Telecoms Customers Churn Analysis

A Capstone Exploratory and Predictive Modeling Project

Zach Kontor
JJ Kailash
Albinson Felix
Sam Khanin

4 May 2025

Table of Content

Table of Figures/Tables

Executive Summary

In an increasingly competitive telecommunications landscape, customer retention is paramount to business survival. This project aims to combat customer churn by developing machine learning models that can identify at-risk customers to then trigger proactive intervention. Customer churn is a critical threat to profitability, sustainability, and growth.

We analyzed over seven thousand customer records to uncover behavioral and demographic patterns influencing churn. Key findings in our exploratory analysis include:

- Lower churn rates amongst customers using value-added services like online security, tech support, and those on longer-term contracts.
- Higher church rates amongst customers on short-term contracts, as well as those primarily using fiber optic internet.

We performed cluster analysis to understand customer segmentation. Which revealed 3 distinct groups, of which cluster 2 was identified as a high churn risk group. This cluster shows that customers who are highly tech-savvy and require high-speed internet typically have a lower commitment threshold.

Additionally, we trained and optimized multiple classification models, including Logistic Regression, Random Forest, XGBoost, Stacking Ensemble, and others. We evaluated each for precision, accuracy, recall and F1-Score. We eventually prioritized recall and precision-optimized models. The recall-optimized model will be used for broad churn detection, where resources are not as limited, focusing on identifying most of the possible churners. The precision-optimized model will be used for more resource-constrained marketing interventions where the business would rather prioritize only customers that will churn instead of spending valuable resources on non-churning clients.

Deployment strategies include:

- CRM Integration or Churn Shield, which utilizes the recall-optimized models to provide early churn risk alerts that customer service agents can utilize for support reasons.
- Dynamic Marketing Dashboards or Churn Intelligence Engine, this will enable the marketing team to tailor their retention campaigns in real time.

Our project highlights predictive accuracy and strategic flexibility to suit business operations and profitability while managing unnecessary expenditures.

Business Understanding

The telecommunications industry is one of the most competitive industries, and customers have numerous alternatives at their fingertips. Many competitors offer the same broad range of services, and as such, maintaining a solid customer base is one of the toughest challenges facing these companies. Customer churn is an alarming component that has direct implications for long-term growth, profitability, and innovation. Oftentimes, acquiring new customers is more expensive than retaining existing ones, making churn mitigation a top priority within this industry.

Objective

The primary objective is to reduce customer churn by accurately identifying customers who are likely to cancel their services. With this insight, the company is now able to launch targeted retention campaigns, such as discounts and bundling services. This means that, ultimately, we can increase the customer lifetime value instead of investing in acquiring customers (CLV vs CAC). Lastly, this will help boost profitability and market competitiveness, ultimately driving innovation.

Problem Statement

Churn rate is defined as the percentage of a customer base that unsubscribes/cancels a service, terminating operation with said service provider. Churn significantly reduces revenue and affects a company’s long-term outlook. This means proactively identifying customers likely to leave and intervening before they churn.

In the telecommunications industry, churn is influenced heavily by contract type, service type, family structure, overall tenure length, and other factors.

Our team will be tasked with developing a data-driven ML-based solution to predict churn and help the business personalize retention strategies at scale.

Project Objectives

- Analyze customer data to identify behavioral and demographic patterns associated with churn.
- Build a machine learning classification model to predict churn risk.
- Evaluate model performance using metrics like precision, recall, F1-score, and AUC-ROC.
- Determine the key drivers of churn to inform business strategy.
- Provide actionable recommendations to reduce churn based on model insights through deployment options.

Success Criteria

The best way to measure success in this project would be to effectively predict the customers most likely to churn and thus implement retention strategies to reduce churn. This will be done by modeling different algorithms and evaluating their accuracy and precision to predict churn based on current user data.

- The predictive model achieves high recall (to catch as many churners as possible) and balanced precision (to avoid too many false alarms).
- The model explains key churn drivers, enabling clear and interpretable business action.
- The solution is deployable and can be integrated into the company's customer relationship management (CRM) system for proactive outreach along with marketing dashboard updates.

Data Understanding

The telecoms customer churn dataset is sourced from Kaggle and represents an IBM sample customer dataset.

Dataset Structure and Dimensions

Shape:
The dataset contains 7,032 records with 21 columns. This gives us a good sample size to analyze customer churn behavior.

Data Types:
The dataset includes both numerical and categorical features

Numerical: *tenure, MonthlyCharges, TotalCharge*s

Categorical: *gender, Partner, Dependents, PhoneService, InternetService*, and several others representing customer demographics and service detail, all were encoded (0's and 1's) to represent these features.

Target Variable: The *churn* variable is the focus of the analysis:

With 5,110 customers stated as not churned and 1,922 customers stated as churned, there is an imbalance in the distribution of the target variable. It is seen that approximately 72% of customers did not churn, whereas 27% are classified as churned. This distribution indicates a class imbalance, which is common in churn prediction. Recognizing this imbalance early allows us to plan for appropriate techniques (such as SMOTE) to mitigate bias in model training, which will be explored and discussed in later sections.

Distribution plots of Tenure and Monthly Charges

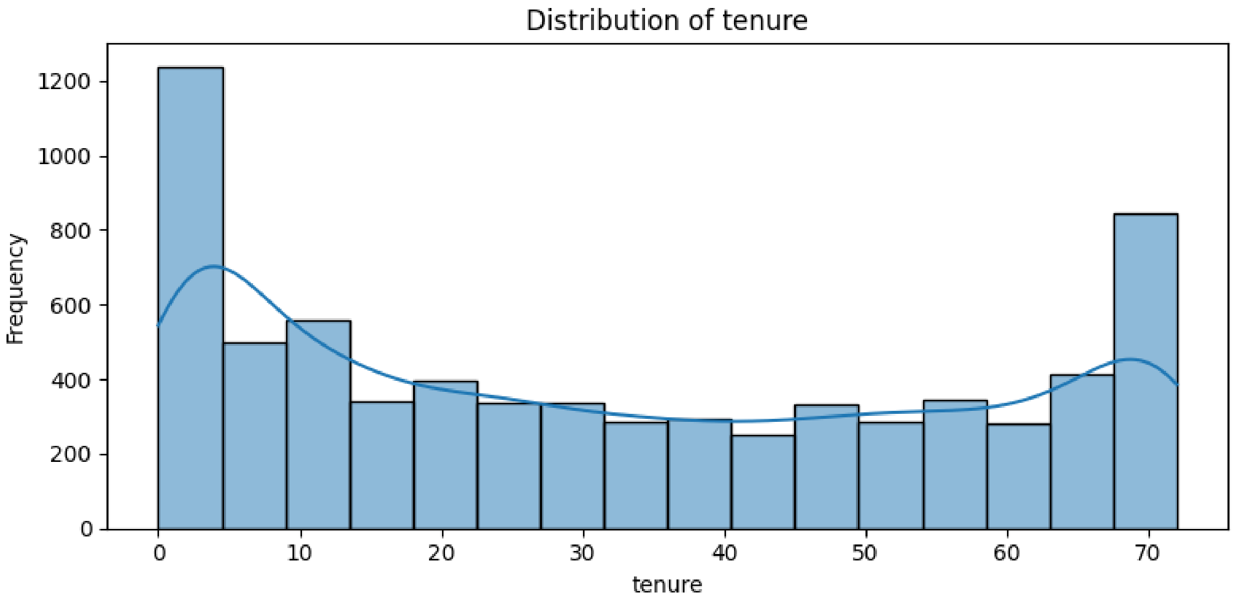

Figure 1: Distribution plot of tenure length of customers

The *tenure* variable exhibits a bimodal and right-skewed distribution. A large number of customers have very short tenures (close to 0 months), indicating a substantial group of new or recently churned users. There is also a notable peak at the upper end (around 70 months), which suggests a second group of long-term, loyal customers. The mid-range is relatively flat, implying that fewer customers stay for intermediate durations. This pattern may reflect differences in customer satisfaction, service type, or contract length.

