New York State Inpatient Healthcare Data Analysis and Cost Prediction

Justin Yong Sik Cho
Columbia University
yc3522@columbia.edu

Yuan-Fang Lin
Columbia University
yl4042@columbia.edu

Jing Qian
Columbia University
jq2282@columbia.edu

Abstract

Healthcare in the United States is increasingly becoming the center of national attention especially due to its quick growth in spending. In fact, healthcare spending in the US grew 4.6% in 2018 surpassing $ 3.6 trillion or $11,172 per person. With 17.7% of the country's Gross Domestic Product spent on healthcare, monitoring the vast amount of data on how people incur costs is increasingly becoming an area of interest. Big data analytic technologies in the modern era enables users to sift through the high volume of healthcare records and extract insights that point to systematic patterns or leverage machine learning to build useful applications. Previous research aimed to test for bias in various fields ranging from healthcare practices to police violence. These works provided guidance on which features to explore to test for bias, while shedding light on potential limitations and difficulties such as data-resolution and underlying distributions that can complicate analyses. Our project aims to build on these findings by....[add point on where our work adds values yc3522 yl4042]. Our research aims to close this gap while focusing on a subset population in New York State. We perform experiments to test for systemic bias across sensitive class labels, identify potential patterns and clusters in the inpatient population, and finally develop a prediction tool which helps users estimate the cost of healthcare given general user information. We make our data, models, and code available on Github[1]*. Through statistical analysis, we found that race, gender and age bias exist significantly in the discharge costs per day and the bias vary from county to county.*

1. Introduction

Healthcare spending over the past decade has grown at a considerable pace in the United States. Especially in New York state, the annual average rate of growth in per-person spending of health-care from 2013 to 2017 was 6.2%, compared with 3.9% nationally. The interest in healthcare has also soon ballooned into the point of popularity. In order to better understand the trends in healthcare, our work focuses on a relatively high-resolution data set based in New York State inpatient discharge records, which offer valuable information such as patient demographics, healthcare geographies, and healthcare cost, all without compromising privacy laws. More specifically we clean the data set, employ statistical tools to select key features, and test for biases among the variables to identify if there are systemic patterns of unequal treatment and costs across different independent groups. We use ANOVA to find significant features that contribute to the mean discharge cost. With ANOVA, we could figure out wether there exist race, gender and age bias in mean discharge costs. Once we find the bias features, we will Tukey method to perform multiple-comparison test and find whether there are significant difference among groups of these bias features. We also aim to develop a tool to help predict healthcare costs given basic demographic and geographic inputs. This tool is designed to help potential patients with limit information about diagnosis and medical information to gain insight into what healthcare costs can potentially amount to. Finally, we package our analyses in a Django web application to enable user interactions, inputs, and visualization of results.

2. Related Works

Bias is a prejudice against any one of the things in the world, no matter person, or group compared with another usually in a way that is considered to be unfair. Bias may exist toward any social group. One's age, gender, gender identity physical abilities, religion, sexual orientation, weight, and many other characteristics are subject to bias. One particular type of bias is called unconscious bias, also known as implicit bias. Unconscious biases are social stereotypes about certain groups of people that individuals form outside their own conscious awareness. The paper, *Implicit Racial/Ethnic Bias Among Health Care Professionals and Its Influence on Health Care Outcomes: A Systematic Review* [**?**] provides a useful summary of investigation to which implicit racial/ethnic bias exists among healthcare professionals and examination the relationships

[1] https://github.com/justinchoys/nyhealth-bigdata

between health care professionals implicit attitudes about racial/ethnic groups and health care outcomes. The research results provide insights into healthcare practices and suggests that most healthcare providers appear to have implicit bias with positive attitudes towards White patients and negative attitudes toward people of color.

Another point we focus on in our project is how geography relates to the bias in healthcare costs. *A Multi-Level Bayesian Analysis of Racial Bias in Police Shootings at the County-Level in the United States, 20112014* [**?**] utilizes a geographically-resolved, multi-level Bayesian model to analyze the extent of racial bias in the shooting of American civilians by police officers in recent years. The research suggests that aside from easily observable bias across Race/Ethnicity or Armed Versus Unarmed by Race/Ethnicity, county-level data are far too coarse to reliably tease apart the conditions that drive racial bias in police shootings. This suggest that although bias may exist in reality, limitations in data and privacy concerns can hinder the ability to find more reliable evidence of such bias.

To predict healthcare charges, *Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation* [**?**] implements several learning methods, including Lasso, Ridge, Elastic Net, Random Forest, Support Vector Machine, Gradient Boosting and Artificial Neural Network to review the literature of healthcare cost prediction. One observation that is highlighted is that [yl4042 what is cost on cost? ¿¿¿] cost on cost prediction performs as well or even better than predictions using clinical data or clinical and cost data. The researchers also point out that the prediction performance based on statistical model is further complicated by the skewed nature of healthcare data, where distributions are strongly skewed with different bias and extreme values can be present, all of which make it inefficient to use small to medium sample sizes if the underlying distribution is not normal. As a result, we chose to first focus on investigating the distribution of data and the multiple bias that exist in the data set before proceeding to implement a prediction model for potential healthcare-charge using high-level user information.

The rest of this report is structured as follows: In Section 3, we describe the data with its features and the preprocessing techniques used to reduce the size of data, and in Section 4, we present the principal methods used to analyze and test for bias in our data set as well as the approach and rationale behind our prediction model. In Section 5, we report on the bias results we found among the data set and several simple comparison of performance of our prediction models. Lastly, Section 6 and 7 covers the software architecture and tech stacks of our application and discuss some potential future works and bottlenecks.

