The Price Is Right? Estimating Medicare Costs with Machine Learning

Sarah Rosston and Samantha Steele

I. INTRODUCTION

Historically, healthcare costs in the United States have been difficult to quantify as there have been no incentives for hospitals to publicize their pricing. However, in 2011, Medicare released the Inpatient Prospective Payments System that for the first time publicized price lists for hospitals across the United States. This dataset reveals vast discrepancies in pricing for similar procedures across different hospitals, even within the same broad geographic region. Consider one particularly prevalent condition, heart failure, which affects more than three million Americans annually. Patients in Baltimore, Maryland are charged on average $\$35,000$ for treatment of heart failure and shock. At a suburban hospital 30 miles away, patients are charged only $\$7,000$, a five factor discrepancy. Figure 1 highlights the broader geographic discrepancies for this disease across the United States.

Although insurance agencies generally negotiate these prices to more standardized payments, the problem is particularly relevant for patients without insurance. They lack the resources to negotiate and may pay exorbitant prices. Consequently, our goal is to identify the predictive factors behind these discrepancies and create a model for more accurately predicting the price of a procedure for the patient and the Medicare system. We explore four potential models for cost prediction: linear regression, regression trees, support vector regression and neural networks. We compare the performance of each model, finding that neural networks most accurately model healthcare prices.

II. RELATED WORK

Previous healthcare studies have predominantly used linear regression to model healthcare costs. Smith et al. [10] attempts to solve a similar problem to ours: predicting payments-to-charge ratios (PCR) for Medicare by hospital. The paper uses a regression model with features including casemix variables, hospital characteristics, and state characteristics. However, their models do not take into account the fine-grained, county level demographic data used in our approach. Using their variable set, Smith et al. modeled PCR with RMSE of 0.17 for Medicare. Penberthy et al. [9] use similar variables to Smith et al. to predict the costs of cancer in individual patients. Penberthy et al. use patient-specific variables including cancer stage, but also factor in county level data including percent of people with below secondary education, percent of population 65 or older, percent of population that is not white, and income data at the zip code level. They use a two-stage least squares model, using many of the population statistics to estimate the number of specialty physicians in the area in the first stage, and predicting costs in the second stage. Medicare costs of cancer are predicted with R^2 values ranging from 0.38 for prostate cancer to 0.49 for breast cancer. Although we also use demographic data, we use it at a different aggregate level, focusing on average hospital costs and producing a better linear regression predictor (see results below).

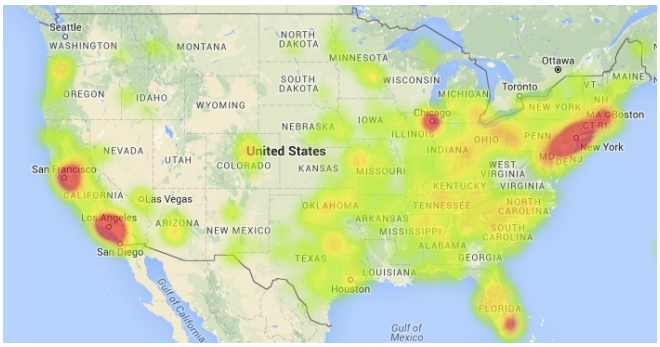

Fig. 1. Heatmap of Prices of Treatment for Heart Failure and Shock Across the United States. Red Areas Represent the Highest Prices.

To our knowledge, no previous papers have applied non-linear methods to the examination of healthcare prices on a per hospital basis. However, on a per-patient basis several papers have used non-linear approaches. Bertsimas et al. [4] show that more complex models are better for cost prediction than linear regression, using clustering and classification trees. They achieve optimal R^2 values of 0.60, although most of their models produced much worse results. Sushmita et al. [11] use Regression Trees, M5 Model Trees, and Random Forests to predict cost data from medical and claims history. As a baseline, they use simple linear regression with previous cost as the predictor, which had a lower RMSE than their more complicated models in most scenarios.

The most successful application of non-linear approaches that we uncovered is Lee et al [7], which improves on the work of Penberthy et al by using Regression Trees and Neural Networks to predict payments for colorectal cancer patients at a single hospital in Korea. Their R^2 values are substantially higher than Penberthys (0.713 for Regression Trees and 0.813 for Neural Networks). However, their work focuses on a tightly controlled patient set at a single hospital, incorporating only patient records. While its conclusions cannot be directly generalized to a highly diverse, large-scale healthcare system like that of the United States, Lee et al is suggestive, as our work confirms, that appropriate non-linear techniques perform better.

