**INTERNSHIP REPORT**

|  |
| --- |
| **End-to-End Analytics Framework for Customer Churn Prediction** |

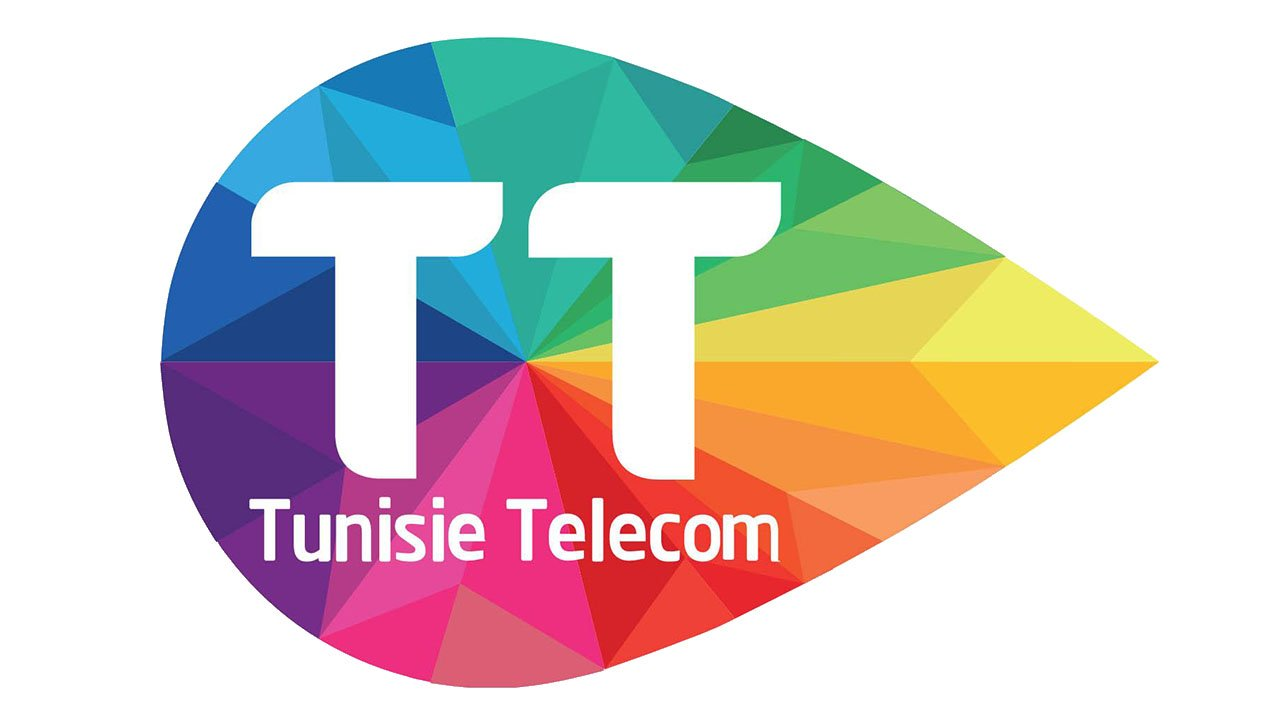
**BY**

**Sabrine Ayadi**

**Supervisor**

**Asma Jemili**

**Tunisie Télécom**



[1. Introduction 2](#_Toc626050957)

[2. Business Understanding 3](#_Toc218362717)

[3. Data Understanding 4](#_Toc443173310)

[3.1. Data source 4](#_Toc1670832368)

[3.2. Data Structure 4](#_Toc1258285463)

[3.3. Variables description 4](#_Toc1238489483)

[4. Data Preparation 5](#_Toc976565052)

[4.1. Data Type Corrections 6](#_Toc1065684771)

[4.2. Handling Missing and Inconsistent Values 6](#_Toc1947415267)

[4.3. Encoding Categorical Variables 7](#_Toc1284108340)

[5. Exploratory Data Analysis (EDA) 8](#_Toc173104710)

[5.1. Univariate Analysis 9](#_Toc693551507)

[5.2. Bivariate Analysis 11](#_Toc910541447)

[5.2.1 Churn Rate by Categorical Variables 11](#_Toc1441766555)

[5.2.2 Churn vs Numerical Variables 15](#_Toc299361238)

[5.2.2 Correlation analysis 17](#_Toc1045154264)

[5.3 Survival-Based Exploratory Data Analysis 18](#_Toc1211527673)

[5.3.1 Kaplan-Meier Survival Curve (Overall) 18](#_Toc1436455631)

[5.3.2 Stratified Survival Curves by Key Categorical Variables 19](#_Toc376879408)

[6. Modeling 21](#_Toc1954595578)

[6.1 Classification model 21](#_Toc144367578)

[6.1.1 Considered Algorithms 21](#_Toc1278060104)

[6.1.2 Handling class imbalance 23](#_Toc1127613874)

[6.1.3 Algorithm selection 23](#_Toc1678206578)

[6.1.4 Selected model vs Baseline model 25](#_Toc616284401)

[Refrences 27](#_Toc1166519467)

## Introduction

The telecom industry is one of the fastest-growing and most competitive sectors globally, playing a crucial role in connecting people and businesses. With increasing competition and evolving customer expectations, telecom companies must focus not only on acquiring new users but also on retaining existing ones. One of the biggest challenges they face in this context is customer churn — the phenomenon where customers discontinue their services and switch to competitors.

Customer churn can occur due to a variety of reasons, including better pricing, service quality, attractive promotional offers from competitors, or dissatisfaction with the current provider. The churn rate directly affects a company's revenue, customer lifetime value, and overall market stability. Retaining existing customers is often more cost-effective than acquiring new ones, which makes churn prediction a critical business objective.

In this project, we aim to address the churn problem by developing a data-driven predictive model that can accurately identify customers who are likely to leave. Furthermore, to go beyond simply identifying who might churn, we will adopt survival analysis techniques to estimate when a customer is most likely to churn — enabling more timely and effective intervention strategies. This end-to-end project follows the CRISP-DM methodology and integrates machine learning, statistical modeling, and dashboard visualization to support decision-making in a real-world telecom context.

## Business Understanding

Customer churn is the act of customers discontinuing their services . It is a major challenge for telecom companies. Retaining existing customers is significantly more cost-effective than acquiring new ones, which makes churn prediction a high-impact business problem. The primary objective of this project is to develop a predictive system that:

* Identifies customers at risk of churning,
* Estimates when churn is likely to occur,
* And supports timely, targeted retention strategies.

The project goes beyond traditional classification models by incorporating **survival analysis**, which enables the estimation of time-to-churn. This provides decision-makers with both the *who* and *when*, enhancing their ability to intervene effectively.

The publicly available **Telco Customer Churn dataset** provided by IBM will be utilized for this project . The dataset comprises over 7,000 customer records, containing demographic information, contract details, subscribed services, billing data, tenure (the length of customer engagement), and a binary indicator of whether a customer has churned. These comprehensive features enable both churn classification and survival analysis approaches.

Despite the simulated nature of the dataset and monthly-level resolution of customer tenure, it offers a valuable foundation to apply both classification and survival analysis techniques.

Business success will be measured by the model’s ability to help reduce churn through earlier and more accurate risk identification. From a technical standpoint, model performance will be evaluated using metrics such as AUC, precision, and recall for classification models, and Concordance Index (C-index) for survival models

## Data Understanding

### 3.1. Data source

The **Telco Customer Churn dataset**, publicly provided by **IBM**, contains a total of **7,043 records**, where each row represents a unique customer. The dataset includes a wide range of features such as **demographic information**, **contract details**, **billing data**, and the **services each customer is subscribed to**. It also contains a **binary target variable**, Churn, indicating whether the customer discontinued the service during the observation period.

### 3.2. Data Structure

The dataset consists of **7,043 entries and 21 columns**. It includes a combination of **categorical**, **numerical**, and **binary** variables. Most features are stored as object types (strings), including some that represent numerical values such as TotalCharges, which will be converted during the data preparation phase. The customerID column serves as a unique identifier and will not be used as a predictive feature.

