Final Project Report

● Graded

Group

Vijayapriya Krishnamurthy
Sukriti Raut
Tushar Kailashchandra Tawanee
View or edit group

Total Points

23 / 25 pts

Question 1

Novelty

4.25 / 5 pts

– **0 pts** Novel, with no issues re: demonstrating contribution

– **0.125 pts** The topic is commonly approached, or the report did not do enough to demonstrate the unique contribution.

– **0.25 pts** The topic is commonly approached, or the report did not do enough to demonstrate the unique contribution.

– **0.5 pts** The topic is commonly approached, or the report did not do enough to demonstrate the unique contribution.

✔ – **0.75 pts** The topic is commonly approached, or the report did not do enough to demonstrate the unique contribution.

– **1 pt** The topic is commonly approached, or the report did not do enough to demonstrate the unique contribution.

– **1.25 pts** The topic is commonly approached, or the report did not do enough to demonstrate the contribution.

– **1.5 pts** The topic is commonly approached, or the report did not do enough to demonstrate the contribution.

– **2 pts** The topic is commonly approached, or the report did not do enough to demonstrate the contribution.

– **2.5 pts** Click here to replace this description.

Question 2

Methodology/Rigor

8.75 / 10 pts

- **– 0 pts** The methodology was reasonable and met the standard of rigor expected of a student finishing ISYE 6740.
- **– 0.125 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 0.25 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 0.5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 0.75 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 1 pt** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- ✔ **– 1.25 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 1.5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 1.75 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 2 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 2.25 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 2.5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 2.75 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 3 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 3.5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 3.75 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 4 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 4.5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.
- **– 5 pts** The methodology had flaws and/or the rigor displayed was not at the level expected of a student finishing ISYE 6740.

Question 3

Scope

5 / 5 pts

✔ **– 0 pts** The work done matched what was expected out of a group of this size.

– 0.125 pts The work done was less than expected for the size of the group.

– 0.25 pts The work done was less than expected for the size of the group.

– 0.5 pts The work done was less than expected for the size of the group.

– 0.75 pts The work done was less than expected for the size of the group.

– 1 pt The work done was less than expected for the size of the group.

– 1.25 pts The work done was less than expected for the size of the group.

– 1.5 pts The work done was less than expected for the size of the group.

– 2 pts The work done was less than expected for the size of the group.

Question 4

Writing Quality

5 / 5 pts

✔ **– 0 pts** The report was clearly written, with good use of plots/figures, and no grammatical issues.

– 0.125 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 0.25 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 0.5 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 0.75 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 1 pt Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 1.25 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 1.5 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 1.75 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 2 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better.

– 2.5 pts Writing was unclear, there were grammatical issues, and/or the report could have used visualizations better and/or the report too short

– 3 pts Writing was difficult to follow

TA1: Solid work overall with great precision TA2: Interesting project, great execution

No questions assigned to the following page.

Project Report
ISyE6740 Computational Data Analytics

Understanding US Hospitals' Price Markup Using Charge-to-cost Ratios

Submitted By
Sukriti Raut
Tushar Tawanee
Priya Krishnamurthy

No questions assigned to the following page.

Contents

Questions assigned to the following page: 1, 2, and 3

Abstract

Markup in hospital charge-to-cost ratio, which represents the ratio of hospital charges to Medicare-allowable costs, is a common measure used in US hospitals. Some hospitals' submitted charges are much higher than others for the same type of service or procedure and have a significant impact on various vulnerable groups. Our study aims to understand this discrepancy better by analyzing the 2019 Medicare hospital cost report using numerous statistical approaches and machine learning models. We hope our findings and the tools we have developed can help the general healthcare consumers, policy makers, hospital administrators, and others to better regulate hospital pricing and make healthcare affordable and accessible for everyone.

Problem Statement

Based on an analysis of 2012 Medicare cost report, fifty US hospitals were found to have charge-to-cost ratios (ratios of hospital charges over Medicare-allowable costs) approximately ten times their Medicare-allowable costs compared to a national average of 3.4 and a mode of 2.41. While these markups do not affect public and private health insurers, hospitals ask certain vulnerable groups like uninsured, out-of-network patients, and casualty and workers' compensation insurers to pay the full charges. Information asymmetry, and lack of price transparency and standard all-payer rate setting, pose a huge financial burden for such vulnerable groups while contributing to the United States' overly expensive and broken healthcare system.

Using machine learning algorithms and data mining tools, our project aims to understand why certain hospitals markups have much higher than their Medicare-allowable costs compared to a much lower national average. We hope our findings will aid in improving price transparency and healthcare finance knowledge among policy makers and healthcare consumers. Our two main research questions are:

1. What factors are contributing most significantly to hospital's charge-to-cost ratios?

2. Does the resulting number of clusters using an unsupervised learning approach (clustering) compare similarly to the actual two groups (high/low) calculated based on charge-to-cost ratios?

Dataset

Data Source

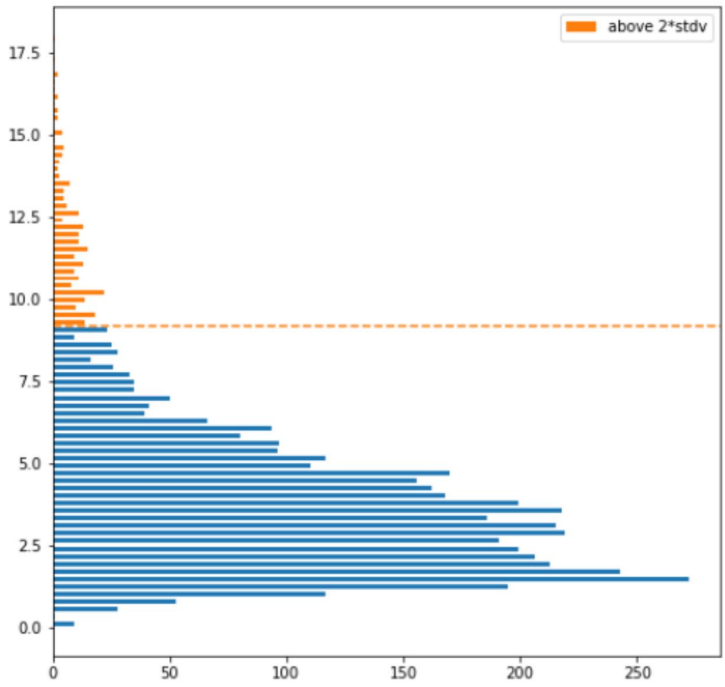
We are analyzing the 2019 Hospital Provider Cost Report that lists various cost, revenue, and expenses-related line items for hospitals in the US. The report is a part of the Healthcare Cost Report Information System (HCRIS) maintained by the Centers for Medicare & Medicaid Services (CMS). Data can be accessed using this link. The full list of 126 variables and definitions can be found here.

