With Characteristics of States to Predict the Growth Rate of Medicare

Github: github link for code
Data Source: enrollment original data sourse; tax related data across state
Author: Hanyu Ma (mhanyu@umich.edu)

Purpose

Medicare provides insurance protection for individuals who are 65 years old and Disabled individuals who have received disability benefits for 24 months. Medicare covers the costs of seniors totals more than $5300, nearly 40% of the median income of individuals age 65 or over. Therefore, without Medicare, many people would struggle to pay for the amount which has been covered by Medicare. For the final project, I utilize the tax data across regions as predictors to predict the future growth rate of medicare.

Data Manipulation

Since the enrollment data of Medicare is from 2013-2024, along with the tax data only before 2021, I pick the intersection of data from 2013-2021. Considering the distinct features of different states, I decide group both the tax and enrollment data by state. What's more, instead of getting mean directly, I calculate the mean with weighted value of the number of returns across state. The details about predictor and outcome variables are attached on the appendix table 1.

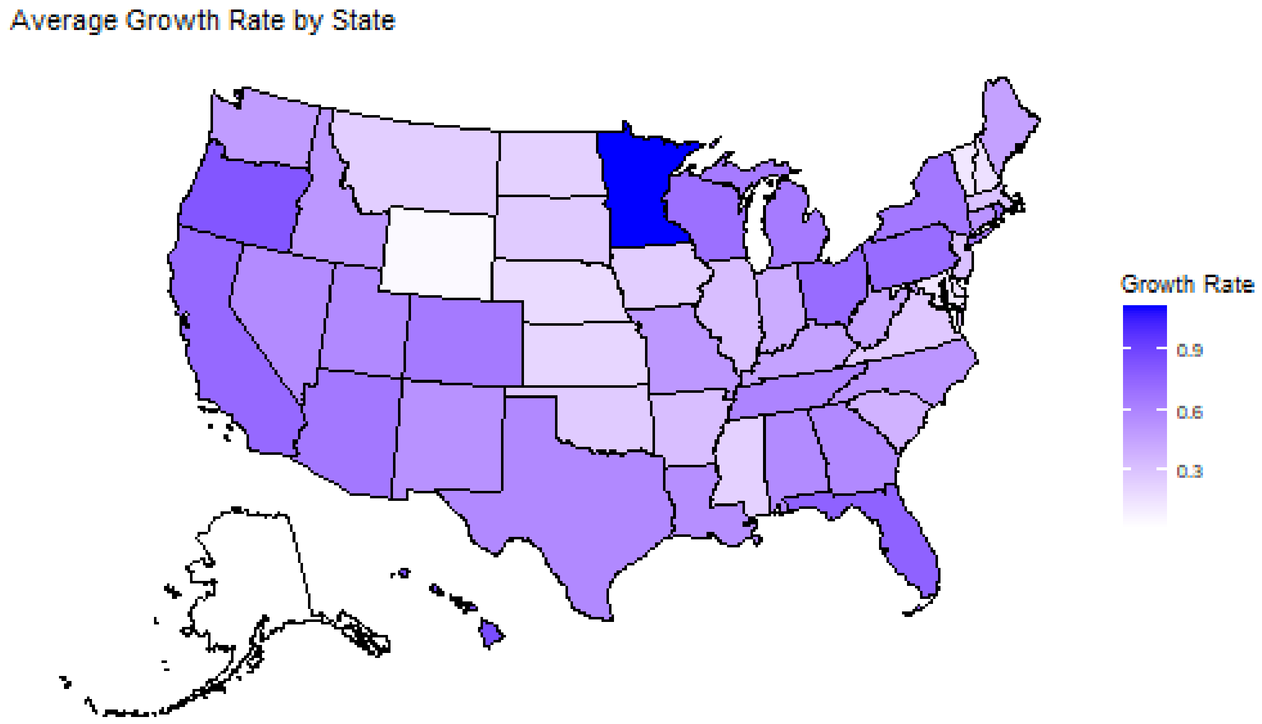

Figure 1: map showing the distribution of growth rate by state

December 2024

Data Description

By far, I drop all the data with missing values and state not in US (PR) and collect the features across states, the left size of data set is 454. From the appendix table 2, for the outcome variable, growth rate, the min value is 0.79%, which the max value is 138.19%.

Especially, for tax-related data, I notice that there is a wide range for all numerical variables, excluding marriage, elderly and growth rate variables. Specifically, for tax, the minimum is 46130, while the maximum is 70306055. There is a extremely huge gap between the minimum and maximum. What's more, features like tax with values from 46139 to 70306055, while features like marriage rate from 16.33% to 46.68%. The model might give disproportionally priority on the tax variable, so I decide to scale all the features.

According to figure 1, the growth rate of Medicare enrollment differs significantly across states. And the growth rate is always more than zero. This is a very positive sign for Medicare insurance development. Therefore, I choose to deal with the data as longitude data since it is obvious that the growth rate changes with states. With the information, I build model to do inference and prediction for growth rate across states.

Statistical Inference

Considering the huge number of states, we use linear mixed effect model to do statistical inference for understanding the relationship between predictor and outcome variables. After scaling all the variables, I apply the linear mixed effect model to my data set to explore the underlying statistical inference relationship. The results is as table 3. It is useful because the model not only control the effect of states, but also does not affected by the large number of states.

After controlling the random effect of state, the model shows that there six variables ("marriage_rate", "elder_rate", "add_med", "edu", "retire", "child") significantly effect the growth rate of Medicare enrollment at 5% significant level. For marriage rate, holding other variables constant, for every 1 standard deviation increase in marriage rate, the growth rate is expected to decrease by 0.98 standard deviation. For elderly rate, every 1 standard deviation increase in elderly rate, resulting that the growth rate is expected to increase by 0.08 units, holding all other predictors constant. Because Medicare exclusively serves for individuals who are 65 years or over, or with disabled, it makes sense that elderly rate has a significantly positive effect on growth rate. For every 1 standard deviation increase in additional Medicare tax, the outcome variable is expected to decrease by 0.42 units, holding all other variables constant. It seems that the growth rate of Medicare beneficiaries is sensitive to additional Medicare tax. For every 1 standard deviation increase in educator expenses, the outcome variable is expected to decrease by 0.41 units, holding all other predictors constant. For every 1 standard deviation increase in retirement savings contribution credit amount, the outcome variable is expected to increase by 0.54 units, holding all other variables constant. This makes sense because the insurance design for elderly individuals who is 65 years old or over. For every 1 standard deviation increase in child, the outcome variable is expected to decrease by 0.25 units, holding all other predictors constant. It is surprising that either the marriage rate or Child and dependent care credit amount has a significantly negative effect on growth rate. I think this is caused by extra expenditure on children so they have to cut the expenditure on insurance. It demonstrates that for most of family, Medicare is not necessary expenditure.

Prediction

For prediction, I utilize the linear regression model, support vector machine, XGBoot and random forest model based on its high accuracy and robustness.I divide the dataset into train set and test set first, utilize models, and calculate the out-of-sample R^2 to estimate each model. To begin with, I drop the year column for each data set because for future data, we have no chance to access the coefficient of its year.

For ordinary least squares for linear regression model, from the test set, the out-of-sample R^2 is 90.48%.

For support vector machine model, based on minimizing the risk of over fitting, especially for smaller and limited datasets, aligning with this data set, the outcome of prediction in this model is excellent. The out-of-sample R^2 reaches at 87.79%.

For XGBoot model, since the data set includes year as the prediction variable, I try this model for its advantage on time series data. The out-of-sample is the highest, which reaches 99.82%. It is very appropriate for the data set.

For random forest model, I apply all the features to the model to predict. The out-of-sample R^2 is 88.59%, which is really high. To further avoid the outfit problems, I use the importance scores to select features. With the method, I pick the top 10 features ("child", "marriage_rate", "edu", "tax", "retire", "add_med", "population", "AGI", "elder_rate", "STATE_MN") for random forest model. This not only significantly reduces the number of

predictor variables, but also maintains the out-of-sample R^2 at 85.05%, which is a relatively high level.
Across the four prediction models, the XGBoot has the highest out-of-sample R^2, reaching at 99.81%. It is extremely fit for the data set.

Summary

For statistical inference, elderly rate and retire savings contribution credit amount, which are elderly indicator, significantly increases the growth rate, while additional tax amount, marriage rate, educator expenditure and Child and dependent care credit amount significantly decreases the growth rate.
For prediction, the XGBoot algorithm comes first, the out-of-sample R^2 reaches 99.82%.

Appendix

Table 1: Variable Description

Variable	Description
growth_rate	(Total Beneficiaries-Original Beneficiaries)/Original Beneficiaries
AGI	Adjust gross income
marriage_rate	Number of joint returns / Number of returns
elder_rate	Number of elderly returns / Number of returns
tax	Taxes paid amount
charity	Taxes paid amount
add_med	Additional Medicare tax
edu	Educator expenses amount
retire	Retirement savings contribution credit amount
child	Child and dependent care credit amount
population	Number of individuals

Table 2: Data Description

	AGI	mar%	eld%	tax	chari	med	edu	retire	child	grow%	population
Min	1.04e+07	0.16	0.17	46130	113629	3017	686	867	584	0.0079	538790
1st Qu.	2.78e+07	0.34	0.24	536231	465669	11067	2574	5959	7952	0.2467	1563180
Median	6.77e+07	0.37	0.25	1426176	1436549	30666	6954	12933	21594	0.4326	3937960
Mean	1.22e+08	0.36	0.25	3907273	2289956	90442	10228	18866	36456	0.4579	6029235
3rd Qu	1.53e+08	0.39	0.27	3863761	2742904	99209	12192	23483	43948	0.6416	7072108
Max	1.15e+09	0.47	0.33	70306055	22685772	1565498	55534	109749	240328	1.3819	39113550

Table 3: Report of Linear Mixed Effect Model

	Estimate	Std..Error	t.value
(Intercept)	-0.01	0.17	-0.05
AGI	0.69	0.41	1.68
marriage_rate	-1.00	0.06	-15.92
elder_rate	0.08	0.02	3.29
tax	0.02	0.04	0.45
charity	0.03	0.13	0.22
add_med	-0.48	0.18	-2.66
edu	-0.14	0.30	-0.48
retire	0.58	0.15	3.78
child	-0.26	0.08	-3.29
population	-0.50	0.41	-1.22