Executive Summary

Erdos Institute Data Science Boot Camp (Summer: May-2024 Cohort)

Project Title: *Exploring the Relationship Between Cancer Occurrence and Contributing Factors in the US Using Machine Learning Algorithms.*

Team members: Priti Singh, Pankaj Singh Dholaniya, Ikenna Nometa, Gbocho Masato Terasaki

Background and Data Set:
In this project, we aim to apply the knowledge gained from the boot camp to analyze real data from the Health Information National Trends Survey (HINTS) conducted by the National Cancer Institute. HINTS gathers nationally representative data on the American public's knowledge, attitudes, and use of cancer- and health-related information. The data is used to track changes in health communication and health information technology, helping to develop more effective communication strategies for diverse populations. We will utilize survey data from the second cycle of HINTS 4, collected between October 2012 and January 2013. This dataset comprises responses from 3,630 individuals residing in various regions across the United States. The data were collected on 357 features. See https://hints.cancer.gov/ for more information.

Study objectives:
Investigate the relationship between cancer incidence and three key factors:1) demographics, including age, gender, income, and geographical location; 2) the utilization of health information technology, including the Internet, for cancer-related education; and 3) medical history. We wanted to identify the features using different feature selection techniques that best predict the cancer outcome and measure the classification models' performance(s).

Data Preprocessing:
We examine the dataset for homogenization. For example, the missing values were originally coded as negative numbers, which were later converted to NAs. These missing values were examined during feature selection. Out of the 357 features, we studied 20 features that could predict cancer across three areas. The list included: Demographic features: *Age, BMI, education, census division, gender, and Income.* Utilization of health information technology: How much attention do you pay to information about cancer from each of the following sources?-*Health news, Internet, local TV, national TV, online news, or printed news.* Health & Medical history- *Diabetes, high blood pressure, heart condition, lung disease, arthritis, depression, general health, own ability to take care of health*

Balancing the Dataset:

To address class imbalance (‘EverHadCancer: No': 2355, ''EverHadCancer: Yes': 323), we used both oversampling techniques (Random Oversampling and SMOTE) to increase minority class instances and undersampling techniques (Random Undersampling) to reduce majority class instances. These techniques balance the class distribution, improving the model's predictive accuracy for both classes.

Feature selection and ranking: We employed Recursive Feature Elimination (RFE) to enhance our predictive modeling accuracy using four machine learning models, including Linear SVC, Decision

Tree, Logistic Regression, and Random Forest. This approach enabled us to robustly identify the most relevant features by aggregating rankings across models. We averaged ranking from all the models to finalize the top features. The top 15 features identified were Age, BMI, various CancerAttention metrics, Census division, Education, Gender, Household Income, and Internet Cancer Info metrics. These features significantly improved our model's predictive performance in terms of accuracy, precision, recall, and F1 score.

Choice of ML model and Hyperparameter Tuning

We chose Random Forests for our dataset as they train quickly with parallel tree building, reduce overfitting by averaging predictions, and offer clear interpretability through feature importance. We performed hyperparameter tuning to optimize our model's performance by systematically searching for the best combinations of parameters. The tuning process involved adjusting the number of estimators, maximum features, maximum depth, minimum samples for splitting and leaf nodes, bootstrap sampling, and splitting criteria.

The best-performing Random Forest model has parameters:
Best parameters found: {'bootstrap': False, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
This allows us to achieve an accuracy of the test model of 99.36%
This model achieves an F-1 score of 0.99 with only 6 False Negative out of 942 predictions.
This comprehensive search aimed to enhance our model's predictive accuracy and generalization capability, ensuring robust and reliable performance across different datasets.

Future directions and acknowledgments:

Some work we wish to do to improve the study include:

1. Evaluation of model performance on new data from subsequent years from HINTS,
2. Exploration of the effects of oversampling vs undersampling on the model accuracy and
3. Development of an app that will automatically predict whether a person has had cancer-based on the model we develop given input features.

We wish to thank:

- Roman Holowinsky, Alec Cott, Steven Gubkin, and the entire Erdös Institute Summer-May-2024 team
- Greg Edwards (our project mentor) and
- NIH's National Cancer Institute's (for HINTS).