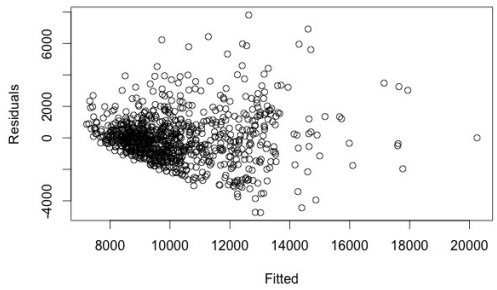

Fig. 2. Residuals vs. Fitted Values Plot

III. DATASET AND PREPROCESSING

While many costs associated with treating disease are not available in public datasets[6], data about medicare costs is freely available from the Center for Medicaid and Medicare. A rich dataset on the Inpatient Prospective Payment System from 2011 chronicles costs of the most common 100 medical procedures from Medicare patients. Overall, the dataset includes information on about 60% of Medicare discharges in that year. To narrow our focus within the dataset, we concentrate on the costs of treating "Heart Failure and Shock with Comorbid Conditions," one of the four most common health issues in the dataset. The data covering Heart Failure and Shock contains information about costs at 2,476 hospitals across all 50 states, accounting for 174,474 discharges and $\$1,826,146,256$ in payments. In addition to payment data, the Center for Medicaid and Medicare has a wealth of data on hospitals. Drawing on this, we supplement our initial dataset with information about each hospital, including type, ownership and patient hospital ratings. The patient hospital ratings datasets reported "Not Available" for a fraction of hospitals. To include this potentially important feature, we limit our dataset to 1,987 hospitals that reported complete survey results.

We also include economic and demographic data drawn from the US Census. Each of these datasets were pre-processed to remove incomplete data and joined to the hospital data on a county-by-county basis. They include data on income, race, education, poverty level, and age distribution. These combined datasets provided 60 potential explanatory variables across the 1,987 hospitals.

A. Heteroscedasticity

From early results, we hypothesized that our data was heteroscedastic. Heteroscedasticity is the property of having non-uniform variance among the elements of the data set. Because linear regression relies on the assumption of homoscedasticity, datasets that violate this assumption tend to perform poorly. To evaluate this conclusion, we plot residuals vs fitted values (Fig 2). While a homoscedastic plot would show evenly distributed values, this cone-shaped data confirms strong heteroscedasticity.

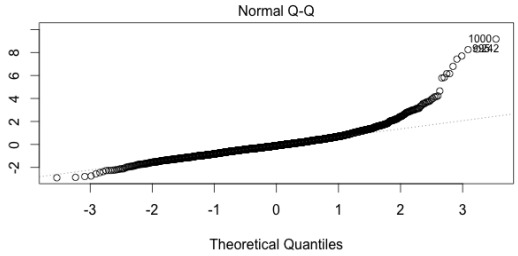

Fig. 3. Normal Q-Q Plot for Linear Fit

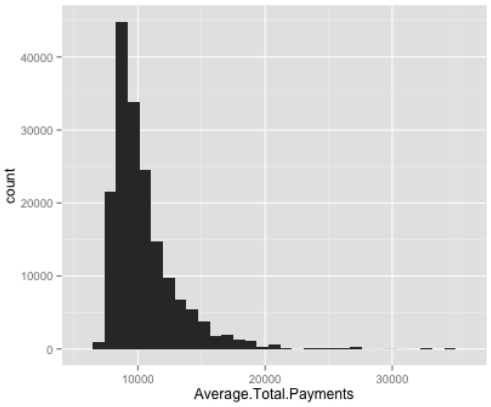

Fig. 4. Histogram of Heart Failure and Shock Costs

We corrected for heteroscedasticity in two ways. We standardized the features of our data set to have zero mean and unit variance. This counteracts variation between features of difference scales. In models that permit weighted training inputs, heteroscedasticity can be corrected by using weights that are inversely proportional to the variance of each point:

$$w_i = \frac{1}{\sigma_i^2}$$

However, The true values of the σ_i are unknown. To determine a proxy, we examined potential root causes of heteroscedasticity. Our data give equal weight to hospitals where data has been collected from only one patient or from over 600. Data backed by fewer patients is likely to have higher variance. Since smaller values of the variable 'Total.Patients' likely correspond to higher variance, we approximate weights with

$$w_i = \#\text{ of patients per datapoint}$$

These values are used on the inputs to both weighted least squares and neural networks.

