Prediction analysis of restaurant shutdown

Helping businesses to survive

Arpit Singh
110162005
arpit.singh@stonybrook.edu

Priya Sharma
110285111
priya.sharma@stonybrook.edu

Rishabh Agrawal
110467487
rishabh.agrawal.1@stonybrook.edu

ABSTRACT

In this paper, we aim to develop prediction models for restaurant closure using various machine learning techniques. We have chosen Yelp business data set and extracted data for 21892 restaurants. Using this, we did extensive analysis to know the impact of various factors on restaurant closure and selected 9 factors viz . price, noise, star rating, no of reviews, specialty count(count of specialties for which restaurant is good at), good for lunch, good for dinner, good for groups and parking to form our feature set. This paper presents various techniques from field of data science and machine learning which can use yelp datasets to predict shutdown of a restaurant business. We treat Prediction of restaurant shutdown as a binary classification problem and compare 7 models to come up with the best one - i) Logistic Regression ii) Bernoulli Naive Bayes iii) Gaussian Naive Bayes iv) Multinomial Naive Bayes v) K-Nearest Neighbor classification vi) Random Forest Classification vii) Support Vector machine. We present time-series analysis to forecast when a restaurant can be expected to shut down in future. We achieved an accuracy of around 80% for most of the modeling techniques applied. This paper includes textual analysis of reviews to find out common words used for describing restaurants with low ratings. Location based analysis of closed and open restaurants has been performed to determine cities /states conducive for restaurant business. This paper can be of interest to restaurant owners, managers as well economists to determine the health of restaurant business.

Categories and Subject Descriptors

D.4.8 [**Performance**]: Modeling and prediction

General Terms

Prediction, Shutdown, Dataset,Restaurant closure,

Keywords

Machine Learning, Data Science, Classification, Analysis,time series analysis, location based analysis

1. INTRODUCTION

Restaurant Industry is one of the huge money minter in the USA. To give an idea of its vastness, total sales in restaurant last year is $ 709 billion[1]. There are about 1 million restaurants across the country employing at least 14 million people. However, a large number of restaurants fail every year and consequently shut down. As per research conducted by Cornell and Michigan State University[2] , 27 % of restaurants close in their first year and as large as staggering 50 % restaurants close after 3rd year. Having said that, prediction analysis of a restaurant closure is an interesting and useful problem. Lack of awareness about how it is performing can be disastrous for any business and could cause it to lose all its business. Yelp provides a website to publish crowd-sourced local business reviews along with a one to five star rating system. It also enables businesses to update their contact information, hours and other features which can help users to make best suitable choices for themselves. This website is widely used in major metropolitan regions of the United States of America. Yelp has provides a business dataset in which we can find out which restaurants have closed down.Today we are available with numerous techniques offered by field of Data Science which can help us to anticipate, if/when a restaurant will run out of business and consequently shutdown, using data made available by yelp.

2. PRIOR WORK

We derived our motivation from work done by Researchers at the University of Maryland Robert H. Smith School of Business[3]. They have used Yelp reviews to predict chances of restaurant shutdown. Their approach involves combining text analysis with restaurant rankings to predict if a restaurant would close within the next three months which fetched 70 percent accuracy. However above work focused only on textual analysis to predict closure, our work has taken multiple features into account like noise, price range, star rating , whether a restaurant is good for lunch dinner etc. to design machine learning models which can predict restaurant closure with around 80% accuracy.

3. DATA PREPARATION AND CLEANING

In Yelp data, there are nearly 21,892 restaurants. A total of 4,334 restaurants have closed down and 17,558 are still

open. We have chosen 9 relevant features namely -review, star rating, price, noise, ambiance, good for lunch, good for group, good for dinner, parking and good count. Feature **Closed** indicates whether restaurant is closed or open, review represents number of review, star represent star ratings, noise indicates noise level divided into four parts. Price range indicates perceived price range of restaurant divided in 4 levels, 1 being least expensive and 4 being most expensive. Good count feature represents number of categories it is good for(viz. groups,kids,lunch,dinner,breakfast,brunch etc) with maximum value of 7. Ambiance ,good for lunch ,good for dinner ,parking and good for groups are binary features with value 1 representing presence of a given feature. The values which are blank are considered to be false.

