Customer Churn Prediction

By Vishnu Singh

Introduction:

Customer churn, a critical concern for businesses across industries, refers to the phenomenon where customers cease using a company's products or services. Understanding and mitigating customer churn is paramount, as it directly impacts a company's revenue and growth. In this analysis, we delve into the factors influencing customer churn, aiming to uncover insights that can aid in retaining valuable customers and ensuring long-term business sustainability.

Problem Statement:

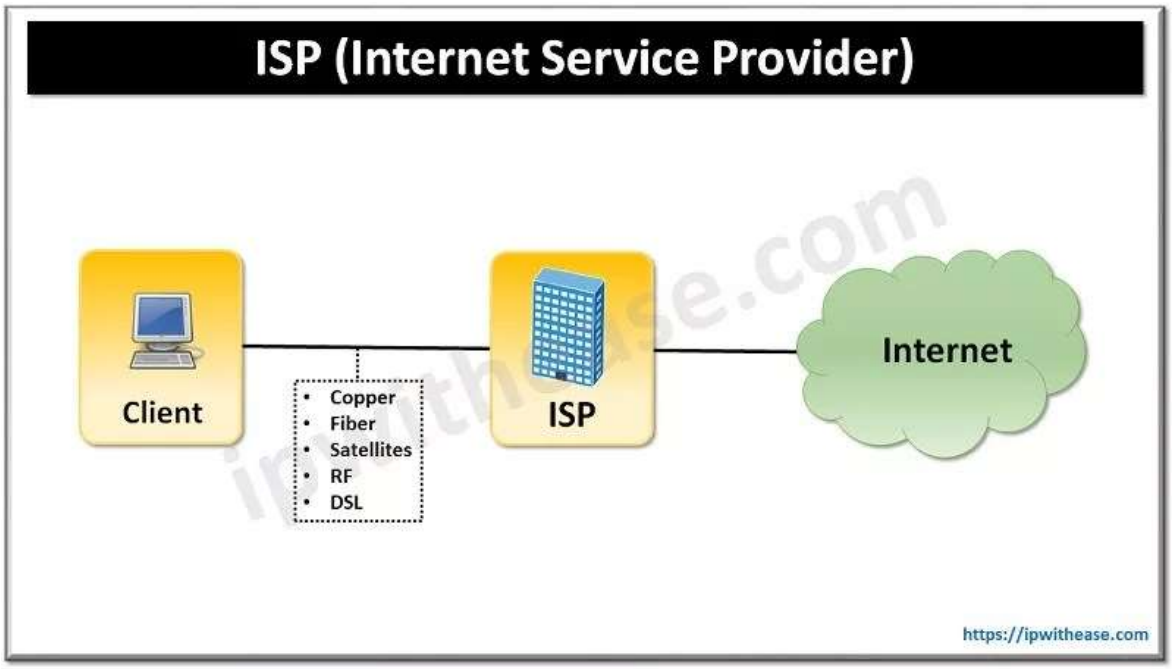
To develop a machine learning model to predict customer churn based on historical customer data. A typical machine learning project pipeline, from data preprocessing to model deployment has to be followed.

Initial Research and Insights:

Based on the columns in the dataset, it appears that this dataset is related to a telecommunications or internet service provider business. The columns likely represent customer information, such as personal details (Name, Age, Gender), location data (Location), subscription-related information (Subscription_Length_Months, Monthly_Bill, Total_Usage_GB), and a target variable indicating whether the customer has churned (Churn).

Business Understanding:

The project appears to be centered around an Internet Service Provider (ISP) business. In the competitive landscape of providing internet services, understanding and addressing customer churn is critical. Customers typically subscribe to a plan and generate monthly bills based on their data usage.



However, customer satisfaction and service quality play a significant role in whether customers decide to continue their subscriptions (Churn). Analyzing customer data, including demographic factors like age and gender, as well as usage patterns, can help the ISP identify trends and patterns that contribute to churn. The goal is to reduce churn by improving service quality, customer support, and tailored offerings, ultimately ensuring the long-term success and growth of the ISP business.