3. Data

The data used was sourced from the New York State Hospital Inpatient Discharges (SPARCS De-Identified): 2017[2] data set which contains detailed patient discharge characteristics, diagnoses, treatments, services, and charges. While the record-level data points represent individual patient discharge instances, the data is already de-identified and so does not contain protected health information under HIPAA. The records are not individually identifiable as well through redaction or aggregation at a lower-resolution scale (e.g. age is categorized into groups rather than specific numbers). Overall, there are 2,343,569 records in the 2017 data set, with each row made up of 34 features including age, gender, zip code, ethnicity, race, and diagnosis.

3.1. Data Preprocessing

To ensure consistency of units across all data records, the first step in data preprocessing was normalizing the healthcare costs by the length of stay for each record. This was driven by the motivation to ensure meaningful comparisons across individual entries and also to ensure that all samples were not missing these features. The next step was dropping rows with missing values in certain columns. While most of the records were not missing values (excluding payment types which had anywhere from one to three payment types per record), we specifically elected to drop records that were missing 'Hospital Service Area', 'Zip Code - 3 digits', and the 'APR Risk of Mortality' fields because the records with missing data of these three columns are no more than 2% of the total data set while the missing input of payments take around 40% of the total.

We also removed columns containing irrelevant and collinear/repeated information. The columns we dropped include 'Discharge Year', 'Abortion Edit Indicator', 'Operating Certificate Number', 'Permanent Facility Id', 'CCS Diagnosis Code', 'CCS Procedure Code', 'APR MDC Code', 'APR Severity of Illness Code' and'APR DRG Code'. We specifically elected to remove the abortion edit indicator because there are no abortion records if we remove the rows with NA values and payment types. The remaining dropped columns were removed due to collinearaity. For example, columns beginning with 'CCS Diagonosis' contained one-to-one correspondences between codes and descriptions. In order to simplify the data set we dropped the codes while keeping the descriptions because the latter is more human-readable.

[2] https://on.ny.gov/2qa8Qlm

3.2. Data Exploration

3.3. Feature Selection

The third step is feature selection driven by data exploration and analyses. Once the data was pared down to 20 columns, there were some interesting patterns in the data. Some columns were numerical, like length of stay while some columns were categorical with cardinality in the hundreds, which hinders the modeling process. To tackle this issue, we factorized all the categorical variables and computed the Pearson correlation matrix among all the variables to inform the feature selection process in Table 9 and Table 10.

The correlation between length of stay and total charges was relatively high around 0.7, which motivated our decision to normalize healthcare costs by the length of stay and use this as the dependent variable. Other than gender, race and age, location of the hospitals also proved to play an important role in predicting costs. We found moderate correlation between and among geographical variables and facilities related variables. Among all these variables, 'Hospital County' had a moderate number of categories and its count distribution was relatively even, which convinced us to use this variable as a representative feature that captured all relevant geographic information. We also expected the diagnosis features to be useful in predicting healthcare costs given. This is because complex diagnoses and procedures are typically associated with higher costs (e.g. various types of cancer) as opposed to simpler diagnoses and procedures such as the common cold. The data set had four medical diagnosis-related variables with each over 200 categories, which made it difficult to model. However, after we found that there was moderate correlation among these variables, we choose 'CCS Diagnosis Description' to represent all medical diagnosis-related information given that this feature was easy to understand and also because it was easy to filter. We also noticed that some diagnoses categories had low frequencies (i.e. rare or infrequent diagnosis) while only a few diagnoses had relatively high frequencies. Because of the disparity in data availability among diagnosis, we decided to limit our bias analysis to the top diagnoses categories.

4. Methods

For the bias analysis, we used a widely used method in statistics, ANOVA. ANOVA refers to analysis of variance, which is based on the law of total variance, where the observed variance in dependent variable is partitioned into componenets attributes to different sources of variation. Specifically, how much of the variance in dependent variable could be explained by the variance between groups in a certain feature and how much could be explained by the variance within groups. For example, while controlling the effects from other variables, if at least two racial groups have the mean costs that statistically significantly different from each other, which means that the variance between racial groups is so strong that race must be a significant contribution to the mean discharge costs. Here significant difference means that the probability that those group-mean difference are due to random chance is less than a pre-specified threshold (we take the conventional threshold, which is p-value = 0.05).

There are several test methods in the family of ANOVA, like MANOVA for multiple dependent variables, ANCOVA for covariance analysis among continuous independent variables. In our case, considering the independent variables are mainly categorical and there is only one dependent variable, we use ANOVA. In ANOVA, a regression model is used to fit the relationship between independent variables and dependent variable. We used R to do the ANOVA bias analysis. We used a linear regression model to fit the mean discharge costs per day and take ten features from previous data preprocessing as independent variables: Hospital County, Age Group, Gender, Race, Ethnity, Type of Admission, Patient Disposition, APR Severity of Illness Description, APR Risk of Mortality and APR Medical Surgical Description. R will print out the p-value of each feature from ANOVA for us using the command $aov(lm(y \sim f(x)))$.

Once we identified what types of biases exist in features (e.g. gender, race, and age) , we explored how the mean charges differ among different groups of the features and tested these differences for statistical significance. As mentioned above, if any two groups of a feature have significant group-mean difference between each other, ANOVA would recognize this feature as an significant feature. For example, if there exists racial bias in the mean discharge costs, we know that at least two racial groups have significantly different mean discharge costs. But what are these racial groups? What do the differences look like? To compute this, we required a multiple-comparison test.

