Telling good restaurants from bad ones: A classification story

Moumita Ghorai

Md Ohiul Islam

Thomas Weinandy

Qing Zang

12 December 2018

Applied Data Mining

Dr. Kevin Lee

Project Report

1. Introduction

Can we determine high ratings of restaurants using data of their business attributes and neighborhoods where they are situated in? The purpose of this paper is investigating this question using classification methods. Ratings of restaurants reflect overall quality of service they provide. Ratings found on websites like Google, Yelp, TripAdvisor, Yellow Pages are some of the sites that are accessed by people who would like to decide services from which business would be the best use of their money. We join Yelp data and IRS (Internal Revenue Services) data to combine effects of business attributes and neighborhood affluence and apply classification methods to predict quality of restaurants of a sample of restaurants in the United States. We use restaurant ratings to create a categorical variable to separate good and bad restaurants. We find that business attributes, geographic location are important predictors of restaurants' ratings. Our analysis also provides comparison of performance of different classification methods available to predict restaurant quality.

2. Data

We will be using a Yelp dataset for this project. The dataset contains 174,567 unique observations, with a wide range of business assessment features like number of reviews, text of the reviews, ratings, number of check-ins, business hours, payment method used, user tips, number of likes, etc. This dataset also contains spatial features such as latitude, longitude, postal code and state. One of the variables in the dataset is about the state of the business (whether the business is open or close) which we plan to use as the dependent variable. We also intend to use a dataset called “Individual Income Tax Zip Code data” with zip code level granularity to connect business attributes of sample organizations with economic and population attributes of

the locality. This dataset contains information about number of returns, which approximates the number of households, number of personal exemptions, which approximates the population, adjusted gross income, wages and salaries and wages of salaries, etc. In our Yelp dataset combined with "Individual Income Tax Zip Code data", we will use business attributes as predictors to predict the business is open or close.

For this study we are using a smaller sample of the whole dataset available on kaggle.com. The data set is granular at individual business level. IRS dataset provided by data.gov is granular at zip code level. The two dataset were joined based on a common spatial identifier, which is US zip code. Since we have tax data of US only, we omit non-US business. We also dropped non-restaurant businesses. After dropping missing values, we randomly selected 1000 observations from the Yelp data.

3. Methodology

3.1 Linear and Quadratic Discriminant Analysis

We have applied both linear and quadratic discriminant analysis with and without cross-validation. We trained our models with 500 of the 1000 observations and made predictions about the 500 observations in the test set. We used ROC curves to show performance of each technique. Figure 1 and 2 depict ROC curves of LDA and QDA without and with cross validation respectively. We can see that QDA performs better than LDA in cross validation in Figure 1, when we do not apply cross validation, but this pattern flips when we apply cross validation; according to Figure 2, LDA performs better than QDA when cross validation is applied.

3.2. Shrinkage methods

Shrinkage methods are widely used to shrink parts of coefficients to be zero. There are three kinds of shrinkage methods we use, which are ridge regression, lasso and elastic net. While ridge regression can't exactly force the coefficients to be zero, lasso and elastic net can. When you have a large set of predictors, lasso and elastic net can exclude trivial coefficients to avoid overfitting problem. Hence, lasso and elastic net can serve as tools to make variable selection (lasso is more popular than elastic net). Note that we are doing a classification problem, which means shrinkage methods should be with logistic rather than OLS regression.

For the procedure, we firstly split the whole data set equally into train and test sets. Then, using 10-fold cross validations, we can find the best lambda. Applying the shrinkage methods with the best lambda on the test set, we can calculate classification test errors. Finally, we fit the whole data set with the best lambda to find the coefficients (see Table 1).

3.3. Logistic regression

The logistic regression is a generalized linear model for classification problem. When you have a dummy variable as dependent variable, logistic regression can estimate how likely that dependent variable is true can be reached. Our procedure to do the logistic regression is as follow. Firstly, we calculate the classification test errors similar to the previous shrinkage methods without doing cross validations. Then we fit the whole data set to form the logistic regression table (see Table 2), from which we can do inference. Finally, we make a reduced logistic regression to calculate classification test errors accordingly. To form the reduced logistic regression, we only keep the variables that have non-zero coefficients in the lasso and have significant coefficients in the logistic regression.

