الجامعه المصرية للتعلم االلكترونى الاهلية

كلية الحاسبات والمعلومات

# Predict Customer Churn

## Using Machine Learning

By

| Name | | Id |
|---|---|---|
| Kareem Gamal Rady | (Team leader) | 2000722 |
| Yostina Samy Dawood | | 2000384 |
| Yostina Maged Nassef | | 2000713 |
| Youssef Tharwat Wadea | | 2000441 |
| Aya Mohammed Abdelnaby | | 2000896 |
| Suhaib Emad Mohamed | | 2001546 |
| Shrouk abdelal Ali | | 2001312 |

Supervised By:

Prof. Ahmed Ezz

Eng. Eman Rizk

" This document is presented as partial fulfillment of the requirements for Bachelor of Information Technology degree "
EELU- Assiut 2024

# Acknowledgment

In the name of Allah, the Most Gracious and the Most Merciful Alhamdulillah, all praise Allah for the strengths and His blessing in completing this Project.

In performing our project, we had to take the help and guideline of some respected persons, who deserve our greatest gratitude. The completion of this project gives us much Pleasure. We would like to present our supervisors:

## Dr. Kamal Hamza

Computer and Information Technology Program Director – EELU

## Dr. Ahmed Ezz

Assistant Professor of Computer Engineering and Information Technology –EELU

## Eng. Eman Rizk

Faculty of Computers and Information Technology- EELU

For giving us a good guideline for the Project throughout numerous consultations.

In addition, special thanks to Eng. Eman Rizk for introducing us to the methodology of work, and whose passion for the “underlying structures” had a lasting effect. We also thank our university EELU for consent to include copyrighted pictures as a part of our graduation project.

We would also like to expand our deepest gratitude to all those who have directly and indirectly guided us in doing this project, especially our classmates and team members themselves who have made valuable comment suggestions on this proposal and inspired us then to improve our project.

# Abstract

Customer churn is a critical concern for the telecommunication industry.

Understanding and predicting customer churn can lead to more effective retention strategies and an increase in profitability. Predicting customer churn allows telecommunication companies to identify potentially dissatisfied customers early on and take proactive measures to retain them. Due to a large client base, the telecom industry generates a large volume of data on a daily basis. Decision makers and business analysts stressed that acquiring new customers is more expensive than retaining existing ones. Business analysts and customer relationship management (CRM) analysts must understand the reasons for customer churn as well as behaviour patterns from existing churn data. This paper proposes a churn prediction model that uses classication and clustering techniques to identify churn customers and provides the factors that contribute to customer churning in the telecom sector. The results presented shows that X Boost and Random Forest achieved higher prediction accuracy when compared to K-Nearest Neighbours, Support Vector Machines and Decision Trees in terms of accuracy, precision, F1-Score and recall

.

# Table of content

**Acknowledgment** .......................................................................................... 2

**Abstract** ........................................................................................................ 4

**Chapter 1: (Introduction)** ...................................................................... **7**

1.1 Introduction.......................................................................................... 8
1.2 Problem definition ............................................................................... 13
1.3 project objectives..................................................................................14
1.4 Motivation .............................................................................................15
1.5 Existing System ....................................................................................16
1.6 Literature Review .................................................................................16

**Chapter 2: (Four Level Analytics)**............................................................ **18**

2.1 Descriptive Analytics ......................................................................... 19
2.2 Diagnostic Analytics..............................................................................19
2.3 Predictive Analytics ..............................................................................19
2.4 Prescriptive Analytics.......................................................................... 20

**Chapter 3: (Participating Technology)…**...................................................**21**

3.1 Using Orange Data Mining ..................................................................22
3.2 Overview of Orange Data Mining...................................................... 22
3.3 Proposed System ..................................................................................22.
3.4 Implementation with Orange............................................................... 23
3.5 Benefits of the Proposed System......................................................... 24

**Chapter 4: (Project Methodology)**........................................................... **26**

4.1 Data understanding...............................................................................27
4.2 Feature Statistics ( Descriptive analysis)............................................ 31
4.3 Pivot Table ............................................................................................33
4.4 Box Plot (Diagnostic Analytics ) .........................................................34
4.5 The Relationship Between Monthly Charges And Churn.................. 36
4.6 The Relationship Between Contract& Churn......................................38
4.7 The Relationship Between Dependents & Churn.................................40
4.8 The Relationship Between Partner & Churn ...................................... 41
4.9 The Relationship Between Payment Method & Churn........................ 43

4.10 The Relationship between Online Security & Churn.................................. 45
4.11 The Relationship between Online Backup & Churn................................... 47
4.12 The Relationship between Phone Service & Churn................................... 49
4.13 The Relationship between Multiple Line & Churn..................................... 51
4.14 Relationship Between Tenure & Churn Analysis Using Box Plot .....................................................52

**Chapter 5: (Application of Cox Regression and Kaplan-Meier in Customer Churn Management)................................................................................................................81**

5.1 Utilizing Cox Regression for Churn Prediction............................................83
5.2 Applying Kaplan-Meier for Customer Retention Analysis........................ 83
5.3 Integration and Implementation...................................................................84
5.4 Case Study: Orange Data Mining Project...................................................84

**Chapter 6:(Future Work and Development of a Customer Churn Prediction Application....................................................................................................85**

6.1 Key Features of the Application ................................................................ 87
6.2 Conclusion.................................................................................................. 88

**Chapter 7:(Conclusion and Discussion of Results) .................................................89**

7.1 Strategic Recommendations........................................................................ 90.
7.2 Kaplan-Meier Estimator...............................................................................91

**Chapter 8: (References)… ...........................................................................94**

8.1 references ....................................................................................95

# Chapter 1
# Introduction

## 1.1 Introduction

Customer retention is one of the primary growth pillars for products with a subscription-based business model. Competition is tough in markets where customers are free to choose from plenty of providers even within one product category. Several bad experiences – or even one – and a customer may quit. And if droves of unsatisfied customers churn at a clip, both material losses and damage to reputation would be enormous. Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided. Churn rate is a health indicator for businesses whose customers are subscribers and paying for services on a recurring basis. Customers (of subscription-driven businesses) opt for a product or a service for a particular period, which can be rather short – say, a month. Thus, a customer stays open for more interesting or advantageous offers. Plus, each time their current commitment ends, customers have a chance to reconsider and choose not to continue with the company. Some natural churn is inevitable, and the figure differs from industry to industry. But having a higher churn figure than that is a definite sign that a business is doing something wrong .Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base that is most vulnerable to churn.

The customer retention system is divided into churn problem and presenting offers. This project focuses on the churn problem, as presenting offers through Survival analysis.

Survival analysis in the context of telecommunications plays a crucial role in understanding customer interactions with services. It helps identify factors influencing their continuity in using services or engaging with the company. Here are some reasons highlighting the importance of survival analysis for customers in telecommunications:

Customer Satisfaction Assessment:

Survival analysis allows examining the extent of customer satisfaction when using telecommunications services. If customers continue to stay, it may reflect their high satisfaction with the provided services.

Identifying Influencing Factors:

Survival analysis helps identify factors significantly affecting customer continuity, whether due to service quality, pricing, or the availability of additional services.

Improving Customer Experience:

Results from survival analysis can be used to enhance the customer experience, such as improving the user interface or providing additional services that meet customer needs.

Enhancing Retention Strategies:

Knowing the influencing factors allows the development of effective strategies to retain customers, whether through special offers or improving customer service.

Guiding Marketing Strategies:

Survival analysis can guide marketing strategies better, targeting customers who might be at risk of churn based on factors influencing retention.

Achieving More Revenue:

Satisfied and retained customers can be a continuous source of revenue. Therefore, focusing on retention can contribute to achieving more revenue in the long term.

Providing offers to customers to retain them and prevent churn is a crucial aspect of customer retention strategies. Here are some offers and plans that companies can provide to their customers to keep them engaged:

1. **Discounts and Special Offers**

- Exclusive Discounts: Offer special discounts to loyal customers or those who have shown signs of leaving.
- Temporary Promotions: Implement limited-time promotions such as "Buy One, Get One Free" deals or percentage discounts.

2. **Loyalty Programs and Rewards**

- Reward Points: Grant reward points that can be redeemed for free products or services.
- Loyalty Program Memberships: Provide additional benefits to members of loyalty programs, such as priority service or exclusive offers.

3. **Enhanced Customer Service**

- Excellent Technical Support: Provide quick and helpful technical support to increase customer satisfaction.
- Listening to Customer Suggestions: Take customer feedback seriously and improve services based on their suggestions.

4. **Personalization and Customization**

- Personalized Offers: Deliver personalized offers based on purchasing behavior and customer

preferences.
- Personalized Thank You Messages: Send personalized thank you messages to maintain customer loyalty.

5. **Improving Product or Service Quality**
- Continuous Updates and Improvements: Ensure that products or services remain advanced and meet changing customer needs.
- Service Guarantees: Offer guarantees for products or services to build trust.

6. **Renewal and Upgrade Offers**
- Renewal Incentives: Provide incentives for customers to renew their subscriptions, such as early renewal discounts.
- Service Upgrades: Offer attractive prices for upgrading to higher-quality services.

7. **Free Trials and Demos**
- Free Trials: Provide free trial periods for new services or products to encourage customers to stay.
- Demo Versions: Offer demo versions for potential customers to try before purchasing.

8. **Personal and Direct Support**
- Personal Meetings: Arrange personal meetings or direct calls with customers to discuss their needs and provide appropriate solutions.
- Training and Support: Offer training sessions to help customers get the most out of products or services.

By implementing these strategies, companies can increase customer satisfaction and reduce churn rates. It's also important to continuously analyze customer data to understand their needs and provide offers that match them.

## 1.2 Problem definition

Managing customer churn is one major challenge facing companies, especially those that offer subscription based services. Customer churn (or customer attrition) is basically the loss of customers, and it is caused by a change in taste, lack of proper customer relationship strategy, change of residence and several other reasons .If businesses can effectively predict customer attrition, they can segment those customers that are highly likely to churn and provide better services to them. Hence, a churn prediction model is developed in this project that uses machine learning techniques such as Logistic Regression, Decision Trees, K-Nearest Neighbors and Support Vector Machine algorithms to assist companies in predicting customers who are most likely to churn. In this way, they can achieve a high customer retention rate and maximize their revenue.

## 1.3 Project objectives

1. Personalization:

   - Customize marketing, recommendations, and services for individual customers.

2. Customer Retention:

   - Identify and retain customers at risk of leaving.

3. Customer Acquisition:

   - Target potential high-value customers effectively.

4. Sales Forecasting:

   - Predict future sales to optimize operations.

5. Customer Lifetime Value (CLV):

   - Estimate the future value of customers for better investment decisions.

6. Fraud Detection:

   - Detect and prevent fraudulent activities.

7. Segmentation:

   - Group customers into segments for targeted marketing.

8. Enhanced Customer Experience:

   - Improve customer satisfaction by anticipating needs.

9. Operational Efficiency:

   - Streamline operations by predicting demand and optimizing resources.

10. Risk Management:

    -Assess and mitigate financial and credit risks.

## 1.4 Motivation

There are many things companies may do wrong, from complicated onboarding when customers aren’t given easy-to-understand information about product usage and its capabilities to poor communication, e.g. the lack of feedback or delayed answers to queries. Even long time clients may feel unappreciated because they don’t get as many bonuses as the new ones. In general, it’s the overall customer experience that defines brand perception and influences how customers recognize value for money of products or services they use. Hence, this project uses churn prediction models to predict customer churn so that the above shortcomings can be overcome for any potential client by assessing their propensity of risk to churn.

## 1.5 Existing System

Many approaches were applied to predict churn in telecom companies. Most of these approaches have used machine learning and data mining. For example, a churn prediction model for prepaid customers was developed in telecom using fuzzy classifiers, Neural Networks, SVM Classifier, Ada Boost & RF techniques, which were compared with a fuzzy nearest-neighbor classifier to predict an accurate set of churners on a real-time dataset of prepaid telecom customers from south Asia. Another model utilized a CRM frame work using neural network and data mining for the prediction of customer behavior in banking. An algorithm was also developed based on click stream data of a website to extract information and tested the predictive power of the model based on data such as number of clicks, repeated visits, repetitive purchases, etc. Nonetheless, these models raised a few concerns which are to be addressed. The main drawbacks of existing systems include:

Most of them were suited only for applying suitable model and taking inference from predictions.None of them focused on the attributes crucial towards customer churn.Focus was more towards comparison rather than attributes determination.

## 1.6 Literature Review

Customer churn refers to the loss of customers over time. Machine learning offers powerful tools for predicting churn, helping businesses to retain customers more effectively.

**Key Concepts**

- Churn Definition: Varies across industries but generally means customers stop using a product or service.

- Features and Data: Key factors include demographics, transaction history, usage patterns, and customer service interactions.

**Machine Learning Techniques**

1. Supervised Learning Models:

   - Logistic Regression: Simple and interpretable.

   - Decision Trees and Random Forests: Handle non-linear relationships well.

- Support Vector Machines (SVM): Effective in high-dimensional spaces.
- Neural Networks: Model complex relationships.

2. Unsupervised Learning Models:

- Clustering: Identifies patterns without labeled data.

3. Ensemble Methods:

- Gradient Boosting Machines (GBM) and XGBoost: Improve accuracy by combining multiple models.

**Advanced Techniques**

- Deep Learning: Uses CNNs and LSTMs to capture temporal dependencies.
- Survival Analysis: Estimates time until churn.
- Natural Language Processing (NLP): Analyzes text data from customer interactions.

**Evaluation Metrics**

- Accuracy, Precision, Recall: Key metrics for model performance.
- ROC-AUC: Measures model's ability to distinguish between churn and non-churn.

**Applications**

- Telecommunications: Analyzing call patterns and service usage.
- E-commerce: Using purchase history and browsing behavior.
- Banking: Based on transaction history and product usage.

**Challenges and Future Directions**

- Data Quality and Integration: Essential for accurate predictions.
- Model Interpretability: Understanding why customers churn.
- Real-time Prediction: Enables immediate intervention.

**Conclusion**

Machine learning enhances customer churn prediction, providing valuable insights for retention strategies. Future research should focus on better data integration, model interpretability, and real-time prediction capabilitie

# Chapter 2
# Four Levels Analytics

### Descriptive Analytics :

Descriptive analytics is the process of using current and historical data to identify trends and relationships. It’s sometimes called the simplest form of data analysis because it describes trends and relationships but doesn’t dig deeper.
Descriptive analytics, or business intelligence, uses historical information to answer the question “What Happened?”

### Uses of Descriptive Analytics:

1. **Understanding Past Performance**: By reviewing historical data, companies can understand how their strategies performed in the past.
2. **Identifying Patterns**: Descriptive analytics can reveal recurring patterns that might not be immediately obvious.
3. **Generating Periodic Reports**: It can be used to create monthly or annual performance reports.
4. **Facilitating Decision-Making**: Clear data visualization helps decision-makers understand the current state of their business easily

### Diagnostic Analytics :

Diagnostic analytics is a form of data analytics that examines data or content to answer the question, “Why did it happen?” It is characterized by techniques such as drill-down, data discovery, data mining, and correlations. It’s used to identify behaviors, trends, and patterns to figure out why certain outcomes have occurred.

### Uses of Diagnostic Analytics:

1. **Identifying Root Causes**: When a specific issue arises, diagnostic analytics can help determine the root cause of the problem.
2. **Analyzing Poor Performance**: If there is a decline in a company’s performance, diagnostic analytics can identify potential reasons.
3. **Data Exploration**: Techniques like drill-down and data discovery help in understanding data deeply and identifying hidden correlations and relationships.
4. **Highlighting Opportunities**: Diagnostic analytics can uncover unexpected improvement opportunities by understanding the underlying causes of data trends.

### Predictive Analytics :

The term predictive analytics refers to the use of statistics and modeling techniques to make predictions about future outcomes and performance. Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events. Predictive analysis can also be used to improve operational efficiencies and reduce risk.

**Uses of Predictive Analytics:**

1. **Forecasting Future Trends**: By identifying patterns in historical data, predictive analytics can forecast future trends and behaviors.
2. **Risk Management**: Predictive models can help in identifying potential risks and taking preventive measures.
3. **Optimizing Resources**: Companies can allocate resources more effectively by predicting future demands.
4. **Improving Customer Insights**: Predictive analytics can help understand customer behavior and predict future actions, enhancing customer relationship management.

**Prescriptive Analytics:**

Prescriptive analytics that focuses on finding the ideal way forward or action necessary for a particular scenario, based on data. Prescriptive analytics uses both descriptive and predictive analytics but the focus here remains on actionable insights rather than data monitoring**.**

**Uses of Prescriptive Analytics:**

1. **Decision Support**: It provides recommendations for actions to achieve desired outcomes.
2. **Optimization**: Helps in optimizing processes and strategies to maximize efficiency and effectiveness.
3. **Scenario Analysis**: Evaluates different scenarios and prescribes the best course of action.
4. **Automated Decision-Making**: In some advanced applications, prescriptive analytics can automate decision-making processes, leading to faster and more accurate actions.

These analytics techniques collectively empower businesses to make informed decisions, optimize operations, and improve overall performance by leveraging data effectively.

# Chapter 3
# Participating Technology

## 3.1 **Using Orange Data Mining**:

Customer churn prediction is a significant challenge for telecommunication companies, as losing customers can lead to substantial financial losses and undermine future growth. To address this challenge, data analysis techniques can be employed to predict churn and take proactive measures to retain customers. One effective tool in this regard is Orange Data Mining, an open-source tool that offers an intuitive, visual interface for data analysis and predictive modeling

## 3.2 **Overview of Orange Data Mining**

Orange is a data analysis software that allows users to perform data analysis using a drag-and-drop visual interface. It provides a variety of tools for exploratory analysis, model building, and evaluation. Thanks to its flexibility and ease of use, anyone can start working with it quickly, regardless of their level of expertise in data science.

## 3.3 **Proposed System**

The proposed system aims to create a predictive model for customer churn using Orange Data Mining. It includes several key stages, from data collection and analysis to model building and evaluation. Here's a detailed explanation of each stage:

### 1. Data Collection

Data is collected from multiple sources within the company, such as:

•Customer databases: containing demographic information such as age, gender, and geographical location.

•Service usage records: including data on call duration, data usage, and number of text messages.

•Customer service interactions: comprising records of complaints, inquiries, and technical support evaluations.

•Billing records: containing information about due payments, paid amounts, and payment delays

### 2. Data Preprocessing

This stage involves data cleaning, feature selection, and feature engineering:

•Data cleaning: dealing with missing values and outliers to ensure data quality.

•Feature selection: identifying the most influential features on customer churn based on field experience and initial analysis.

•Feature engineering: enhancing the data by creating new features such as monthly usage rate, subscription duration, and complaint frequency

### 3. Exploratory Data Analysis (EDA)

Exploratory analysis helps in understanding the data and discovering underlying patterns:
•Graphical visualizations: using charts, histograms, and scatter plots to explore data distributions and feature relationships.
•Descriptive statistics: calculating basic statistics like mean, median, and standard deviation to better understand the data.

### 4. Model Building

Model building involves specific steps using Orange tools:
•Data splitting: dividing the data into training and testing sets (e.g., 80% for training and 20% for testing).
•Algorithm selection: experimenting with various algorithms like Random Forest, Support Vector Machines (SVM), and Logistic Regression.
•Model training: using the training set to build the model using Orange tools

### 5. Model Evaluation

Orange tools are used to evaluate the model's performance and ensure its accuracy:
•Performance metrics: evaluating the model using metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC-ROC).
•Cross-validation: employing cross-validation techniques to ensure model robustness and avoid overfitting.

### 6. Deployment and Monitoring

Once the model is built and evaluated, it is deployed for performance monitoring in a real environment:
•Model deployment: using the model in the actual system to predict customer churn in real-time.
•Continuous monitoring: periodically monitoring the model's performance and retraining it as needed to maintain accuracy

## 3.4 Implementation with Orange

To illustrate how to implement these stages using Orange, the following steps can be followed:

1.**Data Importation**: Import data from CSV or Excel files using the File tool or directly connect to databases using the SQL Table tool.

2.**Data Preprocessing**: Use tools like Select Columns for feature selection, Impute for handling missing values, and Feature Constructor for creating new features.

3.**Exploratory Data Analysis**: Utilize tools like Distributions for data visualization, Box Plot for outlier analysis, Scatter Plot for exploring feature relationships, and Correlation for assessing feature correlations.

4.**Model Building**: Employ tools like Data Sampler for data splitting, various algorithms such as Logistic Regression, Random Forest, and SVM for model building, and Test & Score for model evaluation.

5.**Model Evaluation**: Use tools like Confusion Matrix for visualizing model performance and ROC Analysis for analyzing the ROC curve and AUC.

6.**Deployment**: Export the trained model using the Save Model tool and execute it in the production system using scripting capabilities in Orange for real-time predictions

### 3.5 Benefits of the Proposed System

•Easy-to-use Interface: Orange's visual programming interface makes it accessible to users at different levels of data science expertise.

•Comprehensive Workflow: The system covers the entire data analysis process, from data collection to deployment.

•High Flexibility: Orange's various tools provide high flexibility for experimenting with different models and techniques.

•Real-time Prediction: The deployed model can provide real-time predictions, enabling immediate interventions to retain at-risk customers.

