ORIE 4741 Midterm Project Report

Kathy JaYoung Byun, Raye Liu and Jody Zhu

November 8, 2020

Problem

The goal is to predict the length of stay and total cost of each cancer patient in New York State hospitals from demographics and health data using models such as linear and random forest regression. Our approach is to cluster the patients into “homogeneous” groups that yield better predictions than in the general body. We will also form and examine clusters as a next step of our project.

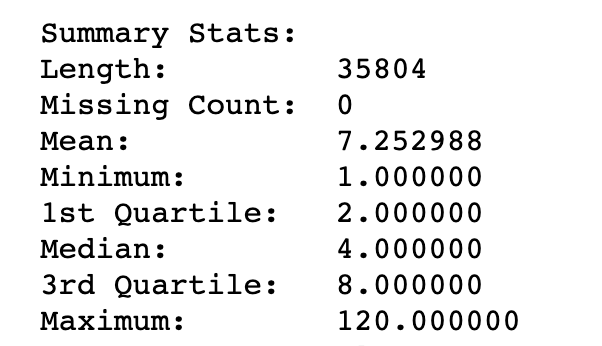
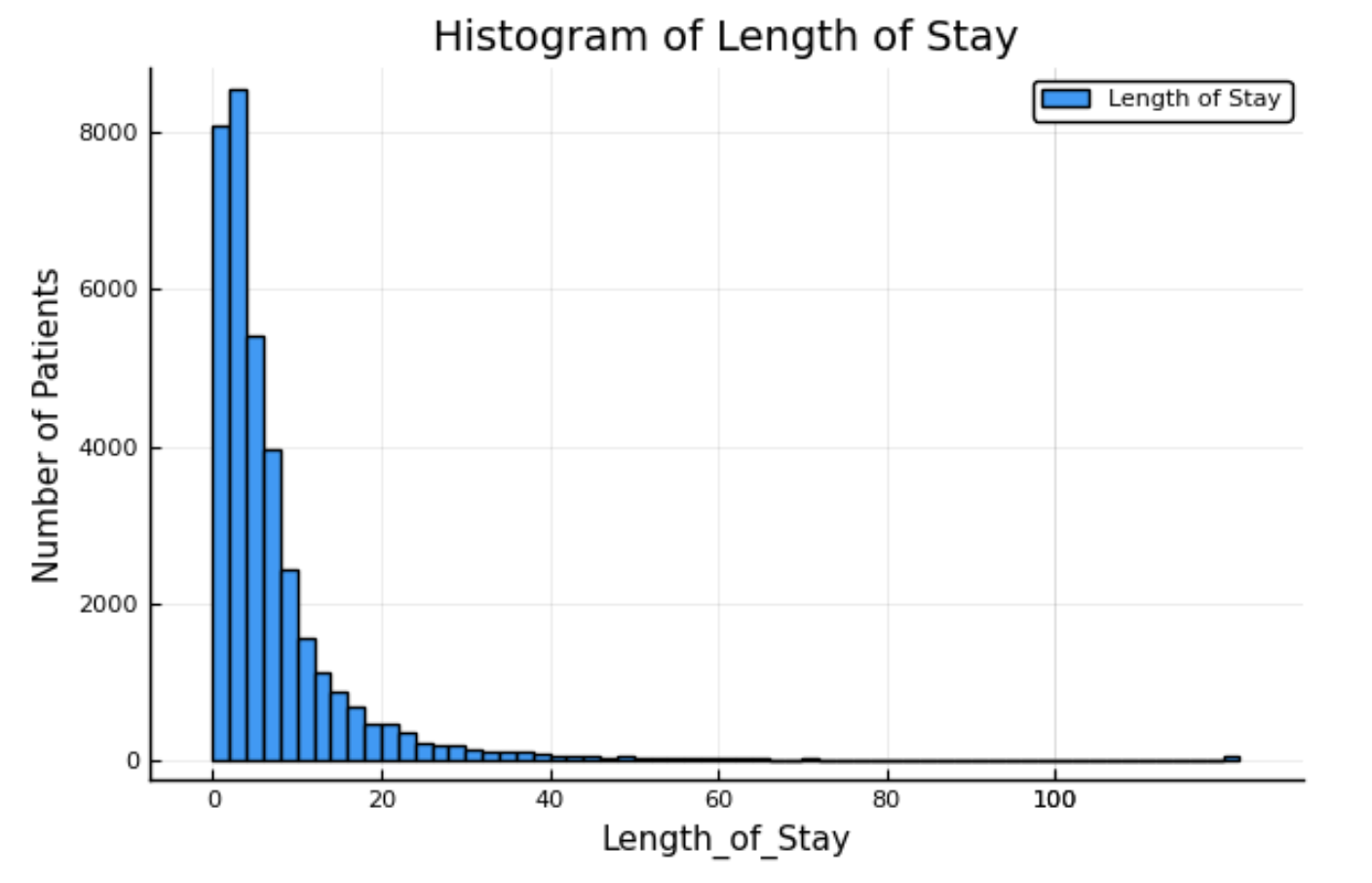
Dataset

The dataset used in this project is New York State’s Statewide Planning and Research Cooperative System’s (SPARCS) Hospital Inpatient Discharges. It contains information about patients discharged from hospitals in New York State in 2012. Some of the fields are race, age group, type of admission, diagnosis, severity index, length of stay, and total charges. To narrow down the scope we are starting with, we will only perform data analysis on cancer patients. The SPARCS\_Cancer data has 35,804 rows/examples and 33 columns/features before preprocessing. There are 420 missing entries in Zip Code, 9,438 in Payment Typology 2, 21,636 in Payment Typology 3, and 0 otherwise. We changed our dataset from MIMIC-III to SPARCS because SPARCS has less missing data and more relevant features. Explained later in Feature Selection, these missing values will not impact our results.

Data Preprocessing

Our target feature **Length of Stay** is continuous. Its values are capped by 120+ which we replaced with 120. APR Severity of Illness and Risk of Mortality are ordinal in nature and rewritten as integers 1 through 4 with 1 as *Minor* and 4 as *Extreme*. For Age Group, its 5 groups are translated into ordinal form in increasing age order. The remaining features are categorical, and the nominal values are converted to numbers using one-hot encoding which creates a column for each possible value and puts a 1 in the applicable column, 0 otherwise.

The following histogram shows the distribution of length of stay in the dataset.



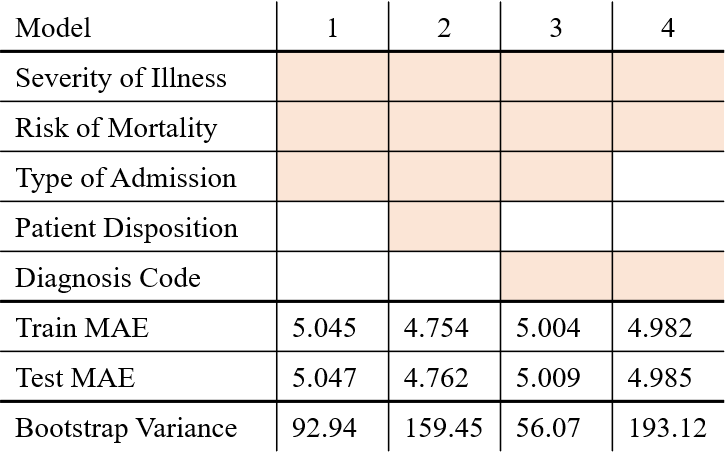
Preliminary Models

We used least squares linear regression to determine which features to include as well as to establish a baseline that can later be used to evaluate the effectiveness of our clusters. Our dataset is split into training and testing sets so that we can measure our models’ prediction accuracy.

Feature Selection

We first manually screened the features. After eliminating those that (1) repeated including the three that had missing entries to avoid collinearity, (2) had over 200 options to avoid being too specific, (3) did not seem relevant like birth weight, or (4) would not be known yet, 12 features remained. Multiple preliminary regression models with a feature left out were then applied to discover that only 5 were significant. 4 different models composed of a combination of these 5 features were then cross-validated to see which was best at predicting length of stay and least likely to overfit.

The results of 5-fold cross validation are summarized in the table below. We wanted to focus on inliers first, so mean absolute error was chosen for its relative insensitivity to outliers.



The last row is the expected variance from out-of-sample error calculated using bootstrap of 500 samples of size 5000. High variance is typically associated with overfitting. While models 2 and 4 have smaller mean absolute error, their variance is much greater than model 3’s. The features we decided to ultimately use are **APR Severity of Illness**, **Risk of Mortality**, **Type of Admission**, and **CCS Diagnosis Code**.

