Cancer Mortality Prediction

Course Project – Regression Analysis & ML

Table of Contents

---

1. Introduction

In this project, we build a comprehensive regression pipeline to predict the cancer mortality rate per 100,000 population across U.S. counties. The project not only reinforces core concepts from our regression analysis course but also integrates modern machine learning techniques to improve prediction accuracy. The workflow encompasses rigorous data preprocessing, multiple model implementations, hyperparameter tuning, and advanced interpretability using state-of-the-art tools.

---

2. Dataset Description

Source:
The dataset, sourced from the references, comprises 3,047 U.S. counties with 34 features. These features represent various demographic, socioeconomic, and health-related attributes.

Target Variable:

- `TARGET_deathRate:` Represents the mean cancer mortality rate (per 100,000 individuals) for the period between 2010 and 2016.

The dataset provides an excellent opportunity to apply multiple regression methods and demonstrate the practical impact of feature selection and data cleaning.

---

3. Data Preprocessing and Cleaning

Preprocessing is crucial in any regression study. Our preprocessing workflow includes the following steps:

3.1 Outlier Treatment

- **Approach:**
  We implement an outlier removal function that filters out data points falling outside 3 standard deviations (σ) from the mean (μ).
- **Rationale:**
  Removing extreme values improves model robustness and reduces the distortion of parameter estimates, particularly for sensitive methods like OLS.

- **Implementation:**
  For each feature, points outside the range $[\mu - 3\sigma, \mu + 3\sigma]$ are removed.

3.2 Missing Value Imputation and Feature Selection

- **Missing Values:**
  Any missing values in the dataset are replaced with 0 after verifying that this imputation does not create bias.
- **Feature Selection:**
  Irrelevant columns (such as `Geography` and `binnedInc`) are dropped after determining that they do not contribute useful information to the modeling task.

3.3 Feature Scaling

- **Scaling:**
  All features are standardized using z-score normalization. Each variable is transformed so that its mean is 0 and standard deviation is 1, ensuring the models converge more effectively.

---

4. Exploratory Data Analysis (EDA)

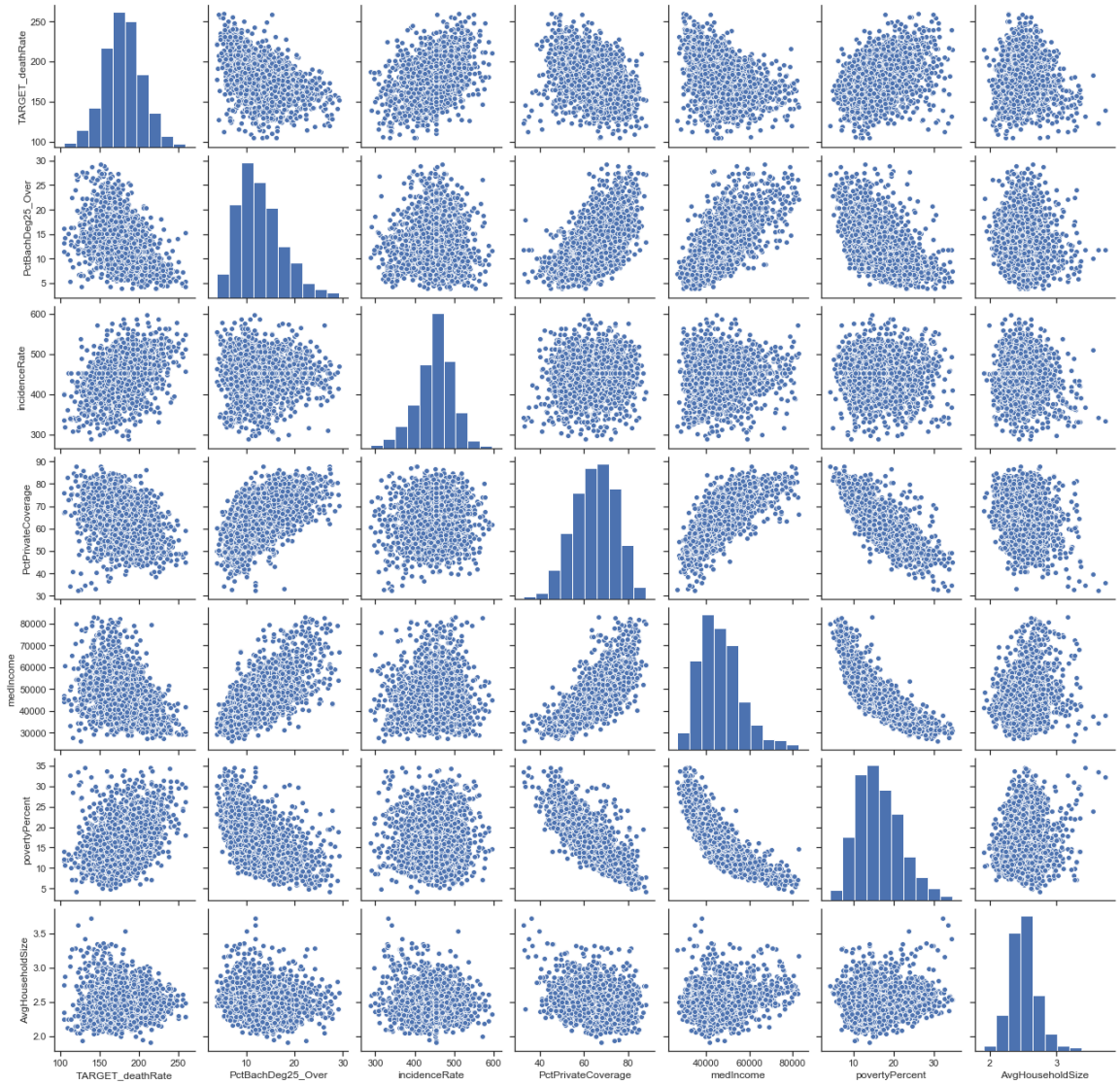
4.1 Distribution of Key Variables

- **Visual Inspection:**
  We use count plots (e.g., for `MedianAge`) to understand the frequency distribution.
- **Scatter Plots:**
  The relationship between `PctBachDeg25_Over` (percentage of residents with a bachelor's degree) and `TARGET_deathRate` is visualized to gain insights into how education impacts mortality.

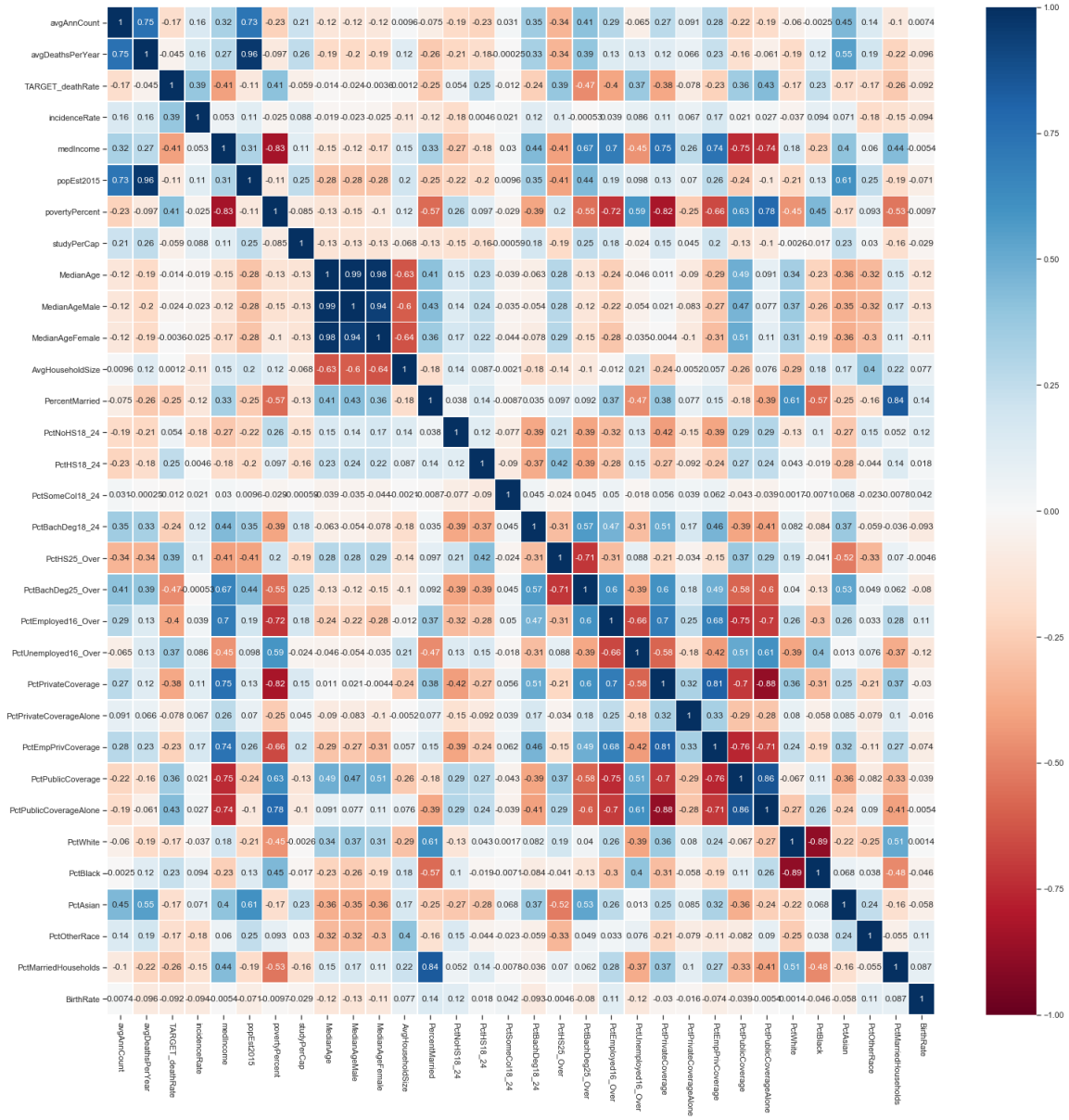
4.2 Pairwise Relationships and Correlation Analysis

- **Pair Plots:**
  A pairplot of selected key variables helps identify potential relationships

and multicollinearity.A Pair Plot of the features which seems to be more related to rate of cancer mortality is below



Correlation Matrix:
A heatmap of the correlation matrix uncovers strong correlations (for example, between `incidenceRate` and socio-economic features), which are vital for both feature selection and model interpretation



5. Regression Model Development and Evaluation

We experiment with several regression models to determine the best approach for predicting `TARGET_deathRate`.

