Predicting Restaurant Health using Yelp data and Government Inspections

Ajay Anand
USC ID:2020698090
University of Southern California
ajayanan@usc.edu

Devershi Purohit
USC ID: 9139405634
University of Southern California
dupurohi@usc.edu

Rajdeep Kaur
USC ID: 9558003248
University of Southern California
kaurr@usc.edu

Rimsha Goomer
USC ID: 3080801166
University of Southern California
goomer@usc.edu

ABSTRACT

This project report presents a comparison study of four supervised machine learning approaches for classifying a restaurant's health. Evaluating a restaurant is important to keep a check on how it is performing and what improvements might be needed. In this project, we start by creating a dataset by combining features from the yelp challenge dataset and the US Government health dataset available online for the city of Las Vegas, Nevada. We then fill in the missing data using web scraping and user-based collaborative filtering. To classy a restaurant as healthy or unhealthy, we use Support Vector Machines, Linear SVC, Decision Trees and Adaboost with decision stumps. The SVM Linear Kernel approach significantly outperforms the remaining three algorithms and achieves a sensitivity of 0.90 and an F1 score of 0.50 on our curated dataset.

1. INTRODUCTION

According to the Centers for Disease Control and Prevention, about one in six Americans (48 million people) get sick, 128,000 are hospitalized, and 3,000 dies of food-borne diseases.

Yelp is used by almost everyone in today's world for sharing their experiences as well as searching for any business (restaurants, hair salon etc) details. Health state of a restaurant is an important factor for customers, governments, and restaurants themselves. This project aims to predict if a restaurant is ‘Healthy’ or ‘Unhealthy’ based on the available features in yelp and US government health data.

Prediction of restaurant health state for each restaurant (based on available the feature data) can help the government to decide the risk of a restaurant. It will ensure public health and safety. Yelp can collaborate with restaurants to provide recommendations to improve their restaurants and user needs. This can provide mutual benefit to both yelp and restaurants.

2. DATASET

There are 3 sources for our dataset: yelp challenge dataset, restaurant inspection results carried out by the U.S government health department, and data extracted for the restaurant by web scraping from Yelp. We selected Las Vegas as the region for inspection.

The first source is yelp challenge dataset for Round 12, out of which this project uses business.json data required to uniquely identify restaurants with business id. Out of total 188,593 businesses, 28,853 businesses belonged to Las Vegas. We have used 48 features for restaurants from this dataset.

The second source is government inspection data for restaurants. For Las Vegas, we used the OpenData platform to get inspection results for Las Vegas carried out by the government from the year 2010-2018. This contained 165,477 rows with multiple entries present for the same restaurant, which represents that the government carried out inspections for a restaurant on multiple dates. We have used 14 features out of 23 available features for restaurant inspection from this dataset.

The third source is web scraping from Yelp. This was needed for two reasons, firstly to fill some percentage of missing feature data in dataset retrieved from Yelp challenge, and secondly to retrieve data for some new useful features for restaurants. The list of features is retrieved from web scraping.

3. METHOD

3.1 Algorithms Used

3.1.1 User-based Collaborative Filtering

It is a technique for determining patterns among various data-sources using similarity metrics like Pearson Correlation, Cosine Similarity.

3.1.2 Web Scraping

Web Scraping is a technique employed to automatically extract large amounts of data from websites given a unique identifier on which to scrap for.

3.1.3 Adaboost

Adaptive Boosting is a meta-learning algorithm which aims to build strong classifiers by combining a set of weak classifiers.

3.1.4 Decision Trees

A decision tree model creates a graph where the nodes are feature columns and edges are the values they can take, and uses to make informed predictions.

3.1.5 Linear SVC

LinearSVC is a support vector classifier with a linear kernel, but has more flexibility in the choice of penalties and loss functions and scales better to large numbers of samples.

3.1.6 SVM Linear Kernel

SVM model represents examples as points in space, such that separate classes are divided by a clear gap as wide as possible.

3.2 Experimental Approach

3.2.1 Data Preparation

We filtered out businesses in Las Vegas (28,853 rows) from all the businesses (188,593 rows) in the yelp_business_json. Another step of filtering was done to extract restaurants specific businesses from the state of Nevada (NV) and the city of Las Vegas (8808 rows). The feature 'attributes' was split into further 39 distinct features and each was represented as a unique integer. A similar preparation was done for the U.S. Government dataset which had inspection grades for 165,477 restaurants in Las Vegas. For various inspection grades for one restaurant at different activity dates, an average inspection grade was calculated. This new US Government dataset (24,338 rows) included 13 additional features along with inspection grades which is the ground truth for us.