The proposed system using Orange Data Mining provides a powerful and effective solution for predicting customer churn in the telecommunications sector. By leveraging Orange's comprehensive tools and user-friendly interface, telecommunication companies can gain valuable insights into customer behavior, accurately predict churn, and implement effective strategies to retain customers. This not only enhances customer loyalty but also significantly contributes to the overall financial health and competitiveness of the company

# Chapter 4
# Project Methodology

**Project Methodology**

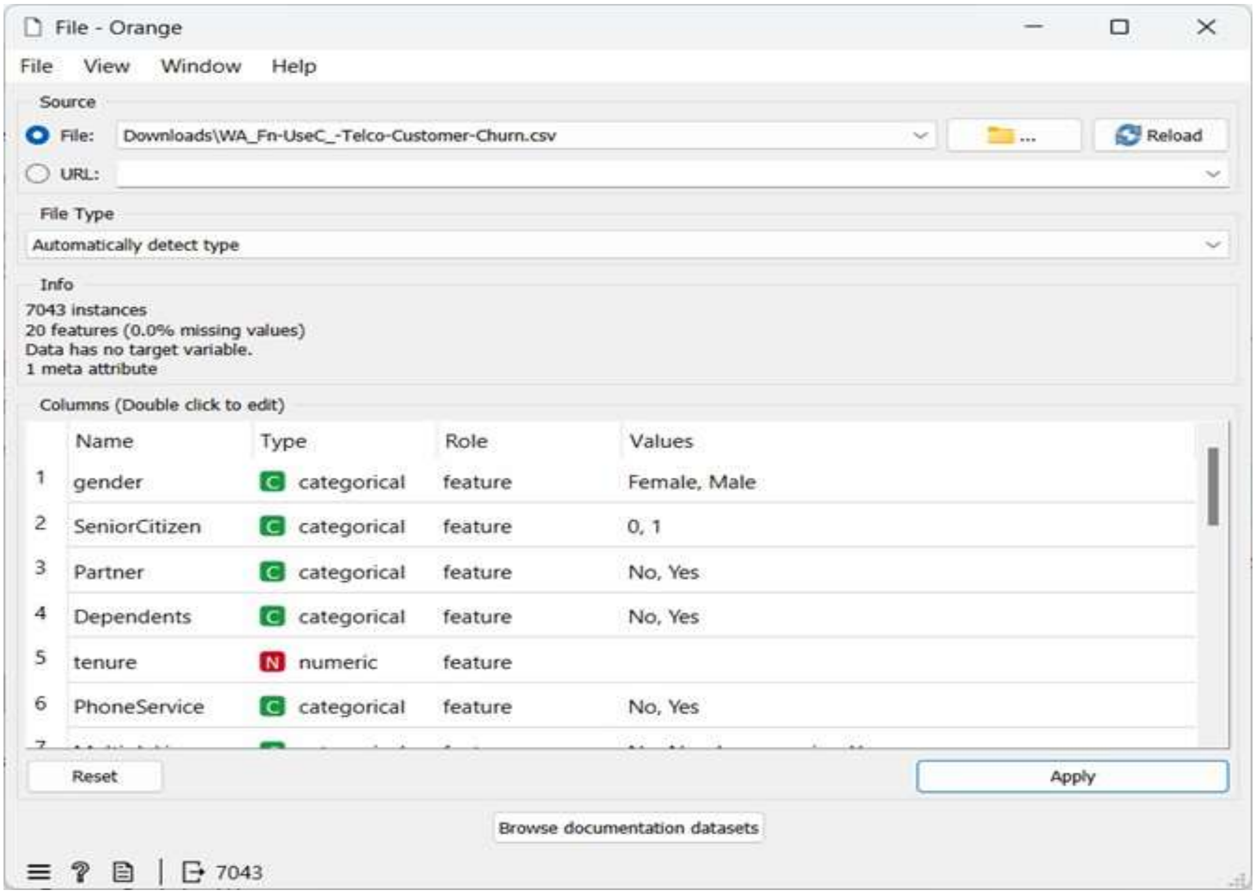
Using model cox regression (Proportional Hazards Model) is used in survival analysis to explore the relationship between the survival time of subjects and one or more predictor variables. In Orange, the Cox regression settings allow for customization to suit different analysis needs.

The Kaplan-Meier plot is a non-parametric statistic used to estimate the survival function from lifetime data. It visualizes the survival probability over time for different groups.

**Data understanding:**

We also would like to understand why the model thinks our customers churn, and for that, we need to be able to interpret the model's predictions. We will use data from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

**Data Description 7043 observations with 20 variables**
**Read data**



Source

● File: Downloads\WA_Fn-UseC_-Telco-Customer-Churn.csv

○ URL:

File Type

Automatically detect type

Info

7043 instances
20 features (0.0% missing values)
Data has no target variable.
1 meta attribute

Columns (Double click to edit)

|  | Name | Type | Role | Values |
|---|---|---|---|---|
| 1 | gender | C categorical | feature | Female, Male |
| 2 | SeniorCitizen | C categorical | feature | 0, 1 |
| 3 | Partner | C categorical | feature | No, Yes |
| 4 | Dependents | C categorical | feature | No, Yes |
| 5 | tenure | N numeric | feature |  |
| 6 | PhoneService | C categorical | feature | No, Yes |

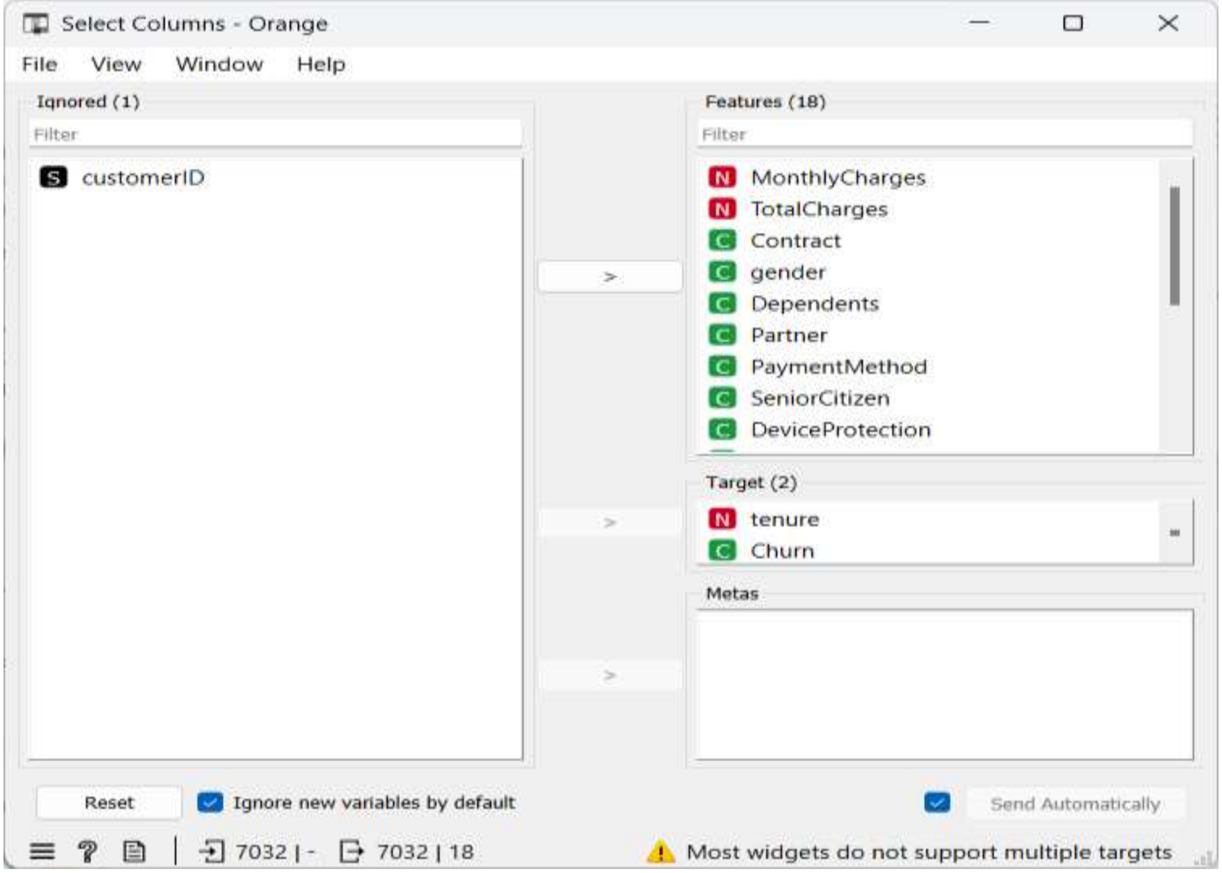
7043

**This table shows some of the variable**:

| **Data filed** | **Description** |
|---|---|

| | |
|---|---|
| Streaming Movies | No, No , Internet Service , Yes |
| Phone Service | Yes , No |
| Internet Service | DSL, Fiber optic. No |
| Streaming TV | No , No Internet Service , Yes |
| Device Protection | No, No , Internet Service , Yes |
| Payment Method | Bank transfer , credit card |
| Device Protection | No , No Internet Service , Yes |
| Multiple Lines | No, No, Phone Service,Yes |
| Tech Support | No , No Internet Service, Yes |
| Online Backup | No , No Internet Service, Yes |
| Time | Month to month . One year . Two year, three year |
| Churn | Yes , No |

**Second Step Select Columns:**



Ignored (1)

- S customerID

Features (18)

- N MonthlyCharges
- N TotalCharges
- C Contract
- C gender
- C Dependents
- C Partner
- C PaymentMethod
- C SeniorCitizen
- C DeviceProtection

Target (2)

- N tenure
- C Churn

Metas

☑ Ignore new variables by default

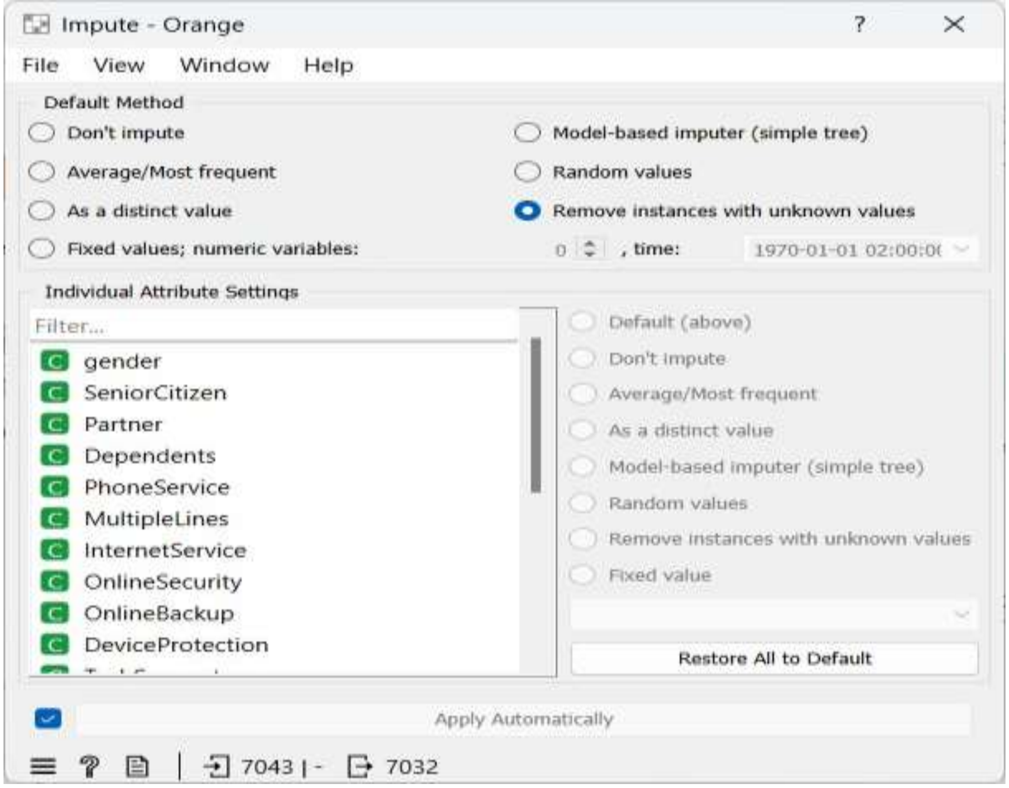
☑ Send Automatically

7032 | - 7032 | 18

⚠ Most widgets do not support multiple targets

In the telecommunication, churn data selection process,Data selection will be carried out on attributes, with the Data selection widget used to select features that will be Used to build the model.in the **Figure** shows the process of Selecting attribute data on 21 attribute data with 2 Attribute as the target, namely Churn and tenure.

### Impute Widget in Orange: Handling Missing Data



Default Method

○ Don't impute
○ Average/Most frequent
○ As a distinct value
○ Fixed values; numeric variables: 0 , time: 1970-01-01 02:00:0(
○ Model-based imputer (simple tree)
○ Random values
● Remove instances with unknown values

Individual Attribute Settings

- gender
- SeniorCitizen
- Partner
- Dependents
- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection

○ Default (above)
○ Don't impute
○ Average/Most frequent
○ As a distinct value
○ Model-based imputer (simple tree)
○ Random values
○ Remove instances with unknown values
○ Fixed value

☑ Apply Automatically

7043 | - 7032

The Impute widget in Orange is a powerful tool designed for handling missing data in your dataset. This tool provides various imputation methods to ensure data integrity before performing further analysis.

### Default Method

- **Do not impute:** No action is taken on missing values.
- **Average/Most Frequent**: Impute with the average for numeric data and the most frequent value for categorical data.
- **As a Distinct Value**: Treat missing values as a separate category.
- **Model-based Imputer (Simple Tree)**: Use a simple tree model to predict and fill in missing values.
- **Random Values**: Replace missing values with random values.
- **Remove Instances with Unknown Values**: Delete rows that contain any missing values.

### Imputation Methods for Individual Attributes

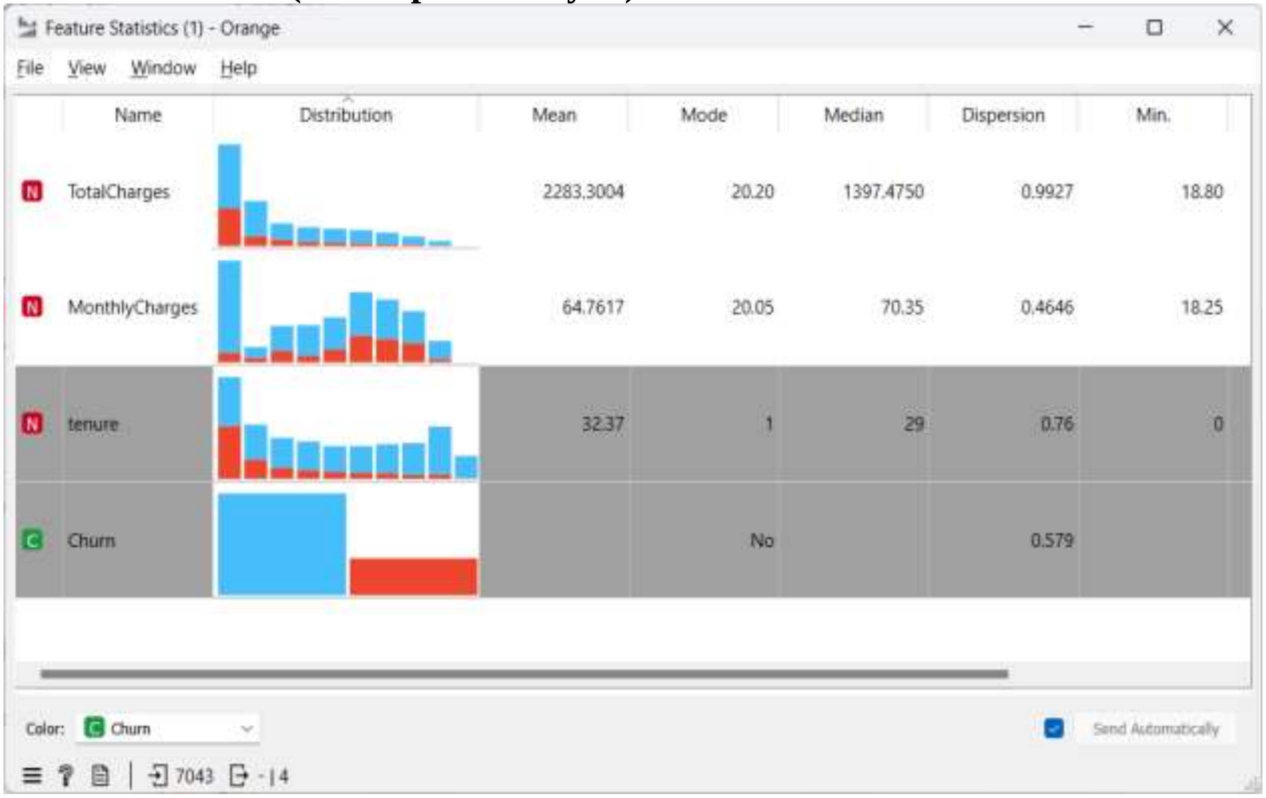
- **Default (Above)**: Use the default method set in the "Default Method" section.

- **Don't Impute**: Leave missing values as they are.
- **Average/Most Frequent**: Impute with the average for numeric data and the most frequent value for categorical data.
- **As a Distinct Value**: Treat missing values as a separate category.
- **Model-based Imputer (Simple Tree)**: Use a simple tree model to fill in missing values.
- **Random Values**: Replace missing values with random values.
- **Remove Instances with Unknown Values**: Delete rows where the selected attribute has missing values.
- **Fixed Value**: Impute with a specific value provided by the user.

### Additional Controls

- **Restore All to Default**: Reset all individual attribute settings to the default method.
- **Apply Automatically**: Automatically apply changes as you make them.

### Feature Statistics ( Descriptive analysis) :



| | Name | Distribution | Mean | Mode | Median | Dispersion | Min. |
|---|---|---|---|---|---|---|---|
| N | TotalCharges | | 2283.3004 | 20.20 | 1397.4750 | 0.9927 | 18.80 |
| N | MonthlyCharges | | 64.7617 | 20.05 | 70.35 | 0.4646 | 18.25 |
| N | tenure | | 32.37 | 1 | 29 | 0.76 | 0 |
| C | Churn | | | No | | 0.579 | |

Color: Churn

☑ Send Automatically

7043 - | 4

in the "Feature Statistics" is **descriptive analysis**. Descriptive analysis aims to describe the basic features of the data and provide simple summaries about the sample and its measures.

The main components of descriptive analysis in this context include:

- **Distribution**: Shows the distribution of values for each feature using visualizations like histograms.
- **Mean**: The average value for each feature.
- **Mode**: The most frequently occurring value in the dataset.
- **Median**: The middle value that separates the higher half from the lower half of the data.
- **Dispersion**: A measure of how spread out the values are around the mean.

**Min**: The smallest value in the dataset

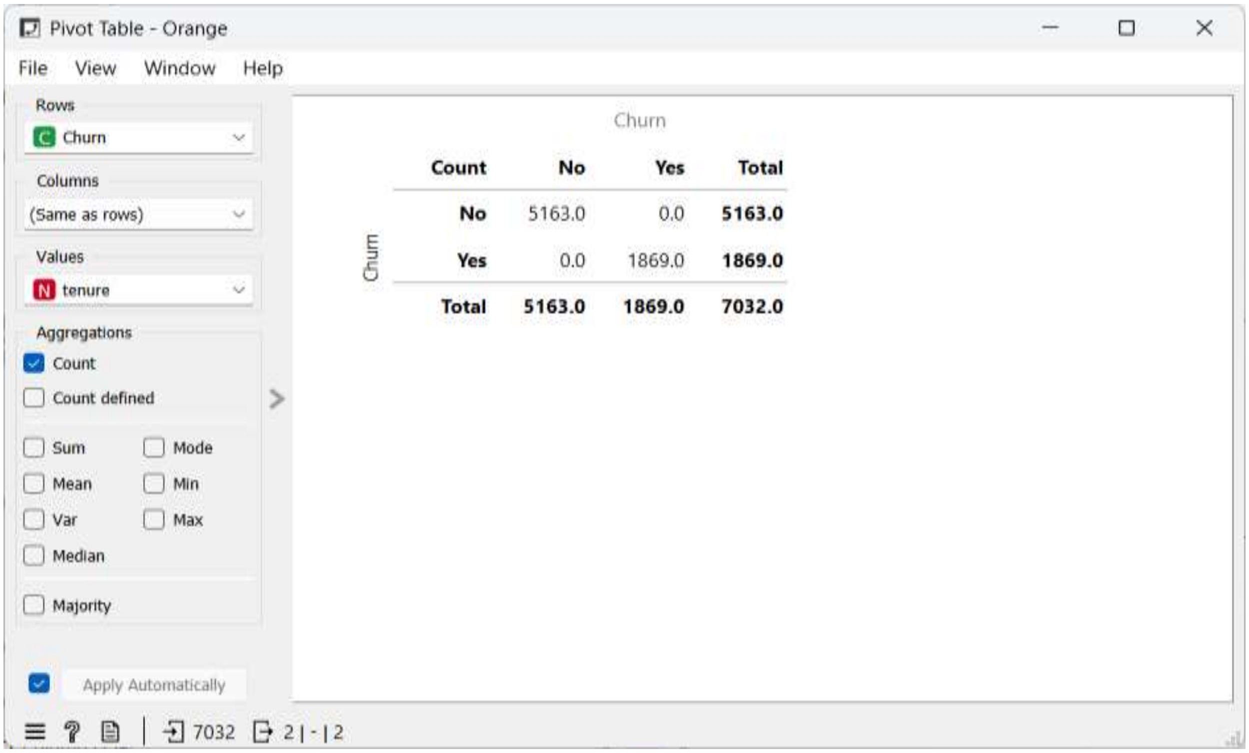
Descriptive analysis is used to understand and summarize the initial data before moving on to more complex analyses such as exploratory data analysis (EDA) or inferential analysis. It helps in identifying the main patterns, distributions, and detecting any anomalies or outliers in the data

**Features Displayed**:

- **Total Charges**:
    - Mean: 2283.3004
    - Mode: 20.20
    - Median: 1397.4750
    - Dispersion: 0.9927
    - Minimum: 18.80
- **Monthly Charges**:|
    - Mean: 64.7617
    - Mode: 20.05
    - Median: 70.35
    - Dispersion: 0.4646
    - Minimum: 18.25
- **tenure**:
    - Mean: 32.37
    - Mode: 1
    - Median: 29
    - Dispersion: 0.76
    - Minimum: 0
- **Churn**:

- ○ Possible values: Yes, No
- ○ Dispersion: 0.579

**Pivot Table :**



Rows: Churn

Columns: (Same as rows)

Values: tenure

Aggregations

☑ Count
☐ Count defined
☐ Sum ☐ Mode
☐ Mean ☐ Min
☐ Var ☐ Max
☐ Median
☐ Majority

☑ Apply Automatically

| Count | No | Yes | Total |
|---|---|---|---|
| No | 5163.0 | 0.0 | 5163.0 |
| Yes | 0.0 | 1869.0 | 1869.0 |
| Total | 5163.0 | 1869.0 | 7032.0 |

Pivot Table **summarizes the data of a more extensive table into a table of statistics. The statistics can include sums, averages, counts, etc. The widget also allows selecting a subset from the table and grouping by row values, which have to be a discrete variable. Data with only numeric variables cannot be displayed in the table.**

## Interpretation of the Table

The rows and columns are both labeled "Churn", which has two possible values: "No" and "Yes".

