Which customers are worth Retaining?

A Machine Learning and Optimization Strategy 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.

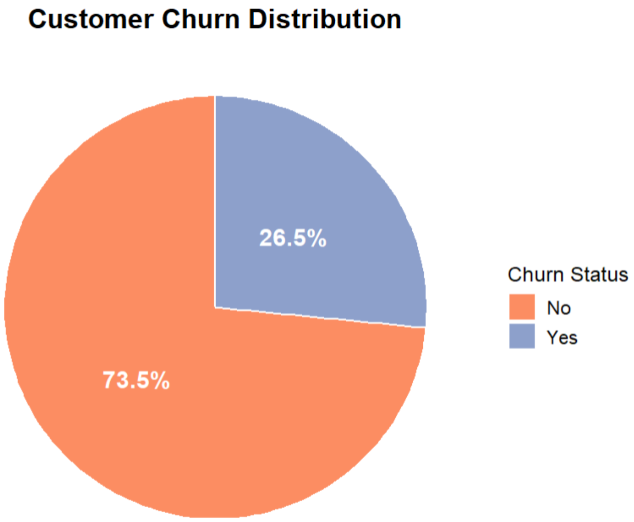
Business Objectives:

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

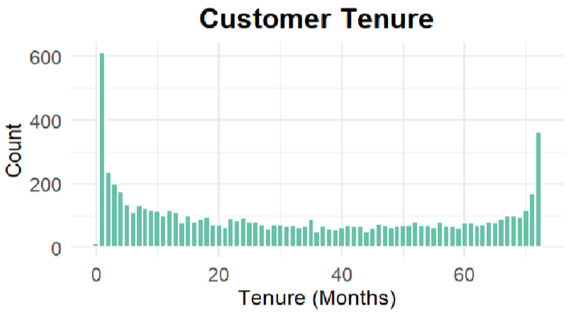
2.1 Summary Statistics

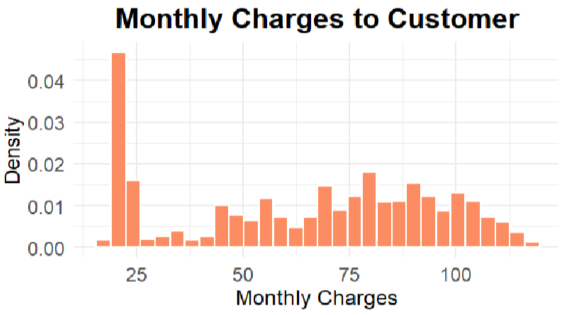


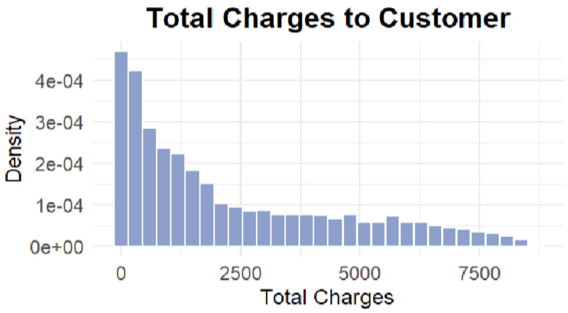
I began with looking at the distribution of churn status among customers in the dataset. Approximately 26.5% of customers have churned while 73.5% have remained loyal.

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.

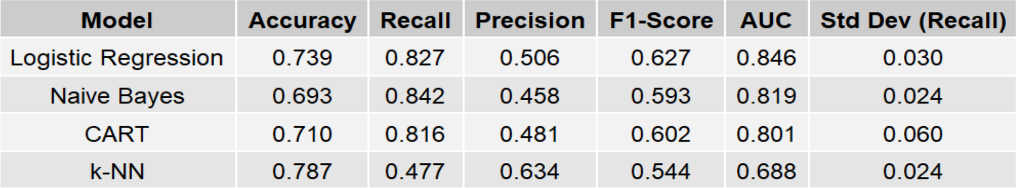
Total Charges:
Highly skewed due to the interaction between tenure and monthly charges. Higher total charges are associated with long-tenured, high-value customers.

