Predictive Analysis and Retention Strategy

Liam Montero Guillemi

2025-07-13

Contents

Executive Summary

This report details the comprehensive analysis conducted on the company's customer data to understand and predict the churn phenomenon. A machine learning model with an 85.8% predictive capacity (AUC) was developed, allowing us not only to identify at-risk customers but also to understand the underlying causes of their behavior.

The analysis reveals that churn is not a random event, but rather a predictable result of contractual factors, tenure, and customer profile. The main vulnerability lies in **new customers with flexible contracts (month-to-month)**, who have a churn rate of 51.4%.

Thanks to our model optimization, we are now able to **identify 82.1% of customers who plan to churn**. This report concludes with a four-pillar strategic plan, based on this evidence, designed to reduce churn, strengthen retention, and estimate **an annual revenue recovery potential of over $143,000**.

Data Preparation and Cleaning

The first fundamental step in any data science project is data preparation. This phase ensures the quality, consistency, and integrity of the dataset, laying the foundation for reliable analysis and modeling. The objective of this task was to load, explore, and clean the "Telco Customer Churn" dataset.

Loading and Initial Inspection

The project began by loading the data from a CSV file. An initial inspection allowed us to understand the structure of the dataset.

The dataset contains **7,043 observations (customers)** and **21 variables**, including customer identifiers, demographic characteristics, contracted services, contractual information, and the target variable, Churn. Each row was verified to correspond to a unique customer, as there were no duplicate customerIDs.

Handling missing values and transforming variables for analysis

Next, we proceed to inspect all the variables, one by one, to see if they have missing values, then we proceed to convert the Yes and No category variables into binary variables for ease of use.

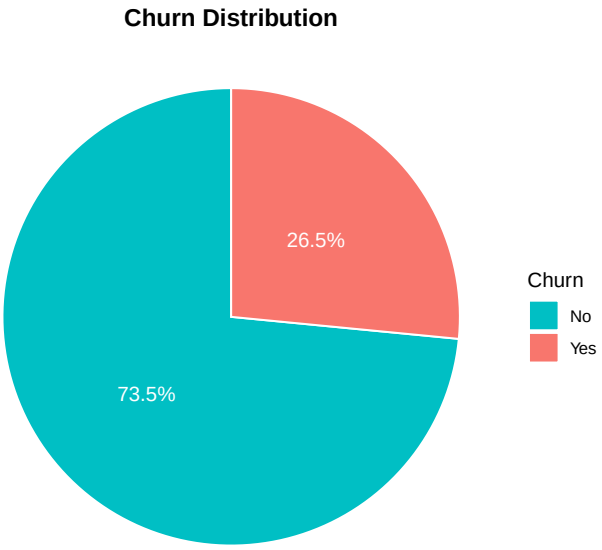
The most interesting thing we found was: + The tenure variable had 11 values that were 0. + The TotalCharges variable had 11 NA values.

Investigating further, I discovered that the 11 NA values in the TotalCharges variable were all related to the 11 0 values in the tenure variable. But looking more closely, these customers have two-year contracts and assigned monthly payment amounts. Therefore, the missing data in the TotalCharges column and the 0 values in the tenure column mean that these customers have already signed a contract with the company but have not yet made their first payment, and therefore, have not yet been with the company for a month.

Based on the above, I assigned 1 to the 0 values in tenure and the value of the MonthlyCharges variable to its respective NA value in the TotalCharges variable to eliminate the NA values and the 0 values.

Exploratory analysis (EDA)

