TELCO CHURN ANALYSIS AND PREDICTION

A data-driven approach to customer retention

JULY 30, 2025

ALEX YEPUR
alexyepur@gmail.com

Contents

1. Executive Summary

In this project, we analyzed a combined dataset containing information about customers of a telecommunications company **(Telco)**, including demographic, geographic, contract, and service usage data, in order to identify patterns associated with customer churn behavior.

The main objectives of the project were to:

- correctly merge several related datasets and perform a soft handling of inconsistent or missing values;
- explore the dataset in depth, visualize behavioral trends and identify hidden patterns relevant to business decision-making;
- design and implement a modular preprocessing pipeline that standardizes data preparation and safeguards against leakage during model training;
- identify the best-performing models and optimize their hyperparameters for accurate prediction of each target variable.

The final merged dataset contained information about **7,043 customers**, each described by **43 features**, including **6 target variables**. All data described customer characteristics, including demographic profile, location, and service usage details.

During analysis, we found that although the dataset covers a broad range of factual and objective variables, it lacks a key element: the customer's subjective experience. The dataset does not capture the customer's empirical experience - that is, direct feedback or behavioral signals such as satisfaction scores or frequency of support requests - which are typically among the strongest predictors of churn and engagement. As a result, we encountered significant limitations when trying to predict several targets based solely on pure facts and the performance of the models was insufficient across many metrics.

The most reliable target turned out to be the **general binary churn flag** whether a customer is likely to leave in the near future.

Using the **XGBClassifier**, we achieved the following performance on this target:

- **F1 Score: 0.652**
- **Accuracy: 0.820**
- **Log Loss: 0.378**

While these metrics fall short of ideal expectations (F1 ~ 0.7 0.75), they are significantly better than baseline guessing.

For comparison, the best baseline gave:

- F1 Score: 0.361
- Accuracy: 0.513
- Log Loss: 0.693

The most important features for churn prediction were **the contract length and the presence of a prepaid period**. However, even the most predictive features mainly reflect financial lock-in rather than true customer satisfaction. Customers tend to stay not because they are satisfied, but because leaving would incur financial penalties. Thus, these features capture retention driven by contractual constraints, not by genuine loyalty. While these insights are valuable and validate expected business logic (longer contracts = more stable customers), **they fail to capture Telco's competitive position or its true value proposition from the customer s perspective.**

2. Introduction

Project Motivation

Since this project was designed as a demonstration, we can define the motivation as if it were a real-world scenario: *"A Telco company has a large customer base and has recently observed a growing churn rate. As each lost customer represents a loss of potential profit, the company turned to a data scientist to identify the possible reasons behind this trend, uncover previously unseen patterns, detect potential anomalies, and provide reasonable recommendations for further actions."*

Objectives

The objective of this study is to identify the factors that have a moderate to strong impact on whether a customer stays or leaves. In addition to manual and visual analysis, machine learning models will be applied to uncover hidden relationships between features and the target variable.
The analysis will aim to segment customers based on their likelihood to churn and highlight areas where intervention may be most effective. The project will be considered complete when we have a clear understanding of the key drivers behind customer churn and actionable strategies to address them.

Target Audience

The target audience of this project, in line with the stated motivation, is the **Customer Retention Department** of Telco. This department is directly responsible for minimizing churn, as its primary objective is to retain the existing customer base and prevent revenue loss. The insights and recommendations derived from this analysis will support their efforts by enabling the development of personalized retention strategies, targeted offers, and the identification of customer segments with a high risk of churn.

3. Methodology

Throughout this project, we aimed to perform a thorough analysis and build several machine learning models for each defined target. Our tasks can be divided into three main categories: **preparatory**, **analytical**, and **modeling** tasks.

Preparatory Tasks

Before starting a project, we needed data, which we sourced from an open dataset on Kaggle. The Telco Customer Churn dataset was provided by the user **blastchar** on the Kaggle platform: https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

License: CC0 1.0 Universal (Public Domain) free to use, modify, and distribute without restrictions.

Analytical Tasks

All core analysis was conducted using **JupyterLab** with a **Python** kernel. For visualizations, we used standard Python libraries such as matplotlib and seaborn, as well as Power BI software. For data exploration and processing, we used built-in functions from scikit-learn and scipy, along with custom-built functions and transformers based on them.

Modeling Tasks

One of the distinct aspects of this project is its multi-target nature. This fact prevents us from defining a single modeling objective upfront, as different targets require different approaches. Binary classification was used to predict the primary target: the churn flag. Multiclass classification was applied to churn category and satisfaction score. Regression was used for predicting churn score and CLTV.

Important note: During the process, we discovered that modeling produced relevant and satisfactory results only for the **churn flag**. All other targets demonstrated weak predictability with the available data. This limitation and its implications are discussed in later sections.

4. Data Overview and Analysis

4.1 Dataset Description

As previously mentioned, the dataset contains information on **7,043 customers**, some of whom have already churned. After merging the source tables and removing duplicated features, we retained **43 features**, including **6 target variables**.

Target Variables:

churn - binary flag indicating whether the customer has left the company.

churn_value - numeric equivalent of churn (0/1).

churn_score - churn risk score, with higher values indicating greater risk.

churn_category - categorical reason for churn if applicable.

satisfaction_score - customer satisfaction rating from 1 to 5.

cltv - Customer Lifetime Value (predicted revenue contribution).

Features:

All remaining features represent factual data - they are not directly predicted but instead reflect subjective characteristics of the customer. All features below are self-descriptive, unless otherwise noted.

1. Metadata

customer_id

2. Customer Demographics

gender, age, under_30, partner, number_of_dependents

senior_citizen - binary flag indicating if the customer is 65 years or older (1 = senior).

3. Geographic Information

city, zip_code, latitude, longitude

population - number of people living in the customer's zip code area (regional density proxy).

4. Contract and Billing Information

tenure_in_months, paperless_billing, payment_method, monthly_charge, total_charges, total_refunds, total_revenue, number_of_referrals

offer - promotional campaign or discount plan received (categorical feature with named offers or "None");

contract - contract duration: month-to-month, one year, or two years.

5. Usage and Service Features

total_extra_data_charges, total_long_distance_charges, avg_monthly_long_distance_charges, avg_monthly_gb_download, phone_service, multiple_lines, internet_type, online_security, online_backup, device_protection_plan, premium_tech_support, streaming_tv, streaming_movies, streaming_music, unlimited_data

4.2 Target Variables Analysis

1. Churn / Churn Value

According to the dataset, **73.4%** (5,163 customers) are still active clients, while **26.6%** (1,869 customers) have churned. This highlights a significant class imbalance however, such a distribution is realistic when 50/50 split would indicate a serious crisis within the company. Therefore, the imbalance here reflects a relatively healthy customer base retention rate, but still has to be addressed on modeling stage.

