Inpatient Total Cost Prediction for NY State Facilities

Soufieh Hakimzadeh Hosseini
Department of Systems Science and Industrial Engineering
State University of New York at Binghamton, Binghamton NY 13902

Abstract: It is known that US healthcare is one of the most expensive industries in the world. The usual sum spent on healthcare per individual in virtually alike nations ($5,697) overall is nearly half of that of the U.S. ($10,966) . Inpatient healthcare costs are typically one of the most expensive healthcare costs in the US. On average, the cost per hospital stay in the US was around $11,700 in 2016, making it one of the most costly forms of healthcare utilization [2]. In order to make informed healthcare decisions, having an idea of expected costs to share with patients and insurers could be valuable to determine type of treatment, expected coverage, and expected end costs of treatment. Here will be shown an effective machine learning model that will predict the cost of inpatient treatment based on key parameters such as APR DRG Description, CSS Procedure Description and more. The ultimate goal of this research is to determine better cost and pay transparency for inpatient facilities for patients.

Keywords: Inpatient, Machine Learning, Health Systems, Cost Transparency

I. Introduction

Healthcare in the US is largely driven by financial incentives between payers and providers. The US healthcare system largely depends on a fee for service model in which the healthcare provider is paid a fee for a particular service rendered. In fact, 70 percent of physician practices report over 75 percent of their revenue comes from fee-for-service reimbursement and 19 percent of hospitals and health systems report the same [3] . This model of healthcare rewards medical practitioners for the volume of service provided regardless of the quality of care [4]. A plethora of monetary programs such as the Hospital Readmissions Reduction Program (HRRP) have been put in place in the healthcare system to guide the system away from the fee for service model into a value based care model, but these programs have often caused more harm than good. Over the course of five years, 20,000 individuals died unnecessarily as a result of the HRRP's use of money to deter doctors from readmitting patients who would have benefited from hospital treatment [5].

To effectively move away from the fee-for-service model into a value based care model, hospitals have been utilizing cost containment programs that highlight high cost patients upon admission and focus efforts on their care. Payers, such as insurance companies, can also use this as a tool to better predict how much a patient would claim for their service. Patients who are choosing to be admitted "electively" could also use this tool to plan their finances. In this project, an inpatient dataset will be analyzed to predict the cost of inpatient treatment to highlight these patients and drive value based care.

II. Background

There have been efforts in the past to create models to predict inpatient costs using clinical data or various model types such as stochastic gradient descent or random forest regressor. Studies such as '*Predicting Inpatient Costs with Admitting Clinical Data'* by William M. Tierney provided valuable insights but focused more on clinical lab data that takes time to process and has an associated cost itself [6]. To get earliest insights on cost prediction, it would be prudent to use data acquired upon initial patient assessment or check-in. The Tierney study was also limited to the internal medicine department of a single urban area and cannot be generalized to form fit for a country, state or even larger region beyond the hospital itself. The data and study itself is also outdated as the data used was from 1989-1999 and since then, key features that impact healthcare decision making have changed. The gap here is the requirement for an up-to-date, holistic dataset that encapsulates all inpatient discharges with features only available after initial patient review.

Studies such as *'Predicting the inpatient hospital cost using a machine learning approach'* by Kulkarni used a more holistic dataset but have not used methods such as lasso regression or decision trees to test higher accuracy in prediction results [7]. These methods will be explored in this project with an updated and holistic dataset.

