Predicting Federally Qualified Health Center Program Funding Levels

Project Overview

Federally qualified health centers (FQHCs) are one of the most essential healthcare providers in the United States, providing care to over 30 million people each year, regardless of their ability to pay. As mission-driven organizations, FQHCs provide high-quality healthcare to uninsured and underserved individuals and families in every state. The Health Resource and Service Administration (HRSA), is that certifies eligibility and compliance for entities to become and maintain FQHC status by awarding Health Center Program funding several times a year. Becoming an FQHC is a long and complex process that requires certain procedures, organizational structures, and government funding experience as they receive a significant federal grant to become an FQHC.

My aim with this project is to predict the estimated Health Center Program funding a new entity could secure given certain information about their patients, community, and organization structure. Health organizations could use this estimate to determine if it is worth going through the process to become a federally-qualified health center, especially to help offset any costs they incur by providing uncompensated care.

High Level Overview of Results

This project presents the culmination of an extensive data analysis and modeling project focused on predicting total health center funding based on various operational and demographic factors. The best-performing model, a Random Forest Regressor, achieved an R2 score of 0.648, demonstrating a strong ability to predict funding levels from complex and diverse inputs. This model notably outperformed others by effectively handling outliers and leveraging important features like total patients, operation hours, and uninsured ratios, which were pivotal in improving the prediction accuracy. The insights derived from this analysis provide a data-driven foundation for decision-making that could inform strategic planning for health organizations interested in applying for Health Center Program funding. The final report includes detailed evaluations of model performance, feature importance analyses, and residual examinations, all of which underscore the robustness and reliability of the predictive models developed.

Data

Currently funded FQHCs are required to provide details about their organizations on an annual basis to HRSA through the Uniform Data System (UDS) report. Data is collected through an online portal and goes through several review processes to ensure no data errors are present. I used the most recently published data, which is from the calendar year 2022.

To view the data source, data reporting requirements and definitions, or other available information about FQHCs, visit this links below:

- Health Center Program Annual UDS Data
- 2022 UDS Manual
- National Awardee Data

The UDS report collects thousands of measures about an organization's operations, including patient demographic details, staffing and utilization, costs of care, operating hours and locations, clinical outcomes and diagnosis details, financial information (funding, insurance, revenue, expenses, etc.).

In order to reduce the size of the dataset for this project, I met with Eric Battis, a former Chief Financial Officer of an FQHC in Arizona, to determine what data is most critical to understand a health center's operations. The list of table I included for the scope of this project include the following:

- *Health Center Info:* Health center name, address, grant number, and project director.
- *Health Center Site Info:* Health center operating and administrative locations.
- *Table 9E:* Non-patient generate revenue including grant and other revenue.
- *Health Center Zip Codes:* Health center operational locations and number of patients served.
- *Table 8A:* Costs of providing care.
- *Table 5:* Personnel, staffing utilization, and total visits by area.
- *Tables 3A and 3B:* Patient demographic details.
- *Table 4:* Other patient characteristics including insurance and special populations.
- *Table 9D:* Revenue generated from patient services.

I imported each of these tables as a separate dataframe. Each was processed with various cleaning and wrangling steps before consolidating data into one dataframe for visualization, preprocessing and modeling.

Data Cleaning and Wrangling

Notebook Links:

- Data Cleaning
- Data Wrangling

1. Cleaning Column Names & Subsetting Data

After importing each dataset as a dataframe, I created a function to iterate over a dataframe dictionary to print out the shape of the dataframe and the column names. I needed to review these to determine which columns would be relevant based on my meeting with Eric.

The nine tables had over 650 columns combined. Additionally, several column names were exceedingly long. For example, one column was called `Medicare (Inclusive of dually eligible and other Title XVIII beneficiaries)-18 and over years old (a)` and I renamed it `Medicare18 and up`. I subsetted each dataframe manually by determining which columns to keep for further analysis. I then renamed columns to keep their details but not be as long.