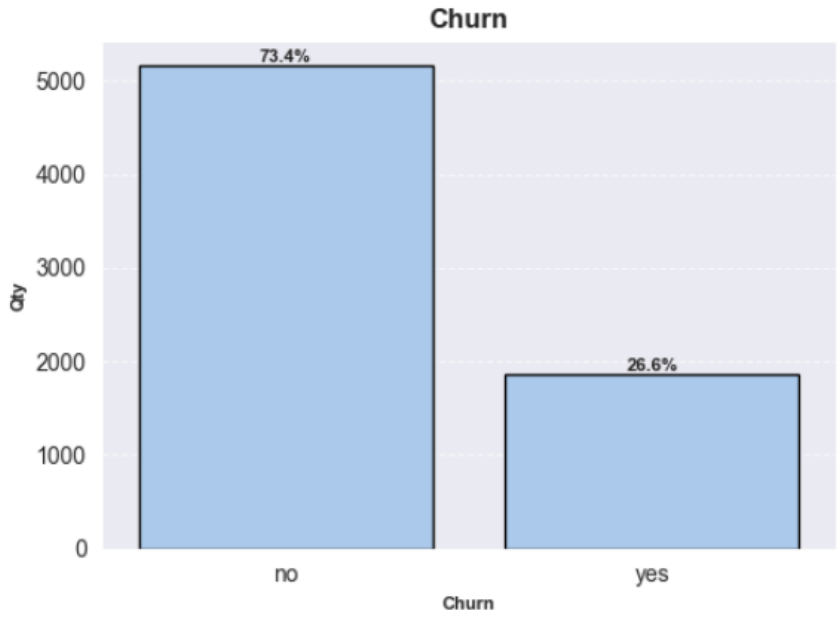

Figure 4.1 Churn Distribution

2. Churn Score

This feature ranges from 0 to 100. The distribution is fairly uniform between 23.2 and 96, with a clear concentration (~17%) in the 68.7 to 77.8 interval. On its own, the churn score tells us no significant insights. However, once grouped and cross-analyzed with the churn target, clear patterns emerge:

- All customers who churned had a churn score above **59.6**.
- Customers with a score above **83.87** churned with 100% certainty.
- Customers with a score below **59.6** never churned.

This allows us to interpret the churn score as a 3-zone risk scale:

- **0** to **59.6**: Stable zone customers remain with 100% probability.
- **59.6** to **83.87**: Risk zone increasing churn probability as score rises.
- **83.87** to **100:** Critical zone customers churn with 100% probability.

Although churn score is conceptually a valuable target, in further steps we found out that our model failed to produce reliable predictions for it. As an alternative, Telco could either continue using the internal methods originally used to generate this score or improve its accuracy by collecting richer empirical data on customer sentiment. If the prediction quality improves, churn score has the potential to become one of the most effective early-warning indicators for churn risk.

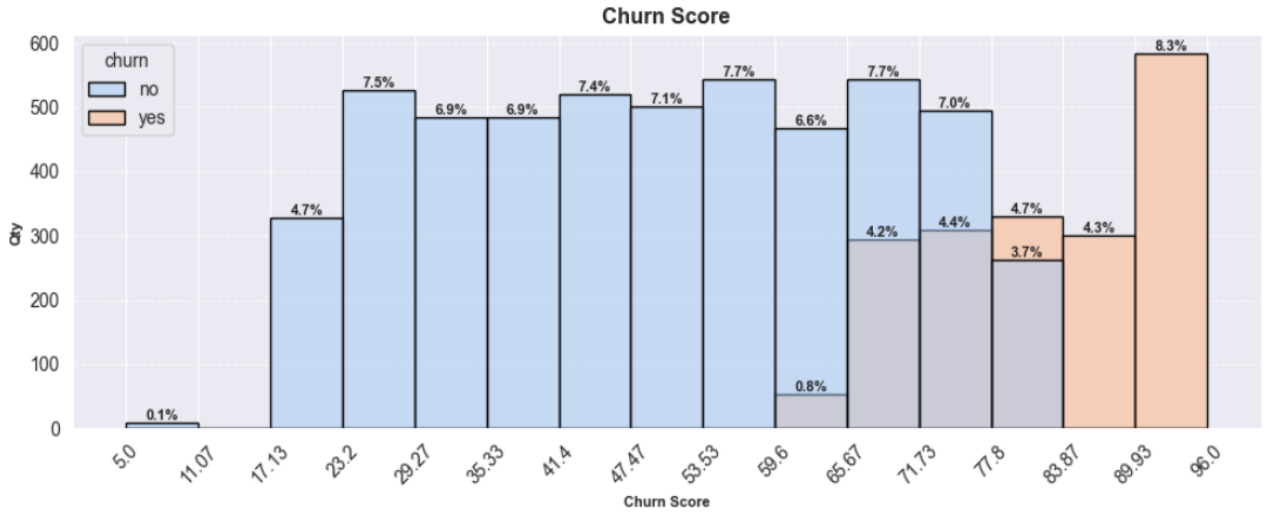

Figure 4.2 Churn Score Distribution by Churn Status

3. Satisfaction Score

A similar pattern is observed with the satisfaction score. Initially, its distribution seems somewhat balanced:

- 20.5% of customers rated their satisfaction as 1-2
- 41.7% gave scores of 4-5
- 37.9% chose 3 indicating borderline satisfaction

This suggests a generally satisfied customer base, though the high proportion of 3s reflects a lack of strong sentiment. When we overlay churn status onto this distribution, a clearer signal emerges:

- Customers with scores 1-2 always churned (100% churn rate)
- Customers with scores 4-5 always stayed (0% churn rate)
- Those with score 3 had a ~16% chance to churn, 84% chance to stay

This again provides a well-segmented structure, allowing clear interpretation. Like churn score, satisfaction score becomes much more valuable when combined with churn data, reinforcing its utility as a potential driver or proxy for churn behavior.

Unfortunately, looking ahead, **we were not able to achieve meaningful results in predicting satisfaction score** using only the available factual data. The solution here is the same as with churn score: if Telco already has internal methods to assign satisfaction scores with such high correlation to churn, **those scores are among the strongest possible indicators** and should be leveraged accordingly.

