Are you Open?

Narenkumar Pandian(npandian@stanford.edu), Vaibhav Aggarwal (vaibhavg@stanford.edu)

Abstract

In this paper we propose a new machine learning approach to predicting whether a business is open or permanently shutdown. This problem is of keen importance because with the dynamic world the information available online becomes stale very fast. The impact of serving stale listing is lost trust. On the other hand it is cost prohibitive to manually verify every business listing for its freshness. Hence we propose using a learning algorithm to classify businesses as shutdown with high probability. We use the Yelp dataset for training and testing. We start with the standard techniques of Logistic Regression and SVM, and then propose our own algorithms EMSVM and EMLOG. Instead of using the traditional RMSE metric we use the more expression Precision-Recall metric to measure performance. This metric allowed us to prioritize one at the expense of the other to fit over use case. At the end we were able to predict shutdown businesses with 86% recall and 61% precision using EMSVM and 85% recall and 73% precision using EMLOG.

I. Introduction

Imagine a scenario where you had a tiring week and the weekend has arrived. You plan an outing with your friends or family and try to look for the perfect restaurant. You find just the place online, an upscale italian restaurant in a nice neighborhood you had always wanted to visit. You grab your coat, get into the car and reach that place; only to find that it has been shutdown. Stale online business listings is a common problem in today's dynamic world. For companies like Yelp, Google and Microsoft it is quite important to keep the business listings fresh. Customers tend to be quite vocal about incorrect business data [8]. At the same time it is extremely costly to manually inspect every business and check whether it is still open or closed every month. We believe that there is an opportunity here to use machine learning to predict which businesses might be shutdown by looking at its various features. Typically indicators like bad reviews, lack of activity over time etc. point to a business going out of business.

The input to our algorithm is business star rating, reviews, checkins, tips and location. We use SVM, Logistic Regression and EM Clustering algorithms to do binary classification. The output of our algorithm is whether a business is 'open' or 'shutdown'. We then go further ahead to use ensembling[10] of these algorithms to develop our own algorithms with better recall. We use Yelp dataset for training and testing our model. We think that this research will help web companies like Yelp, Zomato, Google etc. to find businesses which are shutdown with high enough probability and keep the listings fresh.

2. Related Work

Previous research ranges from the predicting the future ratings of the business[2,3,4] or future popularity of the business [5] to customized predictions[6]. The interesting aspect of all the research is identifying the appropriate feature set [7] and the models that complement the feature set. Hood et.al[5] spent considerable amount of time in generating sets of features to use in their model. In their research, they found that of all the features, the temporal features like days between first and last review, date of the last review etc provided the best prediction results. The researchers used elaborative feature selection process using Univariate feature analysis and greedy feature removal to settle on a definitive feature set for their model.Carbon et.al[2] used spatial feautures along with the review text in developing their model. The project investigated the effect of location and other business attributes vs. quality of food and service (measured indirectly through review sentiment) on the business' rating. Past cs229 project [3] showed how a user's evaluation depends on the surrounding factors and context it is within. In our investigation we decided to use temporal, spatial and assessment features for building the predictive model. Sawant, Pai[4] implemented singular value decomposition, hybrid cascade of K-nearest clustering, weighted bi-partitle graph projection to build a recommendation system based on the yelp data.The problem we are trying to solve is different from the earlier researches, our model to predict whether the business is closed or opened encompasses both the popularity of the business and the influence of environmental/locality factors on the business.In our initial analysis of using few features we realized the importance of the need of a diverse feature set to build a reliable prediction model. The feature set we selected is indicative of quality of the business, local environmental factors and changes to the business over the period of time. Most of the earlier researches [2,3,4] used RMSE or error ratios as the basis for evaluating different machine learning models. In our project we used precision and recall as the reliable indicators to evaluate the models. We build the prediction model based on SVM and logistic regression. To improve on the precision and recall we are proposing a novel ensemble of EM and logistic regression/SVM models.

3. Dataset And Features

3.1 Dataset

We used the Yelp dataset for this problem [1]. The dataset has rich information on businesses related to certain localities. It contains 61,184 businesses, 1.6M reviews and 500,000 tips by 366,000 users. It also contains 481,000 business attributes like ratings, hours, city, category etc. One of the attributes in the dataset is business state which we plan to use as $y^{(i)}$ labels. Out of all businesses 53,725 are open and 7,459 are shutdown.

