**Machine Learning Internship Assessment**

Customer Churn Prediction

Ankita Kanoujiya

**Table of Contents**

* Introduction
* Data Description
* Exploratory Data Analysis (EDA)
* Data Visualization
* Label Encoding
* Outlier Detection and checking the distribution of data
* Check for Class Imbalance
* Data Splitting
* Feature Selection Using Random Forest Feature Importance:
* Feature Scaling
* Model Building: Using Machine Learning
* Model Building: Ensembles of Random Forest
* Hyperparameter Tunning on XGBoost Classifier
* Bayesian optimization in XGBoost Classifier
* Cross-Validation
* Model Building: Using Neural Network
* Model Evaluation
* Conclusion

**Introduction:**

Customer churn refers to the phenomenon where customers stop doing business with a company or cancel their subscription or services. It is a crucial concern for businesses as it can lead to revenue loss and a decline in customer satisfaction**.**

**Problem Statement:**

The problem at hand is to develop a machine learning model that predicts customer churn based on historical customer data. The goal is to identify factors or patterns that contribute to customer attrition and proactively address those issues to retain customers.

**Data Description:**

Dataset consists customer information for a customer churn prediction problem. It includes the following columns:

**CustomerID**: Unique identifier for each customer.

**Name**: Name of the customer.

**Age**: Age of the customer.

**Gender:** Gender of the customer (Male or Female).

**Location**: Location of customer.

**Subscription\_Length\_Months**: The number of months the customer has been subscribed.

**Monthly\_Bill:** Monthly bill amount for the customer.

**Total\_Usage\_GB**: Total usage in gigabytes.

**Churn**: A binary indicator (1 or 0) representing whether the customer has churned (1) or not (0).

**Exploratory Data Analysis (EDA):**

In the initial stage, we conducted an exploratory analysis of the dataset to gain insights into its composition and attributes.

* The dataset comprises details of 100,000 customers, encompassing nine distinct variables.
* Each variable was found to possess the appropriate data type, and a comprehensive examination revealed the absence of any missing data or duplicate records.
* calculated descriptive statistics for each variable, enabling a better understanding of the dataset. Unnecessary columns were also identified and removed, while a heatmap was used to extract further insights.

**Data Visualization:**

* Age Distribution: The age distribution of customers primarily spans the age groups of 20-25, 45-50, and 65-70.
* Churn Distribution: The majority of customers, constituting 50.2%, have not churned, while the remaining 49.8% have churned.
* Gender Distribution: The frequency of female customers surpasses that of male customers, with females making up 50.2% and males 49.8% of the total.
* Subscription Length Months Distribution: The most common choice for customers is a 20-month subscription.

**Label Encoding:**

Categorical variables were encoded to numerical values to enable machine learning algorithms to process them effectively.

* Label Encoding was applied to the 'Gender' variables

**Outlier Detection and checking the distribution of data:**

* Detecting and handling outliers is crucial for accurate model performance.
* We visually examined potential outliers using box plots.
* Fortunately, no significant outliers were spotted, indicating the data's stability

**Checking Distribution of Data:**

* We checked if our data is suitable for modelling by looking at its distribution.
* We used histograms to see how numerical data is spread out.
* All variables were found to be approximately normally distributed

**Check for Class Imbalance:**

* The distribution of the Churn variable is almost balanced, with approximately 50,221 instances in the "Not Churn" category and 49,779 instances in the "Churn" category

**Data Splitting:**

* We split the dataset into two parts: one for training the model and the other for testing its performance.
* This division was done in a 70:30 ratio, allocating 70% of the data for training and 30% for testing

**Feature Selection Using Random Forest Feature Importance:**

* To make the model more efficient and understandable, we determined feature importance.
* Using Random Forest Feature Importance, we ranked features by their impact on the target variable. The most influential features were 'Monthly\_Bill,' 'Total\_Usage\_GB,' 'Age,' and 'Subscription\_Length\_Months’

**Feature Scaling:**

* Feature scaling was applied to ensure all variables were on the same scale, aiding model convergence.
* Standardization was applied to variables such as 'Age', 'Subscription\_Length\_Months', 'Monthly\_Bill', and 'Total\_Usage\_GB'.

**Model Building: Using Machine Learning**

* We tried different machine learning methods like Logistic Regression, Decision Tree, Gaussian Naive Bayes, AdaBoost, Random Forest, and XGBoost.
* We used training and test data to see how well each method worked

**Model Building: Ensembles of Random Forest**

* We tried using ensemble models with Random Forest as the base classifiers, but it didn't lead to a noticeable performance improvement.

**Hyperparameter Tunning on XGBoost Classifier:**

* XGBoost Classifier excelled in performance compared to other methods, but even after fine-tuning through both Randomized Search and Grid Search Cross-Validation, the model's performance didn't see significant enhancements.

**Bayesian optimization in XGBoost Classifier:**

* Even Bayesian Optimization didn't lead to improved model performance.

**Cross-Validation:**

* Cross-validation was performed to validate the model's performance and ensure it generalized well to new data
* Cross-Validation Scores: [0.5042 0.49775 0.5067 0.4991 0.50175]

Mean CV Score (Accuracy): 0.5018999999999999

* Cross-Validation Recall Scores: [0.49030638 0.47900763 0.48583769 0.46966653 0.48533548]

Mean CV Recall Score (Recall): 0.4820307413896887

**Model Building: Using Neural Network**

* We tried building a neural network model, but it didn't improve the model's performance as expected

**Model Evaluation:**

* Train & Test Data Metrics

XGBoost Results:

Model Train Accuracy Test Accuracy F1 Score Recall

XGBoost 0.6546 0.5035 0.493729 0.488056

* Confusion Matrix:

Confusion Matrix for XGBoost:

[[5228 4851]

[5079 4842]]

* The top-left cell (5228) represents the true negatives (TN), which are the cases where the model correctly predicted "No Churn."
* The top-right cell (4851) represents the false positives (FP), which are the cases where the model incorrectly predicted "Churn" when the actual label was "No Churn."
* The bottom-left cell (5079) represents the false negatives (FN), which are the cases where the model incorrectly predicted "No Churn" when the actual label was "Churn."
* The bottom-right cell (4842) represents the true positives (TP), which are the cases where the model correctly predicted "Churn."

**Saving Model:**

* The final XGBoost model was saved as a pickle file for future use

**Conclusion:**

This project encompassed extensive data analysis, preprocessing, and algorithm evaluation for customer churn prediction. The XGBoost Classifier emerged as the top-performing model. Although achieving perfect accuracy and recall is challenging, the findings offer valuable insights for enhancing customer retention and business expansion. Future steps may include gathering more data and employing advanced techniques to boost model performance.