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)

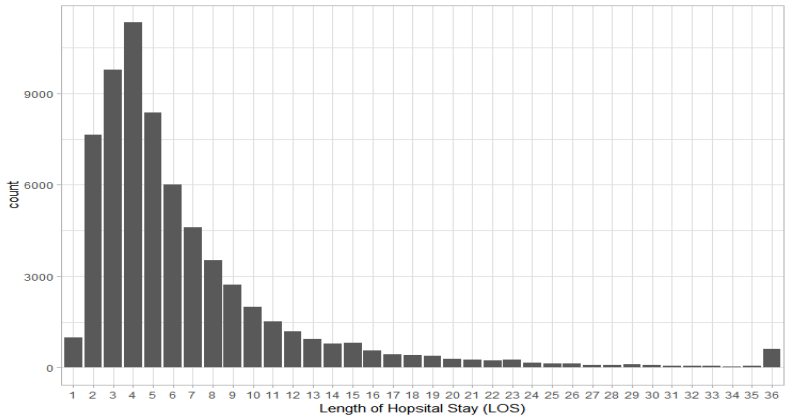
There are eight predictor variables which are related to whether or not a patient has a readmission. Each observation is a patient's status of readmission as well as basic demographic info and hospital records. There are about 70,000 patients.

The ER field tells if a patient has been to the emergency room previously. Most patients have not, but this is their first visit. Because this is only counting ER visits, it is possible that a patient has been to the hospital before for other reasons but not the ER.

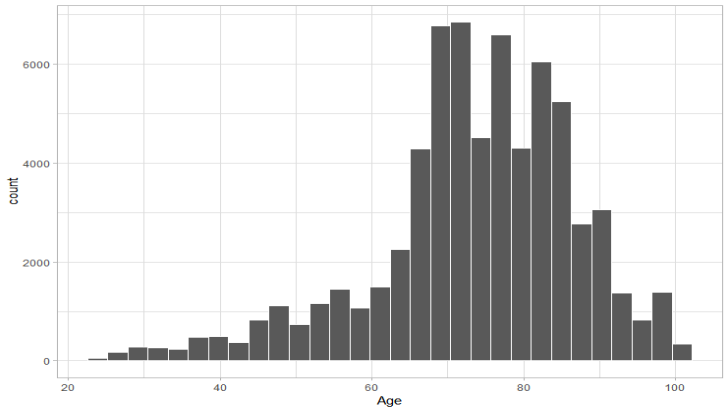
A separate question is a patient would need to go the ER so many times. The ER is not intended to be used for routine visits and this leads to inefficient use of resources – taking time from patients who are there for medical emergencies. The Hospital could investigate these patients with > 4 ER visits.

ER <dbl>	**n** <int>
0	43086
1	16280
2	5286
3	1572
4	438
5	105
6	10
7	3
9	2

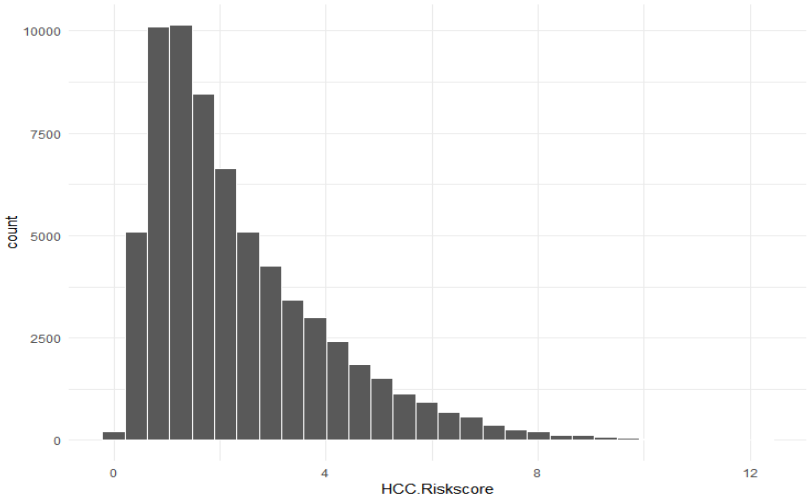
The length of the hospital stay in days is recorded. Most patients spend less than 7 days in the hospital. There are 622 patients that are recorded as staying for exactly 36 days. These are likely patients who have stayed *at least* 36 days. Because this is right-skewed, a log transform was applied.



