ACA Healthcare Enrollments Across United States

Analyzing Influencers, Trends, and Predictions of ACA Healthcare Plan Enrollments

Jie Shen
Master in Data Science
University of Colorado, Boulder
San Mateo, CA, USA
shenjiejie2017@gmail.com

ABSTRACT

Healthcare and healthcare costs are among the most pressing issues in the United States. Using Affordable Care Act related data from CMS.gov (U.S. Centers for Medicare & Medicaid Services), including Issuer Level Enrollment Data (2017–2022) [1] for the Federally-facilitated Marketplace (FFM) and Qualified Health Plan Choice (QHP) Premiums of 2024 [2], the project analyzed across-state patterns and trends in enrollment, demographic characteristics, and premiums. Additionally, the project investigated key factors contributing to ACA healthcare enrollment and reasonably predicted the average 2022 enrollments across states.

The project identified several key trends in health insurance enrollments across states. Enrollment numbers vary significantly between states, and the distribution of enrollments under different issuers is highly right-skewed. States with the highest premiums or significant premium increases tend to have lower total enrollments. Additionally, when predicting enrollments, several factors contribute significantly beyond prior year enrollments. These include the issuer, county-wise total enrollment, length of consumer stays, Federal Poverty Level (FPL) ratio, age ratio, and smoker ratio.

INTRODUCTION

My work experience and problems of healthcare in U.S.: In my role as an actuarial analyst, I have spent many years calculating the funds required by government sector clients—such as cities and counties—to cover retiree healthcare costs. Through this work, I have observed the significant burden that rising healthcare premiums place on these clients. This issue is part of a broader national challenge, as the United States consistently exhibits the highest healthcare costs per GDP globally.

Importance of Monitoring Healthcare Enrollment: Healthcare premiums are intrinsically linked to claim costs, the size of the risk pool, the demographic characteristics of the insured population, and various other factors. One critical metric to monitor in this context is enrollment. Large and stable enrollments benefit insurance companies by expanding the risk pool, which facilitates better risk management and potentially reduces premiums. Conversely, low enrollment rates or significant drops in enrollment within a state may indicate problems such as high premium cost, insurance accessibility, socioeconomic disparities, and many others that should concern the societies.

Existing Solutions: The United States has continually pursued diverse methods to boost healthcare enrollment. These include outreach programs, navigator assistance, streamlined processes, targeted marketing, community partnerships, financial aid, and website improvements. These evolving strategies address this crucial challenge by raising awareness, simplifying enrollment, targeting underserved populations, providing assistance, and improving affordability and accessibility of healthcare coverage

Potential Contribution of This Project: This project, grounded in data mining, aims to provide a different perspective on addressing U.S. ACA healthcare enrollment challenges. Through data mining, it seeks to understand enrollment patterns and contributing factors. Utilizing predictive modeling, the project further endeavors to offer insights into potential future trends, identify key influencers, and evaluate the predictability of enrollments. It is hoped that these insights might inform policymakers' efforts to enhance healthcare accessibility and lower healthcare premium costs from a fresh angle.

RELATED WORK

So much exciting prior work has been done regarding healthcare. Here are just a few examples:

Enhancement of Healthcare Project: For example, The ATHLOS Project [3], funded by the EU's Horizon 2020 Research Program, aims to bolster preventive healthcare using longitudinal studies. These studies offer rich yet complex datasets marked by high volumes and missing values. To tackle these challenges, the project employs Machine Learning, Data Mining, and Data Imputation models. Specifically, it focuses on developing a methodology to interpret aging's impact on health, particularly the Health Status (HS) score, estimating human health. Evaluating various data imputation models, the study highlights their role in enhancing prediction models' efficacy in preventive medicine.

Information from Medical Text Records Project: In the Extraction of Disease Factors from Medical Texts project [4], authors introduce EDFI, a technique enhancing disease factor extraction. It utilizes term proximity information to boost existing techniques without altering their core. Through case studies on medical texts, EDFI markedly enhances diagnosis factor ranking, advancing knowledge-based systems for healthcare decision-making and education. Results confirm its significant improvement over existing methods, suggesting promise for broader application in automatic text annotation of disease factors.

Anomaly Detection project: The Machine Learning Techniques Applied to Anomaly Detection in EGG Signals Project [5] integrates sparse representation of ECG signals with neural network classification. It uses adaptive decomposition with OMP algorithms and QRS complex detection with a neural network. Optimal results were achieved with OMP decomposition and a modified K-SVD dictionary, feeding a three-layer neural network. Evaluation using MIT-BIH Arrhythmia Database affirmed high efficiency in anomaly detection, challenging existing notions on neural network generalization.

