Sample Project – Hospital Readmissions Project Report Template

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

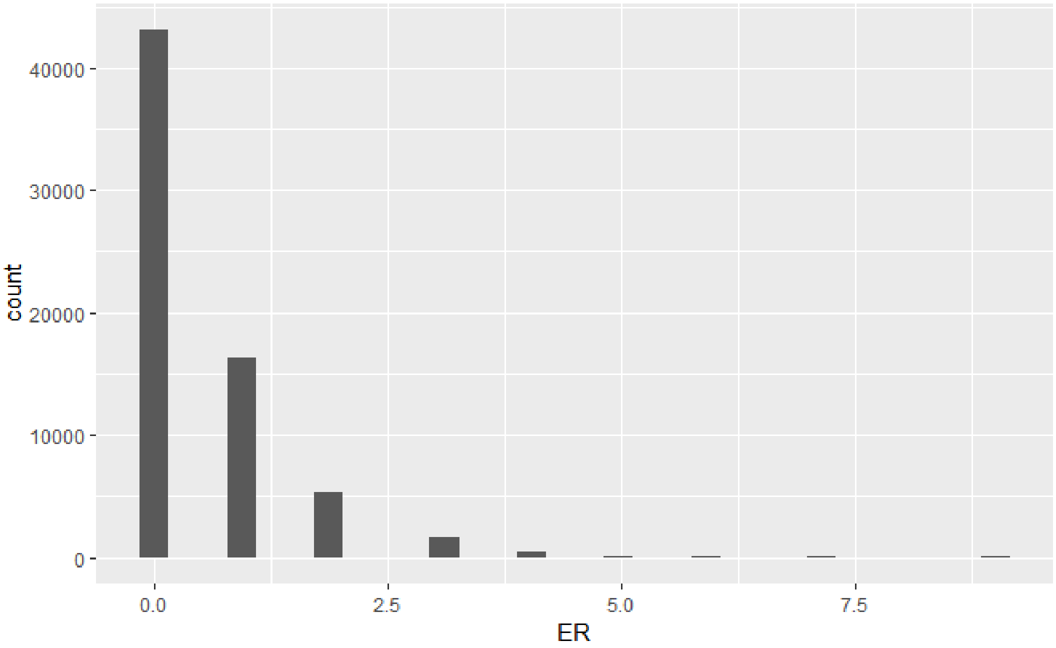
As indicated in the instructions, work on each task should be presented in the designated section for that task.

Task 1 – Perform univariate exploration of the four non-factor variables (6 points)

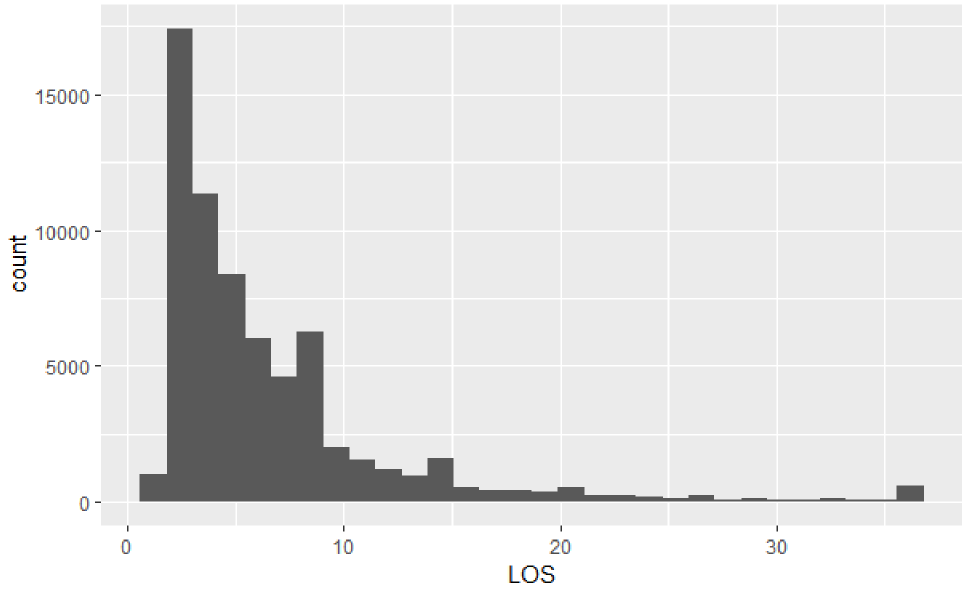
The data consists of 66,782 records and nine features. Each record has a patients gender, race, ER visits, medical class, length of stay, Age, medical cost risk score, and whether or not they have any complications or comorbidities (DRG.Complication).

