DEVELOPING A CLASSIFIER TO DETERMINE HEALTHCARE COMPANIES

By Owen Phillips and Max Rubinstein

INTRODUCTION

Researchers often use NAICS codes to categorize businesses into industries when analyzing business trends. This classification is issued by the Census and revised every five years, and it was last revised in 2012.

However, the NAICS classifications are not useful for all purposes. Often, companies classified in one category often have functions in another.[1] For example, take the company "ARx: Healthcare Business Office Partners." Despite having healthcare in the title of their company, NAICS classifies this company not as a healthcare company, but "Other Management Consulting Services."

Yet perhaps this is the exception and not the rule. To support the importance of this project, we examined the top 25 companies by revenue in the Nashville metro area in ten different NAICS categories. While we deliberately chose categories we thought would have large involvement in the healthcare sector, our results were surprising:

NAICS Codes	Percent Health Related	Descriptions
5112	32%	Software Publishers
5182	52%	Data Processing, Hosting, and Related Services
5191	23%	All Other Information Services
5241	60%	Insurance
5242	20%	Insurance
5412	40%	Accounting Services
5415	48%	Computer Systems
5416	36%	Consulting Services
5419	32%	Marketing Research
5614	56%	Business Support Services

Of these ten categories, approximately 40% of these companies conducted substantial work in the health sector. As a sample of the largest revenue generating companies, we can assume that these businesses also bring in have a disproportionate amount of employment in the area. Thus, industry analysis using the NAICS health classification alone would miss a substantial portion of businesses that do substantial work in the health sector.

Our project therefore seeks to create a classifier to outperform the NAICS classification system. We try to classify businesses in the healthcare industry in particular it involves multiple sectors of the economy, and therefore the misclassification rate is high. This project will hopefully enable researchers to generate better classification for industry analysis.

---

[1] Industries can have multiple NAICS codes associated with them; however, even then, this general observation still holds true.

Ultimately, we were able to generate several classifiers with specificity above 80%, but we had difficulty classifying healthcare companies with more than 70% sensitivity. A part of this is a function of being able to generate words that are exclusive of healthcare companies. Due to these challenges, the classification error rates – particularly of positively identifying healthcare companies – remain high among the methods that we explored. Nevertheless, we ultimate think there is promise in pursuing additional classification methods.

METHODS

We first downloaded industry data from the Nashville, TN area from Moody's analytics. We then hand-classified companies as whether or not they served the healthcare industry from looking at the website. Needless to say, we omitted any companies that did not have a website listed, or whose website was not functioning. Ultimately, we classified 715 unique websites.

We next wrote a web scraper in Python (Ad Hoc.py) that searched for nine strings that we thought off the top of our heads would positively identify healthcare websites. These words were 'Health', 'Dental', 'Optical', 'Medic', 'Treatment', 'Diagnosis', 'Physician', and 'Quality of Care'. The scraper then generated a flat-file with the number of times each word appeared on each webpage.

Next, we sought to develop computer generated lists of keywords to compare this 'ad hoc' list against. To do this, we used R to gather the html data for each homepage. We filtered out websites that didn't work (permission denied, website moved, etc.), which left us with 524 observations (to ensure comparability of the sample, we matched this with the results from our Python scraper). We then created a document-term matrix with all the unigrams, bigrams, and trigrams from the text with the counts associated with each website. This created 8942 n-grams. However, many of these terms appeared very few times, so we removed those terms that appeared in fewer than 0.5% of observations.

We then sought to perform variable selection for our lists of keywords. However, we needed to incorporate this as a part of the cross-validation process. Therefore, we split our data into five folds, and ran the following variable selection procedures on each fold.

