Multivariate Analysis Final Group Project

BIA 652-A (Spring 2017)

Yelp Dataset Analysis

<u>Guided by</u>

Prof. Dr. David Belanger

<u>Team Members</u>

Pradeep Basavaraju

Malhar Inamdar

Samuel Mathew

Table of Contents

Abstract

Yelp provides academic students access to their data to use it in an innovative way and break ground in research. In this paper, we target on the business reviews and star rating for restaurants only. In this project, we are trying to identify the key attributes or features that the consumer is looking for their best dining experience. We are using three different algorithms such as logistic regression, random forest and principal component analysis to create our models. After analyzed the performance of each models, the best model for predicting the ratings from reviews and star rating is the random forest algorithm, which exhibited an accuracy of 82%, which is better than the other algorithms that we used in this project.

1. Introduction

Yelp is one of the largest online searching and reviewing systems for kinds of businesses, including restaurants, shopping, home services etc. Yelp has over 150 million monthly unique visitors and more than 121 million reviews (*based on 2014 data*). Yelp has become one of the key decision making tool for consumers to choose or select better restaurants or a service.

We followed a simple life cycle for the data-mining project (refer figure 1).

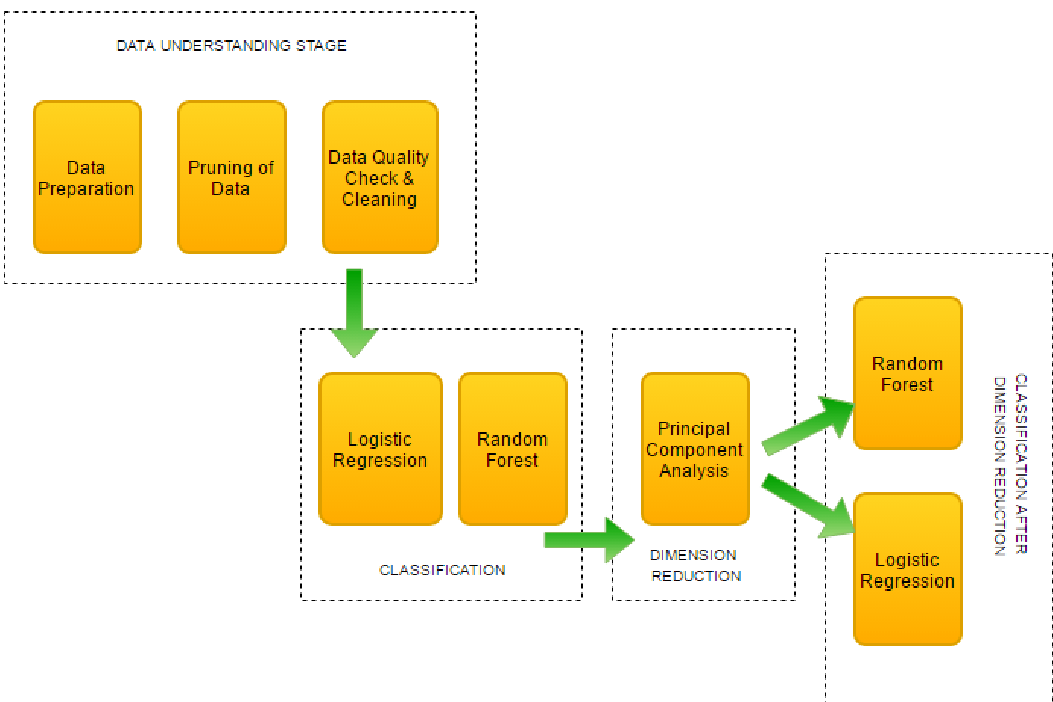


Figure 1: Data Mining Life Cycle

Yelp is one of the best examples of a crowd-sourced review and rating system for the local businesses. The yelp dataset that we have selected is from the “Round 9 Yelp Dataset Challenge”.

2. Understanding the Dataset

Yelp dataset contains over 4.1M reviews and 947 K tips by 1M users for 144K businesses 1.1M business attributes, e.g., hours, parking availability, ambience, Aggregated check-ins over time for each of the 125K businesses, 200,000 pictures from the included businesses and many more attributes. The dataset includes businesses in four different countries: Edinburgh, U.K.; Karlsruhe, Germany; Montreal and Waterloo, Canada; Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, U.S., making it a very versatile dataset. For the purpose of this project, we decided to filter only restaurants that are in United States and we eliminated the rest of the data. By doing so, we were able to reduce the size of the dataset dramatically lower and we were able to easily read the dataset in the tools that we were using to create our models.

