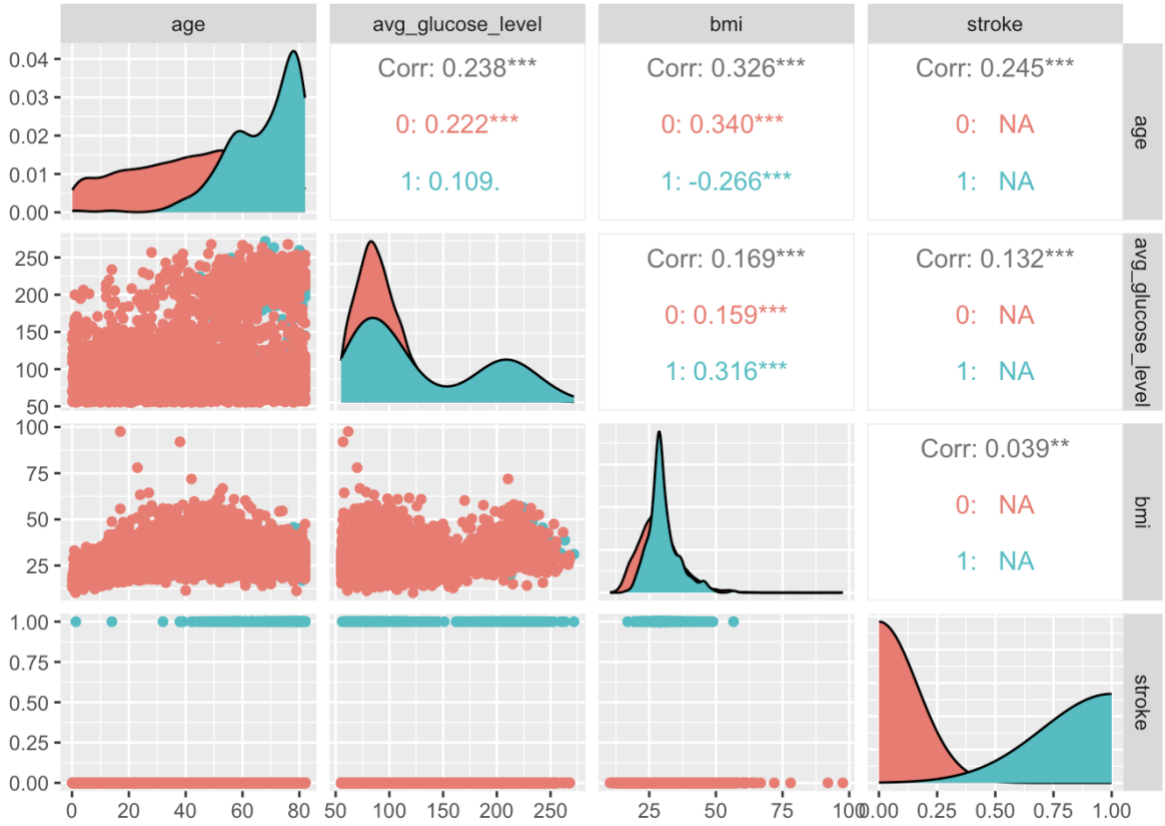
MARCH 25, 2025

STROKE PREDICTION

PROJECT REPORT

DUC ANH NGUYEN
UNIVERSITY OF CENTRAL FLORIDA
STA 5703 DATA MINING METHODOLOGY I

I. Exploratory Data Analysis (EDA)

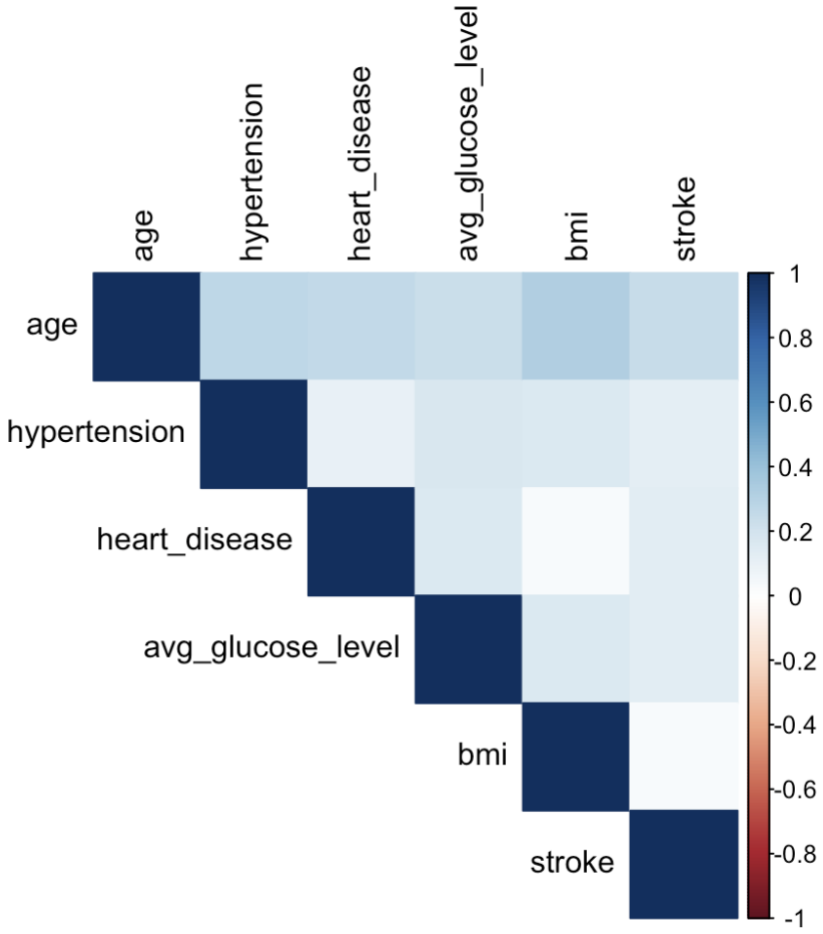


Age and stroke show a moderate correlation (0.245), suggesting that age may be a significant predictor.

Age and BMI (0.326), as well as Age and Glucose Level (0.238), display some correlation, but these are not excessively strong.

The correlation between BMI and stroke (0.039) is weak, indicating that BMI alone might not be a strong predictor of stroke.

The separate correlation values for different stroke groups (0 and 1) suggest differences in how variables interact based on stroke occurrence.



From the heatmap, it's evident that most features exhibit minimal correlation with one another, making them well-suited for regression. Among all features, age shows the strongest correlation with stroke.

```
##        id            gender                age          hypertension
##  Min.   :   67   Length:5110        Min.   : 0.08   Min.   :0.00000
##  1st Qu.:17741   Class :character   1st Qu.:25.00   1st Qu.:0.00000
##  Median :36932   Mode  :character   Median :45.00   Median :0.00000
##  Mean   :36518                      Mean   :43.23   Mean   :0.09746
##  3rd Qu.:54682                      3rd Qu.:61.00   3rd Qu.:0.00000
##  Max.   :72940                      Max.   :82.00   Max.   :1.00000
##  heart_disease     ever_married        work_type         Residence_type
##  Min.   :0.00000   Length:5110        Length:5110        Length:5110
##  1st Qu.:0.00000   Class :character   Class :character   Class :character
##  Median :0.00000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.05401
##  3rd Qu.:0.00000
##  Max.   :1.00000
##  avg_glucose_level      bmi         smoking_status         stroke
##  Min.   : 55.12    Min.   :10.30   Length:5110        Min.   :0.00000
##  1st Qu.: 77.25    1st Qu.:23.80   Class :character   1st Qu.:0.00000
##  Median : 91.89    Median :28.40   Mode  :character   Median :0.00000
##  Mean   :106.15    Mean   :28.89                      Mean   :0.04873
##  3rd Qu.:114.09    3rd Qu.:32.80                      3rd Qu.:0.00000
##  Max.   :271.74    Max.   :97.60                      Max.   :1.00000
```

From the structure and summary of the dataset, we find out that:

- There are 9 input features and 1 outcome(stroke) in the dataset. We don't need the feature ID.
- Some columns include character, which might need to be transformed into factor or number.
- The mean of stroke is 0.04, which means only 4% of the patients have stroke.