The pivot table shows:

**No Churn** (No, No): 5163.0

**Yes Churn** (Yes, Yes): 1869.0

- The total count of records for each category:
  - **No Churn (Row Total)**: 5163.0
  - **Yes Churn (Row Total)**: 1869.0
  - **Overall Total**: 7032.0

### Understanding Key Terms

- **Churn**:
  - Churn in a business context refers to the number or percentage of customers who stop using a company's products or services within a given time period. High churn rates can be a significant problem as they indicate that customers are leaving the company at a high rate.
  - In this data set, "Churn" likely indicates whether a customer has left the service (Yes) or not (No).
- **Rows**:
  - Rows in a pivot table typically represent different categories or groups in the data set. In this case, "Churn" categories are used as rows to show the breakdown of the data by whether customers have churned or not.

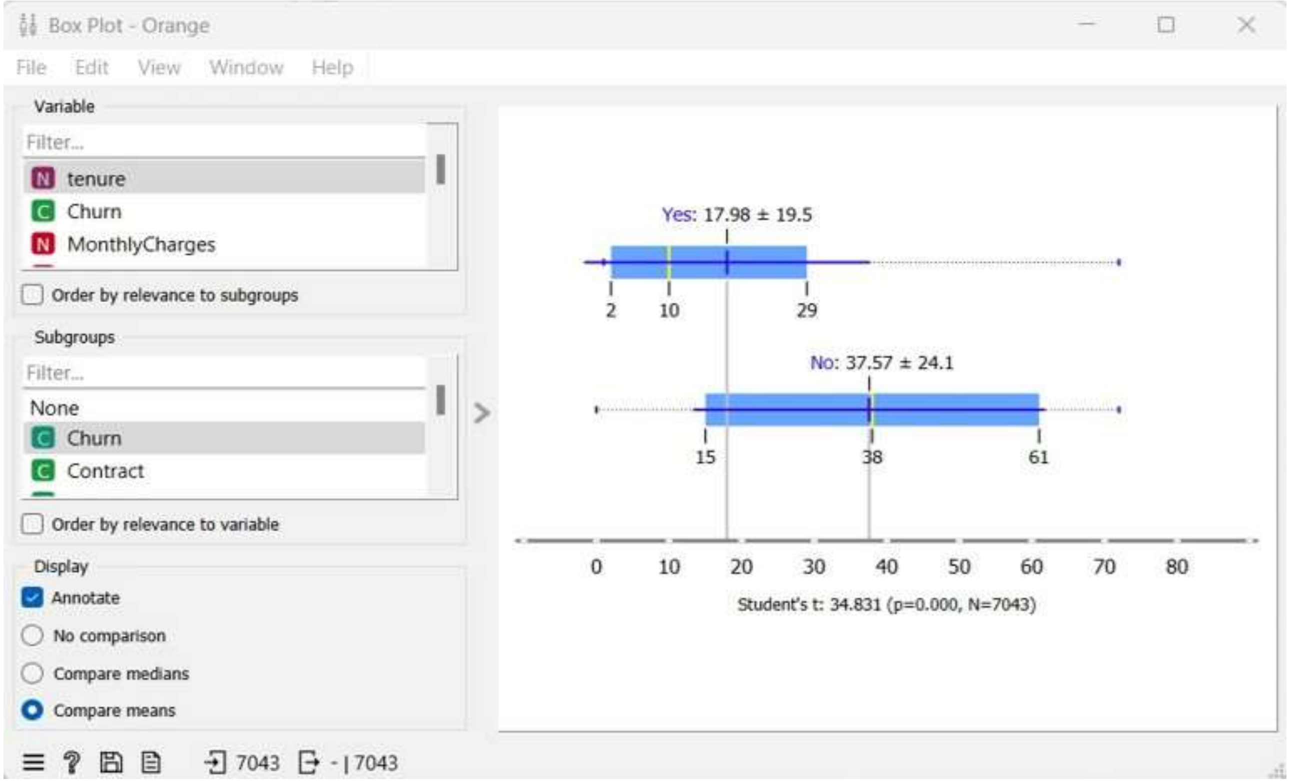
### Using Pivot Tables in Data Mining

- **Exploratory Data Analysis (EDA)**: Pivot tables are often used in the initial stages of data analysis to explore and understand the data. They help identify trends, patterns, and anomalies.
- **Summarization**: Pivot tables can summarize large data sets into meaningful summaries. By aggregating data based on categories, pivot tables make it easier to interpret complex data.
- **Data Preparation**: In data mining, preparing data is crucial. Pivot tables help in cleaning and organizing data, making it ready for further analysis or machine learning modeling.

In this specific pivot table, the goal seems to be to understand the distribution of "tenure" counts across different churn statuses. This information can be valuable for identifying factors that influence customer retention and developing strategies to reduce churn

**Box Plot (Diagnostic Analytics )** :

The Relationship Between tenure and churn :



Variable

Filter...

N tenure

C Churn

N MonthlyCharges

□ Order by relevance to subgroups

Subgroups

Filter...

None

C Churn

C Contract

□ Order by relevance to variable

Display

☑ Annotate

○ No comparison

○ Compare medians

● Compare means

Student's t: 34.831 (p=0.000, N=7043)

A Box Plot is a useful tool in diagnostic analytics for understanding the distribution of data and identifying outliers, variability, and the central tendency of the data set. Here's how you can utilize the Box Plot widget in Orange for diagnostic analytics, particularly focusing on tenure and churn.

The box plot in the image represents the relationship between the variable "tenure" and the variable "Churn". Here's an analysis based on the box plot:

1. **Variables:**
   - Tenure: Likely represents the duration a customer has been with the service provider.
   - Churn: Indicates whether a customer has left (Yes) or stayed (No).

2. **Box Plot Description**:
   - Churn = Yes:
     - Mean tenure: 17.98 ± 19.5 months.
     - Distribution is skewed with most values clustering around a lower range.
     - Indicates that customers who churned typically had a shorter tenure.

- Churn = No:
  - Mean tenure: 37.57 ± 24.1 months.
  - Distribution is more spread out compared to the churned customers.
  - Suggests that customers who stayed have, on average, a longer tenure.

**3. Statistical Test:**

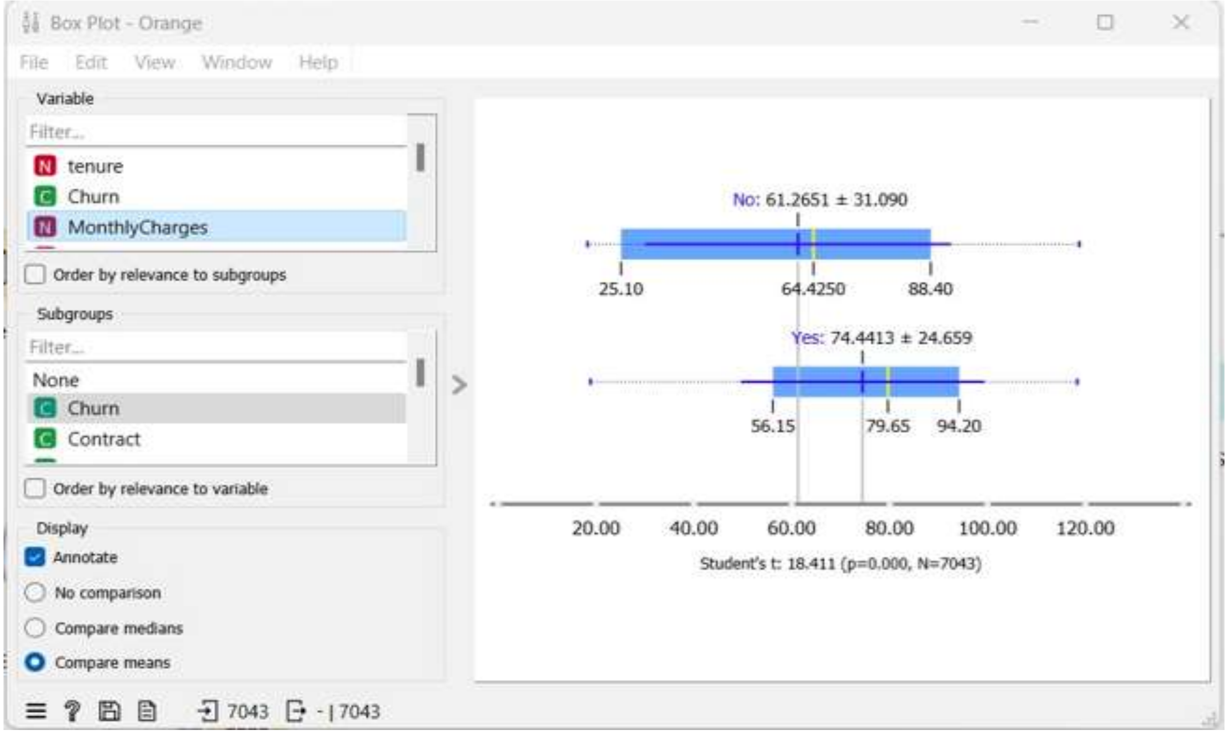
- A Student's t-test was conducted, yielding a p-value of 0.000.
- The test confirms a significant difference between the tenures of customers who churned and those who did not.

**4. Implications:**

- Shorter tenure is associated with a higher likelihood of churn.
- The company may need to focus on retention strategies for newer customers to reduce churn rates.

Overall, the box plot and statistical analysis highlight a significant relationship between tenure and churn, suggesting that longer-tenured customers are less likely to leave the service.

**The Relationship Between Monthly Charges And Churn :**



Variable

- N tenure
- C Churn
- N MonthlyCharges

□ Order by relevance to subgroups

Subgroups

- None
- C Churn
- C Contract

□ Order by relevance to variable

Display

☑ Annotate

○ No comparison

○ Compare medians

● Compare means

Student's t: 18.411 (p=0.000, N=7043)

The box plot illustrates the relationship between "Monthly Charges" and "Churn" status. Here's an analysis based on the box plot:

1. **Variables:**

- Monthly Charges: Represents the amount a customer pays monthly.
- Churn: Indicates whether a customer has left (Yes) or stayed (No).

2. **Box Plot Description:**

- Churn = No:
  - Mean Monthly Charges: $61.27 ± $31.09.
  - The distribution of monthly charges for customers who did not churn ranges approximately from $25.10 to $88.40.
  - The spread indicates that many non-churning customers pay lower monthly charges.

- Churn = Yes:
  - Mean Monthly Charges: $74.44 ± $24.66.
  - The distribution for customers who churned ranges approximately from $56.15 to $94.20.

- Customers who churned tend to have higher monthly charges on average.

3. **Statistical Test:**

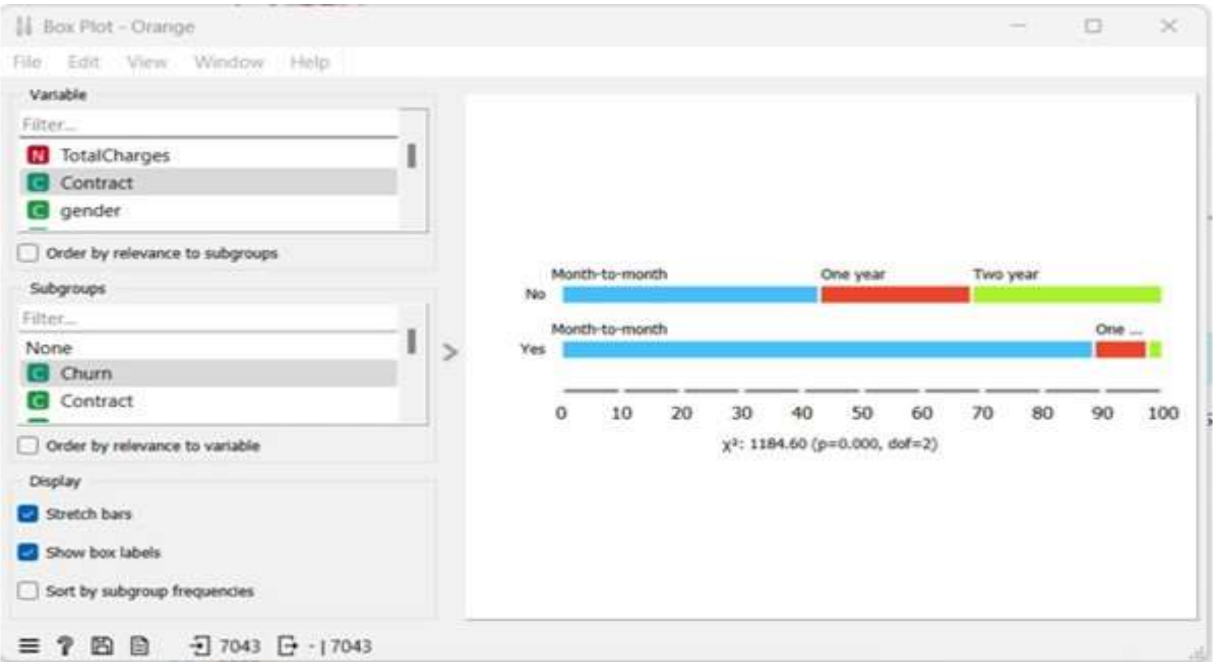
- A Student's t-test was performed, resulting in a p-value of 0.000.
- This p-value indicates a significant difference in monthly charges between customers who churned and those who did not.

4. **Implications:**

- Higher monthly charges are associated with an increased likelihood of churn.
- Customers who pay more are more likely to leave, suggesting that higher charges may be a factor in customer dissatisfaction or financial burden.
-The company might consider reviewing its pricing structure or offering more value to customers with higher monthly charges to reduce churn.

Overall, the box plot and the statistical analysis indicate that there is a significant relationship between monthly charges and churn, with higher charges correlating with a higher likelihood of customer churn.

**The Relationship Between Contract& Churn**



Variable

TotalCharges<br>
Contract<br>
gender

☐ Order by relevance to subgroups

Subgroups

None<br>
Churn<br>
Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

$\chi^2$: 1184.60 (p=0.000, dof=2)

**Variables:**

•Contract: Represents the type of contract customers have, which can be monthly, yearly, or biennial.
•Churn: Indicates whether the customer has canceled their subscription (Yes) or not (No).

**Subgroups in the plot**:

No (without Churn):

•**Monthly**: The highest proportion of customers who have not canceled their subscription have monthly contracts.

•**Yearly and biennial**: The proportion is lower compared to monthly contracts.
Yes (with Churn):

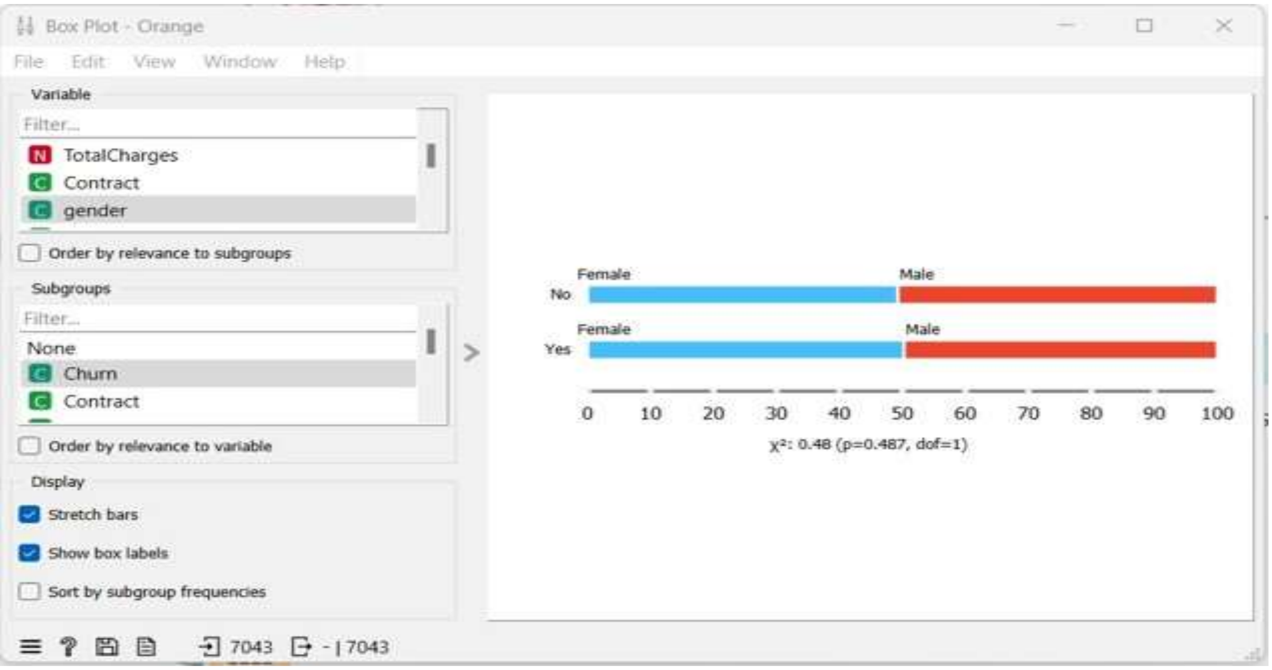
•**Monthly**: The highest proportion of customers who have canceled their subscription have monthly contracts.

•**Yearly and biennial**: The proportion is significantly lower compared to monthly contracts.

The analysis relies on the Chi-Square Test of Independence, used to determine the relationship between variables by comparing the calculated Chi-Square value with a critical value in a table.

If the calculated value exceeds the critical value and the p-value is less than 0.05, there is a statistically significant relationship between the type of contract and subscription cancellation.

### The Relationship Between gender& Churn



Box Plot - Orange

Variable

Filter...

N TotalCharges

C Contract

C gender

□ Order by relevance to subgroups

Subgroups

Filter...

None

C Churn

C Contract

□ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

χ²: 0.48 (p=0.487, dof=1)

7043 - | 7043

The Chi-Square Test of Independence is used to determine whether there is a relationship between two categorical variables. In this analysis, we use it to ascertain whether there is a relationship between gender (male or female) and unsubscription cancellation (Yes or No).

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

**Analysis Steps:**

**1. Observed frequencies**: Actual values from the data regarding gender and subscription cancellation.

**2. Expected frequencies**: Values expected if there is independence between gender and subscription cancellation.

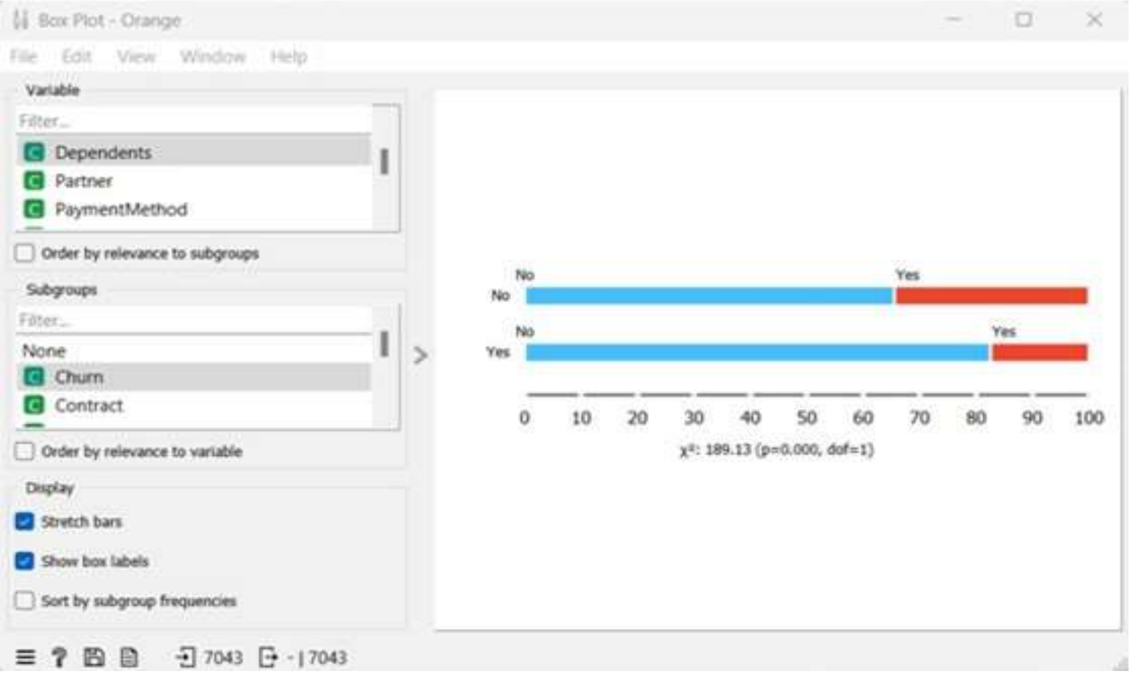
**3. Calculation of Chi-Square value using the equation**:

•Find the difference between each observed and expected pair, square the difference, then divide by the expected value.
•Sum all these values to obtain the Chi-Square value.

**Results:**
•Chi-Square value: $\chi^2 = 0.48$
•p-value: $p = 0.487$

**Interpretation:**
•If the Chi-Square value is sufficiently large or the p-value is less than 0.05, it indicates a statistically significant relationship between gender and subscription cancellation.
•In this example, with $\chi^2 = 0.48$ and $p = 0.487$, both values are higher than 0.05, indicating no significant relationship between gender and subscription cancellation.
Conclusion:
•There is no significant association between gender and subscription cancellation. This implies that cancellation policies are not significantly influenced by the customer's gender, allowing customer retention strategies to remain gender-neutral.
Companies can use this information to understand that their cancellation policies are not greatly affected by the customer's gender, meaning that customer retention strategies can be gender-neutral.

**The Relationship Between Dependents & Churn**



Variable

Filter...

Dependents

Partner

PaymentMethod

□ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

□ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

χ²: 189.13 (p=0.000, dof=1)

**Variable:**
•The main variable selected is "Dependents."

**Subgroups:**
•The data is divided into subgroups based on "Churn" status.

Analysis:

•The box plot shows two main groups:
- Without dependents (No)
- With dependents (Yes)

•Each group is further divided into:
- Churn (Yes)
- No churn (No)

•The Chi-Square equation mentioned in the image is used to analyze the relationship between the variables "Dependents" and "Churn".

•Chi-Square Equation

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where:

- $\chi 2$: The Chi-Square statistic

.

- Oii: The observed frequency in cell i of the contingency table.

- Eii: The expected frequency in cell i of the contingency table.

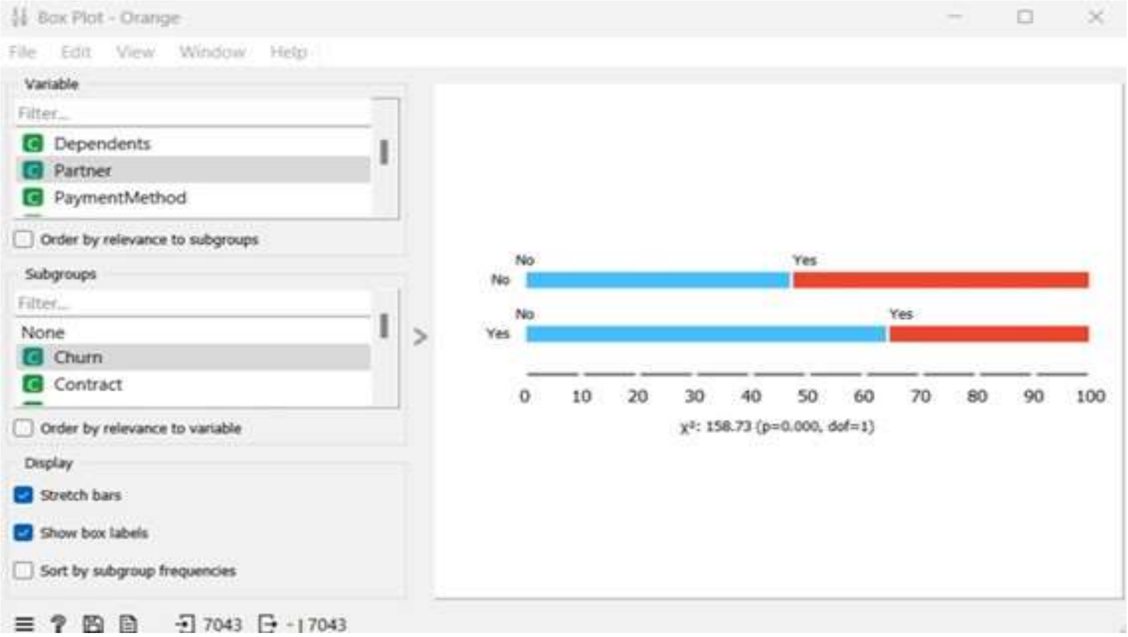
In the image, the calculated Chi-Square value is 189.13189.13189.13 with 1 degree of freedom (DF = 1), and the p-value is 0.0000.0000.000. These results indicate a statistically significant relationship between the variables "Dependents" and "Churn

**Results:**

- Customers without dependents (No) have a higher churn rate.
- Customers with dependents (Yes) have a lower churn rate.

Conclusion :
Having dependents is associated with a lower likelihood of customer churn.

**The Relationship Between Partner & Churn**



Variable

Filter...

- Dependents
- Partner
- PaymentMethod

☐ Order by relevance to subgroups

Subgroups

Filter...

- None
- Churn
- Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 158.73 (p=0.000, dof=1)

7043 - | 7043

1. **Select Variables:**

- **Main Variable**: Select "Partner," which indicates whether customers have a partner or not.
- **Subgroups: Select** "Churn," indicating whether customers have discontinued the service.

2. **Box Plot Visualization**:

- The visualization will display the distribution of customers who churned ("Yes") and those who did not churn ("No") within each category of the "Partner" variable (Yes or No).

3. **Statistical Test:**

- The chi-square test result will be shown, indicating whether there is a significant association between having a partner and churn

**Hypothetical Interpretation (Based on Similar Visualization):**

**•Partner (No):**

A higher percentage of customers without a partner tend to churn compared to those who do not churn.

- **Partner (Yes):**

A lower percentage of customers with a partner tend to churn compared to those who do not churn.

If the chi-square test result shows a significant p-value (e.g., $p < 0.05$), this indicates that there is a statistically significant relationship between having a partner and customer churn.

**Example Results (Hypothetical):**

•**Chi-Square Test Result:** $\chi^2 = 120.45$, $p = 0.000$, dof = 1

This indicates a significant relationship between having a partner and churn.

•**Visualization Interpretation:**

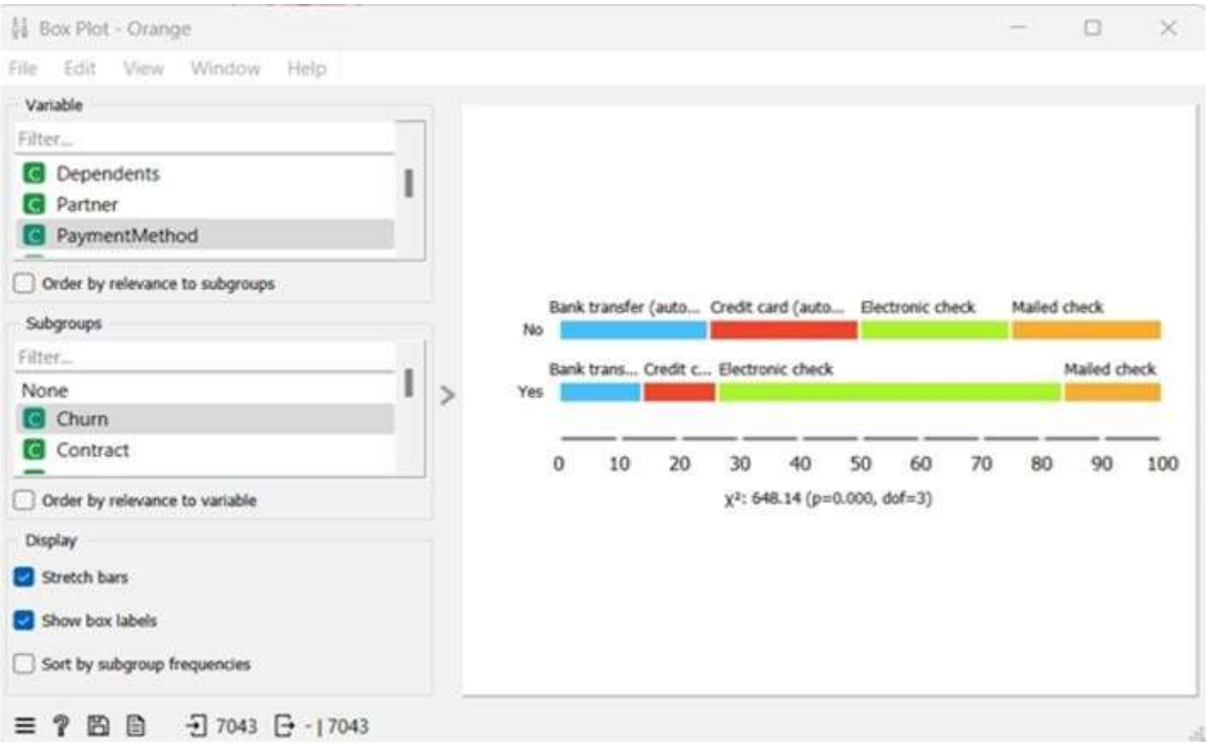
No Partner: A higher proportion of customers without a partner are churning.

With Partner: A lower proportion of customers with a partner are churning.

**Conclusion:**

Customers without a partner are more likely to churn compared to those with a partner. This insight can be valuable for targeting retention strategies towards customers without partners.

### The Relationship Between Payment Method & Churn



Variable: PaymentMethod

□ Order by relevance to subgroups

Subgroups: Churn

□ Order by relevance to variable

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

χ²: 648.14 (p=0.000, dof=3)

**Box Plot Analysis:**

1. **Variable**:

   **Main Variable**: "PaymentMethod," indicating the method customers use to pay.
   **Subgroups**: "Churn," indicating whether customers have discontinued the service.

2. **Box Plot Visualization**:

   The x-axis represents the percentage of customers.
   The y-axis shows the categories for "Payment Method" divided by subgroups (Churn "Yes" or "No").
   Different colors represent different payment methods:
   Blue: Bank transfer (automatic)
   Red: Credit card (automatic)
   Green: Electronic check
   Orange: Mailed check

3. **Statistical Test:**

The chi-square test result is shown below the plot ($\chi^2 = 648.14$, $p = 0.000$, dof = 3), indicating a significant association between the payment method and churn.

**Interpretation:**

**Churn (Yes):**

- **Bank transfer (automatic)**: A relatively smaller percentage of customers who churn use this method.
- **Credit card (automatic):** Also a relatively smaller percentage of customers who churn use this method.
- **Electronic check:** A large percentage of customers who churn use this method.
- **Mailed check:** A moderate percentage of customers who churn use this method.

**Churn (No):**

- **Bank transfer (automatic):** A relatively larger percentage of customers who do not churn use this method.
- **Credit card (automatic):** A relatively larger percentage of customers who do not churn use this method.
- **Electronic check**: A smaller percentage of customers who do not churn use this method.
- **Mailed check**: A moderate percentage of customers who do not churn use this method.

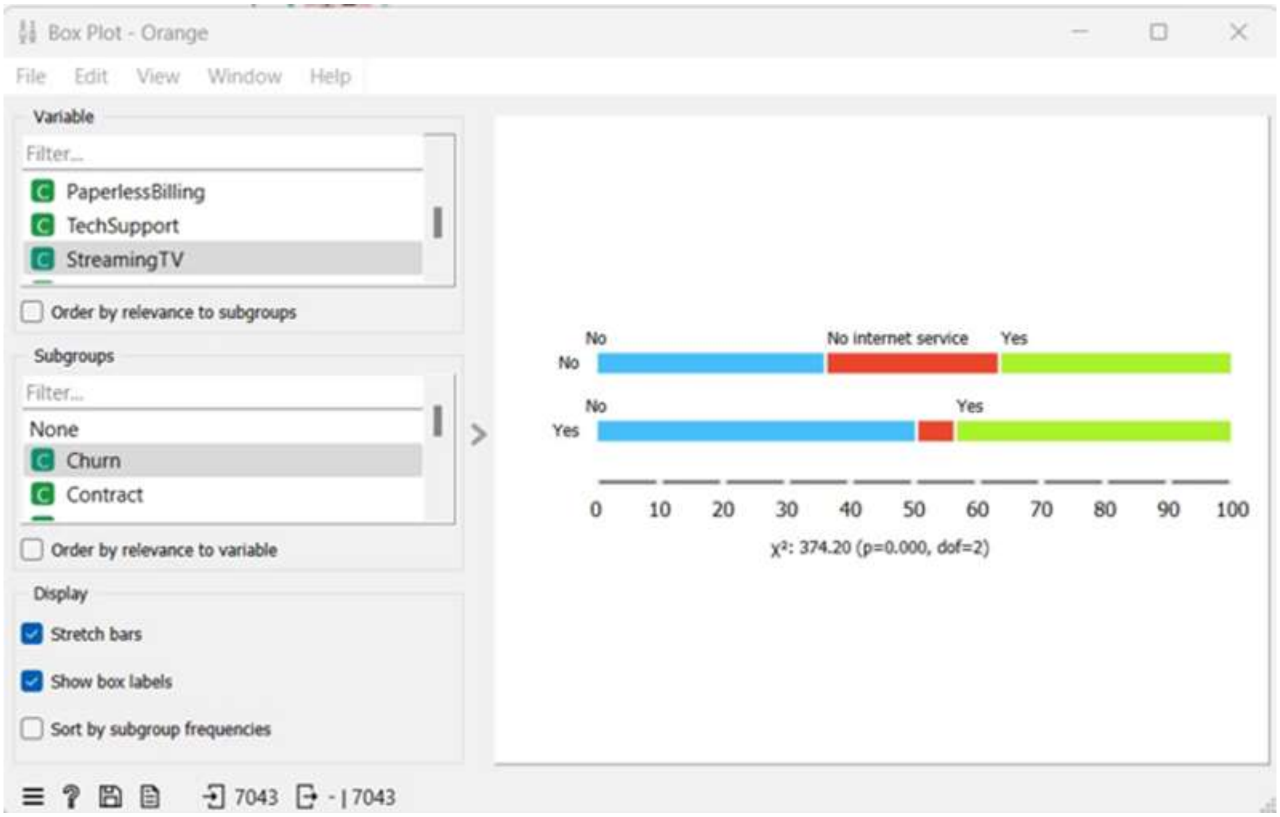
**Statistical Significance**:

- The chi-square test result ($\chi^2 = 648.14$, $p = 0.000$, dof = 3) indicates that there is a statistically significant relationship between the payment method and customer churn.

**Conclusion:**

Customers using electronic checks are more likely to churn compared to those using other payment methods. On the other hand, customers using automatic bank transfers or credit card payments are less likely to churn. This insight can be useful for identifying customers at higher risk of churn based on their payment method and potentially targeting them with retention strategies

**The Relationship between Online Security & Churn**



Variable

Filter...

PaperlessBilling

TechSupport

StreamingTV

□ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

□ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

$\chi^2$: 374.20 (p=0.000, dof=2)

**Variable:**

•The main variable selected is "Online Security."

**Subgroups:**

•The data is divided into subgroups based on "Churn" status.

**Analysis:**

• The box plot shows three main groups:
Without online security (No)
With online security (Yes)
No internet service (No internet service)

Each group is further divided into:
Churn (Yes)
No churn (No)

**Results:**

• Customers without online security (No) have a higher churn rate.

• Customers with online security (Yes) have a lower churn rate.

• Customers with no internet service are shown separately as they do not fit into the categories of having or not having online security.

**Chi-Square Equation**

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

**Where:**

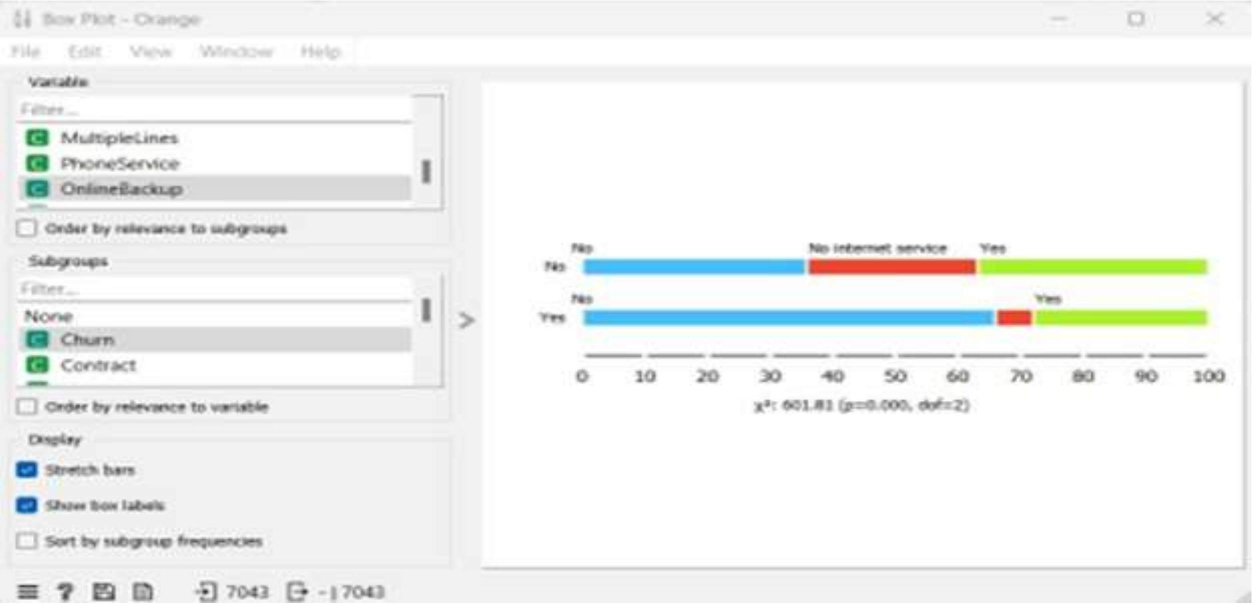
- $\chi 2$: The Chi-Square statistic.
- Oi: The observed frequency in cell i of the contingency table.
- Ei: The expected frequency in cell i of the contingency table.

In the image, the calculated Chi-Square value is 850.0850.0850.0 with 2 degrees of freedom (df = 2), and the p-value is 0.0000.0000.000. These results indicate a statistically significant relationship between the variables “Online Security" and "Churn".

**Conclusion:**

Having online security is associated with a lower likelihood of customer churn. Conversely, not having online security is associated with a higher likelihood of churn. Customers with no internet service are a distinct category and are not directly comparable to those with or without online security.

**The Relationship between Online Backup & Churn**



Variable

Filter...

MultipleLines
PhoneService
OnlineBackup

☐ Order by relevance to subgroups

Subgroups

Filter...

None
Churn
Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

**Variables:**

- **Main Variable:** "Online Backup"
- **Subgroups:** Based on "Churn" status

**Analysis:**

- Groups:

o No Online Backup (No)
o Has Online Backup (Yes)
o No Internet Service (No internet service)

- Each group is divided into:

o Churn (Yes)
o No churn (No)

**Results:**

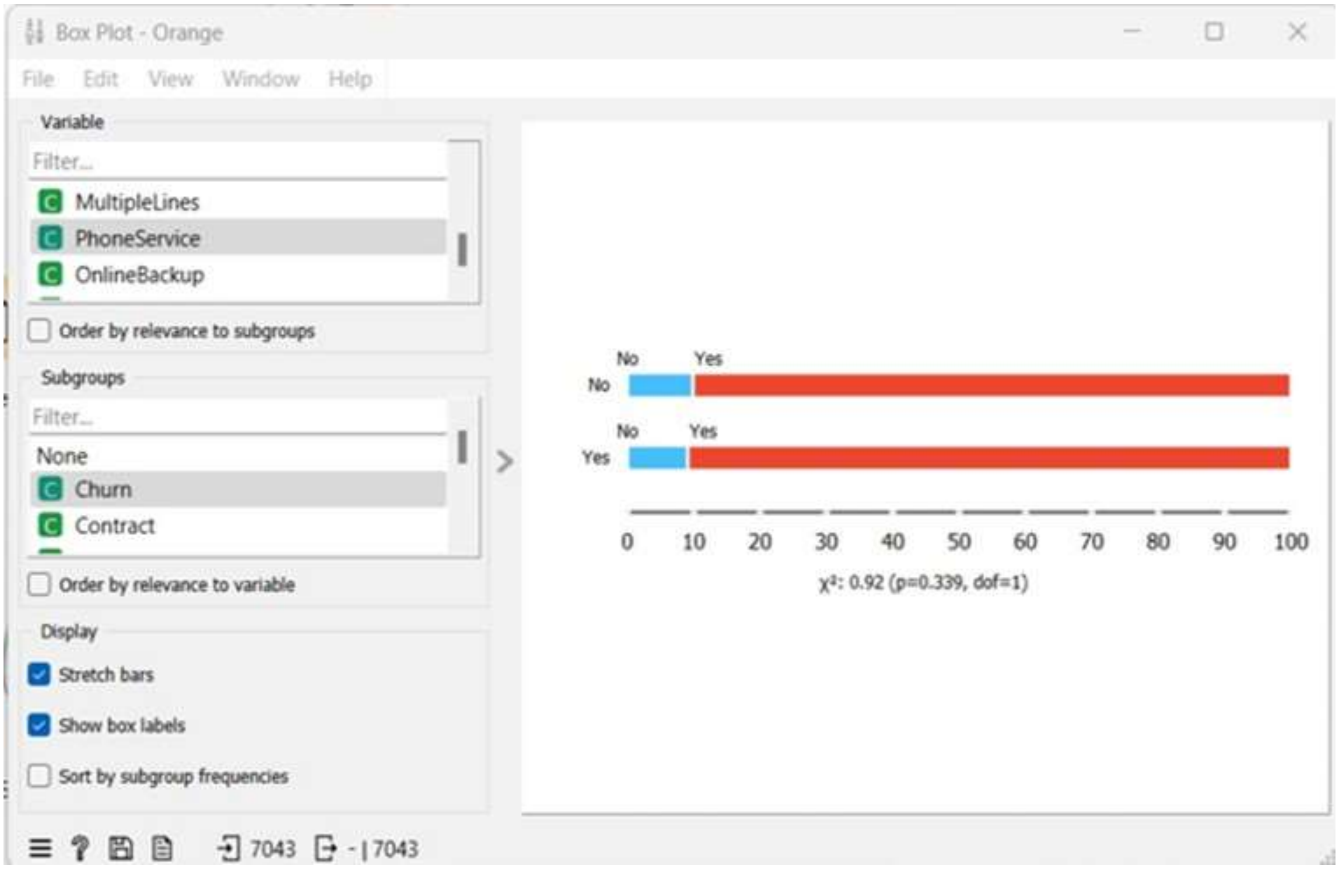
- Customers without online backup (No) have a higher churn rate.
- Customers with online backup (Yes) have a lower churn rate.
- Customers with no internet service are shown separatel

**Statistical Significance:**

- Chi-Square value: 601.8601.8601.8
- Degrees of Freedom (df): 2
- p-value: 0.0000.0000.000

**Conclusion:** Having online backup is associated with a lower likelihood of customer churn. Not having online backup is associated with a higher likelihood of churn. Customers with no internet service are a distinct category.

### The Relationship between Phone Service & Churn



Variable

Filter...

MultipleLines

PhoneService

OnlineBackup

☐ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 0.92 (p=0.339, dof=1)

The box plot visualizes the relationship between phone service subscription and customer churn. It compares the proportion of customers who have churned against those who have not, segmented by whether they have a phone service.

**Key observations:**

• **No Phone Service**: The proportion of customers who have churned is slightly lower than those who have not churned.

