INTERPRETABLE STATISTICAL MODELING OF STUDENT DEPRESSION RISK (GLM & LASSO)

CONTENTS

PROJECT AT A GLANCE

- **Objective:**
  - To evaluate and compare interpretable statistical models (GLM and LASSO) for predicting depression risk in students using large-scale survey data.
- **Scope:**
  - Dataset: ~27,900 anonymized student survey responses.
  - Analysis limited to Steps 1–9: preprocessing, modeling, evaluation, error analysis, and interpretability.
- **Key Outcomes:**
  - Both GLM and LASSO achieved comparable performance (~84.7% accuracy, AUC ~0.92).
  - LASSO favored sensitivity; GLM favored specificity.
  - Feature-level interpretability enabled identification of robust predictors. models perform strongly.

DATASET OVERVIEW:

- Source: Kaggle - Mental Health in Students
- Dataset Link: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset
- Total Records: ~27900 student survey responses
- Target Variable: Depression (Binary: 1 = Yes, 0 = No)
- Class Distribution: ≈ 16336 (Yes) vs 11565 (No)

- **Predictors (22 features):**

a) Demographics (age, gender, city, profession, degree)
b) Family history & lifestyle factors.
c) Mental health indicators (suicidal thoughts, treatment history, stress levels, etc)

- **Why this dataset:**

a) Large & balanced sample (~28k responses, fair distribution)
b) Public, reproducible, and Kaggle verified (100% completeness, credibility, compatibility)
.

Note: Public, reproducible dataset used for methodological evaluation

STEP 1 -DATA INTAKE & VERSIONING

Work:

- Imported the raw student survey dataset (~27,901 records).
- Verified target variable: Depression (Binary → 1 = Yes, 0 = No).
- Checked class distribution: ≈ 16,336 (Yes) vs 11,565 (No).
- Confirmed data integrity (no missing outcome labels).
- No cleaning or transformations at this stage — sanity check only.

STEP 2 -DATA PREPROCESSING

Work:

1. Cleaned column names → machine-friendly variables.
2. Cast character columns to factors.
3. Recoded key binaries to 0/1 (e.g., suicidal thoughts, family history of mental illness).
4. Dropped non-informative columns (ID, city, profession, degree).

Output:

Processed dataset objects in R workspace:

D_model (22 numeric predictors, clean & model-ready)

y (binary outcome: 0 = No Depression, 1 = Depression)

(Dataset now ready for training in Step 3 with LASSO Logistic Regression.)

STEP 3--LOGISTIC REGRESSION WITH LASSO

Work:

- Applied cross-validation (cv.glmnet) to select the optimal penalty (λ), preventing overfitting.
- Trained the final LASSO logistic regression model using the best λ.
- Extracted key predictors strongly associated with depression risk.
- Saved model coefficients for later interpretation.

OUTPUT:

- Cross-validation error curve (λ vs. misclassification error).
- Model coefficient objects saved(coef_df_feature_importance, lasso_top_coefficients)
- Top 10 most important features (LASSO)

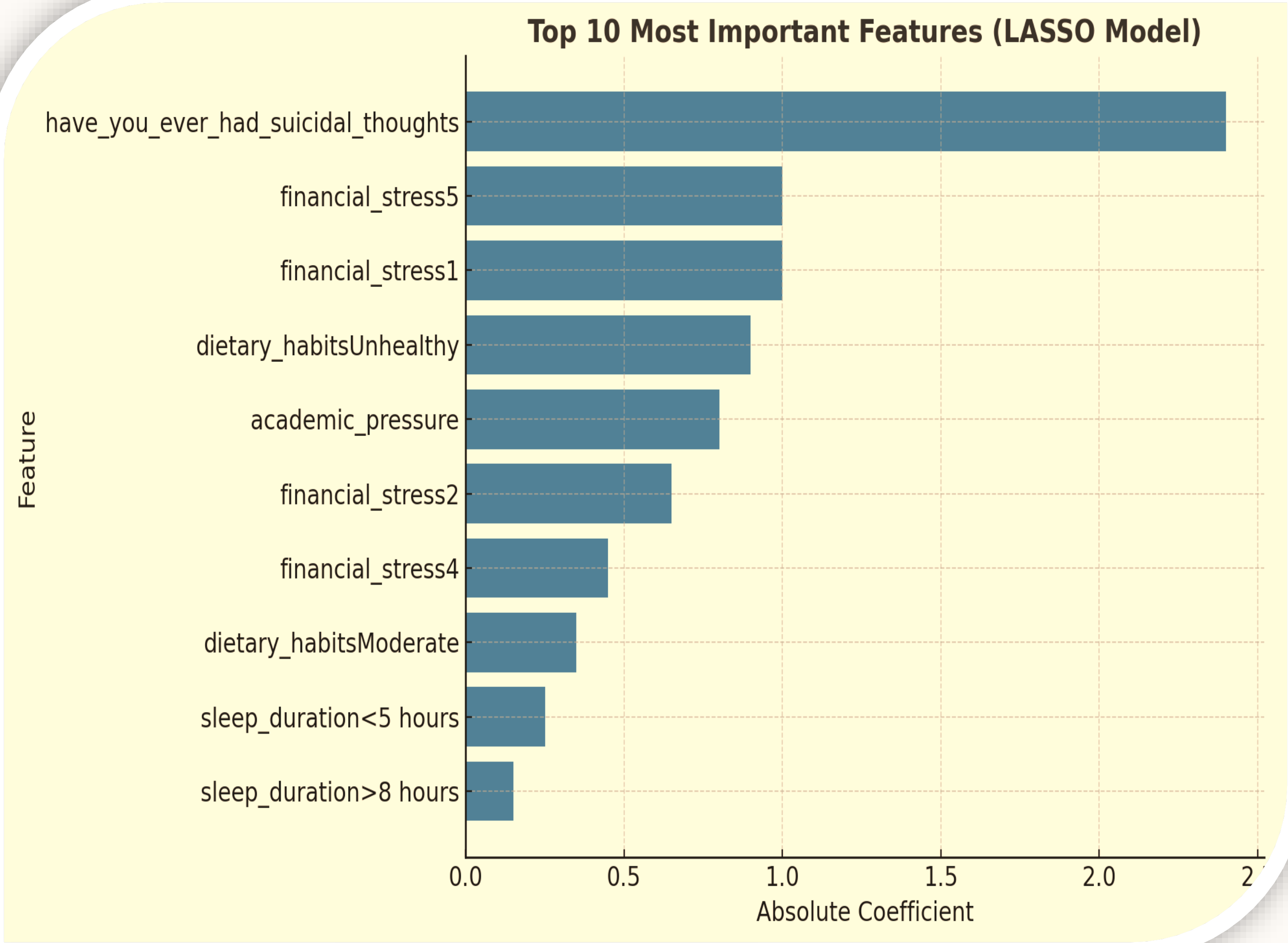


Figure Top 10 predictors selected by LASSO Logistic Regression (absolute coefficient values).

