1. Executive Summary

This project was initiated to address a critical business problem in the telecommunications industry: customer churn. Churn directly impacts long-term profitability, and reducing churn by even a few percentage points can significantly increase revenue retention.

The goal was to build a predictive model that identifies customers at high risk of churning, uncover the top drivers of churn, and deliver actionable insights to reduce customer loss. A dataset of 6,687 customers was analyzed, incorporating variables such as service usage, contract types, billing details, and demographics.

After performing exploratory data analysis, segment-level churn evaluation, and training multiple models on three versions of the dataset, **XGBoost trained on Version 1** emerged as the top-performing model. It achieved an accuracy of **91.13%**, a precision of **85%**, a recall of **81%**, and an F1 score of **83%** for the churn class — making it highly effective in identifying at-risk customers.

Based on model performance and a conservative business assumption that **40%** of flagged customers could be retained, the churn reduction potential is estimated at **7.34%**. This translates to a projected retention of approximately **$319,630 in customer lifetime value**, based on the median spend in the current customer base.

2. Business Problem & Goal

Customer churn poses a significant challenge in the telecommunications industry. High churn rates not only lead to direct revenue loss, but also increase customer acquisition costs, putting pressure on profit margins. Retaining existing customers is far more cost-effective than acquiring new ones, making churn reduction a critical strategic priority.

This project aims to analyze customer behavior, uncover the key factors that contribute to churn, and develop a predictive machine learning model that identifies customers at high risk of leaving. By accurately flagging these individuals, the business can take targeted retention actions to reduce churn, protect customer lifetime value, and improve long-term profitability.

3. Data Overview

(Add UML Diagram)

The dataset used contains 6,687 customer records, each representing an individual subscriber to a telecommunications service. It includes features that describe customer demographics, service usage patterns, billing details, contract type, and support interactions. Key features can be clustered into the following groups, account information, usage metrics, plan attributes, customer demographics, customer service interactions, and target variable which would include status of customer. The UML diagram shown below, highlights key features characterizing each individual customer.

In addition to the original features, feature engineering was performed to extract higher level insights from the raw variables, the features are as such:

- **Engagement Score:** Derived from Call Minutes and data Usage.
- **Support Intensity:** Based on customer services calls relative to tenure.
- **Cost Efficiency:** Monthly charge relative to total usage

- **Billing Pain** and **Usage to Cost Ratio**: Indicators of customer satisfaction vs cost burden.

Overall, the dataset went through a series of preprocessing stages, to handle missing values, outliers, proper encoding of data, and a creation of multiple versions of datasets to evaluate the model's sensitivity to normalization. Multiple versions of the dataset were created, and table x and y shown below show cases the different versions along with data types of variables.

Table 1. Dataset versions

V1	Original dataset with minimal filtering; includes most raw features.
V2	Dataset with a curated subset of selected features based on EDA and correlation.
V3	Version 2 + Engineered features + Collinearity reduced feature set.

Table 2. Dataset Feature Breakdown

Type	Count	Variable names
Numerical	13	Account Length, Local Mins, Total Charges, Age, Customer, Service Calls
Categorical	11	Contract Type, Payment method, Gender, Senior, International plan, Group
Engineered (Numerical)	5	Engagement Score, Cost Efficiency, Support Intensity, Billing pain, and usage to cost ratio.
Target Variable	1	Churn Label

