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)

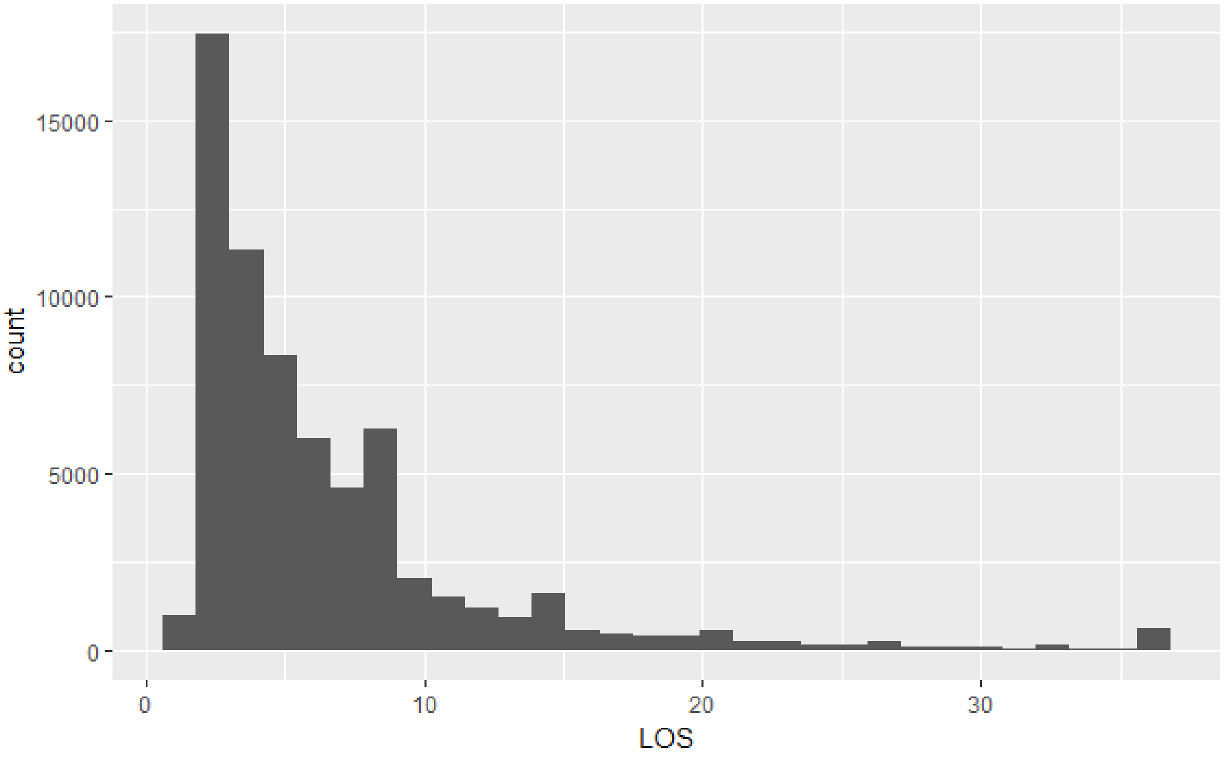
The four non-factor variables are ER, LOS, Age, and HCC.Riskscore. I will discuss each in turn.

ER is a counting variable that ranges from 0 to 9. The following table shows the counts for each value:

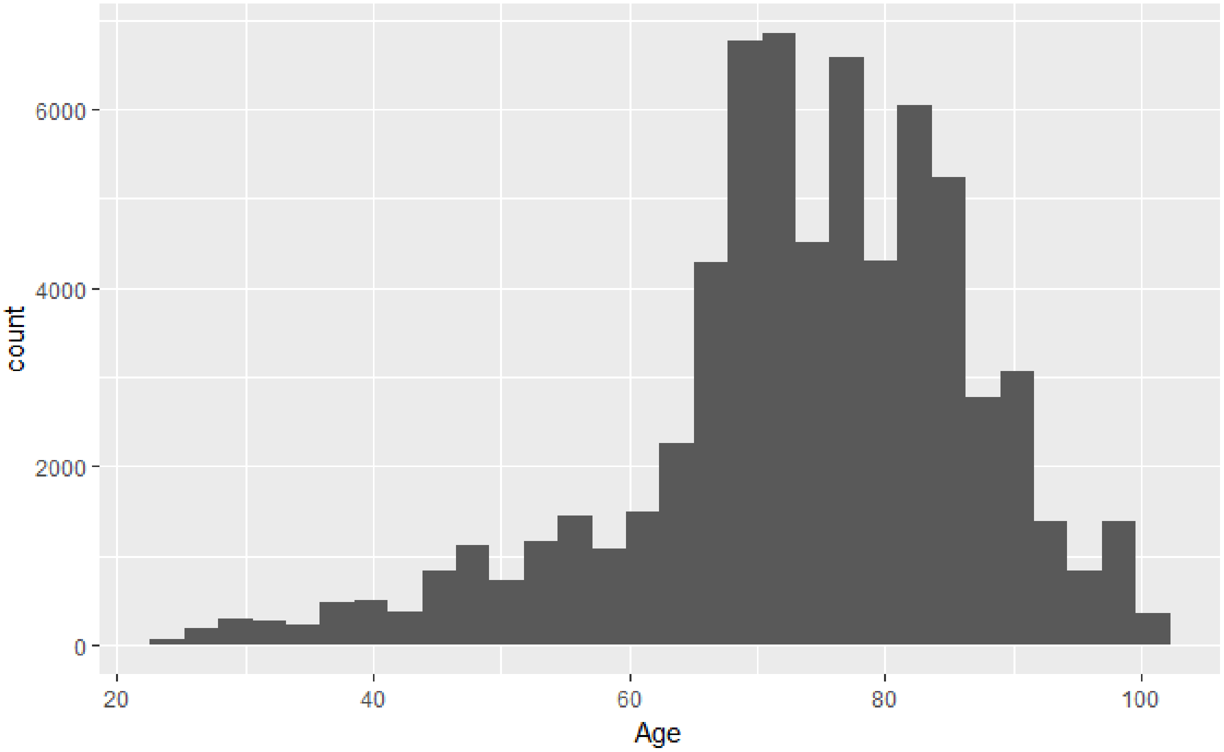
```
    0     1     2     3     4     5     6     7     9 
43086 16280  5286  1572   438   105    10     3     2 
```

It is clear that this distribution is highly skewed. However, as a count variable I would prefer to leave it as is. Also, a log transformation would be problematic with regard to the values at zero.

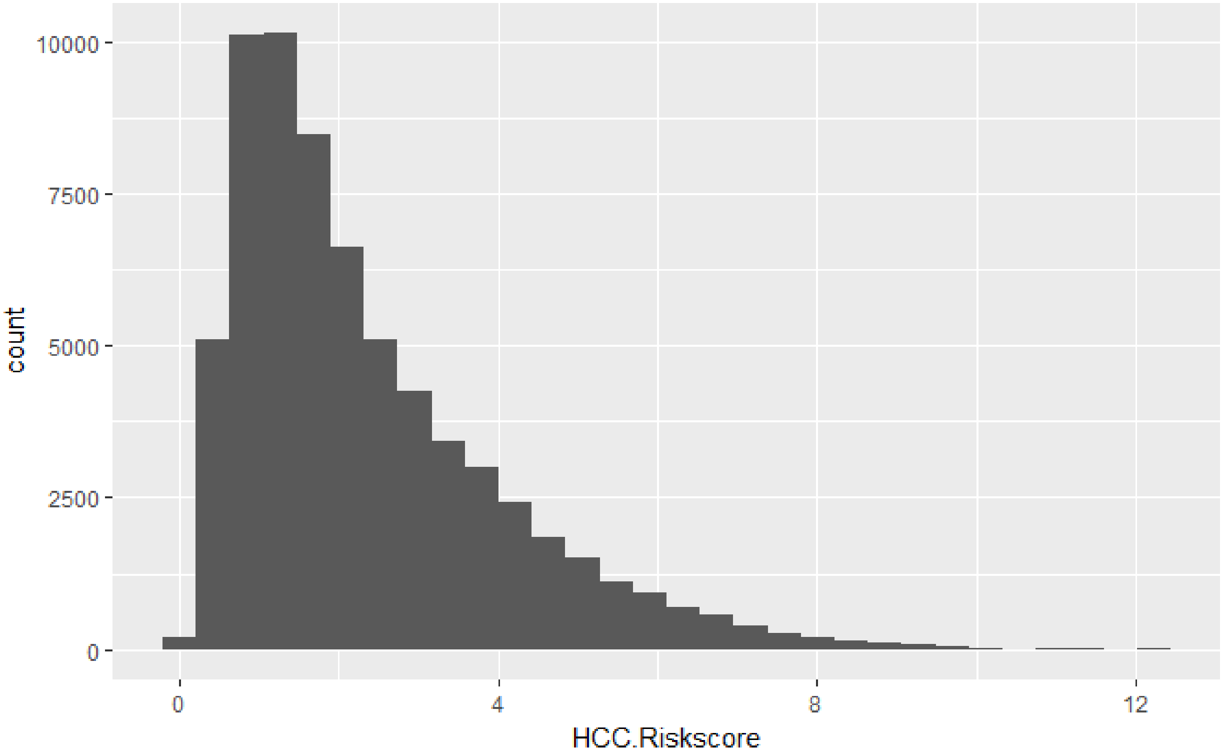
LOS is also a counting variable, but ranges from 1 to 36. The mean is larger than the median, indicating right skewness. The graph below indicates it pronounced and thus a log transformation will be employed.