My Work Based On Prior Works: These projects showcase the impactful applications of data mining in healthcare. From enhancing preventive medicine to improving disease factor extraction and effective anomaly detection, they highlight the transformative potential of advanced data analysis methods in improving healthcare outcomes.

My work takes inspiration and learnings from those prior work in data mining, and then it approaches the topic from a simpler and more personalized perspective, driven by my own interests and a distinct angle. This project focuses on a unique aspect of healthcare: healthcare enrollment. It aims to investigate the factors influencing enrollment, analyze the relationships between these factors, and uncover trends and patterns across states. Additionally, this project seeks to predict future enrollments. Hopefully, this approach could allow for a fresh examination of healthcare enrollment dynamics, offering insights that complement and build upon existing research in the field.

PROPOSED WORK

Datasets: All the 7 datasets are from the U.S. Centers for Medicare & Medicaid Services (CMS). There are two main types of data:

Issuer level enrollment data from 2017 to 2022 [1]: The CMS has prepared public data sets of issuer level enrollment information for the Federally-facilitated Marketplace (FFM). and enrollment platform. The Federally Facilitated Marketplace (FFM) refers to a health insurance exchange established by the United States federal government as part of the implementation of the Affordable Care Act (ACA), also known as Obamacare.

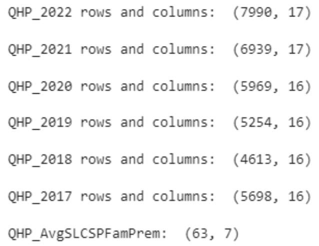
For our project, we chose to use 6 year datasets from 2017 to 2022 (the most current available year data). Each record is for a specific state, county (indicated by county's FIPS code), and issuer (indicated by an issuer's HIOS ID, which is County specific). For each year, there are around 4500 to 8000 records with unique State, County FIPS code, and issuer HIOS ID.

The datasets have the following data fields: State, County (FIPS code), issuer (HIOS ID), average monthly Enrollment (for 2021 & 2022 year), ever enrolled (for 2017 to 2020 year), average number of months enrolled Per Consumer (for 2021 & 2022 year), number of consumers with different Household Income (FPL) (<138%, 138%-250%,250%-400%, above 400%, and unknown), number of consumers with different age ranges, number of consumers with each genders, and number of tobacco use consumers.

QHP 2024 premiums data [2]: The CMS has also prepared public data sets for yearly Qualified Health Plan Choice (QHP) premium information. The QHP plans are plans from HealthCare.gov Marketplaces. Healthcare.gov is the official health insurance marketplace website for the United States, established as part of the Affordable Care Act (ACA), also known as Obamacare. For plan year 2024 (PY24), there are 210 Qualified Health Plan (QHP) issuers in HealthCare.gov Marketplaces. However, to simplify, the 2024 premium I use for this project is only the average county 2024 premiums for specific scenarios - SLCSP QHP premium for a family of four by state and county FIP code. There are only 63 records here. Each record has a unique state and FIPS code.

This dataset has the following data fields: State,County FIP codes, 2024 year average SLCSP premium for a family of four,.premium increase from calendar year 2023 to 2024, 5 year premium increase from calendar year 2020 to 2024.

Combine Data: The project combined 7 datasets (2017-2022 year enrollment data and 2024 year premium data) using key fields: State, FIPS code, and Issuer HIOS ID. Here is a glance of individual data sizes:



```
QHP_2022 rows and columns:  (7990, 17)

QHP_2021 rows and columns:  (6939, 17)

QHP_2020 rows and columns:  (5969, 16)

QHP_2019 rows and columns:  (5254, 16)

QHP_2018 rows and columns:  (4613, 16)

QHP_2017 rows and columns:  (5698, 16)

QHP_AvgSLCSPFamPrem:  (63, 7)
```

Figure 1

Here are a few challenges and key steps:

- *Limited premium data*: Only 63 records of premium data available compared to over around 4500 to 8000 records of issuer level enrollment data for each year. Further investigation was done and found that those SLCSP QHP family premiums are in general only for a few issuers in each county and only for some counties.
- *Inconsistencies in data fields (data field name, content, and data type) across different data sources*: such as 'Issuer HIOS ID' and 'County FIPS Code' formatting ('2020' vs. '02020', vs. '2020.0'). Careful investigation was done to ensure consistency and correct data combining. For example, 'County FIPS Code' was converted to integer type.
- *Changes in data methodology after the year 2020*: with the introduction of the 'Average Monthly Enrollment' field replacing 'Ever enrolled'. Carefully handle this by not combining data fields that are actually different in different years.
- *Non-numerical values in some count data fields:* Those fields were originally in string type and there are lots of missing values indicated as "unknown" or "*" in those counts fields. Replaced those missing values with NA and converted those fields to numeric types. We'll talk about missing values in a later section.