Data and Variables

There are around 6100 rows of data and 126 variables where each line represents observations for a hospital. For our project's purposes, we narrowed the list of variables to 38 features included in Appendix A. The rationale and metrics used to filter down the list is explained in the Methodology section below. Our target variable, charge-to-cost ratio, is derived from reciprocal of "cost to charge ratio" from the dataset.
Charge-to-cost ratio is calculated by dividing hospital's total gross charges by total Medicare allowable cost. The Medicare-allowable cost refers to the cost determined by the CMS to be associated with care for all patients, which includes both direct patient cost and indirect general service cost[1]. Charge-to-cost

Questions assigned to the following page: 1 and 4

variable was converted into a categorical variable indicating ‘high’ (1) and ‘low’ (0) by comparing against a threshold. The threshold of 9.17 was decided based on 2 standard deviations from 2019 mean charge-to-cost ratio as shown in the histogram below. The average charge-to-cost ratio was 3.98, but 251 hospitals had their charge-to-cost ratio higher than the threshold which is more than twice as high.



Methodology

Missing Values

We decided to impute columns with less than 500 missing values. We utilized sklearn’s KNNImputer package that uses the mean value from n_neighbours to impute missing data. Two samples are considered neighbors if the features that neither is missing are close. We compared the average values of variables pre and post data imputation which are comparable.

Exploratory Data Analysis (EDA)

Distributions

By categorizing the features into five major groups – *Identifier, Expense, Patient Care Process, Revenue and Financials* – we initiated the EDA process. We performed EDA using the imputed dataset for the binary Charge-to-cost variable ‘CC_Label’. The original raw data contained significant missing values making it challenging to extract any patterns or distributions. EDA results are summarized based on the groupings of the variables. We did not transform any data at this point to address the skewness. We perform scaling while fitting the respective models.

Questions assigned to the following page: 1, 2, and 4

Identifiers

These categorical variables describe the characteristics of the hospitals. Plotting their bar graphs against the count of binary charge-to-cost ratio variable, we learn that the data are disproportionately distributed. Find the respective plots in Appendix B.

- *Facility type* of Short-term (General and Specialty) Hospitals (STH) and Critical Access Hospitals (CAH) comprise most of the hospitals in the sample. Only STH have a high charge-to-cost ratio compared to CAH. However, they still have a significantly higher proportion of hospitals with low charge-to-cost ratio compared to the high class.

- Similarly, most hospitals in the sample belong to the *provider types* of general short term, general long term, and rehabilitation which have disproportionately higher number of hospitals with low charge-to-cost ratio.

- Most of the hospitals represented in the sample are in an urban area again with a disproportionately higher number of hospitals with a low charge-to-cost ratio.

- 50% of our sample consists of non-profit hospitals and the rest 50% is made up of profit religious hospitals, proprietary-corporation, and governmental hospitals. Proprietary-Corporation category has the highest proportion of hospitals with high charge-to-cost ratio.

Expenses

We built Kernel Density Estimation (KDE) plots using the default gaussian kernel and default bandwidth factor of 1 (see Appendix B) for the numerical variables related to hospital expenses. These are costs that hospitals incur for providing patient care. Based on the density curves grouped charge-to-cost ratio group, the distributions are unimodal with one peak each and mostly skewed to the left. The left-skewness indicates that the mean is less than the median. The *cost of charity care* and *total bad debt expense* are higher for hospitals with low charge-to-cost ratio. All other variables have comparable distributions for hospitals between the two charge-to-cost ratio groups. The left-skewness could be an indication of a heavy tail, but not necessarily any outliers.

Patient Care Process

Similarly, KDE plots (see Appendix B) show unimodal distributions for variables related to hospital operations and patient care process. Hospitals with low charge-to-cost ratio have less skewness in data for these variables: *total bed days available, total days, and total discharges*. Distributions for *FTE – employees on payroll* and *number of beds* are left-skewed for hospitals in both high and low charge-to-cost ratio group.

Revenue

Based on KDE plots (see Appendix B) for revenue-related variables, hospitals with a high charge-to-cost ratio demonstrate less skewed distribution for *inpatient total charges, inpatient revenue, Medicaid charges, outpatient charges, and outpatient revenue* than hospitals with low charge-to-cost ratio. This means that the averages for these low charge-to-cost hospitals will be lower than the high charge-to-cost group. Some variables such as *net income, net income from service to patients, and total other income* have negative values which suggests loss for some hospitals which primarily belong to the low charge-to-cost ratio group.

Questions assigned to the following page: 2 and 4

Finances

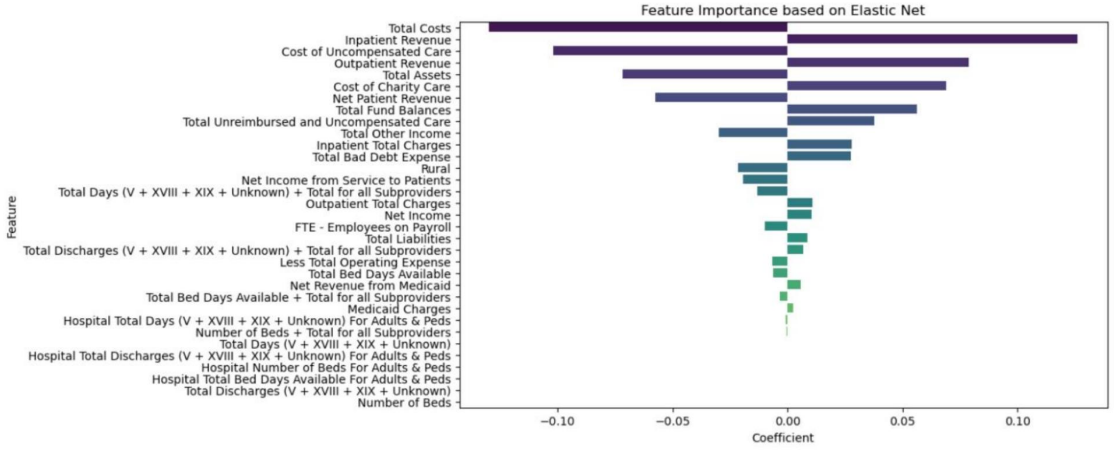
KDE plots of *total assets, total fund balances, and total liabilities* also display some skewness and both positive and negative observations.

Feature Selection

We encountered three problems in the selected dataset as it relates to the features: redundancy, relevance, and data scarcity. Since we had >100 variables, we first performed a subjective review of the variables based on subject matter expertise of a healthcare data analyst and an accountant to remove irrelevant variables. Since we decided to impute only up to 10% of each feature's missing data, that eliminated a few columns with sparse data (> 500 missing entries). We then applied three regularized regressions, Lasso, Ridge and Elastic Net, for a more robust feature selection process. We applied cross validation to determine the optimal regularization parameter (alpha) that maximizes feature selection and evaluated the selected features' performance on test data. A low RMSE value, such as 0.15, suggests that the model's predictions are relatively close to the actual target values on average.

