MINOR PROJECT

END SEM REPORT

on

Customer Churn Rate Prediction

Submitted By :

Name	Roll No	Branch
Khushleen Kaur	A25305221017	CSE
Anmol Kansal	A25305221009	CSE
Sarthak Raheja	A25305221012	CSE
Lakshay Batta	A25305221027	CSE

Under the guidance of

Dr. Aanshi Bhardwaj

Assistant Professor

Amity School of Engineering and Technology

School of Computer Science

Amity School of Engineering and Technology

Amity University, Mohali

2023-24

Approved By

(Dr. Aanshi Bhardwaj)
Project Mentor

(Dr. Harvinder Singh)
Project Coordinator

ABSTRACT

Understanding consumer behaviour and reducing attrition have become critical for long-term success in today's corporate environment. The rate at which customers stop using a company's products or services is known as customer churn, and it is a major problem for many different businesses, especially in fiercely competitive markets like telecoms, finance, subscriptions (OTT), etc. Telecom companies suffer from negative consequences due to yearly turnover rates of 15–25 percent, which makes it necessary to forecast churn rates and put in place efficient retention tactics. This study aims to predict telecom churn rates through the application of machine learning methodologies and historical data. Utilizing methods like data pre-processing, exploratory analysis, and model selection, we seek to identify clients who may be in danger and take proactive measures to retain them. By analyzing data and applying algorithms like Decision Tree (DT), Logistic Regression (LoR), Random Forest (RF), Support Vector Machine (SVM), and Boosting, we hope to increase customer satisfaction and loyalty. Eventually, this will result in increased revenues. Using assessment parameters such as accuracy, recall, precision, and F1 score, we compare the performance of various ML algorithms and conclude that XG Boost is the most effective model for churn prediction. The present study enhances the efficacy of churn prediction tactics and offers valuable perspectives for enterprises seeking to lower attrition rates and improve client retention.

Keywords: customer churn, machine learning, telecom industry, retention tactics, data pre-processing, exploratory analysis, decision tree, logistic regression, random forest, support vector machine, boosting, evaluation metrics, XG Boost, client retention

TABLE OF CONTENTS

Contents

LIST OF TABLES

List of Tables

LIST OF FIGURES

List of Figures

1 Preface

In today's fast-paced and fiercely competitive business landscape, understanding and retaining customers is paramount for sustained success. Customer churn, the rate at which consumers cease using a company's products or services, poses a substantial obstacle for companies in all sectors. It reflects the loss of valuable clientele and can have adverse effects on revenue and market position. To thrive in this dynamic environment, businesses must give top priority to cultivating close bonds with their customers and proactively addressing churn drivers.

The churn rate, often referred to as the rate of attrition, serves as a key metric for assessing customer retention strategies. While achieving a zero churn rate may be unrealistic, successful companies aim to maintain a growth rate that surpasses the churn rate, signaling an expansion of their customer base. This necessitates a deep understanding of customer needs, preferences, and behaviors. By leveraging data analytics and customer insights, businesses can identify patterns and trends associated with churn, enabling them to implement targeted interventions and improve customer retention efforts.

In essence, minimizing churn and fostering customer loyalty are essential components of a sustainable business strategy. In today's competitive environment, organizations may position themselves for long-term success and profitability by consistently adjusting to changing client needs and providing extraordinary experiences.

1.1 Background

Customer turnover is a common occurrence in service businesses due to intense competition. Looking into today's scenario of telecommunication businesses, they have an annual churn rate of 15-20 percent. The main reason behind this is that customers have so many different service providers to choose from and They can quickly move between different service providers. That's the reason Telecommunication businesses face adverse effects hence it's crucial to predict churn rates and methods to retain them.

The paper authored by Wagh et al. tackles the critical challenge of customer churn prediction in the telecom industry, emphasizing the significance of retaining existing customers amidst high acquisition costs. They propose a classification-based approach leveraging ML algorithms like Random Forest, KNN, and decision tree Classifier to develop predictive models for attrition identification and understanding underlying causes. Their study achieves notable success, attaining a remarkable 99 percent accuracy using the random forest classifier. This research not only advances churn prediction accuracy but also extends its applicability across diverse business sectors, fostering improved customer service and churn prevention strategies. [1]

Amin, Adnan, and Anwar investigated the critical issue in telecom industry of customer churn prediction (CCP) and recognized its pivotal role in business sustainability. Traditional CCP models often lack adaptability to dynamic customer behavior changes. Their study introduces an adaptive learning approach, leveraging the Naive Bayes classifier with a Genetic Algorithm-based feature weighting method. Evaluating public datasets, their approach demonstrates significant performance enhancements over baseline classifiers, indicating its effectiveness in improving prediction accuracy and providing timely insights for effective churn prediction strategies. [2]

The paper by Jitendra Maan and Harsh Maan addresses the pressing issue of turnover among customers in an increasingly digitized market, where subscription-based products as well as services proliferate. Recognizing the high cost of bringing in fresh clients compared to keeping the ones that already exist, the study aims to develop a robust churn prediction model. After evaluating various machine learning approaches, XGBoost emerges as probably the most efficient classifier. Emphasizing model interpretability, the paper suggests a unique method to calculate Shapley values, enhancing transparency and understanding of the model's predictions. The focus on explainable machine learning is pivotal for businesses seeking to mitigate churn risks effectively. Through this research, the authors contribute to the field of churn analysis by providing a comprehensive framework for predictive modeling, rooted in both performance and interpretability metrics. [3]

The paper by Ahmad et al. addresses the critical issue of telecom sector customer attrition and the significance it holds for company revenues. It emphasizes the necessity of prediction algorithms to find possible customers and underscores the significance of choosing and feature engineering in model development. By employing ML methods on large-scale data platforms, the research attains an impressive AUC of 93.3 percent. A key innovation lies in integrating Social Network Analysis (SNA) features, which notably enhances model performance. The research compares multiple algorithms, with XGBoost demonstrating superior classification accuracy, thereby contributing to effective churn prediction strategies in telecom.[4]

The study of Prabadevi underscores the critical importance of client retention for businesses, emphasizing the need for early identification of churn to implement proactive measures. It aims to advise on the optimal ML strategy for early churn prediction, utilizing customer data spanning nine months prior to churn. Various algorithms including stochastic gradient booster, random forest, logistics regression, and k-nearest neighbors are tested, with accuracy ranging from 78.1 percent to 83.9 percent. Through comparative analysis, the research identifies the most effective algorithm, offering valuable insights for businesses seeking to enhance their churn prediction strategies using machine learning methods.[5]

The study of Lewlisa Saha et al. highlights the crucial role of predicting churn rates in the telecommunications sector for maximizing profits while retaining customers. It explores various learning strategies to build an accurate churn prediction model, including ensemble learning, traditional classification techniques, and deep learning methods. Evaluation using datasets from Southeast Asian and American telecom markets reveals the superior performance of convolutional neural network (CNN) and artificial neural network (ANN) techniques, achieving accuracies up to 99 percent. This study offers insightful information into effective churn prediction methods, emphasizing the potential of deep learning techniques to improve customer retention strategies in the telecommunications industry.[6]

The paper by Asad Khattak et al. addresses the challenge of customer churn in modern organizations, emphasizing the financial losses incurred when customers switch to alternative service providers due to dissatisfaction. Although deep learning and ML techniques have shown potential in churn prediction, there are still problems with producing precise predictions. Previous research highlighted unexpected outcomes with machine learning classifiers and traditional feature encoding methods. To address these issues, the study proposes a hybrid deep learning model, BiLSTM-CNN, aiming to improve churn prediction accuracy. Experimental results demonstrate the model's effectiveness, achieving a remarkable accuracy of 81 percent on benchmark datasets. This research contributes to advancing churn prediction methodologies and enhancing customer retention strategies in businesses.[7]

The article in question, retracted by Hindawi due to evidence of systematic manipulation of the publication process, purported to present a model for forecasting customer attrition in the telecom industry in China. However, discrepancies in scope, research description, data availability, and peer-review manipulation were uncovered during an investigation by the publisher. Such indicators undermine confidence in the reliability of the article's content, leading to its retraction. Hindawi and Wiley express regret that standard quality checks did not detect these issues prior to publication and have implemented additional measures to uphold research integrity. This retraction notice serves to caution readers about the unreliability of the article's content.[8]

2 Introduction

In today's fiercely competitive market, businesses face constant pressure to not just bring in new clients, but also to maximize revenue streams through strategic customer retention efforts. While customer acquisition and up-selling are vital components of business growth, it's widely recognized that retaining existing clients is both more cost-effective and easier to execute. As such, the attrition rate of customers, often referred to as churn, holds significant importance in the corporate landscape.

