Predicting Covid Medical Absentees

Matthew Grohotolski
Elizabethtown College
GrohotolskiM@etown.edu

***Abstract*—Many medical institutions and insurance companies strive to provide the best customer satisfaction and many people have had a tough time keeping up with personal health during the COVID-19 pandemic. The purpose of this paper is to conduct statistical analysis on recently collected data about medical patients and COVID. After analyzing and making changes to the dataset, several machine learning models are trained and used for predicting the outcome for individuals who skip medical visits due to COVID-19.**

I. Introduction

A. Background

According to a study published in CDC in September 2020 [1], "As of June 30, 2020, an estimated 41% of U.S. adults reported having delayed or avoided medical care during the pandemic because of concerns about COVID-19, including 12% who reported having avoided urgent or emergency care. Medical care delay or avoidance might increase morbidity and mortality risk associated with treatable and preventable health conditions and might contribute to reported excess deaths directly or indirectly related to COVID-19."

B. Data

Our data is extracted from the IPUMS Health Surveys: NHIS, a harmonized set of data covering more than 50 years (1963-present) of the National Health Interview Survey [2]. On average, the survey covers 100,000 persons in 45,000 households each year. Our data is from the survey conducted in 2020 by IPUMS of 37,358 individuals across the USA. Of the 25 total predictor variables examined in the original data, only 10 were used in the modeling process, excluding the predicted variable. Within each of these 10 variables, specific "junk" values had to be dropped and the class values were cleaned. Additionally, all rows with an na value were dropped. For any variables which the class factors are not described for, more information can be found on the IPUMS website where the dataset was sourced from [2]. More details on which variables were chosen and the cleaning process is described below.

C. Variables

Age which is described by the name "AGE", reports the age in years as a quantitative variable. The "junk" values which resulted in row deletions were the values "997" and "998". The resulting bounds for this value after cleaning are ages 18-85.

Employment status which is described by the name "EMPSTAT", reports the employment status as a class variable. The "junk" values which resulted in row deletions were the values "0", "900", "997", "998", and "999". After this, the variable was converted to a factor with two values, 0 if a person is working and 1 if they are not currently employed.

Marital status which is described by the name "MARSTAT", reports the marital status as a class variable. The "junk" values which resulted in row deletions were the values "0" and "99".

Total combined family income which is described by the name "INCFAM07ON", reports the combined family income as a class variable. The "junk" values which resulted in row deletions were the values "96" and "99".

Education level which is described by the name "EDUC", reports the current level of education received as a class variable. The "junk" values which resulted in row deletions were the values "0", "996", "997", "998", and "999".

Health insurance coverage status which is described by the name "HINOTCOVE", reports the health care insurance coverage status as a class variable. The "junk" values which resulted in row deletions were the values "0", "7", "8", and "9".

Race which is described by the name "RACEA", reports the race as a class variable. The "junk" values which resulted in row deletions were the values "900", "970", "980", and "990".

Region which is described by the name "REGION", reports the region the person lives in the U.S. as a class variable. This column did not contain any "junk" values which needed to be deleted.

Sex which is described by the name "SEX", reports the sex of the person as a class variable. The "junk" values which resulted in row deletions were the values "7", "8", and "9". After this, the variable was converted to a factor with two values, 0 if a person male and 1 if they are female.

Urban area status which is described by the name "URBRRL", reports the urban area status of the place the person resides as a class variable. This column did not contain any "junk" values which needed to be deleted.

Lastly, the **medical visit skip status** which is described by the name "CVDDNGCARE", and reports if the person delayed or avoided medical care because of COVID-19 concerns. The "junk" values which resulted in row deletions were the values "0", "7", "8", and "9".

After this, the variable was converted to a factor with two values, 0 for not avoiding medical care, and 1 for delaying or avoiding medical care.

D. Exploratory Data Analysis

Fig. 1 compares the sex of people in the dataset with the medical visit skip status to test whether or not there may be any high correlation. Just from looking at the correspondance between these values, it does not seem like there is high correlation. This is further reinforced by a graph of feature importance amongst predictor variables shown in Fig. 2 that displays the "SEX" variable to have the second least contribution to accuracy gain in models. The feature importance figure was created during the modelling phase and resulted from the random forest model.

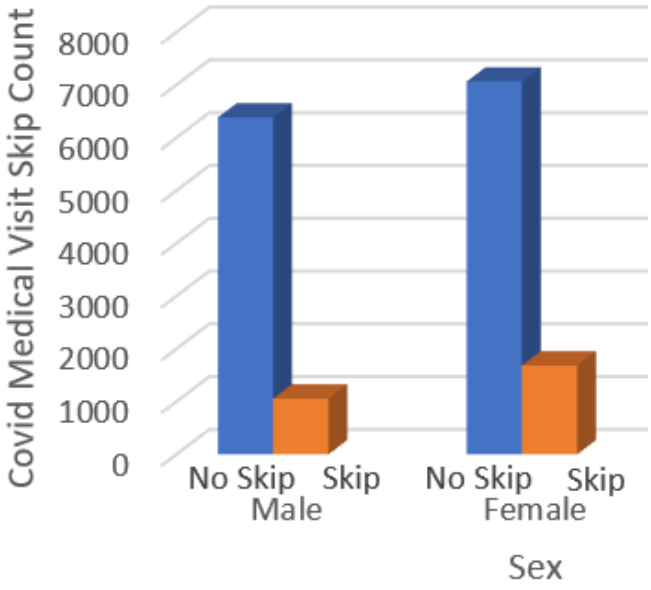

Fig. 1. Gender*Predictor Correlation

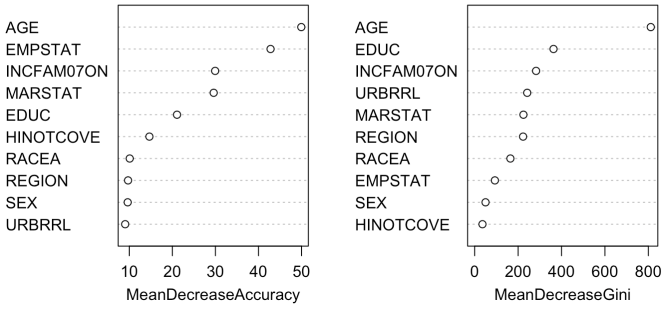

Fig. 2. Feature Importance

Fig. 3 shows a plot of the distribution of the CVDDNGCARE predicted variable. It is apparent that bias exists towards not delaying or avoiding medical care, which could have an impact during modelling.

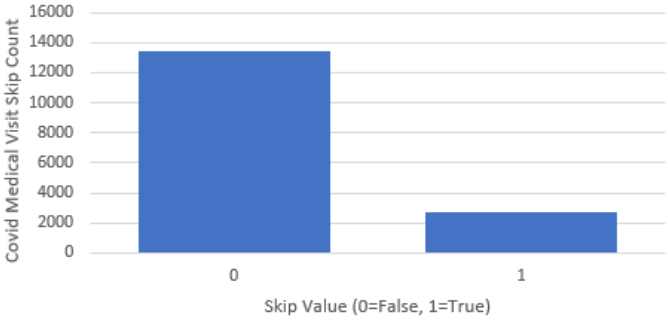

Fig. 3. CVDDNGCARE Distribution

II. Methods & Results

Each of the below models utilized the same simple random sample of 80% of the dataset for training with the remaining 20% for model testing and prediction analysis. Confusion matrices are provided for each model to compare their predictive accuracies.

A. Decision Forest

The first and simplest modelling method to be applied to the dataset was Decision Forest which creates decision and root nodes to predict given a set of explanatory variables. The confusion matrix results are shown in Tbl. I. As seen in the predictions, Decision Forest (DF) has been overfitted to only predict 0, prioritizing accuracy over a properly fitted model. The test accuracy for this model is 83.0677%.

Although overfitting occurred here, DF is able to output an importance graph to identify the most important variables during prediction which is shown in Fig. 2.

TABLE I
Decision Forest Confusion Matrix

	0	1
0	13447	2741
1	0	0

B. Random Forest

Random Forest (RF) was the next model to be fitted to the dataset and generally performed a lot better than the DF model in terms of generalizing the model, since at each break in the prediction tree a random explanatory variable is chosen and given a cutoff, randomizing our results and providing a proper prediction. When taking a look at the confusion matrix shown in Tbl. II, it is clear that while the accuracy is lower than that of DF, it does a better job when predicing response values of 1 while keeping type 1 and type 2 errors low. The test accuracy for this model is 89.68989%.

TABLE II
RANDOM FOREST CONFUSION MATRIX

	0	1
0	13432	1654
1	15	1087

C. Boosting

Boosting is similar to DF except it boosts the model by creating n amount of Decision Forests to use when predicting. For our model, an n value of 5000 was chosen in an attempt to better generalize the predictive responses from the model. A confusion matrix is shown in Tbl. III. The confusion matrix seems to suggest the 5000 Decision Forests prioritized predictive accuracy once again and only predicts 0 for all data in the test set. The test accuracy for this model is 83.0677%.

TABLE III
BOOSTING CONFUSION MATRIX

	0	1
0	13447	2741
1	0	0

D. Ridge Regression

Instead of using decision trees, Ridge Regression (RR) utilizes "shrinkage" to shrink all explanatory variable coefficients towards zero. Cross validation was performed to further optimize the model; A grid of Lambda (λ) values ranging from 0.005 to 4 with an interval of 0.025 were used during training to approximate the best resulting model and is shown in Fig. 4.

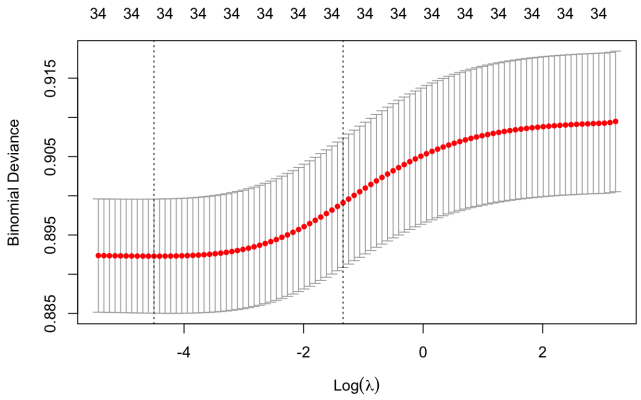

Fig. 4. Ridge Regression λ Optimization Results

After lambda optimization was applied, I used a range of cutoff values which change the cutoff for producing a 1 during prediction. This visualization for this process is shown below:

$$\hat{y} > c = 1$$

Where c is the current cutoff value being used and 0 is used as a default value incase $\hat{y}$ is larger than c. Fig. 5 displays the process of trying different cutoffs and it appears that test accuracy remains the same throughout. The test accuracy for this model is 83.0677%.

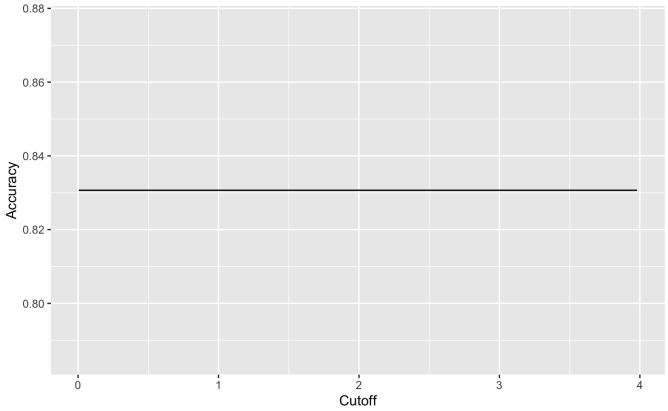

Fig. 5. Ridge Regression Cutoff Optimization Results

E. Lasso Regression

Lasso Regression (LR) works similarly to RR as it also utilizes "shrinkage" to shrink all explanatory variable coefficients. Except coefficients have the possibility to shrink to zero with LR. For training, cross validation was performed to further optimize the model; an identical grid of Lambda (λ) values from RR were used during training to approximate the best resulting model and is shown in Fig. 6.