	Lasso	Ridge	Elastic Net
Best Alpha	0.0001	3.5938	0.001
Best RMSE	0.15463	0.15448	0.1550

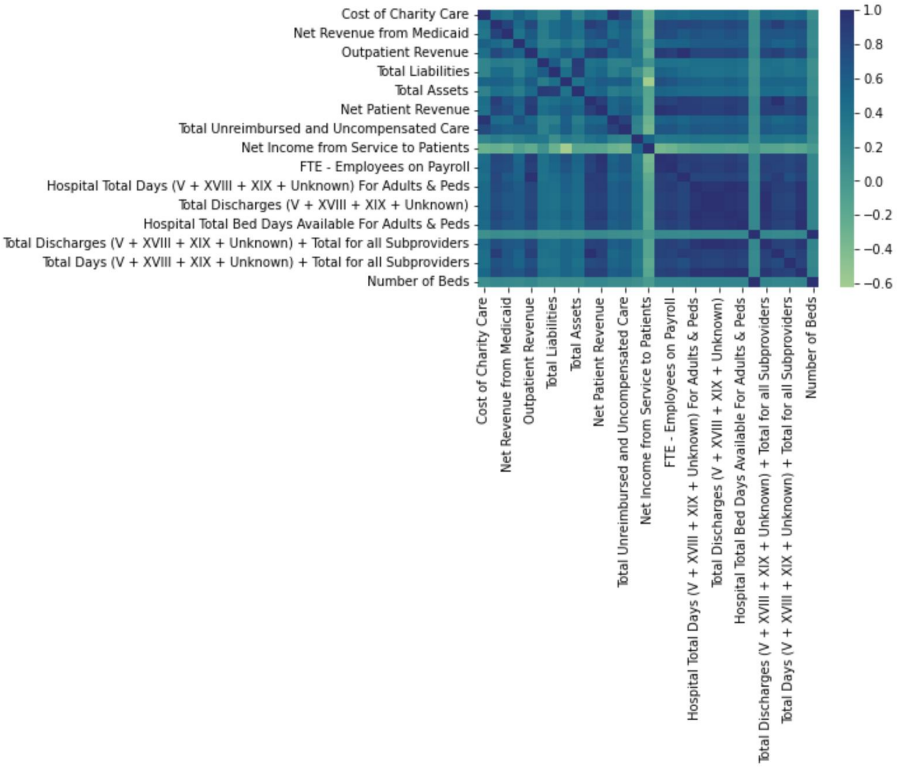
There was an overlap in variables selected by three models. We decided to fit our models using the features selected by elastic net regression as shown in figure below. Elastic net can balance the bias-variance trade-off by finding a middle ground between underfitting and overfitting. Plots for Lasso and Ridge regression are attached in Appendix C.



Questions assigned to the following page: 2 and 4

Collinearity

Multicollinearity seems to be an issue with the dataset as shown in the correlation matrix below. A lot of variables show strong positive correlations and some moderate to weak negative correlations.



Total Other Income and *Net Income from Service* to Patients show moderate negative correlation of -0.618. All other negatively correlated variables have coefficients less than -0.3602. On the other hand, many variables show strongly positive correlation of >0.90 as shown in the table in Appendix D. Including highly correlated variables especially in regression models may result in overfitting and unstable coefficients. Rather than removing variables based on coefficients only, we decided to combine this information with variable selection models to eliminate highly correlated variables.

Question assigned to the following page: 2

Fitting Models

Supervised Learning

We used 12 features based on its importance, selected by Elastic Net Regression, to fit three supervised classification models. The models were trained on the training dataset (80%) and evaluated on the testing dataset (20%).

1. Random Forest Classifier
   We fitted random forest model on the unscaled dataset using RandomForestClassifier library from sklearn and tuned the model using cross validation with various number of trees (50 - 500) and number of splits/nodes (1 - 20). The model selected 166 trees and depth of 15 nodes as the best hyperparameters.

2. KNN Classifier
   Similarly, we fitted KNeighborsClassifier from sklearn on the scaled dataset. The model was tuned using Kfold validation for neighbors ranging from 1 to 25. The best number of nearest neighbors K=3 was identified using cross-validated mean error curve with the lowest mean error (see Appendix E).

3. Logistic Regression
   A simple logistic regression was fitted on the dataset with standardized numerical variables and dummy categorical variables using LogisticRegression from sklearn.

Unsupervised Learning

To answer our second research question, we also explored a few unsupervised algorithms by removing the target variable, charge-to-cost ratio, from the dataset.

1. KMeans clustering using PCA
   Before fitting Kmeans, we first applied PCA for dimension reduction since our dataset includes higher than two dimensions. We applied the PCA algorithm on scaled numerical variables and dummy categorical variables. The model selected five as the best number of components based on cumulative percent of variance explained (see Appendix F). It means that five components explained our threshold variance explained of 80%. We then fitted KMeans clustering using these five PCA components to obtain two clusters each corresponding to high/low charge-to-cost ratio respectively. PCA, however, is designed to work better with numerical data since it involves breaking down its variance structure and categorical variables don't have a variance structure[2].

2. KMeans clustering using Factorial Analysis of Mixed Data (FAMD)
   Thus, to better fit our dataset that has both numerical and categorical variables, we also applied FAM that employs PCA-alike operations for the numerical data and Multiple Correspondence Analysis (MCA)-alike operations for the categorical data within a unified framework[2]. We used 20 components from FAMD to fit Kmeans with two clusters.

Question assigned to the following page: 2

Evaluation

We utilized various metrics to evaluate the performance of our models and assess their validity.

Supervised Learning

All three models have a very low testing error with Logistic regression performing the best followed by Random Forest. Logistic regression tends to work well with a dichotomous or binary outcome variable like in our case.

Random Forest	KNN	Logistic Regression
0.018	0.026	0.013

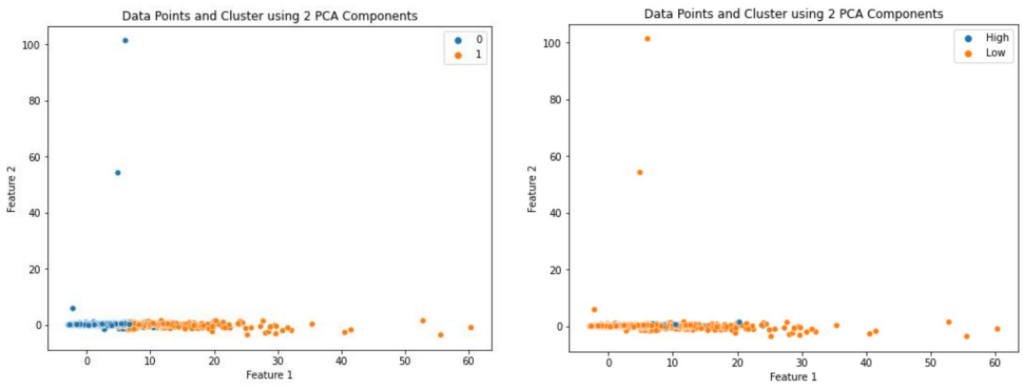
Going beyond testing errors, we also compared evaluation metrics such as precision, recall, fl-score, and support which are better suited for imbalanced classification task like ours where our sample contained significantly higher observations related to 'low' charge-to-cost ratio class. Based on precision, recall, and fl-scores (that combines precision and recall using harmonic mean instead of a simple average where it punishes extreme values[3]), Logistic Regression with values closest to 1 still seems to perform the best classifying both charge-to-cost groups accurately.

precision	Random Forest	KNN	Logistic Regression
0	0.986471	0.981982	0.986547
1	0.903846	0.843137	1
recall	Random Forest	KNN	Logistic Regression
0	0.994318	0.990909	1
1	0.79661	0.728814	0.79661
fl-score	Random Forest	KNN	Logistic Regression
0	0.990379	0.986425	0.993228
1	0.846847	0.781818	0.886792
support	Random Forest	KNN	Logistic Regression
0	880	880	880
1	59	59	59

Unsupervised Learning

The plot on the left shows two clusters formed by the KMeans model using PCA which splits around 5 where hospitals with low charge-to-cost ratio lie on the left side of the split. The plot on the right shows where our true labels lie which are spread across the horizontal line without forming any distinct clusters.

