Group 5 - Neelan Sriranjan, Owusu Boakye Gideon, Vy Manohara
HAD5016H

Stroke Prediction from Lifestyle and Demographic Factors using L2-Regularized Logistic Regression

Introduction

Stroke is a major outcome in cardiovascular disease and remains the second leading cause of disability and mortality among non-communicable disorders [1]. Socioeconomic and demographic factors including employment status, gender, and age have been shown to influence cardiovascular risk through pathways related to stress, social determinants of health, and behavioural patterns [1, 2, 3]. In this project, we examined whether lifestyle and demographic features could be used to predict stroke using logistic regression, and whether a reduced model using a subset of these variables (age, gender, employment) performed as well as a full model including smoking status, marital status, and residence type. The dataset used for this analysis contains a subset of EMR data exploring various aspects of cardiovascular health from 4,254 Ontario patients who visited a primary care provider between January 2017 and July 2017.

Research Question

1) Can a logistic regression model using selected lifestyle features predict stroke in a primary care population in Ontario?
2) Does a reduced feature set perform comparably to a full model (ablation analysis)?

Data Preparation and Cleaning

From the original dataset of 4,254 patients, we removed 965 individuals older than 65 (retirement age) to focus on populations eligible to be in the workforce, aligning with our research question examining employment type and stroke. Individuals who denoted their gender as “Other” were removed due to extremely low cell size (n=1), which could cause unstable coefficients in logistic regression if they were included.

Clinical variables (BMI, glucose, hypertension, heart disease) were excluded from both models as the goal was to isolate the effects of lifestyle and demographic factors on stroke without confounding from existing comorbidities or a history of cardiovascular issues. All remaining categorical variables were label-encoded or one-hot encoded, and transformed into binary variables for analysis. We also removed the “ID” column in this dataset as it does not provide any meaningful contribution to the model, the arbitrary numerical data in this column may skew the relationships between the other numerical variables in the model. There were no missing data from the selected features of interest and no outliers. All features were standardized prior to model fitting.

Two datasets were created: 1) Model 1 (full lifestyle features): age, gender, employment status, smoking status, marital status, and residence type and 2) Model 2 (restricted subset): age, gender, employment status.

Analysis (model selection and justification)

Logistic regression was chosen because the outcome (stroke) is binary and the research objective involved interpreting independent contributions of categorical lifestyle factors, so linear regression would not have been appropriate to use. We could have created a KNN model to select features that affected stroke the most but we pre-selected features to be used in the model based on their clinical relevance, existing literature, and relevance to our research question.

L2 (Ridge) regularization was applied to the model given the small number of variables utilized in the analysis. This was meant to minimize the variance in certain coefficients to ensure better generalization of the data and helped prevent the model from overfitting. Class imbalance was substantial as there were more non-stroke cases than stroke cases in our dataset, so both models were balanced to ensure the algorithm did not default to predicting only the majority class. An ablation analysis was conducted by training both models on the same train/test split to compare the performance of the two models fairly.

Data were split into training (80%) and testing (20%) subsets using a fixed random_state=42 for reproducibility. Each model was evaluated using confusion matrices, precision, recall, F1-score, and accuracy. As stroke was rare in this dataset, F1 and recall were prioritized over accuracy for evaluating the model.

Findings (model evaluation)