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 Machine Learning Model for Churn Prediction

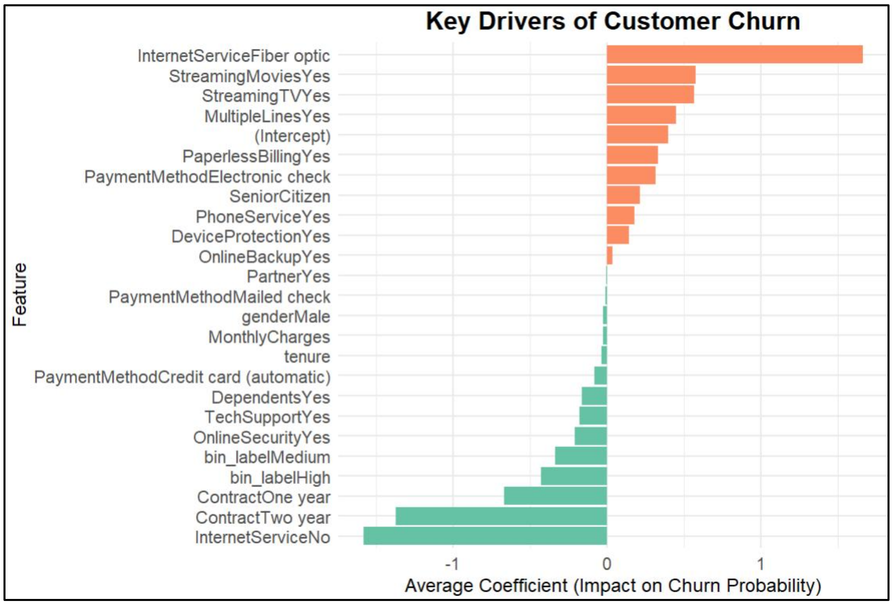
To begin my analysis, I evaluated four different machine learning algorithms: Logistic Regression, Naïve Bayes, Classification and Regression Tree (CART), and K-Nearest Neighbors (k-NN). Each model was trained and validated using cross-validation techniques to ensure robustness and minimize overfitting. The table below summarizes the performance metrics for each model, followed by my rationale for selecting the most suitable model for predicting customer churn.



Model	Accuracy	Recall	Precision	F1-Score	AUC	Std Dev (Recall)
Logistic Regression	0.739	0.827	0.506	0.627	0.846	0.030
Naive Bayes	0.693	0.842	0.458	0.593	0.819	0.024
CART	0.710	0.816	0.481	0.602	0.801	0.060
k-NN	0.787	0.477	0.634	0.544	0.688	0.024

I evaluated the performance of four classification models. Logistic Regression demonstrated the best overall balance, with strong recall (0.827), reasonable precision (0.506), and the highest AUC (0.846). It also showed consistent performance across folds, with a low standard deviation in recall (0.030). Naïve Bayes achieved the highest recall (0.842) but at the cost of lower precision (0.458). CART provided good recall (0.816) with slightly better precision than Naïve Bayes, though it had the highest variability in recall (0.060). While k-Nearest Neighbors (k-NN) yielded the highest accuracy (0.787) and precision (0.634), its low recall (0.477) and moderate variability limited its effectiveness in identifying churners.

To better understand what influences customer churn, I examined the underlying features driving the model's predictions. By analyzing the average impact of each variable on churn probability, we can uncover actionable insights that inform targeted retention strategies. The plot below highlights the most influential factors identified by the logistic regression model.