### 3.3. Variables description

* **customerID**: Unique identifier for each customer
* **gender**: Gender of the customer (Male or Female)
* **SeniorCitizen**: Whether the customer is a senior citizen (1 = Yes, 0 = No)
* **Partner**: Whether the customer has a partner (Yes or No)
* **Dependents**: Whether the customer has dependents (Yes or No)
* **tenure**: Number of months the customer has been with the company
* **PhoneService**: Whether the customer has a phone service
* **MultipleLines**: Whether the customer has multiple phone lines
* **InternetService**: Type of internet service (DSL, Fiber optic, or No)
* **OnlineSecurity**: Whether the customer has online security add-on
* **OnlineBackup**: Whether the customer has online backup add-on
* **DeviceProtection**: Whether the customer has device protection service
* **TechSupport**: Whether the customer has technical support service
* **StreamingTV**: Whether the customer has streaming TV service
* **StreamingMovies**: Whether the customer has streaming movie service
* **Contract**: Type of contract (Month-to-month, One year, Two year)
* **PaperlessBilling**: Whether the customer uses paperless billing
* **PaymentMethod**: Method of payment (e.g., Electronic check, Credit card)
* **MonthlyCharges**: Monthly amount charged to the customer
* **TotalCharges**: Total amount charged to the customer over their tenure
* **Churn**: Target variable — whether the customer has churned (Yes or No)

## Data Preparation

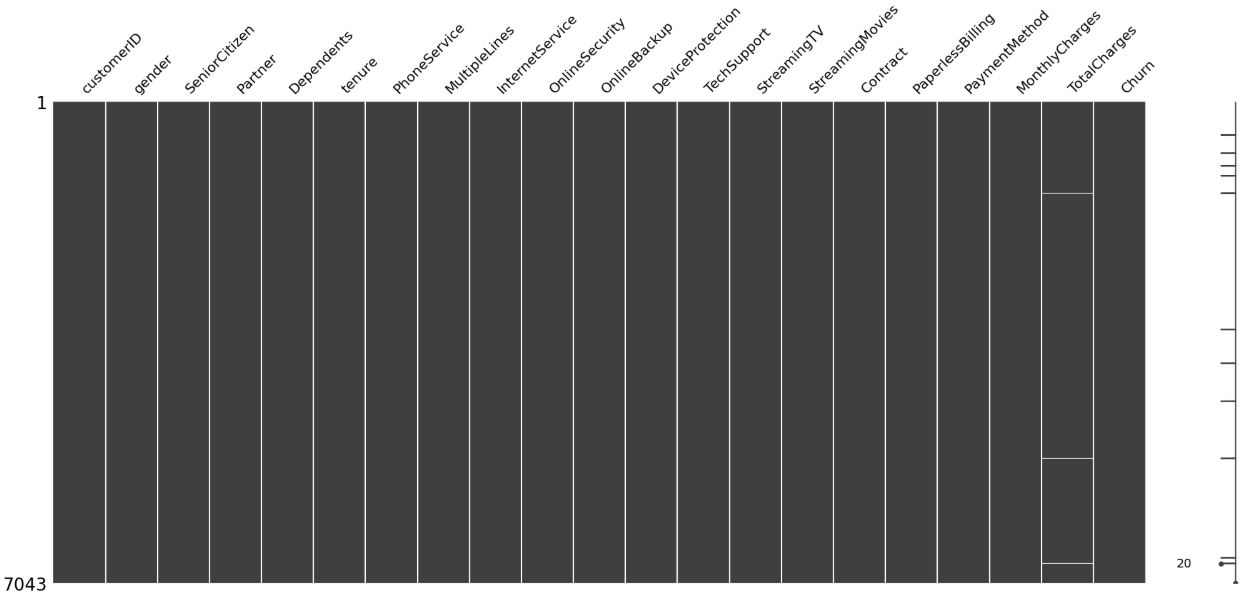
In this section, we present the modifications applied to the dataset, focusing on handling missing values, correcting data types, and resolving inconsistencies to ensure the data is clean, reliable, and ready for analysis

### 4.1. Data Type Corrections

After reviewing the data types, we observed that the Total Charges column was incorrectly stored as an object (string) rather than a numeric type. This issue likely stems from formatting inconsistencies or the presence of non-numeric values. To address this, we converted Total Charges to a numeric format to ensure it can be used in analysis and modeling

### 4.2. Handling Missing and Inconsistent Values

After inspecting missing values using a missing value matrix, we observed that the Total Charges column contained **11 missing entries**. These were associated with customers having a tenure of 0 — meaning they had just signed up and had not yet been billed. To resolve this, we replaced the missing Total Charges values with **0**, reflecting the correct billing status.



Additionally, we identified **inconsistencies in binary categorical variables** with "Yes"/"No" values (e.g., inconsistent casing or formatting). These were standardized to ensure uniform representation across the dataset

### 4.3. Encoding Categorical Variables

To prepare the dataset for modeling, we encoded all categorical variables into numerical format. First, we separated binary "Yes"/"No" variables from multi-category variables.

* **Binary variables** (e.g., Churn, Partner, Paperless Billing) were **label encoded**, with "Yes" mapped to 1 and "No" to 0.
* For **multi-category variables** (e.g., Contract, Payment Method, Internet Service), we applied **one-hot encoding**, dropping the first category to avoid multicollinearity. A mapping dictionary was created to retain the original category context, which documents the dropped base category and how each dummy variable corresponds to its original meaning.

All resulting Boolean columns were then converted to integers (0/1) to ensure compatibility with machine learning models. This step ensured that the dataset was fully numeric and ready for downstream analysis and modeling

**Example: Contract**

|  |  |  |  |
| --- | --- | --- | --- |
| **Original**  **Categories** | **Dropped Category** | **Encoded Columns** | **Interpretation** |
| ['Month-to-month', 'One year', 'Two year'] | Month-to-month | Contract\_One year, Contract\_Two year | e.g., Contract\_Two year = 1 means the customer is on a two-year contract |

**Example: Internet Service**

|  |  |  |  |
| --- | --- | --- | --- |
| **Original Categories** | **Dropped Category** | **Encoded Columns** | **Interpretation** |
| ['DSL', 'Fiber optic', 'No'] | DSL | InternetService\_Fiber optic, InternetService\_No | e.g., InternetService\_No = 1 means no internet service |

**Example: Payment Method**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Original Categories*** | ***Dropped Category*** | ***Encoded Columns*** | ***Interpretation*** |
| ['Bank transfer (automatic)', 'Credit card (automatic)', 'Electronic check', 'Mailed check'] | Bank transfer (automatic) | PaymentMethod\_Credit card (automatic), PaymentMethod\_Electronic check, PaymentMethod\_Mailed check | e.g., PaymentMethod\_Mailed check = 1 means the customer pays by mailed check |

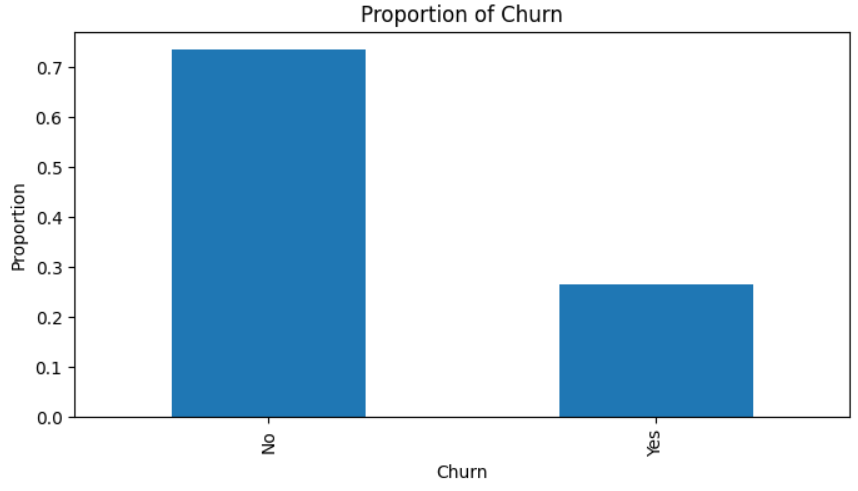
## Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) aims to uncover patterns in customer behavior and identify features strongly associated with churn. This includes both traditional bivariate analysis and time-to-event (survival) analysis to capture churn timing over a customer's lifecycle.

