

**School of Engineering & Technology**

**Asian Institute of Technology**

**AT82.01 - Computer Programming for Data Science and Artificial Intelligence**

**Date: 11-01-2023**

**Churn Classification**

**Submitted To Submitted By:**

**Dr. Chantri Polprasert Anawat Pradabsuk st124039**

**Wut Yee Aung st124377**

**Index**

Contents

[Introduction 1](#_Toc152013189)

[Problem Statement 2](#_Toc152013190)

[Related Works 2](#_Toc152013191)

[Dataset 3](#_Toc152013192)

[Features: 3](#_Toc152013193)

[Target Variable: 4](#_Toc152013194)

[Key Observations: 4](#_Toc152013195)

[Methodology 4](#_Toc152013196)

[Exploratory Data Analysis (EDA) 4](#_Toc152013197)

[Feature Selection 8](#_Toc152013198)

[Selected Features 8](#_Toc152013199)

[Train-Test Split 9](#_Toc152013200)

[Scaling 9](#_Toc152013201)

[Model Selection 10](#_Toc152013202)

[Hyperparameter tunning 10](#_Toc152013203)

[Best Parameters 10](#_Toc152013204)

[Evaluation 11](#_Toc152013205)

[Conclusions and future works 12](#_Toc152013206)

# Introduction

The telecommunications industry operates in a highly competitive landscape, making customer retention a critical factor in its success. One of the key challenges faced by telecom service providers is customer churn. Churn prediction, using a Call Detail Records (CDR) dataset, is the focus of this project. Churn prediction is the process of identifying customers or subscribers who are likely to discontinue their use of a telecommunication service, primarily based on their usage patterns and behavior. The primary objective of this project is to develop a predictive model that can reduce customer churn, thereby helping telecom companies retain valuable customers. Below are the several key concerns why churn prediction is important and interesting for businesses, especially in the telecommunications industry.

**1. Revenue Protection:** Churn prediction is critical for protecting a company's revenue. Every lost customer represents a direct financial loss. By identifying customers at risk of leaving and implementing targeted retention strategies, companies can reduce revenue erosion.

**2. Cost Savings:** Acquiring new customers is often more expensive than retaining existing ones. Churn prediction enables companies to allocate resources more efficiently by focusing on retaining at-risk customers, reducing the need for extensive customer acquisition efforts.

**3. Customer Satisfaction:** High churn rates can be indicative of customer dissatisfaction. Churn prediction provides insights into why customers might be leaving, helping companies address pain points and enhance overall customer satisfaction.

**4. Personalized Customer Engagement:** Churn prediction models can identify the specific factors leading to customer churn. This information allows companies to tailor their engagement strategies, offering personalized solutions and incentives to retain customers.

**5. Enhanced Marketing:** Churn prediction can improve the effectiveness of marketing campaigns by targeting at-risk customers with relevant promotions and offers, increasing the chances of customer retention.

**6. Business Growth:** Reducing churn not only protects existing revenue but can also lead to organic growth. Satisfied, loyal customers are more likely to refer others and become brand advocates.

# Problem Statement

Telecom companies encounter a formidable challenge in curbing customer churn to competitors, a phenomenon that significantly impacts their bottom line. Retaining existing customers emerges as a more cost-effective strategy than acquiring new ones, underscoring the critical importance of accurate churn prediction.

Churn manifests in diverse forms, presenting a multifaceted problem for these companies. Three distinct types—active, passive, and rotational churn—demand nuanced predictive approaches. Active churners make a conscious decision to switch providers, while passive churners undergo service discontinuation initiated by the company. Rotational churn, perhaps the most elusive to predict, unfolds when customers terminate contracts without prior notification.

This predictive challenge is compounded by various influencing factors, notably customer satisfaction, pricing structures, and service quality. The intricate interplay of these elements directly influences revenue and market share, necessitating a proactive strategy to mitigate customer attrition.

In this context, our project seeks to address the intricate landscape of churn prediction. By identifying key drivers and delineating customer segments prone to attrition, we aim to furnish telecom companies with actionable insights. Our objective extends beyond prediction; we aspire to recommend targeted retention strategies, ultimately fostering enhanced loyalty and bolstering profitability.

The underlying premise is rooted in the economic wisdom that retaining existing customers proves more economical than acquiring new ones. Beyond the reduction in marketing and sales costs, existing customers often exhibit increased service usage and possess the potential to bring in additional referrals. This project stands as a proactive response to the imperative of measuring and managing churn, offering telecom companies a robust toolset for sustaining customer loyalty and fortifying their competitive position in the dynamic market landscape.

# Related Works

Churn prediction in the telecommunications industry has encompassed a variety of techniques and approaches. Below are some examples of methodologies employed within the industry.

1. Telco Customer Churn Analysis: EDA & Classification Models

This analysis, viewed from a business perspective, delves into the intricate landscape of Telco customer churn. The study explores exploratory data analysis (EDA) techniques and various classification models. It provides valuable insights into the business implications of churn, likely focusing on revenue implications, customer retention strategies, and the cost-benefit analysis of retaining existing customers versus acquiring new ones.

1. Churn classification model for local telecommunication company based on rough set theory

This study presents a churn classification model tailored to a local telecommunication company. The distinct characteristic lies in its use of rough set theory, which diverges from traditional methods. While akin to our current project's objective, the difference in methodology holds promise for unique insights into churn prediction. Exploring the application of rough set theory in churn classification is expected to provide different perspectives and potentially alternative solutions.

1. Churn Modelling - Bank customer

This dataset contains details of a bank's customers, with the target variable reflecting whether the customer closed their account or remained. While relevant to churn modeling, its limitations in the number of features prompt a deviation from adopting a similar approach in this project.

1. Predicting Credit Card Customer Segmentation

Predicting Credit Card Customer Segmentation: This dataset offers a comprehensive collection of customer information within a consumer credit card portfolio, primarily for predicting customer attrition. It includes a plethora of demographic details, customer-card provider relationships, and spending behavior metrics related to churn decisions. Despite its comprehensive nature, the lack of experience within the banking industry discourages the replication of this approach, leading the project to focus on familiar telecom industry datasets.

# Dataset

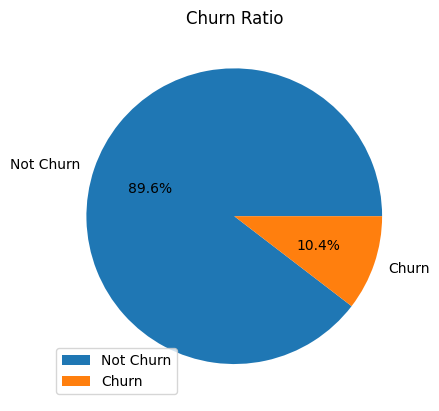
We used the dataset which is published on Kaggle. It contains historical records for 101,174 customers with 17 features.

## Features:

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Features** | **Data Types** | **Description** |
| 1 | Phone Number | object | Sim Number |
| 2 | Account Length | int64 | Active duration of accounts |
| 3 | VMail Message | int64 | Amount of voice mail messages |
| 4 | Day Mins | float64 | Total day call minutes clients have used |
| 5 | Day Calls | int64 | Total day call count clients have made. |
| 6 | Day Charge | float64 | Total day call charged amount clients have made. |
| 7 | Eve Mins | float64 | Total evening call minutes clients have used |
| 8 | Eve Calls | int64 | Total evening call count clients have made. |
| 9 | Eve Charge | float64 | Total evening call charged amount clients have made. |
| 10 | Night Mins | float64 | Total night call minutes clients have used |
| 11 | Night Calls | int64 | Total night call count clients have made. |
| 12 | Night Charge | float64 | Total night call charged amount clients have made. |
| 13 | Intl Mins | float64 | Total internaltional call minutes clients have used |
| 14 | Intl Calls | int64 | Total internaltional call count clients have made. |
| 15 | Intl Charge | float64 | Total internaltional call charged amount clients have made. |
| 16 | CustServ Calls | int64 | Total count of customer service calls made |
| 17 | Churn | Boolean | 1 = churner, 0 = non-churner |

## Target Variable:

Churn: The binary target variable indicating whether a customer has churned (TRUE) or not (FALSE). The ratio of Churners to Non-Churner around 90% to 10%.