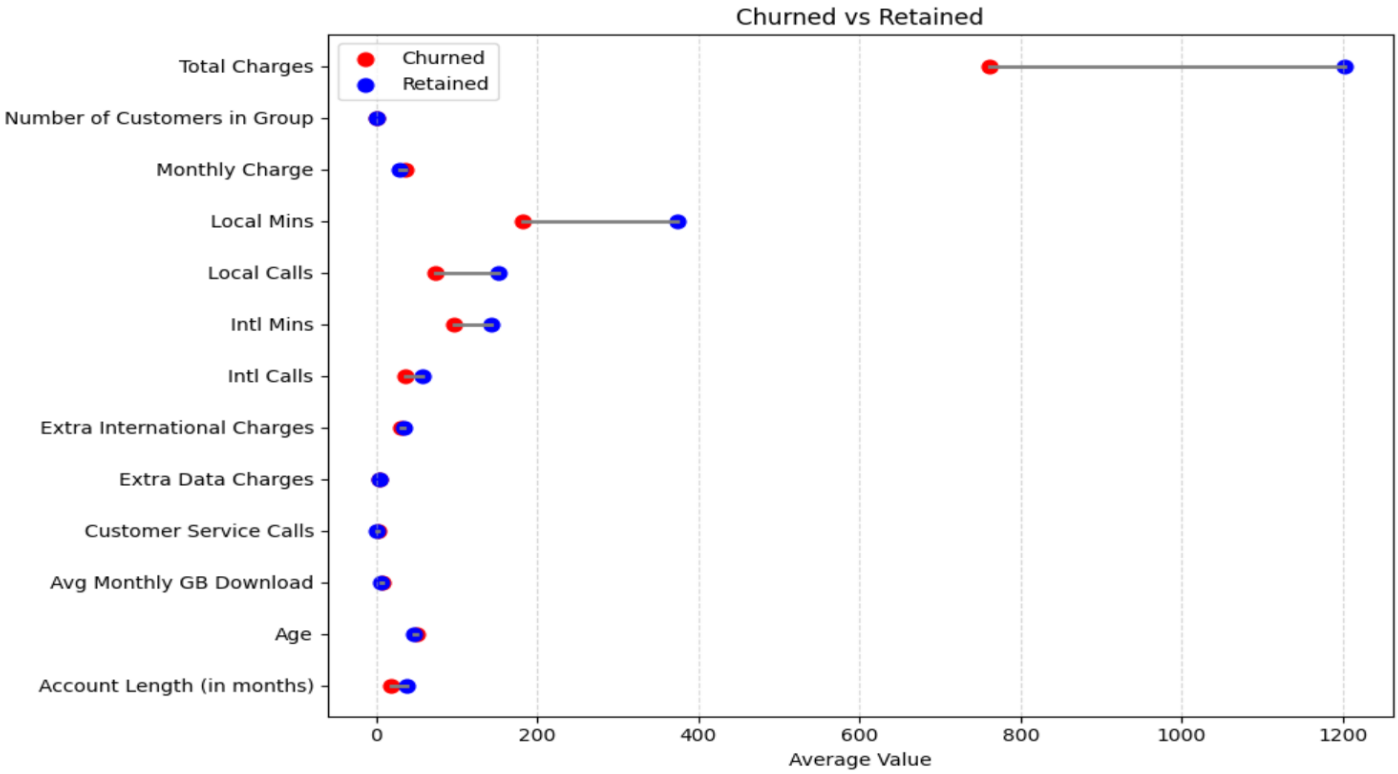
4. Exploratory Data Analysis & Insights

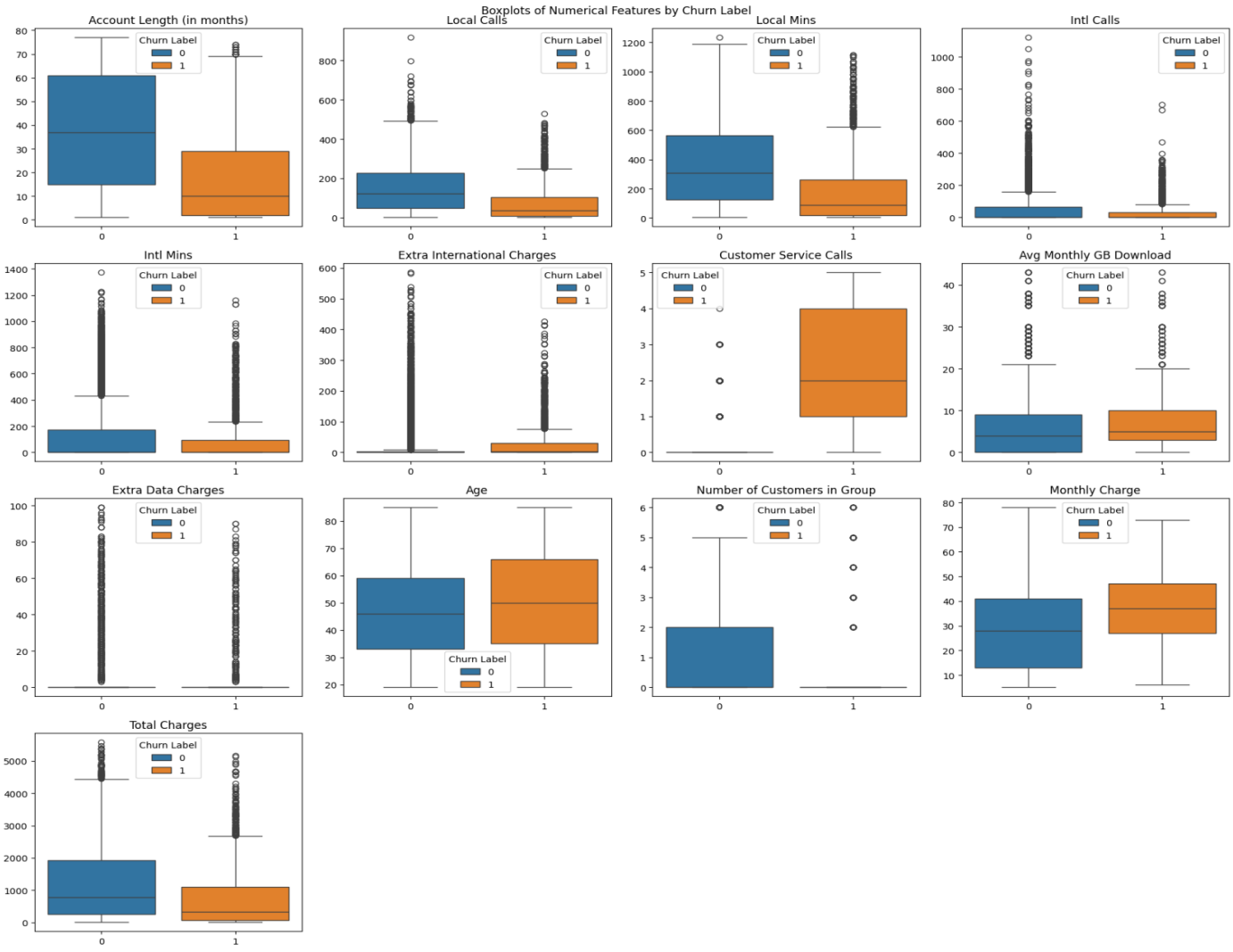
4.1. Stakeholder Questions

a. What are the top drivers of customer churn?
b. What customer segments have the highest churn rate and lifetime value loss?
c. Does the Presence of an international plan or unlimited data plan increase churn?
d. What are the most common churn reasons and how can we address them?
e. Are customers on month-to-month contracts more likely to churn?

4.2. Exploratory Data Analysis – Numerical Data

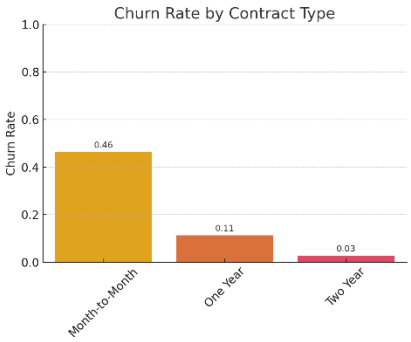
- **Account Length:** churners had shorter account lengths compared to non-churning users who have longer account lengths.
- **Local Calls & Minutes:** Churners generally don't utilize local calling services which indicates disengagement to the product.
- **Monthly GB Download:** On average churners have higher data consumption.
- **Extra International Charges:** While usage is low, churners have a slightly higher charges indicating potential pricing friction.
- **Customer service calls:** A strong churn indicator – users with 2+ calls are far more likely to leave, possibly reflecting dissatisfaction or unresolved issues.
