Exam PA June 16, 2020 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.

Also be sure all the documents you are working on have June 16 attached.

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

> *This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics. In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.*

Task 1 – Edit the data for missing and invalid data (8 points)

Most candidates successfully identified and made adjustments to missing and invalid data. To earn full points, candidates had to make appropriate adjustments and provide clear rationale for their decisions.

- There were 3 unknown/invalid values for the gender variable. Because there were only 3 records out of 10,000 total records, these rows were removed from the data.
- There were 9,691 records with missing values for the weight variable. Because most of the weight data was missing, the variable was removed from the dataset.
- The admin_type_id variable was coded as a numeric variable. Since the numeric values are codes representing categorical data, the variable was changed to a factor variable. There were 1,021 records where the admission type was unavailable, which could be viewed as missing. We do not know if these are missing because of a data collection error or the admission type is routinely unavailable. Because we do not know whether it is missing at random, we should keep it and see whether it being unavailable has predictive power.
- The race variable contained 226 missing values. Similar to the admin_type_id missing data, we do not know if these are missing because of a data collection error or the race is routinely unknown. Because we do not know whether it is missing at random, we should keep the variable and see whether missing race has predictive power. A new race category was created called “Missing.” I also combined the “Asian”, “Hispanic”, and “Other” levels because they each had somewhat low frequencies and similar relationships to the days variable.
- The factor variable levels were reordered so that the most frequent level was first.
- The num_meds variable had values ranging from 1 to 67. It seems unlikely that in a large dataset there would be no individuals that took 0 medications in the prior year. This is suspicious and should be investigated because it could be an indicator of invalid data, but the values look reasonable aside from that, so I will use the variable without alterations.

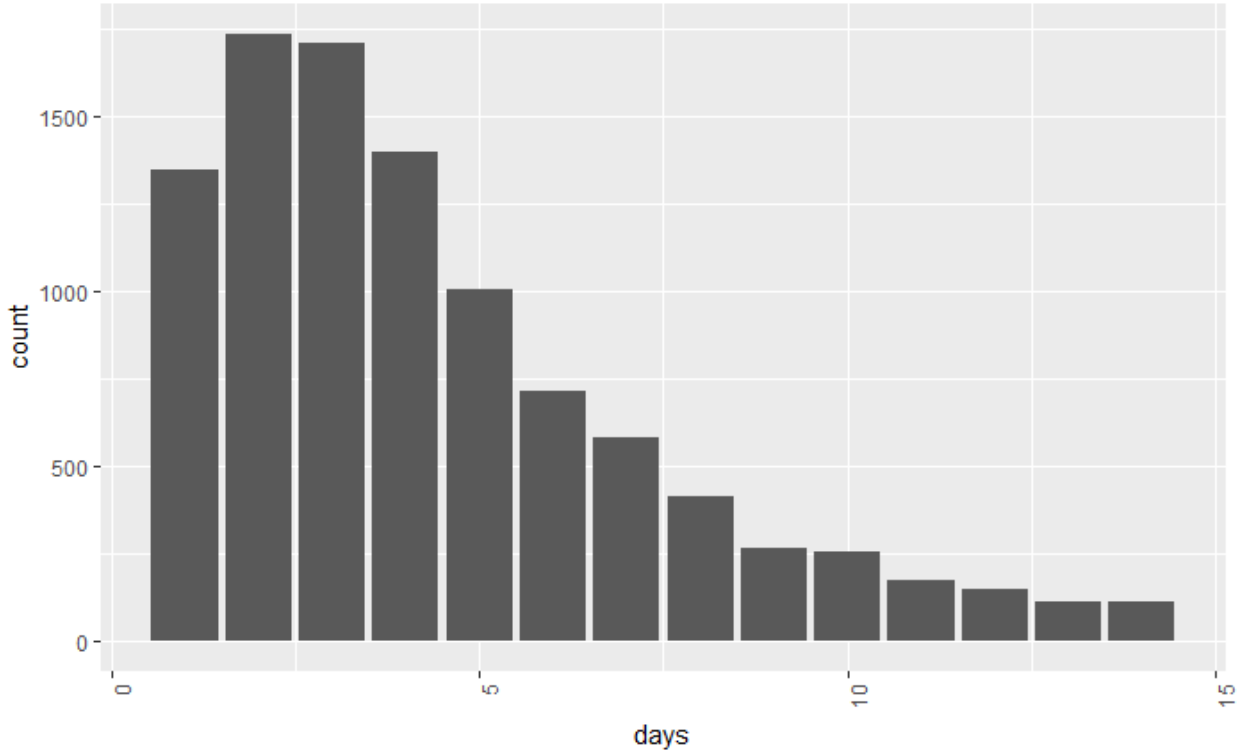
After the changes 9,997 records remained in the dataset.

Task 2 – Explore the data (15 points)

Candidates were expected to use a combination of summary statistics and visualizations for each variable but limit their tables and charts to those that showed the key relationships discussed in the report. The best candidates made insightful observations relating to the business problem when referring to their summary statistics and visualizations. Many candidates failed to adequately explain their reasons for choosing the three predictors.

The target variable is the number of days between admission into and discharge from the hospital. The variable takes on integer values from 1 to 14. The center of the distribution is around 4 to 4.5 days based on the median/mean. From the bar chart below, we can see that the distribution is skewed right with 2-3 days being the most frequent length of stay in the hospital.