2. Missing Values

The UDS report contains 3 different types of missing values that are Missing Not At Random (MNAR). Tables could have one or more of these MNAR types represented by the data entry options below:

1. "-" represents no data entry by health center
2. "--" represents suppressed patient counts between 1-15 to protect patient privacy
3. "---" represents suppressed health center confidential data

I created a function to examine the impact of each of these null types in each dataframe including the number of instances of each MNAR type and percent of the data missing in each column. I then dealt with the missing types differently based on the reason they were missing.

Below is an example output from the function `find_missing_values()`.

```
The health_centers dataframe has the following missing value counts for "-":

           Column Name  Missing Value Count  % of Total
0  HealthCenterZIPCode                    1    0.072993

The health_center_sites dataframe has the following missing value counts for "-":

                   Column Name  Missing Value Count  % of Total
4                   SiteZIPCode                 1400    9.341429
1  TotalWeeklyHoursOfOperation                   17    0.113432
2                      SiteCity                    3    0.020017
0                      SiteName                    2    0.013345
3                     SiteState                    2    0.013345

The health_center_funding dataframe has the following missing value counts for "-":

                  Column Name  Missing Value Count  % of Total
3                   ph_amount                  710   51.824818
0                  mhc_amount                  648   47.299270
2                   ho_amount                  603   44.014599
4  total_other_federal_grants                  494   36.058394
6      total_local_gov_grants                  432   31.532847
5           total_state_grants                 257   18.759124
7         total_private_grants                 175   12.773723
8          total_other_revenue                 159   11.605839
1                  chc_amount                   39    2.846715
```

My strategy for dealing with each missing type was as follows:

1. For the "-" missing type, I replaced these with '0' since there was no data entry by the health center. No entry would mean 0 for that field since it is possible to not have data to enter. For example, the health center funding table had many instances of '-' for `ph_amount` which would be where an entity reports the amount of public housing health center funding they receive. Most health centers do not receive this subtype of health center funding, so they would leave it blank. The health_center site dataframe had this type of missing value for site geographic details, and I dropped these rows later because it could mean the site was not yet operational.
2. The "--" missing type was initially replaced with np.nan values. I decided to replace these values with a random number between 1 and 15 since the "--" was to suppress patient counts between 1-15 to protect patient privacy.
3. The "---" was also replaced with np.nan values. I kept these dataframes separate from the 3 dataframes with the "--" missing types so I could differentiate between the different missing types and impute them appropriately. Details about how I dealt with these missing values will be covered during the preprocessing section.

After dealing with the missing values, I consolidated the initial nine dataframes into 4 dataframes for further wrangling.

3. Creating Summary Data for Consolidation

Two dataframes consisted of several rows of data for each health center. The `Service Area` dataframe consisted of nearly 15,000 rows of data and the `Service Sites` dataframe had approximately 97,100 rows of data. I decided to create aggregate information from these tables to add to the main `Health Centers` dataframe which had details about each entity including the target variable `total_hc_funding`. The new summary data that was added to the `Health Centers` dataframe included zip code counts for each entity representing their total service area, the count of health center sites each entity manages, the total weekly hours of operation across sites for each entity, and the count of cities each entity operated in. I then merged this new aggregated dataframe to the `Health Centers` data.