STEP 4 -- MODEL EVALUATION

Work:

- **Generated predicted probabilities of depression for each student using the trained LASSO model.**
- **Converted probabilities into binary class predictions (threshold = 0.5).**
- **Evaluated performance with a confusion matrix (caret::confusionMatrix) → accuracy, sensitivity, specificity.**
- **Stored raw predictions for later interpretability analysis.**

Output:

- **Confusion Matrix (confusion_matrix_logistic_LASSO_2025.txt).**
- **Evaluation Metrics (model_eval_logistic_LASSO_2025.txt) → see results on right.**
- **Predictions file: predictions_logistic_LASSO_2025.csv.**
- **Interpretable predictions (predictions_logistic_LASSO_2025_interpretable.csv) → used in Step 5.**

	Actual = 0 (Not Depressed)	Actual = 1 (Depressed)
Predicted = 0	9135 → True Negatives (TN)	1835 → False Negatives (FN)
Predicted = 1	2430 → False Positives (FP)	14501 → True Positives (TP)

Confusion matrix showing classification results at threshold = 0.5

```
Model Evaluation Results - Logistic Regression with LASSO
==========================================================
Accuracy:  0.8471
Sensitivity (Recall for depressed):  0.8877
Specificity (Recall for non-depressed):  0.7899
```

Evaluation metrics derived from confusion matrix (Accuracy, Sensitivity, Specificity)

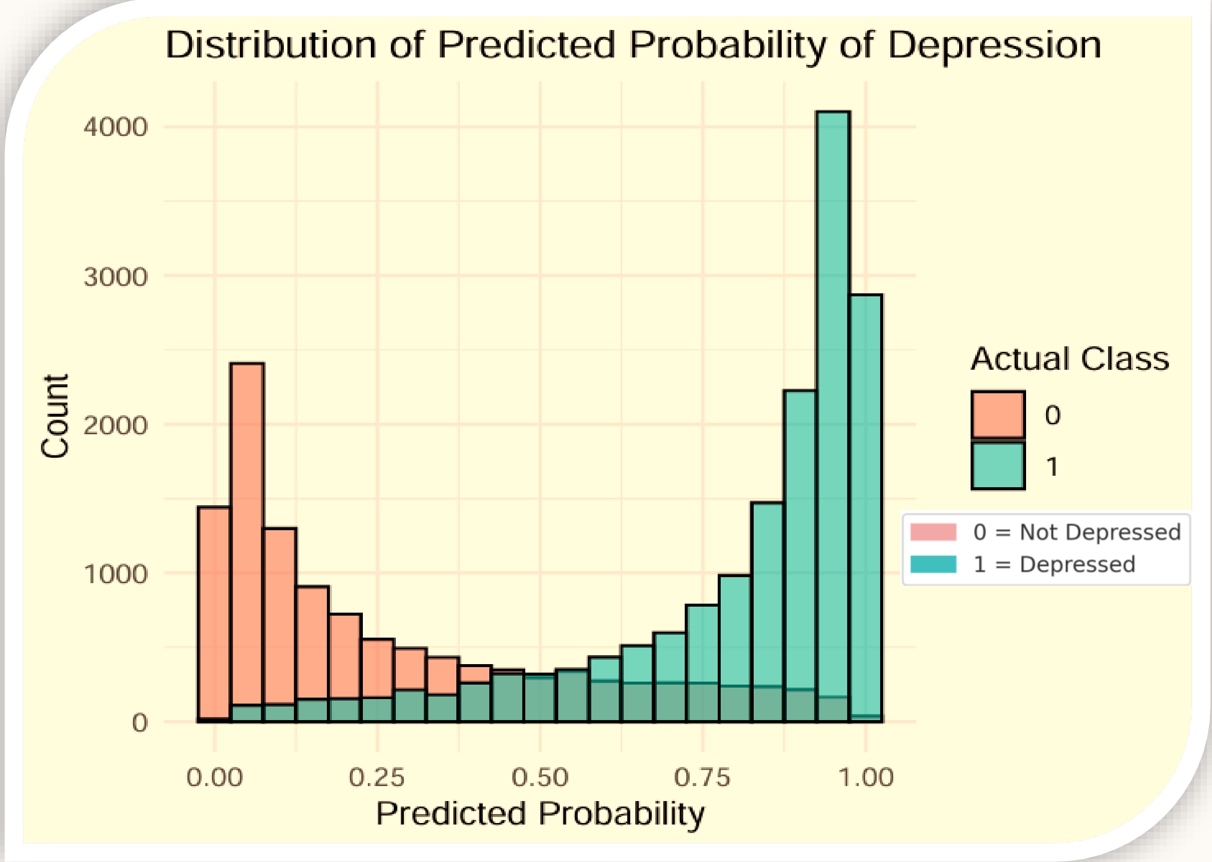
	Actual	Predicted_Class	Probability
0	1	1	0.588900
1	0	0	0.032600
2	0	0	0.044300
3	1	1	0.865300
4	0	1	0.564000
5	0	0	0.017600

Sample predictions with actual outcome, predicted class, and model probability (rounded to 4 decimals)

STEP 5 : ERROR ANALYSIS & MISCLASSIFICATIONS

Work:

- **Compared predicted vs actual depression labels to detect misclassifications.**
- **Classified errors into:**
  - **False Positives (FP): Non-depressed predicted as depressed.**
  - **False Negatives (FN): Depressed predicted as non-depressed.**
- **Analyzed probability distributions to understand model uncertainty.**
- **Visualized overlap between correctly classified and misclassified samples.**

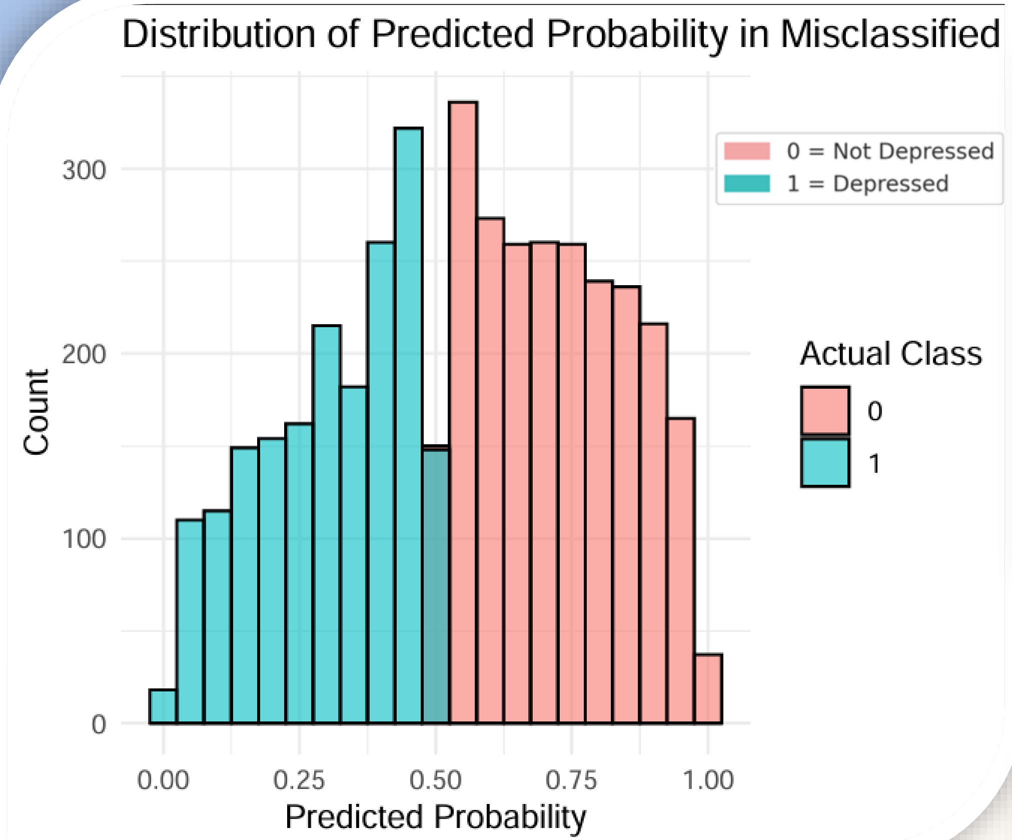


Output:

- **Misclassified cases file – misclassified_cases_logistic_LASSO_2025.xlsx**
- **Distribution of predicted probabilities (with interpretation) → distribution_predicted_probability_depression.png**
- **Distribution of predicted probabilities in misclassified cases (with interpretation) → misclassified_predicted_probability_depression.png**

Interpretations:

- The distribution shows two clear peaks:
  - 0–0.2 (pink): confidently predicted Not Depressed
  - 0.8–1.0 (teal): confidently predicted Depressed
  - Few cases fall in the middle range (0.4–0.6), meaning fewer uncertain predictions.
- This confirms good class separation and strong discriminatory power of the model.