II. Model Selection Methods

1. Best Subset Selection (Cp)

```
## # A tibble: 10 × 4
##     Size Adj_R2    Cp   BIC
##    <int>  <dbl> <dbl> <dbl>
##  1     1 0.0617 84.9  -180.
##  2     2 0.0691 61.1  -197.
##  3     3 0.0757 39.8  -212.
##  4     4 0.0805 24.9  -221.
##  5     5 0.0832 16.7  -223.
##  6     6 0.0850 11.7  -222.
##  7     7 0.0857 10.4  -217.
##  8     8 0.0874  5.52 -216.
##  9     9 0.0879  4.90 -211.
## 10    10 0.0880  5.61 -204.
```

Adjusted R^2 increases as more predictors are added, peaking at 10 predictors (0.088). However, the improvement becomes marginal beyond 6 predictors (0.08499), suggesting diminishing returns. The best trade-off is typically where Adjusted R^2 stabilizes, which appears to be around 6 predictors.

Mallows' Cp values decrease with model complexity, indicating improved fit. A Cp value close to the number of predictors suggests an optimal model, and the 6-predictor model (Cp = 11.72) stands out as a strong candidate. This balance suggests it adequately captures variability without unnecessary complexity.

BIC (Bayesian Information Criterion) favors model simplicity, reaching its lowest value at 5 predictors (-223.14). This suggests the 5-predictor model is the best fit while avoiding overfitting. Beyond this point, BIC increases, indicating that adding more predictors may introduce unnecessary complexity.

2. Forward Selection Using AIC

```
##
## Call:
## lm(formula = stroke ~ 1, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05055 -0.05055 -0.05055 -0.05055  0.94945
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.050554   0.003957   12.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2191 on 3065 degrees of freedom
```

The forward selection method for AIC starts with an intercept-only model and progressively adds variables that improve the model's fit based on the Akaike Information Criterion (AIC). Initially, the model contains only the intercept, which is highly significant (p-value < 2e-16). This indicates that at this stage, the intercept is statistically reliable but does not yet include any predictors. The residual standard error for this model is 0.2191, which indicates the average deviation of observed values from the predicted values.

The forward selection process is designed to gradually introduce predictors into the model, aiming to improve predictive power while minimizing model complexity. In subsequent steps, significant predictors will be added, guided by AIC criteria, to ensure that the final model provides the best balance of fit and complexity.

3. Backward Elimination Using AIC

```
##
## Call:
## lm(formula = stroke ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + bmi, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32854 -0.08059 -0.02248  0.00621  1.02631
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.429e-02  1.680e-02  -1.445 0.148437
## age                       3.029e-03  2.799e-04  10.822  < 2e-16 ***
## hypertension              5.765e-02  1.348e-02   4.276 1.96e-05 ***
## heart_disease             6.639e-02  1.744e-02   3.806 0.000144 ***
## ever_marriedYes          -3.374e-02  1.127e-02  -2.995 0.002767 **
## work_typeGovt_job        -6.340e-02  1.884e-02  -3.365 0.000774 ***
## work_typeNever_worked    -2.910e-02  4.927e-02  -0.591 0.554802
## work_typePrivate         -5.075e-02  1.551e-02  -3.272 0.001078 **
## work_typeSelf-employed   -6.458e-02  1.940e-02  -3.328 0.000885 ***
## avg_glucose_level         2.739e-04  8.791e-05   3.115 0.001855 **
## bmi                      -8.754e-04  5.536e-04  -1.581 0.113913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2093 on 3055 degrees of freedom
## Multiple R-squared:  0.09071,    Adjusted R-squared:  0.08773
## F-statistic: 30.48 on 10 and 3055 DF,  p-value: < 2.2e-16
```

The backward elimination model retains significant predictors: age, hypertension, heart disease, work type, and average glucose level, while excluding BMI. The R-squared value of 0.09071 shows the model explains 9% of variability, with a modest adjusted R-squared of 0.08773, indicating limited explanatory power. The model fits reasonably well with a residual standard error of 0.2093 but suggests room for improvement. The results align with forward selection, highlighting similar key predictors but still indicating potential for enhanced predictive accuracy.