Yelp_Business_Dataset attributes:

business_id	Encrypted business id
city	City
state	State
postal_code	Postal code or Zip code
latitude	latitude
longitude	longitude
stars	star rating, rounded to half-stars
review_count	number of reviews
is_open	{closed, open}
BusinessAcceptsCreditCards	{True, False}
RestaurantsReservations	{True, False}
OutdoorSeating	{True, False}
GoodForKids	{True, False}
NoiseLevel	{quiet, loud, average, very_loud}
RestaurantsTableService	{True, False}
RestaurantsPriceRange2	{Inexpensive, Moderate, Pricey, Ultra High-End}
RestaurantsDelivery	{True, False}
BusinessParking	{None, Garage, Street, Validated, Lot, Valet}
Alcohol	{none, full_bar, beer_and_wine}
Ambience	{Normal, Romantic, Intimate, Classy, Hipster, Touristy, Trendy, Upscale, Casual}

The dataset was provided in a JSON format and the dataset look like the following.

```
{
    "business_id":"encrypted business id",
    "name":"business name",
    "neighborhood":"hood name",
    "address":"full address",
    "city":"city",
    "state":"state -- if applicable --",
    "postal code":"postal code",
    "latitude":latitude,
    "longitude":longitude,
    "stars":star rating, rounded to half-stars,
    "review_count":number of reviews,
    "is_open":0/1 (closed/open),
    "attributes":["an array of strings: each array element is an
attribute"],
    "categories":["an array of strings of business categories"],
    "hours":["an array of strings of business hours"],
    "type": "business"
}
```

The json file provided by Yelp was nested and hierarchical structure with varied length, which weren'[t read to the data frames in R. For e.g. if you take a look at categories field, you will notice that the category will have another nested value like {“Fast Food”, “Restaurants”, “Burgers”}. We then tried to using a python code to convert the JSON into a CSV file but yet again nested structure and varied length led our conversion failure.

```
{
  "business_id": "PK6aSizckHFWk8i0oxt5DA",
  "full_address": "400 Waterfront Dr E\nHomestead\nHomestead, PA 15120",
  "hours": {},
  "open": true,
  "categories": [
    "Burgers",
    "Fast Food",
    "Restaurants"
  ],
  "city": "Homestead",
  "review_count": 5,
  "name": "McDonald's",
  "neighborhoods": [
    "Homestead"
  ],
  "longitude": -79.910032,
  "state": "PA",
  "stars": 2,
```

2.1. Data Preparation

Yelp dataset has one of the most complex nested data structure. It was very difficult to even open these dataset files in either SAS or R Studio easily. We created R script to convert the JSON data into sas7bdat file. Using R studio, we were able to prune the dataset before creating the SAS file. To prune yelp dataset, which contains over 90 categories, we performed the following steps.

- We removed features that are not relevant to predicting the success of a business unit. We removed {name, neighborhood, address, hours, type}

```
yelp_Bus_Restaurants <- yelp_Business_data_tbl %>% select(-
starts_with("hours"),
                              -starts_with("name"),
                              -starts_with("neighborhood"),
                              -starts_with("address")) %>%
        filter(str_detect(categories,"Restaurants"))
```

- We removed all business records that are not in United States (“US”) and are not in a “Restaurant” business category.

```
yelp_Bus_attributes <-  yelp_Business_data_tbl %>% select(-
starts_with("hours"),
                            -starts_with("name"),
                            -starts_with("neighborhood"),
                            -starts_with("address")) %>%
  filter(str_detect(categories,"Restaurants")) %>%
  unnest(attributes) %>%
  select(business_id,attributes)
```

- The attributes field consist of 100s of array of strings and each array element is an attribute. We used several different approaches to identify the relevant attribute that are significant to the decision making process. We then extracted these significant attributes into separate datasets and we converted the value from string to numerical (For e.g. {“True”, “False”} changed to {1, 2}).

```
yelp_Bus_attributes_Kids <-
              yelp_Bus_attributes %>%
      filter(str_detect(attributes,"GoodForKids")) %>%
              unnest(attributes) %>%
              select(business_id,attributes)

yelp_Bus_attributes_Kids <-
rename(yelp_Bus_attributes_Kids,c("attributes"="GoodForKids"))
distinct(yelp_Bus_attributes_Kids, GoodForKids)

a<-yelp_Bus_attributes_Kids$GoodForKids %in% "GoodForKids: True"
      yelp_Bus_attributes_Kids[a,2] <- 2
b<-yelp_Bus_attributes_Kids$GoodForKids %in% "GoodForKids: False"
      yelp_Bus_attributes_Kids[b,2] <- 1
```

- After successfully extracting individual attributes and transforming the data from a string to numeric value, our next step was to merge the individual attribute data file to the main dataset in R.

Note: Due to the large size of the dataset and the volume of data, we had to remove large volume of data in order to easily read/import into the tools that we were using for applying different algorithms.

2.2. Data Quality Check and Cleaning

Data quality check and cleaning is an integral part of data mining or data analysis. Data quality check and cleaning phase is mainly to detect and remove or replace erroneous and inconsistent data from the data.