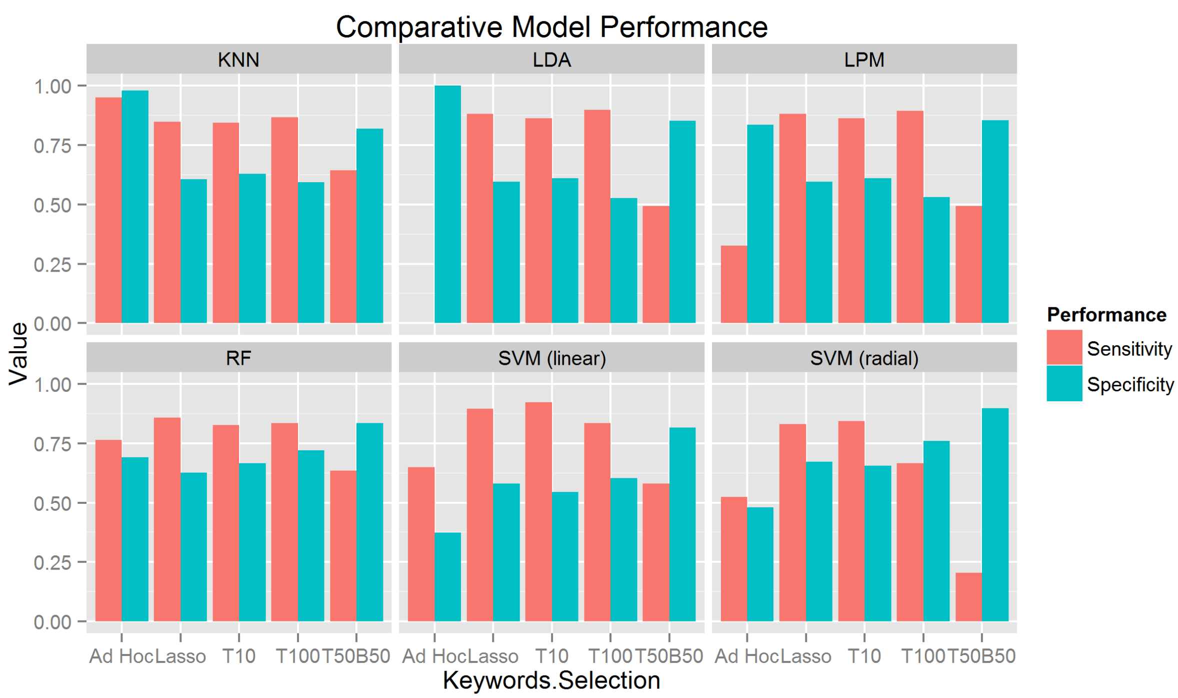
First, we divided the data into two variables: healthcare related websites and non-healthcare related websites, with each n-gram and the count as the observation and ran a chi-squared test on this data frame. Using the magnitude of the residuals, we determined the following lists of keywords[2]:

- **T100:** The top 100 n-grams most associated with healthcare websites
- **T10:** The top 10 n-grams most associated with healthcare websites
- **T50B50:** The top 50 n-grams most associated with healthcare websites, and the 50 n-grams least associated with healthcare websites.
- **Lasso:** We generated this final list by running a lasso on the T100 list, using cross-validation to tune, and using the resulting n-grams in this model as another keyword list.

[2] We also removed any keywords left on the list that were incomprehensible. This occurred at most twice in a list of 100, but occasionally we got a strange character string that was not filtered from the html file.

We then used these lists of keywords on the training data and used them to predict the outcomes of the test data for each fold (and note that each list of keywords was slightly different for each fold). This allowed us to obtain a test error that incorporated the variability of the variable selection process for each model. The lists of keywords can be found in Appendices B, C, and D. The methods we used were K nearest neighbors (KNN), linear discriminant analysis (LDA), linear probability model (LPM), random forests (RF), and support vector machines (SVM) with linear and radial kernels.

RESULTS



As these results indicate, the Lasso, T10, and T100 lists all yielded higher sensitivity than specificity. That is, they were better at determining true positives than true negatives. On the other hand, the models that used the T50B50 list – the only list which used keywords that were not associated with healthcare websites – all had greater specificity relative to sensitivity.

The performance of the ad hoc keywords provides an even more striking contrast to the computer generated keywords. Both the sensitivity and specificity of the ad hoc KNN model was over 95%! Meanwhile, the LDA predicted that all of the sites were healthcare related, and the performance in the rest of the models that used this list was quite lower than the others.

All of the raw numbers, in addition to some additional statistics, can be found in Appendix A.

DISCUSSION

The most striking aspect of these results is the incredibly high performance of KNN with the ad hoc keywords, and the relatively worse performance of all the other models that used this list of keywords.

There are some reasons we might expect higher performance in at least some of the models using the ad hoc keywords. For example, we used five-fold cross validation for the computer generated test errors in all models, but used leave-one-out-cross-validation (LOOCV) for the KNN, LPM, and LDA models using the ad hoc keywords (and out-of-bag error for Random Forests). LOOCV tends to be a less biased estimator of the true error rates; however, the variance is higher. K-fold CV will have higher bias, but lower variance. At least some of the difference in the results here may be a product of these different methods of cross-validation.

However, a more problematic potential source of the differences comes from a theoretical problem: in a sense, generating the list of ad hoc keywords was a misguided exercise in performing variable selection without cross-validation. When we developed our list of keywords, we essentially chose keywords with respect to known outcomes (whether a website would be healthcare related). We therefore may have been able to fool cross-validation down the road by skipping it for this initial procedure.

On the other hand, skipping cross-validation for variable selection should make our results better overall. Other than KNN, the other models using the Ad Hoc keywords performed somewhat worse. We have no great explanation of this finding in combination with the much higher performance of KNN; however, it's worth noting that several of the keywords we chose had very low variance (including quality of care and optical). Another possible explanation may be that these variables helped KNN perform better, but made the fitting more difficult for the other models.

A second striking feature is that adding in keywords negatively associated with healthcare websites significantly affected the performance of the models. In every case, the specificity was greater than the sensitivity using these keywords. Should we therefore have created more models using negative keywords?

This consideration leads to a second theoretical problem. It is simply very difficult to generate convincing keywords that would exclude a company from being healthcare related. While the use of negative keywords is associated here with higher specificity rates, the worry we have is that the selected keywords will simply be an artifact of the sample. The data here actually support this: looking at the 50 keywords least associated with health websites in each of the 5 folds, we found that on average 36.1% of these negative words differed between any two lists. By contrast, on average only 12.7% of the 50 keywords most associated with healthcare varied between any two sets. This high variability reflects that it is much harder to find words exclusive of the healthcare industry than generating with ones indicative of it.

However, it still may be possible to get a convincing set of negative keywords. Unfortunately, there then arises another worry. In this case we know that at least part of this sample is a highly non-random selection from Nashville – we included the top 25 largest companies in 10 different NAICS codes where we thought we were likely to find more healthcare companies. This on its own is problematic, but more generally, given that all the industries are from Nashville, the sample is inevitably non-random compared to the rest of the United States. Moreover, we speculate that the variation across the country in words positively associated with healthcare is likely much less than those negatively associated with healthcare, given the unique industry profile of any given location. Thus, we ultimately did not feel there was sufficient theoretic justification for focusing on these negative keywords much if at all.

CONCLUSIONS

Overall, we were unable to find a classifier that convincingly had sensitivity over 70%. Still, some of the methods we explore here show promise. Some other steps we did not perform, but are worth exploring include use of forward and/or backward variable selection, and using cross-validation to determine an optimal number of keywords to try in a linear model. Given the surprising results using KNN, it also might be worth further investigating use of this model. In particular, if the results we found were in fact a product of using a combination of low variance and high variance predictors, perhaps we could tune a KNN model using some type of forward selection method from a larger set of predictors. Finally, it is also likely worth determining a 'good' list of negative keywords. This would require a truly random sample from a wide portion of the country representing a random sample of industries.

APPENDIX A: DETAILED RESULTS

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Balanced Accuracy	Classifier	Keywords Set
0.85	0.61	0.61	0.85	0.73	KNN	Lasso
0.87	0.59	0.61	0.86	0.73	KNN	T100
0.64	0.82	0.72	0.76	0.73	KNN	T50B50
0.84	0.63	0.62	0.85	0.74	KNN	T10
0.95	0.98	0.97	0.96	0.97	KNN	Ad Hoc
0.88	0.60	0.61	0.88	0.74	LDA	Lasso
0.90	0.53	0.58	0.88	0.71	LDA	T100
0.49	0.85	0.71	0.70	0.67	LDA	T50B50
0.86	0.61	0.61	0.86	0.74	LDA	T10
0.00	1.00	NA	0.58	0.50	LDA	Ad Hoc
0.88	0.60	0.61	0.88	0.74	LPM	Lasso
0.89	0.53	0.58	0.88	0.71	LPM	T100
0.49	0.86	0.71	0.70	0.67	LPM	T50B50
0.86	0.61	0.61	0.86	0.74	LPM	T10
0.33	0.84	0.59	0.63	0.58	LPM	Ad Hoc
0.86	0.63	0.62	0.86	0.74	RF	Lasso
0.84	0.72	0.68	0.86	0.78	RF	T100
0.63	0.84	0.74	0.76	0.74	RF	T50B50
0.83	0.67	0.64	0.84	0.75	RF	T10
0.76	0.69	0.64	0.80	0.73	RF	Ad Hoc
0.89	0.58	0.60	0.89	0.74	SVM (linear)	Lasso
0.84	0.60	0.60	0.84	0.72	SVM (linear)	T100
0.58	0.82	0.69	0.73	0.70	SVM (linear)	T50B50
0.92	0.54	0.59	0.91	0.73	SVM (linear)	T10
0.65	0.37	0.43	0.60	0.51	SVM (linear)	Ad Hoc
0.83	0.67	0.65	0.85	0.75	SVM (radial)	Lasso
0.67	0.76	0.67	0.76	0.71	SVM (radial)	T100
0.21	0.90	0.59	0.61	0.55	SVM (radial)	T50B50
0.84	0.66	0.64	0.85	0.75	SVM (radial)	T10
0.52	0.48	0.42	0.58	0.50	SVM (radial)	Ad Hoc