The four non-factor variables are ER, LOS, Age, and Risk Score.

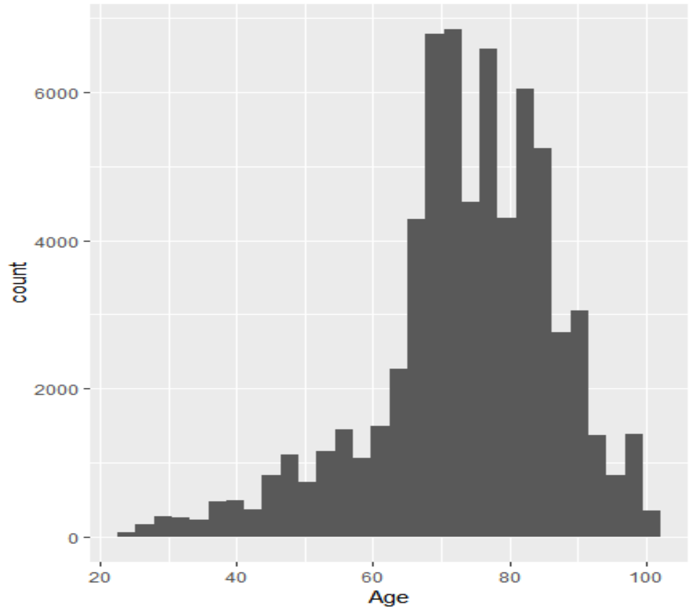
The ER variable has a 43,086 zeros as most patients have not been to the emergency room. This is a counting variable and is right-skewed. Normally I would apply a log transform, but because this has a small range and is a counting variable I will leave it as-is.



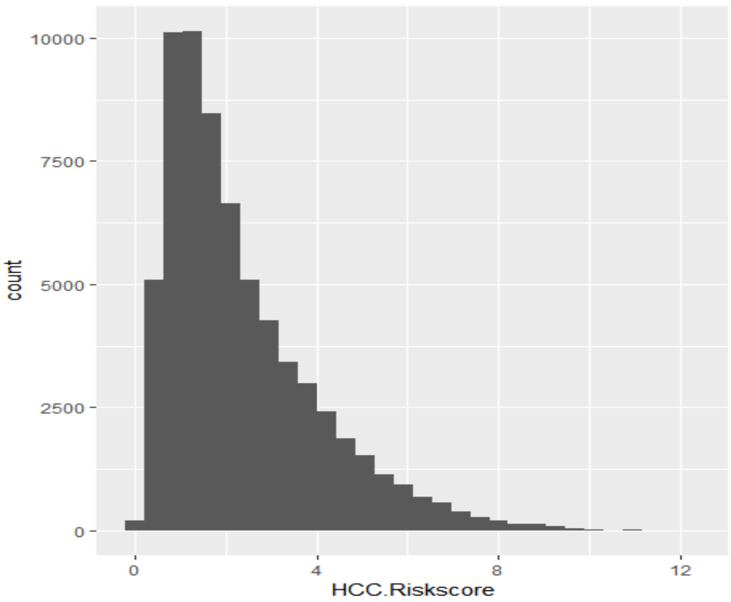
LOS (Length of Stay) is also right skewed and has a median of 6.7. There are a high percentage (622 records) where the LOS is listed as 36. These must be for patients who have stayed for *at least* 36 days in the hospital and so an indicator will be created for LOS = 36.



The Age distribution has a median of 75 as most Medicare recipients are over the age of 65. Because special circumstances are needed for eligibility of patients under the age of 65, an indicator is created to flag these patients.



The risk score has a median of 1.9 and is also right-skewed. A log transform was applied.



After transforming these variables the originals were removed.