We then calculated the common restaurants (614 rows) in both processed Yelp and processed US Government dataset and removed any duplicate entries (562 restaurants x 57 columns). For each restaurant, we scraped out more data values and more features to fill in missing data and expand the current data respectively (562 rows x 68 columns). A restaurant is considered 'Healthy' if it has a score between 90-100 or has an 'A' grade, otherwise, it is considered 'Unhealthy'.

3.2.2 Data Preprocessing

We used User-Based Collaborative Filtering for filling the missing feature values. To determine the similarity between two restaurants, we computed the Pearson correlation coefficient as a similarity metric using seven features like Reviews Count, Current Health Score, Rating etc. We filled the value of the missing features by taking the mode (most frequently occurring value) of that feature from similar users determined using the Pearson Correlation. To standardize the range of independent feature variables in the dataset, min-max feature scaling was used which normalized the data in the range of [0,1]. We split the final dataset into train, validation and test sets in the ratio of 60:20:20. Subsequently, using 3 scoring functions for classification namely chi-squared, f_classif and mutual_info_classif, we selected the best K features to be fed into the model as a dataset.

3.2.3 Machine Learning Task

For each machine learning model, we did the following standard steps -

1. Initialize the hyperparameters

2. Repeat:
    a. Train the model on the train set
    b. Evaluate the model on validation set
    c. Improve the model by fine-tuning the hyperparameters
3. Train the model again using the best parameters on combined train+validation dataset.
4. Make predictions on the test dataset using the final trained model and calculate Sensitivity, Specificity, F1 Score and ROC Curve

4. RESULTS AND DISCUSSIONS

Figure 1 below represents the percentage of missing data initially (53.46%), after web scraping (42.90%) and finally after user-based CF (2.70%). It can be seen that very less data was missing after applying user-based CF (only ~3%).

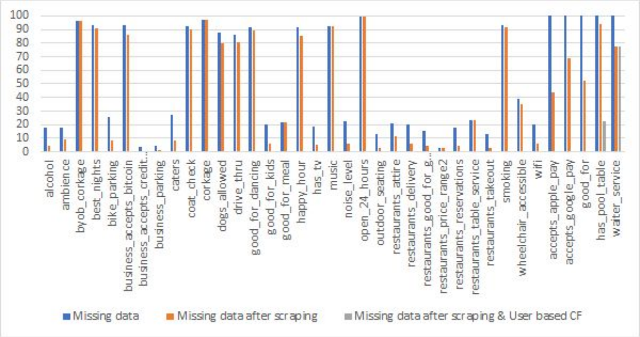

Figure 1. Percentage of Missing Data

The final dataset has 562 restaurants with 44 features, out of which 386 are 'Healthy' restaurants and 176 are 'Unhealthy' restaurants.

We performed SelectKBest features and hyperparameter tuning on each model to select the best features and parameters (Best Results in Table1) on held out train and validation data.

Finally, we ran each model on the held out test data by training on the larger training set (Train Data + Validation Data). Results of each model are in Table2. SVM Linear Kernel model gave the best performance in terms of sensitivity.

Table 1. [No. of features, Best Sensitivity] for SelectKBest

Scoring Function/ ML model	**Adaboost**	**Decision Tree**	**Linear SVC**	**SVM Linear Kernel**
chi2	[4, 0.6585]	[8, 0.6098]	[8, 0.4634]	[32, 1.0]
f_classif	[35, 0.6585]	[5, 0.5123]	[5, 0.4634]	[32, 1.0]
mutual_info _classif	[26, 0.6341]	[8, 0.5123]	[7, 0.4878]	[15, 1.0]

Table 2. Results after training and evaluating on held out test set

Results	**Adaboost**	**Decision Tree**	**Linear SVC**	**SVM Linear Kernel**
[sensitivity, f1score]	[0.4242, 0.475]	[0.5757, 0.5429]	[0.2727, 0.3462]	[0.9091, 0.5000]
Hyperparameters	max_depth = 2, n_estimator = 70, learning_rate = 1	random_state = 100	C: 10, max_iter: 1000	C : 0.01, gamma : 0.01, max_iter: 20