Sample Interpretable Predictions (Confidence - Labeled)

Actual	Predicted_Class	Predicted_Probability	Predicted_Label	Interpretation
1	1	0.588859093	Depressed	Moderate
0	0	0.032595413	Not Depressed	Very Low
0	0	0.044343717	Not Depressed	Very Low
1	1	0.865335786	Depressed	Very High
0	1	0.563952868	Depressed	Moderate
0	0	0.017624436	Not Depressed	Very Low

This table illustrates how the logistic LASSO model predictions are mapped into interpretable labels with confidence levels, helping to identify cases predicted with Very Low, Moderate, or Very High certainty.

Interpretations:

- **Errors are concentrated in the middle probability range (0.3–0.7), showing model uncertainty in ambiguous cases.**
- **Few misclassifications at the extremes (near 0 or 1), where the model is more confident.**
- **Balanced errors across classes → similar misclassification patterns for depressed and non-depressed.**

Preview of Misclassified Predictions

Actual	Predicted_Class	Predicted_Probability	Predicted_Label	Interpretation
0	1	0.563952867558396	Depressed	Moderate
1	0	0.289823134582525	Not Depressed	Low
1	0	0.364274092889944	Not Depressed	Low
1	0	0.256236468274594	Not Depressed	Low
0	1	0.557786957065682	Depressed	Moderate
0	1	0.910652661265129	Depressed	Very High

The misclassified predictions arise from model errors, not dataset label errors. They show cases where the logistic LASSO model predicted incorrectly (*e.g., false positives/false negatives*), mostly in the ambiguous probability range (0.3–0.7)

STEP 6 : FEATURE IMPORTANCE USING LASSO

Work:

- Extracted feature coefficients from the trained LASSO model.
- Reused coefficients from Step 3 model to compute ranked feature importance.
- Ranked features by absolute coefficient values to measure importance.
- Identified the Top 10 predictors most strongly associated with depression risk .
- Highlighted interpretation of key predictors (suicidal thoughts, financial stress, academic pressure, sleep duration, dietary habits).

OUTPUT:

- Visualization of Top 10 important features (top10_lasso_features.png).
- Coefficient objects (already saved in Step 3):
  coef_df_feature_importance and
  lasso_top_coefficients
- Interpretation summary:
  Features with larger absolute coefficients have a stronger effect on depression prediction.

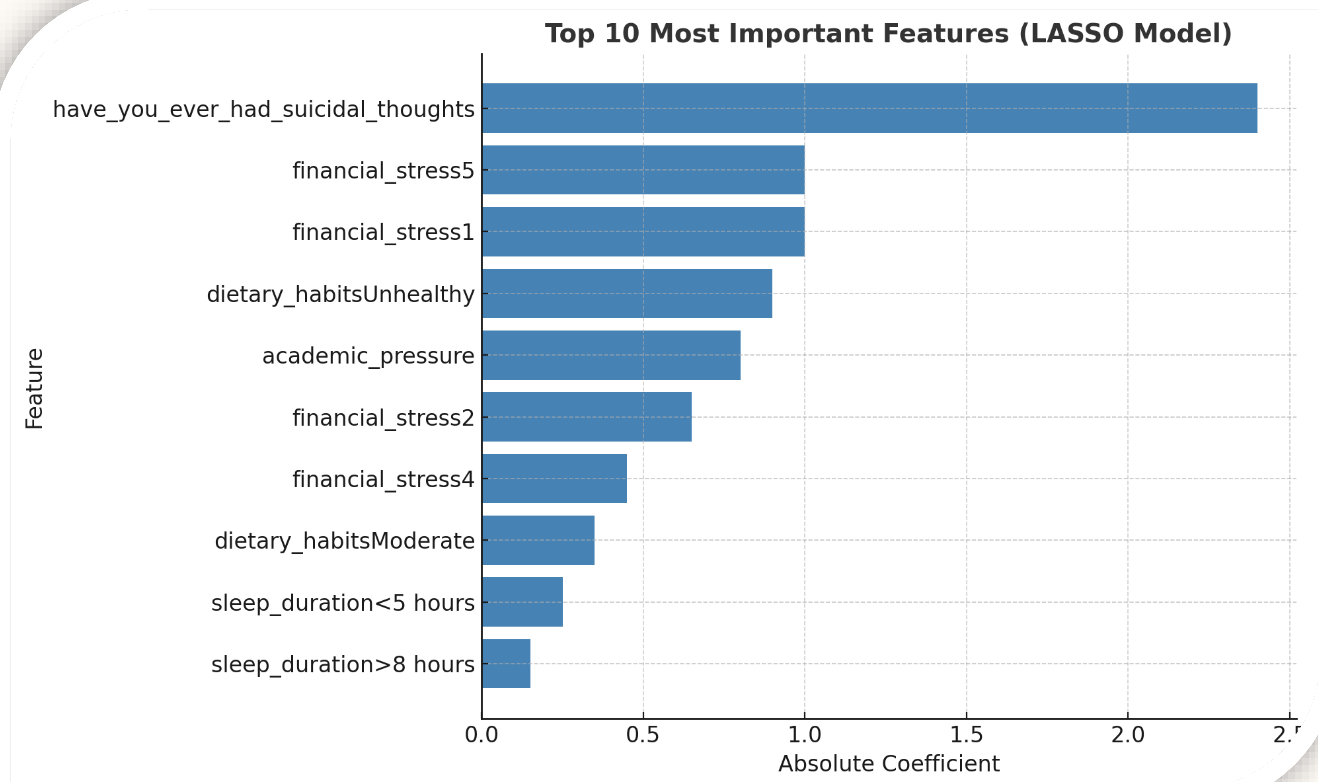


Figure: Top 10 predictors selected by LASSO Logistic Regression (absolute coefficient values).

Interpretation:

- ➢ Suicidal thoughts is the strongest predictor of depression risk.
- ➢ Financial stress (multiple categories) and academic pressure also show high importance.
- ➢ Sleep duration and dietary habits play a moderate but notable role.
- ➢ Larger coefficients indicate stronger influence on depression prediction.

STEP 7 – MODEL EVALUATION (LOGISTIC LASSO)

Work:

- Evaluated the trained Logistic LASSO model on the test dataset.
- Constructed confusion matrix and calculated key metrics:
  - ➢ Accuracy, Precision, Sensitivity (Recall), Specificity, F1-score.
- Stored results in combined CSV file for reproducibility.
- Identified False Positives (FP) and False Negatives (FN) separately.

Metric	Value
Accuracy	84.7%
Precision	85.7%
Sensitivity (Recall)	88.8%
Specificity	79.0%
F1-score	87.2%

Table: Model evaluation metrics for Logistic LASSO (test dataset). Metrics derived from confusion matrix and performance calculations

Output:

- Model evaluation metrics file→
  - ➢ model_evaluation_metrics_LASSO.txt
- Combined results file →
  - ➢ LASSO_results_combined.csv
- False Positive cases →
  - ➢ misclassified_FP_logistic_LASSO_2025.xlsx
- False Negative cases →
  - ➢ misclassified_FN_logistic_LASSO_2025.xlsx

Interpretation Summary:

- ➢ Overall accuracy ~ 84.7% (good performance).
- ➢ Sensitivity (Recall for depressed cases) ~ 0.888 → Model detects majority of depressed students.
- ➢ Specificity (Recall for non-depressed cases) ~ 0.790 → Slightly lower, model sometimes misclassifies non-depressed as depressed.
- ➢ Misclassification analysis shows FN and FP files capture the “error zones,” useful for understanding bias in predictions.

