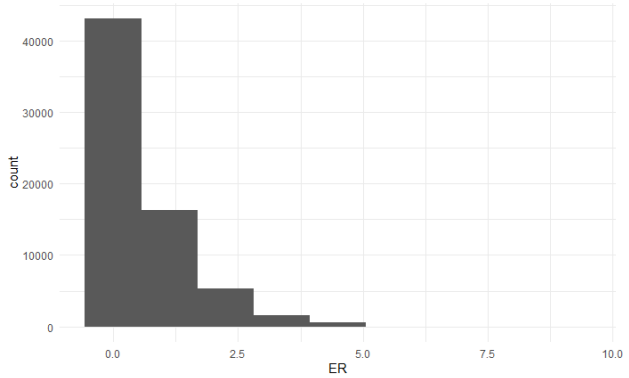
Practice Exam – Hospital Readmissions Solution

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

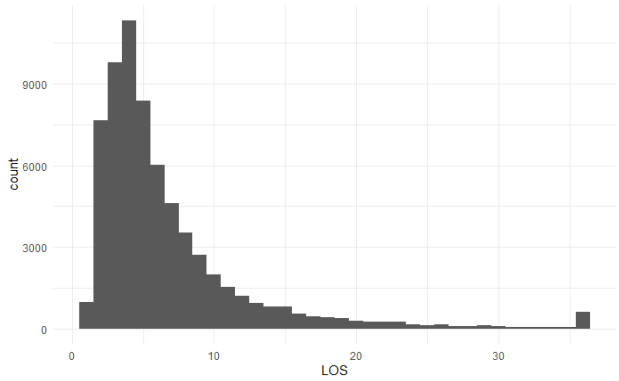
The data consists of readmission status of 66,782 patients along with their demographic and health information. There are 8 predictor variables, including both numeric and factor types.

ER



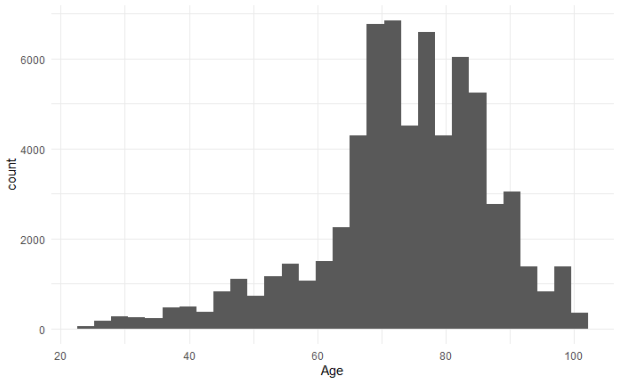
ER is defined as the number of emergency room visits. Although this distribution is positive and right skewed, a transform is not applied because this is a counting variable and only has 6 possible values. The mean is 0.5.

LOS



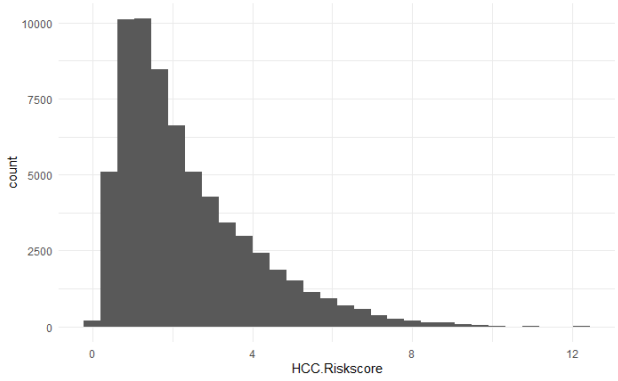
Length of stay is positive, with possible values from 0 to 30. The mean is 6.7 and the median is 5, indicating a right skew. A log transform has been applied.

Age



The age distribution shows that most subjects are older than 65, with a mean age of 73. As the mean is close to the median (75), no log transform has been applied.

HCC.Riskscore



The risk score is also right skewed and positive. The mean of 2.3 and the median of 1.8 indicate skewness, and a log transform was applied to correct the issue.

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

Six patients had a *medical* DRG complication but a *surgical* class. As these records do not fit the data dictionary's definition, data from these 6 patients were removed.