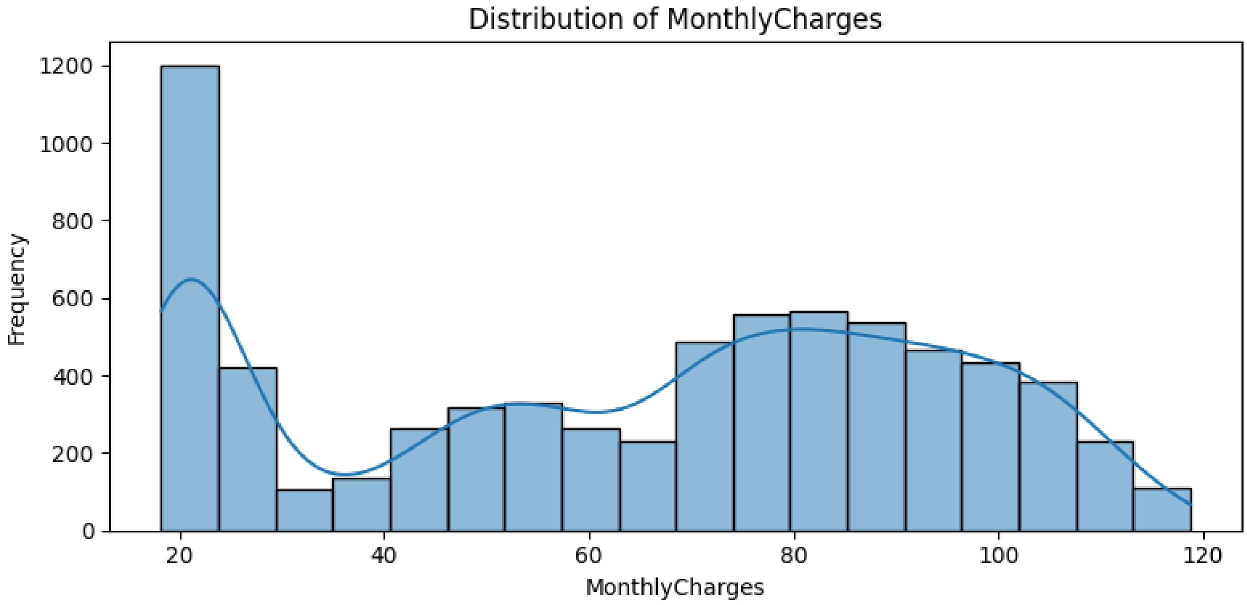

Figure 2: Distribution plot of monthly charges of customers

The *MonthlyCharges* distribution is right-skewed with a heavy concentration of customers paying around $20–$30 per month. However, the distribution also shows multiple small peaks throughout, suggesting a diverse pricing structure with no single dominant charge. Higher charges (above $70) are less frequent but still common, potentially representing customers subscribed to premium services or multiple add-ons. The variability in monthly charges suggests segmentation based on service packages or bundled offerings.

Exploratory Analysis of Factors Influencing Churn

Before developing any machine learning models, we conducted an exploratory data analysis to better understand which customer attributes and service-related factors are associated with higher or lower churn rates. The two dashboards below present key findings from this analysis, offering a visual breakdown of churn behavior across different customer segments and service options.

Service Features and Their Impact on Churn

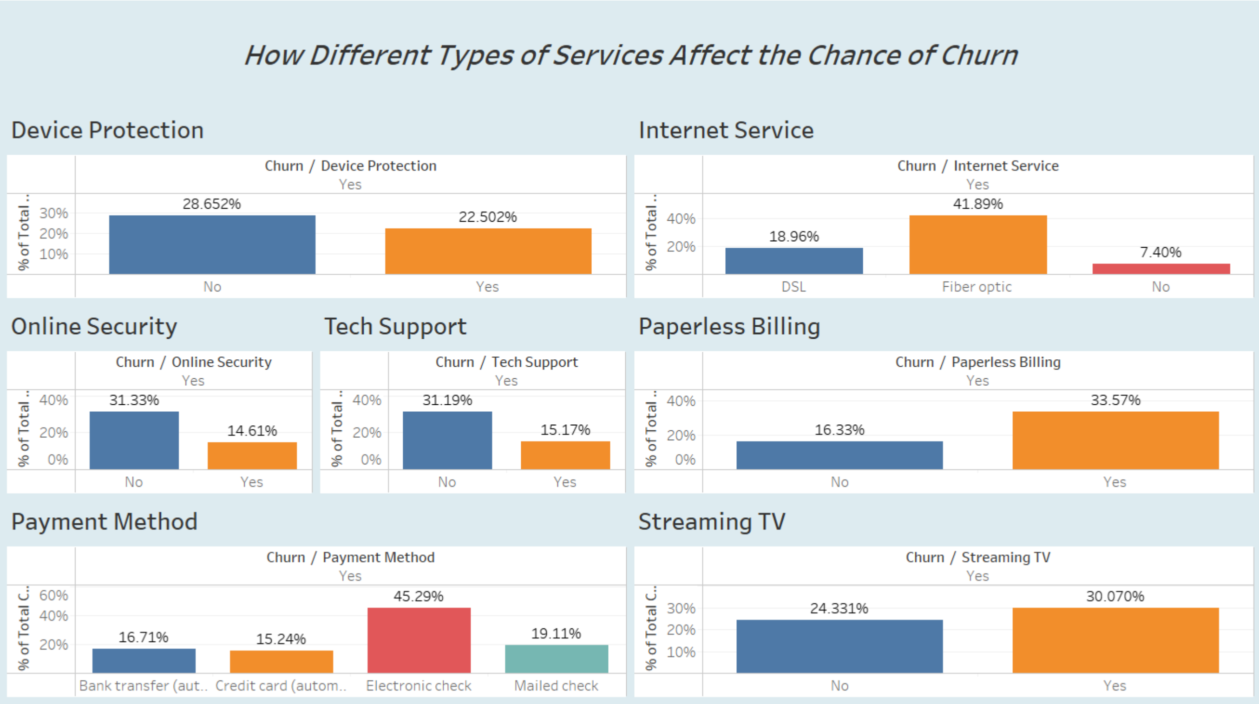

Figure 3: Services and their impact on churn

The first dashboard highlights how different services correlate with customer churn. Several patterns emerge:

- Value-Added Services Reduce Churn: Customers who subscribe to Device Protection, Online Security, and Technical Support are significantly less likely to churn. This suggests that customers who invest in protective or support-related add-ons may feel more secure and satisfied with the services, contributing to stronger loyalty.
- Fiber Optic Internet and Churn: Interestingly, fiber optic users are the most likely to churn among internet service types (41.89%), followed by DSL users. One hypothesis is that tech-savvy customers who opt for high-performance internet may also have higher expectations and are more inclined to switch providers when dissatisfied.
- Paperless Billing and Streaming Services: Customers enrolled in paperless billing (33.57%) and streaming TV (30.07%) show higher churn rates than those who are not. This could again reflect a segment of digital-native or convenience-seeking users who may be more critical of service shortcomings and more comfortable with switching.
- Payment Method Matters: Among the different payment methods, electronic check users exhibit the highest churn rate (45.29%), while customers paying via bank transfer or credit card churn

significantly less. The elevated churn rate among electronic check users might signal less engagement or satisfaction with automated billing or possibly a demographic trend worth deeper analysis.