- Filtered all datasets to only be left with information pertaining to restaurants.
- Explored various features to find which of those are distinguishable for closed/open restaurants using logistic regression.
- Picked up features which had sufficient completeness to be worth being considered as decisive features as input to prediction or classification models.
- A lot of features provide more specifics than needed. In order to reduce dimensionality, we analyzed and grouped similar features. For example: dataset for business comprises of multiple fields related to parking spaces at restaurants. They are- attributes.Parking.lot, attributes.Parking .street, attributes.Parking.garage, attributes.Parking.valet, attributes.Parking. validated. These particulars were not necessary for our purpose. We worked on the data-set to reduce these attributes to two- Parking.available and Parking.not available.
- There are four umbrella characterize where we have grouped restaurant features. 1. Ambiance 2. Music 3. Good for 4. Parking.
- In ambiance , there are 8 features viz: touristy, hipster, romantic, dicey, intimate, trendy, upscale, classy, casual. However, out of 4334 closed restaurant, 70% restaurant has no ambiance feature mentioned and out of 17,558 open restaurant only 50% has atleast one feature mentioned. In the given 8 features, 7 features viz. touristy, hipster, romantic, dicey, intimate, trendy, upscale, classy are present in less than 1% of total restaurants and Casual feature is mentioned in 36% of total restaurant. Thus we will consider ambiance as single binary feature whose value will be 1 if even single ambiance is enabled
- Music feature which consists following features: Live , Video, Background, DJ, Play, Karaoke and Jukebox is mentioned for less than 6% of total restaurants. Thus music feature is not useful to be included for predicting restaurant closure.
- Good for feature consists of sub features: Lunch, Dinner, Kids, Groups, Dancing, Brunch, Late night, Breakfast and Desert. A majority of features like late night, dancing, kids, brunch and desert are enabled for less

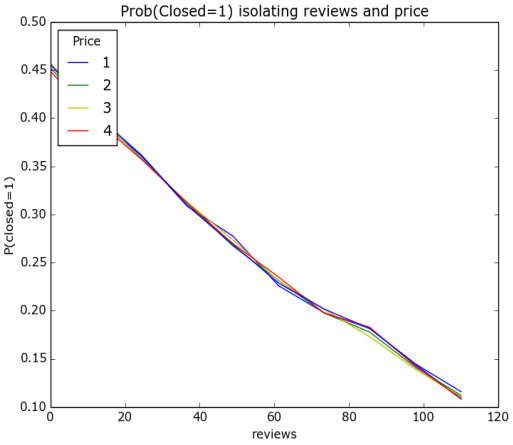


Figure 1: Reviews Vs closure

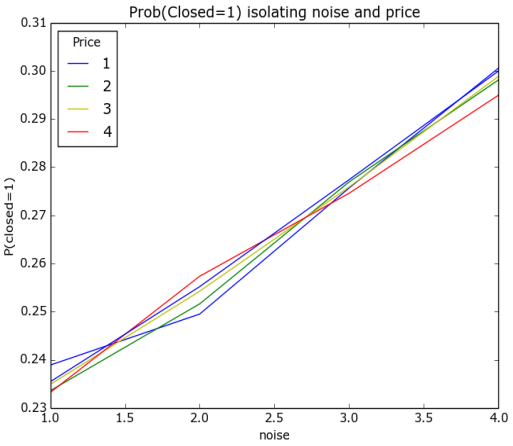


Figure 2: Noise Vs closure

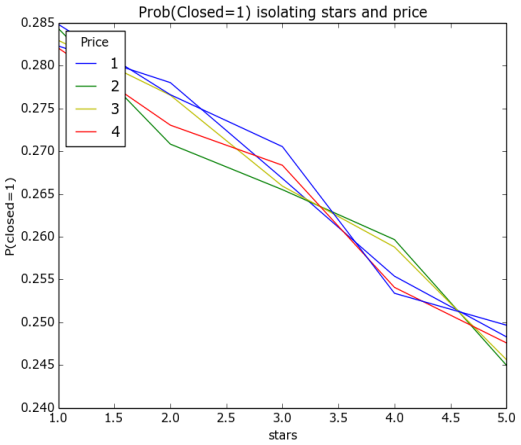


Figure 3: Star Rating Vs Closure

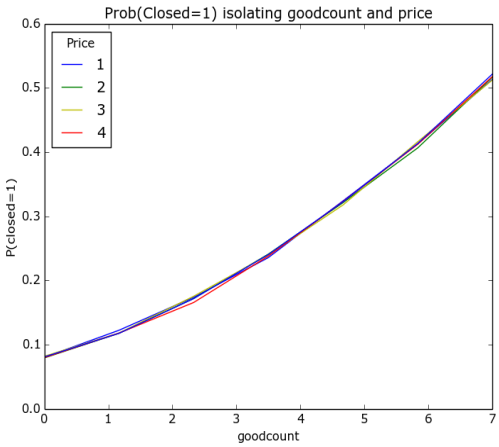


Figure 4: GoodCount Vs Closure

than 1% of restaurants. However, Lunch, Dinner and groups are present in 32%, 25% and 80% of total restaurants. Hence we will consider all these individual sub features as part of our feature set.

