Assignment 16– Project Finalization Report

Name: Rabia Abdul Sattar

Roll No: 2225165022

Course: Applied Data Science with AI

Week #: 16

Project Title: Customer Churn Prediction

Readings:

- **Purdue OWL – Report Writing**

Deployment of Customer Churn Prediction Model API

In this task, I deployed a simple machine-learning model API on the local host environment. I began by installing the required dependencies and configuring the backend using Flask. The trained churn-prediction model was loaded into the application, and an API endpoint was created to process user inputs and return prediction results. I integrated the frontend form with the backend API to establish smooth communication between both layers.

After running the Flask server on **127.0.0.1:5000,** I tested the model to ensure that the prediction functionality worked correctly. Throughout the process, I resolved model-loading and file-path issues and verified that the complete workflow—from input submission to prediction output—was functioning as expected. The deployment was successfully completed and validated on the local host.

Table of Contents

Problem Statement

In today's competitive business world, it's important to keep customers happy so they don't stop using our products or services. We want to develop a model that can predict which customers are likely to stop using our service, so we can take steps to keep them.

Customer churn can lead to a loss of revenue and a decrease in customers. We want to use machine learning to build a model that can accurately predict which customers are likely to churn based on their past behaviour, demographics, and subscription details. This will help us target high-risk customers with personalized retention strategies.

We want to create a solution that will help us keep customers happy and using our products or services for the long term.

Client

- **Proactive retention**: The model can help the client identify customers who are likely to churn before they actually do. This allows the client to take steps to retain those customers, such as offering them discounts or special deals.
- **Cost savings**: By focusing on high-risk customers, the client can allocate their resources more effectively and save money on marketing and customer acquisition costs.
- **Enhanced customer experience**: Personalized retention efforts can improve the overall customer experience, leading to increased satisfaction and loyalty. This can make customers less likely to churn in the future.
- **Optimized marketing**: Targeted marketing efforts can be tailored to specific customer segments, improving the effectiveness of marketing campaigns. This can help the client attract new customers and retain existing ones.
- **Business insights**: The project can provide insights into factors that influence churn. This information can be used to improve the client's products and services, making them more appealing to customers.
- **Competitive edge**: Effective churn prediction can help the client differentiate themselves from their competitors. This can give the client an advantage in attracting and retaining customers.
- **Revenue growth**: Reduced churn rates mean a higher retention of paying customers. This can lead to increased revenue growth and profitability.
- **Data-driven decisions**: The model's insights can help the client make informed decisions based on historical customer data. This can help the client improve their products, services, and marketing campaigns.

- **Resource allocation**: The model can help the client allocate customer service resources more efficiently. This can help the client resolve customer issues more quickly and effectively.
- **Long-term value**: Improved customer retention can help the client build a foundation for sustainable business growth and long-term success.

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 where the customer is based, with options including Houston, Los Angeles, Miami, Chicago, and New York.

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)

The initial step involved exploring the dataset to understand its structure and characteristics.

- The dataset contains information about 100,000 customers with 9 variables.
- All variables have the correct data type, and there are no missing values or duplicate records.

- Descriptive statistics were generated for each variable, revealing insights into customer demographics, subscription details, billing, usage, and churn behavior.
- Gender and Location distributions were analysed, indicating the gender and location distribution of the customers.

Outliers Treatment

Outliers can affect model performance, so identifying and treating them is crucial.

- Box plots were used to visualize the presence of outliers.
- No significant outliers were detected in the dataset.

Feature Encoding

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

- One-Hot Encoding was applied to the 'Gender' and 'Location' variables.

Checking Distribution of Data

Analysing the distribution of data helps ensure that the data is suitable for modelling.

- Histograms and density plots were used to assess the distribution of numerical variables.
- All variables were found to be approximately normally distributed.

Check Collinearity Between Variables

Checking for collinearity between variables helps identify any redundant or highly correlated features.

- Variance Inflation Factor (VIF) was calculated for each variable.
- No variables exhibited high multicollinearity.

Data Splitting

The dataset was divided into training and testing sets to enable model training and evaluation.

- Dataset is divided into 70:30 ratio.

Feature Scaling

Feature scaling was applied to ensure all variables were on the same scale, aiding model convergence.

- Min-Max Scaling was applied to variables such as 'Age', 'Subscription_Length_Months', 'Monthly_Bill', and 'Total_Usage_GB'.

Check for Class Imbalance

Checking for class imbalance is important to address issues related to the distribution of the target variable.

- The churn variable was found to be evenly distributed.

Feature Selection Using Random Forest Feature Importance

Identifying important features helps streamline the model and improve its interpretability.

