Embedding random forest regression model into natural texts resistant to the inflated self-reported data in nursing homes

Jonah Winninghoff *Springboard Data Science*

The U.S. Centers for Medicare and Medicaid Services (CMS) used the star rating system to assist consumers with choosing the nursing home since 2009. As the Covid-19 pandemic took place, one study implied that the self-reported data is inflated due to moral hazard. This inflation problem is attempted to be solved using a predictive model with natural language processing (NLP) techniques. The dataset that contains over 9,000 zip codes obtained from CMS is in use for the modelling algorithm purpose. During the analytic process, Guam and Puerto Rico are excluded due to the presence of outliers. After the NLP predictive model undergoes several tests and investigations, it is in use to interpolate values from Guam and Puerto Rico. Despite the fact that the extreme form of self-reported inflation is successfully mitigated, there are several challenges in this model that need to be addressed.

Keywords: NLP, TF-IDF, Big Data, Gauss Markov Assumptions

September 20, 2021

INTRODUCTION

Developed by U.S. Centers for Medicare and Medicaid Services (CMS) in 2009, the star rating is a system that helps people make decisions in which nursing home the senior residents would live. This system became popular to educate consumers. It is established as inspections; amount of time the nurses spend with residents and quality of care that residents receive. On March 13, 2021, the Covid-19 killed more than 130,000 residents, which accounts for, roughly, 25% of total deaths in the United States. The number of deaths remains underestimated, which is historically tragic. Equally importantly, this pandemic offers key insights that the rating system itself is distorted due to *moral hazard*. For example, the number of Covid-19 deaths at five-star facilities are not significantly different from one-star facilities (Silver-Greenberg and Gebeloff, 2021).

The information in CMS database is not immune to self-report bias, though it is humongous. The self-reported behavior is prone to moral hazard, which is an incentive to deviate from cooperative agreement that the report should be truthful. The incentive to falsify the report is due to business perception of trade-off between the cost of cooperation and that of deviation. The nursing-home industry perceives fraud, rather than as punishment, as the opportunity cost that they are willing to make.

The consumers face *adverse selection*. This term refers to, for example, home nursing staff and employers who have more information than consumers. In other words, the consumers are unable to distinguish different levels of quality services in nursing homes. Increases in monitoring that helps to alleviate the moral hazard can be useful, especially with natural language processing (NLP) artificial intelligence solution. The NLP is an automation to preprocess human languages as unstructured data and make predictions based on data. For this research, the small NLP predictive model is created to process the texts based on resident assessments.

EXPLORATORY DATA ANALYSIS

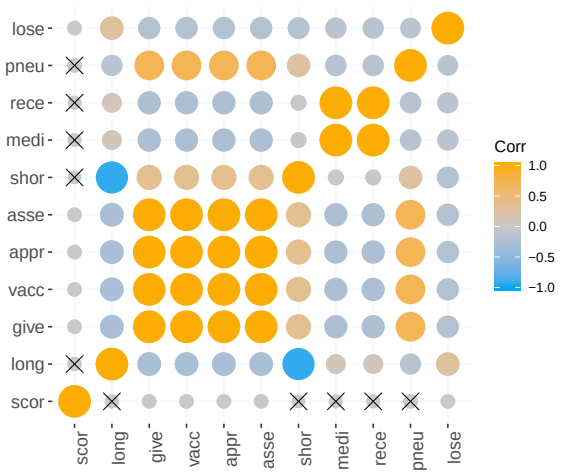
The dataset obtained from CMS database is Minimum Data Set (MDS) Quality Measures. This dataset contains over 15,000 different providers from at least 9,000 zip codes across 50 states plus District of Columbia, Guam and Puerto Rico. For this analysis, Guam and Puerto Rico are excluded due to outliers. Even though the dataset has 23 different columns, however, a few of them are useful to help predict measure quality. The measure description contains natural texts to describe the measurement that may be useful for predictive modeling.