Ⅲ. Methodology

This research method involves a 2015 Statewide Planning and Research Cooperative System (SPARCS) Inpatient De-identified dataset from Health.Data.NYC.gov. This dataset is composed of 35 columns and 2346931 rows from hospitals all over New York State. The variables being reviewed used in this analysis include: 'Hospital Service Area', 'Hospital County', 'Operating Certificate Number', 'Permanent Facility Id', 'Facility Name', 'Age Group', 'Zip Code - 3 digits', 'Gender', 'Race', 'Ethnicity', 'Length of Stay', 'Type of Admission', 'Patient Disposition', 'Discharge Year', 'CCS Diagnosis Code', 'CCS Diagnosis Description', 'CCS Procedure Code', 'CCS Procedure Description', 'APR DRG Code', 'APR DRG Description', 'APR MDC Code', 'APR MDC Description', 'APR Severity of Illness Code', 'APR Severity of Illness Description', 'APR Risk of Mortality', 'APR Medical Surgical Description', 'Payment Typology 1', 'Payment Typology 2', 'Payment Typology 3', 'Birth Weight', 'Abortion Edit Indicator', 'Emergency Department Indicator', 'Total Charges', 'Total Costs', and 'Ratio of Total Costs to Total Charges.

Because of the size of this dataset, pre-processing is a critical step to building our model. From initial review of the dataset, there are clear feature redundancies being captured such as 'CCS Procedure Code' and 'CCS Procedure Description', which are ultimately a numerical and descriptive way to encapture the same feature. After the pre-processing, we are able to train models on the data to identify which is the best predictive model.

A. Model Pre-Processing

Before any initial preprocessing was done, a heat map was generated from the original CSV file to determine the correlations between each set of features. This heat map is shown in figure 1 below.

