Assignment 7 Predictions

Nicholas Jacob

2024-10-29

Looking at Table 1, we see a few missing values for indicators. These will not be utilized in the analysis. We'll impute the time and lab procedures to zeros. Readmitted, admission type and source, and discharge will need to be treated as a factor as well so it will need to be removed.

Table 1: Descriptive Summary of Numeric Variables

variable	n	missing	missing_pct	unique	unique_pct	mean	min	Q1	median	Q3	max	sd
patientID	57855	0	0.00	57855	100.00	4.9e+04	1001	25106.5	49212	73317	97421	2.8e+04
admission_type	57855	0	0.00	8	0.01	2.0e+00	1	1.0	1	3	8	1.5e+00
discharge_disposition	57855	0	0.00	22	0.04	3.0e+00	1	1.0	1	3	28	4.5e+00
admission_source	57855	0	0.00	17	0.03	5.8e+00	1	1.0	7	7	25	4.1e+00
time_in_hospital	57855	4	0.01	15	0.03	4.4e+00	1	2.0	4	6	14	3.0e+00
indicator_level	57855	6	0.01	9975	17.24	5.0e+01	-869	24.8	50	75	999	3.0e+01
indicator_2_level	57855	28809	49.80	7432	12.85	2.5e+01	-682	6.9	19	38	99	2.2e+01
num_lab_procedures	57855	1	0.00	115	0.20	4.3e+01	1	31.0	44	57	132	2.0e+01
num_procedures	57855	0	0.00	7	0.01	1.3e+00	0	0.0	1	2	6	1.7e+00
num_medications	57855	0	0.00	74	0.13	1.6e+01	1	10.0	15	20	81	8.1e+00
number_outpatient	57855	0	0.00	37	0.06	3.9e-01	0	0.0	0	0	42	1.3e+00
number_emergency	57855	0	0.00	30	0.05	2.1e-01	0	0.0	0	0	64	9.6e-01
number_inpatient	57855	0	0.00	19	0.03	6.3e-01	0	0.0	0	1	19	1.2e+00
number_diagnoses	57855	0	0.00	16	0.03	7.4e+00	1	6.0	8	9	16	1.9e+00
readmitted	57855	0	0.00	2	0.00	4.7e-01	0	0.0	0	1	1	5.0e-01

Table 2: Descriptive Summary of Categorical Variables

variable	n	missing	missing_pct	unique	unique_pct	freqRatio	1st mode	first_mode_freq	2nd mode	second_mode_freq
race	57855	1313	2.27	6	0.01	4.07	Caucasian	43515	AfricanAmerican	10694
gender	57855	5	0.01	4	0.01	1.16	Female	31029	Male	26820
age	57855	1	0.00	11	0.02	1.15	[70-80)	14694	[60-70)	12796
payer_code	57855	21515	37.19	4	0.01	1.33	medicare	18985	insurance	14302
medical_specialty	57855	27810	48.07	70	0.12	1.89	InternalMedicine	8460	Emergency/Trauma	4468
diagnosis	57855	11	0.02	668	1.15	1.03	428	3859	414	3746
max_glu_serum	57855	0	0.00	4	0.01	35.40	None	54736	Norm	1546
A1Cresult	57855	0	0.00	4	0.01	10.23	None	48099	>8	4701
metformin	57855	0	0.00	4	0.01	4.43	No	46445	Steady	10478
repaglinide	57855	0	0.00	4	0.01	71.20	No	56962	Steady	800
nateglinide	57855	0	0.00	4	0.01	143.58	No	57432	Steady	400
chlorpropamide	57855	0	0.00	3	0.01	1700.53	No	57818	Steady	34
glimepiride	57855	0	0.00	4	0.01	19.93	No	54814	Steady	2750
acetohexamide	57855	0	0.00	2	0.00	57854.00	No	57854	Steady	1
glipizide	57855	0	0.00	4	0.01	7.64	No	50469	Steady	6610
glyburide	57855	0	0.00	4	0.01	10.28	No	52063	Steady	5066
tolbutamide	57855	0	0.00	2	0.00	3856.00	No	57840	Steady	15
pioglitazone	57855	0	0.00	4	0.01	13.01	No	53531	Steady	4114
rosiglitazone	57855	0	0.00	4	0.01	15.34	No	54166	Steady	3530
acarbose	57855	0	0.00	4	0.01	362.81	No	57687	Steady	159
miglitol	57855	0	0.00	4	0.01	3213.00	No	57834	Steady	18
troglitazone	57855	0	0.00	2	0.00	28926.50	No	57853	Steady	2
tolazamide	57855	0	0.00	2	0.00	3044.00	No	57836	Steady	19
examide	57855	0	0.00	1	0.00	NA	No	57855	NA	NA
citoglipton	57855	0	0.00	1	0.00	NA	No	57855	NA	NA
insulin	57855	0	0.00	4	0.01	1.47	No	26287	Steady	17871
glyburide.metformin	57855	0	0.00	4	0.01	137.73	No	57432	Steady	417
glipizide.metformin	57855	0	0.00	2	0.00	8264.00	No	57848	Steady	7
glimepiride.pioglitazone	57855	0	0.00	1	0.00	NA	No	57855	NA	NA
metformin.rosiglitazone	57855	0	0.00	2	0.00	28926.50	No	57853	Steady	2