B. Outliers

Even from government sources, data about medical costs is often inaccurate or incomplete. [6] Because of these innacuracies, as well as the large number of hospitals in our dataset, we expected to see outliers in costs. A Normal Q-Q Plot, shown in Figure 3. confirms these suspicions, and shows that data in the highest cost quantiles is not normally distributed. The histogram in Figure 4 shows, among mostly evenly distributed data, a few cases where average costs

are above $30,000 for a condition that averages $12,000. These points correspond to the ones identified as outliers in the Normal Q-Q Plot, and indicate either a few very expensives cases or inaccurate data. Because these few data points are so far from the rest of the dataset, they exert high leverage on the model, decreasing its accuracy. To account for outliers, we excluded the top and bottom decile of costs from our analysis as the data points may either be innacurate or include only a small number of patients.

IV. METRICS

To assess the validity of our models, we chose two main metrics. The R^2 coefficient measures correlation between variables to assess goodness-of-fit.

$$\text{Adjusted } R^2 = \frac{\sum_i (y_i - \hat{y_i})^2/(n-p-1)}{\sum_i (y_i - \bar{y})^2/(n-1)}$$

It takes on a value between 0 and 1, with 1 indicating a perfect fit. The R^2 value can only increase as additional features are added. Consequently, we use adjusted R^2, to correct for this bias.

To evaluate our generalization error on the cross-validation set, we use the root mean-square error:

$$\text{RMSE} = \sqrt[2]{\frac{\sum_i ((y_i - \hat{y_i})^2)}{n}}$$

This measures the distance between estimated and actual values and must be compared with the mean value to contextualize scale.

V. MODELS

Our baseline naïve model implemented simple linear regression in R with all 60 features used to predict prices. This model achieved an R^2 value of 0.5482 and a RMSE of 1845.97, in a feature dataset where the average price is $\$12,000$.

A. Model Selection

The initial regression model used all of the variables in our dataset, but many of the coeffecients have high p-values, an indicator that they are not significant and that the model may be overfitting. Feature selection is a clear solution to this overfitting problem and we explored using both forward and backward selection. Because our data set contains a large number of features, some variables are likely correlated. We therefore hypothesize that forward selection would be the better choice because it is biased towards choosing fewer variables. Since one major research goal is to identify the key features that affect hospital price, we investigate using both forward and backward selection. This permits us to examine differences in which features were selected.

To implement feature selection we use the Akaike Information Criterion (AIC). AIC provides a method for comparing models to chose whether to add or subtract a feature, balancing goodness-of-fit with model complexity. The AIC measure for a model is calculated according to the following function

$$\text{AIC} = 2k - 2\ell(\theta)$$

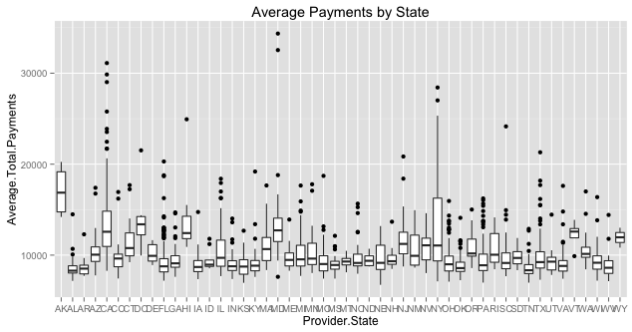

Fig. 5. Boxplot of Average Payments by State

Thus, the optimal θ satisfies the following equation.

$$\hat{\theta} = \arg\min_{\theta} 2k - 2\ell(\theta)$$

In forward selection, features are added greedily until a new feature would no longer decrease AIC. Using forward selection, 23 features were selected from the original set of 60, and using backward selection 28 features were chosen. After weighting average costs by the number of discharges, both forward and backward selection chose significantly more features: 26 and 34 respectively. While the feature set was substantially reduced with model selection, the adjusted R^2 value only changed slightly, increasing to 0.5491 with forward selection and 0.5512 with backward. One reason for the small changes is that the hospital state contributes substantially to the number of features and is a major factor in cost as shown in Figure 5, showing a boxplot marking the 25th and 75th percentiles of cost for each state. We then explored weighted least squares using the weights described in section III. The features selected for the weighted model included more economic data than the unweighted model which included a more even array of economic and demographic data. Weighted least squares significantly improved our estimates, raising the adjusted R^2 value to 0.632 and lowering the RMSE to 1632.87.