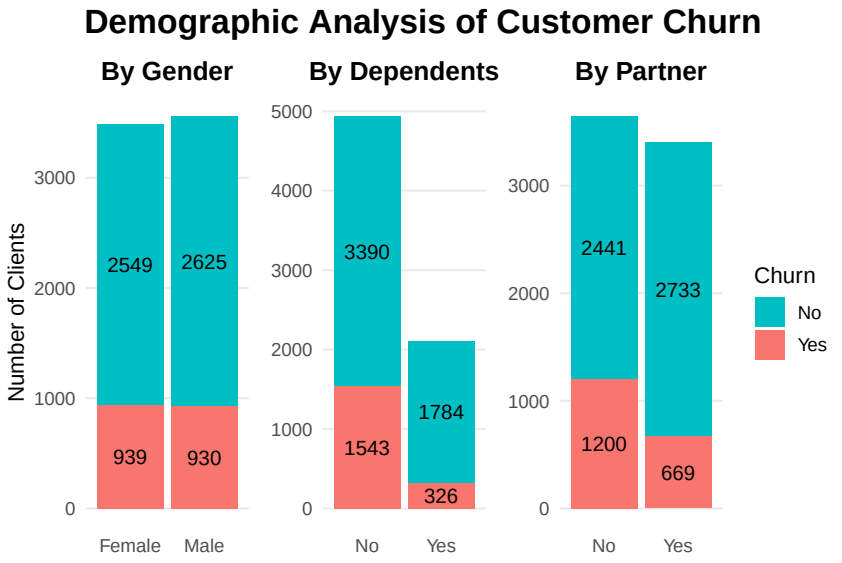
We start by calculating the company's customer churn rate and we can see that the churn rate is a bit high, as **26.5% of customers end up churning**.



Next, we'll look at a demographic graph showing how Churn behaves by gender, whether they have a partner, and whether they have dependents. It shows at a glance that:

- Gender does not influence customer churn
- Customers who do not have a partner or dependents are more likely to churn

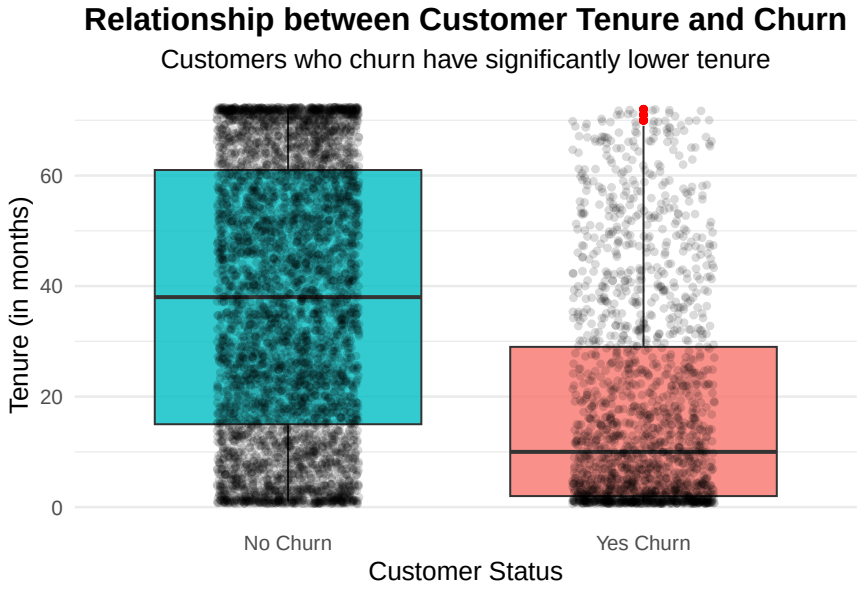
Thanks to this, we can begin to say that customer churn is not random; there are certain characteristics that these customers possess that lead to a higher churn rate.



Distribution of customers who leave (Yes) and stay (No) according to their profile.

Tenure analysis

Now we will see how the seniority of customers in the company behaves so that they decide to leave it:



This visualization compares the distribution of customer tenure (measured in months) for two groups: those who have churned ("Churn") and those who have not ("Non-Churn"). The result is one of the most conclusive findings of the exploratory analysis.

