Spelling and Grammar Exercise

**Instructions:**

1. **Proof-read this document as you would if it was your own writing.**
2. **Compare it to the “Spelling and Grammar Exercise – Solution” file.**
3. **You may wish to skip parts of this. It will take you at least an hour to proofread the entire 20-page document. See tasks 4, and 12 (the Executive Summary) in particular because these tasks use non-technical language.**

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

I made changes to each variable for fix the missing values and data types.

**Gender**

There are three missing values. I just removed these because there are not enough values to make a difference.

**Admit\_type\_id**

This variable came as numeric but is a factor, and so I converted this to the factor data type. I also set the base level to the one which have the most observations, which was Emergency. I noticed that there were 2021 cases where this was “Not Available”. This could either indicate a missing value or it could be intentional. If it is a missing value, then I recommend not removing it because there could be a pattern in the missingness – some patients who have their values omitted may have done so because of medical reasons. I chose to include these values and test if they will be predictive.

**Race**

There were 226 missing values which I grouped into the “other” category. There were also only a few records which had certain levels, for race = Asian, and so I combined these with the other category so that there is more credibility.

There could be a pattern in why these values are missing, which would cause these results to be biased. For instance, when the data was collected, there could be more missing values for certain races than for others due to the surveys which asked patients of their race.

**Weight**

There were 9,688 missing values out of only 10,000 records overall and so I just removed this variable. It did not matter why this variable was missing because there are simply not enough records to use.

**Num\_ip**

I noticed that there were 12 patients who had 9 inpatient visits in the prior 12 months, which seemed high to me. There were also 15 patients who had 8 visits. I chose to leave these as they are but recommend investigating them to be sure that it makes sense to compare them with the other patients in this analysis.

**Num\_diags**

It was surprising to me that there were many patients who had many diagnosis, leading up to 9, where there are 4,914, but then only 2 who had 10 diagnosis. This is described as “Number of diagnoses entered to the system in the twelve months preceding the encounter” and so it could be that these 2 records are errors or that there are more patients who should have 10 diagnosis. I chose to leave these as they are.

|  |
| --- |
|  |
| **num\_diags**  <int> | **n**  <int> |  |  |  |
| 1 | 23 |  |  |  |
| 2 | 96 |  |  |  |
| 3 | 267 |  |  |  |
| 4 | 570 |  |  |  |
| 5 | 1063 |  |  |  |
| 6 | 992 |  |  |  |
| 7 | 996 |  |  |  |
| 8 | 1069 |  |  |  |
| 9 | 4914 |  |  |  |
| 10 | 2 |  |  |  |

Lastly, I converted the factor’s base levels to those which had the most observations. You can see these values below as they are the first value listed. This is Gender = Female, age = [70-80], and so forth.

days gender age race admit\_type\_id metformin insulin readmitted num\_procs num\_meds num\_ip num\_diags

Min. : 1.000 Female:5338 [70-80):2541 Caucasian :7531 1:5289 No :8024 No :4742 NO :5370 Min. :0.000 Min. : 1.00 Min. : 0.0000 Min. : 1.000

1st Qu.: 2.000 Male :4659 [60-70):2228 AfricanAmerican:1848 2:1870 Steady:1799 Steady:2928 >30:3525 1st Qu.:0.000 1st Qu.:10.00 1st Qu.: 0.0000 1st Qu.: 6.000

Median : 4.000 [50-60):1726 Other : 419 3:1817 Up : 114 Down :1198 <30:1102 Median :1.000 Median :15.00 Median : 0.0000 Median : 8.000

Mean : 4.409 [80-90):1676 Hispanic : 199 4:1021 Down : 60 Up :1129 Mean :1.345 Mean :16.16 Mean : 0.6388 Mean : 7.442

3rd Qu.: 6.000 [40-50): 931 3rd Qu.:2.000 3rd Qu.:20.00 3rd Qu.: 1.0000 3rd Qu.: 9.000

Max. :14.000 [30-40): 372 Max. :6.000 Max. :67.00 Max. :21.0000 Max. :16.000

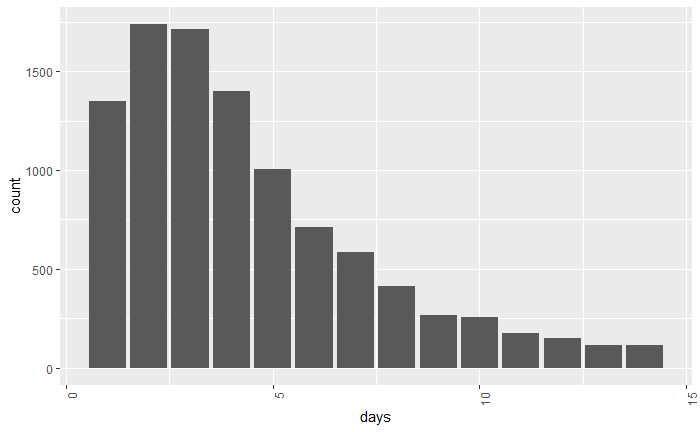
(Other): 523

## Task 2 – Explore the data (15 points)

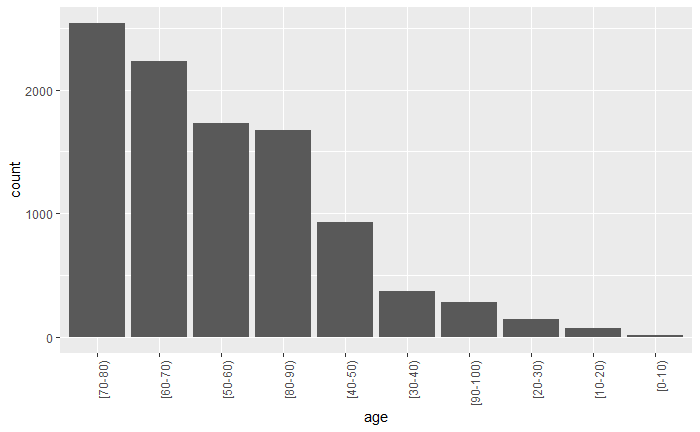
The target variable is the number of days that a patient spends in the hospital after being admitted. This has a mean of 4.4 and a median of 4.0, indicating that it is right-skewed, which you can also see from the histogram below.

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.000 4.000 4.409 6.000 14.000



I expect that the age variable will be predictive of the length of stay. I noticed that older people tend to have longer stays than younger people. Most patients are older, over the age of 50. The median length of stay for people over age 50 is between 3 and 4 days but only between 2 and 3 for patients under 50.



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

1-10 of 10 rows

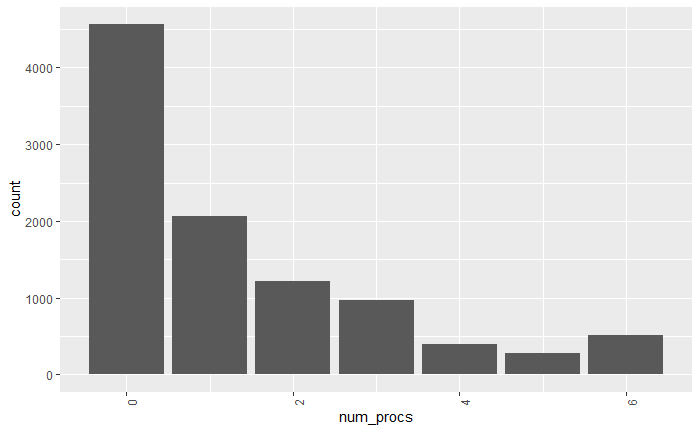
The table above shows that there is a significant difference in the mean length of stay across the age categories.

**Num\_procs**

I expect that the number of procedures will also be predictive. This includes only procedures over the preceding 12 months, and so unhealthy patients will need to have more procedures and might be more likely to have a longer recovery time and length of stay. There is an even spread of observations, with most patients having no procedures. Those who do not have the lowest average length of stay, of 3.78, whereas those who have 5, have an average of 5.4. This makes sense because patients who have more procedures probably need more time to heal after being admitted.

| **num\_procs**  <int> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |  |
| --- | --- | --- | --- | --- |
| 0 | 3.778557 | 3 | 4561 |  |
| 1 | 4.558624 | 4 | 2064 |  |
| 2 | 4.998365 | 4 | 1223 |  |
| 3 | 5.040248 | 4 | 969 |  |
| 4 | 5.631714 | 5 | 391 |  |
| 5 | 5.248175 | 4 | 274 |  |
| 6 | 5.434951 | 5 | 515 |  |

7 rows

The correlation between the number of days and length of stay was 19%. This means that the length of stay increases as the number of procedures increases.

**Readmitted**

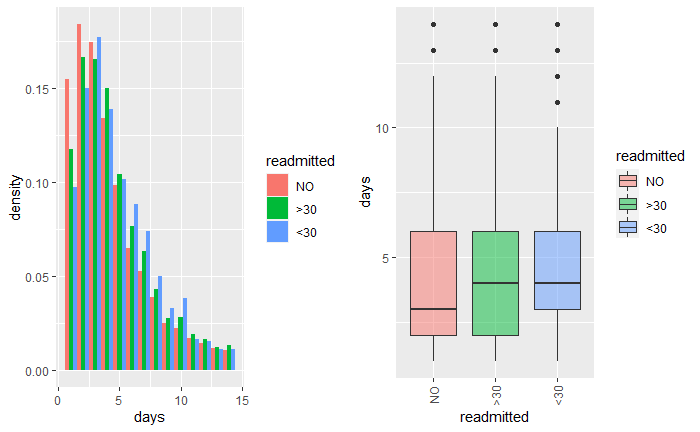
Patients who had a history of being readmitted tended to have longer stays. Patients who had not been readmitted at all had the lowest length of stay at 4.2, whereas patients who had been readmitted in the previous 30 days had a mean of 4.77. There were over 1,000 patients in this group, so I know that this is not just random chance.

| **readmitted**  <fctr> | **mean**  <dbl> | **median**  <dbl> | **n**  <int> |  |
| --- | --- | --- | --- | --- |
| NO | 4.229609 | 3 | 5370 |  |
| >30 | 4.571064 | 4 | 3525 |  |
| <30 | 4.766788 | 4 | 1102 |  |

3 rows

The graphs below show that patients who were readmitted in the prior 30 days have longer stays. The box plot on the right shows the median days as the center line in each box. The blue box (readmitted < 30) is highest and the red box (no readmissions) is lowest.

The histogram on the left shows that the patients who had been readmitted spend more days in the hospital because the blue line is always higher than the orange line.



## Task 3 – Consider two data issues (4 points)

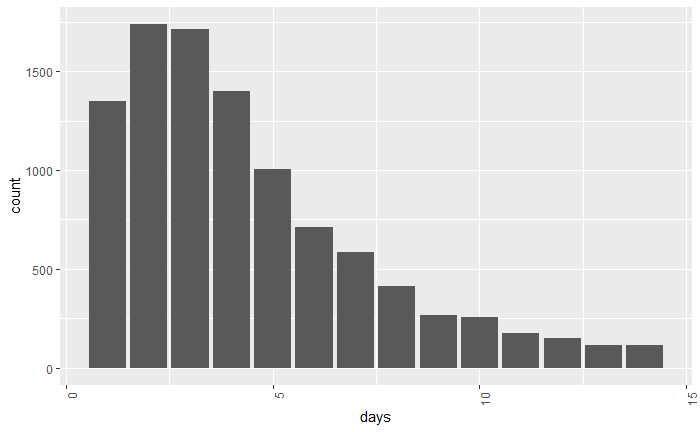
The race variable has potential ethical issues because racial discrimination has been a problem in hospitals in the past.

The client MACH is concerned with the quality of care that patients are receiving. Patients should be getting the best care possible to be as healthy as possible. It may be that including the race variable is useful to this end by making us aware of discrimination practices which may be taking place, and so to include this in the model would help to identify these unfair actions. There may be other problems as well, though. For instance, we did not audit the race data, and it could be that some races have more missing values than others, or that the values were recorded incorrectly for certain races in an unfair manner, and to include this prejudiced variable would introduce prejudice into the model. Finally, it could be race is useful for a medical standpoint and so by including it we can enable MAAC to better care for their patients.

The number of lab procedures seems like it would add additional information to the model, however, we would not have this information until after a patient has stayed at the hospital and so could not use it. If we used the number of labs that were conducted during the current stay, this would leak information from the target variable, the number of days, and would make the results meaningless. Contrast this variable to the num\_procs, which records the number of procedures from the past 12 months. If we could get the number of labs which occurred in the past 12 months then this would be useful.

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

This analysis uses historical records from 10,000 patients who had diabetes who have been readmitted to the hospital. I have performed an exploratory study to resolve issues with the data and to help MAAC to improve patient health. The way that this is being tracked is from the number of days which patients spend in the hospital after being admitted. This is shown below. This length of stay ranges between 0 and 15 days, with an average of about 4 days.



I found three key drivers (or “leading indicators”) of patient length of stay. This info will help the hospital staff to take proactive measures for patients who are currently in the hospital which will help them to recover more quickly and to return to normal life sooner. Three of these are 1) the patient’s age, 2) the number of procedures which they have had in the prior 12 months, and 3) their history of being readmitted.

Patients who are older need to spend longer in the hospital. The table below shows that as the patient’s age increases the number of days that they spend in the hospital also increases. Patients who are older take longer to heal after having a medical procedure and so extra care should be given to elderly patients. Patients who are younger can recover on their own at home and can be sent out of the hospital sooner, thus freeing up time and resources to care for the other patients.

|  |  |
| --- | --- |
| **Age** | **Average Number of Days Spent in Hospital** |
| [0-10) | 3.2 |
| [10-20) | 3.1 |
| [20-30) | 3.5 |
| [30-40) | 3.8 |
| [40-50) | 4.0 |
| [50-60) | 4.1 |
| [60-70) | 4.4 |
| [80-90) | 4.8 |
| [70-80) | 4.7 |
| [90-100) | 4.7 |

Patients who had been readmitted in the prior 30 days had a higher length of stay on average. This was 4.2 days for patients who had never been readmitted but 4.7 days for patients who had been readmitted in the prior 30 days. These patients have underlying issues or more serious cases of diabetes and so by considering their history medical staff and be extra careful when helping these patients to recover.

The data that was provided had issues because there were incorrect and blank values. I had to manually fix these using my best judgement. When there were only a few problems with the record, I just omitted it. When I could, I made changes to fix it.

Two questions which I considered were 1) Are there ethical issues with using a patient’s race in this analysis, and 2) can we attain the number of laboratory procedures for patients? I concluded that adding in the race variable would be unethical if the data was collected in a discriminatory fashion, but also that that it could help doctors to provide better care. For this analysis we chose to include it. Regarding the lab procedures, I advise against including this because MAAC would not have this info in advance.

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

Principal component analysis (PCA) is a dimensionality reduction method which attempts to maintain the information in the data while using fewer variables. It breaks down linearly related (or correlated) variables into principal components, which are linear combinations of the original variables which are uncorrelated. It allows us to use only a subset of these PCs based on the percentage of variation which each explains. First, scaling is applied which subtracts the mean and divides by the standard deviation. This helps to ensure that each variable is given the same amount of weight. Otherwise, variables which have the highest numeric value would have too much influence. Then, each variable is rotated, or multiplied by a scalar, known as the loading. We can use this info to create a “recipe” for each PC, which helps us to interpret it.

**Advantages**

* By reducing the number variables needed, we can use only 2 or three principal components instead of using num\_procs, num\_meds, num\_ip, and num\_diags while still capturing most of the patterns that patients show
* PCA can helps to us to identify groups of patients which have similar characteristics. For instance, num\_procs and num\_meds may be correlated and using a single PC would tell us that often patients which have a high number of process also have a high number of medications. We would find this by looking at the loading factors of the PCS.

**Disadvantages**

* Using the principal component will be less interpretable than using the original variables. Because this business case is asking for insights into patient’s length of stay, this is a big disadvantage

The results of the PCA are below. Each PC explains a percentage of the total variation. If these variables were independent, that is, their correlation was 0, then the cumulative proportion would be 25% for each PC. The fact that the first two PCs only explain 65% of the variation, up from 50% if they were independent, means that this PCA analysis is not able to simplify the data.

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

Each PC is a linear combination of the original variables. A recipe for creating the first PC is 0.56\*num\_procs + 0.68\*num\_meds + 0.12\*num\_ip + 0.47\*num\_diags. This tells us that the largest group of patients have many diagnosis, medications, inpatient visits, and procedures. We know that this is the largest group because it’s the first principal component. You could think of this as a proxy for the patient’s health status: those who are healthy have a low score and those who are less healthy have a higher score.

The second PC has a positive sign the number of inpatient visits and diagnosis, but a negative sign on the number of procedures and medications. This group is patients who may have had an injury or other accident which resulted in a trip to the hospital for their diabetes, but which did not require medications or procedures. We know we know that this represents fewer patients because it is the second PC and only explains 27% of the variation.