4. Stepwise Selection Using AIC

```
##
## Call:
## lm(formula = stroke ~ age + hypertension + heart_disease + ever_married +
##     work_type + avg_glucose_level + bmi, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32854 -0.08059 -0.02248  0.00621  1.02631
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.429e-02  1.680e-02  -1.445 0.148437
## age                       3.029e-03  2.799e-04  10.822  < 2e-16 ***
## hypertension              5.765e-02  1.348e-02   4.276 1.96e-05 ***
## heart_disease             6.639e-02  1.744e-02   3.806 0.000144 ***
## ever_marriedYes          -3.374e-02  1.127e-02  -2.995 0.002767 **
## work_typeGovt_job        -6.340e-02  1.884e-02  -3.365 0.000774 ***
## work_typeNever_worked    -2.910e-02  4.927e-02  -0.591 0.554802
## work_typePrivate         -5.075e-02  1.551e-02  -3.272 0.001078 **
## work_typeSelf-employed   -6.458e-02  1.940e-02  -3.328 0.000885 ***
## avg_glucose_level         2.739e-04  8.791e-05   3.115 0.001855 **
## bmi                      -8.754e-04  5.536e-04  -1.581 0.113913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2093 on 3055 degrees of freedom
## Multiple R-squared:  0.09071,    Adjusted R-squared:  0.08773
## F-statistic: 30.48 on 10 and 3055 DF,  p-value: < 2.2e-16
```

The stepwise selection method, like backward elimination, retains the same significant predictors, including age, hypertension, heart disease, work type, and avg glucose level, while excluding BMI and Never worked. This suggests that stepwise selection agrees with backward elimination in terms of the most important predictors.

The model's explanatory power (R-squared = 9%) remains low, suggesting potential for improvement, perhaps by considering additional variables or more complex modeling techniques.