However, if these scores were synthetically generated or arbitrarily assigned, then the dataset is lacking the **empirical, experience-based input** required for robust prediction. In that case, Telco should consider enhancing its data collection methods with **richer customer experience metrics** - such as interaction logs, complaint frequency, contacts to support or usage anomalies - to turn satisfaction score into key churn-prediction feature.

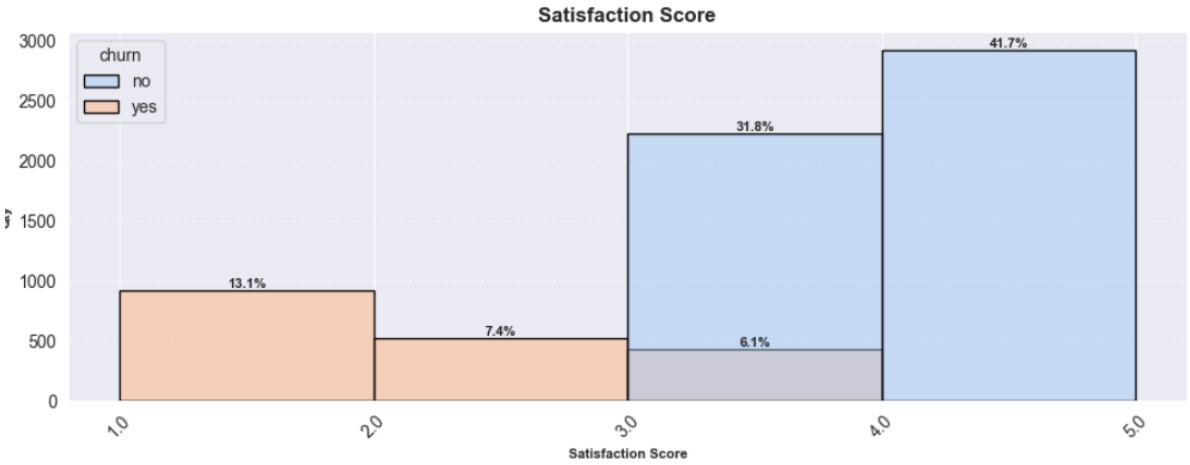

Figure 4.3 Satisfaction Score Distribution by Churn Status

4. Churn Category

Unlike the other target variables, churn category is only relevant for customers who have actually churned. Within this subset, the breakdown is as follows:

- 45% left for a competitor;
- 33% cited dissatisfaction with service or attitude;
- 11.3% left due to pricing concerns;
- 10.7% churned for other unspecified reasons.

Given the high number of borderline cases identified via the satisfaction score, this distribution is both expected and rational. In a competitive market, if a company does not consistently deliver high-quality service and actively monitor customer retention, competitor-related churn becomes inevitable. Customers especially those with neutral or wavering satisfaction are likely to explore alternatives, particularly if competitors offer better pricing, support, or perceived value.

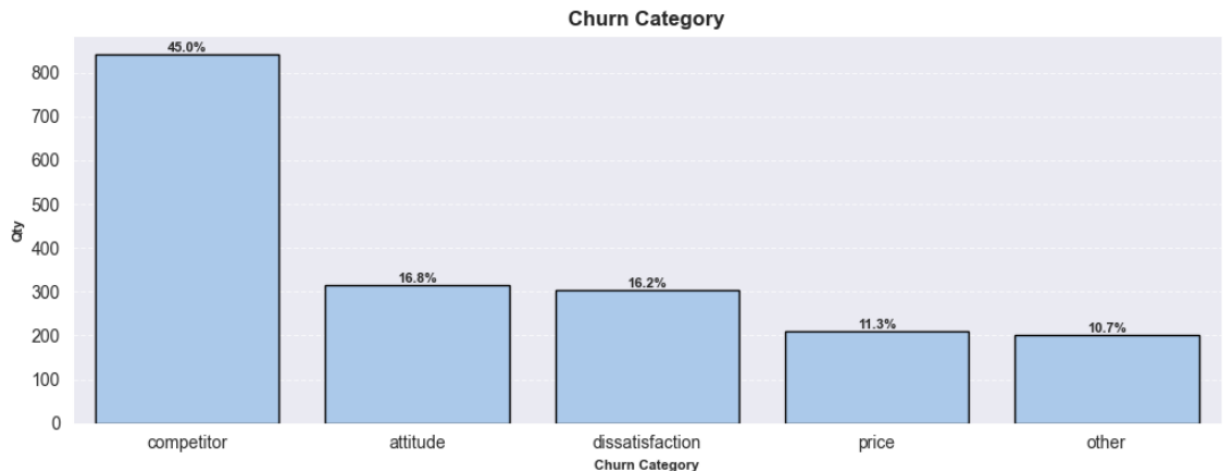

Figure 4.4 Churn Reason Distribution

5. CLTV

In the case of CLTV, it is more difficult to draw immediate conclusions by simply looking at its distribution or its interaction with churn.

The target follows a *roughly normal distribution*, and churn appears to be *evenly distributed* across different CLTV values. This suggests that CLTV, in its current form, does not have a strong direct relationship with churn behavior.

CLTV is heavily influenced by features such as **tenure_in_months**, **monthly_charge**, and **contrac**t which is expected. A long-standing customer on an expensive plan will naturally have a high CLTV, while a new customer on a low-cost plan will have a low CLTV.

As a result, while CLTV is mathematically meaningful and based on intuitive drivers (payment amount and customer lifespan), it offers little additional insight beyond what its input features already convey. Ultimately, this metric reflects the abstract nature of predicting long-term future value it's logical and consistent, but not inherently informative without deeper behavioral or contextual enrichment.

