Predicting Healthcare Expenditure Increase for an Individual from Medicare Data

Bibudh Lahiri
Impetus Technologies
720 University Avenue
Los Gatos, CA 95032
blahiri@impetus.com

Nitin Agarwal
Impetus Infotech (India)
18 Palasia, A.B. Road
Indore, India, 452001
nitin.agarwal@impetus.co.in

ABSTRACT

Healthcare expenditure is a growing concern in the US. In 2012, the total US annual healthcare expenditure reached $2.8 trillion. As a percentage of GDP, it is the highest among all the nations worldwide. In this study we investigate the variability of patient healthcare expenses year-on-year, depending on the different medical conditions patients get diagnosed with, the prescription drugs they consume and the demographic variables. We work with anonymized but publicly available Medicare data, which has more than 114,000 beneficiaries and more than 12,400 features. We address the problem of accurately predicting which beneficiaries' inpatient claim amounts increased between 2008 and 2009, using an ensemble of six different classification algorithms. We achieved a sensitivity (recall) of 80%, an overall accuracy of 77.56% and a precision of 76.46% on the test data set. We demonstrate its benefits to the healthcare stakeholders: the insurance providers for projecting cost and revenue more accurately, the high-risk patients for choosing the right insurance plan, and the healthcare providers for deciding which patients need additional monitoring. Our research shows that kidney conditions, COPD (chronic obstructive pulmonary disease), hypertension, stroke/transient ischemic attack, cancer and osteoporosis are among the most influential conditions behind expenditure increase.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data Mining*; I.2.6 [**ARTIFICIAL INTELLIGENCE**]: Learning—*Concept learning*

General Terms

Algorithms, Experimentation, Management

Keywords

Healthcare, expenditure, binary classification

HI-KDD 2014 New York, New York USA

1. INTRODUCTION

Healthcare expenditure in the US is growing concern. The total annual healthcare expenditure in 2012 was $2.8 trillion, the prescription drug spending accounting for $260.8 billion [1]. As the percentage of GDP (17.9), it is the highest among all nations. The healthcare cost per capita in 2010 was $8,233, which was one-fifth of the personal income per capita ($42,693). However, the huge expenditure does not necessarily buy Americans a better healthcare: in 2010, the number of practicing physicians per 1,000 people was only 2.4; whereas the corresponding number for the OECD countries was 3.1; there were 2.6 hospital beds per 1,000 people in 2009, whereas the OECD average was 3.4. Moreover, a major cause of the huge expense is unnecessary and inefficient measures, which include, but are not limited to: redundant medical tests [2], high cost of patented prescription drugs, infections caught from hospitals and frequent readmissions to hospitals. In 2011, the unnecessary expenditures added up to $476 billion (18%) to $992 billion (37%) of total (2.6 trillion) [3].

We attempt to address the problem of rise in healthcare expenditure by asking the following question: what factors increase an individual's expenses on healthcare? Is it because people develop certain chronic conditions? Is it because of age? Is it due to side-effects of drugs they have been prescribed? More specifically, given an individual's demographic and medical information, how well can we predict whether or not the individual's healthcare expense will rise next year? These questions can be of interest to the patients/beneficiaries, the insurance providers and the healthcare service providers in ways we explain below:

- **Beneficiaries:** If the beneficiaries know in advance that they are under high risk of an increase in expenditure for the next year, they can choose the insurance plans with higher deductible with more confidence. That way, although the annual out-of-pocket expense until the deductible is met would be higher, the beneficiary can benefit from a low monthly premium, and has to pay only 20-30% co-insurance out-of-pocket once the deductible is met.

- **Insurance providers:** The insurance providers can project the cost and revenue for the next year more accurately, given the data about the beneficiaries registered with them.

- **Healthcare service providers:** Hospital admissions and stays are intrinsically expensive. If doctors and hospitals know which patients are the high-risk ones, they can avoid readmission by taking preventive measures, e.g., arranging more frequent check-ups as outpatients, making the patients wear inexpensive sensors when they are discharged to monitor their conditions, etc.

Dataset: We worked on the dataset "CMS 2008-2010 Data Entrepreneurs′ Synthetic Public Use File (DE-SynPUF)" available at [4]. This is an anonymized dataset obtained from Medicare [5], and has data about 6.87 million beneficiaries, their inpatient claims, outpatient claims and prescription drug events. Since inpatient claims often form a significant fraction of an individual's annual healthcare expenses, we analyzed the expenditure on the inpatient claims. We use the other data within the dataset to derive features based on the prescription drugs taken by the beneficiaries, the medical conditions they got diagnosed with as inpatients or even as outpatients, and the chronic conditions they had. One problem of healthcare datasets, in general, is that different aspects of a person's health are often siloed in different databases. This dataset is free from that problem. The dataset is anonymized to protect the privacy of the beneficiaries, but the various parts of the dataset are linked through beneficiary IDs that were created by applying a hash function on the original beneficiary IDs. We present more details of the dataset in Section 3.

Contributions: Our contribution can be summarized as follows:

- We used a large, publicly available dataset with more than 114,000 beneficiaries and their health history for a span of three years (2008-2010) to investigate which ones, among more than 12,400 features, led to increase in expenditure as inpatient between 2008 and 2009. We formulated it as a binary classification problem, and brought down the number of features from 12,400 to 44 through feature selection.
- We experimented with various classification algorithms and finally chose six (gradient boosting machine [16], conditional inference tree [17], neural networks [19], SVM [14], logistic regression and naive Bayes) that performed best on a held-out test data set, and finally applied stacked generalization [20] as the ensemble technique. This achieved a sensitivity (recall) of 80%, an overall accuracy of 77.56% and a precision of 76.46% on the test data set. This way, we showed that it is indeed possible to develop learning models which will predict with great degree of accuracy about whether an individual is going to incur higher or lower healthcare expenditure based on normally collected information.
- Lastly, we identified major factors which are crucial in determining whether an individual is going to incur higher healthcare expenditure going forward.

2. RELATED WORK

Awareness about the US healthcare expenditure, as well as initiatives to contain it, are on rise. Researchers are trying to point out the causes of the continuous increase in healthcare expenditure. McBride [18] pointed out several factors, both outside and inside the healthcare sector, behind it: 1) General inflation, 2) Population growth, 3) Medical inflation and 4) Volume intensity. Factors inside the healthcare sector, like changes in healthcare need of the population because of an aging demographic, insurance-induced demand and increases in producer prices contributed to 4.4% of the 10% average annual change between 1960 and 2005. The remaining increase is due to equipment, procedures and services becoming more expensive with the advancement of technology. The study also compared expenditure in rural areas with that in urban areas, showing that the health spending increase per family for office-based visits was $124 for rural areas but $95 for urban areas: a reason might be people from the rural areas often have to travel to their nearest cities to avail many medical services.

An independent report by Forbes [6] agrees with [18] in that technology is one of the most significant factors to have driven the healthcare cost up over the years. Installing and implementing electronic health records, e.g., take $25,000 per doctor for a system and monthly subscription fees. Administrative expense is another significant factor, and so are lifestyle and chronic conditions. People with three or more chronic conditions form the most expensive 1% of patients who account for 20% of total healthcare expenditure in the US. Many of the chronic conditions are caused by lifestyle: tobacco consumption, inadequate physical exercises, poor diet and excessive alcohol consumption.

3. DETAILS ON DATA

Each of the components of the DE-SynPUF dataset (beneficiaries, inpatient claims, outpatient claims and prescription drug events) is available in 20 different partitions, and we worked with the first partition (about 5% of the data). The volumes of the different parts of the subset we worked on are in Table 1.

Table 1: Volume of DE-SynPUF subset

Entities	Subset explored
Beneficiaries	116k
Inpatient claims	66.7k
Outpatient claims	790.8k
Prescription drug events	5.5 million

We now present some details of each of the components of the dataset (a mode detailed data dictionary is availble at [7]):

1. **Beneficiaries:** We worked with a subset of 114,538 beneficiaries who were registered in both 2008 and 2009 with Medicare. The beneficiary subset provides the gender, date of birth, race, whether the beneficiary had end stage renal disease and whether the beneficiary had any of the following chronic conditions: Alzheimer or related disorders or senility, heart failure, kidney disease, cancer, chronic obstructive pulmonary disease (COPD), depression, diabetes, ischemic heart disease, osteoporosis, rheumatoid arthritis and osteoarthritis, stroke/transient ischemic attack. The beneficiary summary is stored at a per-year level. That helped us to derive features like whether a beneficiary developed

a chronic condition in 2009 which she did not have in 2008, and as we will show later, those derived features turned out to be pretty strong factors behind cost increase. The beneficiary dataset also had information about the total amount that Medicare reimbursed in a year for inpatient and outpatient claims by a beneficiary.

2. **Inpatient claims:** The inpatient claims dataset had, for each claim, the admission and discharge dates for hospitalization episodes, and a list of three types of codes:

   (a) **Diagnosis codes:** These are ICD9 [8] codes for beneficiary's principal or other diagnosis. They capture the physician's opinion of the patient's specific illnesses, signs, symptoms, and complaints. For example, 414.00 is the ICD9 code for Coronary Atherosclerosis.

   (b) **Procedure codes:** These are ICD9 codes for specific health interventions made by medical professionals, e.g., 4513 stands for "Other endoscopy of small intestine".

   (c) **HCPCS codes:** These are CPT codes [9] for tasks and services a medical practitioner may provide to a Medicare patient including medical, surgical and diagnostic services, e.g., 90658 stands for "flu shot".

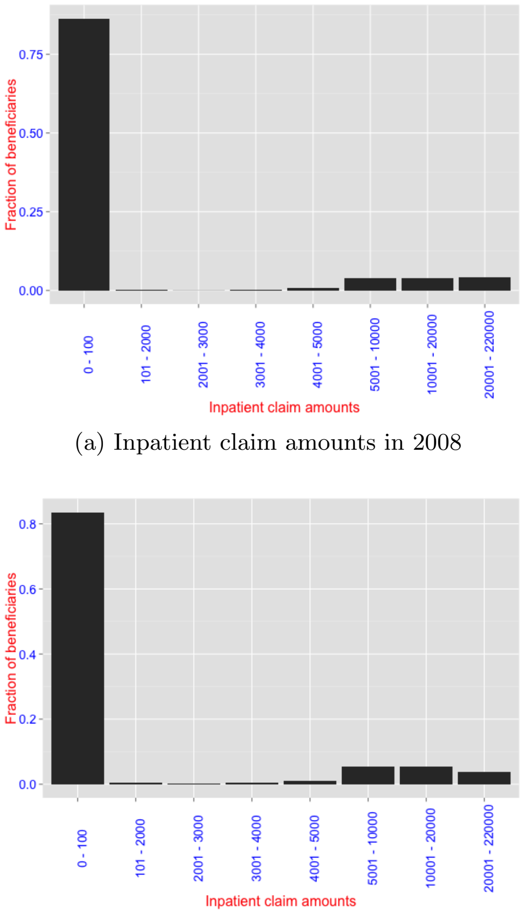
3. **Outpatient claims:** The outpatient claims dataset had, for each claim, a list of diagnosis codes, procedure codes and HCPCS codes, like inpatient claims.

4. **Prescription Drug Events:** In this dataset, each record had a product service ID, which identifies the dispensed drug using a National Drug Code (NDC); the number of units, grams or milliliters dispensed and the number of days' supply of medication.

Before we started working on the classification problem, we did some exploratory analysis of the DE-SynPUF dataset, and we present some of its results now: we had a total of 114,538 beneficiaries who were registered in both 2008 and 2009 with Medicare. 55% of them were female, 45% male. The median age at the start of 2009 was 72. The median number of inpatient claims, among people who did get hospitalized, was one, for both 2008 and 2009. A vast majority of inpatient claim amounts were 0, implying most of the registered beneficiaries never got hospitalized. However, some non-zero values of inpatient claim amounts were very large. In Figures 1a and 1b, we show the distribution of inpatient claim amounts for 2008 and 2009, respectively. We see that although 75-80% of the patients had an inpatient claim amounts between $0 and $100, the fraction of patients with inpatient claim amounts $5,000 and above is not very small. The mean amounts of inpatient claims in these two years are $2,583 and $2,526 respectively. Overall, we see an increase in inpatient claim amounts for 16,248 (14.2%) beneficiaries, while for the remaining 98,290 (85.8%) beneficiaries, it remained same or did not increase.

4. TOWARDS THE MODEL

As the discussion in Section 3 shows, the DE-SynPUF data presented to us a wealth of information, so deciding which features to use took extensive experimentation. The features can be divided into the following five logical groups:



(a) Inpatient claim amounts in 2008

(b) Inpatient claim amounts in 2009

Figure 1: Inpatient claim amounts, showing the overall pattern did not change much between 2008 and 2009

1. **Demographic:** Basic demographic variables like age at the beginning of 2009 and gender.

2. **Chronic conditions:** We derived a set of features based on the history of chronic information: we say a beneficiary *developed* a chronic condition if the beneficiary reportedly did not have that condition in 2008 but had it in 2009. This gave us 11 features.

3. **Diagnosed conditions:** We added the conditions people were diagnosed with in 2008 as inpatients as well as as outpatients as features. This gave us a set of 10,635 distinct conditions.

4. **Drugs taken:** We took the substances in the prescription drugs people took in 2008. This was obtained from an auxiliary data on NDC codes. We used the substance name because the same substance can have different NDC codes, depending on the labeler; e.g., 0615-7593 and 10816-102 are both valid NDC codes for the substance Minoxidil used for hair loss. This gave us 1,806 possible substances.

5. **Financial:** The inpatient claim amount in 2008.

4.1 Feature Selection

The groups described above gave us a total of $2 + 11 + 10635 + 1806 + 1 = 12,455$ different features for $114,538$ beneficiaries, resulting in a matrix with 1.4 billion cells. This matrix was highly sparse. To experiment with different classification algorithms, we computed the information gain [13] of the features arising out of diagnosed conditions and drugs taken (since these two groups contributed the maximum number of features) and took the top 30 features. This brought down the number of features used in the models to $2+11+30+1 = 44$. We show the information gain for the top 20 features in Figure 2. The names of the conditions corresponding to the ICD9 codes are in Table 2. We noticed that although the feature selection was done from the set of diagnosed conditions as well as substances in prescribed drugs, the top 30 features all come from the set of diagnosed conditions. Also, we see in Table 2 that two (401.9 and 401.1) of the top six features are two kinds of hypertension, and two (V58.69 and V58.61) of the top five are conditions developed due to long-term (current) use of other substances, three (280.9, 285.21, 285.9) are different types of anemia: this suggests that there are perhaps groups of conditions that played crucial roles in increasing the expenditure for beneficiaries.

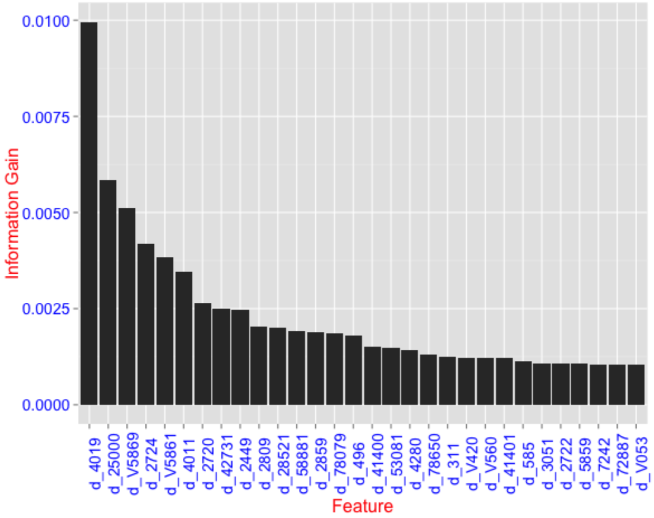


Figure 2: Information gain for top 20 features. The "d_" prefix in feature name indicates these are diagnosed conditions. We have dropped the '.' in the ICD9 codes in the plots to avoid clutter.

Table 2: Top 20 conditions in terms of information gain

ICD9 code	Name
401.9	Unspecified essential hypertension
250.00	Diabetes mellitus without mention of complication
V58.69	Long-term (current) use of other medications
272.4	Hyperlipidemia NEC/NOS
V58.61	Long-term (current) use of anticoagulants
401.1	Benign essential hypertension
272.0	Pure hypercholesterolemia
427.31	Atrial fibrillation
244.9	Unspecified acquired hypothyroidism
280.9	Iron deficiency anemia, unspecified
285.21	Anemia in chronic kidney disease
588.81	Secondary hyperparathyroidism (of renal origin)
285.9	Anemia, unspecified
780.79	Other malaise and fatigue
496	Chronic airway obstruction
414.00	Coronary atherosclerosis of unspecified type
530.81	Esophageal reflux
428.0	Congestive heart failure, unspecified
786.50	Chest pain, unspecified
311	Depressive disorder, not elsewhere classified

4.2 Sampling

As we mentioned in Section 3, the original dataset had 14% positive examples. We started our experiments with an L_1-regularized (LASSO) logistic regression algorithm [10] on an efficiently stored sparse matrix [11]. We varied the regularization parameter, λ, and observed how the overall training error, cross validation error, test error, and false negative rate and false positive rate varied. Since LASSO does feature selection, the number of features selected decreased as λ increased. When the cross-validation error hit minimum (0.1669065), the number of features selected by LASSO was only 44 (out of more than 12,400 features). However, the false negative rate was unacceptably high because of the class imbalance problem - most positive examples were getting labeled as negative by the algorithm since the training set was dominated by the negatives.

In order to deal with this, we took a uniform random sample of 5,000 beneficiaries (without replacement) from the 114,538 beneficiaries, and collected the 44 selected features (discussed in Subsection 4.1) for these 5,000 beneficiaries. Next, we split these 5,000 points into a training set of 4,000 and a test set of 1,000 points. The training set of 4,000 points had positives and negatives in almost the same ratio as the original data, and our goal was to ensure that we get 2,000 points from each class. So, we used Algorithm 1 to create a balanced sample.

Algorithm 1 is based on a simple idea: we split the input dataset, $D = \{(\mathbf{x}, y)\}_1^l$ into two disjoint, exhaustive subsets: D_N is the subset with all negative labels (beneficiaries whose cost did not increase), and D_P is the subset with all positive labels (beneficiaries whose cost increased). D_P occupied much less than half of the dataset D, and D_N occupied much more than half. We wanted to undersample D_N, and oversample D_P, so that the samples taken from either class has size $s = |D|/2$, so that they add up to $|D|$. We took a uniform random sample S_N of size s from D_N without replacement. To create a sample S_P of size $s = 2500$ from D_P of size 16248, we repeated each point from D_P 16248 div 2500 = 6 times, and filled up the remaining $16248 - 2500{\cdot}6 = 1248$ points by randomly sampling the points from D_P without replacement.

Theorem 4.1. *The input and output datasets of Algorithm 1 have the same size, i.e.,* $|D| = |D_B|$.

Proof. Since $D = D_N \bigcup D_P$ and $D_N \bigcap D_P = \phi$, $|D| = |D_N| + |D_P|$. By line 4 and line 1 of Algorithm 1, $|S_N| = s = |D|/2$. Also, by lines 5 and 6, $s = |D_P|q + r$. By line 7, $|S_{P_1}| = |D_P|q$. By line 8, $|S_{P_2}| = r$. By line 9, $|S_P| = |S_{P_1}| + |S_{P_2}| = |D_P|q + r = s$. By line 10, $|D_B| = |S_P| + |S_N| = s + s = |D|$. □

```
Algorithm 1: Sampling-for-Balance($D$)
  Input: $D = \{(\mathbf{x}, y)\}_1^l$: $\mathbf{x}$ is the vector of the predictors
         and $y$ is the response variable.
  Output: $D_B$, a balanced sample with equal number of
          positive and negative examples. The
          algorithm ensures $|D| = |D_B|$
  $s \leftarrow \frac{|D|}{2}$
  $D_N \leftarrow \{(\mathbf{x}, y) \in D | y = -1\}$
  $D_P \leftarrow \{(\mathbf{x}, y) \in D | y = 1\}$
  $S_N \leftarrow$a uniform random sample of size $s$, taken without
    replacement, from $D_N$
  $q \leftarrow s$ div $|D_P|$
  $r \leftarrow s$ mod $|D_P|$
  $S_{P_1} \leftarrow$ a multiset with each element of $D_P$ repeated $q$
    times
  $S_{P_2} \leftarrow$ a uniform random sample of size $r$, taken
    without replacement, from $D_P$
  $S_P \leftarrow$ A multiset with all elements of $S_{P_1}$ and $S_{P_2}$
  $D_B \leftarrow$ A multiset with all elements of $S_P$ and $S_N$
```

5. CLASSIFICATION TASK

Having done the feature selection and the sample balancing as described in Section 4, we applied various classification algorithms. Each classification algorithm was trained on a sample of 4,000 points sampled from the balanced set (of size 5,000) returned by Algorithm 1, and tested on the remaining set of 1,000 points. In order to ensure fair comparison among the algorithms, we ensured that the split of training and test data is always the same. We present here six of the classification algorithms that performed best on the independent test set. We finally created an ensemble of the six algorithms through stacked generalization [20].

5.1 Gradient Boosting Machine

The Gradient Boosting Machine [16] needs two main components for any classification or regression problem: a) a base learner, and b) a differentiable loss function that it aims to optimize eventually. The final model delivered is an *additive* one, given by

$$F(\mathbf{x}; \{\beta_m, \mathbf{a}_m\}_1^M) = \sum_{m=1}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m) \tag{1}$$

where $\mathbf{a}_m$ is the parameter of the hypothesis created in the m^{th} iteration, $h(\mathbf{x}; \mathbf{a}_m)$ is the hypothesis created in the m^{th} iteration, and β_m's are the coefficients of the linear combination.

We used $M = 5,000$ decision trees and included upto 2-way interactions in the trees. We actually performed grid search and cross-validation over the number of iterations and the interaction depth, with the number of iterations starting from 1000, and going up to 10000 in steps of 1000. The interaction depth was varied between 1 and 2. The results are shown in Figure 3. We see that the CV error reduced monotonically as the number of iterations as well as the interaction depth increased. However, when we applied the models with $M = 7000$ and $M = 10000$ back on the (whole of) training and the test data, the results were as in Table 3, which shows that although the training error and training FNR reduced as M was increased from 5000 to 10000, the performance on the test set remained practically same between $M = 5000$ and $M = 7000$, and in fact slightly degraded when we moved to $M = 10000$, implying that GBM probably started overfitting beyond $M = 5000$. Also, it takes longer to train the model as the number of iterations increases, so we chose $M = 5000$ for the final model.

The relative influences of the covariates in GBM (scaled to add up to 100) are shown in Figure 4. Following the notation in [16], the relative influence of the j^{th} covariate, x_j, is given by

$$\hat{I_j}^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{I_j}^2(T_m) \tag{2}$$

where $\hat{I}_j(T_m)$ is the relative influence of x_j in the tree generated in the m^{th} iteration, T_m, and is given by

$$\hat{I_j}^2(T) = \sum_{t=1}^{J-1} \hat{i_t}^2 \mathbf{1}(v_t = j) \tag{3}$$

where the summation is over the $J - 1$ non-terminal nodes of the tree with J terminal ones, v_t is the splitting variable associated with non-terminal node t, and $\hat{i_t}^2$ is the improvement in error as a result of the split at node t. We see that the development of chronic conditions (variables whose names start with "dev_" in Figure 4) like kidney problems, COPD, stroke/transient ischemic attack, cancer, osteoporosis and diabetes are among the most influential factors behind expenditure increase. Among the diagnosed conditions (variables whose names start with "d_"), 401.9 (Unspecified essential hypertension) and 250.00 (Diabetes mellitus without mention of complication) are the most influential ones. Note that the latter two ranked the highest in terms of information gain too, as shown in Table 2. Also, figure 4 shows that demographic variables (bene_sex_ident_cd and age_year2), although used in the model, did not have much of an influence.

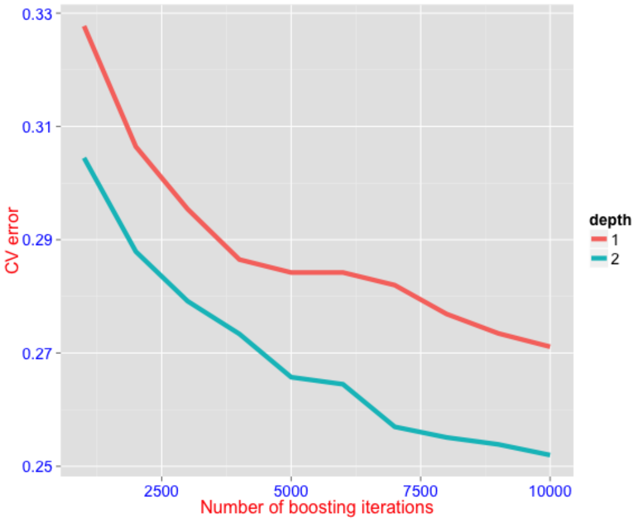


Figure 3: CV error with grid search for hyperparameters of GBM

5.2 Conditional Inference Trees

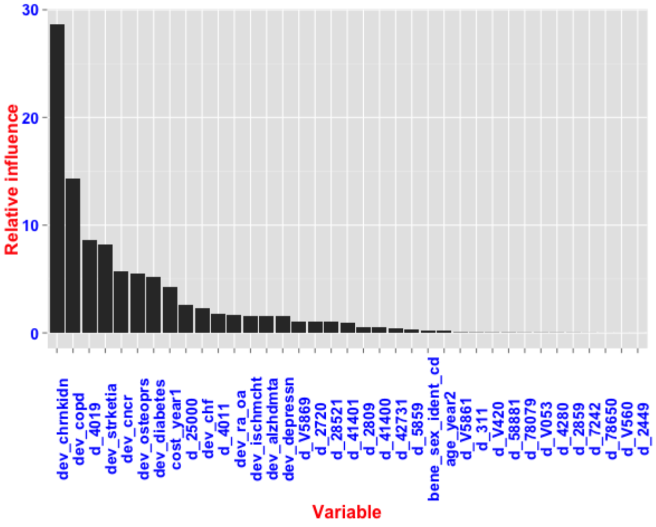


Figure 4: Relative influences of covariates in GBM

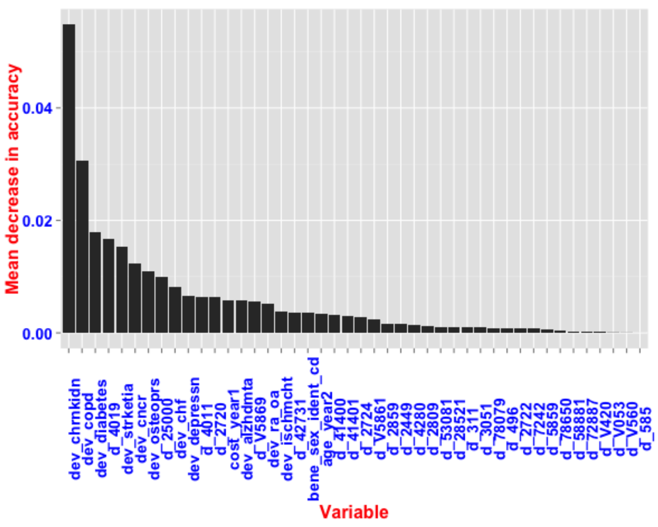


Figure 5: Importance of covariates in conditional inference forest

Conditional Inference Trees [17] combine the idea of recursive partitioning with statistical significance tests to generate decision trees, where the split variables at the non-terminal nodes are chosen by measuring the degree of association between the predictors and the response variables by the test-statistic. The chi-square test of independence [12] is used for measuring the degree of association. The conditional inference trees can be extended to create a forest of such trees, and a measure of variable importance (mean decrease in accuracy) can be obtained from such a forest. When we applied that to our data, we obtained the plot in Figure 5. Although the scales on the Y-axes between this and Figure 4 are different, note the similarity between the ordering of the covariates on the X-axis: we present this as a verification of what one method considers important is also confirmed important by another method.

The performance of the algorithm on the training and the test datasets are listed in Table 4. We see that the performance of this algorithm on the training and the test datasets are pretty comparable, implying this algorithm did not overfit.

5.3 Neural Network

We used a neural network with a single layer of hidden units. Our choice of resticting the number of hidden layers to one is influenced by the well-known result by Cybenko [15], who showed that "arbitray decision regions can be arbitrarily well approximated by continuous feedforward neural networks with only a single internal, hidden layer and any continuous sigmoid nonlinearity". We found the optimal number of units in the hidden layer by grid search, over the range of even values between 2 and 18. The results are summarized in Table 5. Although the values in this table suggest that the result with 18 units is slightly better than that with 12 units, when we applied the model with 18 units back onto the test dataset, its performance (especially the test FNR) was worse than that of the model with 12 units (test error = 0.383, test FNR = 0.352941, test FPR = 0.3877 versus the values in Table 4. Also, it takes longer to train the network as the number of units go up, so we decided to go ahead with 12 units.

5.4 Other algorithms

We also applied SVM [14], logistic regression and naive Bayes, and the performances are listed in Table 4.

5.5 Stacking

We used stacked generalization [20] for creating an ensemble of the classification algorithms we discussed so far. Stacking is a meta-algorithm, where, given a set $\{(\mathbf{x}, y)\}_1^l$ of sample observations, a new set $\{(\mathbf{x}', y)\}_1^l$ is created, where $\mathbf{x}'$ is the set of class labels assigned to $\mathbf{x}$ by a set of classifiers, which are being included in the ensemble. Since we chose the six algorithms mentioned in Sections 5.1 to 5.4 to be included in the ensemble, the attributes of $\mathbf{x}'$ in our example were: `svm_class`, `gbm_class`, `citree_class`, `lr_class`, `nb_class`, `nn_class`, for labels predicted by SVM, gradient boosting machine, conditional inference tree, logistic regression, naive Bayes and neural network respectively. We took the same sample of 5,000 observations as mentioned in Section 4.2, applied Algorithm 1 to create a class-balanced sample out of it, and derived a predicted label for each of the 5,000 points through cross-validation: i.e., we split the 5,000 points into five folds, and used the data in each fold once as a validation set and 4 times as the training set, and derived the labels for the points in a fold when it was used as the validation set.

Once this new dataset $\{(\mathbf{x}', y)\}_1^l$ is created, we split it randomly into two halves (so each had 2,500 points): we used one as the training set, and the other as the test set. We trained a decision tree, after performing a grid search on the the minimum number of observations that should be present in a node to be considered for a split and the maximum depth of any node of the final tree (with the depth of the root node treated as 0). The final contingency table out of these 2,500 test points was as follows:

	PredictedClass		
ActualClass	negative	positive	Row total
negative	932	310	1,242
positive	251	1,007	1,258
Column total	1,183	1,317	2,500

so the recall is $\frac{1007}{1258} = 80.05\%$, the overall accuracy is $\frac{932+1007}{2500} = 77.56\%$, the precision is $\frac{1007}{1317} = 76.46\%$.

6. CONCLUSION

So far, we have obtained notions of relative importance of the different features behind expenditure increase from the plots in Figures 4 and 5. These were variable importance as perceived by the different classification algorithms. In Figure 6, we go back to the original dataset of 114,538 beneficiaries and plot the conditional probabilities of cost increase in presence and absence of different conditions, e.g., the 3rd bar from the left indicates that people who devloped the chronic condition stroke/transient ischemic attack (dev_strketia) had a 39% conditional probability of a cost increase, whereas people who did not develop this condition had only a 13% (1/3rd) conditional probability of a cost increase. The difference between the heights of the red bars and their adjacent green bars are one way to measure how important these covariates are. We see that the conditional probability of cost increase, when these conditions were not present, hovered between 10% and 15%, whereas the conditional probability of cost increase, when these conditions were present, varied from 41% to 20% (except for gender which does not make a significant difference).

What figures 4, 5 and 6 all consistently point out is that the following conditions are really instrumental behind expenditure increase, and hence people with these conditions need careful monitoring to avoid costly hospitalization episodes.

1. **dev_strketia:** Whether the beneficiary had one or more strokes/transient ischemic attacks
2. **dev_chrnkidn:** Whether the beneficiary developed chronic kidney conditions
3. **dev_copd:** Whether the beneficiary developed COPD (chronic obstructive pulmonary disease)
4. **dev_cncr:** Whether the beneficiary developed cancer
5. **d_4019:** Whether the beneficiary got diagnosed with unspecified essential hypertension
6. **d_25000:** Whether the beneficiary got diagnosed with diabetes mellitus
7. **dev_osteoprs:** Whether the beneficiary developed osteoporosis

7. FUTURE WORK

As continuation of this analysis, we plan to take a deeper dive into the data: possibly with some domain expertise. Rather than splitting the beneficiaries into only two groups, we can split them into multiple groups, depending on whether the cost increase was high, medium or low; with some domain expertise, we can identify whether the conditions that people are diagnosed with can be somehow grouped, resulting in a reduction in the number of features.

8. REFERENCES

[1] http://healthaffairs.org/blog/2013/09/18/us-health-spending-growth-projected-to-average-5-8\-percent-annually-through-2022/.

[2] https://www.seattlechildrens.org/about/stories/decreasing-unnecessary-lab-tests-saving-money/.

[3] http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=82/.

[4] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/DE_Syn_PUF.html.

[5] https://www.medicare.gov/.

[6] http://www.forbes.com/sites/realspin/2013/04/03/whos-to-blame-for-our-rising-healthcare-costs/.

[7] http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/SynPUFs/Downloads/SynPUF_Codebook.pdf.

[8] http://www.riversidemd.net/tools/code-search.cfm.

[9] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page.

[10] http://cran.r-project.org/web/packages/glmnet/index.html.

[11] http://cran.r-project.org/web/packages/Matrix/index.html.

[12] http://mathworld.wolfram.com/Chi-SquaredTest.html.

[13] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[14] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[15] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematical Control Signals Systems*, 5(4):455, 1992.

[16] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 2001.

[17] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3), 2006.

[18] T. D. McBride. Why are healthcare expenditures increasing and is there a rural differential? *Rural Policy Brief*, 10(7), 2005.

[19] B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

[20] D. H. Wolpert. Stacked generalization. *Neural Networks,*, 5(2), 1992.

Table 3: Results of grid search for GBM

Iterations(M)	Interaction depth	Trg error	Trg FNR	Trg FPR	CV error	Test error	Test FNR	Test FPR
5000	2	0.26375	0.303	0.2245	0.2657311	0.241	0.2867647	0.233796
7000	2	0.249	0.273	0.225	0.256968	0.246	0.2867647	0.239583
10000	2	0.24025	0.25	0.2305	0.251982	0.252	0.294117	0.24537

Table 4: Summary of performance of all classifiers

Classifier	Hyper/Control parameters	Trg error	Trg FNR	Trg FPR	Min CV error	Test error	Test FNR	Test FPR
Gradient Boosting Machine	bag.fraction = 0.5, 5000 iterations, logistic loss	0.249	0.273	0.225	0.2569680	0.246	0.2867647	0.239583
Conditional Inference Tree	Quadratic form test statistic, Bonferroni correction for p-value	0.25775	0.213	0.3025		0.299	0.25	0.30671
Logistic Regression		0.255	0.281	0.229		0.249	0.2867647	0.2431
SVM	Linear kernel, $C = 1$	0.254	0.2745	0.2335	0.2582514	0.245	0.2867647	0.238426
Neural network	Single hidden layer with 12 units, decay $= 5 \cdot 10^4$, max iterations = 200	0.21325	0.086	0.3405	0.2152645	0.353	0.2721	0.365741
Naive Bayes		0.26575	0.2465	0.285		0.279	0.2720588	0.28

Table 5: Results of grid search for neural network

Number of units in hidden layer	Cross-validation error	Standard deviation in CV error
2	0.2924428	0.07616226
4	0.2310153	0.01854330
6	0.2280071	0.02138528
8	0.2264990	0.01746430
10	0.2209951	0.02012913
12	0.2152645	0.02472898
14	0.2140064	0.01972838
16	0.2207608	0.02240481
18	0.2137507	0.01569324

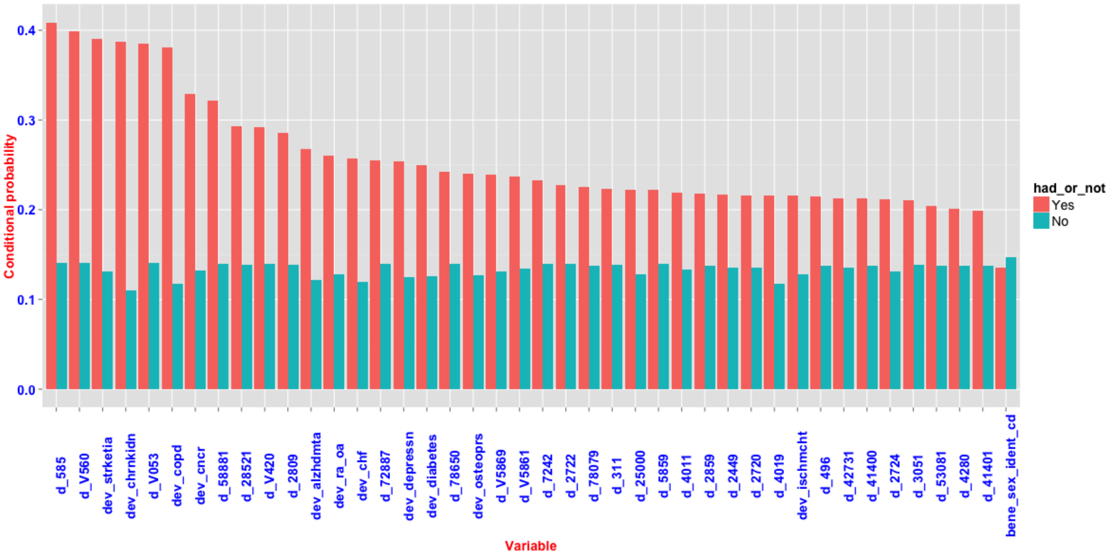


Figure 6: Conditional probabilities of cost increase when conditions are present or absent. Red bars indicate the conditional proabilities when conditions are present, and green bars indicate the conditional proabilities when conditions are absent. For gender (bene_sex_ident_cd), red indicates male.