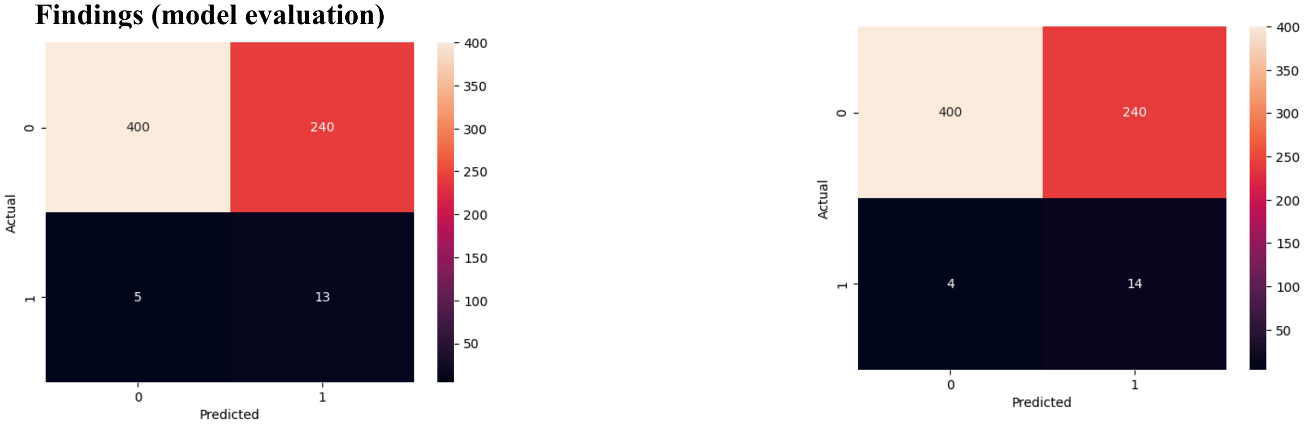


Fig 1. Confusion matrix for Model 1 (left) and Model 2 (right).

Both models performed similarly where the models successfully classified most non-stroke cases but had difficulty with stroke prediction. The models yielded very low precision and F1 scores (~0.05-0.06 and 0.10 respectively in each model) due to severe class imbalance. The restricted model achieved slightly higher recall and F1 despite using fewer variables, suggesting that additional lifestyle factors did not improve stroke prediction performance. Lifestyle factors alone have limited ability to detect stroke events and clinical factors are likely necessary to improve model performance. We also ran the logistic regression model and found that the model cannot estimate effects reliably because stroke cases were very rare (severe imbalance), some employment categories such as “Public Sector” had no stroke cases which makes it difficult to draw a meaningful relationship between the two.

Conclusion

This analysis found that lifestyle and demographic factors alone were insufficient to accurately predict stroke using logistic regression, even with balanced training. Both models demonstrate a moderate ability to rule out stroke, but weak ability to detect true cases. The ablation analysis showed minimal difference between the full and restricted models, suggesting that the additional lifestyle features did not add meaningful predictive value.

Limitations/Future Directions

The primary limitation was extreme class imbalance as there were few stroke cases in the dataset. Excluding individuals over 65 aligned with the research question but likely removed many stroke cases, further exacerbating this imbalance. Future analyses should consider undersampling the majority class (non-stroke), oversampling the minority class (stroke) to improve balance but be cautious of overfitting. Healthcare practitioners should consider analyzing both lifestyle factors and clinical factors such as hypertension, BMI and glucose levels to improve stroke prediction.

Individual Contributions

All team members contributed equally to the code development, analysis, and knowledge products (report and presentation).

References

1. Rexrode, K. M., Madsen, T. E., Yu, A. Y. X., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The Impact of Sex and Gender on Stroke. *Circulation Research*, *130*(4), 512–528. https://doi.org/10.1161/CIRCRESAHA.121.319915
2. Carson, A. P., Rose, K. M., Catellier, D. J., Diez-Roux, A. V., Muntaner, C., & Wyatt, S. B. (2009). Employment Status, Coronary Heart Disease, and Stroke Among Women. *Annals of Epidemiology*, *19*(9), 630–636. https://doi.org/10.1016/j.annepidem.2009.04.008
3. Eshak, E. S., Honjo, K., Iso, H., Ikeda, A., Inoue, M., Sawada, N., & Tsugane, S. (2017). Changes in the Employment Status and Risk of Stroke and Stroke Types. *Stroke (1970)*, *48*(5), 1176–1182. https://doi.org/10.1161/STROKEAHA.117.016967

