Predicting Medicare Costs Using Non-Traditional Metrics

John Louie[1] and Alex Wells[2]

I. INTRODUCTION

In a 2009 piece [1] in The New Yorker, physician-scientist Atul Gawande documented the phenomenon of unwarranted variation – differences in cost that cannot be explained by socioeconomic factors or medical comorbidities alone – in health care costs and delivery. In a particularly stark example, Gawande compares the very similar neighboring cities of El Paso and McAllen, Texas; patients in McAllen have Medicare deductibles several times those of their neighbors in El Paso.

In this paper, we outline our approach to predicting Medicare costs on both hospital referral regions (HRR) and hospital levels. More specifically, we aim to build models to predict an individuals health care costs on the basis of non-traditional metrics by leveraging data from multiple open-source repositories. In doing so, we hope to both improve the accuracy of cost predictions and gain insights into factors responsible for fluctuations in health care costs.

II. METHODS

A. Choosing Our Data Sets

For our project, we focused on using two different sources for our Medicare data sets, the Dartmouth Atlas of Health Care (DAHC) and Medicare.gov. The DAHC provides information on chronically ill patient care, claims based Medicare spending, Medicare population demographics, post discharge events, and more, organized according to state, county, hospital referral region (HRR) and individual hospital levels. Medicare.gov has a dedicated website (data.medicare.gov) that provides a diverse range of data sets from which we chose the publicly available hospital comparison data sets, which track individual hospital attributes and events such as structural measures, complications, readmission rates, payments, value of care, outpatient imaging efficiency, and more. Within both the DAHC and the Medicare.gov data sets, each of the metrics were associated with unique IDs (e.g. a provider ID for a specific hospital or a HRR ID). The data from both DAHC and Medicare.gov contain detailed information on approximately 3,192 hospitals (out of the 5,686 hospitals in the United States [2]), which amounts to over 56% coverage. Table 1 details each of the files that we used for this project.

TABLE I
RAW DATA FILES

Source	Name	Year
DAHC	TA1_demographics.xls	N/A
DAHC	DAP_hrr_data_2011.xls	2011
DAHC	DAP_hospital_data_2010.xls	2010
DAHC	DAP_hospital_data_2011.xls	2011
DAHC	DAP_hospital_data_2012.xls	2012
DAHC	DAP_hospital_data_2013.xls	2013
Medicare.gov	Readmissions Complications and Deaths - Hospital.csv	2014
Medicare.gov	Structural Measures - Hospital.csv	2014
Medicare.gov	Timely and Effective Care - Hospital.csv	2014
Medicare.gov	Healthcare Associated Infections - Hospital.csv	2014
Medicare.gov	Outpatient Imaging Efficiency - Hospital.csv	2014
Medicare.gov	READMISSION REDUCTION.csv	2014
Medicare.gov	Readmissions and Deaths - Hospital.csv	2015
Medicare.gov	Complications - Hospital.csv	2015
Medicare.gov	Structural Measures - Hospital.csv	2015
Medicare.gov	Timely and Effective Care - Hospital.csv	2015
Medicare.gov	Healthcare Associated Infections - Hospital.csv	2015
Medicare.gov	Outpatient Imaging Efficiency - Hospital.csv	2015
Medicare.gov	READMISSION REDUCTION.csv	2015

We elected to use the DAP_hospital_data_*YEAR*.xls files from DAHC for two reasons. First, this particular set of files was one of the few that the DAHC had on the *hospital* level vs. on the HSA, HRR, county or state level. In contrast, the Medicare.gov data sets were almost exclusively at the hospital level. Second, these files represent information collected from chronically ill patients during their "end-of-life" period of care. For the Medicare patient population in particular, chronic illnesses account for 9 out of 10 patient deaths and these patients' last two years of life (i.e. "end-of-life") alone account for 32% of all of Medicare's spending. The chronically ill patient population with Medicare coverage significantly contributes to Medicare's overall costs and is an important population to study and analyze.