These data are for patients who are older, with the age of patients is from 24 to 101 years of age with the median at 75.



The field HCC.Riskscore indicates the patient's overall health status. Higher values indicate greater risk for healthcare costs. Because this is right-skewed, a log transform was applied.



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

The DRG class (Diagnostic Related Group) is a treatment category. This is either medical, surgical, or other. Most patients are medical. The DRG complication field refers to a complication, when something went wrong with the procedure, or a comorbidity, whether or not the patient had a related illness.

If a patient was treated for a medical DRG class, then they can only have a medical complication. Similarly, if a patient was in surgery then they can only have a complication that is surgical. This means that we expect to see an interaction between the two variables because the level of one is related to the level of the other.

The most uncertainty is for those with a DRG complication of "UNGROUP" and a DRG class of "other". We should consult with the Hospital to see if there is anything special about these patients which should be taken into consideration when using these variables.

DRG.Complication	MED	SURG	UNGROUP
MedicalMCC.CC	18104	6	*NA*
MedicalNoC	12310	*NA*	*NA*
Other	5357	3424	564
SurgMCC.CC	*NA*	15468	*NA*
SurgNoC	*NA*	11549	*NA*

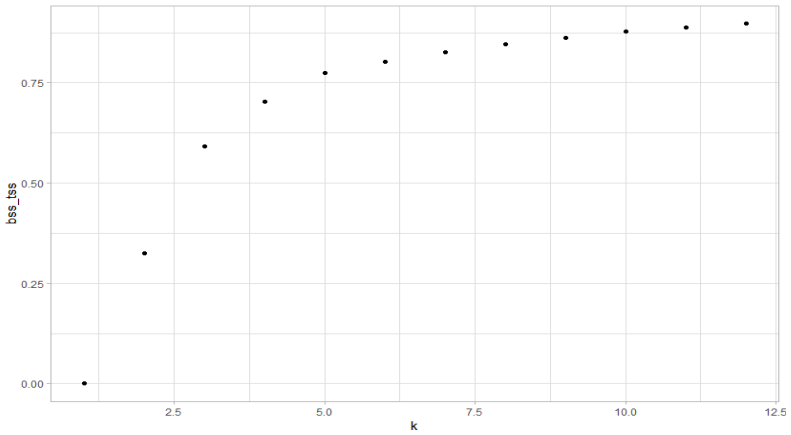
Task 3 – Use observations from cluster analysis to consider a new feature (9 points)

The data given by `LOS and Age` are clustered by the *k*-means method, which aims to partition the patients into *k* groups such that the sum of squares from points to the assigned cluster centers is minimized. K-means assumes that inputs are scaled prior to running the algorithm, which was done.

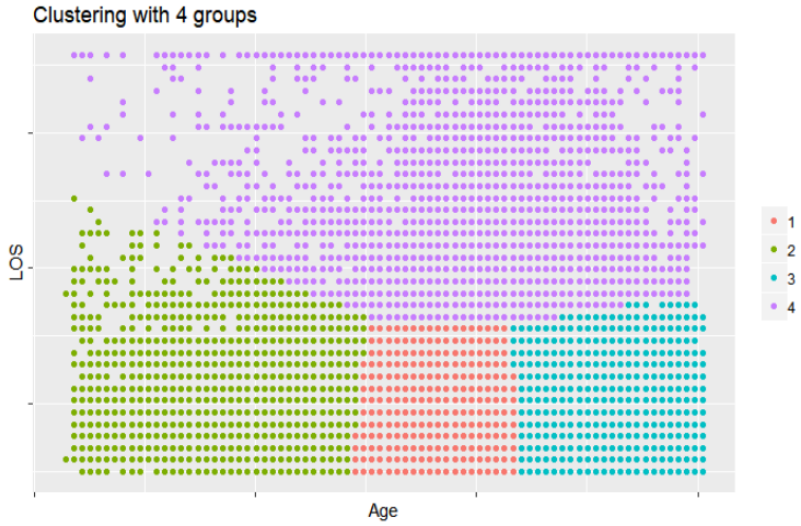
The code which your assistant had set up was looking at the ratio of (total sum of squares – total within cluster sum of squares)/total sum of squares. When k = 1, the total within cluster sum of squares within each cluster is the same as the total sum of squares, and so the ratio is 0. If k = n, then the within cluster sum of squares would be 0 and so the ratio is 1. The idea is to find the optimal value of k which strikes a balance between having clusters that are two small and accurately stratifying the risk of each group of patients.

Finally, to insure that these results are reproducible, I increased the number of times that the algorithm is run to 10 with up to 30 iterations be run. This is the number of times that the algorithm is run. Setting it to 30 means that 30 kmeans centers were found and then the average was taken across each dimension to get the final centers.

The graph below shows the ratio. I chose to go with a value of k = 4.



The graph below shows the different patients in each cluster



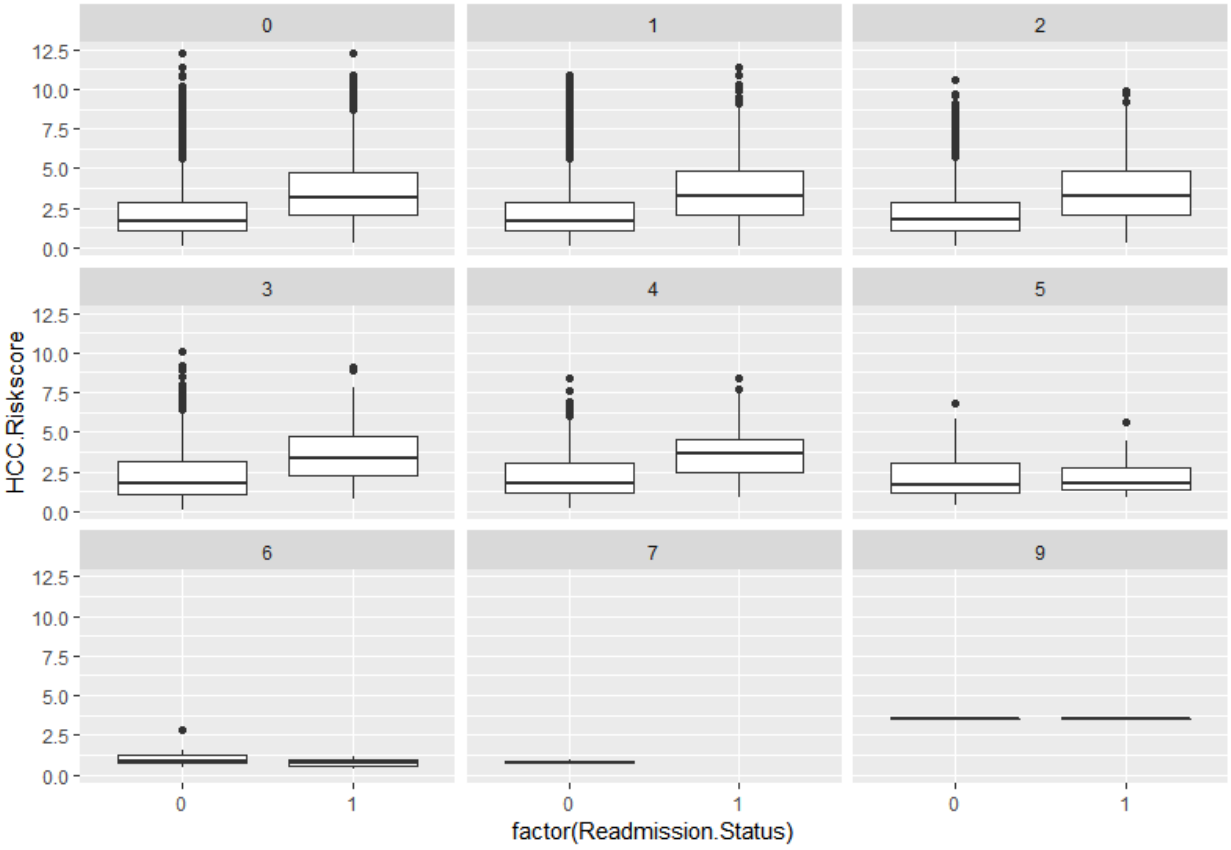
Task 4 –Select an interaction (5 points)

Your assistant thought of including an interaction between race and gender. I did not see significant evidence to suggest that the number of readmissions changed significantly across gender and racial status. Additionally, race is generally a protected class and should not be used in modeling. There are potential ethnical issues with using race in a model, and so this is being excluded from this analysis. Could it be that certain races are being unfairly treated, discriminated against, in the hospital? Could certain races be given preferential treatment? This is a question for the legal department.

The table below shows that there is an interaction between ER visits and risk score. For patients with a high risk score, the effect of ER visits on the likelihood of readmission is different than those with a low risk status. For example, a patient with 5 ER visits that is not readmitted has a median risk score of 1.7 whereas a patient with 4 ER visits has a risk of 1.8. For a patient that *has been* readmitted, these values change to 1.8 and 3.7 respectively.

ER Visits	Not Readmitted Median Risk	Readmitted Median Risk
0	1.7	3.2
1	1.7	3.3
2	1.7	3.3
3	1.8	3.3
4	1.8	3.7
5	1.7	1.8
6	0.9	0.8
7	0.8	*NA*
9	3.5	3.5

Another way of seeing this effect is in the graph below. For with 0 ER visits (upper left), there is a large difference in the median risk score by readmission status. For patients with 5 ER visits, this difference is much smaller.



Task 5 – Select a link function (8 points)

The model objective was to predict the readmission status. A 75% train-test set was used with an equal percentage of readmitted patients in each (13%).

A patient can either be readmitted or not, which means that the response is binary. This suggests that a binomial response distribution is appropriate for modeling.

Because the response is only 0 and 1, a typical regression model will not suffice. The link function transformed the linear predictor from a number into a probability between 0 and 1. The two link fuctions tested were the logit (log(p/(1-p)), and the probit (phi(p) where phi is the CDF of a normal distribution).

As the table below shows, there was no material difference in the test AUC between these link functions and so the logit link was chosen.

Model	Features	Link	Response	test AUC
1	All features except for Race	logit	binomial	0.7442
2	All features except for Race	probit	binomial	0.7428

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

I also tested the above configurations with the Crouchit link (AUC 0.7425), log (did not converge), (AUC 0.7428). Based on this, the logit was best.

Next, I tested if the cluster feature was adding lift to the model. To do this, I a new model excluding the cluster feature (model 3). Compared with model 1, which included this feature, we see that the AUC has decreased from 0.7442 to 0.7438.

Model	Features	Link	Response	test AUC
3	All features except for Race and cluster	logit	binomial	0.7438

This implies that the cluster feature should be included. Additionally, by inspecting the p-values form model 1 we can see that 2 of the 3 clusters have significant p-values less than 0.001 which means that we can reject the null hypothesis that they were generated by random chance. That is, the probability of getting coefficients of this magnitude for the cluster feature if these were random noise is less than 0.001 for two of the three clusters.

term	p.value
los_age_clust2	2.718589e-01
los_age_clust3	1.692850e-03
los_age_clust4	5.611272e-05

Task 7 – Select features (15 points)

The goal of the model is to identify actionable insights which the Hospital can use in order to reduce the number of future readmissions. This means that a black-box model would not be useful. The features used need to be easy to understand, be based on easily-accessible data, and not violate ethical or regulatory issues.