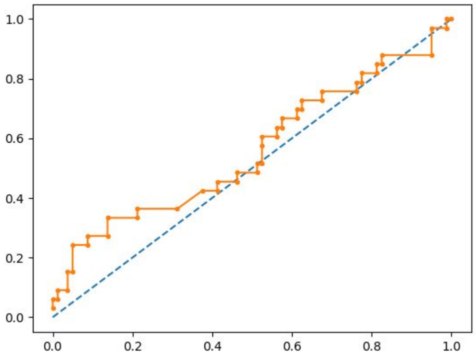

Figure 2. ROC Curve for SVM Linear Kernel

The scope of this project can be extended to any other city to predict health grade (Healthy/Unhealthy) of restaurants and solution to allocate limited government officials/resources for health inspection of restaurants. Further, it can include finding common 'words' and performing text analysis using various NLP techniques on the Yelp review dataset to determine what words may lead to Unhealthy restaurants in the region.

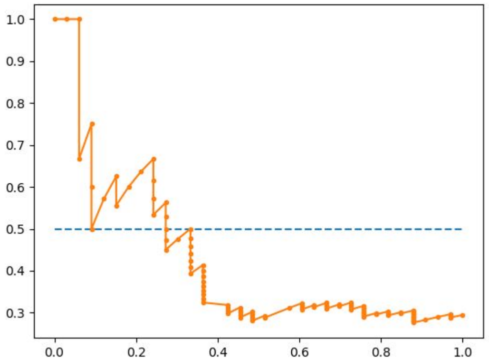

Figure 3. Recall-Precision Curve for SVM Linear Kernel

5. APPENDIX

5.1 Github Repository

The version management and source code were maintained through GIT. The source code and the preprocessed dataset used are available at https://github.com/devership16/INF553-YelpProject

5.2 Individual Contributions

5.2.1 Ajay Anand

Studied research papers on health grade, restaurant features, and public health. Worked on finding common unique restaurants among data gathered from multiple sources. Built yelp adapter to interact with Yelp server, web scraping to extract restaurant features, and fill missing data with scraping. Created metrics to evaluate the performance of all ML models. Designed SVM with linear kernel ML model. Used Los Angeles County as a secondary region, gathered data from multiple sources and prepared dataset.

5.2.2 Devershi Purohit

Built the AdaBoost machine learning model and optimized the hyper-parameters based on Sensitivity (Recall) and F1 Score. Implemented User-Based Collaborative Filtering algorithm using Pearson Correlation, for filling missing feature values for businesses. Wrote a python script for plotting and calculating Area Under the curve for ROC curve and Precision-Recall curve. Wrote a python script for extracting Restaurant specific entries from dataset and handling duplicate business entries.

5.2.3 Rajdeep Kaur

Pre-processed the government health inspection data for Las Vegas. Normalized the grades and counted the number of violations after my analysis of the data. Also, cleaned the data. Prepared data after feature selection using Chi-square. Wrote a python script to run Chi-square and recorded output for different alpha values (six values). Built machine learning model CART - Decision tree to run on the final prepared data and fine-tuned it both in terms of features and parameters to obtain the best f1 score.

5.2.4 Rimsha Goomer

Pre-processed the business.json. Extracted the potential features and defined a standard numeric range for each of them based on their unique values. Designed a schema for the same. Plotted graph between Health Grade v/s Number of Restaurants. Selected Features using the SelectKBest with a different scoring function. Utility file with functions of saving a model (pickle format) from main memory to disk after training and its converse to load the model (pickle format) from disk into main memory. Trained and fine-tuned a Linear SVC Kernel and tested it end-to-end.

6. REFERENCES

[1] Yu, Boya; Zhou, Jiaxu; Zhang, Yi; Cao, Yunong, "Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews," ARXIV, 2017.

[2] Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., Teitel, J., and Silenzio, V. (2017). Deploying nEmesis: Preventing foodborne illness by data mining social media. AI Magazine, 38(1), 37–48. DOI: 10.1609/aimag.v38i1.2711.

[3] J. P. Schomberg, O. L. Haimson, G. R. Hayes, and H. Anton-Culver, "Supplementing public health inspection via social media," PloS one, vol. 11, no. 3, p. e0152117, 2016.

[4] Morgan J. Classification and regression tree analysis. Technical Report No. 1.Boston University of Public Health; 2014.