Population Cost Prediction on Public Healthcare Datasets

Shanu Sushmita
Center for Data Science
Institute of Technology
University of Washington,
Tacoma
sshanu@uw.edu

Stacey Newman
Center for Data Science
Institute of Technology
University of Washington,
Tacoma
newmsc8@uw.edu

James Marquardt
Center for Data Science
Institute of Technology
University of Washington,
Tacoma
jamarq@uw.edu

Prabhu Ram, Virendra
Prasad
Edifecs, Bellevue
[prabhu.ram, virendra.prasad]@edifecs.com

Martine De Cock
Center for Data Science
Institute of Technology
University of Washington,
Tacoma
mdecock@uw.edu

Ankur Teredesai
Center for Data Science
Institute of Technology
University of Washington,
Tacoma
ankurt@uw.edu

ABSTRACT

The increasing availability of digital health records should ideally improve accountability in healthcare. In this context, the study of predictive modeling of healthcare costs forms a foundation for accountable care, at both population and individual patient-level care. In this research we use machine learning algorithms for accurate predictions of healthcare costs on publicly available claims and survey data. Specifically, we investigate the use of the regression trees, M5 model trees and random forest, to predict healthcare costs of individual patients given their prior medical (and cost) history. Overall, three observations showcase the utility of our research: (a) prior healthcare cost alone can be a good indicator for future healthcare cost, (b) M5 model tree technique led to very accurate future healthcare cost prediction, and (c) although state-of-the-art machine learning algorithms are also limited by skewed cost distributions in healthcare, for a large fraction (75%) of population, we were able to predict with higher accuracy using these algorithms. In particular, using M5 model trees we were able to accurately predict costs within less than $125 for 75% of the population when compared to prior techniques. Since models for predicting healthcare costs are often used to ascertain overall population health, our work is useful to evaluate future costs for large segments of disease populations with reasonably low error as demonstrated in our results on real-world publicly available datasets.

1. INTRODUCTION

Increasing cost of healthcare continues to be one of the world's biggest challenges. According to *The Commonwealth Fund*[1], healthcare costs for 2013 in the United States were highest across the world ($8,508 per-capita). Yet, amongst comparable nations, the United States ranks lowest in terms of quality of care [6]. The problem is not unique to the United States though: from the price of medications and the cost of hospital stays to physicians' fees and medical tests, healthcare costs around the world are skyrocketing. Much of this can be attributed to wasteful spending caused by ineffective drugs and duplicate procedures and paperwork, as well as missed disease-prevention opportunities[2] [5]. With a goal of changing this, healthcare reform policies are currently underway, promoting initiatives for managing the overall health of a population while keeping costs manageable[3].

Accurate prediction of healthcare cost is of immense importance to improve accountability in care. The study of healthcare cost analyses is often directed at getting the most accurate estimate of the mean costs of treating the disease, or identifying the patients'/structure characteristics influencing costs and getting an estimate of expected costs [8]. This study is focused on the latter. Statistical methods for estimating the expected healthcare cost have been largely investigated [2, 7], but are still inconclusive and have several open challenges [8]. Furthermore, previous efforts in the literature for algorithmic prediction of healthcare costs are dominated by linear regression and by rule-based approaches, which require a lot of domain knowledge. But more recently, machine learning approaches like clustering and classification are also being explored for this purpose [1, 11].

Healthcare cost prediction is a challenging problem from the data mining perspective as well. It is well recognized that statistical analysis of healthcare cost data poses a number of difficulties [16]. Cost data is characterized by highly non-Gaussian distributions, which poses interesting challenges for prediction and comparison goals. Data on medical expenditures or costs of treatment typically feature a

DH '15 May 18 - 20, 2015, Florence, Italy
Copyright © 2015 ACM 978-1-4503-3492-1/15/05 ...$15.00.
http://dx.doi.org/10.1145/2750511.2750521.

[1] http://www.commonwealthfund.org/publications/fund-reports/2014/jun/mirror-mirror

[2] http://theinstitute.ieee.org/technology-focus/technology-topic/better-health-care-through-data

[3] http://www.hhs.gov/healthcare/facts/timeline

spike at zero and a strongly skewed distribution with a heavy right-hand tail [9]. Another, important challenge is to leverage existing large and varied claims, clinical, and survey data to estimate future healthcare costs, and to take measures in care-management that reduce such costs while improving overall population health.

The goal of this research is to explore machine learning algorithms to provide powerful tools for accurate predictions of healthcare costs. In this paper, we investigate the use of three state-of-the-art machine learning algorithms – regression tree, M5 model tree and random forest, to predict the healthcare costs of individual patients based on data about their previous medical and cost history. To the best of our knowledge, the utility of these algorithms in healthcare costs prediction has not been investigated. We train and test all algorithms on two separate datasets, more specifically a claims (SID) and a survey (MEPS) dataset. Furthermore, we train separate prediction models for four future cost scenarios – three months, six months, nine months and twelve months respectively. That is, given the past, we predict the healthcare cost for the coming three months, six months and so on. We discuss more on this in Section 4. To summarize, the presented research makes the following contributions:

1. Empirical evaluation of regression tree, M5 model tree and random forest to provide powerful tools for accurate predictions of healthcare costs.

