DATA 3461- Mini Project 2, Findings Report

By Tajwar Fahmid

Question 1: Walmart Sales

a) For preprocessing, I converted the ‘Store’ category to string and dropped the date column. Afterwards, I defined the features and split the data into a training and test set.

b) Multiple Linear Regression model:

```
--- MLR Model Interpretation ---
Intercept: $1,211,014.36

Coefficients (in descending order):
Store_4          7.868619e+05
Store_13         7.138299e+05
Store_10         6.530089e+05
Store_20         5.729460e+05
Store_14         5.634710e+05
Store_27         4.982432e+05
Store_2          3.681852e+05
Store_28         1.834467e+05
Store_19         1.568762e+05
Store_24         7.399951e+04
Holiday_Flag     7.276384e+04
Store_23         3.152837e+04
CPI              3.220848e+03
Temperature     -7.851585e+02
Store_6         -1.276754e+04
Unemployment    -2.116503e+04
Fuel_Price      -4.275984e+04
Store_39        -1.024783e+05
Store_12        -1.405201e+05
Store_31        -1.665303e+05
Store_18        -1.827426e+05
Store_11        -2.108171e+05
Store_41        -2.434993e+05
Store_34        -2.731352e+05
Store_22        -2.796512e+05
Store_26        -2.967311e+05
Store_32        -3.189643e+05
Store_35        -3.816067e+05
Store_40        -4.013443e+05
Store_17        -4.199754e+05
Store_45        -6.562988e+05
Store_15        -6.595486e+05
Store_42        -6.869191e+05
Store_8         -7.017592e+05
Store_29        -7.173368e+05
```

```
Store_38      -7.590419e+05
Store_21      -7.977728e+05
Store_25      -8.358209e+05
Store_43      -8.487452e+05
Store_7       -9.176632e+05
Store_33      -9.782133e+05
Store_44      -1.000571e+06
Store_16      -1.008215e+06
Store_37      -1.026154e+06
Store_9       -1.059928e+06
Store_30      -1.120048e+06
Store_36      -1.172546e+06
Store_3       -1.172685e+06
Store_5       -1.268905e+06
```

This was the output that we got from the multiple linear regression model, its includes all the coefficients of all the variables. From our output, we can find a lot of insights:-

- Store 4 is predicted to have approximately $786,862 higher weekly sales than the reference Store 1, when all other factors are equal.
- Store 5 is predicted to have approximately $1,268,905 lower weekly sales than the reference Store 1, all else being equal.
- The model predicts that a week with a major public holiday (when Holiday_Flag is 1) is associated with an average increase of approximately $72,764 in weekly sales across all stores, compared to a non-holiday week, holding other factors constant.

c) Performance metrics of the multiple linear regression model:

--- MLR Model Performance ---
R-squared: 0.9208
RMSE: $159,683.97
MAE: $91,043.93

From the metrics we can see that the R^2 score is very high, confirming that the model, primarily driven by the store identity (as seen in the coefficients), is highly effective at capturing the variation in sales. The difference between RMSE ($159,684) and MAE ($91,044) suggests the presence of some larger errors or outliers in the test set, which disproportionately inflate the RMSE due to the squaring of errors.

d) Ridge Regression model:
The optimal regularization parameter was determined using 5-fold Cross-Validation via RidgeCV. The model tested a large, logarithmically-spaced range of alpha values (from 0.0001 to 10000). The value that minimized the cross-validated Negative Mean Squared Error (MSE), which is equivalent to minimizing the actual Mean Squared Error, was

selected as the best alpha. The Best Alpha value we get is : 0.0060. The performance of the ridge model was similar to that of the linear regression model.

e) Lasso Regression model:
Similar to the Ridge regression model, we use a 5-fold Cross-Validation via LassoCV.The model tested a large range of possible alpha values (from 0.0001 to 10000) and selected the one that minimized the cross-validated Mean Squared Error (MSE).The Best Alpha we get is: 0.5214.

f) Model Performance Comparison:

Model	R-squared	RMSE
Multiple Linear Regression (MLR)	0.920849	159683.968378
Ridge Regression	0.920847	159685.339136
Lasso Regression	0.920847	159685.141169

The performance metrics (R-squared and RMSE) are almost identical across all three models. This indicates that the dataset does not exhibit significant multicollinearity, and the Multiple Linear Regression (MLR) model is highly effective on its own. In this particular scenario, regularization techniques offer minimal additional predictive benefit over the standard MLR.