Choosing this measure description may be perplexing since the zip codes are more useful to predict without requiring any types of monitoring. The simple regression shows that the r-squared score using zip codes as a single variable is 20.9%. In other words, this regressor explains at least 20% of the variance for target variable. The problem with this approach is that this regressor is a *proxy* for race. Modeling the prediction using features correlating to protected classes is a form of indirect discrimination, which is known to be proxy discrimination. If implementing the algorithm, the disparate impact may be occurring (Datta et al., 2017). For example, the nursing home located in a certain zip code may be penalized by this algorithm even if this institution provides high quality of services. This algorithm does not address this moral hazard problem.
This is the reason why the measure descrip-

tion might be a better choice for predictive modeling. To determine if the text could help predict this target variable, the NLP techniques are in use to undertake assessment on the text. Each text is vectorized using term frequency-inverse document frequency (TF-IDF), rather than traditional bag-of-words. As a result, there are 49 regressors with target variable. Under the normal circumstance, the number of columns is much higher. The dataset is relatively thinner because every text in measure description is already standard.

The TF-IDF has its own advantage over its counterpart because this approach standardizes the text. The normalization makes this data more resistant to numeric distortions by downgrading the relative importance of words that appear too often. There is a drawback of TF-IDF that is important to acknowledge. Change in n-gram affects the different predicted outcome using this approach due to the inability to quantify different meanings in n-gram phrases. On the other hand, this approach is well-scalable (Shahmirzadi et al., 2018).

When the TF-IDF approach is applied, the Pearson correlation coefficient including significance test is in use to assess the relationship of variables including target variable. The correlation coefficient is to determine the level of the linear relationship between variables. The significance test is to perform alternative hypothesis against the null hypothesis that the correlation coefficient is not significantly different from zero. The alpha level of chance to reject the null hypothesis is 10%.

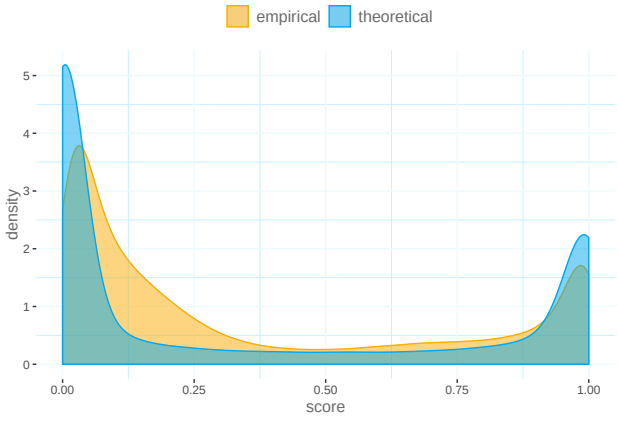


As indicated by the heatmap, the smaller circle size is closer to zero while the bigger circle size is far from zero. The "X" mark means that the null hypothesis failed to reject. In other words, this result is unable to confirm that these terms like "received", "pneumoccal", and "short" are higher than zero. The terms like "lose", "assessed", and "given" correlated to target variable are significantly different from zero, albeit having a weak linear relationship. The certain terms that fail to reject the null hypothesis do not necessarily imply that these terms have no correlation with target variable since the target variable itself is bimodally distributed.

However, this analysis needs to undertake empirical cumulative distribution function in order to determine if the bimodal distribution is parametric. If so, the parametric test can be in use to test against another type of null hypothesis. Given that the target variable is $\in(0,1)$ and cross-sectional, the continuous version of binomial distribution is in use to test the distribution of target variable. Given that both alpha and beta parameters are unknown, the identification formulas for both are:

$$\hat{\alpha} = \bar{x}\left[\frac{\bar{x}(1-\bar{x})}{s^2-1}\right]$$
$$\hat{\beta} = (1-\bar{x})\left[\frac{\bar{x}(1-\bar{x})}{s^2-1}\right]$$

Both formulas help generate beta-distributed random variates to determine if the empirical distribution is identical to or, at least, consistent with theoretical distribution (Sinharay, 2010). For this analysis, the empirical cumulative distribution function (ECDF) is in use to compare theoretical with empirical one. Using this approach is more statistically sensible than histogram due to binning bias. The binning bias refers to the change in the number of bins that alters meanings in this analysis. For example, the histogram shows, with the small size of bins, that the distribution is Gaussian. But when the size of bins increases, this distribution is bimodal. The point in mentioning this is that the analysis needs to infer what presents.



