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

Jonah Winninghoff *Springboard Data Science*

The star rating developed by U.S. Centers for Medicare and Medicaid Services (CMS) since 2009 is a system that helps people make decision which nursing home the senior residents would stay at nursing home. This system is established processed by inspections, amount of time the nurses spend with residents, and quality of care that residents receive. 1: .5, 2: 3, 3: 2, 4: 1, 5: 1, 6: .5

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).

Humongous, although the information in CMS database is, is not immune to survivorship bias and self-report bias. In other words, there is no perfect information on how the data is built. The self-reported behavior is 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.

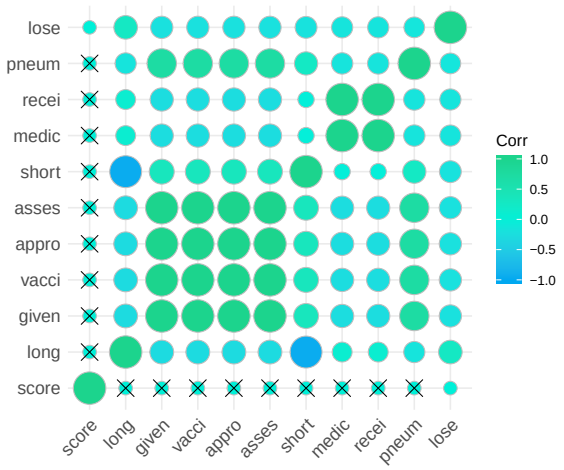
Choosing this measure description may be perplexing since the zip codes are more useful to predict without requiring any types of monitoring. The simple regression shows that the r-squared score using zip codes as a single variable is 20.9%. In other words, this regressor explains at least 20% of the 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 description might be a better choice for predictive modeling. To determine if the text could help predict this target, 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 (BoW). 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 Product-Moment Correlation with significance test is in use to assess the relationship of variables including target variable. This 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 5%.

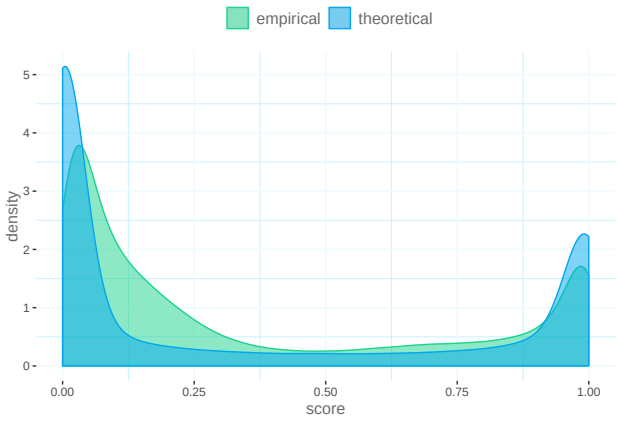


As indicated by the heatmap, the smaller circle size is closer to zero while the bigger circle size is far from zero. The "X" mark means that the null hypothesis failed to reject. In other words, this result is unable to confirm that these terms like "received", "pneumoccal", and "short" are higher than zero. The "lose" term correlated to target variable are significantly different from zero, albeit having a weak linear relationship. That is, this correlation is more likely to be well-calibrated. 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.



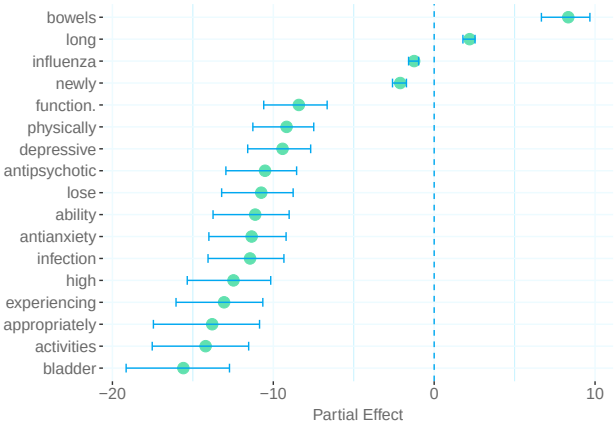
To be clear, this plot is density-based, not ECDF-based. This approach is more interpretable for the broader audience. The plot shows that the empirical distribution is not consistent with Beta distribution. The ECDF analysis makes the same conclusion. This distribution is unlikely to alter when the 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). 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.



Predictive Modeling

Actionable Insights

Future Work

Conclusion

References

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

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