Task 2 – Examine relationships between DRG.Class and DRG.Complication (5 points)

There is redundant info in the two variables DRG class and DRG complication. Patients who have a medical complication must be in the medical class, and patients with a surgical complication must be in the surgical class. There are 6 patients who are in the SURG class but have a medical complication. These records were taken to be an error and were removed. The two variables were then combined into a single variable.

```
  DRG.Complication   MED  SURG UNGROUP

1 MedicalMCC.CC    18104     6      NA
2 MedicalNoC       12310    NA      NA
3 Other             5357  3424     564
4 SurgMCC.CC          NA 15468      NA
5 SurgNoC             NA 11549      NA
```

After making these adjustments, the counts of number of records by each category are below. The original DRG.Complication and DRG.Class variables were removed. The column n is the number of patients.

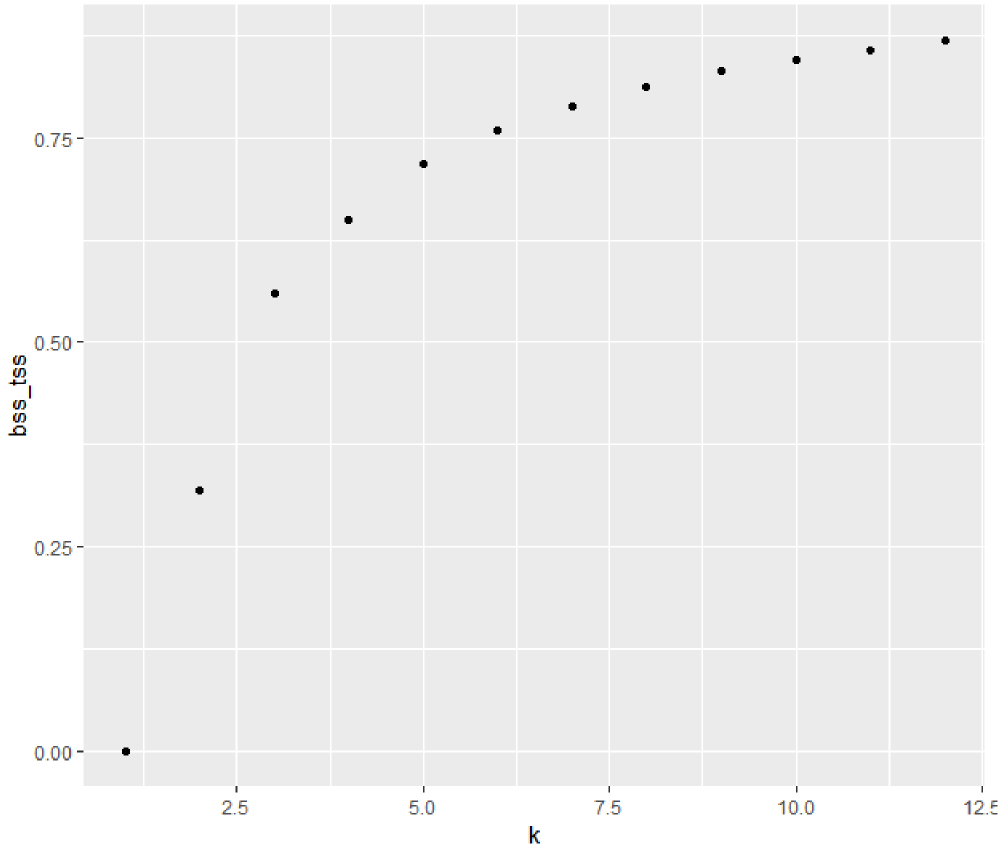
```
  DRG.Complication DRG.Class DRG       n

1 MedicalMCC.CC    MED       YMED  18104
2 MedicalNoC       MED       NMED  12310
3 Other            MED       OTHER  5357
4 Other            SURG      OTHER  3424
5 Other            UNGROUP   OTHER   564
6 SurgMCC.CC       SURG      YSURG 15468
7 SurgNoC          SURG      NSURG 11549
```

Task 3 – Use observations from cluster analysis to consider a new feature (9 points)

Kmeans is a clustering algorithm which attempts to split the observations into a pre-specified number of clusters. This first starts with a random set of cluster centers, and then iteratively adjusts these centers until the total within-cluster-variance is as small as possible. Because this chooses a random starting position, there is randomness in the cluster assignments. Having a bad starting position can cause the algorithm to find a local minimum of variance instead of a global one. Using more starting positions and then taking the average over each dimension of the cluster centers decreases the likelihood of getting stuck like this. I increased the nstarts parameter, which controls this, to 30.