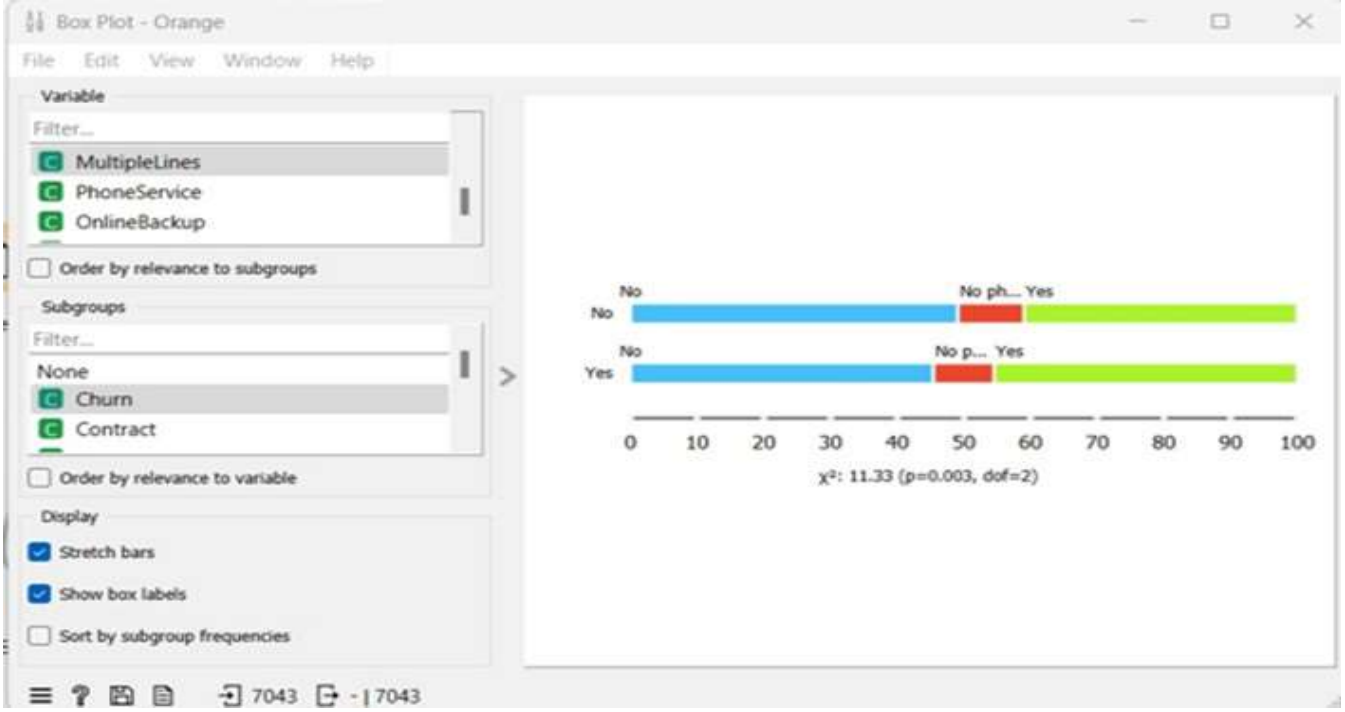
• **Phone Service:** The proportion of churned customers is slightly higher compared to those who have not churned.
This suggests that there is a potential relationship between having a phone service and the likelihood of customer churn. Further analysis might be needed to determine the strength and significance of this relationship.

• **Without Phone Service**: The proportion of customers who churned is slightly lower than those who did not churn.

• **With Phone Service**: The proportion of customers who churned is slightly higher than those who did not churn

• **Conclusion**: There is a potential relationship between having a phone service and an increased likelihood of customer churn

**The Relationship between Multiple Line & Churn**



Variable

Filter...

MultipleLines

PhoneService

OnlineBackup

☐ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

The bar plot visualizes the relationship between having multiple lines and customer churn. It compares the proportion of customers who have churned against those who have not, segmented by whether they have multiple lines.

**Key observations**:

**No Multiple Lines**: The proportion of customers who have churned is higher than those who have not churned.
**Multiple Lines**: The proportion of churned customers is lower compared to those who have not churned.

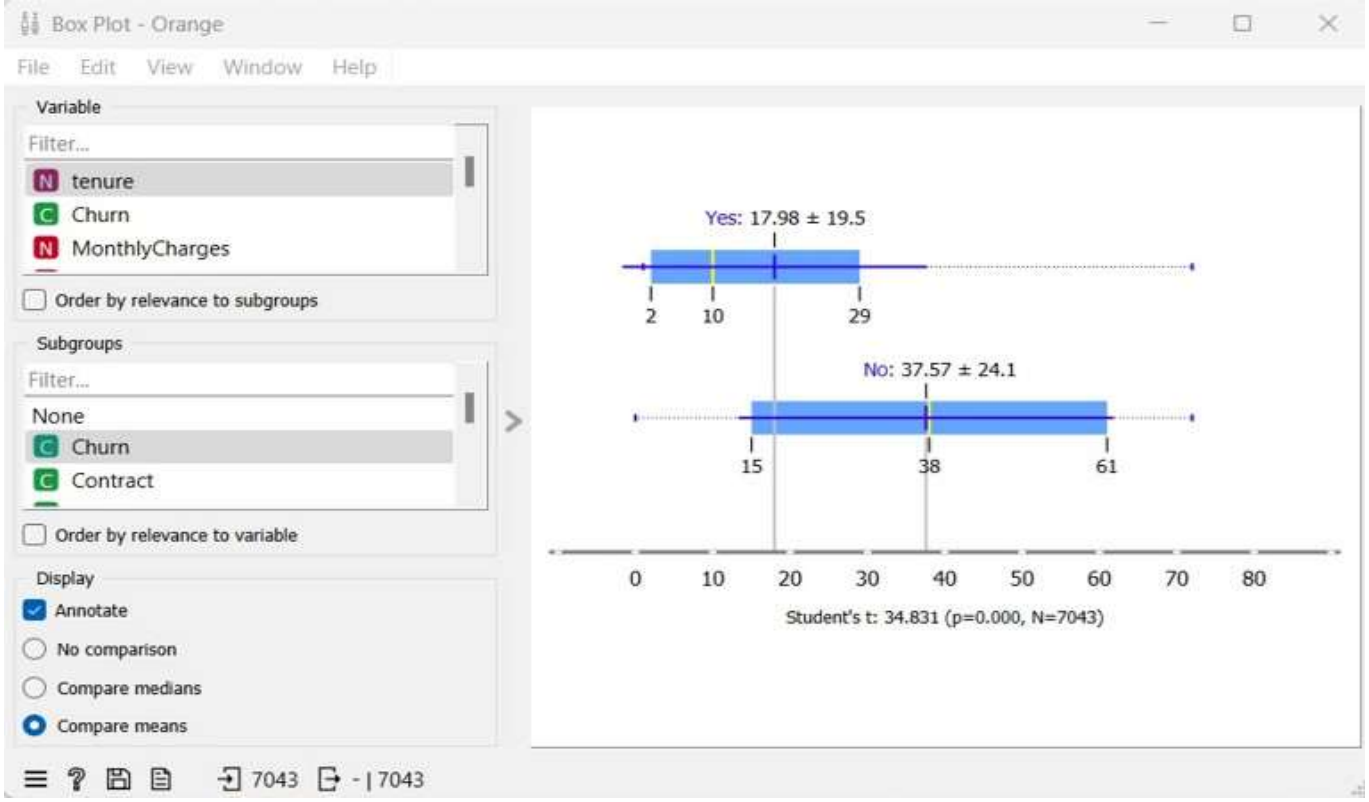
This suggests that having multiple lines may be associated with a reduced likelihood of customer churn. Further analysis might be needed to determine the strength and significance of this relationship.

**Key Points:**
**Without Multiple Lines**: The proportion of customers who churned is higher than those who did not churn.
**With Multiple Lines**: The proportion of customers who churned is lower than those who did not churn
**Chi-Square Test**

The chi-square statistic (χ²) is used to examine the relationship between multiple lines and churn. The observed chi-square value (χ²) is 11.39, with a p-value of 0.003 and degrees of freedom (df) of 2. This indicates a significant association between having multiple lines and the likelihood of churn.

**Conclusion**: There is a potential relationship between having multiple lines and a decreased likelihood of customer churn.

### Relationship Between Tenure & Churn
### Analysis Using Box Plot:



Box Plot - Orange

Variable

Filter...

N tenure

C Churn

N MonthlyCharges

□ Order by relevance to subgroups

Subgroups

Filter...

None

C Churn

C Contract

□ Order by relevance to variable

Display

☑ Annotate

○ No comparison

○ Compare medians

● Compare means

Student's t: 34.831 (p=0.000, N=7043)

**1. Variables:**

- Tenure: Represents the number of months the customer has been subscribed.
- Churn: Indicates whether the customer has canceled their subscription (Yes) or not (No).

**2. Distribution**:

- Without Churn (No):

Average Tenure: 37.57 months ± 24.1.
Range: 15 to 61 months.

- With Churn (Yes):
  Average Tenure: 17.98 months ± 19.5.
  Range: 2 to 29 months.

**3. Statistical Analysis**:

- Independent Sample t-Test:
  t-value: 34.831
  p-value: 0.000
  Sample Size (N): 7043

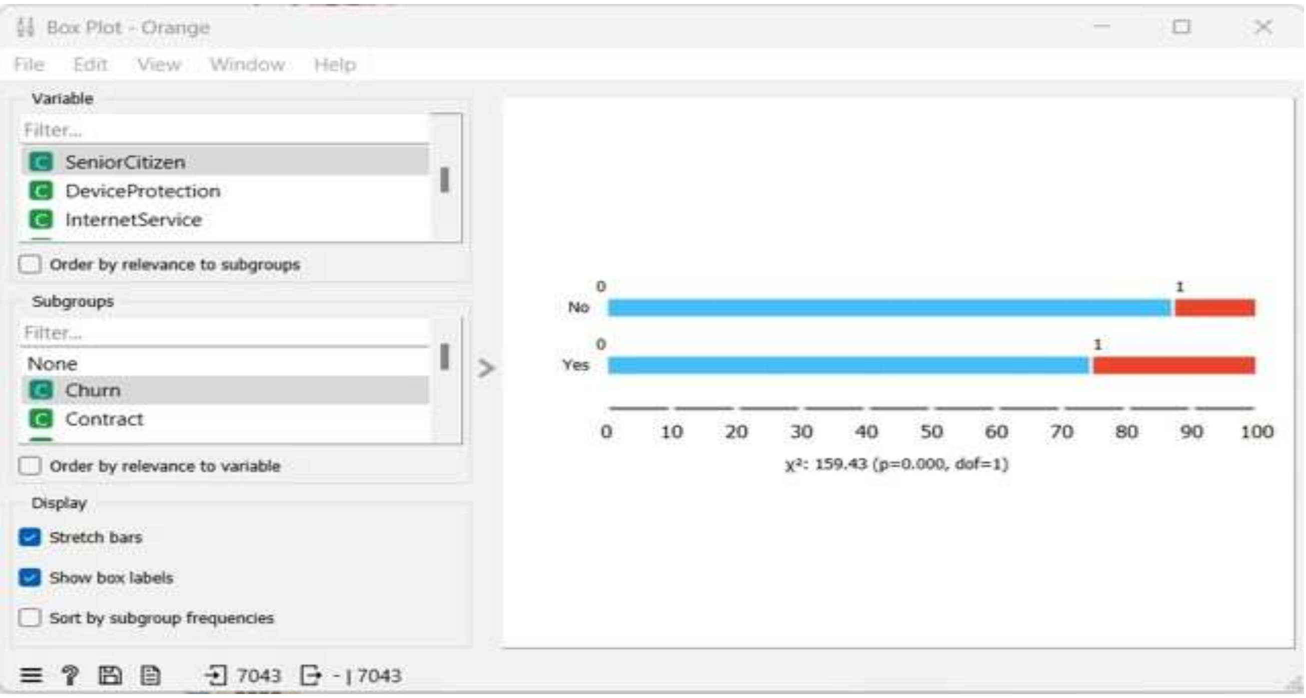
**4. Interpretation:**

- Customers who did not churn (No) have significantly longer tenures compared to those who churned (Yes).
- The relationship between tenure and churn is statistically significant, as the p-value is less than 0.05.

**Conclusion:**

The data shows that customers with longer tenures are less likely to churn compared to those with shorter tenures. This relationship is statistically significant, indicating that tenure is a strong indicator of the likelihood of subscription cancellation.

**Relationship Between Senior Citizen & Churn**
**Analysis Using Box Plot:**



Variable

Filter...

SeniorCitizen

DeviceProtection

InternetService

☐ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

7043 - | 7043

χ²: 159.43 (p=0.000, dof=1)

**1. Variables**:

- Senior Citizen: Indicates whether the customer is a senior citizen (1) or not (0).
- Churn: Indicates whether the customer has canceled their subscription (Yes) or not (No).

**2. Distribution**:

- Without Churn (No):
  Non-Senior Citizens (0): Represented by the blue bar.
  Senior Citizens (1): Represented by the red bar.
- With Churn (Yes):
  Non-Senior Citizens (0): Represented by the blue bar.
  Senior Citizens (1): Represented by the red bar.

**3. Statistical Analysis:**

- Chi-Square Test:
  Chi-Square value ($\chi^2$): 159.43
  p-value: 0.000
  Degrees of Freedom (dof): 1

**4 Interpretation:**

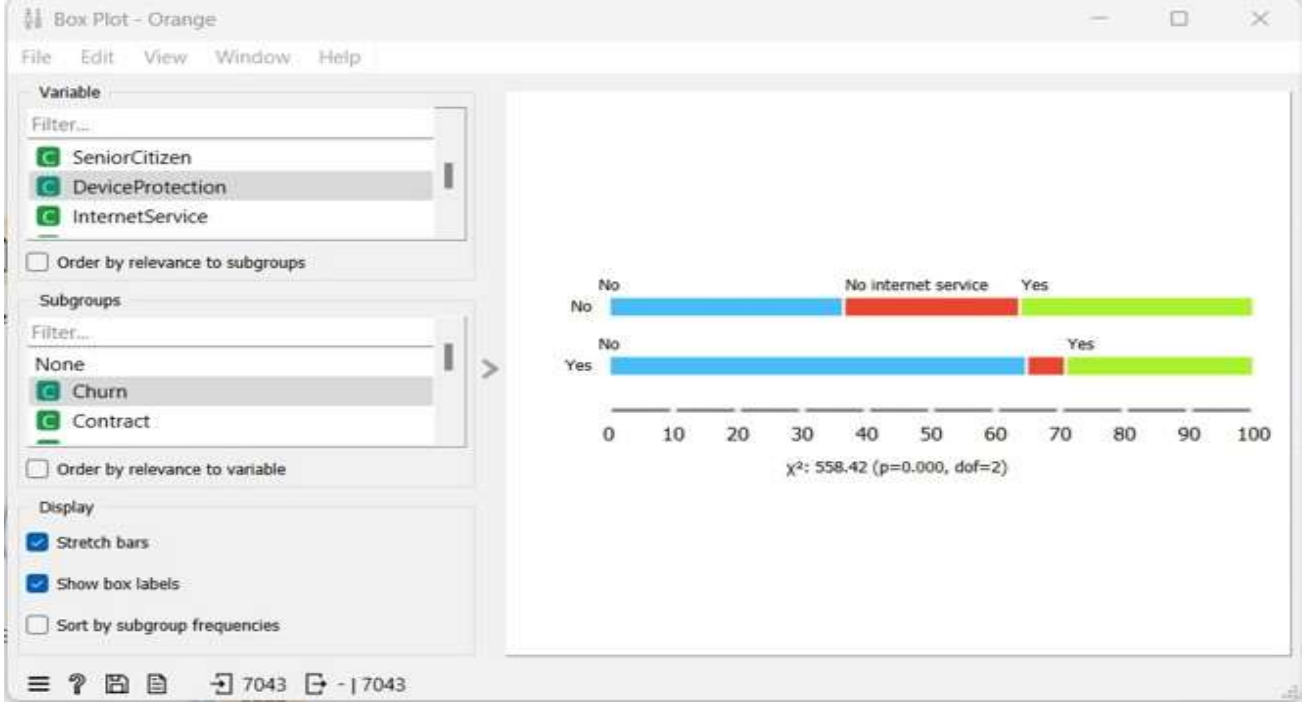
o There is a significant difference in churn rates between senior citizens and non-senior citizens.

o The Chi-Square test results indicate that this relationship is statistically significant, as the p-value is less than 0.05.

**Conclusion:**

The data shows a statistically significant relationship between being a senior citizen and churn. Senior citizens have different churn rates compared to non-senior citizens, suggesting that age group is an important factor to consider when analyzing customer churn.

**Relationship Between Device Protection & Churn Analysis Using Box Plot:**



Box Plot - Orange

Variable

- SeniorCitizen
- DeviceProtection
- InternetService

□ Order by relevance to subgroups

Subgroups

- None
- Churn
- Contract

□ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

7043 - | 7043

**1. Variables:**

o Device Protection: Indicates whether the customer has device protection service (Yes) or not (No), or if they do not have internet service.
o Churn: Indicates whether the customer has canceled their subscription (Yes) or not (No).

**2. Distribution**:

o Without Churn (No):
No Device Protection: Represented by the blue bar.
No Internet Service: Represented by the red bar.
Device Protection: Represented by the green bar.
o With Churn (Yes):
No Device Protection: Represented by the blue bar.
No Internet Service: Represented by the red bar.
Device Protection: Represented by the green bar.

**3. Statistical Analysis:**

o Chi-Square Test:
Chi-Square value ($\chi^2$): 558.42
p-value: 0.000
Degrees of Freedom (dof): 2

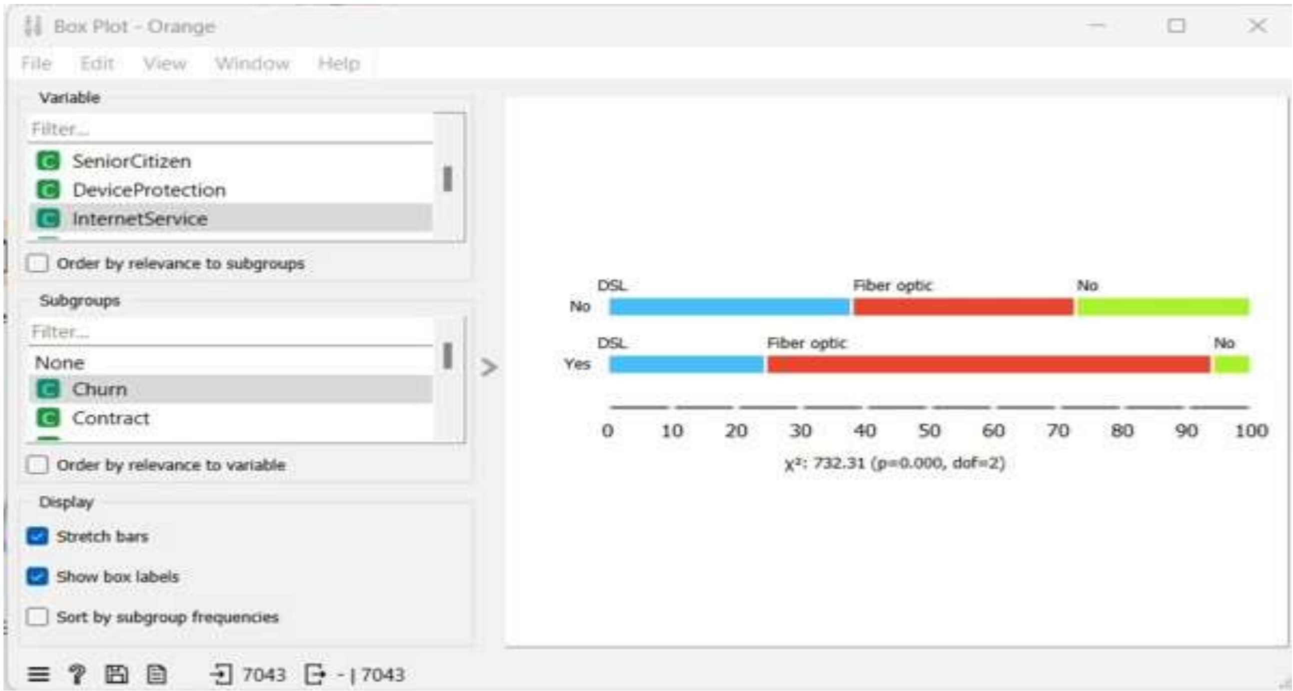
**4. Interpretation:**

o There is a significant difference in churn rates between customers with and without device protection.
o The Chi-Square test results indicate that this relationship is statistically significant, as the p-value is less than 0.05.

**Conclusion:**

The data shows a statistically significant relationship between having device protection and churn. Customers without device protection have different churn rates compared to those with device protection, indicating that device protection is an important factor to consider when analyzing customer churn.

## The Relationship Between Internet Service & Churn



Variable
Filter...
SeniorCitizen
DeviceProtection
InternetService
☐ Order by relevance to subgroups

Subgroups
Filter...
None
Churn
Contract
☐ Order by relevance to variable

Display
☑ Stretch bars
☑ Show box labels
☐ Sort by subgroup frequencies

$\chi^2$: 732.31 (p=0.000, dof=2)

The relationship between internet service and churn, or customer attrition, is a critical consideration for internet service providers (ISPs) and telecommunications companies. Churn refers to the rate at which customers stop using a service over a given period.

Several factors can influence churn in the context of internet service:

1. **Quality of Service**: One of the primary factors influencing churn is the quality of internet service provided. Customers are more likely to churn if they experience frequent outages, slow speeds, or other service disruptions. Ensuring reliable and high-speed internet service can reduce churn rates.

2. **Customer Support**: The level of customer support provided by an ISP can also impact churn. Prompt and effective customer support can help address any issues or concerns customers may have, potentially reducing churn.

