ACA Healthcare Enrollments Across States

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

Jie Shen

June, 2024

Content

- Problem Statement
- Executive Summary
- Related Work
- Proposed Work
- Evaluation
- Timeline & Discussion
- Conclusion & Future Work

image from https://www.elastic.co/industries/healthcare

Problem Statement 1

What is the problem? Why it's important?

- High healthcare costs per GDP in U.S.
- Enrollment is a critical metric in healthcare
  - Impacting risk pool size
  - Reflect Insurance accessibility

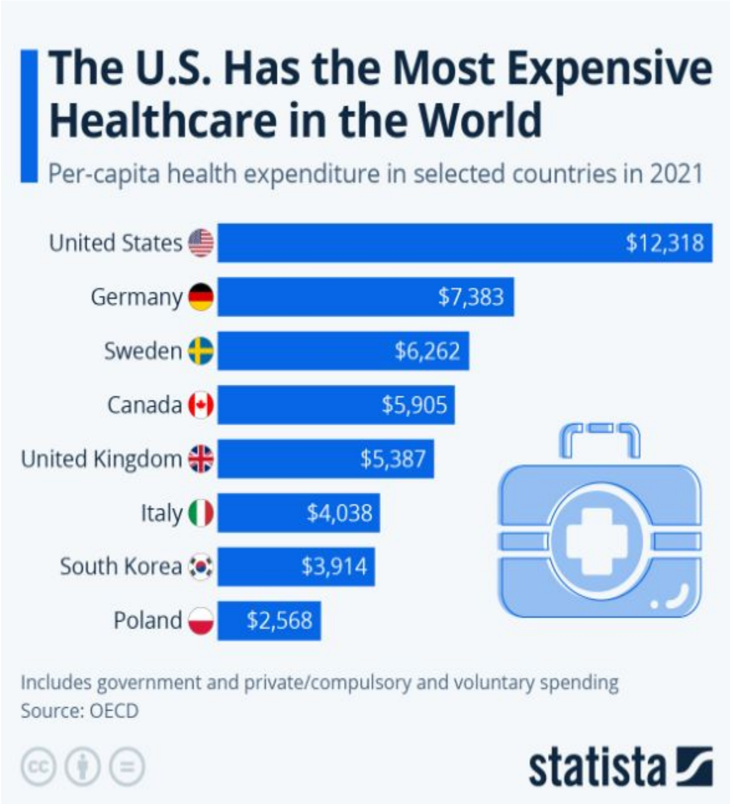


Image from https://www.statista.com/chart/8658/health-spending-per-capita/

Problem Statement 2

Project Aims:

- Discover trends and patterns of enrollments and identify key factors influencing these patterns
- Investigate the main contributors and correlated features of enrollment to understand what drives changes in enrollment numbers
- Build models to predict 2022 enrollment counts

Executive Summary

- **Key Enrollment Trends and Patterns**
  - Enrollment numbers vary significantly across different states.
  - Enrollments under different issuers are highly right-skewed.
- **Main Contributors and Correlated Features**
  - Prior year enrollments
  - Premiums
  - Issuer
  - County-wise total enrollments
  - Length of consumer stays
  - Federal Poverty Level (FPL) ratio
  - Age ratio
  - Smoker ratio
- **Model Predictions**
  - Tree models, MAE, cross-validation, feature importance

Image from https://www.statista.com/chart/8658/health-spending-per-capita/ [1]

Related Work

- **Enhancement of healthcare**: ATHLOS project [2]
- **Information from medical text records**: Extraction of disease factors from medical text project [3]
- **Anomaly Detection**: Machine Learning Techniques Applied to Data Analysis and Anomaly Detection in EGG Signals Project [4]
- **My Work:** Inspired by prior work but unique angle - healthcare enrollment