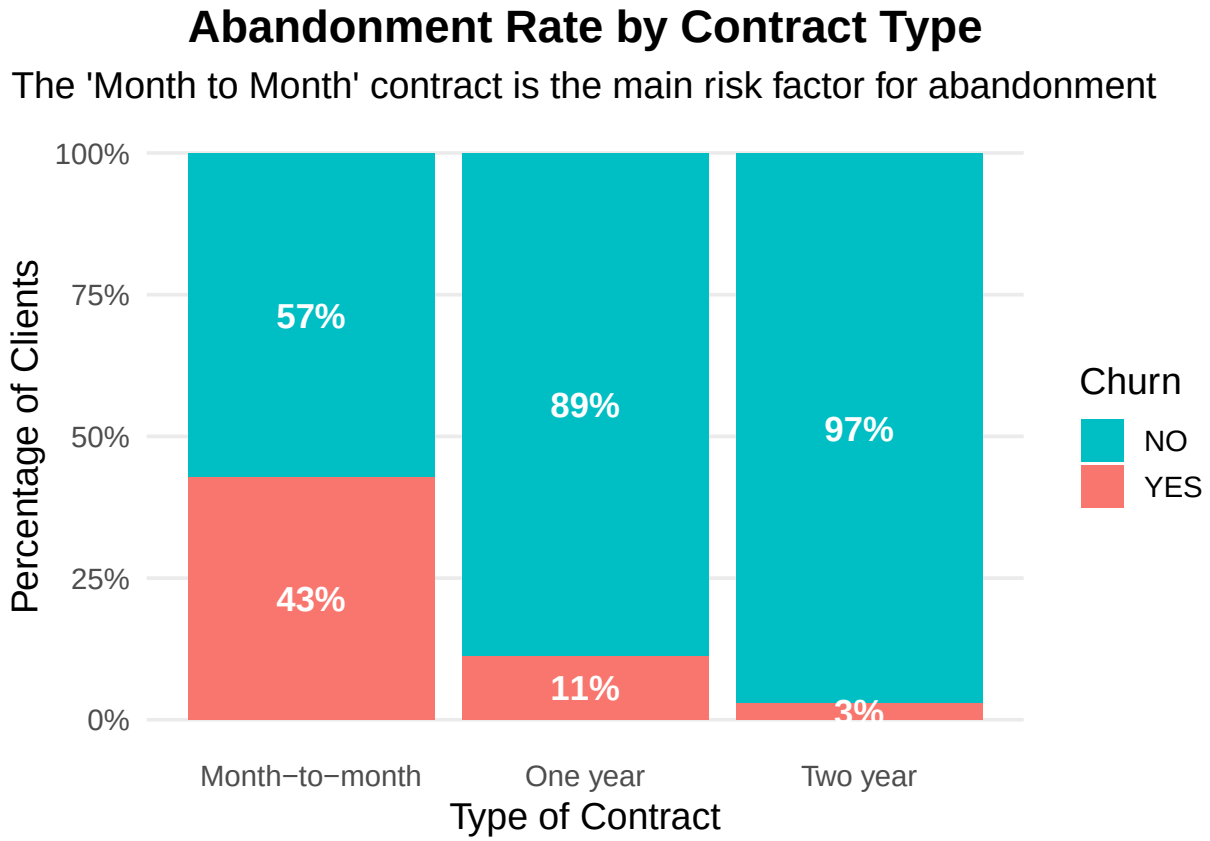
A clear and statistically significant difference is observed between the two groups:

- **Loyal Customers (Non-Churn)**: This group shows a wide tenure distribution centered on high values. The interquartile range (the middle 50% of the data) falls approximately between 15 and 60 months, with a median close to 40 months. This indicates that the loyal customer base is a customer base with a long history with the company.
- **Churning Customers (Yes, Churn)**: In stark contrast, the distribution for this group is heavily skewed to the left (low values). The interquartile range is much more compact and is located at the bottom of the scale, with a median of approximately 10 months. The vast majority of data points are concentrated below 30 months.

Conclusion of the Finding: Churn risk is inversely proportional to customer tenure. The most critical phase for customer retention is their first few months of service. This pattern suggests that company strategies should focus intensely on strengthening the relationship and demonstrating value during the initial (onboarding) period to overcome this vulnerability barrier and guide new customers toward a state of greater loyalty.

Analysis of service conditions

Now we will see how each of the different contract categories influences customer churn:



Analyzing contractual terms provides the most decisive insight of the entire exploratory phase. As illustrated in the percentage bar chart, there is an almost perfect inverse relationship between the length of the contractual commitment and customer loyalty.

- **The High-Risk Segment** (Month-to-Month): Customers without a long-term contract represent the core of the churn problem. With **43%** of these customers canceling their service, this group is not only volatile but also acts as a constant revenue drain. The flexibility offered to them translates directly into a lack of "stickiness" to the service.
- **The Loyalty Segments** (One-Year and Two-Year): By securing a contractual commitment, the churn rate plummets. A one-year contract reduces the rate to **11%**, almost four times less than the monthly plan. The effect is even more pronounced in two-year contracts, where the churn rate is just **3%**, indicating a very high level of loyalty and satisfaction.

