Predicting COVID-19 Vaccine Acceptance Across U.S. States

David Rogers, Franco Valdés Navarro, Bingcheng Wang, Yaning Wu, Guantong Zhang

Introduction

Despite the availability of COVID-19 vaccines, some populations remain hesitant to receive them. Identifying the factors influencing vaccine acceptance is crucial for governments and policymakers to make informed decisions that promote higher vaccination rates and ultimately achieve herd immunity. This study explores various machine learning models using datasets from Carnegie Mellon University's COVIDcast project (Delphi research group) and the Kaiser Family Foundation (KFF).

Data Pre-processing

- This experiment involves 17 variables, which are divided into three categories: Socio-economic factors, Political factors, and Economic factors.
- Among the variables, governor_political_affiliation, state_senate_majority_political_affiliation, and state_house_majority_political_affiliation are categorical variables.
- Due to the small dataset size, several continuous variables, including average_monthly_snap_participants, number_of_births, population, total_private_health_insurance_spending, and unemployment_claims, exhibited skewness. To address this, these variables were log-transformed to reduce skewness.
- After performing log transformations and standardizing all continuous variables, further analysis was conducted.

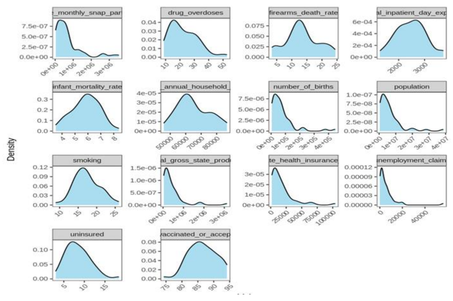

Fig 1 Density Plots Of Numerical Variables

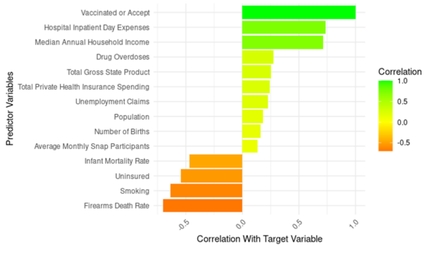

Fig 2 Correlation Between The Dependent Variable And The Independent Variables.

- Individually examining the relationship between each variable and the predicted variable, hospital inpatient day express, median annual household income, and firearms death rate are the most significant, all exceeding 0.70.

Analysis and results

First, we attempted to perform variable selection using best subset selection and stepwise regression methods.

Method	AIC	MSE	Selected Variables
Stepwise Regression	188.05	2.08	uninsured, median_annual_household_income, state_senate_majority_political_affiliation, total_gross_state_product, hospital_inpatient_day_expenses, population
Best Subset Selection (AIC)	186.42	3.37	infant_mortality_rate, median_annual_household_income, firearms_death_rate, state_house_majority_political_affiliation, hospital_inpatient_day_expenses, population
Best Subset Selection (BIC)	192.54	3.66	firearms_death_rate, state_house_majority_political_affiliation, hospital_inpatient_day_expenses

Table 1 Comparison of Variable Selection Methods

- It was observed that Best Subset Selection (AIC) yielded the most ideal results during the variable selection process.. The variables selected by this method demonstrated high correlation with the dependent variable.
- Key variables such as hospital inpatient day express, median annual household income, and firearms death rate were included in the model. This result indicates that the collinearity issue among the variables was effectively mitigated.

Model	Parameter Settings	Results
XGBoost Tuned	max_depth=5, eta=0.1, nrounds=200, gamma=1, colsample_bytree=0.8, min_child_weight=3	MSE = 1.576, R^2 = 0.911
Support Vector Regression (SVR) Tuned	gamma=0.1, cost=10, epsilon=0.1	MSE = 2.691, R^2 = 0.848
Neural Network Tuned	size=7, decay=0.1, maxit=240	MSE = 3.082
Ridge Regression	alpha=0, regularization parameter lambda selected via cross-validation	MSE = 7.042
LASSO Regression	alpha=1, regularization parameter lambda selected via cross-validation	MSE = 7.275
K-Nearest Neighbors (KNN)	k=3 (tuned using 5-fold cross-validation)	MSE = 11.36

Table 2 Comparison of Prediction Methods

From this analysis, it is evident that our tuned XGBoost model achieved the lowest MSE, reaching 1.576, while the R-square, representing the explanatory power of the entire model, reached 0.91, making it the most ideal.

The XGBoost model was fine-tuned using several key hyperparameters across the four variable categories:

- Learning Rate (eta): A smaller learning rate allowed the model to train more gradually, reducing the risk of overfitting.
- Tree Depth (max_depth): A moderate tree depth struck a balance between capturing complexity and avoiding overfitting.
- Column Subsampling (colsample_bytree): Adjusting the ratio of features per tree helped reduce overfitting while maintaining good performance.
- Minimum Child Weight (min_child_weight): Splits were made only when sufficient observations were present in the leaf nodes.