2.3. Missing Values:

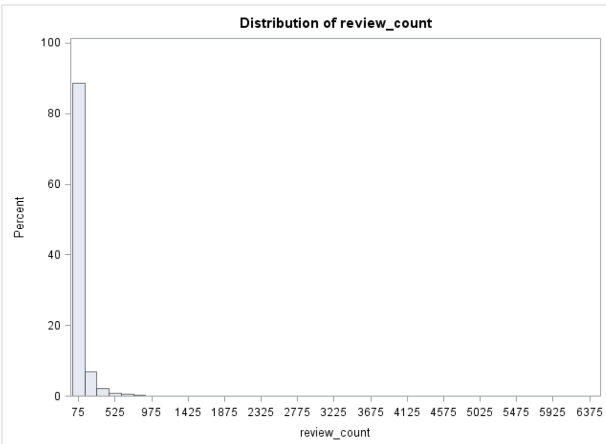
Missing values are a major problem to data analysis. We have noticed that many of the attributes had missing values and we used mode imputation, a widely accepted method for categorical variables to fill in the missing data values.

2.4. Outliers:

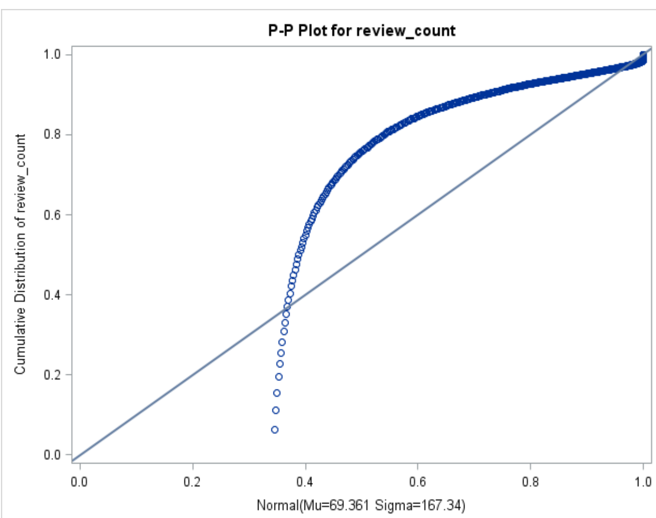
We can verify how many observations data file has and see the names of the variables it contains using PROC CONTENTS. Use PROC FREQ to learn more about categorical variables and to check the distribution of discrete variable. Use PROC UNIVARIATE and CAPABILITY to learn more about continuous variables and its distribution.

From **Appendix 1** we have 30158 observations and 20 variables.

While analyzing each continuous variables, we noticed that the "review count" was right skewed in the distribution histogram. About 95% of the business had a review count which is less than or equal to 500 and nearly 200 business had review count greater than 500.



Basic Statistical Measures			
Location		Variability	
Mean	69.36060	Std Deviation	167.33755
Median	23.00000	Variance	28002
Mode	3.00000	Range	6411
		Interquartile Range	59.00000



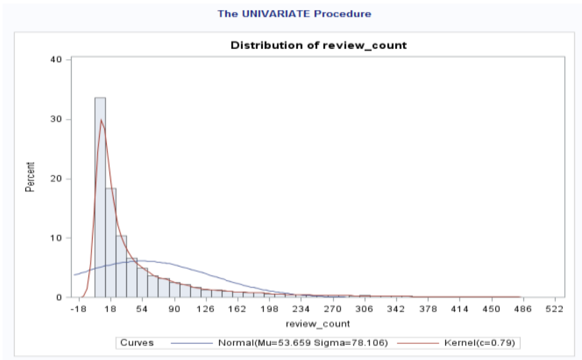
In other words, only very few business had review count beyond 4-digit. We had to eliminate the business records that has review count greater than 500 to reduce the skewness.

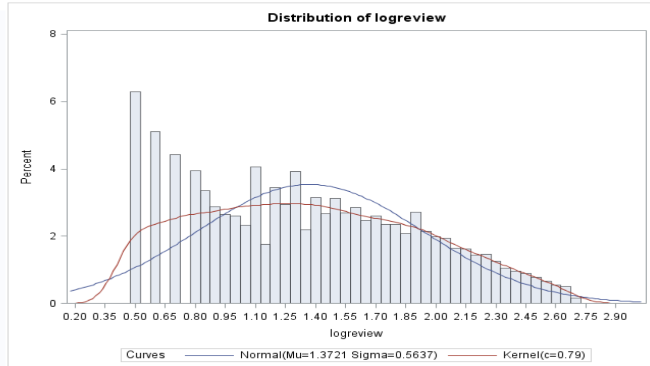
We have noticed that the P-P Plot for review count was not normally distributed. This clearly explains that we should normalize the numerical variable (review count) in order to standardize the scale of effect the variable will have on the result.

We performed a log transformation to normalize the review count variable.

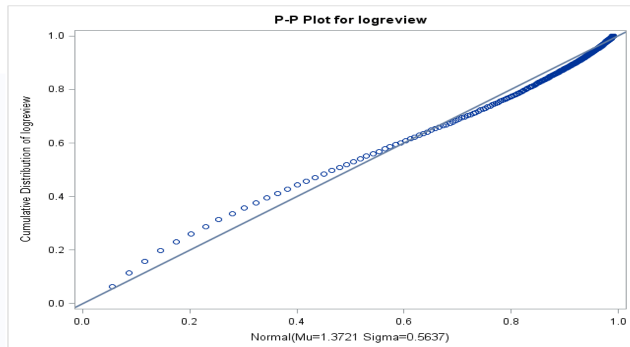
2.5. Transformation:

After removing the review count, which are greater than 500, and applying log transformation techniques, we can notice that the probability distribution is normal and the skewness of data has been reduced.