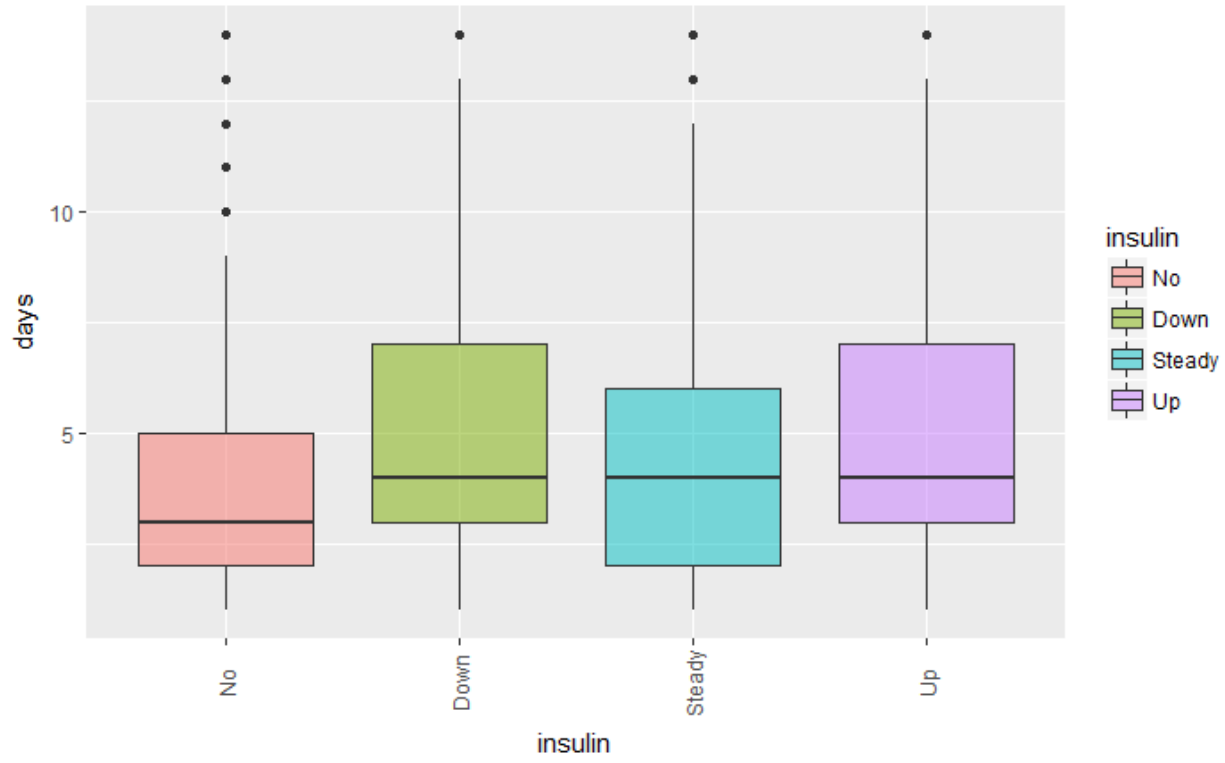
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	2	4	4.409	6	14



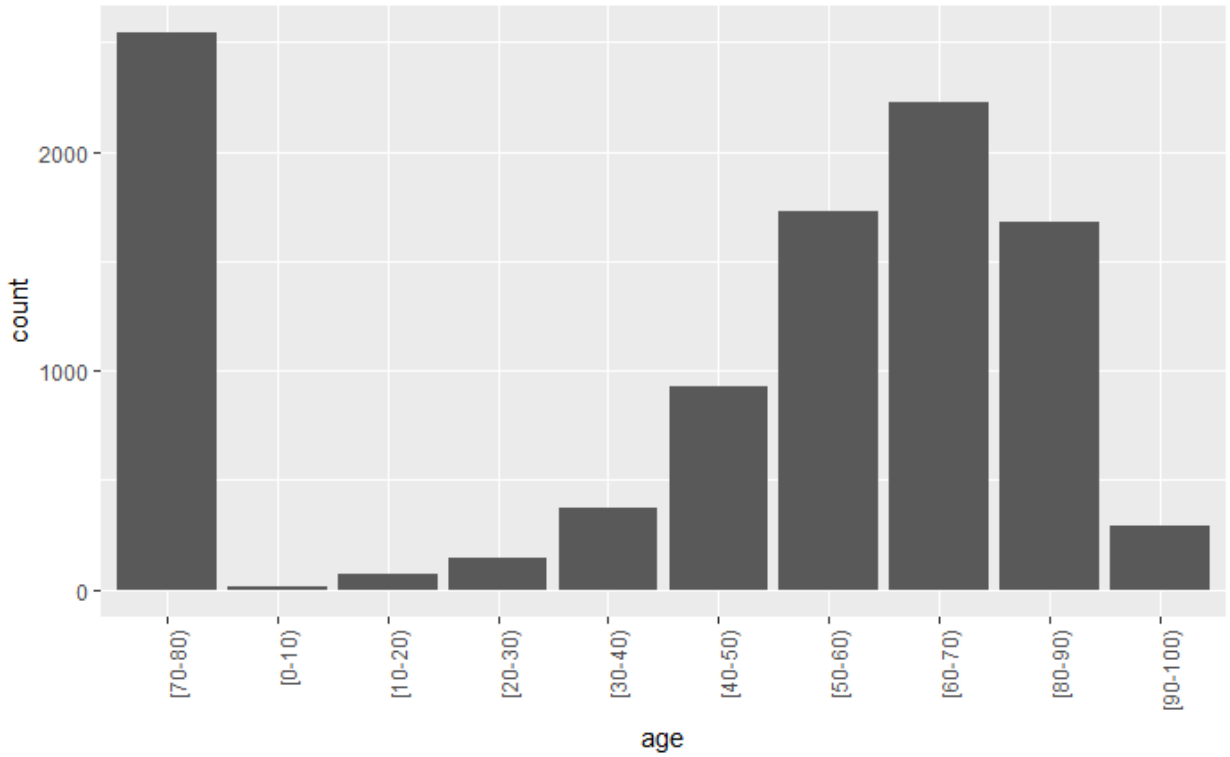
The insulin variable indicates whether, upon admission, insulin was prescribed or there was a change in the dosage. About half of the records did not have insulin prescribed upon admission and these records were admitted on average over a day less than records where insulin was increased upon admission. The boxplots below show that the median and 3rd quartile number of days are also lower when insulin is not prescribed. Changes to insulin dosages also had higher mean days. I selected this variable because (1) each variable level has over 1000 records and a noticeable difference in mean days and (2) It makes

intuitive sense that requiring a medication or change to it upon arrival might lead to a need to monitor a patient over a period of time, increasing the length of stay.

insulin	mean	median	n
No	4.119781	3	4742
Down	4.968280	4	1198
Steady	4.368852	4	2928
Up	5.136404	4	1129