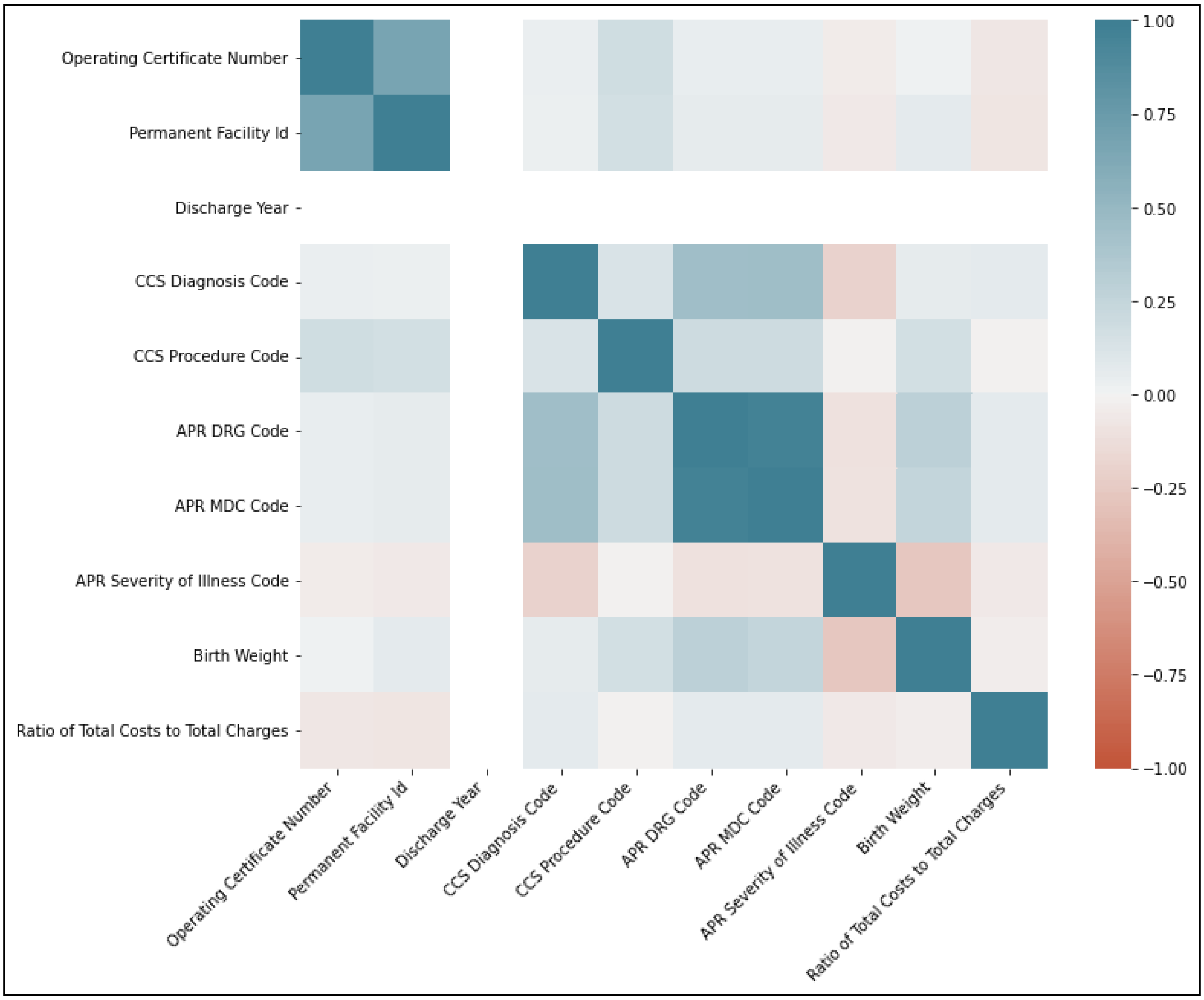


Figure 1: Heat Map of Feature Correlation

One thing to note on this heat map is that “Discharge Year” gives us no information. That is because throughout the entire dataset, discharge year remains constant at 2015. It also identifies a strong positive correlation between features such as Permanent Facility ID and Operating Certificate Number or APR DRG Code and APR MDC Code. This indicates that the information contained in the features may have strong redundancy and could be dropped when we move into feature reduction.

After scoping the initial feature correlation, it was time to identify features that may represent redundant information. The first step in doing this was to investigate if there was a 1:1 mapping between features marked with “codes" and features marked with “descriptions”. The features specifically investigated were, ‘CCS Diagnosis Code' and 'CCS Diagnosis Description', 'CCS Procedure Code' and 'CCS Procedure Description', 'APR DRG Code' and 'APR DRG Description', 'APR MDC Code' and 'APR MDC Description' and lastly, 'APR Severity of Illness Code' and'APR Severity of Illness Description'. For each of these, it was identified that there was a 1:1 mapping between the code and the descriptions provided, which indicated that we could drop one of the features to reduce redundancy. It was decided to retain the descriptive features and drop the numeric features because it simplified pre-processing and retained more information. If numeric codes were kept, preprocessing methods would have ranked one procedure or Diagnosis above another strictly by the value of its numeric code (i.g. The CSS Diagnosis code for Pneumonia is 122 and the CSS Diagnosis code for fluid and electrolyte disorders is 55. Using numeric values would rank pneumonia as a higher value than electrolyte disorders).

Location based redundancy was also dropped. 'Hospital County','Hospital Service Area','Zip Code - 3 digits', and 'Permanent Facility Id' were also dropped. All of these features indicated location based information. This same level of information could be provided by the single feature ‘Facility Name’ which was the only location based feature that was retained. ‘Discharge Year’ was also dropped because it provided us with no information since all the submissions in this feature were from the same year, 2015. ‘'Ratio of Total Costs to Total Charges'’ was also dropped, since this feature would have a direct correlation with the value to be predicted, ‘Total Cost’. ‘Payment Typology 3’ was dropped from the set as well because most of the values were ‘NaN’ and the provided data added no new information to the feature list. All the data shown in ‘Payment Typology 3’ was already listed in ‘Payment Typology 1’ and ‘Payment Typology 2’.

To further clean up the dataset, all the rows with any missing data were removed. This was a relatively small number of data points compared to the entirety of the set. Features such as ‘Total Cost’, ‘Length of Stay’ and ‘Total Charges’ were also converted from string types to float or integer values for further preprocessing. Categorical features that could be put on a numeric scale, such as "APR Severity of Illness Description", were replaced with numerical values such as ‘1,2,3,4’ in associated severity order. Once this was completed for each applicable feature, the dataset was split up into X (all the selected features, dropping ‘total cost’) and Y (‘total cost’).

