DATA SCIENCE INSTITUTE

\WOOLF/

ASSIGNMENT

MACHINE LEARNING 1

Python

REPORT

Pillar1-e

Prepared By

William C. Phiri

PG Dip in Data Science

Student No. 4295158639

+353-87-3502102

chenechoz@gmail.com

Athlone, IRELAND

BACKGROUND:

The data is a marketing campaign data of a skin care clinic associated with its success.

Description of variables-

Success: Response to marketing campaign of Skin Care Clinic which offers both products and services. (1: email Opened, 0: email not opened)

AGE: Age Group of Customer

Recency_Service: Number of days since last service purchase

Recency_Product: Number of days since last product purchase

Bill_Service: Total bill amount for service in last 3 months

Bill_Product: Total bill amount for products in last 3 months

Gender (1: Male, 2: Female)

Note: Answer following questions using entire data and do not create test data.

QUESTIONS

1. Import Email Campaign data. Perform binary logistic regression to model "Success". Interpret sign of each significant variable in the model.
2. Compare performance of Binary Logistic Regression (significant variables) and Naïve Bayes Method (all variables) using area under the ROC curve.
3. Implement binary logistic regression and Support Vector Machines by combining service and product variables.

ASSIGNMENT SUMMARY OVERVIEW:

This project implements a comprehensive machine learning pipeline to predict email campaign success for a skin care clinic marketing campaign. The analysis utilizes proper train/test splitting, extensive visualizations, and compares four different classification algorithms to identify which customers are most likely to open marketing emails.

Key Features:

- Train/Test Split (80/20) for robust model evaluation
- Statistical Significance Testing using statsmodels
- Multiple Classification Algorithms (Logistic Regression, Naive Bayes, SVM)
- Comprehensive Visualizations (EDA, ROC Curves, Confusion Matrices)
- ROC-AUC Analysis for model comparison
- Production-Ready Models with persistence and metadata tracking

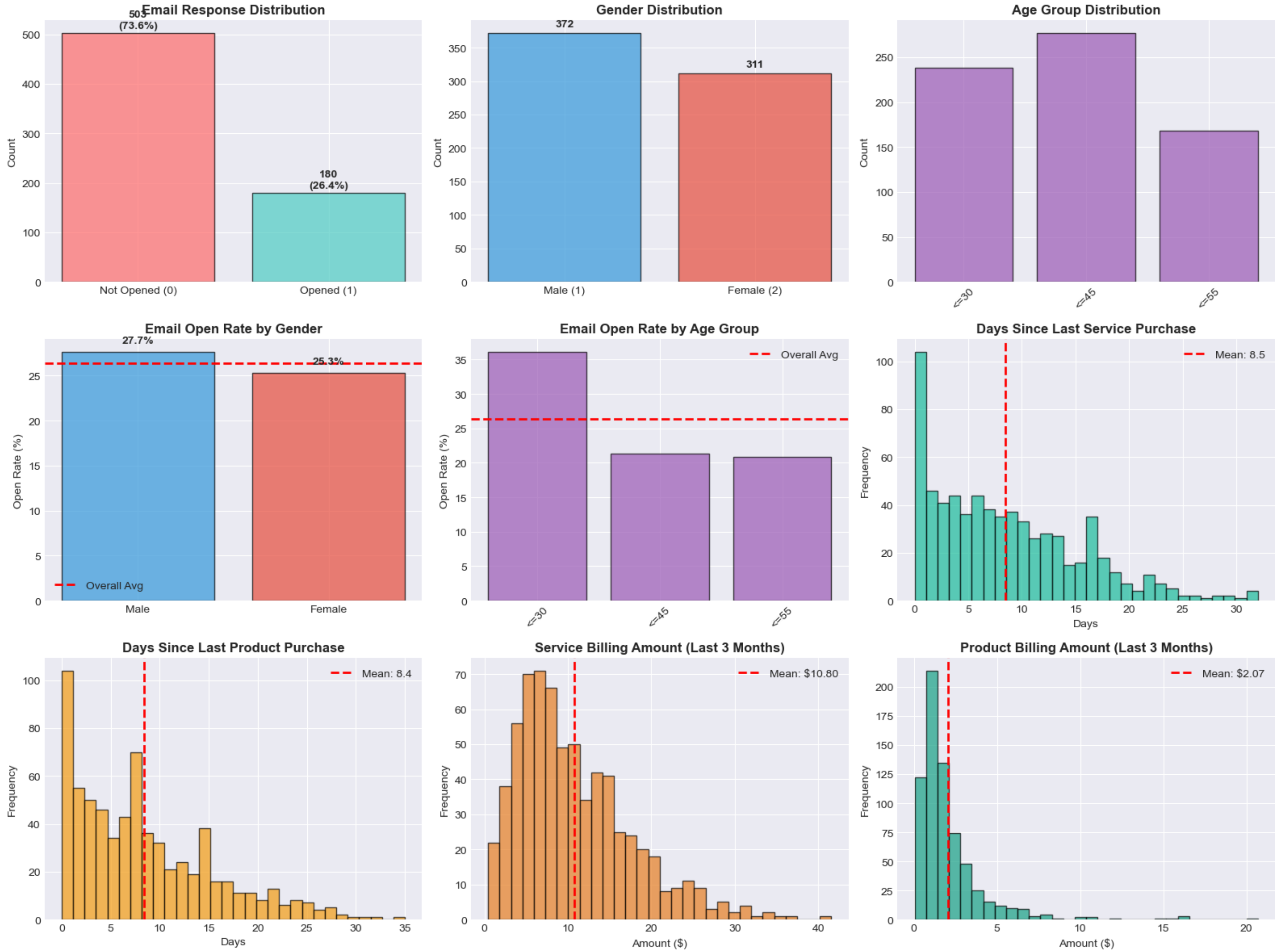
Dataset: 683 customer records with demographics, purchase recency, billing history, and email response data.