The basic idea of multiple-comparison tests is to compute pairwise comparisons between different groups in the variable. It is essentially a t-test, except that it corrects for family-wise error rate because we have multiple groups in the feature instead of just two in t-test. There are several statistical multiple-comparison test methods, like Bonferroni method, Scheffe Method and Tukey's HSD (honestly significant difference) method. The major difference among these multiple-comparison tests is the way they use to correct for family-wise error rate. Here, we used Tukey's HSD (honestly significant difference) method, which is a single-step multiple comparison procedure and statistical test. We choose Tukey's method because it outperforms the other two when the number of groups gets large and also performs well when the number of groups is small. In R, we perform Tukey's test after we get the result from ANOVA

using the command $TukeyHSD(aov())$.

After bias anlysis on the New York State as a whole, we also wonder whether there exists bias among these features especially in race, gender and age for each county, and whether each county has different bias pattern. It seems reasonable considering we found low-to-moderate correlation between race and county in the data-preprocessing part. Also, we found that county is actually a significant contribution to the mean discharge costs. Therefore, we do the whole bias anlaysis for each county the way we do for the New York State.

After choosing to focus on the five categorical features, we first decided to analyze the data through clustering. We randomly picked 100,000 sample points and implemented the K-means algorithm and also the tSNE method for visualization. As expected, in Figure 1 and Figure 2, due to these extremely sparse features and presence of many outliers with extreme values, the within-set-sum-of-squared-error is unacceptably large. We could not claim that there is a stable or meaningful clustering of the sample data, even after using the 3D-tSNE method preceded by a Principal Component Analysis (PCA) step. However, when clustering on a smaller sample size, we can still visually observe clustering as shown in red, light green and dark blue regions in Figure 3.

Another component of our project involved building a prediction model to estimate healthcare costs per day using only high-level user information. This application is designed to be useful in real-world scenarios such as for potential patients who wish to estimate healthcare costs prior to a hospital visit and diagnosis. Leveraging our results from bias detection, we elected to include race, gender and age group information as key input variables that can reliably predict healthcare costs. Taking into account applicability to real-world scenarios that often come with limited information, we decided to add two more features as inputs: 'Zip Code' and 'Type of Admission'. Due to the categorical data type that dominated our features as opposed to the numeric label (healthcare cost per day), we determined that a decision tree model was the best candidate model type to use due to its ability to handle large amounts of categorical inputs. We utilized one-hot encoding to vectorize our categorical input features ('Race', 'Gender', 'Age Group', 'Zip Code', and 'Type of Admission') which resulted in an extremely sparse feature vector. The sparse feature vector implied that a neural network model for the prediction task would be difficult unless the network had sufficient depth and additional network features such as penalizing results if insufficient neurons fired in the training phase. To avoid this pitfall, we instead relied on using gradient boosting techniques, which produces a prediction model with an ensemble of weak prediction models, which typically comprise of decision trees. As an experiment, we tested another approach for the prediction task using a simple artificial neural network despite our initial concerns, which surprisingly yielded moderately better results with almost a 50% reduction in RMSE over the Gradient Boosted Regressor model on the same sample data set. Table 1 shows the resulting RMSE differences between the Gradient Boosted Regressor and two neural networks with different number of layers. We used stochastic gradient descent as the optimizer with a learning rate of 0.1. The neural networks were configured to be a fully-connected network with each layer followed by relu activation functions. The 3-Layer neural network has two hidden layers which respectively has 32 and 8 neurons and the 5-layer neural network has four hidden layers which in each has 32, 16, 8, and 4 neurons.

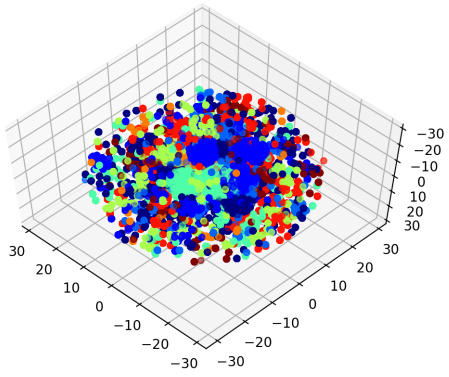

Figure 1. 100K Data visualization with PCA and 3D-tSNE

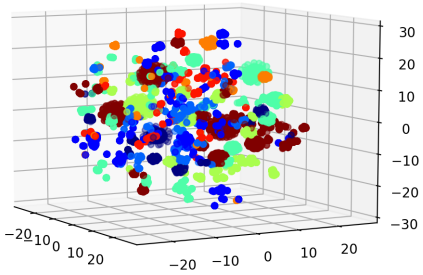

Figure 2. 10K Data visualization with PCA and 3D-tSNE

5. Experiments

We started our data experiments with the bias analysis as mentioned in the Methods section. As mentioned above, after feature selection, we successfully reduced our

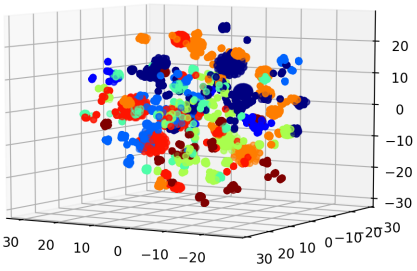

Figure 3. 1K Data visualization with PCA and 3D-tSNE

Variable	DoF	Pr(>F)
county	56	<2.2e-16 ***
age	4	5.4e-10 ***
gender	1	5.9e-09 ***
race	3	<2.2e-16 ***
ethnity	3	<2.2e-16 ***
admission	5	4.7e-14 ***
disposition	18	<2.2e-16 ***
severity	3	<2.2e-16 ***
mortality	3	<2.2e-16 ***
surgical	1	<2.2e-16 ***

Table 1. ANOVA result for Septicemia in the New York State. Significant codes: '***' 0.001.

feature numbers to 10 (Hospital County, Age Group, Gender, Race, Ethnity, Type of Admission, Patient Disposition, APR Severity of Illness Description, APR Risk of Mortality and APR Medical Surgical Description).