Given the paramount importance of customers in driving business success, it becomes imperative for companies to prioritize initiatives aimed at cultivating and preserving client loyalty. However, with a large client base to manage, individually tracking each customer becomes impractical, requiring considerable resources regarding time, money, and effort. In order to solve this challenge, businesses often focus their attention on identifying high-risk clients who are likely to churn prematurely.

In this situation, algorithms for ML emerge as powerful tools for predicting customer churn with high accuracy. Leveraging historical data, these algorithms can effectively analyze patterns and trends to identify potential churn indicators. Using the capabilities of machine learning, companies can obtain insightful knowledge about customer behavior and proactively implement retention strategies.

In our project, we embark on a comprehensive approach to churn prediction by collecting and meticulously cleaning data. Additionally, We used modern machine learning methods such as feature engineering and feature selection to enhance the predictive capabilities of our models. By leveraging a diverse set of techniques including Logistic Regression, DT, SVM, Random Forest, and Boosting, we aim to develop robust churn prediction models that empower businesses to anticipate and mitigate customer churn effectively.

3 Objectives

In our endeavor to reduce customer churn, our initiative is driven by a comprehensive understanding of our customers and their behavior. By delving into past data, we aim to:

- Develop a predictive model for estimating telecommunication customer churn rate.
- Evaluate the effectiveness of various ML algorithms for predicting churn.
- Identify key features or factors influencing customer churn to enhance the accuracy and relevance of the predictive model.
- Assess the model's explainability and interpretability to enhance churn rate prediction.

4 Design

In navigating the challenge of acquiring new clients and retaining existing ones, leveraging customer data from previous years becomes imperative to mitigate churn rates effectively. Given the abundance of available data, identifying the critical factors influencing churn rate prediction becomes a crucial first step.

Within the telecommunications sector, various algorithms offer robust capabilities for accurately forecasting customer turnover rates. These include Random Forest, XG Boost (Boosting Algorithm), and Logistic Regression, each bringing unique strengths to the predictive modeling process.

To operationalize our approach, we outline a systematic workflow depicted in the below proposed system's flowchart. This delineates the stages involved in our churn prediction methodology, supplying a systematic framework for implementation.

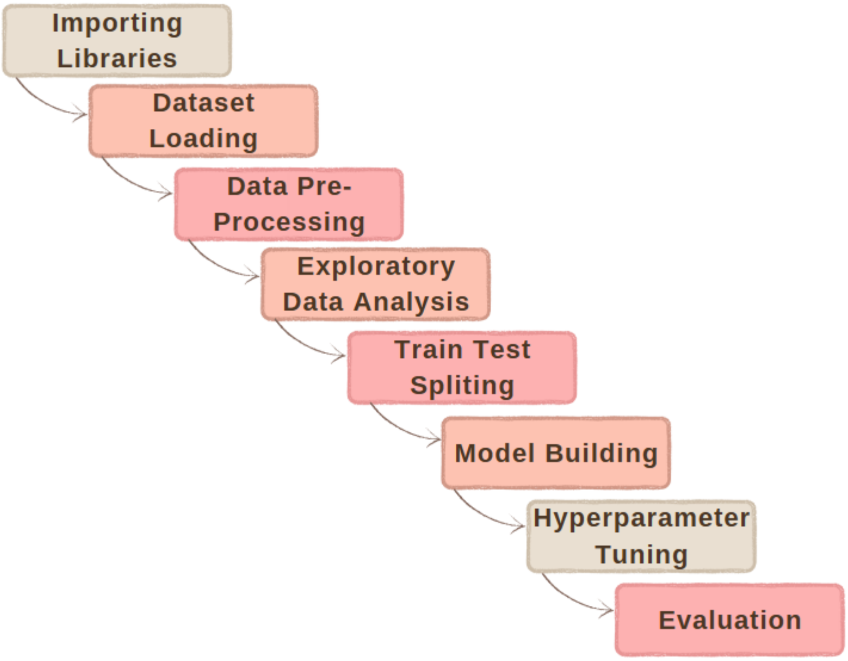


Figure 1: *Flowchart of Machine Learning methodology*

The initial step in the model development process is the utilization of a trained and tested dataset, ensuring the attainment of optimal performance in predicting churn. Data preprocessing constitutes the initial step, involving filtering and standardizing data to facilitate analysis. Following preprocessing, feature selection becomes pivotal in identifying the factors that have the most significant impact to churn prediction. Subsequently, employing techniques such as Random Forest, XG Boost, and Logistic Regression enables the prediction and classification of churn instances. Utilizing the preprocessed dataset, the model undergoes rigorous training and testing to assess its efficacy in monitoring and analyzing client behavior. Hyperparameter tuning further refines model performance, enhancing its predictive accuracy and reliability.

Finally, an in-depth analysis based on the acquired data is conducted to forecast client attrition, providing actionable insights for devising targeted retention strategies. This comprehensive design framework ensures the effectiveness and scalability of our churn prediction approach within the telecommunications domain. Since it can be difficult to get new clients to utilize the company's services, we need the customer data from previous years in order to lower the churn rate. There is an enormous quantity of customer data available, thus what factors are critical to churn rate prediction must be determined.

5 Implementation

During the implementation phase, unprocessed data is converted into a training dataset that is then used for decision-making. In order to determine whether or not a customer would churn, it involves a number of procedures. Anticipating the class enables us to take additional necessary actions. The steps followed in our project are as follows:

5.1 Importing Libraries

In data science and machine learning projects, importing necessary libraries is the first step towards building robust and efficient models. Each library serves a specific purpose and provides functions and tools to handle various tasks involved in data analysis, visualization, model building, and evaluation. Let's explore the role of each library:

1. Pandas:
   - Pandas is a versatile Python package for analysis and data manipulation.
   - It provides data structures like DataFrame and Series, which are essential for working with structured data.
   - With Pandas, you can load data from various file formats, such as CSV, Excel, SQL databases, and more.
   - It offers functions to clean, transform, filter, and aggregate data, making it suitable for data preprocessing tasks.
2. NumPy:
   - A core Python package for numerical computation is called NumPy.
   - Large, multi-dimensional arrays and matrices are supported, and a number of mathematical operations are available for effective manipulation of these arrays.
   - NumPy arrays are the building blocks for many other libraries in the Python data ecosystem, including Pandas and scikit-learn.
   - It enables vectorized operations, which significantly improve the performance of numerical computations compared to traditional loop-based approaches.
3. Matplotlib:
   - Matplotlib is a widely used plotting library in Python for creating static, interactive, and animated visualizations.
   - It has several different charting capabilities that allow you to make histograms, line graphs, scatter plots, bar plots, and more.
   - Matplotlib provides full control over plot customization, allowing users to adjust various aspects such as colors, labels, titles, axes, and annotations.
   - It is highly compatible with NumPy arrays, making it easy to visualize data stored in arrays or Pandas DataFrames.
4. Seaborn:
   - A Matplotlib-based library for statistical data visualisation is Seaborn.
   - It offers a sophisticated drawing tool for creating eye-catching and educational statistical visuals.
   - Seaborn simplifies the process of creating complex visualizations by offering functions for plotting categorical plots, distribution plots, regression plots, and more.
   - It seamlessly integrates with Pandas DataFrames and supports features like color palettes, facet grids, and themes to enhance the visual appeal of plots.
5. Scikit-learn:
   - A flexible machine learning framework based on NumPy, SciPy, and Matplotlib is Scikit-learn.

- It offers a wide range of algorithms for supervised and unsupervised learning, including classification, regression, clustering, dimensionality reduction, and model selection.
- Scikit-learn provides a consistent and user-friendly interface for training, testing, and deploying machine learning models.
- It includes tools for data preprocessing, feature selection, model evaluation, and hyperparameter tuning, making it suitable for end-to-end machine learning workflows.

By importing these libraries into our Python environment, we gain access to a rich set of tools and functionalities that streamline the process of data analysis, visualization, and model building. These libraries form the foundation of data science and machine learning projects, empowering practitioners to explore, analyze, and extract insights from data effectively.

