Machine Learning

Project Increment 1

TEAM MEMBERS:

Seema Shrestha

Dipesh Shrestha

Prathik Raj Kolapudi

Project Title: Churn Prediction for Subscription Services

Team Members: Seema Shrestha, Dipesh Shrestha, Prathik Raj Kolapudi

GitHub Link: https://github.com/Shrestha-Dipesh/ML-Final-Project

# Goals and Objective:

## Motivation

The project's primary objective is to experiment with various binary classification algorithms and develop an effective machine learning model that can predict subscriber churn in the telecom services industry precisely. This prediction will be based on a diverse set of features, including but not limited to gender, tenure, multiple lines, internet service, and more. By effectively identifying subscribers who are at risk of discontinuing their services, companies can proactively implement retention strategies. This proactive approach not only reduces churn rates but also enhances overall service quality and aids in the optimization of business strategies.

## Significance:

In subscription-based industries, to maintain the business sustainability and growth, predicting customer churn is crucial. Determining and retaining customers who are likely to cancel their subscriptions is therefore a critical need, which is addressed by this project by using various machine learning techniques on comprehensive customer data. By emphasizing a priority on client retention over costly acquisition efforts, this project seeks to discover the most relevant statistics related to churn behavior and give businesses predictive insights to reduce churn rates, improve service quality, and increase long-term revenue *(Frankenfield, 2022)*.

## Objectives:

The major objectives of this project are:

1. Analyze correlations among numerous features and importance of each feature towards the prediction to identify the primary factors influencing subscriber churn. To aid with the development of focused relationship strategies, this project seeks to identify critical patterns and behaviors linked to customer churn by examining its relationship with individual as well as collective features.
2. Assess various algorithms’ performance metrices to evaluate and identify the most effective machine learning model in accurately classifying the churn and non-churn customers.
3. Construct a predictive framework and develop a robust machine learning model that can precisely predict subscriber churn by utilizing diverse customer data.

## Features:

**Data Collection and Preprocessing:** Collect historical subscriber dataset from the reliable source. Handle missing values and encode categorical variables.

**Exploratory Data Analysis (EDA):** Visualize and uncover insights, patterns, correlations, and outlier detection within the data to understand subscriber behavior and identify potential indicators of churn.

**Feature Selection Techniques:** Utilize various methods and compare their results to identify the most important features for predicting churn.

**Model Experimentation and Evaluation:** Explore several machine learning models, assess how well they perform with different metrics, and decide which model is best for precise churn prediction.

**Model Training and Hyperparameter Tuning:** Use the dataset to train a selected model, tuning their parameters to improve generalization and predictive ability.

**Model Evaluation and Interpretability:** Evaluate the performance of the model, illustrate the results through ROC curves and confusion metrices, and offer insights into the significance of each feature for easier interpretation and decision-making.

(Banoula, 2023)

## Related Work (Background)

**Kaggle Customer Churn Analysis with Plotly Visualization:**

Using Plotly visualization, Tawfik Elmetwally's analysis on Kaggle offers insightful information about customer churn.  The analysis explores data exploration strategies that make use of Plotly's interactive visualizations. It investigates relationships between different customer attributes and churn, providing a visual representation of the patterns or trends affecting customer attrition. The utilization of Plotly's dynamic charts and graphs have facilitated the comprehension of customer behavior and possible churn indicators by showcasing the various aspects that influence churn rates in the context of telecom services(Elmetwally, 2023).

Link: <https://www.kaggle.com/code/tawfikelmetwally/customer-churn-analysis-plotly-visualization>

**GitHub Customer Churn Prediction Notebook by Himanshu:**

Using a Jupyter Notebook, Himanshu's GitHub project focuses on customer churn prediction. This effort entails putting machine learning models into practice and assessing how well they anticipate churn. The notebook contains feature engineering, applying classification algorithms, and data preprocessing steps. Himanshu have experimented with several models, including random forest, logistic regression, and others, to identify best model for forecasting telecom service churn. To increase prediction precision, the notebook approach also involves model evaluation and hyperparameter tuning(himanshu-03, 2023).

Link: <https://github.com/himanshu-03/Customer-Churn-Prediction/blob/main/src/Customer_Churn_Prediction.ipynb>

## Dataset

**Input**: There are 21 columns (features) with 3 numerical variables, 16 categorical variables and 1 unique identifier and 7043 rows (customers records) in the excel sheet.