We know from common sense that diagnosis and corresponding medical procedures would play an important role in the mean discharge costs. However, the large number of groups in the diagnosis hinders it from being an independent categorical variable for predicting the mean costs. One way to solve this problem is to use hierarchical modeling, which means we first split the categories into several large groups, model and do bias analysis and then the large groups are further broken down into low-level small groups and do the modeling and analysis correspondingly. This method needs deep medical knowledge and in fact, we don't need such intense information to do the bias analysis. Our aim is to find whether bias existing in the mean cost and find significant features as candidates for our predictive model. For the business application, the users won't have the professional diagnosis knowledge to use the predictive model. Therefore, instead of a comprehensive diagnosis hierarchical model, we use a relatively simple approach: do the bias analysis for individual diagnosis category. Considering the diagnosis distribution in the Data Exploration section, this approach is reasonable actually. Although there are over 200 groups in the diagnosis feature, most of the groups have very low frequency.

In the bias analysis, we experimented with top tour diagnosis categories: "Liveborn" (> 9% of the whole data), "Septicemia" (∼5%), "Hypertension" (∼3%) and "Osteoarthritis" (∼3%). It is worth noticing that the patients of 'Liveborn' are newborns, not the mothers. There is only one age category: between 0 and 17 and hence no possible age bias. Therefore, for Septicemia, Hypertension and Osteoarthritis, we did ANOVA with 10 features mentioned above while for "Liveborn", we used 9 features without "Age Group".

In Table. 1, we show the ANOVA result for Septicemia in the New York State. The first column is the 10 variables, the second column is the degree of freedom (DoF) of each variable (DoF equals to the number of groups in a feature minus one. For example, there are 5 age groups and hence the DoF of age is 4.) and the third column is Pr(>F) which is the p-value associated with the F statistic of corresponding variable. As mentioned in the Method section, the null hypothesis that the corresponding variable has no effect on the independent variable (here is the mean discharge cost) is evaluated with regard to this p-value. Since we choose p-value as 0.05 by convention, if a variable has Pr(>F) less than 0.05, we would consider this variable as a significant feature of predicting the mean discharge cost. We could see that for the mean discharge cost of Septicemia in the New York State, all 10 features are significant. Gender, racial, age and geographical/county bias do exist in the mean discharge cost of Septicemia in the New York State. We found similar bias in both diagnosis categories Hypertension and Osteoarthritis. In Liveborn, although there is no age feature and hence no age bias, other 9 features all turned to be significant to the mean discharge cost.

Since there exists gender, racial, age and geographical/county bias in the mean discharge cost of various diagnosis categories in the New York State, we continued the analysis with Tukey's method to find the difference among groups. We started the multiple comparison test on the Liveborn diagnosis regarding race, gender and county (no age, as explained above). The averaged mean discharge cost on Liveborn among all people are around $4300. We found that the female babies spent significantly less money on discharge than male. As to counties, some counties have significant difference while some are not. The difference between any two racial groups are significant, and ordering from low cost to high cost are: white people, black/African American (white+$70), other races (white+$652) and finally multi-racial (white+$2277). More clarifications on the race of "other races" and "multi-racial" are needed to under-

County	White	Black	Other	Multi
Westchester	100	114.76	100	100
Kings	100	100	111.54	130.57
Manhattan	100	89.14	91.59	111.72
Onondaga	100	100	100	100
Bronx	100	100	165.25	143.13

Table 2. Mean dischage costs per day on "Liveborn" of different racial groups inferred from Tukey's method. Take group "White" as the baseline ($100) and scale the cost for other groups. Only show 5 counties for brevity.

County	Male	Female
Westchester	100	92.52
Kings	100	95.06
Manhattan	100	95.93
Onondaga	100	95.04
Bronx	100	96.61

Table 3. Mean dischage costs per day on "Liveborn" of different gender groups inferred from Tukey's method. Take group "Male" as the baseline ($100) and scale the cost for the other group. Only show 5 counties for brevity.

County	$\geq$70	0 - 17	18 - 29	30 - 49	50 - 69
Westchester	100	174.16	114.23	115.71	110.54
Kings	100	81.03	100	100	100
Manhattan	100	100	93.97	96.21	100
Onondaga	100	137.39	100	117.75	115.1
Bronx	100	100	88.39	92.11	100

Table 4. Mean dischage costs per day on "Septicemia" of different age groups inferred from Tukey's method. Take group "$\geq$ 70" as the baseline ($100) and scale the cost for other groups. Only show 5 counties for brevity.

stand why the difference is so large.