STEP 8 – CROSS-VALIDATED EVALUATION: GLM VS LASSO

- ❖ PART A : CROSS-VALIDATED EVALUATION (COMPARISON: GLM vs LASSO)
- ❖ PART B : CROSS-VALIDATED EVALUATION (PERFORMANCE VISUALIZATION)

PART A : CROSS-VALIDATED EVALUATION (COMPARISON: GLM VS LASSO)

Work:

- Performed 5-fold cross-validation on the dataset using both GLM (Generalized Linear Model) and LASSO Logistic Regression.
- Evaluated and compared performance across key metrics:
  - Accuracy, Sensitivity (Recall), Specificity, and AUC (Area Under Curve).
- Compiled fold-wise results into comparison tables for structured evaluation.
- Summarized findings into combined performance statistics to benchmark both models.

Output:

- cv_stats_glm_vs_lasso.xlsx → Detailed fold-wise CV statistics for GLM vs LASSO.

- cv_comparison_glm_vs_lasso.xlsx → Tabular comparison of average metrics between GLM and LASSO.

- cv_comparison_summary_glm_vs_lasso.xlsx → Final summary table highlighting cross-validated averages.

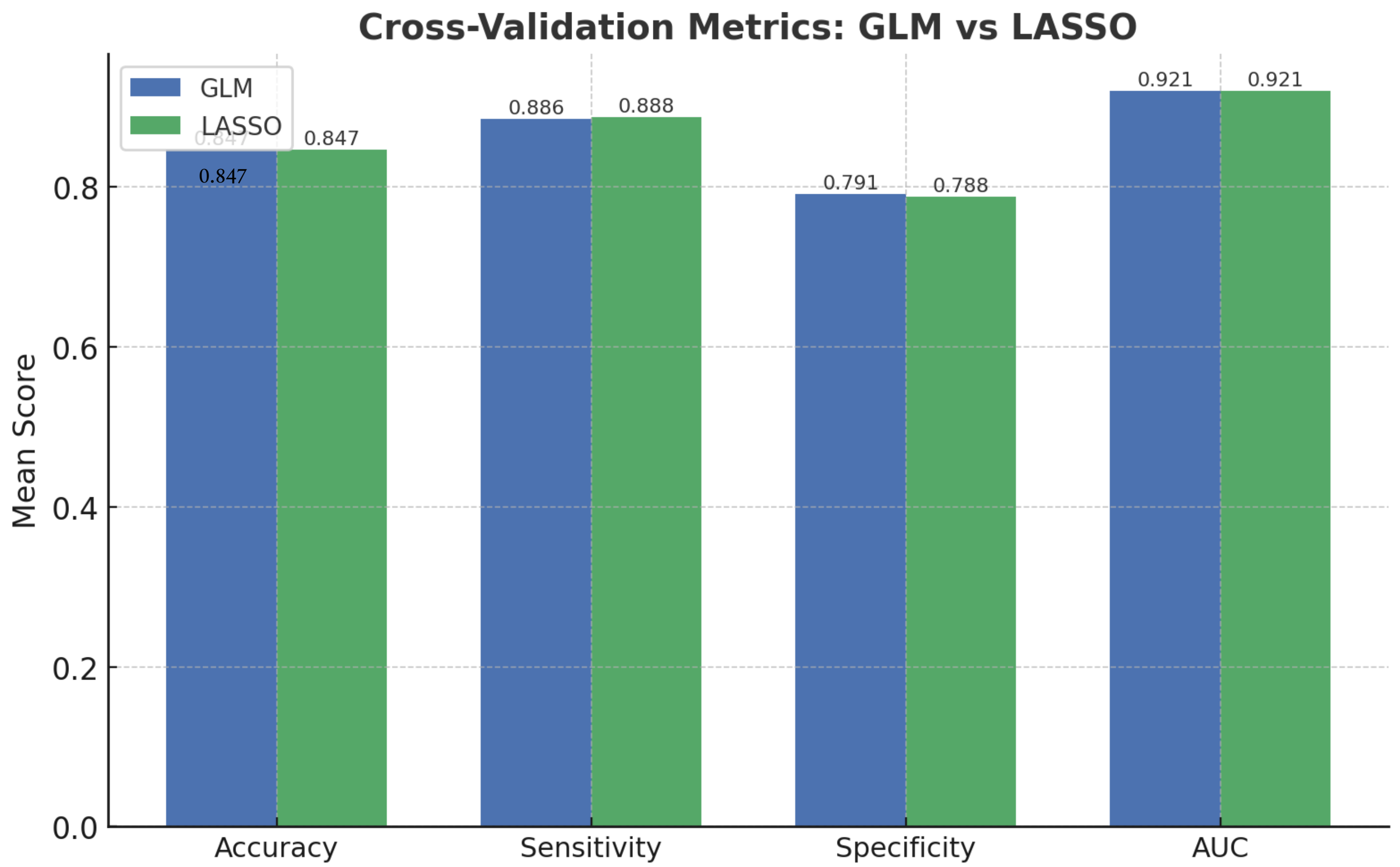
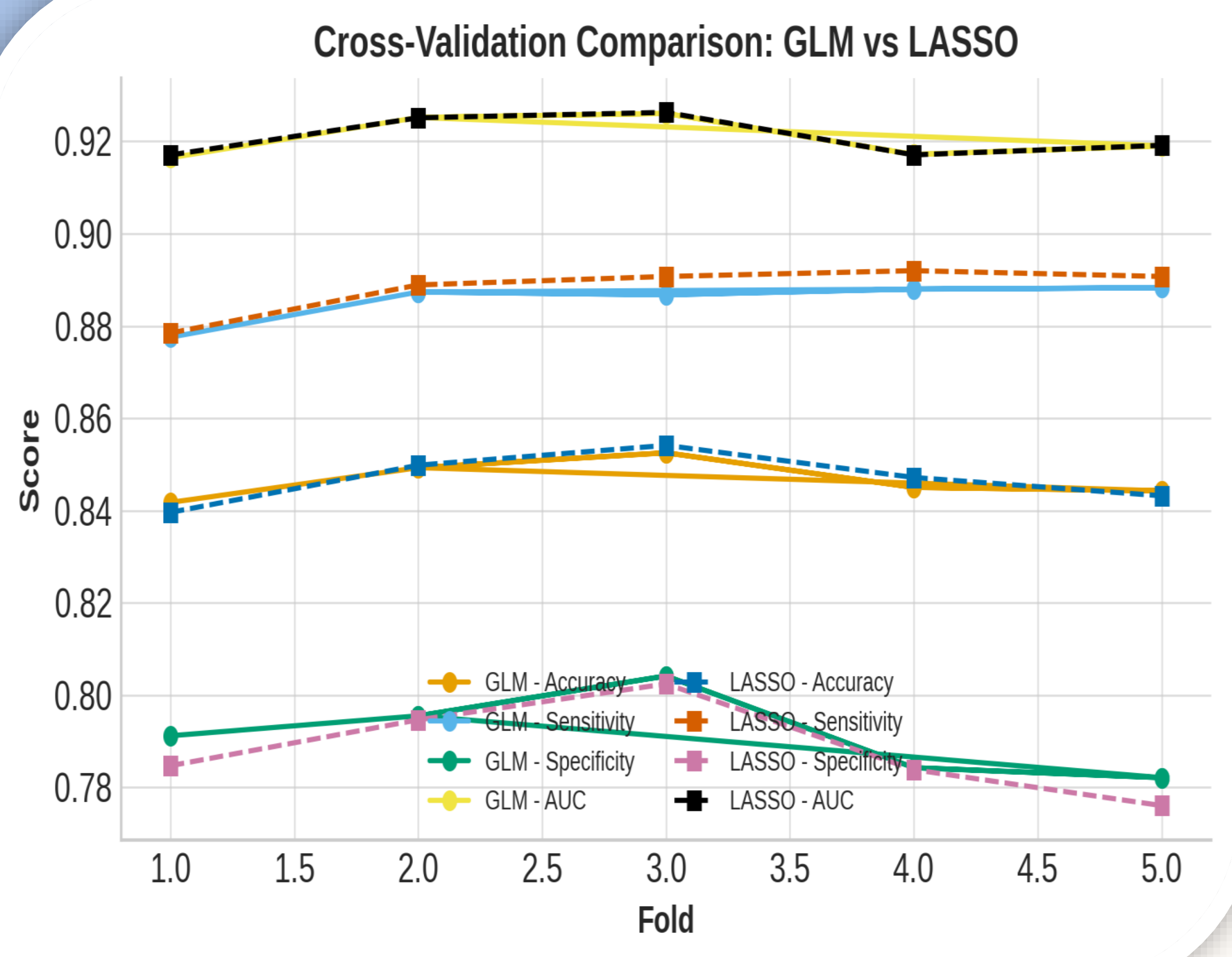
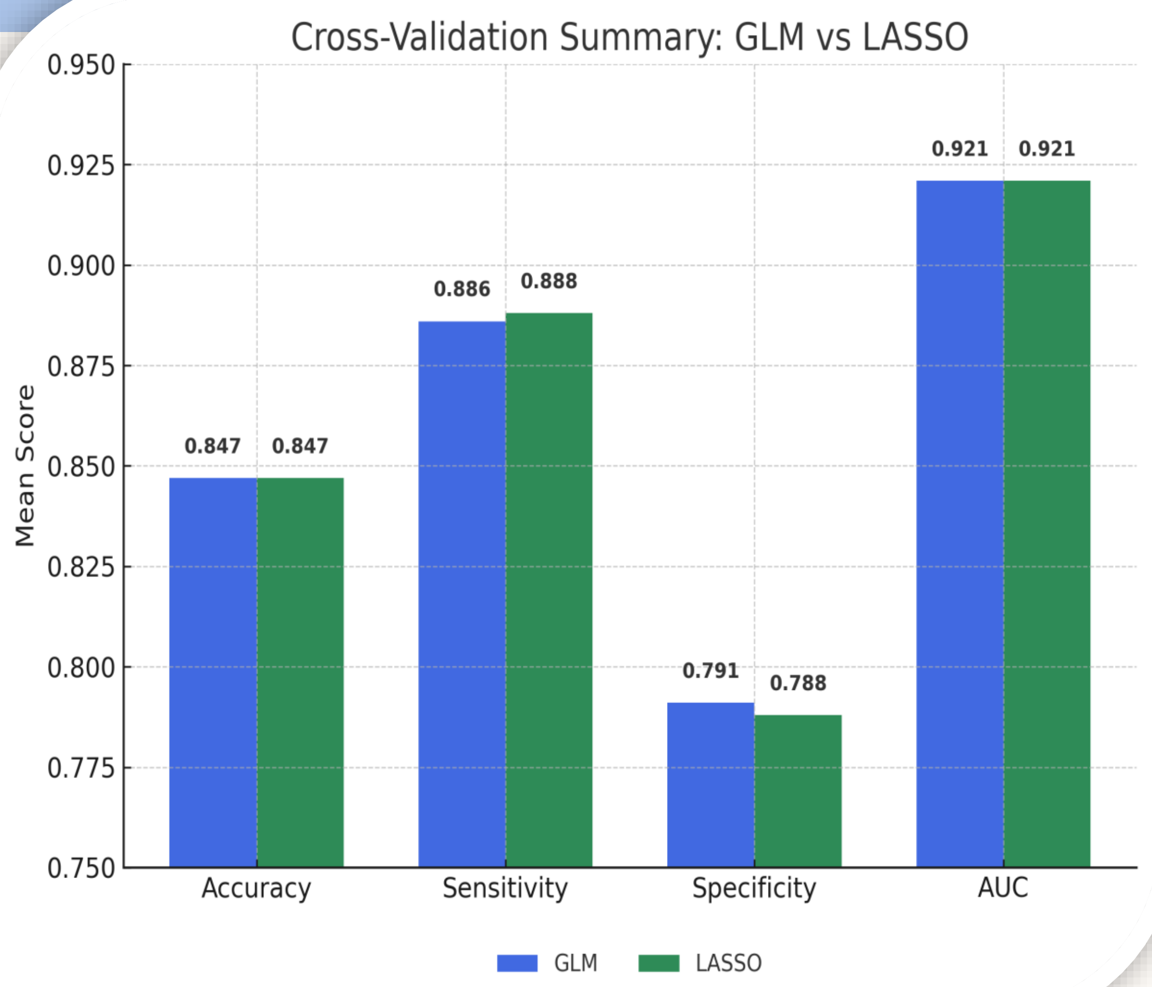


