TravNur
Improving Healthcare through Data Mining
Comp 541 Data Mining | Project 6

Verab Chitchyan, Gabriella de Asis, Emmanoel Dermkrdichyan,
Carson Logston, Ahn (Steven) Nguyen

Modeling Technique

Initially, we tried the Random Forest Classifier with the idea that given the number of patients, this model would be able to determine if a staffing shortage were to occur. However, this resulted in only a 33% accuracy. This was due to a mistaken understanding concerning our target parameter. Our dataset had too many possible solutions because our target parameter states the number of hospitals with shortages, not a classification of shortage status on a hospital. So, we moved on to a linear regression model instead. Our hypothesis was that an increase in COVID patients would result in staffing shortages.

Training and Testing Dataset

We split our dataset into 80% for training and 20% for testing. We trained on 11,392 rows and tested on 2,849.

Parameters

We have three sets of parameters for three different linear regression models. Our first set had three parameters :

- critical staffing shortage one week later (target)
- total number of adult patients hospitalized with confirmed COVID
- total number of pediatric patients hospitalized with confirmed COVID

For the sake of readability, we are shortening parameter names as follows:

- critical staffing shortage one week later = staff shortage
- critical staffing shortage anticipated within a week = anticipated shortage
- total number of adult patients hospitalized with confirmed COVID = confirmed adult cases
- total number of adult patients hospitalized with confirmed and suspected COVID = confirmed and suspected adult cases
- total number of pediatric patients hospitalized with confirmed COVID = confirmed pediatric cases
- total number of pediatric patients hospitalized with confirmed and suspected COVID = confirmed and suspected pediatric cases

For the second model, we added anticipated shortage as a parameter and kept staff shortage as the target. For the third model, we added all six of our original parameters, even though we had initially dropped two of those features in Project 5. We added them just to see how the model would perform with those additional parameters.

Results

For our first model, we only had a 46.7% accuracy. Upon adding the anticipated shortage parameter for the second model, our accuracy went up to 83.7%. Adding the last two parameters for the third model actually brought accuracy down to 83.1%.

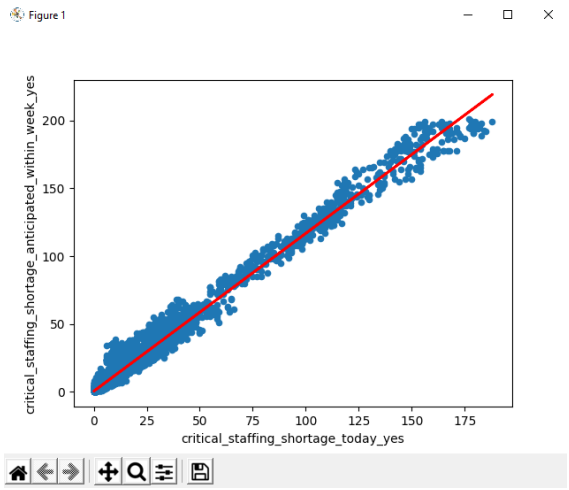

Figure 1: The linear regression prediction model (shown in red) visually represented with the correlation between the target value, critical staff shortage, and the primary attribute, anticipated shortage.

Testing Method

For our testing method, we set the model to Y_pred, with the X_test matrix as the parameter. We then compared the actual Y test matrix to Y_pred using the r2_score function.

Ranking of Models

Based on the accuracy results, we found that the second model, which only contains four attributes, is the most accurate. This makes sense since we are looking at the number of confirmed cases and if a hospital is anticipating a shortage. The third model is still very accurate and we believe that the difference of 0.6% is because not all suspected COVID cases are positive for COVID. The first model, which only has a 46.7% accuracy indicates that COVID cases are not the only reason for staffing shortages. There are more factors such as other illnesses, funding, etc. that affect shortages within state hospitals. This is why adding anticipated shortages as a parameter increased the accuracy by 37%. The anticipated staffing shortages parameter takes those other factors that we do not have access to, into account.

Addressing Objectives

TravNur's business objective is to provide travel nurses to areas facing staffing shortages. Our data mining criteria for success was to predict nurse shortages with a 70% accuracy one week prior to the demand. This model is able to determine shortages a week in advance with an 83% accuracy, which is much higher than what we initially anticipated our model to be. With this model, TravNur will be able to determine areas where shortages will occur and allocate nurses accordingly.

Other Discussions

At the beginning of the semester, we were looking at two datasets: the nationwide hospital survey on staffing shortages, which we used for our model, and a California specific dataset which looked at 71 areas, their population, number of licensed nurses, and nurse/patient ratio. They then classified the area as having a high, medium, or low nurse shortage severity. However, this dataset later became unavailable and we were only able to download information for one day of that dataset. Because this was such a small dataset, we decided to focus solely on the nationwide survey.

Since the random forest classification model did not work on the nationwide dataset, we tried running this model on the California nurse shortage dataset. With this dataset, the RFC was able to give predictions with an accuracy of 83-100% depending on the training and test dataset. We are doubtful that this model would have a 100% accuracy with a larger dataset. Because of its size, we decided to stick with the linear regression model. The shape of the California dataset is (58, 11) once the null values were removed. The training set's shape is (46,3) for the X parameter and (46,) for the Y parameter. The test set's shape is (12,3) for the X parameter and (12,) for the Y parameter. In the future, if the entirety of this dataset becomes available, this model could be used to predict nurse shortages based on population and nurse/patient ratio.