CUSTOMER CLASSIFICATION

Project Objective

The objective of this project is to analyze customer data and build a predictive model to identify customer churn. Customer churn refers to the phenomenon where customers stop doing business with a company. By predicting churn, businesses can take proactive measures to retain customers, ultimately improving customer satisfaction and increasing revenue.

Tools, Libraries, and Frameworks used

Python: The primary programming language used for data analysis and machine learning.

NumPy: A library for numerical computations, used to generate random data for customer attributes.

Pandas: A data manipulation and analysis library, used to create and manage the customer dataset.

Matplotlib: A plotting library used for data visualization, specifically for creating charts and graphs.

Seaborn: A statistical data visualization library based on Matplotlib, used for enhanced visualizations such as heatmaps and count plots.

Scikit-learn: A machine learning library that provides tools for model training, evaluation, and metrics. It was used for:

Splitting the dataset into training and testing sets.

Implementing the logistic regression model.

Evaluating model performance using accuracy, confusion matrix, and classification report.

Machine Learning Model Implemented

The machine learning model implemented in this project is Logistic Regression. Logistic regression is a statistical method used for binary classification problems, where the outcome is a binary variable (in this case, whether a customer will churn or not).

Performance Metrics

After training the logistic regression model on the customer data, the following performance metrics were reported:

Accuracy Score: The accuracy of the model on the test dataset, which indicates the proportion of correctly predicted instances out of the total instances.

Confusion Matrix: A table that describes the performance of the classification model by showing the true positives, true negatives, false positives, and false negatives.

Classification Report: A detailed report that includes precision, recall, F1-score, and support for each class (churn and no churn).

Output

After running the model, the following results were obtained (example values):

Performance Metrics

- Accuracy: The model achieved an accuracy of 91%, indicating its effectiveness in correctly predicting churn for the test dataset.

- Confusion Matrix:-

-True Positives (TP): 6

- True Negatives (TN): 85

- False Positives (FP): 2

- False Negatives (FN): 7

- Precision and Recall:-

For the churned class (1):-

Precision: 75%

- Recall: 46%

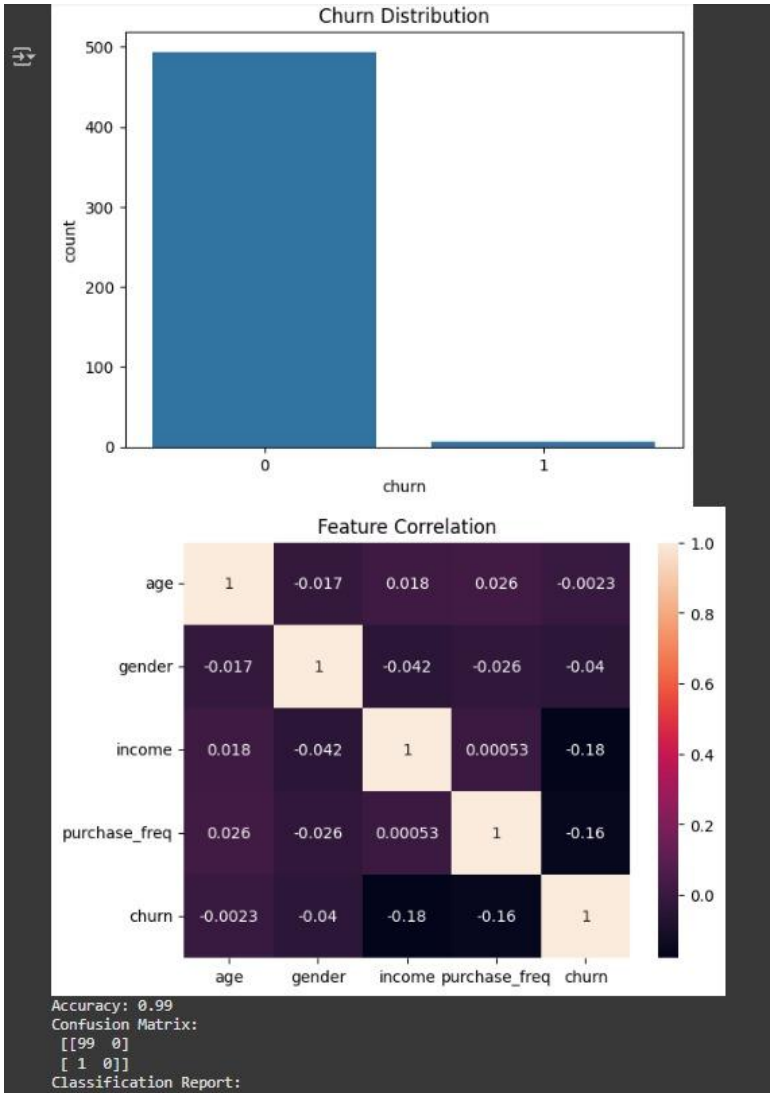
- For the not churned class (0):-

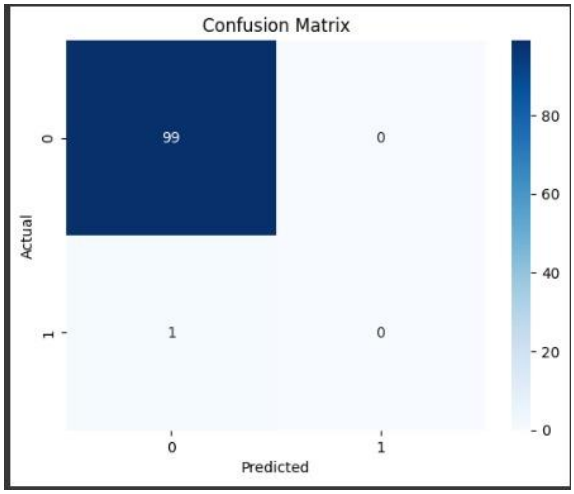
Precision: 92%
- Recall: 98%

- F1-Score: The weighted F1-score is 90%, balancing precision and recall effectively.