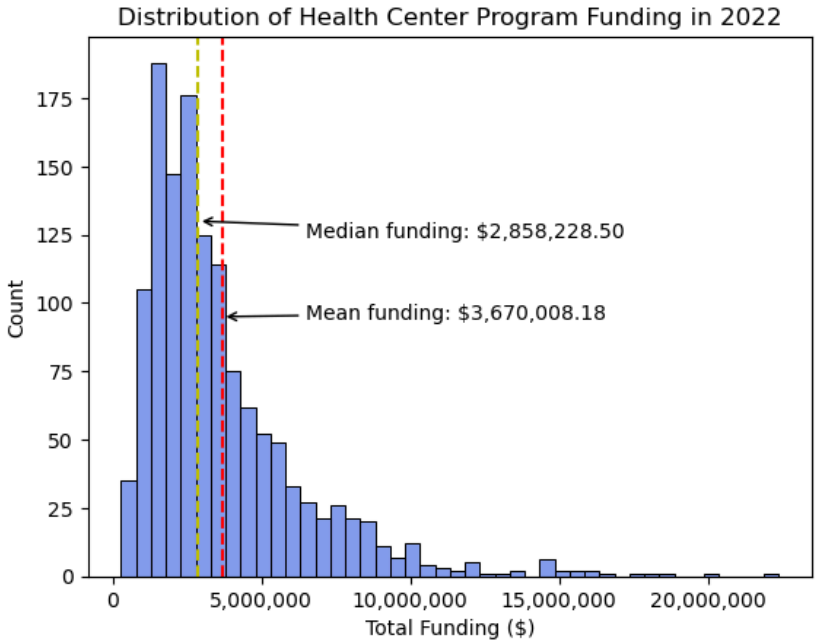
After imputing random numbers ranging from 1-15 for the MNAR type "--", I consolidated the `Health Center Ops Finance` dataframe with `Health Centers`.

Exploratory Data Analysis

Notebook Link:

- Exploratory Data Analysis

During EDA, my primary goal was to look at the distribution of the target variable, `total_hc_funding` and how it interacted with other features. The target variable has a wide distribution, ranging from $275,778 to $22,382,349. Most health centers between receive 1.8M and 4.6M of health center funding annually, with the mean funding level just over 3.6 million. These especially large outliers were not due to errors, some entities do in fact receive nearly 80x the amount of funding as other entities. There are similar distributions with the total number of patients served.



Total patients has a strong positive correlation with total HCP funding, demonstrated with the Pearson's correlation coefficient of 0.730.

I wanted to test several hypotheses with how different features affect funding.

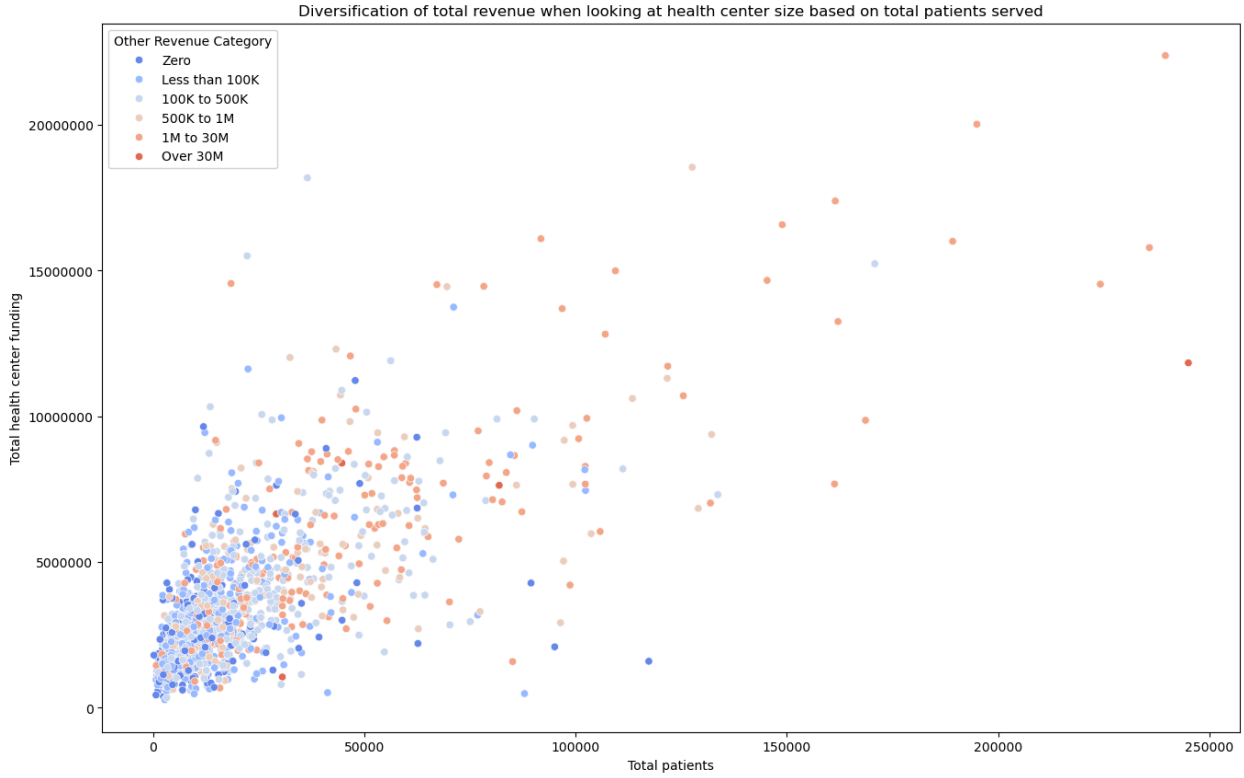
Rural vs. Urban: Urban providers represent 58.7% of the total number of entities and they receive 64.1% of the total available health center funding. Rural providers represent 41.3% of the total number of providers and receive 35.9% of the funding.