Once the X and Y sets were created, the training and test datasets were generated. This X set still contained both numerical and categorical data. In order to standardize the dataset, the categorical and numerical features were broken up and both standardized using appropriate methods (categorical using OneHotEncoder and numerical using StandardScaler). The X training and testing sets were then fit to this standardization and are ready for model fitting.

B. Successful Models

A few various models were investigated for this use case. Because the purpose of this model is to forecast cost, a standard sklearn linear regression model was used as a baseline to compare to other models. Whether the results of this model were successful or not, it is always best practice to investigate other models to see which will outperform the others. The next model that was investigated was Lasso Regression. Even after initial feature reduction done in pre-processing, running computationally heavy models was made incredibly difficult by the large size of the dataset. Lasso Regression was investigated because it encourages a sparse model with fewer parameters; only selecting the parameters that have the most value add. Running this model would also support further variable selection or parameter elimination. The final model that was successfully investigated was the Decision Tree Regression model. One of the challenges identified during pre-processing was the distinction and separation of classification vs numeric features. The decision tree regression model was used because decision trees are used to solve both Regression and Classification, both of which features exist in our dataset.

C. Unsuccessful Models

Due to the size of the dataset and the nature of the machine the model is running on, there are certain disk and RAM restrictions that prevent further model investigation. Some models and methods that were attempted but could not successfully run include, PCA, Random Forest Regressor and Support Vector Regression.

IV. Results

A. Linear Regression Model

When running the Linear Regression model, we received an R^2 value of 0.8814647989662517 and an adjusted R^2 value of 0.8814490876429306. This means that this model was able to capture ~88% of all the variance represented in the dataset. This is also seen in figure 2 below, demonstrating a close and nearly linear relationship between the true and predicted values of Total Cost.