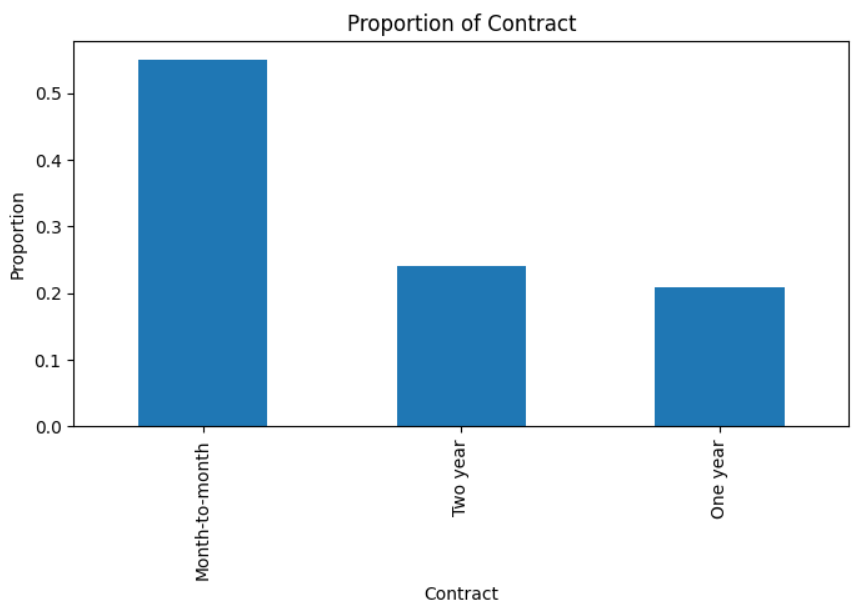
### 5.1. Univariate Analysis

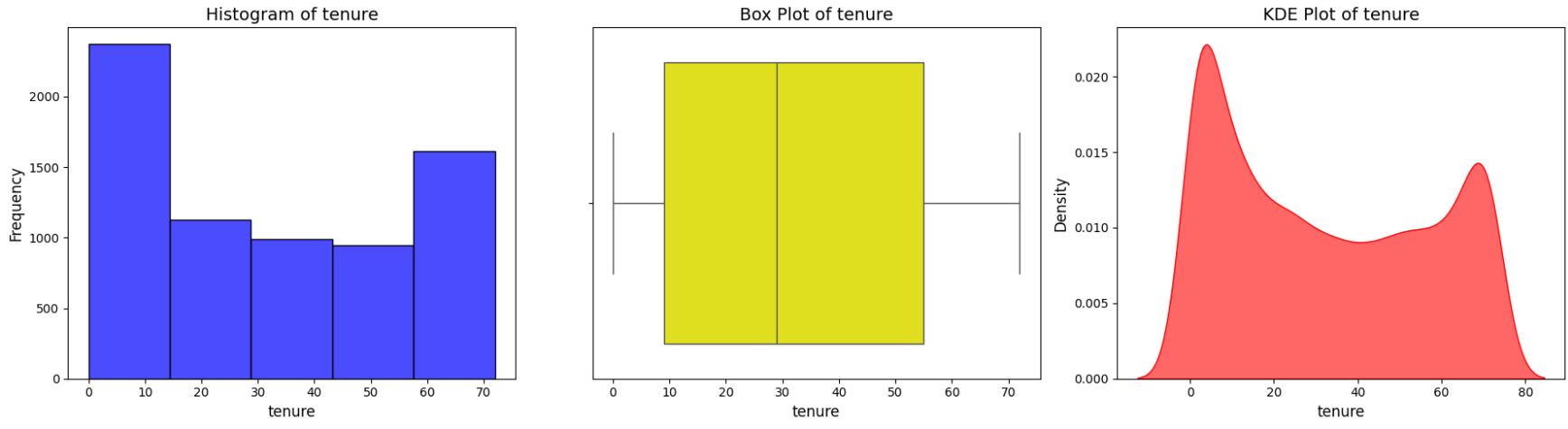
This section tackles individual variables in order to understand the characteristics of the customer base and the target variable (churn). This provides insight into the structure and makeup of the dataset before exploring inter-variable relationships.

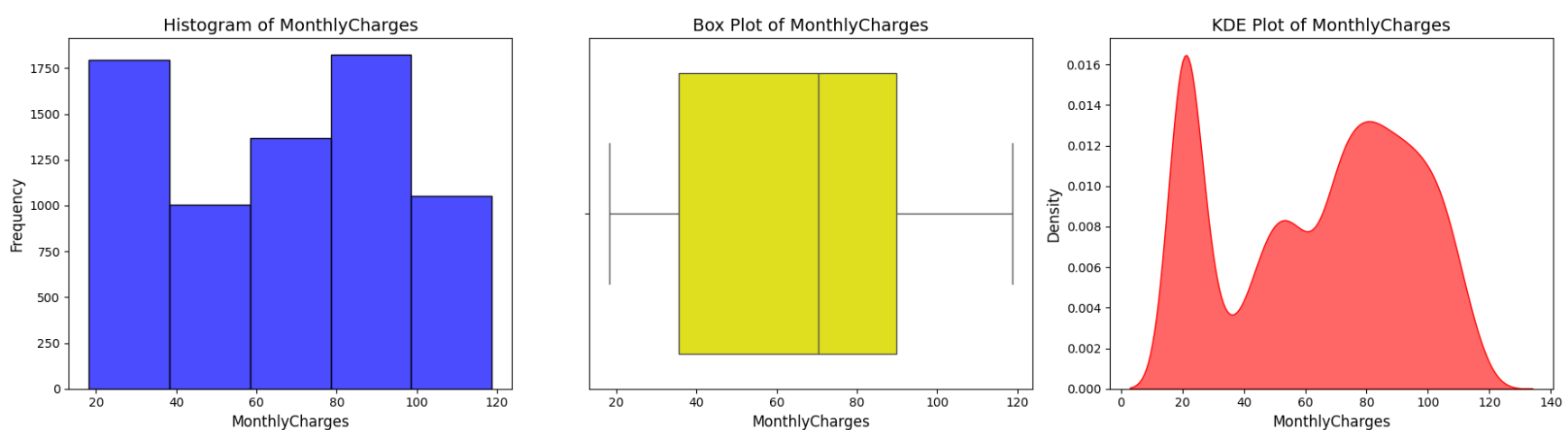
* **Churn:** The Churn variable is the primary target of our analysis . In the dataset, approximately **26%** of customers have churned, while **74%** have stayed. This shows a **moderate class imbalance**, which may affect model performance and should be addressed during model training, potentially through resampling techniques or class weighting. As seen in **Figure below,** the number of non-churned customers significantly outweighs the churned ones.



* **Gender:** The dataset is **balanced in terms of gender**, with approximately **51% male** and **49% female** customers.
* **Contract Type**: Customers can choose between three contract types: **month-to-month**, **one-year**, and **two-year** contracts. In our dataset, most customers (**around 55%**) are on **month-to-month contracts**, followed by **one-year contracts (25%)**, and **two-year contracts (20%). This** distribution, shown in the plot below, shows a clear preference for short-term contracts.



* **Tenure**: It ranges from **0 to 72 months**, with an **average of approximately 33 months** and a **median of 29 months**. As shown in the plots below, the distribution is **right-skewed**, indicating that most customers are relatively new, with fewer customers having longer tenure. The boxplot appears fairly symmetric, with no extreme outliers. The KDE plot confirms this skewed distribution, showing a peak in the early months and a gradual decline over time. 
* **Monthly charges:** It ranges from approximately **$20 to $120**, with an **average of around $65** and a **median of $70** indicating a central tendency toward moderate pricing. As shown in the histogram and KDE plot, the distribution is **roughly uniform**. The boxplot indicates a fairly wide spread in monthly billing amounts, and a few higher values may be considered mild outliers. These variations could reflect differences in service bundles or additional charges.



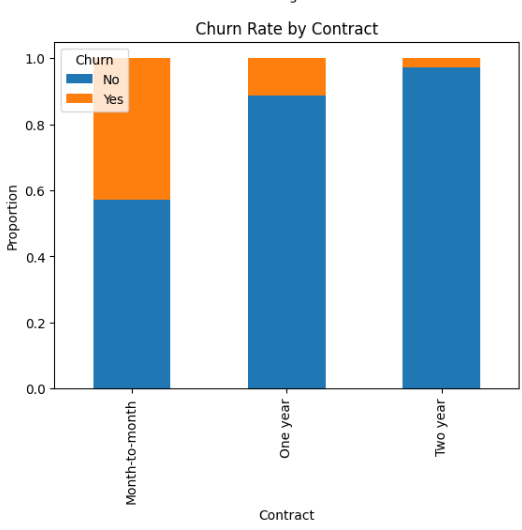
### 5.2. Bivariate Analysis

In this section, we analyze the relationship between key variables and the target variable Churn to identify patterns that may indicate risk factors or predictors of customer attrition. We focus on both categorical and numerical variables using appropriate visualizations and statistical summaries.

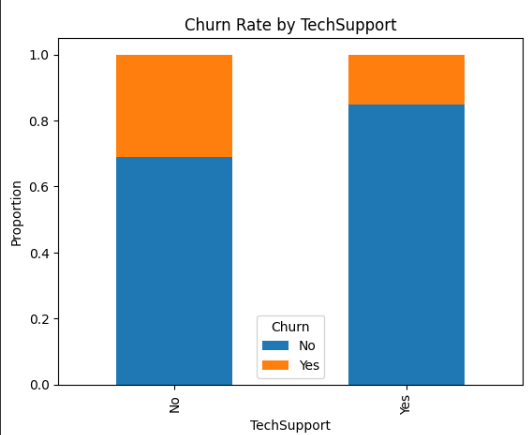
#### 5.2.1 Churn Rate by Categorical Variables

* **Churn rate by contract type :** The churn rate varies significantly across contract types. Customers on **month-to-month contracts** have a much higher churn rate (**43%**) compared to those on **one-year (11%)** or **two-year (3%)** contracts.