The third and fourth PC only account for 34% of the variation. This represents patients who have procedures, medications, and inpatient visits, but few diagnoses, and patients which procedures, inpatient visits, and diagnosis, but few medications.

PC1 PC2 PC3 PC4

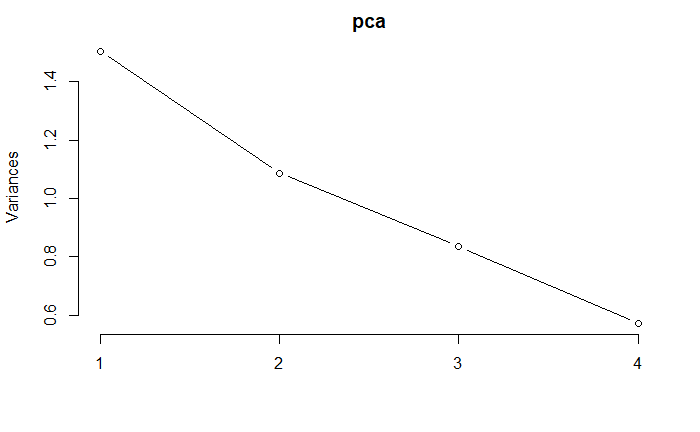
num\_procs 0.5572974 -0.44554566 0.36869207 0.5957977

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

MAAC is more concerned about inference than prediction and so I recommend using the original variables. The first PC only explains 38% of the total variation, which is not very much. If we were to use PCA, I would recommend using at least the first two PCs.



I split the data into training and test sets. This will help me to evaluate the results on new test data and to get an accurate idea of how well it will perform in real life. I compared the mean number of days in the training and test sets and found that it is the same (about 4.4) as in the overall data set.

[1] "TRAIN"

[1] 4.397857

[1] "TEST"

[1] 4.435769

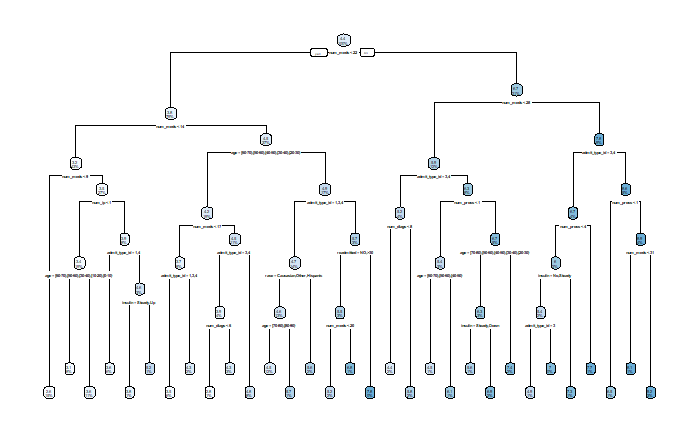
[1] "ALL"

[1] 4.409223

## Task 6 – Construct a decision tree (10 points)

Pruning, also known as cost-complexity pruning, determines a nested sequence of subtrees of the decision tree by recursively snipping off the least important splits, based on the complexity parameter (cp). First, a decision tree that had many leaves is created. Then this is simplified using the pruning algorithm. For each split, the amount by which the model improves on each split, which is measured by the number of days on the right and left branches of the node, are taken into consideration. This is appropriate for this business problem of prediction of length of stay because MAAC is interested in the insights that they can share with hospital managers and having a simpler tree will be more actionable for healthcare providers to use.

The unpruned tree is below.



I use cross-validation to fit trees which have different values of cp and then choose the one which has the lowest error. The lowest error is for a tree that is too complicated for this problem because it has 28 nodes, for a CP value of **0.001223032.**

CP nsplit rel error xerror xstd

1 0.152707521 0 1.0000000 1.0001587 0.02033349

2 0.037698542 1 0.8472925 0.8480251 0.01798971

3 0.019130886 2 0.8095939 0.8107041 0.01729121

4 0.009240202 3 0.7904631 0.7966838 0.01725127

5 0.008676511 4 0.7812228 0.7926694 0.01717010

6 0.005343208 5 0.7725463 0.7839649 0.01699532

7 0.003715982 6 0.7672031 0.7777088 0.01689852

8 0.003568539 7 0.7634871 0.7773847 0.01695395

9 0.003008571 9 0.7563501 0.7762636 0.01695438

10 0.002949588 10 0.7533415 0.7720921 0.01690122

11 0.002883700 11 0.7503919 0.7720237 0.01692143

12 0.002491909 12 0.7475082 0.7707511 0.01692526

13 0.002315013 13 0.7450163 0.7713557 0.01695682

14 0.002125084 14 0.7427013 0.7712409 0.01695252

15 0.001972886 15 0.7405762 0.7716506 0.01695304

16 0.001965305 16 0.7386033 0.7692351 0.01690420

17 0.001753610 17 0.7366380 0.7670100 0.01681699

18 0.001692801 18 0.7348844 0.7666489 0.01679296

19 0.001626374 19 0.7331916 0.7673546 0.01680379

20 0.001603454 20 0.7315652 0.7657792 0.01677920

21 0.001468164 21 0.7299618 0.7633766 0.01671210

22 0.001457705 22 0.7284936 0.7633931 0.01671218

23 0.001455661 23 0.7270359 0.7633931 0.01671218

24 0.001407144 24 0.7255802 0.7634303 0.01672284

25 0.001294707 25 0.7241731 0.7629509 0.01670685

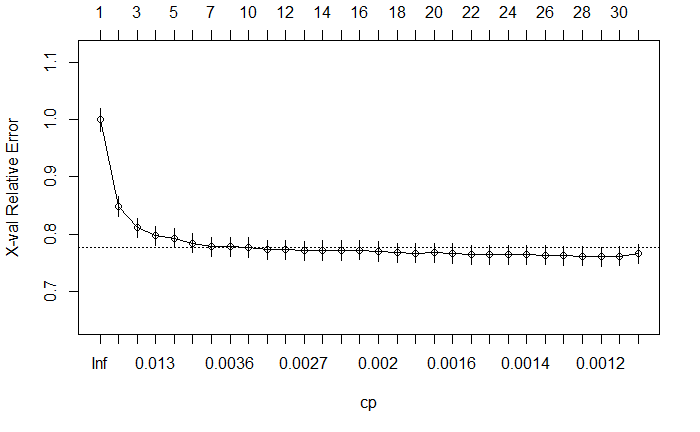
26 0.001273377 26 0.7228784 0.7616959 0.01672131