Strategic Implications: This finding is not a simple correlation; it is a clear sign of causality in customer behavior. The act of signing a term contract is in itself a powerful retention mechanism. Therefore, any effective strategy to combat churn must have as its fundamental pillar the creation of incentives and fluid paths for monthly plan customers to migrate to one- or two-year contracts. This is the point of greatest leverage for positively impacting business results.

Customer Segmentation by Risk Profile

Exploratory analysis showed us that length of service and contract type are the main indicators of churn risk. To make these findings more actionable, the next step was to formalize this logic through customer segmentation. The goal was to group customers into clear and distinct profiles based on their behavior, allowing us to quantify the risk and value of each group.

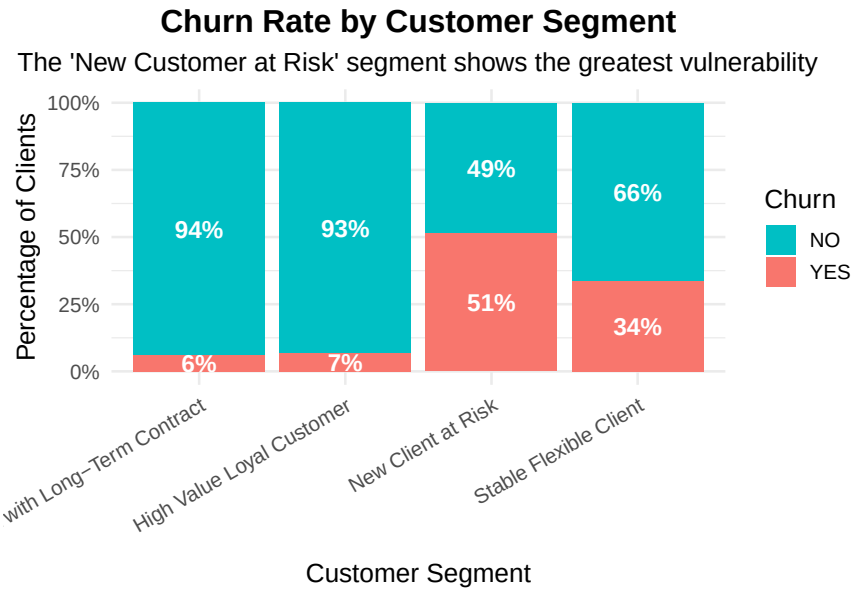
Definition and Creation of Segments

Four main segments were created using a rules-based approach combining customer tenure and contract type.

- New at-Risk Customers: These are customers who have been with the company for 12 months or less and have month-to-month contracts.
- High-Value Loyal Customers: These are customers who have been with the company for more than 24 months and whose contract can be for one or two years.
- Long-Term Contract Customers: These are customers with a one- or two-year contract.
- Flexible Stable Customers: These are customers who have been with the company for more than 12 months and have a month-to-month contract.

Churn Analysis by Segment

Once the segments were created, the churn rate within each segment was analyzed to validate our hypothesis. The following graph demonstrates the effectiveness of this segmentation.



The chart strongly confirms the validity of our segmentation:

- **New Customer at Risk**: This is by far the most problematic group, with a churn rate of 51%. It represents the epicenter of the churn problem.
- **Stable Flexible Customer**: This group, although on a monthly contract, has a more moderate churn rate (34%) due to their longer tenure. It is still a risk, but a lower priority.

- **Loyal High Value Customer and Customer with Term Contract**: These two segments are our pillars of stability, with minimum churn rates of 7% and 6% respectively.

Business Risk Quantification

To understand the financial impact, the value of the most critical segment: "New Customer at Risk" was quantified.

Table 1: Análisis del Segmento de Mayor Prioridad: 'Cliente Nuevo en Riesgo'

Métrica	Valor
Número total de clientes	0
Tasa de abandono específica	NaN%
Ingresos Mensuales en Riesgo (MRR)	$NaN

Segmentation has allowed us to go beyond analyzing isolated variables. We have now identified, named, and quantified a specific group of 1,994 customers who not only have the highest probability of churn but also represent more than $59,615 in monthly revenue at risk. This segment will be the primary target of our retention strategies and the focus of our predictive model.