Question assigned to the following page: 2



Clusters formed by KMeans using FAMD did not show any significant improvement over the first version. The clusters overlap without showing any interpretable patterns.

Question assigned to the following page: 2

Main Findings

Based on our feature selection process, geography (*rural vs urban*) seems to play an important role in predicting charge-to-cost ratio since most of the hospitals in the high charge-to-cost group are in an urban area. Similarly, *hospital provider type* seems to have an important effect in determining charge-to-cost ratio where short-term general hospitals and rehabilitation centers exhibit high charge-to-cost ratio. It makes sense logically that variables categorized as hospital operating expenses such as *total costs, total bad debt expense, less operating expense, and FTE – employees on pay roll*, contribute significantly to charge-to-cost ratio. It is interesting that *bad debt expense* (patient balances to be written off) and *total liabilities* (balance-sheet item that is unrelated to charges or cost) also resulted as important predictors of charge-to-cost ratio. Revenue related variables such as *total other income, inpatient revenue, and net patient revenue* were also identified as important by our feature selection process.

The resulting number of clusters using an unsupervised learning approach (KMeansclustering using PCA) did not compare similarly to the actual two groups (high/low) calculated based on the charge-to-cost ratio using the threshold of 9.17. The PCA KMeans models formed two distinct clusters splitting around threshold 5 but our actual labels of two data groups did not line up in the same pattern. Kmeans using FAMD which also takes categorical variables into account shows a different cluster pattern formation with a horizontal classifier instead of a vertical classifier.

On the other hand, our classification models performed extremely well predicting on the test dataset. Our three classification models, especially logistic regression, can predict, with high accuracy, if a new hospital is likely to have either a low or high charge-to-cost ratio based on its characteristics.

Conclusion & Further Research Opportunities

In summary, we were able to answer our two primary research questions using machine learning models and build some predictive models in the process. We identified significant factors that are important in determining hospital's charge-to-cost ratio using a combination of qualitative and quantitative feature selection process. For future research purposes, we can better understand how the coefficients of various important variables affect the direction of the charge-to-cost ratio (positively or negatively) and the magnitude of change.

We also received enough evidence to reject our hypothesis that the unsupervised modelling approach would also cluster the dataset into two groups to match our actual high/low charge-to-cost ratio groups. As a comparison, we could fit another unsupervised learning algorithm such as one-class SVM anomaly detection algorithm. It could be a better fit because of its ability to differentiate samples of one class by learning from single class samples during training. Applying other feature selection algorithms to fit KMeans again could also improve the clustering.

The results of our project can be a great starting point in understanding the discrepancy in charge-to-cost ratios for hospitals in a more computational approach. Predictive models could also be a useful tool for consumers to compare costs and gap the information asymmetry in healthcare. We excluded *state* from our analysis, but a separate geospatial analysis could also be instrumental to healthcare consumers and policymakers to improve price transparency and promote standardized pricing practices for hospitals.

No questions assigned to the following page.

Appendix

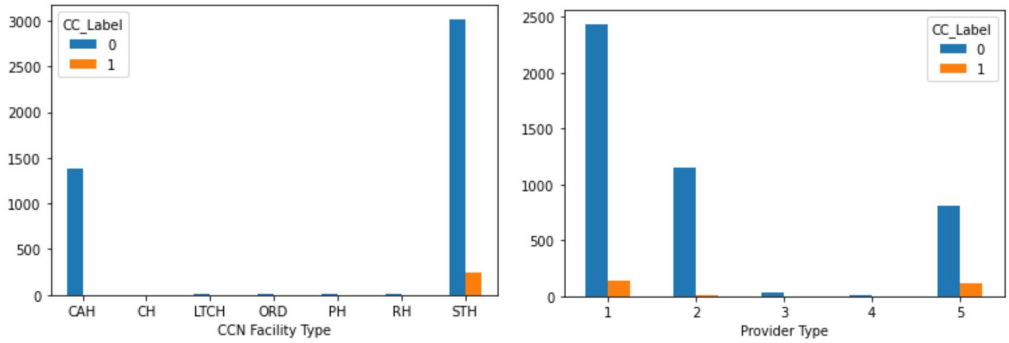
A. Shortlisted Variables

Groupings	Variables
Expense	Cost of Charity Care
Expense	Cost of Uncompensated Care
Expense	Less Total Operating Expense
Expense	Total Bad Debt Expense
Expense	Total Costs
Expense	Total Unreimbursed and Uncompensated Care
Financials	Total Assets
Financials	Total Fund Balances
Financials	Total Liabilities
Identifier	CCN Facility Type
Identifier	Provider Type
Identifier	Rural Versus Urban
Identifier	State Code
Identifier	Type of Control
Patient Care Process	FTE-Employees on Payroll
Patient Care Process	Hospital Number of Beds For Adults & Peds
Patient Care Process	Hospital Total Bed Days Available For Adults & Peds
Patient Care Process	Hospital Total Days (V + XVIII + XIX + Unknown) For Adults & Peds
Patient Care Process	Hospital Total Discharges (V + XVIII + XIX + Unknown) For Adults & Peds
Patient Care Process	Number of Beds
Patient Care Process	Number of Beds + Total for all Subproviders
Patient Care Process	Total Bed Days Available
Patient Care Process	Total Bed Days Available + Total for all Subproviders
Patient Care Process	Total Days (V + XVIII + XIX + Unknown)
Patient Care Process	Total Days (V + XVIII + XIX + Unknown) + Total for all subproviders
Patient Care Process	Total Discharges (V + XVIII + XIX + Unknown)
Patient Care Process	Total Discharges (V + XVIII + XIX + Unknown) + Total for all Subproviders
Revenue	Inpatient Revenue
Revenue	Net Revenue from Medicaid
Revenue	Total Other Income
Revenue	Inpatient Total Charges
Revenue	Medicaid Charges
Revenue	Net Income
Revenue	Net Income from Service to Patients
Revenue	Net Patient Revenue
Revenue	Outpatient Revenue
Revenue	Outpatient Total Charges
Target Variable	Cost To Charge Ratio

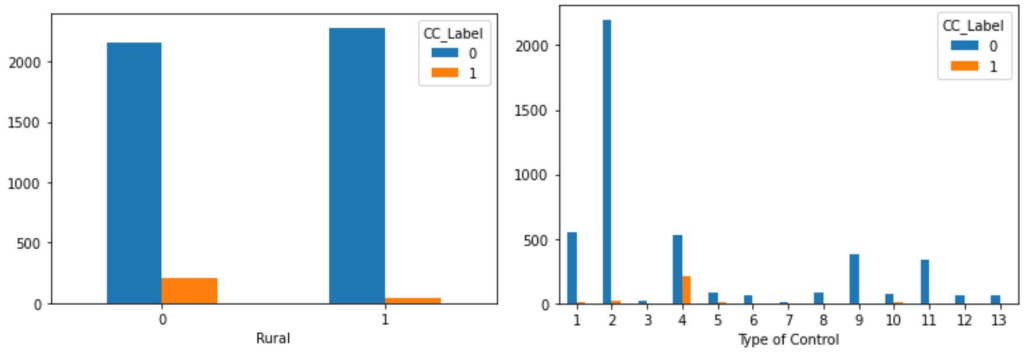
No questions assigned to the following page.