APPENDIX B: TOP 50 NON-HEALTH KEYWORDS

Neg Words F1	Neg Words F2	Neg Words F3	Neg Words F4	Neg Words F5
pet	tax	pet	shop	project
shop	pet	tax	tax	shop

accessori	project	shop	softwar	tax
project	accessori	accessori	farm	softwar
tax	data	safeti	pet	farm
data	comput	farm	safeti	pet
brand	farm	media	gun	safeti
market	protect	bureau	bureau	gun
comput	bureau	gun	plan	bureau
mobil	gun	busi	buy	plan
client	network	comput	air	buy
safeti	cpa	buy	construct	air
english	quickbook	servic	brand	construct
busi	softwar	air	protect	brand
protect	system	quickbook	email	protect
branch	cloud	cpa	busi	email
sale	busi	frame	comput	busi
ticket	email	booth	sale	comput
engag	air	english	control	sale
claim	branch	ticket	engag	control
servic	cabl	translat	english	engag
translat	engag	social	engin	english
buy	frame	sale	booth	engin
veterinari	mobil	data	frame	booth
invoic	booth	softwar	translat	frame
small	network.servic	mobil	invoic	translat
business	custom	god	ticket	invoic
email	invoic	engin	cpa	ticket
system	comput.support	branch	mobil	cpa
strateg	sale	cabl	pump	mobil
compani	servic	pump	system	pump
cabl	control	control	handl	system
construct	english	nois	accessori	handl
cloud	fan	system	nois	accessori
garden	god	cloud	quickbook	nois
strateg.plan	bank	properti	team	quickbook
softwar	pump	social.media	paint	team
bank	manag.servic	arena	manag.servic	paint
social	engin	network.servic	garden	manag.servic
agent	nois	rubi	compani	garden
pest	track	claim	build	compani
rubi	digit	agricultur	account	build
strategi	brand	software	arena	account

power	rubi	protect	pest	arena
arena	veterinari	store	agricultur	pest
museum	account	small	autom	agricultur
autom	develop	pest	claim	autom
network.servic	strateg.plan	custom	automat	claim
store	termin	business	estat	automat
consult	build	email		estat

APPENDIX C: TOP 100 HEALTH KEYWORDS

Pos Keywords F1	Pos Keywords F2	Pos Keywords F3	Pos Keywords F4	Pos Keywords F5
patient	patient	patient	patient	patient
health	health	health	care	care
care	care	care	health	health
medic	medic	medic	medic	medic
center	dental	dental	treatment	treatment
dental	treatment	treatment	center	center
treatment	notic	notic	dental	research
notic	index	index	surgeri	surgeri
surgeri	center	surgeri	index	dentistri
index	physician	physician	notic	healthcar
physician	dentistri	center	physician	clinic
chiropract	healthcar	doctor	clinic	appoint
line	chiropract	vanderbilt	doctor	doctor
vanderbilt	clinic	line	research	therapi
medicin	doctor	healthcar	vanderbilt	medicin
healthcar	surgeri	medicin	healthcar	hospit
clinic	line	research	line	chiropract
dentistri	vanderbilt	dentistri	hospit	cosmet
doctor	cosmet	clinic	chiropract	famili
therapi	research	hospit	dentistri	medic.center
appoint	hospit	specialty	therapi	nurs
specialty	appoint	therapi	famili	smile
nurs	cigna	cigna	medicin	specialty
cosmet	smile	nurs	cosmet	dentist
medic.center	medicin	practic	appoint	diseas
cigna	practic	appoint	sleep	sold
practic	dentist	chiropract	nurs	practic
famili	famili	cosmet	practic	teeth
smile	foot	medic.center	test	foot
test	implant	well	well	pain