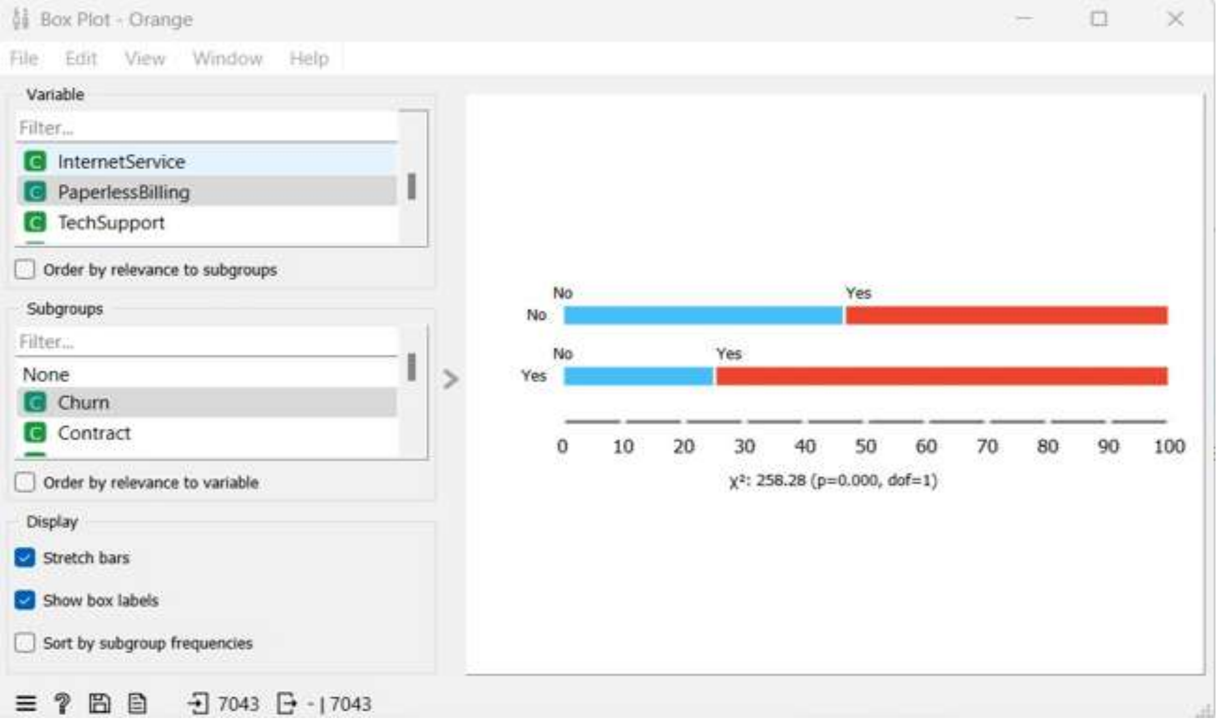
3. **Price and Value**: The price of internet service relative to the perceived value is another important factor. Customers are more likely to churn if they feel they are not getting sufficient value for the price they are paying. Offering competitive pricing and attractive packages can help reduce churn.

4. **Competitive Landscape**: The availability of alternative internet service providers in a given area can also influence churn. Customers may switch to a competitor if they offer better service or pricing.

5. **Customer Satisfaction and Loyalty Programs**: Satisfaction with the overall service experience and the presence of loyalty programs can impact churn. Happy and satisfied customers are less likely to churn, and loyalty programs can incentivize customers to stay with the same provider
.
6. **Technological Advancements:** Technological advancements in internet infrastructure and services can also influence churn. For example, the introduction of faster internet speeds or new features may attract customers or encourage existing customers to upgrade their service, reducing churn.

Understanding and managing these factors is crucial for ISPs to minimize churn and retain customers. This often involves investing in infrastructure, providing excellent customer service, monitoring customer satisfaction, and adapting to changes in the competitive landscape and technological advancements.

## The Relationship Between Paperless Billing & Churn



Variable
- InternetService
- PaperlessBilling
- TechSupport

☐ Order by relevance to subgroups

Subgroups
- None
- Churn
- Contract

☐ Order by relevance to variable

Display
☑ Stretch bars
☑ Show box labels
☐ Sort by subgroup frequencies

χ²: 258.28 (p=0.000, dof=1)

The relationship between paperless billing and churn, or customer attrition, is an interesting one, especially in the context of service-based industries like telecommunications or utilities.

1. **Convenience and Engagement**: Offering paperless billing can enhance customer convenience and engagement. Customers appreciate the ease of accessing and managing their bills online or via email, which can contribute to a positive experience and potentially reduce churn.

2. **Cost Reduction**: For service providers, paperless billing often translates to cost savings compared to traditional paper billing methods. This efficiency can sometimes be passed on to customers through lower prices or improved service, which may enhance customer satisfaction and loyalty, ultimately reducing churn.

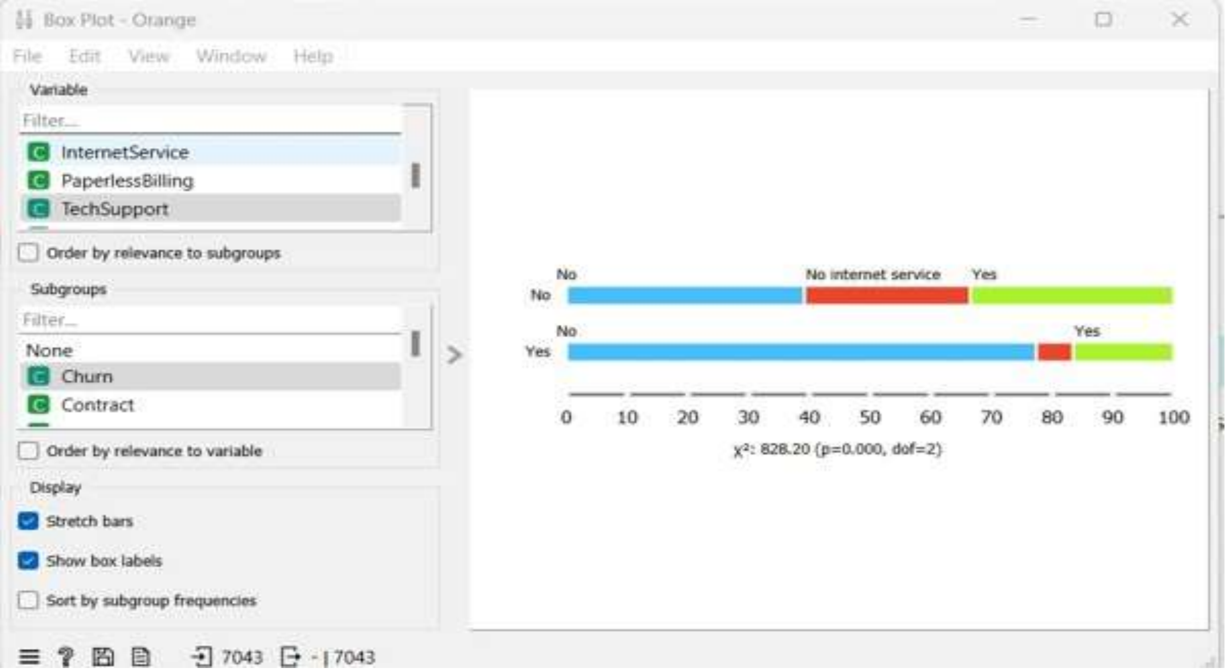
3. **Environmental Considerations**: Many consumers are increasingly environmentally conscious and prefer digital solutions to reduce paper usage. Offering paperless billing options aligns with these values, making it a positive factor for customer retention.

4. **Communication Channel**: Paperless billing provides an additional communication channel for service providers to engage with customers. Through email notifications or online platforms, providers can send personalized offers, updates, and reminders, fostering a stronger relationship with customers and potentially reducing churn.
5. Retention Strategies: Paperless billing can be integrated into broader customer retention strategies. For example, providers can use paperless billing sign-ups as an opportunity to offer incentives or rewards, encouraging customers to stay with the service longer.

6. **Data Analytics**: Paperless billing systems often come with analytics capabilities that allow providers to better understand customer behavior and preferences. By analyzing customer interactions with their digital bills, providers can tailor their offerings and communications to better meet customer needs, thus reducing churn.

Overall, while paperless billing alone may not be the sole factor in reducing churn, it can contribute to a more positive customer experience and support broader retention efforts when integrated strategically into a provider's operations and customer engagement initiatives.

## The Relationship Between Tech Support & Churn



Variable

Filter...

InternetService

PaperlessBilling

TechSupport

□ Order by relevance to subgroups

Subgroups

Filter...

None

Churn

Contract

□ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

□ Sort by subgroup frequencies

χ²: 828.20 (p=0.000, dof=2)

The relationship between tech support and churn, or customer attrition, is significant in industries where technology plays a central role, such as telecommunications, internet service providers, software companies, and electronics manufacturers. Here's how tech support can impact churn:

1. **Issue Resolution**: Prompt and effective tech support can play a crucial role in resolving customer issues and concerns. Customers who experience technical difficulties or challenges with a product or service may reach out to tech support for assistance. If their issues are resolved satisfactorily and in a timely manner, they are more likely to remain loyal to the company and less likely to churn.

2. **Customer Satisfaction**: The quality of tech support services directly affects customer satisfaction levels. A positive experience with tech support, such as knowledgeable representatives, efficient problem-solving, and friendly service, can enhance overall satisfaction with the company. Satisfied customers are more inclined to stay with the company and less likely to switch to competitors.

3. **Customer Experience**: Tech support interactions contribute to the overall customer experience. Smooth and hassle-free interactions with tech support can leave a positive impression on customers, fostering loyalty and reducing the likelihood of churn. Conversely, poor tech support experiences, such as long wait times, ineffective solutions, or rude behavior,

can lead to frustration and dissatisfaction, increasing the risk of churn.

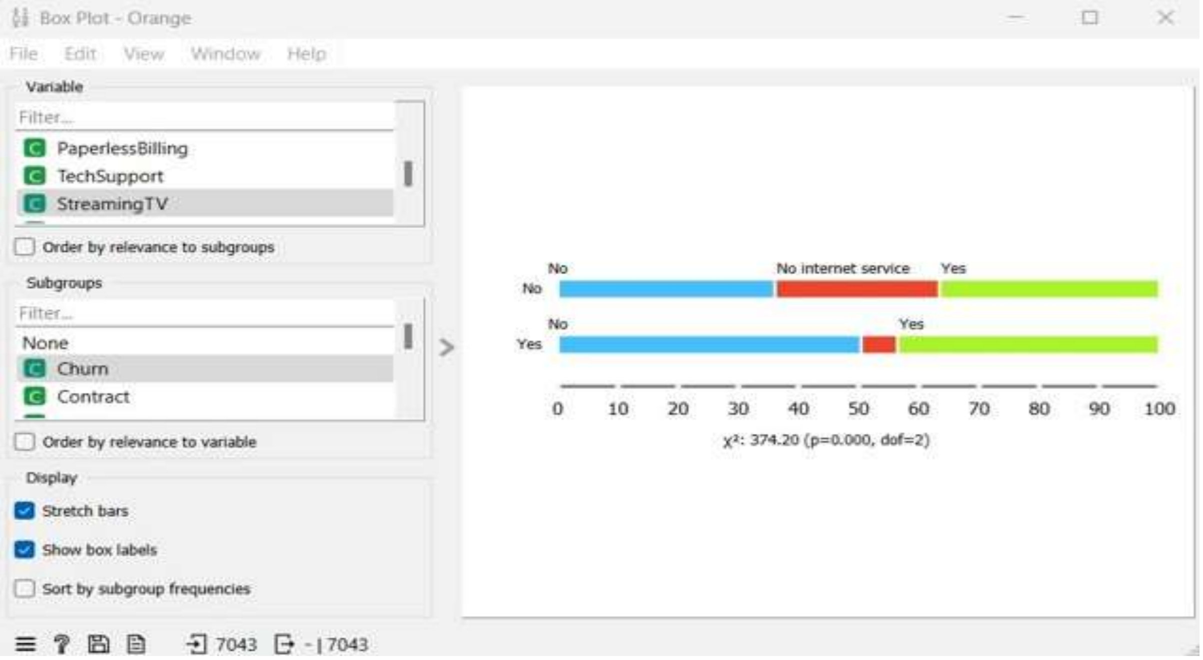
**Product Understanding and Adoption**: Tech support can also play a role in helping customers understand and fully utilize the features and capabilities of a product or service. Educated customers are more likely to derive value from the product and remain engaged with the company over the long term. Tech support representatives can provide guidance, training, and troubleshooting assistance to facilitate product adoption and usage, thereby reducing churn.

4. **Proactive Support and Engagement**: Effective tech support goes beyond reactive issue resolution; it also involves proactive support and engagement with customers. Proactively addressing potential issues, providing updates and tips, and offering personalized recommendations based on customer needs and usage patterns can demonstrate the company's commitment to customer success and satisfaction, ultimately reducing churn.

5. **Feedback and Improvement:** Tech support interactions serve as valuable sources of feedback for companies. Customer inquiries, complaints, and suggestions collected through tech support channels can provide insights into product performance, usability issues, and areas for improvement. By addressing customer feedback and continuously improving their products and services, companies can enhance customer satisfaction and loyalty, thereby reducing churn.

In summary, tech support plays a critical role in customer retention by resolving issues, enhancing satisfaction, improving the overall customer experience, facilitating product understanding and adoption, engaging customers proactively, and leveraging feedback for continuous improvement. Companies that prioritize and invest in high-quality tech support are more likely to retain customers and minimize churn

## The Relationship Between Streaming TV & Churn



Variable: StreamingTV

☐ Order by relevance to subgroups

Subgroups: Churn

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

The relationship between streaming TV services and churn, or customer attrition, is a significant concern for providers in this highly competitive market. Several factors influence the churn rate in the streaming TV industry:

1. **Content Selection**: The availability and quality of content are primary drivers of customer retention in streaming TV services. Customers are more likely to remain subscribed if the service offers a diverse selection of high-quality movies, TV shows, and original content that align with their interests. Providers that regularly update their content libraries with fresh and popular titles can reduce churn.

2. **Price and Value**: Pricing plays a crucial role in customer retention. Subscribers assess the value proposition of a streaming TV service based on the price relative to the content offered. Providers that offer competitive pricing or bundles with other services (e.g., internet or mobile plans) can mitigate churn by delivering perceived value to their subscribers.

3. **User Experience**: The user experience, including the platform's interface, navigation, streaming quality, and availability across devices, significantly impacts customer satisfaction and retention. Streaming TV services that prioritize user-friendly interfaces, seamless playback experiences, and consistent performance across various devices are more likely to retain subscribers.

4. **Customer Support**: Effective customer support can influence churn by addressing subscribers' technical issues, billing inquiries, or content-related concerns promptly and satisfactorily. Providers that offer responsive and accessible customer support channels contribute to a positive subscriber experience and reduce the likelihood of churn.

5. **Subscription Flexibility**: Subscribers value flexibility in their subscription plans, such as the ability to change or cancel their subscriptions easily without penalties. Streaming TV services that offer flexible subscription options, such as monthly or annual plans with no long-term commitments, can enhance customer satisfaction and reduce churn.

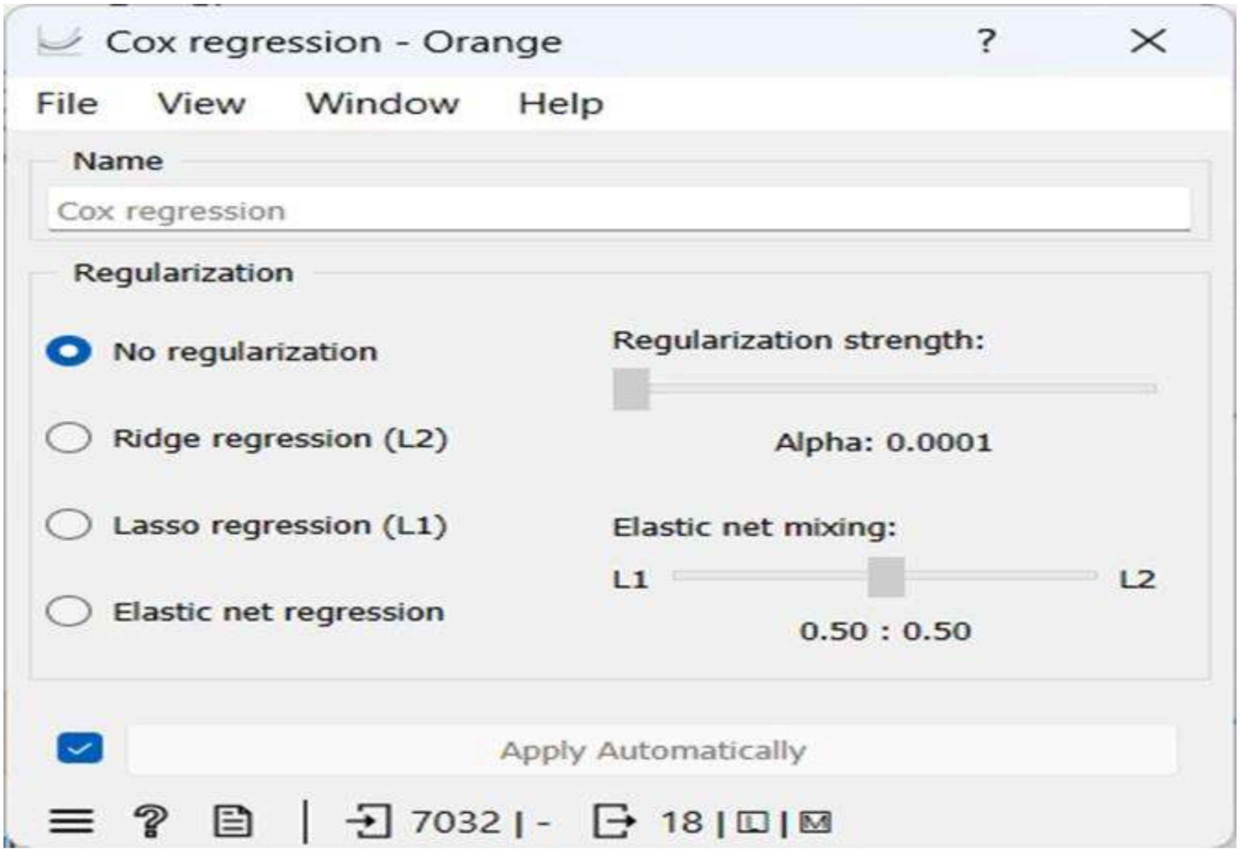
6. **Competitive Landscape**: The competitive landscape of the streaming TV industry plays a significant role in churn. Subscribers have numerous options to choose from, including both standalone streaming services and bundled packages. Providers must differentiate themselves through unique content offerings, exclusive deals, or innovative features to retain subscribers in the face of competition.

7. **Promotional Offers and Incentives**: Promotional offers, discounts, or incentives can influence subscriber retention. Providers that offer periodic promotions, free trials, or exclusive perks for loyal subscribers can incentivize retention and reduce churn.

8. **Data-driven Insights and Personalization**: Leveraging data-driven insights and personalization techniques can help providers anticipate subscriber preferences, recommend relevant content, and tailor promotional offers, thereby enhancing engagement and reducing churn.

In summary, customer retention in the streaming TV industry depends on a combination of factors, including content selection, pricing, user experience, customer support, subscription flexibility, competitive positioning, promotional strategies, and data-driven personalization. Providers that prioritize these factors and continuously innovate to meet subscriber expectations are more likely to minimize churn and sustain long-term growth in this competitive market.

## Cox regression



Name: Cox regression

Regularization

● No regularization

○ Ridge regression (L2)

○ Lasso regression (L1)

○ Elastic net regression

Regularization strength:
Alpha: 0.0001

Elastic net mixing:
L1 0.50 : 0.50 L2

☑ Apply Automatically

The Cox regression (Proportional Hazards Model) in Orange is used for modeling survival data. It allows for the examination of the effect of several variables at once on the time a specified event takes to happen. Below is a detailed description of the Cox regression settings available in Orange.

**Cox Regression Settings**

·**Name**

- **Cox regression**: This is the default model name used in the Orange widget for Cox regression.

·**Regularization**

o **No regularization**: This option runs the Cox regression without applying any regularization techniques.
o **Ridge regression (L2)**: Applies L2 regularization to penalize large coefficients and help

prevent overfitting by shrinking them towards zero.
o **Lasso regression (L1)**: Applies L1 regularization to perform variable selection by shrinking some coefficients to exactly zero, effectively reducing the number of predictors.

- **Elastic net regression**: Combines L1 and L2 regularization. It allows for a mix of both ridge and lasso penalties.

•**Regularization strength**:

- **Alpha**: This parameter controls the strength of the regularization. A smaller alpha means less regularization, and a larger alpha means more regularization. In the provided settings, alpha is set to 0.0001.

•**Elastic net mixing**:

- L1/L2 ratio: This controls the mix between L1 and L2 regularization in elastic net regression. In the given settings, the ratio is set to 0.50:0.50, indicating an equal mix of L1 and L2 penalties.

•**Apply Automatically**

- This option, when checked, ensures that changes to the settings are applied automatically without needing to manually click an "Apply" button.