Demographic and Customer Profile Characteristics

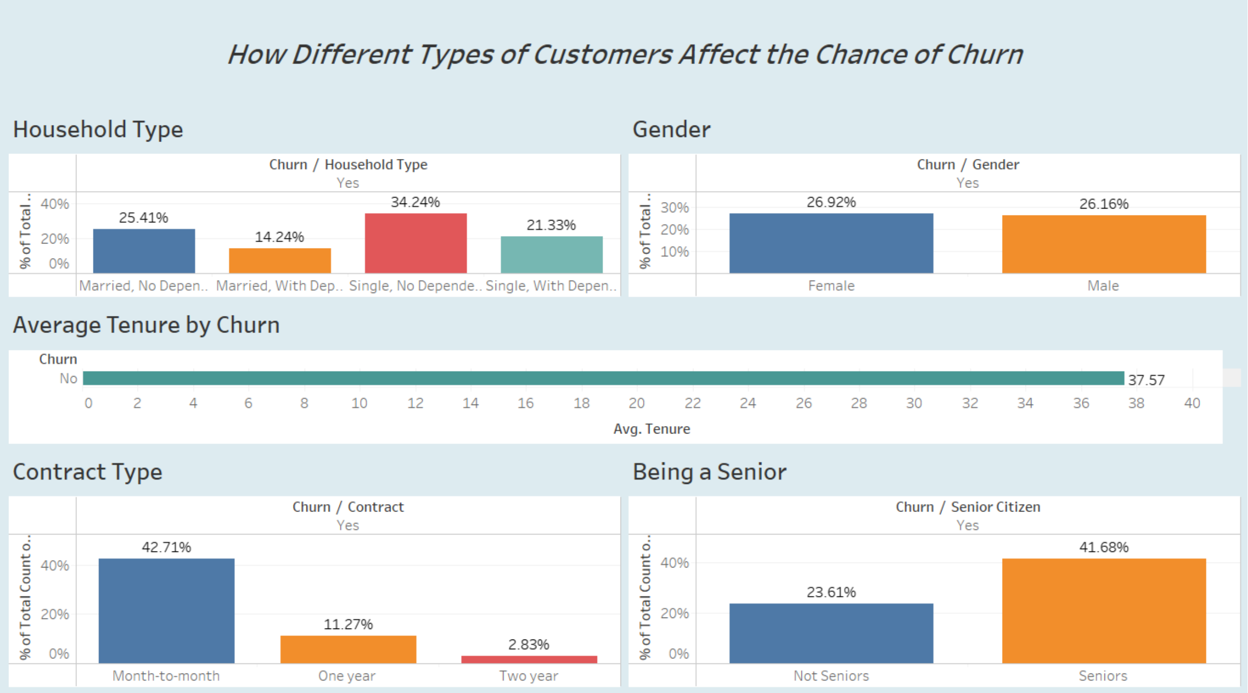

Figure 4: Customer segmentation and churn impact

The second dashboard provides insight into how customer demographics and profiles relate to churn tendencies:

- Household Type: The highest churn rate is observed among single customers with no dependents (34.24%), possibly reflecting a segment with greater mobility and fewer obligations tying them to one provider. In contrast, married customers with dependents show the lowest churn (14.24%), likely due to higher stability and potentially bundled household needs.
- Contract Type: Churn is dramatically higher among month-to-month contract users (42.71%) compared to those on one-year or two-year contracts (11.27% and 2.83%, respectively). This indicates that long-term contractual commitments serve as a strong deterrent against churn, likely due to early termination fees or the hassle of switching.

- Average Tenure: Customers who have not churned tend to have much longer tenures (average of 37.57 months) compared to those who churned (17.98 months). This supports the idea that customer loyalty builds over time and that early churn is a key area to address.
- Age Segment - Seniors vs. Non-Seniors: Senior citizens exhibit a significantly higher churn rate (41.68%) compared to non-seniors (23.61%). This trend may point to unmet needs or usability challenges faced by older customers, such as difficulty navigating digital interfaces, a lack of tailored support, or reduced perceived value. However, this finding also invites scrutiny of the underlying data collection process. It is important to consider whether some instances labeled as "churn" could be due to mortality among senior customers rather than voluntary service cancellation. If this is the case, it may introduce a bias in the churn classification for this age group and suggests that further investigation or data validation is warranted before drawing definitive conclusions.
- Gender: There appears to be no significant difference in churn behavior between genders, as both male and female customers churn at roughly the same rate.

Customer Segmentation via Cluster Analysis

To further enhance our understanding of customer churn behavior, we conducted a cluster analysis to segment customers into distinct groups based on shared characteristics. This unsupervised learning approach complements our predictive models by identifying underlying customer profiles that may not be apparent through individual features alone.

Clustering Methodology

We performed the following steps to develop meaningful customer segments:

- Feature Scaling: Standardized key numerical features—*tenure, MonthlyCharges,* and *TotalCharges,* and—to ensure equal contribution during distance-based clustering.
- Determining Optimal k: Using the Elbow Method, we evaluated the within-cluster sum of squares across values of k ranging from 2 to 10. Based on the inflection point where additional clusters yielded diminishing returns, we identified k = 3 as the optimal number of clusters.
- K-Means Clustering: Implemented the K-Means algorithm with k = 3, excluding the *Churn* variable to maintain an unsupervised approach.

- Dimensionality Reduction and Interpretation: We applied Principal Component Analysis (PCA) to project the data into two dimensions for better cluster visualization:
  - PCA1 – "Higher_Spending": This axis is most influenced by *MonthlyCharges* (0.48), and *TotalCharges* (0.47), with some influence from tenure (0.32) and *StreamingMovies*. It reflects the magnitude of financial engagement with the service.
  - PCA2 – "LongerTenure": Dominated by tenure (0.61) and supported by contributions from *MonthlyCharges*, *TotalCharges*, and long-term contracts. This component captures a customer's loyalty and time with the company.