In the ISP industry, customer retention is essential for sustained profitability. By analyzing the dataset, the ISP can gain insights into customer behavior and preferences, helping them tailor marketing strategies and service enhancements.

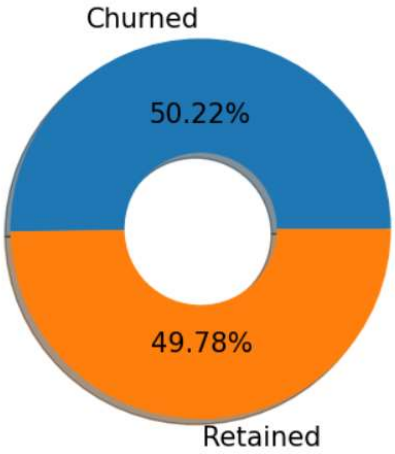
Data Pre-Processing:

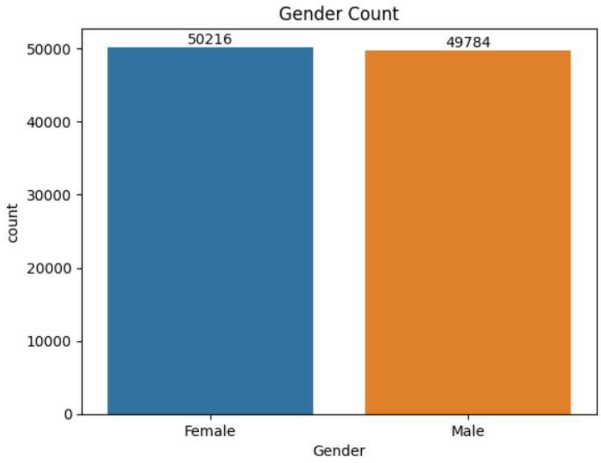
- **Data Size**: The dataset comprises 100,000 rows and 8 columns.
- **Data Types**: It consists of three categorical columns and five numerical columns.
- **Target Variable**: The 'Churn' column serves as the target variable, with values 0 (indicating not churned) and 1 (indicating churned).
- **Exclusion of 'Name'**: The 'Name' column, being a unique identifier, is not useful for the machine learning model and has been dropped.
- **Data Completeness**: The dataset exhibits no missing or null values, ensuring data integrity.
- **Data Uniqueness**: There are no duplicate rows present in the dataset, ensuring each entry is distinct.
- **Missing Values:** No Missing values or duplicate rows were found in the dataset.
- **Problem Type**: This analysis addresses a binary classification problem, focusing on predicting customer churn (0: not churned, 1: churned).

Exploratory Data Analysis:

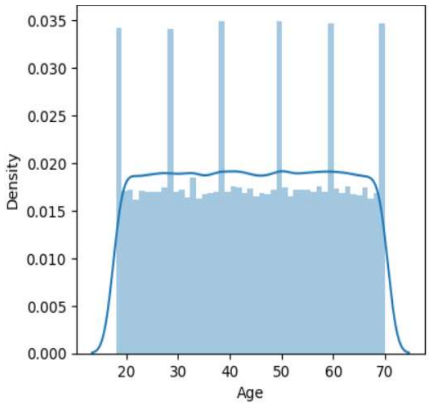
Exploratory Data Analysis (EDA) is a data analysis approach that involves summarizing, visualizing, and understanding the main characteristics of a dataset to uncover patterns, relationships, and insights, aiding in subsequent data modeling and decision-making processes. It helps analysts and data scientists gain a deeper understanding of the data's structure and content.

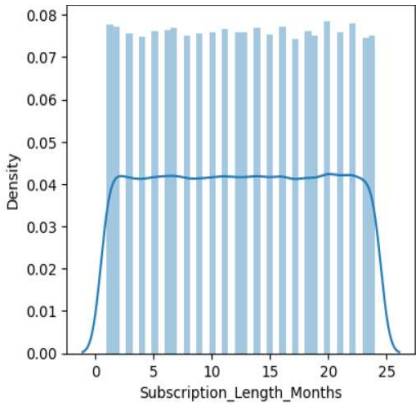
- The number of Churned and non-Churned customers was almost equal.
- The number of customers from each gender are almost equal.
- There are four locations from where customers belong namely Houston, Los Angeles, Miami, Chicago, New York.
- All the locations have almost equal number of customers.
- All the numeric columns like ‘Age’, ‘Monthly Bill’ etc. have uniform distributions.
- The distribution of ages for Churned and non- Churned customers is similar.

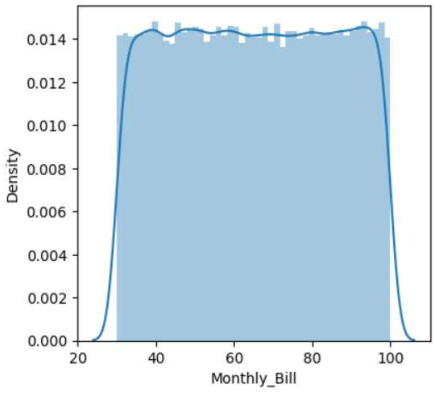


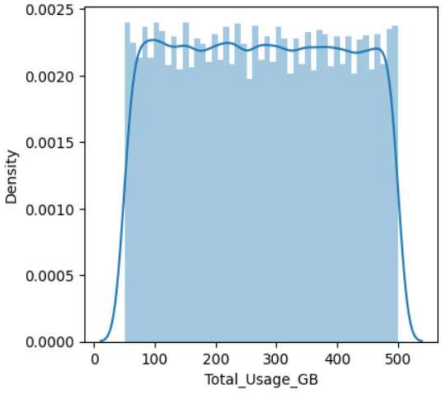


- The number of Churned and non- Churned customer was found to be equal for all locations as well as for both genders.
- No relation was observed between average subscription lengths of Churned and non- Churned customers.
- From statistical analysis, it was found that Location has some effect on customer churn.





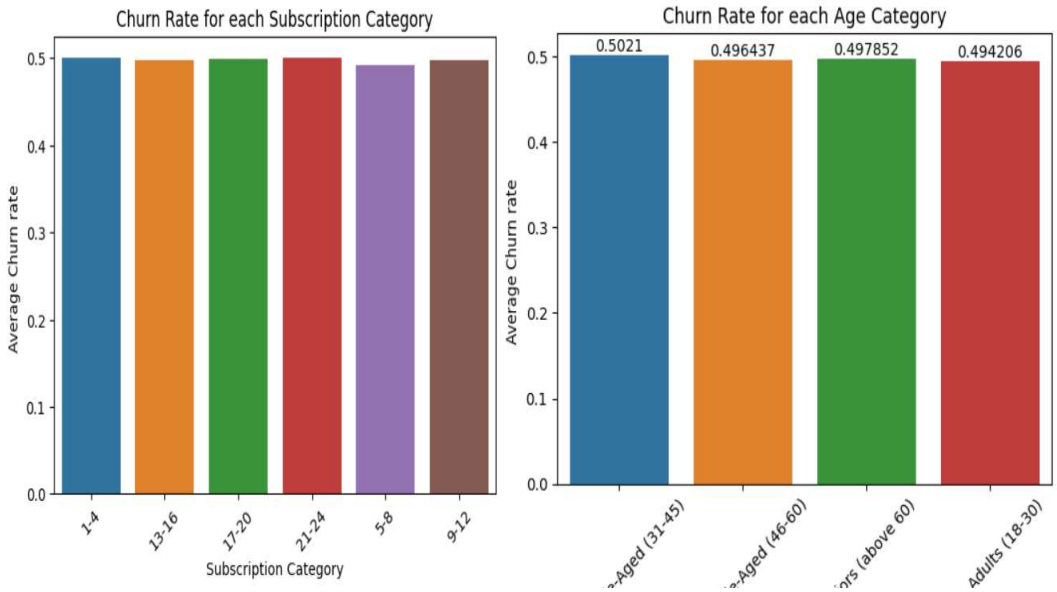




Feature Engineering:

In response to the initial lack of useful insights from EDA, feature engineering was employed to create new features for analysis.

- **Age Category**: A categorical feature 'Age Category' was introduced to group customers based on age ranges, allowing for a more granular analysis of age-related patterns.
- **Subscription Length Category:** Similar to age, 'Subscription Length Months Category' was created to categorize customers based on the duration of their subscriptions.
- **Churn Analysis by Categories:** The impact of churn was examined within these newly created categories to determine if specific age or subscription length groups were more prone to churn.
- **Interaction Feature:** An interaction feature was generated by dividing 'Monthly Bill' by 'Total Usage' to explore whether combined usage and billing patterns had any significant effect on churn.
- **Limited Insights**: Despite these efforts, no substantial new insights or patterns were discovered from the newly engineered features.
- **Feature Pruning:** Consequently, the additional features were deemed uninformative and were removed from the analysis.



Model Selection:

As the problem is a classification problem, the following algorithms were selected.

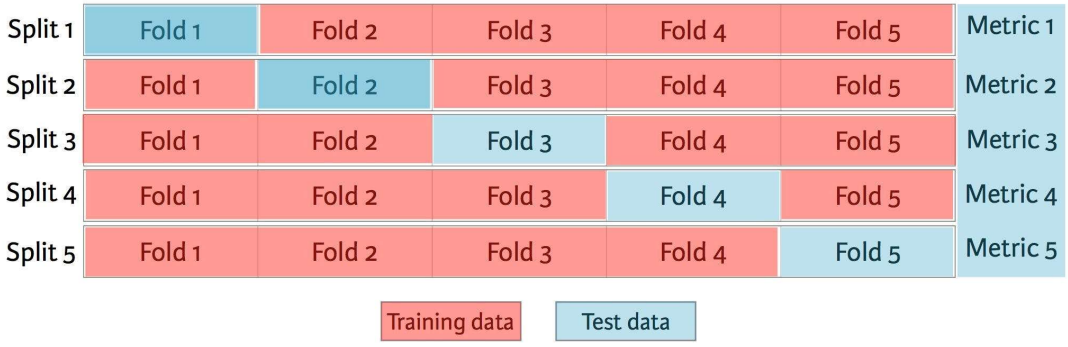
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- XG Boost

All the above mentioned algorithms can be used for binary classification problem but only the one which performs the best on the given dataset will be selected.

5-Fold Cross validation technique was used to determine the mean accuracy of these models.

Cross Validation method:

- Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data.
- It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds.
- Here 5-fold cross validation will be used.



- The performance metric used was accuracy.
- Mean accuracy of all the five iterations for each model was evaluated.

Results of Cross Validation method:

	Model	Mean Accuracy %
0	Logistic Regression	49.922857
1	Decision Tree	50.494286
2	Random Forest	49.818571
3	XG Boost	50.124286

- All the models performed equally unsatisfactory.
- Decision Tree classifier was selected as the final model.
- The poor performance of the models was expected due to no patterns in data and no relation of features with churn.

Model Training:

The final model was trained on the training dataset. This was done using a pipeline which would help us in future during model deployment.

The pipeline consists of two steps:

1. Scaling the numerical features using MinMaxScaler and encoding the categorical features using One Hot Encoder.
2. Training the model on the transformed data.

Performance of the model:

	precision	recall	f1-score	support
0	0.50	0.50	0.50	15152
1	0.49	0.50	0.50	14848
accuracy			0.50	30000
macro avg	0.50	0.50	0.50	30000
weighted avg	0.50	0.50	0.50	30000

To understand the performance metrics, we must be familiar with following terms:

- **True Positives (TP):** It represents the number of positive samples that were correctly identified by the model.
- **True Negatives (TN):** It represents the number of negative samples that were correctly identified by the model.
- **False Positives (FP):** It represents the number of negative samples that were incorrectly classified as positive.
- **False Negatives (FN):** It represents the number of positive samples that were incorrectly classified as negative.

Accuracy:

- Accuracy is a common evaluation metric used in binary classification tasks to measure the overall performance of a machine learning model.
- It represents the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of instances in the dataset.
- Accuracy = (TN + TP) / (TN + FP + FN + TP)

Precision:

- Precision is an evaluation metric used in binary classification to measure the model's ability to correctly identify True Positives among all instances predicted as positive (True Positives + False Positives).
- Precision = TP / (TP+ FP)

Recall:

- Recall is an evaluation metric used in binary classification to measure the model's ability to correctly identify True Positives among all actual positive instances (True Positives + False Negatives).
- Recall = TP / (TP+FN)

F1 Score:

- The F1 score is an evaluation metric used in binary classification that combines both precision and recall into a single score
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

The overall performance of the model was unsatisfactory.

Hyper Parameter Tuning:

The hyper parameter tuning process is aimed to optimize the performance of the machine learning model by systematically searching through a range of hyper parameters to find the best combination.

- **Algorithm Choice:** The Decision Tree Classifier was chosen as the base model for hyper parameter tuning due to its flexibility and suitability for binary classification tasks.
- **Grid Search:** Grid Search Cross-Validation was employed as the tuning technique. It involved specifying a grid of hyper parameters to explore different combinations systematically.
- **Hyper Parameters Tuned:** The hyper parameters tuned included 'criterion' (splitting criterion), 'max_depth' (maximum depth of the tree), 'min_samples_split' (minimum number of samples required to split an internal node), and 'min_samples_leaf' (minimum number of samples required to be at a leaf node).
- **Outcome:** Despite the extensive search for optimal hyper parameters, the tuning process did not yield substantial improvements in the model's performance. The model still achieved an accuracy of approximately 50%, indicating the inherent complexity of the dataset.

0	0.51	0.52	0.51	15152
1	0.50	0.48	0.49	14848
accuracy			0.50	30000
macro avg	0.50	0.50	0.50	30000
weighted avg	0.50	0.50	0.50	30000

There was no significant change in the model performance.

Model Deployment:

- **Model Export:** The trained machine learning model was exported using the 'pickle' library, making it accessible for real-time predictions within the web app
- **Streamlit Web App**: The model was deployed using a Streamlit web application, providing a user-friendly interface for interacting with the predictive model.

- **User Input**: The web app allows users to input relevant features or data points necessary for making predictions.
- **Prediction:** Users can submit their data through the app, and the model, loaded from the exported file, provides predictions based on the provided input.
- **User-Friendly Interface:** The Streamlit app offers an intuitive and straightforward interface, enabling users to access the model's predictive capabilities without requiring expertise in machine learning or programming.
- **Practical Application:** This deployment enhances the practical utility of the model, making it available for real-world decision-making scenarios, such as identifying potential churn among customers in the ISP business.

Conclusion:

- **Data Balance and Uniformity:** The dataset was found to be highly balanced and uniformly distributed across all features, indicating an equal representation of target classes and consistent feature scales.
- **Inconclusive EDA:** Exploratory Data Analysis did not yield significant insights or reveal any discernible trends or patterns in the data, suggesting that the data might be inherently challenging to model.
- **Feature Engineering Efforts:** Despite attempts at feature engineering, no meaningful patterns or relationships were discovered in the data. This suggests that the inherent complexity of the problem may require more sophisticated feature engineering techniques.
- **Model Performance:** Initial models, including a Decision Tree Classifier, achieved only modest accuracy (around 50%), indicating that the dataset might be inherently difficult to predict based on the given features.
- **Hyper Parameter Tuning:** Hyper Parameter tuning was performed, but it did not substantially improve the model's performance, suggesting that the limitations might be related to the data itself.
- **Streamlit Web App:** To make the model accessible and user-friendly, a Streamlit web app was created, allowing users to input data and receive predictions based on the trained model.

Future Work:

- **More Data Collection:** In cases where data is limited or inconclusive, collecting more data, especially diverse or domain-specific data, could potentially improve model performance.
- **Advanced Modelling:** Explore more complex models or ensemble methods to capture intricate patterns that simpler models might miss.
- **Domain Expertise:** Collaborate with domain experts who may provide valuable insights or domain-specific knowledge to enhance the analysis.

Thank You