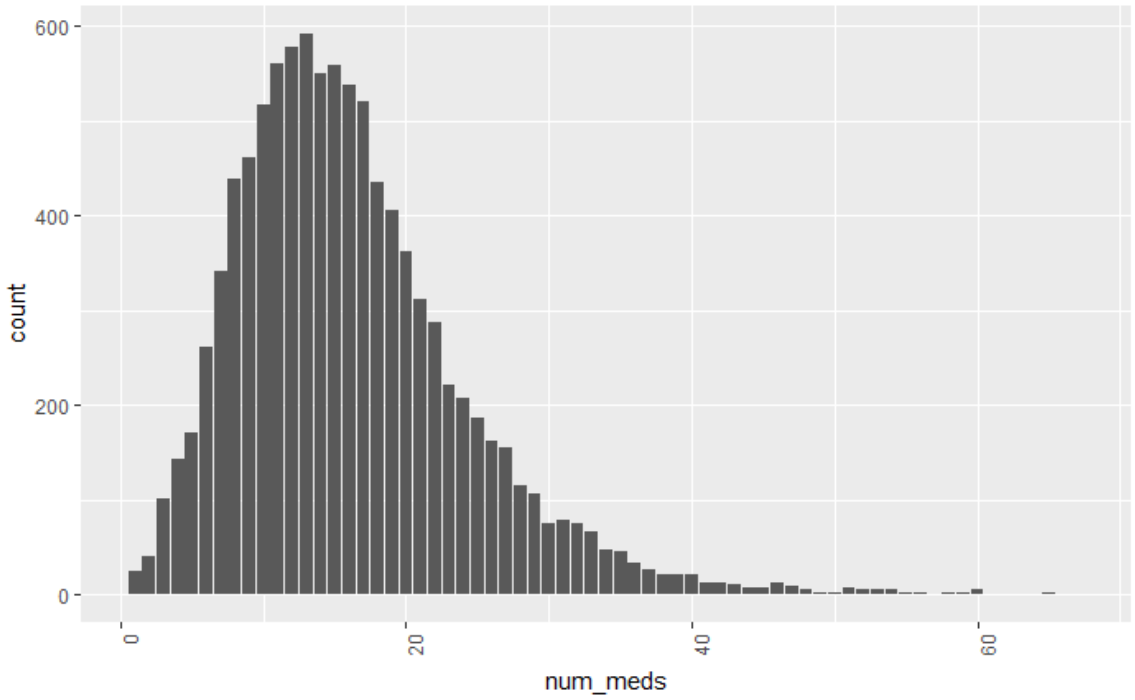
The age variable is a factor variable, and frequency counts can be seen in the bar chart and table below. Most of our data has ages between 50 and 90, with 70-80 containing the most data. If age were numeric, we would say it was skewed left. The table and box plots below show that for the most part as age increases, days increases. There is a sizeable difference (over 1.5 days) between mean days for the age bins with the highest mean days and the lowest mean days. I chose this variable because it makes sense that older patients tend to stay longer since they tend to be in poorer health, and the relationship appears to be strong.

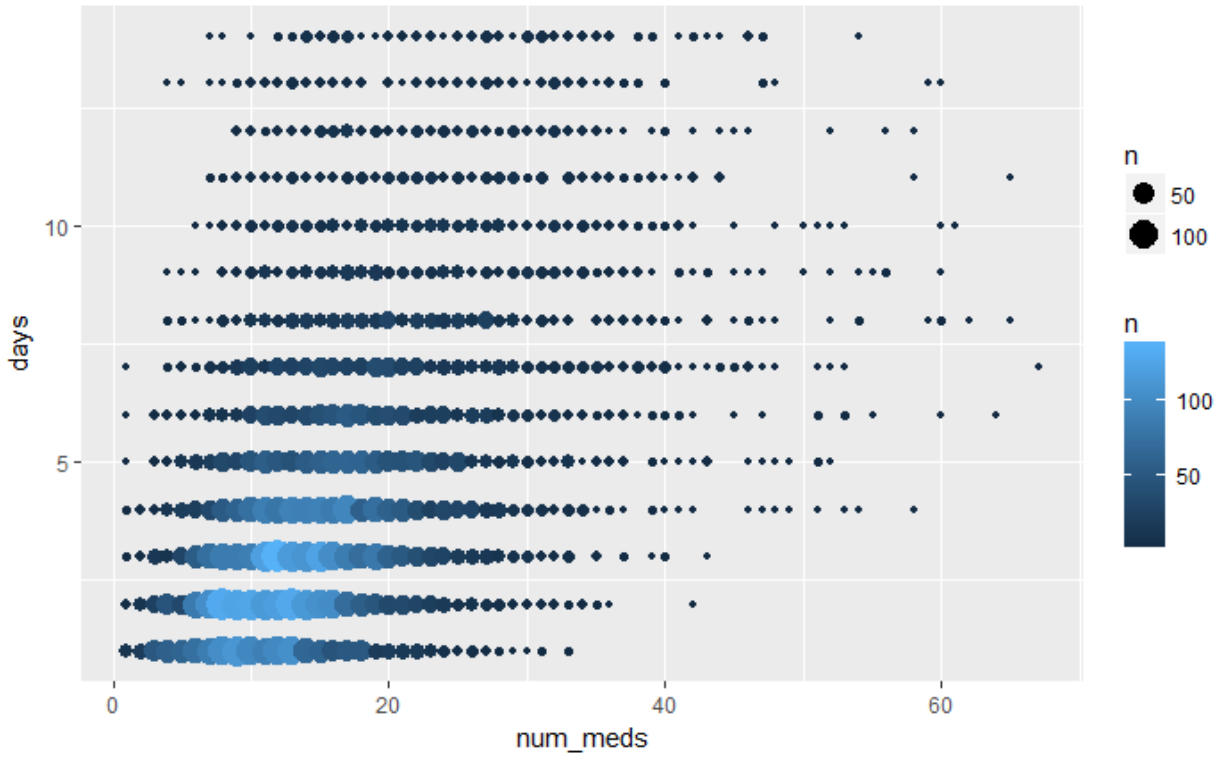


age	mean	median	n
[70-80)	4.658009	4	2541
[0-10)	3.222222	3	18
[10-20)	3.125000	2	72
[20-30)	3.517241	3	145
[30-40)	3.798387	3	372
[40-50)	3.991407	3	931
[50-60)	4.090962	3	1726
[60-70)	4.407989	4	2228
[80-90)	4.817422	4	1676
[90-100)	4.739583	4	288

The num_meds variable takes on integer values from 1 to 67. The center of the distribution is around 15 to 16 based on the median/mean. From the bar chart below, we can see that the distribution is skewed right with 13 being the most frequent number of medications. The correlation between num_meds and days was 0.472, which was the strongest correlation among the numeric variables. The scatterplot, with point size/color based on frequency below shows that as the number of medications increases, the days increases. Like the other variables selected, this relationship makes intuitive sense – patients taking many medications likely have more underlying conditions, have poorer overall health, and may require a longer hospital stay to address those concerns. I selected this variable because the relationship to the target variable was the strongest of the numeric variables and the narrative makes sense.

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
1	10	15	16.16	20	67





Task 3 – Consider two data issues (4 points)

Most candidates successfully identified reasons for considering removing the race variable, but few candidates were able to discuss potential benefits of including the variable. Future candidates should consider the business context more thoroughly when discussing controversial variables. In this case, excluding the race variable could be unethical if it leads to worse care. Lower quality responses mentioned concerns about the race variable due to regulation or privacy laws with little or no explanation.

The race variable presents ethical concerns that should be weighed before using the variable. Historically, racial groups have been mistreated, and efforts to create racial equality continue today.

Hospital administrators intend to use information about important model factors to better manage patient needs. If race turns out to be an important factor, would they apply different treatment plans to different races? This could be seen as discriminatory, but it could also be the ethical choice if it leads to improved care for all races. Failing to take measures to close the gap in length of stay between races could also be seen as discriminatory, since staying at the hospital is expensive, and one race might generally be charged more than another.

Whether or not the race variable is included in the model, users and other stakeholders should make sure races aren't unfairly impacted by the model as it is applied.

Many candidates were unable to describe precisely the problems associated with including the number of laboratory procedures variable.

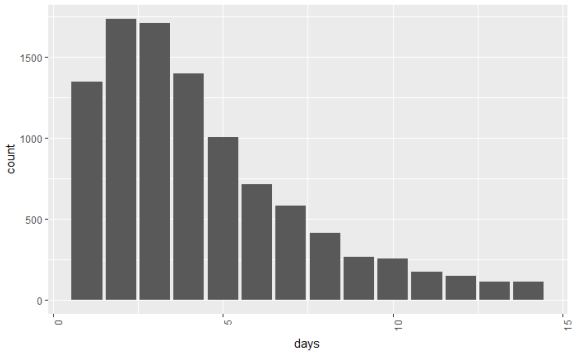
The additional variable that indicates the number of laboratory procedures during the hospital stay should not be included in the model. Typically, variables collected after the time of model application should not be included in a model. Here are two reasons it shouldn't be used:

1. The variable would likely leak information about our target variable, leading to artificially high model performance that could not be realized when the model was used. For example, the number of laboratory procedures might be impacted by the number of days a person is admitted because the hospital might periodically perform lab tests to monitor changes in the patient's health during the hospital stay.
2. Unless hospital administrators know how many laboratory procedures a patient is going to have in the future, they would not be able to use the information in any way. The model cannot be applied at the time of admission if all of the inputs are not known.

Task 4 – Write a data summary for your actuarial manager (6 points)

Most candidates performed well on this task. Lower performing responses often excluded a high-level description of the data source and did not include any visualizations.

The initial data contained 10,000 records based on historical inpatient encounters for patients with diabetes from U.S. hospitals between 1999 and 2008. The dataset contained the following variables about the hospital stay, the patient, their recent treatments, and their treatment upon admission: days (the target variable), gender, age, race, weight, admit_type_id, metformin, insulin, readmitted, num_procs, num_meds, num_ip, num_diags. The distribution of the target variable, skewed to the right, is shown below.



The following are specific issues I explored and adjustments I made.

Completeness and Reasonableness

When reviewing the completeness of the data, I considered the percentage of missing records for each variable and whether having missing values impacted the target variable. Three records were missing for gender and were removed. About 97% of the records were missing weight information, so that variable was removed. The race variable had 226 missing values, but the records did not appear to be missing at random, so I created a new race category called "Missing."

Ethical Concerns

Including the race variable in the model could lead to discriminatory model applications, so we should consider whether to remove the variable from the final model. Before making decisions based on the final model application, we could use the race data to understand whether or not there are any unfair impacts created. We may also want to discuss the issue with legal experts and MACH.

Relevance

Note that the data is limited to diabetes patients, which limits the applicability of this work. I explored the data to see if the variables were appropriate for the problem we are addressing. Descriptive statistics and visualizations were used to analyze the univariate distributions and bivariate (between the variable and target variable) distributions for each variable. Three variables were identified that were likely to predict our target variable, the number of days a patient was admitted to the hospital. Based on my analysis, changes to insulin prescriptions upon admission, higher patient age, and a higher num_meds – the number of distinct medications administered in the prior year – are expected to lead to longer hospital stays.

Additional data preparation steps included recoding variables as factors, combining factor levels, and reordering the factor levels.

Task 5 – Perform a principal components analysis (8 points)

Most candidates were able to describe principal components analysis, but many were unable to describe advantages and disadvantages of using PCA for this problem. The best candidates discussed how correlated variables, centering, and scaling impact PCA. Many candidates thought that PCA dealt with relating the variables to the target variable somehow, leading to poor scores.

Principal components analysis is a method to summarize high dimensional numeric data with fewer dimensions while preserving the spread of the data. It can be particularly helpful when variables are highly correlated. PCA finds orthogonal linear combinations of the input variables (which are typically centered and scaled) called principal components (PCs) that maximize variance to retain as much information as possible. The principal components are ordered according to their variance. The sum of their variances is the total variance explained. It is then common to look at the proportion of variance explained by each principal component to decide how many PCs to use.

Advantages of PCA

PCA could allow us to build a simpler model with fewer features. When exploring data, PCA can help visualize high-dimensional data to explore relationships between variables. PCA can help identify latent variables; in our case a variable, named "overall health" could be based on combinations of our input variables.

Disadvantages of PCA

Using a subset of the principal components results in some information loss. The principal components will be less interpretable than the original variable inputs. Because the hospital administrators want to understand the factors that impact the length of a patients stay, using PCA may not be appropriate for this problem. Although PCA reduces dimensionality in the model, the original variables must still be collected for future predictions, so no efficiency is obtained.