**Dataset link:** <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

**Churn:** This column indicates whether a customer has left the telecommunications company within the last month. It is a binary variable, with values like ‘Yes’ or ‘No’, where ‘Yes’ typically represents customers who have churned.

**Customer Account Information:** Details regarding the customer's account are provided in this section, such as the length of time they have been a customer, the type of their contract, their monthly as well as total bill amount, their preferred payment option, whether they have opted for paperless billing etc.

**Demographic Information:** Customers' demographic data, including their sex, age (converted to binary feature ‘Senior Citizen’) and if they have any partners or dependents, is included in this dataset.

**Services:** The services that each customer has signed up for are included in this dataset. Phone, internet, multiple lines, tech support, device protection, and online backup are some of the services provided.

**Output:** Binary Classification – Predict whether a subscriber will churn (Yes) or stay (No) with the service.

*(Team, 2019)*

## Detail design of Methods

The project employs an extensive approach that involves exploratory data analysis (EDA) to identify patterns and correlations associated with customer churn. Important attributes were identified using feature selection strategies such as mutual information and feature importance ranking. For churn prediction, several machine learning methods were evaluated, including K-Nearest Neighbors (KNN), Random Forest Classifier, and Logistic Regression.

## Analysis

Tenure, total charges, and monthly charges all showed significant correlations, according to EDA. Customers with month-to-month contracts, higher monthly charge, new customers, and those without online security, or tech support all had greater churn rates, according to visualizations. A cleaner dataset for modeling was produced since no outliers were found.

## Implementation

Using Python and tools like Pandas, Scikit-learn, and Plotly, the implementation phase involves implementing the design phase approaches. Snippets of code have been structured to manage model training, evaluation, and preprocessing of data. For improved maintainability and scalability, the coding strategy followed the modular design principles.

## Preliminary Results

Preliminary assessments of the models showed promising performance metrics for KNN (with an accuracy of 0.79 and F1 Score of 0.56) and Naïve Bayes (with an accuracy of 0.73 and F1 Score of 0.53). A feature importance analysis revealed that the following factors have a substantial impact on churn prediction: tenure, contract type, monthly charges, online security, and tech support.

## Project Management

### Implementation status report

* **Work completed:**
  + Description:

Task 1: Data Collection and Preprocessing. Handling Missing Values, Exploratory Data Analysis (EDA), KNN implementation

Task 2: Outlier Detection, Label Encoding, Feature Selection, Naïve Bayes implementation

* + Responsibility:

Task 1: Seema Shrestha

Task 2: Dipesh Shrestha

* + Contributions:

Seema and Dipesh contributed to task 1 and task 2 respectively, with approximately equal participation.

No contributions from Prathik Raj Kolapudi.

* **Work to be completed:**
  + Description:

Implementing further models, Hyperparameter Tuning, Data Training and Evaluation

* + Responsibility:

Implementing Random Forest and AdaBoost, Hyperparameter Tuning (Grid Search): Seema Shrestha

Implementing Decision Trees and XGBoost, Hyperparameter Tuning (Random Search): Dipesh Shrestha

Implementing Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting, Hyperparameter Tuning (Bayesian Optimization), Model Training and Evaluation: Prathik Raj Kolapudi

* **Issues/Concerns**

We might encounter challenges in identifying a highly accurate predictive model with consistent performance in churn prediction. Because of the risk of overfitting and computational demands, we are not sure if we can implement the complex models. Moreover, a lack of participation from a team member is hindering the progress.

# Screenshots

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

# Bibliography

1. Banoula, M. (2023, August 21). *Machine Learning Steps: A Complete Guide*. Retrieved from Simplilearn: https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps
2. Elmetwally, T. (2023, November 14). *customer churn analysis | plotly visualization*. Retrieved from Kaggle: https://www.kaggle.com/code/tawfikelmetwally/customer-churn-analysis-plotly-visualization
3. Frankenfield, J. (2022, May 18). *Churn Rate: What It Means, Examples, and Calculations*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/churnrate.asp
4. himanshu-03. (2023, March 20). *Customer Churn Prediction*. Retrieved from GitHub: https://github.com/himanshu-03/Customer-Churn-Prediction/blob/main/src/Customer\_Churn\_Prediction.ipynb
5. Team, S. (2019, July 11). *Telco Customer Churn (11.1.3+)*. Retrieved from IBM Community: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113