Predicting Alcohol Consumption Using BRFSS Data

Kate Mangubat

March 2025

Introduction

- The Behavioral Risk Factor Surveillance System (BRFSS) is a health related telephone survey that collects data on health behaviors

- Goal
  - Use BRFSS to analyze factors influencing alcohol consumption (DRNK3GE5)

- Investigating the impact of the following:
  - Sleep Duration (SLEPTIM1)
  - Emotional Support (EMTSUPRT)
  - Income (INCOME3)
  - Gender (SEXVAR)

DATA OVERVIEW

- Dataset contains **56,907** observations and **5 variables**.
- Target variable: **DRNK3GE5** (alcohol consumption frequency).
- Other variables: **SLEPTIM1, EMTSUPRT, INCOME3, SEXVAR**.
- **Missing values:**
  - EMTSUPRT: 22,433 missing values.
  - INCOME3: 1 missing value.

PREPROCESSING

- **Handling missing data:**
  - EMTSUPRT: Imputed missing values using median.
  - INCOME3: Dropped single missing value.

- **Feature engineering:**
  - Binned SLEPTIM1 into sleep categories.
  - Created an interaction term between sleep and income.
  - One-hot encoding applied to SEXVAR.

MODELING APPROACH

- **Regression model:** Predicting continuous alcohol consumption values.
- **Train-test split:** 80% training, 20% testing.
- **Feature scaling:** Standardized numerical variables.
- **Evaluation metrics:** Mean Squared Error (MSE), R-squared (R^2).

MODEL PERFORMANCE

Model	MSE	R^2
Linear Regression	1356.43	0.014
Random Forest	1361.41	0.010
Gradient Boosting	1349.33	0.019

MODEL PERFORMANCE

- **Gradient Boosting performed the best**, achieving the lowest MSE and highest R^2.
- The low R^2 values indicate that the features used do not fully explain the variability in alcohol consumption.
- **Feature importance analysis:**
  - Income and sleep duration were significant but weak predictors.
  - Emotional support had a limited direct impact but could interact with other variables.
- **Residual analysis:**
  - The models struggled with extreme values, indicating potential missing factors that influence alcohol consumption.

RESULTS & INSIGHTS

- Key findings:
  - Higher income correlates with increased alcohol consumption.
  - Sleep patterns have a nonlinear relationship with drinking frequency.
  - Emotional support plays a moderating role in alcohol consumption.

- Limitations:
  - Potential biases in self-reported data.
  - Limited external validity due to survey sampling.

CONCLUSIONS

- **Best performing model:** Gradient Boosting Regressor with the lowest MSE and highest R^2.

- **Model improvements:**
  - Additional features such as mental health status, stress levels, or social support could improve accuracy.
  - Hyperparameter tuning may enhance model performance.
  - Explore alternative machine learning techniques for better predictions.