Recommendation System for Restaurants

Anusha Jayasimhan
Fr.C.R.C.E
Bandra
Mumbai

Parikshith Rai
Fr.C.R.C.E
Bandra
Mumbai

Yash Parekh
Fr.C.R.C.E
Bandra
Mumbai

Ojas Patwardhan
Fr.C.R.C.E
Bandra
Mumbai

ABSTRACT
It is often perplexing for a person to decide which restaurant he must visit from a huge range of available options. There have been numerous suggestion frameworks accessible for issues like shopping, online video excitement, recreations, and so forth. Eateries and Dining is one territory where there is a major chance to prescribe feasting choices to clients in light of their inclinations and in addition recorded information. By developing a recommendation system which could help a user to decide which restaurant one should visit, the person can save a lot of his time, efforts and money and thus have a great experience and satisfaction. There are various factors based on which a user makes a decision of visiting a restaurant like the type of cuisine of the restaurant, the location of the restaurant, the ambiance, price range, popularity, ratings, etc. Such information is collected and made available on sites such as Yelp and Zomato. Using well rounded, open source dataset provided by Yelp which provides data not only of the restaurant reviews, but also user-level information on their preferred restaurants the aim is to build an efficient recommendation system for the Yelp users in the form of a software application and thus help them predict whether they will like visiting a restaurant or not by applying machine learning techniques and algorithms.

General Terms
Recommendation systems, Machine learning, Classification

Keywords
Recommendation system, SVM, Yelp dataset, feature selection

1. INTRODUCTION
Nearby business survey and review sites, for example, Yelp and Urbanspoon are an exceptionally prominent goal for a large number of individuals for choosing their eat-outs. Being able to prescribe neighborhood organizations to clients is a usefulness that would be an exceptionally significant expansion to these destinations usefulness.

There is a huge social impact of having a recommendation system for restaurants as it will save a lot of time and money because there is no longer a need to go through numerous web pages or web profiles of the business. The decision making is not only based on the restaurant attributes like facilities and services provided by the restaurants, the location of the restaurant, the quality of food, the popularity of the restaurant, the ambiance of the restaurant,etc but also based on the user preferences. Further it is also important to note that 40% of world population has an Internet facility today compared to 1% in 1995. In this work recommendations are only provided on the restaurant businesses from the Yelp dataset using a linear Support Vector Machine.

The suggestions are presented such that every client will be prescribed eateries according to his inclinations in eateries and the general ubiquity of the eatery. This is done in his territory along with all the restaurant, user and derived features with a discrete yes or no answer of whether he ought to visit the eatery or not.

2. DATA COLLECTION AND PREPROCESSING
The dataset is collected from the Yelp dataset challenge. The data is available in JSON format and is converted into CSV format for the convenience of visualization and better understanding [1]. The size of the data is approximately around 2.6 GB. The data is split into historical, training and testing data as shown in Fig.1. The historical data comprises of derived features like the average business category rating of a particular user to determine the preferences of the user. Yelp clients give evaluations on a 5-point scale, which are mapped to a double yes(4,5)/no(1,2,3) class label. Henceforth, every case in the presented information is a solitary audit with a parallel binary yes/no (1/0) class label [2].

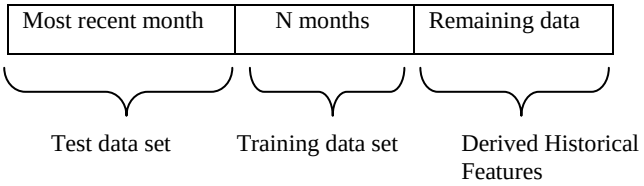

Fig.1: Dataset segregation

Only those tuples in the business data where the category consists of "Restaurant" in its substring have been considered. Also to refine the review data, only the tuples of those users who have a review of more than 20 restaurants have been considered [3]. By doing this, it is possible to understand more about the preferences of the user and build and efficient prediction model according to the users preferences in the historical data The refined business data combined with refined review data was next preprocessed. The dataset to be preprocessed contained features of varied data types which were to be converted to binary data types and normalized in order to improve the performance the SVM algorithm.