PROJECT STRUCTURE:

```
email-campaign-prediction/
├── data/
│   ├── raw/
│   │   └── Email Campaign.csv
│   ├── processed/
│   │   └── email_campaign_processed.csv
│   └── predictions/
├── models/
│   ├── logistic_regression_significant.pkl
│   ├── naive_bayes_model.pkl
│   ├── logistic_regression_combined.pkl
│   ├── svm_model.pkl
│   ├── age_label_encoder.pkl
│   ├── significant_variables.pkl
│   └── model_metadata.json
├── reports/
│   ├── figures/
│   │   ├── 01_exploratory_data_analysis.png
│   │   ├── 02_correlation_matrix.png
│   │   ├── 03_train_test_split.png
│   │   ├── 04_logistic_regression_coefficients.png
│   │   ├── 05_comprehensive_model_comparison.png
│   │   └── 06_confusion_matrices.png
│   └── model_performance_summary.csv
├── notebooks/
│   └── 01_complete_analysis.ipynb
├── environment/
│   ├── requirements.txt
│   └── environment.yml
├── main.py
├── load_and_predict.py
├── .gitignore
└── README.md
```

QUESTION 1: Import Email Campaign data. Perform binary logistic regression to model “Success”. Interpret sign of each significant variable in the model.

<u>Exploratory Data Analysis & Visualizations</u>



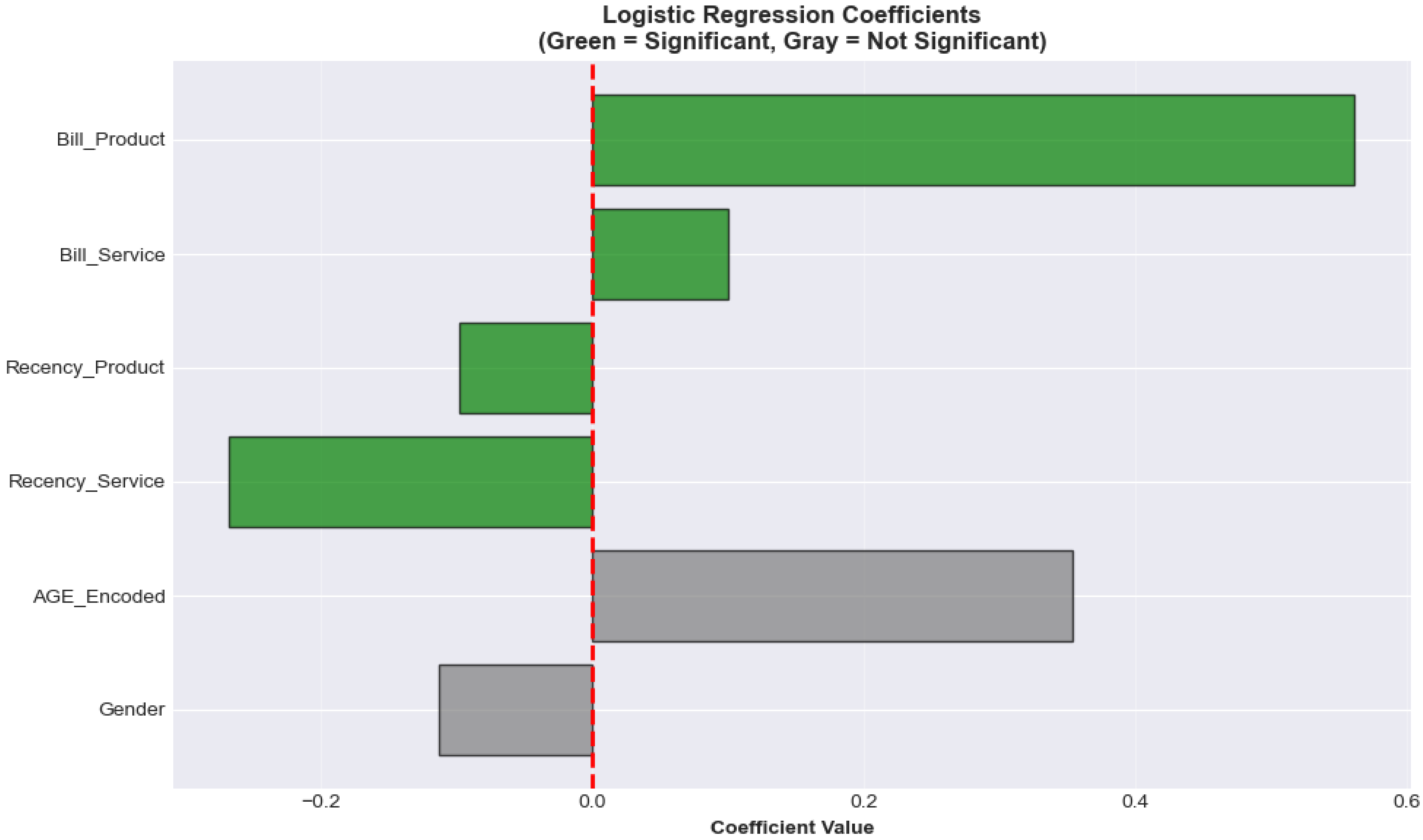
Binary Logistic Regression:

```
================================================================================
📈 QUESTION 1: Binary Logistic Regression with Statistical Testing
================================================================================
Optimization terminated successfully.
         Current function value: 0.384373
         Iterations 7

                           Logit Regression Results
==============================================================================
Dep. Variable:                Success   No. Observations:                  546
Model:                          Logit   Df Residuals:                      539
Method:                           MLE   Df Model:                            6
Date:                Sat, 04 Oct 2025   Pseudo R-squ.:                  0.3338
