8/2/2025

Diabetes Risk Prediction

Alternative Modeling

Kungulio, Seif H.
DATA 650: CAPSTONE PROJECT

Introduction

Comparing predictive models is essential when selecting an approach that not only delivers strong performance metrics but also aligns with public health priorities. For this project, the objective is to enhance early identification of adults at high risk of developing diabetes in Texas using behavioral and clinical survey data. After establishing a baseline with logistic regression, I trained and evaluated additional models to assess whether improved sensitivity and generalization could better support outreach strategies. Balancing recall and interpretability were prioritized to maximize public health impact.

Model Comparison Results

Two models—Logistic Regression and Random Forest—were compared using identical features and preprocessing. Evaluation metrics included accuracy, precision, recall, F1-score, and ROC AUC.

Performance Comparison

Metric	Logistic Regression	Random Forest
Accuracy	0.82	0.85
Precision	0.69	0.72
Recall	0.56	0.66
F1-Score	0.62	0.69
ROC AUC	0.79	0.84

Evaluation and Interpret Results

The Random Forest model consistently outperformed Logistic Regression across all evaluation metrics, particularly in recall—a key priority for this use case. Missing at-risk individuals diminishes the value of a predictive model in public health, where proactive intervention is essential.

While Logistic Regression offers transparency and ease of explanation, the improvement in F1-score and ROC AUC with Random Forest justifies its selection for deployment. Random Forest's strength lies in its ability to model non-linear interactions among features like BMI, general health, and chronic risk load without overfitting, making it more adaptable for complex population-level datasets.

Hyperparameter Tuning

To enhance the Random Forest model, I implemented randomized search over key parameters:

- `n_estimators`: Number of trees in the forest
- `max_depth`: Maximum tree depth
- `min_samples_split`: Minimum samples to split an internal node
- `max_features`: Number of features considered at each split

The tuning process increased recall by ~3% and improved model robustness. Final parameters were selected to prioritize generalizability and interpretability, especially for stakeholders with limited data science expertise.

Generalization Assessment

Generalization was evaluated using:

- 5-fold cross-validation
- Holdout test set comparison
- ROC curve stability across folds

The Random Forest model showed stable recall performance across folds (standard deviation < 0.03) and minimal drop in test performance, confirming that it generalizes well beyond the training data. These results support its suitability for identifying high-risk individuals in unseen populations.

AI Review Summary

I consulted an AI assistant (ChatGPT) to review the modeling decisions. Key questions included:

- Does this model align with public health goals?
- Are the metrics selected appropriate given the 14% class imbalance?
- Are there simpler or more effective tuning or sampling methods?
- Should oversampling be applied now or deferred?

AI Recommendations

- Clarify priority of recall: Explain its importance over accuracy in outreach-focused campaigns.
- Test other tree-based models: Explore XGBoost or LightGBM for further gains.
- Monitor class imbalance: Evaluate whether synthetic oversampling (e.g., SMOTE) might improve sensitivity.
- Expand ensemble strategies: Combine models for potential performance lift.

Summarize the AI Feedback

Based on the AI feedback:

- I reinforced why recall is prioritized: Missing true positives weakens the case for targeted intervention.
- I flagged XGBoost as a next step for future iterations.
- I chose to defer oversampling: While class imbalance exists, the current model performs well without it. Introducing synthetic data may be considered later to boost sensitivity if needed.

Conclusion

The Random Forest model is the recommended approach for predicting diabetes risk among Texas adults. Its improved sensitivity and generalization outperform the logistic regression baseline, and its ability to identify individuals with high probability of developing diabetes supports early intervention goals.

This phase validated that advanced modeling techniques can complement domain-driven features like Chronic Risk Load and Healthcare Barrier Index. Moving forward, I plan to explore boosting methods and test whether hybrid ensemble models can deliver further performance gains without compromising interpretability.

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. 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 tools use only.