Cluster Visualization

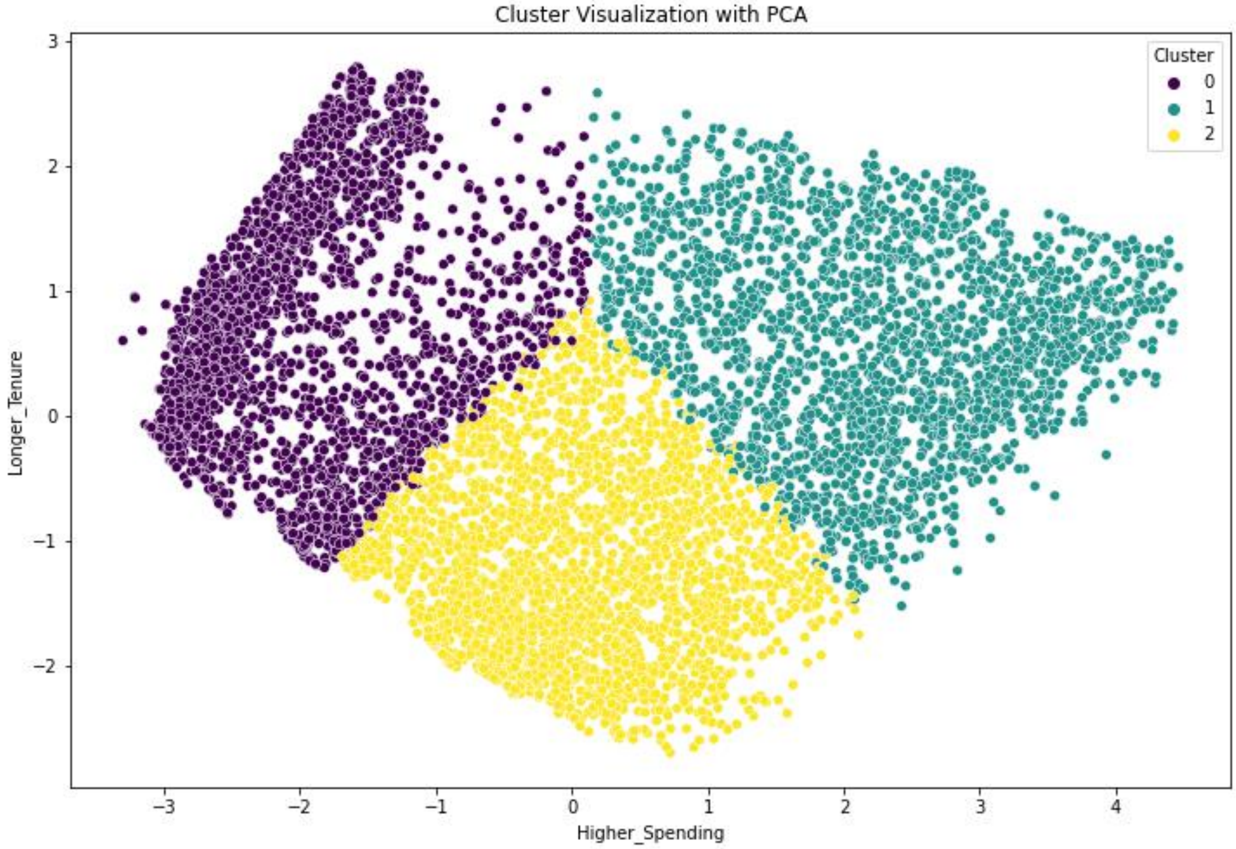

Figure 5: Cluster Analysis

As we can see from the plot, Cluster 0 is spending less and has higher tenure, Cluster 2 has average spending but lower tenures, and Cluster 1 has relatively high spending and tenure. Let's dive deeper into the cluster profiles and insights.

Cluster Profiles and Insights

Cluster 2 – High Churn Risk (48.19%)

Key Characteristics:

- Tenure (mean = 15.3 months): Customers are new, indicating an average tenure well below the overall mean.
- Contract Types: (One year: 10%, Two year: 2%). The majority are on month-to-month contracts (~88%).
- Fiber Optic Internet: 70% of customers use fiber, the highest of all clusters.
- Protection Services Adoption: (Online Security: 22%, Online Backup: 30%, Tech Support: 20%)
- Paperless Billing: Adopted by 70%, suggesting a tech-savvy, self-service-oriented base.
- Household Type: 64% are single (no partner), indicating low household stability.

Profile & Insight:

This segment comprises newly acquired, single subscribers, heavily concentrated in high-speed (fiber optic) plans without corresponding investments in security or technical support services. They favor paperless billing, but lack commitment (nearly all are on flexible contracts) and are light users of value-added services.

Their high monthly costs without perceived added value increase vulnerability to churn, especially in the early customer lifecycle. The lack of "stickiness" is reflected in minimal service protection, with churn risk exacerbated by personal and contract instability.

Cluster 0 – Moderate Churn Risk (15.48%)

Key Characteristics:

- Tenure (mean = 26.6 months): Customers are moderately new.
- MonthlyCharges (mean = $29.5): This is the lowest pricing cluster.
- Streaming Services: Usage of TV and Movie streaming services is negligible (~0%).
- Contract Protection: (Two-year contracts: 29%, One-year contracts: 17%, ~54% still on month-to-month plans.)
- Payment Method: 44% use mailed checks — a traditional and possibly more deliberate billing approach.
- Internet Service: 0% fiber adoption, primarily DSL or no internet service.

Profile & Insight:

This cluster represents budget-conscious customers who use basic service tiers and avoid premium offerings. Despite relatively short tenures, nearly half of them opt for structured contracts, suggesting some level of planning or price-lock incentives. Their traditional payment behavior (mailed checks) and avoidance of fiber plans show a preference for low-cost, low-complexity setups.

While not deeply loyal yet, they aren't as volatile as Cluster 2 — their financial expectations are being met, making them a manageable churn risk with consistent value delivery.

Cluster 1 – Moderate/Low Churn Risk (14.72%)

Key Characteristics:

- Tenure (mean = 58.52 months): The longest-standing customers across all clusters.
- Partner Rate: 71% have a partner, indicating household stability.
- MonthlyCharges (mean = $90.58): These customers are in the highest billing tier.
- Service Usage: (Online Security: 55%, Online Backup: 69%, Device Protection: 70%, Tech Support: 58%
- Streaming Services: (Streaming TV: 73%, Streaming Movies: 73%)
- Contract Type: Two-year plans: 44%, the highest among clusters.
- Fiber Optic Internet: 60% adoption

Profile & Insight:

This is the company’s core loyalty segment. Customers in this group are deeply integrated into the telecom ecosystem: they use nearly every value-added service, commit to long-term contracts, and have the highest engagement in high-speed internet and entertainment services.

Their longer tenure and relationship stability (partnered status) suggest high satisfaction and low likelihood of churn. This cluster represents "gold-tier" customers who derive tangible value from comprehensive service bundles.

Following the cluster analysis, we proceeded with preparing the dataset for machine learning model development, ensuring it was clean, well-structured, and optimized for predictive performance.

Data Preparation

The Telco Customer Churn dataset was carefully prepared to ensure data quality and improve model performance. The preparation process included cleaning, encoding, feature transformation, feature engineering, and data splitting.

Handling Missing Values

Overall, the dataset did not contain many missing values. The only notable issue was with the *TotalCharges* column, which included whitespace entries. Upon inspection, it was found that these entries corresponded to customers with a *tenure* of zero, indicating they were new customers who likely had not yet been billed. These missing values were, therefore, imputed with 0.

Outliers Analysis

A systematic outlier detection approach was implemented for all numeric variables using z-score standardization. This statistical technique quantifies the deviation of each observation from the population mean in standard deviation units, providing a normalized measure of extremity across different scales and distributions.

The conventional threshold of $|z| > 3$ was applied, identifying values that fall beyond three standard deviations from the mean, which statistically represents the 99.7% confidence interval under normal distribution assumptions.

As a result, no customer records were flagged as statistical outliers when the standard z-score threshold of 3 was applied.

Feature Cleaning and Transformation

- Service-related columns containing values such as *"No internet service"* and *"No phone service"* were standardized to *"No"* to simplify categorical levels.
- Binary categorical variables (e.g., *Partner*, *Dependents*, *PhoneService*, *Churn*) were encoded as 0 and 1.
- Multi-category categorical variables, including *InternetService*, *Contract*, and *PaymentMethod*, were one-hot encoded, with one category dropped from each to avoid multicollinearity.

- Boolean columns created during encoding were explicitly converted to integers to ensure consistency.

Feature Engineering

Several new features were introduced to enrich the dataset and capture additional patterns:

- *Household Type*: A new categorical variable created by combining *Partner* and *Dependents*, classifying customers as single or married, with or without dependents.
- *Charges Ratio*: Derived by dividing *TotalCharges* by (*MonthlyCharges* + 1), this feature provides insight into a customer's cumulative payments relative to their monthly bill.
- *Monthly Charges Squared*: A polynomial feature introduced to help capture potential nonlinear effects of *MonthlyCharges*.
- *Tenure Group*: The *tenure* variable was bucketed into ordered categories (e.g., *0–12 months*, *13–24 months*, etc.) to represent different stages in the customer lifecycle.

Preparation

Features were then categorized as numerical or categorical, with separate preprocessing strategies applied:

- Numerical features were standardized using z-score scaling to ensure uniformity.
- Categorical features were transformed using one-hot encoding to convert them into a machine-readable format.

Modeling

Following thorough data preparation, we proceeded to the Modeling phase of the framework, aiming to build robust, interpretable, and generalizable models for predicting customer churn.

Data Splitting

To ensure objective evaluation and model generalization, the dataset was partitioned into three subsets using stratified sampling to maintain the original churn distribution:

- Training set: 75% of the data, used for learning and model fitting
- Validation set: 15%, used for hyperparameter tuning and model selection
- Test set: 10%, held out for final evaluation

This approach mirrors real-world deployment, where unseen data must be accurately predicted.

Class Imbalance Handling

Initial analysis of the *Churn* variable revealed a significant class imbalance: approximately 73% of customers had not churned, while only 27% had. To address this and improve the model's ability to detect minority-class instances, SMOTE (Synthetic Minority Over-sampling Technique) was applied to the training data. SMOTE generates synthetic samples of the minority class to create a more balanced training set and mitigate bias in model learning.

Model Selection Rationale

We selected a diverse mix of models to strike a balance between predictive power, robustness, and interpretability:

- Logistic Regression:
  Chosen for its simplicity and transparency, logistic regression provides clear, interpretable coefficients that help stakeholders understand which factors contribute most to churn. It serves as a baseline model, and its performance is a useful reference point for evaluating more complex approaches.
- Random Forest:
  As a nonparametric ensemble of decision trees, Random Forest is well-suited for modeling nonlinear relationships and handling mixed data types (numerical and categorical). Also robust to noise and outliers, and provides feature importance rankings, which support interpretability. This model is effective when patterns in the data are too complex for linear models to capture.
- XGBoost:
  Known for its state-of-the-art performance on structured/tabular data, XGBoost leverages gradient boosting to build strong learners by sequentially improving weak ones. It handles missing values natively, captures complex interactions, and offers built-in regularization, making it ideal for this churn prediction problem, where subtle patterns may indicate churn risk.
- SGD Classifier:
  The Stochastic Gradient Descent (SGD) Classifier is a scalable linear model trained via gradient descent. It was chosen for its computational efficiency and ability to accommodate large-scale learning, which is useful in production settings or with streaming data. Its flexibility in using different loss functions (e.g., log-loss or hinge) also allows experimentation with linear classification boundaries under different assumptions.

Ensemble Approach: Stacking

To further enhance predictive power, we built a stacking ensemble by:

- Selecting the top three base models based on validation performance: Logistic Regression, Random Forest, XGBoost.
- Using Logistic Regression as a meta-classifier to combine base model outputs

Model Tuning

Each model underwent extensive hyperparameter tuning using GridSearchCV with 5-fold cross-validation to ensure optimal configuration. Our hyperparameter grids were designed to explore:

- Regularization strengths and penalty terms (L1, L2, elastic net)
- Tree depth, number of estimators, and minimum samples per leaf (for ensemble models)
- Learning rates and boosting strategies (XGBoost)
- Class weighting strategies to mitigate the impact of churn class imbalance

Validation Strategy

Model performance was monitored across three stages:

1. Cross-validation scores guided hyperparameter tuning
2. Validation set performance determined model selection
3. Test set evaluation provided final unbiased performance estimates

This three-tiered validation ensures robustness and minimizes overfitting risk.

Model Evaluation Metrics

Our model evaluation strategy employs a multi-metric approach to assess model performance through complementary perspectives comprehensively. Each metric addresses specific business implications in the churn prediction context:

Primary Classification Metrics

- Precision (True Positives ÷ (True Positives + False Positives)) represents prediction efficiency by minimizing false alarms. High precision ensures retention resources target genuine at-risk customers, optimizing intervention budget allocation and preventing wasteful spending on customers who weren't planning to leave.

- Recall (True Positives ÷ (True Positives + False Negatives)) captures detection coverage by minimizing missed churners. This metric is particularly critical as unidentified at-risk customers represent significant revenue loss potential with no opportunity for intervention. Every missed churner translates directly to lost business value.
- F1 Score (2 × (Precision × Recall) ÷ (Precision + Recall)) balances the precision-recall tradeoff in a single metric, particularly valuable when seeking equilibrium between resource efficiency and churn capture rate. This harmonized measure helps avoid overemphasizing either precision or recall when both considerations matter.

Supplementary Evaluation Metrics

- ROC-AUC quantifies model discrimination capability across all possible classification thresholds. This provides a threshold-independent performance assessment, enabling strategic decision-making around operating points based on business priorities. It measures the model's fundamental ability to separate churners from non-churners.
- Accuracy offers a general performance benchmark as the overall correctness ratio (correctly classified instances ÷ total instances). However, we interpret this metric cautiously, given class imbalance considerations in churn prediction, where the majority class can skew results.

Model Evaluation

Model Performance Analysis

Our evaluation strategy focused on analyzing model behavior under two scenarios—recall-optimized and precision-optimized—to reflect different business priorities in churn intervention. The aim was to identify models best suited for deployment based on specific operational needs: maximizing churn detection (recall) or minimizing false alarms (precision).

This dual-scenario evaluation highlights the trade-offs between recall and precision and supports a tailored approach to churn management. By identifying which models excel under each objective, we enable the business to respond more intelligently, whether that involves broadly identifying a larger pool of potentially at-risk customers for early intervention or focusing resources on a more targeted group with a higher likelihood of churn.

This strategy ensures that model selection is not only data-driven but also aligned with real-world constraints such as budget limitations, team capacity, and the cost of customer retention efforts.

Recall-Optimized Scenario: Prioritizing Churn Capture

Table 1: Results of Recall optimized models

Model	**Test Accuracy**	**Test F1**	**Test Precision**	**Test Recall**	**Test ROC AUC**
XGBoost	0.489	0.510	0.342	1.000	0.842
SGD	0.496	0.510	0.344	0.989	0.832
Logistic Regression	0.562	0.541	0.374	0.973	0.847
Random Forest	0.784	0.614	0.585	0.647	0.840
Stacking Ensemble	0.794	0.551	0.654	0.476	0.849

