Yimin Chen

P8106: Data Science II

Midterm Report

**Exploratory analysis and data visualization:**

In this section, use appropriate visualization techniques to explore the dataset and identify any patterns or relationships in the data.

This dataset includes 10000 participants who had their birthweight recorded at the time of their hospital visit. The mean total charges for a patient at discharge across the 49 hospital facilities was $19,787.95. The median total charges for a patient was $6,590.02. For exploratory analysis, we compared the median total charges because there were extreme outliers with very high charges. The median is a better measure of central tendency than the mean. The median total charges varied widely across the 49 facilities included in this analysis. The median total charges for a patient at discharge were highest at Montefiore Medical Center - Henry & Lucy Moses Div ($25,160.36) and lowest at North Central Bronx Hospital ($2,407.01) (Table 1).

Predictor variables included: 1. gender (1 = Male, 0 = Female), 2. race (1 = White, 2 = Asian, 3 = Black, 4 = Hispanic), 3. smoking(Smoking status; 0 = Never smoked, 1 = Former smoker), 4. height(in centimeters), 5. weight(in kilograms), 6. bmi (Body Mass Index; BMI = weight (in kilograms) / height) , 7. hypertension (0 = No, 1 = Yes), 8. diabetes (0 = No, 1 = Yes), 9.Systolic blood pressure (SBP) (in mm/Hg), 10. LDL (low-density lipoprotein) cholesterol (in mg/dL), 11.vaccine (0 = Not vaccinated, 1 = Vaccinated), 12).severity (0 = Not severe, 1= Severe), 13).study (The study (A/B/C) that the participant belongs to)

Outcome/response variable is Time to recovery (recovery\_time), which measures time from COVID-19 infection to recovery in days.

All categorical variables were made into dummy variables.

Three predictors were continuous (length of stay, birthweight, and total costs) as well as the outcome variable (total charges).

**Model training:**

In this section, describe the models you used for predicting time to recovery from COVID-19. State the assumptions made by using the models. Provide a detailed description of the model training procedure and how you obtained the final model.

We fit 5 different models using cross validation to determine which model would fit the data best:

* One linear model,
* Three regularized linear models (Lasso regression, Ridge regression, PCR), and
* One non-linear model (Generalized Additive Model).

/\*

We then use the train function from the caret package to fit a linear regression model to the training data using cross-validation with 10 folds. The trainControl function specifies the cross-validation method, and the method argument specifies the type of model to fit (in this case, linear regression). The resulting model object contains the final model and information about the cross-validation performance.

We can view the summary statistics for the final model using the summary function on the finalModel object within the model object. We can also use the predict function to generate predictions for the test set using the final model, and compute the root mean squared error (RMSE) between the predicted and actual recovery times on the test set.

This approach allows us to obtain a final model that has been trained using cross-validation, which can help to reduce overfitting and improve the generalization performance of the model on new, unseen data.

\*/

Predictor variables included: 1. gender (1 = Male, 0 = Female), 2. race (1 = White, 2 = Asian, 3 = Black, 4 = Hispanic), 3. smoking(Smoking status; 0 = Never smoked, 1 = Former smoker), 4. height(in centimeters), 5. weight(in kilograms), 6. bmi (Body Mass Index; BMI = weight (in kilograms) / height) , 7. hypertension (0 = No, 1 = Yes), 8. diabetes (0 = No, 1 = Yes), 9.Systolic blood pressure (SBP) (in mm/Hg), 10. LDL (low-density lipoprotein) cholesterol (in mg/dL), 11.vaccine (0 = Not vaccinated, 1 = Vaccinated), 12).severity (0 = Not severe, 1= Severe), 13).study (The study (A/B/C) that the participant belongs to)

Outcome/response variable is Time to recovery (recovery\_time), which measures time from COVID-19 infection to recovery in days.

All categorical variables were made into dummy variables.

Three predictors were continuous (length of stay, birthweight, and total costs) as well as the outcome variable (total charges).

Our dataset (92 predictors total, with dummy variables) was partitioned into a training and test data set. Five models were fitted using the training data and the mean squared error (MSE) was calculated for each model using the test data. We measured the mean squared error to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation.

First, we fit a linear model using least squares on all the predictors in the training data. We found that the MSE calculated on the test data was 937604771. By looking at the coefficient values for the linear model, some of the dummy variables were not significant at an alpha value of 0.05 in predicting total charges, including ethnicity, emergency department indicator, and certain CCS diagnosis types.

Next, we fit two models on all the predictor variables using two different techniques that "shrink" the coefficient estimates towards zero, which reduces variance. We fit a ridge regression model on the training data. Alpha was held at a value of 0 and our final lambda value chosen by cross-validation was 5355.389. Using our test data, we found that test error (MSE) was 9790669585. Next, we fit a lasso model on the training data. Alpha was held at a value of 1 and our final lambda value chosen by cross-validation was 28.03162. Our test error was 937116653.

We also fit a principle component regression (PCR) model on the training data with M chosen by cross validation. The PCR method constructs the first M principal components. Our M-value was 87 and our test error was 937635353. A benefit of using PCR is that this method avoids multicollinearity between variables in the dataset. A limitation of PCR is that there is no guarantee that Z­m­ are the best linear combinations of the variables in predicting the response.

Lastly, we fit a GAM model to our training dataset. The benefit of using a GAM model is that it will automatically model non-linear relationships that standard linear regression would miss. S functions were applied to the three continuous variables in the model (total costs, length of stay, and birthweight).

The best model for predicting total charges was PCR. This model was chosen because it has the smallest cross validation RMSE, a value of 28173.68 (Figure 3). For the PCR model, 85.69% of the variance in total charges is explained by the predictors in the model. The mean RMSE value of the lasso regression model (28174.24) and linear regression model (28180.84) were very close to the PCR model (depicted in Figure 4). Overall, GAM had the lowest RMSE (29076.52), which was unexpected because it was the only model to account for non-linear relationships that a standard linear regression model would miss.

For the PCR model, the top 10 variables that played important roles in predicting a patient’s total charges are displayed in Figure 4. These variables were chosen using the varImp() function, which automatically chooses a measure of variable importance that is appropriate for given algorithms. The top five variables included length of stay, total costs, APR risk of mortality (major), APR risk of mortality (minor), and the APR severity of illness code. The greatest importance in predicting total healthcare charges was length of stay in the hospital, which aligns with the linear relationship depicted in Figure 1.

**Results:**

In this section, report the final model that you built for predicting time to recovery from COVID-19. Interpret the results. Assess the model's training/test performance.

Limitations of this analysis included that non-significant variables were included in models and not dropped to improve the model. This was done in an effort to compare models, by first including all predictors in the model building phase. Thus, if we were to drop variables that are neither clinically meaningful nor significant, this may have created more parsimonious and efficient models. Additionally, the highest R^2 value was 85.69%. The remaining variance of total charges are thus unexplained and may be due to variables that were dropped (such as patient’s zip code) or by features not collected in this dataset.

**Conclusion:**

In this section, summarize your findings from the model analysis and discuss the insights gained into predicting time to recovery from COVID-19.

Of all models, the PCR model best predicted total healthcare charges of patients at time of discharge. The PCR model had very similar RMSE values to the linear model and lasso regression model, and thus, they have similar prediction performance. By listing the most important variables in predicting charges, we gained insight into the most important predictors for healthcare charges, particularly gaining knowledge that length of stay and total costs have much greater importance in predicting the charges compared to other variables. Birthweight is the top 6th highest importance, which is expected, as it often recorded in literature. It was of valuable insight to learn that the APR risk of mortality (minor, major, and moderate) has strong importance in determining total charges as well as congenital anomalies and Urgent admission types.

**Appendix**

**Figure 1: Total Charges by Length of Stay in NYC Hospitals**

**Chart, scatter chart

Description automatically generated**

**Figure 2: Total Charges Across Values of Baby Birthweight in NYC Hospitals**

**Chart, scatter chart

Description automatically generated**

**Figure 3: RMSE of GAM, Ridge, Lasso, Linear, and PCR Models**

**Chart, bar chart

Description automatically generated**