Firstly the combined business review file was used and all the features which had boolean true and false values were converted to binary values of 1 and 0. For those features having multiple values, new columns were created consisting of all the possible values of that feature. The values for these columns were filled according to the corresponding value of each tuple in the dataset. For instance, as a first step, the number of distinct categories were identified which turned out to be 309 and hence 309 distinct features were created. In order to extract distinct values, a python dictionary was created and iteratively compared with the values in the dictionary to the values. Example, Mexican category was one of the distinct features we obtained from the category value

mexican. All these distinct features were given the binary values 1 or 0 depending upon whether the corresponding restaurant consisted of that category or not.

The missing values in the data are filled according to the probability that a particular value will occur for that tuple. Similarly, for attribute Alcohol feature in the data the possible values were "full bar", "none", "beer" and "wine". Hence 3 new columns were created which contained binary values for each tuple in the data. The data was then sorted according to the date with the most recent data on the top and this data was stored into a csv file. As the next step, the training, testing and historical data were formed in separate files. The historical data was used to generate the derived categories which included average user rating for each category of the restaurant and this derived feature was generated for each tuple for the training and testing data.

3. IMPLEMENTATION

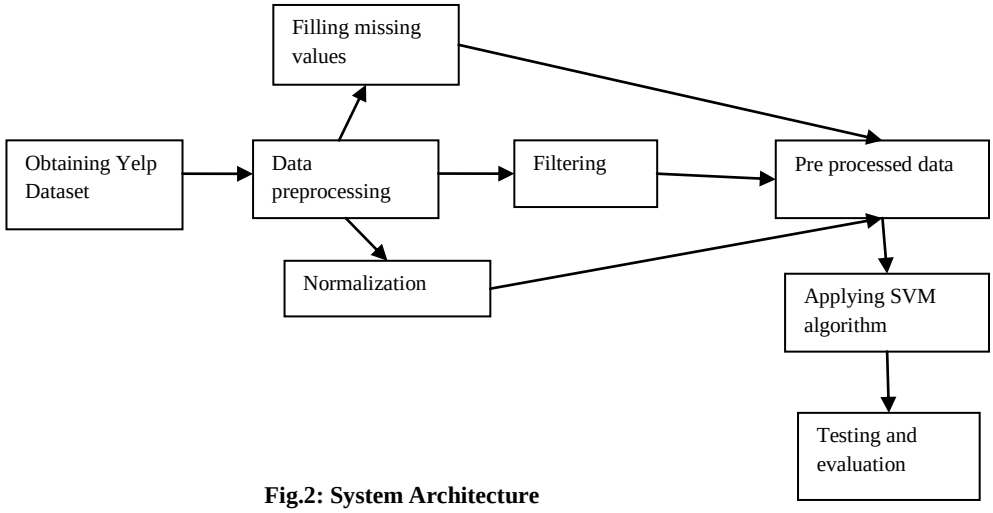

Fig.2: System Architecture

The dataset used was obtained from the yelp website through the Yelp dataset challenge. It consisted of multiple csv files. Segregation of features needed to be done from these files and hence data preprocessing consisted of mainly the following steps as also shown in Fig.2 which represents the system architecture

- Filling missing values
- Normalization
- Filtering

Features with low variance i.e. the features with variance below the threshold were rejected. Univariate feature selection works by selecting the best features based on univariate statistical tests. During the filtering process, only the tuples having "Restaurant" as the value of the category attribute, or as a substring were selected. Also, only the users having more than 20 reviews were considered. After extracting the derived features like restaurant, b&b, breakfast from categories attribute, the corresponding tuples were checked for the derived features obtained. If the categories feature of a tuple had a particular value of the distinct feature, it was given a binary 1 else it was given a binary 0. This was done for normalization of data. The missing values were filled by calculating the probability of a particular word and then assigning that value. The data was separated into training and testing sets with 9,241 & 8,476 tuples respectively.

Removing low variance features: Variance Threshold is a straightforward way to deal with highlight choice. It expels all components whose change doesn't meet some limit. By default, it expels each of the zero-change highlights, i.e. highlights that have a similar incentive in all examples.

Univariate feature selection: Univariate feature selection works by selecting the best features based on univariate statistical tests. It can be seen as a preprocessing step to an estimator.