Next Steps

On the entire training set, the train MAE is 5.004, and test MAE is 5.009. The clusters that we generate should aim to have lower errors than these. We will use k-means and random forest to cluster the patients. Expanding on section 5, we will also use random forest classifier and gradient boosted trees. Besides this benchmark, we will check the variance of length of stay and total cost predictions by our clusters. Cross validation and bootstrap (to measure bias and variance) will continue to be used to prevent under- and overfitting. Specifically to prevent linear models from overfitting, we can decrease the number of features and use less complex models. If models are underfitting, we can increase the number of features. We will explore different loss functions and regularizers as well.

Finally, because health data pertains to life and death, there is the danger of our models becoming weapons of math destruction if used incorrectly. The idea is for predictions to aid hospital management in planning, but doctors should not diagnose or treat patients solely based on our estimates.

Clustering Models

We used LowRankModels.jl package to perform k-means clustering on the dataset. k= 3 was chosen as the initial k value. We included APR Severity of Illness, Risk of Mortality, Type of Admission, and CCS Diagnosis Code as features from the feature selection process above. 15000 rows from 35,804 rows were randomly chosen to make a smaller sample of the dataset because low rank model k-means clustering took more than 40 min to run on the entire dataset.

For k-means clustering, huber loss was used for numerical columns, ordinal hinge loss was used for ordinal columns and hinge loss was used for binary columns. Unit one sparse constraint was used. The results of 5-fold cross validation are summarized in the table below. For all fold, k-means clustering had small training error and test error showing the model generalizes. For fold 3 training error is 0.01 smaller than test error, showing that the model is slightly overfitting.

For each cluster and the entire dataset, we predicted the length of stay of each patient. For cluster 1 training error is smaller than test error, showing that the model is slightly overfitting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Train error | Test error | Length of Stay Mean | Length of Stay  Standard Deviation | Rows |
| Entire dataset | 5.094 | 4.968 | 7.329 | 11.162 | 15000 |
| Cluster 1 | 4.121 | 4.550 | 6.291 | 9.599 | 4784 |
| Cluster 2 | 4.934 | 4.907 | 7.241 | 10.613 | 5042 |
| Cluster 3 | 6.039 | 5.986 | 8.404 | 12.864 | 5042 |

Length of stay standard deviation of cluster 1 and 2 are smaller than the entire dataset. However, standard deviation of cluster 3 is greater than the entire cluster. We compared the sum of test error for the dataset without clustering and with clustering.

Entire dataset MAE \*15000 = 5.025 \* 15000 = 74,520

Cluster 1 MAE \* n1 + Cluster 2 MAE \* n2 + Cluster 3 MAE \* n3 = 77,337.43

The sum of length of stay prediction error with a linear model with clustering is greater than without clustering. The goal was to create clusters of homogeneous cancer patients so that linear prediction performs better on different features including length of stay. We can perform k-means clustering with length of stay as the only feature if the goal is to optimize the clusters for length of stay prediction.

Clustering Models Next Step

For k-means clustering, we can choose the optimal k value using the elbow method. We can perform k-means clustering with length of stay as the only feature if the goal is to optimize the clusters for length of stay prediction.

Focusing on inliers first, so using mean absolute errors which is less sensitive to outliers.

|  |  |  |  |
| --- | --- | --- | --- |
| Without feature... | Train MAE | Test MAE |  |
| **All 13** | **4.684** | **4.735** |  |
| APR\_Severity\_of\_Illness\_Code | 4.714 | 4.780 | keep, worse o/w |
| Age\_Ordinal | 4.684 | 4.736 | little influence |
| Gender\_F | 4.684 | 4.733 | little influence |
| APR\_Risk\_of\_Mortality\_Ordinal | 4.719 | 4.774 | keep, worse o/w |
| Zip\_Code | 4.677 | 4.718 | cross-val |
| Race | 5.072 | 5.055 | similar to ethnicity |
| Ethnicity | 5.066 | 5.047 | cross-val |
| Type\_of\_Admission | 5.070 | 5.057 |  |
| Patient\_Disposition | 5.099 | 5.085 |  |
| CCS\_Diagnosis\_Code | 4.804 | 4.857 | cross-val |
| APR\_Medical\_Surgical\_Description | 4.717 | 4.765 |  |
| Payment\_Typology\_1 | 4.680 | 4.724 | little influence |
| ~~Emergency\_Department\_Indicator~~ | ~~4.689~~ | ~~4.738~~ | ~~little influence~~ |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| sev |  |  | mor | zip | rac | eth | adm | dis | diag | surg |  |  | tr | te |
| y |  |  | y |  | y | y | y | y | y | y |  |  | 4.675 | 4.708 |
| y |  |  | y |  | y |  | y | y | y | y |  |  | 4.718 | 4.693 |
| y |  |  | y |  |  | y | y | y | y | y |  |  | 4.679 | 4.668 |
| y |  |  | y |  |  |  | y | y | y | y |  |  | 4.674 | 4.702 |
| y |  |  | y |  |  |  | y |  |  |  |  |  |  |  |
| y |  |  | y |  |  |  | y | y |  |  |  |  |  |  |
| y |  |  | y |  |  |  | y |  | y |  |  |  |  |  |
| y |  |  | y |  |  |  |  |  | y |  |  |  |  |  |
| y |  |  | y |  |  |  |  |  |  |  |  |  |  |  |

Run kmeans on the whole dataset

<https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>

Kmeans not the bet way

Let’s try something else

The number of patients in the frist 3 clusters

4784

5174

5042

From pst report: quantile loss best, l1 and least squared tried

Multiclass classification:

s

TODO : sample with 15000 rows 3 clusters

If still 2 clusters:

1. Check kmeans
2. Check python cluster function, run with previous lifes?
3. Too fit julia linear model, write clustering function in julia….

35000 rows csv take 6:57 - 7:25 still not doen

https://discourse.julialang.org/t/csv-problem-to-write-big-dataframes/17343/12

4c.lusters but 2 clusters empty

Whole 7000 linear gression: MSE Score: 56.23099012042527

7000 rows and 3 clusters 1 empty

Cluster1 : Score: 28.14944856208815

Cluster 2 Score: 34.29473848650537

61.49423176339637

Y

Kmeans init

Set centroid as an outlier

38 done running by 53

Create a dataframes by copying columns 53

Compared linear regression error with Clustering and without clustering

Size of each cluster is not the same : account for the

Without clustering: squared errors for linear regression

After the clsuteirng: squared errors for linear regression for each cluster

Add squared errors for all the clusters and compare with linear model error without clustering(entire patients)

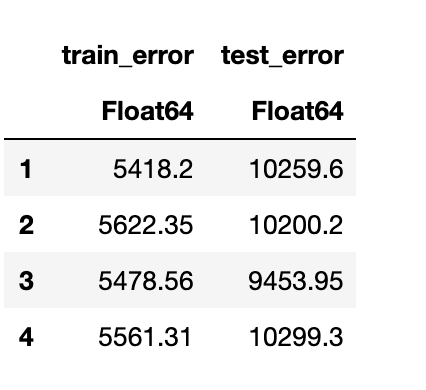
Don’t take the mean, calculate the squared error without errors

I am applying glrm kmeans clustering to the cancer dataset and predicting length of stay in each clusters. I wanted to clarify. When I make the glrm kmeans clusters Should I assume that I know the length of stay when classifying?

If I can assume I already know length of stay, when clustering, can I cluster choosing length of stay as the only feature?

should I include length of stay or not include length of stay?

Glrm different loss function k =3 n=7000 crossvaidation n fold. Not overfitted



Transpose dataframe in julia: DataFrame([[names(df)]; collect.(eachrow(df))], [:column; Symbol.(axes(df, 1))])

<https://stackoverflow.com/questions/37668312/transpose-of-julia-dataframe>

Got values but first cluster is empty

Try with all rows

In the report, you should describe your data set in greater detail. Describe how you plan to avoid over (and under-)fitting, and how you will test the effectiveness of the models you develop. Include a few histograms or other descriptive statistics about the data. How many features and examples are present? How much data is missing or corrupted? How can you tell? You should also run a few preliminary analyses on the data, including perhaps some regressions or other supervised models, describing how you chose which features (and transformations) to use. Finally, explain what remains to be done, and how you plan to develop the project over the rest of the semester.