Basic Statistical Measures			
Location		**Variability**	
Mean	1.372105	**Std Deviation**	0.56367
Median	1.342423	**Variance**	0.31773
Mode	0.477121	**Range**	2.22185
		Interquartile Range	0.89625



3. Classification

3.1 Logistic Regression

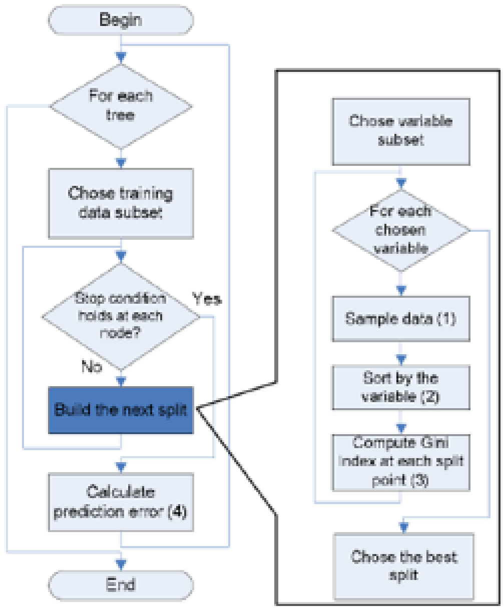
In logistic regression we use a hypothesis class to try to predict the probability that a given example belongs to the “1” class versus the probability that it belongs to the “0” class. Specifically, we will try to learn a function of the form:

$$P(y=1|x)=h_\theta(x)=\frac{1}{1+\exp(-\theta^\top x)}\equiv\sigma(\theta^\top x)$$

The function $\sigma(\theta^T x)$ is often called the “sigmoid” or “logistic” function – it is an S-shaped function that “squashes” the value of $\theta^T x$ into the range [0,1] so that we may interpret $h_\theta(x)$ as a probability. Our goal is to search for a value of θ so that the probability $h_\theta(x)$ is large when *x* belongs to the “1” class and small when *x* belongs to the “0” class.

3.2 Random Forest

Random forest (or random forests) is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.
Random Forest algorithm:



Random Forest Algorithm is one of the most accurate classification algorithms available. It produces a highly accurate classifier for many datasets and can run efficiently on large datasets. It can handle thousands of input variables without variable deletion. One of the most important features of Random forests is that it gives estimates of what variables are important for classification.

4 Dimension Reduction

4.1 Principal Component Analysis

Principal components analysis is a procedure for identifying a smaller number of uncorrelated variables, called "principal components", from a large set of data. The goal of principal components analysis is to explain the maximum amount of variance with the fewest number of principal components.

Principal components analysis are commonly used as the first step in a series of analyses. You can use principal components analysis to reduce the number of variables and avoid multicollinearity, in other words, when you have too many predictors relative to the number of observations.

5 Results and Examples

5.1 Logistic Regression

The area measures discrimination i.e. the ability of test and to correctly classify the star ratings in our case 0.74 (74%) is reasonably a good or fair. The "C" value is equivalent to the well-known measure ROC. The "C" value 0.7 from **Appendix 2** corresponds to the model is good at discriminating the responses.

We can observe that a 77% accuracy in the training data of 100 random samples

From: star	Into: star	Predicted Probability: star=0	Predicted Probability: star=1
1	1	0.4503281608	0.5496718392
1	1	0.1324610982	0.8675389018
1	1	0.138687758	0.861312242
1	1	0.1721824256	0.8278175744
1	1	0.3456578168	0.6543421832
1	1	0.2686139109	0.7313860891
0	1	0.0362881212	0.9637118788
1	1	0.085083165	0.914916835
0	1	0.2985283898	0.7014716102
1	1	0.2336222743	0.7663777257
1	1	0.3338594751	0.6661405249
1	1	0.4133939517	0.5866060483
1	1	0.0109643044	0.9890356956
1	1	0.352335605	0.647664395
1	1	0.2265017966	0.7734982034
1	1	0.1140770221	0.8859229779
1	1	0.2528447041	0.7471552959
1	1	0.0366139314	0.9633860686
1	1	0.0592414421	0.9407585579
1	1	0.4445642211	0.5554357789
1	1	0.1934313749	0.8065686251
1	1	0.4450472649	0.5549527351
1	1	0.3276004611	0.6723995389
0	0	0.5923865103	0.4076134897
1	1	0.0336606395	0.9663393605
1	1	0.0600088095	0.9399911905
1	1	0.2399560587	0.7600439413
0	1	0.0092021152	0.9907978848
0	1	0.1576896273	0.8423103727

Classification: Logistic regression

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of I_star by F_star			
		F_star(From: star)		
	I_star(Into: star)	1	0	Total
	1	75 75.00 76.53 100.00	23 23.00 23.47 92.00	98 98.00
	0	0 0.00 0.00 0.00	2 2.00 100.00 8.00	2 2.00
	Total	75 75.00	25 25.00	100 100.00

5.2 Principal Component Analysis and Logistic regression

In PCA we included all the categorical variables in form of dummy variables. We decided to take all the variables whose eigenvalue is more than or approximately equal to 1 from the correlation matrix. We can notice from the **Appendix 3** output that the principal components together combine 80% cumulative.

