Final Report

Lindsay Knupp

2024-05-14

Dataset

Our data was collected from Centers for Medicare & Medicaid Services on April 03, 2024. It features information about hospitals through annual cost reports in 2020. We were interested in understanding how certain characteristics of hospitals like location, number of full time equivalent employees, or number of beds, for example, affected the hospitals' total operating costs. To perform qualitative analyses, we categorized each hospital as above or below the median.

While the data is roughly annual, certain hospitals reported for different fiscal year lengths. To normalize, we divided some of our variables by the length of their cost reporting period to obtain daily estimates like average number of inpatients per day or average salary expense per day. Variables that are reported as "averages per day" are denoted with the word "average" in the predictor table below. Some hospitals were listed multiple times with distinct reporting periods. We learned that this could correspond to a change in control of the hospital. For example, the hospital could have been sold and transitioned from a voluntary to a governmental hospital. Duplicate hospitals were left in the dataset and a dummy variable, `duplicate` was added to indicate its status.

There were 13 different categories of control ranging from "Voluntary Non-Profit-Church" to "Governmental-Federal". To reduce our number of categories, we re-binned this variable to only include the broad categories: "Voluntary", "Proprietary", and "Governmental". We followed a similar procedure for the 12 different categories of provider type ranging from "Children" to "Cancer" to "General Long Term". In this case, we re-binned provider type to only distinguish between "General" and "Specialized" care. We classified "General Short Term", "General Long Term", and "Religious Non-Medical Health Care Institution" as "General" care and classified "Cancer", "Psychiatric", "Rehabilitation", "Children", "Reserved for Future Use", "Other", "Extended Neoplastic Disease Care", "Indian Health Services", and "Rural Emergency Hospital" as "Specialized" care.

To improve accuracy on methods like Lasso regression, we scaled our numerical variables using the `scale()` function which centered and scaled our data appropriately. The following table includes the ranges of the predictors and response before they were re-scaled.

Variables	Pre-scaled Range	Descriptions
Predictors		
Number of Beds	[1-2,791]	Total number of available beds including adult beds, pediatric beds, birthing room, and newborn ICU beds
FTE employees on payroll	[0.05-26,941.09]	Average number of full time-equivalent employees

Total hospital days	[1-772,819]	Total number of inpatient days (i.e. days all patients spent in the hospital)
Total discharges	[0.0027-462.63]	Average number of discharges including deaths
Total income	[-$6,129,919, $11,516,626]	Average income including net revenue from services given to patients
Total assets	[-$636,856,458, $29,465,487,958]	Total current assets
Salaries	[$128.51, $9,032,294.85]	Average salary expenses
Inpatients	[0.0033-2123.13]	Average number of inpatients
Rural versus Urban	[2487 rural, 3225 urban]	Location of hospital defined as rural or urban
Type of control	[2927 voluntary, 1728 proprietary, 1057 governmental]	Type of control under which hospital is conducted
Type of provider	[4779 general, 993 specialized]	Type of services provided
Duplicate hospital	[132 duplicates, 5580 non duplicates]	Whether or not hospital was listed multiple times
Response		
Total costs	[$2,718.28, $16,000,980.58]	Total hospital costs
Costs bin	[2856 above median, 2856 below median]	Whether or not total hospital costs was above/below median

Qualitative Outcome Analyses

For all of our qualitative outcomes, we were trying to predict whether a hospital's total costs were above or below the median. All of the methods' error rates were comparable except for LDA which had the highest misclassification rate at about 17%. KNN had no consistent choice of an optimal k across simulations and its variability inspired our simulation study.