This below graph shows the values of k on the x-axis and the normalized sum of squares on the y-axis. The goal is to find a set of clusters which strikes a balance between creating homogenous groups of patients that are not too small and not too large. If k = 1, then there is one cluster and all patients are in the same group. If k = n, then each patients has their own cluster. For the purposes of engineering a feature for a model, the clusters need to have enough patients that there is credible data but high enough that the risk is appropriately stratified. Setting k = 4 does a reasonable job from the graph below.



Setting k = 4 allows for at least 7,000 patients in each cluster.

```
Cluster  1      2      3      4
N    23,140 17,155 18,638  7,843
```

Task 4 –Select an interaction (5 points)

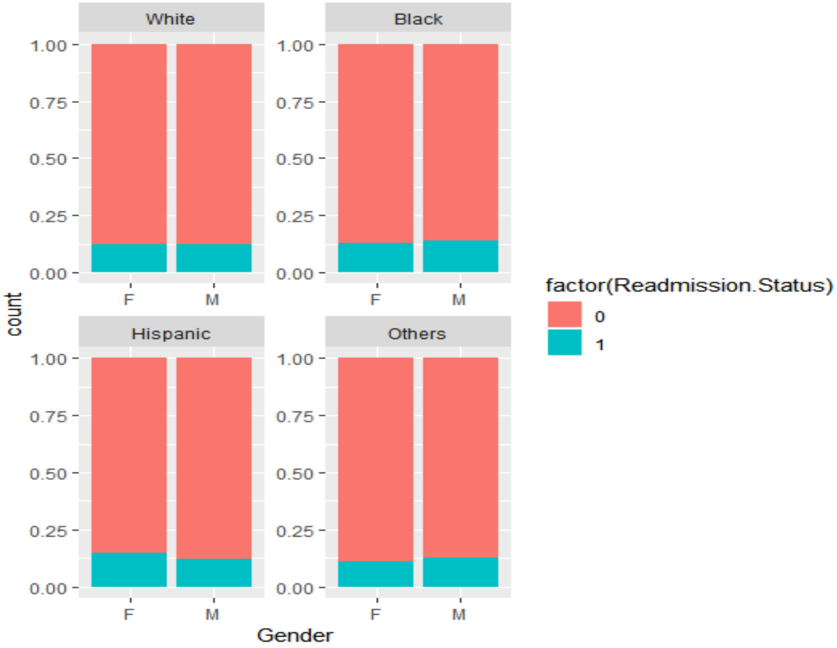
An interaction effect is where the impact of a predictor on the response changes depending on the level of another predictor. There were two interactions considered 1) Gender and Race, and 2) Risk Score and ER.

By looking at the readmission rate by race and gender, we can see that the difference between men and women is less than 0.01 for White, Hispanic, and Others but about 0.03 for Black. This indicates that an interaction is taking place.

```
Race             F     M

1 White    0.125 0.126
2 Black    0.129 0.142
3 Hispanic 0.148 0.120
4 Others   0.109 0.124
```

This same table is presented graphically below.



Because of the ethical and legal implications of using race information in a model, consideration would need to be paid to whether or not this should be included in the final model for the hospital.

The interaction between Risk Score and ER did not appear to be as significant.

Task 5 – Select a link function (8 points)

Before building models, the data was split into training and test sets with 75% serving as a training set and 25% as a test set. The readmission rate was approximately the same (12.5% and 12.8% in train and test respectively). All models were fit using only the training data and not evaluated on the test set except for model selection.

There are two key assumptions to constructing a GLM: 1) the shape of the response distribution and 2) the function which links the linear predictor to the space of the response.

Because readmissions are a binary outcome, only the binomial distribution makes sense. There are a few different link functions which are possible. These models were evaluated based on the AIC and AUC metrics. AIC is a the log of the likelihood penalized by the number of parameters in the model. A higher AIC value is better. AUC is the Area Under the Receiver-Operator Characteristic Curve and can be interpreted as the probability of the model being correct given the value of the outcome. A higher AUC value is better.

For each of these link functions the same set of variables and training data was used. The models were compared based on the test set. The cluster feature was not included. An interaction between race and gender was added.

Logit

This is the default in most software packages and is the canonical link for the binomial distribution. This transformed the linear predictor into the log of the odds, where the odds are the probability of a readmission divided by the complement of this probability.

AIC: 33,856

AUC: 0.7445

Probit

The probic link is the inverse of the standard normal CDF. The AIC of the probit was lower than that of the logit. The AUC was the same as for the logit.

