Harmony Healthcare - Feature Selection

Hibah Arshad, Samantha Nadler, Thomas Walsh

Problem Statement

This project’s objective is to help Harmony HealthCare determine which patient characteristics are the most predictive of ER visits. By identifying which variables, like age, BMI, cholesterol, or chronic conditions, are most closely associated with emergency room visits, the organization can create more efficient methods to cut down on avoidable ER admissions. Patient education, preventative care programs, and customized health treatments are a few examples of these methods. It is our responsibility as student consultants to evaluate the data statistically and offer useful findings that can help in healthcare decision-making.

Per Nigel Glynn et al. (2011), about 27% of patients who are admitted to the emergency room readmit themselves, usually within over a year after their initial visit. When considering factors that contribute to an admitted patient being more likely to return to the emergency room, Glynn et al. concluded that age played an important role in making such a determination. Readmitted patients were an average 67.2 years old, compared to all admitted patients being an average 57.8 years old.

As for diagnoses, respiratory disorders (including severe asthma, bronchitis, pneumonia, and more recently COVID and RSV) were the most common among readmitted patients, taking 22% of readmission causes. Following were nervous disorders (including strokes and seizures) with 17.3% of readmissions, cardiovascular disorders (including heart attacks, heart failure, and hypertension) with 16.5%, and digestive disorders (including dehydration, severe abdominal pain, and IBS) with 11.5%. All in all chronic condtions is a true indicator someone will be readmitted into ER’s.

Data Visualization

The data was first sorted through and cleaned roughly by a Python script as the first part of the project was incomplete. The data visualization was then used to:

- Show the distribution of ER readmission
- Show patient details, such as age and indicators of a chronic illness
- Show the LASSO model coefficient outputs to visualize the most predictive parameters (such as age, blood pressure, and A1C levels)