B. REGRESSION TREE

Because Lee et al., Bertsimas et al., and Sushmita et al. discuss the value of tree based models, we hypothesized that a regression tree model would better explain Medicare costs than linear regression. The model would make sense for predicting medicare cost because some features, such as a hospital's state, could be significant enough to change the weights on every other variable, an interaction not captured by linear regression. To create a regression tree, we use the CART algorithm, which initially groups all nodes together, and then finds the split that minimizes the node impurities as measured by the sum of squared errors. If the reduction in node impurity is above a threshold, the node is split and the process is repeated on the child nodes. Individual regression models are created for each leaf node.

We found that using regression trees did not substantially improve our R^2, shown in Figure 6. The final tree, shown

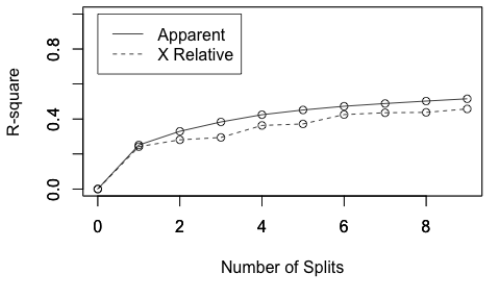

Fig. 6. Test and Train R^2 with a Regression Tree

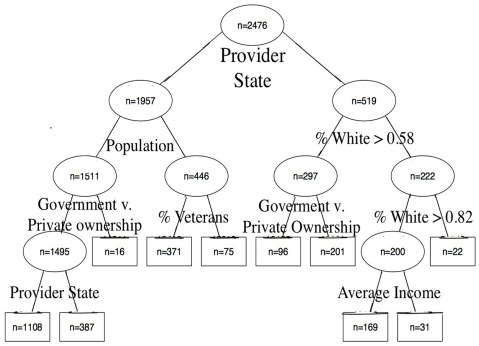

Fig. 7. Regression Tree using CART, n = number of points in each node

in Figure 7, has 10 leaf nodes and has an R^2 value of 0.57 and RMSE of 1534. The tree performed similarly to linear regression, which indicates that the interactions between features may be too complex to model with a decision tree. While other papers found trees useful for cost prediction, they were predicting costs for individual patients and features that make decision trees successful for predicting costs for individuals may not be as relevant for predicting hospital-level costs. While the tree did not provide substantially better results, it did identify state, hospital ownership, and race as some of the strongest factors, which is consistent across all of the models we applied.

C. SUPPORT VECTOR REGRESSION

Despite modifications to our linear regression models and the use of regression trees to better model the data, linear models consistently demonstrated relatively high training and test errors. Our relatively low R^2 values, which are similar to those in multiple studies that use linear models to predict costs on a per-patient basis, suggest that US healthcare costs are nonlinear. Consequently, we explore models that capture this behavior. Support vector regression fits a model such that all data points lie within ϵ of the prediction curve, while maximizing the smoothness of the prediction curve. Support vector regression optimizes the primal:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i - \xi_i^*) \\ &\text{s.t. } y_i - < w, x_i > -b \leq \epsilon + \xi_i \\ &\quad < w, x_i > +b - y_i \leq \epsilon + \xi_i^* \end{aligned}$$

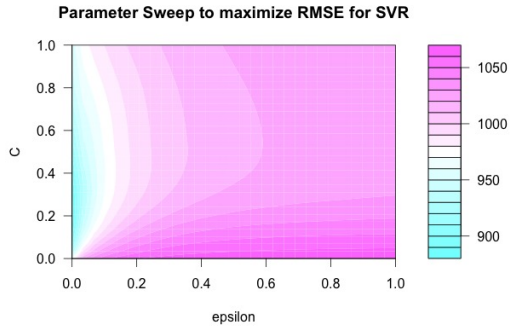

Fig. 8. Color Matrix for the SVR Parameter Sweep

$$\xi_i, \xi_i^* \geq 0.$$

The ξ and ξ^* act as slack variables that permit certain outlier data points to exceed the epsilon curve boundary without forcing a drastic change in the model. Because $||w||$ is non-convex, this primal objective is difficult to optimize. Instead, the Lagrangian dual formulation is used. Finally, a Gaussian kernel permits support vector regression to model non-linearity by mapping the data to a higher dimensional feature space. We conducted a parameter sweep over 0.01, 0.02... 1 to determine the optimal values for for C and ϵ (Figure 8). The optimal parameters of 0.05 and 0.41 produced a RMSE of 922.