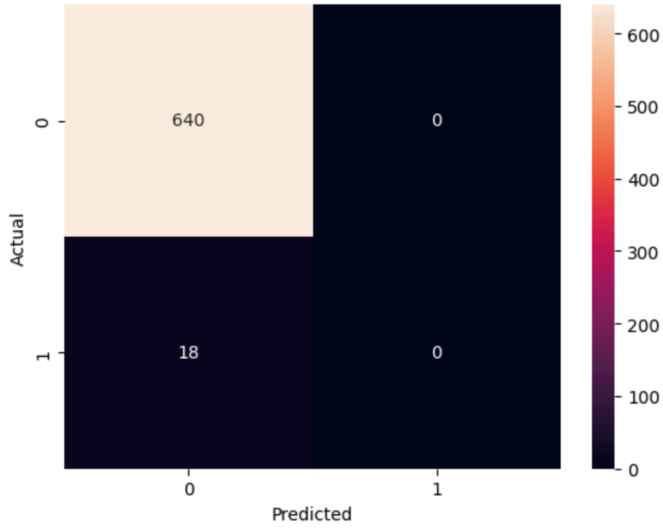
CoLab File:
https://colab.research.google.com/drive/1JdOYlr7tU0uKLgRBjcV2W--7-x0E1Xt-?usp=sharing

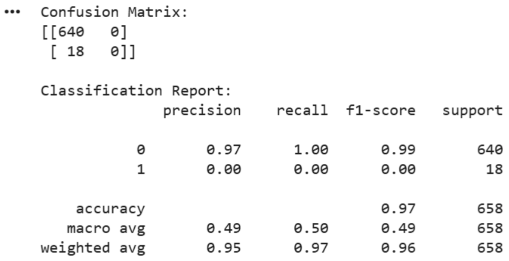
GitHub Link:
https://github.com/GIDEonBoakye/5-HAD5016-S26

Presentation:
https://docs.google.com/presentation/d/1wSyBCVeRuORU6sz3GUEvu492ImdqbQHuR2YPCcVVlqc/edit?usp=sharing

Figures

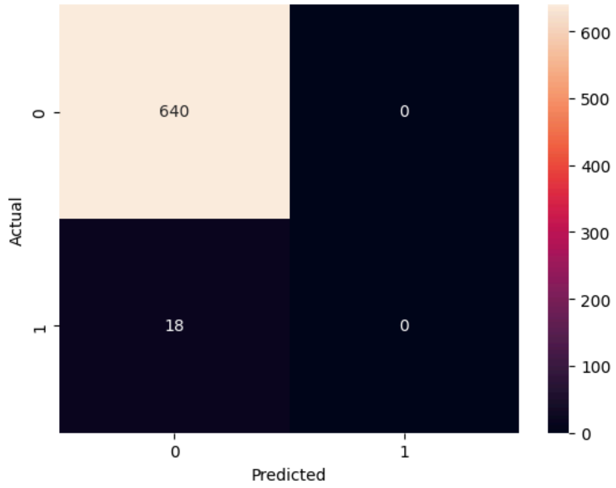
Fig 1. Model 1 confusion matrix (unbalanced) for ALL lifestyle features (age, gender, employment, residence type, marital status, smoking status, stroke)

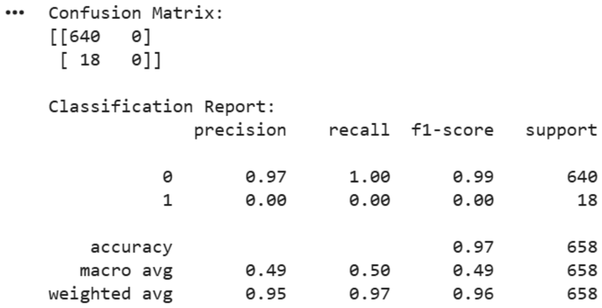




```
Confusion Matrix:
[[640   0]
 [ 18   0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       640
           1       0.00      0.00      0.00        18

    accuracy                           0.97       658
   macro avg       0.49      0.50      0.49       658
weighted avg       0.95      0.97      0.96       658
```

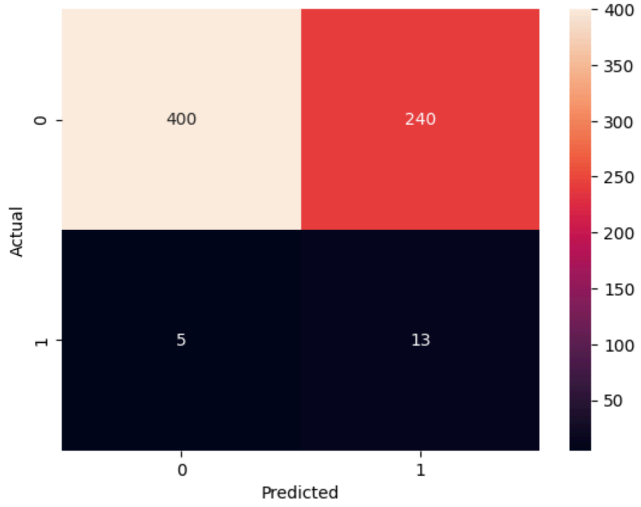
Fig 2. Model 2 confusion matrix (unbalanced) for select lifestyle features (age, gender, employment, stroke)