DRG.Complication	MED	SURG	UNGROUP
MedicalMCC.CC	18,104	6	*NA*
SurgMCC.CC	*NA*	15,468	*NA*
MedicalNoC	12,310	*NA*	*NA*
SurgNoC	*NA*	11,549	*NA*
Other	5,357	3,424	564

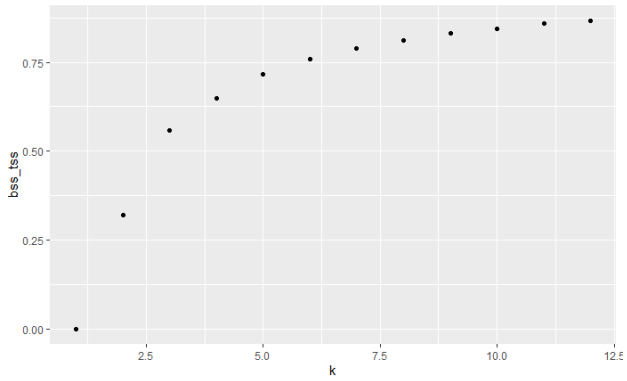
There was a lot of overlap between the DRG class and DRG complication columns. They were therefore combined and placed into either surgical, medical, or "OTHER" values. The resulting record counts are shown below.

DRGClass	Number of Patients
MED	30,414
OTHER	9,345
SURG	27,017

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

K-means is an algorithm which groups patients into clusters sharing similar characteristics. We choose group composition by evaluating the within-cluster deviance across the variables Log (LOS) and Age. K-means is iterative, and initially relies on a random seed. This introduces some randomness into the results because the algorithm can encounter processing difficulties when attempting to identify a local minimum instead of the global minimum in the deviance. The n.start parameter allows the algorithm to be run multiple times; the average cluster centers are used.

The percent of variance which is explained for each of the values of k, the number of clusters, is shown below. Ideally there would be an "elbow", allowing selection of the K in the corner. However, in this case the curve is gradual and so a value of 4 is chosen.



Task 4 –Select an interaction (5 points)

Interaction effects occur when the impact that one variable has depends on the value of another variable.

ER and HCC Riskscore

Being readmitted could be for different reasons due to a person's prior ER visits as well as their health status, measured by HCC.Riskscore. One common reason for frequent ER visits is a behavioral health condition which causes a feeling of dependence. These patients could be physically healthy, and have a

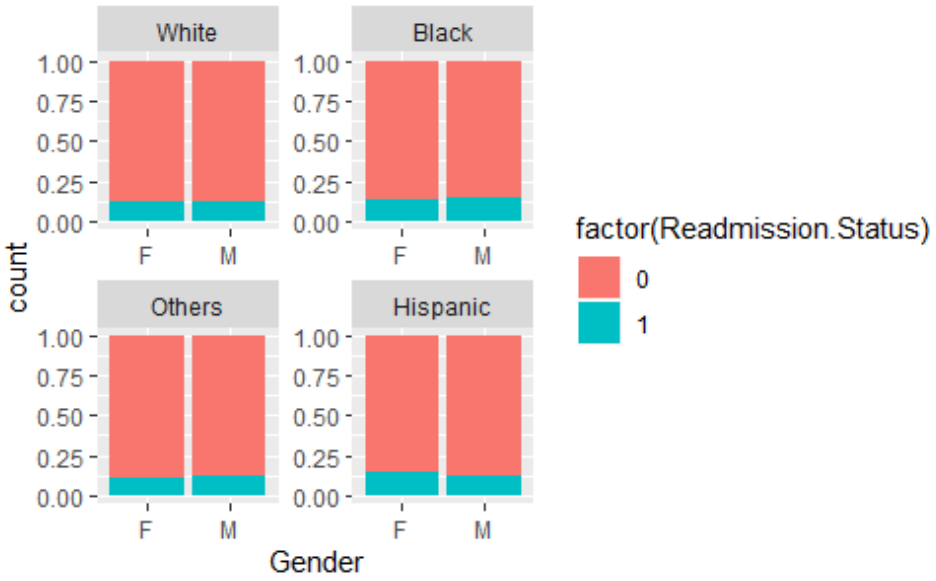
low risk score, and have a different likelihood of readmission from patients who do not have a history of ER visits. No evidence was found to support this hypothesis, however.

Gender and Race

The race and gender variables show some interaction with each other. The difference between the readmission percentages of female vs. male patients is about 0.01 in the White, Black and Other categories, but the gap widens to 0.03 in Hispanic patients. This indicates that an interaction is present.

Here is the percent of readmitted patients by race and gender.