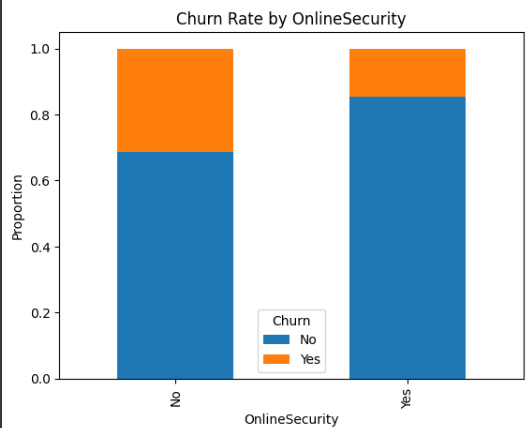
As seen in the grouped bar plot below, churn is most prevalent among those with short-term contracts, indicating that customers who avoid long-term commitments are more likely to leave



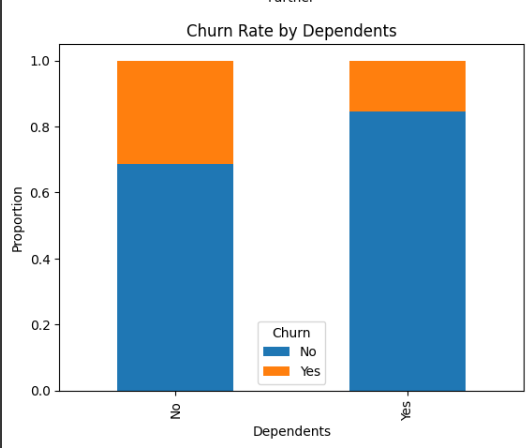
* **Churn rate by Tech support :** One of the clearest patterns is seen with tech support: customers **without tech support churn significantly more** than those with it. This indicates that **access to support services may be a strong deterrent against churn**, possibly due to increased satisfaction or perceived value.



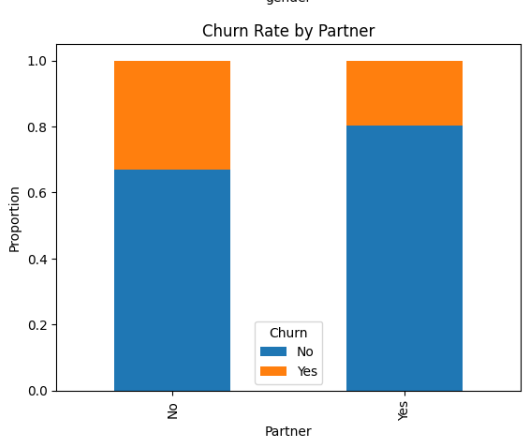
* **Churn rate by Online security :** Customers who **do not subscribe to online security services** show a noticeably higher churn rate. This suggests that **service bundling** may contribute to customer stickiness, as users who add security are more engaged with the platform



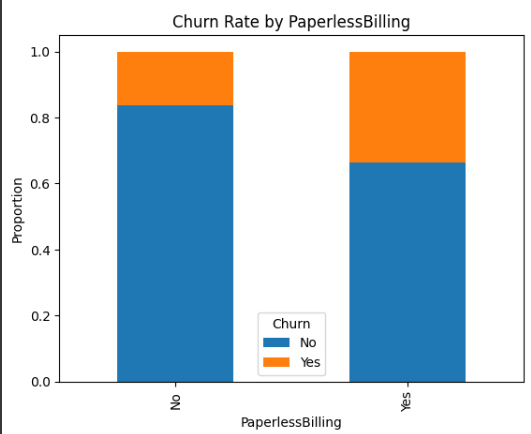
* **Churn rate by Dependents :** Churn appears to be **higher among customers without dependents**. This could indicate that customers with dependents (e.g., families) may be more stable or more likely to keep services for household needs. **Family-oriented customers may exhibit lower churn risk.**



* **Churn rate by Partner :** Customers who **do not have a partner** have a higher churn rate compared to those with partners. This pattern might suggest that **shared household services or commitments** lead to greater customer retention

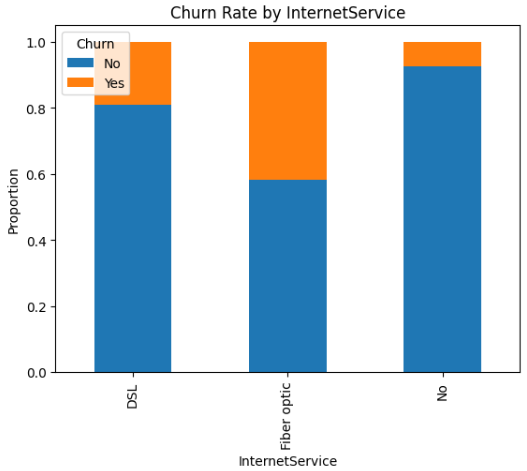


* **Churn rate by paperless billing:** Customers who **opt for paperless billing** exhibit a higher churn rate than those who receive physical bills. This may reflect a more digital-oriented segment that is also more price-conscious or disengaged. Alternatively, customers enrolled in paperless billing may have less frequent contact with the company, potentially reducing retention

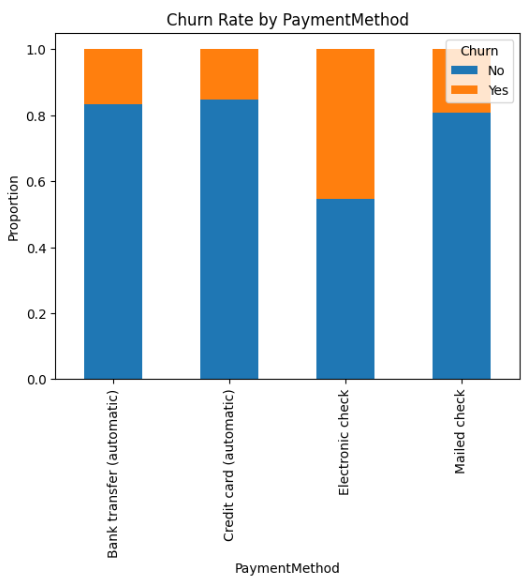


* **Churn rate by Internet service :** Churn rates vary by internet service type. Customers using **Fiber Optic** services tend to churn at a higher rate, potentially due to pricing or performance issues. In contrast, **DSL users show lower churn**, and those **without internet service** are the least likely to churn, possibly because they

use minimal services overall



* **Churn rate by payment method :** The churn rate differs across payment methods. Customers using **electronic check** tend to churn more than those using **credit cards or bank transfers**. This may reflect differences in customer trust, payment automation, or overall satisfaction.

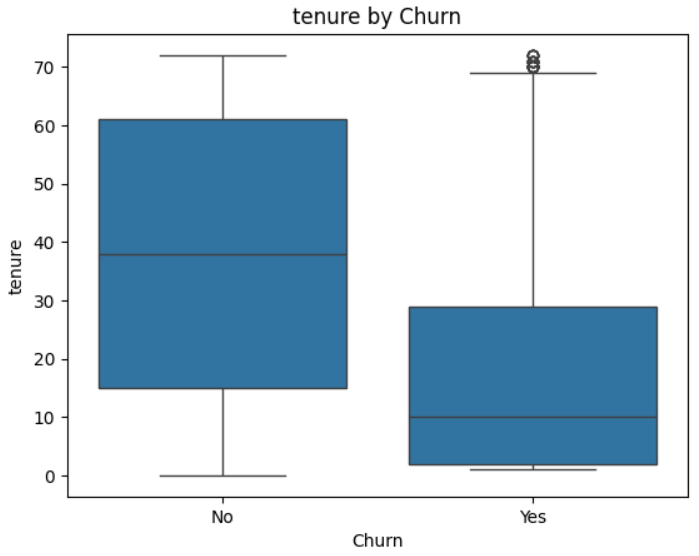


* Many binary service features were reviewed. Those that did not show notable churn rate differences were omitted from detailed visualizations.