Time:                        12:50:38   Log-Likelihood:                -209.87
converged:                       True   LL-Null:                       -315.00
Covariance Type:            nonrobust   LLR p-value:                 1.233e-42
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
const              -1.1513      0.481     -2.393      0.017      -2.094      -0.208
Gender             -0.1127      0.245     -0.460      0.646      -0.593       0.368
AGE_Encoded         0.3538      0.206      1.720      0.086      -0.049       0.757
Recency_Service    -0.2671      0.035     -7.654      0.000      -0.336      -0.199
Recency_Product    -0.0980      0.026     -3.836      0.000      -0.148      -0.048
Bill_Service        0.1003      0.021      4.804      0.000       0.059       0.141
Bill_Product        0.5617      0.090      6.223      0.000       0.385       0.739
===================================================================================
```

```
================================================================================
COEFFICIENT INTERPRETATION
================================================================================
       Variable  Coefficient  Std Error      P-value  Significant
          const    -1.151264   0.481185 1.673126e-02         True
         Gender    -0.112690   0.245107 6.456908e-01        False
    AGE_Encoded     0.353776   0.205732 8.550599e-02        False
Recency_Service    -0.267145   0.034901 1.942249e-14         True
Recency_Product    -0.098002   0.025551 1.252809e-04         True
   Bill_Service     0.100338   0.020886 1.553778e-06         True
   Bill_Product     0.561742   0.090267 4.873945e-10         True
```

The significant variables (***$p < 0.05$***) include the following below:

- Recency_Service,
- Recency_Product,
- Bill_Service,
- Bill_Product


Logistic Regression Coefficients
(Green = Significant, Gray = Not Significant)
Bill_Product
Bill_Service
Recency_Product
Recency_Service
AGE_Encoded
Gender
−0.2
0.0
0.2
0.4
0.6
Coefficient Value