A few of the categorical variables only have one level so they will be removed from the analysis. For race, gender, payer_code, medical_specialty and diagnosis, I'll make a category for NA. That missing age is just going to be made the median.

tried and did not give good results: age, diagnosis, gender

```
trainCombined <- bind_cols(trainNumeric,
                           readmitted = trainFactor$readmitted,
                           insulin = trainFactor$insulin,
                           discharge_disposition = trainFactor$discharge_disposition,
                           admission_source = trainFactor$admission_source,
                           admission_type = trainFactor$admission_type,
                           payer_code = trainFactor$payer_code,
                           A1Cresult=trainFactor$A1Cresult,
                           race = trainFactor$race)

fitglm <- glm(data = trainCombined,
    readmitted ~ .-patientID,
    family = binomial )
summary(fitglm)
```

```
##
## Call:
## glm(formula = readmitted ~ . - patientID, family = binomial,
##     data = trainCombined)
##
## Coefficients:
##                            Estimate Std. Error z value             Pr(>|z|)
## (Intercept)                -0.39725    0.06579   -6.04  0.00000000156349495 ***
## time_in_hospital            0.03600    0.01069    3.37              0.00076 ***
## indicator_level            -0.00140    0.00873   -0.16              0.87280
## num_lab_procedures         -0.00151    0.01048   -0.14              0.88549
## num_procedures             -0.06509    0.00989   -6.58  0.00000000004611833 ***
## num_medications             0.04539    0.01140    3.98  0.00006812596987048 ***
## number_outpatient           0.10948    0.01056   10.36 < 0.0000000000000002 ***
## number_emergency            0.21016    0.01585   13.26 < 0.0000000000000002 ***
## number_inpatient            0.48899    0.01215   40.26 < 0.0000000000000002 ***
## number_diagnoses            0.14955    0.00977   15.30 < 0.0000000000000002 ***
## insulinNo                  -0.19801    0.02933   -6.75  0.00000000001463657 ***
## insulinSteady              -0.15672    0.03012   -5.20  0.00000019543864243 ***
## insulinUp                  -0.06149    0.03617   -1.70              0.08915 .
## discharge_disposition10     1.21238    1.13609    1.07              0.28590
## discharge_disposition12     0.19304    1.42515    0.14              0.89225
## discharge_disposition13    -1.98707    0.20015   -9.93 < 0.0000000000000002 ***
## discharge_disposition14    -2.76291    0.26081  -10.59 < 0.0000000000000002 ***
## discharge_disposition15     0.83215    0.38360    2.17              0.03006 *
## discharge_disposition16     0.02735    0.68701    0.04              0.96825
## discharge_disposition17    -0.76348    0.84402   -0.90              0.36569
## discharge_disposition2      0.01844    0.05944    0.31              0.75645
## discharge_disposition22     0.26498    0.06213    4.27  0.00001996545808802 ***
## discharge_disposition23    -0.46771    0.13147   -3.56              0.00037 ***
## discharge_disposition24    -0.64574    0.42901   -1.51              0.13227
## discharge_disposition25    -0.36599    0.09423   -3.88              0.00010 ***
## discharge_disposition27   -11.22982  112.82459   -0.10              0.92071
## discharge_disposition28     0.34064    0.23610    1.44              0.14908
```

```
