Mapping Mortality: Predicting Cancer Death Rates Across America

Hyunseok Lee, Sahana Krishna Murthy

Background & Data

Cancer mortality continues to pose a significant public health challenge despite advancements in medical treatments. This study aims to investigate how socioeconomic factors (such as median income and poverty rates), healthcare coverage types, and demographic characteristics influence the target death rate due to cancer across different regions. This project focuses on predicting cancer mortality rates across the United States by leveraging various demographic, socioeconomic, and health-related data. We processed and analyzed two datasets—**Cancer Data** (3047 entries with 33 columns) and **Household Data** (3220 entries with 4 columns), merging them into a unified dataset of size 3047 x 39. Using advanced regression models and feature importance analysis, we identified the most influential factors contributing to cancer death rates. The best-performing model, Random Forest, provided insights into the relationships between educational attainment, incidence rates, and cancer mortality.

Data Processing and Exploratory Data Analysis (EDA)

Data Overview

- **Cancer Data**: Includes cancer incidence and mortality statistics.
- **Household Data**: Provides demographic and economic attributes.
- **Merged Dataset**: Combines the above datasets, resulting in 39 features for analysis.

Data Cleaning and Preprocessing

1. **Outlier Detection**:
   - Numeric features: Outliers were removed using the Interquartile Range (IQR) method.
   - Non-numeric features: Missing values were imputed with the mode.
2. **Feature Encoding**:
   - "Geography" was label-encoded for compatibility with machine learning models.
3. **Feature Correlation**:
   - A threshold of 0.3 was applied to identify highly correlated variables with the target variable, target_deathrate.

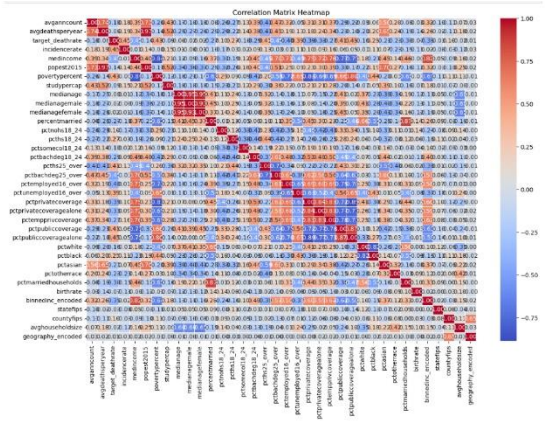

Figure 1 : Correlation between variable table

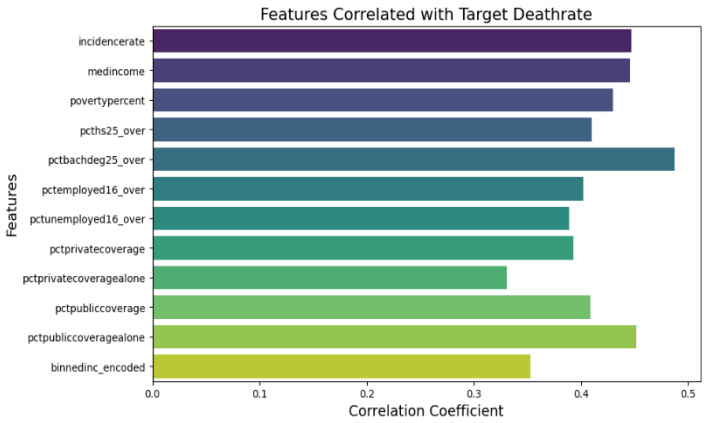

Figure 2 : Highly correlated with target_deathrate

Key Features

- Top 10 features with high correlation:
  - **pctbachdeg25_over**: Percentage of population with a bachelor's degree or higher.
  - **incidencerate**: Cancer incidence rate.
  - Employment and insurance-related variables.

Distribution Analysis

A residual plot highlights the near-normal distribution of errors, indicating reasonable model fit.

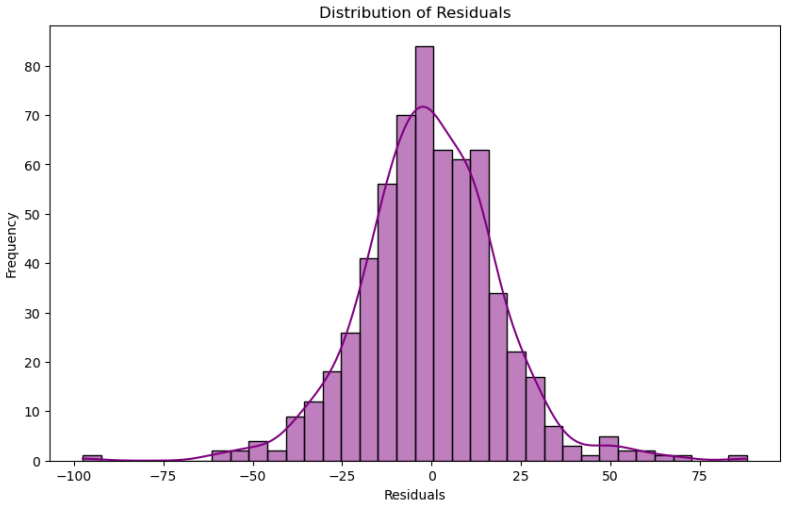

Figure 3 : Distribution of Residual (XGBoost)

Methodology

Machine Learning Models

We experimented with several regression algorithms, comparing their performance based on Root Mean Squared Error (RMSE) and R-squared (R^2) metrics:

1. **Linear Regression**
2. **Lasso Regression**
3. **Ridge Regression**
4. **Regression Tree**
5. **Random Forest**
6. **Partial Least Squares (PLS)**
7. **XGBoost**

Model Selection

Using **GridSearchCV** and **RandomizedSearchCV**, we tuned hyperparameters for optimal performance. The best model identified was the **Random Forest** with the following parameters:

n_estimators: 100 ; min_samples_split: 5 ; min_samples_leaf: 2 ; max_depth: 20

Evaluation Metrics

- Best Model Performance:
  - RMSE: Competitive among all tested models.
  - R^2: Explained 52% of the variance in cancer mortality rates.