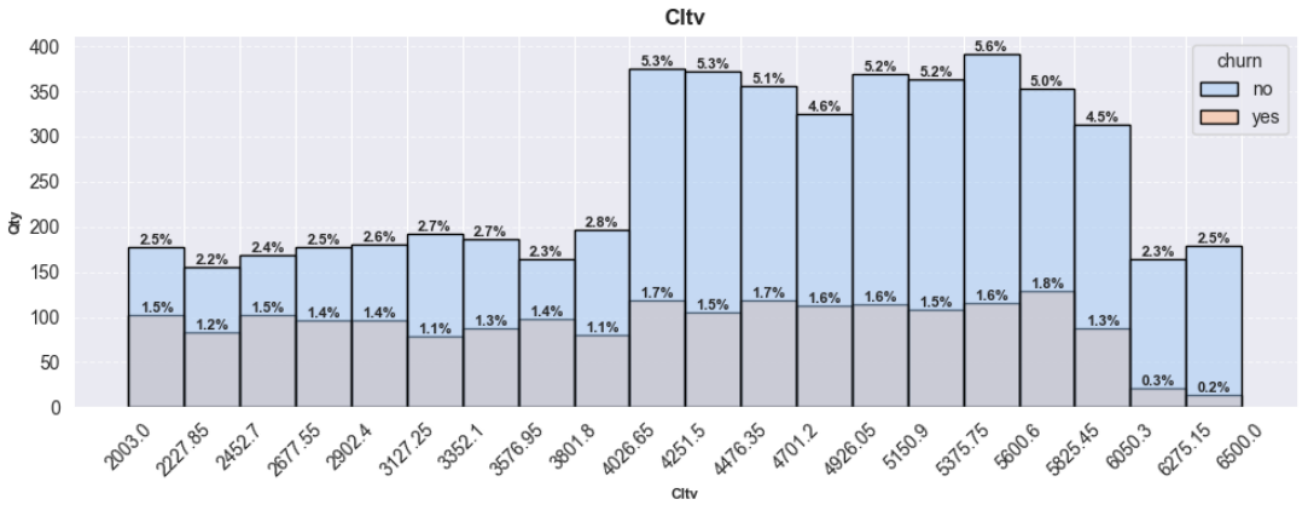

Figure 4.5 CLTV Distribution by Churn Status

4.3 Features Analysis

During the feature analysis, we conducted an initial review of all available variables. Wherever we identified imbalanced distributions, we hypothesized the presence of subtle, non-obvious patterns or behavioral signals. For such features, we performed a more detailed exploration by visualizing their distributions both in the **overall dataset** and specifically **among churned customers**. This approach allowed us to directly assess how certain characteristics differ in the context of churn, without losing vital sight of the general data structure. It provided a clearer view of which factors are more likely to be associated with churn behavior.

1. Contract

One of the most important indicators for churn prediction. The overall distribution shows an expected pattern - customers tend to prefer shorter contracts over longer ones:

- Month-to-month: 55.11%;
- One year: 23.96%;
- Two years: 20.93%.

However, the picture changes dramatically when we look only at the churned customers:

- Month-to-month: 88.55%;
- One year: 8.88%;

- Two years: 2.57%.

Over **88%** of churned customers were on a **month-to-month** contract, despite this group making up just 55% of the total population. This clearly shows that customers on short-term contracts are significantly more likely to churn. The reasons could vary - from fast dissatisfaction to the absence of financial commitment or simply not enough time to get used to the service. Notably, while one-year and two-year contracts are nearly equally represented overall, two-year contracts show over **three times lower** churn than one-year ones.

2. Internet Service Type

One of the unexpected patterns in the dataset emerges when examining churn distribution across internet service types. Distribution across all users shows relatively expected trends:

- Fiber optic: 43.16%
- DSL: 23.44%
- No internet: 21.62%
- Cable: 11.79%

But significant changes occurred among churned users:

- Fiber optic: 66.13%
- DSL: 16.46%
- No internet: 11.40%
- Cable: 6.05%

Clearly, **fiber optic** users churn much more frequently than any other category. This pattern may be driven by several overlapping factors. First, fiber plans are typically positioned as premium products with higher monthly fees, leading to higher customer expectations regarding speed, reliability, and service quality. If those expectations are not met, dissatisfaction may arise faster.

Second, fiber infrastructure is more prevalent in urban or suburban areas, where the market is more competitive and alternative providers are readily available. In such environments, switching costs are lower, and customers are more tend to exploring better deals or promotional offers.

In contrast, **DSL** and **cable** often serve customers in rural or semi-rural areas, where internet options are limited and switching is harder or more costly. These users might also exhibit lower engagement or demand, leading to lower sensitivity to service shortcomings and, as a result, lower churn. This

difference in context creates a strong structural divide in churn behavior based on infrastructure availability and market saturation.

3. Payment Method

Payment method is another strong predictor for churned customers. If we look at general distribution across full dataset, we will see such a picture:

- Electronic check: 33.63%;
- Mailed check: 22.81%;
- Bank transfer (automatic): 21.93%;
- Credit card (automatic): 21.63%.

The distribution is roughly even, except that **electronic check** is used about **1.5x more often** than the others. However, picture changes when we isolate the churned group:

- Electronic check: 57.30%;
- Mailed check: 16.48%;
- Bank transfer: 13.80%;
- Credit card: 12.41%.

Electronic check accounts for **over half of all churned customers** - more than **triple the share** of the next most common method. This cannot be explained solely by its general overall frequency and likely reflects behavioral or demographic traits of this segment. For reasons that may include disengagement, financial instability, or lower digital commitment, e-check users churn significantly more often - even compared to the more conservative mailed check users. While causality is hard to confirm, this pattern suggests that customers selecting e-check require more attention and potentially targeted offers, as they represent a high-risk group from the very beginning of their customer journey.

4. Premium Tech Support

Another interesting trend is observed among customers who use premium tech support. Initially, only 29% of users opt for this service. However, among churned customers, just 16.6% had premium support.

This indicates a nearly twofold reduction in churn probability. For example, from 5,000 customers without premium support, around **30% (1,559)** churned. From the 2,000 who had it, only **15% (310)** did.

This correlation is logical: support access acts as a safety net - customers experiencing issues are more likely to resolve them instead of leaving. Premium support enhances this even further by reducing wait times and providing faster resolutions, improving satisfaction and therefore reducing churn risk.

5. Senior Citizen

Senior customers make up about **16% of the total base** but account for **25% of churned users**. While not as strong as previous factors, the trend is notable. This group may be more churn-prone due to a range of reasons: post-retirement income reductions, advice from younger relatives, lack of digital skills to navigate services, or even shorter planning horizons. Even though the impact is moderate, it should not be ignored.

6. Partner

Another medium-impact feature is whether the customer has a partner. The overall distribution is balanced

- No partner: 51.75%
- With partner: 48.25%

Among churned users, however, only **35.79%** have a partner.

This may point to more stable behavior in partnered households, possibly because decisions are made jointly, affect more than one person, and therefore tend to involve more deliberation, mutual agreement, and resistance to change. Additionally, switching providers may disrupt family routines, introduce uncertainty, or require coordination between multiple users - factors that collectively contribute to a stronger inertia and lower likelihood of churn.

7. Monthly Charges

Analyzing monthly charges variable, we found out that there is a clear positive correlation with churn likelihood. Churned users are heavily concentrated in the **$70-$100** range, while users paying **$20-$50** show significantly lower churn. This is intuitive - higher price implies higher expectations and higher dissatisfaction risk.