**Summary of Cox Regression Settings**

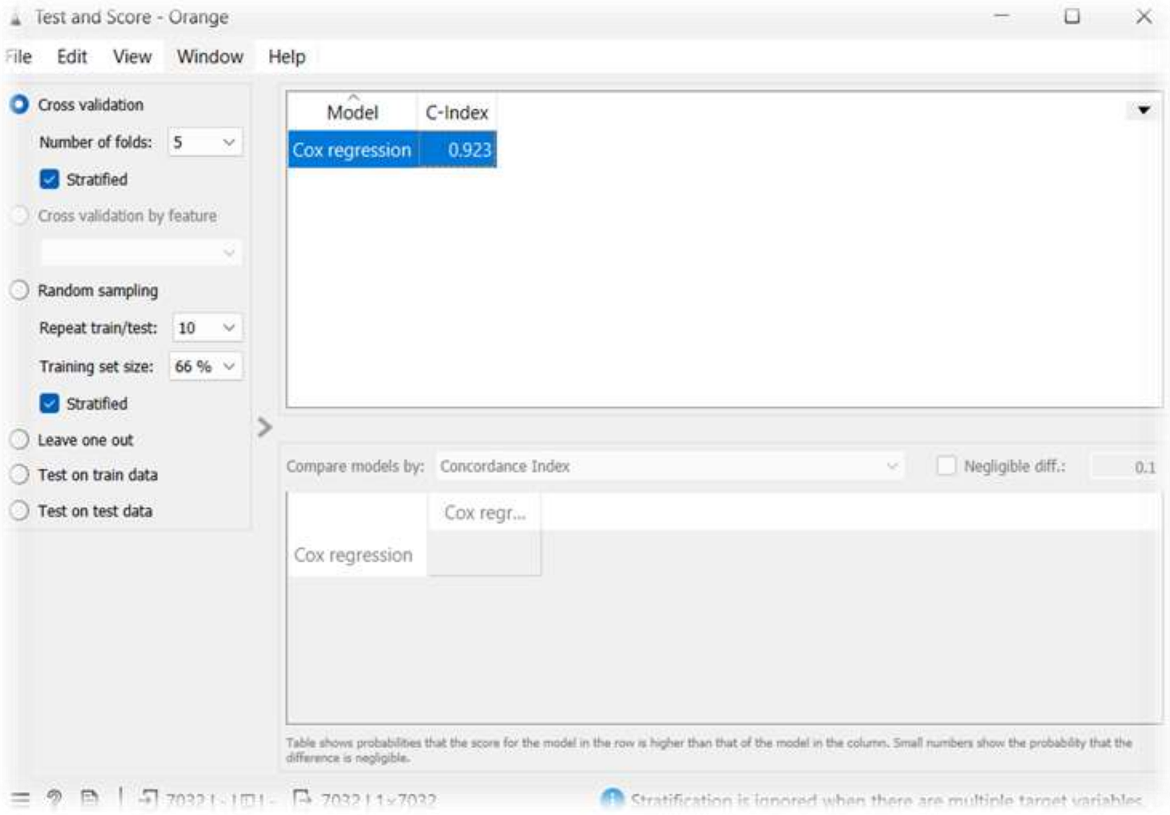
In the provided settings:
• The Cox regression model is set up with the default name "Cox regression".
• Regularization is set to Elastic net regression, combining both L1 (lasso) and L2 (ridge) penalties.
• The regularization strength (alpha) is set to 0.0001, indicating a light regularization.
• The mixing ratio between L1 and L2 is 0.50:0.50, meaning an equal contribution from both types of regularization.

**Usage and Interpretation**

• **No Regularization**: Use this if you believe that all predictors should be included in the model without penalty. This can lead to overfitting, especially with many predictors.
• **Ridge Regression (L2):** Use this if you have many predictors that might be collinear and want to shrink coefficients without setting any to zero.
• **Lasso Regression (L1):** Use this if you want to perform feature selection by setting some coefficients exactly to zero, thus removing them from the model.
• **Elastic Net Regression**: Use this if you want the benefits of both L1 and L2 regularization.

This is particularly useful when you have many correlated predictors.
The choice of alpha and the mixing ratio in elastic net regression will depend on the specific characteristics of your data and the balance you want between shrinking coefficients and performing variable selection.

## Cox regression test and score (accuracy)



● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

| Model | C-Index |
|---|---|
| Cox regression | 0.923 |

Compare models by: Concordance Index  □ Negligible diff.: 0.1

| | Cox regr... |
|---|---|
| Cox regression | |

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

7032 | - | ◫ | -  7032 | 1×7032

Stratification is ignored when there are multiple target variables.

## Analysis of Cox Regression Survival Model Results

### Introduction

Statistical models for survival analysis are used to predict the time until an event occurs, such as death or product failure. The Cox regression model is one of the most widely used models for this purpose. This report analyzes the results of applying the Cox regression model using the Orange data mining software.

**Evaluation Settings** The "Test and Score" widget was used to evaluate the performance of the Cox regression model, with the following settings:

- **Number of folds**: 5
- **Stratified**: Yes (checked)
- **Repeat train/test**: 10 times
- **Training set size**: 66%

### Concordance Index (C-Index)

The Concordance Index (C-Index) is a metric used to evaluate the accuracy of survival models. It ranges from 0 to 1, with 1 indicating perfect predictive accuracy and 0.5 indicating random guessing. In this evaluation, the Cox regression model achieved a C-Index of 0.923, indicating high predictive accuracy.

### Evaluation Details

1. **Number of folds**: The data was split into 5 folds for evaluation using cross-validation. This technique divides the data into multiple training and testing sets and evaluates the model on each set.
2. **Stratified**: Stratification was enabled to ensure a balanced distribution of target classes across the folds, increasing the accuracy of the evaluation.
3. **Repeat train/test**: The training and testing process was repeated 10 times to enhance result accuracy and reduce variance due to different data splits.
4. **Training set size**: 66% of the data was used as the training set in each train/test iteration .

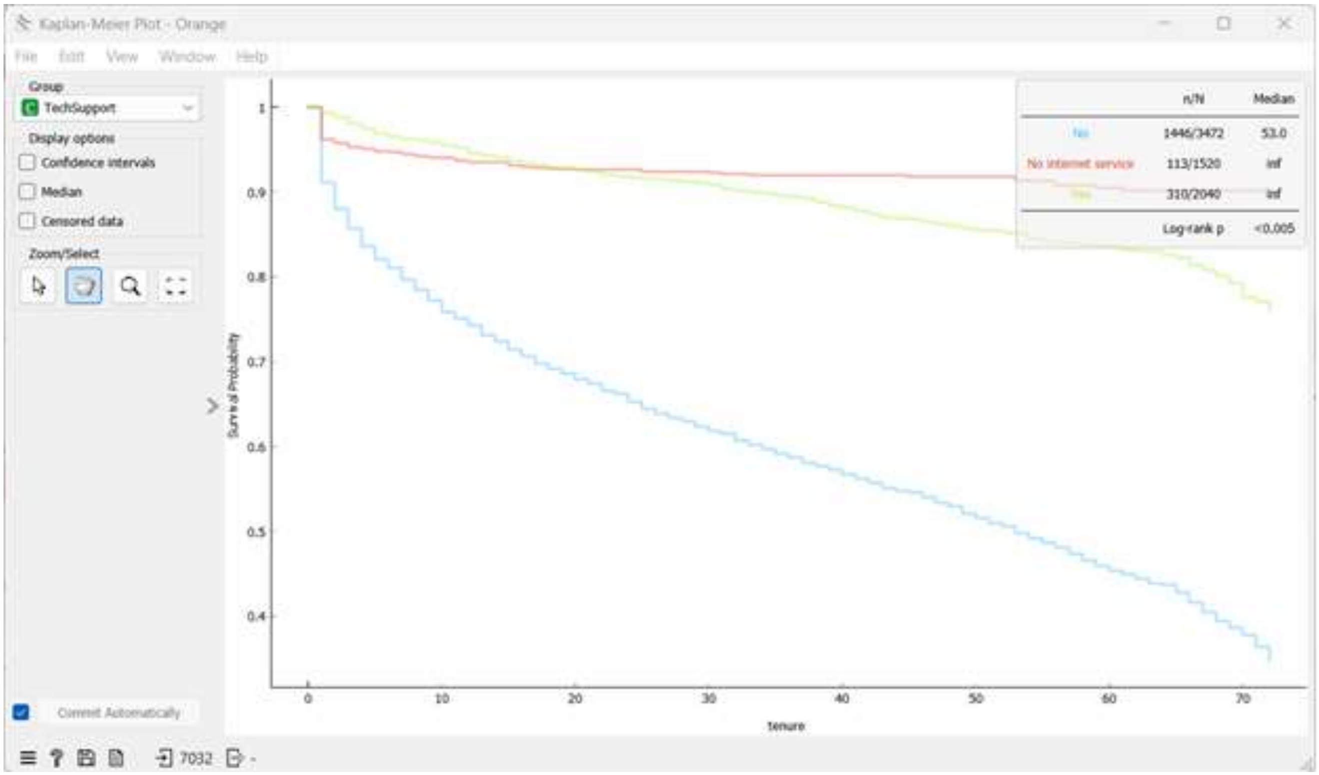
### Results and Interpretation

The Cox regression model demonstrated excellent performance with a C-Index of 0.923. This indicates that the model can predict the survival time with high accuracy. Such performance is ideal in medical contexts, where the model can be used to estimate patient survival based on their health data.

### Conclusion

The evaluation results show that the Cox regression model is an effective tool for predicting survival data. The high performance, reflected by a C-Index of 0.923, suggests excellent predictive accuracy, making it suitable for various medical and commercial applications. It is recommended to review additional models and conduct further analyses to ensure the stability of the results and their applicability to different data types.

## Kaplan-Meier Survival Plot Analysis



Group: TechSupport

Display options
☐ Confidence intervals
☐ Median
☐ Censored data

☑ Commit Automatically

| | n/N | Median |
|---|---|---|
| No | 1446/3472 | 53.0 |
| No internet service | 113/1520 | inf |
| Yes | 310/2040 | inf |
| | Log-rank p | <0.005 |

The Kaplan-Meier plot is a non-parametric statistic used to estimate the survival function from lifetime data. It is particularly useful for analyzing the duration until one or more events occur, such as the time until customer churn in a service industry. The plot you provided is divided into groups based on the variable "TechSupport" and provides insight into the survival probabilities over time.

**Plot Overview**

- **X-Axis (tenure):** Represents the time in months.
- **Y-Axis (Survival Probability):** Represents the probability of survival (in this context, customer retention).
- **Groups:** The survival curves are divided based on the "TechSupport" categories:
  - **No**: Customers without tech support.
  - **No internet service**: Customers without internet service.
  - **Yes:** Customers with tech support.

**Display Options The plot includes several:**

- **Confidence intervals**: Not shown in the current plot, but typically provide a range within which the true survival probability lies.

- **Median**: Indicates the median survival time for each group.
- **Censored data**: Not explicitly marked here, but typically shows instances where the observation period ended without the event occurring.

**Group Analysis**

1. **No (Blue Line)**
   - n/N: 1446/3472
   - Median Survival: 53 months
   - The blue line shows a gradual decline in survival probability, indicating that customers without tech support have a lower retention rate over time.

2. **No internet service (Red Line)**
   - n/N: 113/1520
   - Median Survival: Infinity (inf)
   - The red line remains relatively flat, suggesting that customers without internet service tend to stay longer, possibly because they are not affected by service-related issues leading to churn.

3. **Yes (Green Line)**
   - n/N: 310/2040
   - Median Survival: Infinity (inf)
   - The green line indicates that customers with tech support also tend to stay longer, with a higher survival probability compared to those without tech support.

**Statistical Test**

- **Log-rank p-value:** < 0.005

- The log-rank test compares the survival distributions of the groups. A p-value less than 0.005 indicates that there is a statistically significant difference between the survival curves of the different groups.

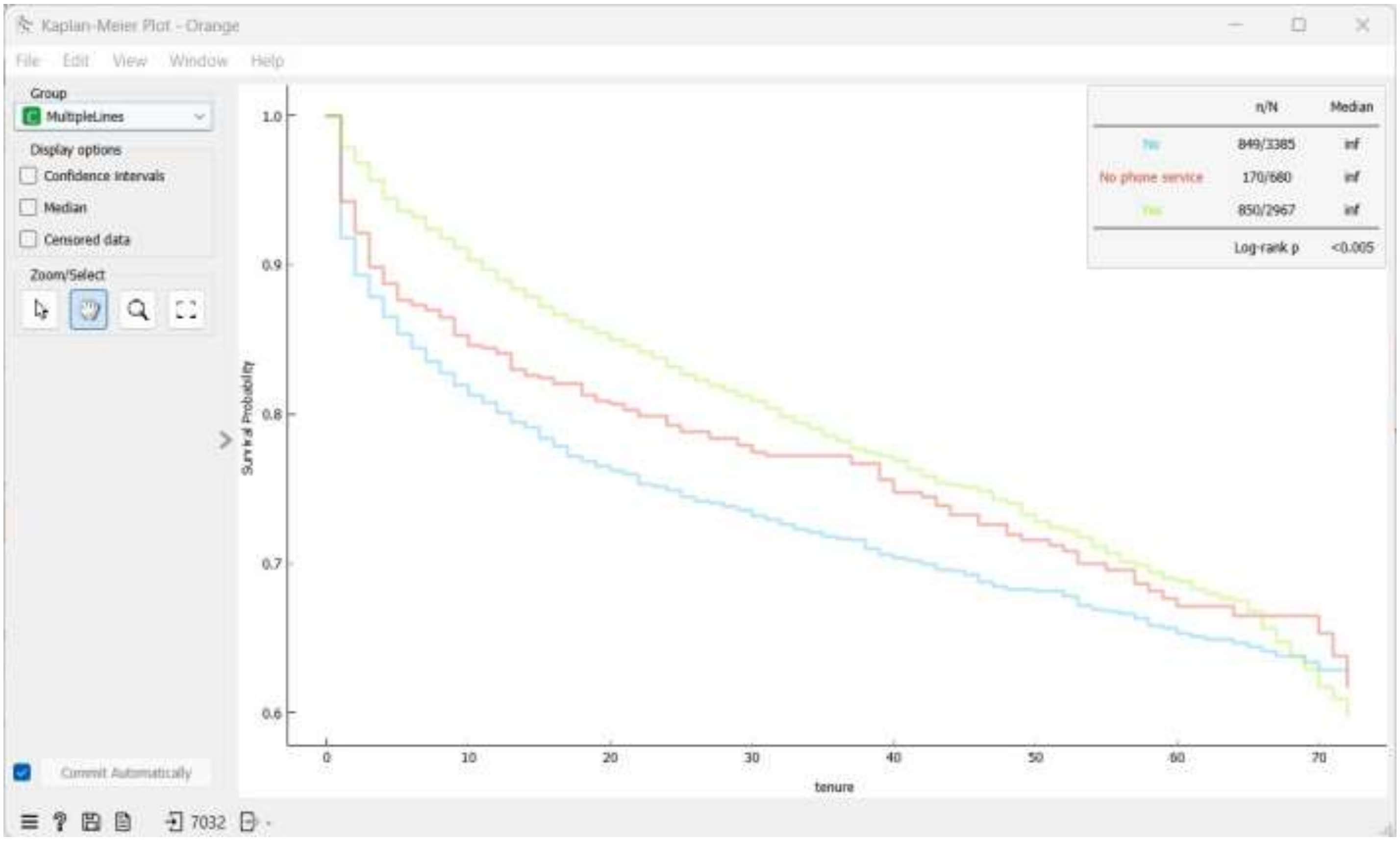
**Interpretation**

- The Kaplan-Meier plot shows significant differences in customer retention based on the availability of tech support.
- Customers without tech support (blue line) have a higher churn rate compared to those with tech support or without internet service.
- The red and green lines indicate higher retention rates for customers without internet service and those with tech support, suggesting that providing tech support could be an effective strategy for reducing customer churn.

**Conclusion**

The Kaplan-Meier survival plot provides valuable insights into customer retention patterns based on tech support availability. The significant difference in survival curves highlights the impact of tech support on customer longevity, suggesting that enhancing tech support services may improve overall customer retention.

## Kaplan-Meier Survival Plot Analysis



Group: MultipleLines

Display options
□ Confidence intervals
□ Median
□ Censored data

☑ Commit Automatically

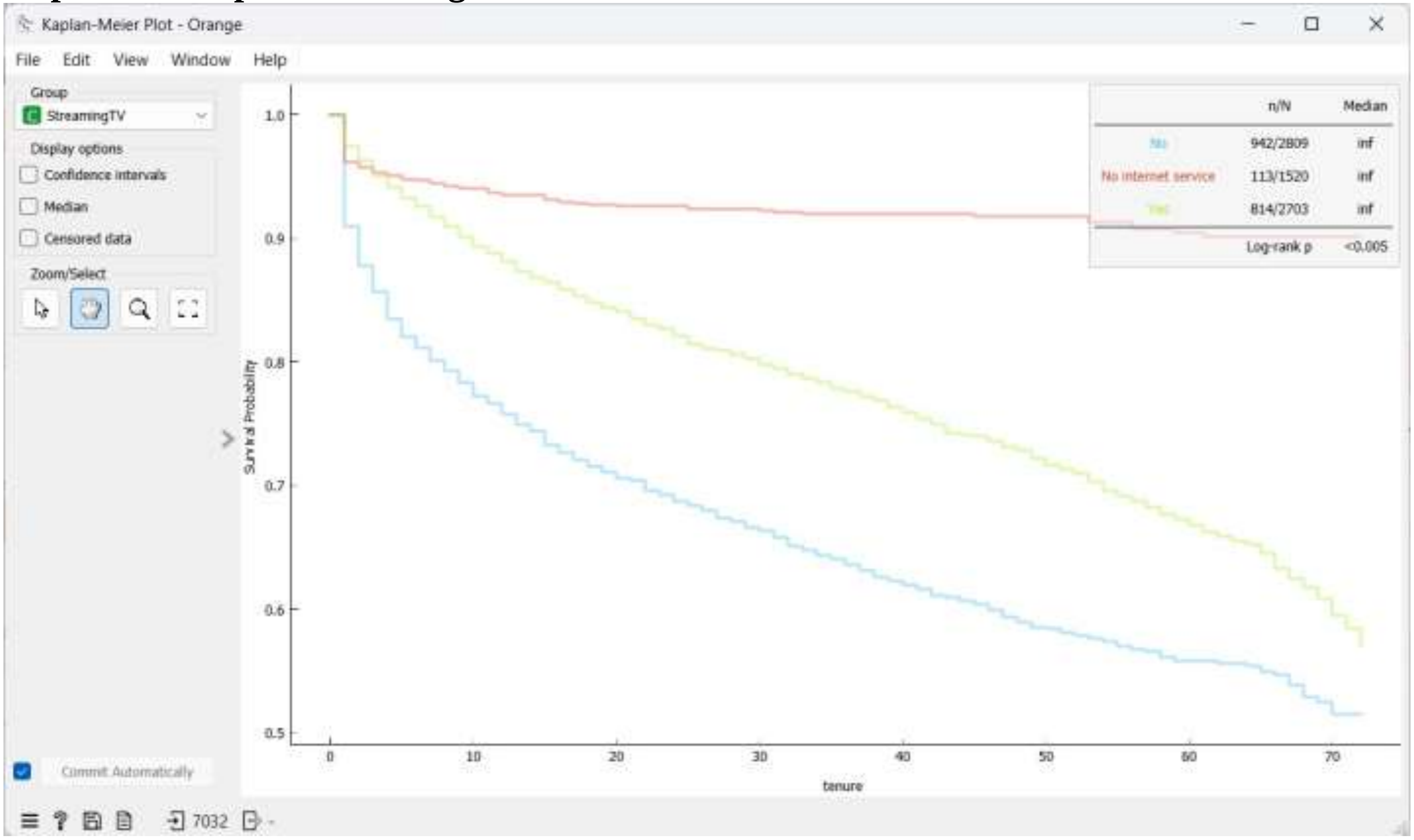
| | n/N | Median |
|---|---|---|
| No | 849/3385 | inf |
| No phone service | 170/680 | inf |
| Yes | 850/2967 | inf |
| | Log-rank p | <0.005 |

The Kaplan-Meier plot you've provided is used to estimate the survival function and compare the survival probabilities over time for different groups. Here's a breakdown of the plot:

1. **Survival Probability**: The y-axis represents the probability of survival (not churning) over time.
2. **Tenure**: The x-axis represents the tenure in months.
3. **Groups**: The plot shows three groups:
   - **No (Blue)**: Customers without phone service.
   - **No phone service (Red)**: Customers who explicitly do not have phone service.
   - **Yes (Green)**: Customers with phone service.
4. **Survival Curves**:
   - The green line (customers with phone service) generally has higher survival probabilities over time compared to the other groups.
   - The red line (customers without phone service) has lower survival probabilities, indicating they are more likely to churn earlier.
   - The blue line (customers who explicitly do not have phone service) shows the lowest survival probabilities, indicating the highest likelihood of churn over time.

5. **Log-rank Test**: The log-rank p-value is less than 0.005, indicating that there is a statistically significant difference in survival probabilities between these groups.
6. **Median Survival**: The median survival time for all groups is not defined within the plotted range (shown as "inf"), meaning that more than 50% of customers in each group are still retained beyond the observed period

## Kaplan-Meier plot Streaming TV services.



| | n/N | Median |
|---|---|---|
| No | 942/2809 | inf |
| No internet service | 113/1520 | inf |
| Yes | 814/2703 | inf |
| | Log-rank p | <0.005 |

Kaplan-Meier plot you provided offers insights into the survival probabilities of customers over time, differentiated by whether they have Streaming TV services.