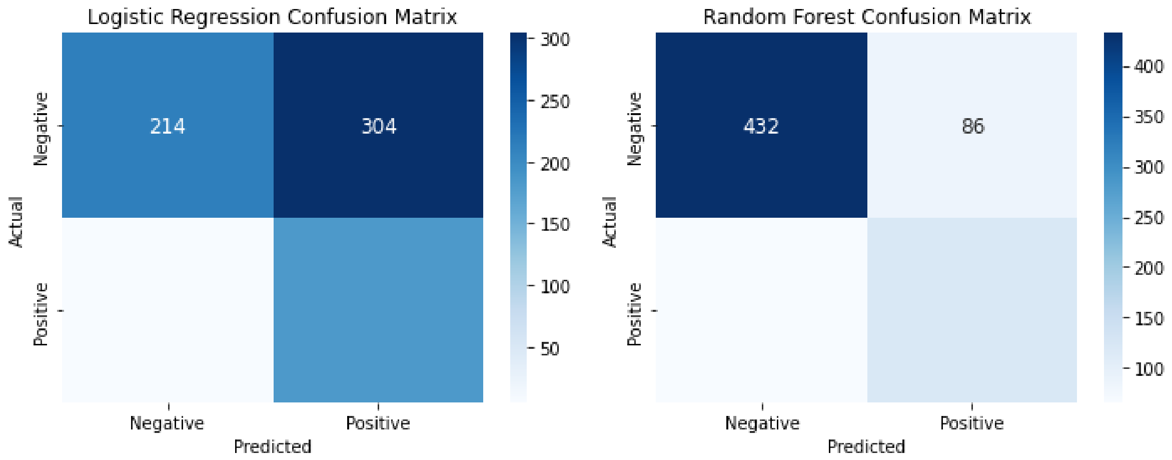

Figure 6: Logistic Regression and Random Forest Confusion Matrices

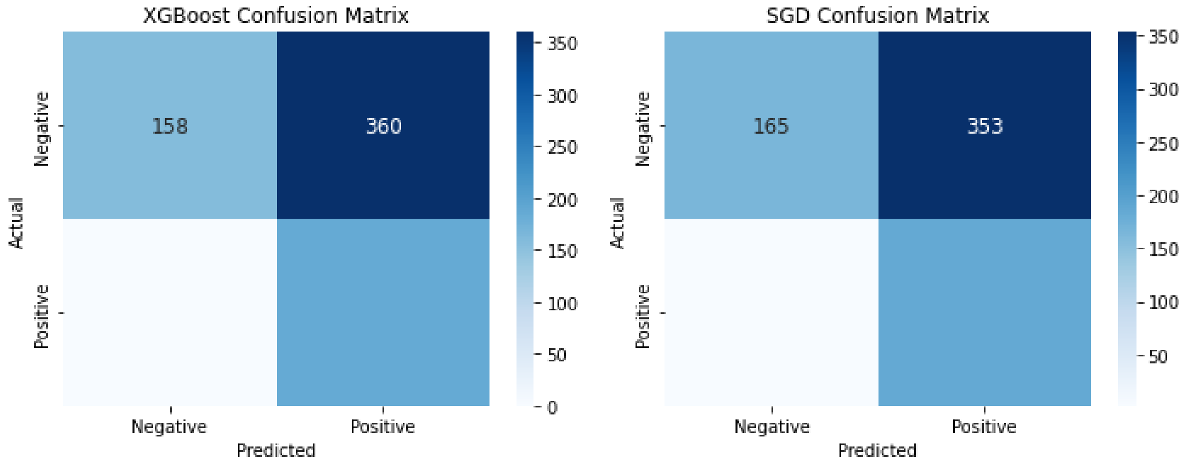

Figure 7: XGBoost and SGD confusion Matrices

When recall was the primary objective—ensuring we captured as many churners as possible—the Logistic Regression model emerged as the most reliable candidate for deployment. While XGBoost achieved perfect recall (1.0), it did so at a high cost in precision (0.342), resulting in excessive false positives. Logistic Regression, by contrast, maintained a recall of 0.973 with improved precision (0.374), producing the highest F1 score among the models with high recall.

This trade-off is significant: Logistic Regression enables us to intervene on nearly all potential churners while still limiting wasted resources on false positives. Additionally, its interpretability makes it easy to communicate insights to stakeholders and implement targeted retention strategies.

Precision-Optimized Scenario: Minimizing False Alarms

Table 2: Results of Precision optimized models

Model	Test Accuracy	Test F1	Test Precision	Test Recall	Test ROC AUC
Stacking Ensemble	0.809	0.582	0.691	0.503	0.854
XGBoost	0.791	0.593	0.615	0.572	0.846
Random Forest	0.790	0.593	0.610	0.578	0.833
SGD	0.765	0.636	0.539	0.775	0.848
Logistic Regression	0.760	0.629	0.534	0.765	0.852

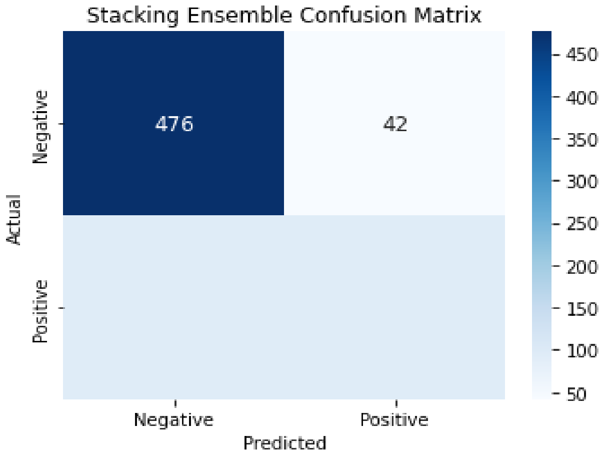

Figure 8: Stacking Ensemble Confusion Matrix

In the precision-focused setting, where the goal is to conserve retention resources and avoid acting on false alarms, the Stacking Ensemble model delivered the best results. It achieved the highest precision (0.691) and overall accuracy (0.809), meaning fewer false positives while still identifying over half of the churners (recall = 0.503).

This makes the Stacked Ensemble an ideal choice for deployments where false positives incur real business costs, such as retention discounts or personalized outreach efforts. Compared to single models like Random Forest or XGBoost, the ensemble benefits from leveraging multiple perspectives and achieving robust predictions through meta-learning.

Feature Importance Analysis

The feature importance analysis reveals consistent patterns across both modeling scenarios:

Key Churn Predictors from Logistic Regression

1. Contract Type: Two-year contracts show strong negative coefficients (reducing churn likelihood), with odds ratios of 0.52 and 0.6, indicating customers on longer contracts are approximately half as likely to churn.
2. Tenure Factors: Both raw tenure and the TenureGroup_0-12 feature appear consistently important. The positive coefficient for TenureGroup_0-12 confirms that newer customers are more prone to churn.
3. Internet Service Features: Fiber optic service shows a positive association with churn (odds ratio >1), while no internet service correlates with reduced churn (odds ratio <1).
4. Payment Method: The Electronic check payment method positively correlates with churn across both models.

Insights from Random Forest Importance

The Random Forest importance scores provide complementary insights:

1. Financial Metrics: The engineered Charges_Ratio feature ranks consistently high, revealing that the relationship between total and monthly charges significantly influences churn behavior.
2. Monthly Charges: Both direct (MonthlyCharges) and transformed (MonthlyCharges_Squared) features rank in the top features, suggesting complex relationships between pricing and customer retention.

Model Selection Considerations

The modeling evaluation demonstrates that no single model is optimal across all metrics, making it critical to match the model to the business objective. Logistic Regression offers an excellent recall-optimized option with interpretability, while the Stacking Ensemble excels at precision and minimizing false positives. Together, these models form a flexible and effective toolkit for deployment in customer churn prediction.

Business Interpretation

From the models, we were able to get clear and actionable business insights. Primarily, we observed that the high recall model would be most effective for early churn prevention, as it produced more false positives, meaning resources could be spent on non-churning customers who would have stayed nonetheless. Secondly, the high precision model was more ideal for cost-effective and direct targeting of customers who will churn. It should be noted that we considered removing obvious non-churners from our models, like those with long tenures or long-term contracts, but we opted not to do this because of the size of the dataset.

As a business, we identified high-impact predictors such as tenure, value-added services and contract lengths. This allows for better business strategic actions going forward, such as promoting long-term contracts and bundling services. Ultimately, these insights will be turned into deployment methods to optimize business operations, including CRM integration, marketing dashboards, and customer segmentation.

Error Analysis

Despite strong results, we observed several challenges and limitations for the business operations.

- High Number of False Positives and Negatives: Since we optimized for both recall and precision, our models produced a high number of false values, which ultimately led to wasted resources or retention efforts. These could give unnecessary discounts to already loyal customers. While we do believe that the cost evaluation of retaining customers outweighs the cost of the efforts, it should still be noted.
- Feature Bias: Features such as payment method, billing preferences, and senior citizens could create ethical concerns. As these features could speak more to socioeconomic status and mortality thus we should evaluate and caution whether we do want to target these customers in retention efforts.

- Real-Time Limitation: In a real-world environment, it might be more effective to score (regression) customer churn potential in real time, as their churn risk is always changing. This could be solved with the CRM integration, where customer data is always backed up and updated, specifically features such as tenure as well as contract length conversion as high-impact features need to always be checked and updated. This is also the case with evaluating the cluster analysis of customer data.