Features Analysis Summary

Based on our analysis of the most influential features affecting churn, we can draw several conclusions and formulate targeted recommendations.

The highest-risk customer segment can be described as a **single senior** on a **month-to-month contract** using **fiber optic internet**, paying via **e-check**, and spending over **$70 per month**, despite having **opted out of premium tech support**.

Of course, this is a generalized profile, and each factor on its own can significantly impact churn. Here, one of the strongest individual predictors of churn is the choice of **month-to-month** billing. This often reflects an initial lack of long-term commitment or confidence in the service, especially since long-term contracts tend to be more cost-effective. Customers choosing monthly plans should be prioritized for **conversion campaigns** to long-term contracts - offering additional benefits such as price reductions, extra features, or bonuses for loyalty. Extending their contract duration would directly improve retention and help stabilize a group otherwise prone to churn.

Additionally, **fiber optic** users require extra attention. While fiber offers superior speed and performance, it also comes with higher expectations and price sensitivity. The company must ensure that its infrastructure and support can consistently meet the demands of this premium tier internet. Otherwise, customers disappointed by the service-to-cost ratio are likely to leave or keep being unsatisfied for prolonged period of time. For this segment, it is advisable to conduct **pre-installation consultations** to assess whether fiber is truly necessary for the customer's usage needs. If expectations are mismatched, consider offering more appropriate plans or free bundling add-ons like service guarantees or priority support.

For other customer features, such as those identified by age group, marital status, or payment method, it's more difficult to assign direct actions - but they may still signal elevated risk. For example, seniors, customers without partners, or those who choose e-check payments have shown disproportionately higher churn rates in our dataset. These factors might reflect financial caution, behavioral inertia, or digital disconnection, each of which can create friction in maintaining long-term engagement. One effective strategy could be to offer **free access to premium tech support** for such high-risk profiles. In our dataset, this service was associated with a 50% reduction in churn, suggesting that closer guidance, better responsiveness, and more proactive service delivery, directly address the pain points of these users. By identifying and supporting these segments early, we can minimize avoidable departures and foster greater satisfaction and loyalty across the board.

5. Feature Engineering

During the feature engineering phase, three new features were introduced: **did_renew_contract**, **has_active_prepaid_period**, **and services_count**.

In addition, some existing variables were restructured using binning and mapping, and unnecessary columns were removed via dropper components. However, this section will focus exclusively on the newly created features that potentially carry novel information.

1. did_renew_contract

This binary flag indicates whether a customer renewed their contract at least once (1) or only paid for a single billing cycle without renewal (0). According to our analysis: **87.6%** of customers renewed their subscription at least once. Among them, only 53% were on a month-to-month contract.

Conversely, **13.4%** of customers never renewed. Interestingly, within this subgroup, the distribution by contract type was counterintuitive: the majority **(69%)** were on month-to-month contracts. This may reflect either a large proportion of new customers who have not yet had the opportunity to renew because their initial contract period is still ongoing, or a significant level of dissatisfaction among these users, despite the longer commitment. Supporting the latter, 62.9% of these non-renewing users ended up churning. This flag can thus serve as a key retention signal and should be closely monitored by customer success or retention teams.

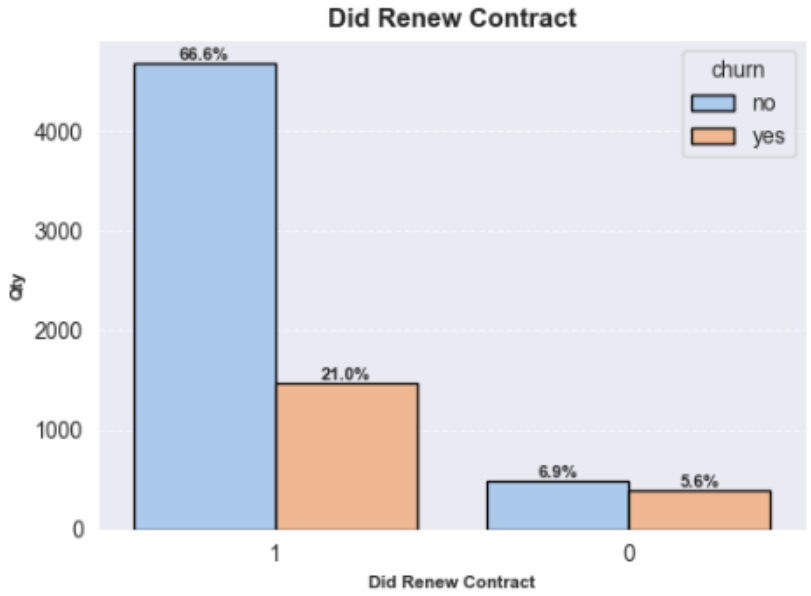

Figure 4.6 Contract Renovation Distribution by Churn Status

2. has_active_prepaid_period

This feature represents a binary flag indicating whether the customer is currently within a prepaid period. For instance, a customer on a 12-month contract with only 5 months of tenure receives a value of 1, while a customer with the same contract but 12 months of tenure would receive a 0. Customers on month-to-month plans always receive 0 by definition.

Our initial hypothesis was that users who had prepaid for their service - especially on long-term contracts - would almost never churn during the active paid period, with at most a few rare exceptions. However, the data contradicted this assumption: **7.3%** of customers with an ongoing prepaid period still churned, which translates to nearly **190 users**.

More surprisingly, a deeper analysis of their tenure revealed a strong concentration of these churners among customers with very long service durations - **over 50 months**. This implies that they were long-standing, previously loyal clients who had renewed their contracts multiple times over the years. Despite that, they chose to leave abruptly before the end of a paid period, incurring a financial loss in doing so. Such behavior indicates the presence of a strong, disruptive trigger.

Further investigation into this subgroup revealed that over **half of them** had received **no special offers** or targeted retention benefits. Additionally, **43%** of them cited **“competitor”** as the reason for leaving.

Although small in absolute numbers, this group is analytically significant. Their behavior exposes a critical gap in the company’s retention strategy: even the most loyal customers are at risk of abrupt churn if they feel ignored or unrewarded - especially in the face of appealing competitor offers. This finding emphasizes the necessity of reinforcing loyalty with ongoing value, personalized incentives, and proactive engagement. Failing to do so not only lead to silent, avoidable losses, but also erodes the very foundation of long-term customer value.