Age ranges from 24 to 101. The median and mean are similar and the graph below indicates no particular pattern that would require an adjustment. Medicare is available to those under 65 only under special circumstances. I have added an indicator variable that identifies those under 65 so that their estimated readmission probability can differ.



HCC.Riskscore ranges from 0.079 to 12.206 and there may be some right skewness, which is confirmed by the plot below. A log transformation will be used.



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

The following indicates how the observations line up with these two variables:

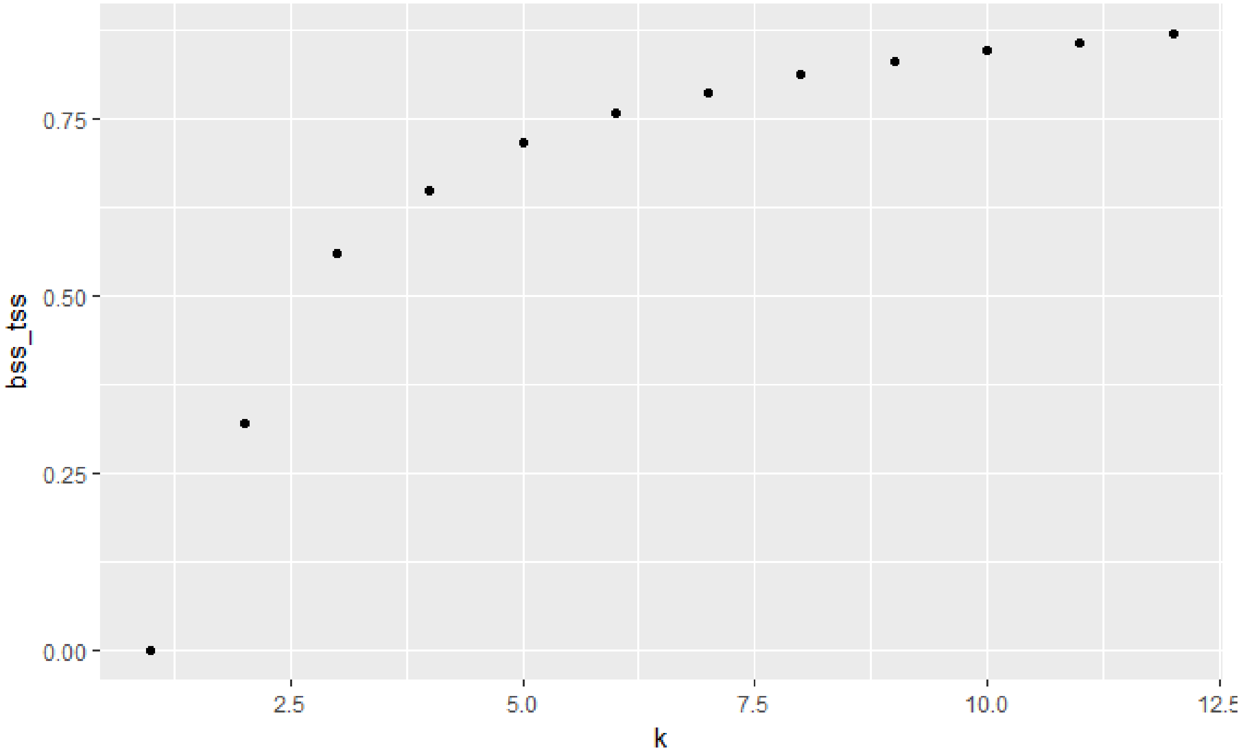
```
         MedicalMCC.CC MedicalNoC Other SurgMCC.CC SurgNoC
  MED            18104      12310  5357          0       0
  SURG               6          0  3424      15468   11549
  UNGROUP            0          0   564          0       0
```