The following code in R separates the response (y) and predictor (x) variables. It uses mean to deal with missing data. It uses cross-validation and `glmnet` for LASSO. The output selected characteristics and their coefficients.

```r
library(readr)
HH_Data <- read_csv("C:/Users/nadle/Downloads/cleanedDataV2.csv")
```

```
## Rows: 2821 Columns: 139
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (95): EHR.Sex, Ethnicity, Language, Race, Refugee, Sex.at.Birth, Usual....
## dbl  (41): UDS.Qualifying.Encounter.Count, Active.Medications, Alcohol.Asses...
## lgl   (1): Public.Housing
## date  (2): Age, Obesity
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(HH_Data)
```

```
## # A tibble: 6 × 139
##   Age        EHR.Sex Ethnicity               Language Race  Refugee Sex.at.Birth
##   <date>     <chr>   <chr>                   <chr>    <chr> <chr>   <chr>
## 1 1900-01-27 female  another hispanic, lati… spanish  white ignore  female
## 2 1900-01-21 female  another hispanic, lati… spanish  amer… ignore  female
## 3 1900-02-17 female  another hispanic, lati… spanish  unre… ignore  female
## 4 1900-02-10 male    not hispanic, latino/a… english  blac… ignore  male
## 5 1900-02-17 female  another hispanic, lati… spanish  white ignore  female
## 6 1900-02-08 female  puerto rican            english  blac… ignore  female
## # ℹ 132 more variables: Usual.Location <chr>, Usual.Provider <chr>,
## #   Veteran.Status <chr>, Zip <chr>, Most.Recent.Encounter.Type <chr>,
## #   UDS.Qualifying.Encounter.Count <dbl>, UDS.Homelessness.Status <chr>,
## #   Active.Medications <dbl>, ACE.ARB.Med.Name <chr>,
## #   Alcohol.Assessment.Code <chr>, Alcohol.Assessment.Result <dbl>,
## #   Anti.HTN.Med.Name <chr>, Patient.Appointment.No.Show.Count <dbl>,
## #   Patient.Appointment.No.Show.Rate.. <dbl>, AST.Result <chr>, …
```

```
# Still run the following two lines
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# HH_Data$`ED Episode Admit Last 6 Mths` <- as.numeric(HH_Data$`ED Episode Admit Last 6 Mths`)
# HH_Data <- HH_Data %>% replace_na(list(`ED Episode Admit Last 6 Mths` = 0))
#
# # Remove columns with more than 50% missing values
# col_missing <- colMeans(is.na(HH_Data))
# HH_Data <- HH_Data[, names(HH_Data)[col_missing <= 0.5]]
#
# # Remove rows with more than 10% missing values
# row_missing <- rowMeans(is.na(HH_Data))
# HH_Data <- HH_Data[row_missing <= 0.1, ]
#
# # View the cleaned dataset
# View(HH_Data)
```

```
y <- HH_Data$Admission
x <- HH_Data %>%
  select(where(is.numeric), -Admission) %>%
  as.matrix()
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.4.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

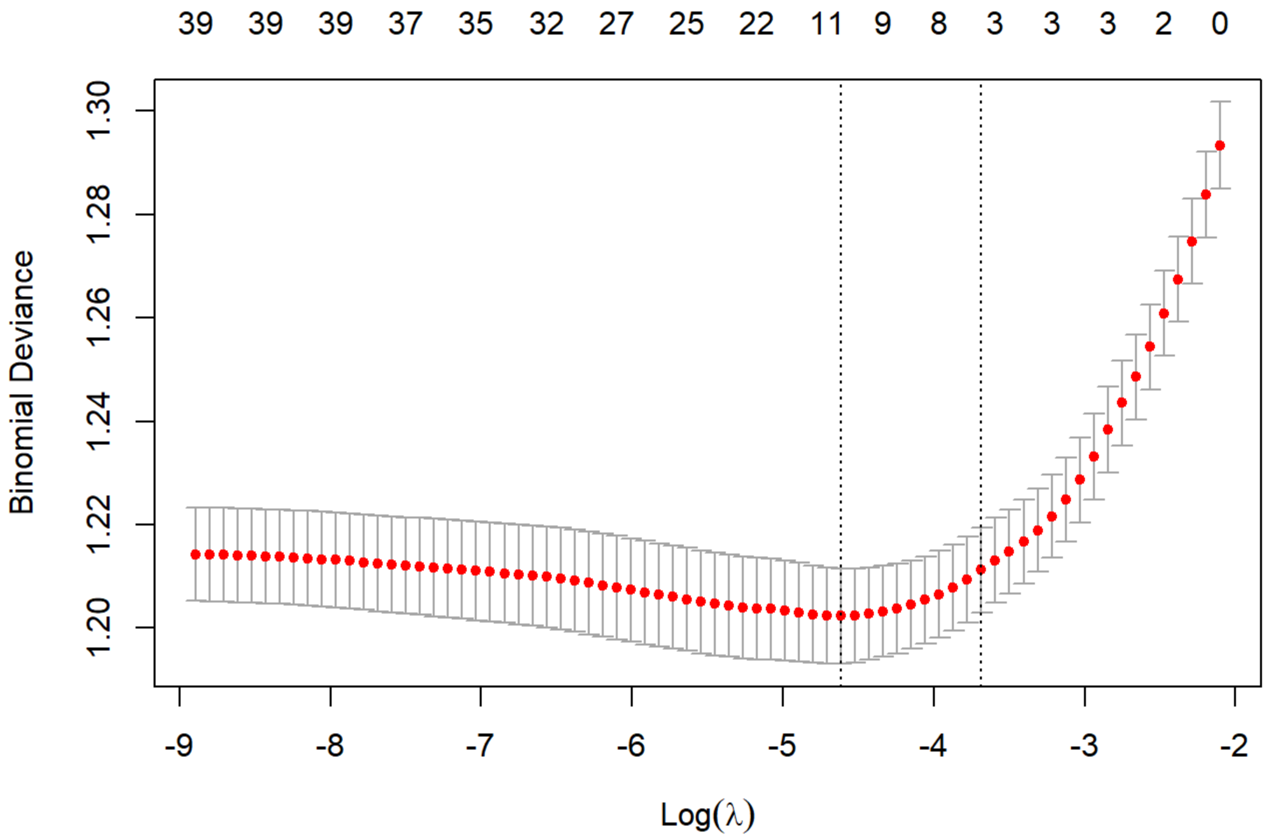
```
## Loaded glmnet 4.1-8
```

```
set.seed(42)

lasso_model <- cv.glmnet(x, y, alpha = 1, family = "binomial")
best_lambda <- lasso_model$lambda.min
cat("Best lambda (from cross-validation): ", best_lambda, "\n")
```