Race	F	M
White	0.12	0.13
Black	0.13	0.14
Others	0.11	0.12
Hispanic	0.15	0.12



Task 5 – Select a link function (8 points)

I split the data into training and test sets, in which 75% of patients were used for training the model and the other 25% were used for testing. As shown below, the readmission rate was about 13% in both data sets.

Readmission Rates

- TRAINING: 0.125
- TESTING: 0.128

A GLM allows for multiple inputs to be mapped to predict a single output or 'response'. In this case, the response is binary and therefore a binomial response distribution is the optimal choice. The model predictors are related to predicted readmissions through a link function. Because the result should be a probability between 0 and 1, choices of link functions are limited to those with a domain of (0,1).

As log link allows for outputs outside of 0 and 1, it was immediately ruled out.

The model was tried on the test set for the other links and its performance, based on the AUC score, was compared. The goal of this analysis is to achieve results better than the LACE index (0.7).

Logit

Logit is the most common link function for binary classification and is the canonical link for the binomial family. It is also the default in R and is commonly used in logistic regression.

Area under the curve: 0.7451

Probit

Probit is the inverse cumulative distribution of a standard normal random variable. The AUC the same as the logit.

Area under the curve: 0.7451

Cauchit

Cauchit has a worse (lower) AUC than the Logit or Probit and therefore was not considered further.

Area under the curve: 0.7449

ClogLog

Area under the curve: 0.745

Based on the above results, I chose to use the Logit link.

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

Using Log(LOS) and Age

The Variables for Log Length of Stay and Age are both highly significant. The AUC is 0.7451.

```
Call:
glm(formula = Readmission.Status ~ . + Gender * Race - los_age_clust,
```

```
    family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3200  -0.5625  -0.3925  -0.2592   3.5082

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.735001   0.093088 -29.381  < 2e-16 ***
GenderM               0.003635   0.031228   0.116  0.90732
RaceBlack            -0.013293   0.060167  -0.221  0.82514
RaceOthers           -0.107187   0.111979  -0.957  0.33846
RaceHispanic          0.074722   0.132926   0.562  0.57403
ER                    0.004250   0.017273   0.246  0.80565
Age                  -0.005727   0.001062  -5.391    7e-08 ***
log_LOS               0.052907   0.020472   2.584  0.00976 **
log_riskscore         1.300394   0.023607  55.085  < 2e-16 ***
DRGClassOTHER         0.117817   0.042141   2.796  0.00518 **
DRGClassSURG          0.017028   0.030677   0.555  0.57886
GenderM:RaceBlack     0.109082   0.090315   1.208  0.22713
GenderM:RaceOthers    0.158532   0.159044   0.997  0.31887
GenderM:RaceHispanic  0.062118   0.205592   0.302  0.76254
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37785  on 50081  degrees of freedom
Residual deviance: 33825  on 50068  degrees of freedom
AIC: 33853

Number of Fisher Scoring iterations: 5
```

Using Cluster based on Log (LOS) and Age

After removing LOS and Age variables and using a Clustered version instead, the fourth cluster does not appear to be significant. The AUC is worse at 0.447, and so the original variables are used instead of the cluster.

```
Call:
glm(formula = Readmission.Status ~ . + Gender * Race - log_LOS -
    Age, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2868  -0.5623  -0.3933  -0.2598   3.5316

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.969577   0.054325 -54.663  < 2e-16 ***
GenderM               0.008473   0.031214   0.271 0.786039
RaceBlack             0.001758   0.060085   0.029 0.976661
RaceOthers           -0.096143   0.111856  -0.860 0.390052
RaceHispanic          0.085040   0.132865   0.640 0.522141
ER                    0.004465   0.017267   0.259 0.795942
log_riskscore         1.297804   0.023429  55.392  < 2e-16 ***
DRGClassOTHER         0.120186   0.042068   2.857 0.004278 **
DRGClassSURG          0.016588   0.030672   0.541 0.588623
los_age_clust2       -0.109377   0.049624  -2.204 0.027515 *
los_age_clust3       -0.190585   0.049949  -3.816 0.000136 ***
los_age_clust4       -0.043013   0.048849  -0.881 0.378578
GenderM:RaceBlack     0.111004   0.090261   1.230 0.218771
GenderM:RaceOthers    0.152611   0.158976   0.960 0.337073
GenderM:RaceHispanic  0.067647   0.205504   0.329 0.742023
---
```

Task 7 – Select features (15 points)

Because several variables from the model proved to be insignificant, they have been removed. Removing variables individually is not reliable because the model changes as each variable is eliminated. A better approach is to use a stepwise procedure.