The six cases with Class = SURG and Complication = MedicalMCC.CC seem to be an error. They will be deleted. Otherwise, there appears to be seven distinct categories. A new variable, DRG, is created that has those seven levels. (*Note to candidates: There are several ways to do this in R. An alternative is to export the current version of the dataset to Excel, perform the manipulations there, save as a csv file, and then load back into RStudio. This may create a new variable that represents the row number, which should be deleted.*)

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

A cluster analysis attempts to partition the observations into *k* subsets. It is impossible to examine every possible arrangement of the observations and thus an algorithm is employed to intelligently search for the optimal partition. However, it is likely that the search will find a local, not a global, optimum. Because the algorithm starts with a random selection of initial cluster centers, the results will vary based on that random selection. The nstart parameter in the kmeans function controls the number of different random selections that are used. This improves the chances of finding a better local optimum. A value of 20 to 50 for nstart is recommended. I elected to use 20 to save time.

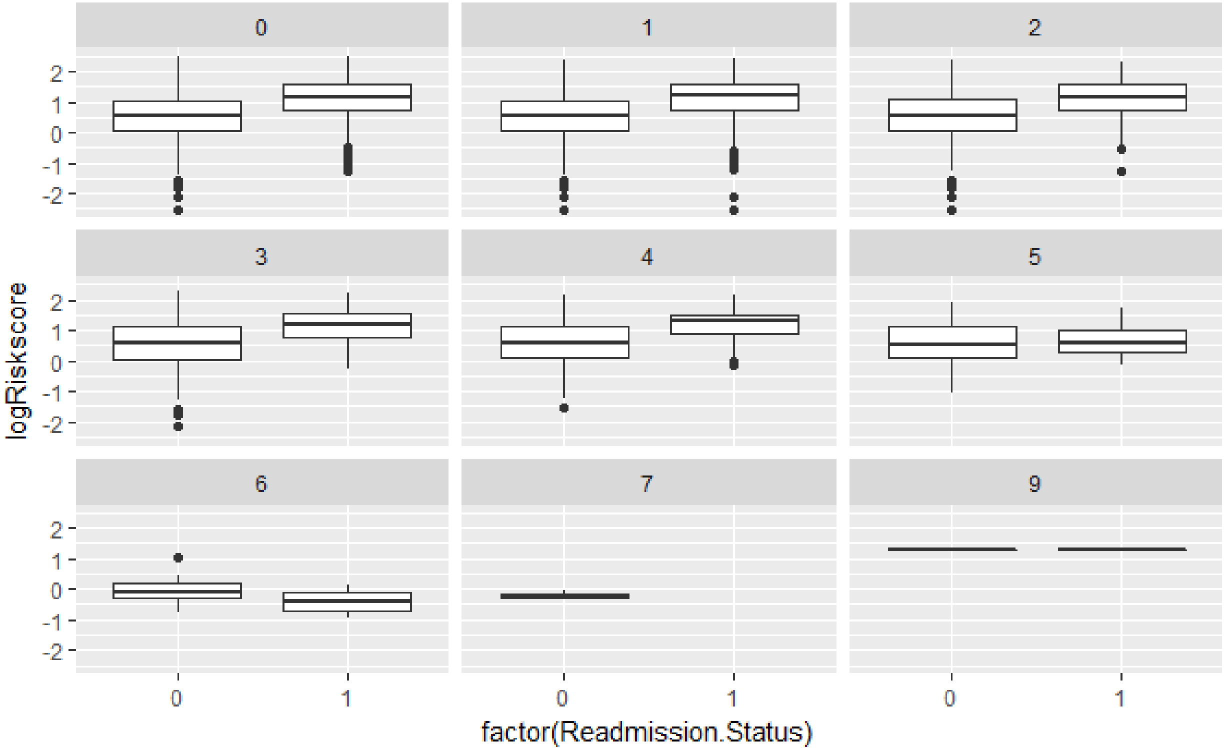
Running the cluster algorithm with from 1 to 12 clusters produced the following elbow plot:



There is no obvious “elbow” in this elbow plot, but appears the gains drop off after 5 clusters, so that is the number that will be used. (*Note to candidates: Any value from 3 to 6 could be justified here.*) A new variable is created that places each observation in one of these clusters.

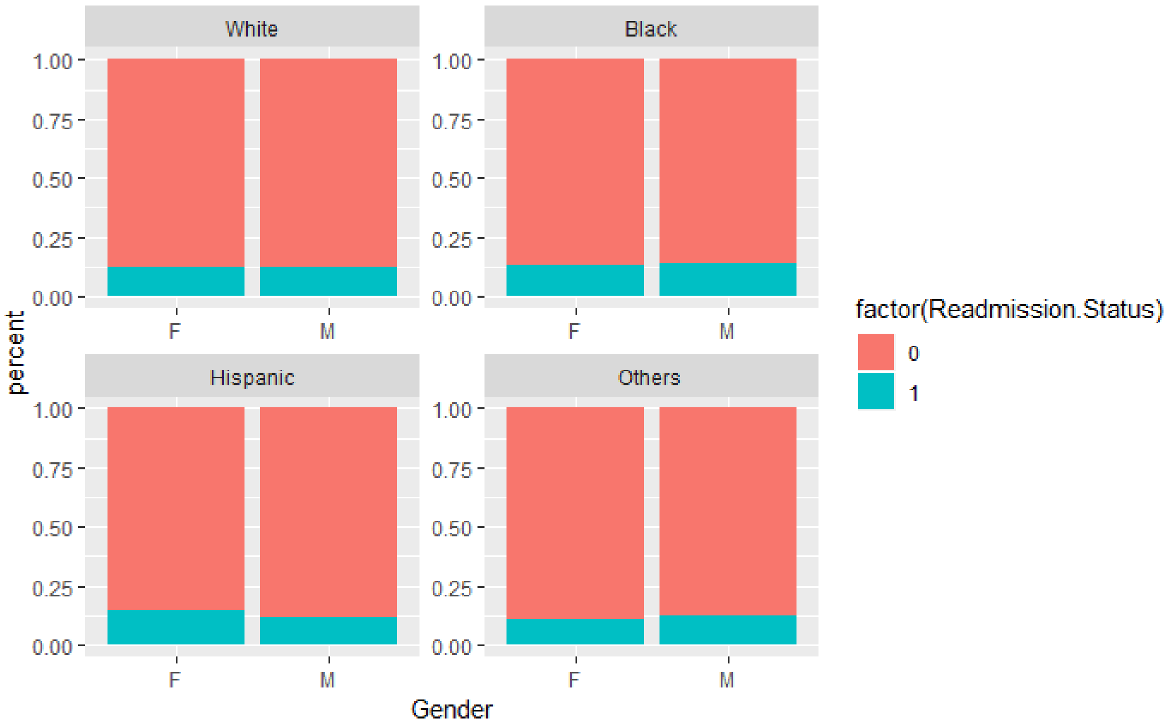
Task 4 –Select an interaction (5 points)

My first choice for a possible interaction is between ER and logRiskscore. An interaction means that the effect on the target of one variable is influenced by the level of another variable. It may be that those with more prior emergency room visits will have greater sensitivity to how logRiskscore relates to readmission. The following plot can help with that determination.



While it appears that there are some differences once the number of ER visits exceeds 4, these cases are a very small part of the sample. Hence this interaction is unlikely to be useful.

I next tried Race and Gender. I am not sufficiently well-versed in the socio-economics of hospital usage, but perhaps there are some interaction effects here. The plot below shows that black males are more likely to be readmitted than black females while the reverse is true for Hispanics. I will use this one in subsequent models.



Task 5 – Select a link function (8 points)

Prior to fitting a GLM, I split the data into a training set (75%) and a testing set (25%). To check that the two sets are representative, I note that the training set has a 12.70% readmission rate while the testing set has 12.29%, while this does not exactly match, it is close enough.

With the binomial model (the only one appropriate when the target is zero or one), there are five link functions available in the glm package: logit, probit, cauchit, log, and cloglog. I ruled out the log link as it does not ensure that responses will be in the zero to one range. From my prior studies it appears that the logit function is the one most commonly used. Also, it is the default choice in R and is the canonical link function.

Fitting the binomial/logit model using all variables plus the identified interaction produced the following results:

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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.3460  -0.5643  -0.3911  -0.2544   3.0541  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -2.603997   0.136923 -19.018  < 2e-16 ***
GenderM              -0.021739   0.031186  -0.697  0.48575    
RaceBlack             0.050490   0.059038   0.855  0.39244    
RaceHispanic          0.142477   0.129136   1.103  0.26989    
RaceOthers           -0.127669   0.113611  -1.124  0.26112    
```

```
ER                    -0.005826   0.017218  -0.338  0.73508    
Age                   -0.007438   0.001617  -4.598 4.26e-06 ***
logLOS                 0.065528   0.020390   3.214  0.00131 ** 
logRiskscore           1.330490   0.023717  56.100  < 2e-16 ***
Under65               -0.064840   0.057945  -1.119  0.26314    
DRGMed.NoC            -0.040421   0.042587  -0.949  0.34254    
DRGOtherMED            0.125416   0.054738   2.291  0.02195 *  
DRGOtherSURG           0.103201   0.065984   1.564  0.11781    
DRGSurg.C              0.007461   0.039946   0.187  0.85184    
DRGSurg.NoC           -0.010876   0.043732  -0.249  0.80360    
DRGUNGROUP             0.134489   0.139704   0.963  0.33571    
GenderM:RaceBlack      0.003736   0.090470   0.041  0.96706    
GenderM:RaceHispanic  -0.321541   0.213049  -1.509  0.13124    
GenderM:RaceOthers     0.237410   0.158452   1.498  0.13405    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38117  on 50081  degrees of freedom
Residual deviance: 33955  on 50063  degrees of freedom
AIC: 33993
```

It appears that only a few of the features are statistically significant. That will be investigated later. I also note that the AUC value is 0.7324. This already exceeds the value of 0.70 for the LACE index, so this approach has promise.

I tried the probit link function, which is the inverse cumulative normal distribution, which like logit is appropriate in that it maps to (0, 1). While the cauchit and cloglog links also map to (0,1), the connection to the normal distribution may make this one easier to explain. The log link has the advantage of ease of interpretation (the coefficients relate to percentage changes in the target probability) but runs the risk of making predictions outside the (0,1) interval. Because our goal is more about prediction than interpretation, I elected not to try this link function. The same variables are significant. The AIC value is 33,966, which is better than that for the logit model. The AUC is identical at 0.7324. As I result, I will use the probit link function.

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

In Task 3, a feature was created using cluster analysis that may be a useful replacement for logLOS and Age. The probit regression from the previous task was rerun, adding this variable and removing the original two variables. Given that Age and logLOS were significant in the original model, it is unlikely that an improvement will be seen.

The new model has an AIC of 33,990 and an AUC of 0.7321. It did not perform as well as the previous model and so the cluster variable will not be used.

Task 7 – Select features (15 points)

Many of the features were not statistically significant. It is rarely a good idea to remove multiple features at one time as the removal of one feature can lead to others becoming important. One of the easiest ways to remove features is to apply a procedure such as stepAIC. It is automated and keeps removing features until all that remain are significant by the selected measure. A drawback is that it treats factor variables as a single feature and thus either retains or removes all levels. I see from the output that one level of DRG is significant with respect to the base and I would like the opportunity to use that level for prediction.

One approach to this issue *(Note to candidates: an alternative approach is provided at the end of the Rmd file)* is to binarize the factor variables. Now each level is its own feature and stepAIC can be performed. With this approach, two factor levels that are significant with respect to the base level but have essentially the same coefficient will not be combined into a single level.

After binarization, the probit model produces the same results as before, but now I can run stepAIC on this model. The function removes all features except logLOS, Age, logRiskscore, and the OtherSURG and OtherMED levels of DRG. I next run and test the model using only these variables. The results are:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.5092975  0.0469985 -32.114  < 2e-16 ***
logLOS        0.0343635  0.0112216   3.062   0.0022 ** 
Age          -0.0037101  0.0005667  -6.547 5.88e-11 ***
logRiskscore  0.7115770  0.0124458  57.174  < 2e-16 ***
DRGOtherSURG  0.0558636  0.0341427   1.636   0.1018    
DRGOtherMED   0.0710677  0.0273260   2.601   0.0093 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38117  on 50081  degrees of freedom
Residual deviance: 33938  on 50076  degrees of freedom
AIC: 33950
```