Figure: Comparison of mean cross-validation metrics between GLM and LASSO (5-fold CV). Values represent average performance across folds.

Interpretation:

LASSO shows slightly higher sensitivity, while ***GLM performs marginally better in specificity***. Overall, both models achieve very similar cross-validated performance.





Interpretation:

- Both GLM and LASSO show stable performance across 5 folds.
- Accuracy & AUC remain consistently high (~0.84–0.92), indicating reliability.
- Sensitivity is slightly better for LASSO → captures more depressed cases.
- Specificity is slightly better for GLM → fewer false positives.
- *Variability across folds is minimal → models are robust and not overfitting.*

Interpretation:

- **Accuracy:** Both GLM and LASSO ~84.7% (similar performance).
- **Sensitivity:** *LASSO* slightly higher → *better at detecting depressed cases*.
- **Specificity:** *GLM* slightly higher → *fewer false positives*.
- **AUC:** Both very strong (0.921) → excellent discrimination ability.

Overall: Both models perform almost equally, with LASSO favouring sensitivity and GLM favouring specificity.

PART B : CROSS-VALIDATED EVALUATION (PERFORMANCE VISUALIZATION)

Work:

- Performed 5-fold cross-validation for both GLM and LASSO models.
- Visualized fold-wise distributions using boxplots.
- Compared four key metrics: Accuracy, Sensitivity (Recall), Specificity, and AUC.
- Boxplots highlight variability across folds and model stability.

