Risk Prediction of Asthma Development in Youth using Statistical Learning Models

Introduction

- Asthma is a chronic respiratory disease that disproportionately affects youth in the U.S.
- There is currently no known cure for asthma, making early detection and prevention highly desirable.

Research Objectives

- Develop statistical learning models that can accurately predict the risk of a youth developing asthma using easily accessible survey data.

Methodology

Data Preparation

- Data was obtained from the National Health Interview Survey: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.

Data Pre-Processing

- Violin plots show the distribution of age and family poverty ratio of the children with and without asthma. A lower family poverty ratio means that the family is more impoverished.

Model Construction

Circular Bar Chart of variables associated with asthma development. Each bar in a variable is arranged from the lowest to the highest proportion of having asthma going clockwise.

Comparison of AUC scores under the Receiver Operating Characteristic (ROC) curves of multiple statistical learning algorithms. The AUC under the ROC curves indicates the performance of the statistical learning algorithm.

- No sampling.
- Undersampling.
- SMOTE-Tomek (undersampling and oversampling).
- Oversampling.

Results

Conclusions

- Undersampling was the best sampling technique and Random Forest, Logistic Regression, and Support Vector Machine were the best performing models.

Future Work

- Given the relatively modest AUC scores across all models and sampling techniques, future work could focus on leveraging larger datasets to improve model performance.
- Furthermore, hyperparameter tuning for each model may be implemented to further improve their performance.

References

Data was obtained from the National Health Interview Survey: https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm.