BUSINESS PREDICTION WITH YELP'S HELP

A PREPRINT

Om Patel
Computing and Information
University of Pittsburgh
Pittsburgh, PA 15260
omp8@pitt.edu

Vamshi Bussa
Computing and Information
University of Pittsburgh
Pittsburgh, PA 15260
vab58@pitt.edu

Xuejiao Feng
Computing and Information
University of Pittsburgh
Pittsburgh, PA 15260
xuf9@pitt.edu

April 24, 2020

ABSTRACT

The paper intends to learn a trustworthy model for deciding the success factor of various kinds of small business such as restaurants, bars and sports goods shops in cities like New York, Pittsburgh and many more. This is done given we have a review data set of pre-existing business ratings in cities in the USA. Using algorithms like Light Gradient Boosted Machines we have found test accuracy as high as 70, but we are sure that with data having more predictors and using models like Neural Networks, we can get better scores and place more trust on the models for real life usage.

1 Hypothesis

In a human civilisation like a city, several aspects of business and human concentration distribution result in different success rate in certain area and different failure rates in certain areas for small businesses like restaurants, shops and bars.

2 Introduction

We took up a basic problem faced by many entrepreneurs who are willing to start various types of businesses like restaurants, bars, clubs, sports stores etc. at specific locations. They also hope that with time, their business will be a successful one/high rated. When they start looking at a specific area to set up, say a restaurant, the typical question that arises is "What kind of a restaurant can I set up in this area?" or "What specific stuff do daily users like/not like in this area". The risk in such businesses is certainly high and it helps to get a good solid answer to their questions and relieve the pressure on their heads.

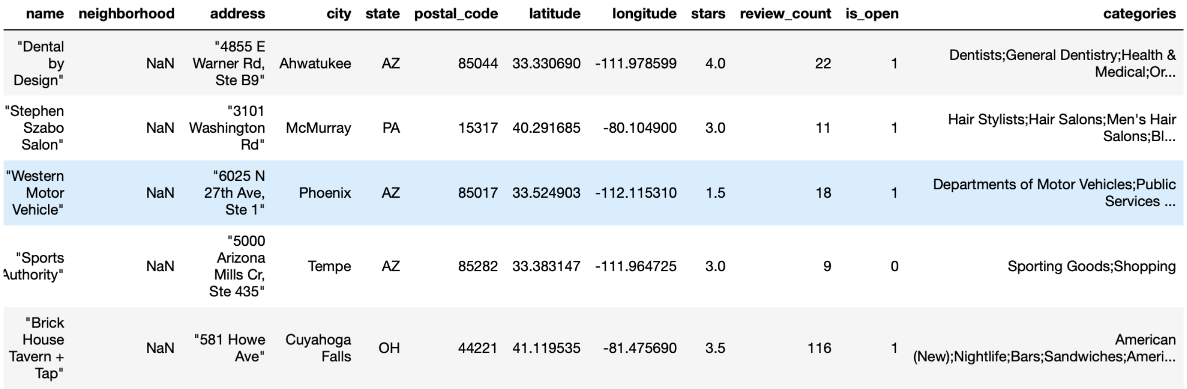
Hence, with the Yelp business data (cited later), we tried to solve the problem of predefined success in business and reduce risk by predicting, based on customer reviews and business ratings, which kind businesses usually do good in which areas and also give more specific details of such businesses. The Yelp data set is huge and several factors which are discussed next were taken care of when filtering our data

3 Data Preparation

Our project is revolves around the Yelp Business Dataset which can be downloaded at

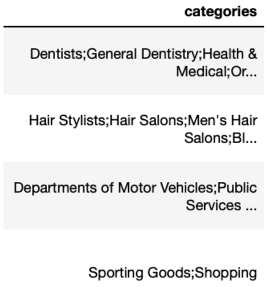
`https://www.yelp.com/dataset/download`

The dataset consists of several individual tables - business(contains business data including location data, attributes, and categories), review(contains full review text data including the user_id that wrote the review and the business_id the review is written for), user (user data including the user's friend mapping and all the metadata associated with the user)



