Ερώτημα 1:

1. Number of records in the dataset: 12330
2. Percentage of users who made a purchase: 15.474452554744525 %
3. Accuracy of a model always predicting users won't purchase: 84.52554744525548

Ερώτημα 5:

1. Ευστοχία στο σύνολο εκπαίδευσης: 0.8767234387672344
2. Ευστοχία στο σύνολο δοκιμής: 0.8745606920789403

- Πώς ερμηνεύετε τον πίνακα σύγχυσης

3083	**41**
423	**152**

Ο πίνακας σύγχυσης που δόθηκε παραπάνω μπορεί να ερμηνευτεί ως εξής:

- **3083 True Negatives (TN)**:
  - Αυτά είναι τα δείγματα που ανήκουν στην αρνητική κλάση και προβλέφθηκαν σωστά ως αρνητικά από το μοντέλο.
- **41 False Positives (FP)**:
  - Αυτά είναι τα δείγματα που ανήκουν στην αρνητική κλάση αλλά προβλέφθηκαν λανθασμένα ως θετικά από το μοντέλο. Δηλαδή, προβλέφθηκαν ότι θα αγοραστούν, αλλά στην πραγματικότητα δεν αγοράστηκαν.
- **423 False Negatives (FN)**:
  - Αυτά είναι τα δείγματα που ανήκουν στη θετική κλάση αλλά προβλέφθηκαν λανθασμένα ως αρνητικά από το μοντέλο. Δηλαδή, προβλέφθηκαν ότι δεν θα αγοραστούν, αλλά στην πραγματικότητα αγοράστηκαν.
- **152 True Positives (TP)**:
  - Αυτά είναι τα δείγματα που ανήκουν στη θετική κλάση και προβλέφθηκαν σωστά ως θετικά από το μοντέλο. Δηλαδή, προβλέφθηκαν ότι θα αγοραστούν και πραγματικά αγοράστηκαν.

Βάσει αυτών των αριθμών, μπορούμε να εκτιμήσουμε την απόδοση του μοντέλου, τις πιθανές αδυναμίες του και τα σημεία που μπορούν να βελτιωθούν.

Τι τροποποιήσεις ή επιπλέον πειράματα θα υλοποιούσατε ώστε να βελτιώσετε το μοντέλο σας;

Για να βελτιώσουμε το μοντέλο μας και να μειώσουμε τον αριθμό των False Positives, μπορούμε να εφαρμόσουμε τις παρακάτω τεχνικές

- **Χρήση Υπερδειγματοληψίας (Oversampling)**:
    - Η υπερδειγματοληψία αφορά την αύξηση του αριθμού των δειγμάτων στη λιγότερο αντιπροσωπευόμενη κλάση του συνόλου δεδομένων, είτε με την αντιγραφή υπαρχόντων δειγμάτων είτε με τη δημιουργία νέων δειγμάτων μέσω άλλων τεχνικών
    - Αυτό βοηθάει να αποφευχθεί το πρόβλημα της ανισορροπίας κλάσεων και μπορεί να βελτιώσει την ικανότητα γενίκευσης του μοντέλου, καθώς προσφέρει περισσότερα δείγματα για την εκπαίδευση.
- **Κανονικοποίηση ή Αφαίρεση Χαρακτηριστικών (Feature Normalization or Feature Selection)**:
    - Η κανονικοποίηση των χαρακτηριστικών στην ίδια κλίμακα μπορεί να βοηθήσει το μοντέλο να συγκλίνει πιο γρήγορα και να βελτιώσει την απόδοσή του.
    - Η αφαίρεση λιγότερο σημαντικών χαρακτηριστικών μπορεί να μειώσει τον θόρυβο στα δεδομένα και να βελτιώσει την ικανότητα γενίκευσης του μοντέλου.
- **Χρήση Υποδειγματοληψίας (Undersampling):**
    - Η διαδικασία υποδειγματοληψίας περιλαμβάνει την απόρριψη τμημάτων της υπερπροσπαθούσας κλάσης ή την τυχαία επιλογή ενός υποσυνόλου από την υπερπροσπαθούσα κλάση, ώστε να είναι παρόμοιος με τον αριθμό δειγμάτων της λιγότερο αντιπροσωπευόμενης κλάσης.