hca	bill	hca	smile	surgic
sold	diseas	pain	foot	test
diseas	therapi	foot	cigna	implant
allergi	pain	famili	medic.center	medicar
find	medicar	pediatr	pain	well
pediatr	teeth	test	hca	pediatr
pain	medic.center	diseas	specialty	locat
well	well	dentist	hear	massag
hospit	new.patient	bill	allergi	skin
medicar	communiti	medicar	dentist	oral
teeth	pediatr	smile	pediatr	new.patient
implant	hca	sold	sold	allergi
dentist	sleep	hear	oral	cancer
oral	allergi	addict	teeth	comfort
bill	specialty	teeth	new.patient	tristar
hear	oral	allergi	cancer	dentur
hill	hill	breast	skin	drug
addict	procedur	sleep	breast	ankl
breast	nurs	juliet	children	home.care
children	addict	implant	diseas	cosmet.dentistri
juliet	locat	massag	massag	arthriti
locat	dental.implant	new.patient	surgic	laboratori
surgic	cosmet.dentistri	procedur	health.care	qualiti
new.patient	dentur	drug	implant	dental.implant
cancer	healthi	oral	bill	orthodont
rehabilit	surgic	hormon	parent	testimoni
hormon	ankl	health.care	ankl	sleep
tristar	dds	healthi	treat	patholog
skin	tristar	ankl	juliet	patients
staff	treat	parent	one	treat
procedur	hear	patholog	disord	bill
mt.juliet	find	park	hormon	feel
pharmaci	bio	comfort	procedur	health.care
dental.implant	juliet	disord	live	patient.form
home.care	one	find	offic	rehabilit
oncolog	arthriti	oncolog	healthi	staff
patients	patient.form	mt.juliet	dds	market.research
laser	patients	treat	rehabilit	ehr
dentur	feel	arthriti	feel	endodont
find.doctor	patholog	psycholog	park	laser
patholog	children	radiolog	laser	pharmaci

cosmet.dentistri	market.research	locat	testimoni	procedur
radiolog	orthodont	children	staff	disord
dds	offic	dentur	tristar	crown
foot	health.care	patient.form	addict	dds
heal	heal	restor	locat	live
health.well	pharmaci	bodi	market.research	restor
psycholog	drug	cosmet.dentistri	heal	bio
health.care	make.appoint	rehabilit	arthriti	make.appoint
qualiti	rehabilit	market.research	home.care	oncolog
comfort	laboratori	hill	mt.juliet	psycholog
parent	staff	dds	patients	region
healthi	dentistry	health.well	feet	offic
drug	feet	alway	health.well	murfreesboro
massag	murfreesboro	cancer	orthodont	communiti
patient.form	nashvill	eye	patient.form	appointment
testimoni	bone	feet	comfort	chiropractor
facil	breast	practition	lab	dermatolog
practition	ehr	feel	hill	radiolog
hundr	radiolog	testimoni	cosmet.dentistri	joint
one	asthma	one	bodi	physic
joint	parent	appointment	dental.implant	patient.care
appointment	find.doctor	endodont	find.doctor	find
chiropractor	oncolog	find.doctor	appointment	conveni
dermatolog	patient.center	heal	bone	vanderbilt
endodont	cancer	laboratori	dentistry	assist.live
adult	testimoni	oak	drug	patient
form	Crown	patients	patient	care
patient	patient	skin	care	

Note that T50B50 is simply the top 50 from this list combined with the previous list, and T10 is simply the top ten keywords from this list. The lasso list simply runs a regression on this list to determine those that it predicts are most associated with healthcare websites and eliminates other variables.

APPENDIX D: LASSO KEYWORDS

Lasso F1	Lasso F2	Lasso F3	Lasso F4	Lasso F5
patient	Patient	patient	patient	patient
health	health	medic	care	care
treatment	care	treatment	treatment	treatment
healthcar	treatment	healthcar	healthcar	dentistri
therapi	physician	dentistri	dentistri	healthcar
cosmet	dentistri	therapi	therapi	clinic
practic	healthcar	practic	cosmet	therapi
pain	cosmet	well	practic	cosmet
dentist	practic	pain	well	dentist
children	therapi	bill	dentist	practic
new.patient	new.patient	allergi	new.patient	surgic
staff	communiti	new.patient	children	massag
patients	procedur	procedur	massag	new.patient
dds	healthi	healthi	procedur	home.care
health.well	dds	comfort	live	treat
	feel	children	dds	feel
	children	bodi	feel	health.care
	dentistry	dds	heal	ehr
	bone	health.well	health.well	dds
	ehr	feel	bodi	assist.live
			bone	
			dentistry	