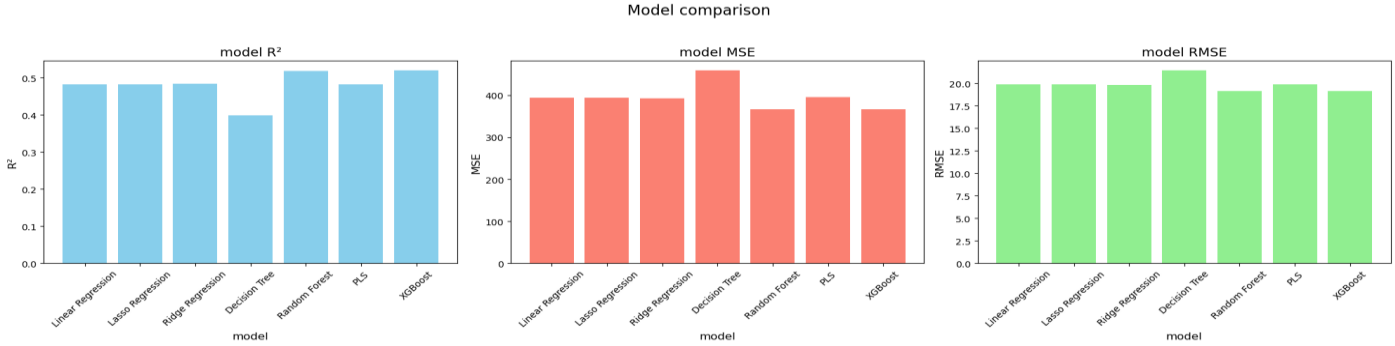

Figure 4: Model RMSE, MSE, R^2 comparison

Results and Analysis

Feature Importance

Feature importance analysis from the Random Forest model (see Figure 5) revealed the following key predictors of cancer mortality:

1. **Percentage of population with a bachelor's degree or higher** ("pctbachdeg25_over"): Education emerged as the most significant factor.
2. **Cancer incidence rate** ("incidence rate"): Positively correlated with mortality rates.
3. **Median income** and **target_deathrate** also played substantial roles.

Insights

The results underscore the importance of education and preventive health measures. Regions with higher educational attainment generally reported lower cancer mortality rates, suggesting that public health interventions targeting education could significantly impact mortality reduction.

Visual Representation

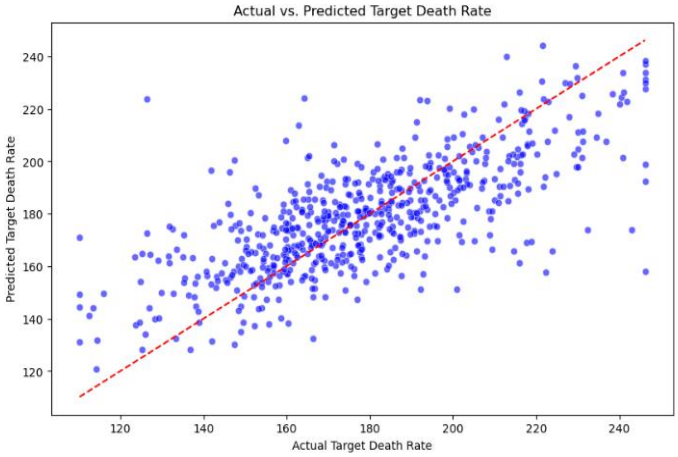

Figure 5: Residual distribution plot shows minimal skewness

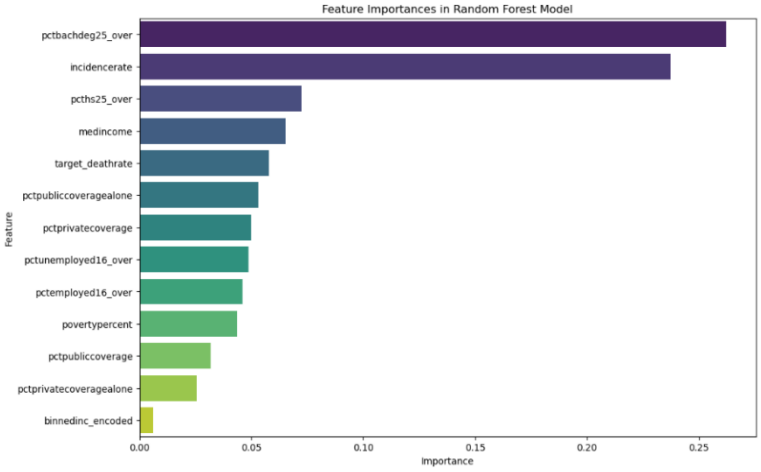

Figure 6: Feature importance plot by rank

Conclusion

1. **Best Model**: Random Forest proved to be the most effective model in predicting cancer mortality rates, capturing the relationships between demographic and health factors.
2. **Key Insights**:
   - Increasing educational attainment and lowering cancer incidence rates could significantly reduce mortality.
3. **Limitations**:
   - The model explains only 52% of the variance, indicating the need for additional data.
   - Geographical factors and skewed demographic distributions (e.g., race) limit generalizability.

Lessons Learned

- A multi-faceted approach is essential to understand and predict health outcomes.
- Incorporating more granular geographical data and addressing data skewness could enhance model performance.

This project highlights the utility of machine learning in public health, emphasizing actionable interventions based on data-driven insights.