- **Number of Customers in a group:** Churners tend to be enrolled in plans individually whereas others have more than 1 members associated in their plan.
- **Monthly Charges:** Churners typically spend more on average on a monthly basis, indicating potential dissatisfaction with the pricing of services.
- **Total Charges:** Similarly due to their low lifecycle, churners tend to spend less in total indicating low customer lifetime value.

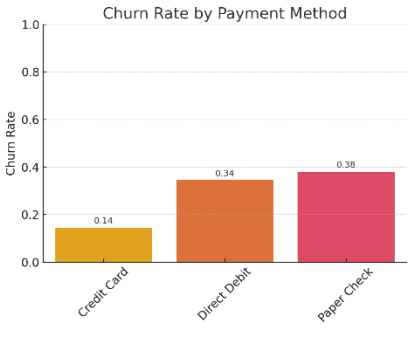

Churned vs Retained
Churned
Retained
Total Charges
Number of Customers in Group
Monthly Charge
Local Mins
Local Calls
Intl Mins
Intl Calls
Extra International Charges
Extra Data Charges
Customer Service Calls
Avg Monthly GB Download
Age
Account Length (in months)
0
200
400
600
800
1000
1200
Average Value

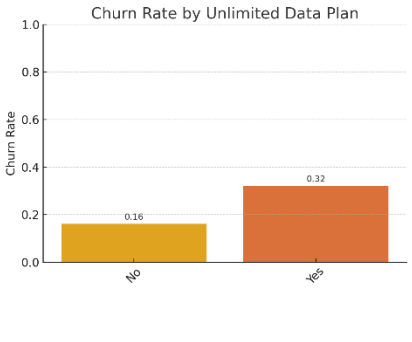

Boxplots of Numerical Features by Churn Label
Account Length (in months)
Local Calls
Local Mins
Intl Calls
Intl Mins
Extra International Charges
Customer Service Calls
Avg Monthly GB Download
Extra Data Charges
Age
Number of Customers in Group
Monthly Charge
Total Charges
Churn Label
0
1

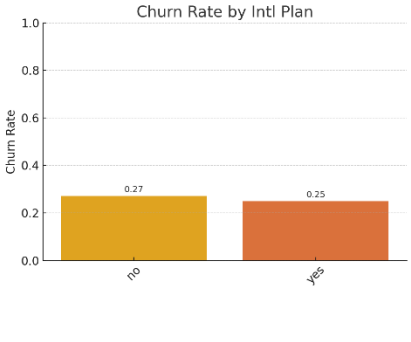
4.3. Exploratory data analysis – Categorical Data