## discharge_disposition3   -0.05779    0.02756   -2.10              0.03602 *
## discharge_disposition4    0.05189    0.09692    0.54              0.59242
## discharge_disposition5    0.18039    0.07865    2.29              0.02181 *
## discharge_disposition6    0.20479    0.02741    7.47  0.00000000000008000 ***
## discharge_disposition7    0.05715    0.10714    0.53              0.59375
## discharge_disposition8    0.57520    0.29469    1.95              0.05095 .
## discharge_disposition9   -0.97366    0.67031   -1.45              0.14635
## admission_source10       -0.54989    1.18432   -0.46              0.64242
## admission_source11      -10.76829  196.96769   -0.05              0.95640
## admission_source13      -11.01835  196.96768   -0.06              0.95539
## admission_source14      -10.81118  196.96768   -0.05              0.95623
## admission_source17       -0.28220    0.05190   -5.44  0.00000005417649934 ***
## admission_source2        -0.30145    0.09148   -3.30              0.00098 ***
## admission_source20        0.91105    0.22359    4.07  0.00004607170312568 ***
## admission_source22       -0.08929    0.79927   -0.11              0.91105
## admission_source25      -11.05952  139.07558   -0.08              0.93662
## admission_source3         0.12130    0.27291    0.44              0.65670
## admission_source4        -0.48228    0.05879   -8.20  0.00000000000000023 ***
## admission_source5        -0.14162    0.10653   -1.33              0.18371
## admission_source6        -0.48536    0.07322   -6.63  0.00000000003380063 ***
## admission_source7         0.19703    0.03238    6.08  0.00000000117033128 ***
## admission_source8        -0.10175    0.66707   -0.15              0.87877
## admission_source9        -0.72334    0.32742   -2.21              0.02716 *
## admission_type2           0.12184    0.03050    3.99  0.00006483348165841 ***
## admission_type3          -0.05086    0.03762   -1.35              0.17636
## admission_type4           1.20643    0.92968    1.30              0.19440
## admission_type5           0.20107    0.05524    3.64              0.00027 ***
## admission_type6           0.62067    0.05508   11.27 < 0.0000000000000002 ***
## admission_type7         -11.05071   69.10779   -0.16              0.87296
## admission_type8          -0.36713    0.16036   -2.29              0.02206 *
## payer_codemedicare        0.18798    0.02416    7.78  0.00000000000000730 ***
## payer_codeselfpay         0.22815    0.04251    5.37  0.00000007991814401 ***
## payer_codeNA              0.22157    0.02473    8.96 < 0.0000000000000002 ***
## A1Cresult>8               0.05925    0.05487    1.08              0.28020
## A1CresultNone             0.13013    0.04660    2.79              0.00523 **
## A1CresultNorm            -0.09066    0.05971   -1.52              0.12892
## raceAsian                -0.28146    0.11037   -2.55              0.01077 *
## raceCaucasian             0.09045    0.02340    3.87              0.00011 ***
## raceHispanic             -0.20813    0.06845   -3.04              0.00236 **
## raceOther                -0.23054    0.07588   -3.04              0.00238 **
## raceNA                   -0.24802    0.06563   -3.78              0.00016 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80003  on 57854  degrees of freedom
## Residual deviance: 74311  on 57787  degrees of freedom
## AIC: 74447
##
## Number of Fisher Scoring iterations: 10
```

Looking at these results, we see lots of interesting characteristics.