AIC uses a penalized log likelihood and lower results are better. It adjusts for the fit of a model depending on the number of parameters included. Models with many parameters almost always have a lower error rate and a higher log likelihood, but risk overfitting. The penalty terms in AIC helps to correct for overfitting.

One disadvantage of this methodology relates to handling of factor levels; either all are included or all are completely removed. To fix this, I first encoded the factors Race and Gender into binary indicators and then ran a Step AIC.

The AIC starts at 33,847.6 and then improves to 33,840 after applying a Step AIC. This results in variables log_LOS, DRGCLlassOTHER, Age, and log_riskscore remaining.

All of the variables are highly significant.

```
Call:
glm(formula = Readmission.Status ~ Age + log_LOS + log_riskscore +
    DRGClassOTHER + RaceBlack + RaceHispanic, family = binomial(link = "logit"),
    data = train_bin)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3187  -0.5624  -0.3926  -0.2594   3.5119

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.720301   0.088953 -30.581  < 2e-16 ***
Age           -0.005783   0.001056  -5.474 4.39e-08 ***
log_LOS        0.052715   0.020466   2.576   0.0100 *
log_riskscore  1.300703   0.023601  55.113  < 2e-16 ***
DRGClassOTHER  0.110062   0.039645   2.776   0.0055 **
RaceBlack      0.034392   0.045184   0.761   0.4466
RaceHispanic   0.100989   0.101443   0.996   0.3195
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37785  on 50081  degrees of freedom
Residual deviance: 33829  on 50075  degrees of freedom
AIC: 33843

Number of Fisher Scoring iterations: 5
```

While the AIC has increased, the AUC has decreased slightly to 0.7453. Compared with the LACE index of 0.70, this model represents an improvement in performance.

Task 8 – Interpret the model (6 points)

I retrained the model on the entire data set to ensure that the coefficients are more stable.

```
Call:
glm(formula = Readmission.Status ~ Age + log_LOS + log_riskscore +
    DRGClassOTHER + RaceBlack + RaceHispanic, family = binomial(link = "probit"),
    data = readmission_binarized)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2570  -0.5747  -0.3953  -0.2420   3.9790

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.511383   0.041502 -36.417  < 2e-16 ***
Age          -0.003620   0.000497  -7.284 3.25e-13 ***
log_LOS       0.033235   0.009729   3.416 0.000635 ***
log_riskscore 0.700327   0.010727  65.289  < 2e-16 ***
DRGClassOTHER 0.059406   0.018642   3.187 0.001439 **
RaceBlack     0.018695   0.021326   0.877 0.380676
RaceHispanic  0.048057   0.047324   1.015 0.309880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50559  on 66775  degrees of freedom
Residual deviance: 45116  on 66769  degrees of freedom
AIC: 45130

Number of Fisher Scoring iterations: 5
```

The model uses a logit link. This means that the coefficient sign matches the direction of the effect that the variable has on readmission status, for example:

- Older patients are less likely to be readmitted;
- Patients who stay in the hospital longer are more likely to be readmitted;
- Those with a higher risk score are more likely to be readmitted; and,
- Patients with a DRG Class of OTHER are at higher risk of being readmitted, as are Blacks and Hispanics.

Task 9 – Set the cutoff (9 points)

Each of the intervention scenarios results in a different cost. If no intervention is applied, then the total penalty fees would be $210,225, the same result as having a cutoff of 1. If we apply intervention in all cases, we get the same result as having a cutoff of 0, with a cost of $133,152. By using predictive analytics, with a cutoff set at 0.08, the cost is only $105,952, a savings of $28,000.

Cutoff	Cost
1	$ 210,225.00
0.4	$ 203,799.00

0.3	$ 180,869.00
0.2	$ 141,967.00
0.1	$ 108,038.00
0.09	$ 106,426.00
0.08	$ 105,952.00
0.075	$ 106,227.00
0.07	$ 106,855.00
0.05	$ 110,386.00

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

Lasso and regularized regression

Advantages

- Selects features automatically, which may lead to more consistent results.
- Can still use the Logit link function and binomial response distribution.
- Often leads to higher performance than a stepwise method.

Disadvantages

- Can be more complicated to implement because the variables need to be scaled.
- Includes all the disadvantages of using a GLM (sensitive to outliers, interaction effects need to be manually added).

Classification tree using cost-complexity pruning

Advantages

- Able to detect interaction effects (such as between race and gender, risk score).
- Able to select features.
- Able to handle non-linearities.
- Handles missing values (assistant already handled for current case).
- Handles outlying values (assistant already removed for current case).