Do urban providers receive more funding than rural providers after controlling for patient volume? I conducted an ANCOVA statistical test to isolate the effect of the categorical variable (Urban or Rural) on the dependent variable (total health center funding). I want to control for patient volume because the total number of patients is highly correlated with the total health center funding. After controlling for patient counts, there is not a

statistically significant difference in the total health center funding Urban providers receive compared to Rural providers.

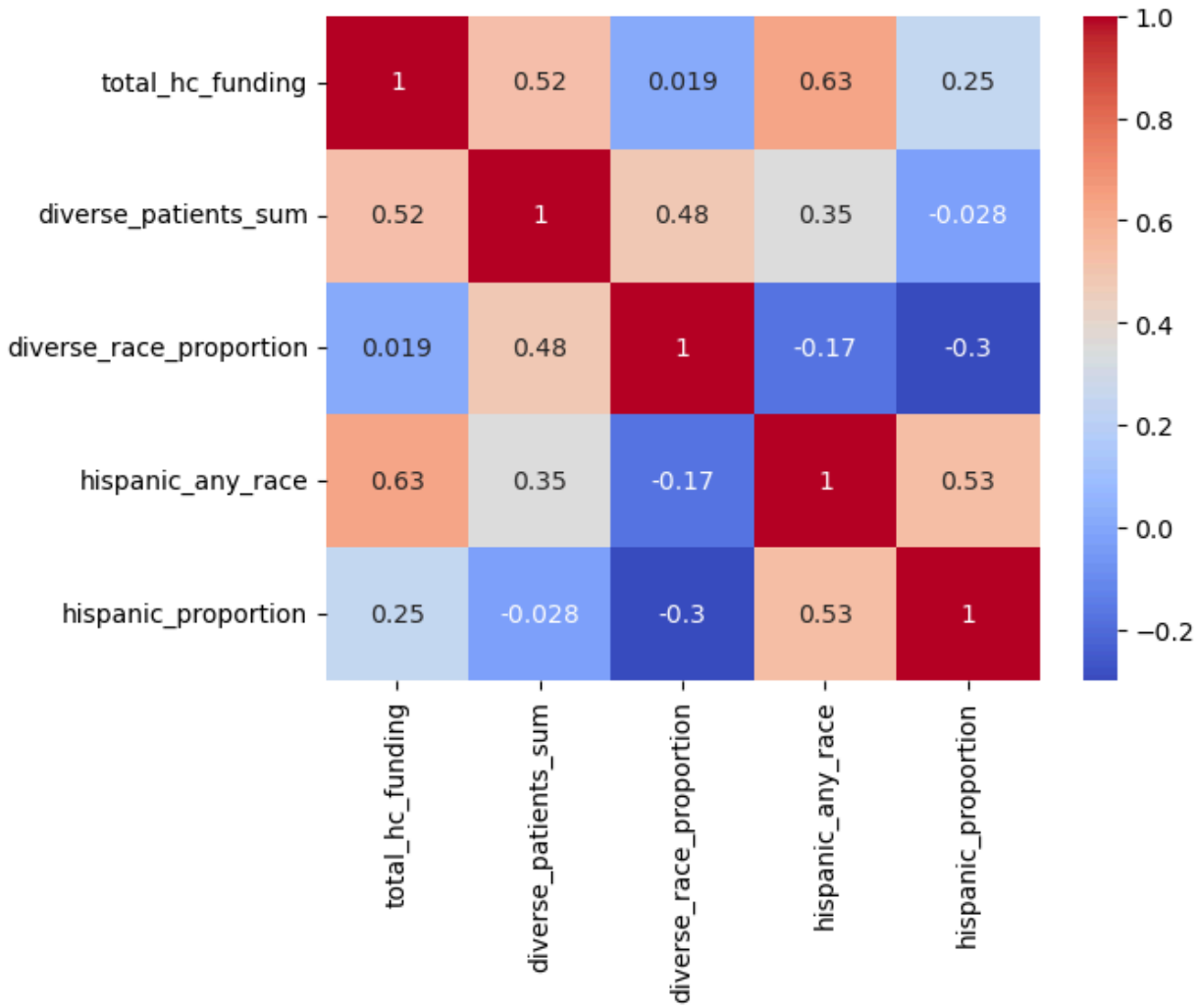
Other Non-patient Revenue:
The larger health centers become, the more diversified their revenue streams appear. As some health centers increase in patient size, they seem to be more likely to have larger amounts of funding coming from other non-patient and non-grant revenue streams. However, we did see in the table earlier that this is not always the case. Some health centers that have massive amounts of total other revenue, like East Boston Neighborhood Health Center only serve a small number of patients, just over 80,000 in their case.