The PCA Analysis yielded the following output:

```
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.2267 1.0426 0.9141 0.7568
Proportion of Variance 0.3762 0.2717 0.2089 0.1432
Cumulative Proportion  0.3762 0.6479 0.8568 1.0000

                PC1         PC2         PC3         PC4
num_procs 0.5572974 -0.44554566  0.36869207  0.5957977
num_meds  0.6739567 -0.05424897  0.09615285 -0.7304752
num_ip    0.1178855  0.79602911  0.58013950  0.1260112
num_diags 0.4704307  0.40605883 -0.71990204  0.3091152
```

The first table shows the proportion of variance explained by each component. PC1 has about 38% of the total variance. The bottom row of that same table shows the cumulative variance explained. If we used 3 PCs in our model, we would retain about 86% of the information.

The second table shows the coefficients applied to the input variables to create the principal components. The size and sign of the coefficients indicates the relative influence each input variable has on the PC. For PC1, the number of medications, the number of procedures, and the number of diagnoses have similar influence for higher values in each while the number of inpatient visits has relatively little impact.

Using only the first principal component in our model would result in significant information loss since it only explains 38% of the variance. For this reason, additional PCs should be included or the original input

variables should be used instead. If we need to include 3 PCs to retain avoid losing a lot of information, it might be better to keep 4 input variables that are more easily interpreted.

Task 6 – Construct a decision tree (10 points)

This task was made up of several very specific subtasks, and it is important that candidates make sure they are performing the exact task requested. Many candidates seemed to ignore detailed requests in the problem statement, resulting in poorer scores. Some candidates did not explain why pruning was important in the context of the business problem. When choosing the optimal CP parameter, alternative approaches were able to earn full points as long as the approach was justified. Where specifically requested or necessary for facilitating discussion (e.g. discussing the CP table or plot), candidates should include R output in their report. Candidates were expected to interpret all leaves of the pruned tree.

When a decision tree is trained, it can become very large and include splits that are not particularly valuable for predictions on new data. When examining the fit of a tree, it is a good idea to try to prune a tree when it has many splits that do not improve performance. Pruning reduces the size of the tree, hopefully removing less valuable splits from the tree. This process reduces overfitting the tree on the training data, can lead to better predictions, and results in a simpler, more interpretable tree.

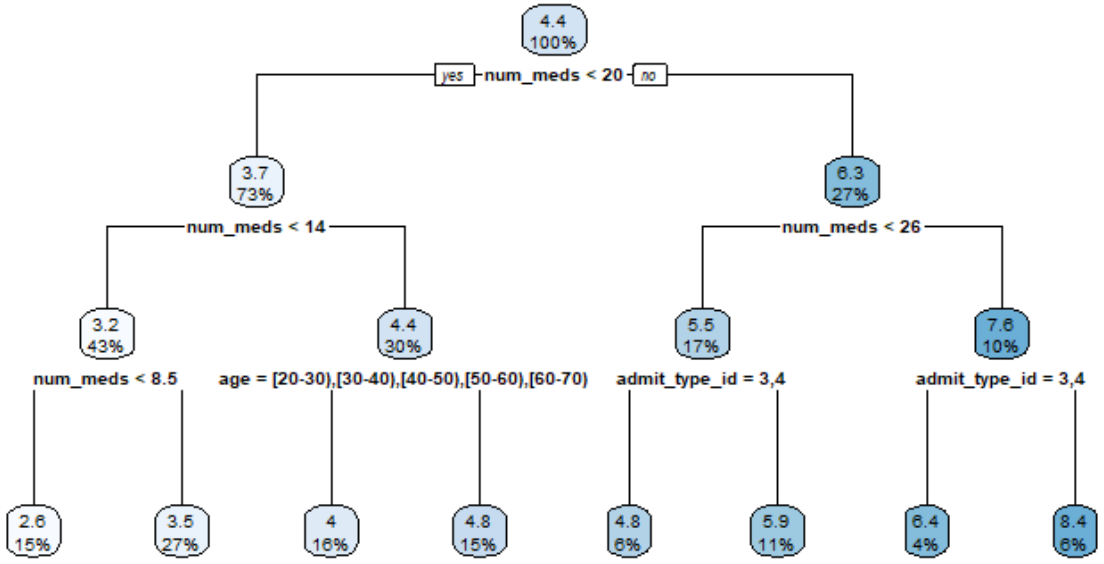
To help the hospital administrators understand the factors that lead to longer hospital stays, it is important that our tree can be understood and that we ignore any relationships that are based on noise in the training data. For these reasons, pruning should be used.

The following output is from the initial unpruned tree. The optimal CP is the one that minimizes the cross validation error (in the xerror column). Row 20 accomplishes that, with CP = 0.001430991. Pruning with this CP value will result in a tree with 19 splits and so 20 leaves.

```
            CP nsplit rel error    xerror       xstd
1  0.147557542      0 1.0000000 1.0001409 0.02036690
2  0.029268050      1 0.8524425 0.8530852 0.01820828
3  0.027590674      2 0.8231744 0.8350616 0.01803185
4  0.010035051      3 0.7955837 0.8000857 0.01744791
5  0.008589466      4 0.7855487 0.7940982 0.01744242
6  0.004761404      5 0.7769592 0.7831189 0.01717749
7  0.004292069      6 0.7721978 0.7836060 0.01726199
8  0.004091581      7 0.7679057 0.7828297 0.01727611
9  0.003207097      8 0.7638142 0.7806302 0.01726371
10 0.002987113      9 0.7606071 0.7772366 0.01720296
11 0.002222715     10 0.7576200 0.7763915 0.01726294
12 0.002145076     11 0.7553972 0.7771849 0.01728728
13 0.002064277     12 0.7532522 0.7761244 0.01727037
14 0.002034127     13 0.7511879 0.7774116 0.01729694
15 0.001916314     14 0.7491538 0.7768608 0.01724022
16 0.001886978     15 0.7472374 0.7771581 0.01724854
17 0.001801968     16 0.7453505 0.7766402 0.01723072
18 0.001668200     17 0.7435485 0.7746664 0.01719913
19 0.001498365     18 0.7418803 0.7740650 0.01719737
20 0.001430991     19 0.7403819 0.7740525 0.01717004
21 0.001253460     20 0.7389509 0.7757195 0.01719489
22 0.001229875     21 0.7376975 0.7783302 0.01727428
23 0.001140899     22 0.7364676 0.7779563 0.01723795
24 0.001139919     23 0.7353267 0.7799776 0.01727255
```

```
25 0.001135262     24 0.7341868 0.7802674 0.01728302
26 0.001079637     25 0.7330515 0.7797102 0.01727112
27 0.001026098     27 0.7308923 0.7810684 0.01726792
28 0.001019631     28 0.7298662 0.7816605 0.01727265
29 0.001000000     30 0.7278269 0.7820143 0.01724219
```

Using CP = 0.0042 to prune the tree results in a tree with 8 leaves.



The table below shows the Pearson goodness-of-fit statistic for the original (unpruned) tree and the pruned tree with 8 leaves on the training and test data. The statistic is a way to measure the error between the predicted values and the actual values, so a smaller value is better. Based on the table below, the original tree performed better than the pruned tree.

Tree	Statistic on Train Data	Statistic on Test Data
Original	1.501790	1.459844
Pruned	1.582546	1.487181

The pruned tree allows for 8 possible predicted values, 1 for each of the leaves. The tree can be interpreted as 8 series of if statements. The possibilities are summarized below for the 8 leaves pictured above from left to right. Note that predicted days are rounded based on decision tree image.

1. If num_meds < 8.5, predict 2.6 days
2. If 8.5 <= num_meds < 14, predict 3.5 days
3. If 14 <= num_meds < 20 and age in [20,70), predict 4 days
4. If 14 <= num_meds < 20 and age < 20 or age > 70, predict 4.8 days

5. If 20 <= num_meds < 26 and admit_type_id in [3,4], predict 4.8 days
6. If 20 <= num_meds < 26 and admit_type_id not in [3,4], predict 5.9 days
7. If num_meds >= 26 and admit_type_id in [3,4], predict 6.4 days
8. If num_meds >= 26 and admit_type_id not in [3,4], predict 8.4 days

According to the pruned tree, a distinguishing factor for predicting the length of hospital stay is the number of medications. Depending on the number of medications, age and admit_type_id may also be distinguishing factors.

Task 7 – Construct a generalized linear model (7 points)

The most successful candidates were able to discuss their distribution choices in light of the data structure and the business problem. Some candidates failed to point out that using the PCs may make it more difficult for the hospital administrators to understand the drivers of longer hospital stays.

Binomial distributions are typically used when there are only two outcomes, so it would not be a good fit. Gamma is used for non-negative continuous variables. Although the target variable we have is discrete, it could also be seen as continuous since patients are actually discharged at various points in the day, so the gamma distribution could be a viable alternative.

GLM	Statistic on Train Data	Statistic on Test Data
All variables except PC	1.569789	1.445401
With PC and without original numeric variables	1.631226	1.499745

The GLM with the PC in place of the numeric variables performed slightly worse, and using the PC instead would make interpreting the factors that lead to longer stays more difficult. I will use the model without the PC since it will be easier to gain a better understanding of the factors driving length of stay and has better performance.

Task 8 – Perform feature selection with lasso regression (4 points)

Some candidates did not know whether a higher or lower Pearson goodness-of-fit statistic was good or bad. Candidates were expected to make a clear model recommendation and justify their choice. Better candidates went beyond comparing the performance and variables of the GLM and LASSO models and discussed how both were good or bad for the specific business problem. Either method could be justified and receive full credit.

The features used by the model are:

- Age = [80-90)
- admit_type_id = 2
- num_meds
- num_diags

GLM	Statistic on Train Data	Statistic on Test Data
GLM selected in Task 7	1.569789	1.445401
LASSO Model	1.607803	1.503696

The output from the GLM selected in task 7 is below.

```
Call:
glm(formula = days ~ . - PC1, family = poisson(link = "log"),
    data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7009  -1.0116  -0.2918   0.5739   4.7992

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.6871798  0.0318652  21.565  < 2e-16 ***
genderMale          -0.0262939  0.0116099  -2.265 0.023527 *
age[0-10)            0.1191066  0.1636182   0.728 0.466641
age[10-20)          -0.0111956  0.0787686  -0.142 0.886975
age[20-30)          -0.1024500  0.0518593  -1.976 0.048207 *
age[30-40)          -0.0980179  0.0352703  -2.779 0.005452 **
age[40-50)          -0.0968176  0.0225696  -4.290 1.79e-05 ***
age[50-60)          -0.1230116  0.0182452  -6.742 1.56e-11 ***
age[60-70)          -0.0764664  0.0165057  -4.633 3.61e-06 ***
age[80-90)           0.0743523  0.0173622   4.282 1.85e-05 ***
age[90-100)          0.1207857  0.0342164   3.530 0.000415 ***
raceMissing          0.0732532  0.0379248   1.932 0.053417 .
raceAfricanAmerican  0.1042902  0.0150314   6.938 3.97e-12 ***
raceOther            0.0757885  0.0301442   2.514 0.011930 *
admit_type_id2       0.1049588  0.0151845   6.912 4.77e-12 ***
admit_type_id3      -0.0756297  0.0162349  -4.658 3.19e-06 ***
admit_type_id4      -0.0396198  0.0205325  -1.930 0.053655 .
metforminDown        0.0614142  0.0705831   0.870 0.384247
metforminSteady     -0.0107102  0.0152937  -0.700 0.483739
metforminUp          0.1406112  0.0480649   2.925 0.003440 **
insulinDown          0.0273141  0.0185431   1.473 0.140751
insulinSteady       -0.0152972  0.0137588  -1.112 0.266220
insulinUp            0.0224528  0.0187638   1.197 0.231463
readmitted<30        0.0530252  0.0186097   2.849 0.004381 **
readmitted>30        0.0295589  0.0126609   2.335 0.019561 *
num_procs            0.0100826  0.0036437   2.767 0.005655 **
num_meds             0.0305111  0.0007091  43.027  < 2e-16 ***
num_ip               0.0136492  0.0043966   3.105 0.001906 **
num_diags            0.0320346  0.0035064   9.136  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 13255  on 6999  degrees of freedom
Residual deviance: 10038  on 6971  degrees of freedom
AIC: 32111

Number of Fisher Scoring iterations: 5
```

The LASSO coefficients are shown below:

```
genderMale          .
age[0-10)           .
age[10-20)          .
age[20-30)          .
```

```
age[30-40)          .
age[40-50)          .
age[50-60)          .
age[60-70)          .
age[80-90)          0.008628822
age[90-100)         .
raceMissing         .
raceAfricanAmerican .
raceOther           .
admit_type_id2      0.016065072
admit_type_id3      .
admit_type_id4      .
metforminDown       .
metforminSteady     .
metforminUp         .
insulinDown         .
insulinSteady       .
insulinUp           .
readmitted<30       .
readmitted>30       .
num_procs           .
num_meds            0.027345435
num_ip              .
num_diags           0.018298035
```

The LASSO model performed slightly worse than the GLM from task 7, but there is more to consider for our business problem, which focused on interpretability. Both the LASSO model and the GLM from task 7 are easy to interpret. Clearly, the LASSO model is much simpler since it removed several features that were used in the GLM from task 7. One limitation for the LASSO model arises from the way the LASSO model binarizes the factor variables and thus can shrink individual factor level coefficients to zero. You can see that occurred with many of the age bins. The LASSO model will give the same prediction for a patient in their 50s as one in their 90s. While that result might be simpler, it does not make intuitive sense, whereas the coefficients from the GLM in task 7 suggested a longer stay as age increased (starting at age 50). That problem with the LASSO model could be avoided by changing the age variable into a numeric variable (perhaps by using the midpoint of each bin). That being said, the simplification of the model among the other features will help narrow the list of important factors for hospital administrators, so I recommend the LASSO model for this business problem.

Task 9 – Discuss the bias-variance tradeoff (7 points)

Many candidates mixed up bias and variance, or were unable to relate variance to overfitting and bias to underfitting. Better candidates explained how model complexity could refer to both the model type and the features included.

Bias is the expected loss caused by the model not being complex enough to capture the signal in the data. Variance is the expected loss from the model being too complex and overfitting to the training data.

We typically think of the expected loss as Bias + Variance + Unavoidable error. When building models, we are trying to minimize this expected loss, but to do so we often need to find a balance between bias and variance. Models with low bias tend to have higher variance and vice versa.

Without regularization, coefficients are found that maximize the likelihood function. This results in models that may not be optimal because coefficients are found even for features that may not be important. This process results in models that tend to overfit to the training data; they have high variance. LASSO penalizes models that have large coefficients to the extent that it can shrink coefficients of unhelpful predictors to zero. This is essentially trading some of the high variance from our non-regularized model for increased bias, which can potentially reduce the overall error.

With high variance (overfitting), the model will perform better on the training set than on a test set. With high bias (underfitting), the model will perform poorly on both the training set and the test set. When evaluating a single model, using a test set will help detect whether we have high variance because we can see a difference between the training and test set performance. When comparing models with different levels of complexity, comparing the test set performance and selecting the best performing model can also help us select the model design with the least total error.

Task 10 – Consider the final model (4 points)

Most candidates were able to identify advantages and disadvantages of GLMs vs decision trees, but many struggled to do the same with GLMs vs LASSO.

Advantage of a GLM vs Decision Tree

Consider the case where an increase in the number of medications produces an increase in the expected length of admission. A decision tree separates a numeric variable such as the number of medications into buckets. For the tree to capture the true relationship, it needs to split on the same variable many times, creating a very complex tree that would be difficult to interpret. On the other hand, a GLM can fit to this relationship with a single coefficient that summarizes how the expected length of admission increases with each unit increase of number of medications (provided there is a simple functional form that describes the relationship). Because our data includes numeric variables, a GLM might capture the nature of the true relationship while being more interpretable.

Disadvantage of a GLM vs Decision Tree

GLMs do not capture the effects of variable interactions automatically. If the right interactions aren't explicitly coded in the GLM, the model may be unable to fit the data well. A decision tree will automatically create variable interactions as it is trained. For example, the pruned decision tree built earlier found an interaction between the number of medications and the age of the patient. No such interaction was even tried with the GLM.

Advantage of a GLM without removing features vs LASSO regularization

The GLM can retain insignificant factor levels that might be dropped by the LASSO model. This can lead to improved interpretability via comparison of factor levels. The LASSO model has to binarize the factor variables and can shrink some individual factor level coefficients to zero. In the LASSO model created earlier, this occurred with many of the age bins. The LASSO model will give the same prediction for a patient in their 50s as one in their 90s because both levels were dropped by the model. While that result might be simpler, it does not make intuitive sense, whereas the coefficients from the GLM in task 7 suggested a longer stay as age increased (starting at age 50).

Disadvantage of a GLM without removing features vs LASSO regularization

Building a GLM without removing features can lead to a model that is overfit to the training data because coefficients will be found even for features that are not important.

Task 11 – Interpret the model for the client (7 points)

Candidates were expected to know the application of the coefficients would be multiplicative. Many candidates struggled to explain how to use the model using language appropriate for the client.

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.6885918  0.0265186  25.966  < 2e-16 ***
genderMale          -0.0296000  0.0097062  -3.050 0.002292 **
age[0-10)            0.1174630  0.1323599   0.887 0.374836
age[10-20)           0.0042256  0.0685219   0.062 0.950828
age[20-30)          -0.0987227  0.0458631  -2.153 0.031354 *
age[30-40)          -0.0974386  0.0286530  -3.401 0.000672 ***
age[40-50)          -0.0962409  0.0190828  -5.043 4.58e-07 ***
age[50-60)          -0.1312465  0.0151951  -8.637  < 2e-16 ***
age[60-70)          -0.0806118  0.0137196  -5.876 4.21e-09 ***
age[80-90)           0.0655843  0.0145593   4.505 6.65e-06 ***
age[90-100)          0.1199149  0.0288651   4.154 3.26e-05 ***
raceMissing          0.0358093  0.0319222   1.122 0.261962
raceAfricanAmerican  0.0929807  0.0126331   7.360 1.84e-13 ***
raceOther            0.0600796  0.0252317   2.381 0.017260 *
admit_type_id2       0.1010454  0.0126950   7.959 1.73e-15 ***
admit_type_id3      -0.0965367  0.0137104  -7.041 1.91e-12 ***
admit_type_id4      -0.0367625  0.0169053  -2.175 0.029659 *
metforminDown        0.0944203  0.0585558   1.612 0.106857
metforminSteady     -0.0111176  0.0127279  -0.873 0.382398
metforminUp          0.1691376  0.0402240   4.205 2.61e-05 ***
insulinDown         -0.0010478  0.0154190  -0.068 0.945821
insulinSteady       -0.0224321  0.0115379  -1.944 0.051871 .
insulinUp            0.0235433  0.0155238   1.517 0.129370
readmitted<30        0.0605620  0.0157021   3.857 0.000115 ***
readmitted>30        0.0350362  0.0105465   3.322 0.000894 ***
num_procs            0.0122596  0.0030399   4.033 5.51e-05 ***
num_meds             0.0311730  0.0005943  52.453  < 2e-16 ***
num_ip               0.0160014  0.0037233   4.298 1.73e-05 ***
num_diags            0.0316686  0.0029282  10.815  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Positive coefficients for an input variable increase the predicted length of admission, while negative coefficients decrease it, with numbers further from zero having larger effects. These model coefficients can be translated into factors that can be multiplied together to determine a patient’s predicted length of admission, where 1.99 days (interpreted from the intercept) are predicted before applying any factors. The table below illustrates how the coefficients are interpreted for different types of items. If a patient is male, their length of admission on average will be 97% of the length of admission for a female. For the last four items, the multiplier is applied per procedure, medication, etc. A patient with 3 procedures in the prior year would have, on average, a length of admission that is 101.2% the length of admission for a patient with only 2 procedures in the prior year.

