7/26/2025

Diabetes Risk Prediction

Baseline Modeling

Kungulio, Seif H.

DATA 650: CAPSTONE PROJECT

Establishing a Baseline Model to Predict Diabetes Risk

Introduction

Establishing a baseline model is a critical early milestone in predictive analytics. It helps evaluate how well current data and features can identify individuals at risk and serves as a benchmark for assessing future model improvements. For this project, the objective of public health is to predict which Texas adults have a 30% or higher likelihood of developing diabetes within five years, enabling proactive outreach by agencies like the Texas Department of State Health Services.

Model Selection Justification

The target variable, “Diabetes_binary”, is binary (0 = no diabetes, 1 = diagnosed diabetes). Given its interpretability, transparency, and suitability for categorical outcomes, logistic regression was selected as the baseline model. It allows for easy communication with stakeholders like public health directors and offers insight into how each predictor contributes to diabetes risk.

Train and Test the Model

The dataset—comprising over 253,000 responses—was split into training (80%) and testing (20%) sets using a random stratified method to preserve the class distribution of the target variable. The logistic regression model was trained on the following preprocessed features:

- **Demographic:** Age group, Sex, Income, Education
- **Clinical/Health:** BMI (scaled), GenHlth, DiffWalk
- **Behavioral:** Smoking, Physical activity, Alcohol consumption
- **Engineered Features:**
  - **“Chronic_Risk_Load”** (sum of high blood pressure, high cholesterol, stroke, heart disease/attack)
  - “**Healthcare_Barrier_Index**” (access barriers like no insurance or high cost)
  - Binned “MentHlth” and “PhysHlth”

Performance Evaluation

The logistic regression model produced the following metrics on the test set:

- Accuracy: 0.82
- Precision: 0.69
- Recall: 0.56
- F1 Score: 0.62
- ROC AUC: 0.79

These results suggest strong baseline performance, particularly in overall classification accuracy. However, the recall score indicates that the model misses a portion of true diabetes cases, which is critical for the project's goal of proactive intervention.

Interpret the Results

From a public health lens, recall (sensitivity) is more important than precision. While the baseline model is solid overall, missing individuals who are truly at risk of diabetes limits its effectiveness for preventive care. Many of the retained features—such as BMI, general health, and chronic conditions—were found to be statistically significant contributors to prediction of diabetes.

Use AI to Review Your Work

I consulted an AI tool (DeepSeek.com) to evaluate modeling choices and assumptions. Key prompts included:

- Is logistic regression suitable given the project's public health goals?
- Should other models (e.g., decision trees, random forests) be considered to improve recall?
- Are evaluation metrics appropriate given class imbalance (only \~14% diabetic)?
- Are the engineered features appropriate for this domain?

Summarize the AI Feedback

The AI confirmed that logistic regression is a good starting point and that recall is the metric most aligned with the public health use case. It suggested testing tree-based models like Random Forests to potentially boost sensitivity. The tool also recommended refining how class imbalance is handled—e.g., with SMOTE resampling or weighted loss functions—and emphasized validating engineered features for both interpretability and predictive contribution.

Conclusion

This modeling phase confirms that logistic regression provides a strong, interpretable foundation for predicting diabetes risk using BRFSS data. However, recall remains an area for improvement. Future work will explore ensemble and tree-based models to improve sensitivity without compromising interpretability. Incorporating AI feedback early also helped refine variable choices and confirm that our modeling direction aligns with public health goals, especially for targeted interventions across high-risk communities.

References

- American Diabetes Association. (2023). The economic costs of diabetes in the U.S. https://www.diabetes.org/resources/statistics/cost-diabetes
- Centers for Disease Control and Prevention. (2024). Behavioral Risk Factor Surveillance System survey data. U.S. Department of Health and Human Services. https://www.cdc.gov/brfss/index.html
- Harvard T.H. Chan School of Public Health. (2021). Predictive analytics in chronic disease management. https://www.hsph.harvard.edu
- OpenAI. (2025). AI-powered review of data preparation and exploration. Internal tool use only.
- Texas Department of State Health Services. (2023). Diabetes in Texas. https://www.dshs.texas.gov