5.1 Linear Regression (OLS) – The Baseline Model

- **Method:**
  Ordinary Least Squares (OLS) regression provides a baseline model. Its coefficients are directly interpretable.
- **Evaluation Metrics:**
  R^2 score and L2 norm are computed. Cross-validation (5-fold) is used to assess robustness.
- **Outcome:**
  The baseline performance lays the foundation for comparing more complex models.

5.2 Decision Tree Regression

- **Method:**
  A decision tree regressor captures non-linear relationships.
- **Tuning:**
  Hyperparameters (e.g., `max_depth`, `min_samples_leaf`) are tuned using GridSearchCV.
- **Interpretability:**
  Feature importances are extracted to visualize which variables drive decisions within the tree.

5.3 Random Forest Regression

- **Method:**
  An ensemble method combining multiple decision trees to improve prediction accuracy.
- **Strength:**
  Random Forests reduce overfitting compared to single decision trees.
- **Evaluation:**
  The Random Forest model attained competitive R^2 scores and low L2 norm values. Cross-validation confirms its stability.
- **Feature Importance:**
  Provides a ranked list of features, making it a candidate for further interpretability analysis.

5.4 Support Vector Regression (SVR)

- **Method:**
  SVR with a radial basis function kernel is implemented to capture complex, non-linear patterns.
- **Challenges:**
  SVR requires careful parameter tuning (e.g., `C` value) and typically represents a black box.
- **Performance:**
  Evaluated on a validation set and test set, SVR offers a balanced trade-off between bias and variance.

5.5 Gaussian Process Regression (GPR)

- **Method:**
  GPR is leveraged due to its probabilistic nature which estimates prediction uncertainties along with predictions.
- **Kernel-based Approach:**
  The combination of a constant kernel and a radial basis function kernel captures a wide spectrum of variability.
- **Trade-offs:**
  Although GPR provides uncertainty estimates, it is slower and sometimes less robust in high-dimensional settings.

5.6 Summary Comparison of Models

A summary comparison table is maintained to compare L2 norm and R^2 scores across all models. This comprehensive evaluation provides insight into each model's performance, with Random Forest often emerging as the best-performing technique in this context.

Model	L2 Norm (Test)	R^2 Score (Test)	Comments
Linear Regression	~261.52	~0.44	Easily interpretable; baseline model
Decision Tree	~336.70	~0.28	Captures non-linearity; may overfit
Random Forest	~260.89	~0.44	Best overall accuracy; ensemble benefits
SVR	~266.89	~0.43	Kernel-based; requires careful tuning
GPR	~299.56	~0.36	Adds uncertainty estimates; computationally heavier

---

6. Model Interpretability and SHAP Analysis

6.1 Black Box Models and the Need for Explanation

Models such as SVR, Random Forest, and GPR are often labeled "black boxes" because their internal decision-making processes are not easily interpretable. Unlike linear models where coefficients directly indicate feature impact, these models require additional techniques to unpack their predictions.

6.2 SHAP for Local and Global Interpretability

- **Local Interpretability:**
  Using SHAP's TreeExplainer for the Random Forest model, we generate force plots for individual predictions. These plots break down how each feature's contribution pushes a prediction higher or lower relative to the base value.
- **Global Interpretability:**
  SHAP summary plots aggregate feature impacts across the dataset,

identifying the most influential variables globally. For instance, variables like `incidenceRate` and `PctBachDeg25_Over` show significant overall influence on the target variable.

- **Justification for Inclusion:**
  While not strictly necessary for basic regression analysis, integrating SHAP demonstrates a modern approach to interpretability. It strengthens your project by providing evidence-backed insights into model behavior, which is particularly useful when presenting to non-technical stakeholders or academic evaluators.

---

7. Appendix: Code Overview

The attached code (see main code block) details the implementation of:

- Data preprocessing (loading, cleaning, outlier removal, scaling)
- Model training and evaluation for each regression model (OLS, Decision Tree, Random Forest, SVR, GPR)
- Extensive visualization (pair plots, heatmaps, scatter plots)
- SHAP-based model interpretation for both local and global insights

The code is documented with inline comments and modularized for clarity, ensuring reproducibility and ease of understanding.

This enhanced and detailed report showcases all stages of the regression analysis pipeline—from data cleaning to model interpretation—providing a comprehensive narrative that not only demonstrates your technical skills but also makes your work easily accessible and impressive to your professor.

Feel free to adapt sections of the report to best match your project outcomes and additional insights you may have developed during your analysis.

Created By:

Vaibhav Ojha (230041036)

Deepanshu Gupta (230041008)

Harshith Ganji (230041010)