3.4. Decision tree

A decision tree is a non-parametric machine learning method that uses a splitting rule to divide the predictor space into several simple regions into a "tree" format to make a prediction for a given observation. We use a classification tree technique to analyze our dataset. To perform this method, we train half of our dataset and make prediction on the other half. This test error rate of this model is 38.2%.

One of the drawbacks of this method is that this may overfit the data. Therefore, we "prune" the tree with fewer splits based on minimum cross validation error rate. After "pruning" the model gives better predictions with 33.6% error rate.

3.5. Bagging

The decision tree technique may suffer from high variance. Therefore, we use Bagging to reduce the variance of the learning method. Bagging uses several bootstrapped training-samples to train the data to create an ensemble of different trees. The prediction for each observation is then obtained by averaging (or taking a majority vote) over all the fitted decision trees.

3.6. Random Forest

Random forest also uses the bootstrap training-samples to create several decision trees. The only difference from bagging is that this method uses a random sample of predictors is as split candidates from the full set of predictors. Since bagging uses the same set of predictors everytime, this may result in highly correlated trees. In case of such highly correlated trees bagging might not lead to a substantial reduction in variance. Random forest can provide a solution to this problem by 'decorrelating' the trees using random samples of predictors.

3.7 Support Vector Machines

We then ran a Support Vector Machine (SVM) for a classification model to predict when

a restaurant on Yelp has a high or low rating. We used a 10-fold cross validation using linear, radial basis and polynomial basis kernels under a variety of parameter values. The range of values for the parameters of cost, gamma and degree are shown in Table 3. The results of this methodology are included in Table 4 and show the optimal parameter values for each kernel type that minimizes the error. It also shows the error rate and number of support vectors. Overall, the SVM with the lowest error rate is a radial basis kernel with cost of 1 and gamma of 0.5 that results in an error rate of 32.6% with 922 support vectors.

4. Aggregate Results

The first discussion of results is around the predictive capacity of our models. The error rates across models (shown in Table 5 and Figure 3) range from a minimum of 32.6% under the SVM with radial basis kernel to a maximum of reduced logistic regression of 46.8%. The mean error rate was 37.3%. Overall the top three models, ranked by lowest error rates, were SVM with radial basis kernel, LDA, and SVM with linear kernel. If we predict each restaurant however as the dominant class "good" we have an error rate of 36.1%. This demonstrates the difficulty of the problem where our best model is only 3. percentage points better than our baseline.

Next, we turn to inference. Some of the methods as discussed before lend insight into which predictors are more relevant. These results are compared in Table 6 that are drawn from the shrinkage methods, logistic regression and random forests. We see good restaurants are associated with having a high review count (i.e. popular), being in areas with many tax return (i.e. populous), being in areas with high total taxable incomes (i.e. wealthy), staying open on Saturday, closing on Sunday and being a cafe. Even if we cannot predict good restaurants with much accuracy, we can at least identify these positive characteristics.