For logistic regression with PCA, 80% accuracy is observed on training data of 100 random samples.

From: star	Into: star	Predicted Probability: star=0	Predicted Probability: star=1
1	1	0.0602502113	0.9397497887
0	1	0.0584731722	0.9415268278
1	1	0.4420499241	0.5579500759
1	1	0.0710085665	0.9289914335
1	1	0.1760895565	0.8239104435
1	1	0.0704262901	0.9295737099
0	1	0.3093472191	0.6906527809
1	1	0.4676910778	0.5323089222
1	1	0.3413262273	0.6586737727
1	1	0.2767916856	0.7232083144
1	1	0.302565224	0.697434776
1	1	0.4343845245	0.5656154755
0	1	0.0096605242	0.9903394758
1	1	0.0611071632	0.9388928368
1	1	0.1818413509	0.8181586491
1	1	0.3592324374	0.6407675626
1	1	0.0804228908	0.9195771092
1	1	0.1130376047	0.8869623953
0	1	0.4111857106	0.5888142894
0	1	0.1771617455	0.8228382545
1	1	0.1199949166	0.8800050834
1	1	0.2167781467	0.7832218533
1	1	0.0597705399	0.9402294601
1	1	0.024024848	0.975975152
1	1	0.1524870762	0.8475129238
1	1	0.3245061339	0.6754938661
0	1	0.321912068	0.678087932
1	1	0.1525776844	0.8474223156
1	0	0.5536329954	0.4463670046
0	1	0.1923528519	0.8076471481
1	1	0.1271251992	0.8728748008
0	1	0.3935022844	0.6064977156

Classification: Logistic regression with PCA

The FREQ Procedure

Frequency / Percent / Row Pct / Col Pct	Table of I_star by F_star			
		F_star(From: star)		
	I_star(Into: star)	1	0	Total
	1	79 79.00 79.80 100.00	20 20.00 20.20 95.24	99 99.00
	0	0 0.00 0.00 0.00	1 1.00 100.00 4.76	1 1.00
	Total	79 79.00	21 21.00	100 100.00

5.3 Random Forest Classifier

The main reason why we decided to implement Random Forest classifier is that it gives us the variables in order of their importance. For example, **Appendix 4** table shows us that if anyone wants to open a new restaurant then ‘Parking’, ‘Ambience’, ‘Restaurant Reservations’ are the attributes which needs to be given more importance over ‘Restaurant Delivery’, ‘Price Range’ or ‘Good for kids’

Use PROC HPFOREST in SAS for Random Forest classification. We specified our target as star variable (i.e. the overall stars). In our dataset there are mostly categorical variables, hence we specify the level as nominal. We input all the categorical variables sequentially.
The model has 100 % - 18.9 % = 81.1% accuracy. An 82% accuracy for 100 random samples tested manually

This classifier performed pretty well. The accuracy achieved was 83 % without PCA with all the selected categorical variables.

	star	Predicted: star0	Predicted: star1	Into: star	Warnings
1	1	0.110615888	0.889384112	1	
2	1	0.13125435	0.86874565	1	
3	0	0.4496614034	0.5503385966	1	
4	0	0.1998440989	0.8001559011	1	
5	1	0.2197678348	0.7802321652	1	
6	1	0.0365817894	0.9634182106	1	
7	1	0.0944949344	0.9055050656	1	
8	1	0.2400969716	0.7599030284	1	
9	1	0.3444734307	0.6555265693	1	
10	1	0.2193864788	0.7806135212	1	
11	1	0.0697481223	0.9302518777	1	
12	1	0.1563407643	0.8436592357	1	
13	1	0.3472286072	0.6527713928	1	
14	1	0.2576199269	0.7423800731	1	
15	1	0.1660180581	0.8339819419	1	
16	0	0.1859848267	0.8140151733	1	
17	1	0.313407992	0.686592008	1	
18	0	0.2800830828	0.7199169172	1	
19	1	0.1603466652	0.8396533348	1	
20	1	0.1213831954	0.8786168046	1	
21	1	0.3570558013	0.6429441987	1	
22	1	0.1238356903	0.8761643097	1	
23	1	0.028751734	0.971248266	1	
24	0	0.1269109095	0.8730890905	1	
25	1	0.1300424895	0.8699575105	1	
26	1	0.1670193122	0.8329806878	1	
27	0	0.3186223326	0.6813776674	1	
28	1	0.170379283	0.829620717	1	
29	1	0.4028638731	0.5971361269	1	
30	0	0.3182956264	0.6817043736	1	

Random Forest

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of star by I_star			
		I_star(Into: star)		
	star	1	0	Total
	1	79 79.00 100.00 81.44	0 0.00 0.00 0.00	79 79.00
	0	18 18.00 85.71 18.56	3 3.00 14.29 100.00	21 21.00
	Total	97 97.00	3 3.00	100 100.00

6 Conclusion

We have created different classification models. Based on the observation, we can interpret the accuracy of our model is as follows:

- Logistic Regression - 77% accuracy.
- Logistic Regression with PCA - 80 % accuracy
- Random Forest - 82% accuracy

We can conclude that the model helps us to identify the attributes that helps to contribute to the success of the business (in the order of its importance such as parking, review count, credit card, ambiance etc.) and to eliminate the attributes, which are least significant.

Appendix:

Appendix 1: Data quality check and cleaning

The SAS System

The CONTENTS Procedure

Data Set Name	YELP.YELPDATA_BUSINESS	Observations	30158
Member Type	DATA	Variables	20
Engine	V9	Indexes	0
Created	04/28/2017 17:43:43	Observation Length	200
Last Modified	04/28/2017 17:43:43	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information	
Data Set Page Size	65536
Number of Data Set Pages	93
First Data Page	1
Max Obs per Page	327
Obs in First Data Page	312

#	Variable	Type	Len
10	Alcohol	Num	8
11	Ambience	Num	8
13	BusinessAcceptsCreditCards	Num	8
14	GoodForkids	Num	8
15	NoiseLevel	Num	8
16	OutdoorSeating	Num	8
12	Parking	Num	8
18	PriceRange	Num	8
17	RestaurantsDelivery	Num	8
19	RestaurantsReservations	Num	8
20	RestaurantsTableService	Num	8
1	business_id	Char	22
2	city	Char	36
9	is_open	Num	8
5	latitude	Num	8
6	longitude	Num	8
4	postal_code	Char	8
8	review_count	Num	8
7	stars	Num	8
3	state	Char	3

Parking	Frequency	Percent	Cumulative Frequency	Cumulative Percent
None	13036	43.23	13036	43.23
Garage	1726	5.72	14762	48.95
Street	3173	10.52	17935	59.47
Validated	37	0.12	17972	59.59
Lot	12018	39.85	29990	99.44
Valet	168	0.56	30158	100.00

Appendix 2: Logistic Regression

Classification: Logistic regression

The LOGISTIC Procedure

Model Information	
Data Set	WORK.YELPDATA_BUSINESS1
Response Variable	star
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

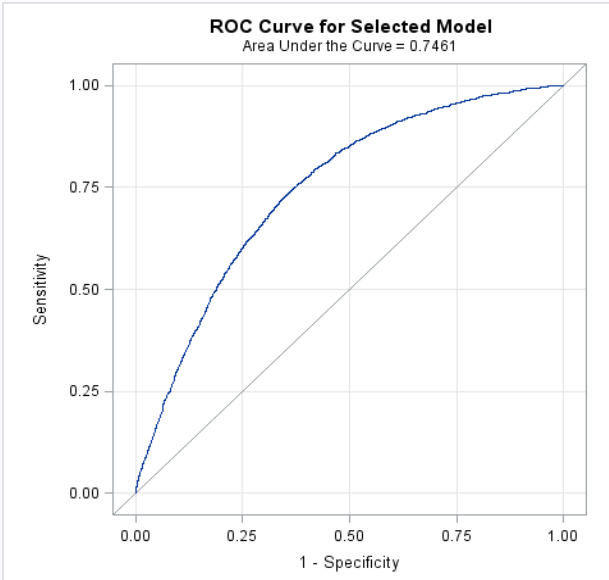
Number of Observations Read	29621
Number of Observations Used	29621

Response Profile		
Ordered Value	star	Total Frequency
1	0	5608
2	1	24013

Probability modeled is star=0.

Summary of Stepwise Selection							
	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	Parking		5	1	1840.9374		<.0001
2	logreview		1	2	510.2990		<.0001
3	NoiseLevel		3	3	342.8573		<.0001
4	Alcohol		2	4	216.4231		<.0001
5	BusinessAcceptsCredi		1	5	149.2635		<.0001
6	RestaurantsReservati		1	6	111.9927		<.0001
7	Ambience		8	7	108.7987		<.0001
8	RestaurantsTableServ		1	8	60.2708		<.0001
9	OutdoorSeating		1	9	62.2360		<.0001
10	is_open		1	10	27.2844		<.0001
11	RestaurantsDelivery		1	11	23.3045		<.0001
12	GoodForkids		1	12	9.2266		0.0024
13	PriceRange		3	13	8.3359		0.0396

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.5243	0.1517	100.9621	<.0001
logreview		1	-0.6851	0.0473	210.1851	<.0001
is_open	1	1	0.2026	0.0384	27.8934	<.0001
Alcohol	2	1	-0.5964	0.0681	76.6088	<.0001
Alcohol	3	1	-0.0481	0.0520	0.8545	0.3553
Ambience	1	1	-0.7717	0.3318	5.4103	0.0200
Ambience	2	1	-2.2364	0.7137	9.8187	0.0017
Ambience	3	1	-2.1962	0.5077	18.7140	<.0001
Ambience	4	1	-1.1462	0.2911	15.5035	<.0001
Ambience	5	1	0.6889	0.2533	7.3967	0.0065
Ambience	6	1	-0.9703	0.2025	22.9518	<.0001
Ambience	7	1	-0.6110	0.5259	1.3499	0.2453
Ambience	8	1	-0.2604	0.0496	27.5609	<.0001
Parking	1	1	-0.1015	0.0889	1.3022	0.2538
Parking	2	1	-1.0946	0.0834	172.2815	<.0001
Parking	3	1	-1.2594	0.7471	2.8416	0.0919
Parking	4	1	-0.4609	0.0453	103.6463	<.0001
Parking	5	1	-1.2926	0.5202	6.1734	0.0130
BusinessAcceptsCredi	2	1	1.4763	0.1306	127.7088	<.0001
GoodForkids	2	1	-0.1671	0.0524	10.1563	0.0014
NoiseLevel	2	1	0.8668	0.0746	134.9621	<.0001
NoiseLevel	3	1	0.2687	0.0463	33.6221	<.0001
NoiseLevel	4	1	1.3321	0.0981	184.3171	<.0001
OutdoorSeating	2	1	-0.2888	0.0374	59.6347	<.0001
RestaurantsDelivery	2	1	-0.2125	0.0428	24.6316	<.0001
PriceRange	2	1	0.0526	0.0367	2.0596	0.1512
PriceRange	3	1	-0.2460	0.1229	4.0064	0.0453
PriceRange	4	1	0.1605	0.2070	0.6011	0.4382
RestaurantsReservati	2	1	-0.3470	0.0509	46.3890	<.0001
RestaurantsTableServ	2	1	-0.3319	0.0380	76.1686	<.0001

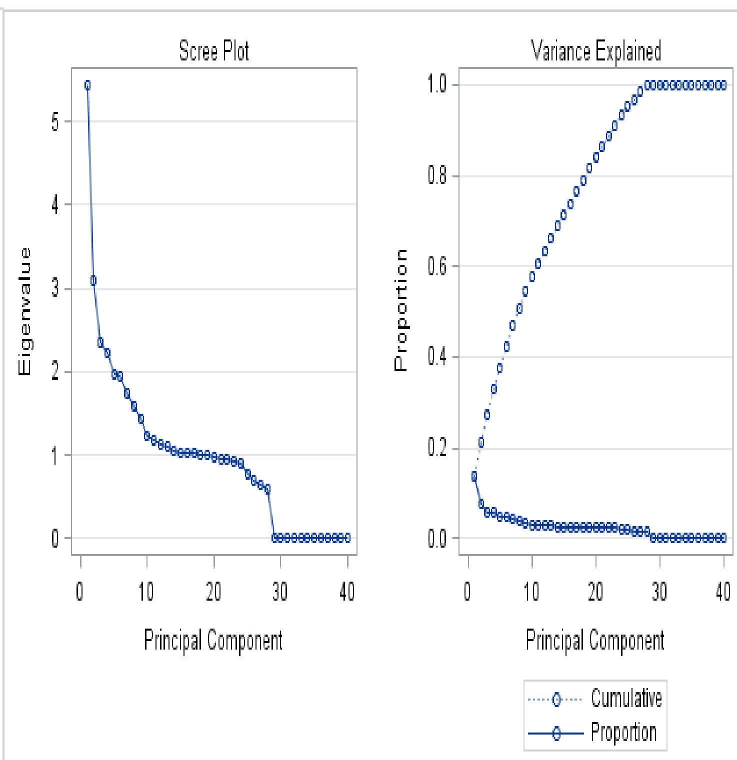


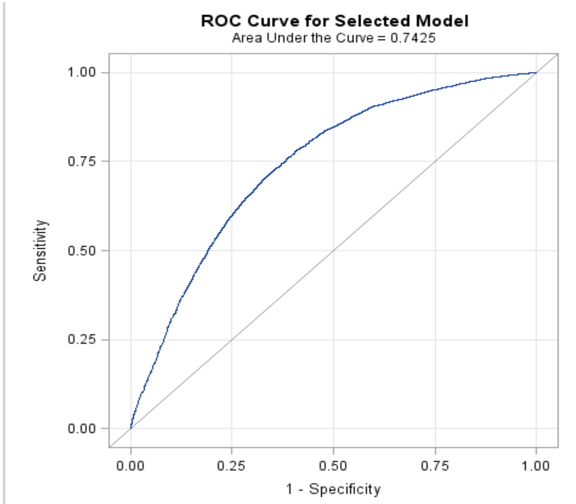
Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.6	Somers' D	0.492
Percent Discordant	25.4	Gamma	0.493
Percent Tied	0.1	Tau-a	0.151
Pairs	134664904	c	0.746

Appendix 3: Principal Component Analysis

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.44096907	2.34897093	0.1360	0.1360
2	3.09199814	0.74778041	0.0773	0.2133
3	2.34421773	0.12439422	0.0586	0.2719
4	2.21982351	0.25527363	0.0555	0.3274
5	1.96454988	0.00639286	0.0491	0.3765
6	1.95815702	0.20326613	0.0490	0.4255
7	1.75489089	0.16154311	0.0439	0.4694
8	1.59334778	0.14976215	0.0398	0.5092
9	1.44358564	0.20319982	0.0361	0.5453
10	1.24038582	0.04930559	0.0310	0.5763
11	1.19108022	0.06560365	0.0298	0.6061
12	1.12547658	0.02668009	0.0281	0.6342
13	1.09879649	0.03805625	0.0275	0.6617
14	1.06074024	0.02788308	0.0265	0.6882
15	1.03285716	0.01035749	0.0258	0.7140
16	1.02249967	0.00385380	0.0256	0.7396
17	1.01864587	0.01112928	0.0255	0.7651
18	1.00751659	0.01639711	0.0252	0.7902
19	0.99111948	0.01882163	0.0248	0.8150





Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.7532	0.0629	143.4721	<.0001
logreview	1	-0.7249	0.0460	248.2553	<.0001
Prin1	1	-0.2027	0.00928	477.1302	<.0001
Prin2	1	0.1292	0.0129	100.1552	<.0001
Prin5	1	-0.1709	0.0136	157.5136	<.0001
Prin6	1	0.2334	0.0138	286.9716	<.0001
Prin7	1	0.0374	0.0125	8.9336	0.0028
Prin8	1	-0.1517	0.0140	117.4174	<.0001
Prin9	1	-0.0721	0.0153	22.1266	<.0001
Prin10	1	-0.1063	0.0168	39.9473	<.0001
Prin11	1	0.2474	0.0191	166.9767	<.0001
Prin12	1	-0.1041	0.0180	33.4737	<.0001
Prin14	1	-0.0804	0.0161	24.8293	<.0001
Prin16	1	0.0437	0.0160	7.4830	0.0062

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.2	Somers' D	0.485
Percent Discordant	25.7	Gamma	0.485
Percent Tied	0.1	Tau-a	0.149
Pairs	134664904	c	0.743

Appendix 4: Random Forest

Model Information		
Parameter	Value	
Minimum Category Size	30	(Default)
Leaf Size	6	
Maximum Depth	50	
Maximum Trees	500	
Minimum Category Size	5	(Default)
Variables to Try	4	
Alpha	0.05	
Exhaustive	5000	(Default)
Leaf Fraction	0.001	(Default)
Inbag Fraction	0.6	
Node Size	100000	(Default)
Prune Fraction	0	(Default)
Prune Threshold	0.1	(Default)
Rows of Sequence to Skip	5	(Default)
Split Criterion	.	Gini
Missing Value Handling	.	Valid value

Number of Observations	
Type	N
Number of Observations Read	29621
Number of Observations Used	29621

Baseline Fit Statistics	
Statistic	Value
Average Square Error	0.153
Misclassification Rate	0.189

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
Parking	413	0.012127	0.007855	0.024253	0.015864
logreview	508	0.006725	0.003585	0.013450	0.008021
Ambience	260	0.005731	0.003582	0.011463	0.007375
RestaurantsReservations	212	0.002998	0.001853	0.005996	0.003948
BusinessAcceptsCreditCards	153	0.001696	0.001049	0.003391	0.002149
NoiseLevel	473	0.002126	0.001044	0.004252	0.002543
OutdoorSeating	277	0.001801	0.000984	0.003602	0.002167
RestaurantsTableService	252	0.001364	0.000712	0.002728	0.001633
Alcohol	224	0.001063	0.000556	0.002126	0.001301
is_open	159	0.000792	0.000364	0.001585	0.000908
RestaurantsDelivery	232	0.000556	0.000160	0.001113	0.000538
GoodForkids	172	0.000366	0.000068	0.000733	0.000311
PriceRange	220	0.000465	0.000035	0.000931	0.000328

References

Dzieciolowski, A. (n.d.). *Demystifying Random Forests.* Retrieved from https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/dzieciolowski-randomforests.pdf

Education, U. -I. (n.d.). *SAS Annotated Output*. Retrieved from http://stats.idre.ucla.edu/sas/output/

Know, S. T. (n.d.). Retrieved from Overview: LOGISTIC Procedure: https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect001.htm

Know, S. T. (n.d.). *Syntax: PRINCOMP Procedure*. Retrieved from https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#princomp_toc.htm

Know, S. T. (n.d.). *The SURVEYSELECT Procedure*. Retrieved from https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#surveyselect_toc.htm

Praveenkumar Kondikoppa, S.-J. P. (n.d.). Modeling Business Trends based on Yelp Review Data. 1.

Stanford, U. o. (n.d.). *Supervised Learning and Optimization*. Retrieved from Logistic Regression: http://ufldl.stanford.edu/tutorial/supervised/LogisticRegression/

Wicklin, R. (n.d.). *Techniques for scoring a regression model in SAS*. Retrieved from http://blogs.sas.com/content/iml/2014/02/19/scoring-a-regression-model-in-sas.html

Yelp. (n.d.). *About Us*. Retrieved from Yelp: https://www.yelp.com/about