After evaluating all parameter combinations, the optimal hyperparameters were found to be nrounds=200, max_depth=5, eta=0.1, gamma=1, colsample_bytree=0.8, and min_child_weight=3.

Finally, grid search was applied on the training set to train models and evaluate their performance for each combination of hyperparameters.

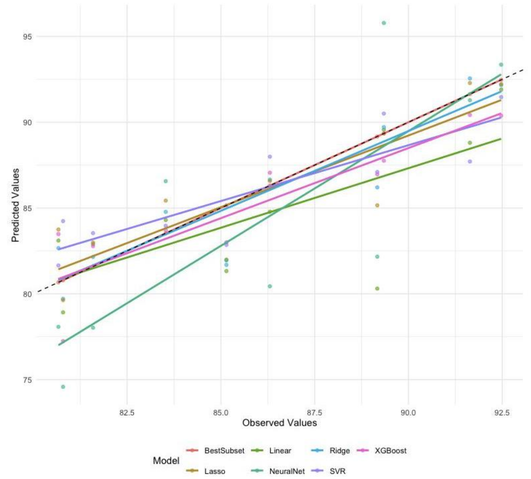

Fig 3 Observed VS. Predict Values Across Models

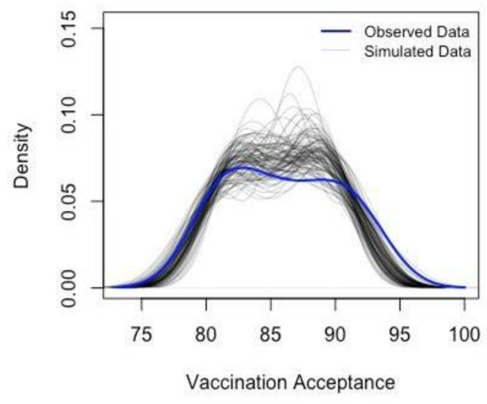

Fig 4 Tuned Check Plot: Tuned XGBoost

According to this two plots, we can see that:

- XGBoost (purple line) shows predictions that are closest to the diagonal line, indicating it has the highest accuracy among the models. BestSubset, Ridge, and Linear Regression also show relatively good alignment with the observed values, though with slight deviations.
- NeuralNet (green line) and SVR (yellow line) deviate more significantly from the diagonal line, indicating poorer predictive accuracy compared to XGBoost.
- Model Performance Consistency: XGBoost demonstrates consistent performance across different ranges of observed values, while NeuralNet and SVR struggle, especially at lower observed values.
- The results from Fig 4 reaffirm that the tuned XGBoost model is both accurate and robust, with minor variability in predictions for some data points.

Methods

- This project focused on achieving two main goals: explanation and prediction. The dataset was split into a training set (80%) and a test set (20%) to ensure that model performance could be reliably evaluated. Cross-validation was employed on the training data to optimize hyperparameters and assess model stability, while the mean squared error (MSE) on the test set was used as the primary metric to compare the predictive performance of different models.
- To identify key predictors, the project employed variable selection techniques, including stepwise regression and best subset selection. Stepwise regression relied on the Akaike Information Criterion (AIC) to balance model complexity and goodness of fit, while best subset selection exhaustively evaluated variable combinations to minimize AIC or BIC. These methods highlighted the importance of socio-economic and political factors in explaining the response variable.
- For prediction, the project explored a variety of regression and machine learning models, including LASSO, ridge regression, K-Nearest Neighbors (KNN), Support Vector Regression (SVR), XGBoost, and neural networks. Regularization-based methods such as LASSO and ridge regression were particularly useful for handling multicollinearity and reducing overfitting. KNN and SVR offered non-linear modeling capabilities, while neural networks were utilized to capture complex interactions, though they required substantial tuning. Among these, XGBoost demonstrated the best performance, achieving the lowest MSE and the most accurate predictions, making it the optimal choice for this dataset.

Conclusion

- The analysis demonstrated the effectiveness of machine learning models, particularly XGBoost, in predicting COVID-19 vaccination acceptance based on socio-economic and healthcare variables.
- Significant predictors included hospital inpatient day express, median annual household income, and firearms death rate, highlighting the interplay of socio-economic and political factors in vaccination behavior.
- While best subset selection provided useful insights, its performance was less consistent compared to XGBoost.
- Future work may involve exploring non-linear models or incorporating additional predictors to enhance prediction accuracy and understanding.
- Given the small dataset and the large number of variables, more data may be required in the future to ensure the stability and accuracy of the model.
- This project underscores the value of data-driven approaches in addressing public health challenges.

References

1.https://delphi.cmu.edu/covidcast/?date=20241130
2. https://www.kff.org/statedata/custom/