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

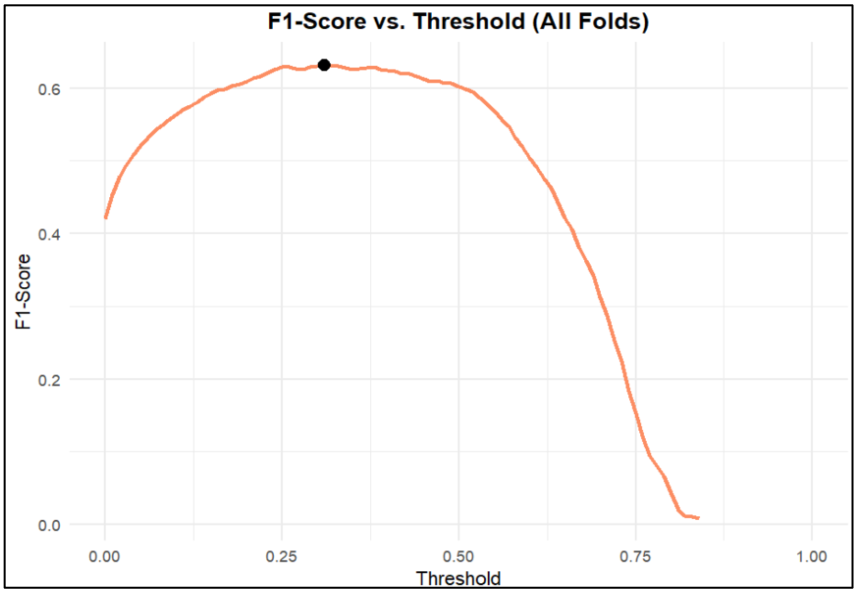
Insight	Description
Long-Term Contracts Reduce Churn	Customers with 1-year and especially 2-year contracts are significantly less likely to churn, confirming the effectiveness of contract-based retention.
Fiber Optic Internet Increases Churn	Customers subscribed to fiber optic internet show the highest churn risk. This group may require targeted interventions such as service quality improvements or loyalty incentives.

Insight	Description
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 Increase Churn Slightly	Customers who use streaming services 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.

3.2 Threshold and Trade-Off Analysis

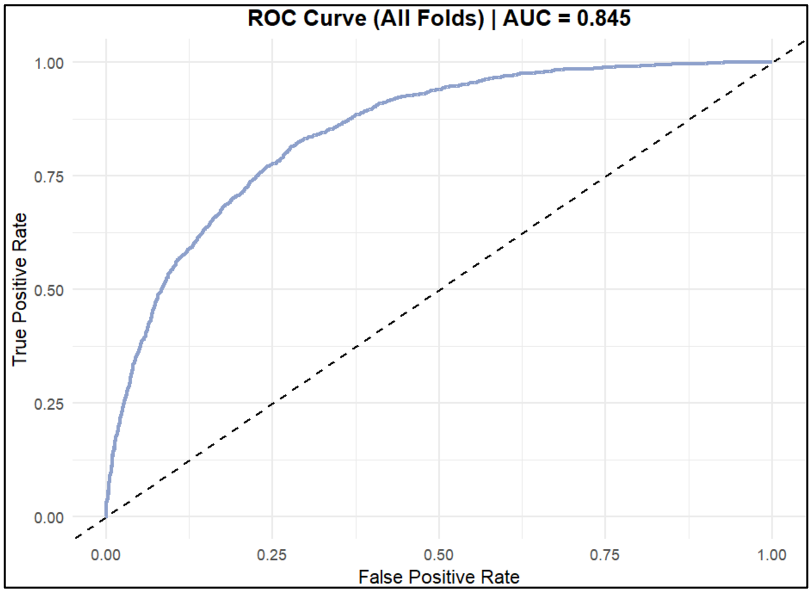
After selecting logistic regression as the most suitable model based on its strong overall performance, I evaluated how it performs in real-world decision-making by analyzing its classification outcomes. To do this effectively, I explored the trade-offs involved in choosing the right classification threshold — a decision point that determines whether a customer is flagged as likely to churn.

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.

3.3 Model Performance Evaluation

Based on this trade-off analysis, a threshold of approximately 0.25 was selected to strike a balance between maximizing recall (capturing most at-risk customers) and limiting false positives, aligning with business objectives focused on early churn intervention. With this threshold applied, the confusion matrix below illustrates how the model performs in classifying customers as churners or non-churners, providing insight into the practical outcomes of this decision.