Do health centers with more total other revenue receive more HCP funding compared to health centers with less other revenue? I conducted an ANCOVA statistical test and controlled for patient size. I also conducted a Logit Regression test to compare HCP funding between entities broken up into 4 quantiles. Both tests demonstrate that health centers that have higher other non-patient non-grant revenue have more total HCP funding. Of note however is that other revenue only explains 6% of the variance in HCP funding.



SDOH Proportions:
Health centers are generally supposed to increase access to care and other critical services to a wide range of vulnerable populations, including, but not limited to, school children, the elderly, pregnant women and infants, immigrants, minority populations, the LGBTQ+ community, people with disabilities, military veterans, migrant and seasonal farm workers, people experiencing homelessness, residents of public housing and people with limited English proficiency. I looked at the distribution of each of these populations served across entities and how correlated each feature was to total HCP funding. I also created new columns to see if higher proportions of vulnerable populations were correlated with more funding.

The correlation matrices demonstrated that the number of racially diverse patients, the number of hispanic patients, and the number of other groups mentioned above are positively correlated with health center funding. Below is the matrix for race and ethnicity total numbers and proportions and how they related to funding.



Although most ratios didn't show a strong relationship with funding, they are essential for understanding equity and representation within health services. They do not strongly influence funding amounts but they are critical for policy-making, resource allocation, and community engagement strategies. Data is not available about entities that were not selected for funding, however these ratios likely play a large role in scoring new entities based on the community impact they could have in their proposed service areas. Data from this analysis could serve as a good reference point for potential new agencies to compare their impact to funded entities in their state.

I conducted logistic statistical tests to test whether agencies serving higher proportions of each population type were more likely to receive more health center funding compared to agencies with lower proportions. For the groups that had a wider distribution range, I split the data into quantiles (0-25th, 25th-50th, 50th-75th, 75th-100th). For data that had extremely skewed distributions (like mostly 0), I split the entities based on being above or below the median.

The following proportions showed a statistically significant difference in total HCP funding between quantile groups or binary groups:

- Diverse race
- Hispanic
- Low income
- Poverty
- Medicaid
- Limited English proficiency

- Private insurance
- Uninsured
- Medicare
- Homeless
- Veterans
- Public housing
- Migrants
- School-based
- Public insurance

As the proportions increased, the amount of funding entities had also increased for most entities. The exception was entities serving higher than the median proportion of veterans receiving statistically significantly less funding compared to other entities. This could be due to having access to other types of special funding reserved for veteran-facing agencies, therefore these agencies are not prioritized for HCP funding.