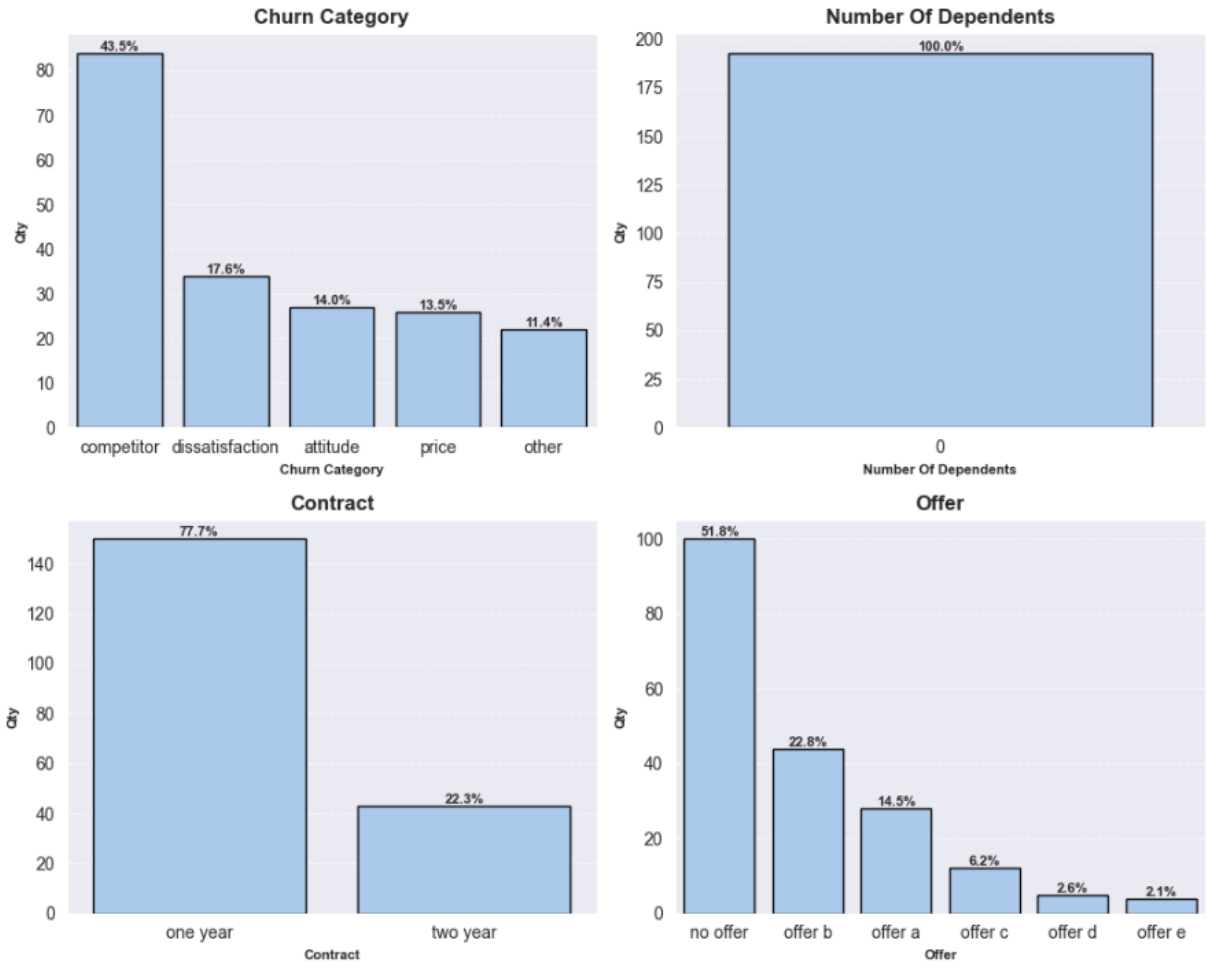

Figure 4.7 Features Distribution Among Churned Customers with Active Prepaid Period

3. services_count

This feature represents the number of telecom services a customer is subscribed to, ranging from 1 to 11. The distribution is quite interesting: the largest share - around **17%** - uses only **one service**. Other significant segments are those using between 4 and 8 services, each comprising about 10% of the customer base. Only **2.3%** of customers are subscribed to all **11 services.**

What makes this feature especially insightful is the churn distribution across different service counts. Less than 10% of users with only one service churned. Surprisingly, the lowest churn ratio was observed among those subscribed to all 11 services - **95.2%** of them remained. In contrast, starting from **3 services**, the churn rate increases sharply: almost **half** of the customers with 3 services have churned. As the number of services increases beyond this point, churn begins to gradually decline - e.g., for those with 6 services, roughly one-third churned.

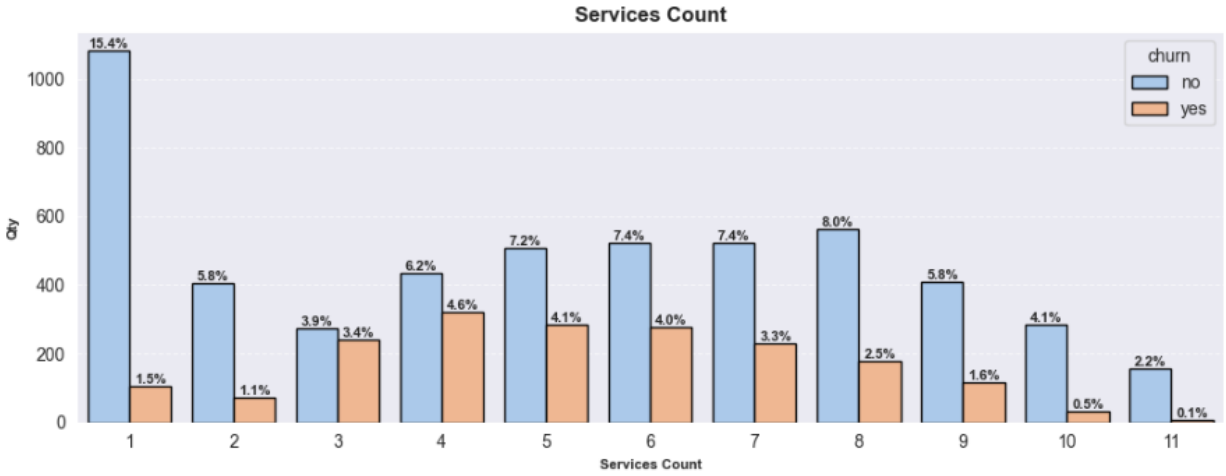

Figure 4.8 Services Count Distribution by Churn Status

This creates an unusual trend: customers with either very few (1–2) or the maximum number of services (10–11) tend to be the most loyal, while those in the middle range (3–8 services) show the highest churn risk. This pattern likely reflects a trade-off between convenience and perceived value for money. Customers with only one service are likely to incur minimal costs and thus feel little need to leave. Those with 3–5 services face significantly higher bills but may not yet feel a strong attachment to the company, making them more open to switching providers - especially if they perceive better deals elsewhere. As the number of services increases toward the upper end, loyalty tends to increase—probably because switching would mean a major disruption in convenience, requiring the customer to find alternatives for multiple services. *These insights can be highly valuable for the retention team: they help in identifying unstable customer segments in advance and designing targeted offers - for example, discounts on bundled services - to retain those in the mid-range, who are most at risk.*