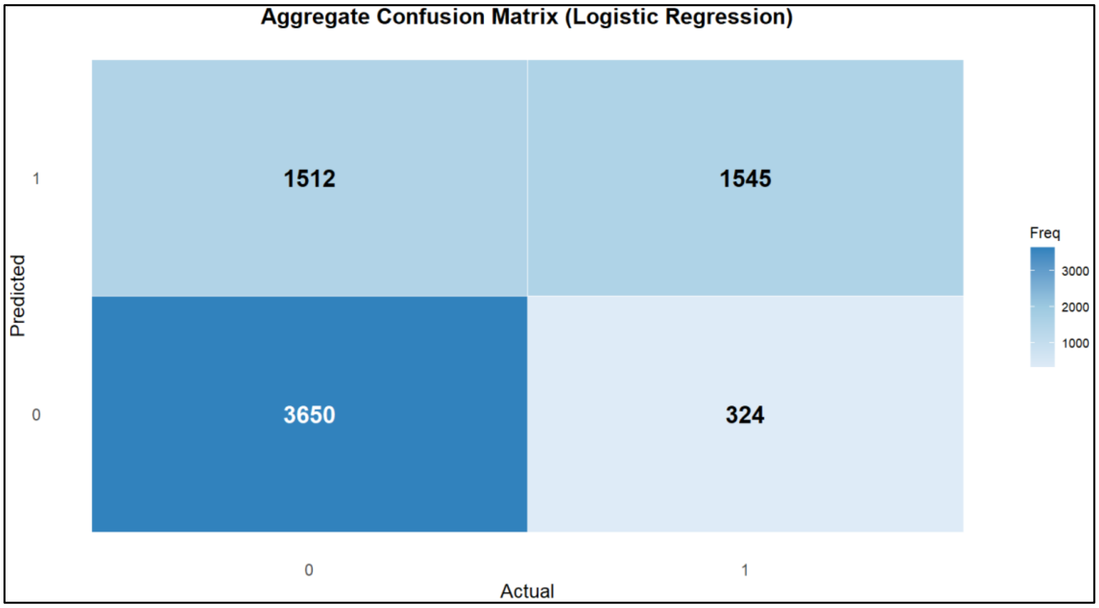


Figure 3.2 shows the confusion matrix of the logistic regression model.

Key Insights:

- The model correctly identified **1,545** churners (**True Positives**).
- It also correctly identified **3,650** non-churners (**True Negatives**).
- However, there were **1,512 false positives** (non-churners predicted as churners).
- There were **324 false negatives** (actual churners who were missed by the model).

Business Implication:

- While the model captures a large share of actual churners, the number of false positives suggest that some retention efforts may be spent on customers unlikely to churn.
- On the positive side, the low false negative count indicates that relatively few actual churners are slipping through the cracks, which is critical for effective retention

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.

Final Threshold Selection and Model Performance

After analyzing the trade-offs, The threshold of 0.25 provides a balanced cut-off point for identifying churners. This choice balances business priorities with model performance, while being informed by the ROC Curve which achieved an AUC of 0.845 — indicating the model has strong discriminatory power.

Why this threshold?

- High Recall (0.827)
  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 (0.506):
  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 (0.627)
  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 **83% 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.

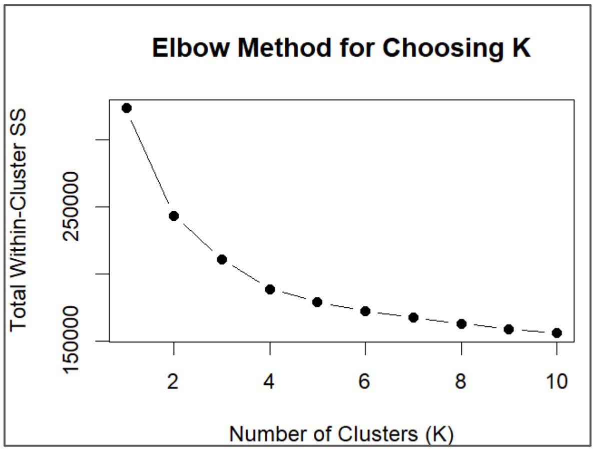
Having established churn risk, we now turn to understanding which types of customers exhibit these risks.