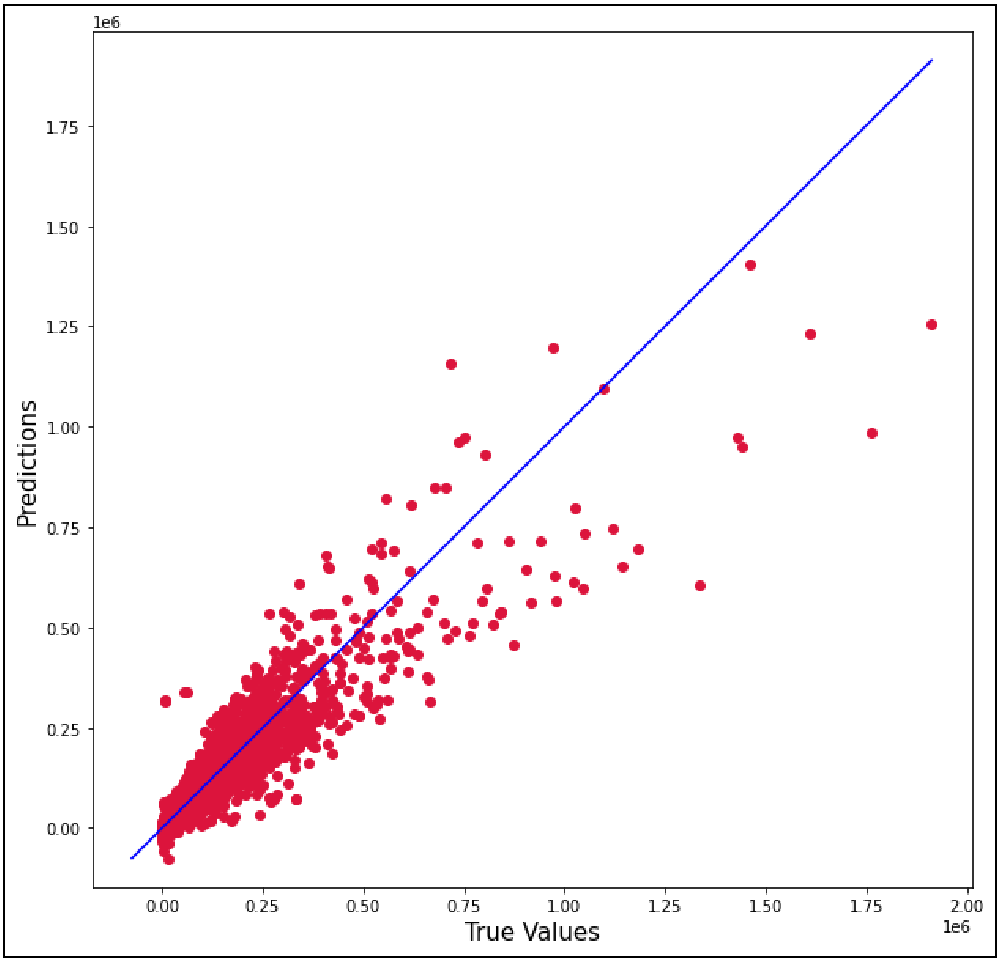


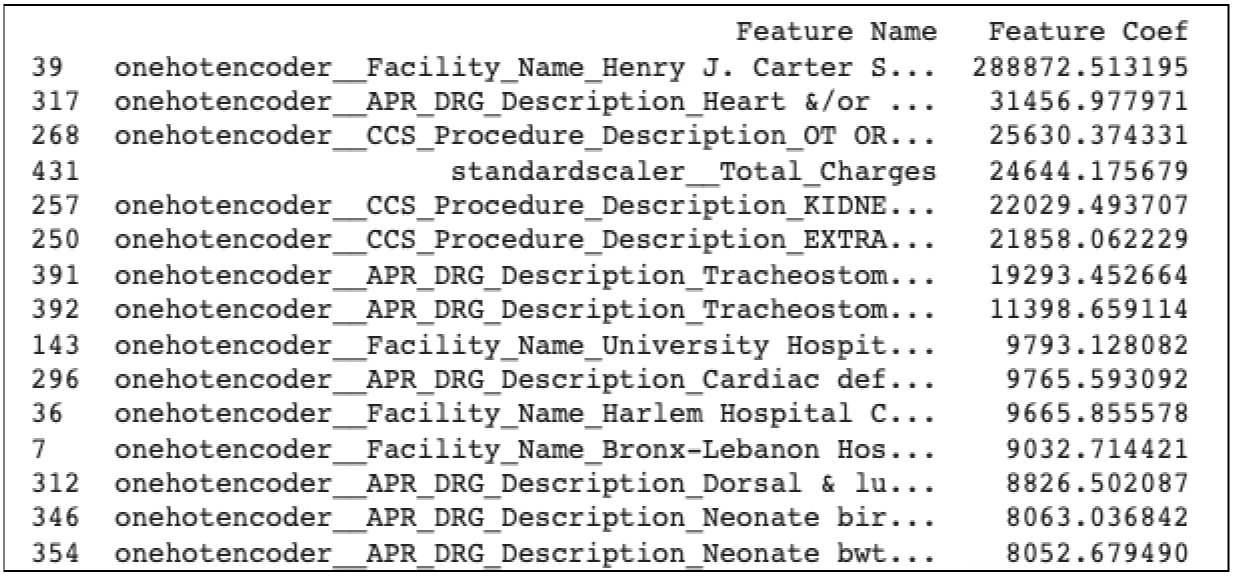
Figure 2: True vs Fitted Graph of Total Cost Using Linear Regression

While this initial baseline model did a good job of value capture, other methods were also investigated to see if it is possible to reach 90% variance capture in the dataset.

B. Lasso Regression Model

After running this model, the best forecasting score for running lasso was 0.862101. This was less than the baseline R^2 score for linear regression of 0.8814647989662517 so this model would not be the ideal model for use.