III. Regularization and Cross-Validation

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                 -6.42441743
## genderMale                   .
## genderOther                  .
## age                          0.05554360
## hypertension                 0.24856914
## heart_disease                0.24016114
## ever_marriedYes              .
## work_typeGovt_job            .
## work_typeNever_worked        .
## work_typePrivate             .
## work_typeSelf-employed       .
## Residence_typeUrban          .
## avg_glucose_level            0.00296555
## bmi                          .
## smoking_statusnever smoked   .
## smoking_statussmokes         .
## smoking_statusUnknown        .
```

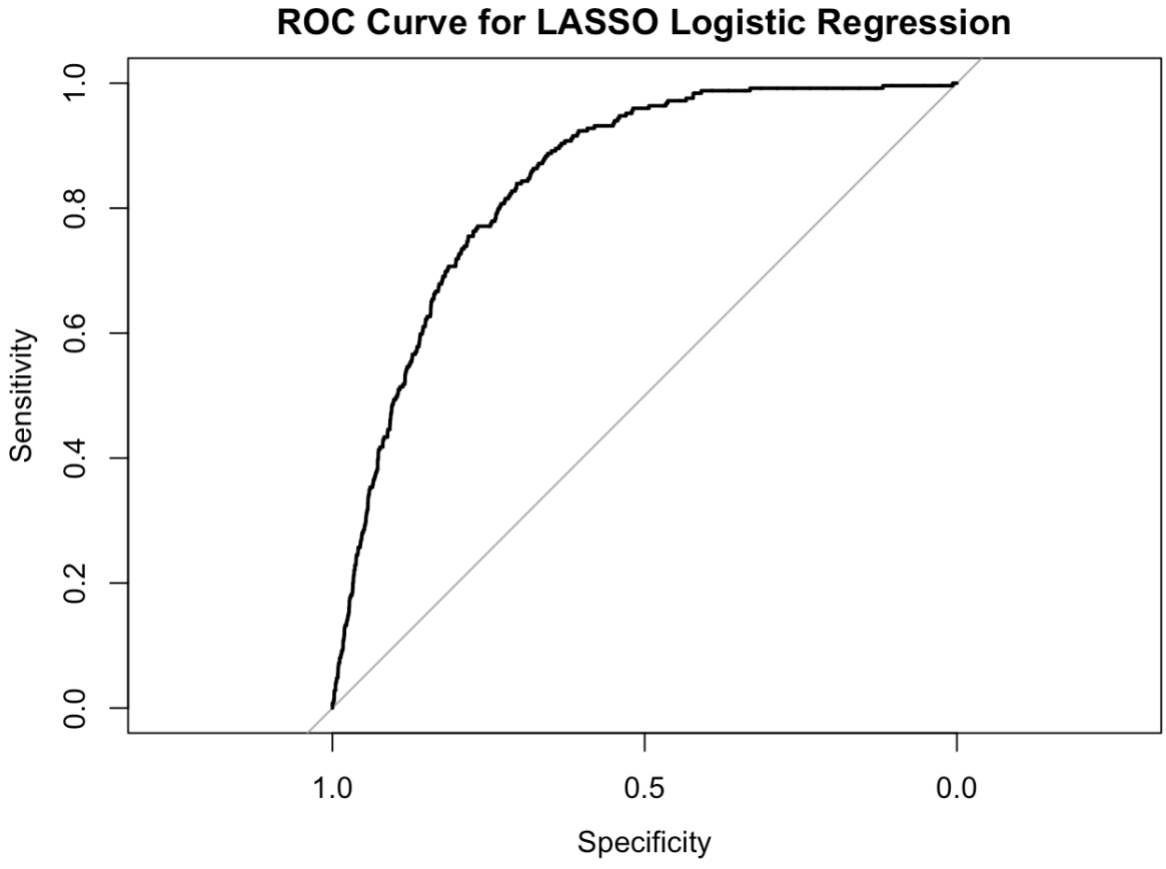
The logistic regression model shows that age, hypertension, heart disease, and average glucose level are significant factors in predicting stroke risk. As age increases, the likelihood of having a stroke also increases. Both hypertension and heart disease raise the risk, while glucose levels have a small positive effect. Other variables like gender, work type, and smoking status were excluded from the model, indicating they are less important in predicting stroke risk.

