Predictive Modeling of Smoking Behavior Using Health Metrics: A Data Science Approach

Sophie Lin, Sudiksha Kalepu

Introduction

Smoking is a global public health issue associated with numerous chronic diseases, including cardiovascular conditions, respiratory ailments, and various cancers. Predictive models utilizing health metrics can offer insights into smoking behaviors, enabling proactive interventions. This paper explores the development of a predictive model to classify smoking status based on health examination data, focusing on feature engineering, model selection, and evaluation.

Data and Preprocessing

The dataset consists of 15,000 training samples and 10,000 test samples, with features including anthropometric measurements (e.g., height, weight, waist circumference), clinical data (e.g., blood pressure, fasting blood sugar, cholesterol, triglycerides, HDL, LDL, hemoglobin, serum creatinine, AST, ALT, GTP), sensory measurements (e.g., eyesight and hearing), and lifestyle indicators (e.g., dental caries, smoking status). The target variable, `smoking`, is binary: 1 for smokers and 0 for non-smokers.

1. Exploratory Data Analysis

Exploratory data analysis (EDA) revealed that approximately 63.1% of the training data were non-smokers, while 36.9% were smokers. All features were complete, with no missing values, making preprocessing simpler. Descriptive statistics and visualizations were used to understand feature distributions and correlations.

2. Feature Engineering

To enhance predictive performance, new features were engineered:

- **BMI Calculation:** Derived as weight divided by height squared. BMI serves as a proxy for overall health and lifestyle choices.
- **Blood Pressure Categories:** Systolic blood pressure was categorized into clinical risk groups (normal, pre-hypertensive, hypertensive stages).
- **Age Groups:** Stratified into young, middle, senior, and elderly age groups.
- **Health Ratios:** Cholesterol to HDL, LDL to HDL, and AST to ALT were calculated as markers of cardiovascular and liver health.
- **Blood pressure difference:** Calculated as the difference between systolic and diastolic (relaxation) pressures.
- **Vision asymmetry:** Computed as the absolute difference between eyesight measurements for the left and right eyes.

3. Data Preprocessing

Several steps were followed to prepare the data for modeling:

- **Separate Features and Target:** The target variable, `smoking`, was separated from the features in the training data. Features such as `id`, `bp_category`, and `age_group`, which were engineered but not relevant for modeling, were dropped.
- **Ensure Train and Test Consistency:** The test dataset was adjusted to match the feature columns in the training set. This was done by adding missing columns to the test dataset and setting their values to 0. Any extra columns in the test dataset were ignored.
- **Feature Scaling:** All numerical features were standardized using `StandardScaler` to ensure consistent scaling and prevent bias from features with larger ranges.
- **Categorical Variable Encoding:** Categorical variables were encoded using the `pandas.get_dummies` function to create one-hot encoded representations, ensuring they were ready for machine learning models.

Modeling and Evaluation

Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were trained and evaluated using a training-validation split (80-20). The primary evaluation metric was the area under the receiver operating characteristic curve (ROC-AUC), complemented by cross-validation to assess model stability. Results included:

- **Logistic Regression:** ROC-AUC of 0.8658; cross-validation mean ROC-AUC of 0.8689.
- **Random Forest:** ROC-AUC of 0.8743; cross-validation mean ROC-AUC of 0.8784.
- **XGBoost:** ROC-AUC of 0.8784; cross-validation mean ROC-AUC of 0.8782.

XGBoost emerged as the best-performing model, with its capability to capture complex feature interactions reflected in its feature importance analysis. Key predictive factors included BMI, cholesterol-HDL ratio, and systolic blood pressure.

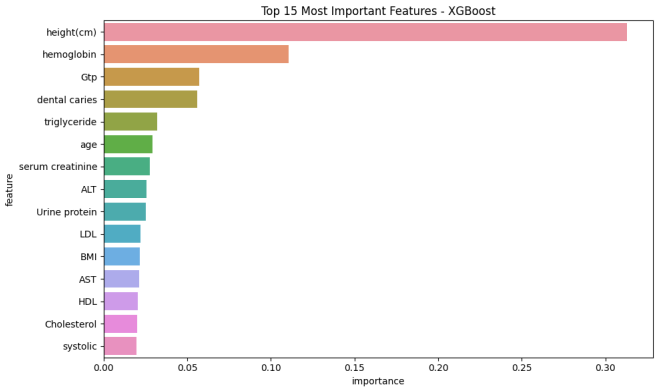

Figure 1: Top 15 Most Important Features (XGBoost)

Results and Insights

Visualizations revealed significant differences in health metrics between smokers and non-smokers:

- Smokers generally had higher BMI and systolic blood pressure, indicating poorer cardiovascular health.
- Cholesterol levels and LDL-to-HDL ratios were elevated in smokers, correlating with increased cardiovascular risks.

The correlation matrix further highlighted interdependencies among clinical metrics, such as strong correlations between cholesterol and triglyceride levels.

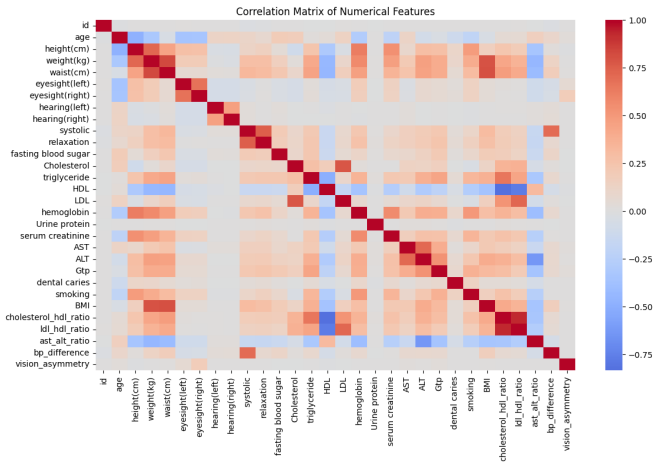

Figure 2: Correlation Matrix of Numerical Features

Conclusion and Future Work

The study demonstrates that predictive modeling can effectively classify smoking behavior using health metrics. XGBoost's robust performance underscores the value of incorporating engineered features and leveraging advanced machine learning techniques. Future work could explore:

1. **Temporal Analysis:** Longitudinal health data could provide insights into causal relationships.
2. **Interventions:** Predictive models could guide targeted health campaigns aimed at smoking cessation.
3. **Expansion of Features:** Integrating lifestyle data (e.g., diet, exercise) and genetic markers could enhance model precision.

The findings reaffirm the utility of data science in addressing public health challenges, emphasizing its role in preventative care and policy-making.

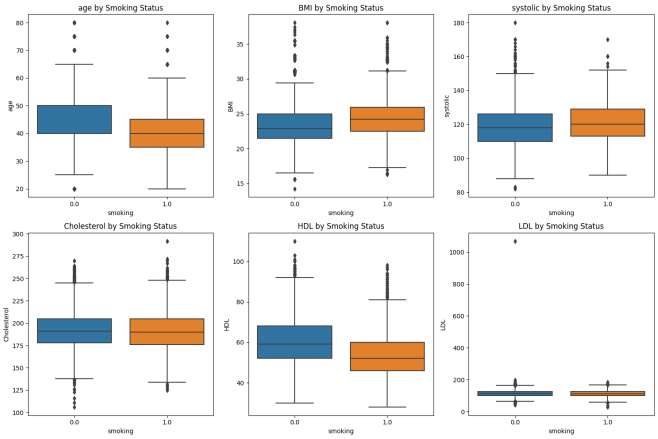

Figure 3: Key Features by Smoking Status

Figures