Deployment

Recognizing that a singular intervention strategy is insufficient for addressing customer churn, we propose a dual machine learning model deployment strategy. This approach offers optimal flexibility in applying various model applications:

- Recall-Optimized Model (e.g., XGBoost or Logistic Regression) for High-Stakes Churn Prevention – Shield feature in a given CRM

  In scenarios where missing a churner is highly detrimental (e.g., high-LTV customers, contract renegotiations, or executive retention programs), we prioritize recall to ensure that nearly all potential churners are flagged. False positives are tolerable here, as the cost of a missed opportunity outweighs the intervention cost.

- Precision-Optimized Model (e.g., Stacking Ensemble) for Resource-Constrained or Campaign-Oriented Interventions – Business Resource Allocation

  For use cases like targeted marketing promotions, where the goal is to efficiently convert at-risk customers with limited incentives, high precision ensures we reach the most likely churners without overspending on customers unlikely to leave. This model minimizes unnecessary outreach, improving campaign ROI.

Summary of Optimal Model Usage

- Use the Recall-Optimized Logistic Regression model when maximizing customer retention coverage is critical and the cost of unnecessary outreach is acceptable.
- Deploy the Precision-Optimized Stacking Ensemble for targeted, budget-sensitive initiatives where efficiency and ROI of interventions are paramount.

This flexible strategy ensures that we not only predict churn effectively but also act on these predictions through tangible business integrations to address customer churn.

Tangible Business Integrations
CRM integration

- When dealing with customer requests in a given CRM environment, there is a churn shield feature added to each customer. Aiding the telecoms representative in prioritizing certain customer requests and claims. (i.e., a customer with high churn probability gets service priority or tailored strategies based on their vulnerabilities)

Business resource allocation

- Marketing dashboards and customer analysis are based on churn propensity. Marketing strategies and business promotions are tailored to aid in customer retention.

Listed are the frameworks of potential deployments that align with the above integrations.

Deployment 1 *- Customer Churn Shield CRM Feature*

Solution

Build a Churn Shield Feature for a business's CRM, identifying at-risk customers in their first 6–12 months. Using our recall-tuned model to flag high-risk customers, and would automatically recommend the business to:

- Nudge customers to adopt Online Security, Tech Support, Streaming and other services that keep customers longer, avoiding churn.
- Onboarding bundles with loyalty points, free trials, and guided walkthroughs
- Tiered milestone rewards: retain customers past 3/6/12-month checkpoints

Execution

- Integrate a churn flag into the CRM
- Automated outreach flows based on the flag
- Measure churn drop post-intervention across treatment vs. control

Monitoring

Weekly Review: Cross-functional “Churn Shield Stand-up” reviews dashboard, experiments, and open incidents to assign actions.

***Deployment 2** - Churn Intelligence Engine powering marketing dashboards*

Solution

Marketing dashboards and customer analysis are based on churn propensity and our precision-optimized stacking model that predicts churn with high confidence. Thus, marketing strategies and business promotions are tailored to aid in customer retention.

Execution

Score every customer weekly/monthly and rank by churn-risk percentile, and feed the top 20% at-risk customers into:

- Customer-service prioritization.
- Marketing automation for dynamic discounts and proactive check-ins.
- CRM retention workflow.
- Assign risk tier (High / Medium / Low) and write back to the warehouse and dashboards.
- Dashboards monitor conversion, offer-level ROI

Monitoring

Feedback loops - model retrained quarterly using updated metrics based on previous performance to increase market capture and customer retention.

Project Summary

This project combined customer segmentation and predictive modeling to uncover actionable insights for churn reduction. Using clustering analysis, we identified three distinct customer groups with varying levels of churn risk—48.19%, 15.48%, and 14.72%, respectively. The highest-risk segment is characterized by new, single subscribers with fiber internet but low adoption of support services.

To address different business objectives, we implemented a dual-model strategy:

- Logistic Regression, optimized for recall, achieved 97.3%, enabling broad identification of at-risk customers.

- Stacking Ensemble, optimized for precision, reached 69.1%, allowing for more efficient, resource-conscious interventions.

Across models, the most influential predictors of churn included:

- Short contract durations
- Early customer tenure (0–12 months)
- Fiber internet without bundled support features
- Electronic payment methods (especially electronic checks)
- Discrepancies between total and monthly charges (Charges Ratio)

Despite limitations such as static data and limited behavioral signals, our analysis provides a framework and deployment strategies for reducing churn through data-driven segmentation and tailored interventions.

Recommendations

Strategic Recommendations based on the Cluster Analysis

1. Prioritize Retention for Cluster 2 (High Churn Risk – 48.19%)

- Bundle high-churn services (fiber optic internet) with underutilized offerings such as Online Security, Backup, and Tech Support—currently used by only 20–30% of this segment.
- Promote contract conversion incentives to reduce month-to-month reliance (~88% currently on flexible contracts) and boost commitment levels.
- Deploy educational campaigns focused on demonstrating the long-term value of service add-ons to increase perceived benefit among new users with minimal tenure.
- Offer onboarding promotions or guided setup for single, digital-native users (64% single, 70% using paperless billing), who may churn early without intervention.

2. Transition Early-Tenure Users Toward Cluster 1 Behavior

- Launch engagement programs designed to:
    - Introduce streaming and support services early
    - Encourage bundling to mimic the adoption behavior of Cluster 1 (where over 55% use each protection feature and 73% stream content)

- Implement milestone-based incentives to retain users through critical early-tenure thresholds (e.g., 6 or 12 months) when churn risk is highest.

3. Strengthen Cluster 0 (Moderate Churn Risk – 15.48%) Retention with Low-Cost Enhancements

- For value-driven customers paying the lowest monthly rates, consider small-cost service upgrades or rewards programs for long-term contracts.
- Use their traditional payment preferences (44% mailed checks) to offer analog-friendly communications and engagement methods.

4. Reinforce Loyalty in Cluster 1 (Low Churn Risk – 14.72%)

- Offer exclusive loyalty bundles or tiered rewards to retain this high-value segment.
- Maintain satisfaction by proactively ensuring continued service quality and minimal disruption for fiber and streaming offerings (both at 60–70% adoption).

5. Integrate Cluster Segmentation into CRM Strategy

- Map churn intervention tactics to cluster-specific pain points and behavioral profiles.
- Use cluster assignments to prioritize customer support ticket routing, marketing personalization, and outreach frequency based on churn probability and revenue potential.

Recommendations based on Modeling

1. Contract Length

- Contract duration is a dominant churn predictor. Offer graduated incentives (e.g., service credits, premium features) for moving from flexible to longer-term contracts.

2. Payment Method Optimization

- Customers using electronic checks exhibit higher churn. Streamline this payment experience and incentivize auto-pay enrollment to reduce friction.

3. Charges Ratio Insights

- High Charges_Ratio values signal churn risk. Improve billing transparency and better align promotional pricing with long-term rates to mitigate “sticker shock.”

4. Targeted Support for New Customers

- Both models flag 0–12-month tenure as high-risk. Implement specialized onboarding and milestone campaigns to reinforce engagement early in the lifecycle.

5. Model Deployment & Monitoring Strategy

- Use the Logistic Regression model (optimized for recall) to identify at-risk users for intensive retention efforts.
- Use the Stacking Ensemble (optimized for precision) for cost-efficient, high-confidence marketing interventions.
- Establish continuous monitoring through A/B testing, tracking real-world prediction errors, and scheduling quarterly retraining with updated data.

6. Ethical Application of Predictive Features

- Define clear usage protocols for sensitive variables (e.g., payment method, senior status).
- Ensure fairness in intervention design, avoiding disproportionate targeting of vulnerable customer groups.
- Introduce internal audits to monitor equity in retention campaign outcomes.

Limitations

This project presents several important limitations that should be considered when interpreting the results and recommendations:

1. Temporal Constraints:

   The dataset provided a static snapshot of customer data, limiting our ability to capture seasonal churn trends or long-term behavioral shifts. Consequently, predictions may not reflect patterns that emerge over time or during specific billing or service cycles.