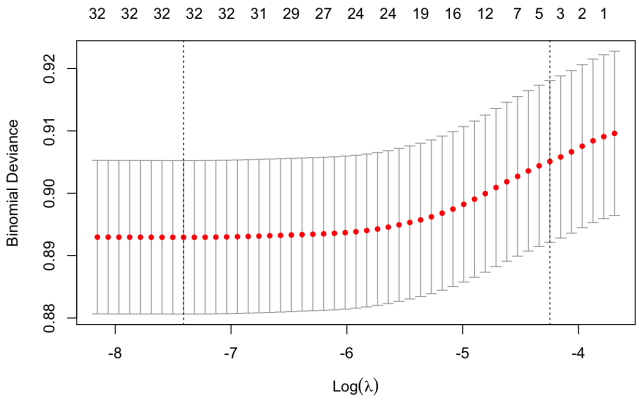

Fig. 6. Lasso Regression λ Optimization Results

After lambda optimization was applied, an identical set of cutoff values were used similar to RR. Fig. 7 displays the process of trying different cutoffs and it appears that test accuracy remains the same throughout. The test accuracy for this model is 83.0677%.

III. DISCUSSION

A. Best Model

The most effective model when predicting was Random Forest which obtained not only the highest test accuracy

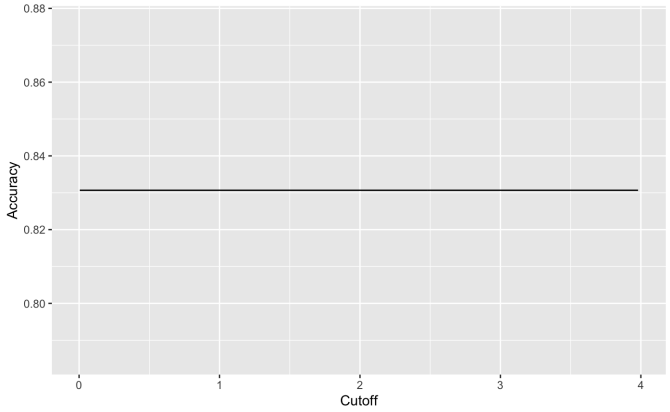

Fig. 7. Lasso Regression Cutoff Optimization Results

of 89.68989%, but was also generalized enough to begin producing prediction values of 1. For these reasons, it is sufficient to say this model would perform well in real-world scenarios. One example of how this could be helpful is if a medical center sends a follow-up phone call during the COVID-19 pandemic to check on patients who have been predicted to either delay or avoid medical care. This could have the potential to increase profit margins and improve the business in a positive way.

B. Limitations

A few of the models were over-generalized including Decision Forest, Boosting, and both Ridge and Lasso Regression. These models always chose to predict 0 which is most likely caused by our y value's distribution shown in Fig. 3. One potential fix for this would be to have the distribution between 0 and 1 for the predicted value be approximately equal to each other.

IV. R Code

The R code used in the study is located in a GitHub file. View references for the link [3].

References

[1] M. Czeisler, K. Marynak, K. E. Clarke, Z. Salah, I. Shakya, J. M. Thierry, N. Ali, H. McMillan, J. F. Wiley, M. D. Weaver, C. A. Czeisler, S. M. Rajaratnam, and M. E. Howard, "Delay or avoidance of medical care because of covid-19–related concerns," 2020. [Online]. Available: https://www.cdc.gov/mmwr/volumes/69/wr/mm6936a4.htm

[2] B. A. Lynn, R. A. Julia, R. Drew, L. Miriam, and W. C. Kari, "Ipums usa: Version 6.4 [dataset]," 2019. [Online]. Available: https://doi.org/10.18128/D070.V6.4

[3] M. Grohotolski, "Ds315 covid project," 2021. [Online]. Available: https://github.com/MattBcool/DS315-Covid-Project