- Parking feature covers: Valet, Garage, Street, Lot and Validate. At least 58% of closed restaurants and 55% of open restaurants has at least one type of parking available. Thus we will consider parking as binary feature whose value will be 1 if any kind of parking is available

4. EXPLORATORY DATA ANALYSIS USING LOGISTIC REGRESSION

We have applied logistic regression to obtain relationship between probability of restaurant being closed and other features in dataset. For purpose of logistic regression we took price as a dummy variable and ran logistic regression for rest of variables viz reviews, noise, good count and stars.

The given variables displayed strong relationship with restaurant being closed.

From figure 1, it is clear that number of reviews have negative correlation with probability of closure. The probability of restaurant closure reduces drastically as number of review increases.

From figure 2, it is evident there is a close relationship between a restaurant closure and noise level. It appears that restaurants with high noise level tend to close more. However variation ranges from 23% to 30%. Although it does not makes a huge impact but there is small positive correlation between restaurant closure and noise level. With increasing noise level, probability of closing increases.

From figure 3 we can see that there is small impact of star rating with restaurant closure. Since probability of restaurant closure varies between 28% to 24%. As evident, restaurant with higher star rating tends to survive more than restaurants with lower star rating. Price does not seems to affect restaurant closure. Owing to small variation range of probability of closing over full star rating, it can be assumed that star rating has small impact on restaurant closure.

In figure 4 graph depicts that a restaurant which has large number of good facilities available seems to close more. However it might be because of increased cost that restaurant offering large number of good facilities cannot remain open. It is observed that restaurant offering with none, 1 or 2 facilities are open and probability of closing is fairly low. However as soon as good count goes beyond 4, probability of closing increases. Figure 6 depicts logistic regression results and provides correlation coefficients between various factors and probability of restaurant closure. It is evident that number of reviews and star rating bear negative correlation with probability of restaurant closure.

	precision	recall	f1-score	support
0	0.77	0.97	0.86	2765
1	0.68	0.19	0.30	651
2	0.53	0.13	0.21	280
3	0.35	0.10	0.15	62
4	0.11	0.12	0.12	8
5	0.00	0.00	0.00	0
avg / total	0.73	0.75	0.70	3766

Figure 5:Classification Report for time series

5. TIME SERIES ANALYSIS

We did a time series analysis to predict when a restaurant would close. Approach followed is as follows:

1. Filter restaurant businesses in review data.
2. Find last date of review for each business id in review. We are treating this as closing date.
3. Select closed businesses in the review set.
4. Create review data with 12 month aggregation. Take average of star ratings and count of all reviews for each business.
5. Aggregation will create classes, these are years before closing.
6. Split training and testing set.
7. Apply multifold validation on training set to fit best model.
8. Use training data to train the selected model and predict on test data.
9. Resultant class is year before closing or last working year

ML model is used for multiclass classification to predict duration before closure: **Random Forest**.

Accuracy of multiclass classification using random forest : 75.49%
. Figure 5 depicts classification report per class for time series classification.

Table 1 summarizing statistics is given below:

Table 1: Statistics of Time Series

Metrics	Results
Samples in training data	15072
Samples in test data	3766
Correct Predictions	2846
Accuracy	75.49%
Recall score	0.7549
Precision score	0.7262

Logit Regression Results

Dep. Variable:	closed	No. Observations:	21892
Model:	Logit	Df Residuals:	21882
Method:	MLE	Df Model:	9
Date:	Wed, 02 Dec 2015	Pseudo R-squ.:	0.09364
Time:	20:55:14	Log-Likelihood:	-9872.9
converged:	True	LL-Null:	-10893.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[95.0% Conf.	Int.]
reviews	-0.0152	0.001	-16.878	0.000	-0.017	-0.013
stars	-0.2113	0.018	-11.603	0.000	-0.247	-0.176
price	0.0005	0.025	0.020	0.984	-0.048	0.049
noise	-0.0926	0.020	-4.637	0.000	-0.132	-0.053
ambience	-1.2361	0.056	-22.047	0.000	-1.346	-1.126
lunch	0.2047	0.056	3.642	0.000	0.095	0.315
group	-0.5041	0.063	-8.049	0.000	-0.627	-0.381
dinner	1.1622	0.060	19.320	0.000	1.044	1.280
parking	0.8250	0.050	16.615	0.000	0.728	0.922
goodcount	0.0964	0.039	2.483	0.013	0.020	0.172

Figure 6:Logistic Regression Summary

6. LOCATION BASED ANALYSIS

Location based analysis of closed as well open restaurant is done. Figure 7(a) contains a map containing the location of open restaurants. Larger dot indicates larger frequency. A large number of open restaurants are present in New York and California region indicating that industry thrives in big cities. Apart from New York and California, Arizona and North Carolina cities seem to be good places where a large number of open restaurants reside.

Fig 7(b) shows cities where closed restaurants are present. It seems, closed restaurants are mainly present in smaller cities and far flung areas. Number of closed restaurants in New York is very low. Number of closed restaurants seem to be high in Pennsylvania, Ohio and few cities of Arizona region indicating that restaurant business in not thriving in those regions.

