Assignment: Applying Logistic Regression to a Dataset

Link: https://www.datacamp.com/tutorial/understanding-logistic-regression-python

Objective:

In this assignment, you will select a dataset, perform necessary data preprocessing, apply logistic regression to build a classification model, and evaluate its performance. The goal is to understand how logistic regression works in practice and gain hands-on experience with implementing it.

Instructions:

1. **Dataset Selection**:
    - Choose a dataset suitable for binary classification (e.g., predicting whether a customer will make a purchase, whether a patient has a disease, or whether an email is spam or not).
    - Possible sources include:
        - UCI Machine Learning Repository
        - Kaggle Datasets
        - Built-in datasets in libraries like `sklearn.datasets`
2. **Data Preparation**:
    - Load the dataset and explore its structure.
    - Perform any necessary data cleaning:
        - Handle missing values.
        - Convert categorical variables to numerical representations if needed (using encoding techniques like one-hot encoding or label encoding).
    - Standardize or normalize the data if necessary.
3. **Exploratory Data Analysis (EDA)**:
    - Visualize the data to understand the distribution of classes and key features.
    - Check for class imbalance (e.g., if there are significantly more examples in one class than the other).
    - Create plots to investigate relationships between features and the target variable (e.g., histograms, box plots, correlation heatmaps).
4. **Applying Logistic Regression**:
    - Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
    - Implement logistic regression using a library like `scikit-learn`.
    - Train the model on the training set and obtain predictions on the testing set.
5. **Evaluation**:
    - Evaluate the model using appropriate metrics:
        - Accuracy
        - Precision
        - Recall
        - F1-score
        - Confusion matrix
    - If the dataset is imbalanced, focus on precision, recall, and F1-score rather than accuracy.

    - Plot the ROC curve and calculate the AUC score to assess the model's performance.
6. **Analysis & Interpretation**:
    - Interpret the model’s coefficients to understand which features contribute most to the predictions.
    - Reflect on the model's performance based on the evaluation metrics.
    - Discuss any potential limitations of your model and possible ways to improve it (e.g., feature engineering, handling class imbalance).

Bonus (Optional):

- Try adjusting hyperparameters (e.g., regularization strength) to improve the model.
- Experiment with feature selection techniques to identify which features are most important for your logistic regression model.