Supplementary Material for: Machine learning-based nomogram for predicting depressive symptoms in women: a cross-sectional study in Guangdong province, China

Jia-Min Chen, Mei Rao, Yu-Ting Wei, Qiong-Gui Zhou, Jun-Long Tao, Shi-Bin Wang, Bo Bi

S1. Measurements of Variables

Body mass index (BMI) is calculated as weight/height (kg/m^2). We grouped adults as underweight with BMI < 18.5 kg/m^2, normal with BMI < 24.0 kg/m^2 but ≥ 18.5, overweight with BMI < 28.0 kg/m^2 but ≥ 24.0 kg/m^2, and obese with BMI ≥ 28 kg/m^2, based on the reference standard of Chinese BMI[1].

Chronic diseases were identified through a questionnaire: "Have you ever been informed by a doctor, nurse, or other health professional that you have a chronic disease?" Participants were given a table listing various chronic diseases, which investigators read aloud in the local dialect to aid recall. Diagnosed conditions were classified according to the International Classification of Diseases, 10th revision. They included hypertension, diabetes, ischemic heart disease, cerebrovascular disease, chronic obstructive pulmonary disease, cancer, hyperlipidemia, arthritis, chronic low back pain, anemia, liver diseases, chronic gastroenteritis, cataract/glaucoma, gout, Parkinson's disease, chronic nephritis, and urolithiasis, epilepsy, and tuberculosis.

The PSQI sleep-related variables were categorized as follows: sleep latency (≤ 15 minutes, 16 – 30 minutes, 31 – 60 minutes, > 60 minutes); sleep duration (≥ 7 hours, 6 – 6.99 hours, 5 – 5.99 hours, < 5 hours); and sleep efficiency (> 85%, 75 – 84%, 65 – 74%, < 65%), according to established classification criteria[2–4].

Social-demographic and lifestyle factors were gathered: place of residence (urban or rural), marital status (single/never married, married, or divorced/widowed/separated), and level of education (≤ 9 years, 9 to 15 years, or ≥ 16 years). Furthermore, the average monthly income was classified as follows: low (< 3500 yuan/month), medium (3500 to 6000 yuan/month), and high income (> 6000 yuan/month). A current smoker was described as smoking at least one cigarette per day and for six months or longer. Current alcohol drinker was designated as consuming on average one time or more of standard alcoholic drinks a week. Tea-drinking habits were delimited as a person who drank tea at least four times per week. Physical exercise frequency was categorized as hardly exercise, 1 to 3 times per month, 1 to 2 times per week and 3 or more times per week.

S2. Hyperparameter Grid

In this study, hyperparameter optimization for the LightGBM, Support Vector Machine (SVM), and XGBoost models was conducted using a systematic grid search approach. Grid search is an exhaustive method that evaluates all predefined hyperparameter combinations within a specified search space to identify the configuration yielding the highest performance on validation data. For each model, the hyperparameter search space was defined based on common practices in machine learning literature to ensure methodological rigor and reproducibility.

For LightGBM, the search space included five key hyperparameters: n_estimators (100, 200, 300, 400, 500), learning_rate (0.001, 0.01, 0.1), max_depth (5, 10, 15, 20), num_leaves (20, 31, 40, 50), and min_child_samples (10, 20, 30, 40). These ranges were selected to balance model complexity and generalization, as excessively large values for num_leaves or max_depth could lead to overfitting, while smaller values might underfit the data. For SVM, the search grid covered C (0.1, 1, 10, 100), kernel ('rbf', 'linear'), and gamma ('scale', 'auto'), reflecting typical configurations for balancing margin width and classification error. The XGBoost parameters included n_estimators (50, 100, 200, 300), learning_rate (0.01, 0.1, 0.2, 0.3), max_depth (3, 5, 7, 9), subsample (0.8, 0.9, 1.0), and colsample_bytree (0.8, 0.9, 1.0), which are widely recommended to control tree growth and feature sampling.

The evaluation process utilized 5-fold cross-validation on the training dataset to mitigate overfitting and ensure robustness. Each hyperparameter combination was trained on four folds and validated on the remaining fold, with performance metrics (e.g., accuracy or F1-score) averaged across all folds. The optimal hyperparameters were selected based on the highest mean cross-validation score. For instance, LightGBM achieved peak performance at n_estimators=200, learning_rate=0.01, max_depth=10, num_leaves=31, and min_child_samples=30, while SVM performed best with C=1.0, kernel='rbf', and gamma='scale'. XGBoost's optimal configuration included n_estimators=100, learning_rate=0.1, max_depth=3, subsample=1.0, and colsample_bytree=1.0.