Output:

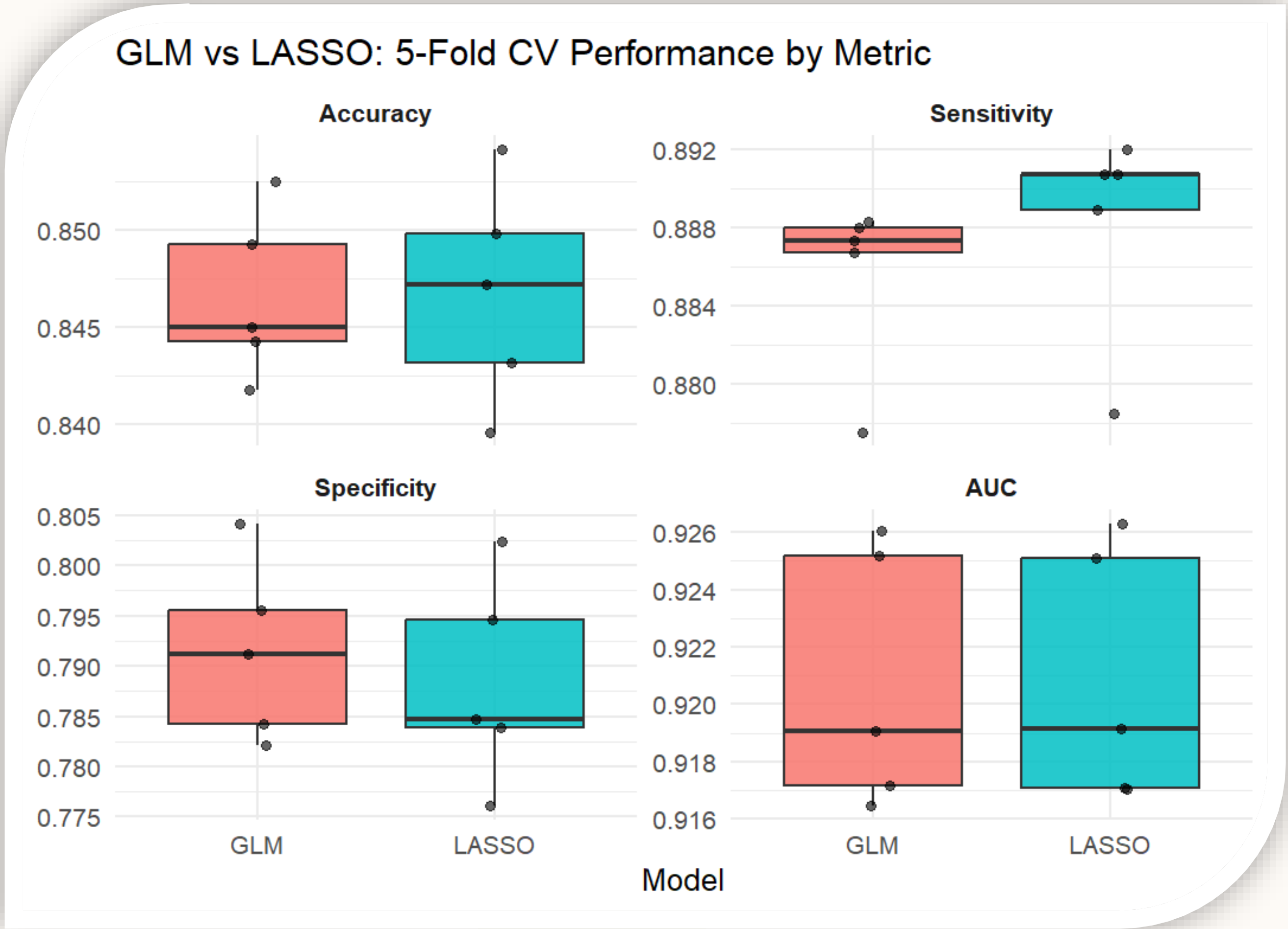
- Faceted Boxplot:

  Image:step8F_glm_vs_lasso_faceted_boxplot.png

- Combined view of Accuracy, Sensitivity, Specificity, and AUC for both models

Summary:

Both GLM and LASSO show comparable performance, with LASSO favoring sensitivity and GLM favoring specificity.



Interpretation:

- ➢ Accuracy & AUC: Nearly identical and consistently high (~0.84–0.92), indicating strong overall performance.
- ➢ Sensitivity: LASSO slightly higher → better at detecting depressed cases.
- ➢ Specificity: GLM slightly higher → fewer false positives.
- ➢ Overall: Both models are robust, with minimal variability across folds.

STEP 9 – GLM & LASSO INTERPRETABILITY

Work :

- Goal: Translate model outputs into interpretable evidence about which features increase or decrease depression risk.
- Dataset: ~27,901 students, 14 predictors (cleaned & standardized).
- Methods applied:
  - GLM (Logistic Regression): Produced Odds Ratios (OR, 95% CI, p-values).
  - LASSO Logistic Regression: Selected strongest predictors with direction (risk/protective).
- Comparison: Evaluated agreement vs disagreement between GLM and LASSO.
- Next step prep: Generated a shortlist (~16 predictors) for downstream validation.

OUTPUTS AND VISUALIZATION

Outputs (GLM Logistic Regression):

GLM (Logistic Regression):

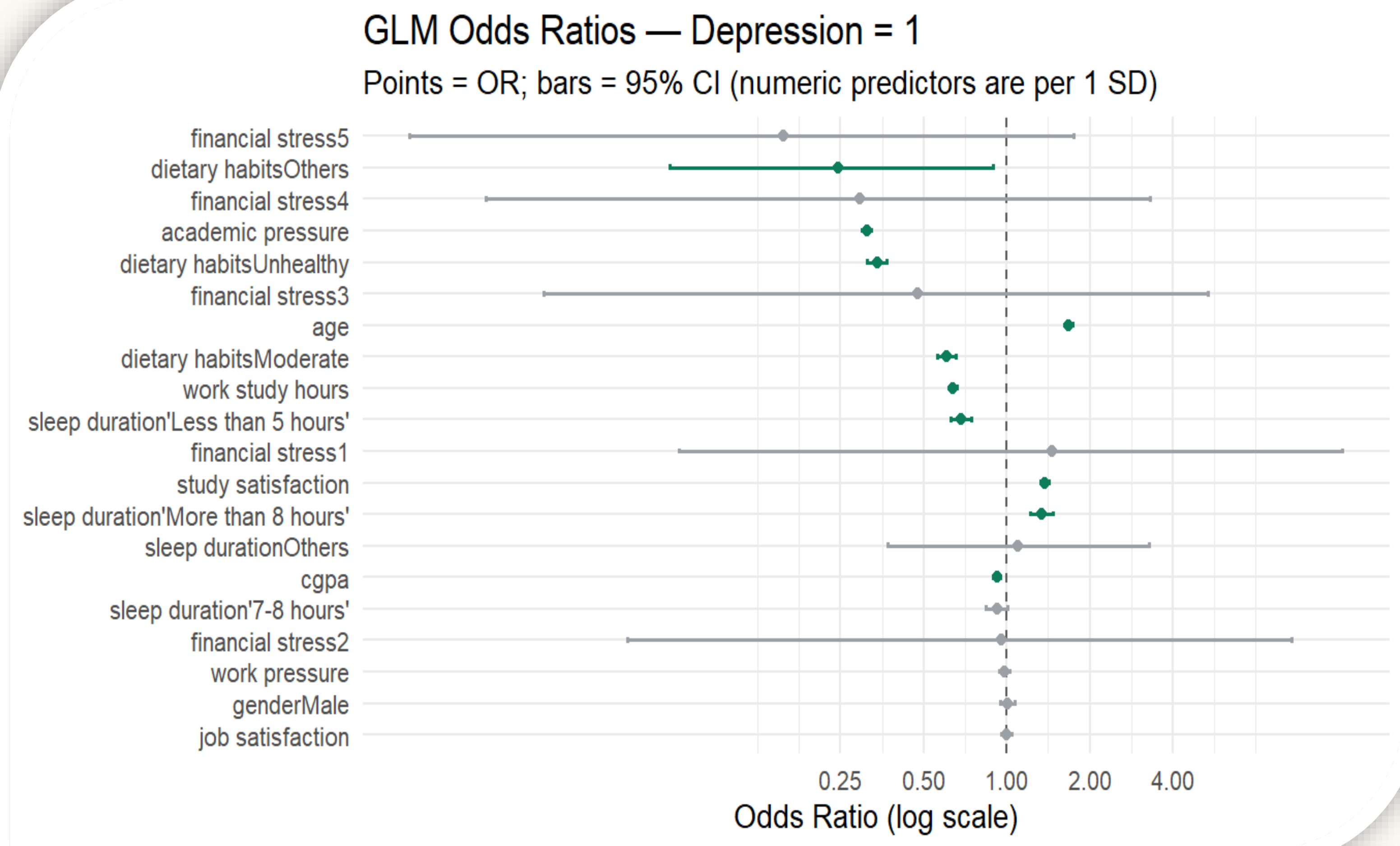
- ➢ Odds Ratios table (step9_glm_odds_ratios.xlsx)
- ➢ Forest plot of ORs with 95% CI (step9_glm_or_forest_white_.png)