6. Preprocessing and Modeling

6.1 Initial Model Benchmark

One of the core principles guiding our work was to ensure modularity, reproducibility, and leakage prevention. To achieve this, we implemented a pipeline architecture using the scikit-learn library. This pipeline was designed to operate on a dataset that is nearly raw. Before applying it, we manually performed minimal corrections: we removed duplicate columns, fixed data entry errors, and verified the absence of missing values.

The main stages of our pipeline were as follows:

1) Outlier Handling

At this step, we analyzed the skewness and kurtosis of numerical features to assess distributional imbalance. Features showing unacceptable levels were handled individually, depending on their structure and nature. We applied various methods such as discretization, flagging, and power transformations (e.g. Yeo-Johnson) to address these issues. To maintain modularity, all transformations were implemented as standalone transformers within a sub-pipeline.

2) Feature Engineering

Following the same modular structure, we applied engineered transformations beyond those described earlier. Unlike the previous step, this one involved logical feature transformations (based on domain knowledge) and converting certain variables into modeling-ready numeric representations. As it was previously mentioned, all steps were encapsulated into reusable transformers and included as a feature engineering sub-pipeline steps.

3) Encoding and Scaling

Once feature transformations were completed, we ensured that all data was numeric and scaled. This is a necessary step before applying most ML algorithms. To maintain structure and avoid information leakage, we created two consecutive Column Transformers:

- the encoding transformer was responsible for converting categorical variables depending on their type: binary, discrete, high-cardinality, ordinal, or nominal.
- the scaling transformer then standardized all numerical variables (excluding binary) using StandardScaler, ensuring all features were on a consistent scale.

Although this two-step encoding and scaling process deviates from the more common practice of applying all transformations within a single unified feature transformer, it provides a critical advantage: it ensures that encoded categorical features are also standardized. Otherwise, their encoded values (e.g., ordinal levels or frequency counts) would remain in varying numeric ranges, potentially introducing artificial scale differences that mislead the model. This layered design guarantees that all features, whether original or engineered, contribute in a consistent manner.

4) Multicollinearity Filtering

To improve model generalization, we implemented a custom transformer that drops multicollinear features. Using Variance Inflation Factor scores, we identified strongly interdependent variables. We cross-validated this with a correlation matrix to decide which features should be removed, giving higher priority to more general features. Also, as a result of this step, the newly engineered **services_count** feature was excluded due to its high correlation with other, more valuable and informative features.

5) Feature Selection

As a last preprocessing step, we dropped features with weak or no predictive signal. We assessed feature importance using multiple statistical measures: correlation, chi-squared, ANOVA and mutual information. For each target, we prepared both a strict and soft selection of candidate features. Both were tested in model evaluation, and the best-performing variant was retained.

6) Final Estimator

The final step in the pipeline was the machine learning model. Model choice was guided by task type (regression, binary, or multiclass classification) and evaluated through internal benchmarks using the complete preprocessing pipeline. Hyperparameter tuning was handled via RandomizedSearchCV.

Throughout modeling, we tested a broad selection of algorithms and defined extensive parameter grids to ensure robust optimization. This allowed us to identify the most performant models tailored to each specific prediction task.

6.2 Modeling Structure

Before moving on to the modeling steps, it's important to highlight a key fact of this project: the presence of **five distinct target variables**. This situation forced us to make an early architectural decision

- whether we would treat some of the targets as features for other models, or strictly use them as prediction endpoints. This section is dedicated to that reasoning.

Our primary focus was on predicting churn, and all decisions were made in relation to this core target. Starting with the obvious, the variable **churn_category** could not be used as a feature because it directly reveals the churn status: all non-churned clients are labeled as "Still a customer" in this column, which would introduce direct data leakage. Therefore, churn_category should only be predicted after churn is determined.

Next, we considered **cltv** and **churn_score**. Both are predictive constructs by nature and, importantly, not directly observable or collectible from the customer. Because of this, they cannot be used as input features, and must be treated as separate prediction targets.

The variable **satisfaction_score** initially appeared to be a potentially powerful feature, since it directly reflects customer satisfaction. However, our early EDA revealed an almost perfectly deterministic relationship between satisfaction_score and churn: scores of 1-2 led to 100% churn, and scores of 4-5 led to 100% retention. Including such a feature in a churn model would clearly result in data leakage. This was empirically confirmed during experiments: simply adding this feature caused model performance to exceed 95% F1 without any hyperparameter tuning - an unrealistic and artificially inflated result. Since we have no insight into how satisfaction_score was originally derived by Telco, we made the decision to treat it as a separate prediction target.

These conclusions directly shaped our modeling sequence. We first attempted to predict **satisfaction_score** and **churn_score** using only the observable, non-target features available for a customer. If successful, the idea was to feed those generated values as OOF features (out-of-fold predictions) into the churn model - thus simulating a real-world proxy of Telco's possible internal scoring systems. However, during modeling, it became clear that no approach yielded strong enough performance in predicting either satisfaction_score or churn_score. Although results surpassed baseline accuracy, they remained far from acceptable for practical use. We ran numerous experiments - tuning the aggressiveness of feature selection, changing hyperparameters, switching between regression and classification tasks, reducing target diversity, and more - but none produced reliable results.

At this point, we revisited the nature of our available features and realized a key limitation: the dataset consists almost entirely of objective, factual variables - such as location, tenure, service usage, and

demographics. These features lack the nuance needed to model subjective user sentiment or behavioral tendencies. Crucially, the dataset contains no data about support interactions, customer feedback, complaint history, or other experience-based signals. Without this contextual information, it's almost impossible to infer real customer satisfaction or internal churn risk. This insight suggests that improving satisfaction or churn score prediction would require expanding the dataset to include subjective, behavior-driven features. Such additions potentially could significantly boost model quality - and consequently improve churn prediction downstream.