Using lasso did give us additional insight into the dataset and allowed us to see the features that have the highest impact on the prediction variable of “Total Cost”. We are able to identify these features and use them in future models to support feature reduction. As seen in figure 3 and 4 below, the top 15 features were highlighted and compared coefficient values in the bar graph to understand scale of impact.



```
                                           Feature Name    Feature Coef
39   onehotencoder__Facility_Name_Henry J. Carter S...  288872.513195
317  onehotencoder__APR_DRG_Description_Heart &/or ...   31456.977971
268  onehotencoder__CCS_Procedure_Description_OT OR...   25630.374331
431                     standardscaler__Total_Charges   24644.175679
257  onehotencoder__CCS_Procedure_Description_KIDNE...   22029.493707
250  onehotencoder__CCS_Procedure_Description_EXTRA...   21858.062229
391  onehotencoder__APR_DRG_Description_Tracheostom...   19293.452664
392  onehotencoder__APR_DRG_Description_Tracheostom...   11398.659114
143  onehotencoder__Facility_Name_University Hospit...    9793.128082
296  onehotencoder__APR_DRG_Description_Cardiac def...    9765.593092
36   onehotencoder__Facility_Name_Harlem Hospital C...    9665.855578
7    onehotencoder__Facility_Name_Bronx-Lebanon Hos...    9032.714421
312  onehotencoder__APR_DRG_Description_Dorsal & lu...    8826.502087
346  onehotencoder__APR_DRG_Description_Neonate bir...    8063.036842
354  onehotencoder__APR_DRG_Description_Neonate bwt...    8052.679490
```

Figure 3: Top 15 Features and Feature Instances with the Highest Feature Coefficients