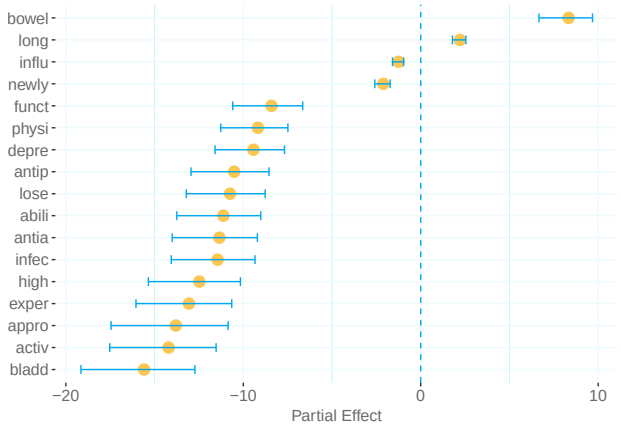
To be clear, this plot is density-based, not ECDF-based. This approach is more interpretable for the broader audience. The plot shows that the empirical distribution is not consistent with Beta distribution. The ECDF analysis makes the same conclusion. This distribution is unlikely to alter when the size increases since the dataset is enormous. The sampling size is over 200 thousand. In other words, the target variable is nonparametrically distributed, so the nonparametric test is useful to test null hypothesis that two groups are within the same distribution.

More specifically, the nonparametric test is to evaluate five terms that are previously not different from zero at the significance level, which are "long", "short", "medication", "received", and "pneumococcal." The null hypothesis is that the average measure quality remains the same if each term either includes or excludes. The alpha level for this null hypothesis is 5%, however, with Bonferroni correction, the alpha level lowers to 1% due to five statistical tests. As a result, each term may have the relationship with measure quality at the significance level. This result is confident enough to conclude that the causality might exist.

The multiple ordinary least squares (OLS) regression is capable of inferring the causal linkage between target variable and regressors. However, this regression is dependent on Gauss Markov assumptions, as the following: linearity in parameters (1), no perfect collinearity (2), zero conditional mean error (3), homoskedasticity and no serial correlation (4), and normality of the error (5) (Goldstein, 2021). But as indicated by Breusch-Pagan test, the p-value is below the alpha level, which is 5% in violation of homoskedasticity assumption. In other words, this result shows that heteroskedasticity exists.

Some regressors are perfectly collinear, so the number of regressors automates to reduce 49 to 17 regressors. When the robust OLS regresses target variable on regressors, the result shows that the p-value of the F-statistics is significantly smaller than the alpha level equal to 1%. The null hypothesis states that all regressors equal zero is rejected. In other words, this model performs better than the intercept-only model. Furthermore, the p-

value of each term is significantly smaller than the alpha level equal to 1%. Every parameter, *ceteris paribus*, affects the level of measure quality at the significance level. Both r-squared score and adjusted r-squared score are 90.95%. The first score refers to the regressors explaining at least 90% of the variance for target variable while the second score shows that the increase in the number of parameters does not change the value of this score.



The partial effect refers to one-unit change in regressor that affects target variable. This effect is a coefficient of regressor given by partial derivative of the expected value of target variable in respect to this regressor. As indicated by the plot, "bowels" and "long" terms have, *ceteris paribus*, positive impact on target variable while terms like "bladder" and "activities" have negative impact. Intuitively, the measure quality score with these terms appeared in text is lower than without these. In other words, this model establishes as a baseline for predictive modeling, which can be useful for NLP artificial intelligence solutions.

Predictive Modeling

Prior to creating and investigating four different models, the features in the dataset have undergone a preprocessor by separating training, validation, and testing sets. The training set is a seen dataset used to fit parameters in the given model while the validation set is an unseen dataset used to test if this model is able to fit well outside training set and tune hyperparameters. The hyperparameters refer to the values that regulate the learning process of the given model. For example, the hyperparameter limits the number of parameters in the given model. The testing set is a hidden dataset used when the given model is confident enough to apply and if it is capable of generalization. This set can be used once.

Not only that, but the principal component analysis (PCA) is also applied to compare models with others that do not use this approach. The PCA is a dimensional reduction technique that can reduce the number of columns and minimize information loss through linear transformation at the same time. Unlikely to t-stochastic neighbor embedding, PCA preserves the explained variance information. PCA can identify each component, from the largest to smallest. For example, the first 12 components explain 92.24% of the variance while the first 15 components explain 99.38%.

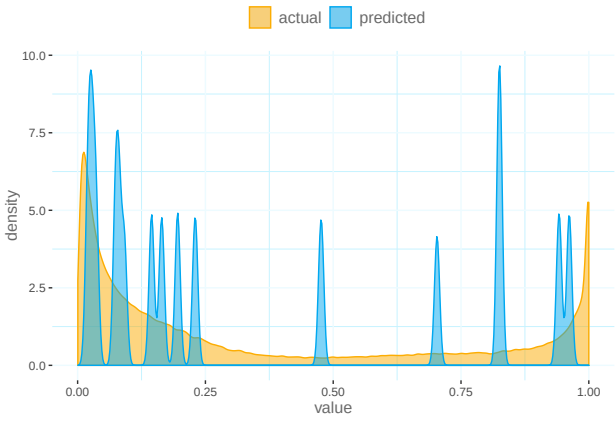
The assumption is that the number between 12 and 15 components is sufficient to outperform the models without using PCA approach. The OLS model tests against this assumption. However, the surprising result suggests that the model is at optimal level if using 18 components. The validation set further rejects this assumption. For example, both r-squared and adjusted r-squared scores equal 90.91% while mean absolute error (MAE) and root mean square error (RMSE) are equal to 0.07 and 0.11, respectively. Having 18 components not only optimize this model but this model is well-calibrated. In other words, the chance of predicting the measure

quality equal to 50% may be either 57% or 43%. But what is more surprising, the model without using PCA shows that both r-squared and adjusted r-squared scores including MAE and RMSE are identical to that without using them.

The lasso regression, random forest regression, and light gradient boosting being tested against training set and validation set offer some interesting insights. For example, the lasso regression model underperforms, being compared to the previous model even with its optimized hyperparameter. The random forest regression (RFR) and light gradient boosting (LGB) models perform as well as OLS. The Bayesian optimization is in use for both lasso regression and light gradient boosting models. This tool that can be in use for particular non-convex or derivative-free optimization problems through constructing a probabilistic model with objective function as a guider (i.e., r-squared score)(Adams, 2012). With this tool, the Lasso regression remains underperformed, and the light gradient boosting model remains at the same performance level. What makes this analysis more perplexing is that OLS, RFR, and LGB have identical scores even either with or without PCA.

Investigating the models are necessary, so the behaviors can be better understood. The investigation process includes evaluating RFR model using PCA against testing set, creating ECDF, assessing regressors, comparing this model using PCA with another model not using that, and testing this model against simulated data. This process offers some intriguing insights.

The RFR model using PCA is tested against the testing set. The result is, both r-squared and adjusted r-squared scores are little higher than these in the validation set. The MAE and RMSE are 0.07 and 0.11, respectively, not different from these in the validation set. As indicated by these metrics, this model does not suffer from being overfitted, which is not remarkable. The RFR is not prone to overfitting, unlikely to AdaBoost and gradient boosting models. For example, the RFR is generated to create a number of decision trees using random subsets of dataset and features while these boosting models create iterative trees (Kaynig-Fittkau et al., 2021).



