CMS Public Use File

A Machine Learning Case Study

Analysis & Results

Table of Contents

[Introduction 1](#_Toc49958310)

[Dataset 2](#_Toc49958311)

[Feature Selection 3](#_Toc49958312)

[Human Feature Selection 3](#_Toc49958313)

[Computer Feature Selection 3](#_Toc49958314)

[Model Performance 3](#_Toc49958315)

[Conclusion 3](#_Toc49958316)

# Introduction

The CMS data provides county and statewide Medicare and Medicaid information. There are a few target variables (*Actual Per Capita Cost*, *Average HCC Score, etc.*) to choose from and a variety of ‘per capita' explanatory variables (*Ambulance Events Per 1000 Beneficiaries*, *Imaging Events Per 1000 Beneficiaries*, etc.) **Essentially, we want to see if a human can outperform a machine in predicting the *Actual Per Capita Cost* target variable.**

Here is the competition:

* The human is only allowed to use graphics to visually pick the 5 most important explanatory variables
* The computer is allowed to use penalized regression and decision trees to pick the 5 most important explanatory variables
* Two multiple linear regression models will be created
  + One with the 5 human chosen variables
  + One with the 5 computer chosen variables
* The species with the higher R^2 wins

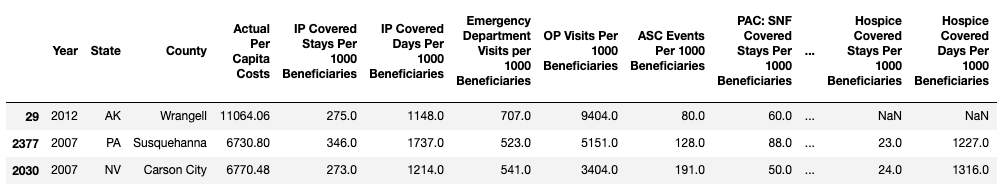
Git Hub Repo – <https://github.com/Data-Science-Link/CMS_Public_Use_File>

Jupyter Notebook – <https://nbviewer.jupyter.org/github/Data-Science-Link/CMS_Public_Use_File/blob/master/CMS_Public_Use_File_Analysis.ipynb>

# Dataset

The most salient variables in the transformed dataset are:

* Year
* State
* County
* Actual Per Capita Costs – The cost of healthcare Medicare/Medicaid services per person in a given state or county
* “….” Per 1000 Beneficiaries – Per capita estimates of healthcare service usage



*Figure 1: Sample of dataset*

# Feature Selection

Feature selection is a process by which you choose which explanatory variables are most important for predicting your target variable. This is important when you have an overwhelming number of explanatory variables. If you include too many of them you can face the issue of multi-collinearity and the curse of dimensionality. In this experiment, we compare the effectiveness of using the human eye vs. the computer’s hardware to perform feature selection.

## Human Feature Selection

To reiterate, in this competition the human is only allowed to use graphics to choose the top 5 explanatory variables. There are two types of graphics generated in this section. The first (Figures 2 & 3) graph per capita cost over time for each state. The state lines are colored according to the value of a given explanatory variable. In Figure 2 you can see that per capita cost is correlated with E&M Events. Conversely, in Figure 3 you can see that per capita cost is not correlated with Hospice Covered Stays. A graph was created for each of the explanatory variables. By visually inspecting all resulting graphs, five were chosen to be the best explanatory variables. These variables are listed below.