5.2 Dataset Loading

Prior to beginning the implementation, we must collect the data. The web platform ”Kaggle” provides the ”Telecom Churn Prediction” dataset, which is employed in our project.[9] The dataset comprises customer attributes such as:

- Customer ID,
- gender: indicating whether the client is a male or a female,
- SeniorCitizen: indicating if the customer is a senior citizen or not (1, 0),
- Partner: Indicates if the client has a partner (Yes, No),
- Dependents: whether or not the customer has dependents,
- tenure: the length of time the customer has been a customer of the business,
- PhoneService: whether or not the customer has a phone service (Yes, No);
- MultipleLines if the client has more than one line (Yes, No, No phone service),
- InternetService customer’s internet service provider (DSL, Fiber optic, No),
- OnlineSecurity based on the customer’s status (yes, no, or no internet service)

There are a total of 21 features (columns) and 7043 entries of customers (rows). The ”Churn” column is the target feature. The dataset, typically stored in a CSV (Comma Separated Values) file, is loaded using the Pandas library which allows us to read the CSV file into a DataFrame, a tabular data structure that resembles a spreadsheet. This DataFrame serves as the primary data container for our analysis.

5.3 Data Pre-processing

Dataset loading and preprocessing are crucial steps to guarantee that the data is in a suitable format for analysis and model training. Let’s delve into each step:

1. Handling Missing Values:
   - Missing values, also known as NaN (Not a Number) or null values, can occur in datasets due to various reasons such as data entry errors or incomplete information.

- In our project, we address missing values by using the dropna() function provided by Pandas. This function removes rows containing any missing values, ensuring that there are no incomplete data or unreliable records in our dataset.
- By dropping missing values, we maintain data integrity and avoid potential biases or inaccuracies in our analysis and predictions.

2. Adjusting Data Types:
   - The data type of each column in the dataset determines how the data is stored and processed by the computer.
   - In telecom churn prediction, certain columns may contain numeric values represented as strings (e.g., 'TotalCharges') due to formatting issues or mixed data types.
   - To perform the numerical computations and analysis on such columns, we apply the pd.to_numeric() function to transform them into numerical data types. This ensures that numeric columns are interpreted correctly and can be used for mathematical operations and model training.

3. Handling Categorical Variables:
   - Categorical variables represent qualitative attributes such as gender, internet service type, or customer segment.
   - Usually, ML algorithms demand numerical input, so we need to provide a numeric format for categorical variables.
   - One common approach is to use one-hot encoding, also known as dummy encoding, to create binary columns in the categorical variable for every category. This is achieved using the pd.get_dummies() function in Pandas.
   - Transforming categorical variables into a format that ML algorithms can comprehend and process efficiently by turning them into dummy variables.

4. Normalization:
   - To normalize or standardize the range of independent variables or features in a dataset, this is an essential machine learning preprocessing step. A technique for normalization is min-max scaling.
   - Min-max scaling involves transforming the values of features to fit within a specified range, typically between 0 and 1. This normalization process is achieved by rescaling each feature's values based on its minimum and maximum values in the dataset. The formula for min-max scaling is as follows:

     $$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

     where:
     - X is the original value of the feature.
     - X_{min} is the lowest value of the dataset's feature.
     - X_{max} is the highest value of the dataset's feature.
     - X_{scaled} is the scaled value of the feature within the specified range.
   - By applying min-max scaling, the variety of values for each feature is compressed to a consistent scale, which can restrict traits with bigger magnitudes from taking the forefront of the learning process. This confirms that every feature contributes equally to the analysis and avoids issues related to varying scales among different features.

- Min-max scaling is especially helpful for algorithms that rely on distance calculations or gradient descent optimization, such as K-nearest neighbors (KNN) and support vector machines (SVM). Normalizing the features makes these algorithms more rapidly convergent and improves their performance.
- Min-max scaling ensures that the features are uniformly distributed within the specified range, improving the dataset's training suitability of machine learning models. This preprocessing step enhances the model's ability to learn about relationships and trends of the data, ultimately resulting in forecasts that are more accurate of customer churn.

By performing dataset preprocessing in our telecom churn prediction project, we ensure that our data is clean, structured, and suitable for analysis and modeling. These preprocessing steps lay the foundation for building accurate predictive models that can identify factors contributing to churn and help telecom companies take proactive measures to retain customers.

5.4 Exploratory Data Analysis

An essential in every data analysis project is exploratory data analysis (EDA), which provides knowledge about the patterns and correlations discovered in the dataset. In our project, EDA involves the following key components:

- Correlation Analysis:
  - Correlation analysis is conducted to understand the connection between the desired variable (churn) and other predictor variables (features) in the dataset.
  - This analysis helps to identify which features are positively or negatively correlated with churn, indicating their potential influence on customer attrition.
  - Understanding the correlations can guide feature selection and model building, as features with high correlations could have a stronger predictive power for churn.

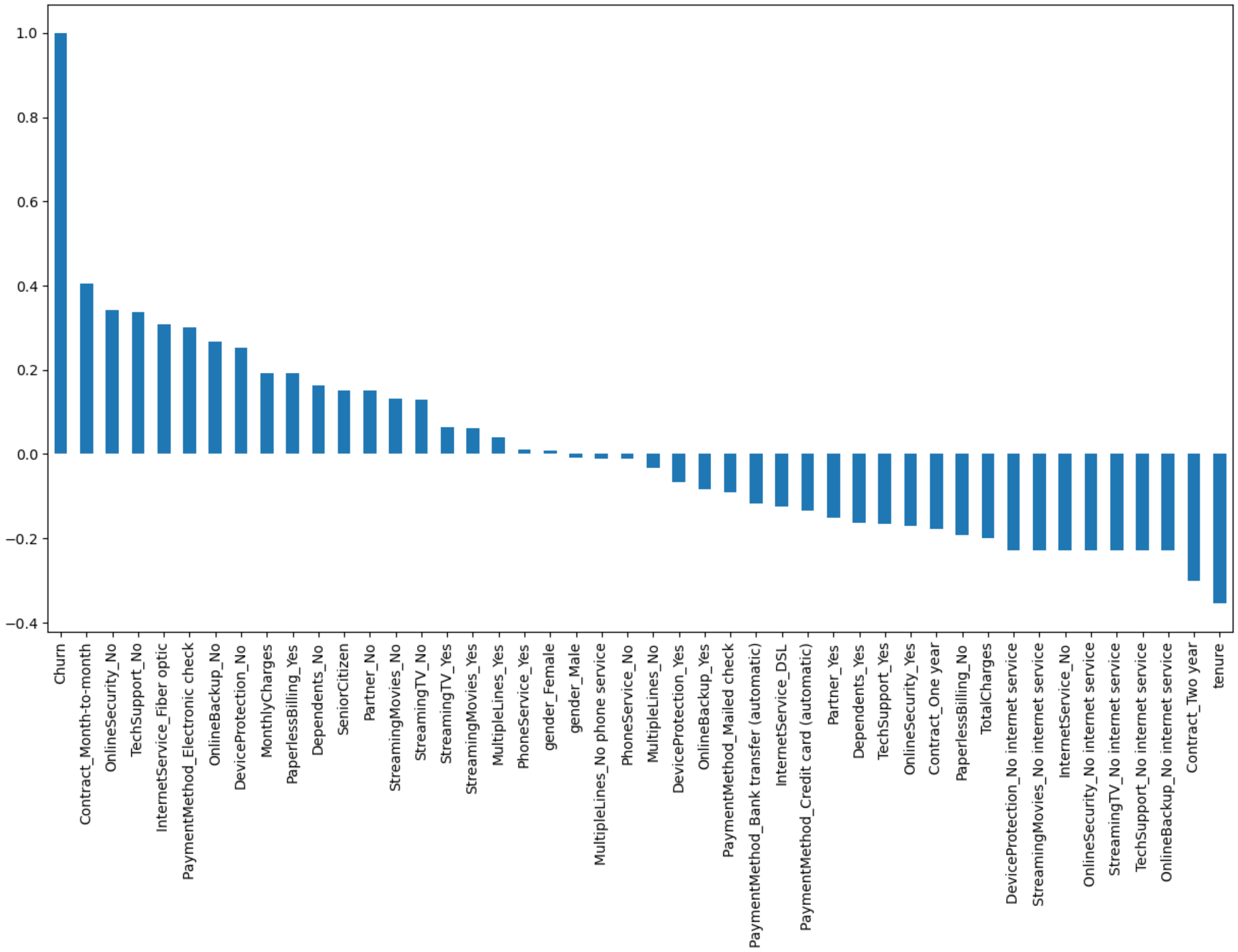
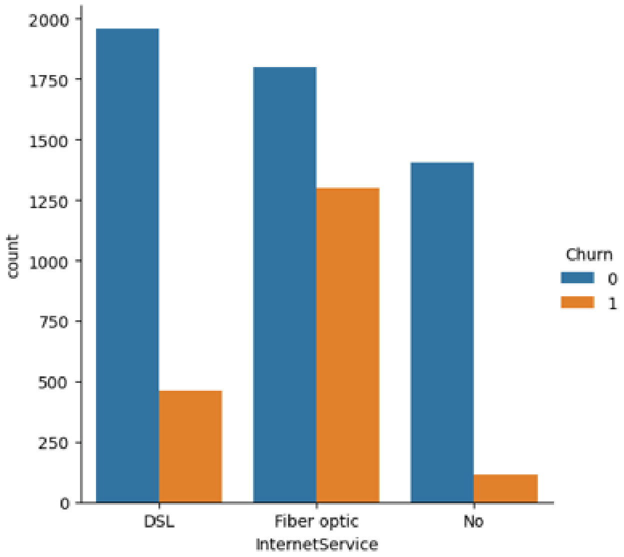
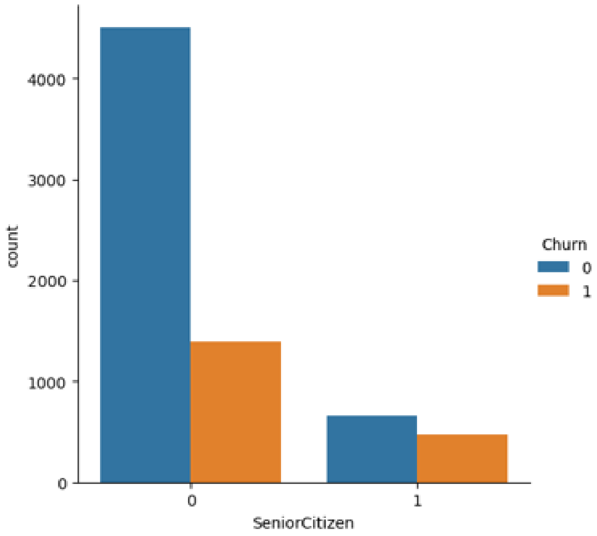
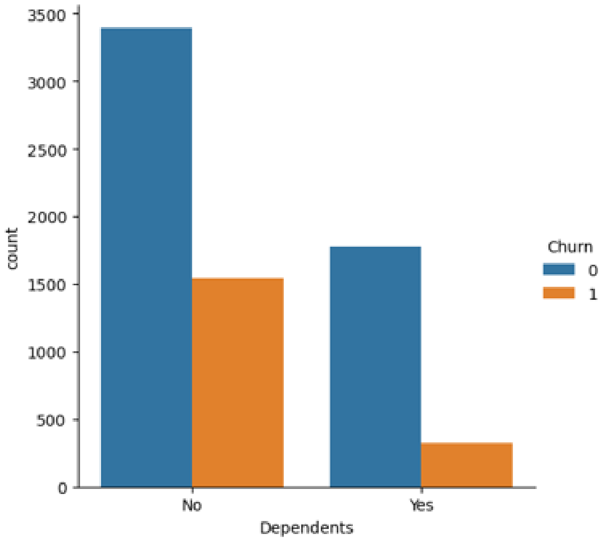