B. Preprocessing Data sets

In order to utilize any of the aforementioned data sets, we needed to perform an extensive amount of preprocessing. First, we organized the different data sets by their respective levels; for example, we grouped all the hospital level data sets for DAHC together because each file utilized the same unique provider ID. Using a single raw data set, which contains information for around 3,000 hospitals in a single year, would provide poor learning opportunities. Thus, when creating our training sets, we combined data from multiple years. We then selected features from our DAHC data sets (columns in the files correspond to our features) and from our Medicare.gov data sets (measure IDs correspond to our features) and created training sets corresponding to the DAHC HRR data (one with demographic data to serve as a baseline estimator and another with HRR data from 2011), DAHC hospital data over multiple years and Medicare.gov

[1] Stanford University, *Computer Science* (jwlouie@stanford.edu)
[2] Stanford University, *Biomedical Informatics* (awells2@stanford.edu)

hospital data over multiple years. Each training example within each training set is specified by a unique provider or HRR ID.

For our four training sets, we used the following labels ($y^{(i)}$'s): 1 & 2) DAHC baseline and HRR estimators: *"Price, age, sex & race-adjusted Total Medicare reimbursements per enrollee (Parts A and B) (2011)"*, 3) DAHC hospital level estimator: *"Total Medicare spending"* for Medicare spending per decedent by site of care during the last two years of life (deaths occurring over all four years), and 4) Medicare.gov hospital level estimator: *"Spending per Hospital Patient with Medicare (Medicare Spending per Beneficiary)"*.

Once we created our training sets (the code to do so is found on our Github page), we found that many of the training examples, especially those from Medicare.gov, were missing information. In order to address this issue, we used the following different strategies:

1) Missing Feature Thresholding: In our first preliminary approach, we omitted any training examples that failed to have a certain percentage of features (say 60% or 50%) filled. After this initial thresholding, for the remaining training examples that had missing features, we replaced them with the mean over that feature's column. This filling method was used for both our DAHC and Medicare.gov training sets.

2) Item-Item Collaborative Filtering: Our subsequent data filling approach was to use an item-item collaborative filtering recommender system, which we wrote with the guidance of CS 246's Recommender System lecture notes [3]. For this algorithm, we used the Pearson correlation coefficient and when filling in an entry, we considered at most 100 nearest neighbors. If after the recommender system step we still had missing data (i.e. for a given entry the nearest neighbors were neighbors with that entry missing), we again filled the entry with the mean over the feature column. This filling method was only used with our Medicare.gov training set (in future we could apply it to the DAHC hospital training set which had little data missing).

In addition to the above strategies for addressing missing data, we utilized variance thresholding in order to remove features that yielded little to no variance to our data set. After performing these preprocessing steps, our training sets were ready to be used with various learning models.

C. Learning Algorithms

All of the learning algorithms we used were implementations found in Python's scikit-learn package [4].

1) Supervised Learning Algorithms: For our datasets, we used multiple different supervised algorithms for both classification and regression.

a) Classification: For classification, we used both logistic regression and linear discriminant analysis (LDA). Logistic regression was used on our Medicare.gov training set to predict whether or not the expected Medicare cost for an individual hospital was above or below the national average. We also used multi-class LDA to predict the expected Medicare cost quantile for an individual hospital.

b) Regression: We decided to use the following three algorithms to predict Medicare costs: (1) Linear regression, (2) Kernelized Support Vector Machine, and (3) Gradient Boosting. Each of these methods was used to predict expected Medicare cost on both the HRR and hospital levels.

2) Unsupervised Learning Algorithms: In addition to using supervised learning techniques like those mentioned above, we decided to also use the following unsupervised algorithms: (1) k-Means clustering, (2) Principal Component Analysis (PCA) and (3) various Manifold Learning techniques:

a) k-Means Clustering: We ran k-means on our initial DAHC and Medicare.gov training sets in an attempt to try and determine whether there were any overall patterns or trends in our data.

b) Principal Component Analysis (PCA) and Manifold Learning: Our primary reason behind leveraging both PCA and Manifold Learning was to help us visualize our high-dimensional data. Both PCA and Manifold Learning allowed us to project our data onto two dimensional plots; however, PCA makes the assumption that there is inherent linearity in the data while Manifold Learning attempts to reveal non-linear structures in the data. The following Manifold Learning algorithms were used in our visualization: (1) Locally Linear Embedding (LLE), (2) Local Tangent Space Alignment (LTSA), (3) Hessian Locally Linear Embedding, (4) Modified Locally Linear Embedding, (5) Isomap, (4) Multi-dimensional Scaling (MDS), (5) Spectral Embedding and (6) t-distributed Stochastic Neighbor Embedding (t-SNE).

D. Validation Methods

For each of our supervised learning techniques we used k-fold cross validation, where $k = 7$. In addition, we generated learning curves, which show the cross validation and training scores for an estimator, to help us visualize our model's performance with varying sized training sets. Learning curves allow us to determine how much we benefit from adding more training data and whether the estimator suffers more from a variance error or a bias error. For our k-Means Clustering, we evaluated the "quality" our clustering by utilized the silhouette coefficient, which is a method for evaluating clusters when the true cluster designations are unknown.

III. Results

A. Medicare.gov Hospital Compare Datasets

The training set constructed using Medicare.gov's Hospital Compare data sets generally performed poorly with all of the supervised learning algorithms that were attempted. Table 2 contains the performance of the Medicare.gov training set with linear regression, SVM and gradient boosting with k-folds cross validation ($k = 7$).

Because the labels for this particular data set represented how much a hospital's spending deviated from the national average (the average being represented with a 1, with under spending being less than 1 and over spending being greater than 1), we leveraged logistic regression to determine whether we could accurately classify an example as over or under spending. With k-folds cross validation ($k = 7$), our mean classification accuracy was 0.59847865.

Unfortunately, because of the poor performance of our regression and classification models on the Medicare.gov training set, we elected to omit further results.

TABLE II
MEDICARE.GOV ESTIMATOR RESULTS

Model	Residual Sum of Squares	R^2 Score	Explained Variance Score
Linear Regression	0.00430412	0.11957590	0.13182519
SVM	0.00451941	0.06599637	0.08464518
Gradient Boosting	0.00377542	0.22417090	0.23331060

B. Dartmouth Atlas of Health Care: HRR Datasets

a) Baseline Models: Projected Medicare cost is commonly based on attributes such as age, gender, and ethnicity. Using demographic features, we created three supervised baseline estimators of Medicare cost on the HRR level. To determine how well the baseline models performed on the test set, we calculated the residual sum of squares, R^2 score, and explained variance score for each model. We also plotted the training and test set deviance as a function of boosting iteration, as well as the most important features used for regression. The results of this analysis are shown in Table 3 and Figure 1.

TABLE III
HRR BASELINE ESTIMATOR RESULTS

Model	Residual Sum of Squares	R^2 Score	Explained Variance Score
Linear Regression	50,447,836	-26.28506791	-25.9796321
SVM	2,026,808	-0.09621391	-0.08512193
Gradient Boosting	1,540,374	.16687749	.19183267

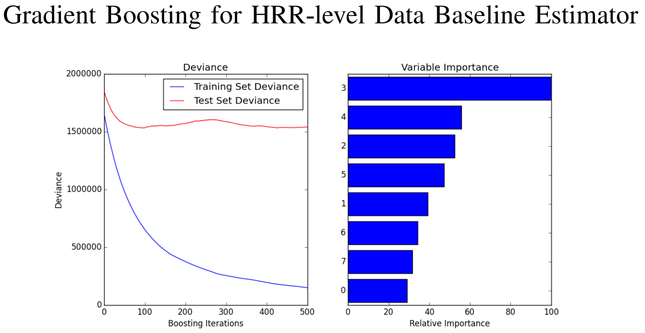

Fig. 1. Training/Test Set Deviance and Most Important Features for HRR Baseline Estimator Regression

b) Models Using Non-Traditional Features: In order to predict the expected Medicare cost on the HRR level, we utilized the same three supervised models and evaluated them using the same three metrics as mentioned above for our baseline model. The results are shown in Table 4 and Figure 2 below. (Note that in our plot of important features, we limited the number of features displayed to only include the top 20).

TABLE IV
HRR REGRESSION RESULTS

Model	Residual Sum of Squares	R^2 Score	Explained Variance Score
Linear Regression	759,045	.58946504	.59364512
SVM	706,283	.61800151	.6248149
Gradient Boosting	595,374	.67798742	.68243715

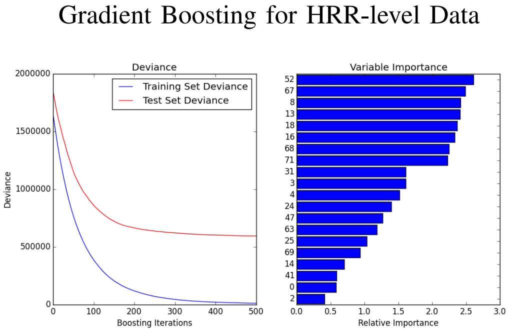

Fig. 2. Training/Test Set Deviance and Most Important Features for HRR Regression

C. Dartmouth Atlas of Health Care: Hospital Level Datasets

Using hospital level datasets from the Dartmouth Atlas of Health Care, we created some additional models to both predict the quantile rank and the raw expected Medicare cost of an individual hospital.

a) Quantile Rank: To predict how expensive an individual hospital's cost is relative to other hospitals in the United States, we used multi-class LDA with $k = 3$ fold cross validation. Each hospital in the training set was grouped by a quantile (either quartile or decile) based its average Medicare cost. We then predicted which quantile group a hospital from the test set belongs and kept track of the overall accuracy and error of the model. The results are shown in Table 5.

TABLE V
LINEAR DISCRIMINANT ANALYSIS RESULTS

Quantile Grouping	Accuracy	Error
Quartiles	.665	1.10
Deciles	.440	1.72

The LDA model was able to predict the quartile or decile rank of an individual hospital with reasonable accuracy and error. The model correctly predicted the quartile and decile rank of a hospital with 66.5% and 44.0% accuracy respectively. We also saw that when the model incorrectly predicted the quartile or decile, it was usually only off by 1 or 2 quantile groupings.

b) Expected Medicare Cost Models: Next, we attempted to predict the expected Medicare cost of an individual hospital based on non-traditional features from the DAHC. Again, we leveraged the same three models as before with the same performance metrics with $k = 7$ fold cross validation. The results are shown in the Table 6 and Figure 3 (the boosting plot is again only for the top 20 features).

TABLE VI
HOSPITAL REGRESSION RESULTS

Model	Residual Sum of Squares	R^2 Score	Explained Variance Score
Linear Regression	77,321,554	.7843928	.78697434
SVM	36,016,307	.8932150	.89335357
Gradient Boosting	45,902,257	.8655878	.86390434

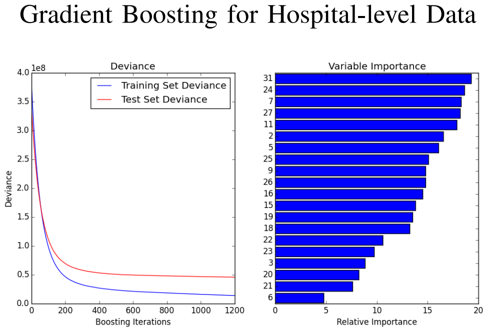

Fig. 3. Training/Test Set Deviance and Most Important Features for Hospital Level Regression

c) k-Means, PCA and Manifold Learning: For our DAHC hospital level data set, we performed *k*-means with cluster sizes of 3 to 8. As mentioned before, the silhouette coefficient, which is on a scale of -1 to 1, was used to evaluate the quality of our clusters. The silhouette coefficients for $k = 3$ to $k = 8$ are shown in Table 7.

TABLE VII
DAHC HOSPITAL LEVEL *k*-MEANS CLUSTERS

Number of Clusters	Silhouette Coefficient
k = 3	0.493468
k = 4	0.464665
k = 5	0.449925
k = 6	0.424073
k = 7	0.398933
k = 8	0.382540

In order to visualize our clusters, we decided to combine our *k*-means clustering labels with both PCA's and multiple Manifold Learning algorithms' visualizations. By projecting our data onto two dimensions ($\mathbb{R}^{39} \rightarrow \mathbb{R}^2$), we were able to roughly see how the data was clustered. The result of PCA and *k*-means ($k = 5$) clustering is shown in Figure 4, while the Manifold Learning techniques and *k*-means ($k = 5$) are shown in Figure 5.

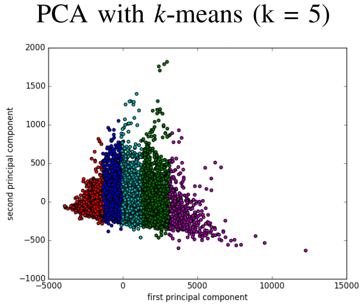

Fig. 4. Hospital level data projected onto two dimensions using PCA

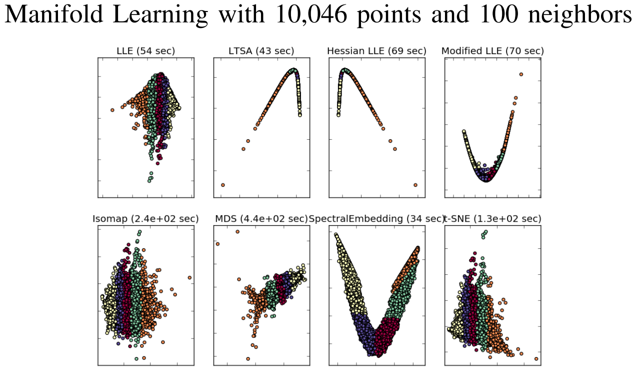

Fig. 5. Hospital level data projected onto two dimensions using various Manifold Learning techniques

d) Hospital Level Learning Curves: Because our training set for our DAHC hospital level data was significantly larger than any of our other training sets (≈ 10,000 vs. ≈ 5,000), we also elected to produce learning curves for our three regression models in Figures 6, 7 and 8 for linear regression, SVM and gradient boosting, respectively.

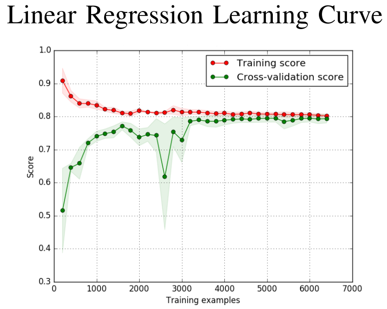

Fig. 6. Linear regression learning curve produced from Hospital level data with training sets of varying size

IV. ANALYSIS

a) HRR-level Data: Our first baseline HRR estimator using traditional metrics such as demographics and ethnicity performed poorly and proved very ineffective at predicting Medicare costs when compared to the estimator based on non-traditional features. From our analysis, some of the most important features for our non-traditional estimator were number of *ambulatory cases*, *readmission rate*, *physicians per 100,000 residents*, and *critical care physicians per*

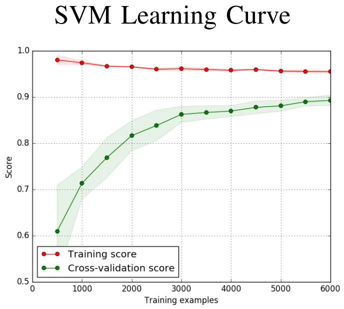

Fig. 7. SVM learning curve produced from Hospital level data with training sets of varying size

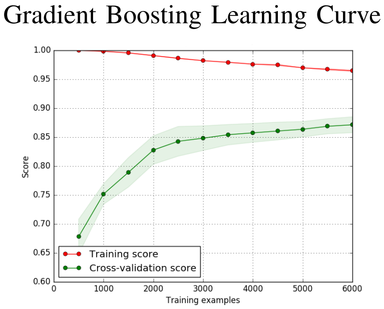

Fig. 8. Gradient boosting learning curve produced from Hospital level data with training sets of varying sizes

100,000 residents. These results suggests that quality of initial care and health care system complexity (e.g. number of physicians per resident) may play a larger role in determining a region's Medicare costs than individual demographics. As a result, enforcing more thorough initial health screens to reduce readmission rates (and possibly ambulatory cases) may lead to a decrease in Medicare costs.

b) Hospital Level Data: We found that models trained on our data set consisting of non-traditional metrics at the hospital level predicted expected Medicare cost very well. On the individual hospital level, some of the most predictive features include: *percent of deaths occurring in hospital*, *medical and surgical unit days per patient*, *medical specialist visits per patient*, and *number of beds*. A higher percent of deaths occurring in hospitals may be indicative of the hospital's reception of more extreme cases of chronic illness, (e.g. treatment of eczema vs. renal dialysis), which may in turn, require more medical and surgical unit days per patient and medical specialist visits per patient. Additionally, hospitals receive more extreme medical cases usually through a referral process because they are larger (either in staff, which usually translates to a higher patient capacity) and more equipped to handle medical complications that may arise from severe medical conditions. These patients eventually end up with high medical and surgical costs before the end of their lives. As with the HRR level finds, we see that a system's complexity is correlated with Medicare costs.

c) Learning Curves: The learning curve graph for our linear regression model shows that the training and validation scores are plateauing to a value of approximately 0.8. This particular graph pattern shows that our model is currently suffering from higher bias than desirable, meaning we are underfitting the data. In order to address this issue, we could introduce greater complexity to our estimator; however, this visual phenomenon may be indicating that the data may not have a strictly linear relationship. It is also worth noting that increasing our training set size for linear regression would provide little improvement on the training and validation scores. With this in mind, understanding the visuals provided by the Manifold Learning plots may be informative in understanding the underlying non-linear structure to our data that may explain the observed. plateauing behavior.

The learning curves for our gradient boosting and SVM models show that we are achieving validation scores of around 0.9. Because the training scores are above our validation scores for both of these models, we can see that their generalization and performance can be further improved with additional training examples.

V. CONCLUSIONS & FUTURE DIRECTIONS

Our results indicate that Medicare costs can be estimated with reasonable accuracy using non-traditional metrics associated with individual hospitals or HRRs. Additionally, models trained on non-traditional features were significantly better at predicting Medicare costs than models trained on demographic information alone. As a result, our analysis suggests that Medicare costs are more strongly influenced by location of care (HRR region or hospital) than they are by individual demographics. Therefore, non-traditional metrics, such as those used in our project, should be included in order to assure accurate and realistic Medicare cost predictions for patients.

Looking forward, we hope to next construct a predictive model for individual patients. The ability to predict the expected cost of admission to a specific hospital based on an individuals symptoms could be beneficial and more reflective of cost than predictions at the hospital level. We also hope to improve our current learning models by utilizing methods like grid search to find optimal parameters for our models and incorporating more data into our training sets.

CODE

All the code and datasets used are located at: https://github.com/jlouie95618/cs229-project.

ACKNOWLEDGMENT

We would like to thank the entire CS 229 teaching staff for this opportunity and especially Professor John Duchi for his investment in providing an incredibly challenging and rewarding offering this quarter.

REFERENCES

[1] Gawande, Atul. ”The Cost Conundrum.” The New Yorker. 25 May 2009.
[2] ”Fast Facts on US Hospitals.” American Hospital Association. Jan 2016. Web.
[3] Leskovec, Jurij, Anand Rajaraman, and Jeffrey D. Ullman. ”Recommendation Systems.” Mining Massive Datasets. Stanford University.
[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, 2011.