Employer Health Insurance Premium Prediction

Elliott Lui

1 Introduction

The US spends 15.2% of its GDP on health care, more than any other country, and the cost of health insurance is rising faster than wages or inflation. Per year, employers spend $500 billion on health premiums for their employees. One important question these employers must always consider is if the coverage they are getting is worth what they are paying. The health insurance providers have their proprietary actuarial methods and complex models to determine these premiums, but they are hidden from the public.

The purpose of this project is to explore the use of machine learning algorithms to predict the prices of annual health insurance premiums given the specifications of the contract and the company's demographics. That is, given a health insurance contract and information about a company's employees, can we accurately predict how much it will cost per year? Using SVM, multinomial Naive Bayes, and a decision tree classifier, we can try to predict the premium costs of a contract, as well as determine the optimal set of features using feature selection.

2 Data

2.1 Dataset

The data set used for this project encompasses over 5,500 anonymous companies representing over 3 million lives. It comes from the Kaiser Family Foundation, which conducts an annual survey to collect this data to create the annual Health Benefits Report. A vector of 1199 variables represents each company, or data point. To put into relevant context, some of these features include industry, size of firm, percentage of workforce age 26 or lower, copayment amount for prescription drugs, coinsurance rate for hospital visit, and percent of workforce earning $21,000 or less, which conceptually seem quite relevant. On the other hand, there are features such as "Does HDP use Copay, Coinsurance, or Both for Generic Drugs" and "Firm offers this PPO Plan Last Year" that seem irrelevant in insurance premium pricing.

2.2 Training Data

Because there are three different types of health insurance plans, and each type of plan has slightly different features, we come up with three sets of training data to predict three sets of prices. We are trying to predict the annual premium cost for a company for the categories of PPO, HMO, and HDP. Some companies provide more than one type of plan, but because there are slightly different features, we treated the company as separate data points in each plan. For example, if company A offered both PPO and HMO plans, we put the relevant feature vectors of A in both the PPO training data and HMO training data.

2.3 Feature Selection

To quantitatively determine which features are most relevant, we ran forward search first using the SVM classifier with a Gaussian to maximize 10-fold cross validation, and then with the other classification algorithms. Interestingly but perhaps not too surprising, the top features made sense conceptually, consisting of variables representing coinsurance amounts for surgeries, hospital visits, and drugs.

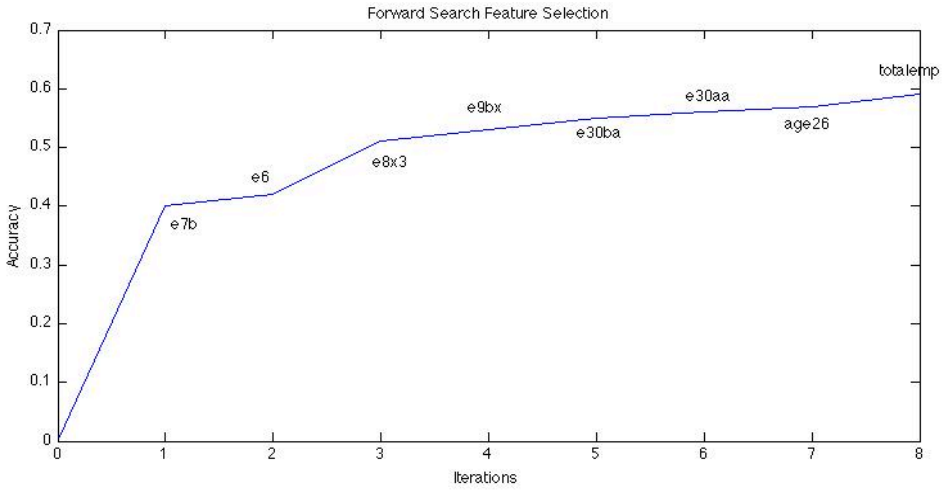


Figure 1. Variables added to feature set through forward search

Improvements in accuracy stopped after just 8 iterations for the SVM as seen in Figure 1. The forward search algorithm resulted in a set of just 8 features, down from the 1199 in the raw format as seen in Figure 2. Furthermore, after initially running with just SVM, running with multinomial Naïve Bayes, and decision tree classification yielded the same top 8 features with variances of at most three features.