Construcción del Modelo Predictivo

After identifying the key factors and segments, the next step was to develop a machine learning model capable of predicting the probability of churn for each individual customer. The goal was to create a proactive tool that would allow the business to intervene before a customer decides to leave. Generalized Logistic Regression (GLM) was chosen for its robustness and high interpretability.

The modeling process was divided into three phases: 1. Building a Baseline Model: To establish a performance benchmark. 2. Feature Engineering and Model Improvement: To increase its predictive power. 3. Decision Threshold Optimization: To align the model with business objectives.

Preparing for Modeling

Before training, the data was split into a training set (80%) and a test set (20%). A tidymodels recipe was used to encapsulate all preprocessing steps (creating dummy variables for categorical variables and normalizing numerical variables), ensuring a robust and reproducible process.

Phase 1: Baseline Logistic Regression Model

A first GLM model was trained with the base recipe to establish our benchmark.

The base model achieved a solid initial result, with an AUC of 0.856, indicating that the selected variables already had good predictive power.

Phase 2: Improvement Through Feature Engineering

To improve performance, an advanced recipe was developed that included new features:

- Average_month_value: Total spend divided by tenure, to capture the relative value of the customer.
- Interaction terms between tenure and contract.
- Polynomial terms for MonthlyCharges and tenure to capture nonlinear relationships.

The GLM model was retrained using this new recipe. The result was a slight but consistent improvement in predictive ability, achieving an **AUC of 0.858**. This advanced model was selected as our final model.

Phase 3: Decision Threshold Optimization

A predictive model generates a probability (from 0 to 1). By default, it is classified as "Churn" if the probability is greater than 0.5. However, this threshold is not always optimal for business objectives. For a retention strategy, it is preferable to **identify as many customers as possible who are going to churn (maximize Recall)**, even if that means contacting some who were not going to churn (lower Accuracy).

The ROC curve of the advanced model was analyzed to find the threshold that maximizes Youden's J Index, a method that seeks the best balance between sensitivity (Recall) and specificity.

The optimal threshold was set at **0.2415**. This value will be used as the cutoff point for all operational decisions, ensuring that the model is aligned with the proactive retention strategy. With the completion of this task, we obtained a model that was not only accurate but also strategically calibrated.

Other observations

It's worth noting that the GLM model wasn't the only one I used to train the model. I also used random forests, XGBoost, and deep learning with Keras and TensorFlow, but none of these performed as well as the GLM, indicating that the data had a linear relationship. I decided not to include them in the report code due to their poor performance, but they are worth mentioning.

Evaluation and Interpretation of the Predictive Model

The selected model, a Generalized Logistic Regression (GLM) model with advanced engineering features, demonstrated exceptional predictive capability, with an Area Under the ROC Curve (AUC) of 0.858. More importantly, by strategically optimizing the decision threshold to 0.2415, we transformed the model from a predictive tool to a business action tool. This optimization resulted in a 48% increase in the recall rate, allowing us to identify 82.1% of customers planning to abandon the service. Analysis of the determining factors reveals that contract type, tenure, and payment method are the pillars that explain customer behavior.

Model Performance Evaluation

The robustness of the model was validated on a test dataset. Performance was analyzed by comparing a standard decision threshold (0.5) with an optimized threshold (0.2415), selected to maximize the Youden J Index, achieving a superior balance between capturing customers who abandon (Sensitivity/Recall) and correctly classifying those who remain (Specificity).