To this end, the Race variable was discarded based on this last point. Because of interpretability issues, the cluster feature was not used. An indicator feature was created for members who had a hospital stay of over 36 days. This is because these patients actually represent patients who had a stay *of at least* 36 days.

The final features selected are below. The AUC for this model on the test set was 0.7437.

Feature	Coefficient
(Intercept)	-2.685
Age	-0.006
GenderM	0.028
log(LOS)	0.043
log(HCC.Riskscore)	1.556
DRG.ComplicationMedicalNoC	-0.088
DRG.ComplicationOther	0.108
DRG.ComplicationSurgMCC.CC	0.038
DRG.ComplicationSurgNoC	0.042
DRG.ClassSURG	-0.073
DRG.ClassUNGROUP	-0.128
ER	0.062
HCC.Riskscore	-0.079
Indicator if LOS = 36	0.140
ER:HCC.Riskscore	-0.015

Task 8 – Interpret the model (6 points)

The final model has an AUC of 0.7442 which is comparable to the LACE index of 0.75.

Note: I would have run this model on the entire data set, but after going through these tedious interpretations on the test set I read that I should use the entire data set. As a point, interpretation based on the set that it was trained on, if using the full data set, would likely be overfitting. Because the model will be tested on unseen data this provides a reasonable interpretation. Given that there are 70,000 observations the testing set is large enough to not have significant testing variance.

The above coefficients are on the scale of the linear predictor after applying the logit transform. This means that the interpretation applies to the log-odds of the probability of being readmitted. By taking the exponent of the coefficients we convert from the log-odds to the odds.

The odds of being readmitted is the probability of being readmitted divided by the probability of not being readmitted. This means that the p, the probability, is close to 0.5, then the log odds is 1.

- If a person is male, then the log-odds of being readmitted increases by 2.9%.
- For each additional day of time in the hospital (Length Of Stay), the log-odds of being readmitted increases by a factor of 1.044.

Feature	Exponent of Coefficient
(Intercept)	0.068
Age	0.994
GenderM	1.029
log(LOS)	1.044
log(HCC.Riskscore)	4.741
DRG.ComplicationMedicalNoC	0.916
DRG.ComplicationOther	1.114
DRG.ComplicationSurgMCC.CC	1.039
DRG.ComplicationSurgNoC	1.042
DRG.ClassSURG	0.929
DRG.ClassUNGROUP	0.880
ER	1.064
HCC.Riskscore	0.924
Indicator if LOS = 36	1.151
ER:HCC.Riskscore	0.985

As you can see, the interpretation here is not very clear. An easier way to interpret the model is to look at example cases. If I had more time, I would create new rows in the data with different feature values and then look at the model average across all input values in order to create partial dependence (aka, marginal effects) plots.

Here are three examples.

The first is a male of age 20. The second is a female of age 70. The third is a female of age 20. Notice that the other variables stay consistent. On the right is the model's prediction that this patient would be readmitted to the hospital. This shows that men or women over the age of 70, with an average risk score of 5, who have spent between 10-4 days in the hospital, are less likely to be readmitted than another patient with similar characteristics who is around age 20.

Gender	ER	DRG.Class	LOS	Age	HCC.Riskscore	DRG.Complication	ReadmissionProbability
M	0	MED	4	20	5	MedicalMCC.CC	0.35
F	0	SURG	10	70	5	MedicalMCC.CC	0.28
F	0	SURG	10	20	5	MedicalMCC.CC	0.34
F	0	SURG	12	20	10	SurgMCC.CC	0.52
M	0	SURG	12	20	10	SurgMCC.CC	0.52
F	0	SURG	12	90	10	SurgMCC.CC	0.41

The most likely patients to be readmitted tend to

- Have a higher risk score
- Have had a surgical complication
- Have had a surgical procedure
- Have spent more time in the hospital
- Are younger

In order to interpret the relative importance of each input I fit a separate model with scaled coefficients (all values between 0 and 1) and then compared the size of the model coefficients. This used the same features as the model above. The ranking of these is below in order of importance.