Disadvantages

- Weak predictive power.
- Easy for trees to overfit without bagging or boosting.

Random forest

Advantages

- Easy to train.
- High predictive power.
- Handles missing values, outliers, non-linearities.

- Automatic feature selection.

Disadvantages

- Difficult to interpret (irrelevant in current case as hospital prioritizes performance).
- High complexity can be computationally demanding.

Task 11 – Executive summary (20 points)

Being readmitted to a hospital results from a health provider failure. It is expensive for the hospital and unpleasant for the patient. We can use data and predictive analytics to help health care providers offer better care for patients before they are discharged, saving time and money for both patients and hospitals. The CMS hospital reduction program provides economic incentives to help achieve this goal, and we can increase profitability by reducing the penalties imposed by excess readmissions. For larger hospitals, which have higher total revenue, these savings measures can make a significant impact.

We collected data on patients, including whether they had been readmitted. Using predictive analytics, we identified which patients were at a high risk of readmission. The current method of assessing this risk using LACE was not as effective as the model we constructed based on the statistical measure of AUC (0.744 vs. 0.70).

The method used is a Generalized Linear Model, a popular choice in the health insurance industry. It allows for multiple inputs to be considered simultaneously. The LACE index only considers a limited number of variables, including length of stay, acuity of admission, comorbidity, and emergency visits. Our model takes into account patient risk scores, gender, race, length of stay, classes of diagnoses and comorbidities.

Total cost savings were found to be about $28,000 for a single hospital with 66,782 patients (13% readmissions). This takes into consideration currently available intervention programs.

For a cost of $2 per patient, hospitals may help prevent readmissions by providing counseling and individual instructions to discharged patients. The question of how extensive such a program should be can be precisely answered using real data.

In this study, no intervention resulted in $210,225 of penalty fees. The cost of full intervention, sending every patient through this program, would be $133,552. Even though this action would avoid fees, administrative costs would need to still be paid.

Using our model, we created a third option that resulted in costs of $108,038 (a savings of $25,514). Larger hospitals might realize even greater savings.

While our priority was to reduce overall costs, we also acquired some insight into causes of readmissions. We found that readmitted patients tend to:

- Be older;
- Have longer hospital stays before discharge;

- Have higher risk scores;
- Be in non-medical and non-surgical diagnosis related groups (DRGs); and,
- Be Black or Hispanic.

We recognize that collection of racial demographic information may not be possible or may create legal issues. Therefore, we created an alternative model that excluded these attributes.

We started with data on about 67,000 patients, including their risk scores, ages, genders, races, and medical histories. We applied appropriate transformations to optimize the model.

We considered how patients' diagnoses may be related to their complicated medical histories and made adjustments to simplify these relationships. The data of 6 patients with inaccurate medical records were removed.

We considered how patients may have things in common and used a clustering algorithm to look for patterns between length of stay and age. Using a k-means algorithm, we ran several tests using different starting configurations. These results were not statistically significant and so were not used.

Often reasons for readmission are dependent on age, gender, racial demographic, and/or risk score. Using statistical analysis, we examined possible relationships between risk scores, number of prior ER visits, gender, and race and found an interaction between gender and race.

The choice of GLM was the result of tests using five different models. We looked at different combinations of models and chose the best one based on a blind holdout test. We then used an automatic algorithm known as Stepwise selection to remove unnecessary inputs. This removed variables that were not statistically significant and improved the model's penalized log-likelihood (AIC).

Once we had a strong statistical model, we then fine-tuned it to maximize profitability by reducing medical costs. We considered costs associated with a CMS discharge penalty as well as the administration of a patient intervention program. The values used were a $25 penalty and a $2 intervention. Depending on your business needs, we can customize these results.

To further increase your profitability, we can use a higher-performing model. For this situation, regularized regression or boosted trees (a random forest) would likely lead to additional cost savings. If you are looking for a more interpretable result, a decision tree may be a good choice.

This analysis shows that hospitals can save money and provide better care by treating patients who are at a high risk for readmission. Hospital readmissions are not random. Using historical data, we can accurately predict their occurrence and thus prevent or reduce them.

Our model used racial information, although collecting and using this data may cause legal complications. Our model recommendations were not tested for regularized regression, random forests, or decision trees. We based our results on an elderly population, with an average age of about 65 years old. Use of a younger population would require readjustment of our model.