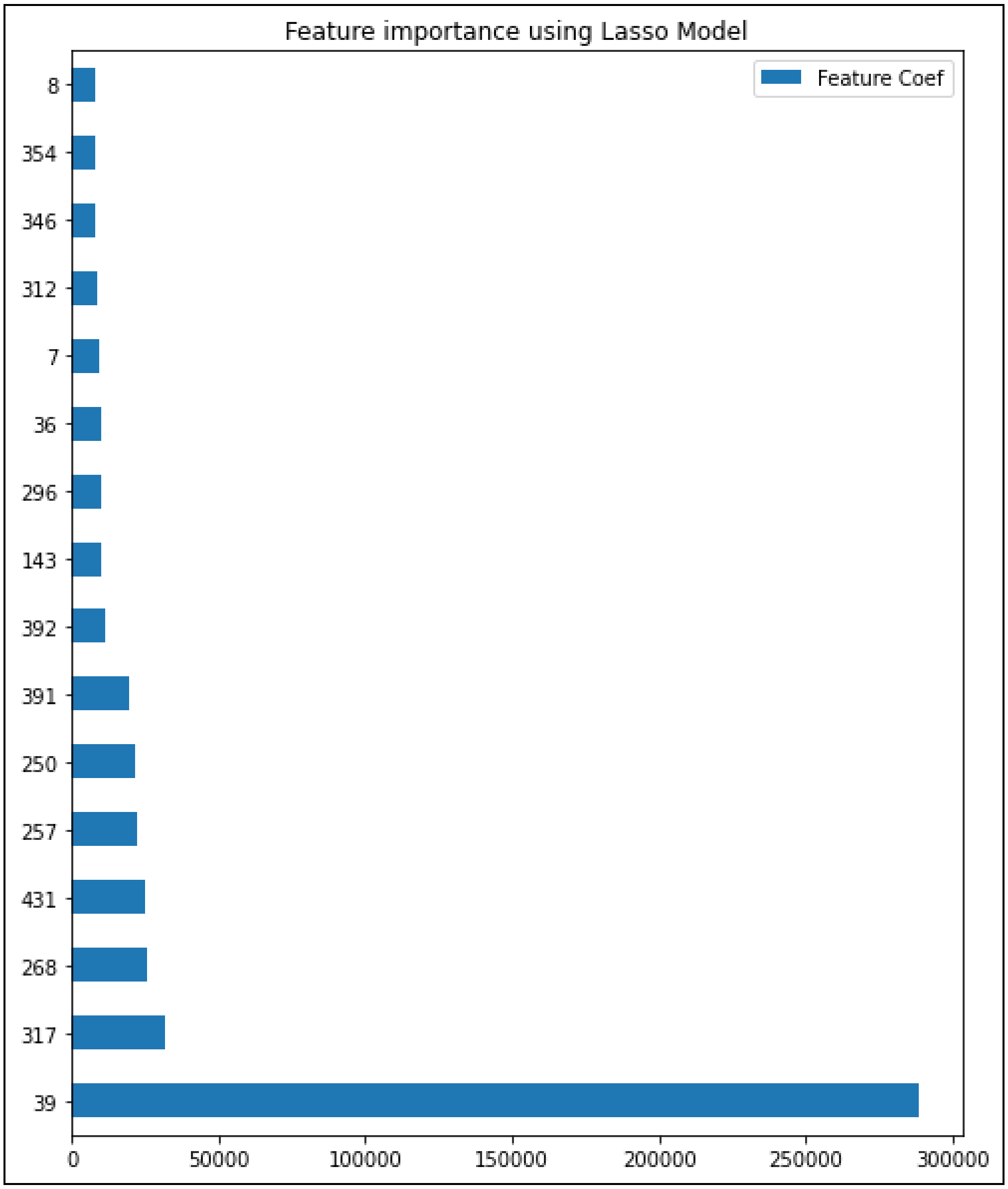


Figure 4: Bar Graph of Top 15 Features and Feature Instances

From this visualization, you can see that features such as Facility Name, APR DRG Description, and CSS Procedure Description have the highest impact on the model. The bar graph's Y axis aligns to the features listed in figure 3. For instance, the highest impact feature, 39, aligns to 'onehotencoder__Facility_Name_Henry J. Carter'.

C. Decision Tree Regression Model

Running the decision tree model resulted in an R^2 value of 0.9327174367246217 and adjusted R^2 value of 0.9327085187146443. This means that ~ 93% of all the variance of the dataset is captured in this model. This is captured in Figure 5 below, True vs Fitted Graph of Total Cost Using Decision Tree Regression Model.