The relative risk score is the most important factor followed also by age. One key question is whether or not the risk score takes age into consideration already.

Input	**Importance Rank**
Relative risk score	1
Age	2
DRG.ComplicationOther	3
LOS	4
ER	5
ER interaction with RiskScore	6
Surgical	7
DRG.ClassUNGROUP	8
DRG.ComplicationMedicalNoC	9
DRG.ComplicationSurgMCC.CC	10
DRG.ComplicationSurgNoC	11
GenderM	12
Have spent > 36 days in hospital	13

Task 9 – Set the cutoff (9 points)

Because the goal is to reduce patients who are readmitted we want the sensitivity (true positive rate) to be high. This is the percentage of patients that are readmitted that are predicted to be. The cutoff was set to 0.30 which resulted in a sensitivity of 95%.

For the intervention program, consider this scenario where there are 100 patients. About 13 (12.5%) of these patients will be readmissions. The detection rate is the percentage of patients who are predicted to be readmitted by the model.

Number of Patients	100
Detection Rate	8%
Sensitivity	95%

With a cost of $25 per readmitted patient, the total costs would be 13*$25 = $325.

When using the model which has a 95% sensitivity the total cost would be $39.00. Here is a summary of the calculations.

	Patients Readmitted	Interventions	Not Readmitted	Total Cost
No Intervention	13	0	87	$ 325.00
Intervention for all patients	13	13	0	$ 200.00
Strategic Intervention	13	12.41	0.59	$ 39.60

As you can see, using the model leads to a cost savings of almost $300.

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

There are other model types which could lead to better results and cost savings.

• LASSO regularized regression

Pros

1) Simplifies coefficients by using L1 penalization to perform variable selection
2) Allows for more features to be tested which could improve performance

Cons

1) Slightly more complex to implement
2) Does not handle outliers
3) Does not handle non-linearities
4) Interaction terms need to be hand-coded
5) Strict assumptions (residuals uncorrelated, error terms follow certain distribution, etc)

• Classification tree using cost-complexity pruning

Pros

1) Minimizes cost function directly which could lead to more dollar savings
2) Simple to interpret
3) Detects interactions effects
4) Detects nonlinearities
5) Handles missing values

Cons

1) Easily overfits
2) Higher variance than linear model which means that results can change upon retraining the model
3) Weakly predictive since each observation prediction is an average of other observations in that tree's terminal node

• Random forest

Pros

1) Detects interaction effects

2) Handles non-linearities
3) Handles missing data
4) Better predictive power than using a single tree
5) Results are more stable over retraining due to boosting over many trees reducing the overall variance
6) Can be interpreted through partial-dependence plots and feature importance
7) Can optimize over cost function directly

Cons

1) Less interpretable than other models
2) Requires more data. 70,000 records is probably sufficient but if deploying in other smaller hospitals this would be an issue
3) Longer time to train
4)

Task 11 – Executive summary (20 points)

The Centers for Medicare and Medicaid Services (CMS) financial penalties for hospitals which re-admit patients after an inpatient stay. In order to decrease their costs, hospitals need to reduce the amount of penalty fees which they pay. In a perfect world, they would just reduce the number of patient readmissions, but this is easier said than done as there are many complex health factors which determine when someone needs to be readmitted.

A solution is early intervention for patients who are likely to be readmitted. If we knew which patients were likely to be readmitted, we could have them meet 1-on-1 with a care provider, have their doctor follow-up with a phone call a week after they have left the hospital, or receive supplemental care. This analysis uses predictive analytics in order to predict which patients are likely to be readmitted and thus allow for this strategic intervention.