name	neighborhood	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories
"Dental by Design"	NaN	"4855 E Warner Rd, Ste B9"	Ahwatukee	AZ	85044	33.330690	-111.978599	4.0	22	1	Dentists;General Dentistry;Health & Medical;Or...
"Stephen Szabo Salon"	NaN	"3101 Washington Rd"	McMurray	PA	15317	40.291685	-80.104900	3.0	11	1	Hair Stylists;Hair Salons;Men's Hair Salons;Bl...
"Western Motor Vehicle"	NaN	"6025 N 27th Ave, Ste 1"	Phoenix	AZ	85017	33.524903	-112.115310	1.5	18	1	Departments of Motor Vehicles;Public Services ...
"Sports Authority"	NaN	"5000 Arizona Mills Cr, Ste 435"	Tempe	AZ	85282	33.383147	-111.964725	3.0	9	0	Sporting Goods;Shopping
"Brick House Tavern + Tap"	NaN	"581 Howe Ave"	Cuyahoga Falls	OH	44221	41.119535	-81.475690	3.5	116	1	American (New);Nightlife;Bars;Sandwiches;Ameri...

Figure 1: Raw Data table.



categories
Dentists;General Dentistry;Health & Medical;Or...
Hair Stylists;Hair Salons;Men's Hair Salons;Bl...
Departments of Motor Vehicles;Public Services ...
Sporting Goods;Shopping

Figure 2: Dictionary.

and several other tables. Clearly the dataset is big and our approach to using it in a model was to not make complex models. We also may not know how much the training attributes are related to the predictor attribute.

Hence we decided to use the business table for our model input as it was the most relevant to what we were trying to do. Before any data modification, the table looked like the one given in Figure 1.

Moving forward, the data is clearly dirty and cannot be used directly for learning and testing. So we followed the following to processes of data cleaning depending on what the data was used for.

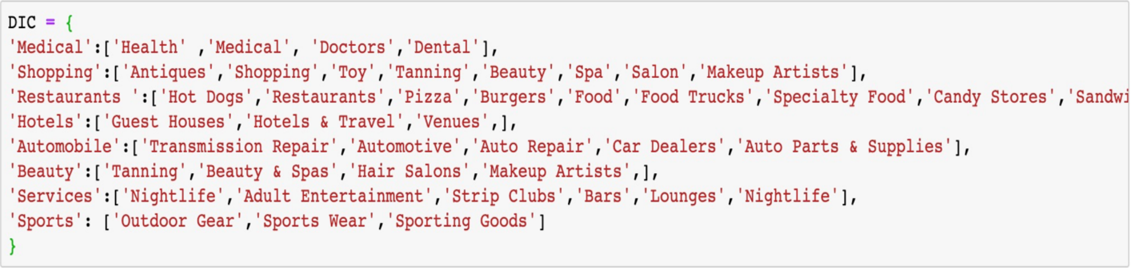
3.1 Data cleaning for Visualisation

When processing the business data, we found that part of the data was null, including postal_code, longitude, latitude, state and city attribute. Also, because the post_code will not affect the data analysis, we ignored that part. In order to clear the other null value, we fulfill those part with a specific value '999' and then drop those data.

3.2 Data cleaning for Modelling

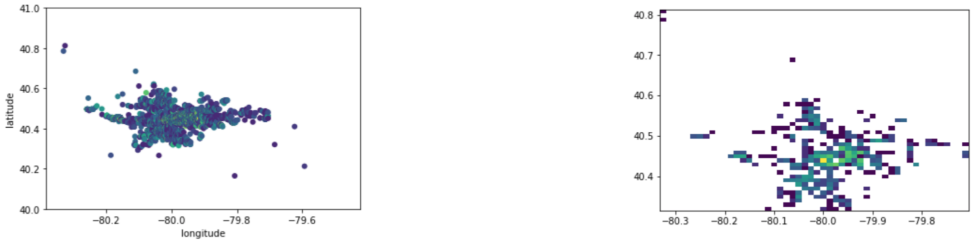
Categories Besides the normal filtering we did in visualization part, According to the problem statement, the column in categories is a necessary attribute in our main categorical. The 'categories' columns gives us a hint of what kind of a business it is. But the given values look Figure **??** So we did some word string matching and mapped these values to a single word value among the dictionary key in Figure 3 and then converted to categorical value. The dataset contains around 1.7 million tuples initially. So we did this data cleaning on Google TPU for faster processing to get 1.4 million tuples in return.

Ratings We also rounded of the star rating values (for e.g 4.5 changes to 5) so that the model is not short of data. Having lesser classes also help to learn a better model



```
DIC = {
'Medical':['Health' ,'Medical', 'Doctors','Dental'],
'Shopping':['Antiques','Shopping','Toy','Tanning','Beauty','Spa','Salon','Makeup Artists'],
'Restaurants ':['Hot Dogs','Restaurants','Pizza','Burgers','Food','Food Trucks','Specialty Food','Candy Stores','Sandwi
'Hotels':['Guest Houses','Hotels & Travel','Venues',],
'Automobile':['Transmission Repair','Automotive','Auto Repair','Car Dealers','Auto Parts & Supplies'],
'Beauty':['Tanning','Beauty & Spas','Hair Salons','Makeup Artists',],
'Services':['Nightlife','Adult Entertainment','Strip Clubs','Bars','Lounges','Nightlife'],
'Sports': ['Outdoor Gear','Sports Wear','Sporting Goods']
}
```

Figure 3: Dictionary.



(a) Most Popular restaurants in Pittburgh

(b) Heat map of restaurant density

Figure 4

4 Visualization

The data we download from 'Yelp Dataset' contains over 4.7 million reviews, 156 thousands business, and 12 metropolitan areas. Spanning over decades of data and attributes, the dataset provide a countless number of questions to be asked. The dataset contains several types of information including business, users, checkins, reviews, etc. In order to get a better understanding the actual situation of a city, we want to find its characteristics of different neighborhood and its cuisines. Since we are familiar with Pittsburgh,PA, which is rich in business and encompasses different types of business models. The majority of the project focuses on the restaurants in Pittsburgh, so we screened out the restaurants in Pittsburgh and show the data virtualization. Through this, we can have an intuitive understanding and feeling of the distribution of catering in Pittsburgh, and take this as a reference to continued follow-up research.

Pittsburgh has 6355 restaurants and 62 different neighborhoods. We selected the most popular restaurant (the restaurant which has more reviews) and restaurant density in a neighborhood and shows them on heat maps. The restaurant density map of Pittsburgh (Figure 4) shows that the restaurant are widely distributed in different neighborhood and a lot of the neighborhood restaurants in the center of the city are more intensive and also more popular which can found from the most popular restaurant in Pittsburgh. Combining the two pictures, we can see that the center of Pittsburgh, as a business gathering area, not only has many restaurant, and they are also widely praised.

To better visualize and explore the Pittsburgh neighborhoods a bit more, we found 10 neighborhoods(Figure 6) which has the largest number of restaurants and the most popular restaurants. The map on the left shows the 'most restaurant', we can find that the area which latitude is around -80 (downtown) has more restaurants in Pittsburgh, that represent the same trend compared with the heat map. The second map on the right shows the top 10 neighborhoods with the most popular restaurant which means those restaurants have the highest average review count. In comparison the two maps, only Shadyside, Strip District, Downtown and Lawrenceville are on both the list. These four neighborhoods are the 'Best Neighborhoods for Eating in Pittsburgh'.

5 Model Learning

Choosing the algorithm is always one of the most important part of solving a problem in the field of data science, identifying the type of problem is key task in the process of choosing a model. Our problem here is a multinomial classification problem so we decided to implement multinomial logistic regression on our model later on we also implemented multinomial logistic regression with parameters, lightgbm with hyperopt.

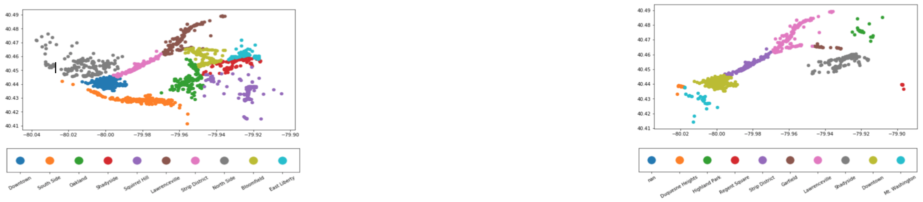
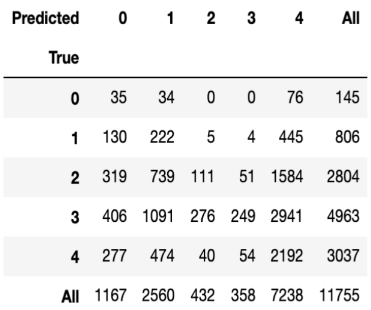


Figure 5: Pittsburgh Neighborhood Restaurants



Predicted	0	1	2	3	4	All
True						
0	35	34	0	0	76	145
1	130	222	5	4	445	806
2	319	739	111	51	1584	2804
3	406	1091	276	249	2941	4963
4	277	474	40	54	2192	3037
All	1167	2560	432	358	7238	11755

Figure 6: Confusion Matrix in Naive Bayes model

Initially we also tried models like like Naïve Bayes, Decision Trees and observed which are more generalizable.The above models if trained on data such as business location, type, number of positive and negative user reviews, workplace timings with the predictor being overall business rating, we see the model giving us good ratings as outputs for business type in particular areas that will do well in that area and bad ratings for business specifications that will not do well.

5.1 Model Performances

Multinomial Naive Bayes

Starting with Multinomial Naïve Bayes, we first calculated the confusion matrix of our predictions. Looking at the predictions, we see that major part of the star rating predictions is 4 and around 4. This makes it very difficult to predict the other ratings. As a result, the accuracy of this model came to be was around, 0.2389621437686091, Figure 6 above shows the Confusion Matrix

Multinomial Logistic Regression with parameters

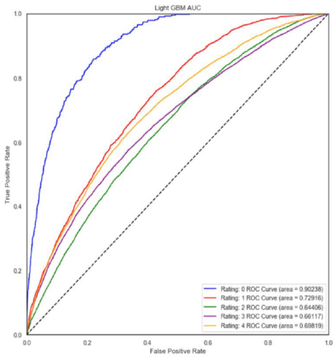


Figure 7: Train Data ROC AUC plot

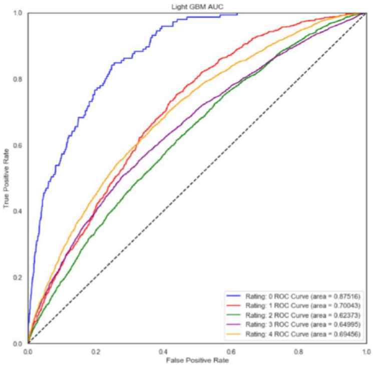


Figure 8: Test Data ROC AUC plotl

Since our accuracy is very low, we applied Multinomial Logistic regression with the parameters (_random_state=0, multi_class='multinomial', solver='newton-cg', penalty = 'l2', C = 2, max_iter = 100) Applying Multinomial Logistic Regression: Our accuracy is 0.4361548277328796

Light Gradient Boosted Machine (LightGBM)

Clearly the accuracies we are getting are no good. Hence, we move on to applying LightGBM on the same dataset given the boosting techniques are more efficient for uneven data. After modelling the lightgbm model, we get the AUC for train and test data (80-20 ratio) as displayed. Both the test and the train AUC curves have good scores for rating 0. This is probably because there is low data for that rating and people usually do not rate businesses so badly. While the other ratings have accuracies of around 0.68. This is a very good improvement from our previous models We used the following parameter values for the model after manual hyperparameter tuning.

The AUC plots are gven in Figure 7 and Figure 8

Looking at the graphs we can deduce that the Test and Train accuracies improved from the previous results i.e. after categorical filtering.

5.2 Shapley Additive explanations (SHAP)

SHAP – SHapley Additive exPlanations – explains the output of any machine learning model using Shapley values. SHAP belongs to the family of "additive feature attribution methods". This means that SHAP assigns a value to each feature for each prediction (i.e. feature attribution); the higher the value, the larger the feature's attribution to the specific prediction. It also means that the sum of these values should be close to the original model prediction.

SHAP has actually unified six existing feature attribution methods (including the LIME method) and it theoretically guarantees that SHAP is the only additive feature attribution method with three desirable properties: Local accuracy,Missingness,Consistency.

Take a look at the shapley plots(Figure 10 and 9): The first plot which describes the impact of different features on the output and also the impact of categories on the review_count which in fact are the important values that are looked into while solving our real problem.

For comparison, a multi-prediction force plot is shown in Figures 11, 12, and 13. It is a combination of many individual force plots that are rotated 90 degrees and stacked horizontally. This figure below for 500 observations. We took snapshots of 50th , 250th, 500th observation.

In our problem the shapley graphs explains that at a particular observation how different features impact the output and explains it's impact. Features in our model are (latitude,longitude,review-count,ratings,categories). We can also see the general distribution of data and about two-thirds of the displayed data gives higher output values which is clearly not uniform data.

6 Conclusion

To conclude we can say that the models we learned may not look much reliable with accuracies revolving around 70%. But, given the low number of predictors we had, we can definitiely say that it is a promising hypothesis and several

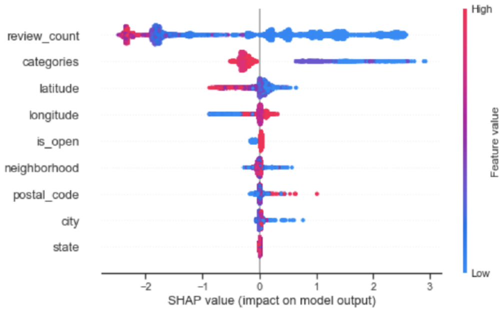


Figure 9

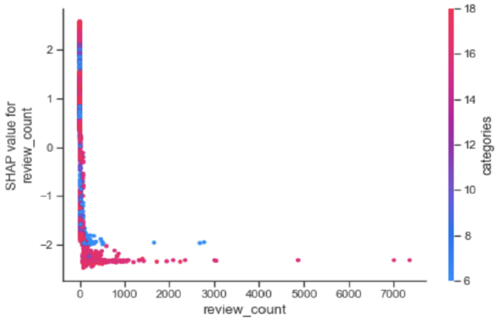


Figure 10

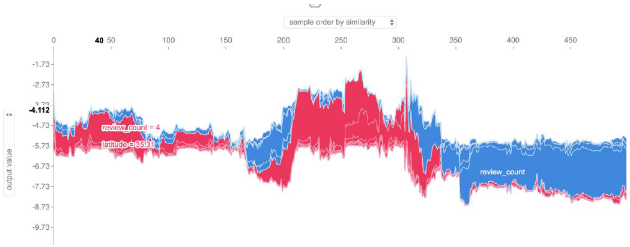


Figure 11

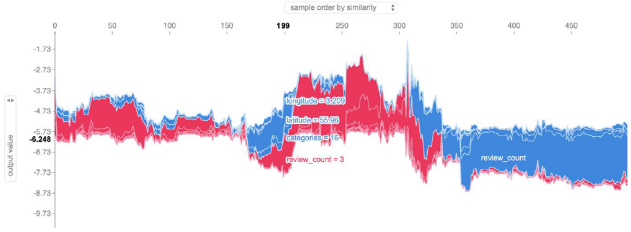


Figure 12

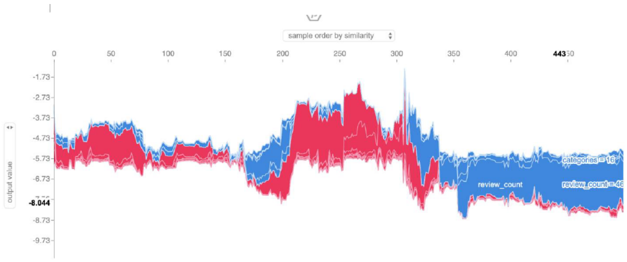


Figure 13

different learning algorithms may give better accuracies. Following this, our hypothesis can be considered true provided the simple models we have used to prove it.

References

[1] Yelp Dataset from 2019 Yelp Dataset challange . From *https://www.yelp.com/dataset/challenge*