```
## Best lambda (from cross-validation):  0.009912212
```

```
plot(lasso_model)
```



```
coef_lasso <- predict(lasso_model, type = "coefficients", s = best_lambda)
coef_lasso_df <- as.data.frame(as.matrix(coef_lasso))
colnames(coef_lasso_df) <- "coefficient"
coef_lasso_df$feature <- rownames(coef_lasso_df)

coef_lasso_df <- subset(coef_lasso_df)#, feature != "(Intercept)" & coefficient != 0)
coef_lasso_df[order(abs(coef_lasso_df$coefficient), decreasing = TRUE), ][1:20, ]
```

```
##                                      coefficient
## (Intercept)                        -2.2242630140
## Patient.HCC.Risk.Total.Risk        -0.0446898215
## Active.Medications                  0.0443919593
## Primary.Care.Encounter.Count        0.0331995417
## SDOH.Assessment.Count               0.0146984831
## Patient.Appointment.No.Show.Rate..  0.0146875578
## Depression.Screening.Count.Past.Yr  0.0086325980
## eGFR.Result                         0.0067602679
## Most.Recent.BMI.Value               0.0037749693
## UDS.Qualifying.Encounter.Count      0.0032103970
## COVID.19.Immunization.Code         -0.0006886961
## Fasting.Glucose.Test.Result        -0.0006487510
## Alcohol.Assessment.Result           0.0000000000
## Patient.Appointment.No.Show.Count   0.0000000000
## Blood.Pressure.Systolic             0.0000000000
## Blood.Pressure.Diastolic            0.0000000000
## Patient.Medicaid.Risk.Total.Risk    0.0000000000
## Patient.Medicaid.Risk.Risk.Gap      0.0000000000
## Cholesterol.Result                  0.0000000000
## Dental.Encounter.Count              0.0000000000
##                                                                  feature
## (Intercept)                                                  (Intercept)
## Patient.HCC.Risk.Total.Risk               Patient.HCC.Risk.Total.Risk
## Active.Medications                                  Active.Medications
## Primary.Care.Encounter.Count             Primary.Care.Encounter.Count
## SDOH.Assessment.Count                           SDOH.Assessment.Count
## Patient.Appointment.No.Show.Rate.. Patient.Appointment.No.Show.Rate..
## Depression.Screening.Count.Past.Yr Depression.Screening.Count.Past.Yr
## eGFR.Result                                                  eGFR.Result
## Most.Recent.BMI.Value                           Most.Recent.BMI.Value
## UDS.Qualifying.Encounter.Count         UDS.Qualifying.Encounter.Count
## COVID.19.Immunization.Code                 COVID.19.Immunization.Code
## Fasting.Glucose.Test.Result               Fasting.Glucose.Test.Result
## Alcohol.Assessment.Result                   Alcohol.Assessment.Result
## Patient.Appointment.No.Show.Count   Patient.Appointment.No.Show.Count
## Blood.Pressure.Systolic                       Blood.Pressure.Systolic
## Blood.Pressure.Diastolic                     Blood.Pressure.Diastolic
## Patient.Medicaid.Risk.Total.Risk     Patient.Medicaid.Risk.Total.Risk
## Patient.Medicaid.Risk.Risk.Gap         Patient.Medicaid.Risk.Risk.Gap
## Cholesterol.Result                                  Cholesterol.Result
## Dental.Encounter.Count                         Dental.Encounter.Count
```

Are the Results Significant?

By reducing irrelevant variable coefficients to zero, LASSO regression produced a clear list of relevant predictors. Our model kept characteristics such as:

- HCC Risk Score: Chronic Condition Risk
- Active Medications
- Missed Appointments
- BMI

- Mental Metrics
- Covid 19 Vaccine Status

The importance of these findings was supported by their consistency with published medical research on readmission risk. In order to prevent overfitting, we also employed cross-validation to determine the ideal lambda value.

Discussion

Our results were consistent with the background research. Higher ER use was linked to chronic illnesses. Research also showed that Covid 19 was a predictor. Social and mental variables, like depression from the lasso also had a moderate impact. We plan on looking into variables including a patient’s age for future research as for now the column was corrupted.

Contributions

Thomas: Background Research, slides

Hibah: Updating final R markdown, testing lasso with logistic, slides

Samantha: Lasso code, working with professor, slides