As shown by the correlation matrix in Data Preprocessing section, there is low-to-moderate correlation between race and county. Also, previous ANOVA showed that county is a significant feature to the mean discharge cost. If we do our analysis with the New York State as a whole, some geographical patterns may get smoothed. We would like to know how the racial, gender and age bias vary from place to place. Therefore, we filtered our data by its county and did the ANOVA and Tukey's method for each county. However, when we did the bias analysis for each county, a new problem appeared: because of the data splitting, some of the counties may lack feature diversification and hence the original model would not work. For example, in county Otsego, there is no female Liveborn records in the state health system in 2017. Then gender should not be considered as a feature if we do a regression model for the Liveborn data in Otsego. Similar situation appear on features Surgical, Ethnity and others. To make a consistent comparison of the bias across all the counties, we used a reduced model which includes 5 features (Gender, Race, Patient Disposition, APR Severity and APR risk of Mortality). For the diagonsis Septicemia, Hypertension and Osteoarthritis, we acutally had 6, with an extra feature "Age Group".

The output of Tukey's method in R is the difference between two group means and the p-value of this difference. For better understanding the difference between groups in features, we showed the scaled cost instead of the difference from Tukey. Table. 2 showed the mean discharge costs per day on "Liveborn" of different racial groups inferred from Tukey's method. We took group "White" as the baseline and set the mean discharge costs for White to $100 per day. We took the pairwise comparison result between three other groups: "Black", "Other" and "Multi" with "White" from Tukey's method's output. If the difference is significant, we scaled the mean discharge costs of that group regarding White is $100. For example, the scaled black = original black/original white * 100 if there is significant difference between "White" and "Black". If the difference is not significant, we just set the scaled value as $100. If there is no data in certain group in the county, we set the value as $0. We did the analysis for all the counties and just show random 5 in this table. We could see from Table. 2 that each county did have different racial bias. In county Onondaga, we did not observe significant difference between other three racial groups and white people in "Liveborn" in 2017. In Westchester, for every $100 white people spent on Liveborn, black people spent $115 while in Manhattan, black people only spent $89.

Similarly, in Table. 3, we showed the gender bias in "Liveborn" for each county. Here we take male as the baseline. The pattern is quite homogeneous here: female spent less money on mean discharge costs in all the counties. If we look into other diagonsis, the same bias exist too. In fact, some counties don't have female records for certain diagnosis in 2017. It is an interesting phenomenon. Does it suggest that female are healthier? Or do the health facilities tend to perform excessive treatment on males? Or there is undertreatment on females? We may need more research to get this answer.

To illustrate the age bias in the mean discharge cost, we chose the second top diagnosis "Septicemia" because there is no age bias in the top one "Liveborn". Using the similar calculation way, we could see that the age bias differ from county to county.

We have included in Table 1 a summary of the performance between the two approaches to implementing a healthcare cost prediction model. The results indicate that in contrast to our expectations, neural networks yielded better performance as shown by the significantly lower RMSE value. All models used in the experiment are trained on the

	GBTRegressor	NN(3-Layer)	NN(5-Layer)
RMSE	12401.1	**6939.7012**	6940.5825

Table 5. RMSE of GBTRegressor (maxDepth=10, maxIter=30), 3-Layer Neural Network and 5-Layer Neural Network. All trained on 1M data sample with 80/20 train-test split

	NN(3-Layer)	NN(5-Layer)
lr = 0.1	6939.7012	6940.5825
lr = 0.05	6939.1870	6939.2915
lr = 0.001	**5785.8926**	10235.0606

Table 6. RMSE of neural network models with 60 epochs, 32 as batch_size and SGD as optimizer with varying learning rates

	NN(3-Layer)	NN(5-Layer)
SGD	5785.8926	10235.0606
Adam	**5705.9321**	9835.3730
AdaGrad	6492.8960	6149.6025

Table 7. RMSE of neural network models with 60 epochs, 32 as batch_size and 0.001 as learning rate with varying optimzers

same subset of 1 million records randomly sampled from the SPARCS 2017 data set. As we can see, neural network depth does not appear to have much impact on prediction performance. This may be driven by the sparsity of the vectorized features. Without a truly deep neural network it does not seem likely that incremental increases in network depth will result in significant performance improvements. We also conducted experiments to test the performance of the neural networks using different hyper-parameters, including learning rate and type of optimizer. Table 2 summarizes the results from varying the learning rate hyper-parameter and suggest that learning rate can significantly affect the supervised learning process and resulting performance. Similarly, Table 3 shows the the result from varying the optimizer, indicating that the Adam optimizer yielded the best performance for the 3-layer network. Finally in Table 4, we show the impact of varying maximum tree depth and iteration on the Gradient Boosted Regressor model. For all models, the selected features were vectorized into 63 ($\approx 2^6$) dimensional records, which explains how the model with maximum depth 10 tended to perform better.

6. System Overview

As shown in Figure 5, we created a front-end website which enables users to visualize various results based on fields and features of their choosing. There are a total of three tabs Home, Dashboard, Prediction and Heatmap, the latter three which focus on the findings and results of our project. Figure 6. illustrates the Home tab which includes a simple introduction to our website application and the data set. For the remaining three tabs, the web application retrieves the information collected from the user and submit a SQL or PySpark Job to BigQuery or Google Cloud Platform Cluster respectively. The Prediction tab, which uses the prediction model, is mainly driven by a model we trained using the SPARC 2017 dataset. Once the user submits the relevant information, ... [need to rephrase what you are trying to say here yf4042] benefited from PySpark Machine Learning model for Gradient Boosting Regressor and PyTorch for neural network. Figure 7 shows how the Dashboard tab provides users an overview to the data across different user-selected fields. The pie chart and bar chart update in real time based on cursor location over bars/pies and therefore provide an interactive experiencing in analyzing the data. Figure 8 covers the Prediction tab which enables users to input their own information to get two values. The first value returned by the web application is the simple average healthcare cost per day queried from the data set. The second value returned is the predicted result according to our model, which can handle instances and inputs where the SPARC 2017 data set have no historical records to refer to. Finally, Figure 9 demonstrates the Heatmap tab which is a user-friendly visualization of New York state by counties, colored by the differences in average charges per day across user-selected fields such as race, gender, or age groups for a particular diagnosis. One of the potential bottlenecks to our application is scalability. The current design is able to retrieve the model predicted healthcare cost by submitting a job to a cluster on Google Cloud Platform. This presents a potential scalability issue where multiple users may submit multiple jobs and thus cause significant delays in cluster computing as the cluster is not elastic. In order to alleviate this problem, a potential solution may be to scaling out the clusters and leveraging a load-balancer to separate and direct jobs to idle or available cluster instances.