Question 2: framingham dataset

a) There were a few columns with missing values, I used a variation of median and mode imputation for continuous and binary/ordinal features respectively to fill out the null spaces.

b) The final features selected using the Backward Selection model were the following: ['male', 'age', 'cigsPerDay', 'prevalentStroke', 'totChol', 'sysBP', 'glucose']. This approach allowed us to keep the model really simple and yet retain important features that had a statistically significant relationship with the outcome.

c) Performance of the Logistic Regression model (Threshold: 0.5)

- Accuracy: 0.8479
- Sensitivity (Recall): 0.0620
- Specificity: 0.9889

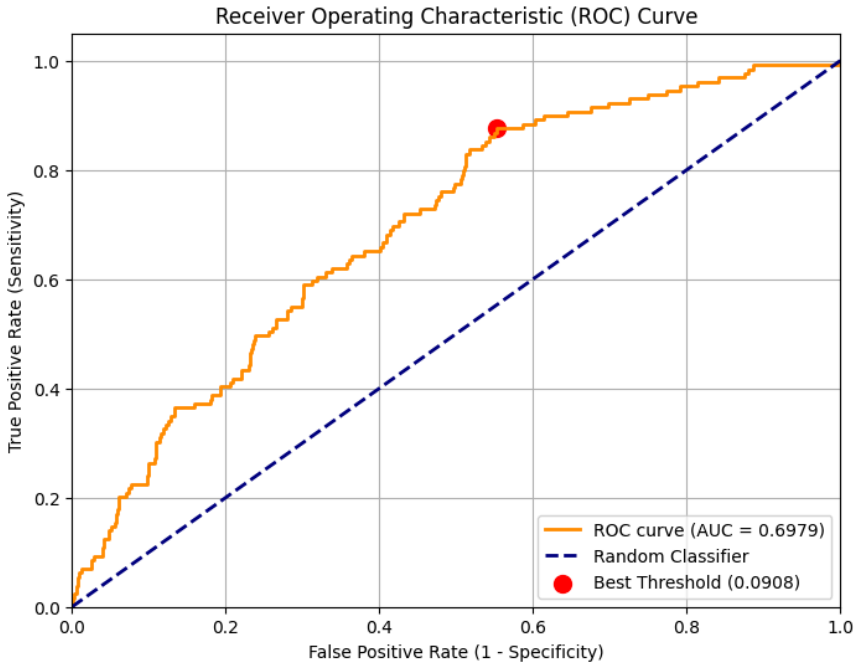
Confusion Matrix:

- True Negatives -> 711
- False Positives -> 8
- False Negatives -> 121
- True Positives -> 8

From this confusion matrix we can see that there is a significant class imbalance and there is a chance that our model is biased towards the majority class. This makes the model very reliable when it predicts "No CHD," but it leads to an unacceptably high rate of False Negatives (FN=121), meaning it misses the vast majority of patients who are actually at risk of developing heart disease.

(d) Optimal Threshold Results
Optimal Threshold (Maximizing Youden's J): 0.0908
Model Performance at Optimal Threshold:
Sensitivity: 0.8760
Specificity: 0.4465

(e) ROC curve and Area Under the Curve (AUC): 0.6979



The model's overall predictive power is fair, with the Area Under the Curve (AUC) measuring 0.6979.