---

4. Customer Segmentation

Now that I have a trained ml model for relatively good classification of customer churn. 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 began my segmentation using an Elbow Method to determine the optimal number of customer groups.

4.1 Determining Optimal Number of Segments (Elbow Method)

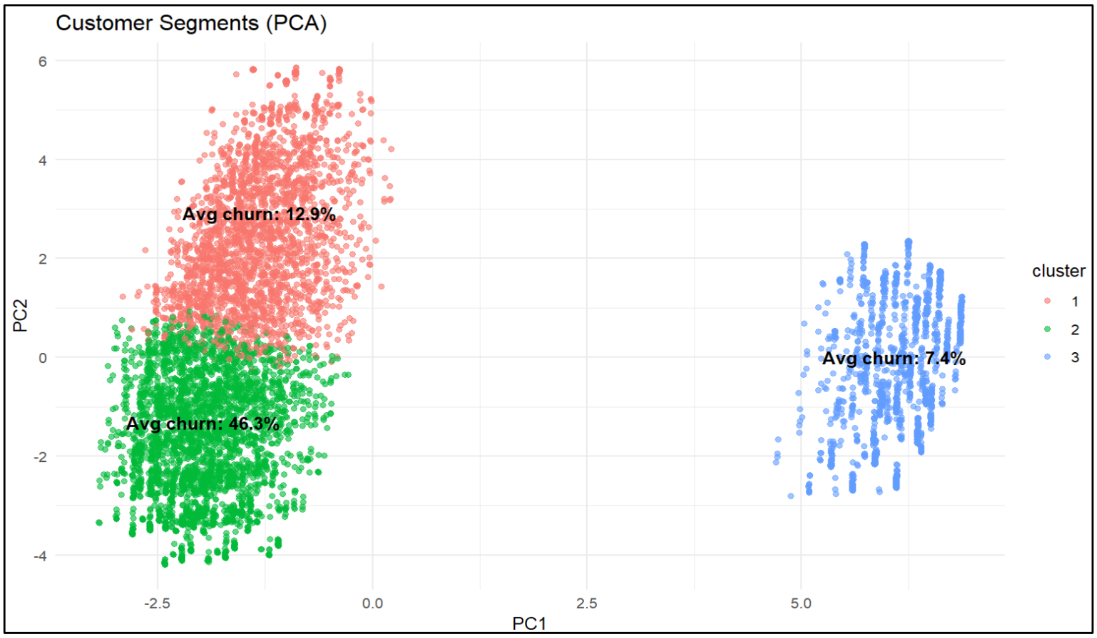


To determine the optimal number of customer segments, I applied the Elbow Method, a widely used technique in unsupervised learning.

As shown in the plot, there is a noticeable inflection point at K = 3, where the WSS drops steeply before starting to level off. This suggests that increasing beyond three segments yields diminishing returns in reducing within-segment variance.

This segmentation allows for actionable customer profiling while maintaining interpretability and focus for downstream retention strategies and targeted interventions.