7. Conclusion

[jq2282 add results]

8. Appendix

1. Correlation Matrix

Training Data Size	Maximum Depth	Maximum Iteration	RMSE
10K	10	30	9679.85
	10	50	**8890.98**
	20	50	10207.5
	20	30	10091.7
100K	10	30	12049.1
	10	50	12574.1
	20	50	**11889.6**
	20	30	12182.7
1M	10	30	12688.1
	10	50	**12167.1**
	20	50	12609.1
	20	30	12293.9

Table 8. RMSE for different hyperparameters of Gradient Boosted Regressor

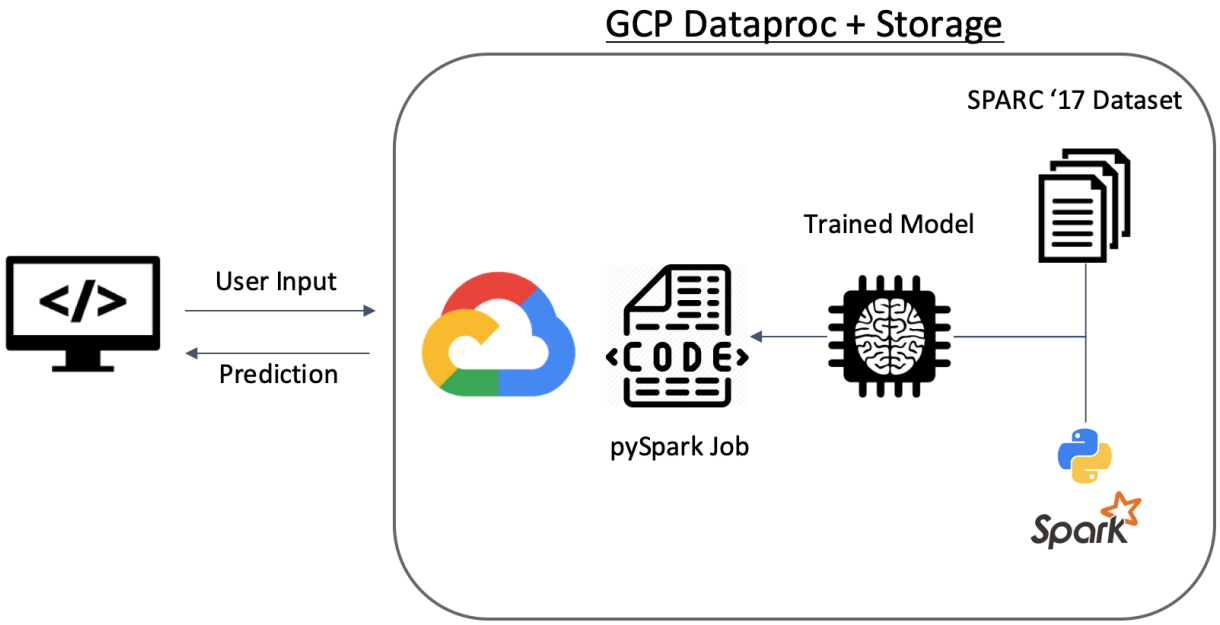

Figure 4. Application/Software Architecture

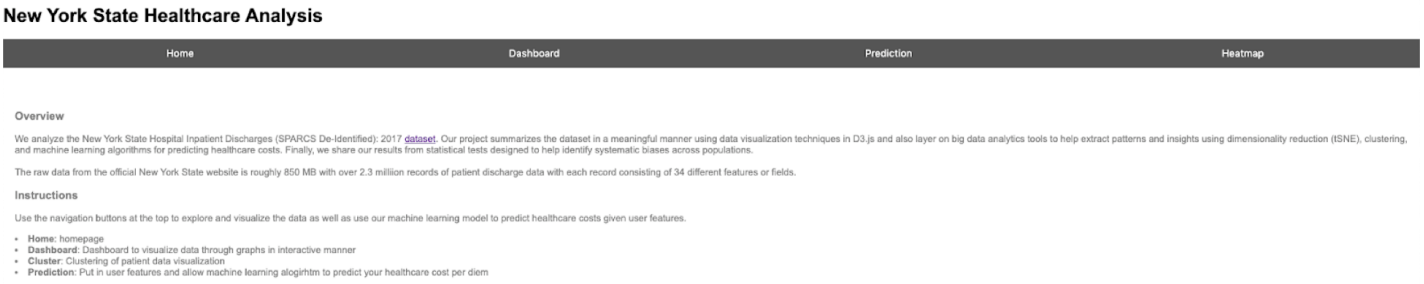

Figure 5. Homepage

Figure 6. Dashboard tab for data overview

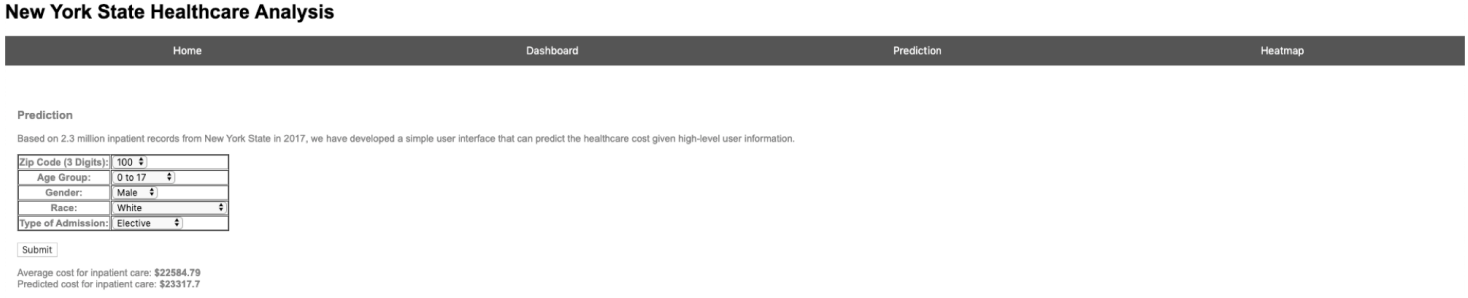

Figure 7. Prediction tab for potential healthcare charge according to high-level user information with our model

