Which hospital to go?

ORIE 5741
Ruobing Shui rs2579
Hehong Li hl778
Ruize Ren rr779

Agenda

- Project overview
- Dataset introduction and pre-processing
- Preliminary analysis and effectiveness
- Model selection
- Future work

Project Overview

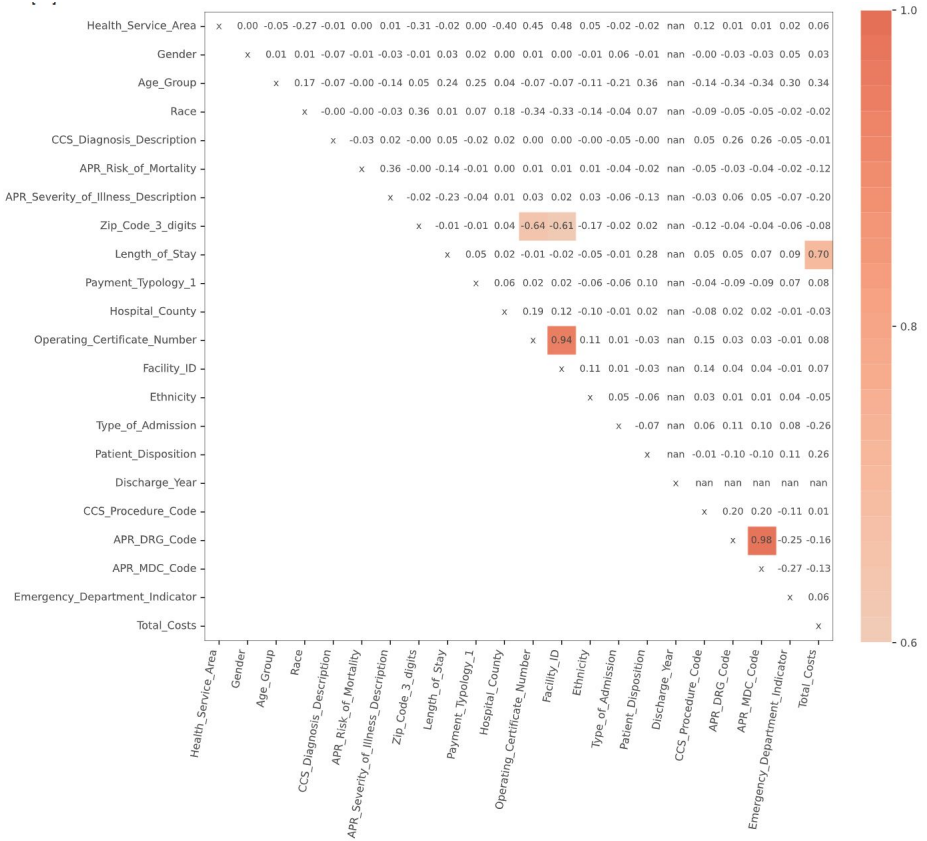
- Background
  - Hard to choose appropriate hospital
    - Cost money
    - Cost time
    - Lack information
- Purpose
  - Assist residents in New York State on choosing the best hospital to go to when they are experiencing medical concerns.
- Outcome
  - Recommend Quantile regression method to generate model to predict which hospital will cost least when people get Septicemia.

Dataset Introduction & Pre-processing

- Dataset description
  - Dataset source: Hospital Inpatient Discharges (SPARCS De-Identified): 2012
  - 2544543 rows and 35 columns
  - Patient and hospital information
- Pre-processing
  - Data cleaning
    - Drop the rows which have null values
  - Feature selection
  - Feature transformation

Dataset Pre-processing

- Pre-processing
  - Data cleaning
  - Feature selection
    - Using random forest model
    - Result: 5 highest correlation features
      - Length of stay
      - Age group
      - Patient disposition
      - Facility ID
      - Operating certificate number



Dataset Pre-processing

- Pre-processing
  - Data cleaning
  - Feature selection
  - Feature transformation
    - Categorical Data
      - Apply the one-hot method to all the categorical data except multiple payment method.
      - For multiple payment method solutions, we use the many-hot method to label the feature.
    - Miscellaneous Data
      - The length of stay data is hybrid with numerical and text data.

Preliminary Analysis

- Train and Test Split
  - 80% → Train dataset, 20% → Test dataset
- Linear least squares function
- Analysis:
  - Too much categories
    - Shrink dataset→ Diseases: Septicemia
  - More complicated model
    - Quantile, Ridge regression

Data type	Train_MSE	Test_MSE
many_hot	2.8e8	4.4e8
one_hot	3.5e8	5.1e8
numerical	4.8e8	6.8e8
many_hot+one_hot	3.5e8	5.1e8
many_hot+one_hot+numerical	6.9e8	4.9e8