7. TEXTUAL ANALYSIS OF REVIEW DATA

Basic idea is to use review text to analyze common trends of these businesses. To be able to do so we need to come up with text analysis model, which will help in trends mining. Our work on reviews is based on n-grams and frequency distribution approach.

Problems trying to solve:
Problem 1: Finding out what people care for the most in business category which leads to differences in star ratings. Finding these attributes for a business category will be helpful for all the businesses in that category.
Problem 2: Finding out what people like or dislike in a business entity based on user review and review ratings: Finding these attributes will be helpful for a business unit to understand its weakness and strength.
Figure 8(a) shows textual analysis of general trends in restaurant category. Figure 8(b) shows review trends for most popular restaurant.

8. PREDICTION MODELING OF RESTAURANT CLOSURE

We have used following supervised learning algorithms to train our prediction models:

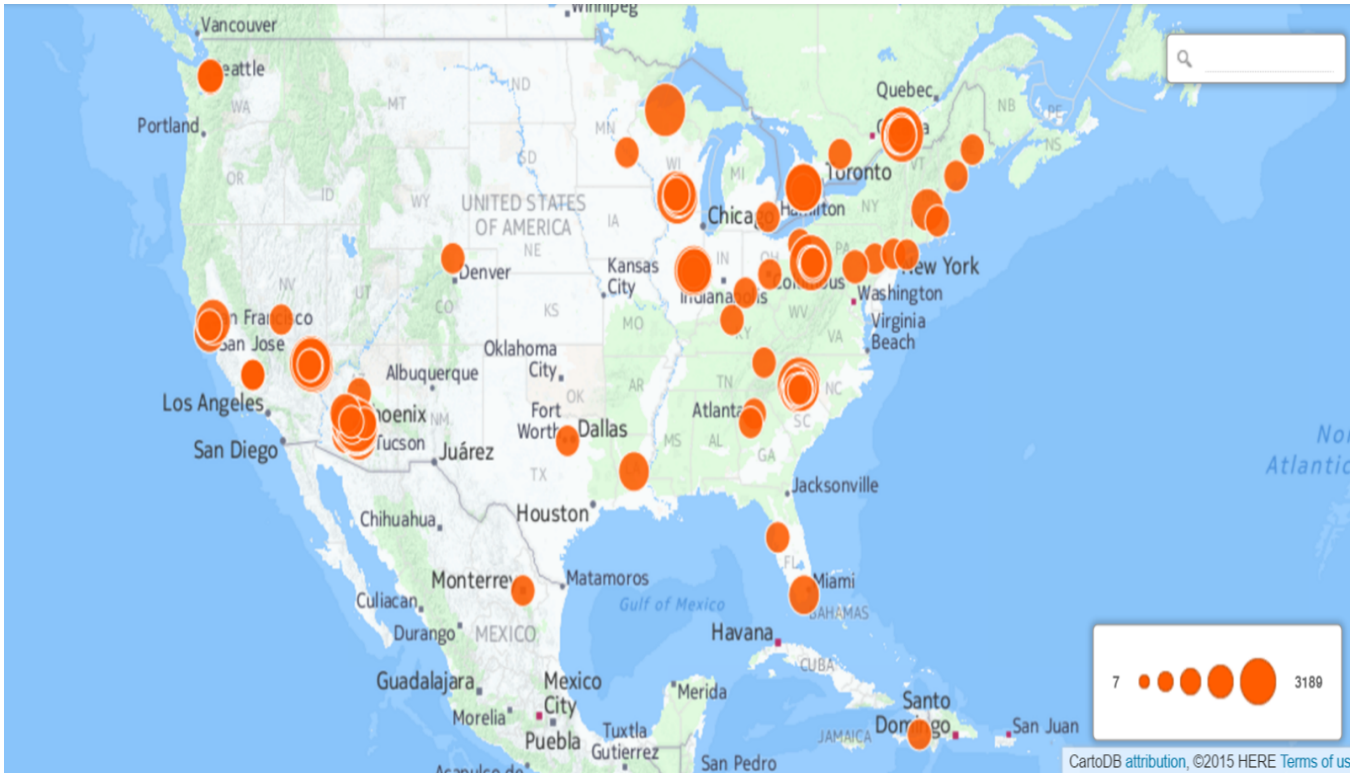
8.1 Logistic regression

This is a classification algorithm used to estimate a binary variable using a given set of independent variable(s). This computes conditional probability P(C|F) where F is the feature vector as described in previous sections and S is either 0 or 1 to represent closed and open restaurant respectively.

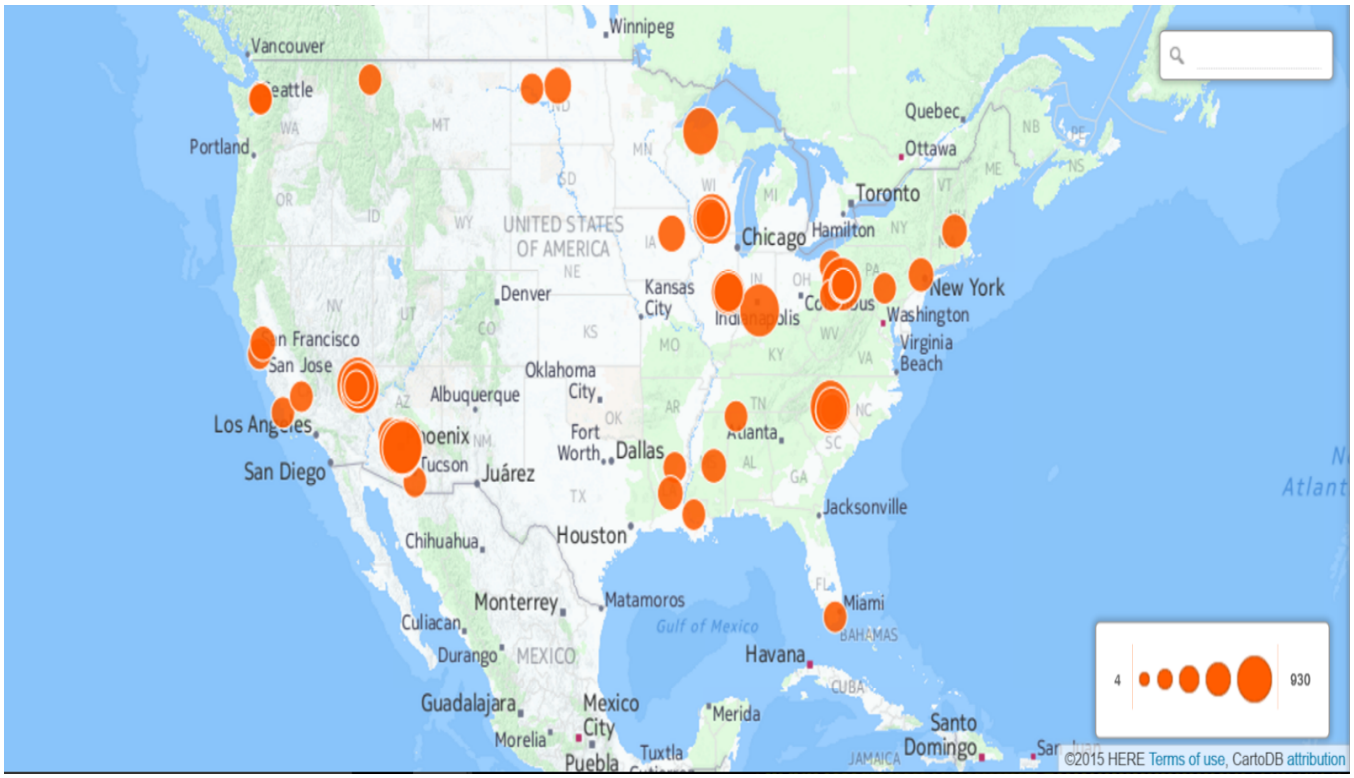
8.2 Naive Bayes classification

Naive Bayes classifier is a model which classifies given data set into various classes, characterized by feature vectors. It makes use of assumption that the value of a particular feature is independent of the value of any other feature. We use this model for binary classification of restaurants into Closed and Open. The results obtained would show how strong a set of features are to determine probability of restaurant shutdown.

We used following three kinds of Naive Bayesian classification models:



(a) Open Restaurants

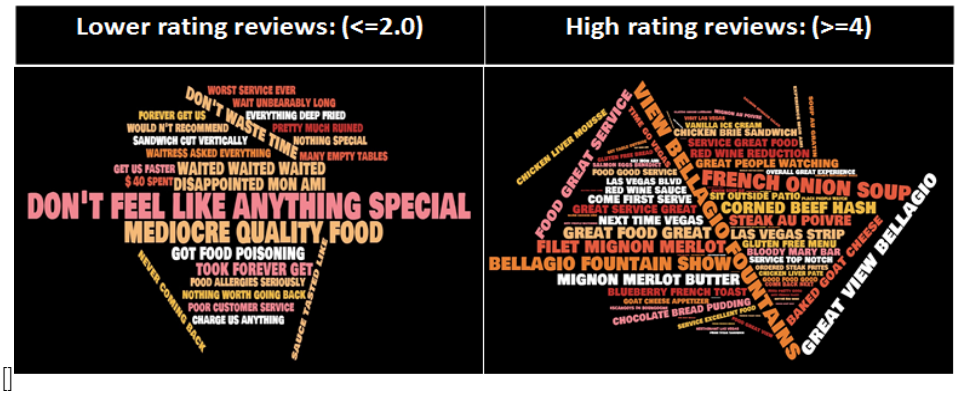


(b) Closed Restaurants

Figure 7:Location based analysis of restaurant closure

	Common Patterns in higher rating	Common Patterns in lower rating
1	Great Food	Wrong order served.
2	Great BBQ	Delay in serving food or not delivered the order at all.
3	Super friendly staff	Fast food chains are having maximum complaints.
4	Service top notch	Poor food quality, cold Food.
5	Good food in great price	Got food poising.
6	Food always fresh	Staff doesn't seem care.
7	Fresh Air	Phone ringing but nobody receiving call.

*[]



	Common Patterns in higher rating	Common Patterns in lower rating
1	View Bellagio fountain across the street	Longer wait time and empty tables
2	French onion soup	Medicare food quality lead to food poisoning
3	Baked goat cheese	Poor Customer Service
4	Filet mignon merlot	Charges are high
5	Corned Beef Hash	Everything is deep fried

*[]

Figure 8(a)General Trends(Top figure) (b) Restaurant Specific Trends(Bottom figure):Textual analysis of review data

8.2.1 *Bernoulli Naive Bayes classification*

In Bernoulli Naive Bayes classification, features are considered as independent Boolean variable. In this method of classification, training and classification data is assumed to be distributed as per multivariate Bernoulli distributions. This method requires samples to be represented in format of binary valued feature vectors. Often the first step in implementing above classifier is to convert data in binary format.

8.2.2 *Gaussian Naive Bayes classification*

This classifier works by assuming that data being dealt with is continuous and hence it can be distributed using Gaussian distribution. Then,$P(X = V|C)$ where p is the probability distribution, can be computed by plugging "V" into the equation for a Normal distribution and calculating values of respective values of probability.

8.2.3 *Multinomial Naive Bayes classification*

This classifier makes use of naive Bayes algorithm for multinomially distributed data.

8.3 K-Nearest Neighbor classification

Using k=5 closest training points based on features selected, this algorithm predicts the status of a restaurant to be the one with majority votes among k-nearest neighbors and assigns it a value obtained by weighted average of these.

8.4 Random forest classification

In Random Forest, we grow multiple trees to classify a new object based on attributes.Each tree gives a classification and we say the tree **votes** for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.

8.5 Support Vector Machine classifier

Support vector machines are supervised learning models which analyze data and recognize patterns. It separates data using some optimal method and perform classification and regression analysis SVM is representation of points in space mapped to set of feature vectors and divided in categories by widest gap possible. Testing sets are predicted to which group they belong

8.6 Cross Validation Accuracy

Cross Validation is technique to benchmark the reliability of results of Machine learning techniques. It is used as model validation technique and judge the accuracy of prediction model. It involves partitioning of data in multiple rounds such that every data set is used both as testing as well as training set. Multiple rounds of cross validation are performed using different partition sets and net result is average over all rounds. It is helpful in evaluating the model prediction performance.

8.7 Receiver Operating Characteristics

ROC curve represents performance of classifiers in which True positive rate (TPR) is plotted against false positive rate (FPR). It helps in judging the right models for data set presented. A good ROC curve characteristic is more near to

left border than the top border. The area under the curve is measure of accuracy. The closer the area is to value 1, the stronger is the prediction model.

9. RESULTS AND FINDINGS

The accuracy of various Machine learning models and their respective ROC is presented in table below:

Table 2: Accuracy of various classifier tested

Classifier	Accuracy	ROC
Logistic Regression	80.42%	0.61
Bernoulli Classifier	78.27%	0.52
Gaussian Classifier	74.65%	0.56
Multinomial Classifier	48.97%	0.48
K-NN Classifier	77.48%	0.58
Random Forest Classifier	78.49%	0.58
SVM Classifier	80.57%	0.57

Above results show that all classifiers except Multinomial Naive Bayes give accuracy not less than 75% which is satisfactory. However, for all these classifiers the area under ROC curve is far less than 1. Area under the ROC curve (AROC) is a single number summary of performance. A low AROC indicates that the predictive power of classifier is weak. Figure 9-15 presents ROC Curves for all classifiers ,we have used for prediction. It is interesting to note that although the accuracy of most of classifiers is good, their AROC is not acceptable. Possible reasons could be:

- Number of restaurants which are a part of yelp rounds to 22000. Our dataset under examination is restricted only to these many restaurants which is not a very good figure for data analysis.
- Number of restaurants open is roughly 4 times the restaurants closed. This huge difference leads to less randomization in dataset which can lead to biased results.

Cross Validation Results are calculated over SVM. We applied fivefold cross validation over SVM and calculated the respective accuracy. Net accuracy is average of all accuracies obtained across folds.