27 0.001242786 27 0.7216050 0.7611078 0.01670085

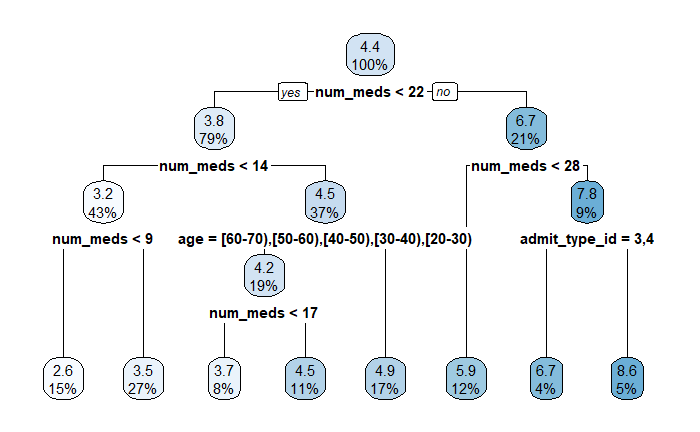
28 **0.001223032 28 0.7203622 0.7603318 0.01669433**

29 0.001002113 29 0.7191392 0.7609538 0.01674067

30 0.001000000 30 0.7181371 0.7651169 0.01687691



Instead of using this minimum CP value, I chose to use 0.003568539 because it has a tree which uses only 7 nodes.



To interpret this tree, I start at the top and move down. I move left if the patient meets the given criteria and I move right if not. After going through these questions, there is a predicted number of days. The predicted number of days spent in the hospital are

* 2.6 days when num\_meds < 9
* 3.5 days when with 9 >= num\_meds < 14
* 3.7 days when 14 <= num\_meds < 17, and age is between 20 and 70
* 4.5 days when 17 >= num\_meds < 22, and age is between 20 and 70
* 4.9 days when num\_meds >= 14, and age is not between 20 and 70
* 5.9 days when num\_meds between 23 and 27
* 6.7 days when num\_meds > 28 and admit\_type\_id is either Elective or Not Available
* 8.6 days when num\_meds > 28 and admit\_type\_id is either Urgent or Emergency

n= 7000

node), split, n, deviance, yval

\* denotes terminal node

1) root 7000 62304.970 4.397857

2) num\_meds< 21.5 5559 36673.060 3.804281

4) num\_meds< 13.5 2985 14840.800 3.200670

8) num\_meds< 8.5 1069 3351.667 2.612722 \*

9) num\_meds>=8.5 1916 10913.420 3.528706 \*

5) num\_meds>=13.5 2574 19483.450 4.504274

10) age=[60-70),[50-60),[40-50),[30-40),[20-30) 1351 9365.500 4.162102

20) num\_meds< 16.5 568 3306.394 3.676056 \*

21) num\_meds>=16.5 783 5827.581 4.514687 \*

11) age=[70-80),[80-90),[90-100),[10-20),[0-10) 1223 9785.045 4.882257 \*

3) num\_meds>=21.5 1441 16117.470 6.687717

6) num\_meds< 27.5 843 8468.833 5.921708 \*

7) num\_meds>=27.5 598 6456.691 7.767559

14) admit\_type\_id=3,4 262 2721.958 6.690840 \*

15) admit\_type\_id=1,2 336 3194.143 8.607143 \*

I used the Pearson Goodness of Fit statistic to evaluate the model. This assumes that the squared error divided by the predicted value has a chi-squared distribution. A higher value of this statistic is worse because the residuals are larger; conversely, a lower value is better because it means that the residuals are smaller. You might ask why I am not using a metric such as R^2 or RMSE or MAE. The reason is because we are modeling a counting value, the number of days spent in the hospital. This gives the data specific properties. For instance, the likelihood that a person spends an additional day in the hospital decreases as the number of days increases. RMSE or MAE does not take this into consideration.

The Pearson Goodness of Fit stat for the training and test sets are below. This is higher (worse) on the test sets than on the training sets, as is always the case. After pruning the test stat is only slightly worse, from 1.56 to 1.60. This is a good value considering that the tree has 7 nodes instead of 28.

**Before pruning:**

Train 1.452799

Test 1.55681

**After pruning:**

Train 1.539727

Test 1.596192

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

The target variable, the number of days, is positive and right skewed. It only takes on discrete values in the data, although it is technically possible that a patient could stay for a fractional number of days, so any response distribution which matches these criteria is possible. The binomial is not useful because the data takes more than two values. The Poisson is the best choice because this models a counting variable. The Gamma is also possible because it is right skewed and positive. In the context of the business problem, it’s important that the predictions always be positive because hospital staff would not be able to make sense of a negative number of days spent in the hospital.

The Goodness of fit stats are calculated below. The GLM which has the principal component instead of the original variables has a higher (worse) test statistic. The principal component would also be more difficult to interpret. This implies that using the first model is better both for performance and for ease of interpretation.

**GLM using original variables instead of principal component:**

Train 1.525184

Test 1.558293

**GLM using the principal component instead of the original variables:**

Train 1.587999

Test 1.618514

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

The features used in the lasso are

genderMale

age[50-60)

age[80-90)

age[90-100)

raceAfricanAmerican

admit\_type\_id2

admit\_type\_id3

readmitted<30

num\_meds

num\_ip

num\_diags

The results from the GLM in task 7 are below. This had a lot of variables but had a lower (better) value of the Pearson goodness of fit stat (1.56) as compared to the Lasso, which had a value of 1.57; however, MAAC is concerned about ease of interpretation rather than performance. The second model, the lasso, removed several variables, which will be easier to explain to medical personnel. I recommend that this model be used. One aspect of this Lasso which may be difficult to explain is the fact that only certain ages are considered: if the patient is between 50-60, 80-90, or 90-100, rather than having a different coefficient for each age.

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.7179703 0.0316229 22.704 < 2e-16 \*\*\*

genderMale -0.0348400 0.0116236 -2.997 0.002723 \*\*

age[60-70) -0.0805106 0.0165719 -4.858 1.18e-06 \*\*\*

age[50-60) -0.1342346 0.0181664 -7.389 1.48e-13 \*\*\*

age[80-90) 0.0648094 0.0174044 3.724 0.000196 \*\*\*

age[40-50) -0.1042567 0.0225005 -4.634 3.59e-06 \*\*\*

age[30-40) -0.0929152 0.0342933 -2.709 0.006740 \*\*

age[90-100) 0.1144602 0.0347736 3.292 0.000996 \*\*\*

age[20-30) -0.1015377 0.0580091 -1.750 0.080054 .

age[10-20) 0.0457339 0.0747208 0.612 0.540495

age[0-10) 0.0302069 0.1679155 0.180 0.857236

raceAfricanAmerican 0.1156573 0.0150857 7.667 1.77e-14 \*\*\*

raceOther 0.0509221 0.0291368 1.748 0.080518 .

raceHispanic 0.0876614 0.0421711 2.079 0.037644 \*

admit\_type\_id2 0.1254435 0.0152688 8.216 < 2e-16 \*\*\*

admit\_type\_id3 -0.0858340 0.0163343 -5.255 1.48e-07 \*\*\*

admit\_type\_id4 -0.0230973 0.0200092 -1.154 0.248363

metforminSteady -0.0156711 0.0152461 -1.028 0.304007

metforminUp 0.1526204 0.0472482 3.230 0.001237 \*\*

metforminDown 0.0674198 0.0716793 0.941 0.346922

insulinSteady -0.0222760 0.0137654 -1.618 0.105608

insulinDown -0.0109830 0.0186137 -0.590 0.555157

insulinUp 0.0237683 0.0188068 1.264 0.206296

readmitted>30 0.0423494 0.0125898 3.364 0.000769 \*\*\*

readmitted<30 0.0793006 0.0189584 4.183 2.88e-05 \*\*\*

num\_procs 0.0112080 0.0036594 3.063 0.002193 \*\*

num\_meds 0.0308537 0.0007074 43.618 < 2e-16 \*\*\*

num\_ip 0.0140475 0.0044002 3.192 0.001411 \*\*

num\_diags 0.0273086 0.0035108 7.779 7.34e-15 \*\*\*

**GLM from Task 7:**

Train 1.525184

Test 1.558293

(Intercept) .

genderMale -0.0005247084

age[60-70) .

age[50-60) -0.0318520898

age[80-90) 0.0609863438

age[40-50) .

age[30-40) .

age[90-100) 0.0271444979

age[20-30) .

age[10-20) .

age[0-10) .

raceAfricanAmerican 0.0408150599

raceOther .

raceHispanic .

admit\_type\_id2 0.0879293247

admit\_type\_id3 -0.0365864924

admit\_type\_id4 .

metforminSteady .

metforminUp .

metforminDown .

insulinSteady .

insulinDown .

insulinUp .

readmitted>30 .

readmitted<30 0.0014827045

num\_procs .

num\_meds 0.0296963221

num\_ip 0.0025137279

num\_diags 0.0223963381

**Lasso from Task 8**

Train 1.541731

Test 1.572154

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

Bias is the difference between the expected value of the model and the expected value of the target. This can be thought of as the “difference from the center of the target”.

Variance is the amount by which the predicted values change when the input data changes. This is just the statistical variance of the predicted values.

The bias-variance tradeoff says that the root mean squared error can be decomposed into three parts

1. The bias squared – Model which have high bias (underfitting) tend to have low variance
2. The variance – Models which have high variance (overfitting) often have low bias
3. Irreducible error – This is random noise which no model can completely remove

The lasso controls the model’s bias and variance using the lambda parameter. This imposes a penalty in the log likelihood so that models the coefficients are either removed or made smaller or larger. Models which have fewer variables have lower variance but higher bias because they are less flexible; conversely, models which have low bias but higher variance have more variables. But looking at many values of lambda and choosing the one which has the lowest error, the GLM can be optimized.

If we did not split into training and test sets, then we would not be able to measure the bias accurately. This is because the model would fit exactly to the data that it was trained on, but if new data was used, then the results would be far worse. By splitting into training and test sets, I can accurately estimate the model error (the Pearson goodness of fit stat in this case), and thus decide on the model which will accurately reflect reality.

## Task 10 – Consider the final model (4 points)

One advantage to using a GLM instead of a tree for this problem is that the predictions will change gradually as the input variables change, because a GLM uses coefficients for each variable instead of yes/no questions in a decision tree, which results in stepwise predictions that can be difficult to explain. Hospital staff may be more comfortable seeing their patient’s projected length of stay change gradually rather than suddenly.

One disadvantage is that the decision tree handles missing values automatically, which was a problem in our data that I needed to fix manually. The GLM does not handle missing values automatically and so this will cost addition time if the model was to be used. MAAC should consider the cost of hiring an actuary’s time in building this model in production when making this decision.