B. EDA plots

i. Identifiers



Provider Type: The number listed best corresponds with the type of services provided. 1 = General Short Term, 2 = General Long Term, 3 = Cancer, 4 = Psychiatric, 5 = Rehabilitation

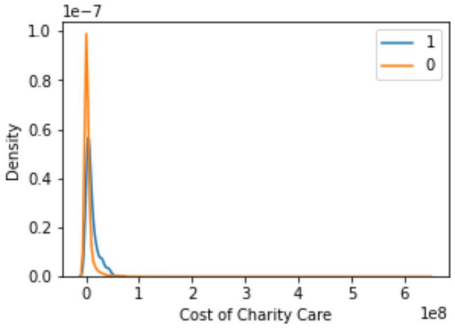


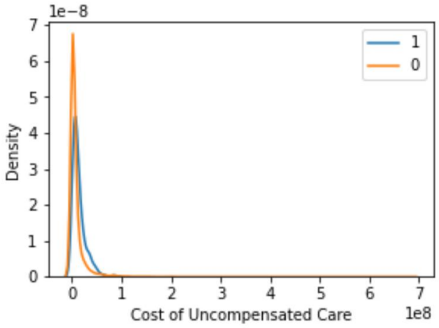
Rural=0 (rural) and Rural=1 (urban)

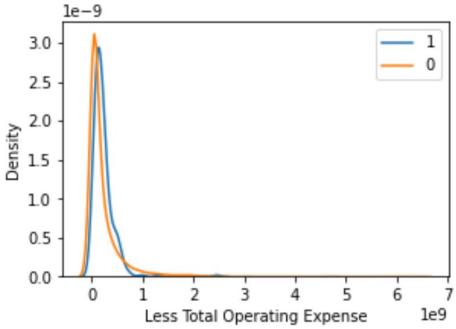
Indicates the type of control or auspices under which the hospital is conducted as indicated: 1 = Voluntary Nonprofit- Church, 2 = Voluntary Nonprofit-Other, 3 = Proprietary- Individual, 4 = Proprietary-Corporation, 5 = Proprietary- Partnership, 6 = Proprietary-Other, 7 = Governmental- Federal, 8 = Governmental-City-County, 9 = Governmental-County, 10 = Governmental-State, 11 = Governmental-Hospital District, 12 = Governmental-City, 13 = Governmental-Other.

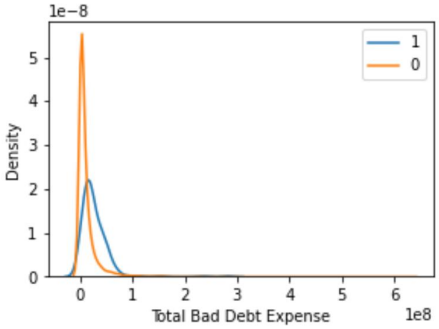
No questions assigned to the following page.

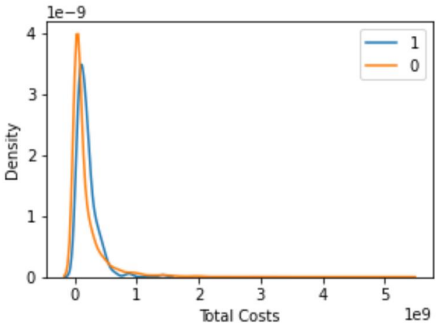
ii. Expenses

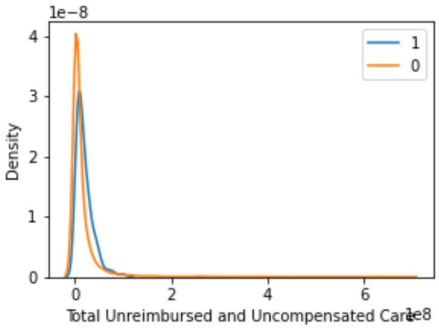

1e−7
1.0
0.8
0.6
0.4
0.2
0.0
Density
1
0
0 1 2 3 4 5 6
Cost of Charity Care
1e8


1e−8
7
6
5
4
3
2
1
0
Density
1
0
0 1 2 3 4 5 6 7
Cost of Uncompensated Care
1e8


1e−9
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
1
0
0 1 2 3 4 5 6 7
Less Total Operating Expense
1e9