For this project we combined the business, users, review and tip dataset and picked diverse feature set for developing the machine learning model. The structure of the different datasets was available in individual json files. We co-related the data in the files, joined, denormalized and stored the data in the mongo database. The database was used as the central repository for initial data analysis, data pre-processing and feature extraction.For this project we used

3.2 Feature Set

Assessment Features	**Spatial Features**	**Temporal Features**	**Vocabulary Features**
- Number of reviews for a business. - Average rating for the business. - Number of user checkins.	- Latitude. - Longitude. - State.	- Number of days since last review. - Number of days since last tip.	- Tokenized review text.

3.3 Skew

We observed that the data was heavily skewed towards open businesses. This was impacting training adversely and most algorithms were converging to labeling everything as negative to reach high accuracy. In order to remove the data skewness we randomly shuffled the data related to business open and picked equal number of positive and negative examples.

4. Methods

4.1 Naive Bayes

Naive Bayes is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions between the features.

$$p(C_k|x_1,\dots,x_n) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

4.2 Logistic Regression

Logistic regression measures the relationship between the state of business and the feature set by estimating probabilities using the following logistic function and update equation [11].

$$g(z) = \frac{1}{1+e^{-z}} \qquad \theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta(x^{(i)})\right)x_j^{(i)}$$

4.3 SVM

In the next step we trained an SVM classifier with the basic features. SVM optimizes for best fitting of a hyperplane in high dimension feature space using the following minimization objective.

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i$$

$$\text{subject to} \quad y^{(i)}(\mathbf{w}^T\phi(x^{(i)}) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

with Gaussian kernel:

$$K(x^{(i)}, x^{(j)}) = e^{(-\gamma \| x^{(i)} - x^{(j)} \|^2)}$$

4.4 EM Algorithm

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely. Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step.

(M-step) Update the parameters:

$$\begin{aligned}\phi_j &:= \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}, \\ \mu_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}, \\ \Sigma_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}\end{aligned}$$

(E-step) For each i, j, set

$$w_j^{(i)} := p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma)$$

4.5 EMLOG (New Algorithm)

This technique ensembles Logistic Regression with EM algorithm. D_M here is the Mahalanobis distance of the test vector from the nearest gaussian probability distribution as generated by EM algorithm. It is used to adjust the output of logistic regression by ɸ when the distance is less than T. The final classification output is generated using the standard logistic regression classification technique.

$$D_M(\underline{x}) = \sqrt{(\underline{x} - \underline{\mu})^T S^{-1} (\underline{x} - \underline{\mu})}.$$

$$f(x^{(i)}) = g(x^{(i)}) + ɸ * 1\{D_M(x^{(i)}) < T\}$$

$$y^{(i)} = 1\{ f(x^{(i)}) > 0.5 \}$$

4.6 EMSVM (New Algorithm)

This technique ensembles Logistic Regression with EM algorithm. It uses a very similar approach to EMLOG except that instead of using $g(x^{(i)})$ from logistic regression it uses scores from SVM.

5. Results

5.1 Error Metric

5.1.1 Precision

Precision measures the ratio of number of actual positive samples with correct classification to the number of samples classified positive. $y^{h(i)}$ is the classification and $y^{(i)}$ is the actual label.

$$\text{Precision} = \sum 1\{y^{h(i)} = 1, y^{(i)} = 1\} / \sum 1\{y^{h(i)} = 1\}$$

5.1.2 Recall

Recall measures the ratio of number of actual positive samples with correct classification to the total number of positive samples.

$$\text{Recall} = \sum 1\{y^{h(i)} = 1, y^{(i)} = 1\} / \sum 1\{y^{(i)} = 1\}$$

5.1.3 FMeasure

Fmeasure is the weighted average of precision and recall. It is used to generate a combined error metric to track a single error. We use a beta = 0.5 for increasing the weight of recall. We want to have higher recall at a reasonable precision in order to identify as many businesses at risk as possible.

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

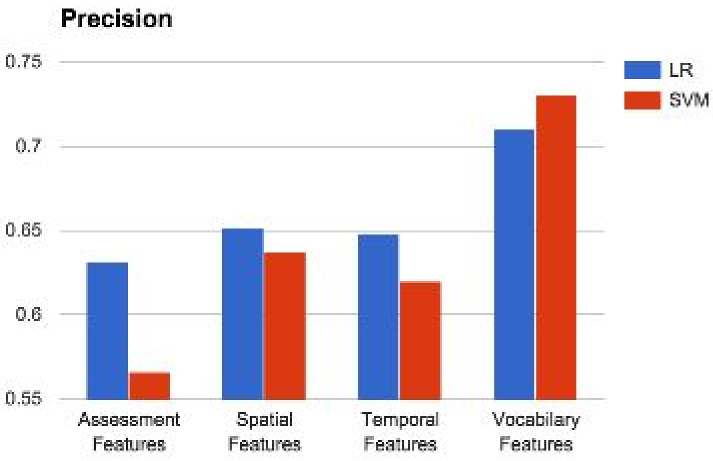
5.1 Baseline

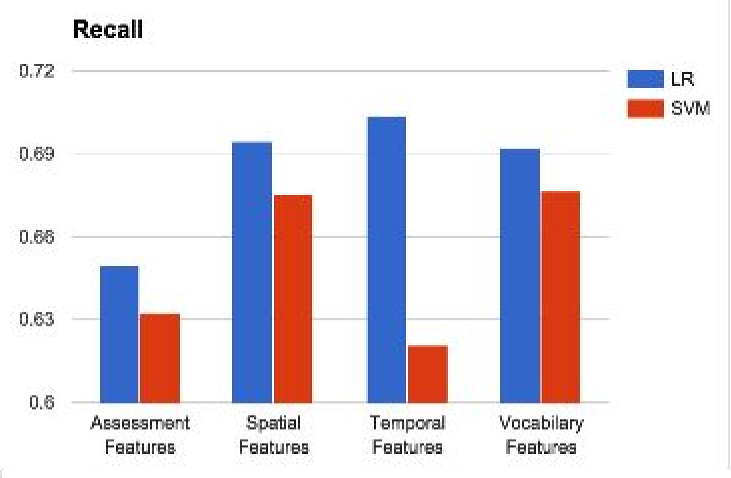
We first trained a simple Naive Bayes classifier using basic features of ‘Business Rating’ and ‘Review count’ to check the linear separability of the labels using these features. This was our attempt to create a simple baseline which we can improve over time. We tried k-fold cross validation technique with both raw data and unskewed data set. Our baseline is 12% precision and 9% recall on raw dataset and 48% precision and 56% recall on unskewed dataset.

Scenarios	**Precision**	**Recall**
Raw data	0.057143	0.063291
Unskewed data	0.49819	0.7759
Raw data, k-fold cross validation, k=3	0.12445	0.092589
Unskewed data, k-fold cross validation, k=3	0.48662	0.5698

5.2 Feature Improvement

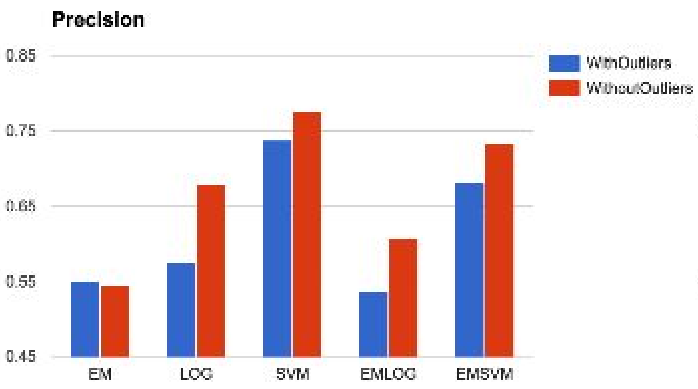
After generating the baseline we looked at the training and testing errors. Both the training and testing errors were very high which implied a bias problem. In order to reduce bias in our estimator we added more features to our dataset. This had a very positive effect our our precision and recall. The following figure shows how the precision improved from approx. 50% to 73% by using spatial, temporal and vocabulary features. The overall recall also improved from 60% to 68%. This analysis proves that the range of features we generated had incremental effect of improvement on the precision recall.