One advantage to using a Lasso is that it is easier to interpret because it removes variables using the penalty term by setting coefficients equal to zero. This can make the results easier for doctors to understand.

One disadvantage to using the Lasso in R is that the log link function is not supported and so we could only use the identity link function. The actuaries who would be hired to train this model would need to take this into consideration.

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

I reran the GLM from task 7 on the entire data set.

This model can be interpreted using a simple formula. We most with the most common patient and make a prediction for the number of days which they will spend in the hospital. Then depending on their characteristics, this number either increases or decreases.

The most common patient is female between the ages of 70-80, Caucasian, has no metformin, insulin, or history of being readmitted, and has been admitted for an emergency. This patient has a predicted length of stay of 2.2 days. Then, for men, we decrease this prediction by multiplying by 0.97. The other pieces of information about the patient produce similar changes. The coefficients below which have a negative sign on the estimate decrease this prediction and those with a positive sign increase it. Then we multiply by 1.012 to the nth power, where n is the number of procedures which the patient has had in the past 12 months. There are similar relationships for the number of procedures, number of inpatient visits, and number of diagnoses.

Call:

glm(formula = days ~ . - PC1, family = poisson(link = "log"),

data = data.all)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.8113 -0.9898 -0.2784 0.5766 4.8714

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.7978752 0.1383465 5.767 8.06e-09 \*\*\*

genderMale -0.0280296 0.0103461 -2.709 0.006744 \*\*

age[10-20) -0.0983412 0.1542494 -0.638 0.523769

age[20-30) -0.2302232 0.1454871 -1.582 0.113551

age[30-40) -0.2458890 0.1407981 -1.746 0.080743 .

age[40-50) -0.2329239 0.1390500 -1.675 0.093913 .

age[50-60) -0.2670869 0.1386212 -1.927 0.054012 .

age[60-70) -0.2123259 0.1384994 -1.533 0.125265

age[70-80) -0.1467226 0.1384676 -1.060 0.289319

age[80-90) -0.0794199 0.1385948 -0.573 0.566620

age[90-100) -0.0146799 0.1410411 -0.104 0.917104

raceAfricanAmerican 0.0901276 0.0130461 6.908 4.90e-12 \*\*\*

raceHispanic 0.0710824 0.0387262 1.836 0.066430 .

raceOther 0.0283174 0.0449429 0.630 0.528646

raceAsian 0.0877105 0.0637325 1.376 0.168751

admit\_type\_id2 0.1034508 0.0128909 8.025 1.01e-15 \*\*\*

admit\_type\_id3 -0.0962379 0.0139251 -6.911 4.81e-12 \*\*\*

metforminSteady -0.0099422 0.0135395 -0.734 0.462760

metforminUp 0.1625055 0.0432158 3.760 0.000170 \*\*\*

metforminDown 0.1319025 0.0651327 2.025 0.042854 \*

insulinSteady -0.0280576 0.0123218 -2.277 0.022782 \*

insulinDown -0.0065038 0.0163592 -0.398 0.690952

insulinUp 0.0266025 0.0165736 1.605 0.108469

readmitted>30 0.0336618 0.0112498 2.992 0.002770 \*\*

readmitted<30 0.0636384 0.0166012 3.833 0.000126 \*\*\*

num\_procs 0.0120259 0.0032266 3.727 0.000194 \*\*\*

num\_meds 0.0310800 0.0006329 49.109 < 2e-16 \*\*\*

num\_ip 0.0134967 0.0039234 3.440 0.000582 \*\*\*

num\_diags 0.0361976 0.0031564 11.468 < 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: 16548 on 8770 degrees of freedom

Residual deviance: 12335 on 8742 degrees of freedom

AIC: 40052

Number of Fisher Scoring iterations: 5

## Task 12 – Executive summary (20 points)

Our client, Merged and Acquired Clinics and Hospitals (MACH), has hired us to help their hospital executives gain a better understanding of the factors that drive inpatient length of stay. We used predictive analytics to identify the reasons why some patients are sent back home quickly, and other patients need to spend several days in the hospital. We used historical inpatient encounters for patients with diabetes for U.S. hospital between the years of 1999 and 2008. Any conclusions of this report are limited to the population of this data set and would change if applied to a different population.

Each encounter includes the length of stay in hospital, measured in days as well as their gender, age, race, weight, whether they were admitted for an emergency, elective, or non-elective reason, any changes to their metformin or insulin prescriptions, their medical history in the last 12 months such as the number of readmissions in the past 30 days, number of procedures, medications, inpatient visits, and diagnosis.

We can provide you with a table that can be used by your healthcare staff to predict how long a patient will spend in the hospital. This uses a simple spreadsheet, and based on the patient’s characteristics, adjusts their length of stay. You could use this either to predict which patients are on course to have a long stay and then to intervene, or you could just interpret these results directly. You can see below that different factors either increase or decrease the length of stay. We start with a predicted of 2.2 days. This represents the most common patient. Then, if they are male, multiply the 2.2 by 0.97. If they are between the ages of 10 and 20, decrease the estimate by multiplying by 0.91. If they are between ages 20 and 30, decrease again by multiplying by 0.79. Each of these variables is interpreted in the same way. For each procedure that the patient has multiply by 1.07.

