Assignment 6 – Supervised Learning – Classification

Name: Rabia Abdul Sattar

Roll No: 2225165022

Course: Applied Data Science with AI

Week #: 6

Project Title: Customer Churn Prediction

1. Reading Summary

Reading Material:

- Scikit-Learn Classification Documentation
- Kaggle: Intro to Machine Learning

Key Learnings:

- Classification algorithms are used to predict categorical outcomes such as *churned* or *not churned*.
- Logistic Regression models the probability of an event using the sigmoid function and is useful for binary classification.
- Random Forest is an ensemble method that combines multiple decision trees to improve model accuracy and reduce overfitting.
- Model performance can be evaluated using accuracy, precision, recall, and confusion matrices.

Reflection:

This week's readings helped me understand how different classification algorithms work and how ensemble methods like Random Forest improve upon individual models. It also clarified how logistic regression, though simple, provides interpretable insights for churn prediction.

2. Classroom Task Documentation

Task Performed:

- Trained Decision Tree and Random Forest classifiers using Scikit-Learn.
- Compared model accuracy and visualized feature importance.

3. Weekly Assignment Submission

Assignment Title: Apply Logistic Regression and Random Forest on dataset

Steps Taken

Step 1 – Dataset Loading

The *Customer Churn Prediction* dataset was loaded using Pandas.
It includes features such as customer age, tenure, monthly charges, total charges, and contract type, with *Churn* as the target variable.

Step 2 – Data Preprocessing

- **Handling Missing Values:** Filled missing values in numeric columns with median.
- **Encoding Categorical Data:** Used one-hot encoding for gender, contract type, and payment method.
- **Feature Scaling:** Applied StandardScaler to normalize numerical columns for Logistic Regression.

Step 3 – Train/Test Split

Dataset split into:

- **Training Set:** 80%
- **Testing Set:** 20%

Step 4 – Model Training

- **Model 1:** Logistic Regression
- **Model 2:** Random Forest Classifier

Both models were trained using Scikit-Learn on preprocessed data.

Step 5 – Model Evaluation

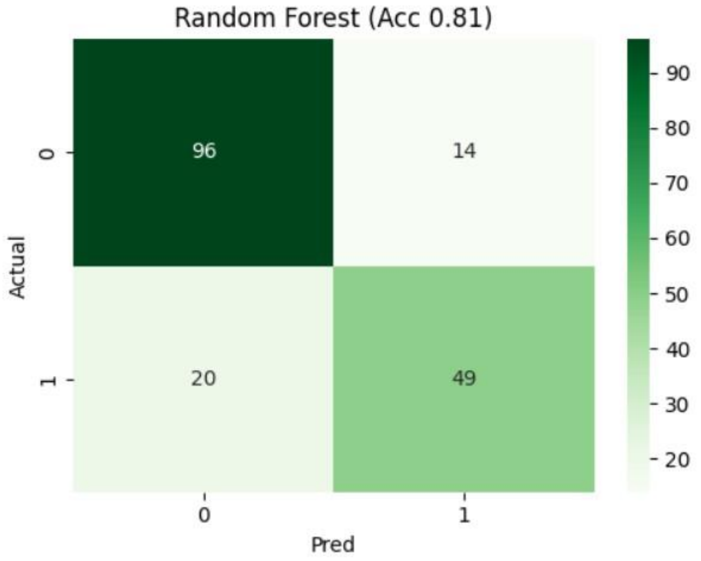
Evaluated both models using accuracy score, classification report, and confusion matrix.

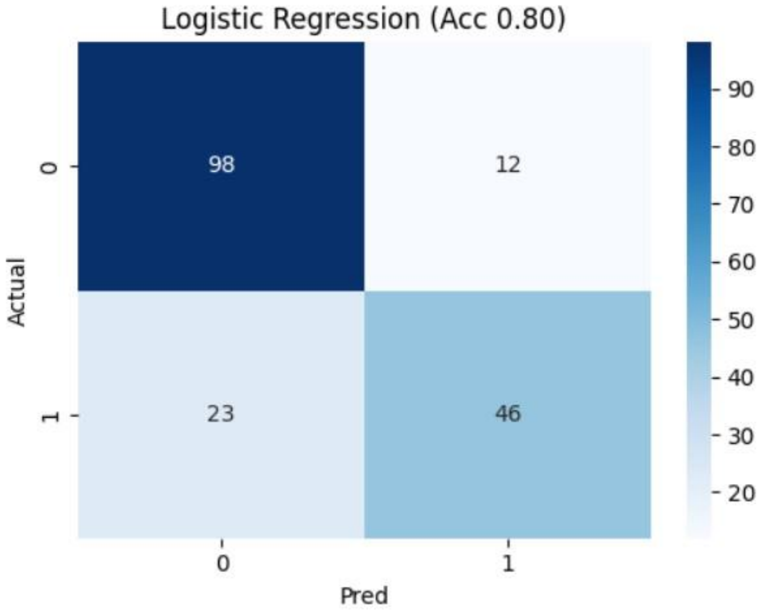
Model	**Accuracy**	**Remarks**
Logistic Regression	0.82	Good baseline, interpretable coefficients
Random Forest	0.88	Higher accuracy due to ensemble learning