AIC: 33,840

AUC: 0.7445

Cauchit

This is a complex function as it is the inverse of a standard Cauchy distribution. The AIC was better but the AUC was worse.

AIC: 34,381

AUC: 0.7445

Log

The log link would not transform the linear predictor into the range of 0,1, which mean that the model could generate predictions that do not make sense as a readmission rate. For this reason the log link was not used.

Complementary log-log

The AIC was worse and the AUC were worse (lower) for the complementary log-log model.

AIC: 33,877

AUC: 0.7445

In general, when two performance metrics disagree it can be unreliable to choose a final model based on only one metric while ignoring the other. For this reason both the cloglog and cauchit will be considered further. An exception is when the change in a performance metric is large.

Task 6 – Decide on the factor variable from Task 3 (5 points)

In task 3, a cluster feature was created from the Age and Length of Stay variables. In order to test this feature for statistical significance two models were fit including and excluding this feature. The cloglog distribution was used. When the cluster was included the original variables were removed.

The result of the model including the cluster feature is below. The AUC decreased significantly and so the cluster feature was not used.

AIC: 33,888

AUC: 0.7442

Additionally, the cauchit distribution was also tested with the above configuration and found a similar decrease in the AUC.

Task 7 – Select features (15 points)

Many of the p-values on the features are not statistically significant. Using a feature selection algorithm such as step AIC can remove features which do not improve the model. Because there are factors included such as gender and DRG, using step AIC on the model directly would lead to either removing all levels of the factor, some of which may be useful, or keeping all levels, some of which may *not* be useful. A solution is to first create dummy 0/1 variables for each of the levels of the categorical factors and then to run step AIC. This allows to separate hypothesis tests to eliminate individual factor levels.

After performing this manipulation and retraining the model the AIC and AUC when using the Cauchit link was 34,363 and 0.7353, which is are both worse (lower) than what was previously seen. This was after removing variables which were not statistically significant. The remaining variables were DRGNSURG, DRGYMED, DRGOTHER, Age, low_LOS, and log_riskscore.

This model was overfitting. I went back and changed the link to probit, reran the step AIC, and the AUC and AIC were 0.7358 and 33,825. If these are compared with the pre-variable selection metrics for the probit model (AIC 33,840 and AUC 0.7445) this difference is smaller, indicating that there is less overfitting. The AIC actually improved (increased) by 15 points and the AUC decreased only slightly. This was selected as the final model.

Note that the final AUC of 0.7445 based on the unseen test data is an improvement over the LACE model of 0.70.

After retraining on the entire data set the results are below.

```
Call:
glm(formula = Readmission.Status ~ DRGNSURG + DRGYMED + DRGOTHER +
    Age + log_LOS + log_riskscore, family = binomial(link = "probit"),
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2416  -0.5738  -0.3967  -0.2448   3.9563

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.5037421  0.0338578 -44.413  < 2e-16 ***
DRGNSURG       0.0279032  0.0154138   1.810 0.070254 .
DRGYMED        0.0048279  0.0133707   0.361 0.718037
DRGOTHER       0.0555313  0.0163595   3.394 0.000688 ***
Age           -0.0035952  0.0004016  -8.952  < 2e-16 ***
log_LOS        0.0284720  0.0079189   3.595 0.000324 ***
log_riskscore  0.6925409  0.0087475  79.170  < 2e-16 ***
---
```

Task 8 – Interpret the model (6 points)

The objective is to predict readmissions and so interpretation is not the main priority. Interpreting a glm with a probit link function is not as straight-forward as reading off the coefficients. One means of doing so is by considering the average patient and observing how the predicted readmission rate changes as the inputs vary.

```
===  =============  =======  =====
Age  log_riskscore  log_LOS  y_hat
===  =============  =======  =====
 75          0.625    1.609  0.098
 20          0.625    1.609  0.136
 75          2.079    1.609  0.387
 75          0.625    2.996  0.105
===  =============  =======  =====
```

This factors impact whether or not a patient is readmitted. The first record is the average patient. Each numeric variable was summarized with the median. Each categorical variable was summarized with the mode. This shows the following:

- Being younger (increases rate)
- Having a higher risk score (increases rate)

- Having longer hospital stays (increases rate)

Task 9 – Set the cutoff (9 points)

Task 10 – Consider alternative models and model construction techniques (12 points)

Task 11 – Executive summary (20 points)