Looking ahead, attempts to predict the other 2 targets - **churn_category** and **cltv** - also produced unsatisfactory results. Predicting churn reason (category) inherently depends on understanding why a customer left, which again ties back to satisfaction and perception - areas where we lack data. As for CLTV, predicting it proved to be significantly more complex due to the complete absence of temporal context in the dataset. By its very nature, CLTV is a long-horizon target - often spanning several months or even years - and making any reliable forecast without granular time-series history becomes irrelevant. We ultimately concluded that CLTV should be viewed not as a static label, but as a dynamic and continuously evolving metric. As such, any attempt to model it properly would require a fundamentally different data architecture, one built around temporal sequences, customer lifecycle events, and regular behavioral signals over time. Model testing for CLTV prediction consistently yielded R2 scores around 0.2, indicating that the models captured virtually no meaningful variance. This suggests that, given the available features, the task of predicting CLTV lacks value in its current form.

Given these findings, we were forced to pivot and develop the churn model without using any of the other targets as features. Fortunately, despite the dataset still having its limitations, we managed to achieve acceptable performance for this particular task. After benchmarking several binary classifiers, the best result came from an **XGBClassifier**, optimized via grid search and trained with a strict feature selection strategy. This model achieved an **F1 score** of approximately **0.65**, which, although below the typical business threshold of 0.70–0.75 for production-ready churn models, is still a reasonable outcome - especially considering that it more than doubled the performance of the baseline model.

	Model	F1 Score	Accuracy	Precision	Recall	ROC AUC	Log Loss
0	XGBClassifier	0.652	0.820	0.708	0.604	0.881	0.378
1	LinearSVC	0.631	0.810	0.688	0.582	NaN	NaN
2	LogisticRegression	0.629	0.811	0.696	0.573	0.871	0.391

Figure 8.1 Top-3 Models After Parameters Tunning

For a dataset based on real customer data, this is a solid result, likely enabled by the strong correlations we uncovered during earlier stages of the analysis. While this score still falls short of the ideal target, it provides a strong foundation. With the inclusion of additional data reflecting customer experience, we expect further performance improvements to be achievable.

8.3 Features Importance

The most important features for the model, as expected, were **contract types**. This aligns with the common understanding that long-term customers tend to be more stable, often held back from leaving due to financial penalties for early termination or simply a psychological reluctance to leave thing incomplete. Another significant factor was the **presence of an active prepaid period**, which logically reduces the likelihood of imminent churn. Although some churners with prepaid periods exist, they remain a minority and represent an anomaly rather than the norm.

Among other relatively important features were **payment method**, **tenure**, and the **number of referrals**, which also align with expectations. The payment method e-check was highlighted during the analysis as strongly associated with higher churn risk, a hypothesis the model confirmed. Tenure, previously unconsidered as a valuable feature, demonstrated that loyalty tends to increase over time if customers are not dissatisfied - though its influence remains moderate and not absolute. Lastly, the number of referrals reflects an indirect measure of customer satisfaction with the company's services, which naturally encourages retention.

	Feature	Importance
0	contract_two year	0.208867
1	has_active_prepaid_period	0.164184
2	contract_one year	0.134422
3	payment_method_e-check	0.087526
4	binned_tenure	0.077236
5	number_of_referrals	0.059120

Figure 8.2 Features Importance

7. Conclusions and Recommendations

After conducting an in-depth analysis of the Telco dataset and testing several predictive models, we can now summarize the key findings and practical takeaways.

1. Key Drivers of Customer Churn:

The most influential factors behind customer churn turned out to be, quite predictably, the type of contract. Short-term contracts, especially monthly ones, are strongly associated with higher churn rates. This is understandable: long-term customers are more stable, often due to financial penalties for early termination or simply due to habit and the psychological barrier of leaving something unfinished. Similarly, the presence of an active prepaid period also turned out to be important - again, not surprising, as customers with time left on their contract are less likely to leave immediately. Although exceptions do exist, they are relatively rare and represent an anti-pattern.

Other notable contributors included the method of payment, tenure, and the number of referrals. The fact that the payment method e-check is associated with higher churn had already been observed during the EDA phase, and the model's feature importance scores confirmed that assumption. Another meaningful factor was the length of the customer's relationship with the company: longer tenure is generally associated with increased loyalty, which makes sense until the moment the client becomes disappointed. Still, the correlation is not absolute, and the feature's importance remains relatively low. Finally, the number of referrals - while not the strongest variable - indirectly captures customer satisfaction. People who refer others are more likely to be happy with the service, which in turn decreases their likelihood of leaving.

2. Model Performance

The resulting model demonstrated reasonably good performance, with a weighted **F1-score of approximately 0.65**. While not exceptional, this score indicates fair generalization given the limitations of the dataset. The absence of direct feedback or sentiment-related variables made it difficult to capture the emotional or experiential side of the churn decision. In this regard, one of the main conclusions of the project is the need for better customer data collection - particularly behavioral and satisfaction indicators that could provide early signals of dissatisfaction or churn.

3. Strategic Recommendations

When it comes to recommendations, the primary focus should be on improving Telco's customer engagement system. As our analysis revealed, a significant portion of churn is driven by dissatisfaction with service quality or company attitude, which increases tendency to switch to competitors. These observations point toward a systemic issue - a lack of attention to customer needs. This hypothesis is supported by the fact that some vulnerable customers groups did not receive targeted offers, had no loyalty reinforcement, and likely experienced low service quality relative to high payments.

Another confirmation of that hypothesis is the structure of the dataset itself. Despite containing extensive customer metadata, it lacks essential inputs related to customer experience - such as satisfaction ratings collected via apps or calls, or cumulative data from support interactions. These types of behavioral feedback should be tracked continuously. Integrating them into the pipeline would significantly enhance predictive power of any model and allow the company to monitor dissatisfaction in real time. For example, a spike in support interactions could signal risk and trigger intervention. With such data, target prediction models would become not only more accurate but also more reasonable.

In addition to data improvements, Telco should expand its offering strategy. Customers in high-risk segments would benefit from more personalized offers, such as bundled services, loyalty bonuses, or contract flexibility. One particularly effective measure could be providing premium support access to these groups for no extra charge. This relatively simple feature could have an outsized impact on churn reduction as it has shown already how it can reduce churn rate by 50%.

In summary, Telco should focus on ***systematically tracking customer sentiment****,* ***implementing proactive measures for at-risk groups****,* ***delivering more attractive and responsive offers****, and* ***ensuring consistent service quality****. These changes would not only reduce churn but also strengthen long-term customer loyalty and allow Telco to have much more control of their position in highly competitive market.*