Create Ratio Demographic Features*:* The dataset has counties of varying sizes. So for prediction, direct comparisons between counts (such as number of smokers, number of males) are less meaningful. For prediction, adding for example female and male counts will get total enrollments. So, it also doesn't make sense to use those as features.

The project transformed those categorical counts (such as gender, age, income, and smoking status) into ratios. This allows for more intuitive comparisons and facilitates predictive modeling without redundancy.

Create Total Enrollment Features: The project created 3 total enrollment features under an issuer, county, and state. Those total enrollment features offered valuable insights into the dynamics of pool size, which directly impact risk management and enrollment stability. Recognizing the predictive potential of previous year total enrollment data, the project incorporated these features into both the exploratory data analysis (EDA) and prediction processes. Notably, the introduction of total enrollments, particularly 2021 total county-wise enrollment, has revealed a substantial correlation with 2022 enrollment patterns, significantly enhancing the predictive accuracy.

Data Warehousing*:* The project structured the data differently to suit distinct purposes:

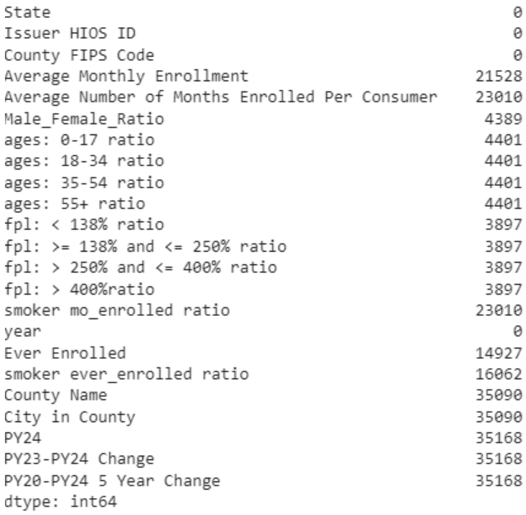
Data for EDA: To facilitate exploratory data analysis (EDA), we introduced a "year" column, enabling the examination of trends over time across various features. This involved stacking different years' data together to comprehensively analyze temporal patterns. By aggregating data at the state and year levels, as well as solely at the year level, we gained insights into the evolving trends of state-specific features over the years and the overall trends across the entire U.S. This approach allowed for a nuanced understanding of how different variables fluctuate over time and provided valuable context for interpreting the dataset's dynamics and informing subsequent analyses and predictions.

Data for prediction: For predictive modeling, each year's data serves as a feature. Rather than stacking different yearly datasets, each year's data is appended as separate features (columns). Additionally, although the goal is to predict 2022 average state-wise enrollments, the project does not aggregate data across states before modeling. This decision ensures an adequate amount of data for accurate predictions. Consequently, prediction tasks are first conducted at the granular level of each state, county (FIPS code), and issuer (HIOS ID) combination. And then, predictions are aggregated to derive state-level averages or other relevant metrics post-modeling.

Missing Values*:* There are lots of missing values and they originated from two distinct sources.

From individual data: For example, certain issuers under certain counties lack specific information such as gender, income, or smoking status.

*From data integration***:** For example, premium information may be absent for many counties and issuers. Also, there may be instances where enrollment data for a specific issuer under a certain county is available for one year but not for another year. It is common for issuers available in one year to be unavailable in another year. In fact, there are around 21% of issuers in 2022 not present in 2021.



```
State                                               0
Issuer HIOS ID                                      0
County FIPS Code                                    0
Average Monthly Enrollment                      21528
Average Number of Months Enrolled Per Consumer  23010
Male_Female_Ratio                                4389
ages: 0-17 ratio                                 4401
ages: 18-34 ratio                                4401
ages: 35-54 ratio                                4401
ages: 55+ ratio                                  4401
fpl: < 138% ratio                                3897
fpl: >= 138% and <= 250% ratio                   3897
fpl: > 250% and <= 400% ratio                    3897
fpl: > 400%ratio                                 3897
smoker mo_enrolled ratio                        23010
year                                                0
Ever Enrolled                                   14927
smoker ever_enrolled ratio                      16062
County Name                                     35090
City in County                                  35090
PY24                                            35168
PY23-PY24 Change                                35168
PY20-PY24 5 Year Change                         35168
dtype: int64
```

Figure 2

Above is the missing value output for data before aggregation. The good news is that at the individual level, there are no missing values for 'Average Monthly Enrollment' and 'Ever Enrolled'. To illustrate, we have 14,927 missing values for 'Ever Enrolled'. Those are the total records of 2021 & 2022 (14,929 records in raw data minus 2 records with non-numerical FIP codes that were deleted). Recall that the 'Ever Enrolled' field was replaced by "Average Monthly Enrollment" in the year 2021, so year 2021 & 2022 records are all supposed to have missing values in 'Ever Enrolled'. Similarly, we have 21528 missing values of 'Average Monthly Enrollment', which is the number of records from 2017 to 2020. There are no missing values in 'Average Monthly Enrollment'.