Table 2: Model Metric Comparison

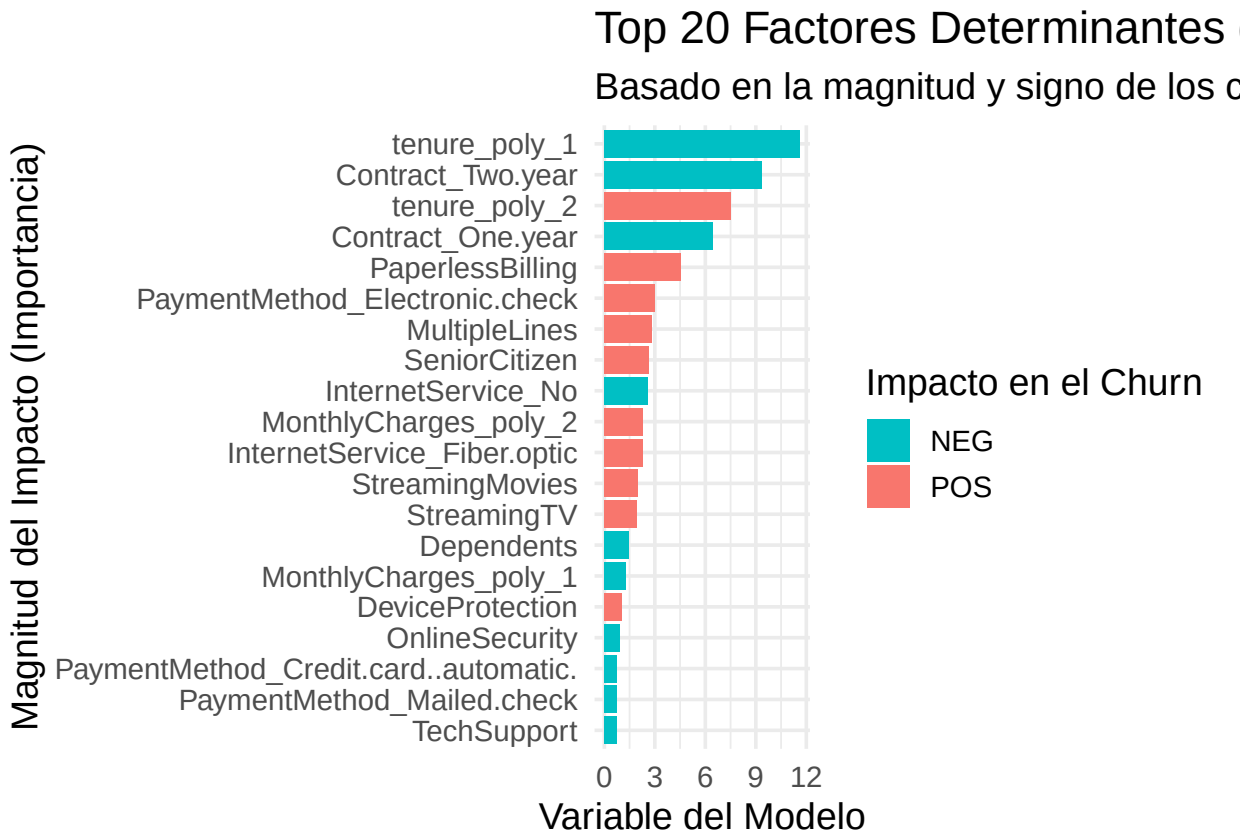
Metrics	Base Model (Threshold 0.5)	Optimized Model (Threshold 0.2415)	Strategic Impact of Optimization
ROC AUC	0.858	0.858	No change (intrinsic predictive ability)
Recall (Sensibilidad)	0.555	0.821	+47.9% (¡Key Strategic Success!)
Precision	0.682	0.506	-25.8% (Calculated compensation)
F1-Score	0.612	0.626	+2.3% (Better overall balance)
Accuracy	0.818	0.747	-8.7% (Non-priority metric)

Evaluation Analysis:

- Predictive Power (AUC): An AUC of 0.858 confirms that the model is highly effective at differentiating between churn-prone and churn-free customer profiles. It is a robust and reliable model.
- Impact of Threshold Optimization: The model's true value is unlocked with the optimized threshold. By lowering it to 0.2415, we consciously accepted lower accuracy (more "false alarms") in exchange for a monumental leap in recall. We are now able to identify 8 out of 10 customers who are actually going to churn. For a business, it is much more profitable to contact a customer who was not going to churn by mistake than to not contact one who is. This optimization perfectly aligns the model with the business objectives of proactive retention.

Churn Driver Analysis

To understand the underlying causes of churn, the importance and direction of each variable within the final predictive model were analyzed. Figure 1 visually summarizes the 20 most determining factors, providing a clear roadmap for strategic interventions.



Variable importance chart showing the impact (positive or negative) of the main predictors of churn.