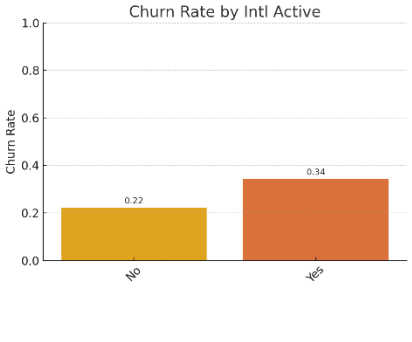
- **Contract Type:** Churn is heavily concentrated among month-to-month users. Longer term contracts have significantly lower churn rates
- **Group:** Customers enrolled in group plans have drastically lower churn rate than solo users.
- **International Active:** users with an active international plan are more likely to churn.
- **Payment method:** Direct Debit and Paper check users have the highest churn rates; credit card users show more stability and are more likely to stay enrolled with eh service.
- **Senior:** Older customers are more likely to churn, likely due to service complexity or evolving needs.'
- **Unlimited data plan:** High churn rate among unlimited plan users suggests unmet expectations or dissatisfaction with the plan selected.

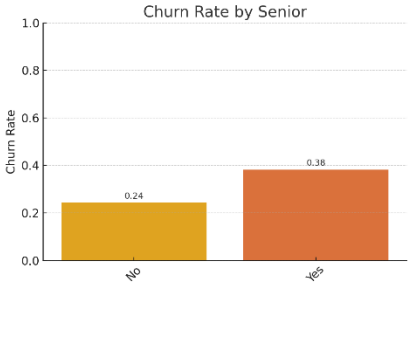

Churn Rate by Contract Type
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.46
0.11
0.03
Month-to-Month
One Year
Two Year

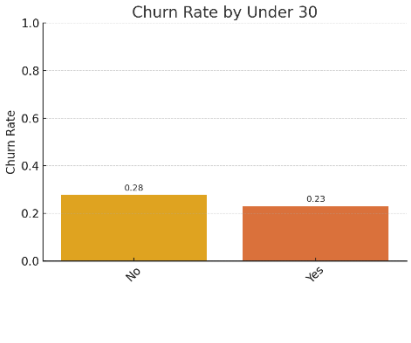

Churn Rate by Payment Method
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.14
0.34
0.38
Credit Card
Direct Debit
Paper Check

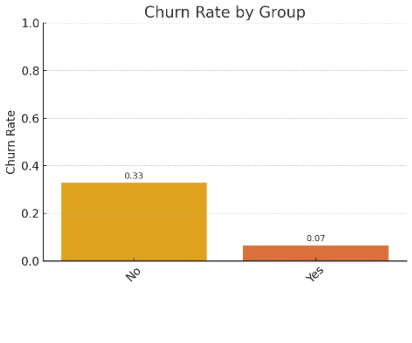

Churn Rate by Unlimited Data Plan
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.16
0.32
No
Yes

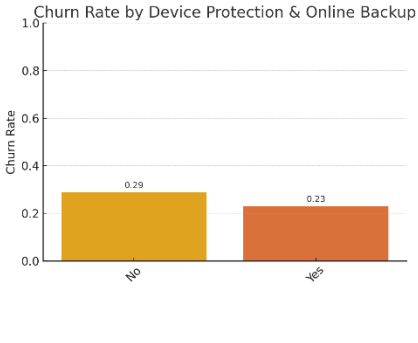

Churn Rate by Intl Plan
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.27
0.25
no
yes

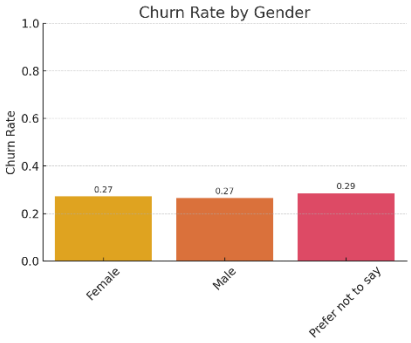

Churn Rate by Intl Active
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.22
0.34
No
Yes