|  |
| --- |
| **Interpretation** |
| Start with a prediction of 2.2 days |
| If the patient is Male, multiply by 0.97 |
| If the patient is age is between 10 and 20, multiply by 0.91 |
| If the patient is age is between 20 and 30, multiply by 0.79 |
| If the patient is age is between 30 and 40, multiply by 0.78 |
| If the patient is age is between 40 and 50, multiply by 0.79 |
| If the patient is age is between 5 and 60, multiply by 0.77 |
| If the patient is age is between 60 and 70, multiply by 0.81 |
| If the patient is age is between 70 and 80, multiply by 0.86 |
| If the patient is age is between 80 and 90, multiply by 0.92 |
| If the patient is age is between 90 and 100, multiply by 0.99 |
| If the patient's race is African American, multiply by 1.09 |
| If the patient's race is Hispanic, multiply by 1.07 |
| If the patient's race is Other, multiply by 1.03 |
| If the patient's race is Asian, multiply by 1.09 |
| If the patient's readmission was Urgent, multiply by 1.11 |
| If the patient's readmission was Elective, multiply by 0.91 |
| If the patient's Metformin was Steady, multiply by 0.99 |
| If the patient's Metformin was Up, multiply by 1.18 |
| If the patient's Metformin was Down, multiply by 1.14 |
| If the patient's Insulin was Steady, multiply by 0.97 |
| If the patient's Insulin was Down, multiply by 0.99 |
| If the patient's Insulin was Up, multiply by 1.03 |
| If the patient had not been readmitted in the last 30 days, multiply by 1.03 |
| If the patient had been readmitted in the last 30 days, multiply by 1.07 |
| For each procedure that the patient has had, multiply by 1.01 |
| For each medication that the patient has had, multiply by 1.03 |
| For each inpatient visit that the patient has had, multiply by 1.01 |
| For each diagnosis that the patient has had, multiply by 1.04 |

To ensure that the quality of this analysis is the best that it can be, we performed integrity checks of the data prior to beginning the analysis. This involved correcting errors in the data and removed values which were incomplete. There were a few patients who had missing records for their gender, and these were removed. The patient’s race was not included for 226 cases, and so because this is too many to ignore, we included them as a separate group. We were supplied with the patient’s weight but did not use this because it was not included for most patient records. This is a serious problem with the data which MAAC should examine.

We used visual and statistical methods to look for patterns. We found that most patients who are readmitted spend about 4 days in the hospital. We looked at the other patient info to see if there was a clear difference in the type of patient who had a long stay in the hospital as compared to a patient who had a short stay.

We found that patients who had a recent hospital stay, within the last 30 days, had a longer stay on average than other patients. There were 1,102 patients in this group, which is a significant number. It may be worth considering if these patients have something different about them, such as other chronic illnesses, and including this info within the analysis. Perhaps your medical team would have more insight here that could help.

|  |  |  |
| --- | --- | --- |
|  | **Average Days in Hospital** | **Number of Patients** |
| No readmission history | 4.2 | 5,370 |
| History of readmission | 4.6 | 3,525 |
| Recent readmission | 4.8 | 1,102 |

There are ethical concerns with using the patient’s race in the model. It may be the case that this info helps healthcare providers to prevent discrimination from occurring, or it could be that this info creates a bias in the model which gives people different levels of quality of care depending on their race, which would be discriminatory. This has legal risks because of potential lawsuits.

We concluded that the other info in the data was complete and informative for the patient’s length of stay. We recommend against using the number of laboratory procedures, which my assistant mentioned, because this would not be possible to collect from patients in advance of knowing their hospital stay.

We sent our actuarial manager a summary of all the data steps. You can review this information with them to follow up on any of the above steps. This summary contains our process of cleaning the data and would be useful for repeating this analysis. There are also three findings here related to what factors predict a patient’s length of stay. These were 1) their age, 2) the number of procedures that they’ve had in the prior 12 months, and 3) their history of being readmitted.

One way to find which factors are related to the length of the hospital stay is by looking at the correlations. When two things are correlated, it means that one increase as the other increases and vice versa for decreasing. When the number of medications which a patient had increased, their length of stay did as well. It is important to remember that correlation does not imply causation. It could be, for instance, that patients who have more medications have other underlying health issues which are causing their increased length of stay. We recommend using the results from our predictive model, which includes these other factors and adjusts for them.

You can simplify the info that you collect on a patient’s history down into a single number. This would give you the option of using a single score that would consider the patient’s history instead of using the separate columns in the data for these values. The advantage would be that the set of rules would be simpler to implement but more difficult for your hospital staff to understand. The recipe for this is

(0.56)(Number of Procedures) + (0.67)(Number of Medications) + (0.12)(Number of Inpatient Visits) + (0.47)(Number of Diagnosis)

We investigated several alternate model approaches so that the results are as reliable as possible. We created experiments with each of these methods and chose the one which has the best match of the actual patient readmission patterns. In other words, we are only showing you the best results from our modeling. We also looked at decision trees and penalized regression models.

You can be confident that these results will work in real life because they have been tested using a scientific approach known as training-testing validation. This used 70% of the patients as a training set and the remaining 30% were held out as a blind test set. We evaluated each of the models based on this test set and ended up selected a generalized linear model (GLM) because it had the best result.

We interpreted this model into the table of rules which were included earlier.

MAAC may be interested in building a model which has better predictive power. We considered this. You would need to look at the tradeoff between interpretability, which is how explainable the results would be to your hospital staff, versus the predictive power. We chose a model which is easy to interpret; however, you could consider more powerful models as well. There are other advantages and disadvantages to using a different model. For example, some of the data cleaning work that we needed to do could be automated if using a tree-based model.

Your next steps should consider several factors. You may wish to follow up with your hospital staff regarding these issues:

* How may these results change for non-diabetes patients?
* What additional data could be collected on these patients? Would the type of prescription, for example, be useful to know so that doctors could see if certain medications are causing longer hospital stays?
* Could the more recent data be collected? This study was based from 1999-2008, but medical records have changed in the last 12 years and so this study may be out of date.

In conclusion, we identified the factors which determine how long a patient will spend in the hospital after being readmitted. We present these results to you so that you can take proactive action in caring for your patients in the best way possible.