Patients at high risk for readmission tend to

Have a higher risk score

- Have had a surgical complication
- Have had a surgical procedure
- Have spent more time in the hospital
- Are younger

There is an interaction effect between the patient's risk score and the number of ER visits. This means that as a patient's risk score increases their likelihood of being readmitted changes depending on the number of times that they have been to the ER previously.

The patient's gender had little impact. The relative importance of who will be readmitted is below. Hospitals should ask if their patients who are about to leave meet any of these criteria. The most surprising finding below is that the risk of readmission changes substantially when a patient is in an unknown DRG class. This is to say that they DRG complication is "Other" and their DRG class is

"UNGROUP". Perhaps your physicians could shed some light on to why this would be the case. It was not clear from the data which type of patient these represented.

Unhealthy patients are more likely to be readmitted. This is seen below from the Relative Risk Score being the most predictive factor. This could indicate that patients are leaving the hospital before they have fully recovered. Patients who are less healthy may require longer on overage to heal after surgery, for instance. Perhaps these patients should be allowed longer time in the hospital.

Input	**Importance Rank**
Relative risk score	1
Age	2
DRG.ComplicationOther	3
LOS	4
ER	5
ER interaction with RiskScore	6
Surgical	7
DRG.ClassUNGROUP	8
DRG.ComplicationMedicalNoC	9
DRG.ComplicationSurgMCC.CC	10
DRG.ComplicationSurgNoC	11
GenderM	12
Have spent > 36 days in hospital	13

The population sampled was generally older and there is a correlation between age and health status. One question which I had was related to this. Is the relative risk score already accounting for patient age and gender status? I have in the data description that this is a predictor of healthcare costs.

We compared three different scenarios for intervention and found that using this predictive model would save 88% of revenue relative to not using an intervention at all. If a $2 intervention was applied to all patients, where any patient who had intervention would not be readmitted, the cost savings would only by 38%.

The table below summarizes what the costs would be for 100 patients.

	Patients Readmitted	**Interventions**	**Not Readmitted**	**Total Cost**
No Intervention	13	0	87	$ 325.00
Intervention for all patients	13	13	0	$ 200.00
Strategic Intervention	13	12.41	0.59	$ 39.60

This strategic intervention could be implemented within your database in order to triage those patients who are about to leave the hospital and allow for strategic intervention. This model is statistically better than LACE (AUC of 0.70 vs. AUC of 0.74), which means that it will save more money.

This model is based on a sample of 70,000 patients. There were eight variables that are related to whether a patient is a readmission. These were the patient's gender, length of hospital stay, relative risk

score, DRG class, whether they have had a cormorbidity or complication, and other factors which the hospital has readily available for each patient.

I picked up the process where my assistant had left off. I first inspected the data for any errors and made a few transformations. I created new features for interaction effects between risk score and ER visits. These were to indicate if a patient had stayed in the hospital for more than 36 days, as well as a log-transform of the LOS and ER variables.

I noticed that there were a large number of patients with many ER visits. Is there anything special about these patients which should be taken into consideration? I found it surprising that a patient would have a non-scheduled appointment as this will generally be more expensive than a scheduled one.

Most patient's were older, with ages from 24 – 101 and the median at 75. These results will likely not apply to pediatric (patients under the age of 18) as there were no records of these patients in the data.

These results depend on the patient's risk score, which I understand is an estimated quantity. To the extent that this would change the results of this model need to be revised. If a patient suddenly went from having a high risk score to a lower risk score then this would not be reflected in this model.

Next I fit several models and compared them based on performance. There were about 10 different models tested to insure that the most dollar savings were be realized. Once a final model was chosen based on statistical measures I re-optimized it to maximize cost savings based on the $2/patient intervention initiative. If the costs of patient intervention change then this section of the analysis should be revisited to insure that the model is set up for max economic value.