Raw Data

Variable Information

Variable	Label	Measurement Level
id	Unique Identifier	
resp2011	Previous Response	Scale
response	Response Category All Firms	Nominal
respoffr	Response Category Offering Firms	Nominal
respnoff	Response Category Non-offering Firms	Nominal
respa6	Response Category A6	Nominal
panel2	Panel Group	Nominal

Feature Set

Variable	Label
e7b	Coinsurance Rate for an Outpatient Surgery Episode
e6	Copayment for Office Visit
e8x3	Coinsurance for Hospital Admission
e9bx	Maximum Out of Pocket
e30ba	Preferred Drug Copay
e30aa	Generic Drug Copay
age26	Percent Workface < 26
totalemp	Total Employee Count

Figure 2. Raw data to relevant features

3 Predicting Insurance Premiums

3.1 Bucketing Target Values

Since the premium prices are continuous, and we wanted to leverage the power of SVM and other classification algorithms, it was necessary to bucket these values. In order to make

meaningful predictions, it would make sense to limit the maximum error to what is acceptable in practical situations. Year to year changes in insurance premiums max out at about 30%, so we bucketed our data such that each of N buckets corresponds to a maximum price differential of 30%. For example, a bucket with a minimum value of $10,000 would have a maximum value of $13,000.

3.2 SVM

We trained a multi-class SVM to classify which of N buckets a given data point belonged in, with one SVM trained for each type of plan. When training the SVMs, we experimented using Gaussian and linear kernels, the results shown in figure 3. In the Gaussian kernel, $K(x,z) = e^{-(x-z)^2/(2\sigma^2)}$ whereas in the linear kernel $K(x,z) = x^T z$.
The parameter σ for the Gaussian kernel was tuned by using a grid-based search to maximize 10-fold cross validation. We made sure to not overestimate this parameter to avoid it performing like a linear kernel. Underestimating it on the other hand would make it too sensitive to noise in the training data.

We created the multi-class SVM by training a SVM for the upper and lower limit for each bucket. Each SVM predicts whether the price of a contract should be greater than or less than a bucket value; the final classification is then made by a majority vote by the SVMs. In addition to using different kernels, we also experimented with different bucket sizes of 15%, 30%, and 45% to see its effect on accuracy.

3.3 Multinomial Naïve Bayes

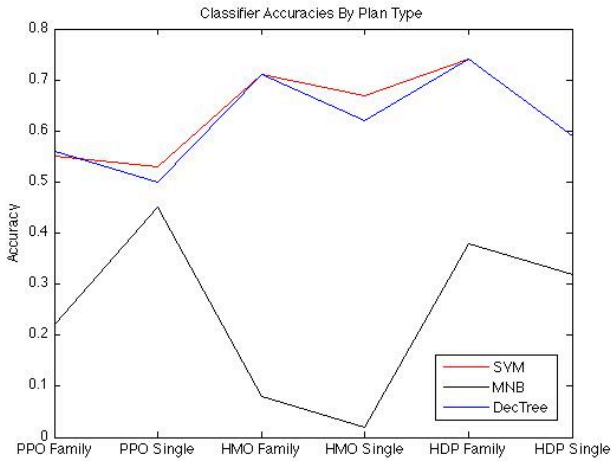
Multinomial Naïve Bayes models the distribution of feature values as a multinomial. The Naïve Bayes assumption is that each feature is generated independently of every other. Using the same methodology described above, we trained multiple binary multinomial Naïve Bayes classifiers. We wanted to keep consistent with our previous method of multi-class classification, and it performed much better than multiclass Naïve Bayes on its own.

3.4 Decision Tree

Though it wasn't explicitly taught in class, decision tree learning seemed like a good model for this data. One hypothesis for the low accuracy of the models could be the combination of continuous and categorical variables. Of the 8 features in our feature set, 5 are categorical – even copay amounts were discretized into ranges. Decision tree classifiers are able to handle categorical variables especially well, so it made sense to train this classifier.

4 Results

The results of our three models are summarized in figure 3. Each model was run on three datasets corresponding to PPO, HMO, and HDP plans. Within each



category we see the accuracy of predicting two values – the premium cost for family and the premium cost for an individual. The PPO training data contained 696 rows, the HMO 247, and HDP 213.

Algorithm	Overall Accuracy
SVM	.6317
Multinomial Naiive Bayes	.245
Decision Tree	.62

Figure 3. Overall Performance of Models

4.1 SVM Results

With the accepted 30% bucket size standard for this project, the SVM predicted premium prices with the highest accuracy among the three models at a rate of .632 (+/- .02). We see increasing bucket sizes improves classification accuracy; intuitively, it makes sense that it is easier to classify data into larger ranges.

Bucket Size	Accuracy	
	Linear Kernel	Gaussian Kernel
15%	.449	.508
30%	.593	.632
45%	.765	.738

Figure 4. SVM Accuracy

The two different kernels on the other hand didn't display any significant differences in performance. Though we made sure to tune our Gaussian parameter σ to not overestimate and behave linearly, it still didn't produce a substantial difference.

4.2 Multinomial Naïve Bayes Results

At an average accuracy of .245, multinomial Naïve Bayes exhibited the worst performance among our models. For all three categories (PPO, HMO, HDP), Naïve Bayes was the only algorithm to pick the company's industry as a feature in forward search. Conceptually it would make sense that a company in a labor intensive industry like construction would command higher premium prices than a company with young, highly skilled workers like a tech company. At first glance at the raw data, the poor performance could be attributed to a violation of the Naïve Bayes assumption; however, the features it selected were all independent from a conceptual standpoint.

4.3 Decision Tree Results

The decision tree classifier performed almost as well as the SVM overall with an accuracy rate of .62. Figure 5 shows a sub-tree of the decision tree that was ultimately constructed from our training data (the whole tree would be too large to display).

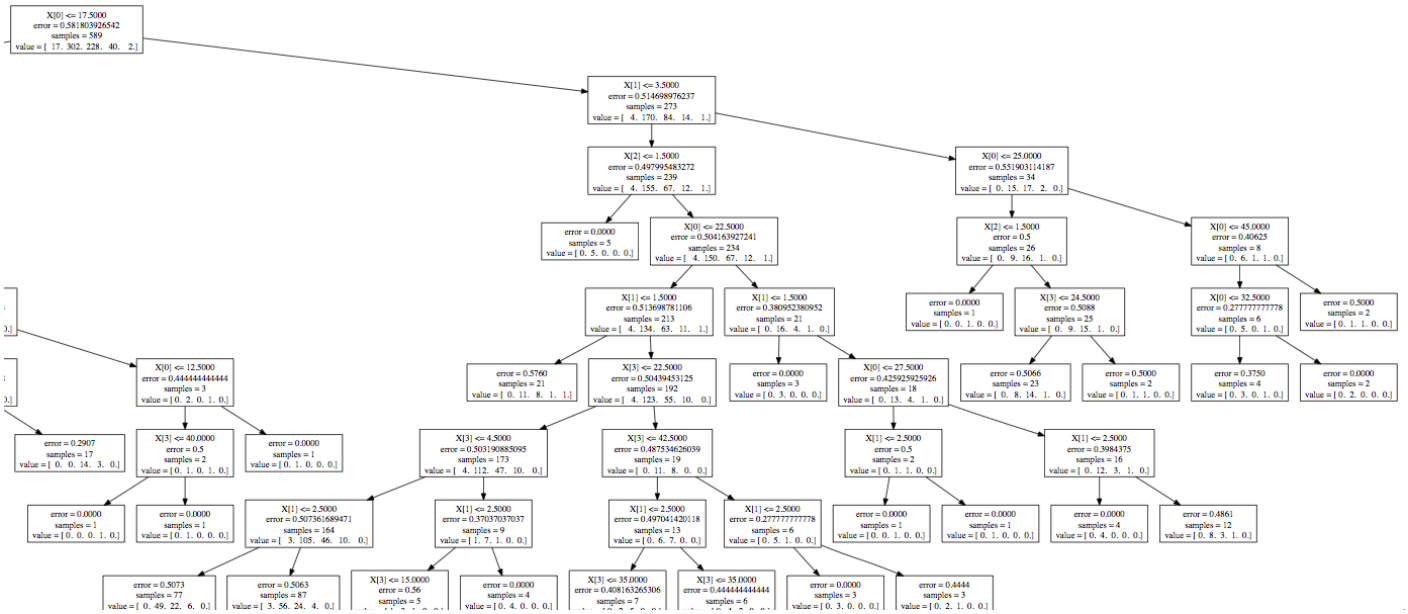


Figure 5. Sub-tree of the Decision Tree Classifier

The tree contained 130 nodes representing possible “decisions”, the number of samples with that “decision”, and its error. Because these decisions are binary <= operations, the model seemed to work on the categorical nature of our features.

5 Discussion

The results from our models didn’t predict insurance premium costs very well. While the accuracy rates weren’t abysmally low, there was much left to be desired. The low accuracy across all models suggests that the feature vectors don’t encompass all pertinent information. In fact, the data doesn’t account for the strength of the health plan network. For example, two identical data points can have differing physician networks – one could include a top-notch institution like Stanford hospital while the other includes only small clinics with fewer doctors. The former would obviously command a higher premium price, but such information was not included in the data set, presumably because it is extremely difficult to quantify. The forward search component of the analysis was quite successful in choosing the most important features. That is, without supervision, it was able to select the features such as copays for drugs and hospital visits that make sense in determining the price of health insurance premiums. Perhaps with key additional information like network strength, we could have predicted premium prices at a much higher accuracy.

References

[1] N Chapados, Y Bengio, P Vincent, J Gohsn, C Dugas, I Takeuchi, L Meng. Estimating Car Insurance Premia: a Case Study in High Dimensional Data Inference. *Advances in Neural Information Processing Systems* (2002)

[2] D Biggs, B Ville, E Suen. A Method of Choosing Multi-way Partitions for Classification and Decision Trees. *Journal of Applied Statistics,* 18(1):49-62

[3] G Cass. An Exploratory Technique for Investigating large quantities of categorical Data. *Applied Statistics,* 29(2):119-127