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. There are 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

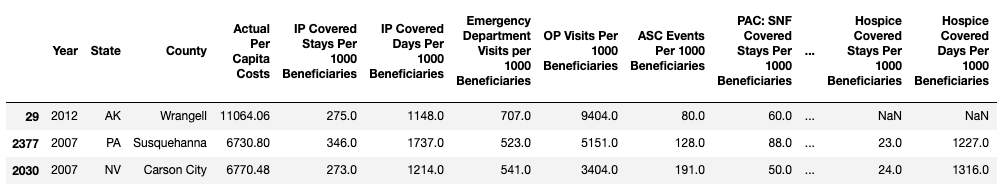
Git Hub Repo –

Jupyter Notebook –

# 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

Choose which explanatory variables have the most predictive power

## Human Feature Selection

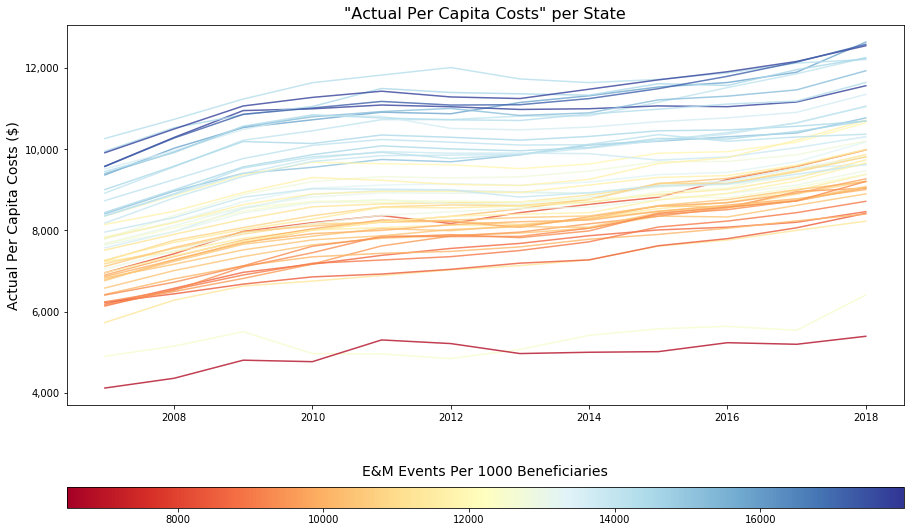
Again, the human is only allowed to use graphics to choose the top 5 explanatory variables.

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

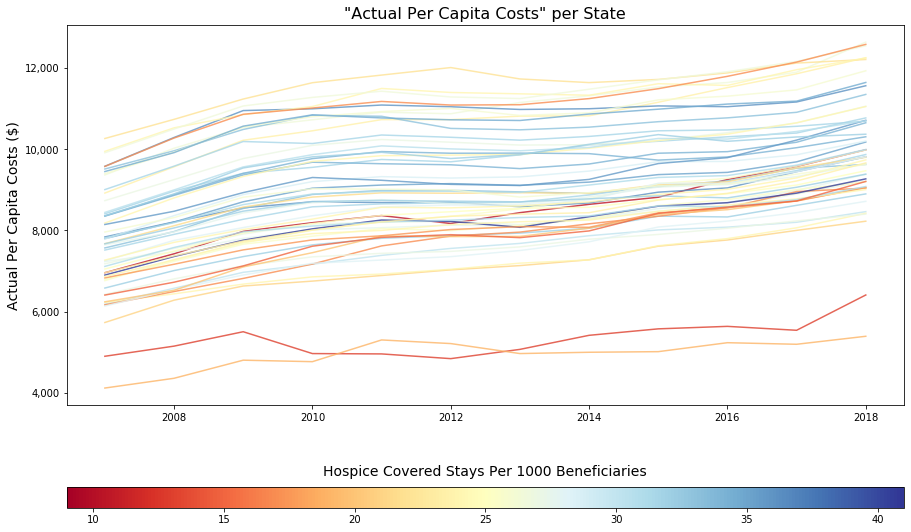
* IP Covered Stays Per 1000 Beneficiaries
* PAC: SNF Covered Stays Per 1000 Beneficiaries
* E&M Events Per 1000 Beneficiaries
* *Imaging Events Per 1000 Beneficiaries*
* Emergency Department Visits per 1000 Beneficiaries