Interpretation of the chart reveals two opposing forces governing customer loyalty:

1. **Retention Factors (Negative Impact)**: The variables in blue represent the pillars of our loyal customer base. The model conclusively identifies tenure and one- and two-year contracts as the most potent risk reducers. This quantitatively demonstrates that customer lifetime value is maximized through building long-term relationships and solidifying contractual commitments.
2. **Risk Factors (Positive Impact)**: The variables in red are the red flags our model has learned to identify. Factors such as paperless billing and e-check payments emerge as significant drivers of churn. This doesn't mean that these characteristics are inherently bad, but rather that they are associated with a more volatile customer profile or a higher-friction customer experience. Similarly, variables such as MultipleLines and InternetService_Fiber.optic indicate that customers with more expensive or complex services require special attention, as they are more prone to churn.

The model allows us to move from a general overview of churn to a precise diagnosis of its causes. The following strategic recommendations are directly based on the evidence presented in this analysis.

Strategic Plan and Recommendations

The culmination of this project is not the model itself, but the action plan derived from its findings. This section translates the data intelligence obtained into a tangible and actionable business strategy to reduce customer churn.

Final Diagnosis: Predictable and Concentrated Churn

The comprehensive data analysis, validated by a predictive model with an **85.8% accuracy (AUC)**, reveals a fundamental conclusion: customer churn is not a random event, but a predictable result of specific factors. Thanks to the strategic optimization of our model, we are now able to **correctly identify 82.1%** of all customers planning to leave the service, giving us an unprecedented opportunity for proactive intervention.

The company's core problem is a systematic "churn" of customers in its first months, concentrated among those with **month-to-month** contracts. Our segmentation analysis identified a critical group, the **"New Customer at Risk"**, comprising 1,994 customers with an alarming churn rate of **51.4%** and representing **$59,615 in Monthly Recurring Revenue (MRR)** at risk.

The Four Pillars of the Retention Strategy

We propose a four-pillar action plan, where each recommendation is directly linked to quantitative evidence from our predictive model.

Pillar 1: Turn Flexibility into Engagement

- **The Evidence**: The model identifies long-term contracts (One-year Contract, Two-year Contract) as the most powerful factors in reducing churn. The lack of a long-term contract is the main driver of churn.
- **The Strategy**: Implement proactive migration campaigns to move customers from the Month-to-Month plan to term contracts, using the model's risk score to prioritize offers.
- **Recommended Tactics**:
  - 1-Year Migration Offer: For customers with high risk scores, offer a permanent 10% discount on their bill when signing for 12 months.
  - 2-Year Migration Offer: Offer a higher incentive, such as a free month of service or the free addition of OnlineSecurity or TechSupport, services that the model also identifies as loyalty drivers.

Pillar 2: Ensure Success in the First 100 Days

- **The Evidence**: Tenure is the strongest predictor of loyalty. Churn risk is highest early in the customer relationship.
- **The Strategy**: Design and implement an automated onboarding program ("Customer Journey") to ensure a seamless initial experience and quickly demonstrate value.
- **Recommended Tactics**:
  - **Week 1**: Welcome email and post-installation confirmation SMS, offering easy access to guides and support.
  - **Month 1**: Proactive communication ("Did you know…"), showing how to get more out of their plan (e.g., "Manage your account from our app").
  - **Month 3**: Small gesture of appreciation for their loyalty (e.g., a small discount or a one-time bonus).

Pillar 3: Eliminate Friction in the Payment Process

- **The Evidence**: The model identifies e-check payments as a **significant risk factor**. The manual process of paying each month is an avoidable cause of churn.
- **The Strategy**: Aggressively incentivize the adoption of automatic payment methods to eliminate this friction.
- **Recommended Tactics**:
  - **Auto-Pay Discount**: Launch a clear offer: a **flat $5/month discount** for any customer who switches to direct debit or credit card.
  - **Ease of Switching**: Implement a one-click feature in the customer portal to instantly switch payment methods.

Pillar 4: Increase "Stickiness" with Smart Services

- **The Evidence**: The model shows that customers with more complex services like Fiber Optic are more prone to churn, while those with value-added services like TechSupport and OnlineSecurity are more loyal.
- **The Strategy**: Use the model's risk score to make cross-sell offers that increase customer integration into our ecosystem.
- **Recommended Tactics**:
  - **For High-Risk Fiber Customers**: Proactively offer a TechSupport or OnlineSecurity package with a special discount for the first 3 months.
  - **Value-Based Communications**: Promote these services not as an extra cost, but as an "experience enhancement" that guarantees peace of mind and security.