#### 5.2.2 Churn vs Numerical Variables

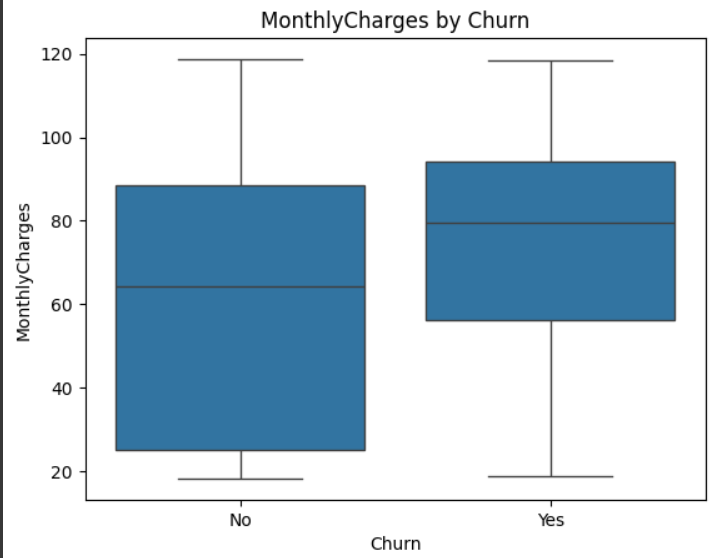
* **Tenure by churn:** The variable tenure, representing the number of months a customer has been with the company, shows a strong relationship with churn.
* For customers who **churned** , the **median tenure** is around **10 months**, and most exited within **30 months**.
* In contrast, **non-churned** customers have a **median tenure** of approximately **38 months**, with an interquartile range spanning **15 to 61 months**.
* A small number of outliers churned after long periods (close to 70 months), but these are rare.

This indicates that **churn is more common among newer customers**, emphasizing the need for early engagement and retention strategies.



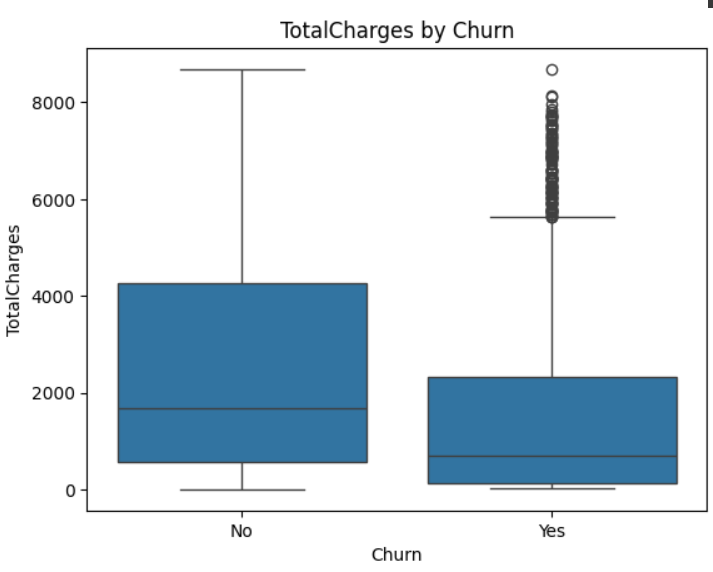
* **Monthly charges by churn:** There is a clear difference in MonthlyCharges between churned and retained customers.
* Churned customers have a **higher median monthly charge** of around **$80**, with most paying between **$60 and $90**.
* Non-churners have a **median of about $60**, with their distribution more evenly spread across the price range.

These findings suggest that **customers with higher bills may be more likely to churn**, possibly due to affordability concerns or perceived lack of value.



* **Total charges :** TotalCharges is notably lower among customers who have churned, since they have spent less time with the service.
* The **median total charge** for churned customers is approximately **$1,000**, with most paying **less than $2,000**.
* Non-churned customers show a **median of around $2,000**, and many have paid significantly more, including some exceeding **$6,000**.

This reinforces that **longer-tenure customers contribute more revenue**, underlining the financial impact of early churn.



#### 5.2.2 Correlation analysis

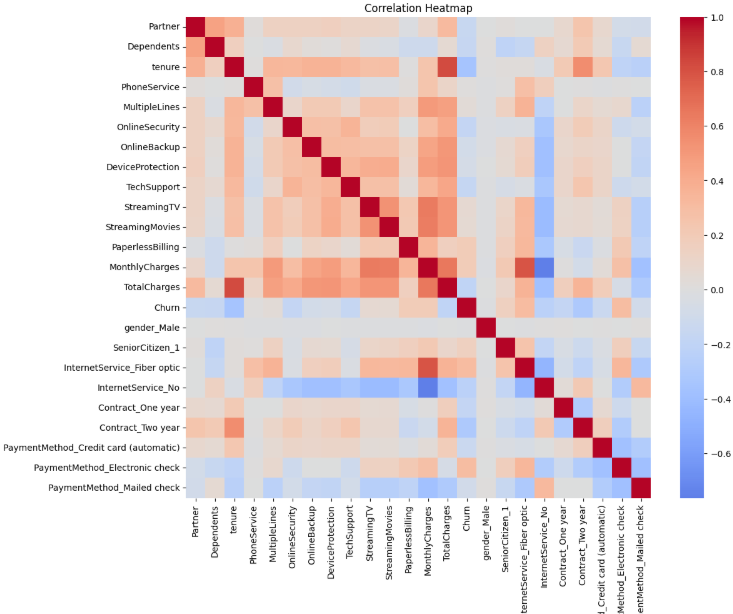
To identify meaningful relationships between variables, a **correlation matrix** was computed using the **fully encoded dataset**, including the binary target variable Churn (encoded as 1 for churned, 0 for retained).

Several features showed moderate to strong correlations with churn:

* Variables such as tenure, Contract\_TwoYear, OnlineSecurity\_Yes, and TechSupport\_Yes were **negatively correlated** with churn, indicating that longer service duration and access to certain services are associated with retention.
* In contrast, features like Contract\_Month-to-Month, PaperlessBilling\_Yes, and PaymentMethod\_ElectronicCheck were **positively correlated** with churn, suggesting these customer segments are more likely to leave.

Due to a **strong correlation (r ≈ 0.83)** between TotalCharges and tenure, the variable TotalCharges was **excluded** from further analysis to avoid redundancy and multicollinearity.

To streamline the modeling process, **feature selection** was performed by retaining only variables with an **absolute correlation of at least 0.05** with the target. This threshold balances interpretability and relevance, helping reduce noise from weakly related features while maintaining important predictors.



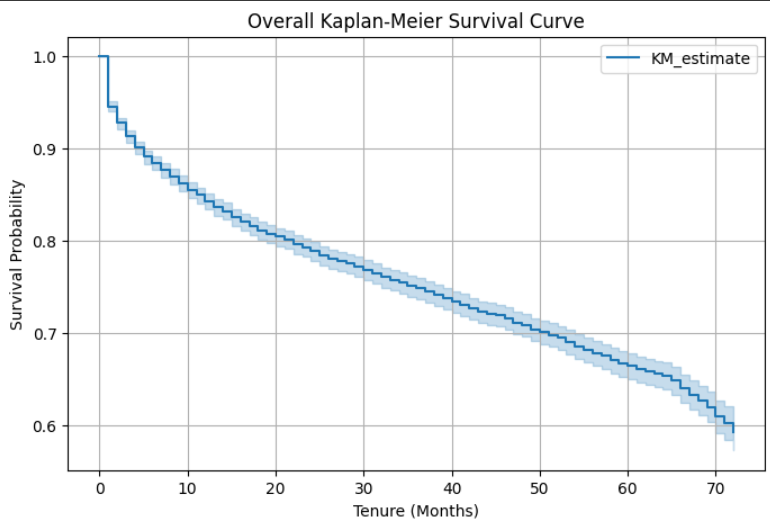
### 5.3 Survival-Based Exploratory Data Analysis

In this section, we explore the dataset from a **survival analysis perspective**, where the time-to-event is represented by tenure, and the event of interest is customer churn. Customers who have not churned are treated as **censored observations**. This step helps assess the distribution of survival times and how they vary across different customer segments

#### 5.3.1 Kaplan-Meier Survival Curve (Overall)

The curve shows a **sharp decline in survival probability during the first 20 months**, indicating that **most churn occurs early in the customer lifecycle**.