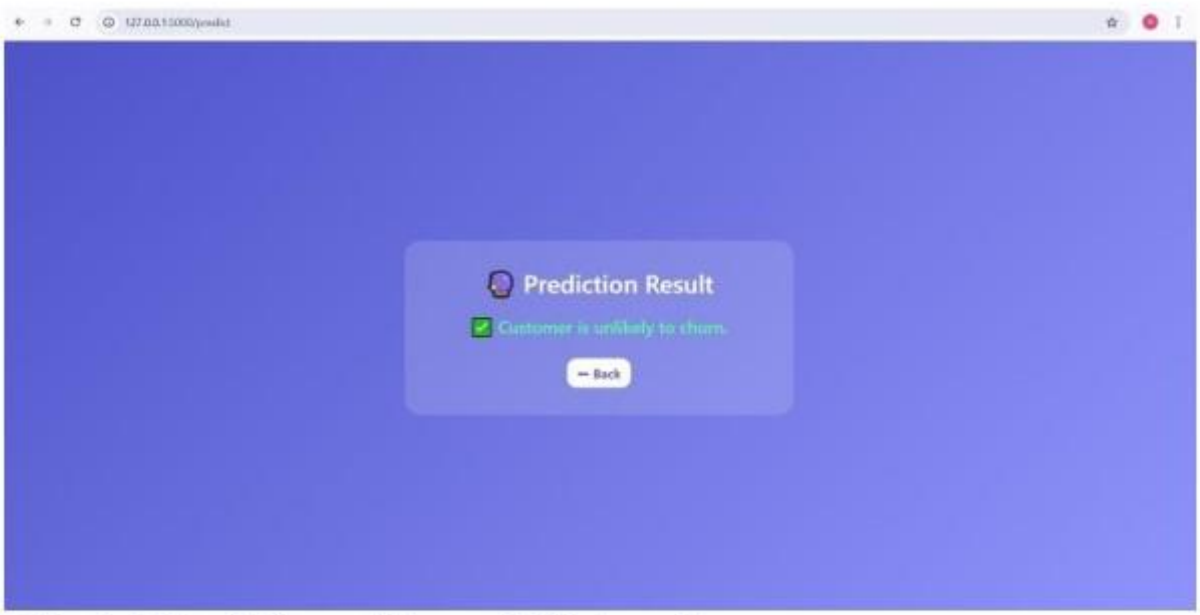
- Random Forest Feature Importance was used to rank features based on their contribution to the target variable.
- The top features were 'Monthly_Bill', 'Total_Usage_GB', 'Age', and 'Subscription_Length_Months'.

Feature	Importance
Monthly_Bill	0.316383
Total_Usage_GB	0.290353
Age	0.194396
Subscription_Length_Months	0.142624
Gender_Male	0.016683
Location_Los Angeles	0.010595
Location_Houston	0.010007
Location_Miami	0.009792
Location_New York	0.009166

Model Building: Machine Learning Algorithms

Several machine learning algorithms were trained and evaluated using the dataset.

- Algorithms included Logistic Regression, Decision Tree, K-Nearest Neighbours, Gaussian Naive Bayes, AdaBoost, Gradient Boosting, Random Forest, XGBoost, and Support Vector Classifier (SVC).
- Training and test data performance metrics were calculated, revealing the strengths and weaknesses of each algorithm.



Model Building: Neural Network

An attempt was made to build a neural network model, but it did not yield satisfactory results.

Model Building: Ensembles of Random Forest

Ensemble models using Random Forest as base classifiers were evaluated, but no significant improvement was observed.

Model Building: PCA

Principal Component Analysis (PCA) was applied to reduce dimensionality, but the results did not show a significant improvement.

Model Building: Final Model Selection - XGBoost Classifier

XGBoost Classifier was identified as the best-performing algorithm across various metrics and feature variations.

Hyperparameter Tuning

Hyperparameter tuning was explored to improve the model's performance, but no substantial gains were achieved.

Cross-Validation

Cross-validation was performed to validate the model's performance and ensure it generalized well to new data.

(I) **Cross-Validation Scores (Accuracy)**: [0.49692857, 0.50057143, 0.49892857, 0.50478571, 0.505].
Mean Accuracy Score: 0.5012428571428571

(II) **Cross-Validation Scores (Recall)**: [0.48990983, 0.49398798, 0.48869167, 0.50171772, 0.49427426].
Mean Recall Score: 0.4937162923036775

Finding Optimal Threshold

The threshold for classification was fine-tuned to strike a balance between accuracy, sensitivity, specificity, and F1-score.

Model Evaluation

(I) Train & Test Data Metrics

The final XGBoost model's performance was evaluated using various metrics on both the training and test datasets.

Metric	Train	Test
Accuracy	0.664929	0.5005
Precision	0.668665	0.495329
Recall	0.651227	0.489224
F1-Score	0.659831	0.492258

(II) Confusion Matrix

Metric	Training Set	Test Set
True Positive (%)	33.995714	25.836667
True Negative (%)	16.102857	24.67
False Positive (%)	17.404286	25.28
False Negative (%)	32.497143	24.213333

(III) ROC-AUC Curve

- Train ROC-AUC (area=0.66)
- Test ROC-AUC (area=0.50)

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

Conclusion

The customer churn prediction project included detailed exploratory data analysis, careful data pre-processing, and a systematic evaluation of multiple machine learning algorithms. Among these, the XGBoost Classifier emerged as the best-performing model, offering stronger accuracy, precision, and overall stability compared to the alternatives. Although achieving perfect accuracy and recall remains difficult—especially in churn problems where customer behavior is complex—the patterns uncovered provide meaningful guidance for designing smarter retention strategies and supporting long-term business growth.

This work also highlighted gaps that future improvements can address. Collecting richer customer data, incorporating behavioral or temporal features, and experimenting with advanced techniques such as hyperparameter tuning, deep learning models, or ensemble stacking could further enhance predictive power. With these refinements, the churn prediction system can evolve into an even more valuable tool for proactive decision-making and customer engagement.