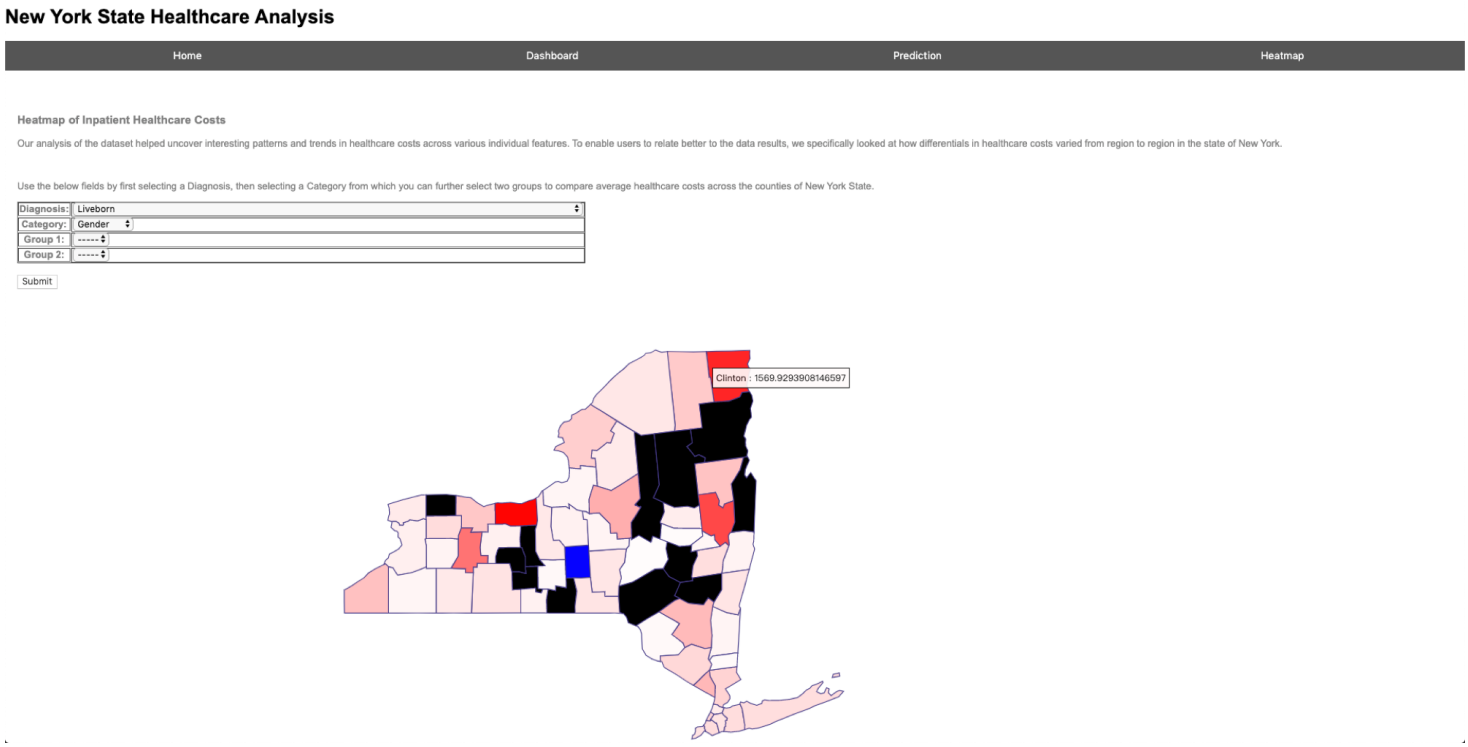

Figure 8. Heatmap tab for user-friendly visualization

	Area	County	Facility	Age	Gender	Race	Ethnicity	Admission	Disposition	Diagnosis
Area	1.0000	0.4459	0.2133	0.0107	0.0065	-0.2907	-0.1767	0.0034	0.0106	0.0099
County	0.4459	1.0000	0.4288	0.0161	0.0043	-0.2904	-0.1389	0.0207	0.0018	-0.0184
Facility	0.2133	0.4288	1.0000	0.0189	-0.0072	-0.1464	-0.0932	-0.0155	-0.0132	0.0082
Age	0.0107	0.0161	0.0189	1.0000	-0.0338	-0.0179	0.0322	0.2041	0.0098	-0.0683
Gender	0.0065	0.0043	-0.0072	-0.0338	1.0000	0.0091	0.0014	-0.0249	-0.0397	0.0062
Race	-0.2907	-0.2904	-0.1464	-0.0179	0.0091	1.0000	0.3662	0.0355	-0.0506	-0.0196
Ethnicity	-0.1767	-0.1389	-0.0932	0.0322	0.0014	0.3662	1.0000	0.0651	-0.0397	-0.0273
Admission	0.0034	0.0207	-0.0155	0.2041	-0.0249	0.0355	0.0651	1.0000	-0.0639	-0.1522
Disposition	0.0106	0.0018	-0.0132	0.0098	-0.0397	-0.0506	-0.0397	-0.0639	1.0000	0.0319
Diagnosis	0.0099	-0.0184	0.0082	-0.0683	0.0062	-0.0196	-0.0273	-0.1522	0.0319	1.0000
Procedure	-0.0156	-0.0592	-0.0301	-0.0027	-0.0419	0.0075	0.0095	-0.0767	0.0030	0.1865
APR DRG	0.0168	-0.0280	0.0072	-0.0912	-0.0843	-0.0245	-0.0299	-0.1211	0.0581	0.5188
APR MDC	-0.0204	-0.0270	-0.0181	0.0508	0.0324	0.0288	0.0417	0.0311	-0.0365	0.2529
Severity	-0.0086	-0.0065	-0.0023	0.0882	0.0069	0.0223	0.0235	0.0660	-0.0160	-0.0326
Mortality	0.0469	0.0427	0.0007	0.1297	-0.0804	-0.0949	-0.0756	-0.0223	0.2841	-0.0332
Surgical	0.0454	-0.0164	0.0095	-0.1692	0.0357	-0.0415	-0.0368	-0.2382	0.0211	0.1888
Birth Weight	-0.0273	-0.0234	-0.0224	0.3165	-0.0324	0.0645	0.0714	0.2371	-0.0935	-0.1242
Emergency	0.0183	0.0505	0.0295	-0.0889	-0.0650	-0.0145	-0.0483	-0.0666	0.1202	-0.0226