**Sample report:** [**https://github.com/MattD18/Healthcare-Information-System/blob/master/midterm\_report.pdf**](https://github.com/MattD18/Healthcare-Information-System/blob/master/midterm_report.pdf)

3 parts

1. Jody : length of stay prediction:
   1. cancer: hospitalizable disease
   2. Linear regression with least squared
   3. Picking features, multiple models, current bad scores
   4. Think about cross validation, look up bootstrap
2. Total cost prediction
   1. Cancer:
3. Clustering
   1. Cancer
   2. Low rank model:
   3. Random forest
   4. Evaluate cluster using prediction models 1 and 2

GLRM kmeans clustering

Chose features using random forest regression feature importance / linear regression sklearn

Used random forest regression feature importance as weights of the features

k=4

Kmean init?

Glrm with quadloss

Performed

Glrm with different loss functions

Glrm with different loss functions and weights

Barplot for variance in each cluster and predict length of stay in each cluster

Only 2 clusters

good\_features = [

:Age\_Group, # nominal

:Zip\_Code, # nominal

:Gender, # nominal

:Race, # nominal

:Ethnicity, # nominal

:Type\_of\_Admission, # nominal

:Patient\_Disposition, # nominal

:CCS\_Diagnosis\_Code, # nominal

:APR\_Severity\_of\_Illness\_Code, # ordinal

:APR\_Risk\_of\_Mortality, # nominal

:APR\_Medical\_Surgical\_Description, # nominal

:Payment\_Typology\_1, # nominal

:Emergency\_Department\_Indicator # nominal

]

Pyjulia

Use dataframe in julia in python notebook

[PyCall.jl](https://github.com/JuliaPy/PyCall.jl)

Jupyter notebook switch kernels

<https://github.com/JuliaPy/PyCall.jl>

<https://blog.jupyter.org/i-python-you-r-we-julia-baf064ca1fb6>

10/28 TODO

* Do package update bc glrm package updated
* Then push my kmeans init code
* Csv write very slow, use pycol instead of writing csv files if cross validation fast
* If cross validation slow / subsample 10k rows

Look for julia environments

Package update for

Or package update

If I update julia package, it will force update other packages

<https://julialang.github.io/Pkg.jl/v1.1/environments/> : start new project

Midterm report TODO

1. Kidney cancer dataset / subsampel 1000 rows from original datset
2. Choose features using linear regression weights, highest weights
3. Init\_kmeans for glrm
4. Get the clusters
5. How to prevent overfitting : boots strap cross validation
6. Glrm with different loss function for different features
   1. Weight the loss function with linear regression weights or random forest weights
7. Get regression code working for cancer dataset / fit the cluster to jody’s model
8. Variance within clusters
9. and decision tree prediction within clusters: write my linear regression 3120

Csv write is slow

Final report TODO

Questions to ask Chengrun

1. Not sure how to answer questions for glrm: go to OH
2. Not sure if I should try predicting total cost as well?: talk to teammates
3. What is the grading criteria for the algorithm development project? / does the performance of the clustering matter?
   1. Explain why it works or why it doesn’t work
   2. One model for each clusters
   3. Prediction on each cluster doesn’t do better than entire dataset
   4. Alanyze Why the performance is not good
   5. Random forest importance feature and low rank model different models nto sure if will perform better
   6. When feature importance from linear regression :
   7. Clustering losses:
   8. Most of them are : logistic loss for categorical to do kmeans
   9. Different loss functions for different features
   10. If most features ordinal or categorical; random forest regression,
   11. Explain why I made certain decisions more important than good prediction result, and explaining weakness of rsults
   12. Subsample on cancer dataset
4. How to choose k number of clusters
5. 30k rows can I subsample 5k rows to do the clustering?

1.do data analysis processing

2. Answer questions

3. Fit it into glrm:

Not sure how to answer questions for glrm: go to OH

4. try predicting total cost as well

Heart Attack Linear Models.ipynb

:x discharge identifier column

Questtions to Prof.udell

* Init\_kmeans still error…., do I need to update the low rank model package to accept the change?

What if our clustering models are not useful in prediction? If we justify our approach and explain possible improvements will it satisfy the 4741 algorithm development project requirements?

<https://discourse.julialang.org/t/mutate-a-new-variable-with-row-numbers/30985/2>

<https://discourse.julialang.org/t/mutate-a-new-variable-with-row-numbers/30985/2>