Predicting Vaccine Acceptance in the United States

Ellie Lai, Ana Mafla, Eric Ricci, Makino Xin
Course: 36-600, Overview of Statistical Learning and Modeling

Introduction

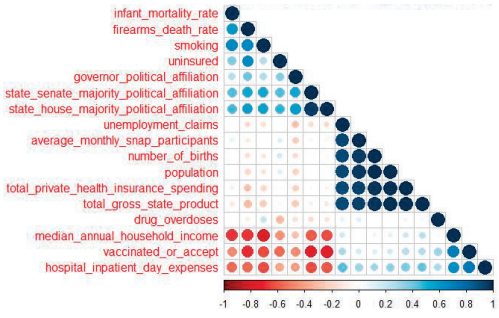
With the introduction of the COVID vaccine in 2020, a milestone was reached in the fight against COVID-19. However, the vaccination process has been slow-going, with vaccine acceptance varying across different geographical regions of the United States. This dataset was collected from Carnegie Mellon University's COVIDcast project by the Delphi research group and from the Kaiser Family Foundation. The goal of this project is to model vaccine acceptance percentages given various demographic variables across all 50 states.

Data

This dataset consists of 16 predictor variables, 13 quantitative and three qualitative, for 50 observations. Logarithmic and square-root transformations are performed on eight predictor variables to reduce skew and ensure better visualization.

Health-Related	Population-Related	Government-Related
• *Uninsured* • *Total private insurance spending* • *Drug overdose death rate* • *Smoking* • *Hospital in-patient expenses*	• *Births* • *Population* • *Infant mortality rate* • *Firearm death rate* • *Unemployment claims* • *Household income*	• *Governor affiliation* • *Senate majority* • *House majority* • *Gross state product* • *SNAP monthly participants*

Our response variable is the percentage of vaccine acceptance among survey respondents, vaccinated_or_accept. We see that some government-related factors and health-related factors, such as senate majority, house majority, and smoking are negatively correlated to vaccination acceptance. On the other hand, median household income and hospital in-patient expenses have a positive correlation with vaccination acceptance.



We also see that several variables are linearly related with one another. By doing PCA analysis, an eight-dimensional subspace is able to account for 96.6% of the variance with minimal loss.

Figure 1. Correlation plot of predictor and response variables

Methods & Analysis

The data is split such that 70% of the data is used for training and 30% used for testing. We apply 4 models for statistical learning: linear regression, linear regression with subset selection, random forest, and KNN. The mean-squared error (MSE) is calculated for all models, with random forest regression having the lowest MSE.

Model	Test-Set MSE
Linear Regression (LR)	4.295
LR w/subset analysis	3.580
Random Forest	1.817
KNN	5.780

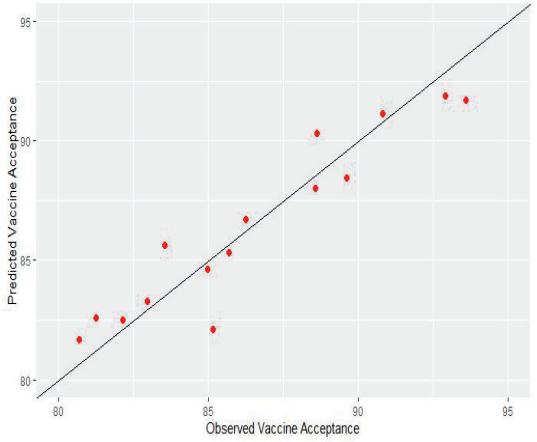


Figure 2. Predicted vs. observed vaccine response graph generated from random forest regression

We find that the best model to predict vaccine acceptance is random forest regression. We determine that the most important variables to predict the acceptance of the vaccine were generally population-related or politically-related. This analysis demonstrates two forces that can predict COVID acceptance or rejection: outbreak severity and politicization. Regions that were hit harder with COVID (as measured by population-related healthcare spending) are more likely to accept the vaccine, while regions that didn't are less likely to. Another factor that affects vaccination acceptance is local political affiliations, which demonstrates the extent of politicization with regards to COVID today.

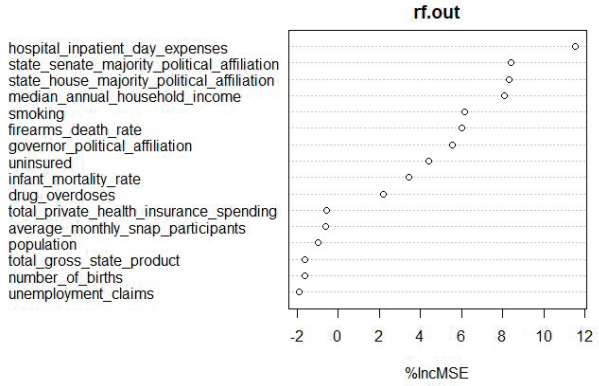


Figure 3. Predictor variable significance contributions

Conclusion

Overall, we find that several predictor variables that are generally population-related or politically-related have a strong correlation with vaccine acceptance. This correlation is further seen by successful prediction of vaccine acceptance rate via a random forest regression model, where factors such as senate majority and hospital in-patient expenses play a role in determining vaccine acceptance. For future studies, it may be beneficial to include more predictor variables and observations to learn factors that negatively correlate with the response variable. Additionally, finding a way to address potential bias present with the survey respondents may find different factors that affect vaccine acceptance.

References

Carnegie Mellon University, Delphi Group. COVIDcast: https://delphi.cmu.edu/covidcast/, 2020.