However, this result is skeptical, so the ECDF is created to compare the actual value with predicted value produced by the RFR model. As indicated by the density-based plot, the distribution shape of predicted values is different from that of actual values. The plot may suggest that either at least one regressor or this model is discrete. When the simulated data tests against this model, the predicted values turn into uniform distribution. In other words, the values in regressors are not various enough to predict continuous values. This analysis also concludes that the irreducible error exists due to a lack of various values in regressors. This error is reducible if undertaking data wrangling. For example, there is the algorithm that can identify potential linkages between this dataset and another dataset from the outside to determine if both datasets are joinable.

Besides this, the interesting part in this analysis is that the RFR models with and without PCA do not behave the same way even if both r-squared and adjusted r-squared scores are identical. The F-distribution is in use to test the null hypothesis of two models that both variances are not different from each other at the significance level. The finding is that null hypothesis is rejected at the alpha level equal to 1%. But the F-distribution shows that there is no significant difference between the variances of OLS models with and without PCA.

Actionable Insights

As mentioned earlier, Guam and Puerto Rico are excluded from this analysis. The model is in use to interpolate both. After the interpolation is applied, the predicted values for Guam and Puerto Rico compare with actual values for all other states. If Guam and Puerto Rico are not different from other states at the significance level, this analysis can conclude that this model is likely capable of resisting to self-reported inflation. It may be useful to help reduce other implicit costs. But there is the most fundamental approach of statistics that the hypothesis testing proves the existence of difference, rather than disproves it, against Type I error. Using the null hypothesis to prove no difference is inaccurate due to a chance of Type II error or *Beta*. By principal, the statistical test can either reject or fail to reject the null hypothesis, nothing more. For that reason, the roles of null and alternative hypotheses are reversed.

By definition, the null hypothesis is two groups different from each other while the alternative hypothesis is both groups not different from each other.

$$H_0 = |\mu_1 - \mu_2| > \delta$$
$$H_A = |\mu_1 - \mu_2| \leq \delta$$

If the null hypothesis is rejected at the given alpha level, both groups are not different from each other. To disprove the difference, this approach needs the Student's t-test (Streiner, 2003).

$$t(df) = \frac{(M_1 - M_2) - \delta}{S_{M_1 - M_2}}$$

Before this test is applied, due to the extremely uneven balance of dataset, the sampling size for all states excluding Guam and Puerto Rico is reduced to 70 using the simple random sampling generator. The actual sampling size for Guam and Puerto Rico is 30. Prior to using the model, the mean score difference between two groups is 0.37. The standard error for both groups is 0.01. After the model is applied, the mean score difference between two groups is 0.33. To disprove the difference between two groups, the t-statistic value must be higher than the critical value for the alpha level equal to 0.05. The critical value is 1.66. As a result, The t-statistic equal to 1.83 is past the critical value. This result rejects the null hypothesis that both groups are different from each other.

In other words, this model may be useful to mitigate the serious form of self-reported inflation. However, this model is, as discussed earlier, unable to offer the continuous predicted values due to due to a lack of various values in regressors. This model may not eliminate inflation in regressors as well. This analysis does not make conclusion, for that reason, with full of confidence. This model may have real effects on the rating system as long as if this model deploys exclusively for more research, without making any implementations. But this model does hold potential value in it by improving public trust.

Future Work

As mentioned earlier, the model should be deployed for research purposes without making any implementations. This process allows one to reevaluate the model using additional labels like, for example, the difference between high and low frequency of inspections over time. In any ideal situation, the randomized controlled trials (RCTs) over time are in use to determine if the model has real impacts using the number of inspections as a metric since the in-person inspection holds the greatest weight of CMS rating system (Silver-Greenberg, 2021). However, in the anticipated future, conducting the experiment might not be possible.