For EDA and for prediction, the project dealt with missing values differently.

Missing values in EDA and aggregations:

The project refrained from filling in those missing values for EDA because preserving the true distribution and characteristics of the original dataset for EDA is crucial. This approach ensures that any patterns or insights gleaned from the analysis are based on the actual data and not influenced by imputation methods.

Furthermore, when aggregating data, such as calculating averages for a state, the project carefully distinguishes between missing values and zero values. Missing values are typically excluded from calculations to avoid skewing the results, while zero values may hold significance and should sometimes be included in the computation.

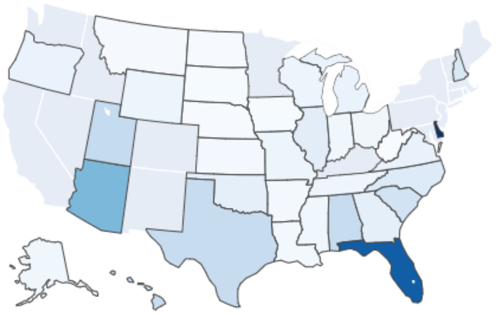
Missing values in prediction: In prediction, machine learning models typically cannot handle missing values directly, requiring them to be filled or excluded before modeling. Fortunately, at the individual level, there are no missing values in the enrollment field ('Ever Enrolled' in 2017 to 2020 and 'Average Monthly Enrollment' in 2021 & 2022). Because the goal is to project 2022 enrollment, records without 2022 enrollment are automatically excluded during the preparation of prediction data by starting with 2022 data and left joining other years' data.

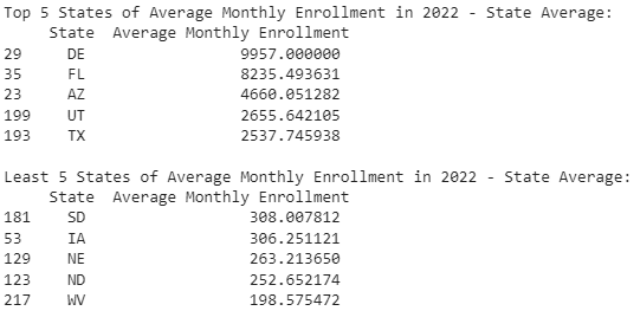
For other fields, such as demographic features, many counties or county/issuer combinations within each state have missing values. To address this, the project first fills missing values with the state average. Next, for states where all missing values in these fields occur (amounting to only 37 records), the rows are deleted.

Statistical Analysis and Findings: Various statistical analysis and visualizations were performed and the following and some key analysis and findings:.

Enrollment Patterns and Trends:

- For 2022, there are 33 states (and 2448 Counties) in data. For the average state enrollments, state to state are quite different, with highest of almost 10,000 and lowest only around 200 in 2022. Below is the state-wise 2022 total enrollment map, the darker the color, the higher the average enrollment. Also, the top and least 5 states.



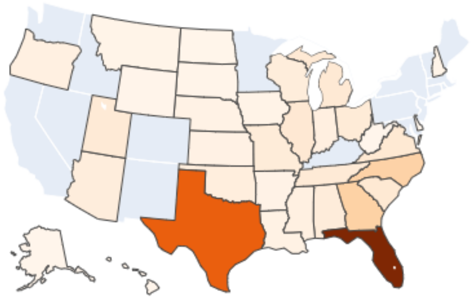


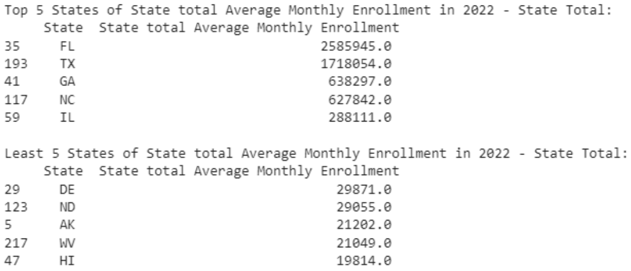
```
Top 5 States of Average Monthly Enrollment in 2022 - State Average:
     State  Average Monthly Enrollment
29      DE                 9957.000000
35      FL                 8235.493631
23      AZ                 4660.051282
199     UT                 2655.642105
193     TX                 2537.745938

Least 5 States of Average Monthly Enrollment in 2022 - State Average:
     State  Average Monthly Enrollment
181     SD                  308.007812
53      IA                  306.251121
129     NE                  263.213650
123     ND                  252.652174
217     WV                  198.575472
```

