Data Sources

- Data Source: From local machine as I have downloaded the dataset from email.

Data Description

- Size of Dataset: 10000 rows, 16 columns.

Features:
-Gender: Gender define the classification of customer on the basis of gender which values are either male or female.
- Age: Define the age of customers.
- Tenure: Define the time in months that customers have been using internet.
- MonthlyCharges: Charges for customers for internet service.
- TotalCharges: This is the charge for internet connection plus service charge plus monthly fee.
- Contract: Define the customer's contract for internet service. Once the contract completed customers has to do repay for service continuation.
- PaymentMethod: This is the payment channel that customers use to do payment for services.
- InternetService: This is the service that customers using for the internet services like optic fiber etc.
- OnlineSecurity: Define the customers status for who are using online security of the internet service.
- OnlineBackup: Define the customers who are having online backup of their services.
- DeviceProtection: Customers who have been using deviceprotection feature of the internet service.
- TechnicalSupport: Defines customers who have been taking technical support of the service.
- StreamingTV: Customers who are using the feature StreamingTV of the service.
- StreamingMovies: Customers who have been using the feature StreamingMovies of the service.

Target Variable: Churn: Which is the dependent variable which values are depend on the values of independent variables.

Data Preprocessing

Data Cleaning

- Handling Missing Values: We had no missing values in our dataset so nothing to do here.).
- Outliers: No outliers in our dataset..
- Data Transformation: Performed Scaling, normalization, encoding categorical variables and manual encoding on required variables. .

Exploratory Data Analysis (EDA)

- Summary Statistics: Created ss of dataset and got Mean, median, standard deviation of features and create ss of categorical variables also..
- Visualizations: Create visualizations like Histograms, box plots, scatter plots, correlation matrix to understand data distributions and relationships.

Model Selection and Training

Model Choice

- Algorithms: I chose two algorithms Random forest and XGBoost first but finalize Random Forest as it is better for classification problem like we have.

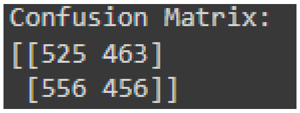
Model Training

- Training Set: 80 % of the whole dataset.
- Validation Set: 20 % of the dataset.

Model Evaluation

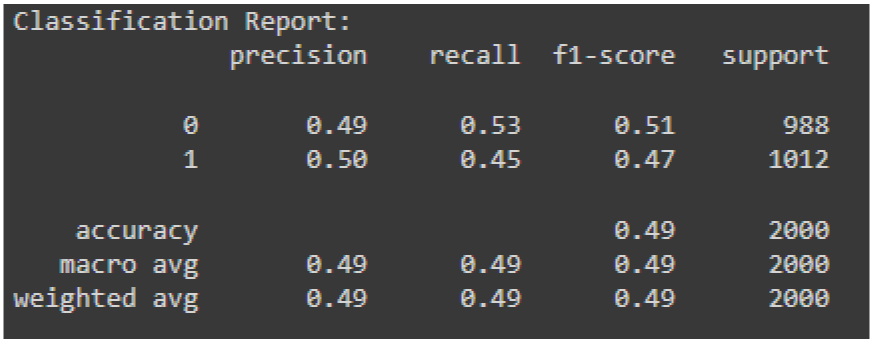
Metrics

- Accuracy: Overall accuracy of the model is 49 %.
- Precision: for 0 0.49,for 1 0.50
- Recall,: For 0 0.53, for 1 0.45
  F1-Score: For 0 0.51, for 1 0.47
- Confusion Matrix:



```
Confusion Matrix:
[[525 463]
 [556 456]]
```

- Classification. Report:



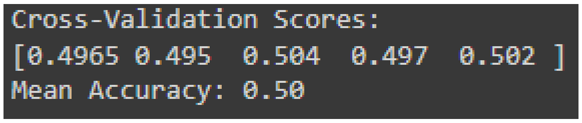
```
Classification Report:
              precision    recall  f1-score   support

           0       0.49      0.53      0.51       988
           1       0.50      0.45      0.47      1012

    accuracy                           0.49      2000
   macro avg       0.49      0.49      0.49      2000
weighted avg       0.49      0.49      0.49      2000
```

Cross-Validation

- Method: I have applied k-Fold Cross-Validation technique.



```
Cross-Validation Scores:
[0.4965 0.495  0.504  0.497  0.502 ]
Mean Accuracy: 0.50
```

Deployment

Model Export

- Format: I export trained model creating pickle file.
- Environment: Local Machine

Integration

- API: Create api end point using flask.
- User Interface: HTML.
- Container: Docker