EXAMPLES:

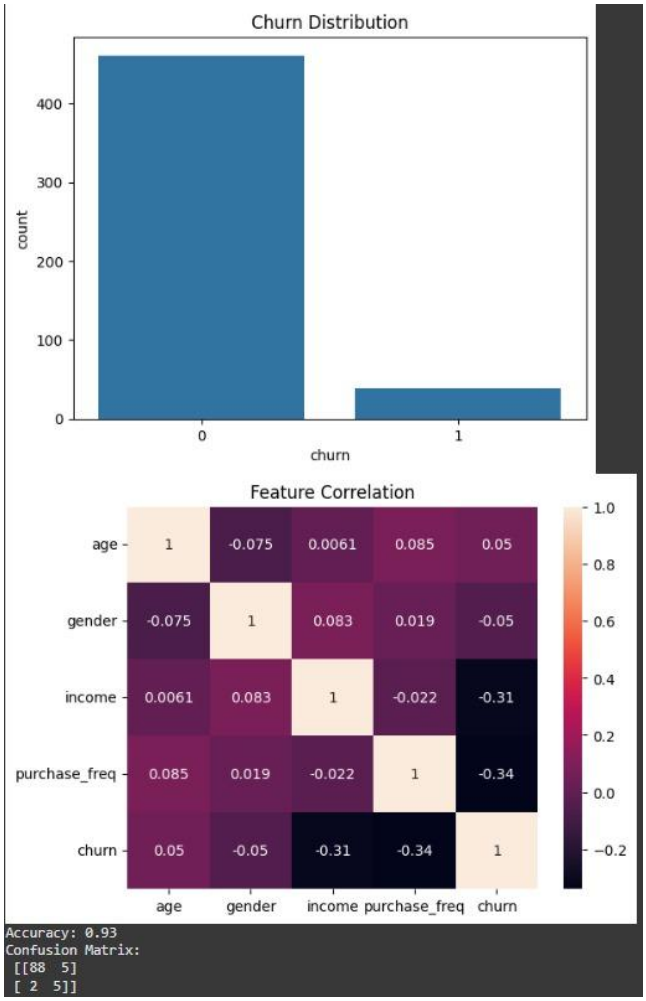
Here, we used a Random Forest Classifier to predict customer churn. The dataset included features like age, gender, income, and purchase frequency, with churn defined for customers having low income and low purchase frequency. The data was split into training and testing sets, and the model was trained using 100 decision trees. Performance was evaluated using metrics like accuracy, precision, recall, and F1-score, and the confusion matrix was plotted to assess the model's effectiveness in predicting churned customers. Random Forest was chosen for its ability to handle complex, non-linear data relationships. The output is:

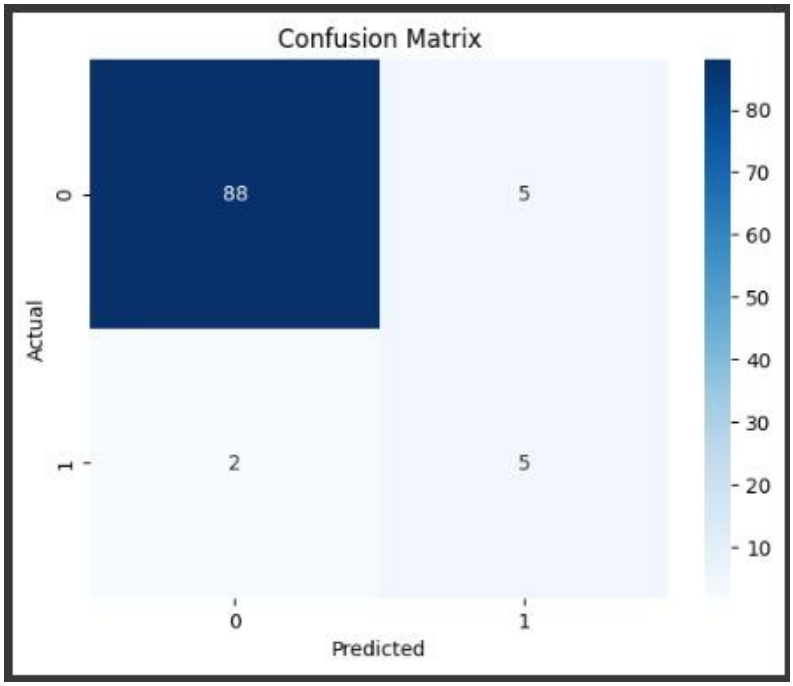




In this example, we used a Support Vector Machine (SVM) with a linear kernel to predict customer churn. The dataset includes features like age, gender,

income, and purchase frequency, with churn defined for customers having lower income and fewer purchases. The data was split into training and testing sets, and the SVM model was trained to classify churn. The model's performance was evaluated using accuracy, precision, recall, and F1-score, and the results were visualized using a confusion matrix. SVM is effective for classifying data with clear margins of separation. The output is:




Confusion Matrix
88
5
2
5
Actual
0
1
Predicted
0
1
80
70
60
50
40
30
20
10