Appendix

Figure 1: ROC curve for LDA and QDA without cross-validation

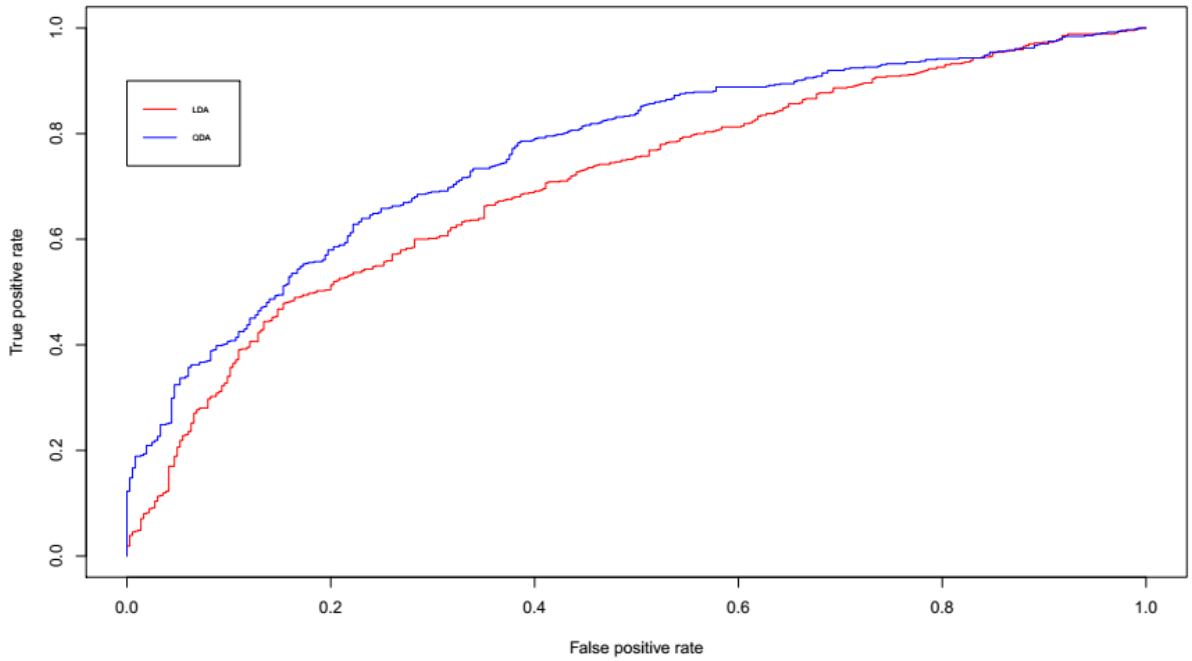

Figure 2: ROC curve for LDA and QDA with cross-validation

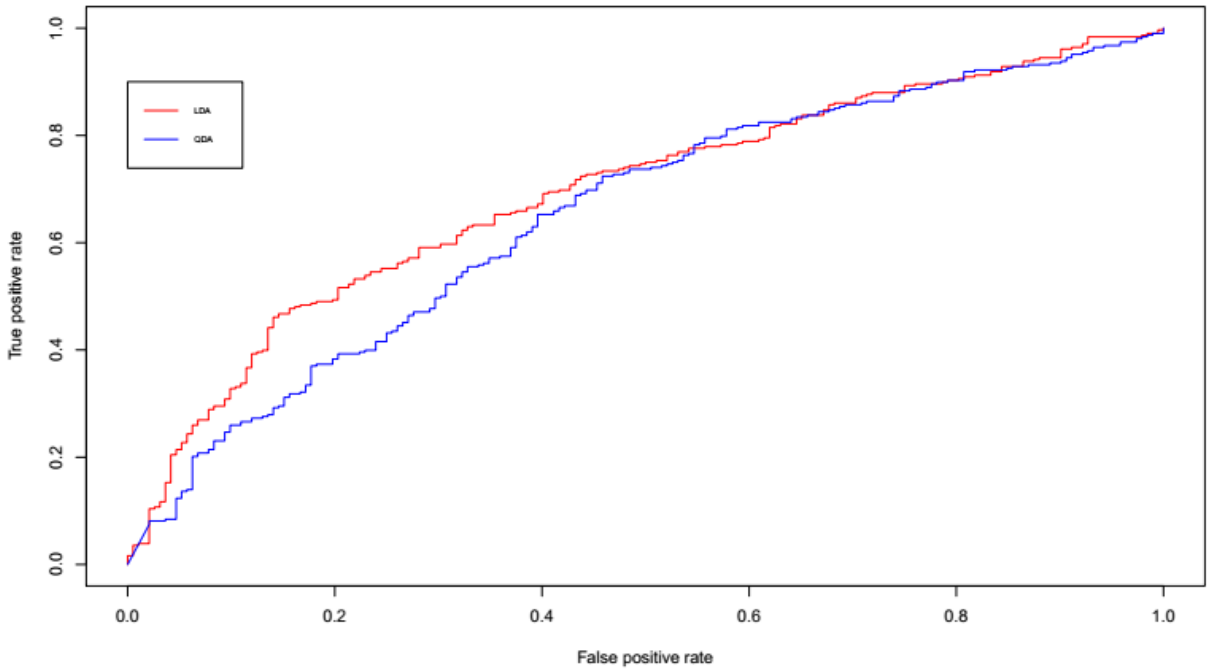

Table 1 Coefficients of Shrinkage Methods

	coef.ridge.final	coef.lasso.final	coef.elasticnet.final
(Intercept)	-9.626665e-03	-3.156275e-01	-2.835260e-01
stateBW	4.470014e-01	4.995703e-01	5.043438e-01
stateIL	-2.422933e-01	.	.
stateNC	-8.705552e-02	.	.
stateNV	-4.406216e-02	.	.
stateOH	8.040498e-02	4.653928e-02	6.383124e-02
statePA	2.385179e-01	2.620551e-01	2.748445e-01
stateSC	-1.044966e-01	.	.
stateWI	2.597271e-01	1.711928e-01	2.112669e-01
review_count	1.439198e-03	3.254224e-03	2.838337e-03
adjusted_gross_income	1.283872e-08	.	.
number_of_returns_with_total_income	-8.408370e-06	.	.
total_income_amount	1.388375e-08	.	.
number_of_returns_with_salaries_wages	-9.944803e-06	-1.470331e-05	-1.485638e-05
salaries_wages_amount	-2.290078e-08	.	.
number_of_returns_with_salaries_wages.1	1.081512e-05	.	.
taxable_interest_amount	3.977330e-06	.	9.356619e-07
ordinary_dividend_amount	1.309646e-06	9.169949e-07	1.054928e-06
business_or_personal_net_income	2.192340e-06	.	7.747667e-07
unemployment_compensation_amount	-6.688967e-06	.	.
taxable_social_security_amount	1.448331e-07	.	.
educator_expenses_amount	-8.709352e-05	.	.
real_estate_tax_amount	2.788267e-06	7.191676e-06	5.404916e-06
sat_open	6.121938e-01	1.168061e+00	1.030950e+00
sun_open	-3.748114e-02	-4.548456e-01	-3.225660e-01
cafe	5.765466e-01	8.295362e-01	8.010975e-01
dinerbuffet	1.435190e-02	.	.
blvd_ave	-1.101924e-01	-4.125457e-02	-5.004876e-02

Table 2 Logistic Regression Table

	Dependent variable: pop_or_not				
	estimate	**std.error**	**statistic**	**p.value**	
(Intercept)	-5.68E-01	3.37E-01	-1.684329	9.21E-02	
stateBW	9.65E-01	6.03E-01	1.602123	1.09E-01	
stateIL	-9.87E-02	5.97E-01	-0.165304	8.69E-01	
stateNC	1.86E-02	2.91E-01	0.063828	9.49E-01	
stateNV	1.67E-01	3.73E-01	0.447826	6.54E-01	
stateOH	4.36E-01	3.67E-01	1.187049	2.35E-01	
statePA	5.64E-01	4.59E-01	1.230016	2.19E-01	
stateSC	7.46E-02	8.49E-01	0.087876	9.30E-01	
stateWI	6.77E-01	5.66E-01	1.196845	2.31E-01	
review_count	4.77E-03	9.86E-04	4.842583	1.28E-06	***
adjusted_gross	-3.97E-05	4.11E-05	-0.967026	3.34E-01	
number_of_ret	-1.08E-04	2.34E-04	-0.463454	6.43E-01	
total_income_a	4.00E-05	4.10E-05	0.974978	3.30E-01	
number_of_ret	1.01E-04	2.58E-04	0.391756	6.95E-01	
salaries_wages	-1.39E-06	1.78E-06	-0.783192	4.34E-01	
number_of_ret	1.31E-04	1.81E-04	0.726298	4.68E-01	
taxable_interes	-6.28E-06	2.24E-05	-0.280807	7.79E-01	
ordinary_divid	-1.12E-06	6.62E-06	-0.169575	8.65E-01	
business_or_pe	3.93E-06	1.31E-05	0.300899	7.63E-01	
unemployment	-1.21E-05	1.19E-04	-0.101504	9.19E-01	
taxable_social_	-1.86E-06	1.75E-05	-0.106381	9.15E-01	
educator_expe	1.80E-03	3.06E-03	0.589628	5.55E-01	
real_estate_ta	-9.09E-06	2.01E-05	-0.452632	6.51E-01	
sat_open	2.30E+00	4.31E-01	5.350743	8.76E-08	***
sun_open	-1.63E+00	4.16E-01	-3.916337	8.99E-05	***
cafe	1.13E+00	3.03E-01	3.72497	1.95E-04	***
dinerbuffet	-2.36E-02	1.51E-01	-0.156455	8.76E-01	
blvd_ave	-1.72E-01	1.61E-01	-1.070356	2.84E-01	

Note: Significant levels 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Parameter Ranges