The AIC has improved from 33,966 to 33,950. I note that one of the variables is not significant at the 10% level. It was retained because AIC is more likely to retain a variable than using hypothesis tests. (*Note to candidates: A reasonable additional step here would be to note that the coefficients of DRGOtherSURG and DRGOtherMED are similar. There may be improvement by combining them into a single factor level.*) The AUC is slightly improved at 0.7334. This will be my final model and is demonstrated, for this dataset, to be superior to the LACE index.

Task 8 – Interpret the model (6 points)

I first ran the model using the full dataset. The results are:

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -1.5026235  0.0407530 -36.871  < 2e-16 ***
logLOS        0.0329948  0.0097365   3.389 0.000702 ***
Age          -0.0036999  0.0004916  -7.526 5.25e-14 ***
logRiskscore  0.7008229  0.0107283  65.325  < 2e-16 ***
DRGOtherSURG  0.0560715  0.0293707   1.909 0.056250 .  
DRGOtherMED   0.0691546  0.0237996   2.906 0.003664 ** 
```

Coefficients in a probit model are difficult to interpret. One way to get a feel for the coefficients is to consider the median patient. This patient has LOS = 5, Age = 75, HCC.Riskscore = 1.866. And the majority are not in the two indicated DRG levels. We can see how a change in these factors affects the estimated probability of readmission as indicated in the following table (*Note to candidates: This could also be done in in Excel.*):

LOS	Age	HCC.Riskscore	DRGOtherSURG	DRGOtherMED	Probability
5	75	1.866	0	0	0.09855
6	75	1.866	0	0	0.09960
5	80	1.866	0	0	0.09538
5	75	2.053	0	0	0.11068
5	75	1.866	1	0	0.10864
5	75	1.866	0	1	0.11110

We see that an extra day in the hospital only increases the readmission probability by 0.1%. An extra 5 years of age lowers it by 0.3%. Increasing the risk score by 10% increases the probability by 1.2%, being in the OtherSURG group adds 1.0%, and being in the OtherMED group adds 1.3%.

In addition to providing predictions for use in deciding interventions, this may also inform the hospitals regarding actions they can take during hospitalization that may reduce readmissions. For example, is there anything in the procedures in the OtherSURG and OtherMED categories that could be changed to reduce readmissions?

Task 9 – Set the cutoff (9 points)

Setting the cutoff for predicting a readmission at 0.5 produces only 21 predictions of readmission. A lower cutoff may provide more savings. Any change in the cutoff leads to a tradeoff with regard to the two types of prediction error. One way to optimize the tradeoff is to associate costs with making errors.

Using the approach provided by the hospitals, if we do nothing, there is a cost of 25 associated with a readmission. In our dataset there are 8,409 readmissions for a total cost of 210,225. This is equivalent to setting the cutoff at 1 (predict that no one will be readmitted).

If all patients receive the intervention, the cost is 2*66,776 = 133,552. This is equivalent to setting the cutoff at 0 (predict that everyone will be readmitted and thus in need of an intervention).

If our model is used, the hospital would spend 2 on each patient predicted to be readmitted. Those patients will not incur additional costs. With our current cutoff of 0.5, there are 21 such patients and a cost of 42. The others will still incur a cost of 25 if readmitted. There are 8,399 of them for a cost of 209,975. The total cost of the program is 210,017. The following table presents the cost at various cutoff values.

Cutoff	Cost
1	210,225
0.4	203,795
0.3	180,658
0.2	142,093
0.1	107,934
0.09	106,238
0.08	106,002
0.075	106,100
0.07	106,866
0.05	110,423
0	133,552

Thus, at a cutoff of 0.08 there is an estimated cost of 106,002. With this cutoff, there will be 39,301 predicted readmissions, which is 59% of the cases. This seems like a lot, but we are spending 2 to save 25, so there is a reason to over predict. Of the 8,409 who were readmitted, 7,313 will incur the savings of 23.

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

There are alternatives to the approache taken here. The following are three such alternatives along with advantages and disadvantages.

LASSO Regularized Regression

Advantages:

- Through the use of the model matrix, binarization is always done and each factor level treated as a separate feature, helpful since the data included several factor variables.
- Variable selection is automatic, using cross-validation to minimize prediction error rather than a proxy such as AIC or hypothesis tests. This carries some risk, though, because optimizing prediction error may not optimize AUC or cost-weighted outcomes from the confusion matrix.

Disadvantages:

- Because the variables are scaled, the estimated coefficients are difficult to interpret. This is a minor issue here as we are more interested in prediction than interpretation.
- The R package works with a limited number of models. Hence, at least with R, there is less flexibility in model choice.