```
          Reference
Prediction    0    1
         0 4861  249
         1    0    0
```

True Negatives (TN): 4861 – The model correctly predicted no stroke (negative class) for 4861 instances.
False Positives (FP): 249 – The model incorrectly predicted a stroke (positive class) when there was none for 249 instances.
False Negatives (FN): 0 – The model did not miss any actual stroke cases (there were no false negatives).
True Positives (TP): 0 – The model did not correctly predict any stroke cases (there were no true positives).

Sensitivity is 0, which means the model is not detecting any positive cases (stroke). This is problematic, especially in medical predictions.
Specificity is 0.9503, which shows that the model is very good at identifying negative cases (no stroke).

The AUC of 0.8439 suggests the model has a good ability to discriminate between classes, but improvements are needed to improve sensitivity (detection of strokes).



The ROC curve is above the diagonal (random classifier), indicating that the model is better than random guessing.

The curve rises sharply at first, meaning the model correctly identifies strokes with high sensitivity at lower false positive rates.

However, as specificity increases, the gain in sensitivity diminishes, showing that some positive cases are missed as threshold tightens.

IV. Model Performance Evaluation

1. Compare different model selection methods and justify the final model choice.

Best Subset Selection:

Adjusted R^2 increases with the addition of more predictors, peaking at 10 predictors (Adjusted R^2 = 0.088). However, the increase becomes marginal after 6 predictors (Adjusted R^2 = 0.08499), suggesting diminishing returns. The 6-predictor model strikes the best trade-off between explanatory power and complexity.

Mallows' Cp: For the 6-predictor model (Cp = 11.72), this indicates a good fit with minimal complexity. The 6-predictor model provides an effective balance between model performance and complexity.

BIC (Bayesian Information Criterion): BIC reaches its lowest value at 5 predictors (-223.14), indicating that the 5-predictor model offers the best balance between fit and simplicity. Adding more predictors beyond this point results in higher BIC values, indicating unnecessary complexity.

Stepwise Selection (Forward, Backward, and Mixed):

Both the stepwise selection and backward elimination methods align in identifying the most important predictors. In the Forward AIC method, the model initially only includes the intercept, which is highly significant, meaning the model does not yet include any predictors. The intercept's statistical reliability suggests a solid foundation, but additional predictors must be added in subsequent steps of the forward selection process to build a more meaningful predictive model.

The R-squared value of 9% indicates that the model explains a modest portion of the variability, leaving room for improvement, possibly through additional predictors or more advanced modeling techniques.

Lasso Logistic Regression with Cross-Validation:
Cross-validation was employed for Lasso logistic regression, which played a crucial role in selecting the final model. Lasso not only helps with variable selection by penalizing the inclusion of less important predictors but also regularizes the model to avoid overfitting. Cross-validation ensures that the model is evaluated across different subsets of the data, providing a robust estimate of its performance and enhancing its ability to generalize to new data.

The Lasso model retained the following significant predictors: age, hypertension, heart disease, and average glucose level. These predictors were deemed essential for predicting stroke risk. Other variables like gender, work type, BMI, and smoking status were excluded from the model, indicating that they did not contribute meaningfully to stroke prediction after regularization.

Justification for the Final Model:
The final model choice is justified by combining the results from best subset selection, stepwise selection, and Lasso logistic regression with cross-validation. The 6-predictor model provides a good balance between model complexity and fit, as indicated by the Adjusted R^2 and Mallows' Cp. Lasso logistic regression helps refine the model by penalizing less important predictors, and its use of cross-validation ensures robustness and mitigates overfitting. The final model, which includes age, hypertension, heart disease, and average glucose level, effectively predicts stroke risk while maintaining generalizability.

2. Discuss potential overfitting issues and how cross-validation helps mitigate them.
Overfitting occurs when a model learns not only the true underlying patterns in the data but also the noise and random fluctuations, leading to excellent performance on the training data but poor generalization to new, unseen data. In this case, overfitting might happen if the model is too complex, incorporating too many predictors or overemphasizing certain features, leading to high specificity but zero sensitivity (failing to detect any positive cases like strokes).

Cross-validation helps mitigate overfitting by splitting the dataset into multiple folds, training the model on one subset, and validating it on another. This process ensures that the model is evaluated on data it hasn't seen before, providing a more reliable estimate of its performance. By using cross-validation, you can assess the model's ability to generalize and adjust it accordingly to prevent overfitting. It also allows for fine-tuning the model's parameters and regularization techniques to strike the right balance between bias and variance, ensuring that the model performs well both on training and unseen data.

3. Present key takeaways from the model selection process.

Model Complexity vs. Fit: A balance between model complexity and fit is crucial. The 6-predictor model, identified through best subset selection and Mallows' Cp, provides an optimal balance, capturing variability without overcomplicating the model.

Importance of Cross-Validation: Cross-validation ensures the model generalizes well, preventing overfitting and improving its robustness by evaluating its performance on different data subsets.

Variable Selection: Techniques like Lasso and stepwise selection identified key predictors (e.g., age, hypertension, heart disease, and glucose levels) as significant, while reducing noise by excluding irrelevant variables.

Model Performance: The model demonstrated good discrimination with an AUC of 0.8439, but sensitivity improvements are necessary, highlighting the importance of fine-tuning for detecting positive cases (strokes).