NPHA Doctor Visits

Alize Ispahani
Manasi Acharya
Namrata De
Sahar Saiyed

July 18, 2024

Contents

1 Introduction **3**

2 Project Goals **4**
2.1 Intended Audience: 4

3 Summary of the Dataset **5**
3.1 Key Features 5
3.2 Why is this dataset is interesting? 5
3.3 Dataset Creation 5
3.4 Data Processing and Quality 6
3.5 Metadata and Protected Attributes 6

4 Design Decisions **7**
4.1 Data Visualization and Analysis 7

5 Challenges Faced **8**

6 Reflection on the Process **9**
6.1 Achievements 9
6.2 Lessons Learned 9
6.3 Future Improvements 9

1 Introduction

Our goal is to deliver meaningful insights into the health and well-being of older adults, utilizing data from NPHA dataset. We aspire for our app to aid policymakers and insurance providers in making well-informed decisions concerning the well-being of older adults in the USA.

2 Project Goals

The primary objective of this project is to provide insights into healthcare utilization among older Americans. The following questions are addressed through our dashboard:

1. **Attributes Leading to Doctor Visits**: Identify which attributes are associated with a higher frequency of doctor visits.
2. **Bias in Demographic Features**: Assess whether the dataset is biased in terms of race and gender distribution.
3. **Relation Between Health Features**: Explore the relationship between physical and mental health.
4. **Impact of Dropping Sensitive Features**: Evaluate if excluding sensitive features like race and gender affects feature importance.
5. **Overall Health Status Based on Demographics**: Determine the overall health status across different demographic factors.
6. **Low Doctor Visits Despite Poor Health**: Identify subsections of the population with poor health but low doctor visit frequency to help policymakers provide targeted benefits.

2.1 Intended Audience:

The primary audience for this analysis includes:

- **Healthcare Policymakers**: To aid in developing informed healthcare policies.
- **Insurers**: To help understand healthcare utilization patterns and potentially optimize insurance offerings.

3 Summary of the Dataset

The dataset originates from the University of Michigan National Poll on Healthy Aging (NPHA), which collects data on health, healthcare, and policy issues affecting Americans aged 50 and older. The dataset used for this project is a subset containing 714 records with 14 attributes, intended for testing and developing machine learning algorithms to predict the number of doctor visits per year.

3.1 Key Features

The dataset includes the following attributes:

- **Number of Doctors Visited**: Target variable indicating the count of different doctors seen in a year.
- **Age**: There is only one category of age in our dataset.
- **Physical Health, Mental Health, Dental Health**: Rated on a scale from Excellent to Poor, with an option for "Refused."
- **Employment**: Indicates employment status, such as full-time, part-time, retired, or not working.
- **Various sleep disturbance indicators**: Binary indicators for various sleep disturbances.
- **Prescription Sleep Medication**: Indicates regular, occasional, or no use of sleep medication.
- **Race**: Categorized as White, Black, Other, Hispanic, 2+ Races, with options for 'Refused' or 'Not asked'.
- **Gender**: Categorized as Male, Female, with options for 'Refused' or 'Not asked'.

3.2 Why is this dataset is interesting?

This dataset provides valuable insights into how older Americans utilize healthcare facilities. By predicting the number of doctor visits, we can better understand health-related behaviors and improve healthcare services for seniors. Analyzing demographic impacts on healthcare utilization can also inform more equitable policy decisions.

3.3 Dataset Creation

- **Creators**: Preeti N. Malani, Jeffrey Kullgren, and Erica Solway from the University of Michigan.
- **Date Collected**: April 2017.
- **Funding**: AARP and Michigan Medicine, the University of Michigan's academic medical center.
- **Collection Method**: Conducted by the GfK Group using the KnowledgePanel, a probability-based web panel representative of the US population.

3.4 Data Processing and Quality

- **Processing**: Dropped 'Age' feature before training since there is only one age group.
- **Quality**: The dataset is reported to have no missing values or known errors. Attributes are clearly defined, ensuring easy understanding and analysis.

3.5 Metadata and Protected Attributes

The dataset includes documentation detailing variable descriptions and survey methodology but lacks a README file. Protected attributes include race, gender and age which require careful handling due to privacy and ethical considerations.

4 Design Decisions

4.1 Data Visualization and Analysis

- **Visualization Tools**: Balsamiq (prototyping), Streamlit (Dashboard), Python (Matplotlib, Plotly)
- **Machine Learning Models**: Implemented to predict the number of doctor visits and analyze feature importance.(Random Forest)
- **Bias Analysis**: Conducted to ensure fair representation of demographic features.

5 Challenges Faced

- **Data Imbalance**: The dataset showed an imbalance in demographic features, particularly race and gender.
- **Computationally Exhaustive**: LIME and SHAP graphs take ample amount of time to reload, so optimization was needed.
- **Sensitive Information Handling**: Ensuring ethical use and analysis of protected attributes was paramount.

6 Reflection on the Process

6.1 Achievements

- Successfully created a dashboard providing actionable insights for healthcare policymakers and insurers.
- Identified key factors influencing doctor visits and health status among older adults.
- Addressed potential biases in the dataset and evaluated the impact of sensitive features on predictions.

6.2 Lessons Learned

- **Importance of Data Quality**: Ensuring data is free of errors and well-documented is crucial for reliable analysis.
- **Ethical Considerations**: Handling sensitive information with care and transparency is essential.
- **Iterative Approach**: The importance of refining analysis and models through multiple iterations to achieve accurate results.

6.3 Future Improvements

- **Enhanced Models**: Developing more sophisticated models to improve prediction accuracy.
- **Broader Applications**: Applying similar methodologies to other age groups or health datasets.
- **Policy Impact**: Collaborating with policymakers to translate insights into tangible healthcare improvements.
- Using Chi-Square Statistic, we checked that out of the demographic features, Race and Employment have significantly uneven distribution. Therefore, we tried to introduce some Random Over Sampling to ensure that data distribution is equal for all categories. As a future scope of this project, we would like to use this synthetically generated dataset with evenly distributed classes for training the prediction model and analyze how the model performance is affected.