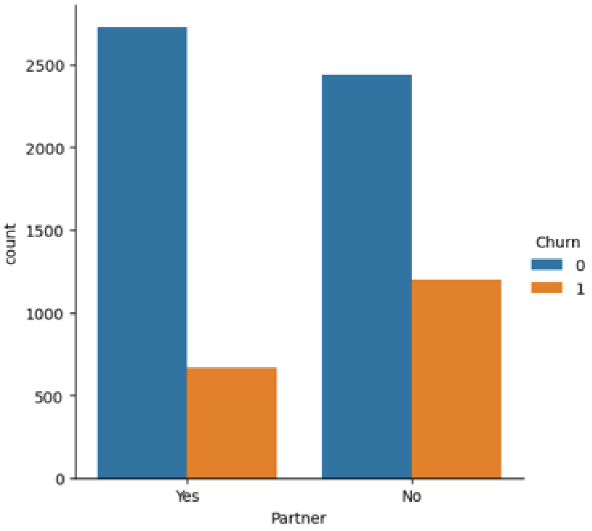
Figure 2: *Correlation of churn with all features*

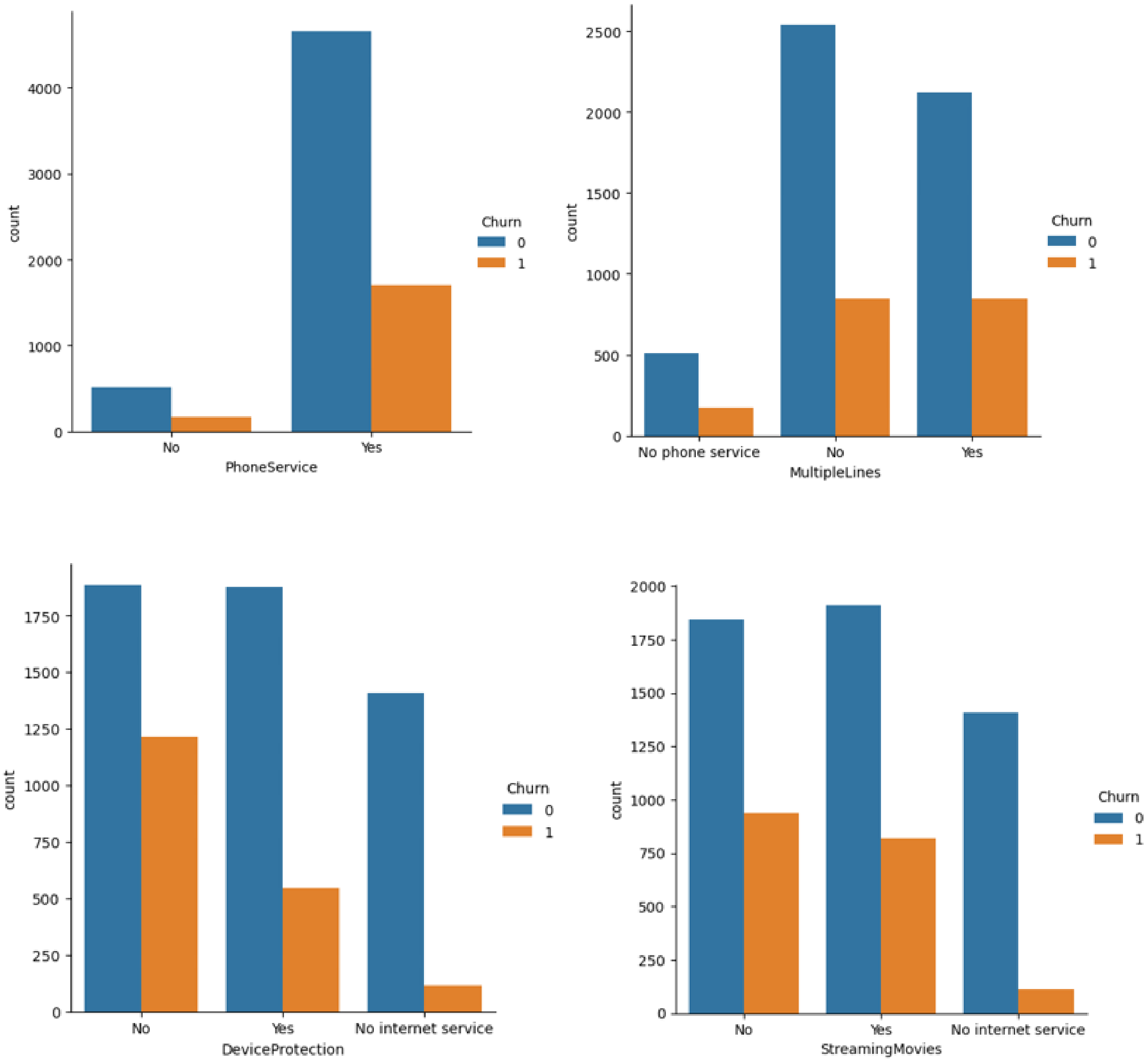
- Visualization of Categorical Variables:
  - Count plots and other visualizations are used to explore the distribution of categorical variables concerning churn.
  - These graphics offer perceptions into how different categorical variables (e.g., gender, senior citizen status, internet service type) are distributed among churned and non-churned customers.
  - By visually inspecting these distributions, we can determine any possible patterns or trends that may influence churn behavior.