Figure 3

- For the 2022 total state enrollments, state to state are quite different too. with the highest of almost 2.5 million and lowest only around 20,000. Likewise, below is the 2022 state-wise total enrollment map and the top and the least 5 states.





```
Top 5 States of State total Average Monthly Enrollment in 2022 - State Total:
     State  State total Average Monthly Enrollment
35     FL                              2585945.0
193    TX                              1718054.0
41     GA                               638297.0
117    NC                               627842.0
59     IL                               288111.0

Least 5 States of State total Average Monthly Enrollment in 2022 - State Total:
     State  State total Average Monthly Enrollment
29     DE                                29871.0
123    ND                                29055.0
5      AK                                21202.0
217    WV                                21049.0
47     HI                                19814.0
```

Figure 4

- For total U.S. average and total enrollments, the patterns are similar in other years.
- For overall enrollment trends in the U.S., and for both state total and state average, there are no obvious fluctuations from 2017 to 2022 in enrollment. The enrollment remained relatively stable during this period. However, there are some states having enrollments in one year but not in another (Ex: NV, NE, PA, NJ in 2017 but not in 2022)
- There are 213 unique issuer IDs in 2022 data, for total enrollment under specific issuer ID, the distribution is quite right screwed with many around 15,000 enrollment and some big outlier enrollment of above 700,000. Below is the distribution for the year 2022. The 25% is around 4,300, the median is around 14,000, 75% is around 45,000 and the maximum is around 750,000. This pattern is very similar in all the years (2017 to 2022).

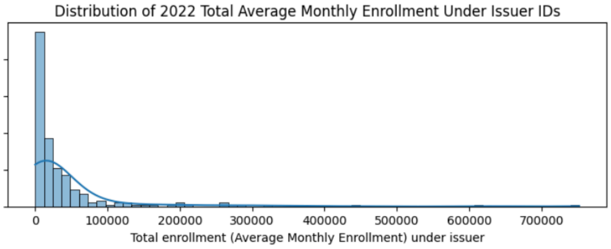
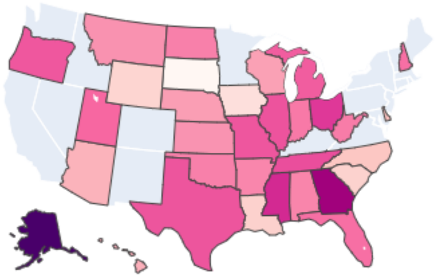
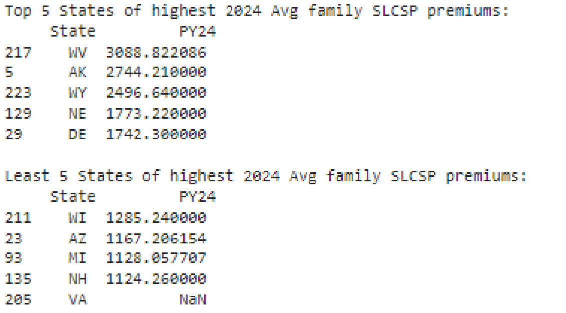


Figure 5

QHP 2024 SLCSP Family Premiums

QHP 2024 SLCSP state average family premiums are quite different also, with highest almost $3,089 and lowest around $1,125 per month. Below is the map for 2024 average family premiums. Also, the top and least 5 states.





```
Top 5 States of highest 2024 Avg family SLCSP premiums:
     State         PY24
217     WV  3088.822086
5       AK  2744.210000
223     WY  2496.640000
129     NE  1773.220000
29      DE  1742.300000

Least 5 States of highest 2024 Avg family SLCSP premiums:
     State         PY24
211     WI  1285.240000
23      AZ  1167.206154
93      MI  1128.057707
135     NH  1124.260000
205     VA          NaN
```

Figure 6

Factors contributing or correlated to healthcare enrollment:

The project used visual and statistical analysis, heat map between features (Spearman), and modeling to analyze relationship between features and the enrollments. Here are some key findings:

- As expected, from the heat map and modeling, the prior year enrollments are key influencers. With the closer the year, the more influence.
- It seems lots of states in the top 5 states with the highest premium or premium increase are states with lower enrollments. So, there might be a negative correlation between premium or premium increase and enrollments.
- From the visualization, top 5 analysis, and modeling, there is a negative relationship between the average months

consumers stay and enrollment. For example, the Spearman correlation is -0.3. This could mean that consumers with more options tend to stay for a shorter duration and may switch plans more frequently.

- Spearman correlation of some demographic features to healthcare enrollment :
  - Age ratios: Age ratios in all ranges have good correlation with enrollment,, with age 17 (0.32) age 18-34 (0.49!), age 35-54 (-0.25), and age 55+(-0.24). It could mean the main consumers for the affordable plans are young people from 18-34 and kids.
  - Smoker ratios: smoker ratio has a big correlation of 0.41 with enrollment.
  - Percent FPL ratios: FPL < 138 has a good positive correlation of 0.22 to the enrollment. This is expected because the plan is supposed to serve lower income people. FPL 138 to 250 has a good negative correlation of -0.3 to enrollment. This could not be the target consumers. However, FPL 400+ also has a good positive correlation of 0.25. This is a bit unexpected because those are richer people.

Model Selection:

Chosen of tree models: Data analysis and visualization, primarily through scatter plots of features and the response variable, reveal that most relationships are clearly non-linear. This makes linear regression models unsuitable for our data. Tree models, however, are well-suited for this scenario as they excel in capturing non-linear relationships through recursive partitioning and approximation on small segments of the data.

Additionally, determining the relevant features to include based solely on data analysis in this data is especially challenging. Tree models address this issue effectively by handling irrelevant features well. By tuning hyperparameters like tree depth, tree models can exclude unimportant features gracefully, making them a robust choice for our analysis.

Support Vector Machines (SVM) and deep learning are other options, but they have limitations. SVMs are better suited for classification, require feature scaling, and are sensitive to categorical feature labeling, making them less ideal for this dataset with features like State, County code, and Issuer IDs that have many unique values. Deep learning models, while powerful, maybe too complex for our needs.

Data Preprocessing for modeling:

Feature scaling: An advantage of choosing tree models is that they do not require feature scaling since they are not distance-based. Thus, we will start with tree models and consider other options if necessary.

Categorical variable encoding: The only string type is "State." The project used label encoding based on the state-wise average of the dependent variable values.

Outliers: Tree models are usually robust to outliers. Only one extreme record was excluded.

Hyperparameters tuning: For each model, we performed 4-fold cross-validation to identify the optimal hyperparameters. To enhance the efficiency of the search, we employed random search instead of grid search. The results indicated that tree models perform better with deeper depths. This improved performance could be attributed to the fact that, except for the "State" variable, all features are numerical.

EVALUATION

Models used**:** As discussed in model selection, only tree models are used. Simple decision tree isn't chosen because trees by themselves are weak learners and tend to overfit easily. Instead, we'll use the following ensembling trees:

Random Forest: utilizes a bagging ensemble method. It randomly samples a subset of training data with replacement, a technique known as "Bootstrap," grows a tree, and then aggregates the results. The essence of Random Forest lies in two key strategies: bagging, involving the random sampling of data, and decorrelation, achieved through the random sampling of features.
Gradient Boost: unlike Random Forest, employs the Boosting ensemble method, which emphasizes different weights on trees to enhance each tree's learning capability. Typically, Boosting trees begin with an initial function, f(x), and iteratively update weights based on performance errors. Ultimately, the results from various trees are aggregated based on their respective weights to derive the final outcome.

In the case of Gradient Boosting, it diverges from other Boosting methods such as AdaBoost by utilizing the negative gradient of a loss function rather than solely fitting the residuals. This approach is advantageous because it efficiently selects the direction of steepest descent in terms of reducing the loss function. As a result, Gradient Boosting optimizes the model's performance by iteratively improving upon the shortcomings identified during each iteration.

XGBoost: another tree model utilizing the Boosting method, shares similarities with Gradient Boosting but offers several distinct advantages. It provides built-in support for regularization techniques like L1 and L2 regularization, which effectively prevent overfitting by penalizing complex models. Moreover, XGBoost is engineered for speed and efficiency, leveraging various optimization techniques such as parallelization and cache-aware computing. These optimizations enable XGBoost to handle large datasets and complex models with exceptional efficiency and scalability.

Evaluation Method, Metric: Our task involves predicting the average enrollment for each state. Aggregating data by state initially would result in insufficient data for prediction, given the limited number of years and states. Thus, we maintain our data un-aggregated, ensuring a robust dataset with around 4500 to 7000 records for each year from 2017 to 2022, providing ample data for prediction.

Following prediction on each county and issuer combination, we aggregate the results to derive state-wise averages. We then compare the actual and predicted state averages using appropriate metrics.

For our regression problem, metrics like mean squared error or mean absolute error serve as suitable evaluation measures. We've chosen mean absolute error due to its intuitive interpretation, providing a clear understanding of the exact discrepancies in averages. Furthermore, its resilience to outliers, compared to MSE, ensures a more robust assessment of our model's performance across all states.

