Report on Predicting H1N1 and Seasonal Flu Vaccine Uptake with Comprehensive Exploratory Data Analysis (EDA)

Introduction

The objective of this study is to predict the likelihood of individuals receiving vaccinations for H1N1 flu (h1n1_vaccine) and seasonal flu (seasonal_vaccine) using data from the 2009 National H1N1 Flu Survey conducted by the CDC. This report presents a detailed exploration of the dataset, including data preprocessing, model building, evaluation, and insights derived from exploratory data analysis (EDA).

Methodology

Data Exploration:

- Conducted an extensive exploration of the dataset to understand its structure, features, and distributions.
- Examined missing values, data types, and distributions of categorical variables to gain insights into the data.

Data Preprocessing:

- Performed preprocessing steps such as handling missing values (imputation), encoding categorical variables, and splitting the data into training and testing sets.

- Ensured data quality and consistency through appropriate preprocessing techniques.

Model Building:

- Employed CatBoostClassifier, a gradient boosting algorithm designed for categorical features, for modeling.
- Trained separate models for predicting h1n1_vaccine and seasonal_vaccine probabilities to capture distinct vaccination behaviors.

Model Evaluation:

- Evaluated the models based on the area under the receiver operating characteristic curve (ROC AUC) for each target variable.
- Calculated the mean ROC AUC score across both target variables to determine the overall model performance.

Exploratory Data Analysis (EDA)

Data Loading and Overview:

- Loaded the dataset using Pandas and conducted a comprehensive overview.
- Obtained summary statistics of numerical features and displayed the first few rows of the dataset for initial insights.
- Missing Values Analysis:
- Conducted a thorough analysis of missing values in each column to assess data completeness.
- Applied appropriate imputation strategies such as mean or mode imputation to handle missing values effectively.

Categorical and Numerical Analysis:

- Analyzed categorical features including sex, age group, education, and employment status to understand their distributions and relationships with vaccination behavior.
- Explored numerical features such as household size, concern levels, and knowledge scores for patterns and correlations with vaccination outcomes.

Visualization Techniques:

- Utilized a variety of visualization techniques including stacked bar charts, count plots, and catplots to visually represent relationships between different variables and vaccination status.
- Visualized demographic patterns, vaccination rates, and opinions regarding vaccine effectiveness to identify trends and insights.

Feature Importance Analysis:

Conducted feature importance analysis to identify the most influential demographic and behavioral factors affecting vaccine uptake.
Extracted insights from feature importance analysis to inform targeted interventions and vaccination campaigns.

Results

- Initial Model Performance: The initial CatBoost models exhibited reasonable ROC AUC scores for predicting both h1n1_vaccine and seasonal_vaccine probabilities.
- Bagging Approach: Employed a bagging approach to enhance model robustness. Ten CatBoost models were trained with different random seeds, and predictions were aggregated to improve prediction accuracy.

- Final Model Performance: The bagged CatBoost models demonstrated enhanced performance compared to individual models, as evidenced by higher ROC AUC scores and improved predictive accuracy.

Conclusion

The predictive models developed in this study offer valuable insights into the likelihood of individuals receiving H1N1 and seasonal flu vaccines. By leveraging demographic, socioeconomic, and behavioral factors, healthcare organizations and policymakers can tailor vaccination campaigns and allocate resources effectively to mitigate the spread of flu viruses.

Recommendations

- Feature Importance Analysis: Further delve into feature importance analysis to pinpoint the most critical demographic and behavioral factors influencing vaccine uptake, enabling more targeted interventions and outreach strategies.
- Real-Time Monitoring: Establish mechanisms for continuous model updating with new data to ensure ongoing relevance and enable real-time monitoring of vaccination trends, particularly during flu seasons or pandemics.
- Collaboration: Foster collaboration with public health authorities and healthcare providers to integrate predictive models into vaccination planning and distribution strategies, facilitating data-driven decision-making and resource allocation.
- By implementing these recommendations, the predictive models can serve as invaluable tools in promoting public health and preventing the spread of flu viruses, ultimately contributing to improved vaccination coverage and disease prevention efforts.