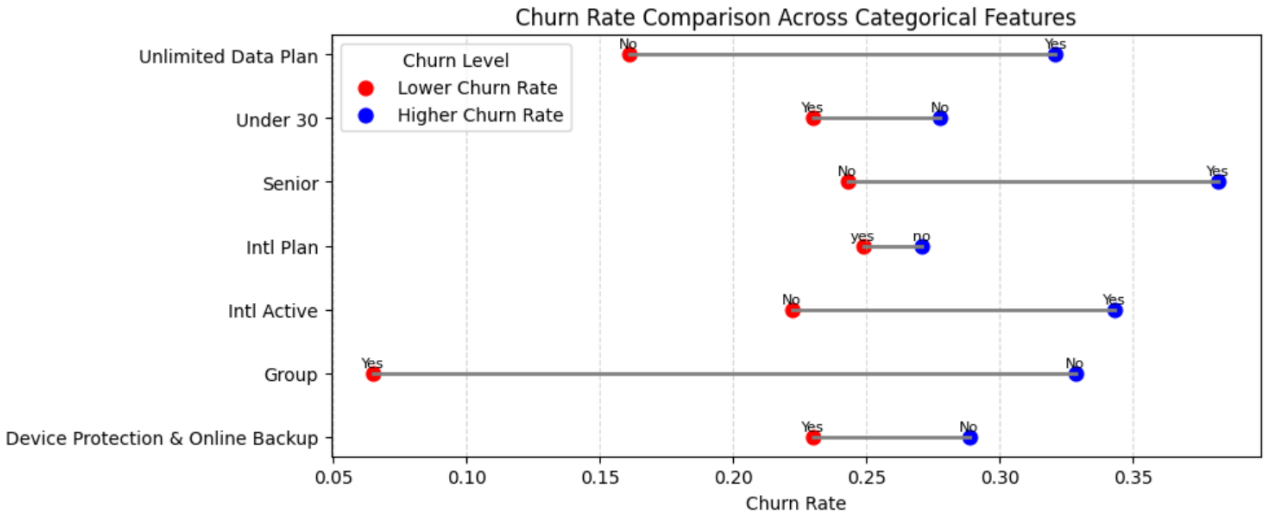

Churn Rate by Senior
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.24
0.38
No
Yes


Churn Rate by Under 30
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.28
0.23
No
Yes


Churn Rate by Group
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.33
0.07
No
Yes


Churn Rate by Device Protection & Online Backup
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.29
0.23
No
Yes


Churn Rate by Gender
Churn Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.27
0.27
0.29
Female
Male
Prefer not to say



4.4. Correlation Insights – Top Drivers of churn

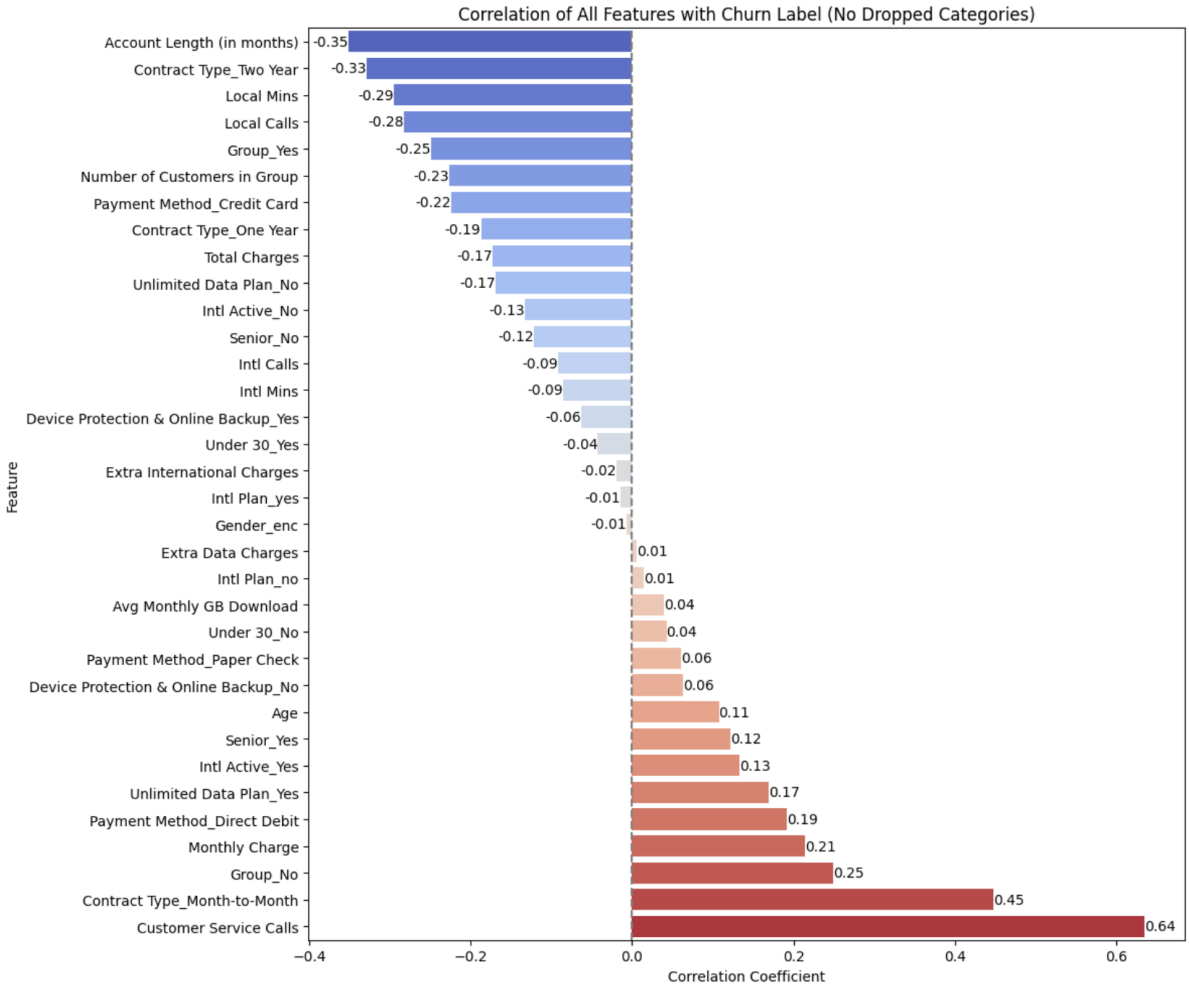
Features with positive correlation are more likely to contribute towards churn, and negative correlation are less likely to contribute to churn.

Numerical Features

- Account Length: Longer tenure reduces churn
- Local Mins: Higher usage indicates engagement which reduces churn
- Local Calls: Higher usage indicates engagement which reduces churn
- Total Charges: Measurement for lifetime value.
- Monthly Charges: Higher values linked to higher churn
- Customer Service Calls: Most predictive churn driver

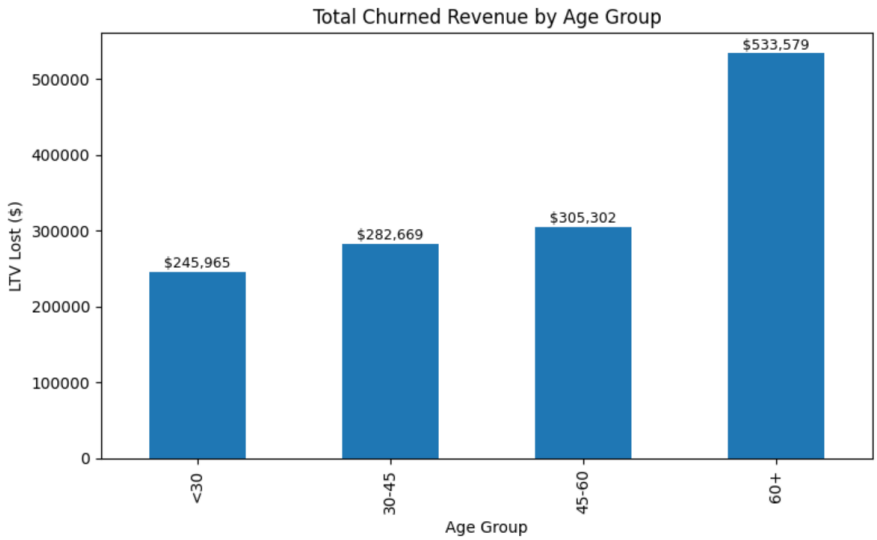
Categorical Features

- Contract Type: A strong churn predictor highlighting that month-to-month users strongly increase churn
- Group: Group Members are more loyal
- Number of Customers in a Group: Quantifies group effect
- Payment Method: Paper & Direct Debit users churn more
- Unlimited Data Plan: Users with the plan churn more
- International Active: Active international users churn more
- Senior: are slightly more prone to churn



Evaluating the analysis performed and insights derived from the conducting visual EDA & feature correlation analysis, the key drivers of churn are the following: Account Length, Local calls, Local Mins, Total Charges, Monthly Charges, Customer service calls, Contract type, Group, Number of customers in a group, payment method, unlimited data plan, international active, and senior.

4.5. Customer segmentation of highest churners and lifetime value loss



Customers of age 45 and above contribute to a significant LTV loss of more than 300,000$.

4.6. International Plan vs Unlimited Data plan effect on churn rate

		Total Customers	Churned Customers	Total LTV ($)	LTV Lost ($)	Churn Rate
International Plan	No	6036	1634	6570876	1259934	0.270709
	Yes	651	162	676198	107581	0.248848
Unlimited data plan	No	4494	1443	5848706	1148233	0.321095
	Yes	2193	353	1398368	219282	0.160967

International plan does not affect churn since churn rate is quite marginal, and customers who use the plan represent a small segment of the total customer base. Whereas customers with an unlimited data plan have a churn rate of 32.1% and have life time value (LTV) loss of more than 1 million dollars, it is an important feature to look at and consider.

4.7. Most Common Churn reasons & Potential Retention Strategies

There are 4 major categories where we can cluster the reasons into:

1. **Competitive Offers**

- 805 churned customers, $622,757 LTV lost
- Reason: Competitors provided better pricing, devices, data, or speeds.
- Strategy: Launch targeted retention offers, promote device upgrades, and offer exclusive bundles to stay competitive.