The alternative method is the observational study capable of estimating the accurate value using the econometric approach. The first step to conduct the analysis begins by writing the causal model of behaviors and its underlying assumptions instead of searching for correlations. The causal model may include, for example, the hours of care services spent by nurses per day with instrumental variable as such as hourly wage rate, number of inspections per year, number of deaths in nursing home per year over time multiplied by the dummy variable for either using or not using the model while the endogenous variable is the star rating system reviewed by the consumers rather than CMS rating system.

Having this information requires undergoing the data wrangling process using record linkage technique. This algorithm is, as mentioned earlier, to determine if the dataset from the outside can merge with the CMS database. This is a kind of process that helps to establish the causal inference of what kind of effects the predictive model might have on the nursing home industry. This algorithm is also in use to increase the number of features in data that helps this model predicting continuous values.

Not only that, but this analysis also intends to further assess the behavior of model, for example, using Box-Cox, Yeo-Johnson, and logarithmic transformations to determine if one of these approaches enhances its capability resistant to self-reported inflation. There are many tasks that need to be done in order to make the NLP artificial intelligence solution work.

Conclusion

In this analysis, the evaluation on the RFR model performance using TF-IDF vectorizaton and PCA dimensional reduction methods shows that this model can fit well in training, validation, and testing sets. Also, this model is well-calibrated. For example, the r-squared and adjusted r-squared scores are approximately 91% while MAE and RMSE are approximately 0.07 and 0.11, respectively. The variance of RFR model with PCA comparing with the variance of this model without PCA is different at the significance level. This model is generalizable to unseen data.

Before the MDS Quality Measure dataset is in use to train the model, this data excludes Puerto Rico and Guam. When the model is trained, Puerto Rico and Guam are in use for interpolation. This interpolation of this model concludes that there is no mean score difference between Puerto Rico and all states in the United States at the significance. Intuitively, this model has ability to predict the measure quality resistant to extreme form of self-reported inflation.

However, this model should be deployed for more research because it is not production-ready. For example, RCT needs to be in use

in order to determine if this model is useful. Another way to assess this model is the econometric approach with record linkage by comparing predicted values with actual values. The transformations in scores might be in consideration for this research to determine if this process increases resistance to self-reported inflation. Currently, the understanding of this model behavior is limited.

References

Adams, Ryan P. "Practical Bayesian Optimization of Machine Learning Algorithms." School of Engineering and Applied Sciences Harvard University, 2012.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. "Proxy Discrimination in Data-Driven Systems." *Theory and Experiments with Learnt Programs*, arXiv, 25 Jul. 2017. https://arxiv.org/abs/1707.08120

Goldstein, Nathan. "Lecture 1. Foundations of Microeconometrics." *Microeconometrics*, Zanvyl Krieger School of Arts and Sciences Johns Hopkins University, 2021.

Kaynig-Fattkau, V., Blitzstein, J., and Pfister, H. "CS109 - Data Science." *Decision Trees*, Harvard University, 2021. https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=c22cbde8-94dd-42ad-86ef-091448ad02e4

Shahmirzadi, O., Lugowski, A., and Younge, K. "Text Similarity in Vector Space Models: A Comparative Study." arXiv, 24 Sep. 2018. https://arxiv.org/pdf/1810.00664.pdf

Silver-Greenberg, Jessica, and Robert Gebeloff. "Maggots, Rape and Yet Five Stars: How U.S. Ratings of Nursing Homes Mislead the Public." *How U.S. Ratings of Nursing Homes Mislead the Public*, The New York Times, 13 Mar. 2021.

Sinharay, S. "Coninuous Probability Distributions." *The International Encyclopedia of Education*, Elsevier Science, 2010. https://doi.org/10.1016/B978-0-08-044894-7.01720-6

Streiner, David L. "Unicorns Do Exist: A Tutorial on"Proving" the Null Hypothesis." *Research Methods in Psychiatry*, The Canadian Journal of Psychiatry, Dec. 2003.