Proposed Work 1 - Data Sources & Data Integration

- **Data Sources**
  - 2017-2022 Issuer level enrollment data & 2024 QHP Avg. family premium
  - All from ACA (know as Obama Care) healthcare plans from CMS.gov (U.S. Centers for Medicare & Medicaid Services)
- **Data Integration:**
  - Merge 7 data sources (6 year enrollment data & 1 premium data)
  - Inconsistencies in data fields across data files (Ex: “2020” vs. “02020” vs. “2020.0”)
  - Data methodology change: ‘Ever Enrolled’ changed to ‘Avg. Monthly Enrolled after 2020

Proposed Work 2 - Data Warehouse & Missing Values

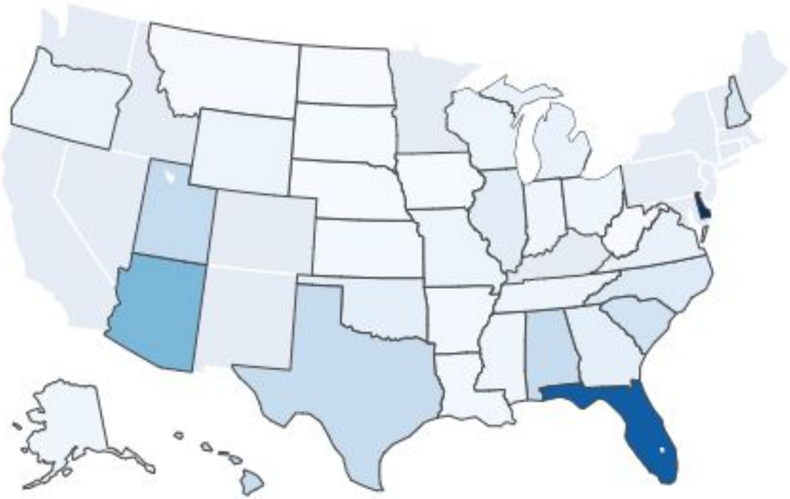
- **Data Warehouse**
  - **Data for EDA**: Added "year" column to analyze trends by stacking yearly data, aggregated at state and year levels
  - **Data for Prediction**: Treated each year as separate features

- **Missing values:**
  - **From 2 Sources**: Individual data & data integration
  - **In EDA**: Not to fill to ensuring accurate analysis
  - **In Prediction**: Model cannot accept . First fill with state averages, then deleted rows with entirely missing state data
  - **In Aggregation**: zero vs. missing values treated differently in avg. calulation

Proposed Work 3 - Enrollment State-wise Variations

- **Enrollments varies a lots across states:**
  - 2022 Avg. state enrollments range from ~200 to ~10,000
  - 2022 Total state enrollments range from ~20,000 to ~2.5 million
- **2022 Top 5 Highest states**
  - Avg - DE,FL, AZ, UT, TX
  - Total - FL, TX, GA, NC, IL
- **2022 Top 5 lowest states**
  - Avg - SD, IA, NE, ND, WV
  - Total - DE, ND, AK, WV, HI

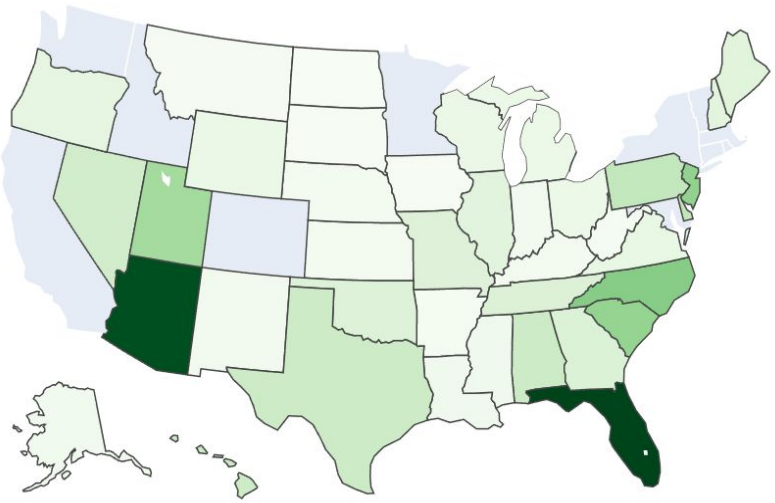
2022 Avg. State-wise Enrollments



Proposed Work 4 - Enrollment Trends Over Years

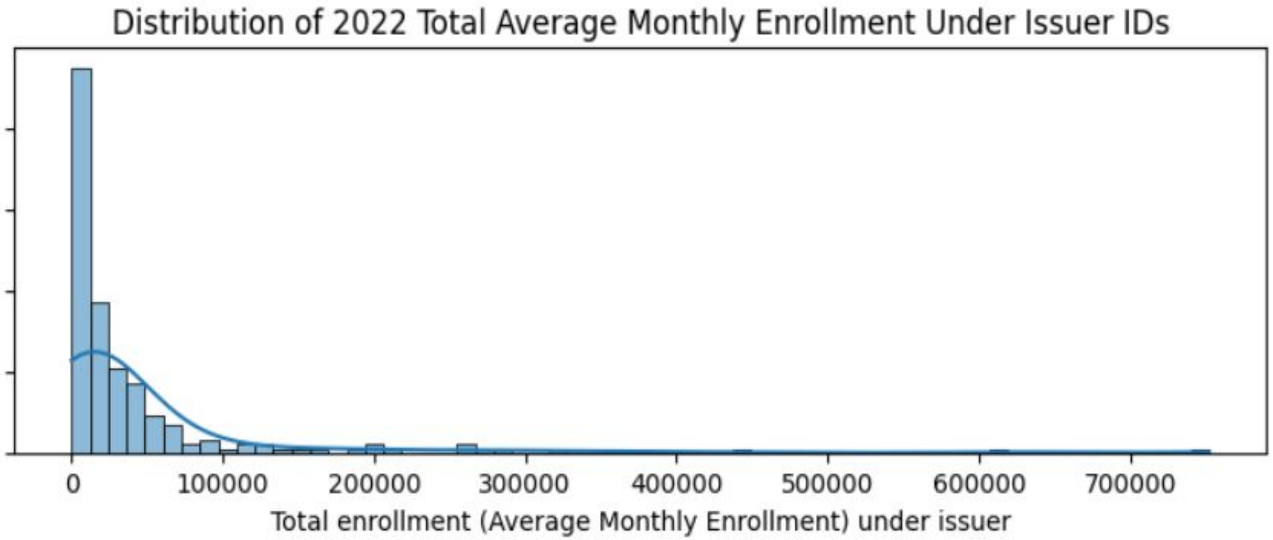
- Both state average and state total enrollments remain relatively stable across yrs.
- Similar patterns observed across all years with slight changes
- There are some states having enrollments in one year but not in another (Ex: NV, NE, PA, NJ in 2017 but not in 2022)

2017 Avg. State-wise Enrollments



Proposed Work 5 - Enrollments Under Issuers

- Distributions of total enrollments under issuers are quite right-skewed across all years
- Many around 15,000 enrollments.
- Outliers with enrollments above 700,000.
- Quartiles: 25% (~4,300), Median (~14,000), 75% (~45,000), Maximum (~750,000).