Visualization:

- Forest Plot: Odds Ratios with 95% CI for key predictors.
- Excel Preview: Snapshot of OR table (variable, OR, 95% CI, p-value).

Interpretation:

- OR > 1 → Risk-increasing (e.g., age, financial stress, academic pressure).
- OR < 1 → Protective (e.g., 7–8 hrs sleep, study satisfaction).
- GLM highlights both lifestyle and psychosocial factors as significant predictors.



parameter	OR	CI_low	CI_high	p_value
financial_stress5	0.155	0.007	1.749	0.141
dietary_habitsOthers	0.244	0.06	0.893	0.036
financial_stress4	0.293	0.013	3.312	0.333
academic_pressure	0.311	0.3	0.322	0.0
dietary_habitsUnhealthy	0.339	0.313	0.367	0.0
financial_stress3	0.475	0.021	5.367	0.557

Outputs (LASSO Logistic Regression):

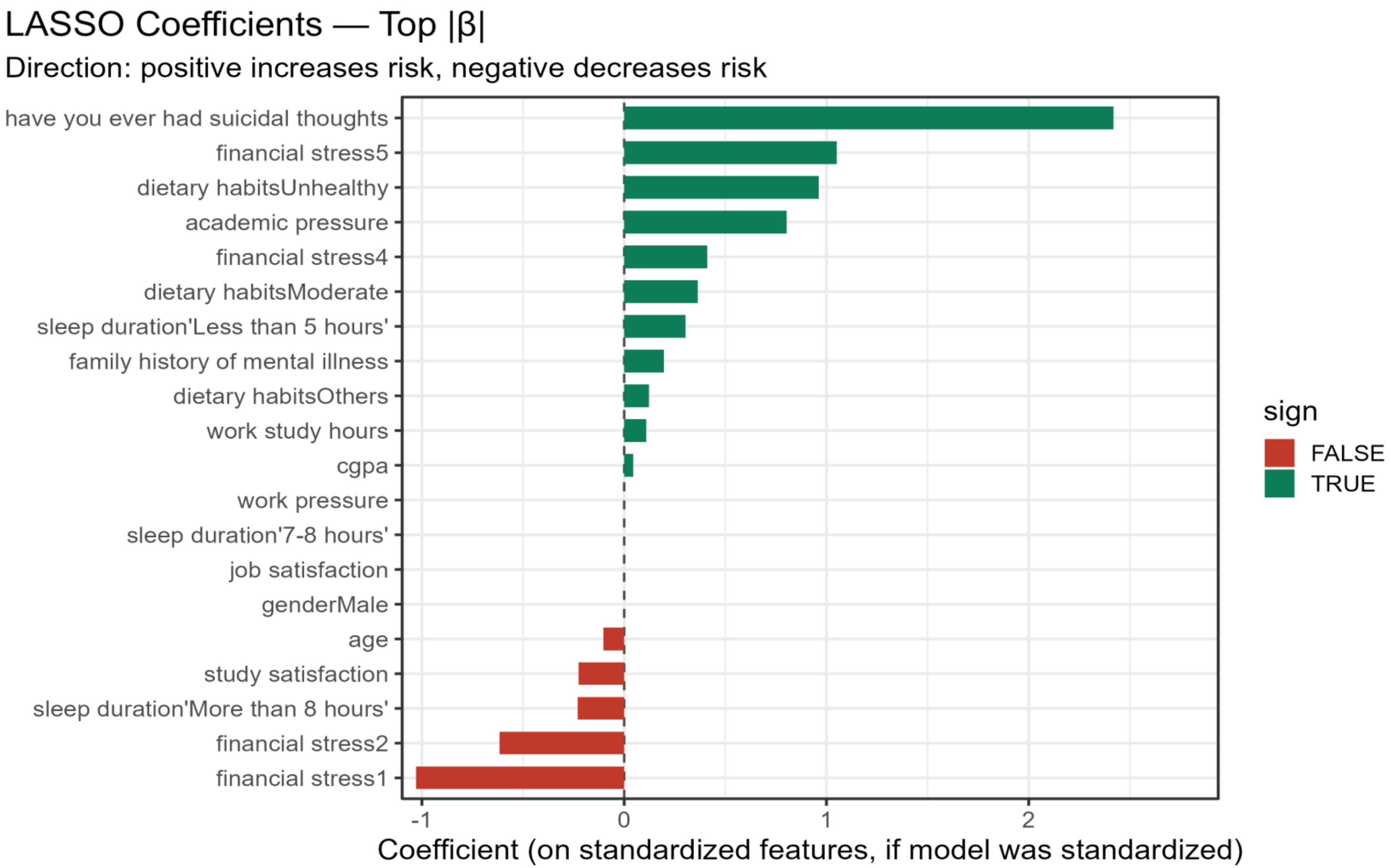
- Coefficient table → step9_lasso_coef_tidy.xlsx
- Bar plot of top coefficients → step9_lasso_coef_bar_wide.png

Visualization:

- Bar Chart (LASSO): Top coefficients (positive = risk-increasing, negative = protective).
- Excel Preview: Snapshot of coefficient table (first 6 rows).

Interpretation:

- Positive coefficients → Risk-increasing (e.g., suicidal thoughts, financial stress, unhealthy dietary habits).
- Negative coefficients → Protective (e.g., balanced sleep duration, satisfaction factors).
- LASSO highlights the strongest predictors and filters out weaker signals.



parameter	coef
have you ever had suicidal thoughts	2.42
financial stress5	1.052
financial stress1	-1.029
dietary habitsUnhealthy	0.963
academic pressure	0.804
financial stress2	-0.615

AGREEMENT ANALYSIS

Outputs Generated:

- GLM–LASSO comparison table → step9_glm_lasso_comparison.xlsx
- Scatter plot of agreement → step9_glm_lasso_agreement_scatter.png
- Summary counts → step9_model_agreement_summary.xlsx
- Disagreements list → step9_direction_disagreements.xlsx

Visualization:

- Scatter Plot: GLM log(OR) vs LASSO coefficient
  - ➢ Green = agreement in direction
  - ➢ Red = disagreement
- Excel Preview: Snapshot of summary table (agreement counts, disagreements).