Classification Tree using Cost-Complexity Pruning

Advantages:

- Trees tend to be easy to explain and present, provided there aren't too many branches. This is a small advantage for this particular problem because we are more interested in prediction.
- Variables are removed automatically (by not showing up in the tree), allowing interpretation to focus on the most significant factors.
- It more easily adapts to non-linear relationships, especially those with natural discontinuities, such as might be suggested with age 65 and Medicare.
- It automatically captures interaction effects.

Disadvantages:

- Even with cost-complexity pruning there is relatively high danger of overfitting.
- The resulting tree can be highly dependent on the training set (high variance).
- The splits made for continuous or count variables, which are abundant in this data, can seem arbitrary and suggest large discontinuities in results where little difference may exist.

Random Forest

Advantages:

- Reduces overfitting and variance by allowing results from multiple trees to be combined.

- Uses cross-validation to set the tuning parameter.
- Inherits third and fourth advantages above for trees.

Disadvantages:

- Difficult to interpret. This is less of a disadvantage given that this problem is chiefly about prediction.
- Takes considerably longer to run and analyze compared to other methods.
- Difficult to implement. With the other methods a simple spreadsheet can be constructed from which patient predictions can be made.

Task 11 – Executive summary (20 points)

With the implementation of the Hospital Readmission Reduction Program in 2010, hospitals have a significant motivation to reduce patient readmission rates. Currently, hospitals use the LACE index, reported to produce an AUC (area under the curve, a commonly used measure of predictive ability) of 0.70. We have been asked to use patient-level data supplied by a group of hospitals to obtain a model with superior predictive power compared to the LACE index.

We cleaned the data to remove any records with missing values. Two of the variables, DGR.Class and DRG.Complication, appear to overlap, with many combinations not being possible or self-consistent. Six records with inconsistent values for these variables were removed. To remove the overlap, a single new variable with seven levels was created. While several other adjustments were considered, the only additional one was to use the logarithms of two of the variables, LOS (length of stay) and HCC.Riskscore.

The supplied data includes race and age. Given that this model might be used to decide which patients receive additional support, it is possible that these variables would be considered inappropriate. Age appears in the recommended model and this should be reviewed prior to implementation.

A generalized linear model (GLM) was selected for making the predictions. It is a flexible model that (1) ensures that for each patient, the prediction is a number between 0 and 1 representing the estimated probability that patient will be readmitted, (2) allows for controlled removal of variables that lack predictive power, and (3) indicates the direction of the influence of each variable that is retained. The GLM does assume that directional effects are monotonic, however, and does not handle non-linear relationships very well.

After investigating several forms of the GLM and variable choices, a model using the following variables provided the most accurate predictions on a random train-test partition:

- Length of hospital stay – Longer stays increase the probability of readmission
- Age – Older patients have a lower probability of readmission
- HCC.Riskscore – Patients with higher risk scores have a higher probability of readmission
- DRG.Class and DRG. Complication – Those with Complication = Other and Class = MED or SURG have a higher probability of readmission than those with other pairs of values.

The AUC for this model is 0.73, an improvement over the 0.70 of the LACE index.

We were asked to investigate a cost/benefit approach to adopting this predictive model. For the 66,776 patients in our sample, without using the predictive model, 8,409 would be readmitted at a cost of

210,225. Using our predictive model, 39,301 will be predicted to be readmitted and receive an intervention, preventing readmission of 7,313 of the 8,409. The total cost of using the model and selective intervention is 106,002, for a saving of 104,223. We are not able to determine the savings that would have been produced using the LACE index.

Given that the proposed model uses relative few variables that are easily obtained and outperforms the LACE index, we recommend additional discussion and testing of this model to further validate its potential use. Also, while the main goal was to identify patients more likely to be readmitted, the results may also be useful with regard to investigating proactive steps that may reduce readmissions.