Model Performance:

```
================================================================
MODEL PERFORMANCE - LOGISTIC REGRESSION (Significant Variables)
================================================================

📊 TRAINING SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.86      0.93      0.89       402
      Opened       0.74      0.56      0.64       144

    accuracy                           0.83       546
   macro avg       0.80      0.75      0.77       546
weighted avg       0.83      0.83      0.83       546


📊 TEST SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.78      0.93      0.85       101
      Opened       0.59      0.28      0.38        36

    accuracy                           0.76       137
   macro avg       0.69      0.60      0.61       137
weighted avg       0.73      0.76      0.73       137
```

When we look at this models performance analysis we can conclude the following, the Logistic Regression model achieves **76%** overall accuracy on the test set but shows imbalanced performance across classes. While it excels at identifying customers who won't open emails (**93%** recall, **78%** precision), it struggles with the target class of email openers, achieving only 28% recall and 59% precision. This means the model is highly conservative, correctly identifying fewer than 1 in 3 actual email openers while minimizing false positives. The **7%** drop in accuracy from training (**83%**) to test (**76%**) suggests mild overfitting. For business deployment, the model is currently best suited for exclusion campaigns (identifying who NOT to target) rather than positive targeting (identifying who WILL engage).

QUESTION 2: Compare performance of Binary Logistic Regression (significant variables) and Naïve Bayes Method (all variables) using area under the ROC curve.

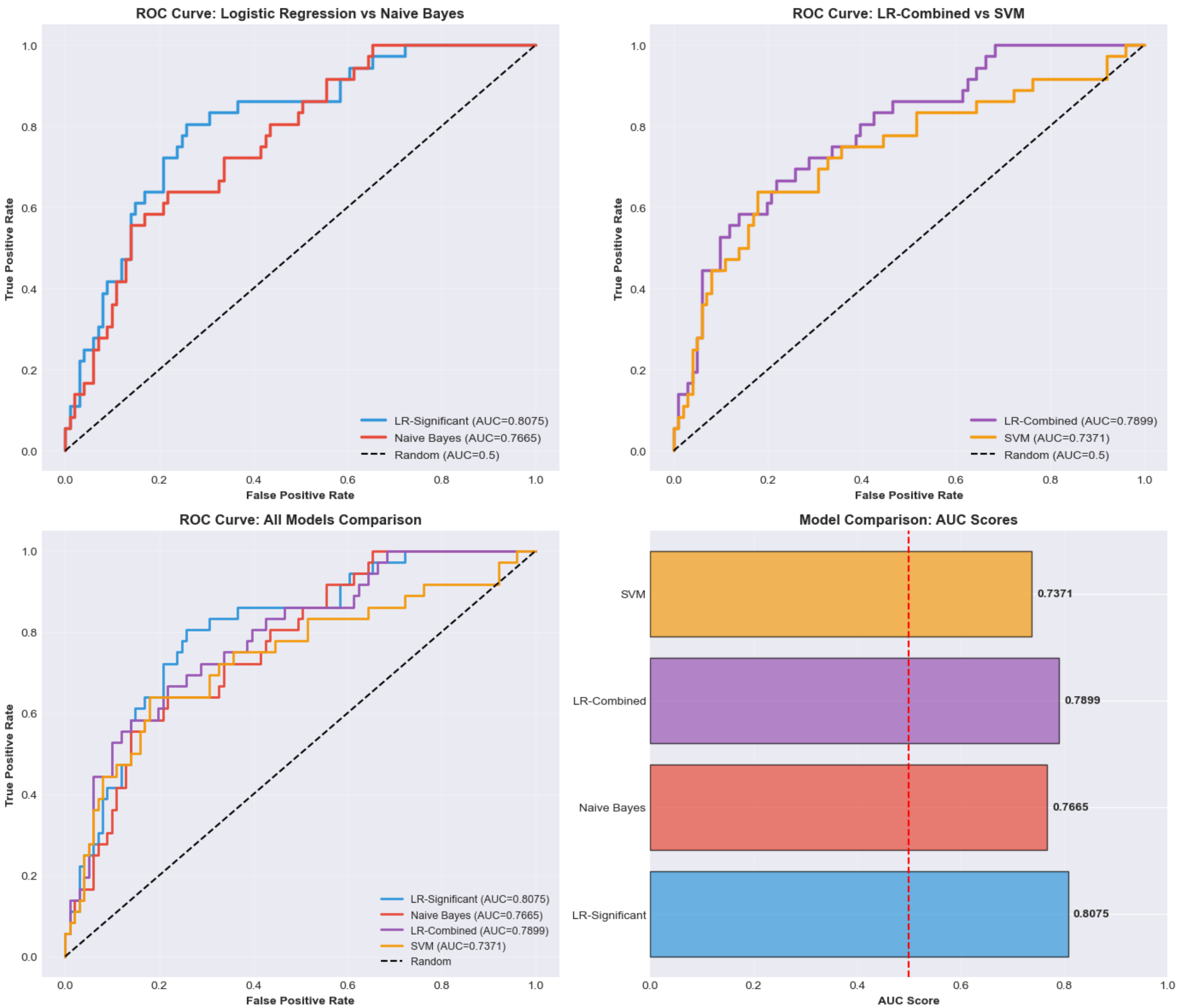
```
============================================================
📈 QUESTION 2: Naive Bayes Classification
============================================================

📊 TRAINING SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.80      0.92      0.86       402
      Opened       0.63      0.36      0.46       144

    accuracy                           0.77       546
   macro avg       0.71      0.64      0.66       546
weighted avg       0.76      0.77      0.75       546


📊 TEST SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.78      0.92      0.85       101
      Opened       0.56      0.28      0.37        36

    accuracy                           0.75       137
   macro avg       0.67      0.60      0.61       137
weighted avg       0.72      0.75      0.72       137
```