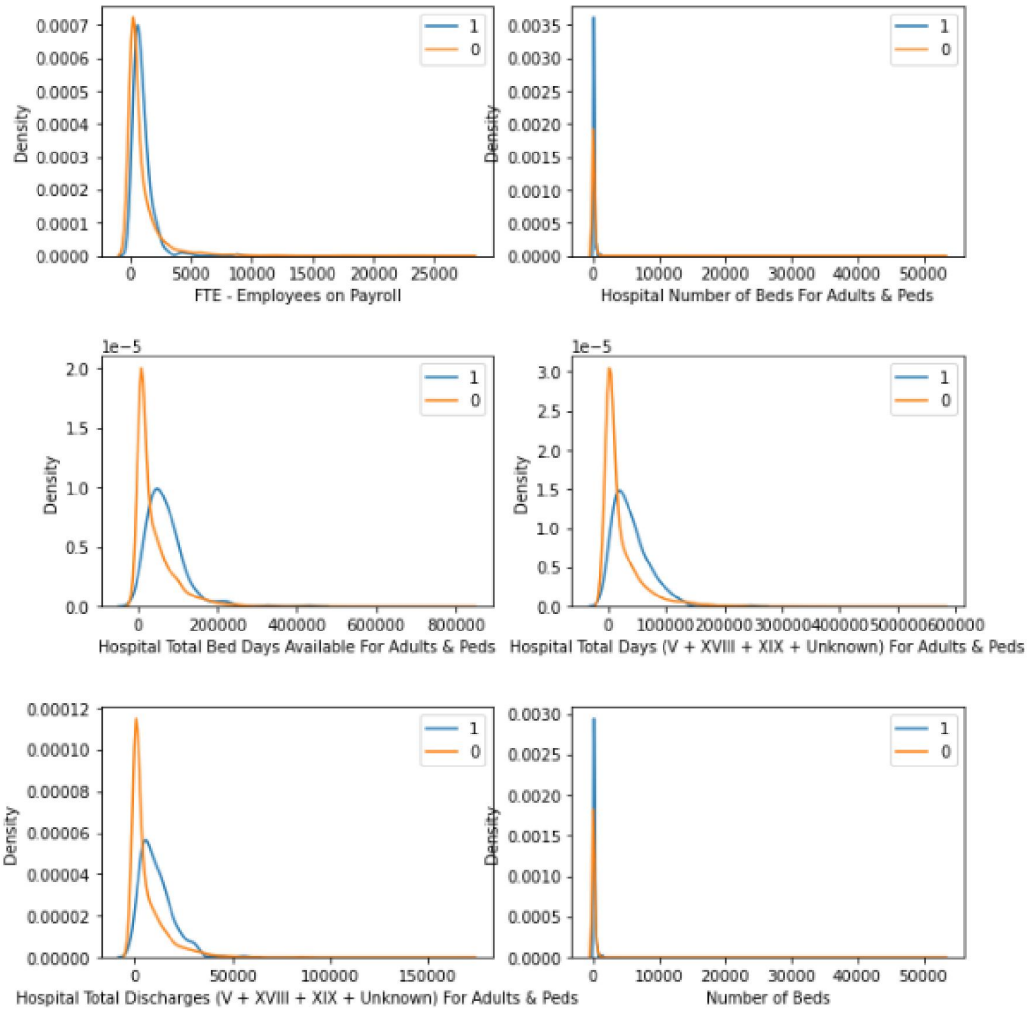

1e−8
5
4
3
2
1
0
Density
1
0
0 1 2 3 4 5 6
Total Bad Debt Expense
1e8


1e−9
4
3
2
1
0
Density
1
0
0 1 2 3 4 5
Total Costs
1e9

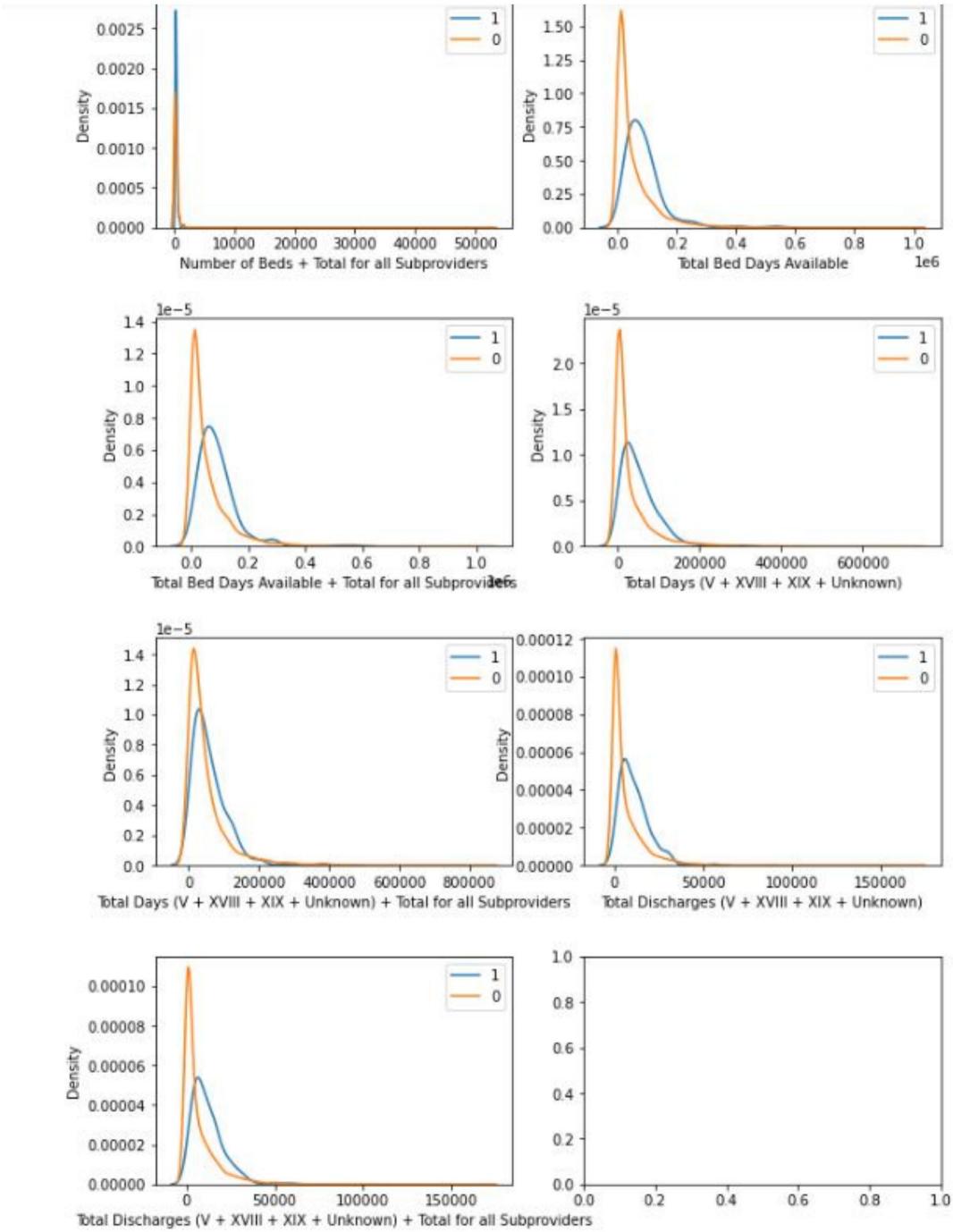

1e−8
4
3
2
1
0
Density
1
0
0 2 4 6
Total Unreimbursed and Uncompensated Care
1e8

No questions assigned to the following page.