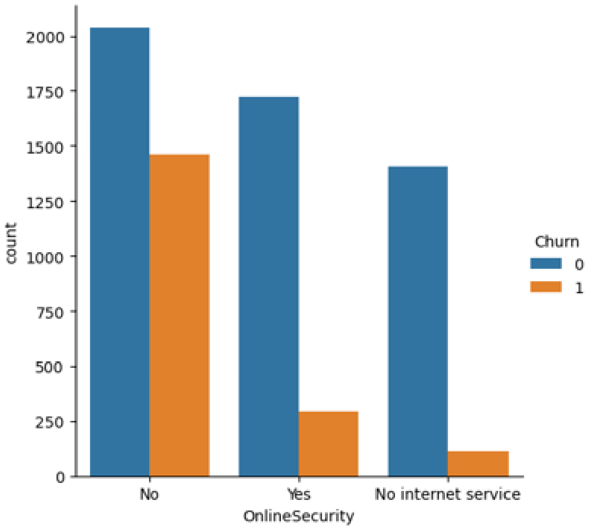


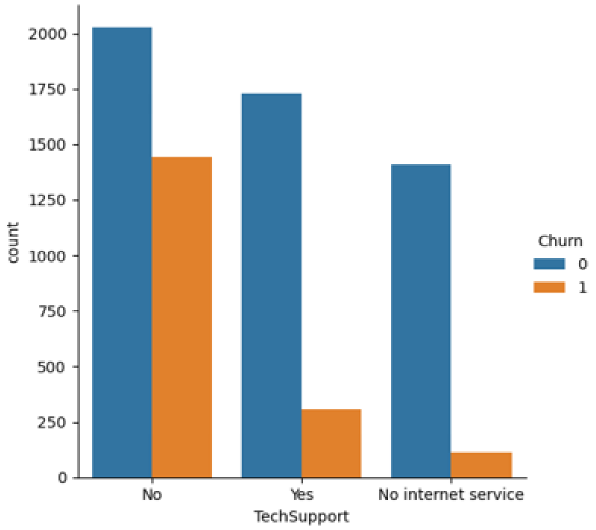


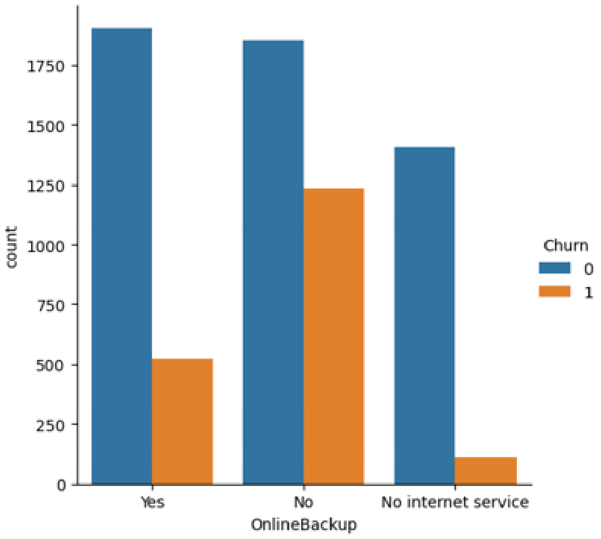


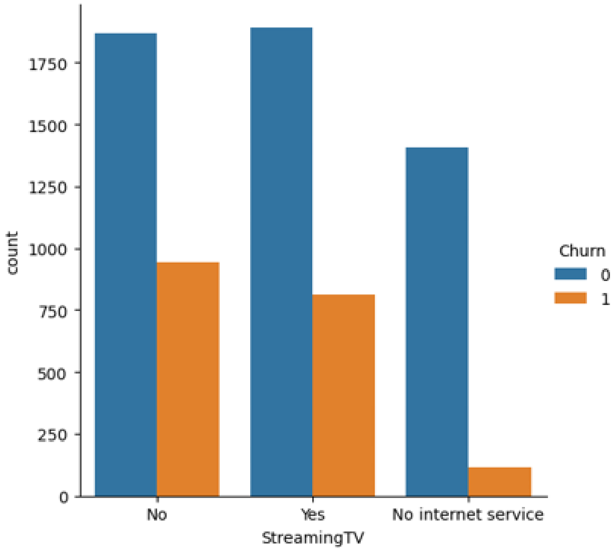


count
4000
3000
2000
1000
0
No
Yes
PhoneService
Churn
0
1
count
2500
2000
1500
1000
500
0
No phone service
No
Yes
MultipleLines
Churn
0
1
count
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
DeviceProtection
Churn
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
StreamingMovies
Churn
0
1


count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
OnlineSecurity
Churn
0
1


count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
TechSupport
Churn
0
1


count
1750
1500
1250
1000
750
500
250
0
Yes
No
No internet service
OnlineBackup
Churn
0
1


count
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
StreamingTV
Churn
0
1

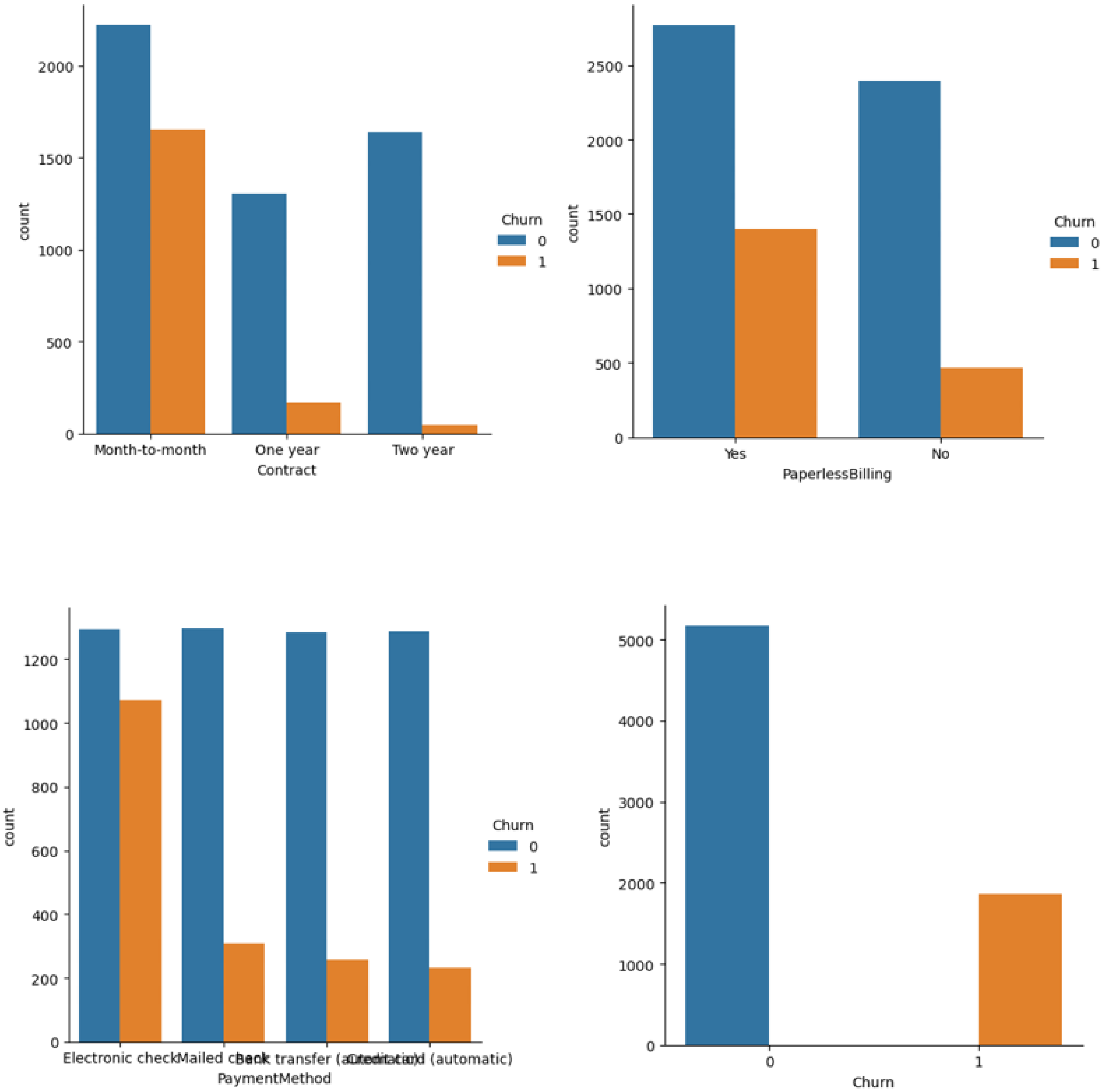


Figure 3: *Graph between count of churn values v/s features*

In summary, EDA involves exploring the dataset through correlation analysis and visualizations to uncover insights into the relationship between churn and other variables. These insights serve as a foundation for feature selection, model building, and ultimately, the development of efficient methods for keeping clients and reducing churn rates.