4.2 Customer Segmentation Visualization (PCA)

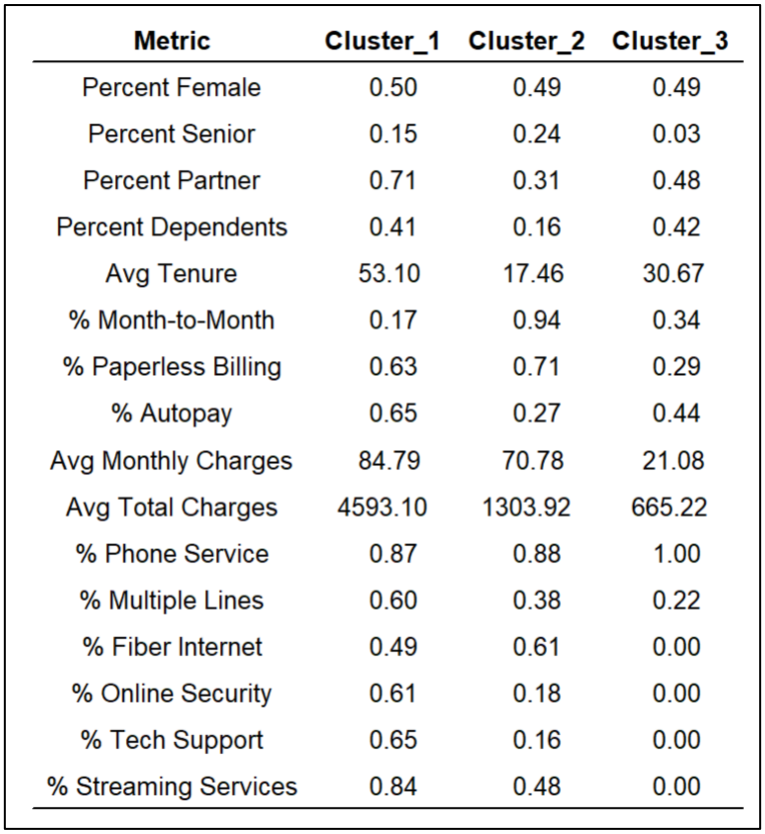


To uncover patterns in customer behavior, I performed K-means clustering using a variety of customer attributes such as contract type, tenure, monthly charges, and usage patterns. The objective was to group customers with similar characteristics into distinct, actionable segments.

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	0.50	0.49	0.49
Percent Senior	0.15	0.24	0.03
Percent Partner	0.71	0.31	0.48
Percent Dependents	0.41	0.16	0.42
Avg Tenure	53.10	17.46	30.67
% Month-to-Month	0.17	0.94	0.34
% Paperless Billing	0.63	0.71	0.29
% Autopay	0.65	0.27	0.44
Avg Monthly Charges	84.79	70.78	21.08
Avg Total Charges	4593.10	1303.92	665.22
% Phone Service	0.87	0.88	1.00
% Multiple Lines	0.60	0.38	0.22
% Fiber Internet	0.49	0.61	0.00
% Online Security	0.61	0.18	0.00
% Tech Support	0.65	0.16	0.00
% Streaming Services	0.84	0.48	0.00

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

Segment 1 – "Moderate" (≈11 % churn)

- Very long-tenured (Avg. tenure ≈ 53 months)
- Highest-value cohort (Avg. total charges ≈ $4,593)

Segment 2 – "High Risk" (≈ 46% churn)

- Heavy Fiber Internet usage.
- Predominantly month-to-month contracts.
- Shortest tenure (Avg. ≈ 16 months).

Segment 3 – "Loyal" (≈ 7.4% churn)

- Long tenure (~31 months)
- Lowest monthly bills
- Almost universally on Phone Service

Equipped with distinct customer personas and their corresponding churn probabilities, I shifted focus toward formulating a targeted, cost-effective retention strategy. By leveraging both customer segmentation and churn propensity scores, I evaluated how to allocate limited retention resources for maximum return on investment. The following analysis presents a comparative view of two strategic approaches — one rule-based and one optimization-driven — to determine which customers should be prioritized 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 is applied as part of retention offers. (estimated parameter)
Marketing Cost	$5	Average Marketing cost per customer targeted
Cost of Service	40%	Service delivery cost as a percentage of CLV
Acceptance Rate	70%	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 > retention cost***, without considering budget limitations.

Metric	Value
Total Customers	7,032
Identified for Retention	2,052
Total Expected Benefit	$628,463
Retention Cost	$437,273
Total Expected Profit	**$191,189**

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.

Segment-Based Budget Allocation (Includes customers' likelihood of churn segments)

In this scenario, the $250,000 budget is distributed across customer segments with ~93% of the budget allocated to segment 2 (high risk of churn customers). ~12% to segment 1 on (moderate risk) but high value customers and ~1.6% to segment 3 (low value, low churn).