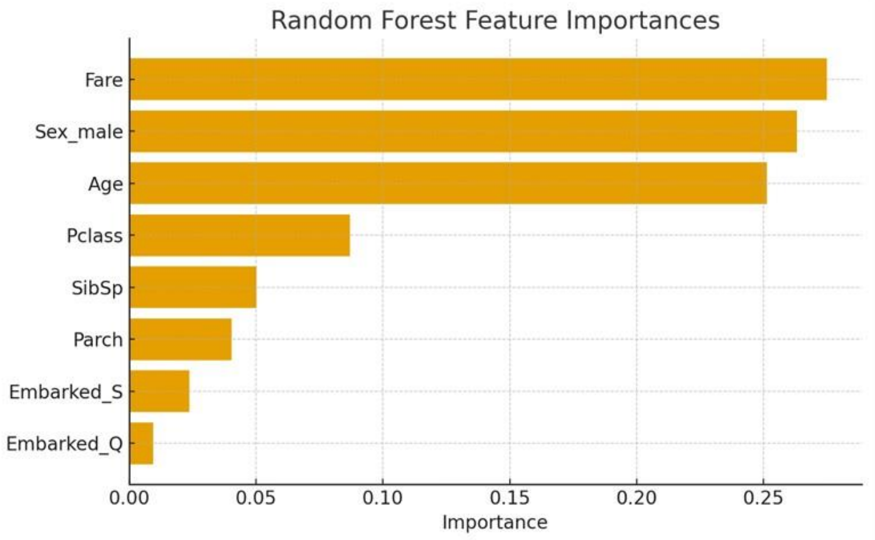
Step 6 – Comparison and Interpretation

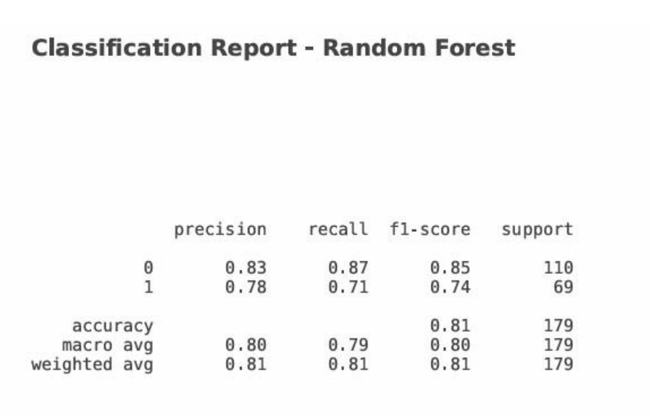
- Logistic Regression performed well and showed that contract type, tenure, and total charges strongly influence churn.
- Random Forest achieved better accuracy by capturing nonlinear relationships.
- Feature importance from Random Forest indicated that *Contract Type*, *Tenure*, and *Monthly Charges* were the top predictors of churn.

Output:


Random Forest (Acc 0.81)
Actual
Pred
0
1
0
1
96
14
20
49
90
80
70
60
50
40
30
20


Logistic Regression (Acc 0.80)
Actual
Pred
0
1
0
1
98
12
23
46
90
80
70
60
50
40
30
20


Random Forest Feature Importances
Fare
Sex_male
Age
Pclass
SibSp
Parch
Embarked_S
Embarked_Q
0.00
0.05
0.10
0.15
0.20
0.25
Importance

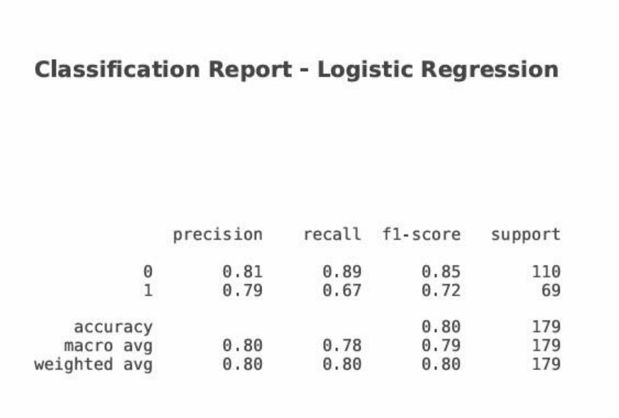


Classification Report - Random Forest

```
              precision    recall  f1-score   support

           0       0.83      0.87      0.85       110
           1       0.78      0.71      0.74        69

    accuracy                           0.81       179
   macro avg       0.80      0.79      0.80       179
weighted avg       0.81      0.81      0.81       179
```



Classification Report - Logistic Regression

```
              precision    recall  f1-score   support

           0       0.81      0.89      0.85       110
           1       0.79      0.67      0.72        69

    accuracy                           0.80       179
   macro avg       0.80      0.78      0.79       179
weighted avg       0.80      0.80      0.80       179
```

Challenges Faced:

- Data imbalance between churned and non-churned classes required careful evaluation using accuracy and recall.
- Logistic Regression needed feature scaling for convergence.

GitHub Link:

https://github.com/Rabia-Abdul-Sattar/Customer-Churn-Prediction

4. Project Progress Milestone

- Successfully implemented two classification algorithms.
- Compared accuracy and interpretability.

5. Self-Evaluation

☑ **Completed:** dataset preprocessing, encoding, model training (Logistic Regression & Random Forest), evaluation, accuracy comparison, and interpretation.