iii. Patient Care Process

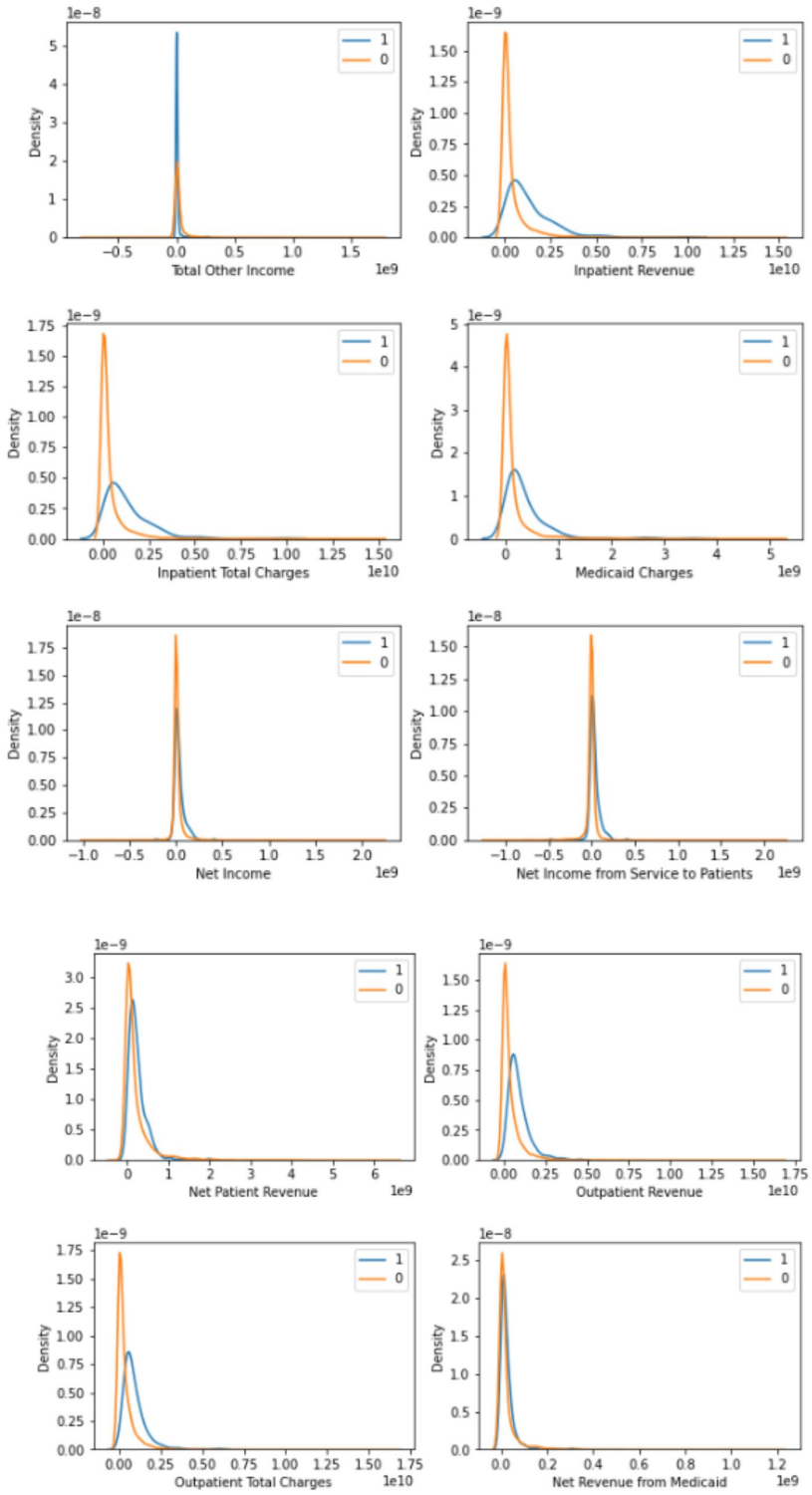


No questions assigned to the following page.


Density
Number of Beds + Total for all Subproviders
Total Bed Days Available
1e6
1e−5
Total Bed Days Available + Total for all Subproviders
Total Days (V + XVIII + XIX + Unknown)
Total Days (V + XVIII + XIX + Unknown) + Total for all Subproviders
Total Discharges (V + XVIII + XIX + Unknown)
Total Discharges (V + XVIII + XIX + Unknown) + Total for all Subproviders
1
0

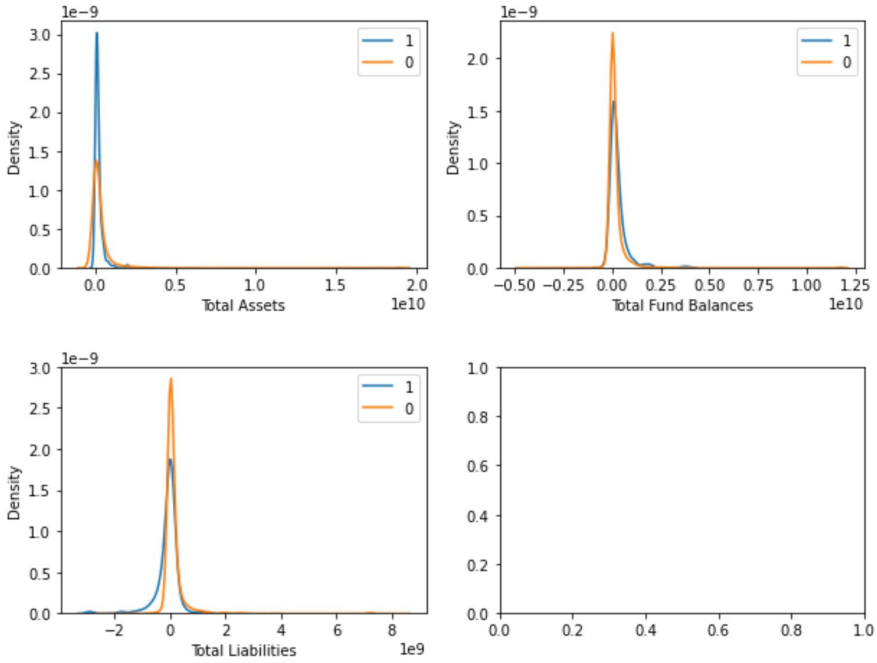
No questions assigned to the following page.

iv. Revenue


Total Other Income
Inpatient Revenue
Inpatient Total Charges
Medicaid Charges
Net Income
Net Income from Service to Patients
Net Patient Revenue
Outpatient Revenue
Outpatient Total Charges
Net Revenue from Medicaid
Density
1
0

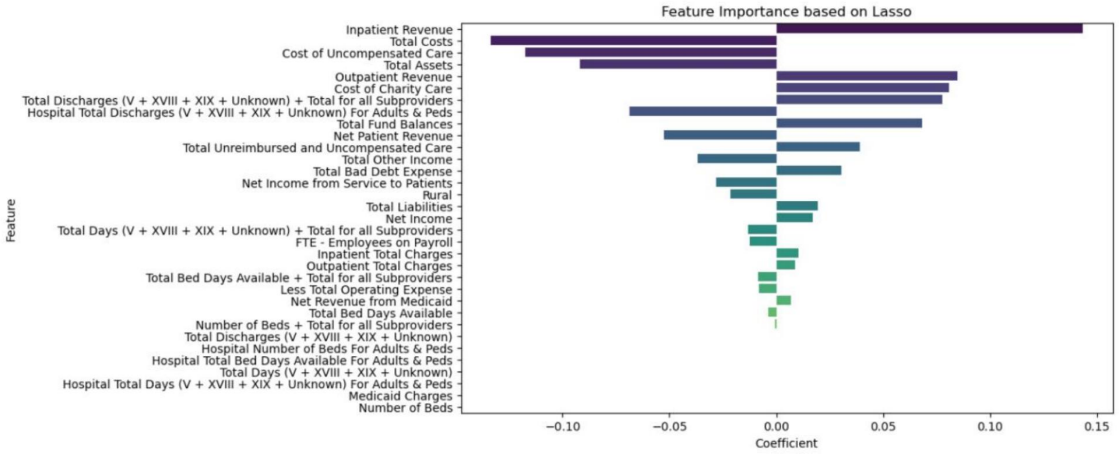
No questions assigned to the following page.

v. Finances



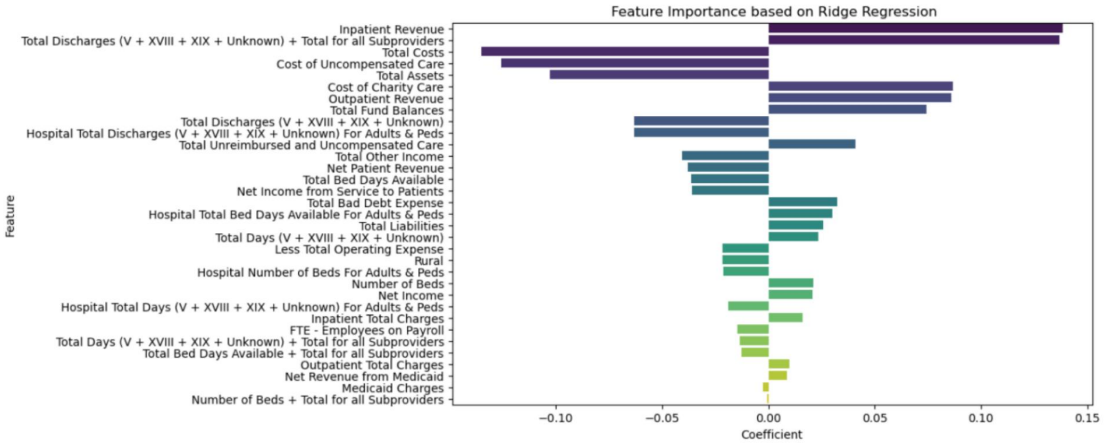
C. Feature selection

Lasso:



No questions assigned to the following page.