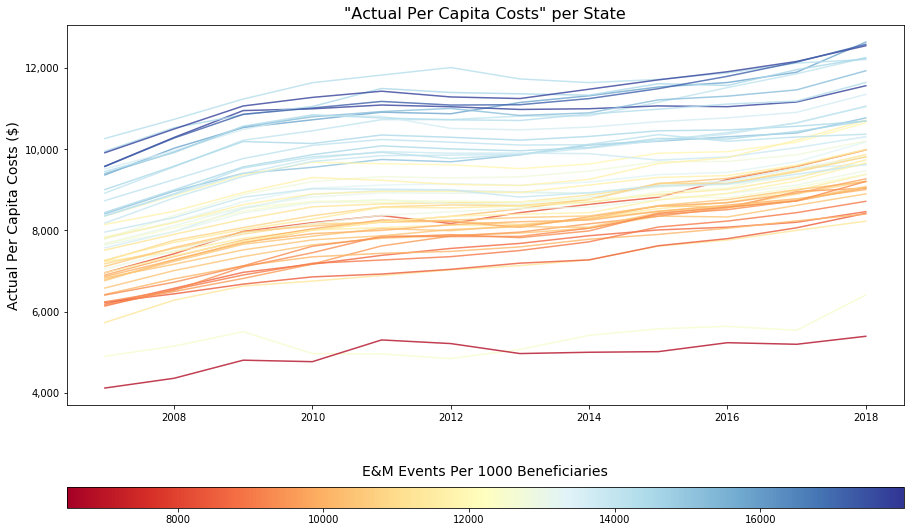
**Top 5 Variables According to Time Series Plot:**

* IP Covered Stays Per 1000 Beneficiaries
* PAC: SNF Covered Stays Per 1000 Beneficiaries
* E&M Events Per 1000 Beneficiaries
* *Imaging Events Per 1000 Beneficiaries*
* Emergency Department Visits per 1000 Beneficiaries

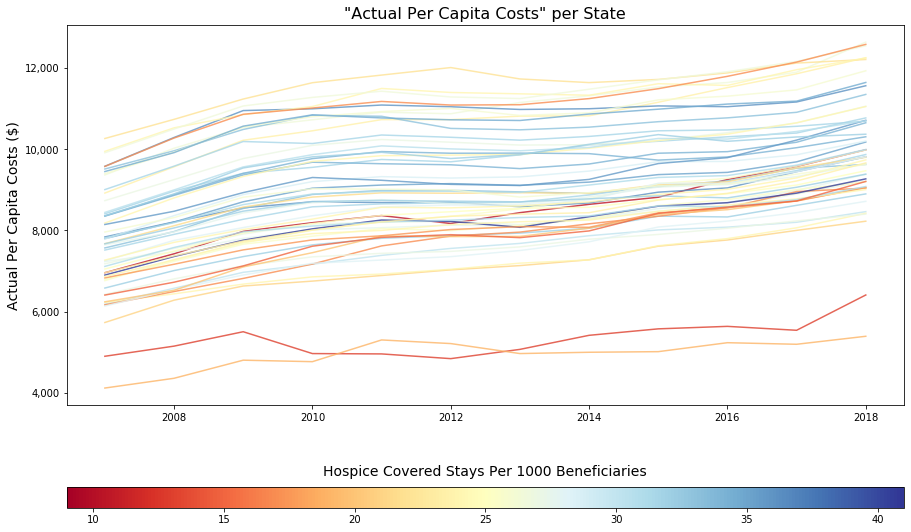
The second type of graph is a correlation bar plot of all variables against per capita cost (See Figure 4). The five variables with the highest correlation to per capita cost were chosen and are listed below.

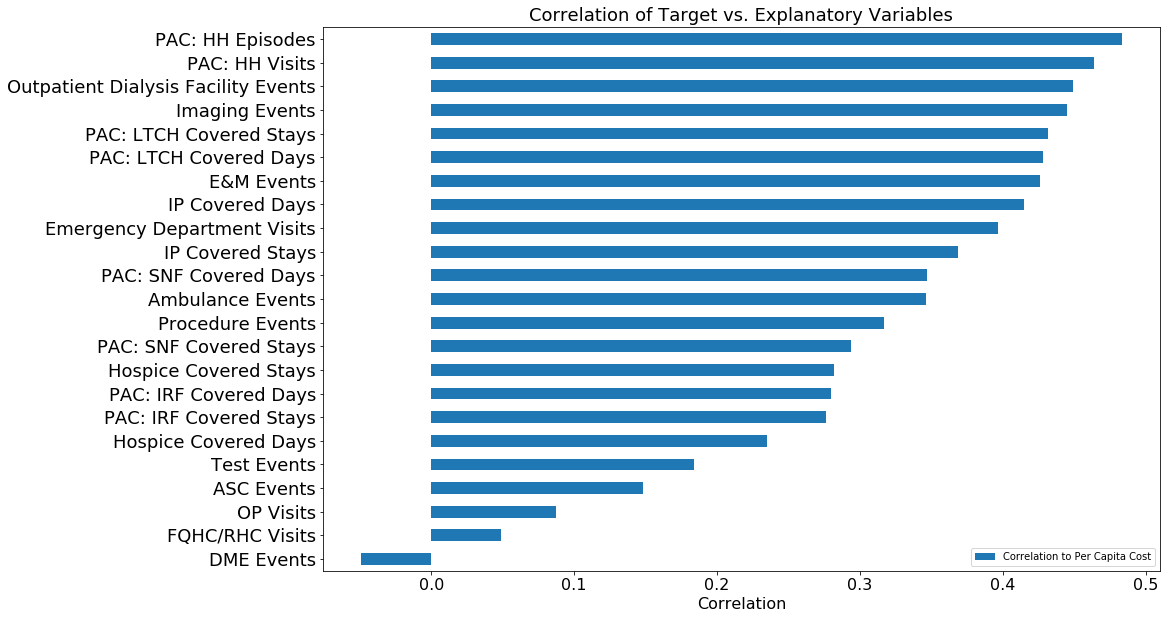
**Top 5 Variables According to Correlation Plot:**

* PAC: HH Episodes Per 1000 Beneficiaries
* PAC: HH Visits Per 1000 Beneficiaries
* Outpatient Dialysis Facility Events Per 1000 Beneficiaries
* *Imaging Events Per 1000 Beneficiaries*
* PAC: LTCH Covered Stays Per 1000 Beneficiaries



*Figure 2: National growth of per capita cost; High correlation to E&M Events*

*Figure 3: National growth of per capita cost; Low correlation to Hospice Stays*



*Figure 4: Per Capita Cost correlation to all other variables*

## Computer Feature Selection

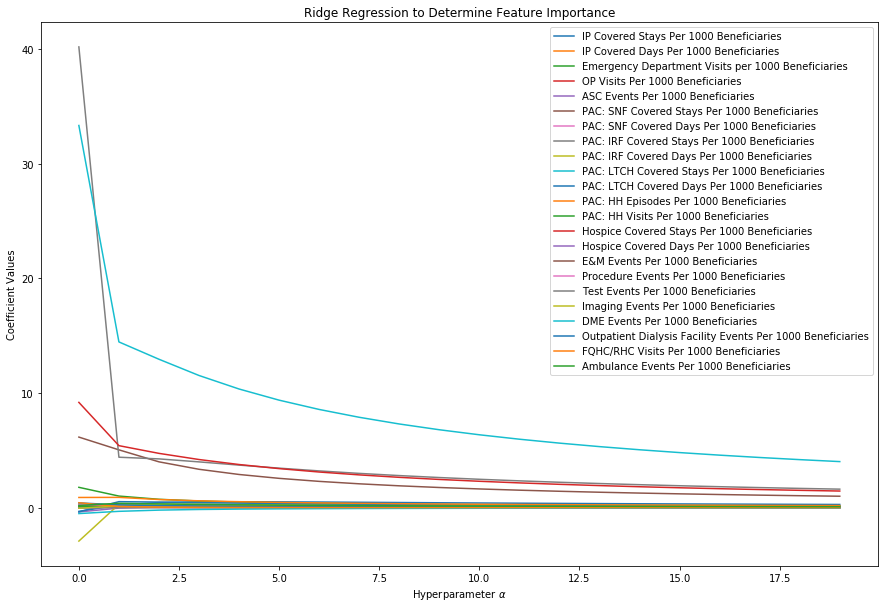
To reiterate, in this competition the computer is allowed to use any method to choose the top 5 explanatory variables. Two methods were explored. The first is penalized linear regression. A ridge regression model was created and the penalization hyperparameter increased to observe which variables had the most predictive power (see Figures 5 & 6). The second method was to use an ensembled decision tree (Random Forest). This random forest model outputted feature importance (see Figure 7). The top five explanatory variables for each of these methods are detailed below.

**Top 5 Variables According to Ridge Regression:**

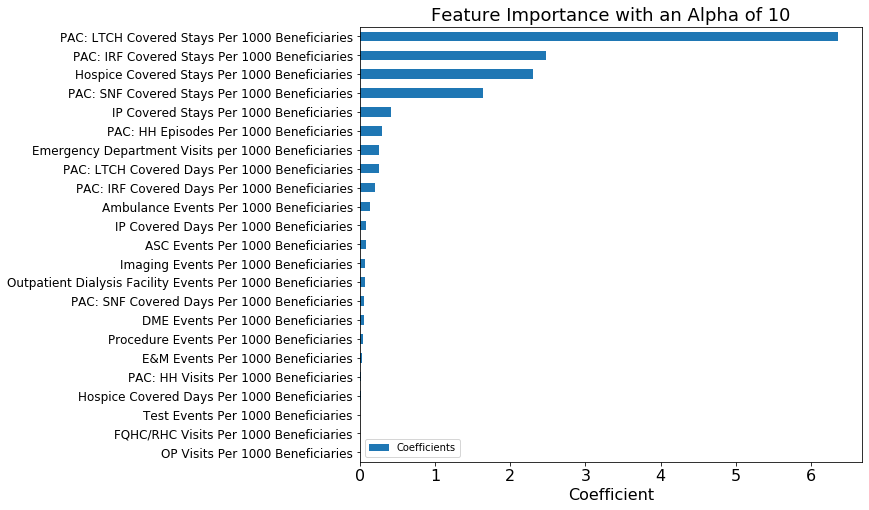
* PAC: LTCH Covered Stays Per 1000 Beneficiaries
* PAC: IRF Covered Stays Per 1000 Beneficiaries
* Hospice Covered Stays Per 1000 Beneficiaries
* PAC: SNF Covered Stays Per 1000 Beneficiaries
* IP Covered Stays Per 1000 Beneficiaries

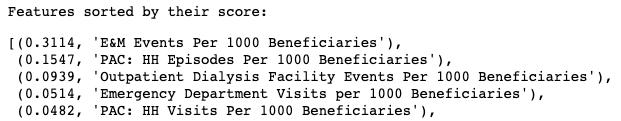
**Top 5 Variables According to Random Forest:**

* E&M Events Per 1000 Beneficiaries
* PAC: HH Episodes Per 1000 Beneficiaries
* Outpatient Dialysis Facility Events Per 1000 Beneficiaries
* PAC: HH Visits Per 1000 Beneficiaries
* Emergency Department Visits per 1000 Beneficiaries



*Figure 5: Ridge regression using all per capita explanatory variables*

*Figure 6: Ridge regression feature importance using an alpha of 10*

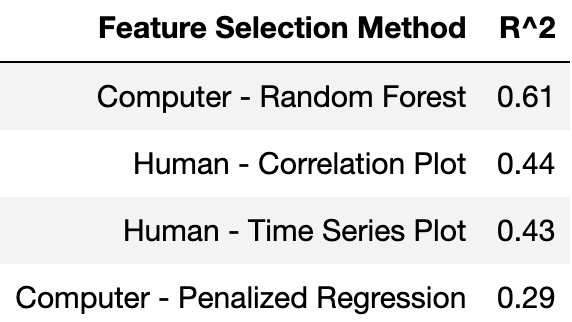


*Figure 7: Random forest feature importance*

# Model Performance

In the end, the random forest feature selection method worked the best. A multiple linear regression model using the random forest specified features outperformed all other models (see Table 1). Human feature selection facilitated by graphics outperformed the penalized regression model. Despite this human “success”, it should be noted that the resulting R squared of 0.44 is quite poor.

*Table 1: Model performance using various feature selection techniques*

**

# Conclusion

According to the rules of our competition **the Computer is the winner.**

In reality though, the strategy of just including more data is the real winner. A multiple linear regression model will all per capita variables (all the data), outcompeted the others with an R squared of 0.76. In this case, multicollinearity and model stability is not a concern. Including all variables generates a model with the highest predictive power. The conclusion is that, feature selection should be left to the computers and that oftentimes more data means more predictive power.