Model Performance:

Overall, the model performances are good, with Random Forest exhibiting the strongest performance among the tree models (~80 MAE). This superiority can be attributed to the inherent complexity of gradient boosting and XGBoost algorithms, which feature a larger number of hyperparameters and are more susceptible to overfitting. Below are the mean absolute error figures for all three models, along with a visualization illustrating the actual versus predicted state average enrollment for the year 2022, specifically using the Random Forest.

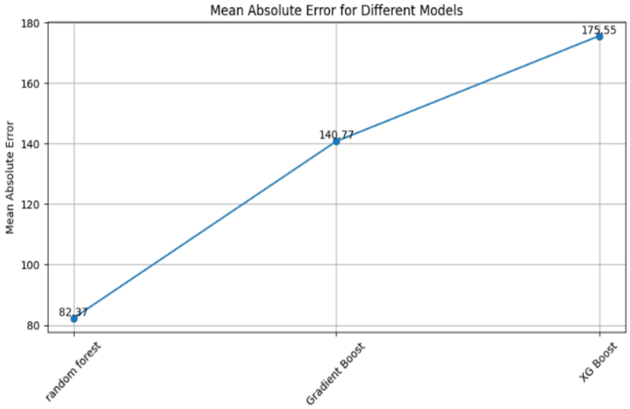


Figure 7

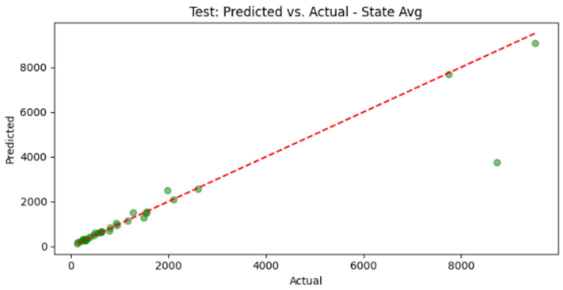


Figure 8

Feature importance: For tree-based models, feature importance was conducted. For instance, below is the feature importance of the Random Forest. The feature importance is consistent with EDA observation. Notably, prior year enrollment emerged as a top contributor. County-wise total enrollment, FPL ratios, duration of consumer stays, age ratio, and smoker ratios also featured prominently, aligning with our EDA findings.

However, Issuer ID is a top contributor in tree but not in EDA. The reason can be the large amount of issuer data available across years (around 300 for each year), allowing the model to gain significant insights during training. Interestingly, while our EDA indicated a relationship between premiums and enrollment, premiums did not emerge as a key contributor in trees. This reason can be the limited availability of 2024 premium data (only 63 records for county and issuer combinations). This sparsity of data limited the model to learn valuable information.

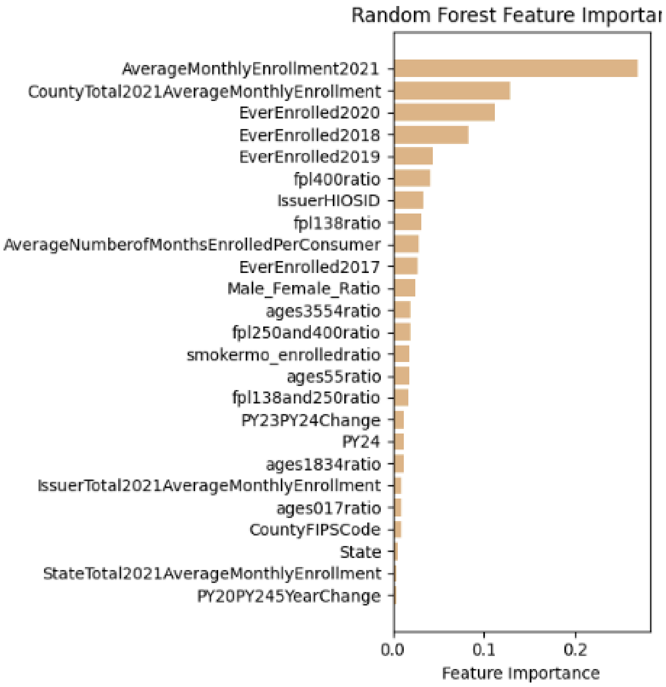


Figure 9

DISCUSSION

Timeoline: The project deadline is 5 weeks, but I aim to complete it within 3 weeks to be safe

- Week 1: project topic, dataset
- Week 2: EDA and a quick model run
- Week 3: final report
- Glady, now is 3rd week and I am on track

Potential Challenges and learnings**:** At the beginning of the project, a couple of challenges emerged. Merging data from different years, integrating enrollment and premium data, and navigating methodology changes in enrollment data posed complexities. Missing data, particularly in demographic information, alongside inconsistencies in available identifiers across years, added further hurdles. Also, enrollment prediction feasibility was hindered by the lack of freely available data.