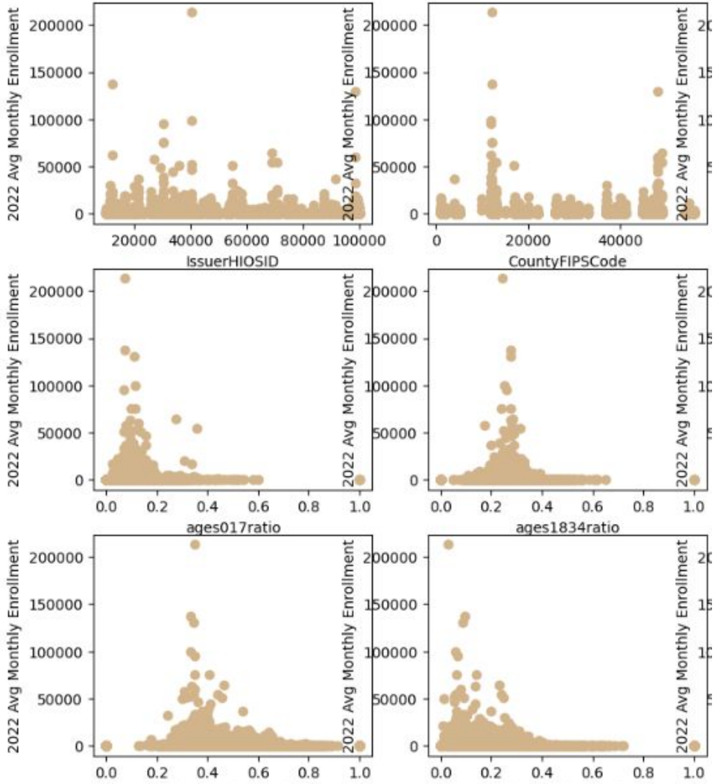
Proposed Work 6 - Factors Influencing Enrollments & Correlations

- **Prior year enrollments**: strongly influence, especially recent years
- **Premiums:** low enrollment states usually have high avg. premiums
- **Avg. months consumers stay**: Neg. Spearman correlation of -0.3
- **Demographic (Spearman) Correlations with Enrollment**:
  - Age Ratios: positive correlation with young ages 17 (0.32), 18-34 (0.49); negative correlation with older ages 35-54 (-0.25), and 55+ (-0.24).
  - Smoker Ratios: positive correlation (0.41).
  - Percent FPL Ratios: positive correlation for FPL < 138 (0.22) and FPL 400+ (0.25)*; negative for FPL 138-250 (-0.3).

*A little unexpected and may need more investigation

Proposed Work 7 - Model Selection

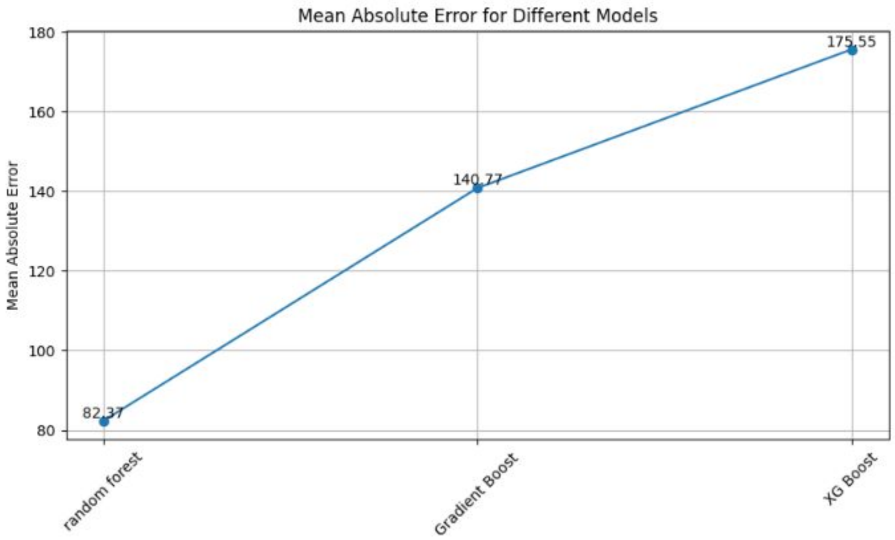
- Based on data, tree based models are chosen:
  - nonlinearity
  - relevant vs. irrelevant of features not obvious
- Random Forest, Gradient Boost, XGBoost (no scaling, insensitive to outliers, relatively insensitive to label encoding)



