REPORT

Himanshu, MDS202327

Task 1

Steps followed:

1. Dropped the CustomerID column
2. Encoded the three qualitative features i.e. Gender, PromotionResponse, EmailOptIn using LabelEncoder() and OneHotEncoder()
3. Splitted the encoded data into train and test data (70-30)
4. Fitted the AdaBoostClassifier() using GridSearchCV with 5-fold cross-validation on both type of encoded data having hyper-parameters as
    a. estimator - A weak classifier
        i. DecisionTreeClassifier(max_depth=1)
        ii. RandomForestClassifier(max_depth=1)
        iii. GaussianNB()
    b. learning_rate - 0.5, 1, 1.5
    c. algorithm - SAMME, SAMME.R
5. Fitted the RandomForestClassifier() using GridSearchCV with 5-fold cross-validation on both type of encoded data having hyper-parameters as
    a. n_estimators – 50, 100, 200, 500
    b. criterion – gini, entropy, log_loss
    c. max_features – sqrt, log2, None
6. Computed the Confusion Matrix, Overall Accuracy, Precision and Recall

Results:

Model	Encoding	Overall Accuracy	Precision	Recall
AdaBoost	Label	0.543	0.573	0.646
AdaBoost	One Hot	0.510	0.556	0.512
RandomForest	Label	0.527	0.573	0.524
RandomForest	One Hot	0.527	0.568	0.561

Conclusion:

AdaBoost fitted on label encoded data gave the best results across the categories

*Hyper parameter range was chosen short because of computation constraint!

Task 2

Sub Task A: Gender

Steps followed:

1. Dropped the InvoiceID column
2. Encoded the three qualitative features i.e. CustomerType, ProductType, PaymentType and Branch using OneHotEncoder()
3. Splitted the encoded data into train and test data (70-30)
4. Fitted the DecisionTreeClassifier() using GridSearchCV with 5-fold cross-validation on both type of encoded data having hyper-parameters as
    a. criterion – gini, entropy, log_loss
    b. max_depth – 3,4,5,6,7,8,9,10
5. Fitted the RandomForestClassifier() using GridSearchCV with 5-fold cross-validation on both type of encoded data having hyper-parameters as
    a. n_estimators – 50, 100, 200, 500
    b. criterion – gini, entropy, log_loss
    c. max_features – sqrt, log2, None
6. Computed the Confusion Matrix, Overall Accuracy, Precision and Recall

Results:

Model	Parameter	Overall Accuracy	Precision	Recall
DecisionTree	max_depth=10	0.530	0.511	0.648
RandomForest	n_estimators=200	0.513	0.497	0.572

Conclusion:

Decision Tree fitted on label encoded data gave the best results across the categories

Sub Task B: Rating

Steps followed:

1. Dropped the InvoiceID column
2. Encoded the three qualitative features i.e. CustomerType, ProductType, PaymentType and Branch using OneHotEncoder()
3. Splitted the encoded data into train and test data (70-30)
4. Fitted the DecisionTreeRegressor() using GridSearchCV with 5-fold cross-validation on both type of encoded data having hyper-parameters as
    a. criterion – squared_error, absolute_error, poisson
    b. max_depth – 1,2,3,4,5,6,7,8,9,10
5. Fitted the LinearRegression()
6. Computed the Mean Absolute Error

Results:

Model	Parameter	MAE
DecisionTree	max_depth=1	1.516
LinearRegression	default	1.526

Conclusion:

Decision stump fitted on label encoded data gave the best result.

Note:

1. I chose absolute error as it felt more natural to compare the ratings.
2. A decision stump getting lower MAE shows a simple 1-question model works better