### Simplifying State Healthcare Rate Prediction with L1 Regularization

#### Surya Gopalakrishnan for MET CS777

**My dataset source:** <https://healthdata.gov/CMS/2024-Child-and-Adult-Health-Care-Quality-Measures-/pahp-sncm/about_data>

**2024 Child and Adult Health Care Quality Measures Quality**

Performance rates on publicly reported health care quality measures in the CMS Medicaid/CHIP Child and Adult Core Sets, for 2024 reporting.

Source format : CSV (9.1 MB) with 11,101 rows of data

#### Some preliminary details about the table:

|  |  |  |
| --- | --- | --- |
| **Columns I will use** | **What is it about** | **Notes** |
| State | The U.S. state reporting the data | Can be one-hot encoded or used to analyze regional patterns |
| Domain | The healthcare domain or category of the measure (e.g., Behavioral Health, Maternal Health) | Important for grouping types of care |
| Reporting Program | Indicates if the measure belongs to Child Core Set or Adult Core Set | Helps compare program impact |
| Measure Name - Not Required when we have an abbreviation | Full name of the performance measure | Usually too granular; better to use Measure Abbreviation |
| Measure Abbreviation | Short code for the measure (e.g., FUM-CH, CCW-AD) | Good categorical feature- for dimensionality reduction |
| Measure Type | Indicates whether higher or lower rates are better (direction of improvement) | Crucial to interpret rate properly, also serves as a categorical feature |
| Population | Population group for the measure (e.g., Medicaid, CHIP, Dual Eligible) | Important for understanding population context |
| Methodology | Data collection methodology (e.g., Administrative, Hybrid) | May affect reported rates |
| Core Set Year (Optionally) | Year of the reporting core set (e.g., 2024) | Can be treated as categorical or numerical |
| **State Rate** ( This is the target variable (performance rate) the model aims to predict) | Indicates if the state’s performance rate |  |

I will not be using all the rows for the project. Instead will take into account some important columns like the measure type – this is a challenge as some rates are better when lower (e.g., hospital readmissions), while others are better when higher (e.g., follow-up after emergency visits)

#### Research Question:

**How well can we estimate state healthcare performance rates using minimal input variables?**

I do not have a categorical target (rate), so I will try to use Lasso Regression that shrinks less important feature coefficients.

My approach will be,

* Convert categorical features like domain, measure type using one hot encoding
* Apply feature selection – L1 Lasso
* Train Linear Regression
* Evaluate the model

**Goal:** Apply Lasso to create a lightweight model that keeps only the strongest features.

#### Machine Learning Model

I plan to use **Lasso Regression (a linear regression model with L1 regularization**) for this project. Since the target variable, State Rate, is continuous (percentage rate), regression is appropriate. Lasso is particularly useful because it performs feature selection by shrinking less important feature coefficients to zero, resulting in a simpler, more interpretable model. This is especially valuable given the complexity of the dataset and the presence of multiple categorical variables such as Domain, Reporting Program, and Measure Type. Prior to modeling, I will apply **one-hot encoding** to convert categorical variables into a numerical format.

#### Expected Outcomes:

The primary goal is to understand how well we can predict state healthcare performance rates using available features like Reporting Program (Child vs Adult Core Set), Domain, and Measure Type, among others. By applying Lasso regression, I expect to **identify the most influential factors affecting performance rates** and build a lightweight **predictive model that excludes irrelevant features**. This can provide insights into which programs or domains tend to have better or worse performance and assist policymakers in targeting improvements.

#### Evaluation:

I will evaluate the model's performance using standard regression metrics:

**R² Score** to measure how well the model explains variance in the performance rates.

**Root Mean Squared Error (RMSE)** to quantify the prediction error in the same units as the rates.

I plan to split the same dataset into training and testing subsets to assess generalization.

from sklearn.model\_selection import train\_test\_split

train, test = train\_test\_split(data, test\_size=0.2, random\_state=42)

Given the complexity and variability of healthcare measures (some rates are better when higher, others better when lower), I expect moderate predictive performance. The goal is not only accuracy but also interpretability, with Lasso helping to highlight key features.