It should be noted that the pseudo R-squared values (which demonstrate the amount of variance explained by each of these measures) ranged from 0.5% to 3.9%, so none were particularly strong predictors of total HCP funding.

Pre-Processing

Notebook Link:

- Pre-Processing

During preprocessing, identifier columns that were not required for analysis, such as `BHMISID` and `HealthCenterName`, were removed. The `HealthCenterState` feature was encoded using frequency encoding to reflect the potential influence of state population sizes on funding allocations. This approach ensures that the categorical data can be appropriately used in distance-based modeling techniques. I also removed the features I created that resulted in SDOH bins from my statistical testing during EDA.

To address missing data, which constitutes approximately 43% of our dataset due to confidentiality choices by health centers, two strategies were adopted. First, columns with missing data were excluded from one version of the dataset while maintaining indicators for withheld operations and financial data. Secondly, another version of the dataset utilized `MICE` for imputing missing values, with both datasets prepared for comparative model performance evaluation.

For scaling, numerical data was segregated from categorical data and processed using two scaling techniques. The `RobustScaler` method was employed to mitigate the influence of outliers by adjusting according to the median and interquartile ranges, suitable for data with extreme values. Additionally, the `PowerTransformer` method was used to normalize the data distribution, which enhances certain modeling techniques' effectiveness. Each scaling method was applied to prepare four distinct datasets for subsequent modeling steps.

Categorical variables were transformed using dummy encoding for nominal features and integer encoding for binary or ordinal features. This process ensures that all data presented to the models are numerical and appropriately formatted for analysis.

I also created a separate train-test set for my target variable, `total_hc_funding` where I rounded each row to the nearest 500,000. I did this because knowing the exact amount of funding is not necessarily the goal. I wanted to see if rounding improved model performance.

The final step involved integrating the scaled numerical data with the encoded categorical data, creating a comprehensive dataset ready for detailed analysis and modeling. This methodical approach to data preparation ensures that the models developed are robust, reliable, and well-suited for predictive accuracy.

Modeling and Model Evaluation

Notebook Link:

- Modeling

For each train-test split, I tested 2 different models Linear Regression and Random Forest Regressor. I also conducted PCA on the four different train-test splits and incorporated the PCA columns into the datasets for modeling.

I created an empty dataframe to keep track of model scores, including R2, Mean Absolute Error, and Root Mean Squared Error. I also used cross validation for each model and calculated the mean scores for each of the earlier mentioned metrics.

I sorted the model scores by Mean Absolute Error and compared the best 2 performing Linear Regression models and the best 2 Random Forest models.

Summary of Top Models

Random Forest Models

Reduced Robust Rounded Data:

- Train RMSE: 630,797
- Test RMSE: 1,585,783
- Train MAE: 412,195
- Test MAE: 1,110,124
- Train R^2: 0.948
- Test R^2: 0.648

The model demonstrates a strong fit on the training data but shows some overfitting as indicated by the drop in R^2 from training to testing. The robust handling of outliers in the rounded dataset might have contributed to the better performance on the test set.

Reduced Power Data:

- Train RMSE: 768,621
- Test RMSE: 1,589,978
- Train MAE: 547,509
- Test MAE: 1,110,669
- Train R^2: 0.923
- Test R^2: 0.646

This model also fits well on the training data with a slightly worse performance on the test data compared to the rounded dataset model. It indicates a consistent performance across different preprocessing strategies.

Linear Regression Models

Imputed Power Rounded Data:

- Train RMSE: 1,687,938
- Test RMSE: 1,642,445
- Train MAE: 1,195,219
- Test MAE: 1,213,404
- Train R^2: 0.628
- Test R^2: 0.623

The linear regression model on rounded data with power transformation shows minimal overfitting. The lower R^2 values compared to RF models indicate less ability to explain the variance in the data.