Ridge Regression:



D. Correlation coefficients for highly correlated variables

Cost of Charity Care	
Cost of Uncompensated Care	0.963099
FTE - Employees on Payroll	
Less Total Operating Expense	0.96248
Net Patient Revenue	0.948763
Total Days (V + XVIII + XIX + Unknown)	0.909555
Hospital Total Bed Days Available For Adults &...	
Hospital Total Discharges (V + XVIII + XIX + U...	0.941214
Hospital Total Days (V + XVIII + XIX + Unknown...	
Hospital Total Bed Days Available For Adults &...	0.969016
Hospital Total Discharges (V + XVIII + XIX + U...	0.953896
Inpatient Total Charges	
Inpatient Revenue	0.99631
Net Patient Revenue	
Less Total Operating Expense	0.977955
Number of Beds	
Hospital Number of Beds For Adults & Peds	0.998772
Number of Beds + Total for all Subproviders	

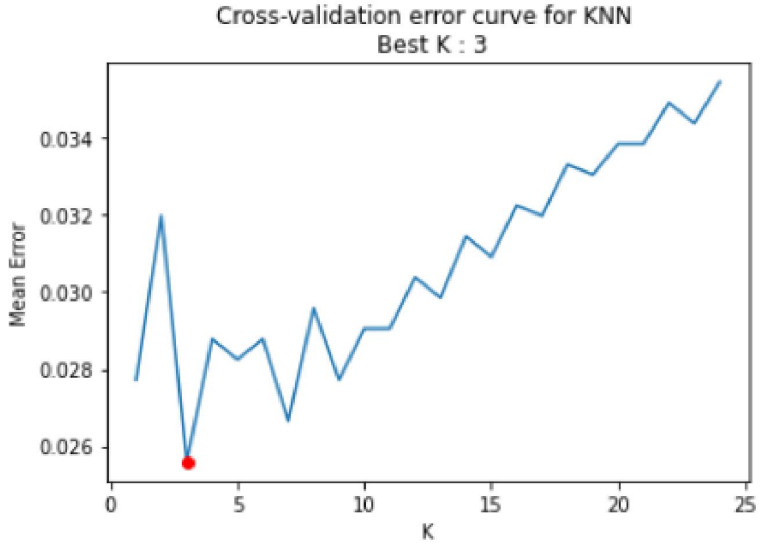
No questions assigned to the following page.

Hospital Number of Beds For Adults & Peds	0.985167
Number of Beds	0.986902
Outpatient Revenue	
Less Total Operating Expense	0.900015
Net Patient Revenue	0.904938
Outpatient Total Charges	
Outpatient Revenue	0.981285
Total Assets	
Total Fund Balances	0.903422
Total Bed Days Available	
Hospital Total Bed Days Available For Adults &...	0.992599
Hospital Total Days (V + XVIII + XIX + Unknown...	0.969216
Hospital Total Discharges (V + XVIII + XIX + U...	0.949891
Total Discharges (V + XVIII + XIX + Unknown)	0.949891
Total Discharges (V + XVIII + XIX + Unknown) +...	0.95118
Total Bed Days Available + Total for all Subpr...	
Hospital Total Bed Days Available For Adults &...	0.959881
Hospital Total Days (V + XVIII + XIX + Unknown...	0.939285
Hospital Total Discharges (V + XVIII + XIX + U...	0.917712
Total Bed Days Available	0.966673
Total Days (V + XVIII + XIX + Unknown)	0.944747
Total Discharges (V + XVIII + XIX + Unknown)	0.917712
Total Discharges (V + XVIII + XIX + Unknown) +...	0.924797
Total Costs	
FTE - Employees on Payroll	0.950711
Hospital Total Days (V + XVIII + XIX + Unknown...	0.902376
Less Total Operating Expense	0.977224
Net Patient Revenue	0.973057
Total Days (V + XVIII + XIX + Unknown)	0.911937
Total Days (V + XVIII + XIX + Unknown)	
Hospital Total Bed Days Available For Adults &...	0.962998
Hospital Total Days (V + XVIII + XIX + Unknown...	0.991335
Hospital Total Discharges (V + XVIII + XIX + U...	0.959188
Net Patient Revenue	0.900031
Total Bed Days Available	0.976419
Total Discharges (V + XVIII + XIX + Unknown)	0.959188
Total Discharges (V + XVIII + XIX + Unknown) +...	0.959608
Total Discharges (V + XVIII + XIX + Unknown)	
Hospital Total Bed Days Available For Adults &...	0.941214
Hospital Total Days (V + XVIII + XIX + Unknown...	0.953896
Total Discharges (V + XVIII + XIX + Unknown) +...	0.998434
Total Discharges (V + XVIII + XIX + Unknown) +...	

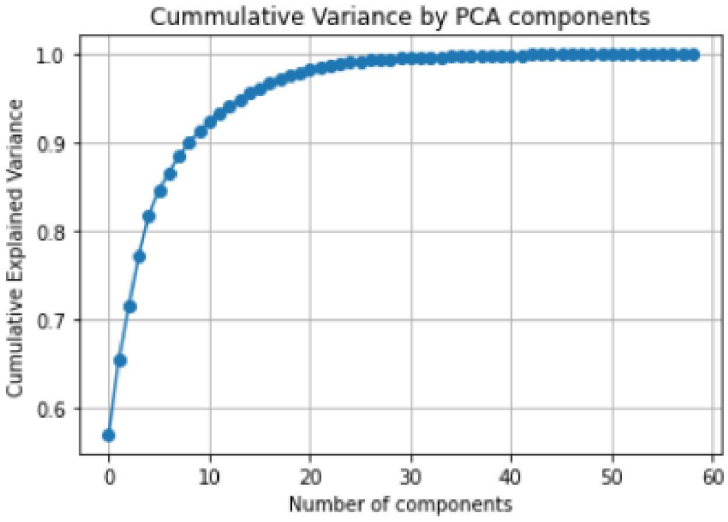
No questions assigned to the following page.

Hospital Total Bed Days Available For Adults &...	0.942356
Hospital Total Days (V + XVIII + XIX + Unknown...	0.954406
Hospital Total Discharges (V + XVIII + XIX + U...	0.998434

E.



F.



No questions assigned to the following page.

References

1. Bai G, Anderson GF. Extreme Markup: The Fifty US Hospitals With The Highest Charge-To-Cost Ratios. Health Aff (Millwood). 2015 Jun;34(6):922-8. doi: 10.1377/hlthaff.2014.1414. PMID: 26056196.
2. Can PCA be used for categorical variables? example & alternatives. Statistics Globe. (2023, June 2). https://statisticsglobe.com/pca-categorical-variables
3. Precision and recall: How to evaluate your classification model. Built In. (n.d.). https://builtin.com/data-science/precision-and-recall