7/20/2025

Diabetes Risk Prediction

Data Preparation

Kungulio, Seif H.
DATA 650: CAPSTONE PROJECT

Improving Predictive Power Through Feature Engineering and Selection

Introduction

Transforming and engineering features are essential steps in building a high-performing and interpretable machine learning model. These processes help improve consistency, emphasize meaningful relationships, and align data with the intended analytical or business objectives. For the Diabetes Health Indicators dataset, thoughtful feature preparation enhances our ability to predict individuals with a higher likelihood of having diabetes, supporting both public health initiatives and personalized intervention strategies.

Initial Transformation and Engineering Plan

Before incorporating AI feedback, the transformation strategy focused on:

- **Encoding Binary Variables:** Variables such as “HighBP”, “HighChol”, “Smoker”, and “PhysActivity” are already binary (0/1). These were retained as-is but reviewed for clarity and labeling.
- **Binning Age:** Age (originally a categorical bracket from 1 to 13) was transformed into broader life stages (e.g., 18–34, 35–54, 55+), improving interpretability while minimizing granularity loss.
- **Scaling Numerical Features:** Continuous variables such as “BMI”, “MentHlth”, and “PhysHlth” were standardized to support algorithms sensitive to feature magnitude (e.g., logistic regression, SVM).
- **Re-labeling Ordinal Features:** Variables like “GenHlth”, “Education”, and “Income”, though numeric, represent ordered categories. These were relabeled with meaningful descriptors to support interpretation and feature selection.

This transformation plan aimed to enhance data quality, reduce modeling complexity, and promote clearer downstream insights.

Feature Engineering

Two engineered features were introduced based on domain relevance and exploration trends:

- **Chronic Risk Load**
    - **Logic:** Combining conditions such as “HighBP”, “HighChol”, “Stroke”, and “HeartDiseaseorAttack” captures underlying cardiovascular/metabolic risk.
    - **Construction:** Summed the binary presence of these four conditions for everyone.
    - **Impact:** Provides a composite health risk score that correlates with diabetes likelihood and facilitates stratified risk modeling
- **Healthcare Barrier Index**
    - **Logic:** Indicators like lack of insurance (“AnyHealthcare”) or unaffordability (“NoDocbcCost”) influence preventive care and early diagnosis.
    - **Construction:** Combined these binary features into an index of healthcare accessibility (e.g., 0 = no barrier, 1 = one barrier, 2 = both barriers).
    - **Impact:** Highlights socioeconomic and systemic barriers that may be critical for population health interventions.

Feature Evaluation and Selection

The evaluation strategy followed three principles:

- **Correlation Checks:** Variables such as “MentHlth” and “PhysHlth” showed moderate correlation but captured distinct dimensions of wellbeing, so both were retained.
- **Variance Assessment:** Features with minimal variance (e.g., “CholCheck” with nearly all entries being 1) were flagged for potential exclusion based on model performance.
- **Domain Alignment:** Variables with direct relevance to diabetes etiology and social determinants (e.g., “BMI”, “Income”, “Education”, “DiffWalk”) were prioritized regardless of immediate statistical correlation.

As a result:

- The original “CholCheck” variable was flagged for limited utility due to class imbalance.
- Engineered features were retained due to their strong theoretical justification and potential for improved discrimination in modeling.

Use AI to Review My Work

An AI assistant (chatGPT) was consulted to evaluate the completeness and appropriateness of the feature preparation steps. Prompts used included:

- Should BMI, MentHlth, or PhysHlth be transformed or binned to reduce skewness?
- Are any variables redundant due to high correlation?
- Do the engineered features align with clinical logic and business objectives?
- Should ordinal variables be treated as numeric or recoded as factors?

Summarize the AI Feedback

Key AI feedback included:

- Binning Mental and Physical Health: Suggested binning "MentHlth" and "PhysHlth" into categories (e.g., 0 days, 1–10, 11–20, 21–30) to mitigate skew and improve interpretability.
- Flagging Placeholders: Recommended checking and flagging special values like 77, 88, 99 that may represent missing or coded entries in health-related variables.
- Separating Risk Index Dimensions: Advised testing "Chronic Risk Load" as both an aggregated index and its individual components for improved model explainability.

Actions Taken:

- Binning was applied to "MentHlth" and "PhysHlth" to reduce right skew.
- Placeholder values were confirmed absent in this dataset, so no additional imputation was required.
- Both the combined index and individual condition indicators were retained for model testing.

Final Feature Set and Rationale

The final feature set includes:

- **Transformed/Binned:** “Age_group”, “BMI_scaled”, “MentHlth_binned”, “PhysHlth_binned”
- **Engineered:** “Chronic_Risk_Load”, “Healthcare_Barrier_Index”
- **Domain-Informed:** “Sex”, “GenHlth”, “Education”, “Income”, “Smoker”, “PhysActivity”, “DiffWalk”, “HvyAlcoholConsump”
- **Target:** “Diabetes_binary”

These selections were made to maximize both interpretability and model precision. Rather than relying solely on statistical criteria, emphasis was placed on public health relevance and predictive power.

Conclusion

Feature preparation was instrumental in surfacing the predictive signals buried within the health indicators dataset. From thoughtful binning and scaling to strategically engineered features, each decision supported model clarity and relevance. AI-driven feedback added objectivity and highlighted nuanced adjustments that might have been overlooked. This refined dataset is now well-positioned for robust modeling aimed at diabetes prediction and public health insight.

References

- Centers for Disease Control and Prevention. (2023). Chronic Disease Indicators Overview. www.cdc.gov
- World Health Organization. (2022). Diabetes: Facts and Figures. www.who.int
- Harvard School of Public Health. (2021). Predictive Analytics in Chronic Disease Management. [www.hsph.harvard.edu](http://hsph.harvard.edu)
- OpenAI. AI-Powered Review of Data Preparation and Exploration. Internal Tool Use Only, 2025.