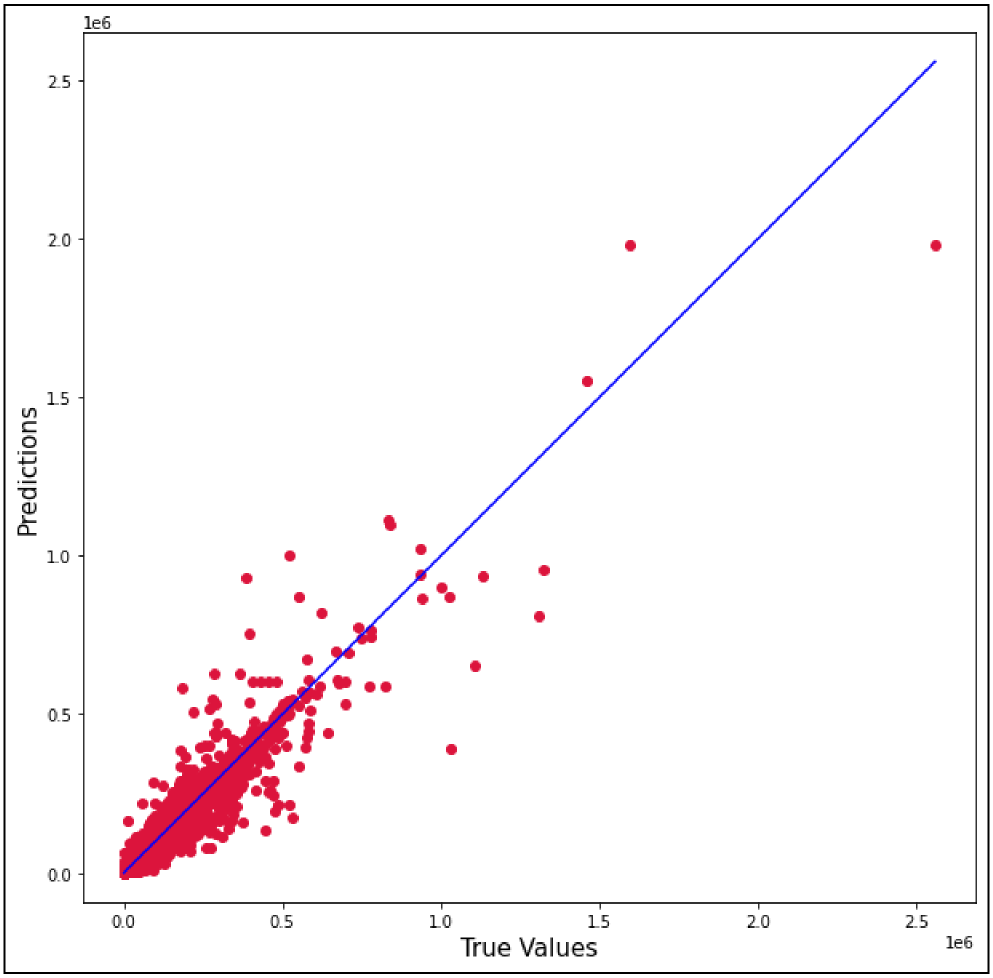


Figure 5: True vs Fitted Graph of Total Cost Using Decision Tree Regression

Comparing Figure 5 to Figure 2, both are excellent models with strong data concentration on the lower left corner and forming a regressed linear line. The distinction in Figure 5 is that there is a more distinct linear relationship between the true and predicted values, which indicates a more accurate model.

V. Conclusions and Discussion

Ultimately,through this project, a model was found that was able to predict the total cost of an inpatient stay with 0.9327174367246217 accuracy. The decision tree regression model was found to be the best out of the 3 successful models, with an R^2 value of 0.9327174367246217. This study had computational limitations which prevented investigation of additional models such as random forest regression or support vector regression. In further iterations of this project, additional models and further preprocessing using PCA could improve the accuracy of the model. This model also did not account for recurring patients and admissions which could be identified and used to improve model accuracy and understand the total cost of treatment with recurring visits. Temporal correlation for recurrent patients or distinct groups of people could be used to identify if there are trends between features such as CSS diagnosis description and readmission resulting in increased total cost. Further model improvement could include utilizing web scraping to increase the number of value added features. Certain features that could be more value added to the cost prediction include if the hospital is a learning hospital or not, the average experience level of the hospital staff, median income of the hospital service area and more. These features could be pulled from web scraping and a similar correlation investigation could be performed to identify if these features would be value added to the model or not.

Being able to predict a patient's total cost of care upon admission provides cost transparency and puts power back into the hands of the patient to make decisions regarding their physical and financial health. It can also support providers and facilities to move away from the fee for service model into the value based care model of healthcare. Models such as these explored in this investigation can close the cost transparency gap that is so detrimental in the US healthcare space today.

References

[1]. "How does health spending in the US compared to other countries?." 23 Dec. 2020, https://www.healthsystemtracker.org/chart-collection/health-spending-u-s-compare-countries/. Accessed 19 Nov. 2021.

[2]. "National Inpatient Hospital Costs: The Most Expensive Conditions ...." 9 Jul. 2020, https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp. Accessed 19 Nov. 2021.

[3]. "Healthcare Reimbursement Still Largely Fee-for-Service Driven." 26 Mar. 2020, https://revcycleintelligence.com/news/healthcare-reimbursement-still-largely-fee-for-service-driven. Accessed 12 Dec. 2021.

[4]. "What is fee-for-service? | healthinsurance.org." https://www.healthinsurance.org/glossary/fee-for-service/. Accessed 12 Dec. 2021.

[5]. "The Deadly Consequences Of Financial Incentives In Healthcare." 28 Jan. 2019, https://www.forbes.com/sites/robertpearl/2019/01/28/financial-incentives/. Accessed 12 Dec. 2021.

[6]. "Predicting inpatient costs with admitting clinical data - PubMed." https://pubmed.ncbi.nlm.nih.gov/7823640/. Accessed 12 Dec. 2021.

[7]. "Predicting the inpatient hospital cost using a machine learning ...." 30 Dec. 2020, https://www.emerald.com/insight/content/doi/10.1108/IJIS-09-2020-0175/full/html. Accessed 12 Dec. 2021.