Input	Input Type	Coefficient	Interpreted Coefficient
If the patient is male	Categorical	-0.0296000	0.970834
For each additional procedure in the prior year	Numeric	0.0122596	1.012335

Task 12 – Executive summary (20 points)

Rather than restating information from prior tasks, candidates were expected to alter their messaging for the intended audience. Often this includes avoiding overly technical language, discussing topics at a different level of detail, and translating performance metrics to be more meaningful to the reader. Brief discussions about approaches attempted are acceptable, but candidates should avoid lengthy discussion about models or techniques that were not ultimately selected. The best candidates were able to incorporate the business context of the problem throughout their summary.

To: Merged and Acquired Clinics and Hospitals Executives

From: Actuarial Analyst

You have asked us to build a model that yields insights about the factors driving the length of inpatient hospital stays so the hospital administrators can better understand and manage patient needs. We were supplied with 10,000 observations based on historical inpatient encounters for patients with diabetes from U.S. hospitals between 1999 and 2008. Each observation contained information about the hospital stay, the patient, their recent treatments, and their treatment upon admission. The model we constructed identifies information that can be used to predict the length of inpatient visits for diabetes patients. The model will not be relevant for patients that do not have diabetes. To build a model that generalizes well for all patients, data about other types of patients should be obtained.

Prior to building the model, the data were reviewed for completeness, reasonableness, ethical concerns, and relevance. The variables included the hospital stay length in days, demographic information, the type of hospital admission, history of medical activity in the prior 12 months, and information about diabetes medication changes upon admission (metformin and insulin). The weight variable was discarded because it contained mostly missing values. Observations missing gender information were removed. A separate category was created to address any missing race information.

MACH should weigh the risks of using the model with the race variable included. The hospital system could be inviting legal action if decisions based on a model, with or without race included, are viewed as discriminatory. To mitigate the risk, additional work could be performed to make sure the races are not unfairly impacted by decisions based on the model.

After modifying a few features to prepare them for modeling, I tried a variety of models to see which would best explain the factors affecting the length of hospital stays. Each model was calibrated using 70% of the data and then its performance was measured using the other 30%. This process helps identify models that adequately capture the patterns in the data and generalize well to new data. Every model we built predicted the length of the inpatient visit in days. Many of the models had similar performance. When methods used to simplify the model by decreasing the number of inputs were

attempted, they led to decreased performance and did not make it much easier to determine the factors leading to longer hospital stays, so they were not used in our final model.

The selected model is a generalized linear model. It had the best performance while offering insights into the factors affecting the length of stay.

> *The intent of the following is to justify that the model is useful. There are many approaches for doing this, but it is important to consider that the audience likely needs positive proof that the model is worth using.*

Rather than build a model, we could have simply used the average length of admission from the 70% of data used for training (4.4 days) as a prediction. When this approach is compared to the recommended model on unseen data reserved for performance measurement, the predictions from the recommended model lead to a 25% reduction in error. The selected model displays distinctions among patients that lead to better predictions on the length of stay.

The model coefficients can be used to gain insights about the factors affecting a patient's length of stay. The model starts with a baseline predicted length of stay for each patient of 1.99 days. Then, it applies the factors below based on the patient data. Note that values have been rounded. Multiplying by factors greater than 1 increase the predicted length of stay, while multiplying by factors less than 1 decrease the predicted length of stay.

If the patient is Male	Multiply by 0.97
If age 0 to 9	Multiply by 1.12
If age 10 to 19	Multiply by 1.00
If age 20 to 29	Multiply by 0.91
If age 30 to 39	Multiply by 0.91
If age 40 to 49	Multiply by 0.91
If age 50 to 59	Multiply by 0.88
If age 60 to 69	Multiply by 0.92
If age 80 to 89	Multiply by 1.07
If age 90 to 99	Multiply by 1.13
If race is missing	Multiply by 1.04
If race is African American	Multiply by 1.10
If race is Asian, Hispanic, or Other	Multiply by 1.06
If admit type is urgent	Multiply by 1.11
If admit type is elective	Multiply by 0.91
If admit type is not available	Multiply by 0.96
If metformin dosage decreased upon admission	Multiply by 1.10
If metformin prescription exists but dosage unchanged upon admission	Multiply by 0.99
If metformin dosage increased upon admission	Multiply by 1.18
If insulin dosage decreased upon admission	Multiply by 1.00
If insulin prescription exists but dosage unchanged upon admission	Multiply by 0.98
If insulin dosage increased upon admission	Multiply by 1.02
If patient is being readmitted within 30 days of their last inpatient visit	Multiply by 1.06
If patient is being readmitted but it has been more than 30 days since their last inpatient visit	Multiply by 1.04

If the patient has had n procedures in the prior 12 months	Multiply by 1.01^n
If the patient has taken n medications in the prior 12 months	Multiply by 1.03^n
If the patient has had n inpatient visits in the prior 12 months	Multiply by 1.02^n
If the patient has had n diagnoses in the prior 12 months	Multiply by 1.03^n

Many of the factors affecting the length of stay make intuitive sense. Starting at age 50, the length of stay on average increases as age increases, which is not surprising as older patients tend to have declining health. Increased treatments (procedures, medications, inpatient visits, and diagnoses) in the prior year also led to longer stays on average.

As a next step, I recommend discussing how the hospital intends to manage care differently as a result of the model. Then we can analyze impacts to protected groups to ensure the model is fairly applied.