Scatter plot sample of a few features to dependent variable -2022 enrollment

Evaluation 1

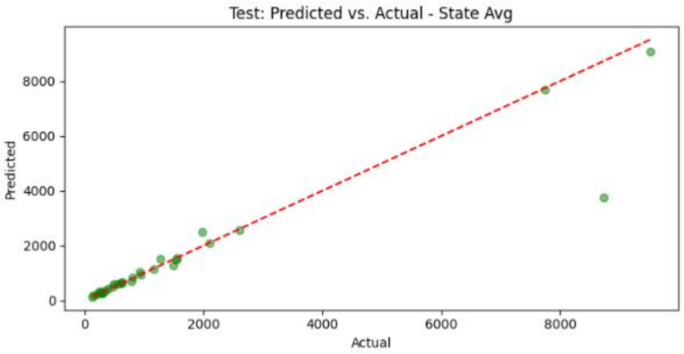
- **Parameter Tuning**: Cross-validation and random search (with mostly numerical field, deep trees and large bin numbers work better)
- **Evaluation metric**: MAE (for regression, intuitive interpretation and resilience to outliers)
- **Evaluation Results:** Random Forest outperformed the others



Evaluation 2

Why Random Forest outperformed?

- Less hyperparameter to tune
- Simpler and less overfitting



Random Forest predicted vs. actual

Predictions from worst to best- State Average%:

	State	Actual	Predicted	Percentage_Error
29	AZ	8733.285714	3757.739286	0.569722
2	NE	248.904762	333.136508	0.338410
9	KS	152.152542	200.780508	0.319600
28	TX	1983.327731	2496.022479	0.258502
7	SD	242.153846	296.169231	0.223062
26	UT	1274.590909	1523.411364	0.195216
11	OH	498.585366	591.488110	0.186333
22	IL	1496.687500	1287.935938	0.139476
20	OK	925.156250	1054.071875	0.139345
15	WY	792.357143	684.966071	0.135534
5	MT	305.750000	270.274432	0.116028
18	MI	555.562500	616.153906	0.109063
10	MS	394.955556	431.933333	0.093625
19	GA	639.721739	677.207826	0.058597
14	TN	474.906977	498.459593	0.049594
21	NC	1540.988889	1468.943611	0.046753
3	AR	289.442623	275.997951	0.046450
0	WV	134.500000	128.280682	0.046240
30	FL	9518.940299	9094.724254	0.044565
13	MO	616.972603	643.002397	0.042190
16	OR	622.300000	645.450000	0.037201
8	AK	1164.333333	1129.275000	0.030110
17	WI	805.095238	828.885714	0.029550
1	LA	945.274510	924.328431	0.022159
4	ND	215.250000	210.956250	0.019948
6	IA	299.878049	294.282317	0.018660
27	AL	2613.388889	2568.859722	0.017039
24	HI	7750.000000	7678.375000	0.009242
12	IN	357.500000	354.685417	0.007873
23	NH	1552.333333	1559.941667	0.004901
25	SC	2099.405405	2090.147297	0.004410

Random Forest predicted vs. actual

Timeline & Discussion

Week 1	Week 2	Week 3
- Project topic - Data collection	- Data merging - Data cleaning - EDA - Quick modeling	- Modeling - Evaluation

- Lots of time was spent on data collection, data integration, data cleaning, data sanity check, and data warehousing/arrangement

🙌Now in week 3 and everything is on track!

Conclusion & Future Work

- **Conclusion:**
  - Findings: enrollment pattern, trend, and contribution factors
  - Prediction
- **Future work**
  - **Label Encoding:** relabeling Issuer ID and County FIPS code (based on enrollment avg. similar to state label encoding)
  - **Healthcare Knowledge:** medical code/terminologies/policies

References

[1] https://www.statista.com/chart/8658/health-spending-per-capita/

[2] 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

[3] 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

[4] 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