Predicting Medicare Costs Using Non-Traditional Metrics

John Louie, Alex Wells

Introduction & Motivation

Our aim was to build models to predict an individual's healthcare costs on the basis of non-traditional patient metrics by leveraging open source hospital data. In doing so, we hope to both improve the accuracy of patient cost predictions and gain insights into factors responsible for fluctuations in healthcare costs.

Methods

Datasets

- Dartmouth Atlas of Health Care (DAHC): Hospital and hospital referral region (HRR) level
  - Combined years 2010 to 2013
- Medicare.gov "Hospital Compare": Hospital level
  - Combined years 2014 and 2015

Preprocessing

- Data Completion Methods/Filling Missing Data
  - Feature Mean
  - Item-Item Collaborative Filtering
    - Pearson correlation coefficient
  - Thresholding based on missing data
    - Removed training examples that had *over* 50% of features missing

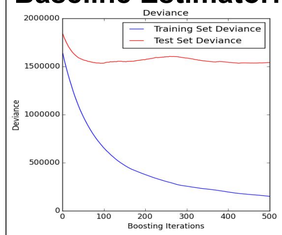
Algorithms

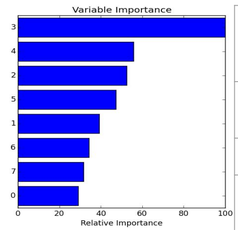
- Supervised Learning
  - DAHC Data (both HRR and hospital levels)
    - Linear Regression
    - Kernelized Support Vector Machines
    - Gradient Boosting
  - Medicare.gov Data
    - Logistic Regression
- Unsupervised Learning (*visuals to the right*)
  - k-means Clustering
  - PCA
  - Manifold Learning
- Feature Selection/Validation Methods
  - Variance Thresholding
  - K-folds Cross Validation
  - Learning Curves (*visuals to the right*)

Note: Results from our Medicare.gov data set are omitted because the training set constructed failed to achieve reasonable learning, relative to the DAHC training set.

Results

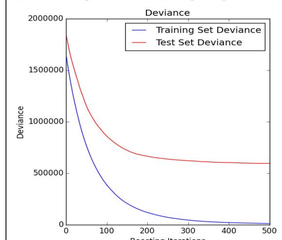
Baseline Estimator:

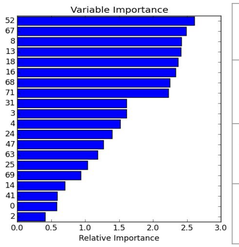




Model	Residual Sum of Squares	R^2 Score	Explained Variance
Linear Regression	50,447,836	-26.28506791	-25.9796321
SVM	2,026,808	-0.09621391	-0.08512193
Gradient Boosting	1,540,374	.16687749	.19183267

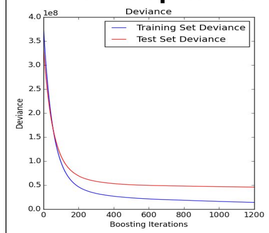
DAHC HRR-level:

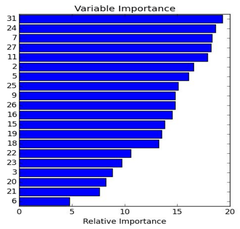




Model	Residual Sum of Squares	R^2 Score	Explained Variance
Linear Regression	759,045	.58946504	.59364512
SVM	706,283	.61800151	.6248149
Gradient Boosting	595,374	.67798742	.68243715

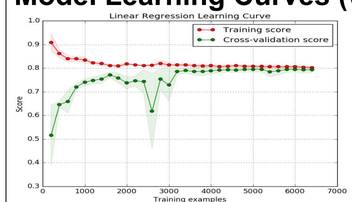
DAHC Hospital-level:

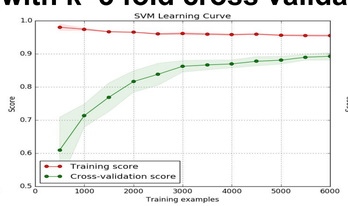


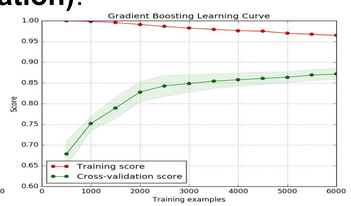


Model	Residual Sum of Squares	R^2 Score	Explained Variance
Linear Regression	72,389,847	.78537125	.78679061
SVM	36,016,307	.8932150	.89335357
Gradient Boosting	45,902,257	.86558786	.86390434

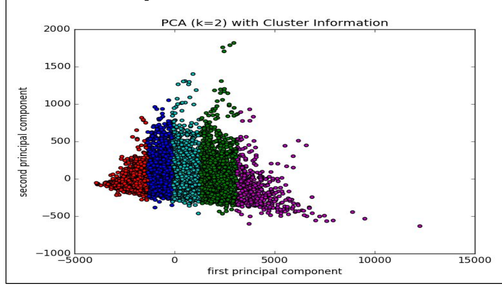
Model Learning Curves (with k=3 fold cross validation):

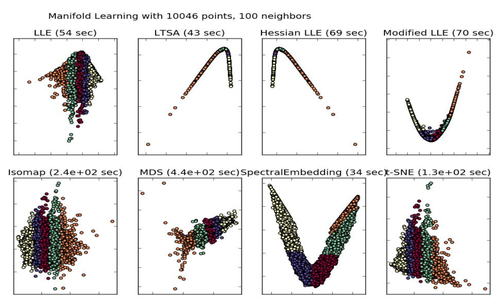






Visual Representations of Our Data:





Discussion

DAHC Hospital Referral Region Data:

We first sought to predict the cost of an average hospital in a given Hospital Referral Region (HRR). We implemented a baseline estimator based on traditional metrics/features such as demographics and ethnicity. This estimator performed poorly and proved very ineffective at predicting Medicare costs when compared to the estimator based on non-traditional features. Some of the most important features for this estimator were *number of ambulatory cases*, *readmission rate*, *physicians per 100,000 residents*, and *critical care physicians per 100,000 residents*. This suggests that quality of initial care and a health care system's complexity may play a larger role in determining Medicare costs than individual demographics.

DAHC Hospital 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*.

Learning Curves:

Our plotted learning curves for our three supervised models show that our gradient boosting and SVM models' generalization and performance can be further improved with more training examples. More training examples could potentially be retrieved and incorporated from previous years' data sets. Our linear model's performance is plateauing, indicating that we may need to incorporate greater complexity into our model.

Conclusion & Future Directions

Our results indicate that Medicare costs can be estimated with reasonable accuracy using non-traditional metrics associated with individual hospitals or hospital referral regions.

Looking forward, we hope to next construct a predictive model on the individual patient level. The ability to predict the expected cost of admission to a specific hospital based on an individual's symptoms could be extremely beneficial. In addition, we hope to improve our current learning models by utilizing methods like grid search to find the optimal parameters for our models (e.g. SVM) and we hope to find more data that we could potentially include in our training sets.