Analysis in Medicare Provider Utilization and Payment Data

From the Prospectives of Average Difference between Submitted and Charged Medicare Amount from Physician in California

Kai-yu Chen - August 7, 2018

[1]

[1] Picture reference: https://seniortubs.com/cost/medicare-and-medicaid-coverage

Introduction

With the growth of older population, the need for Medicare is increasing. It is important to understand how Medicare works and what factors would contribute to the difference in Medicare. In this analysis, I would use statistical method and machine learning to investigate what factors would influence the difference between average Medicare submitted amount and average amount physicians charged in the provider type of Neurosurgery, Cardiac surgery, Vascular surgery, Nurse Anesthetist (CRNA), and Thoracic Surgery, in the state of California.

Data Preprocessing

Data Set Information and Subset

The data set contains roughly 10 million samples and 26 features. After dissecting the data set into the State of California, The dataset became 700 thousands samples. Several features have been removed before preprocessing because they are not factors that I am interested in looking into based on domain knowledge and I would like to reduce potential noises from the data.

Missing Value Imputation

Missing value would be one of the important issue to be dealt during data preprocessing steps. Generally, there will be multiple strategy to deal with missing value, including imputation by mean, median, mode, and probability distribution of the data. Here, after removing unimportant features, only feature of Gender of the Provider has missing value. The result showed that gender in male accounts for almost 74% of the data, it is unlikely that missing value missed at random. Generally, more information should be looked into to before removing missing value, yet missing values account less 20% of the feature (missing at 6%), therefore these missing values were just removed without further imputation.

Correlations Among Features

To determine the degree of relationships exist between two variables, correlation matrix was performed. the matrix showed that Average Submitted Charged Amount, Average Medicare Allowed Amount, Average Medicare Payment Amount, and Average Medicare Standardized Amount have relatively higher positive correlation to Average Medicare Difference, with correlation coefficient of 0.98, 0.71, 0.7, and 0.7 respectively.

Converting Categorical Variables into Dummy Variables

All the categorial variables in the dataset have been converted dummy variables because I am going to apply the dataset to regression models. Keep noted that our question is " predicting Average Medicare difference", which our target variable is numerical and a regression model should be built. Converting categorical variables into dummies would allow us to build a regression model without too much hurdles. In addition, dummies improves efficiency by saving space and computational complexity.

Top 5 Highest Average Medicare Difference in Providers
Average Medicare Difference (AMD) is referred to the difference between Average Medicare Allowed Amount (AMAA) and Average Submitted Charge Amount (ASCA). The top 5 providers that have highest AMD are: Thoracic Surgery, Neurosurgery, Cardiac Surgery, Vascular Surgery, and CRNA. Among all, Thoracic Surgery has the highest AMD, with $1192 USD per beneficiary.

Average Medicare Difference by City and Zip Code
It is also interesting to see if AMD differs among cities or zip code. The result showed that City of San Diego (zip code 92093) has the highest AMD, with roughly $17248 USD per beneficiary.

Top Procedures in Provider Type
In Thoracic Surgery, heart surgery has the highest AMD, with $17248 USD per beneficiary. In Neurosurgery, the procedure of repair of bulging of blood vessel (aneurysm) in brain has the highest AMD, with $11521 USD per beneficiary. In Cardiac Surgery, the procedure of insertion of vena cava filter by endovascular approach (including radiological supervision and interpretation) has the highest AMD, with $30758 USD per beneficiary. In Vascular Surgery, the procedure of removal of plaque and insertion of stents into artery in one leg, endovascular, accessed through the skin or open procedure, has the highest AMD, with $23740 USD. Lastly, in CRNA, the procedure of anesthesia for procedure on heart and great blood vessels on heart-lung machine, age 1 year or older, or re-operation more than 1 month after original procedure, has the highest AMD, with $5002 per beneficiary.

The information behind these numbers indicate that any of these procedures involved would be expected to have larger difference between the amount that physician actually charged and submitted and the amount that Medicare generally allows. The difference would be expected to be paid by either the provider or beneficiary according to In-Network or Out-of-Network. More information should be provided to land a conclusion.

Training Dataset and Testing Dataset

The dataset was split into training dataset and testing dataset at ratio of 4:1.

Regression Models

Baseline Model: Linear Regression

A linear regression model was built as a baseline model. The result of the model showed a R-squared, an indicator of how good the model fit to the data, is -2.56. Here, the coefficient R-squared is defined as (1 - u/v), where u is the residual sum of squares ((y_true - y_pred) ** 2).sum() and v is the total sum of squares ((y_true - y_true.mean()) ** 2).sum(). The number was negative because the model is arbitrarily worse, and the Mean Squared Error is unreasonably high. The model performs awful and linear model may not be a good tool to predict the data because the data might not be linear separable. Thus, other models are built to improve the performance

Elastic Net Regressor

Elastic Net is a combination of Lasso Regression(l1) and Ridge Regression (l2). To prevent overfitting of the model, l1 and l2 norms are penalized based on parameter λ. Some of coefficients of unimportant variables would shrink to zero by shrinking the beta coefficient. Alpha mediates the amount of penalty applied to the data, if alpha equals to zero, then we have a ridge regression, if alpha equals to one, then we have a lasso regression.

After 10-fold cross validation, the optimal parameters for l1_ratio and alpha is 0.1 and 0.000012 respectively. The parameters were used to build a Elastic Net Regressor. After penalizing unimportant variables, the model reduced MSE to 0.001, with R-Squared of 0.59. The Elastic Net far better fit to the data compared to linear regression model.

The Elastic Net showed that Average Medicare Payment Amount is the most important feature, followed by Average Medicare Standardized Amount and Heart Surgery in providers of Thoracic Surgery.

Random Forest Regressor

Random Forest is a collection of decision trees on different random sub-datasets to prevent overfitting and averages out the performance from every tree to improve the predictive accuracy overall. Here, there are 200 trees in this collection, with minimum samples of 2 per leaf, and The minimum number of samples required to split per internal node of 15. The result showed that MSE was reduced to 0.0009, and R-Squared was increased to 0.64.

The Random Forest Regressor showed that Average Medicare Payment Amount and Average Medicare Standardized Amount are the most two important features.

eXtreme Gradient Boosting

Neural Network

A MLP regressor was built with 3 hidden layers, each layer consists of 30 nodes. The result showed that with implementing neural network, the MSE was 0.001, and the R-Squared was 0.56.