Imputed Power Data with PCA:

- Train RMSE: 1,687,190
- Test RMSE: 1,642,862
- Train MAE: 1,194,952
- Test MAE: 1,215,413
- Train R^2: 0.629
- Test R^2: 0.622

This model slightly underperforms compared to its rounded counterpart but still maintains similar metrics, showcasing the effect of PCA on model performance.

Evaluation

The Random Forest models outperform Linear Regression models in terms of RMSE and MAE, highlighting their effectiveness in handling complex patterns and outliers within the.

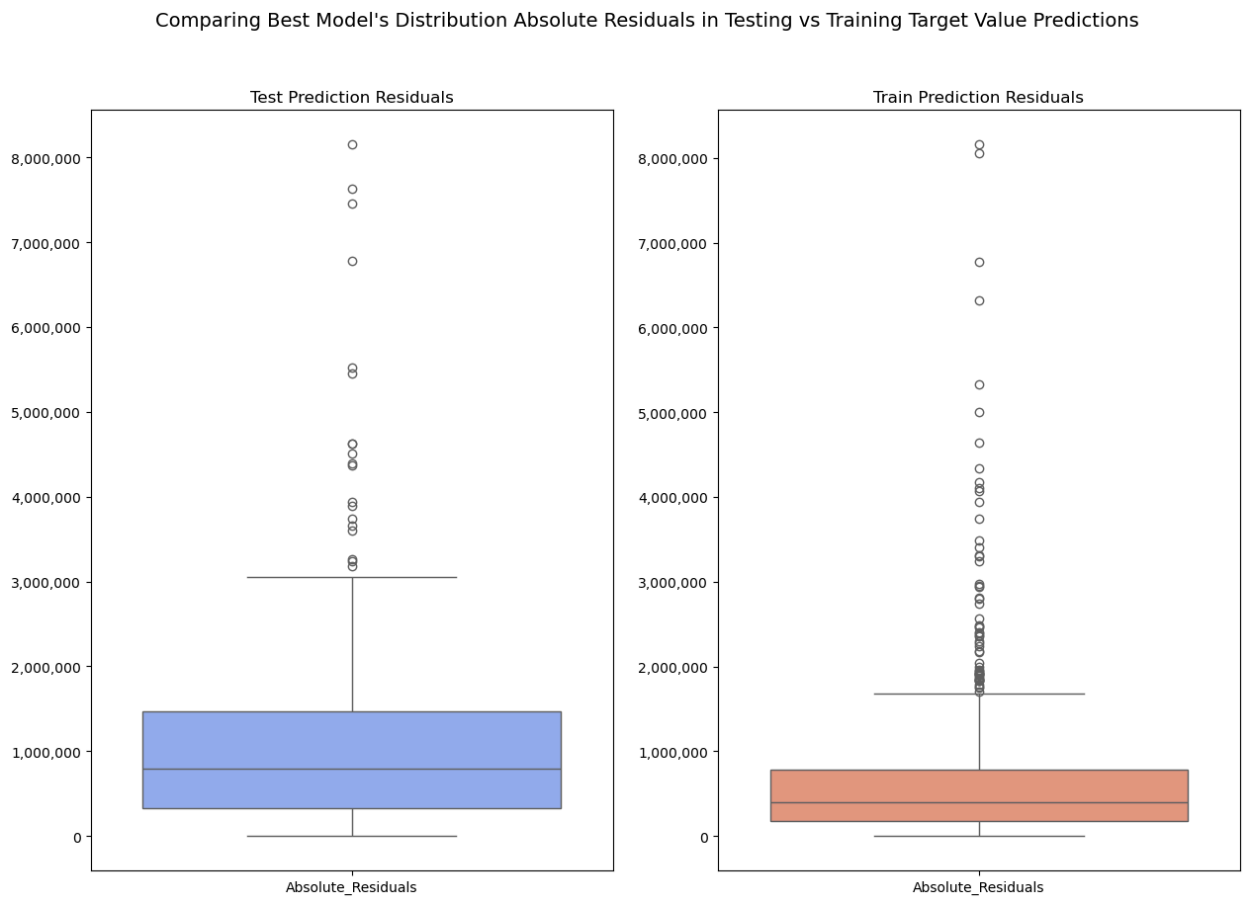
The best random forest model explains about 64% of the variance seen in the data. After some additional fine tuning, the best parameters for this model were as follows:

- bootstrap: True
- criterion: friedman_mse
- max_depth: None
- max_features: 20
- min_samples_leaf: 4
- min_samples_split: 3
- n_estimators: 200
- n_jobs: -1

The best forest model's features and their importance can be seen below:

Feature	Importance
Total Patients	44.05%
Total Weekly Hours of Operation	14.31%
Total Grant Funding	9.23%
Total Sites	4.27%
Uninsured Ratio	3.59%
Diverse Race Proportion	2.33%
Total Other Revenue	1.99%
Zip Code Count	1.94%
Site City Count	1.74%
Funding MHC	1.54%
Medicaid Ratio	1.39%
Private Insurance Ratio	1.25%
Hispanic Proportion	1.25%
FPL 100 Below Ratio	1.11%
Grants to Revenue Ratio	1.02%
State Frequency Encoding	0.96%
Funding HO	0.95%
Medicare 18 Up Ratio	0.95%
Poverty Ratio	0.92%
Total 18 Up Ratio	0.88%

Of the test set predictions, 50% were off by 800,000 or less. The average absolute residual is approximately 1.1 million, which indicates the average error magnitude between the predicted and actual funding. The smallest residual was off by 631. The worst prediction was off by 8,153,076. Below are box plots demonstrating the distribution of absolute residuals between the test predictions and the training predictions.



Future Improvements

State-Level Data: I joined the residuals data with the original test data. I then grouped the results by state. Different states show different levels of prediction accuracy. For example, states like FL and GA show very high mean residuals and a broad standard deviation, suggesting potential model inconsistencies or unique state-related factors not captured by the model.

Some states like MS have relatively lower mean residuals, which could be due to more consistent data, fewer outliers, or features that align well with the model's strengths. In contrast, states like RI and TN show very high residuals, which might be worth investigating for data anomalies or specific regional characteristics affecting funding. States with the highest residuals or greatest variability might benefit from a more in-depth, state-specific analysis to tailor the model or address unique local factors.

Also, not all states were represented in the test data or the training data. Ensuring representative proportions in each split may improve future model performance.

A level of complexity not currently captures are state-related factors. For example future iterations of this project could include state-level demographic and fiscal data. For example, including things like the state uninsured rate, population, low income, and diversity as well as the amount of Medicaid/Medicare funding could further segment data into groups.

Additional Entity Data: Additionally, I found another dataset from HRSA that provides more details about each entity, including:

- Detailed site categories (hospital, school, other clinic)
- HRSA region
- Location types (permanent, seasonal, mobile van)
- Dates for when sites were added to scope
- Organization type (Federal Tax Exempt of U.S. Government entity)
- Proximity to U.S.-Mexico border

I am particularly interested in the dates field because this could inform how long agencies have received HCP funding for. Potentially the agencies that receive the most funding have had the funding for longer periods of time.

Clustering Methods: One of the challenges I experienced was not being able to find any major groupings between health centers. Another future improvement could be to conduct KMeans clustering as an unsupervised method to group the data.

Classification: One of my thoughts would be to turn this into a classification problem instead of a regression problem. The exact amount of funding is less important than gauging a general range of funding an entity could potentially receive.

Other Experts: Finally, I could meet with other industry experts, especially policy makers who are more knowledgeable about Health Center Program funding and get their thoughts on the biggest factors that influence the funding an entity receives.

Credits & Thanks 🤗

I want to thank my mentor Jyant Mahara for coaching me throughout this process and making it feel less daunting.

I want to thank Eric Battis for taking time to help me prioritize which features to start with for this modeling project.

Also, thank you HRSA for providing such great data!