2. Empirical evaluation of the regression tree, M5 model tree and random forest on two kinds of datasets – claims and survey data.

3. Empirical evaluation of the regression tree, M5 model tree and random forest for four different future scenarios – three months, six months, nine months and twelve months.

The rest of the paper is organized as follows: In Section 2 we discuss the related background. Two datasets (SID and MEPS) are described in Section 3. Four future cost prediction scenarios are described in Section 4. The three algorithms that we investigate in this study are explained in Section 5. The performance of these algorithms is discussed in Section 6. In Section 7, we show our online cost prediction service where these models are currently deployed. Finally, in Section 8 we conclude with our overall findings.

2. RELATED WORK

Methods: Previously proposed cost prediction models often used rule based methods and multiple linear regression (MLR) models. The challenge with the rule based methods (e.g. [10]) is that they require a lot of domain knowledge, which is not easily available and is often expensive. MLR models are powerful tools for capturing the relationships between the exploratory variables and the dependent variable, but, working with several independent variables often causes the multicollinearity problem, which is caused by the presence of significant correlations among the predictors [18]. In addition, their performance is challenged by the skewed healthcare data. Healthcare cost data typically feature a spike at zero, and a strongly skewed distribution with a heavy right-hand tail [9]. As a result, the prediction models are posed with the challenge of an extreme value situation. It is known that regression models are sensitive to extreme values and likely to be inefficient in small to medium sample sizes if the underlying distribution is not normal [16]. In the past, several advanced statistical methods have been proposed to accommodate the skewness observed in healthcare data, such as General Linear Models (GLM) [13], mixture models based on mixtures of parametric models [17], Markov models [15] etc. For a comprehensive comparison of previously proposed statistical methods for healthcare cost prediction, we refer to the review paper by Mihaylova et al. [16].

The development of accurate healthcare cost prediction models using machine learning methods has been more recent. Bertsimas et al. [1] utilize classification tree and clustering algorithms to provide predictions of healthcare costs in the third year by applying data mining methods to medical and cost data from the first two years. While Lahiri et al. [11], investigate classification algorithms to predict whether an individual is going to incur higher or lower healthcare expenditure. In this paper, we investigate three additional state-of-the-art machine learning algorithms – regression tree, M5 model tree and random forest. To the best of our knowledge, their utility for the healthcare costs prediction problem has not been investigated before.

Evaluation: In previous studies, the models would often estimate the mean cost of the given sampling distribution with a certain confidence interval. Estimation methods utilize observed data as inputs, and do not consider new observations (i.e., observations not in the sample). As a result in these studies, only in-sample data were used to report predictive performance of the methods. In addition, because the relations found by chance in the estimation sample will not replicate, the predictive performance of the model often deteriorates significantly when applied to new data [2]. Therefore, it is more appropriate to express the predictive performance of a method based on out-of-sample experiments (that is, use data that the method has not seen) [1]. In this paper, we divide the data into three parts: a training sample, a validation sample and a testing sample. We measured the performance of the prediction algorithms using the root mean square error (RMSE), mean absolute error (MAE), and the prediction error quantiles. These measures are discussed in detail in Section 6.

Data: The underlying data for building the predictive models often come from claims data, clinical data and/or self-reported survey data (i.e., questionnaires). These datasets are sometimes used separately (e.g., CDPS[4] uses claims data, or PRA[5] uses self-reported data) or as a combination of one or more datasets (e.g., Dorr's algorithm from Care Management Plus[6] uses claims and questionnaire data). Although the predictive power of claims data is often challenged, its utility has been established through several dedicated studies [1, 20]. Furthermore, in the context of building a healthcare cost prediction system for individuals, being able to leverage claims and survey data is beneficial since often the privacy concerns associated with clinical data (such as lab results, vitals, etc.) are far more constrained than those associated with claims and survey data. In this study, we separately train and test our algorithms on claims (SID) as well as survey data (MEPS).

[4] http://cdps.ucsd.edu

[5] https://www.highmarkblueshield.com/pdf_file/ger_binder/pra-overview.pdf

[6] http://caremanagementplus.org

3. DATASET AND FEATURES

In this section we describe the two datasets we use in our study, namely SID and MEPS. In addition, we provide information on the feature set, pre-processing steps, and descriptive statistics of these datasets.

3.1 SID Dataset – Claims Data

The Healthcare Cost and Utilization Project (HCUP) provides Washington State Inpatient Database (SID). In this study we use data from the years 2009 to 2012. The dataset is comprised of a combination of clinical and claims data, collected through a partnership between HCUP and the Washington State Department of Health. This dataset contains several feature sets – diagnoses, procedures, admission charges, comorbidities, revenue codes for specific services, hospital specific data, and injury descriptions. A particularly useful feature of this dataset is the presence of a unique patients' identifier that allows the admissions to be tracked across a given year. More detailed information about the data and its features is available on the HCUP-SID website[7].

In order to build predictive models, we focus on a set of features described as useful in [1]. These include grouper codes for diagnoses and procedures, utilization code units, revenue codes, comorbidities, number of chronic conditions, age, gender, and race. A summary of the feature set used in this study is shown in Table 1.

Variable Number	Description
1-3	Demographics
4-6	Patients' history
7-271	Diagnosis Groups
272-504	Procedures
505-816	Revenue Codes
817-845	Comorbidities
846	Previous Year cost

Table 1: Summary of the features used from the SID dataset.

During the pre-processing stage, the raw SID data was transformed to the individual level from the admission level. Every row in the raw SID data corresponds to a single admission (in the hospital) and its associated information (cost, diagnosis, length of stay, etc.). It is possible that an individual may have several hospital admissions in the given training period, thus leading to several rows in the data. To predict the future healthcare cost of an individual, we use medical and cost information of previous (all) admissions of that individual. In order to transform data from admission to individual level, we summed the values over multiple admissions for diagnosis, procedure, utilization, and cost variables for an individual. For the comorbidity variables, a logical OR was applied. For demographic variables, the value present in the future time window was retained.

In Table 2 we provide descriptive statistics of the SID data after pre-processing. It can be seen that the total number of unique beneficiaries is very large – over 1.5 million. Overall, there are more admissions of female (≈ 60%) than male (≈ 40%) patients. The average healthcare cost of an individual is approximately $50,000, reflecting the issue of increase in the healthcare cost (also discussed in Section 1).

Statistics	**2009-2012**
Unique Beneficiaries	1,884,223
Mean Cost (in dollars)	46,598.35
Mean Age (SD)	45.01 (27.88)
Males	39.71%
Females	60.29%

Table 2: Descriptive statistics from the SID data.

3.2 MEPS Dataset – Survey Data

The Medical Expenditure Panel Survey (MEPS) dataset consists of data extracted from responses to panel surveys given to households and their employers, medical providers, and insurance providers over two year periods[8]. The survey is intended to be representative of the national civilian non-institutionalized population of the United States. Data is available for two year periods from 1996 to 2012. This study utilizes the data from 2004-2012. Data is collected from households[9] for each collection period in a series of five interviews. From those respondents, a sample of their medical providers is contacted in order to provide more accurate medical information. For the purpose of this study, we focus our efforts on the data available in the public MEPS dataset. These are primarily demographic, diagnosis, and care provider outcome based variables. The summary of the feature set used in this study is shown in Table 3.

Variable Number	Description
1-226	Diagnosis Group
227	Previous Year Cost
228-231	Demographics
232-235	Patients' History

Table 3: Summary of the features used from the MEPS dataset.

For the MEPS dataset, we aggregate the data grouping by beneficiary. The feature vector for each beneficiary becomes the number of times a diagnosis was reported during the first year of the survey, their demographics, and their previous and future costs.

Statistics	**2004-2012**
Unique Beneficiaries	128,312
Households	43,490
Mean Cost (in dollars)	6,518
Mean Age (SD)	34.34 (22.92)
Males	23.09%
Females	9.08%
Unspecified Gender	67.83%

Table 4: Initial statistics from the MEPS data.

Table 4 shows initial statistics about the MEPS data. On an average there are three members per family (households). When compared to the SID dataset, the number of features are low in the MEPS data (see Tables 1 and 3), thus suggesting richer and detailed level of information available in the SID dataset. The average cost (mean cost) of a household

[7] http://www.hcup-us.ahrq.gov/sidoverview.jsp

[8] http://meps.ahrq.gov/mepsweb/

[9] Household here means all members of a house (family)

is much lower when compared to the SID data average cost (Table 2). This is expected because SID data is exclusively about patients who visited hospitals, and hospital visit cost are often higher. Another difference between SID and MEPS data is the distribution of male and female. The SID data shows more balance in terms of gender distribution, whereas the distribution is not very clear in the MEPS data due to a large number of unspecified genders.

In summary, the SID data is specifically about hospital admissions (in-patients), while MEPS data contains data about a broader range of medical events (not necessarily in-patient). This is an interesting advantage of the MEPS data over the SID data, and an interesting aspect of our study. On the other hand, the MEPS data is noisier than the SID data because of the way it is collected.

4. PROBLEM DESCRIPTION

The goal of this research is to predict the future healthcare cost of individuals based on their past medical and cost information. More formally, we treat the cost prediction problem as a supervised machine learning problem. The input consists of an attribute vector $\mathbf{x} = (x_1, x_2, x_3......x_p)$ that contains all available data about the individual, including general demographics such as age and gender, as well as specific clinical and claims data (including cost) for the training period. The output attribute that we aim to predict is the cost y for the result period.

We define *four* future scenarios for predicting cost in this study, that is, predicting future cost for three months, six months, nine months and twelve months. The algorithms (described in Section 5) are trained and tested for the following four scenarios:

P1: We use the history (medical, demographic and cost) of the first three months to predict the cost of the following nine months. Patients with at least one admit in the previous three months were used for the experiments.

P2: We use the history (medical, demographic and cost) of the first six months to predict the cost of the following six months. Patients with at least one admit in the previous six months were used for the experiments.

P3: We use the history (medical, demographic and cost) of the first nine months to predict the cost of the following three months. Patients with at least one admit in the previous nine months were used for the experiments.

P4: We use the history (medical, demographic and cost) of the first twelve months to predict the cost of the following year (twelve months). Patients with at least one admit in the previous year were used for the experiments.

The problem scenarios **P1, P2,** and **P3** are trained and tested using the SID dataset, while for **P4** scenario, we used the MEPS dataset only. This is because, a major downside of the publically available SID dataset is that it is not possible to track individuals across multiple years. Therefore, learning from the previous year to predict healthcare cost for the coming year was not possible with the SID data as in the fourth scenario **P4**. To overcome this limitation, we performed experiments for the fourth scenario using the MEPS dataset. Furthermore, to derive meaningful comparisons between models trained using SID data and MEPS data, we used same features when available in both datasets, or similar features (capturing similar information) in the absence of identical features.

It should be noted that the goal of this research is neither to compare the utility of the SID and MEPS dataset, nor to compare the performance of algorithms on different datasets. Instead, the objective is to perform an empirical evaluation of regression tree, M5 model tree and random forests for the purpose of healthcare cost prediction. These algorithms are trained using the SID and MEPS data depending on the problem scenario at hand.

5. METHODS

In this section we describe the machine learning algorithms including the four baseline models used in this study.

5.1 Baseline Algorithms

1. **Average Baseline:** For this baseline model, we calculate the mean μ of all the individuals' cost within the training set for the training period. The mean (μ) is then multiplied by a factor (k) to make the prediction for all individuals in the test set. For **P1**, the predicted cost = $3 \times \mu$, for **P2**, the predicted cost = $1 \times \mu$, for **P3**, the predicted cost = $0.33 \times \mu$, and for **P4**, the predicted cost = $1 \times \mu$.

2. **Previous Cost Regression (PCR):** For this baseline, a linear regression model is fitted using only the previous cost during the training period as a predictor variable. This model is then used to predict the cost for the testing period.

3. **Multiple Linear Regression (MLR):** We use a multiple linear regression model to predict the cost using a p - dimensional vector of predictive variables. The difference between PCR and MLR is that all features (as shown Table 1 and 3) from SID and MEPS data were used to train the MLR models, while only 'cost' variable was used in the PCR models. We use MLR as an additional baseline due to its extensive use in the literature of healthcare cost prediction.

4. **Generalized Linear Model (GLM):** GLM is a generalization of ordinary least squares regression that relaxes the assumption that the distribution of the response variable be normal[10]. A GLM will consist of a linear predictor (as in ordinary least squares linear regression), a link function that describes how the mean depends on the linear predictor, and a variance function that describes how the variance depends on the mean. For our experiments with GLMs, we assume a Poisson distribution and select the log link function.

5.2 Regression Tree (RTree)

When the data has lots of features which interact in a nonlinear way, assembling a single global model (multiple linear regression) can be very difficult, and confusing. An alternative approach to nonlinear regression is to partition the space into smaller regions, where the interactions are

[10] http://en.wikipedia.org/wiki/Generalized_linear_model

more manageable. The goal of a regression tree is to predict a response y (cost in our case) from inputs $x_1, x_2, ...x_p$. This is done by growing a binary tree. At each internal node in the tree, a test is applied to one of the inputs, for example x_i. Depending on the outcome of the test, the left or the right sub-branch of the tree is then selected. Eventually a leaf node is reached, where the prediction is made. For this study, we used an implementation of classification and regression tree algorithm (CART)[4] in R. The minimum deviance (mean squared error) was used as the test parameter for proceeding with a new split. That is, adding a node (feature selection) should reduce the error by at least a certain amount. We tested the performance of regression trees using different complexity parameters (cp=0.01, 0.001, 0.0005). In Table 5 we report the best performing tree with cp set to 0.0005 value.

5.3 Random Forest Regression (RFR)

Random forest regression is an ensemble learning method that operates by constructing a multitude of regression trees at training time and outputting the mean prediction of the individual trees for new observations. Each tree is constructed using a random sample of the row (observation) and column (feature) space of the original dataset. This has the effect of correcting the tendency of individual regression trees to overfit the training data [3]. For this work we utilize the implementation of random forests in the R, and grow all regression trees without pruning[12].

5.4 M5 Model Tree (M5)

M5 model trees are a generalization of the CART model. The structure of a M5 model tree follows that of decision tree, but has multiple linear regression models at the leaf nodes, making the model a combination of piecewise linear functions. The algorithm for the training of a model tree breaks the input space of the training data through a recursive partitioning process similar to the one used in CART. After partitioning, linear regression models can be fit on the leaf nodes, making the resulting regression model locally accurate [19].

All the models, including baseline models and three machine learning algorithms are trained using R.

6. RESULTS

We measure the performance of the prediction algorithms using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), with a lower error indicating a better performance. We test our models using 10-fold cross validation. Traditionally, R^2 or adjusted R^2 have been used to evaluate healthcare cost prediction models, but there are some drawbacks to their use. While R^2 does measure how well a model fits the training data, the metric is not a true indicator of how well a model will predict unseen observations [1]. Therefore, we use MAE and RMSE in this study. Both metrics measure the predictive quality of a model, with RMSE being more sensitive to the outliers. Formally, MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i|,$$

and RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}},$$

where $\hat{y}_i$ is the predicted value, y_i the true value, and n is the number of observations in the sample.

In addition to the single value metrics described, we measure the performance of the algorithms by examining the absolute prediction error distribution. As models for predicting healthcare costs are often used to predict the overall healthcare cost of a population, it is useful to evaluate whether large segments of a test set are predicted with reasonably low error.

The RMSE and MAE results for the four future scenarios **P1, P2, P3** and **P4** are shown in Table 5. In addition to results for the regression trees, M5 model tree and random forest, we include results for all four baseline algorithms – average, previous cost, linear regression and GLM (described in Section 5).

Algorithm	RMSE ($)	MAE ($)
P1 (SID Data)		
AB	127,156	115,875
PCR	64,414	27,669
MLR	92,064	29,007
GLM	152,779	81,497
RTree	67,664	26,480
M5	79,790	20,715
RFR (n =50)	66,340	25,655
P2 (SID Data)		
AB	72,367	51,970
PCR	53,405	19,633
MLR	126,062	22,448
GLM	157,785	79,443
RTree	67,344	19,057
M5	77,242	14,607
RFR (n =50)	52,581	17,563
P3 (SID Data)		
AB	69,239	23,854
PCR	61,259	12,710
MLR	82,361	18,652
GLM	233,733	80,485
RTree	72,888	11,671
M5	68,864	7,647
RFR (n = 50)	66,066	11,293
P4 (MEPS Data)		
AB	36,328	12,474
PCR	34,715	11,387
MLR	33,631	10,597
GLM	37,397	8,886
RTree	34,434	10,568
M5	35,476	8,112
RFR (n= 50)	33,618	10,060

Table 5: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) using SID and MEPS data. Where, AB = Average Baseline, PCB = Previous Cost Regression, MLR = Multiple Linear Regression Baseline, GLM = Generalized Linear Models, RTree = Regression Tree, M5 = M5 Model Tree, and RFR = Random Forest. For RFR, n is the number of trees, and for GLM, we assume a Poisson distribution and use the log link function.

As can be seen in Table 5, for future scenario **P1**, all three models outperformed (lower prediction errors) the baseline models in terms of MAE. Among them, M5 model tree had the lowest MAE. With respect to the RMSE values, all three algorithms performed better than average, multiple linear regression and GLM baselines, but, none could outperform the *previous cost* baseline (PCR).

For future scenario **P2**, all three models performed better than the baseline models with respect to MAE values, and the M5 model tree was the best performing model again. With respect to RMSE, random forest was the best performing model, but, regression and M5 model tree could not outperform the *previous cost* baseline (PCR).

In case of **P3** scenario, mean absolute prediction errors (MAE) by M5 model trees were very low when compared to all baseline models. Regression tree and random forest were the next best performing models. As seen in previous scenarios (**P1, P2**), RMSE values were again very high for all models when compared to MAE values, GLM being the worst model. None of the machine learning algorithms (regression tree, random forest and M5 model trees) could outperform the *previous cost* baseline (PCR) when compared against RMSE values.

Finally, for the **P4** scenario, M5 model tree again had the lowest MAE values, but the performance of the GLM baseline was also comparable. Random forest had the lowest RMSE values but the difference was not large when compared to other algorithms. RMSE values of the regression tree, M5 model tree and all four baselines were quite close (similar error values). It should be noted that results for **P4** were obtained using MEPS data, while results for **P1, P2** and **P3** were obtained using SID data.

Overall, three key observations can be made from the performance results shown in Table 5. First, *previous cost regression (PCR)* is a strong baseline, and therefore previous healthcare cost alone can be a good indicator for future healthcare cost. Second, among the three machine learning algorithms (regression tree, random forest and M5 model tree), M5 model tree consistently performed best and achieved a substantially lower MAE than strong PCR baseline for predicting future healthcare cost in all scenarios. Third, while all three machine learning algorithms (regression tree, random forest and M5 model tree) outperform the PCR baseline with respect to MAE, their comparative performance with respect to RMSE is far less convincing. This might indicate that while the three algorithms do really well on average, they are prone to higher variance and make much larger errors than PCR baseline.

In order to further investigate the error values (in Table 5), we looked at the quartiles of the absolute error (deviation from the actual cost) distribution for all the algorithms. The error distribution results are shown in Figure 1, 2, 3 and 4. The x axis shows the proportion of the population, and the y-axis shows the error distribution in dollar amount. For visualization and discussion purposes, we include error distribution upto 75 percentile only. Complete distribution (upto 100 percentile) can be seen in the Table[11] link provided.

For all four scenarios, it can be seen that for 75% of the test data, the error on the predicted cost by the models was less than $30,000. In particular, for the **P3** scenario, the error is as low as $125 using M5 model. This result is promising because for a large fraction of the population (75%), we were able to predict with higher accuracy using these algorithms. This is an interesting observation because from an accountable care organizations (ACOs) perceptive, predicting aggregated cost for population can be very useful. For the **P2, P3** and **P4** scenario, the error values were between $0 - $7,000 dollars for 50% of test data (See Figure 2, 3 and 4). Although the results are promising, but they also reflect the challenges of modeling skewed distribution of healthcare cost data because of large overall RMSE and MAE values (also discussed in Section 2), and that modern machine learning algorithms are not immune to this issue. It might be possible to improve the performance of these algorithms by removing extreme values. This could be an interesting problem to investigate in the future.

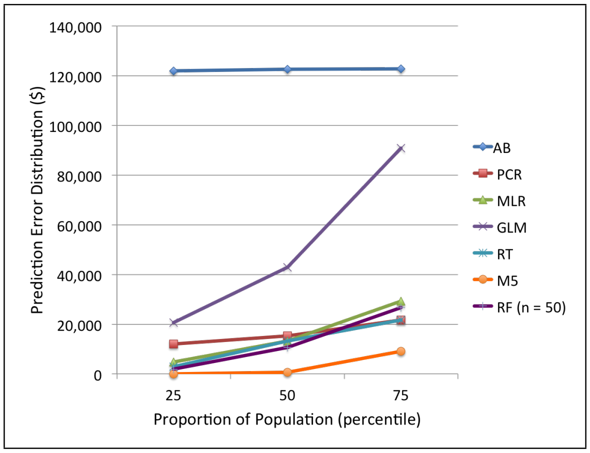


Figure 1: Absolute Error distribution summary for models in **P1** scenario. The x axis shows the proportion of the population, and the y-axis shows the error distribution in dollar amount. Here, AB = Average Baseline, PCR = Previous Cost Regression (Baseline), MLR = Multiple Linear Regression (Baseline), GLM = Generalized Linear Models (Baseline), RTree = Regression Tree, M5 = M5 Model Tree, and RFR = Random Forest Regression. For RFR, n is the number of trees, and for GLM, we assume a Poisson distribution and use the log link function.

7. HEALTHCARE SCALABLE COST PREDICTION ENGINE

The models described in previous sections are deployed on HealthSCOPE (Healthcare Scalable COst Prediction Engine)[12], which is a framework for exploring historical and present day healthcare costs as well as for predicting future costs. HealthSCOPE can be used by individuals to estimate their healthcare costs in the coming year. In addition, HealthSCOPE supports a population based view for actuaries and insurers who want to estimate the future costs of a population based on historical claims data, a typical scenario for accountable care organizations (ACOs).

Using our interactive data mining framework, users can view claims (sample files are provided for demo purposes), use HealthSCOPE to predict costs for the upcoming year, interactively select from a set of possible medical conditions, understand the factors that contribute to the cost, and compare costs against historical averages (See Figure

[11] http://tinyurl.com/ErrorDistributionTable

[12] http://healthscope.cloudapp.net/hscope-dev/aco/

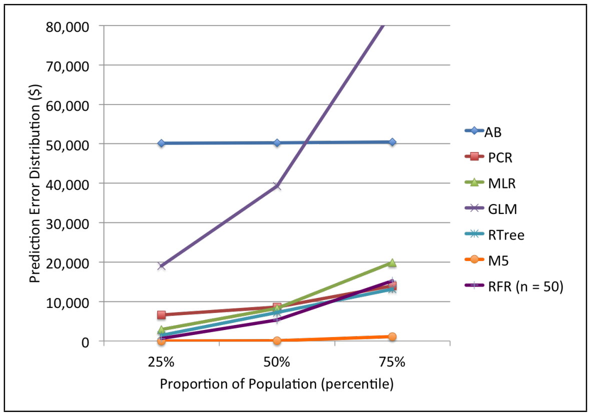


Figure 2: Absolute Error distribution summary for models in **P2** scenario. The x axis shows the proportion of the population, and the y-axis shows the error distribution in dollar amount. Here, AB = Average Baseline, PCR = Previous Cost Regression (Baseline), MLR = Multiple Linear Regression (Baseline), GLM = Generalized Linear Models (Baseline), RTree = Regression Tree, M5 = M5 Model Tree, and RFR = Random Forest Regression. For RFR, n is the number of trees, and for GLM, we assume a Poisson distribution and use the log link function.

5). The back-end system contains cloud based prediction services hosted on the Microsoft Azure infrastructure that allow the easy deployment of models encoded in Predictive Model Markup Language (PMML) and trained using either Spark MLLib or various non-distributed environments. More details about the underlying framework and individual modules can found here [14].

8. CONCLUSION

Accurate prediction of healthcare cost is of immense importance to improve accountability in care. As a result, the analysis of healthcare costs has become an important part of both experimental and epidemiological research. The goal of the presented research was to investigate modern machine learning algorithms for the task of predictions of future healthcare costs. In this paper, we investigated three algorithms – regression tree, M5 model tree and random forest using claims and survey data. In addition, empirical evaluation of these algorithms for four different future cost prediction scenarios – three months, six months, nine months and twelve months were also performed. Overall, three key observations were made during this study. First, previous healthcare cost alone can be a good indicator for future healthcare cost. Second, M5 model tree shows potential for solving future healthcare cost prediction problems. Third, state-of-the-art machine learning algorithms are also limited by the skewed distribution of healthcare cost data. However, for a large fraction (75%) of the population, we were able to predict with higher accuracy using these algorithms. As models for predicting healthcare costs is often used to predict the overall healthcare cost of a population, it is useful to evaluate whether large segments of a test set are predicted with reasonably low error. As continuation of this analysis, we plan to take a deeper dive into the data and explore ways to improve the performance of algorithms, may be through feature selection, or by filtering the outliers that are several orders larger than the other samples when calculating the errors (in particular RMSE). In addition, identifying (automatically) which are the difficult cases (individuals, sub-population, etc) to predict accurately (have large errors), and build dedicated models for those cases.

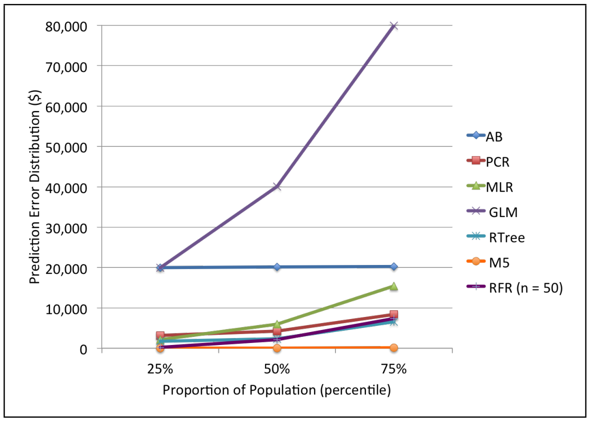


Figure 3: Absolute Error distribution summary for models in **P3** scenario. The x axis shows the proportion of the population, and the y-axis shows the error distribution in dollar amount. Here, AB = Average Baseline, PCR = Previous Cost Regression (Baseline), MLR = Multiple Linear Regression (Baseline), GLM = Generalized Linear Models (Baseline), RTree = Regression Tree, M5 = M5 Model Tree, and RFR = Random Forest Regression. For RFR, n is the number of trees, and for GLM, we assume a Poisson distribution and use the log link function.

Acknowledgements

We would like to thank Dr. George Wu, David Hazel and Senjuti Roy Basu for their helpful suggestions and feedback. We would also like to thank the Azure For Research[13] program from Microsoft Research Connections for the compute resources grant enabling us to use Azure infrastructure for this research, and Edifecs Inc., for their generous support to the Center for Data Science, and Zementis for the donation of licenses for their ADAPA software.

9. REFERENCES

[1] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang. Algorithmic prediction of health-care costs. *Operations Research*, 56(6):1382–1392, 2008.

[2] M. Bilger and W. G. Manning. Measuring overfitting in non-linear models: A new method and an application to health expenditures. *International Journal of Health Economics*, 24(1):75–85, 2015.

[3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth Publishing Company, 1984.

[13] http://azure4research.com/

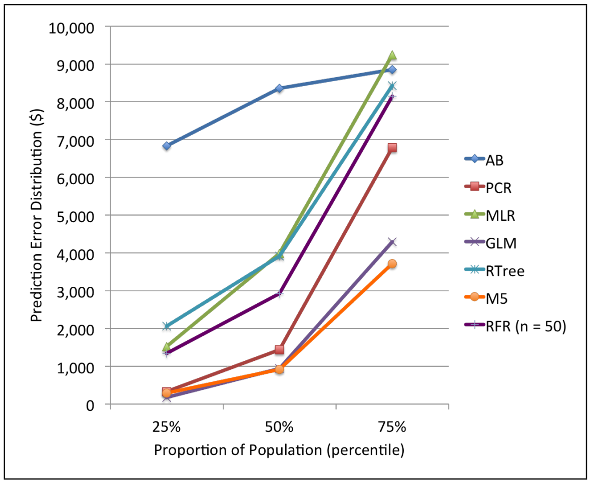


Figure 4: Absolute Error distribution summary for models in **P4** scenario. The x axis shows the proportion of the population, and the y-axis shows the error distribution in dollar amount. Here, AB = Average Baseline, PCR = Previous Cost Regression (Baseline), MLR = Multiple Linear Regression (Baseline), GLM = Generalized Linear Models (Baseline), RTree = Regression Tree, M5 = M5 Model Tree, and RFR = Random Forest Regression. For RFR, n is the number of trees, and for GLM, we assume a Poisson distribution and use the log link function.

[5] V. Chandola, S. R. Sukumar, and J. C. Schryver. Knowledge discovery from massive healthcare claims data. In *Proceedings of the 19th ACM SIGKDD*, pages 1312–1320, 2013.

[6] N. R. Council and I. of Medicine. *U.S. Health in International Perspective: Shorter Lives, Poorer Health*. The National Academies Press, 2013.

[7] P. Diehr, D. Yanez, A. Ash, M. Hornbrook, and D. Y. Lin1. Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, 20(1):125–44, 2007.

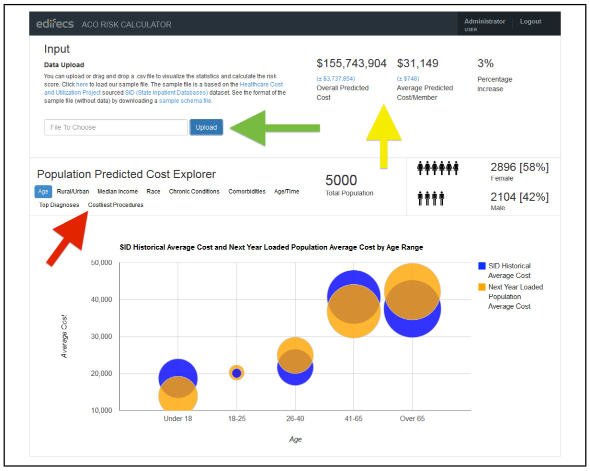
[8] D. Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti, and E. Pagano. Regression models for analyzing costs and their determinants in health care: an introductory review. *International Journal of Quality Health Care*, 23(3):331–41, 2011.

[9] A. Jones. Models For Health Care. Technical report, HEDG, c/o Department of Economics, University of York, Jan. 2010.

[10] R. Kronick, T. P. Gilmer, T. Dreyfus, and T. G. Ganiats. Cdps-medicare: The chronic illness and disability payment system modified to predict expenditures for medicare beneficiaries. Technical report, 2002.

[11] C. B. Lahiri and N. Agarwal. Predicting healthcare expenditure increase for an individual from medicare data. In *Proceedings of the ACM SIGKDD Workshop on Health Informatics*, 2014.

[12] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.



You can upload or drag and drop a .csv file to visualize the statistics and calculate the risk score. Click here to load our sample file. The sample file is a based on the Healthcare Cost and Utilization Project sourced SID (State Inpatient Databases) dataset. See the format of the sample file (without data) by downloading a sample schema file.

Figure 5: Screenshot showing previous and predicted costs for the group of individuals across different age groups. The green arrow indicates the form for uploading healthcare data to be evaluated, the red arrow indicates navigation options for different population visualizations, and the yellow arrow indicates population cost predictions. In the current visualization, it can be seen that for the individuals over age 65, future healthcare cost (yellow bubble) is predicted to be higher that the previous year cost (blue bubble).

[13] W. G. Manning, A. Basu, and J. Mullahy. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24(3):465–488, 2005.

[14] J. Marquardt, S. Newman, D. Hattarki, R. Srinivasan, S. Sushmita, P. Ram, V. Prasad, D. Hazel, A. Ramesh, M. D. Cock, and A. Teredesai. Healthscope: An interactive distributed data mining framework for scalable prediction of healthcare costs. In *In Proceedings of the IEEE ICDM*, 2014.

[15] A. H. Marshall, B. Shaw, and S. I. McClean. Estimating the costs for a group of geriatric patients using the Coxian phase-type distribution. *Statistics in Medicine*, 26:2716–2729, 2007.

[16] B. Mihaylova, A. Briggs, A. O'Hagan, and S. G. Thompson. Review of statistical methods for analysing healthcare resources and costs. *Health Economics*, 20(8):897–916, 2011.

[17] J. Mullahy. Heterogeneity, excess zeros, and the structure of count data models. *Journal of Applied Econometrics*, 12(3):337–50, 1997.

[18] C. V. PATRICHE, R. G. PIRNAU, and B. ROÅđCA. Comparing linear regression and regression trees for spatial modelling of soil reaction in dobrovĂČÅč basin (eastern romania). *Bulletin UASVM Agriculture*, 68(1):264–271, 2011.

[19] J. R. Quinlan. Learning with continuous classes. In *In Proceedings of AI*, pages 343–348. Adams and Sterling, 1992.

[20] Y. Zhao, A. S. Ash, R. P. Ellis, J. Z. Ayanian, G. C. Pope, B. Bowen, and L. Weyuke. Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Journal of Medical Care*, 43(1):34–43, 2005.