KNN

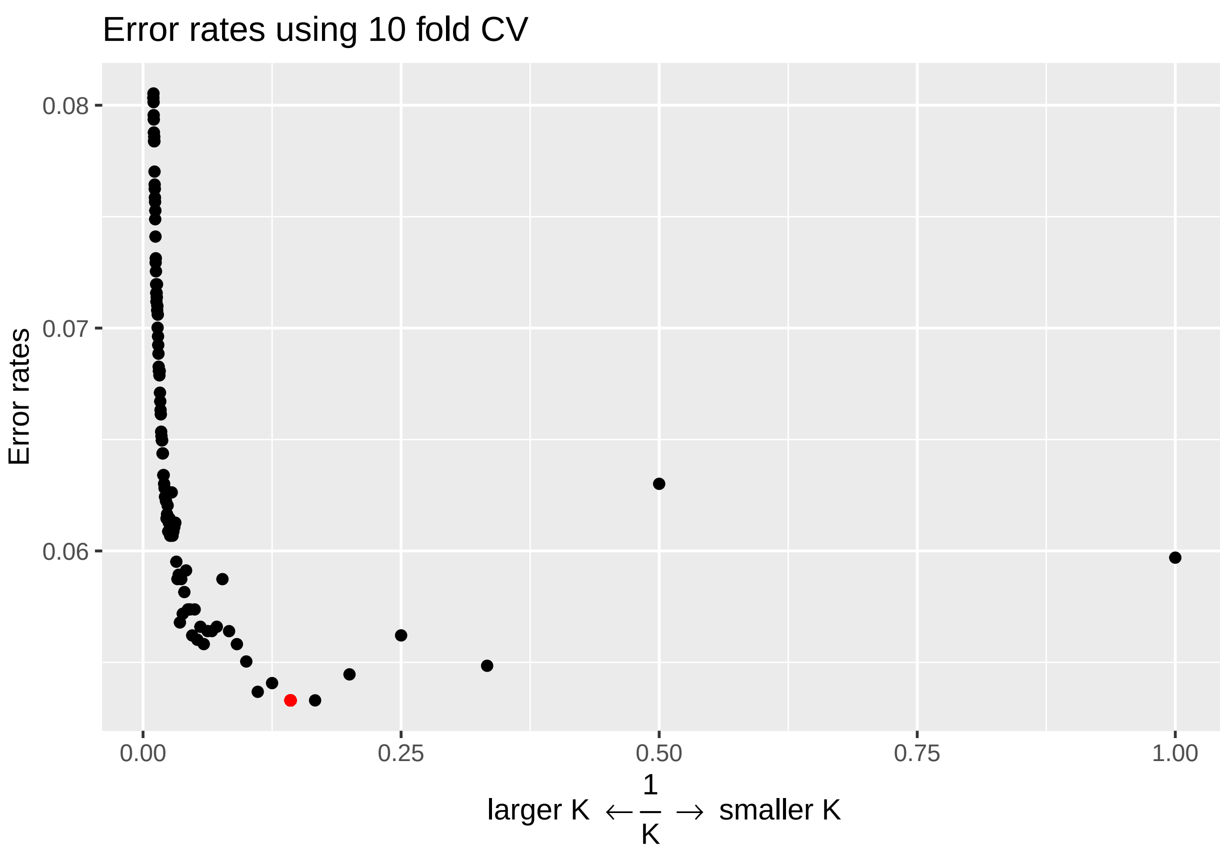
Assumptions

We assumed that hospitals with similar predictor values have similar total costs.

Results

We used 10 fold cross validation to first choose an optimal number of neighbors, k and found $k = 7$ to be optimal with an error rate of 0.0533 using Euclidean distance. The true error rate with $k = 7$ was 0.0490 and the true/false positive and negative rates are summarized in the table below. When plotting the cross validation error rates against the chosen k, we see a condensed U shape. This may suggest that large k suffers from high inaccuracy but too small k can lead to overfitting.

Classification Rates	Values
True positive	0.91200
True negative	0.99300
False positive	0.00722
False negative	0.08840



Multiple Logistic Regression

Assumptions

We assume that our predictors are not correlated with one another.

Results

The coefficients and standard errors associated with our model can be found below. We found that `total_discharges`, `total_assets`, `salaries`, `rural`, and `provider_bin_Specialized` were the most statistically significant. Further, most predictors increased the probability of a hospital's total costs being above the median; unsurprisingly, `salaries` stood out the most. A one unit increase in `salaries` increased the log odds of an above median classification by 50.27. It also produced a z statistic of 21.001 providing strong evidence of an association between salaries and total costs.

Coefficients	Estimate	Std. Error	z value	p-value
(Intercept)	18.98	0.86	21.96	0.00
number_of_beds	-0.84	0.68	-1.23	0.22
fte_employees_on_payroll	0.68	0.33	2.02	0.04
total_days	-4.44	5.80	-0.77	0.44
total_discharges	4.94	0.75	6.55	0.00
total_income	0.94	0.36	2.58	0.01
total_assets	1.76	0.52	3.35	0.00
salaries	50.27	2.39	21.00	0.00
inpatients	3.62	5.91	0.61	0.54
rural	-1.22	0.20	-6.16	0.00
control_bin_Governmental	-0.46	0.24	-1.94	0.05
control_bin_Proprietary	0.44	0.23	1.93	0.05
provider_bin_Specialized	-4.02	0.40	-9.92	0.00
duplicate	0.56	0.53	1.06	0.29

Our estimated and true error rates were pretty close to one another with our cross validation error of 0.073 and true test error of 0.0333.

Classification Rates	Values
True positive	0.9460
True negative	0.9890
False positive	0.0108
False negative	0.0544

Multiple Logistic Regression with Transformations

Assumptions

We again assume that our predictors are not correlated with one another.

Results

We decided to transform `total_income`,`fte_employees_on_payroll`, `salaries`, and `total_days` to experiment with how less significant predictors in conjunction with `salaries` affected the response. We computed polynomial models up to degree 2 for `total_days` and interaction terms between `total_income` & `fte_employees_on_payroll` and between `fte_employees_on_payroll` & `salaries`.

With our smaller model, all of our new predictors were statistically significant with extreme z statistics. However, it is important to note that the standard errors associated with each coefficient were extremely high suggesting a poor fit.

Coefficients	Estimate	Std. Error	z value	p-value
(Intercept)	6.583056e+14	971394.8	677691118	0

total_income	1.743355e+14	1667476.0	104550530	0
fte_employees_on_payroll	7.442677e+14	3289614.4	226247710	0
salaries	2.737959e+15	3725566.4	734910894	0
total_days	2.542708e+14	2645218.8	96124663	0
total_days_sq	-8.915542e+13	360959.8	-246995412	0
income_emp	-2.830918e+13	259889.2	-108927893	0
sal_emp	-2.512173e+14	333727.6	-752761545	0

Compared to our original multiple logistic regression, our true error rate shot up from 0.033 to 0.158 and our cross validation error rate shot up from 0.073 to 0.130. Interestingly enough though, the model perfectly predicted hospitals whose total costs were below the median with a true negative rate of 100%. But, it did misclassify hospitals whose total costs were above the median with a false negative rate of 30.6%. Overall, the transformations performed worse than our original multiple logistic model.

Classification Rates	Values
True positive	0.694
True negative	1.000
False positive	0.000
False negative	0.306

LDA

Assumptions

We assume that our predictors are drawn from a multivariate normal distribution and both classes share a common covariance matrix.

Results

As stated in the introduction, LDA performed the worst with a cross validation error rate of 0.175 and a true error rate of 0.165. This may suggest that our original assumption of our predictors being sampled from a multivariate normal was incorrect or that our classes do not share a common covariance matrix. Further, it is clear that a linear decision boundary is not sufficient to classify hospitals' total costs.

Classification Rates	Values
True positive	0.7620
True negative	0.9130
False positive	0.0866
False negative	0.2380

QDA

Assumptions

We still assume that our predictors are drawn from a multivariate normal distribution but drop the assumption that both classes share a common covariance matrix.

Results

QDA did not perform much better than LDA with a cross validation error of 0.115 and a true error rate of 0.119. However, compared to LDA, QDA did a much better job of accurately classifying hospitals whose total costs were below the median reducing the false positive rate from 8.7% to 1.4%. The false negative rates stayed fairly consistent hovering around $\sim 20\%$ in both methods. Overall, QDA is adequate at predicting our class labels.

Classification Rates	Values
True positive	0.7820
True negative	0.9860
False positive	0.0144
False negative	0.2180

Naive Bayes with Gaussian kernel

Assumptions

We assume that our predictors are not correlated with one another and are drawn from a multivariate normal distribution given the target class.

Results

With a Gaussian kernel, the naive Bayes classifier was comparable to QDA. This classifier produced a cross validation error rate of 0.127 and a true error rate of 0.137. The misclassification rates were also extremely similar and can be summarized in the table below.

Classification Rates	Values
True positive	0.7520
True negative	0.9820
False positive	0.0181
False negative	0.2480

Naive Bayes with Kernel Density Estimation

Assumptions

We still assume that our predictors are not correlated with one another but drop the normal distribution assumption.

Results

Without assuming normality, the Bayes Classifier performs much better with a cross validation error rate of 0.073 and a true error rate of 0.0806.

Classification Rates	Values
True positive	0.8670
True negative	0.9750

False positive	0.0253
False negative	0.1330

Decision Tree with Pruning

Assumptions

We make no assumptions on the structure of the data.

Results

We pruned the tree using 5 terminal nodes which was found to be the optimal number of nodes using cross validation. The cross validation error rate was 0.051 compared to the true error rate of 0.0595. We also noticed that the split at the root node immediately determines how each hospital will be classified. If `salaries < -0.305`, then the hospital will be classified as having total costs below the median; otherwise, the hospital will be classified as above the median.

The false positive rate was extremely low at 1.4% while the false negative rate was slightly higer at 10%.

Classification Rates	Values
True positive	0.8980
True negative	0.9860
False positive	0.0144
False negative	0.1020

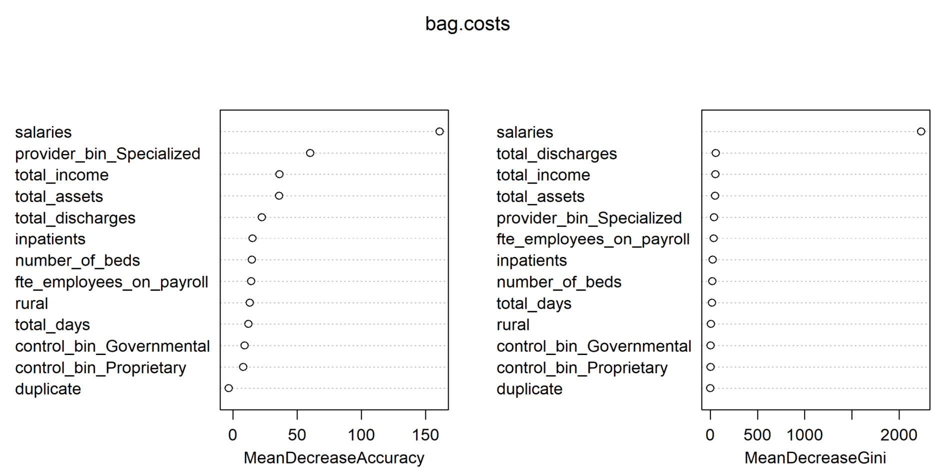
Bagging

Assumptions

We again make no assumptions on the structure of the data.

Results

Bagging reduced our cross-validation error to 0.038 compared to the decision tree, a factor of about 1.3. Looking at the importance plot, `salaries` is the most importance variable. If it were to be removed from the tree, an average of 161 hospitals would be misclassified, given by the mean decrease in accuracy. Further, it's mean decrease of the Gini index is 2234.21.



Bagging also had low misclassification rates with a false positive rate of 1.81% and a false negative rate of 4.08%.

Classification Rates	Values
True positive	0.9590
True negative	0.9820
False positive	0.0181
False negative	0.0408

Random Forest

Assumptions

We again make no assumptions on the structure of the data.

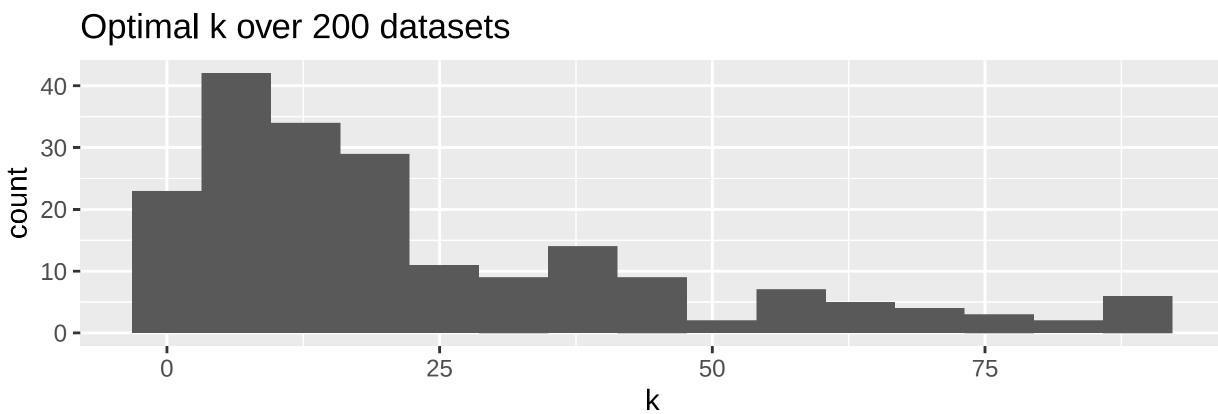
Results

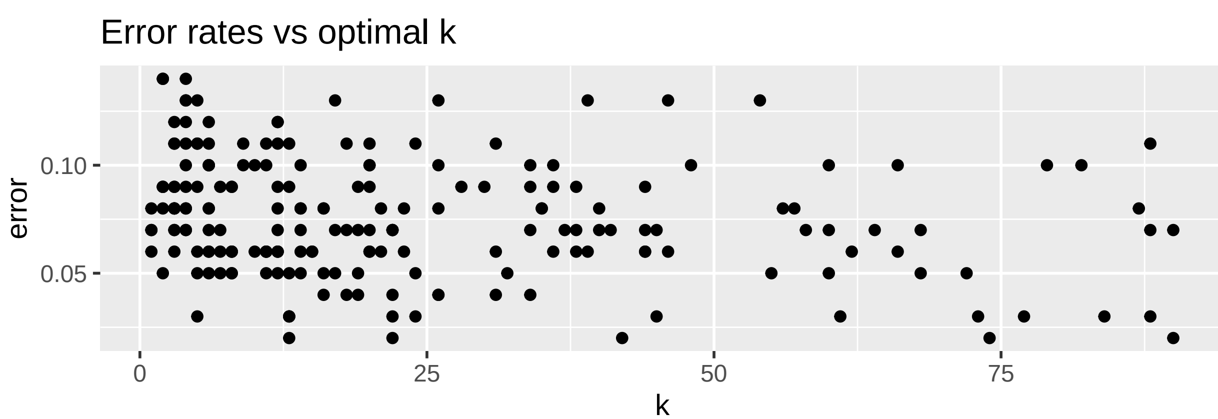
Simulation Study

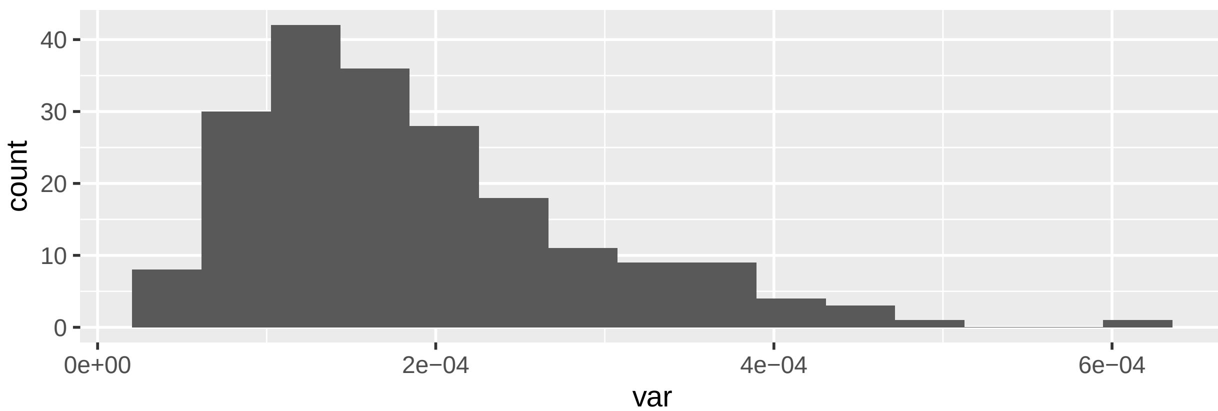
We were interested in understanding how our data affected the optimal choice of k in the k -nearest neighbors algorithm. We already experienced some variability when running our model through on different computers. Therefore, we wanted to see if more simulated datasets would produce the same variablility.

To replicate our 13 predictors and 2 response variables, we used a package called `faux` to simulate our numerical predictors from a multivariate normal distribution.

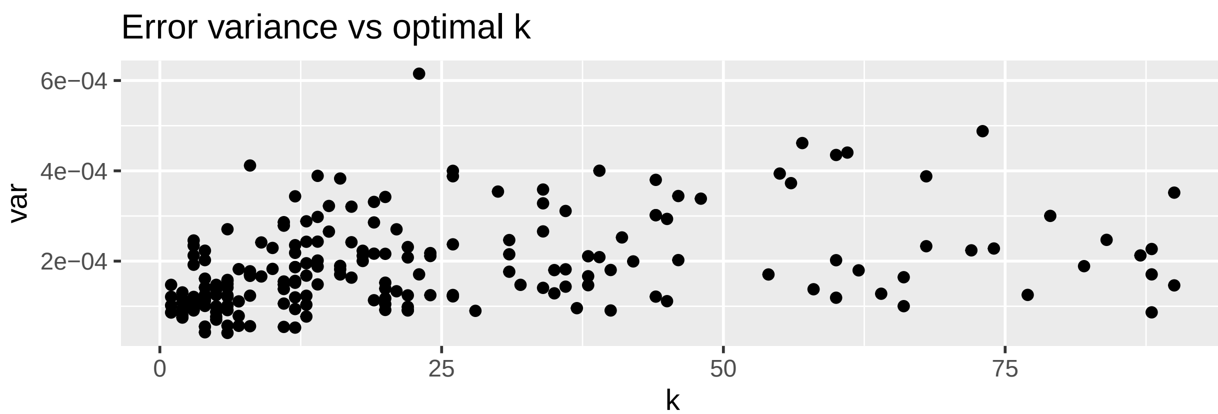
We used a standard normal distribution to replicate our 13 predictors.







count
40
30
20
10
0
0e+00
2e−04
4e−04
6e−04
var



Error variance vs optimal k
var
6e−04
4e−04
2e−04
0
25
50
75
k