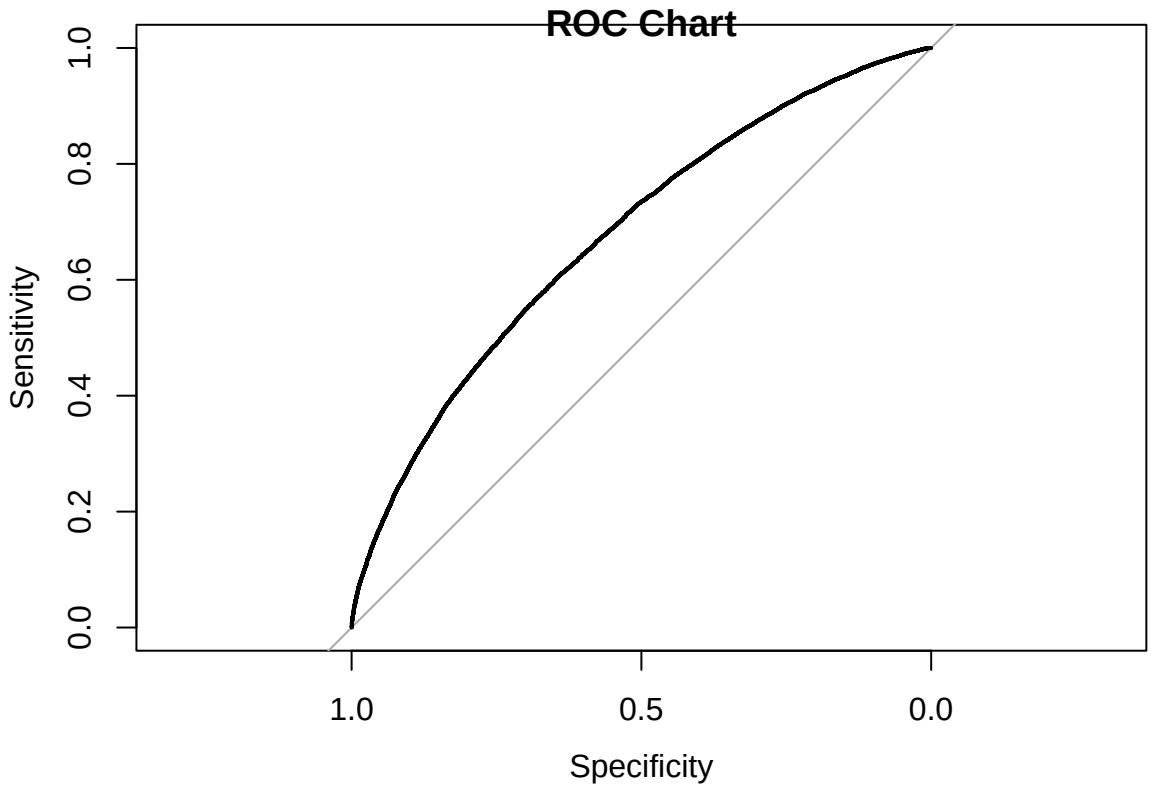
1. `payer_code`: While payer code appears significant, the added values to the probability of readmission

are very similar. However we should be aware of the levels, insurance, medicare, selfpay, NA. But the one held out, insurance, is the one we need to pay attention to. We see here that if the patient is insured, they will be much less likely, (on the order of ~20%) readmitted.

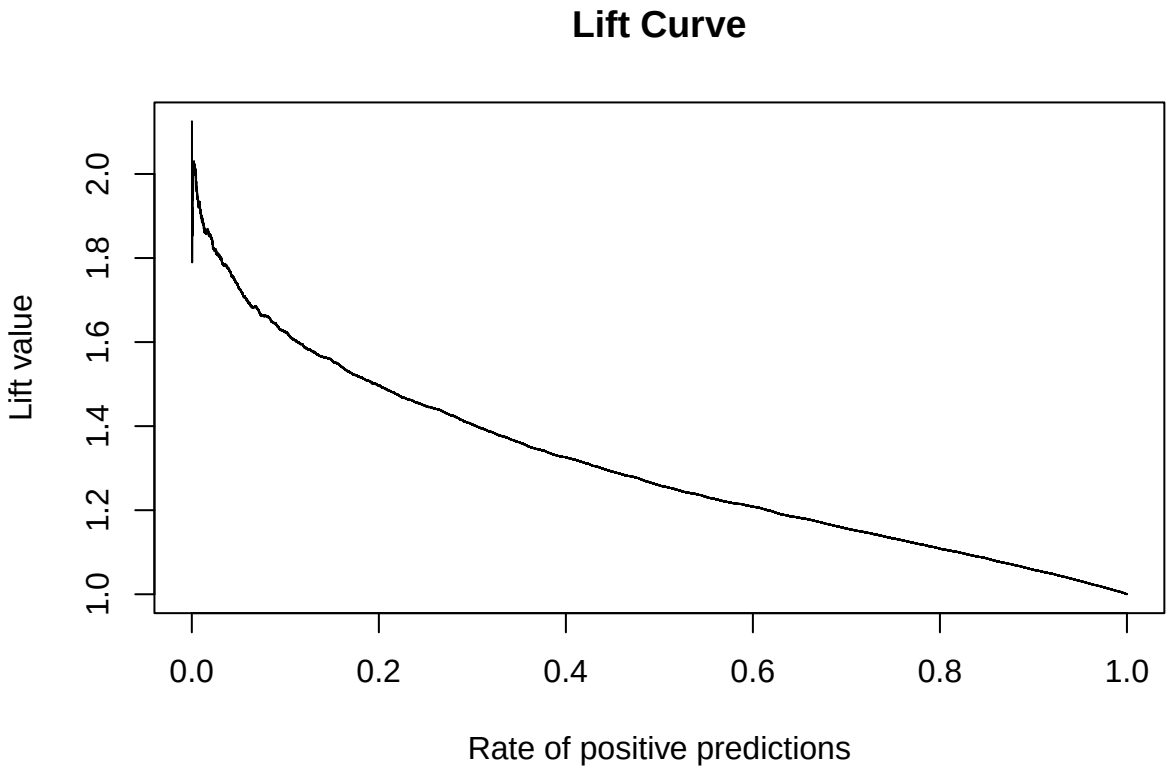
2. `race`: Race is significant at all levels. Again, we look at the levels, AfricanAmerican, Asian, Caucasian, Hispanic, Other, NA. We see here that African American is used as the baseline and all other races give a negative effect to readmission while being Caucasian give a slight positive effect. This is curious and should be explored further as one does not expect skin color to change healthcare.

3. `number_inpatient`: Lastly, we see that the number of previous trips to the hospital give a strong indication of readmission. Of course this is expected as the patient is clearly sicker and has required multiple previous visits to the hospital.

```
pred <- predict(fitglm, type = "response", newdata = trainCombined)
roc.curve<- roc(trainCombined$readmitted,pred, ci = T)
plot(roc.curve)
title("ROC Chart")
```




```
rocr_pred_test <- prediction(pred,trainCombined$readmitted)
perf <- performance(rocr_pred_test, "lift","rpp")
plot(perf, main = "Lift Curve")
```




```
threshold <- 0.5
confusionMatrix(factor(pred >threshold),factor(trainCombined$readmitted==1),positive = "TRUE")
```