5.5 Train and Test Splitting

The training set and the testing set comprise the dataset. These two separate subsets correspond to machine learning's training and testing phases. This separation is essential for accurately assessing the model's performance and preventing over-fitting, a situation in which the model learns the training data excessively well but is unable to generalize to new data.

The process typically involves the following steps:

1. Dataset Splitting: The training and testing sets are a pair of mutually exclusive subsets of the labeled instances that make up the dataset. The testing set is used to evaluate the model's performance after it has been trained utilizing the training set.

2. train_test_split Function: This function, often provided by machine learning libraries like sci-kit-learn, facilitates the random splitting of the dataset into sets for testing and training. It ensures that the distribution of the data points is uniform between the two subsets, preserving the original distribution of classes or labels if the dataset is imbalanced.

3. Training Set: The training set constitutes the majority of the dataset, typically around 70-80 percent. To generate predictions on new, unknown data, the model discovers relations and trends in the training set.

4. Testing Set: The testing set comprises the remaining portion of the dataset, usually around 20-30 percent. It remains untouched during the training process and is used solely for evaluating the trained model's performance. The accuracy and generalization abilities of the model can be evaluated by comparing its predictions to the true labels in the testing set.

Overall, the train and test splitting phase is fundamental in the machine learning workflow, enabling the assessment of model efficacy.

5.6 Model Building

To create predictive models for telecom churn prediction, we use a variety of ML methods during the model-building stage. Boosting algorithms, LR, DT, RF, and SVM are some of these algorithms. These algorithms all have distinctive qualities that might help them work well for the given purpose.

1. Logistic Regression: An approach that is commonly applied to binary classification problems, such as churn prediction, is logistic regression. It simulates a binary outcome's likelihood depending on one or more predictor variables.

2. Random Forest: An ensemble learning technique called Random Forest creates a variety of decision trees during training and outputs the class that is the class's (classification) mode of the individual trees.

3. Support Vector Machine (SVM): For classification tasks, SVM is an effective algorithm. It operates by locating the feature space hyperplane that best divides the classes.

4. Decision Tree: Decision trees are simple yet efficient algorithms that partition the area of features into regions considering the predominant class in each zone.

5. AdaBoost: AdaBoost is a boosting method that builds a powerful classifier by blending several weak classifiers. It sequentially trains a series of weak learners, each focusing on the instances that were misclassified by the previous ones.

6. XGBoost: XGBoost is an advanced implementation of gradient boosting that has gained popularity because of its efficiency and accuracy. It builds a series of decision trees and optimizes their performance using gradient descent.

After training, the above mentioned model's performance is evaluated based on the accuracy score.

5.7 Hyperparameter tuning

During the hyperparameter tuning phase, we employed grid search to systematically search for the ideal setting combination for each ML model. This process involves specifying a grid of hyperparameter values to investigate and assess the model's effectiveness for each combination using cross-validation.

- Grid Search: GridSearchCV, a module from the scikit-learn library, is utilized to perform grid search. It thoroughly examines each potential combination given as input as a dictionary of hyperparameters and the ranges that correspond to them. Cross-validation is utilized to train and assess the model for each combination. Typically, this is done using k-fold cross-validation, in which the dataset is divided into k-subsets or folds, and the model is trained k times, using a different fold as the validation set each time.

- Cross-Validation: Cross-validation helps to lessen the possibility of overfitting by providing a more robust estimate of the model's performance. It involves partitioning the dataset between training and validation sets multiple times, and the average performance across all folds is used to evaluate the model. Grid search, combined with cross-validation, allows for a comprehensive assessment of each hyperparameter combination's performance.

- Best Hyperparameters: After the grid search is complete, the hyperparameter combination that yields the highest performance metric (such as accuracy, recall, precision, or F1-score) is the ideal set of hyperparameters chosen based on the validation set for the model.

By performing grid search, the hyperparameters of the model are fine-tuned to improve the model's performance and ultimately build an accurate and reliable predictive model for the task at hand.

5.8 Model Evaluation

Different evaluation metrics are available to determine which model performs better than the others.

- Accuracy: Accuracy represents the percentage of correctly classified instances among all the instances. In churn prediction, accuracy shows the proportion of correctly identified churned and non-churned customers out of the total customer base. A high accuracy implies that the model is effective in distinguishing between churners and non-churners.

- Precision: Precision calculates the proportion of true positives (correctly predicted churners) among every case anticipated as positive (predicted churners). In churn prediction, precision indicates the model's accuracy in identifying churners, minimizing false positives. High precision indicates that when the model predicts a customer will churn, it's highly likely to be correct.

- Recall: The percentage of true positives (accurately predicted churners) among all actual positive cases (actual churners) is measured by recall, which is occasionally called sensitivity or true positive rate. In churn prediction, recall indicates the model's capacity to seize all churners, minimizing false negatives. High recall means that the model is able to recognise most of the churners in the dataset.
- F1 Score: The harmonic mean(H.M.) of the two: recall and precision, is the F1 score. It provides a single metric that balances both precision and recall. In churn prediction, the F1 score represents the overall accuracy of the model in determining churners while controlling both false positives (FP) and false negatives (FN). A high F1 score indicates that the model has a good balance with precision and recall, ensuring it is reliable for predicting churn.

These measures are essential for assessing how well churn prediction models work. Accuracy gives a broad picture of the model's performance; still precision, recall, and F1 score shed light on the model's capacity to distinguish between churners and non-churners, which is a crucial distinction for companies looking to successfully retain clients.

5.9 SHAP Implementation

SHAP (SHapley Additive exPlanations) is a method used to interpret machine learning models by explaining the contribution of each feature to the model's predictions.It provides insights into the model's decision-making process and helps in understanding the importance of different features.

- Significance of SHAP: In the realm of churn rate prediction, SHAP plays a crucial role in enhancing model interpretability.By elucidating why a model makes certain predictions, SHAP values empower businesses to make informed decisions regarding customer retention strategies.
- Integration of SHAP into Churn Rate Prediction: Churn rate prediction models are often complex, making it difficult to understand how they arrive at their predictions.Interpretability is vital as it enables stakeholders to comprehend the factors influencing churn and devise effective retention strategies.

  SHAP can be seamlessly integrated into churn rate prediction models to provide transparent insights into model predictions. By leveraging SHAP values, businesses can gain a deeper understanding of the factors contributing to churn and tailor retention efforts accordingly.
- Implementation of SHAP in the Project: SHAP values are computed for each feature to determine their impact on model predictions.These values are derived using an algorithm based on cooperative game theory, providing a clear understanding of feature importance. SHAP values can be visualized through summary plots, offering an overview of each feature's influence on predictions.Individual SHAP plots provide detailed insights into specific predictions, aiding in the identification of key drivers behind churn.

5.10 Streamlit App

In contemporary business environments, understanding customer churn, the phenomenon where customers cease engagement with a company, is crucial for sustained success. Predicting and addressing churn effectively can significantly impact revenue and customer satisfaction. In this project, we introduce a Churn Rate Prediction Application built using Streamlit, a Python library tailored for rapid development of web applications.

- Overview of Streamlit: Streamlit is an open-source Python library designed for the swift creation of web applications for data science and ML projects. The framework allows developers to craft interactive web applications directly from Python scripts, minimizing the need for traditional front-end development.
- Streamlit's Key Attributes:
  - Ease of Use: Streamlit offers an intuitive interface, enabling developers to focus on application logic without delving into complex web development.
  - Rapid Prototyping: Developers can swiftly prototype and deploy applications for ML models and data visualizations.
  - Customization: Streamlit provides customizable components such as sliders and buttons, facilitating the creation of interactive user interfaces.
  - Integration with ML Libraries: The framework seamlessly integrates with popular machine learning libraries like scikit-learn and TensorFlow, simplifying the incorporation of ML models into applications.
- Implementation of Churn Rate Prediction Application:

  Our Churn Rate Prediction Application, powered by Streamlit, aims to offer businesses an intuitive platform for forecasting customer churn. Users can input customer data via the application, which then employs a machine learning model to predict churn likelihood.
- Application Features:
  - Online Prediction Mode: Users can input individual customer details, including demographics and service usage, to receive real-time churn predictions.
  - Batch Prediction Mode: The application allows users to upload CSV files containing customer data, enabling simultaneous predictions for multiple customers.
  - Interactive Visualization: Predictions are presented using interactive visualizations, providing insights into churn determinants.
  - Integration of SHAP for Interpretability: To enhance model interpretability, we integrated the SHAP (SHapley Additive exPlanations) library into our application. SHAP values illuminate the contribution of each feature to prediction outcomes, offering valuable insights into churn drivers.
- Relevance to Churn Rate Prediction Project:
  - Application Significance: The Churn Rate Prediction Application addresses the critical business need to forecast and mitigate customer churn effectively. Leveraging Streamlit's capabilities, we developed a user-friendly tool empowering businesses to adopt data-driven decision-making and implement targeted retention strategies.
  - Advantages Over Traditional Approaches: In contrast to traditional churn analysis methods that entail intricate data processing and manual model deployment, our application streamlines the process. Featuring an intuitive interface and real-time prediction capabilities, the application offers an efficient and accessible solution for churn rate prediction.

6 Result and Discussion

The optimal model for churn prediction in our telecommunication project is ”XGBoost,” which was determined after several ML algorithms were evaluated using a range of performance criteria, such as precision, accuracy, recall, and F1 score.

Here's a summary of the evaluation results for each algorithm:

Table 1: Evaluation of ML Algorithms

Algorithm	Accuracy	Tuned Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.82018	0.81378	0.66	0.54	0.59
Random Forest	0.78962	0.80881	0.67	0.48	0.56
SVM	0.81876	0.82089	0.71	0.52	0.60
DT	0.74129	0.80454	0.66	0.52	0.58
ADABOOST	0.81592	0.82231	0.70	0.55	0.61
XGBOOST	0.80597	0.82231	0.70	0.55	0.62

These results show that XGBoost performs competitively across all measures, with a strong F1 score of 0.62, which implies a solid equalization of recall and precision.

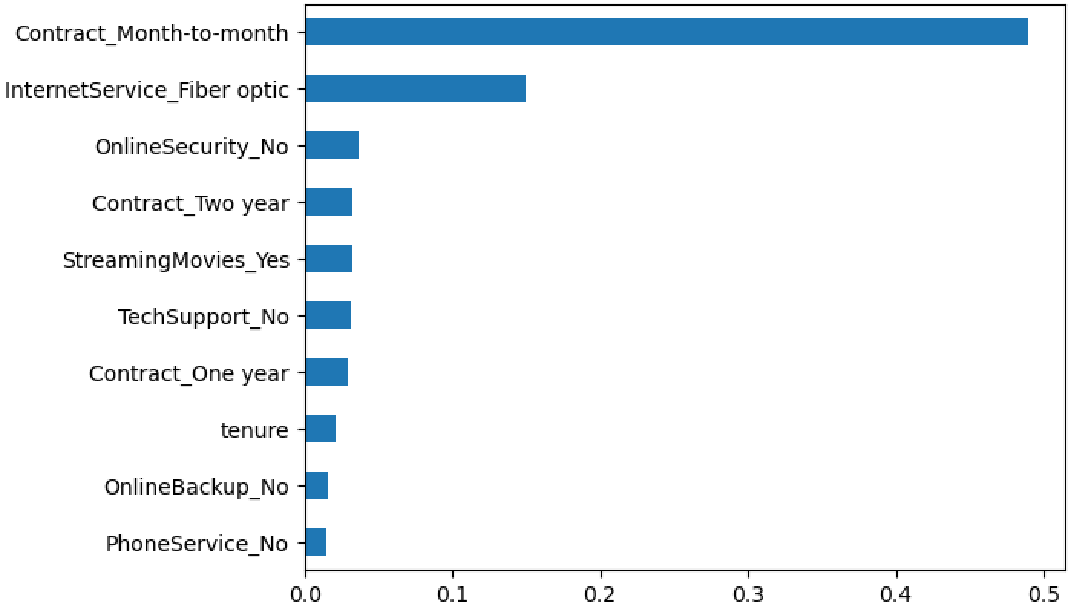


Figure 4: *Graph shows the importance of features*

Furthermore, Figure 4 illustrates the significance of the data, giving insights into the features that are most important in forecasting churn. This information can guide further analysis and decision-making in refining the churn prediction model.

Overall, the evaluation results and feature importance analysis support the selection of XGBoost as the optimal model for churn prediction in our telecom project.

Feature Importance tells us how important each feature is to the model prediction in general but doesn't tell us how features tend to increase or decrease the prediction or if we had a classification problem it would not tell us how the features change the probability of a positive prediction. So, we used SHAP plots which are as follows:

- Waterfall Plot: The SHAP waterfall plot offers a transparent and intuitive way to interpret model predictions in churn rate prediction scenarios. It empowers stakeholders to grasp the underlying factors driving churn and provides actionable insights for targeted retention efforts.

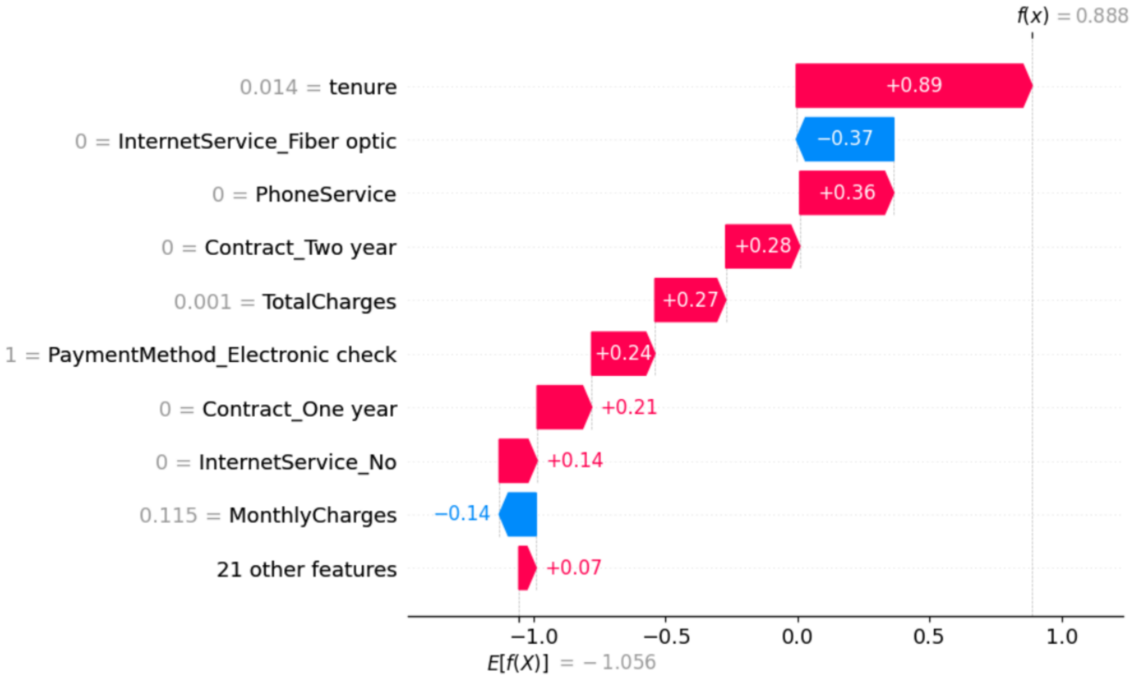


Figure 5: *SHAP Waterfall Plot*

- Force Plot: The SHAP force plot provides an insightful view of individual predictions, helping stakeholders understand the rationale behind the model's decisions. It facilitates informed decision-making by highlighting the drivers of churn for specific customers, thereby guiding targeted intervention strategies.

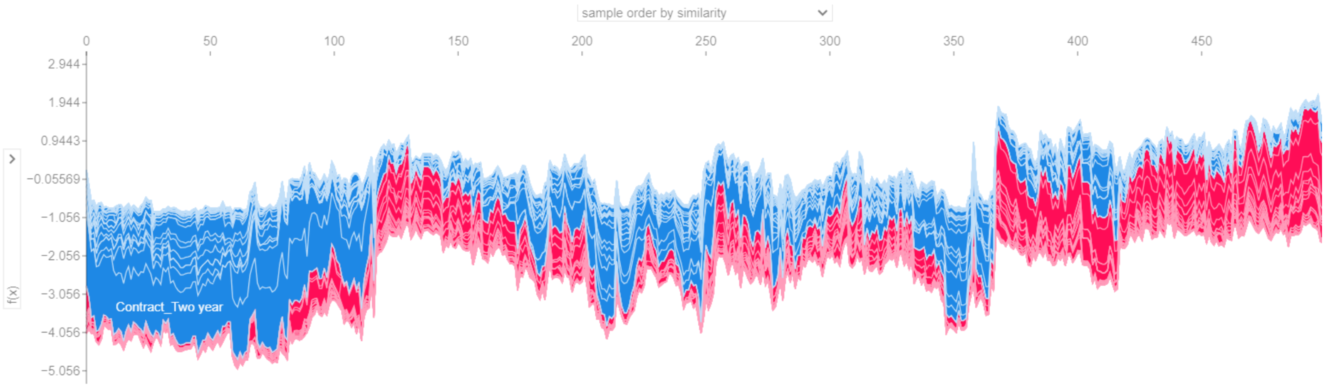


Figure 6: *Force Plot*

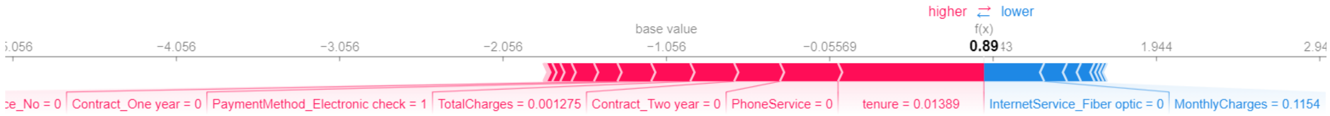


Figure 7: *Force Plot*

- Dependence Plot: The SHAP dependence plot facilitates the exploration of feature effects on model predictions, aiding in the understanding of complex relationships in churn rate prediction. It empowers stakeholders to uncover insights into customer behavior and inform strategic decisions aimed at reducing churn rates.

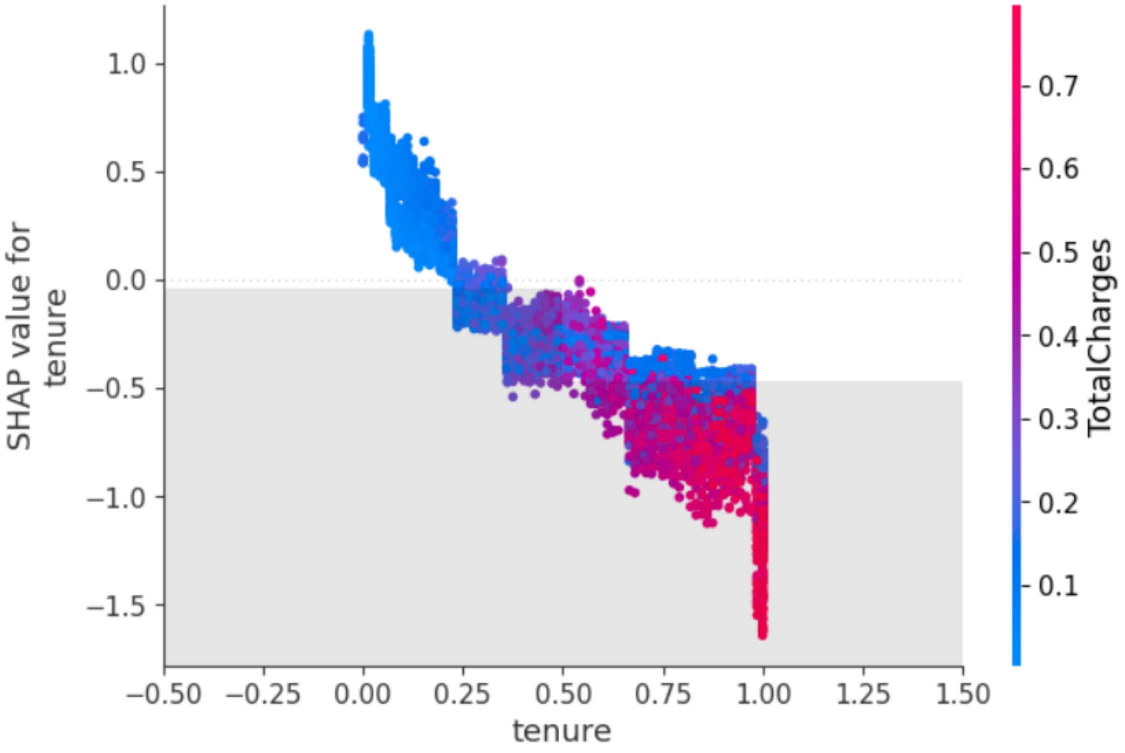


Figure 8: *Dependence Plot*

- BeeSwarm Plot: The SHAP beeswarm plot offers a comprehensive view of feature importance, aiding in the identification of key drivers of churn in the dataset. It facilitates informed decision-making by highlighting features with consistent impact on churn predictions, thereby guiding strategic efforts to improve customer retention.

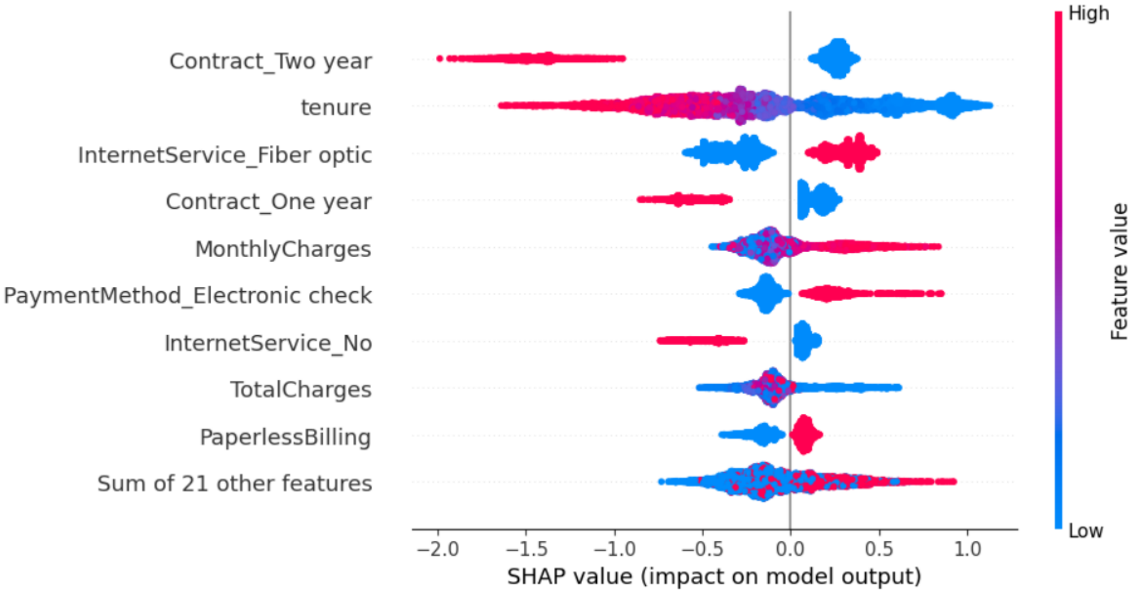


Figure 9: *BeeSwarm Plot*

References

[1] S. K. Wagh, A. A. Andhale, K. S. Wagh, J. R. Pansare, S. P. Ambadekar, and S. H. Gawande, "Customer churn prediction in telecom sector using machine learning techniques," *Results in Control and Optimization*, vol. 14, 3 2024.

[2] A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and naïve bayes," *Applied Soft Computing*, vol. 137, 4 2023.

[3] J. Maan and H. Maan, "Customer churn prediction model using explainable machine learning," *International Journal of Computer Science Trends and Technology*, vol. 11, 2023.

[4] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *Journal of Big Data*, vol. 6, 12 2019.

[5] B. Prabadevi, R. Shalini, and B. R. Kavitha, "Customer churning analysis using machine learning algorithms," *International Journal of Intelligent Networks*, vol. 4, pp. 145–154, 1 2023.

[6] L. Saha, H. K. Tripathy, T. Gaber, H. El-Gohary, and E. S. M. El-kenawy, "Deep churn prediction method for telecommunication industry," *Sustainability (Switzerland)*, vol. 15, 3 2023.

[7] A. Khattak, Z. Mehak, H. Ahmad, M. U. Asghar, M. Z. Asghar, and A. Khan, "Customer churn prediction using composite deep learning technique," *Scientific Reports*, vol. 13, 12 2023.

[8] "Retraction: A prediction model of customer churn considering customer value: An empirical research of telecom industry in china (discrete dynamics in nature and society (2021) 2021 (7160527) doi: 10.1155/2023/7160527)," 2023.

[9] Atindrabandi, "Wa_fn-usec_-telco-customer-churn," 2019.