QUESTION 3: Implement binary logistic regression and Support Vector Machines by combining service and product variables.

```
================================================================================
📈 QUESTION 3: Support Vector Machine with Combined Features
================================================================================

📊 Creating Combined Features...
   ✅ Total_Recency = Recency_Service + Recency_Product
   ✅ Total_Bill = Bill_Service + Bill_Product
   ✅ Recency_Ratio = Recency_Service / (Recency_Product + 1)
   ✅ Bill_Ratio = Bill_Service / (Bill_Product + 1)

🔄 Training Logistic Regression (Combined Features)...

📊 Logistic Regression (Combined) - TEST SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.79      0.94      0.86       101
      Opened       0.65      0.31      0.42        36

    accuracy                           0.77       137
   macro avg       0.72      0.62      0.64       137
weighted avg       0.75      0.77      0.74       137


🔄 Training Support Vector Machine (RBF Kernel)...
```

```
📊 SVM - TEST SET PERFORMANCE:
              precision    recall  f1-score   support

  Not Opened       0.79      0.95      0.86       101
      Opened       0.67      0.28      0.39        36

    accuracy                           0.77       137
   macro avg       0.73      0.61      0.63       137
weighted avg       0.76      0.77      0.74       137
```

COMPREHENSIVE MODEL COMPARISON WITH ROC CURVES



From the ROC curves and AUC scores we can deduce the following outcome across the models.

Model	Features	AUC	Accuracy	Precision	Recall	F1-Score	Rank
Logistic Regression (Significant)	***4***	***0.81***	***0.76***	***0.59***	***0.28***	***0.38***	***1***
Logistic Regression (Combined)	6	0.79	0.77	0.65	0.31	0.42	2
Naive Bayes (All Variables)	6	0.77	0.75	0.56	0.28	0.37	3
SVM (RBF Kernel)	6	0.74	0.77	0.67	0.28	0.39	4

The best performing model is clearly the **Logistic Regression** retaining the significant variables seeing as it has the highest AUC score.

```
================================================================================
MODEL PERFORMANCE SUMMARY TABLE
================================================================================

                            Model  Features       AUC  Accuracy  Precision   Recall  F1-Score  Rank
Logistic Regression (Significant)         4  0.807481  0.759124   0.588235 0.277778  0.377358     1
   Logistic Regression (Combined)         6  0.789879  0.773723   0.647059 0.305556  0.415094     2
      Naive Bayes (All Variables)         6  0.766502  0.751825   0.555556 0.277778  0.370370     3
                 SVM (RBF Kernel)         6  0.737074  0.773723   0.666667 0.277778  0.392157     4

================================================================================
🏆 BEST PERFORMING MODEL
================================================================================
   Model: Logistic Regression (Significant)
   AUC Score: 0.8075
   Accuracy: 0.7591
   Precision: 0.5882
   Recall: 0.2778
   F1-Score: 0.3774
```