Table 3: Multifold accuracy

Folds	Accuracy
1	80.32%
2	80.17%
3	79.65%
4	80.84%
5	80.01%

The overall accuracy is 80.20% which is a good sign to be sure about our results.

We also did a location-based analysis for closed restaurants. However, the results don't give a very strong correlation with

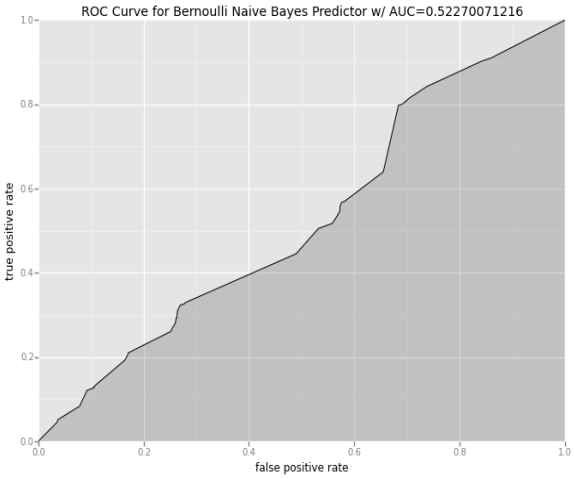


Figure 9:ROC-Bernoulli Classifier

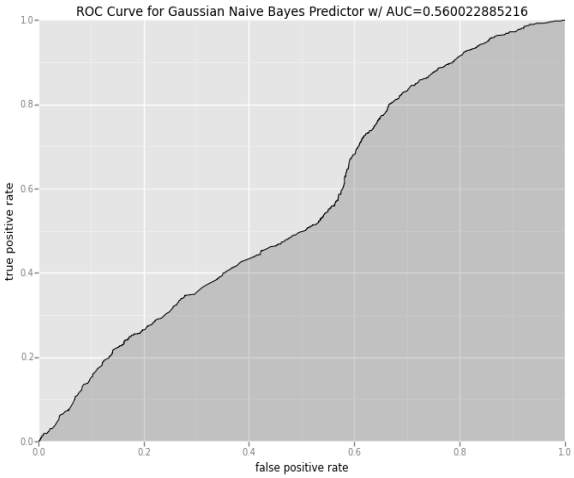


Figure 10:ROC-Gaussian Classifier

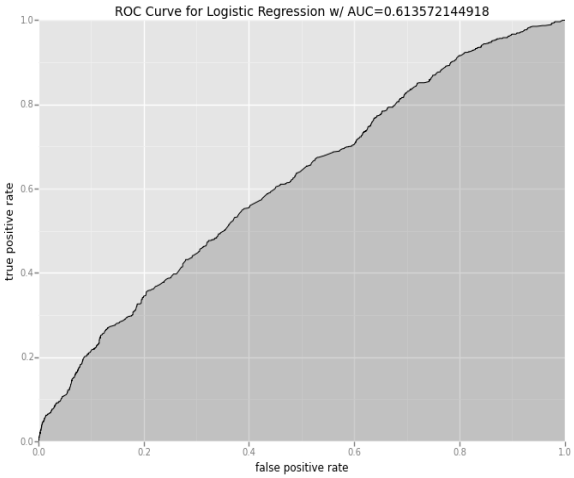


Figure 11:ROC-Logistic Regression

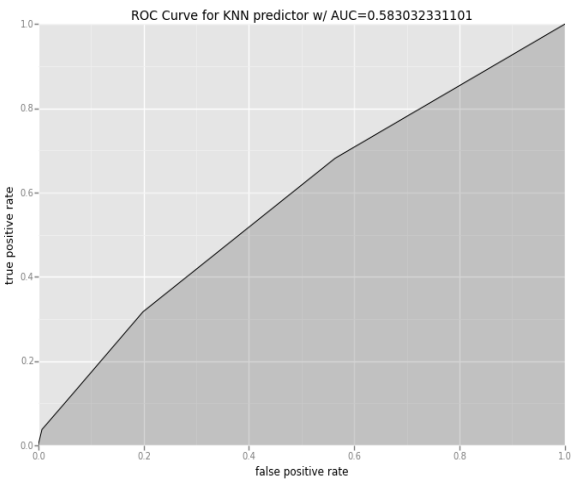


Figure 12ROC-KNN Classifier

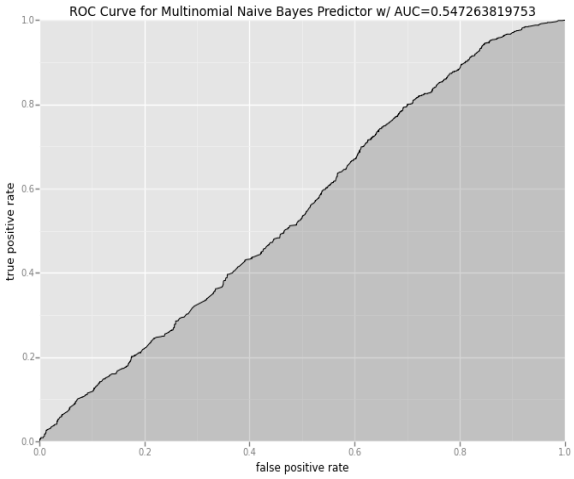


Figure 13:ROC-MNB Classifier

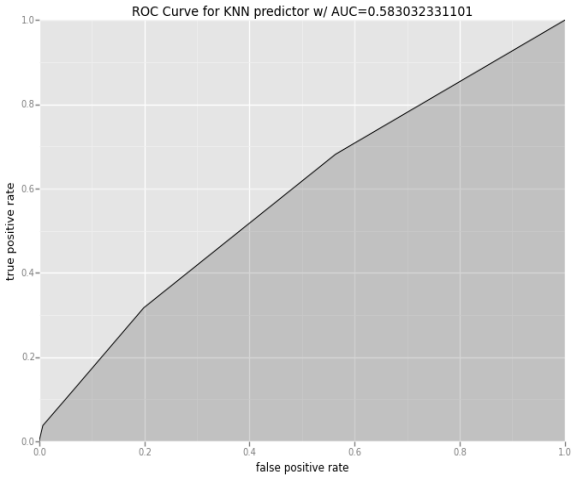


Figure 14:ROC-Random Forest

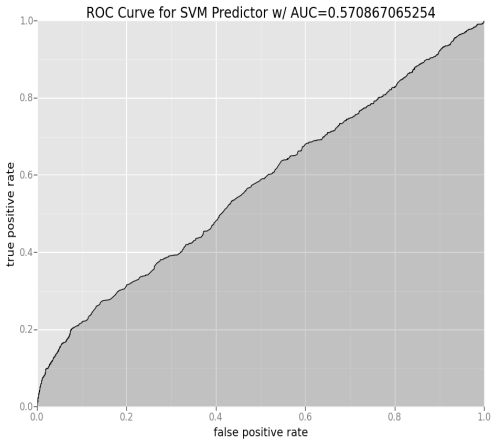


Figure 15:ROC-SVM

location, it can be said that urban areas have more number of successful restaurants than non-urban areas.

The time-series analysis described above shows that review count and star ratings can be used to predict closure year of a restaurant with an accuracy as good as atleast 75%.

10. CONCLUSION

- Based on restaurant data, it is possible to predict restaurant closure with a high level of accuracy. As we have shown nearly 80% of prediction accuracy can be achieved using different machine learning models.
- Various methods, like time series , location based analysis , review data analysis can be converged to get reliable prediction about restaurant closure.
- Our studies are relevant in context of all other business units. Applied techniques can be generalized to be applicable for any business unit health prediction
- This study suggests that yelp data set can be used to predict the business health. We have used yelp data to just get the business state (open or close) in future but also the duration remaining before closing.
- We have also showed that user reviews can be used for predicting most important characteristics from user point of view. We gathered general attributes of a business as well as the important observations for a single business unit in its good and bad reviews.

11. FUTURE WORK

- A larger feature set consisting of composite feature set can be used for better prediction
- Multiple types of business data like reviews, credibility of user providing feedback can be taken into account to get holistic set of features.
- Restaurant closure can be predicted location wise in order to determine local factors affecting restaurant closure.
- Impact of user review can be measured using sentimental analysis to know how it correlates with restaurant closure.
- Closure prediction engine can be converted to feedback engine which will inform restaurants if its feature set has started to lie in closed restaurants domain.
- Receiver Operating Characteristic can be improved by providing bigger set of data.

12. REFERENCES

[1] $http://www.restaurant.org/News-Research/Research/Facts-at-a-Glance$
[2] $http://pitchforksoptional.com/wp-content/uploads/2011/12/restaurantsfail.pdf$
[3] $http://www.bizjournals.com/washington/blog/top-shelf/2014/12/umd-researchers-use-yelp-reviews-to-predict.html?page=all$
[4] $https://dato.com/learn/gallery/notebooks/intro-regression.html$

[5] $http://aimotion.blogspot.com/2011/11/machine-learning-with-python-logistic.html$

[6] $http://blog.yhathq.com/posts/logistic-regression-and-python.html$

[7] $http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/$

[8] $http://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/$

[9] $http://hunch.net/?p=21/$

[10] $http://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/$

[11] $https://www.yelp.com/academic_dataset/$

[12] $https://cs.uwaterloo.ca/\ nasghar/886.pdf$

[13] $http://www.acm.org/publications/article-templates/sig-alternate-sample.pdf$