D. NEURAL NETWORKS

Neural networks are known to exhibit superior behavior in modeling non-linear problems. We utilize a feedforward neural network that implements backpropagation via the Levenberg-Marquardt algorithm in the Matlab Neural Networks Toolkit. The Levenberg-Marquardt algorithm varies between gradient descent and the Gauss-Newton method via the following update rule:

$$\theta_i := \theta_i + \gamma$$

where γ is the solution to:

$$(J^T J + \lambda diag(J^T J))\gamma = J^T[y - f(\theta)].$$

The value of lambda causes the algorithm to vacillate between gradient descent and Gauss-Newton like behavior. This permits the algorithm to use the efficiency of Gauss-Newton near local minima and the robustness of gradient descent (Gavin, 2015).

The neural network is composed of an input layer, one to two hidden layers and an output layer as shown in figure X. Initially, a mean square error target cost function was used:

$$J(\theta) = \tfrac{1}{n}\sum_i e_i$$

.

However, this model tended to overfit the training data across a variety of parameters (discussed below). Consequently, a regularization parameter was introduced to correct this high variance.

$$J(\theta) = \gamma(\tfrac{1}{n}\sum_i e_i) + (1-\gamma)(\tfrac{1}{n}\sum_i w_i^2)$$

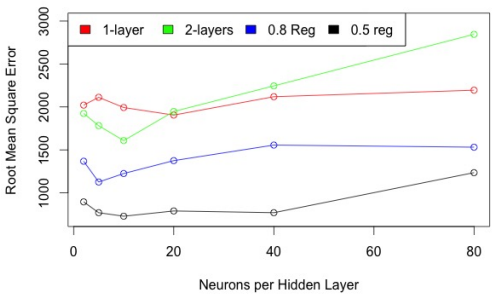

Fig. 9. The RMSE for Neural Networks with 2 to 80 Neurons per layer. Network configurations are show for 1 hidden layer, 2 hidden layers, 2 hidden layers with 0.8 regulation parameter, and 2 hidden layers with 0.5 regulation parameters

	1-Layer		2-layer		0.3 Reg		0.5 Reg		0.8 Reg	
Neurons	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
2	0.78	0.60	0.75	0.61	0.78	0.67	0.72	0.61	0.70	0.69
5	0.70	0.64	0.74	0.67	0.76	0.74	0.78	0.74	0.73	0.67
10	0.82	0.63	0.82	0.68	0.72	0.67	0.78	0.73	0.72	0.70
20	0.84	0.68	0.76	0.69	0.81	0.58	0.75	0.67	0.69	0.70
40	0.8	0.59	0.81	0.48	0.81	0.68	0.79	0.66	0.76	0.71
80	0.88	0.55	0.81	0.48	0.73	0.55	0.74	0.57	0.78	0.56

Fig. 10. R^2 values for the training and test set under various parameter configurations

This regularization parameter penalizes large weights and tends to correct overfitting.

E. PARAMETER FITTING

The hyper-parameters of a neural network must be tuned to optimize performance. In our model, the key hyper-parameters are the number of hidden layers, the number of neurons per layer and the regularization factor. One to two hidden layers are standard for most neural networks, and networks with at least one layer can reproduce any non-linear function given sufficient neurons. However, a second hidden layer can decrease the number of required neurons and remove neuron interdependencies that may interfere with fitting (Tamura, 1997, Chester 1990). Consequently, we hypothesized that two hidden layers would give optimal performance. Our results in figured 9 and 10 validated that theory.

The network was tested on configurations of 2, 5, 10, 20, 40 and 80 neurons per layer. Larger concentrations of neurons tended to overfit the data, leading to almost perfect fit in the training data and high generalization errors (Figure 7). 5-10 neurons per layer were ideal for most configurations. A regularization parameter significantly improved the RMSE and reduced overfitting. A parameter sweep of 0.1, 0.2, ... 0.9 was conducted to determine the optimal regularization. For summary purposes, results for a subset of regularization parameters are shown in Figures 9 and 10. The 0.5 regularization value produced both the lowest RMSE and highest adjusted R^2 values. Consequently, our optimal neural network consisted of 2 hidden layers each with 10 neurons and 0.5 regularization.