Feature scaling: It can be viewed as a preprocessing venture to an estimator. Feature scaled information is scaled to a settled range - for the most part, 0 to 1. It suppresses the standard deviations and thus minimize the outliers Min-Max scaling equation A Min-Max scaling is typically done as shown in equation (1):

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Filling Missing values and categorizing data: The missing values were handled by taking the probability of the occurrence of that particular value of the feature. Example, if a feature had 3 possible values the probability would be 1/3 which is 0.33. The feature attribute alcohol had 3 possible values viz. full bar, none and beer and wine. If the value of this attribute was missing for a tuple, 0.33 was used as the default value for all the columns full bar, none and beer and wine. In case of Boolean values True/ False, 0.5 was used as the default value to fill missing values. 3 algorithms namely

Linear SVM, SVM with rbf Kernel and decision tree algorithm and the performance of each algorithm is evaluated and compared.

3.1 Results

3.3.1 Linear SVM

A Support Vector Machine (SVM) is a discriminative classifier defined by a decision boundary or a separating hyperplane. Given labeled training data (supervised learning) the algorithm outputs an optimal hyper-plane which categorizes new tuples. Linear SVM is a quick machine learning (data mining) calculation for taking care of multiclass grouping issues from ultra-vast informational collections that execute a unique exclusive adaptation of a cutting plane calculation for classifying tuples into class labels The planar boundary is essentially linear. The examinations with other known SVM models plainly demonstrate its unrivaled execution when high accuracy is required. The most important features of linear SVM are as follows:

- Efficiency in dealing with extra large data sets.
- Can perform classification for multiclass category problems
- Prevents overfitting

Table 1: Linear SVM on Yelp Dataset

Accuracy	Precision	Recall	f1-score
68.99	67.89	96.55	80.59

3.3.2 SVM with rbf kernel

SVM with rbf kernel also draws a decision boundary through the data points but it does so using nonlinear but normal curves through the data points. The curve function is given by an a radial basis function. This algorithm tries to fit the data around the decision boundary so that it improves the classification accuracy by constructing a non linear boundary.

Table 2: SVM with rbf kernel on Yelp Dataset

Accuracy	Precision	Recall	f1-score
67.80	73.23	99.39	80.46

3.3.3 Decision tree

Decision tree is one of the supervised learning algorithms (having a pre-defined class variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this algorithmic technique, the population or sample is split into multiple homogeneous sets (sub-populations) based on most significant differentiator in input features. The below results are obtained by applying ID3 algorithm.

Table 3: ID3 algorithm on Yelp Dataset

Accuracy	Precision	Recall	f1-score
61.46	56.37	71.78	71.31

3.3.4 Comparison of Results

Table 4: Results with training set: 9241 and testing data: 8476

	Linear SVM	SVM with rbf kernel	Decision tree
Accuracy	68.99	67.80	61.46
Precision	67.89	73.23	56.37
Recall	96.55	99.39	71.78
F1 score	80.59	80.46	71.31

Table 5: Results with training set: 11741 and testing data: 3000

	Linear SVM	SVM with rbf kernel	Decision tree
Accuracy	68.93	68.60	60.93
Precision	67.74	70.24	56.33
Recall	96.39	98.09	70.02
F1 score	80.51	80.62	70.47

4. CONCLUSION AND FUTURE WORK

SVM algorithm performs best in case of a large dataset by giving us the best accuracy and performance in comparison to the other algorithms that were compared for the Yelp academic dataset. The reason for this is that SVM doesn't much have to deal overfitting issues. When the size of the data was increased it was found that SVM with RBF kernel performs better than linear SVM algorithm. In the case of decision tree algorithm the model becomes a victim of overfitting and fails to provide good results for high dimensional dataset. Thus, the recommendation system uses SVM at its core.

For the future, collaborative filtering can be used to further improve the predictions. It can also involve trying to identify stronger features beyond what is available in the datasets, as well as investing in an approach to gather training and evaluation data from alternate means. Further the training data set may also be increased to train the model better and provide better results in terms of accuracy. Future work could also include attempting to distinguish more grounded components past what is accessible in the datasets, and in addition contributing in a way to deal with accumulated preparing and assessment of information. The recommendation system need not be limited to restaurants but can be extended for other systems and businesses as well.

5. REFERENCES

[1] Yelp, "www.yelp.com/academicdataset," *Yelp academic dataset*, 2016.

[2] A. Gandhe, "Restaurant recommendation system," *cs229.stanford.edu*, 2015.

[3] "A preference-based restaurant recommendation system for individuals and groups," *www.cs.cornell.edu*, 2013.