# 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 |

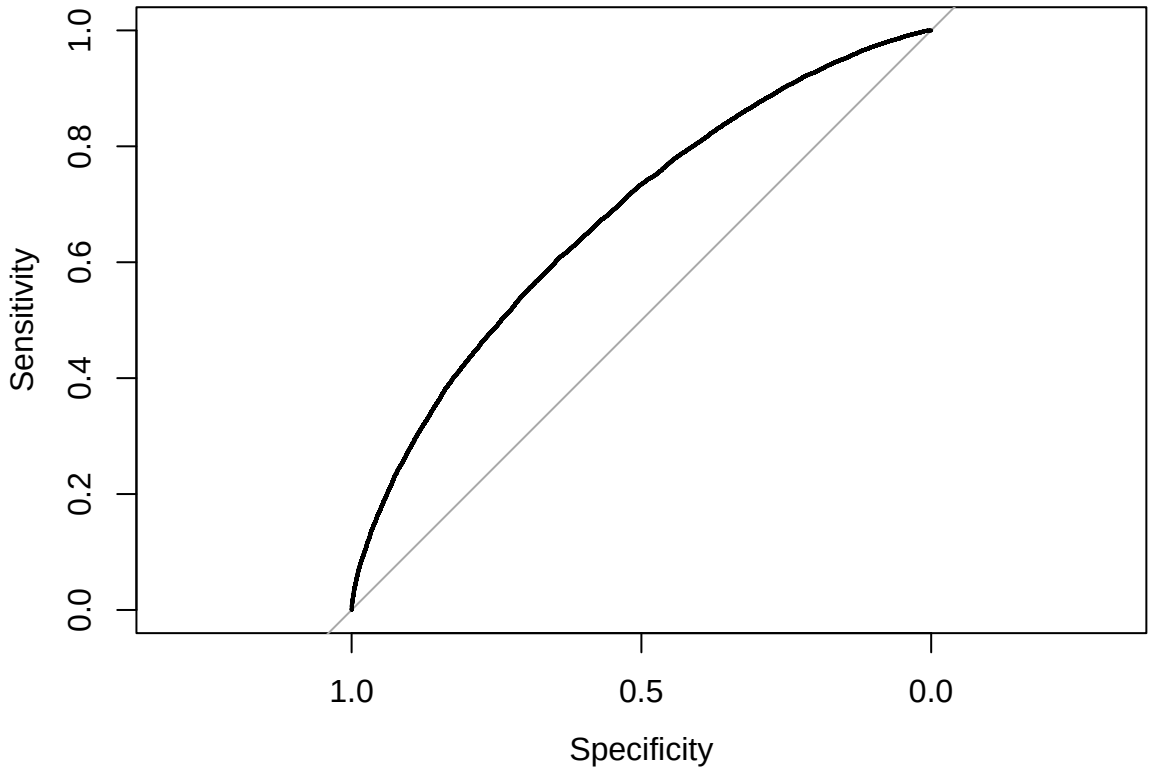
Table 2: Descriptive Summary of Categorical Variables

| 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

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



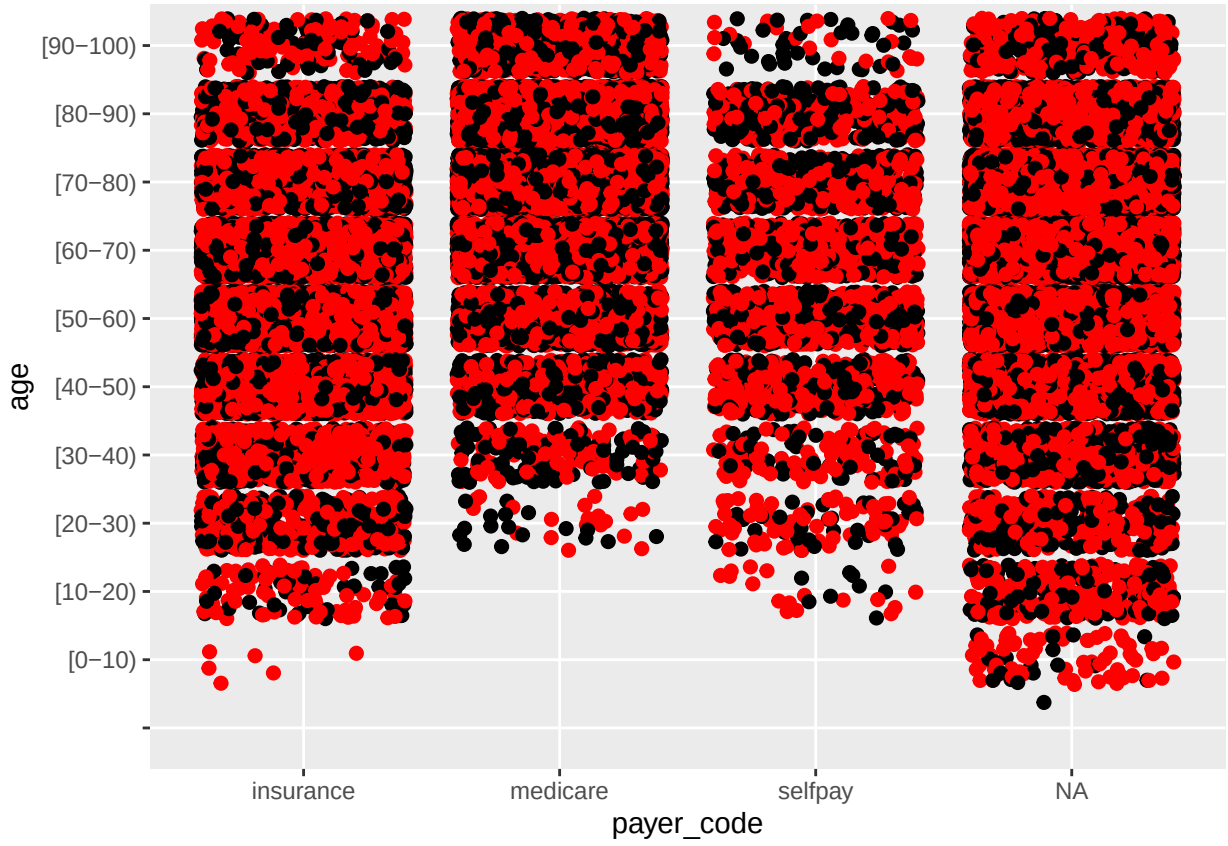

```
threshold <- 0.5
confusionMatrix(factor(pred >threshold),factor(trainCombined$readmitted==1),positive = "TRUE")
```

```
## 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
##
```

```
##
##
##    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 | 0.626 | 0.239 |
| ridge (logreg) | glmnet | glmnet | lambda | 0.012 | 0.627 | 0.24 |
| lasso | glmnet | glmnet | lambda | 0 | 0.626 | 0.238 |
| decision tree | rpart | rpart | cp | 0 | | |
| 0.622 | 0.232 | | | | | |