After identifying the best hyperparameters, final models were retrained on the entire training dataset and evaluated on an independent test set.

S3. SHAP-based Feature Selection Comparison

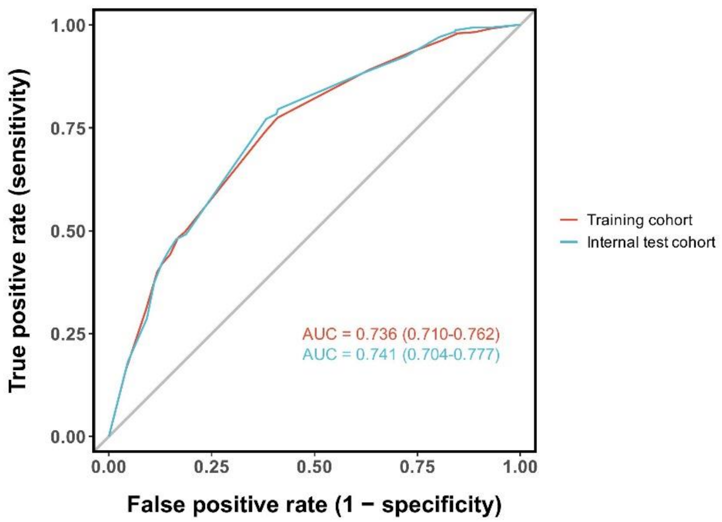

Figure 1 The receiver operating characteristic (ROC) curves evaluating the predictive performance of nomograms incorporating the top three SHAP-ranked features.

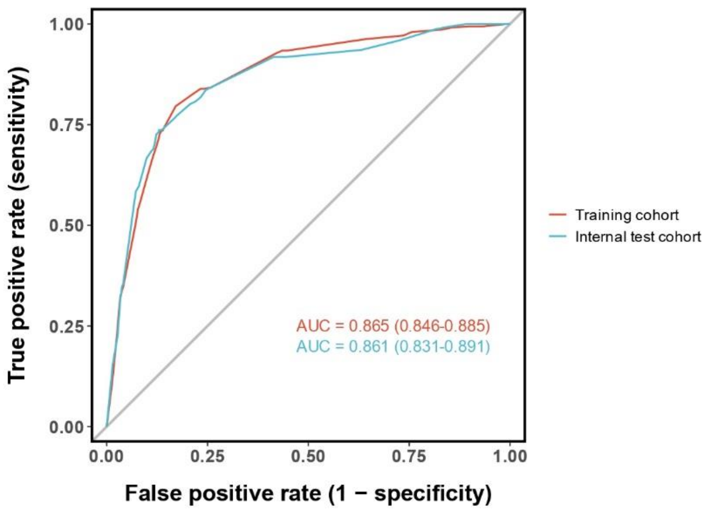

Figure 2 The receiver operating characteristic (ROC) curves evaluating the predictive performance of nomograms incorporating the top four SHAP-ranked features.

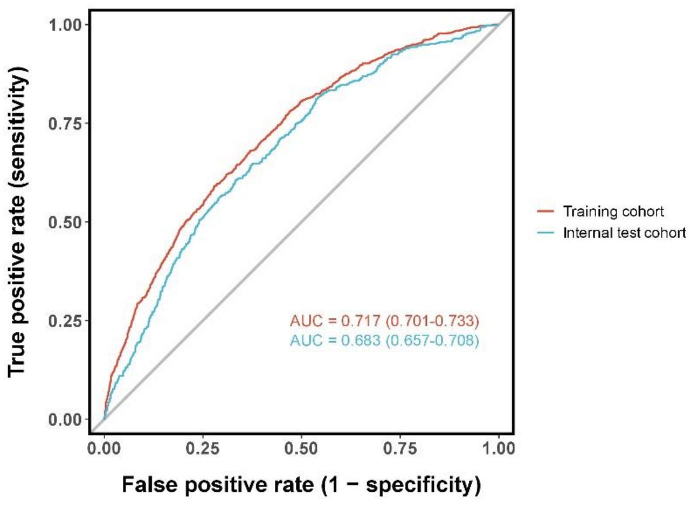

Figure 3 The receiver operating characteristic (ROC) curves evaluating the predictive performance of nomograms incorporating 16 variables SHAP-ranked features.

S4. Interactive Digital Version of Nomogram

https://hnmc.shinyapps.io/MyDynNomApp/