Metric	Value	Retention Cost	Expected Benefit	Expected Profit
Retention Budget	$250,000			
Retention Target — Segment 1	78 customers	$30,349	$47,088	$16,738
Retention Target — Segment 2	1,445 customers	$232,966	$385,667	$152,700
Retention Target — Segment 3	51 customers	$4,132	$6,955	$2,823
Total		**$267,448**	**$439,710**	**$172,261**

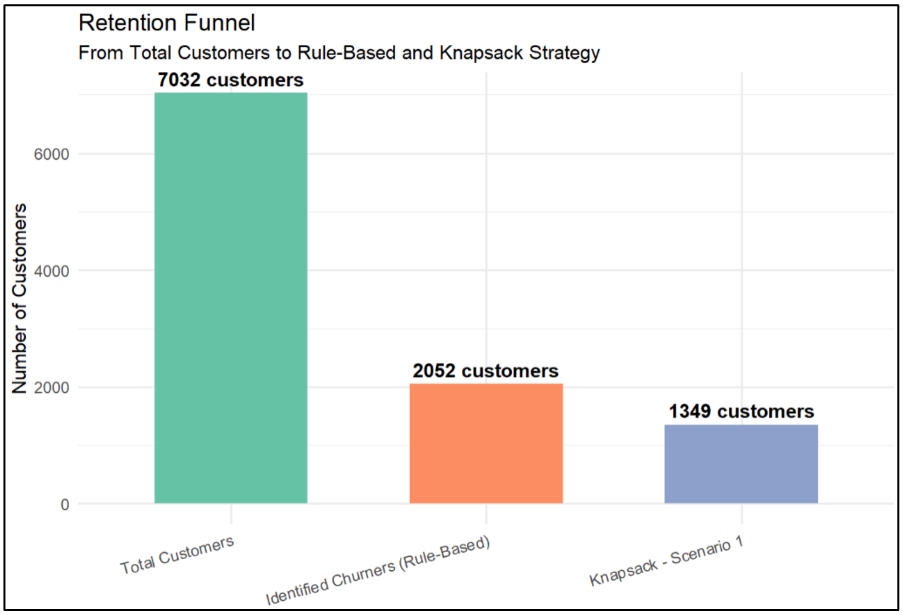
Strategic Insights

- **Segment 2-** (high churn risk, moderate value): Retained the most customers and profit
- **Segment 1-** (moderate churn, high value): Fewer but higher-value customers retained
- **Segment 3-** (low value, low churn): Still included — probably due to very low cost and positive margin

5.3. Comparative Results

Retention Funnel

- Starting with 7,032 total customers, the churn model first identifies 2,052 customers as likely churners based on predicted probabilities and expected benefits (heuristic approach).
- Applying the Knapsack optimization strategy, only 1,349 customers are selected as the optimal retention targets from 3 segments. These are the customers who provide a balance of high net expected profit while staying within budget.



This highlights the importance of not only predicting churn but also strategically selecting which customers to retain given limited resources. While the baseline heuristic approach (Approach 1) identified a large pool of 2,052 potentially profitable churners, it delivered a return on investment ROI of approximately 43.7%. In contrast, the knapsack optimization method (Approach 2) selected a more refined group of 1,349 customers, achieving a significantly higher ROI of 64.4%. This represents a 20.7 percentage point uplift in ROI, demonstrating how data-driven optimization can dramatically improve the efficiency and profitability of customer retention strategies.

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 a higher return on investment compared to the Rule-Based model — 64.4% ROI versus 43.7%. This improvement stems from targeting customers who offer the highest expected profit per dollar spent, especially those in high-risk, high-value segments.

Future Considerations:

While this solution provides a solid foundation for a data-driven retention strategy, several enhancements can increase its long-term effectiveness:

1. Tailored Retention Incentives
   Move beyond uniform discounting by designing offers customized to customer segments or individual churn risk profiles. Personalization can improve response rates and increase ROI on retention spending.
2. Integration of Predicted CLV
   Enhance prioritization by incorporating machine learning–based Customer Lifetime Value predictions. This allows the business to focus efforts on customers with the greatest potential future value, not just short-term profitability.

By embedding this analytical framework into ongoing operations, the business can scale a smarter, more sustainable retention strategy — one that adapts to evolving customer behavior while maximizing long-term revenue and customer loyalty.