After this initial drop, the curve begins to **flatten out**, suggesting that customers who stay beyond this period are increasingly likely to remain. On average, a customer has around a 60% chance of still being with us after 70 months

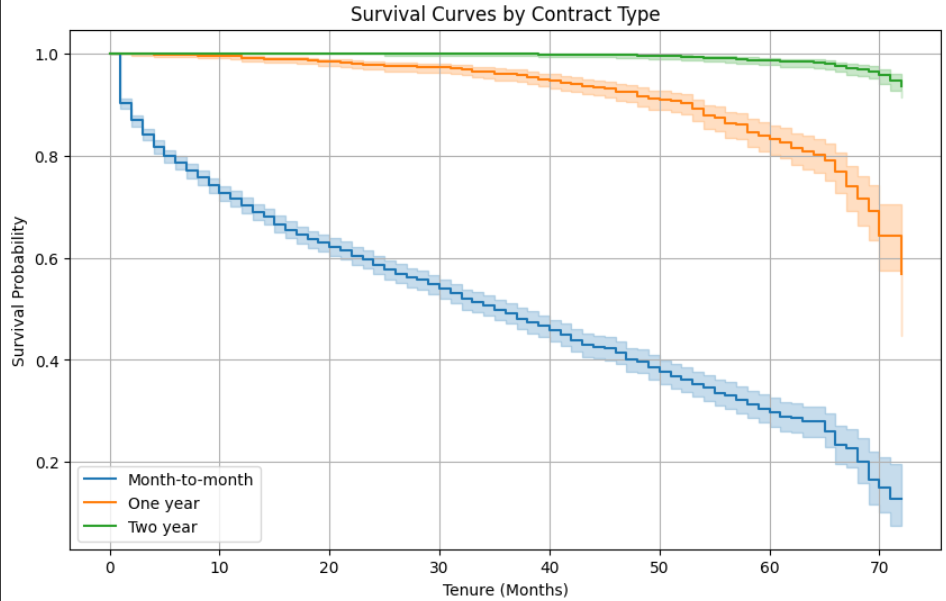


#### 5.3.2 Stratified Survival Curves by Key Categorical Variables

* **By Contract Type:**

The **log-rank test** yielded a **p-value < 0.05**, leading us to **reject the null hypothesis** that the survival functions for the different contract types are identical. This confirms that **contract type has a statistically significant impact on customer retention**.

As illustrated by the **Kaplan-Meier curves**, customers on **Month-to-Month contracts** experience a much **steeper decline in survival probability**, especially in the first 20 months. This indicates a **faster and higher churn rate** compared to those on longer-term plans. In contrast, customers with **Two-Year contracts** show the **highest retention**, with over **90% still active after 60 months**. This highlights the role of long-term contracts in **significantly reducing churn** and encouraging longer customer lifetimes

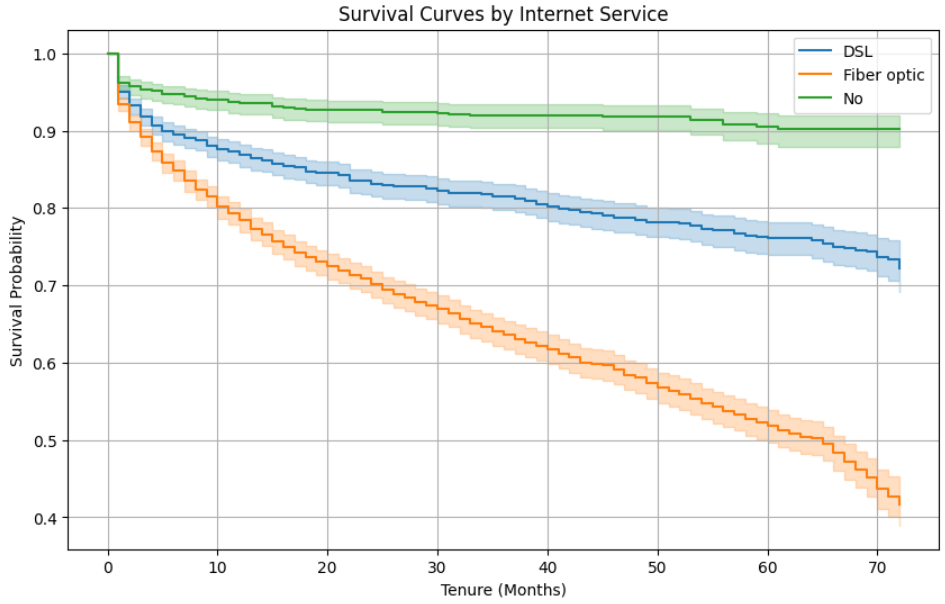


* **By Internet Service**

The **log-rank test** yielded a **p-value < 0.05**, indicating a **statistically significant difference** in survival curves across the different internet service types. This allows us to **reject the null hypothesis** that all groups share the same survival function.

According to the **Kaplan-Meier curves**, customers with **Fiber Optic internet** show a **steep decline in survival probability**, reflecting a **faster and higher churn rate** compared to other service types.

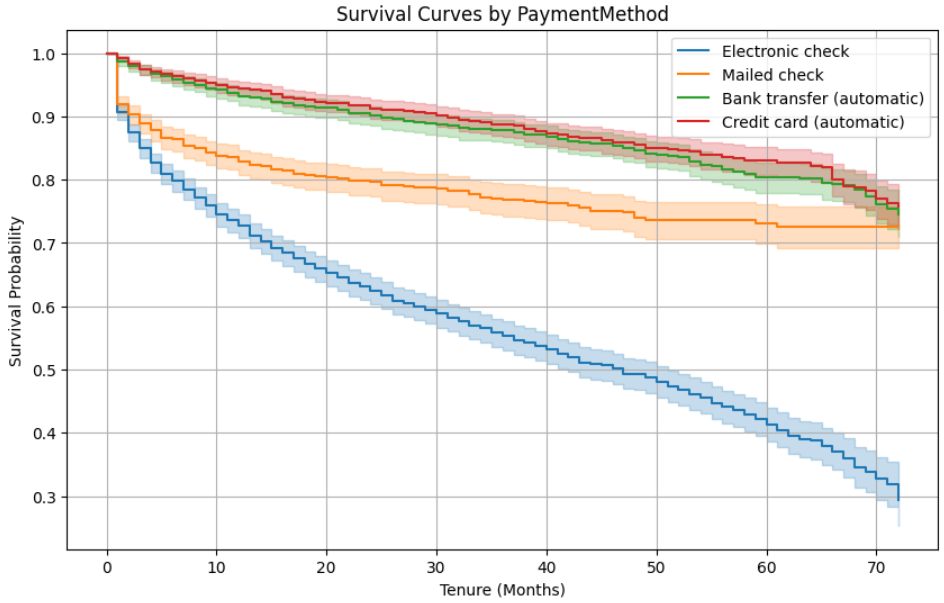
In contrast, customers with **no internet service** demonstrate the **highest retention**, with over **90% remaining active after 70 months**. This suggests that customers not using internet services are less likely to churn, possibly due to simpler or more stable service needs.



* **By Payment Method**

The **log-rank test** indicates that **Electronic Check** and **Mailed Check** users have significantly different survival patterns compared to other groups (**p-value < 0.05**). However, the survival curves for **Bank Transfer** and **Credit Card** users are statistically similar (**p-value > 0.05**), meaning we **fail to reject the null hypothesis** for those two groups.

The **Kaplan-Meier curves** show that customers who pay via **Electronic Check** experience a **steep decline in survival probability**, suggesting a **higher and faster churn rate** among this segment. On the other hand, customers who use **Bank Transfer** or **Credit Card** methods exhibit **much higher retention**, with over **70% still active after 70 months**, indicating these payment types may be associated with more stable, long-term customers.



## Modeling

### 6.1 Classification model

The objective of this classification model is to predict churn based on features extracted from teleco customer churn dataset . The goal is to accurately classify churners versus non-churners using supervised machine learning techniques.

#### 6.1.1 Considered Algorithms

We began the classification modeling by training a **Logistic Regression model** as our baseline, given that our target variable is categorical. Training was conducted on the **original imbalanced dataset**, as revealed during EDA, where non-churners significantly outnumber churners. This baseline helps establish a reference point to assess the impact of balancing techniques and more advanced models.

To identify the most effective model for our classification task, we tested and compared four additional algorithms: Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost. These models vary in complexity, interpretability, and ability to capture non-linear patterns, making them valuable for a comprehensive comparison.

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | Description | Reasons to use | Strengths for this project |
| Logistic Regression | Predicts probabilities for binary outcomes | Simple ,interpretable, and easy to implement | Interpretable, strong baseline for churn detection |
| SVM ( Support Vector Machine) | Finds the optimal hyperplane for classification | High-dimensional space suitability , robust for classification and regression | Robust to overfitting, good churn recall, can separate churners in complex spaces |
| Decision Tree | Splits data into branches for decision making | Easy to visualize , handles both numerical and categorical | Visualizable, interpretable, helps communicate churn drivers clearly |
| Random Forest | Constructs multiple decision trees for robust prediction | Reduces overfitting, handles large datasets well | Better generalization than a single tree, handles class imbalance reasonably well |
| XGBoost | Builds models in a sequential way, where each new tree corrects the errors of the previous ones. | High accuracy, efficient, built-in imbalance control | Often yields top results in churn problems, supports boosting weak signals in complex data |

#### 6.1.2 Handling class imbalance

During the exploratory data analysis (EDA), we identified a significant class imbalance, with non-churners constituting the majority class and churners the minority. Addressing this imbalance is crucial because models trained on imbalanced data can achieve deceptively high accuracy by favoring the majority class, yet often fail to correctly identify minority class instances. This leads to biased predictions and overlooked critical insights.

We employed two separate balancing strategies based on the specific algorithms. Although tree-based models like Random Forest and XGBoost can naturally handle class imbalance to some extent, training them directly on imbalanced data may still result in poor performance on the minority class, especially in recall and AUC. In our case, the class distribution was 73% to 27%, which is considered moderate imbalance. We found that applying SMOTE prior to training improved the models’ ability to correctly identify minority class instances. Therefore, we chose to oversample the minority class using SMOTE specifically for the tree-based models.

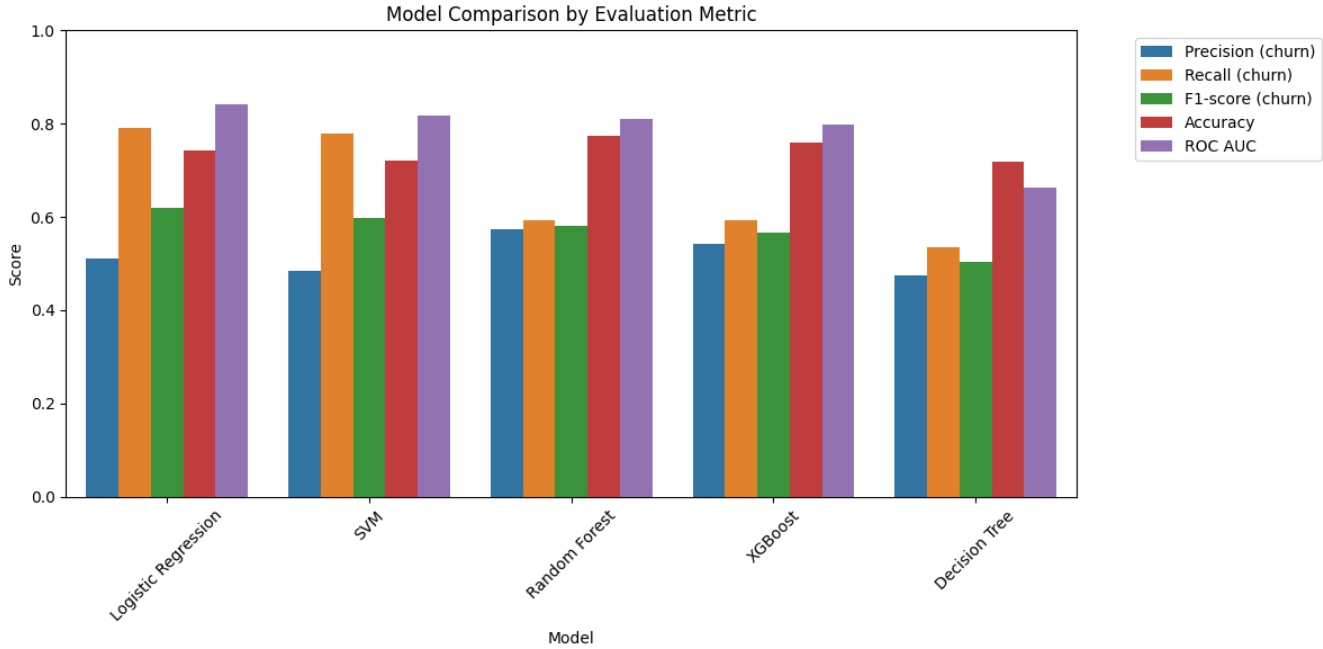
For models such as Logistic Regression and Support Vector Machine (SVM), we applied class\_weight='balanced' to compensate for class imbalance. This approach modifies the loss function to penalize misclassification of minority class instances more heavily. Unlike oversampling techniques such as SMOTE, class weighting does not introduce synthetic data, reducing the risk of overfitting and maintaining the original data distribution.

#### 6.1.3 Algorithm selection

Following model training on the balanced dataset, we compared their performance across several key classification metrics. The chart below presents a side-by-side comparison of the five classification models based on key evaluation metrics: precision, recall, F1-score (focused on the churn class), accuracy, and ROC AUC. These metrics are also summarized numerically in the accompanying table.

* **Logistic Regression** achieved the **highest recall with 0.791** and **ROC AUC of 0.841**, indicating its strong ability to detect churners and distinguish between classes. Its **F1-score of 0.620** further reflects a good balance between precision and recall.
* **Random Forest** outperformed all models in **precision with 0.573** and **accuracy of 0.775** , making it highly reliable when it predicts churn, though it had slightly lower recall.
* **XGBoost** showed **moderate yet balanced performance** across all metrics, while **SVM** reached **similar recall to logistic regression (0.779)** but had lower precision and ROC AUC.
* **Decision Tree** performed the weakest overall, with the **lowest recall with 0.535** and **ROC AUC of 0.662** , indicating difficulty in identifying churners effectively.

In summary, **Logistic Regression** and **Random Forest** emerged as the top-performing models, with the choice between them depending on whether the priority is **maximizing recall** (detecting more churners) or **increasing precision** (reducing false positives).



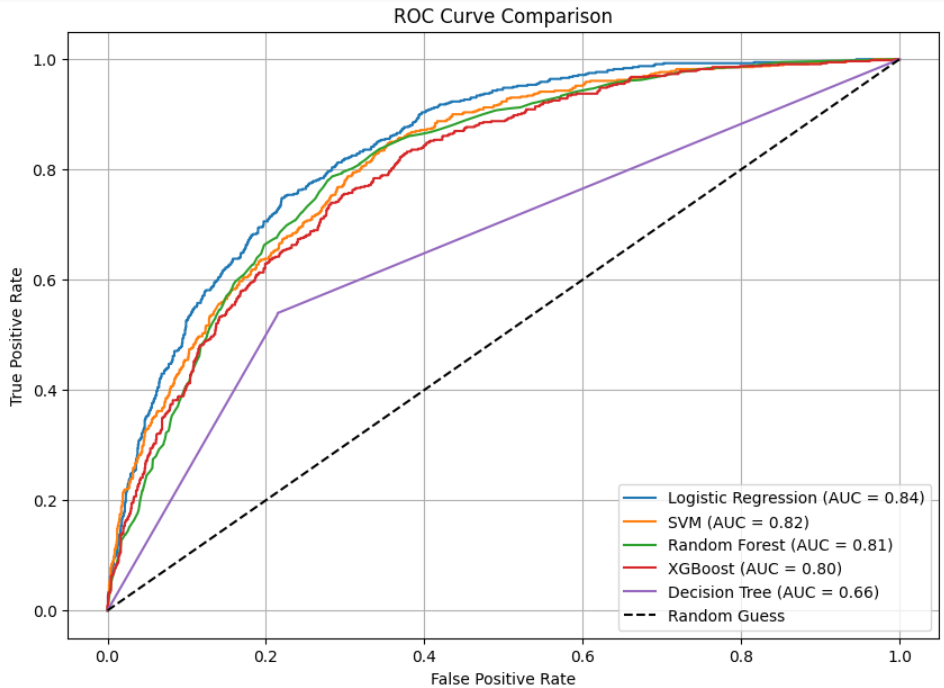
In addition to the classification metrics, the ROC (Receiver Operating Characteristic) curves were plotted to evaluate the models' ability to distinguish between churners and non-churners across different classification thresholds.

A model with a **higher ROC curve** and a **larger Area Under the Curve (AUC)** indicates better discriminative power.

As shown in the ROC plot, **Logistic Regression** consistently demonstrated the strongest performance, with an **AUC of 0.841**, followed closely by **Random Forest (0.810)** and **SVM (0.817)**.

**XGBoost** showed moderate separation ability (**AUC = 0.798**), while the **Decision Tree lagged behind with the lowest AUC (0.662)**, indicating limited capability to separate churners from non-churners effectively.

Overall, the ROC curves support the findings from the evaluation metrics: **Logistic Regression** has the most reliable performance in distinguishing between classes, making it particularly effective when decision thresholds need to be adjusted (e.g., to increase recall or precision based on business needs).



