Model Selection and Rationale:

I selected logistic regression as the baseline model to predict diabetes risk among Texas adults, given the binary nature of the target variable ("Diabetes_binary"). Logistic regression was chosen for its interpretability and ability to communicate meaningful health insights to stakeholders in public health, such as the Texas Department of State Health Services. The model provides transparency into how each predictor—like general health, BMI, or access barriers—contributes to diabetes risk, aligning well with the communication and decision-making needs of health officials. Since the dataset includes both categorical and numeric data, logistic regression fits well when features are appropriately preprocessed and scaled.

Model Training and Testing:

The dataset, sourced from the Behavioral Risk Factor Surveillance System (BRFSS), included over 253,000 adult responses. I used an 80/20 stratified train-test split to ensure the class distribution (approx. 14% diabetes) was preserved. Categorical variables (e.g., Sex, Education) were encoded, and numeric variables such as BMI were scaled. The model also incorporated engineered features such as:

- Chronic_Risk_Load – a sum of comorbid conditions like stroke, high cholesterol, and heart disease
- Healthcare_Barrier_Index – a composite measure of access issues, including insurance status and healthcare costs
- Binned versions of mental and physical health days to reduce skew and highlight health burdens

The logistic regression model was trained on this feature set to identify adults with a predicted diabetes risk ≥30% within five years.

Performance Evaluation:

The model produced the following results on the test data:

- Accuracy: 82%
- Precision: 0.69
- Recall: 0.56
- F1 Score: 0.62
- ROC AUC: 0.79

While the overall accuracy is strong, recall is a key concern in this case. Since the primary goal is to proactively identify individuals at risk of developing diabetes, the model's

moderate recall suggests that some at-risk individuals may still go undetected. Improving recall without dramatically reducing precision will be the next area of focus.

Result Interpretation:

From a public health standpoint, failing to identify someone truly at risk (false negatives) has greater consequences than false alarms. Therefore, recall should be prioritized—even at the expense of some precision. The model's interpretability has already helped pinpoint key health and behavioral predictors (e.g., BMI, self-rated general health, chronic conditions), which are actionable and relatable to public health programs. Still, improvements in sensitivity will be necessary to better support preventive outreach strategies.

AI Feedback:

I used DeepSeek.com to assess my modeling strategy and variable engineering. The AI tool confirmed that logistic regression was appropriate for the first iteration due to its simplicity and transparency. It also recommended:

- Trying Random Forests to capture nonlinear relationships
- Using SMOTE resampling or weighted loss functions to handle class imbalance more effectively
- Verifying that engineered features like Chronic_Risk_Load contribute meaningfully to model performance and remain interpretable

This feedback validated my current path while suggesting thoughtful next steps to refine the model.

What I'd Like Feedback On:

I'd appreciate peer thoughts on the following:

- Given the moderate recall, should I move directly to more complex models like Random Forest, or fine-tune the logistic regression further (e.g., threshold tuning or class weighting)?
- Are there other engineered features—perhaps domain-specific—that you'd recommend for improving the model's sensitivity to risk?
- Is my tradeoff between recall and interpretability reasonable for a public health project, or should I explore even deeper models now?

Looking forward to your thoughts and any additional strategies you've used to improve health-related classification models.