Problem Set 3 Part 2: Logistic Regression

In this problem set we analyze the relationship between online ads and purchase behavior. In particular, we want to classify which online users are likely to purchase a certain product after being exposed to an online ad. Use the dataset "**purchase.csv**"

Tasks:

1. Explore the data in detail and describe your findings (use exploratory data analysis).

2. Logistic regression model:

    a. Fit a logistic regression model with all predictor variables (response: **Purchased**; predictors: **Gender**, **Age**, **EstimatedSalary**).
    b. Please explain if you would recommend to exclude a predictor variable from your model (from task 2a). Update your model if necessary.
    c. Use your updated model and predict the probability that an online user will purchase a product. Classify an online user as purchaser (with label "**Yes**") if the predicted probability of the purchase exceeds:

        c1): 0.4 (i.e. threshold = 0.4)
        c2): 0.5 (i.e. threshold = 0.5)
        c3): 0.7 (i.e. threshold = 0.7).

    Otherwise classify the user as non-purchaser (with label "**No**").

    d. Compute the **confusion matrix** for every threshold (c1), c2) and c3)) in order to determine how many observations were correctly or incorrectly classified. Furthermore, use the results from the confusion matrix and create the following variables:

        true positive;
        true negative;
        false positive and
        false negative.

    Use these variables to calculate the following measures:

    Accuracy,
    Precision (what proportion of positive identifications was actually correct?),
    Recall (what proportion of actual positives was identified correctly) and the
    F1 score (measure of a test's accuracy)

    for the thresholds in c1), c2) and c3). Which threshold would you recommend?

    Precision is defined as the number of true positives over the number of true positives plus the number of false positives.
    Recall is defined as the number of true positives over the number of true positives plus the number of false negatives.
    These two quantities are related to the **F1** score, which is defined as the harmonic mean of precision and recall: F1 = 2* ((Precision * Recall) / (Precision + Recall)).

    e. Fit the logistic regression model using a **training** data set. Compute the confusion matrix and accuracy for the held out data (**test data**). Use a threshold of 0.5.