## Key Observations:

* The dataset covers various dimensions of customer behavior, including call patterns, usage charges, and engagement with voicemail services.
* The target variable 'Churn' provides a binary indicator of customer attrition, which is vital for our predictive modeling.

# Methodology

In this section, we present a comprehensive overview of the dataset through summary statistics, aiming to understand the distribution and characteristics of the features. The dataset comprises 101,174 records, each associated with various telecommunications-related metrics.

## Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a critical phase in the data analysis process where the main objective is to understand the structure, patterns, and characteristics of the dataset. EDA involves a combination of statistical and visual methods to derive insights, identify patterns, and uncover hidden relationships within the data.

**Visualization:** The bar chart analysis reveals that a significant portion of our dataset exhibits positively skewed distributions across various features as seen in figure-1. This skewness indicates a distinct pattern in the data distribution that is essential to understand. Here's an explanation of why certain features are positively skewed:

**Interpretation:** Positive skewness often arises due to the natural variation in the data. Certain characteristics might naturally lead to higher values for a subset of observations, causing the distribution to extend towards the right.

**Data Collection Constraints:** The measurement limitations or constraints during data collection can contribute to positive skewness. For instance, upper bounds on certain metrics might lead to a concentration of values below the limit.

**Behavioral Patterns:**

Consumer Behavior: In the context of telecom industry data, certain features related to customer usage or engagement may exhibit positive skewness. For example, a majority of customers might fall within a usage pattern that aligns with lower values, while a smaller subset engages in more intensive usage.

**Outlier Presence:**

Extreme Values: The presence of outliers with higher values can heavily influence the mean, pulling it towards the right. These outliers might represent exceptional cases or unique scenarios that contribute to the positive skewness. As in figure-2, we found a lot of outliers there.

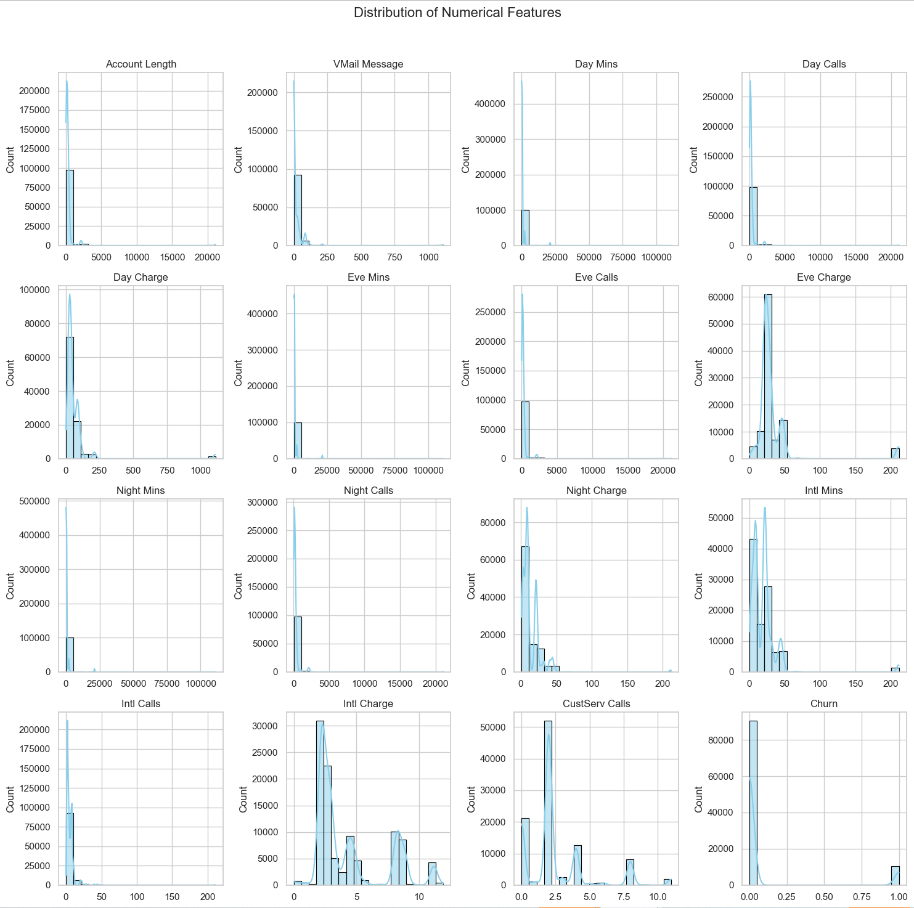
****

Figure-1: Distribution of features, most are positively skewed

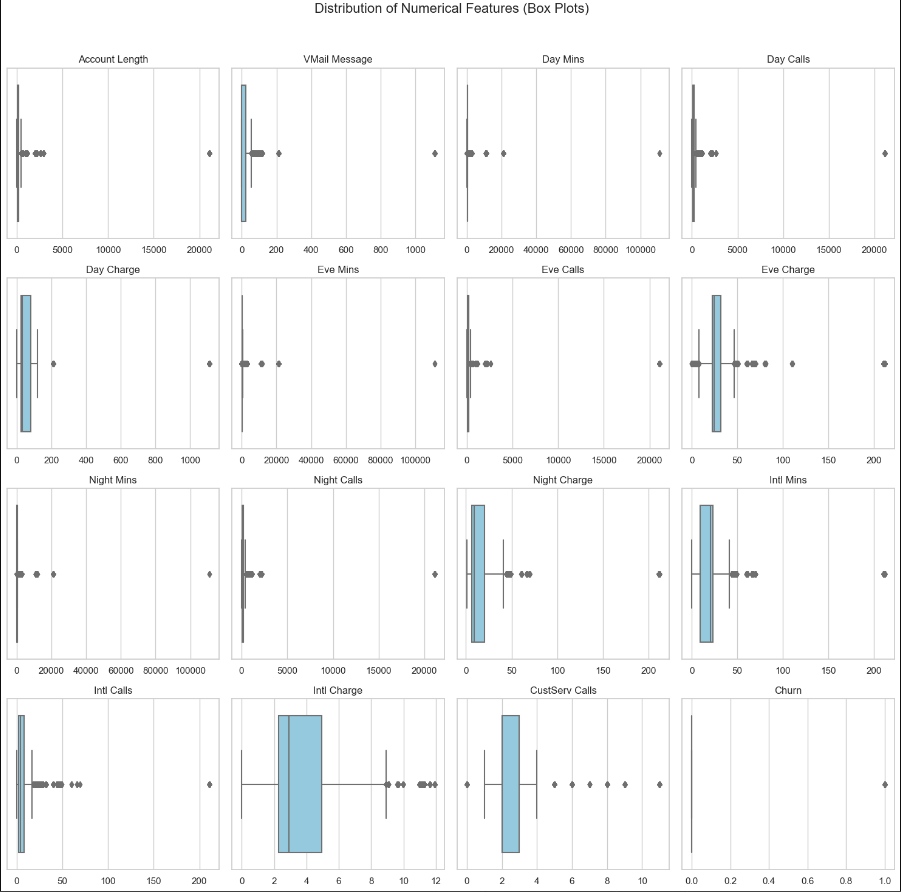
****

Figure-2: Distribution of features in box plots, found outlier in every feature

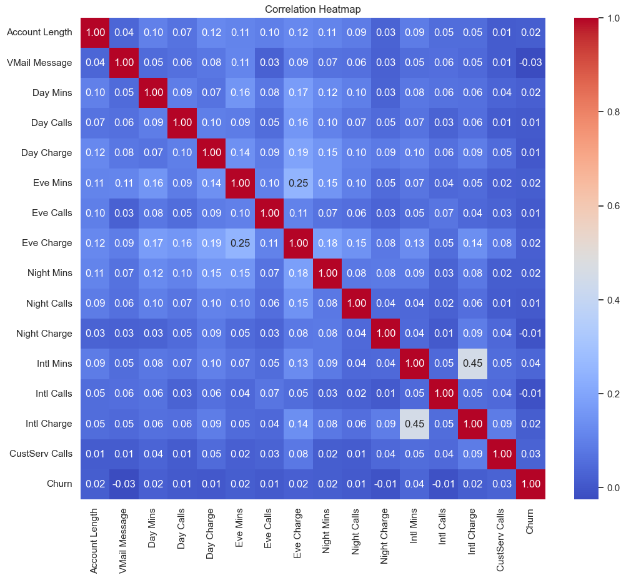
****

Figure-3: Correlation heatmap of features

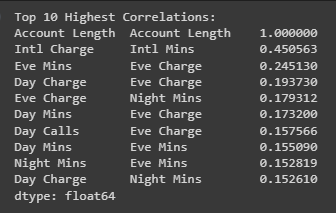
****

Figure-4: Top 10 Highest Correlations

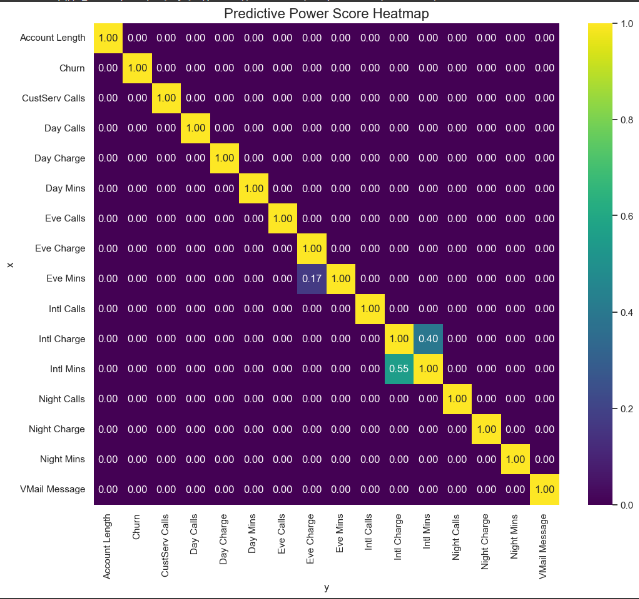
****

Figure-5: Predictive Power Score Heatmap

## Feature Selection

Among 16 features excluding target, we exclude some feature which aren’t much related to transactional data. Among transactional features, we only select total numbers of minutes and call count, we don’t use charge amount because in telecommunication industry, call charged are fixed per minute. In terms of churn, less usage will have high potential to become churner. So that we only focus on the total minutes in term of usage and total number of calls in term of frequency. In addition we consider account length to identify shorten users.

## Selected Features

* **Age on Network (Service Duration):** The inclusion of service duration is essential for understanding customer loyalty and satisfaction over time. PPScore analysis indicates a significant predictive power, suggesting that the service duration is a strong indicator of potential churn. This aligns with the business goal of retaining long-term customers by addressing their evolving needs and preferences.
* **Voicemail Message:** Despite a lower PPScore, the inclusion of voicemail messages is justified as it reflects an additional service that may contribute to overall customer satisfaction. While not the strongest predictor, considering voicemail usage is still relevant for tailoring services based on specific customer preferences.
* **Total Minutes (Total Mins) and Total Calls (Total Calls) for Day/Evening/Night/International:** Aggregating usage metrics across different time periods and call types is crucial for gaining a comprehensive view of customer behavior. PPScore analysis indicates a strong predictive power, emphasizing the significance of these features in understanding usage patterns that may influence churn.
* **Number of Calls to Customer Service:** This feature, with a notable PPScore, directly correlates with customer satisfaction and issue resolution. A higher number of calls to customer service is a strong predictor of churn, emphasizing the need to address customer concerns promptly. Analyzing the reasons behind these calls can guide strategies for enhancing customer support and minimizing churn risk.

## Train-Test Split

We split the testing set with 0.33 ratio and use stratify focusing on target column to make sure that we have same ratio of classes in both the training and testing sets.

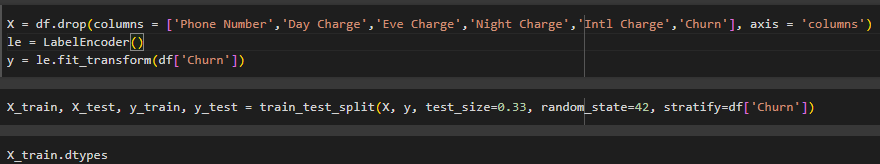


Figure-6: Split the test-train data

## Scaling

In preparation for model training, the feature matrices are standardized using the StandardScaler from scikit-learn. Standardization is a crucial preprocessing step that ensures all features share a consistent scale. This process involves transforming the data to have a mean of 0 and a standard deviation of 1.

* **Consistent Scale:** Standardization is essential when dealing with features that have different units or scales. It guarantees that all features contribute equally to the model, preventing those with larger scales from overshadowing others.
* **Enhanced Model Stability:** Machine learning models, especially those relying on distance metrics or optimization algorithms, benefit from standardized features. It promotes stable model training and convergence, contributing to reliable and accurate predictions.
* **Uniform Treatment of Training and Test Data:** The scaling parameters learned from the training data are applied consistently to the test data, ensuring a seamless transition and maintaining the integrity of the preprocessing pipeline.

Incorporating feature standardization into the workflow improves the overall performance and robustness of the machine learning models, setting the stage for effective predictive analytics.

## Model Selection

In the field of churn prediction, various models have been employed individually to forecast customer attrition. However, our approach is distinct as it combines the strengths of different models. By utilizing an odd number of models (typically 3 or 5), we aim to avoid tie results and enhance the predictive accuracy. These models will incorporate unique sets of features, with some features being common across all models. Below is the process flow and sample of how the model is working.

1. Generate Predictions from Each Model:
   * Each model independently generates predictions, where a value of 1 indicates "Churn," and 0 denotes "Not Churn."
2. Calculate the Average Prediction:
   * The average of predictions from each model is computed.
3. Apply the Sigmoid Function:
   * The average prediction is passed through a sigmoid function, transforming the value into a range between 0 and 1.
4. Final Churn Classification:
   * If the result of the sigmoid function is greater than or equal to 0.5, the final classification is "Churn"; otherwise, it is "Not Churn".

Below is the sample calculation methodology of our churn prediction.



## Hyperparameter tunning

All the selected features are equally used in each model to generate the prediction. The effectiveness of a model relies significantly on its hyperparameters, underscoring the importance of identifying values that yield optimal results. Consequently, our study implements hyperparameter tuning to elevate model performance through the exploration of various value sets.

## Best Parameters

|  |  |  |
| --- | --- | --- |
| *Logistic Regression*   * C: 0.001 * Penalty: L2 * Solver: lbfgs | *KNN*   * Metric: Euclidean * n\_nighbors: 5 * weights: distance | *Random Forest*   * max\_features: sqrt * n\_estimators: 100 |
| *Gradient Boosting*   * learning\_rate: 0.01 * loss: exponential * max\_dept: 1 * max\_features: log2 * min\_samples\_leaf: 0.1 * min\_samples\_split: 0.1 * n\_estimators: 5 * subsample: 0.5 | *XGBoost*   * learning\_rate: 0.1 * max\_depth: 10 * n\_estimators: 180 * random\_state: 42 |  |

# Evaluation

The evaluation results based on the predictive result and actual labels is presented through confusion matrices and use Accuracy, Precision, Recall and F1-Score for performance measurement.

|  |  |
| --- | --- |
| **Training Score** |  |
|  |  |
| **Feature Importance** | |
| **Testing Score** |  |
|  |  |

# Conclusions and future works

In conclusion, our model has achieved an accuracy score of 96% with an F1-Score of 0.98 using the public dataset on Kaggle. All the features we selected demonstrate a significant impact on classifying churn or non-churn instances. Among these features, 'Vmail-Message' exhibits the highest score for feature importance. In a real scenario comparison, customers who are likely to churn tend to have a high volume of voicemail messages, potentially indicating that they have removed the SIM card from their mobile device.

As the future work, it will be very useful if we combine with customer segmentation based on predicted lifetime value. This could lead to more nuanced customer targeting and retention strategies. In addition, it will be required to add additional features such as subscription revenue, average revenue per user (ARPU), data usage, visitor location register (VLR) if we want to use our churn prediction model in the real business.