3. **Customer Service Experience**

- 354 churned customers, $242,087 LTV lost
- Reason: Dissatisfaction with the attitude and professionalism of support staff.
- Strategy: Invest in communication and conflict resolution training, enhance support across all channels, and implement AI/chatbot tools to reduce repetitive inquiries.

3. **Price & Value Perception**

- 172 churned customers, $136,472 LTV lost
- Reason: Customers feel they are overpaying or encountering unexpected charges.
- Strategy: Provide transparent pricing, introduce capped or rollover data plans, and deploy alerts and savings recommendations for heavy users.

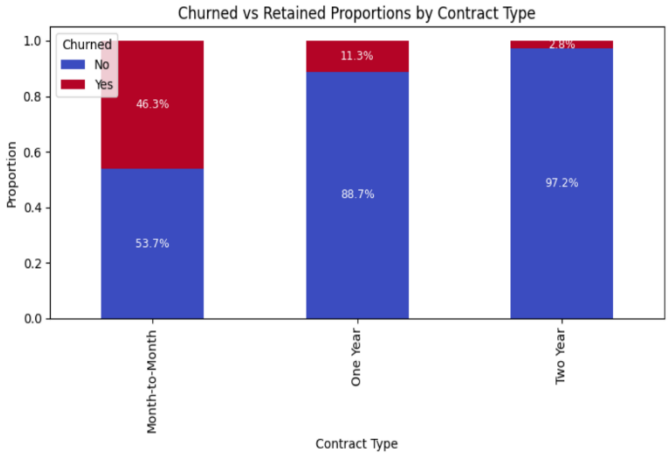
4. **Product & Network Performance**

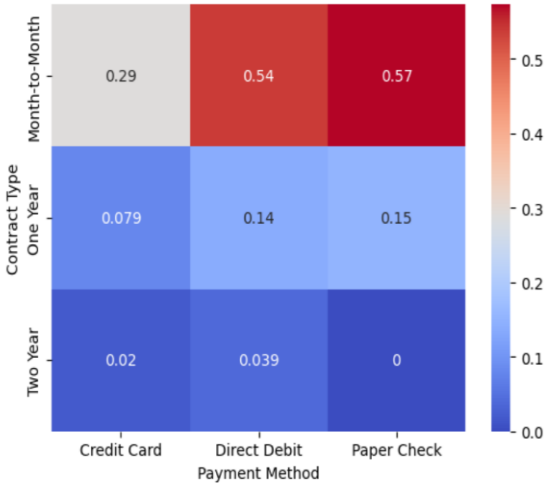
- 205 churned customers, $159,969 LTV lost
- Reason: Poor product satisfaction and unreliable network service.
- Strategy: Improve infrastructure and coverage, run regular satisfaction surveys, and offer temporary plan upgrades during issue resolution.

Overall, majority of churn is driven by competition, poor customer support experience, and pricing.

4.8. Contract type and Payment Method Churn rate analysis

Customers on month-to-month contracts exhibit the highest churn rates. Additionally, those who use direct debit or paper checks for payments are more likely to churn.





4.9. EDA Summary

Upon conducting a deep analysis and answering stakeholder questions, the data reveals that the following key features play a significant role in predicting churn. The features are Account Length, Local calls, Local Mins, Total Charges, Monthly Charges, Customer service calls, Contract type, Group, Number of customers in a group, payment method, unlimited data plan, international active, and senior.

5. Modeling Approach

The goal of this phase is to build a machine learning model capable of reliably capture future churners. The focus was to identify at-risk customers early to enable proactive retention strategies.

5.1. Objective

A binary classification model was developed to distinguish between churned and retained customers. The primary business goal was to maximize the correct identification of churners to reduce revenue loss and improve retention efficiency.

5.2. Data preparation

As shown in Table 1 of Chapter 3, three versions of the dataset were used to evaluate model performance. Data preprocessing was a crucial step to ensure the dataset was suitable for training machine learning models. This process included handling missing, duplicate, or unknown values through imputation, encoding categorical variables into numerical formats, standardizing features, and splitting the data using a 70/30 stratified sampling approach for training and evaluation.

5.3. Model Selection & Evaluation Strategy

Three models were selected for evaluation:

- **Logistic Regression** - used for interpretability and baseline comparison

- **Random Forest** - used for strong general performance and feature importance insights.
- **XGBoost** - used for optimized accuracy and robust handling of structured data.

To evaluate model performance, four key metrics were used: Accuracy, Precision, Recall, and F1 Score — each offering a different lens on how effectively the model identifies churners. While accuracy reflects overall correctness, precision ensures that retention resources are focused only on customers truly at risk. Recall measures the model's ability to catch actual churners in time, reducing lost revenue. The F1 Score balances both precision and recall, representing the model's overall reliability in churn prediction. Together, these metrics confirm that the model is both efficient and actionable — identifying high-risk customers with confidence while minimizing wasted effort.

6. Results Summary

Data Version	Model	Scaled	Model Accuracy	Precision (1)	Recall (1)	F1 Score (1)
V1	Logistic Regression	No	0.8968	0.85	0.75	0.8
		Yes	0.8963	0.85	0.75	0.8
	Random Forest	No	0.9083	0.86	0.78	0.82
		Yes	0.9073	0.86	0.78	0.82
	XGBoost	No	0.9113	0.85	0.81	0.83
		Yes	0.9113	0.85	0.81	0.83
V2	Logistic Regression	No	0.8943	0.85	0.74	0.79
		Yes	0.8953	0.85	0.74	0.79
	Random Forest	No	0.8943	0.84	0.76	0.79
		Yes	0.8923	0.83	0.75	0.79
	XGBoost	No	0.8993	0.85	0.76	0.8
		Yes	0.8978	0.84	0.77	0.8
V3	Logistic Regression	No	0.8993	0.85	0.76	0.8
		Yes	0.8993	0.85	0.76	0.8
	Random Forest	No	0.8938	0.83	0.76	0.79
		Yes	0.8938	0.83	0.76	0.79
	XGBoost	No	0.8933	0.84	0.75	0.79
		Yes	0.8933	0.84	0.75	0.79

The best overall model was XGBoost trained and tested on version1 of the dataset which outperformed all models. These results obtained can be translated into real business value as such:

- Precision of 85% means that most customers flagged as likely to churn truly are, which helps ensure that retention efforts and incentives are not wasted on customers who were unlikely to leave.
- Recall of 81% means that the model is successfully identifying the majority of churners before they leave, giving the business a meaningful chance to intervene.
- The F1 Score of 83% reflects a strong balance between the two — making the model reliable and consistent in practice.

Together, these metrics make the model both effective and efficient in a real-world churn prevention strategy.

7. Business Recommendations

Based on the results of the churn analysis and predictive modeling, the following recommendations are proposed to reduce churn, retain high-value customers, and improve long-term revenue outcomes.

a. Focus retention efforts on high-risk segments

Use the trained model to score customers weekly or monthly and **prioritize interventions** for those flagged as high risk. Begin with:

- Month-to-month contract users
- Solo users (not in group plans)
- Customers using Direct Debit or Paper Check
- Users with high monthly charges and frequent support interactions

b. Design targeted retention programs

- Offer loyalty incentives or upgrades to customers with rising churn scores
- Create contract conversion offers for flexible-plan customers
- Promote group plans via discounts or referral programs
- Engage customers who make multiple service calls with personalized outreach

c. Improve onboarding & Early engagement

Since short-tenure customers are more likely to churn:

- Strengthen the onboarding experience
- Send usage tips, feature introductions, and value education in the first 90 days
- Monitor support interactions closely during this phase

d. Integrate churn scoring into operations

Deploy the churn model through an internal dashboard or CRM system. Flag customers above a certain churn risk threshold and route them to the customer success or marketing teams for follow-up.

e. Track model impact over time

Measure the performance of retention campaigns driven by the churn model:

- Track reduction in churn rate among contacted customers
- Monitor LTV preserved by intervention
- Continue tuning the model with updated data

8. Estimated Business Impact

If implemented, the model is estimated to reduce churn by **7.34%**, protecting an estimated **$319,630** in customer lifetime value — a meaningful financial impact for a data-driven retention initiative.

Churn Context	Model Results
• Current churn rate: 26.85% • Total Customer base: 6,687 • Median customer lifetime value (LTV): $647 • Average LTV: $1,083.75	• Best performing: XGBoost (V1) • Recall (Churners Identified): 81% • Precision (Accuracy of flagged churners): 85% • Estimated Churn reduction: 7.34%

Impact Calculation:

- **Churned Customers in Dataset**: 1,853
- **Customers Potentially Saved (81% recall × 7.34%)**: ≈ 490 Customers saved.
- **Estimated Value Retained (Median LTV × 490)**: **$317,030** (conservative estimate)

To sum up, if model is deployed at scale, it has the potential to:

- Reduce churn by over 7%
- Retain **hundreds of customers per cycle.**
- Preserve **hundreds of thousands of dollars** in recurring revenue
- Enable smarter, **data-driven retention campaigns**

9. Project Summary

This project successfully combined exploratory analysis, advanced machine learning, and strategic business interpretation to address the challenge of customer churn.

Through detailed EDA and segmentation, we identified the behavioral, financial, and demographic drivers behind churn. A predictive churn model — led by XGBoost — was developed and tuned to high performance, correctly identifying over 81% of churners while maintaining strong precision.

The model was not only accurate, but actionable. Its deployment enables proactive retention, targeted interventions, and the preservation of significant customer lifetime value.

With an estimated **$317,000** in preventable churn loss captured in this analysis, this project proves that data science can directly protect revenue and drive smarter customer engagement.