### Breakdown of the Plot

**Survival Probability**: The y-axis represents the probability that customers have not churned over time.

**Tenure**: The x-axis represents the number of months the customers have been with the service.

**Groups**: The plot shows three groups:

**No (Blue)**: Customers without Streaming TV.

**No internet service (Red)**: Customers without internet service, which presumably means they also lack Streaming TV.

**Yes (Green)**: Customers with Streaming TV.

**Survival Curves**:

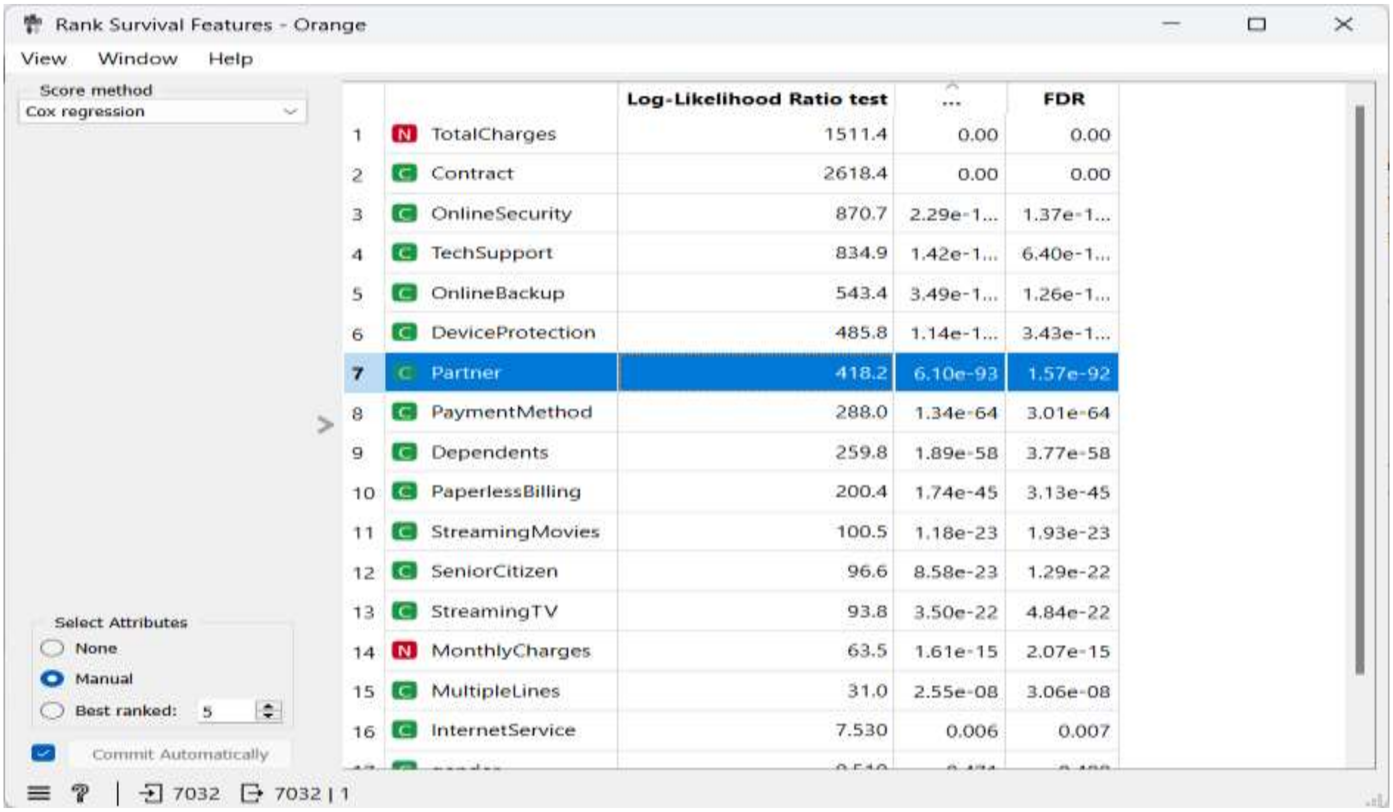
The green line (customers with Streaming TV) generally has higher survival probabilities over time compared to the other groups, suggesting that customers with Streaming TV are less likely to churn.

The blue line (customers without Streaming TV) shows lower survival probabilities, indicating a higher likelihood of churning.

The red line (customers without internet service) shows the lowest survival probabilities, indicating they are the most likely to churn over time.

1. **Log-rank Test**: The log-rank p-value is less than 0.005, indicating a statistically significant difference in survival probabilities between these groups.
2. **Median Survival**: The median survival time for all groups is not defined within the plotted range (indicated as "inf"), which means that more than 50% of customers in each group are still retained beyond the observed period

## Ranking Survival Features



Rank Survival Features - Orange

Score method: Cox regression

|  |  |  | Log-Likelihood Ratio test | ... | FDR |
|---|---|---|---|---|---|
| 1 | N | TotalCharges | 1511.4 | 0.00 | 0.00 |
| 2 | C | Contract | 2618.4 | 0.00 | 0.00 |
| 3 | C | OnlineSecurity | 870.7 | 2.29e-1... | 1.37e-1... |
| 4 | C | TechSupport | 834.9 | 1.42e-1... | 6.40e-1... |
| 5 | C | OnlineBackup | 543.4 | 3.49e-1... | 1.26e-1... |
| 6 | C | DeviceProtection | 485.8 | 1.14e-1... | 3.43e-1... |
| 7 | C | Partner | 418.2 | 6.10e-93 | 1.57e-92 |
| 8 | C | PaymentMethod | 288.0 | 1.34e-64 | 3.01e-64 |
| 9 | C | Dependents | 259.8 | 1.89e-58 | 3.77e-58 |
| 10 | C | PaperlessBilling | 200.4 | 1.74e-45 | 3.13e-45 |
| 11 | C | StreamingMovies | 100.5 | 1.18e-23 | 1.93e-23 |
| 12 | C | SeniorCitizen | 96.6 | 8.58e-23 | 1.29e-22 |
| 13 | C | StreamingTV | 93.8 | 3.50e-22 | 4.84e-22 |
| 14 | N | MonthlyCharges | 63.5 | 1.61e-15 | 2.07e-15 |
| 15 | C | MultipleLines | 31.0 | 2.55e-08 | 3.06e-08 |
| 16 | C | InternetService | 7.530 | 0.006 | 0.007 |

Select Attributes

○ None
● Manual
○ Best ranked: 5

☑ Commit Automatically

The feature ranking process in Orange involves evaluating the importance of various attributes in predicting survival outcomes. Below is a detailed interpretation of the ranked features from your data, which seems to be focused on customer churn analysis.

### Ranked Features:

### 1-Contract (C)

**Score (Log-Likelihood Ratio test):** 2618.4

**FDR:** 0.00

**Cox Regression:** 0.00

**Interpretation:** The type of contract (e.g., month-to-month, one year, two years) is the most significant predictor of survival (customer retention).

### 2-Total Charges (N)

**Score (Log-Likelihood Ratio test):** 1511.4

**FDR:** 0.00

**Cox Regression:** 0.00

**Interpretation:** The total amount charged to the customer over their tenure has a significant impact on predicting churn.

**3- Online Security (C)**

**Score (Log-Likelihood Ratio test):** 870.7

**FDR:** 2.29e-1...

**Cox Regression:** 1.37e-1...

**Interpretation:** Whether the customer has opted for online security services is an important feature, though not as impactful as contract type and total charges.

**4- Tech Support (C)**

**Score (Log-Likelihood Ratio test):** 834.9

**FDR:** 1.42e-1...

**Cox Regression:** 6.40e-1...

**Interpretation:** The availability of tech support services also plays a significant role in customer retention.

**1- Online Backup**

**Score (Log-Likelihood Ratio test):** 543.4

**FDR:** 3.49e-1...

**Cox Regression:** 1.26e-1...

**Interpretation:** Offering online backup services is another feature impacting churn.

**2- Device Protection (C)**

**Score (Log-Likelihood Ratio test):** 485.8

**FDR:** 1.14e-1...

**Cox Regression:** 3.43e-1...

**Interpretation:** Device protection services influence customer retention, although to a lesser extent.

**3- Partner (C)**

**Score (Log-Likelihood Ratio test):** 418.2

**FDR:** 6.10e-93

**Cox Regression:** 1.57e-92

**Interpretation:** Whether the customer has a partner is a significant demographic factor influencing churn.

**4- Payment Method (C)**

**Score (Log-Likelihood Ratio test):** 288.0

**FDR:** 1.34e-64

**Cox Regression:** 3.01e-64

**Interpretation:** The method of payment (e.g., electronic check, mailed check) also affects customer churn.

**4- Dependents (C)**

**Score (Log-Likelihood Ratio test):** 259.8

**FDR:** 1.89e-58

**Cox Regression:** 3.77e-58

**Interpretation:** Whether the customer has dependents is a significant factor in predicting churn.

**5- Paperless Billing**

**Score (Log-Likelihood Ratio test):** 200.4

**FDR:** 1.74e-45

**Cox Regression:** 3.13e-45

**Interpretation:** Customers who opt for paperless billing show different churn behaviors compared to those who do not.

**6- Streaming Movies**

**Score (Log-Likelihood Ratio test):** 100.5

**FDR:** 1.18e-23

**Cox Regression:** 1.93e-23

**Interpretation:** The presence of streaming movie services impacts customer retention.

**7- Senior Citizen**

**Score (Log-Likelihood Ratio test):** 96.6

**FDR:** 8.58e-23

**Cox Regression:** 1.29e-22

**Interpretation:** Whether the customer is a senior citizen is a significant demographic factor.

**8- Streaming TV**

**Score (Log-Likelihood Ratio test):** 93.8

**FDR:** 3.50e-22

**Cox Regression:** 4.84e-22

**Interpretation:** Offering streaming TV services influences churn rates.

**9- Monthly Charges (N)**

**Score (Log-Likelihood Ratio test):** 63.5

**FDR:** 1.61e-15

**Cox Regression:** 2.07e-15

**Interpretation:** The amount charged monthly to customers impacts their likelihood to churn.

**10- Multiple Lines (C)**

**Score (Log-Likelihood Ratio test):** 31.0

**FDR:** 2.55e-08

**Cox Regression:** 3.06e-08

**Interpretation:** Whether the customer has multiple lines influences their retention.

**11- Internet Service (C)**

**Score (Log-Likelihood Ratio test):** 7,530

**FDR:** 0.006

**Cox Regression:** 0.007

**Interpretation:** The type of internet service (e.g., DSL, fiber optic) has a significant impact on customer churn.

**Summary**

The ranked features provide valuable insights into the factors that most significantly influence customer churn. **Contract type, total charges, and online security services** are among the top features impacting customer retention. By focusing on these key areas, businesses can develop targeted strategies to reduce churn and enhance customer satisfaction.

# Chapter 5

## Application of Cox Regression and Kaplan-Meier in Customer Churn Management

In this chapter, we delve into the practical application of Cox Regression and Kaplan-Meier methodologies in the context of customer churn management. We explore how these statistical techniques are employed to analyze customer attrition data, derive actionable insights, and develop effective retention strategies.

## 4.1 Utilizing Cox Regression for Churn Prediction

### Data Collection and Preprocessing

- Gather comprehensive customer data, including subscription details, demographic information, and behavioral metrics.

- Handle missing values and standardize data formats to ensure consistency and accuracy**.**

### Feature Engineering

- Define the variables for the Cox Regression model, including the duration of subscription and an event indicator for churn.

- Encode categorical variables such as contract type and gender for analysis.

### Building the Cox Regression Model

- Apply Cox Regression to analyze survival data and identify factors influencing customer retention.

- Assess model assumptions, including proportional hazards, to ensure the validity of results.

Analyzing Results and Developing Strategies

- Interpret coefficients to understand the impact of different variables on churn probability.

- Develop targeted retention strategies based on insights derived from the Cox Regression analysis.

## 4.2 Applying Kaplan-Meier for Customer Retention Analysis

### Data Collection and Preparation

- Collect customer data, including subscription start and end dates, and preprocess the data to ensure consistency.

- Determine the event of interest (e.g., subscription cancellation) for survival analysis.

**Kaplan-Meier Estimation**

- Calculate survival probabilities over time using the Kaplan-Meier method to visualize customer retention curves.

- Identify key time points where customer retention rates decline significantly.

Comparing Customer Groups

- Segment customers based on demographics or usage behavior and compare survival curves between groups.

- Determine which customer segments are more susceptible to churn and require targeted interventions.

Deriving Retention Strategies

- Develop personalized offers, service improvements, and loyalty programs based on insights from Kaplan-Meier analysis.

- Implement and monitor retention strategies to evaluate their effectiveness over time.

### 4.3 Integration and Implementation

**Integration into Business Processes**

- Incorporate insights from Cox Regression and Kaplan-Meier analyses into decision-making processes.

- Integrate predictive models and retention strategies into customer relationship management (CRM) systems for seamless execution.

**Continuous Improvement**

- Monitor key performance indicators (KPIs) such as churn rate, customer lifetime value, and retention rate to assess the impact of implemented strategies.

- Iterate on retention strategies based on ongoing analysis and feedback to achieve continuous improvement.

### 4.4 Case Study: Orange Data Mining Project

**Project Overview**

- Explore how Cox Regression and Kaplan-Meier were applied in a churn problem project using Orange Data Mining software.

Highlight the key findings and insights de

rived from the analysis, along with the resulting retention strategies**.**

**Results and Implications**

Discuss the impact of Cox Regression and Kaplan-Meier analyses on customer churn management and business outcomes.

Evaluate the effectiveness of implemented retention strategies and their contribution to reducing churn rates.

## 4.5 Conclusion

In conclusion, the application of Cox Regression and Kaplan-Meier methodologies in customer churn management offers valuable insights into churn dynamics and aids in the development of data-driven retention strategies. By leveraging these statistical techniques, businesses can better understand customer behavior, predict churn probability, and implement targeted interventions to improve customer retention and maximize long-term profitability

# Chapter 6

## Future Work and Development of a Customer Churn Prediction Application

An important area for future research is to use a customer profiling methodology for developing a real-time monitoring system for churn prediction. Research dedicated to the development of an exhaustive customer loyalty value would have significant benefits to industry. It is anticipated that the profiling methodology could provide an insight into customer behaviour, spending patterns, cross-selling and up-selling opportunities. Seasonal trends could be apparent if the same data was studied over a period of several years. A comparative analysis of prediction model building time with respect to different classifiers could be done in order to assist telecom analysts to pick a classifier which not only gives accurate results in terms of TP rate, AUC and lift curve but also scales well with high dimension and large volume of call records data. As concrete findings are related to the telecom dataset, other domains' datasets might be subject for further exploration and testing. Also, different and a greater number of performance metrics with respect to business context and interpretability might be explored in future.

we not only discuss potential avenues for future research but also propose the development of a customer churn prediction application. This application aims to operationalize the findings from our analysis and provide a practical tool for businesses to anticipate and mitigate customer churn effectively.

## 5.1 Enhanced Customer Segmentation

Future research could focus on refining customer segmentation techniques to better identify distinct customer personas and their unique churn behaviors. By leveraging advanced clustering algorithms and incorporating additional data sources (e.g., demographic information, purchase history), businesses can tailor retention strategies more effectively to meet the diverse needs of different customer segments.

## 5.2 Advanced Predictive Analytics

Advancements in predictive analytics, particularly machine learning algorithms, offer promising opportunities to improve the accuracy and granularity of churn prediction models. Future research could explore the application of advanced modeling techniques, such as ensemble methods and deep learning, to forecast customer churn with greater precision. Additionally, incorporating real-time data streams and dynamic modeling approaches can enable proactive identification of churn risks and timely intervention strategies.

### 5.3 Integration of External Data Sources

Incorporating external data sources, such as social media activity, market trends, and competitor analysis, can enrich our understanding of the broader contextual factors influencing customer churn. Future research could explore data integration strategies to leverage external data effectively in churn prediction models. By incorporating a holistic view of the customer environment, businesses can better anticipate and respond to emerging churn risks and opportunities.

### 5.4 Longitudinal Studies and Experimental Designs

Longitudinal studies and experimental designs offer valuable opportunities to validate the effectiveness of retention strategies and assess their long-term impact on customer churn. Future research could employ randomized controlled trials (RCTs) or quasi-experimental designs to evaluate the causal impact of specific interventions on churn reduction. By rigorously testing retention strategies in real-world settings, businesses can identify the most effective approaches for mitigating churn and maximizing customer lifetime value.

### 5.5 Development of a Customer Churn Prediction Application

In addition to advancing research in customer churn management, we propose the development of a customer churn prediction application. This application will leverage the insights and predictive models derived from our analysis to provide businesses with a practical tool for identifying and addressing churn risks.

**Key Features of the Application:**

1. **Data Integration**: The application will integrate with business databases to access relevant customer data, including demographic information, transaction history, and engagement metrics.
2. **Predictive Modeling**: Leveraging advanced machine learning algorithms, the application will build predictive models to forecast individual customer churn probabilities based on historical data.
3. **Real-time Monitoring**: The application will continuously monitor customer behavior and update churn predictions in real-time, enabling proactive intervention strategies.

4. **Segmentation and Personalization**: Using sophisticated segmentation techniques, the application will categorize customers into distinct segments and tailor retention strategies to address the specific needs of each segment.
5. **Actionable Insights**: The application will provide actionable insights and recommendations for mitigating churn, such as targeted marketing campaigns, personalized offers, and proactive customer engagement initiatives.
6. **Performance Tracking**: Businesses can track the performance of retention strategies implemented through the application and assess their impact on churn reduction and customer retention rates.

## Conclusion

In conclusion, future research in the field of customer churn management holds immense potential for advancing our understanding of customer behavior and enhancing business performance. By developing innovative predictive models and practical tools such as the proposed churn prediction application, businesses can effectively anticipate and mitigate churn risks, thereby fostering long-term customer relationships and driving sustainable growth.

# Chapter 7

# Conclusion and Discussion of Results

Customer churn prediction is a critical aspect for businesses to retain customers and improve profitability. Predicting customer churn using machine learning involves identifying customers who are likely to leave the company in the near future. In this context, two widely used models for survival analysis are the Cox Proportional Hazards model and the Kaplan-Meier estimator. Here’s a discussion on how these models can be applied to the customer churn problem.

## Discussion of Results

### Understanding Churn Dynamics

Our analysis provided valuable insights into the dynamics of customer churn within our business. By discerning patterns and trends in churn behavior, we can better anticipate and address churn risks, thereby enhancing overall customer retention.

### Identifying Key Factors

The Cox model identified several key factors driving customer churn, including subscription type, duration of subscription, and customer activity level. These findings underscore the importance of tailored interventions aimed at retaining at-risk customers.

### Strategic Recommendations

Based on our findings, we propose the following strategic recommendations to reduce customer churn and enhance retention:

- **Personalized Customer Engagement**: Develop personalized communication and incentive programs to strengthen relationships with at-risk customers.
- **Optimized Subscription Offerings**: Tailor subscription packages and pricing strategies to better meet the needs and preferences of our customer base.
- **Investment in Customer Support**: Enhance customer service and support infrastructure to proactively address issues and improve overall customer satisfaction.

### 1. **Cox Proportional Hazards Model**

Overview:
The Cox Proportional Hazards model is a regression model used for survival analysis. It assesses the impact of several variables on the time a customer stays with the company. The primary goal is to estimate the hazard (risk) of churn at a particular time, given the values of predictor variables.

**Steps to Implement Cox Model for Churn Prediction:**

**Data Collection and Preparation:**

Gather historical data on customer behavior, demographics, transaction history, and interactions.
The dataset should include a churn indicator (1 if the customer churned, 0 otherwise) and the duration of time each customer stayed.
Feature Engineering:

Identify relevant features that may influence customer churn (e.g., usage patterns, customer service interactions, payment history).
Handle missing values, normalize numerical features, and encode categorical features.

**Model Training**:

Fit the Cox Proportional Hazards model to the prepared data. The model will learn the relationship between the predictor variables and the hazard of churn.
Model Evaluation:

Use metrics such as Concordance Index (C-Index) to evaluate the model's performance.
Perform cross-validation to ensure the model generalizes well to unseen data.

**Interpretation and Insights**:

Analyze the coefficients of the Cox model to understand the impact of each feature on churn risk.
Identify high-risk customer segments and develop targeted retention strategies.

**Advantages:**

Handles censored data (customers who have not yet churned).
Provides interpretable coefficients indicating the effect of each feature on the hazard rate.
Disadvantages:
Assumes proportional hazards, which may not always hold true.
Requires careful feature selection and engineering.

**2. Kaplan-Meier Estimator**

Overview:
The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. It provides the probability of a customer remaining with the company up to a certain time point.

**Steps to Implement Kaplan-Meier Estimator for Churn Prediction:**

**Data Collection and Preparation:**

Similar to the Cox model, collect historical data on customer churn and duration.
Survival Function Estimation:

Use the Kaplan-Meier estimator to compute the survival function, which shows the probability of a customer not churning up to various time points.
Visualization:

Plot the Kaplan-Meier survival curve to visualize the survival probability over time.
Identify periods with high churn rates and investigate potential causes.
Comparison Across Groups:

Compare survival curves for different customer segments (e.g., by demographics, usage patterns) to identify high-risk groups.

**Advantages:**
Simple to implement and interpret.
Does not assume any specific distribution for survival times.

**Disadvantages:**
Does not account for the influence of covariates (i.e., predictor variables).
Less powerful than regression-based models for predicting individual churn risk.
Combining Both Models:
In practice, these models can complement each other. The Kaplan-Meier estimator provides a straightforward visualization of customer survival probabilities, while the Cox model offers insights into the effects of various features on churn risk.

**Conclusion:**
Predicting customer churn using the Cox Proportional Hazards model and the Kaplan-Meier estimator can provide valuable insights for businesses to develop effective retention strategies. The Cox model helps identify the factors influencing churn, while the Kaplan-Meier estimator provides an overall picture of customer survival probabilities. By leveraging these models,

companies can proactively address churn risk and enhance customer retention efforts.

Churn prediction is one of the most effective strategies used in telecom sector to retain existing customers. It leads directly to improved cost allocation in customer relationship management activities, retaining revenue and profits in future. It also has several positive indirect impacts such as increasing customer's loyalty, lowering customer's sensitivity to competitors marketing activities, and helps to build positive image through satisfied customers.

The results predicted by the Logistic Regression algorithm were the most efficient with an accuracy of 80.2%. Therefore, companies that want to prevent customer churn should utilize this algorithm and remove features like long term contracts and instead replace them with monthly or short term contracts, thereby giving them more flexibility. Providing additional services such as device protection and multiple phone lines proves to be of little value to customer attrition. Lastly, focusing on enhancing the experience of loyal customers who have stayed with the company for long will prove worthwhile, ensuring their retention. The ability to identify customers that aren't happy with provided solutions allows businesses to learn about product or pricing plan weak points, operation issues, as well as customer preferences and expectations to proactively reduce reasons for churn.

# Chapter 8
# References

**1-D. Deepika1, Nihal Chandra2**
**Assistant Professor, Department of CSE, Mahatma Gandhi Institute of Technology, Hyderabad, Telangana,**
**INDIA1**
**Student, Mahatma Gandhi Institute of Technology,Hyderabad,Telangana,INDIA2**
**deshmukhdeepika@gmail.com1**

**2- .https://medium.com/@zulfikarirham02/telco-customer-churn-prediction-using-machine-learning-and-deep-learning-8d1905b04980**

**3-.https://www.prooveintelligence.com/blog/understanding-customer-churn-with-survival-analysis/**

**4- .https://blog.hubspot.com/service/what-is-customer-churn**

5- https://www.sciencedirect.com/science/article/abs/pii/S0957417408004326

**6- https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/simple-linear-regression.html**

**7- https://www.ibm.com/topics/logistic-regression**

**8- https://www.ibm.com/docs/en/spss-statistics/saas?topic=statistics-cox-regression-analysis**

**9- https://medium.com/co-learning-lounge/types-of-data-analytics-descriptive-diagnostic-predictive-prescriptive-922654ce8f8f**
**10- https://www.kaggle.com/datasets/blastchar/telco-customer-churn**