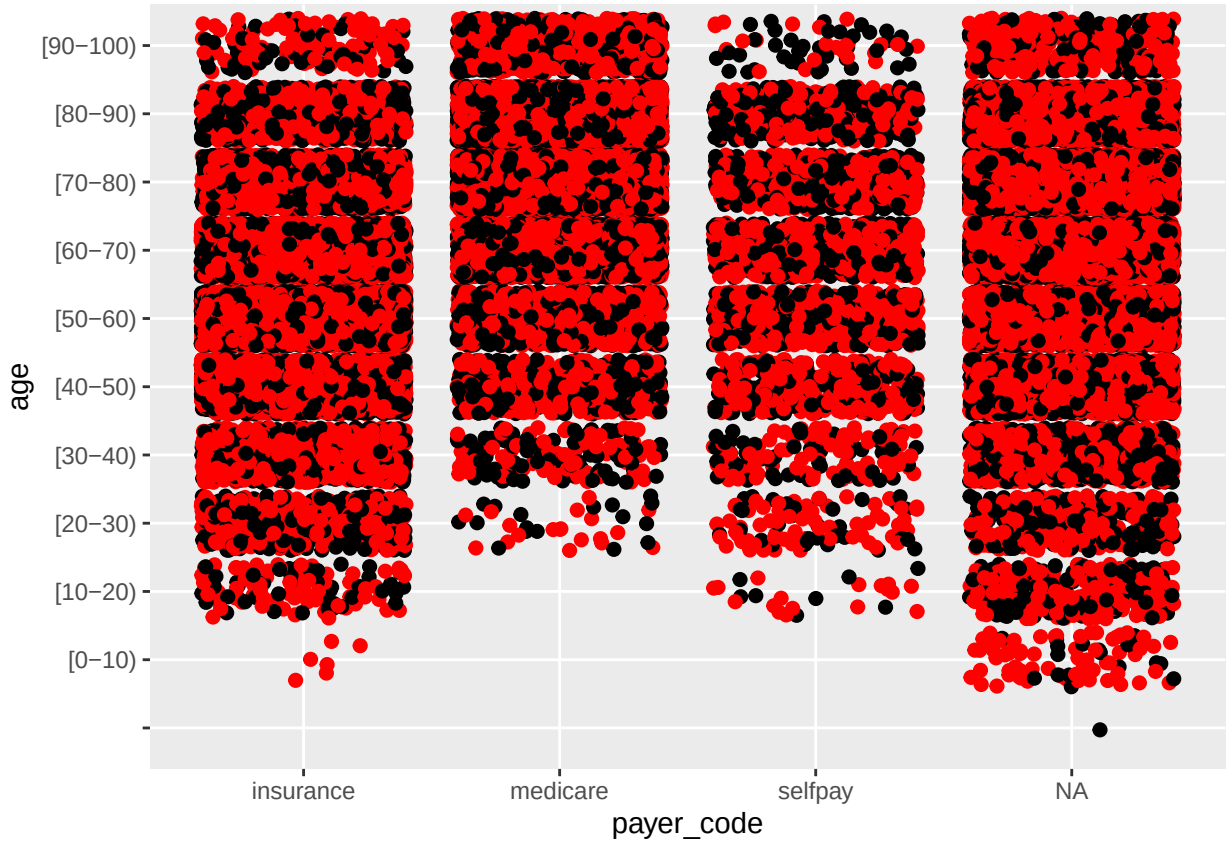
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction FALSE  TRUE
##      FALSE 23487 14391
##      TRUE   7145 12832
##
##                Accuracy : 0.628
##                  95% CI : (0.624, 0.632)
##     No Information Rate : 0.529
##     P-Value [Acc > NIR] : <0.0000000000000002
##
##                   Kappa : 0.242
##
##  Mcnemar's Test P-Value : <0.0000000000000002
##
##             Sensitivity : 0.471
##             Specificity : 0.767
##          Pos Pred Value : 0.642
##          Neg Pred Value : 0.620
##              Prevalence : 0.471
##          Detection Rate : 0.222
##    Detection Prevalence : 0.345
```

```
##       Balanced Accuracy : 0.619
##
##        'Positive' Class : TRUE
##
```

Here we have included our three metrics of the logistic regression. We have created the ROC chart, it shows us that we can identify the patients more likely of a readmitted above the random guess. We see the percentage that we have correctly captured. The lift chart shows us the ratio of the cumulative percent captured by using the model to the percentage that would happen with just random chance. Lastly we include the confusion matrix and all it's outputs. We see are values of correct identifiers and the false positives and false negatives.

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |-------------------------|
##
##
## Total Observations in Table:  36340
##
##
##                            | trainCombined$payer_code
## trainCombined$readmitted | insurance |  medicare |   selfpay | Row Total |
## -------------------------|-----------|-----------|-----------|-----------|
##                        0 |      8121 |      9581 |      1594 |     19296 |
## -------------------------|-----------|-----------|-----------|-----------|
##                        1 |      6181 |      9404 |      1459 |     17044 |
## -------------------------|-----------|-----------|-----------|-----------|
##             Column Total |     14302 |     18985 |      3053 |     36340 |
## -------------------------|-----------|-----------|-----------|-----------|
##
##
```

```
ggplot(data = train, aes(x = payer_code, y = age, colour = as.factor(readmitted)))+
  geom_jitter(size=2) +
   scale_color_manual(values = c("red", "black")) +
  theme(legend.position = "none")
```



Summarize the Models

Model	Method	Package	Hyperparameter	Selection	Accuracy	Kappa
logreg	glm	stats	NA	NA	0.627	0.239
ridge (logreg)	glmnet	glmnet	lambda	0.012	0.626	0.237
lasso	glmnet	glmnet	lambda	0	0.626	0.239
decision tree	rpart	rpart	cp	0	0.621	0.232
nearest neighbors	knn	class	k	15	0.579	0.15