Estimating Potential Impact and Return on Investment (ROI)

To quantify the value of these interventions, we made a conservative estimate for the most vulnerable segment: **"New Customers at Risk"**.

Table 3: Estimación del Impacto Financiero de las Estrategias de Retención

Métrica	Valor
Segmento Analizado	Cliente Nuevo en Riesgo
Tamaño del Segmento	1994 clientes
Tasa de Abandono Actual	51.4%
Hipótesis de Reducción	20%
Clientes Retenidos Adicionalmente (por ciclo)	~ 205
Ingresos Mensuales Salvados (MRR)	$11,935
Impacto Anualizado Estimado (ARR)	$143,220

Implementing this plan has the potential to retain approximately **205 additional customers** per cycle, which translates to nearly **$143,220 in annual revenue recovered** from this segment alone. The return on investment is extremely high, as most of these tactics can be automated.

Conclusion and Implementation Roadmap

This project has achieved a fundamental transformation in the way the company addresses customer churn. We have moved from a general understanding of the problem to an **accurate, quantitative, and actionable diagnosis**, supported by a robust predictive model.

The key finding is that churn is not a monolith, but a mosaic of predictable behaviors. We have demonstrated that by focusing on the **right factors (contract, tenure, and payment method)** and the **right segments ("New Customer at Risk")**, we can move from a reactive retention strategy to a **proactive and intelligent retention culture**.

The developed predictive tool is not just a technical artifact; it is a **strategic asset** that, if used correctly, can generate substantial and sustainable financial value, estimated at more than **$143,000 annually** by intervening in the most vulnerable segment alone. The true success of this project will be measured by the organization's ability to integrate these findings into its daily operations.

Next Steps: A Roadmap for Successful Implementation

To ensure that the results of this analysis translate into real and lasting impact, the following implementation roadmap is recommended, divided into three phases:

Phase 1: Operational Deployment and Technology Integration (Short Term: 1-2 Months)

The first step is to make the model accessible and useful to frontline teams (retention, marketing, customer service).

- **Action 1.1: CRM Integration**: Collaborate with the IT team to integrate the final predictive model with the Customer Relationship Management (CRM) system. The goal is for each customer record to automatically display two new fields:
  - churn_score: The probability of churn (from 0.00 to 1.00).
  - retention_alert: A flag (e.g., "Yes"/"No") that is triggered if the churn_score exceeds the optimized threshold of 0.2415.
- **Action 1.2: Team Training**: Conduct training sessions with the retention and customer service teams to explain what the risk score means and how they should use it to prioritize their calls and offers.

Phase 2: Validation and Learning through a Pilot Project (Medium-Term: 3-6 Months)

Before a large-scale rollout, it is crucial to validate the recommended strategies in a controlled environment to measure their actual effectiveness and optimize offers.

- **Action 2.1: A/B Test Design**: Select a cohort of 1,000 customers marked with alert_retention = "Yes." Randomly divide them into:
  - Treatment Group (500 customers): They will receive the proactive offers from Pillar 1 (contract migration) and Pillar 3 (auto-pay discount).
  - Control Group (500 customers): They will not receive any proactive offers.
- **Action 2.2: Measurement and Analysis**: After a 3-month period, compare the churn rate between the two groups. The goal is to confirm that the intervention reduces churn in a statistically significant way and calculate the actual ROI of the offers.

Phase 3: Continuous Monitoring and Model Evolution (Long Term: Continuous)

A machine learning model is not static; its performance can decline over time as customer behavior or market conditions change (a phenomenon known as "model drift").

- **Action 3.1: Create a Performance Dashboard**: Develop a dashboard (e.g., in Power BI, Tableau, or R Shiny) that monitors in real time:
  - The overall churn rate and by key segments.
  - The model's performance (AUC, Recall) over time.
  - The adoption rate of retention offers.
- **Action 3.2: Retraining Plan**: Establish a plan to retrain the model every 6 to 12 months with new data to ensure it remains accurate and relevant, adjusting both the coefficients and the decision threshold if necessary.

This project is coming to an end, but the journey toward a fully data-driven organization is just beginning. It has been a pleasure collaborating on the development of this solution, and I am confident that, with the implementation of these strategies, the company will see a significant impact on its retention metrics. I appreciate the opportunity and the support provided throughout the process, and I remain at your disposal for future discussions and collaborations.