ORIE 4741 Midterm Project Report

Kathy Byun, Jody, Zhu, Raye Liu

November 7, 2020

1 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.

2 Dataset

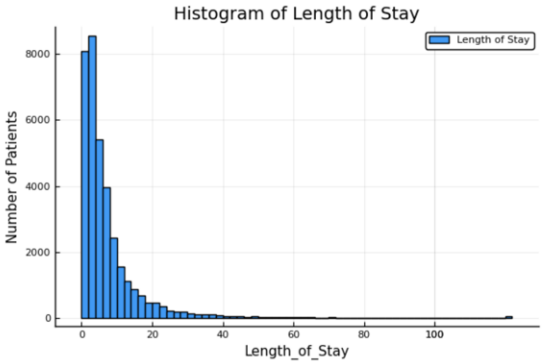
The dataset used in this project is New York State's Statewide Planning and Research Cooperative System's (SPARCS) Hospital Inpatient Discharges. It contains information about patients discharged from hospitals in New York State in 2012. Some of the fields are race, age group, type of admission, diagnosis, 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. As we will see later in this report, these missing values will not impact our results.

3 Data Preprocessing

Our first target feature Length of Stay is continuous. Its values are capped by 120+ which we replaced with 120. The other target feature Total Cost is continuous as well, ranging from 1,562.44 to 2,193,723.15, with a mean of 5,8879.05. 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 histogram in Figure 1 on the next page shows the distribution of Length of Stay in the dataset.

4 Preliminary model: Linear Regression

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.
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




```
Summary Stats:
Length:          35804
Missing Count:   0
Mean:            7.252988
Minimum:         1.000000
1st Quartile:    2.000000
Median:          4.000000
3rd Quartile:    8.000000
Maximum:         120.000000
```

Figure 1: Length of Stay Statistics

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.

Model	1	2	3	4
Severity of Illness				
Risk of Mortality				
Type of Admission				
Patient Disposition				
Diagnosis Code				
Train MAE	5.045	4.754	5.004	4.982
Test MAE	5.047	4.762	5.009	4.985
Bootstrap Variance	92.94	159.45	56.07	193.12

Figure 2: Cross-validation Statistics

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 ultimately decided on for linear regression are APR Severity of Illness, Risk of Mortality, Type of Admission and CCS Diagnosis Code.

5 Random Forest

Random decision forests are powerful methods for classification and regression, and they correct for decision trees' habit of overfitting to their training set, and generally outperform decision trees, but their accuracy is lower than gradient boosted trees. Therefore, we will start with the random forest mode. In this part, we use random forest regressor to predict the total cost of patients as response variable. For features, we dropped Payment Typology 2 and 3 since their influence to the training result is relatively trivial in the presence of Payment Typology 1, as well as they included most of their inputs as null.
To find appropriate hyperparameters of the random forest regressor, we run 4 different models with distinct hyperparameters inputs: 1) bootstrap = True, max_features = "auto"; 2) bootstrap = True, max_features = "log2"; 3) bootstrap = False, max_features = "auto"; 4) bootstrap = False, max_features = "log2". Parameter "bootstrap" stands for whether sampling with bootstrap or not, while "max_features" is the number of features to consider when looking for the best split, if set to "auto", then the all features are considered. Besides all other parameters are default values. Then we run the four models on a different number of trees: n_estimators = 100, 150, 200, 250, 500, 1000. Repeat this process for 10 iterations. The best R^2 is approximately 0.7559, from the bootstrap sampling model with all features included and

250 trees (excluding 50 because random forest handles the bias-variance trade-off better with a relatively large number of trees). We display here the R^2 trend of the 10th iterations:

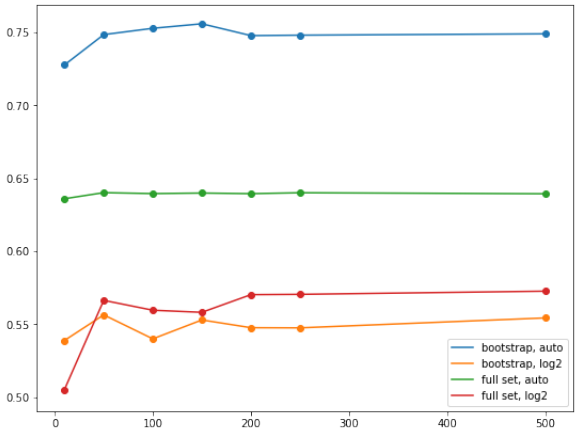


Figure 3: Iteration 10 R^2 Comparison

We also get the feature importance from the random forest model. The top 10 most influential features are displayed below:

	columns	importance
3	Length.of.Stay	0.549700
282	APR.DRG.Code.4	0.124110
374	APR.Medical.Surgical.Description.Surgical	0.045906
6	Hospital.County.Manhattan	0.037793
2	Age.Ordinal	0.014987
9	Type.of.Admission.Elective	0.014565
1	APR.Risk.of.Mortality	0.013860
0	APR.Severity.of.Illness.Code	0.013085
122	CCS.Procedure.Code.50	0.008754
18	Patient.Disposition.Expired	0.007105

Figure 4: Feature Importance

6 Next Steps

On the entire training set for linear regression, 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.