Kernel	Cost range	Gamma range	Degree range
Linear	(0.0001, 25)	--	--
Radial Basis	(0.0001, 100000)	(0.125, 10)	--
Polynomial Basis	(0.01, 10)	(0.25, 5)	(3, 5)

Table 4: Support Vector Machine Results

Kernel	Optimal cost	Optimal gamma	Optimal degree	Error rate	Number of SV's
Linear	0.1	--	--	0.329	747
Radial Basis	1	0.5	--	0.326	922
Polynomial Basis	0.01	1	3	0.354	615

Table 5: Error Rates Across Models

LDA	QDA	LDA CV	QDA CV	Ridge	Lasso	Elastic Net	Logit
32.7%	34.9%	35.0%	42.4%	38.8%	36.6%	36.6%	44.6%
Logit (reduced)	**Decision Tree**	**Pruned Decision Tree**	**Bagging**	**Random Forest**	**SVM (linear)**	**SVM (radial)**	**SVM (poly.)**
46.8%	38.2%	33.6%	38.0%	37.0%	32.9%	32.6%	35.4%

Figure 3: Error Rates Across Models

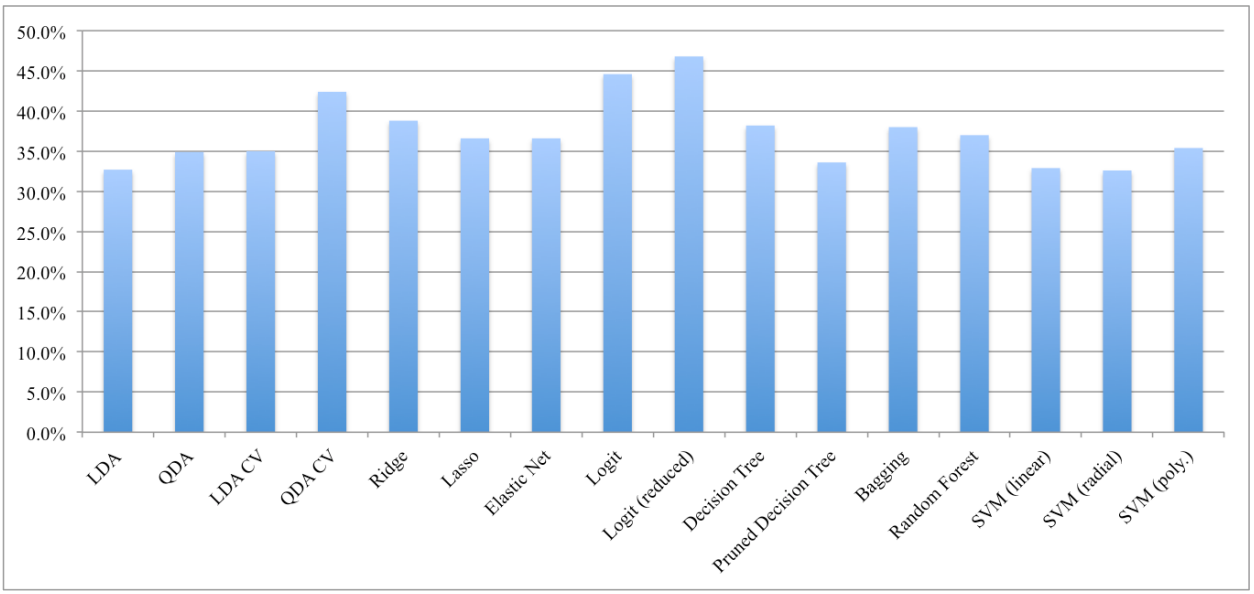

Table 6: Feature Representation

Shrinkage Methods*	**Logistic Regression****	**Random Forests*****
review_count (+)	review_count (+)	review_count
...	...	adjusted_gross_income
...	...	number_of_returns_with_total_income
...	...	total_income_amount
number_of_returns_with_salaries_wages (-)	...	number_of_returns_with_salaries_wages
sat_open (+)	sat_open (+)	
sun_open (-)	sun_open (-)	
cafe (+)	cafe (+)	
ordinary_divend_amount (+)		
business_or_personal_net_income (+)		
**Variables remaining in Ridge, Lasso, and Elastic Net estimates*	***Variables significant at the 0.1 level or less*	***Top five variables ranked in order of importance