#### 6.1.4 Selected model vs Baseline model

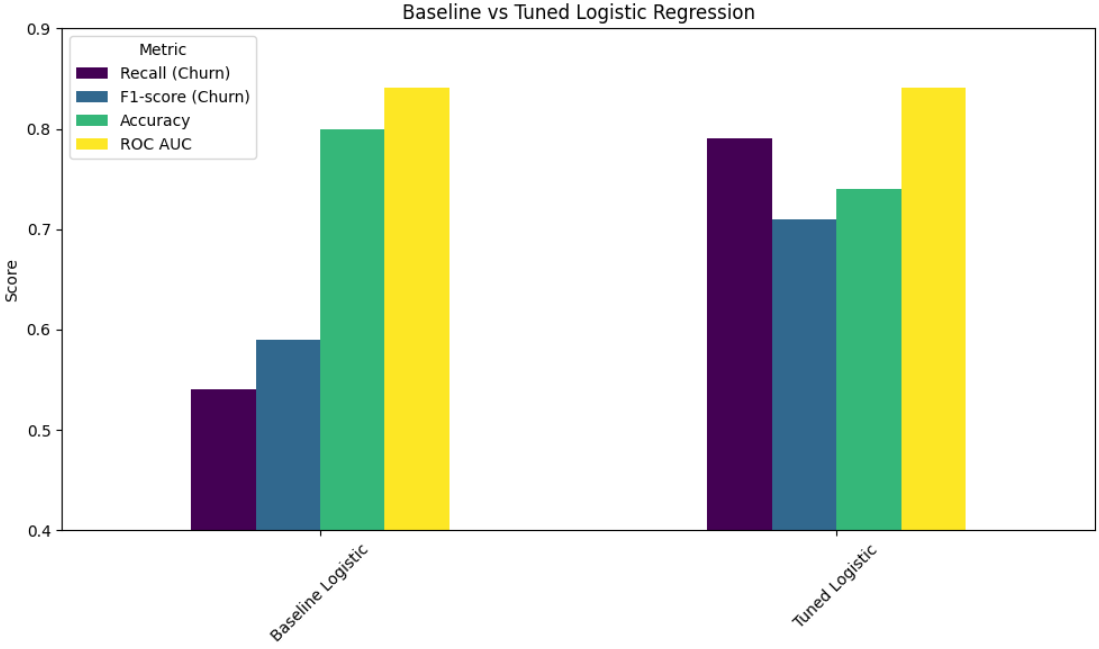
The histogram below compares the baseline model to the chosen model across key classification metrics.

While the **baseline model** achieved a higher overall **accuracy of 0.80**, this came at the expense of correctly identifying churners, with a relatively low **recall of 0.54**. In contrast, the **chosen model** prioritized recall, achieving a significantly higher **recall of 0.79**, making it more effective at identifying customers at risk of churning.

Although this gain in recall led to a slight drop in **precision by 0.14** indicating more false positives , the **F1-score** for the churn class improved from **0.59 to 0.62**, reflecting a better overall trade-off. Both models achieved similar **ROC AUC scores of 0.84**, indicating comparable ability to distinguish between churners and non-churners across thresholds.

Additionally, to ensure the robustness of our chosen model, we applied **5-fold cross-validation** during hyperparameter tuning. This helped mitigate the risk of overfitting and ensured that the model’s performance was consistent across different subsets of the training data.

Overall, despite a slightly lower accuracy, the chosen model offers **greater business value** by substantially improving churn detection—a critical factor in customer retention strategies



### 6.2 Survival model

The objective of this survival model is to predict the **time until customer churn** using features from the Telco Customer Churn dataset. This model provides a time-to-event prediction, offering insights into **when** churn is likely to occur. This enables more timely and targeted customer retention strategies.

#### 6.2.1 Modeling Approach

For predicting the time until customer churn, we employed the **Cox Proportional Hazards (CoxPH) model**, a semi-parametric survival analysis method widely used in time-to-event modeling. The Cox model estimates the effect of multiple covariates on the hazard, or risk, of the event (churn) occurring at any point in time, without requiring assumptions about the baseline hazard function.

This approach is well-suited for our data, as it can handle censored observations—customers who have not yet churned—and provides interpretable hazard ratios that quantify the impact of each feature on churn risk. The model assumes proportional hazards, meaning the effect of the covariates on the hazard is constant over time.

The CoxPH model was implemented using the lifelines Python library, which fits the model via maximum partial likelihood estimation.

#### 6.2.2 Model Fitting and Evaluation

The Cox Proportional Hazards model was fitted using the training dataset, with the tenure column as the duration variable and the churn column as the event indicator. The fitting process used **maximum partial likelihood estimation**, which estimates the effects of covariates by focusing on the order of observed events rather than their exact timing. This approach allows estimation of regression coefficients without specifying the baseline hazard function, making the Cox model flexible and capable of handling censored data effectively.

To assess model performance, we calculated the **concordance index (C-index)**, which measures the model’s ability to correctly rank survival times based on predicted risk scores. A C-index closer to 1 indicates better predictive discrimination.

The model achieves 0.93 C-index on training and 0.75 on held-out test data. This indicates some overfitting but still acceptable predictive performance. For deployment, the simpler unpenalized Cox model was retained for stability.

Prior to evaluating model performance, the proportional hazards assumption was checked and addressed in both the training and test datasets. As part of this process, numeric features were scaled to ensure consistent variable ranges to reduce non-linearity and improve model fit. These preprocessing steps helped satisfy model assumptions and contributed to a reliable estimation of the concordance index (C-index).

## Model deployment

Following the training and evaluation phase, both the classification and survival models were deployed within a Flask-based web application. The purpose of this deployment was to provide business users with an interactive decision-support tool, enabling them to explore predictive insights and take timely actions to reduce churn.

### 7.1 Data input and Processing

The application accepts customer data in the form of a CSV file. Uploaded datasets are validated to ensure that all required fields are present (e.g., tenure, MonthlyCharges, Contract, InternetService, PaymentMethod). Once uploaded, the data undergoes preprocessing: categorical variables are encoded into numerical features, binary responses are converted into 0/1 format, missing values such as *TotalCharges* are reconstructed, and numerical attributes are standardized before being passed to the survival model.

### 7.2 Dashboard Design

The web application is structured into four main tabs, each serving a specific analytical purpose:

* **Overview**: Provides a high-level summary of churn trends, revenue at risk, average customer lifetime, and the distribution of customers across risk categories. It also highlights projected financial losses and potential savings under different retention strategies.
* **Classification**: Presents churn predictions in binary form (*churn vs. no churn*). This tab includes customer-level churn probabilities, segmentation into low, medium, and high-risk groups, and the identification of top churn drivers through feature importance analysis.
* **Survival**: Extends the analysis by focusing on the timing of churn. Using the Cox Proportional Hazards model, this tab displays survival probability curves, hazard rates, and median survival time. It also identifies critical intervention windows, allowing the business to prioritize customers likely to churn in the short term.
* **Recommendations**: Translates predictive insights into actionable strategies. This tab provides a customer-level table containing identifiers, churn probabilities, survival metrics, and tailored retention recommendations. Recommendations are prioritized by urgency and adapted to customer profiles (e.g., contract type, service usage, payment method).

### 7.3 Technical Implementation

From a technical perspective, the deployment was carried out using the Flask framework, with Jinja2 for templating and Matplotlib for visualization. The models were serialized using joblib and stored as separate files:

* churn\_model.pkl for the classification model
* cox\_model.pkl for the survival model
* scaler.pkl for feature scaling

Each tab of the dashboard corresponds to a dedicated Flask route (/overview, /classification, /survival, /recommendations). Upon uploading data, the predictions are computed once and cached, ensuring consistent outputs across all views. Visualizations such as churn distributions, feature importance plots, and survival curves are dynamically generated and embedded into the interface.

The recommendation engine combines model outputs with business rules to generate tailored strategies. For example, high-risk customers may trigger immediate retention calls with discounts, while medium-risk customers may receive loyalty offers or personalized communication. Service type, contract type, and payment method are also incorporated into the decision logic, ensuring that recommendations are both personalized and actionable.

### 7.4 Business Value

By integrating both classification and survival models into a single platform, the web application provides an end-to-end solution for churn management. Executives gain a strategic overview of churn risk, analysts can explore churn drivers in depth, managers benefit from time-to-churn insights, and frontline teams receive customer-level retention strategies. This deployment ensures that advanced analytics are translated into practical, business-ready outcomes.

## Refrences

* [Churn Reduction: Data-Driven Telecom BI Strategy | Bold BI](https://www.boldbi.com/resources/blog/bi-strategy-for-telecom-churn-reduction/)
* [(PDF) “Churn Prediction in Telecommunication Industry”](https://www.researchgate.net/publication/369783546_Churn_Prediction_in_Telecommunication_Industry)
* [Machine Learning Algorithm: When to Use Which One in 2025 | Label Your Data](https://labelyourdata.com/articles/how-to-choose-a-machine-learning-algorithm)