Mean Charges	0.0416	0.0634	0.0595	-0.0561	-0.0100	-0.0194	-0.0047	-0.0421	0.0329	0.0354

Table 9. Correlation matrix of factorized 19 features. Part 1.

	Procedure	APR DRG	APR MDC	Severity	Mortality	Surgical	Birth Weight	Emergency	Mean Charges
Area	-0.0156	0.0168	-0.0204	-0.0086	0.0469	0.0454	-0.0273	0.0183	0.0416
County	-0.0592	-0.0280	-0.0270	-0.0065	0.0427	-0.0164	-0.0234	0.0505	0.0634
Facility	-0.0301	0.0072	-0.0181	-0.0023	0.0007	0.0095	-0.0224	0.0295	0.0595
Age	-0.0027	-0.0912	0.0508	0.0882	0.1297	-0.1692	0.3165	-0.0889	-0.0561
Gender	-0.0419	-0.0843	0.0324	0.0069	-0.0804	0.0357	-0.0324	-0.0650	-0.0100
Race	0.0075	-0.0245	0.0288	0.0223	-0.0949	-0.0415	0.0645	-0.0145	-0.0194
Ethnicity	0.0095	-0.0299	0.0417	0.0235	-0.0756	-0.0368	0.0714	-0.0483	-0.0047
Admission	-0.0767	-0.1211	0.0311	0.0660	-0.0223	-0.2382	0.2371	-0.0666	-0.0421
Disposition	0.0030	0.0581	-0.0365	-0.0160	0.2841	0.0211	-0.0935	0.1202	0.0329
Diagnosis	0.1865	0.5188	0.2529	-0.0326	-0.0332	0.1888	-0.1242	-0.0226	0.0354
Procedure	1.0000	0.3269	0.0946	0.0474	-0.0252	0.4386	0.0259	-0.1851	0.0264
APR DRG	0.3269	1.0000	0.1808	-0.0279	0.0263	0.3607	-0.1066	-0.0093	0.0624
APR MDC	0.0946	0.1808	1.0000	0.0530	-0.2322	0.0242	0.1690	-0.2186	-0.0348
Severity	0.0474	-0.0279	0.0530	1.0000	0.1184	0.0554	0.1121	-0.1071	-0.0178
Mortality	-0.0252	0.0263	-0.2322	0.1184	1.0000	-0.0461	-0.1435	0.3199	0.0690
Surgical	0.4386	0.3607	0.0242	0.0554	-0.0461	1.0000	-0.1178	-0.2508	0.0699
Birth Weight	0.0259	-0.1066	0.1690	0.1121	-0.1435	-0.1178	1.0000	-0.2442	-0.0030
Emergency	-0.1851	-0.0093	-0.2186	-0.1071	0.3199	-0.2508	-0.2442	1.0000	0.0401
Mean Charges	0.0264	0.0624	-0.0348	-0.0178	0.0690	0.0699	-0.0030	0.0401	1.0000

Table 10. Correlation matrix of factorized 19 features. Part 2.