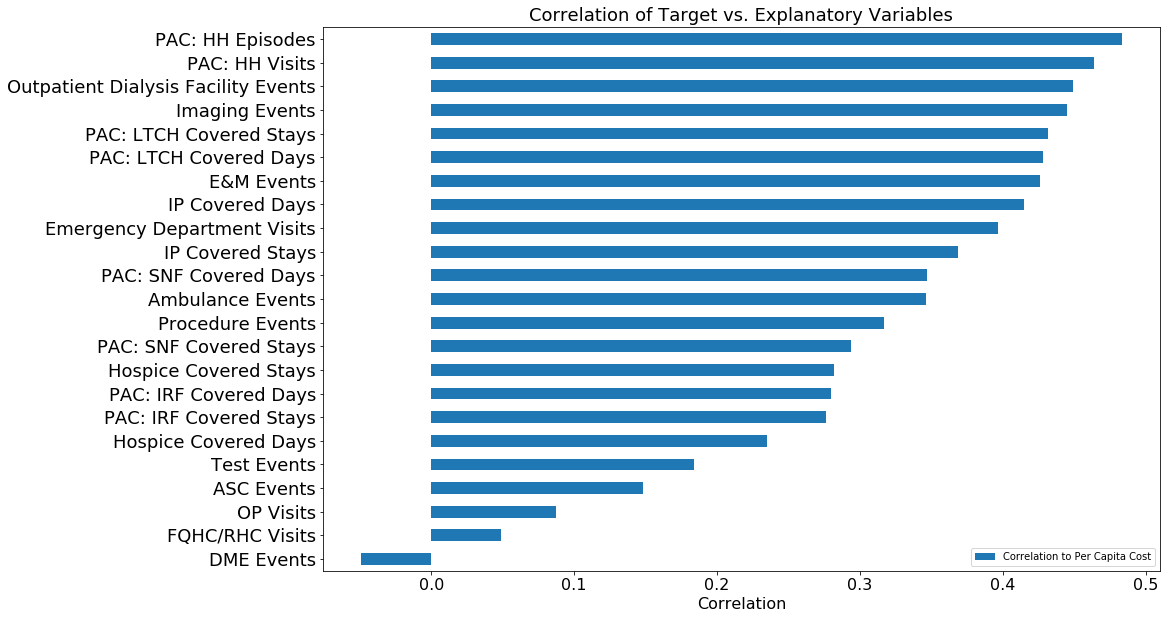
**Top 5 Variables According to Correlation Plot:**

* 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 :*

*Figure :*



*Figure :*

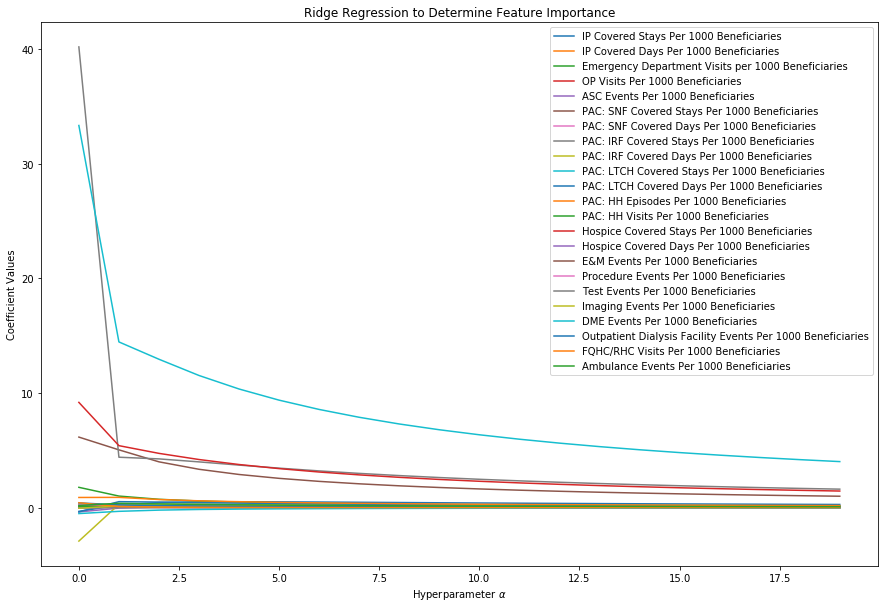
## Computer Feature Selection

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

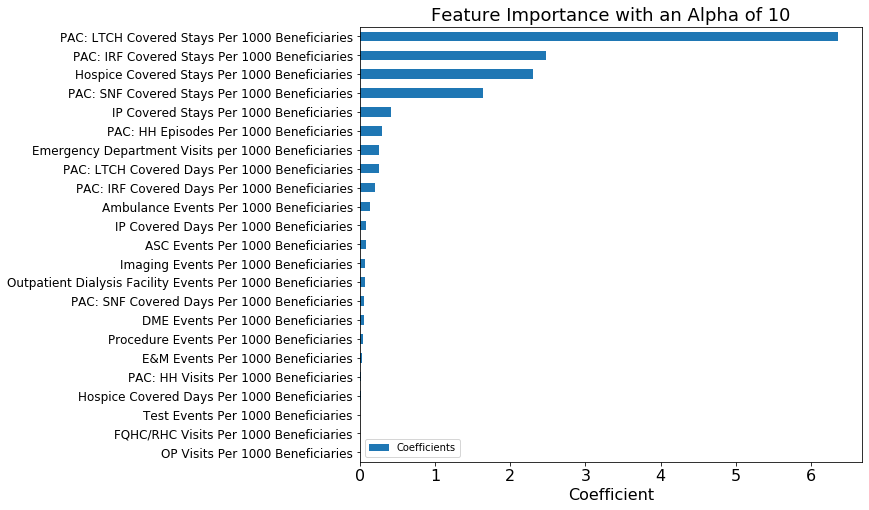
* PAC: LTCH Covered Stays Per 1000 Beneficiaries
* PAC: IRF Covered Stays Per 1000 Beneficiaries
* Hospice Covered Stays Per 1000 Beneficiaries
* PAC: SNF Covered Stays Per 1000 Beneficiaries
* IP Covered Stays Per 1000 Beneficiaries

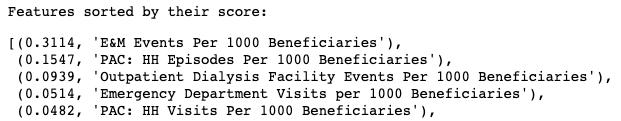
**Top 5 Variables According to Random Forest:**

* 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 :*

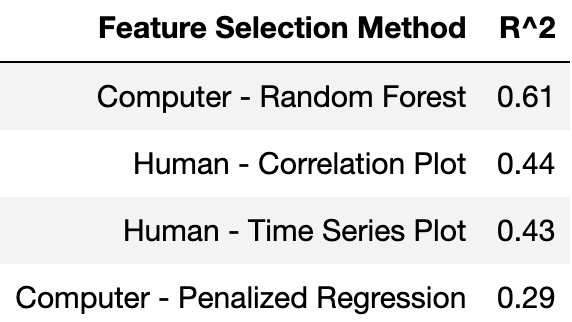
*Figure :*



*Figure :*

# Model Performance

*Table :*

**

# 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. When creating a multiple linear regression model will all per capita variables, it outcompeted the others (R^2 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 in many cases, models with more data give more predictive power.