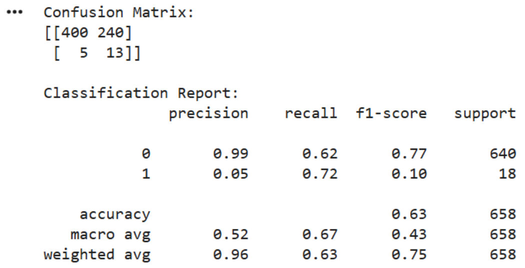




```
Confusion Matrix:
[[640   0]
 [ 18   0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       640
           1       0.00      0.00      0.00        18

    accuracy                           0.97       658
   macro avg       0.49      0.50      0.49       658
weighted avg       0.95      0.97      0.96       658
```

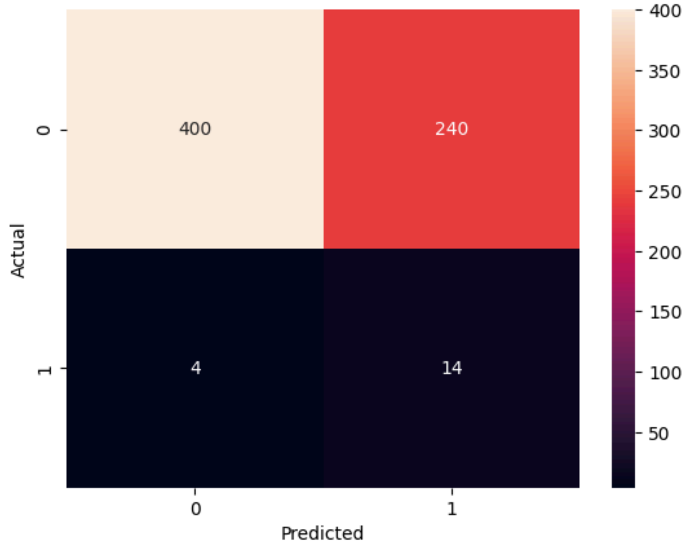
Fig 3. Model 1 confusion matrix (balanced) for ALL lifestyle features (age, gender, employment, residence type, marital status, smoking status, stroke)

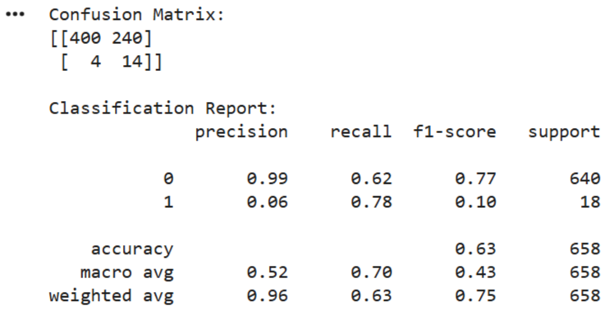




```
••• Confusion Matrix:
    [[400 240]
     [  5  13]]

    Classification Report:
                  precision    recall  f1-score   support

               0       0.99      0.62      0.77       640
               1       0.05      0.72      0.10        18

        accuracy                           0.63       658
       macro avg       0.52      0.67      0.43       658
    weighted avg       0.96      0.63      0.75       658
```

Fig 4. Model 2 confusion matrix (balanced) for select lifestyle features (age, gender, employment, stroke)





```
••• Confusion Matrix:
    [[400 240]
     [  4  14]]

    Classification Report:
                  precision    recall  f1-score   support

               0       0.99      0.62      0.77       640
               1       0.06      0.78      0.10        18

        accuracy                           0.63       658
       macro avg       0.52      0.70      0.43       658
    weighted avg       0.96      0.63      0.75       658
```