As an actuarial analyst with years of experience, I've come to recognize that a significant portion of any project revolves around acquiring, merging, and cleaning data, alongside conducting thorough data sanity checks. In this project, I've learned that data mining projects share a similar workflow. Much effort is dedicated to data integration, prompting key questions such as: What data do we need to acquire? What domain knowledge is necessary to understand the data better or to aid in data validation? How can we ensure the accuracy of our data and conduct effective data sanity checks? What insights can we glean from missing data, and how can we address these gaps? Additionally, how can we organize and structure our data effectively to support various analytical tasks? By addressing these questions early on, we can streamline our data processing efforts and set a strong foundation for our project's success.

Alternative Approaches/Backup Plan**:** Despite facing significant challenges initially, I devised an alternative strategy to focus on exploratory analysis, anomaly detection, and unsupervised learning techniques like clustering. However, as the project progressed, it became apparent that the original approach was yielding promising results, negating the need to resort to the contingency plan.

CONCLUSION

Project Summary: Healthcare and healthcare costs are critical issues in the United States. This project utilized data related to the Affordable Care Act (ACA) from CMS.gov, specifically the Issuer Level Enrollment Data from 2017 to 2022 and the Qualified Health Plan Choice Premiums of 2024. The analysis focused on examining across-state patterns and trends in enrollment, demographic characteristics, and premiums. The project also aimed to identify key factors influencing ACA healthcare enrollment and predict the average state enrollments for 2022.

Key findings: The project uncovered several important trends in health insurance enrollments across different states. Enrollment numbers vary widely between states, and the distribution of enrollments among different issuers is highly right-skewed. States with higher premiums or significant premium increases generally have lower total enrollments. Additionally, when predicting enrollments, several factors beyond prior year enrollments were found to be significant. These include the issuer, county-wise total enrollment, length of consumer stays, Federal Poverty Level (FPL) ratio, age ratio, and smoker ratio. These insights help to understand the dynamics of ACA enrollments and the factors that significantly impact them, providing a foundation for better policy-making and strategic planning in healthcare.

Future works:
Issuer ID and County FIPS code were converted to numerical types for data merging and subsequently used in modeling. Ideally, these should be relabeled using label encoding, potentially based on the average dependent variable (2022 enrollment) for Issuer ID or County FIPS code, similar to the approach used for state label encoding.

However, due to time constraints, this update was not implemented in the current project. Tree models, which perform well with non-linear relationships and smaller sections of data, can manage this type of encoding reasonably relatively well. For future work, relabeling these variables could enhance model performance. Additionally, support vector machines (SVM) and deep learning models, which are more sensitive to categorical feature encoding, could be explored to further improve predictive accuracy.

Content knowledge is essential for any data mining project, particularly in healthcare. Without sufficient healthcare knowledge, identifying valuable data sources and understanding complex medical codes and terminology is challenging. Conversely, data mining and machine learning are excellent at detecting patterns, identifying trends, and making predictions, with applications such as increasing diagnostic accuracy, enabling robotic surgeries, identifying drug development candidates, and determining optimal treatments. As someone interested in combining data science with healthcare, I believe it is crucial to deepen my understanding of healthcare while exploring data science applications in this field.

REFERENCES

[1] Issuer enrollment data from CMS.org: https://www.cms.gov/marketplace/resources/data/issuer-level-enrollment-data

[2] Qualified Health Plan Choice and Premiums in HealthCare.gov States form CMS.org: https://www.cms.gov/marketplace/resources/data/qualified-health-plan-choice-premiums-healthcaregov-states

[3] Anagnostou, P., Tasoulis, S., Vrahatis, A. G., Georgakopoulos, S., Prina, M., Ayuso-Mateos, J. L., … Panagiotakos, D. (2021). Enhancing the Human Health Status Prediction: The ATHLOS Project. *Applied Artificial Intelligence*, *35*(11), 834–856.: https://doi.org/10.1080/08839514.2021.1935591

[4] Liu, R. L., Tung, S. Y., & Lu, Y. L. (2015). Extraction of Disease Factors from Medical Texts. *Applied Artificial Intelligence*, *29*(1), 49–65. https://doi.org/10.1080/08839514.2014.962281

[5] Andrysiak, T. (2016). Machine Learning Techniques Applied to Data Analysis and Anomaly Detection in ECG Signals. *Applied Artificial Intelligence*, *30*(6), 610–634. https://doi.org/10.1080/08839514.2016.1193720