Question 3: bmi dataset

a) One vs rest model:

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	3.00
1	0.00	0.00	0.00	4.00
2	0.50	0.71	0.59	14.00
3	0.00	0.00	0.00	14.00
4	0.56	0.77	0.65	26.00
5	0.89	1.00	0.94	39.00
Accuracy	**0.69**	**0.69**	**0.69**	**0.69**
Macro Avg	**0.32**	**0.41**	**0.36**	**100.00**
Weighted Avg	**0.56**	**0.69**	**0.62**	**100.00**

The One-vs-Rest (OvR) model achieved an Accuracy of 69.00%, making it the poorest performer of the three multiclass strategies. While OvR trained six separate binary classifiers to distinguish each BMI class from the rest, the method struggled significantly with the decision boundaries for the less represented classes (0, 1, and 3), resulting in an F1-score of 0.00 for these categories. The Macro Avg F1-score was low (0.3622), indicating that the OvR approach was not well-suited for this problem, as it failed to build effective, high-quality boundaries across all classes simultaneously.

b) One vs one model:

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.000000	0.000000	0.000000	3.00
1	0.000000	0.000000	0.000000	4.00
2	0.736842	1.000000	0.848485	14.00
3	0.928571	0.928571	0.928571	14.00
4	0.960000	0.923077	0.941176	26.00
5	0.974359	0.974359	0.974359	39.00
Accuracy	**0.890000**	**0.890000**	**0.890000**	**0.89**
Macro Avg	**0.599962**	**0.637668**	**0.615432**	**100.00**
Weighted Avg	**0.862758**	**0.890000**	**0.873494**	**100.00**

The One-vs-One (OvO) model achieved the highest accuracy of 89.00% among the three methods, proving to be the most effective strategy for the BMI classification problem. By training separate classifiers for every unique pair of classes, OvO successfully established clear decision boundaries, leading to excellent performance across the distinct BMI categories. This approach resulted in a strong Weighted Avg F1-score of 0.8735, demonstrating high predictive reliability and superior overall quality compared to OvR and Softmax.

c) Softmax function:

Classification Report:

Class	Precision	Recall	F1-Score	Support
0	0.00	0.00	0.00	3.00
1	0.00	0.00	0.00	4.00
2	0.74	1.00	0.85	14.00
3	1.00	0.57	0.73	14.00
4	0.81	0.96	0.88	26.00
5	0.97	0.97	0.97	39.00
Accuracy	**0.85**	**0.85**	**0.85**	**0.85**
Macro Avg	**0.59**	**0.58**	**0.57**	**100.00**
Weighted Avg	**0.83**	**0.85**	**0.83**	**100.00**

The Softmax (Multinomial) model achieved the second-highest accuracy of 85.00%, modeling the probability of belonging to all six BMI classes simultaneously using a single generalized linear function. This approach was highly effective, yielding a strong Weighted Avg F1-score of 0.8287, which indicates high overall predictive quality. However, by imposing a single global boundary structure, Softmax proved marginally less robust than the OvO method, slightly struggling with precision and recall on specific intermediate classes (like Class 3) and thus performing less optimally when forced to distinguish all categories at once.

d) Comparison of models:

Model	Accuracy	Macro Avg F1-score	Weighted Avg F1-score
OvO (One-vs-One)	0.89	0.615432	0.873494
Softmax (Multinomial)	0.85	0.571218	0.828676
OvR (One-vs-Rest)	0.69	0.362193	0.616601

The One-vs-One (OvO) model is the best performer across all key metrics:

- Accuracy: OvO has the highest overall accuracy (89%), meaning it correctly classifies the BMI index more often than the other two models.
- F1-Score: The high Macro F1-score (0.6154) indicates that OvO performs better across the classes, handling the individual BMI categories more effectively than Softmax and significantly better than OvR.

The superior performance of OvO suggests that for this specific dataset, classifying the BMI index is easier when the model focuses on linear boundaries between pairs of adjacent classes rather than trying to define a single class against all others (OvR) or modeling all class probabilities simultaneously (Softmax).