Interpretation:

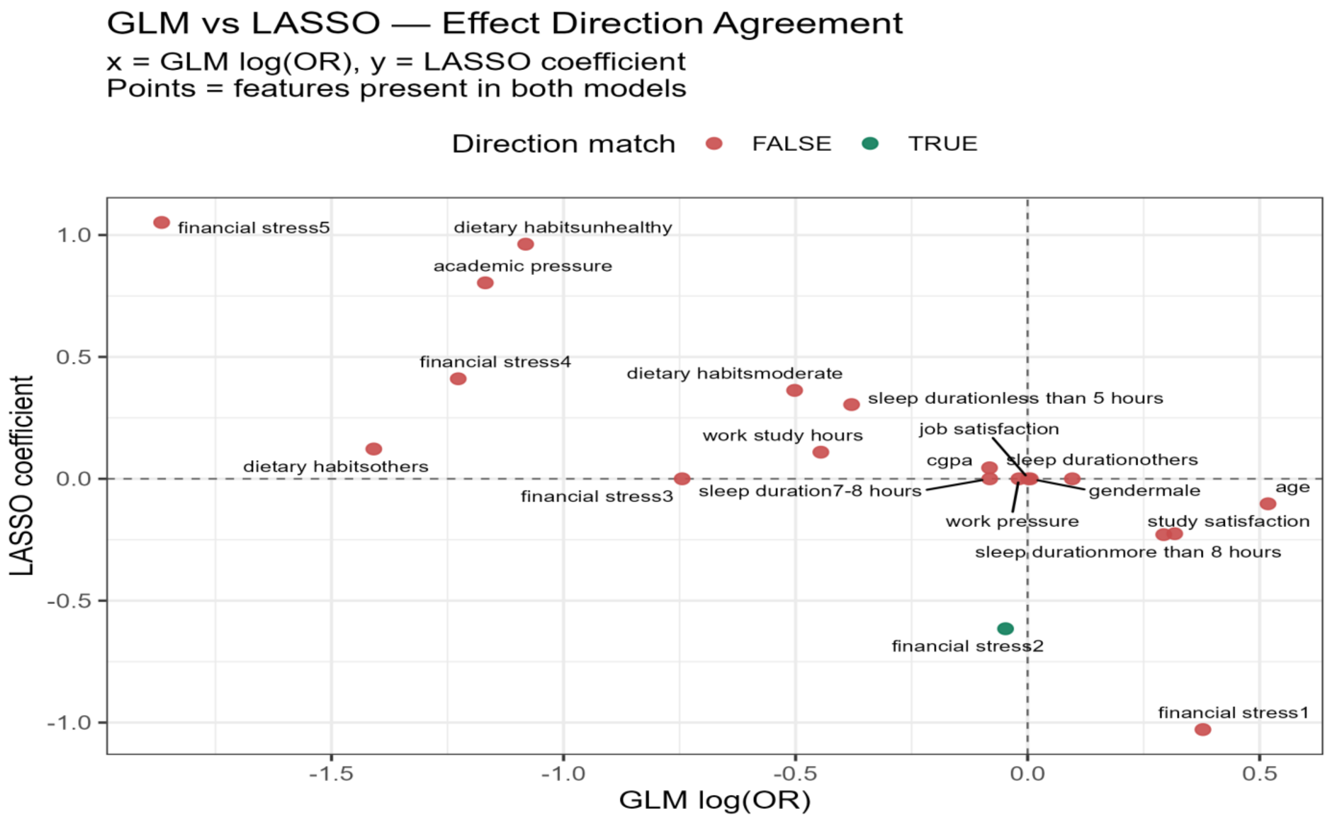
➢ Agreement (1 predictor): Both models match direction → strongest evidence.

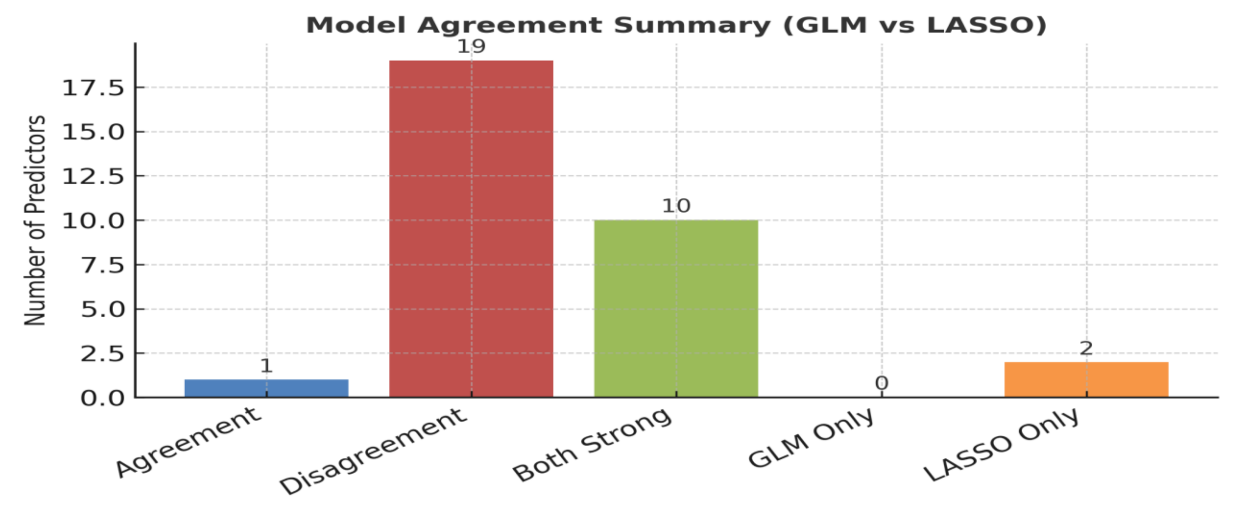
➢ Disagreement (19 predictors): Models give opposite directions; likely due to collinearity or dataset noise.

➢ Both Strong (10 predictors): Predictors consistently important across both models.

➢ GLM Only (0), LASSO Only (2): Minimal exclusive selections; shows good overlap.

➢ Key Takeaway: Despite disagreements, the overlap of strong predictors gives us a reliable shortlist for Step 10 validation





Source : step9_model_agreement_summary.csv

Parameter	GLM Direction	LASSO Direction	Agreement	Abs(GLM Effect)	Abs(LASSO Effect)
financial stress5	↓ risk	↑ risk	False	1.866	1.052
dietary habitsunhealthy	↓ risk	↑ risk	False	1.082	0.963
academic pressure	↓ risk	↑ risk	False	1.169	0.804
financial stress4	↓ risk	↑ risk	False	1.227	0.41
dietary habitsothers	↓ risk	↑ risk	False	1.409	0.122
financial stress1	↑ risk	↓ risk	False	0.378	1.029
dietary habitsmoderate	↓ risk	↑ risk	False	0.502	0.363
sleep durationless than 5 hours	↓ risk	↑ risk	False	0.379	0.304

Source: step9_direction_disagreements.xlsx (Step 9 Outputs – Interpretability)

Interpretation:

- **Disagreements occur due to collinearity and model penalty differences between GLM and LASSO.**
- **Abs(GLM Effect) and Abs(LASSO Effect) show the strength of predictor influence in each model (ignoring direction).**
- **Differences in these values help explain why some predictors show directional disagreements between GLM and LASSO.**
- **Example: Financial stress5 has Abs(GLM Effect = 1.866, LASSO Effect = 1.052) → both models say it's influential, but GLM gives it higher weight.**
- **Example: Dietary habits. Others has GLM stronger (1.409) vs LASSO weaker (0.122) → GLM finds it more important, but LASSO doesn't.**

DELIVERABLES & SUMMARY OF PROJECT

Key Outputs :

- GLM & LASSO used for predictor interpretability.
- Odds Ratios (GLM) and Coefficients (LASSO) identified.
- Agreement & Disagreement analysis performed.

SUMMARY :

- Completed a comparative modeling and interpretability study using GLM and LASSO.
- Demonstrated stable performance and clear trade-offs between models.
- Identified a robust set of predictors through convergent interpretability analysis.