Έγινε προσπάθεια και με τις 3 τεχνικές που αναφερθηκαν χωρίς ωστόσο κάποια βελτίωση παρά μόνο χειροτέρευση

```
1 usage
def feature_normalization():
    df = pd.read_csv("project2_dataset.csv")
    X_train, X_test, y_train, y_test = linear_transformation(df)

    # Apply undersampling to the training set
    undersampler = RandomUnderSampler(random_state=42)
    X_train_resampled, y_train_resampled = undersampler.fit_resample(X_train, y_train)

    # Feature selection
    model = LogisticRegression(max_iter=1000)
    model.fit(X_train_resampled, y_train_resampled)
    coefficients = model.coef_[0]
    important_features_indices = [i for i, coef in enumerate(coefficients) if abs(coef) > 0.1]
    X_train_resampled_selected = X_train_resampled[:, important_features_indices]
    X_test_selected = X_test[:, important_features_indices]

    model = CustomLogisticRegression()
    model.fit(X_train_resampled_selected, y_train_resampled)

    y_train_pred = model.predict(X_train_resampled_selected)
    y_test_pred = model.predict(X_test_selected)

    # Model evaluation
    train_accuracy = accuracy_score(y_train_resampled, y_train_pred)
    test_accuracy = accuracy_score(y_test, y_test_pred)

    print("Ευστοχία στο σύνολο εκπαίδευσης:", train_accuracy)
    print("Ευστοχία στο σύνολο δοκιμής:", test_accuracy)

    confusion_mat = confusion_matrix(y_test, y_test_pred)
    print("Πίνακας Σύγχυσης:")
    print(confusion_mat)
```

```
Ευστοχία στο σύνολο εκπαίδευσης: 0.7764441110277569
Ευστοχία στο σύνολο δοκιμής: 0.7829143011624764
Πίνακας Σύγχυσης:
[[2461  663]
 [ 140  435]]
```

```python
def oversampling():
    df = pd.read_csv("project2_dataset.csv")
    X_train, X_test, y_train, y_test = linear_transformation(df)

    # Apply oversampling to the training set
    oversampler = RandomOverSampler(random_state=42)
    X_train_resampled, y_train_resampled = oversampler.fit_resample(X_train, y_train)

    model = CustomLogisticRegression()
    model.fit(X_train_resampled, y_train_resampled)

    y_train_pred = model.predict(X_train_resampled)
    y_test_pred = model.predict(X_test)

    # Model evaluation
    train_accuracy = accuracy_score(y_train_resampled, y_train_pred)
    test_accuracy = accuracy_score(y_test, y_test_pred)

    print("Ευστοχία στο σύνολο εκπαίδευσης:", train_accuracy)
    print("Ευστοχία στο σύνολο δοκιμής:", test_accuracy)

    confusion_mat = confusion_matrix(y_test, y_test_pred)
    print("Πίνακας Σύγχυσης:")
    print(confusion_mat)
```

```
Ευστοχία στο σύνολο εκπαίδευσης: 0.8050835845437107
Ευστοχία στο σύνολο δοκιμής: 0.8534739118680724
Πίνακας Σύγχυσης:
[[2734  390]
 [ 152  423]]
```

```python
def undersampling():
    df = pd.read_csv("project2_dataset.csv")
    X_train, X_test, y_train, y_test = linear_transformation(df)

    # Apply undersampling to the training set
    undersampler = RandomUnderSampler(random_state=42)
    X_train_resampled, y_train_resampled = undersampler.fit_resample(X_train, y_train)

    model = CustomLogisticRegression()
    model.fit(X_train_resampled, y_train_resampled)

    y_train_pred = model.predict(X_train_resampled)
    y_test_pred = model.predict(X_test)

    # Model evaluation
    train_accuracy = accuracy_score(y_train_resampled, y_train_pred)
    test_accuracy = accuracy_score(y_test, y_test_pred)

    print("Ευστοχία στο σύνολο εκπαίδευσης:", train_accuracy)
    print("Ευστοχία στο σύνολο δοκιμής:", test_accuracy)

    confusion_mat = confusion_matrix(y_test, y_test_pred)
    print("Πίνακας Σύγχυσης:")
    print(confusion_mat)
```

```
Ευστοχία στο σύνολο εκπαίδευσης: 0.7771942985746436
Ευστοχία στο σύνολο δοκιμής: 0.7853473911868073
Πίνακας Σύγχυσης:
[[2464  660]
 [ 134  441]]
```