2. Limited Behavioral Signals:

   Our models primarily relied on static customer attributes (e.g., tenure, contract type), rather than dynamic indicators such as service usage frequency, engagement with digital tools, or customer support interactions. This restricts the ability to detect early signs of dissatisfaction or disengagement.

3. Incomplete Demographic Representation:

   Although the analysis raised ethical considerations around sensitive features (e.g., payment method, senior citizen status), the dataset lacked sufficient demographic granularity (e.g., income, ethnicity, or geographic diversity) to comprehensively assess potential model biases or equity implications.

4. Theoretical Deployment Strategy:

   While we proposed deployment and monitoring framework, it remains untested in real-world environments. The absence of live A/B testing or customer feedback limits our ability to assess the true effectiveness of interventions across different segments.

Appendix

Code snippets

Model hyperparameter grids:

```
MODEL_PARAMS = {
    "Logistic Regression": {
        'model__C': [0.01, 0.1, 1.0, 10.0, 100.0],
        'model__class_weight': [None, 'balanced', {0: 1, 1: 3}, {0: 1, 1: 5}],
        'model__solver': ['liblinear', 'saga'],
        'model__penalty': ['l1', 'l2']
    },
    "Random Forest": {
        'model__n_estimators': [100, 200, 500],
        'model__max_depth': [None, 15, 25, 35],
        'model__min_samples_split': [2, 5, 10],
        'model__min_samples_leaf': [1, 2, 4],
        'model__class_weight': [None, 'balanced', 'balanced_subsample']
    },
    "XGBoost": {
        'model__n_estimators': [100, 200, 300],
        'model__learning_rate': [0.01, 0.05, 0.1],
        'model__max_depth': [3, 5, 7],
        'model__subsample': [0.8, 0.9, 1.0],
        'model__colsample_bytree': [0.8, 0.9, 1.0],
        'model__scale_pos_weight': [1, 3, 5, 7]
    },
    "SGD": {
        "model__alpha": [0.0001, 0.001, 0.01, 0.1],
        "model__penalty": ["l2", "l1", "elasticnet"],
        "model__loss": ["log_loss"],
        "model__max_iter": [1000, 2000],
        "model__learning_rate": ["optimal", "adaptive"],
        "model__class_weight": [None, 'balanced', {0: 1, 1: 3}]
    }
}
```

Model Evaluation:

```
for name, model in models.items():
        # Make predictions on test set
        y_pred = model.predict(X_holdout)
        y_proba = model.predict_proba(X_holdout)[:, 1]

        # Calculate metrics
        metrics = {
            'accuracy': accuracy_score(y_holdout, y_pred),
            'f1': f1_score(y_holdout, y_pred),
            'precision': precision_score(y_holdout, y_pred),
            'recall': recall_score(y_holdout, y_pred),
            'roc_auc': roc_auc_score(y_holdout, y_proba)
        }

        # Print results
        print(f"\nTest Metrics for {name}:")
        for metric, value in metrics.items():
            print(f"{metric.capitalize()}: {value:.4f}")

        # Print classification report
        print(f"\nClassification Report for {name} on Test Set:")
        print(classification_report(y_holdout, y_pred))

        # Store test results for DataFrame
        test_results.append((
            name,
            metrics['accuracy'],
            metrics['f1'],
            metrics['precision'],
            metrics['recall'],
            metrics['roc_auc']
        ))
```

Clustering:

```
# Scale numerical features
numerical_cols = ['tenure', 'MonthlyCharges', 'TotalCharges', 'AverageMonthlySpend']
scaler = StandardScaler()
data_scaled = data.copy()
data_scaled[numerical_cols] = scaler.fit_transform(data[numerical_cols])

# 2. Determine optimal number of clusters using the Elbow method
inertia = []
k_range = range(2, 11)
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_scaled.drop(columns=['Churn']))
    inertia.append(kmeans.inertia_)

plt.figure(figsize=(10, 6))
plt.plot(k_range, inertia, marker='o')
plt.title('Elbow Method For Optimal k')
plt.xlabel('Number of clusters')
plt.ylabel('Inertia')
plt.show()

# 3. Apply K-means with optimal k
k = 3  # Update this based on your elbow plot
kmeans = KMeans(n_clusters=k, random_state=42)
data_scaled['Cluster'] = kmeans.fit_predict(data_scaled.drop(columns=['Churn']))
# 5. Visualize clusters with PCA for dimensionality reduction
```

```
pca = PCA(n_components=2)
pca_result = pca.fit_transform(data_scaled.drop(columns=['Churn', 'Cluster']))
data_scaled['PCA1'] = pca_result[:, 0]
data_scaled['PCA2'] = pca_result[:, 1]

# Analyze component loadings
features = data_scaled.drop(columns=['Churn', 'Cluster', 'PCA1', 'PCA2']).columns
loadings = pd.DataFrame(
    pca.components_.T,
    columns=['PC1', 'PC2'],
    index=features
)


# Display top contributors to each component
print("\nTop 5 Contributors to Each Principal Component:")
for i, pc in enumerate(['PC1', 'PC2']):
    print(f"\n{pc} Top Contributors:")
    print(loadings[pc].abs().sort_values(ascending=False).head(5))
# Analyze clusters using original (unscaled) data
original_cluster_analysis = data.groupby('Cluster').mean().sort_values('Churn', ascending=False)
```

Python Libraries Used

Pandas, Numpy, Sklearn, Matplotlib, Seaborn, Xgboost, Imblearn

Data Dictionary

Table 3: Data Dictionary

Column Name	Description	Non-Null Count	Distinct Count	Min	Max	Avg	Std Dev
customerID	Unique identifier for each customer.	7043	7043	N/A	N/A	N/A	N/A
gender	Whether the customer is a male or a female.	7043	2	N/A	N/A	N/A	N/A
SeniorCitizen	Whether the customer is a senior citizen.	7043	2	0	1	0.16	0.37
Partner	Whether the customer has a partner.	7043	2	N/A	N/A	N/A	N/A
Dependents	Whether the customer has dependents.	7043	2	N/A	N/A	N/A	N/A
tenure	Number of months the customer has stayed with the company.	7043	73	0	72	32.37	24.56

PhoneService	Whether the customer has a phone service.	7043	2	N/A	N/A	N/A	N/A
MultipleLines	Whether the customer has multiple lines.	7043	3	N/A	N/A	N/A	N/A
InternetService	The customer's internet service provider.	7043	3	N/A	N/A	N/A	N/A
OnlineSecurity	Whether the customer has online security.	7043	3	N/A	N/A	N/A	N/A
OnlineBackup	Whether the customer has online backup.	7043	3	N/A	N/A	N/A	N/A
DeviceProtection	Whether the customer has device protection.	7043	3	N/A	N/A	N/A	N/A
TechSupport	Whether the customer has tech support.	7043	3	N/A	N/A	N/A	N/A
StreamingTV	Whether the customer has streaming TV.	7043	3	N/A	N/A	N/A	N/A
StreamingMovies	Whether the customer has streaming movies.	7043	3	N/A	N/A	N/A	N/A
Contract	The contract term of the customer.	7043	3	N/A	N/A	N/A	N/A
PaperlessBilling	Whether the customer has paperless billing.	7043	2	N/A	N/A	N/A	N/A
PaymentMethod	The customer's payment method.	7043	4	N/A	N/A	N/A	N/A
MonthlyCharges	The amount charged to the customer monthly.	7043	1585	18.25	118.75	64.76	30.09

TotalCharges	The total amount charged to the customer.	7043	6531	N/A	N/A	N/A	N/A
Churn	Whether the customer churned. Target variable.	7043	2	N/A	N/A	N/A	N/A