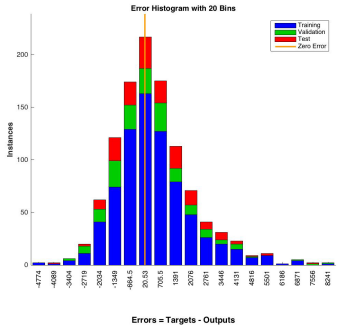

Fig. 11. Error Histogram For Optimal Neural Network

VI. RESULTS

For linear regression, regression trees and support vector regression, a random sampling of data was used to create a 20% holdout cross-validation test set. For neural networks, random sampling created a 15% validation set and 15% test set. As shown below, support vector regression and neural networks consistently outperformed our linear models.

Model	R^2	RMSE
Lin Reg	0.55	1846
W. Lin Reg	0.63	1632
Reg. Trees	.57	1534
SVR	0.68	922
NN	0.74	755

This confirms our hypothesis that US healthcare data is non-linear. Hospitals have a high degree of flexibility in pricing their chargemasters, a master list of prices for individual services that is used to negotiate for reimbursement with healthcare insurers and other payers, including Medicare. This variability likely contributes to the non-linearity of hospital pricing data. Our optimal neural network produced a RMSE of only 755, roughly 6.5% of the mean (Figure 11).

VII. CONCLUSION AND FUTURE WORK

Due to the non-linear structure of medicare data, non-linear models including support vector regression and neural networks are better at predicting hospital prices than linear regression and regression trees. Neural networks had the highest R^2 and lowest RMSE of all of the models we tried. We hypothesize that non-linear models better capture the complex interactions that lead to pricing, and that neural networks are more successful than SVR because SVR is limited to capturing linear hyperplanes in high dimensional spaces. Despite their different outputs, our models agree that state, hospital ownership, and race are among the most highly influential factors in determining the medicare costs related to heart failure and shock. In future work, modeling the impact of hospital market density and predicting the availability of cardiac doctors could improve our models by providing better supply-side data. We also hope to explore how our results generalize to a wider range of hospital conditions publicized by Medicare.

REFERENCES

[1] Bertsimas, Dimitris, et al. "Algorithmic prediction of health-care costs." Operations Research 56.6 (2008): 1382-1392.

[2] Booth, Ash, Enrico Gerding, and Frank McGroarty. "Predicting equity market price impact with performance weighted ensembles of random forests." Computational Intelligence for Financial Engineering and Economics (CIFEr), 2104 IEEE Conference on. IEEE, 2014.

[3] Dunn, D. L., A. Rosenblatt, D. A. Taira, E. Latimer, J. Bertko, T. Stoiber, P. Braun, S. Busch. 2002. A comparative analysis of methods of health risk assessment. Society of Actuaries Monograph M-HB96-1.

[4] Bertsimas, et al. "Algorithmic Predication of Healthcar Costs" Operations Research, Vol 56, no 6, November-December 2008 pp 1382-1392.

[5] Chester, Daniel L. "Why two hidden layers are better than one." Proceedings of the international joint conference on neural networks. Vol. 1. 1990.

[6] Hancock, Jay. "Attention, Shoppers: Prices For 70 Health Care Procedures Now Online!" NPR. 26 February 2015

[7] Lee, SM, Kang, JO, and Suh, YM. "Comparison of Hospital Charge Prediction Models for Colorectal Cancer Patients: Neural Network vs. Decision Tree Models" Journal of Korean Medical Science., 19 October 2004 pp 677-681.

[8] Mackenzie, Andrew, et al. Predictive Healthcare Cost Modeling Using Regression. Issue 1. Society of Actuaries Conference, 2013.

[9] Penberthy, et al. "Predictors of Medicare Costs in Elderly Beneficiaries with Breast, Colorectal, Lung, or Prostate Cancer." Health Care Management Science, 1999.

[10] Smith, et al. "Predicting Inpatient Hospital Payments in the United States: a retrospective analysis." BMC Health Services Research, 2015.

[11] Sushmita, Shanu, et al. "Population Cost Prediction on Public Healthcare Datasets." Proceedings of the 5th International Conference on Digital Health 2015. ACM, 2015.

[12] Tamura, Shin'ichi, and Masahiko Tateishi. "Capabilities of a four-layered feedforward neural network: four layers versus three." Neural Networks, IEEE Transactions on 8.2 (1997): 251-255.

[13] Therneau, Terry, Atkinson, Beth, and Ripley, Brian. rpart: Recursive Partitioning and Regression Trees. https://cran.r-project.org/web/packages/rpart/index.html