptions:

Parameter	Value	Description
Average Gross Profit	40%	Gross profit margin on retained customers
Retention Discount	15%	The discount applied as part of retention offers
Marketing Cost	$5	Average Marketing cost per customer targeted
Cost of Service	40%	Service delivery cost as a percentage of CLV
Acceptance Rate	40%	Percentage of targeted customers expected to accept the retention offer

Financial Formulas Used

- Expected Net Benefit = (CLV − Service Cost) × Churn Propensity × Acceptance Rate − Retention Cost
- Expected Gross Benefit = (CLV − Service Cost) × Churn Propensity × Acceptance Rate
- Retention Cost = (CLV × Retention Discount %) + Marketing Cost
- Service Cost = CLV × Cost of Service %

5.1. Baseline heuristic approach

Approach 1 — Baseline heuristic approach (without budget constraint)
This approach targets all customers if their **expected benefit** exceeds their **retention cost**, without considering budget limitations.

Metric	Value
Total Customers	7,032
Identified for Retention	2,620
Total Expected Benefit	$887,798
Retention Cost	$632,057
Total Expected Profit	$255,741

This approach provides a quick and intuitive way to select customers but assumes that the business has **no budget constraints**, which may not reflect operational realities.

5.2 Knapsack Optimization Strategy

Approach 2 — Knapsack Optimization
This approach solves a binary knapsack optimization problem to select the subset of customers that maximizes expected net profit while adhering to a fixed retention budget.

Scenario 1: Full Optimization (No segmentation)

Metric	Value
Retention Budget	$250,000
Identified for Retention	775 customers
Total Expected Benefit	$387,769
Retention Cost	$250,000
Total Expected Profit	$137,769

Scenario 2: Cluster-Based Budget Allocation (Includes customers' likelihood of churn)
In this scenario, the $250,000 budget is distributed across customer segments with 70% of the budget allocated to segment 1 (high risk of churn customers) and 30% to segment 3 on (moderate risk) but high value customers.

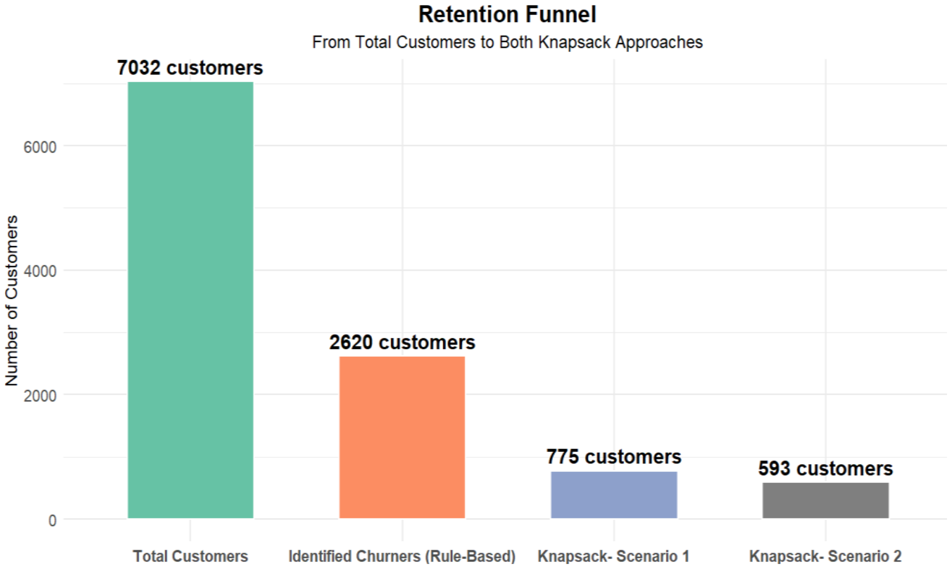
Metric	Value
Retention Budget	$250,000
Retention Target — Cluster 1	487 customers
Retention Target — Cluster 3	106 customers
Total Expected Benefit	$372,468
Retention Cost	$250,000
Total Expected Profit	$122,468

By splitting the budget across clusters, the model sacrifices ~11% of potential net profit but achieves additional strategic objectives such as maintaining market share among moderate-risk customers.

5.3. Comparative Results

Retention Funnel

- Starting with 7,032 total customers, the churn model first identifies 2,620 customers as likely churners based on predicted probabilities and expected benefits (heuristic approach).
- Applying the final Knapsack (scenario 2), only 593 customers are selected as the optimal retention targets from 2 segments. These are the customers who provide a balance of high net expected profit and maintain market share, while staying within budget.



This highlights the importance of not only predicting churn but also strategically selecting which customers to retain given limited resources. The next section summarizes the key insights and actionable recommendations derived from this analysis.

6. Final Summary and Recommendations

This project successfully combines machine learning and optimization techniques to develop a targeted and cost-effective customer retention strategy. By integrating churn prediction with a budget-constrained selection framework, the solution addresses not only the challenge of identifying at-risk customers but also the crucial question of *which customers are worth retaining* under real-world financial constraints.

Key Takeaways:

- The Logistic Regression churn model produced actionable churn probabilities with strong recall, ensuring that most actual churners are successfully captured.
- The customer segmentation revealed distinct behavioral patterns, enabling the design of tailored retention strategies across different customer groups.
- The Rule-Based approach, while effective for initial scoping and exploratory analysis, fails to account for budget limitations and can lead to inefficient resource allocation.
- The Knapsack Optimization framework balances profitability and budget constraints, achieving approximately 54% higher profit per dollar spent compared to the Rule-Based model by focusing on high-value, high-risk customers.

Recommendation:

Based on the findings, I recommend that the business adopt a data-driven customer retention strategy, combining machine learning for churn prediction with an optimization-based retention framework.

This approach:

- Maximizes net expected profit, preventing unnecessary spending on low-return or loyal customers.
- Focuses retention efforts on strategically important, high-risk, and high-value customers.
- Provides flexibility to adapt to dynamic budget scenarios and changing business conditions.

Overall, Knapsack Optimization aligns financial objectives with operational capabilities, making it a scalable and sustainable strategy to improve customer lifetime value and reduce churn-related revenue loss.

Future Considerations:

While this solution establishes a strong foundation, the following improvements could further enhance the retention strategy:

1. Personalized Retention Offers
   Move beyond generic discounts by tailoring incentives according to cluster characteristics or individual churn probabilities.
2. Model Monitoring and Updates
   Regularly update churn models and optimization inputs (e.g., acceptance rates, cost structures) to reflect evolving customer behavior.
3. CLV Segmentation Integration
   Incorporate predicted Customer Lifetime Value (CLV) into the optimization process to better prioritize customers with the greatest long-term revenue potential.

By embedding this data-driven approach into the organization’s retention strategy, the business can achieve sustainable profitability while strengthening customer relationships and brand loyalty.