Table: The results of training error and test error

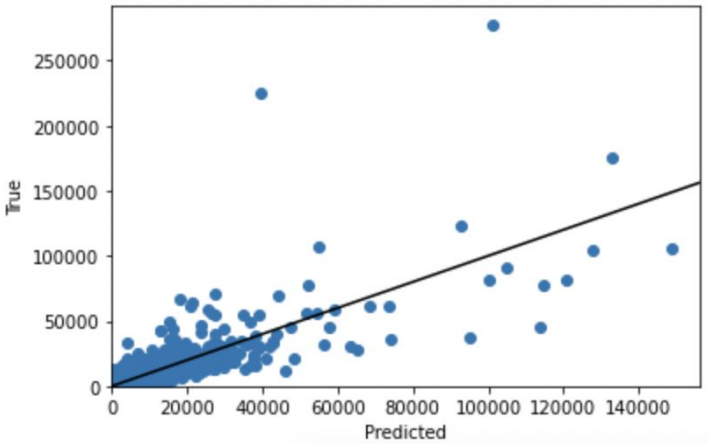


Figure: Result for all combined features

Model Selection

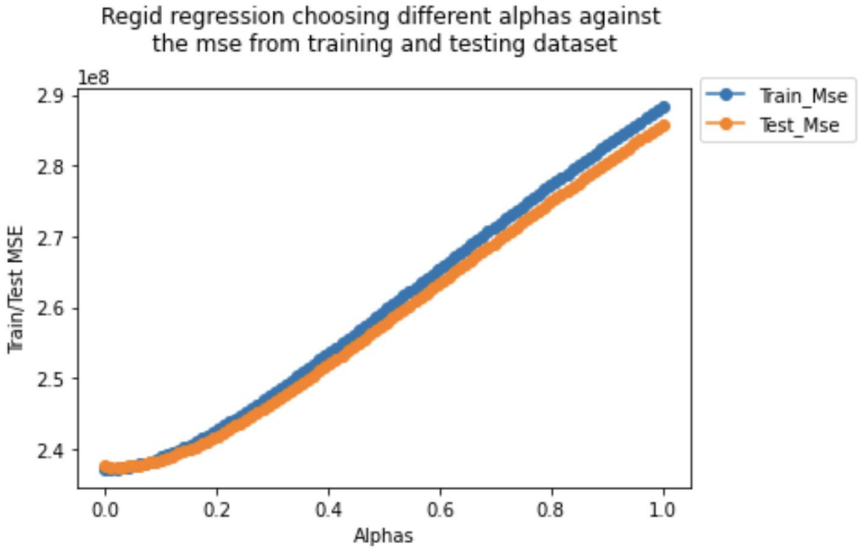
Quantile Regression

- Top 3 features - Length_of_Stay, Patient_Disposition, Operating_Certificate_Number
- Disease Selection - Septicemia
- Randomly drop 99% data points
- Best model evaluated by Test_MSE: quantile = 0.42
- Train_MSE: 2.5e8
- Test_MSE: 2.3e8

Model Selection

Rigid Regression

- Top 5 features
- Disease Selection -> Septicemia
- Best Alpha=0.041
- Best model evaluating Test_MSE: alpha = 0.041
- Train_MSE: 2.4e8
- Test_MSE: 2.4e8

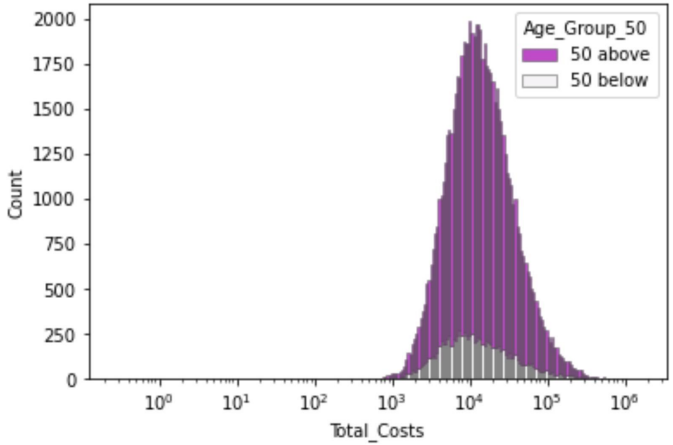


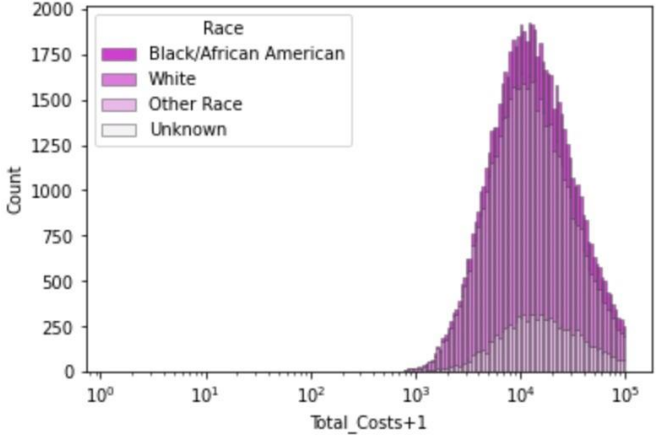
Linear least squares < Rigid Regression < Quantile Regression (best)

Results and Analysis

- Linear least squares (poor) < Rigid Regression < Quantile Regression (best)
- Significant Feature: length of stay, operating certificate number, patient disposition, facility id , ccs () procedure code.
- Insignificant Feature: gender, ... individual properties
- Benefits
    - Patients
    - Medical treatment environment

Fairess metrics





Group	25%	50%	75%	Mean total cost
Age below 50	5720	11059	24293	24664
Age above 50	7095	13023	25179	22600
Black / African American	8656	16412	32232	28646
White	6299	11525	22173	19961
Other Race	8535	15867	31962	29000

Limitation and Future Work

- Limitation:
  - The dimension of the consideration is single
  - There are information bias in the dataset and a lot of outliers.
  - The dataset lack the specific hospital names.
  - The computer capacity is not big enough for us to run the whole dataset when doing preliminary analysis.
- Future work
  - Consider the quality of medical treatment and success rate to give more fitable recommended hospital given by the patients' information.
  - Segment each disease in the dataset and put the segments into three kinds according to the emergency level.
  - Using computer with larger capacity.
  - Making a user friendly website or online intelligent chat bots to launch our model into real products to help residents.