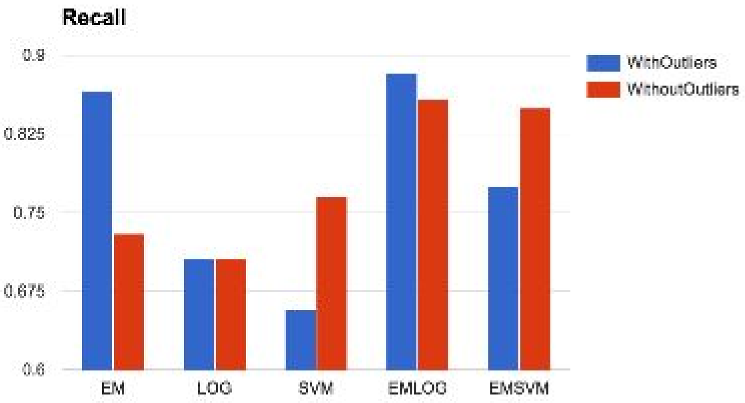




5.3 Data Distribution Impact on Learning

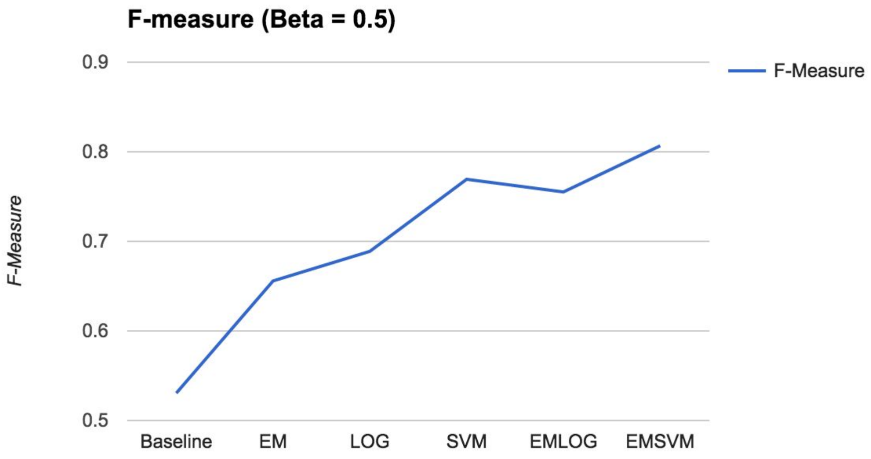
We also studied the performance improvements under different data distributions. During our initial data analysis we observed that lot of business with high star rating (4,5) that were closed were because of renovation or moving to another location etc. This confirmed the basic intuition on the dataset that Yelp does not capture financial or the macro economic picture regarding a business. But yelp provides information on the quality of the business and possibly influence of local environmental factors on the business. Hence we classified all closed businesses in our dataset with very positive ratings as outliers and tested both with and without outliers. We in general observed that the classification performance was 10% better without outliers for all algorithms.





5.4 Algorithmic Improvements

Finally we compare all algorithms performance using F-measure which takes weighted harmonic average of precision recall. We notice that EMSVM algorithm has the best performance of 81% overall. EMLOG and SVM also perform nicely with 78% and 76% score respectively. We prove that our novel approach of ensembling EM and SVM algorithms into EMSVM has higher payoff than those algorithms individually.



6. Conclusion

In this paper we researched multiple machine learning algorithms to identify shutdown businesses. We studied the impact of different classes of features on performance. We found that spatial and temporal features gave a really good boost to accuracy of our algorithms. We also studied the impact of different data distribution on performance and showed that removing outliers improves performance. Out of all the various algorithms we tested, our own novel machine learning algorithm EMSVM have the best overall performance with 85% recall at 73.5% precision. We optimized for high recall with reasonably high precision. This enables the web companies to identify most of the businesses at high risk of closure.

VI. References

[1] Yelp Dataset Challenge, http://www.yelp.com/dataset_challenge.
[2] Kyle Carbon,Kacyn Fujii,Prasanth Veerina., Applications of Machine Learning to Predict Yelp Ratings http://cs229.stanford.edu/proj2014/Kyle%20Carbon,%20Kacyn%20Fujii,%20Prasanth%20Veerina,%20Applications%20Of%20Machine%20Learning%20To%20Predict%20Yelp%20Ratings.pdf
[3] Jeff Han,Justin Kuang,Derek Lim., Predicting Yelp Ratings From Business and User Characteristics

http://cs229.stanford.edu/proj2014/Jeff%20Han,%20Justin%20Kuang,%20Derek%20Lim,%20Predicting%20Yelp%20Ratings%20From%20Business%20and%20User%20Characteristics.pdf
[4] Sumedh Sawant,Gina Pai ., Yelp Food Recommendation system
http://cs229.stanford.edu/proj2013/SawantPai-YelpFoodRecommendationSystem.pdf
[5] Bryan Hood, Victor Hwang and Jennifer King., Inferring Future Business Attention
http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_InferringFuture.pdf
[6] Alexis Weill, Thomas Palomares,Arnaud Guille., Yelp Personalized Reviews
http://cs229.stanford.edu/proj2014/Alexis%20Weill,%20Thomas%20Palomares,%20Arnaud%20Guille,%20Yelp%20Personalized%20Reviews.pdf
[7] Pedro Domingos.,A Few Useful Things to Know about Machine Learning
https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
[8] Incorrect Business Address Listing http://www.yelp.com/topic/denton-incorrect-business-address-listing
[9] Mahanalobis Distance https://en.wikipedia.org/wiki/Mahalanobis_distance
[10] Ensemble learning https://en.wikipedia.org/wiki/Ensemble_learning
[11] Logistic Regression http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf