Predictive Success of a Restaurant Using Data Science

Abhinav Soni
Northwestern University
AbhinavSoni2012@u.northwestern.edu

Abstract. This paper intends to tap the power of classification techniques to predict the success of restaurant. Decision Tree is one of the classification technique which is used for this particular use case. But the logic used here is applicable to many existing small businesses and people intending to start a business. There are many ingredients in a food item served at a restaurant but the item is best when the right ingredients are used in right proportion and cooked to right temperature. A good chef is essential to a good restaurant but many successful chefs have failed to make a successful restaurant. Restaurant, unlike a food item, is a business and it will only succeed if its customers are happy. We don't need a psychic to predict what will make a customer happy, however, we can use our psychology/common sense to easily identify the essential things to run a restaurant. 99.9% of restaurant owners in the world have the basic understanding of how to run their business, yet, not all of them are successful. Which means, we need a psychic who can tell us what the customer is thinking and what is that one thing / group of things that will make him/her happy. This paper intends to reveal exactly that, but, with the help of Decision Tree classification technique.

The likes and dislikes of a customer from previous experience is already available on Yelp.com. This data has been massaged to fit into the magic box called WEKA which will provide the recipe of success by using its complex algorithms. Reviews from Yelp.com have been broken into attributes like Portion size, food quality, location, customer service, etc. and then a rating has been assigned / value for each of these attributes. Combination of these attributes and their values for one restaurant will form one record and 180 records like this are used to find the key success strategy. To make the exercise and results meaningful, the data has been collected from restaurants that mostly serve one type of food and are located in particular area. Due care has been taken to cover the different type of customers like college students, business customers, family customers, etc. In addition to this, the dependability of the algorithm is also discussed so as to give an understanding on how reliable the results are. Further research on popular food items is also discussed in section 10, to get an idea of what is the most popular food item among people in that particular neighborhood.

I. INTRODUCTION

The objective of this project is to find the predictive success of a restaurant business based on the data collected from Yelp reviews. There are various factors that contribute to success of a restaurant like quality of food, customer service, ambience of restaurant, location, price, waiting time, portion of food (quantity) and menu size. **Good mood is associated with good food and hence good memories.** Restaurant business is unique because a successful restaurant becomes a tradition down the line and people like to visit their favorite restaurant even if they have moved into another suburb. **Even in an extremely competitive market, good restaurants thrive**. One reason is that there is lot of customization and personal touch that one can offer to its customer, in this line of business.

- For example Mamoun's falafel in New York City is thriving despite the competition from surrounding restaurants. The Mamoun's is a popular narrow hole-in-the-wall type of restaurant, with a line out the door most of the times.
- A fifteen minute Sushi meal could cost $300 to $500 per person, in the famous Sukiyabashi Jiro's restaurant in Minato, Japan. It is hard to get a reservation even one month in advance. President Obama made sure he dined at Jiro's restaurant and he did, with Japanese Prime Minister, when he visited Japan.

This paper aims to identify the key factors that lead to such success and also the factors that could lead to failure of popular restaurants, based on the data collected from reviews. These factors shall be used to establish a new restaurant as well.

II. DATA MINING APPLICATIONS

The AlchemyAPI software is used to identify the most popular and least popular menu item, from the information in reviews. AlchemyAPI offers AlchemyLanguage function as part of its text analysis service, which uses sophisticated natural language processing techniques to analyze content and add high-level semantic information. AlchemyAPI provides the ability to extract entity-level sentiment (positive or negative statements). Using sentiment analysis can help identity the content that refers to an entity in a positive or negative manner. To find the most and least popular menu item, five reviews are used as text and sentiment analysis is used in Alchemy to find the positive or negative sentiment associated with the item.

Weka software is used to run the Decision Tree algorithm on data. Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. The algorithms can either be applied directly to a dataset or called from your own Java code.

III. DATA UNDERSTNDING

To collect the data about restaurants, Yelp.com was used for collecting information from reviews. Sample of a restaurant review on Yelp is provided below. As we can see below review provides information about ambiance (whole in the wall), food quality (awesome), parking (street), portion size (big), etc. about the restaurant.

Figure 1.1: Screen shot of review on yelp.com

Yelp provides 'star' ratings for restaurants based on cumulative reviews. For purpose of this project, a five star rating is interpreted as excellent and two star as poor customer feedback. The 'star' rating is derived from the average rating of all the reviews for that restaurant. The 'star' rating is used as a class variable and the various attributes (service, food portion, parking, menu option, location, etc.) that lead to the rating are used as test data set.

Data is used from restaurants mostly offering Indian in New Brunswick and Somerset cities of New Jersey, so that the competition will be factored in. New Brunswick has many college students because of Rutgers University and professional people because of New Brunswick downtown area. The neighboring Somerset city, covers for restaurants with non- college customers. Hence the test data set sufficiently represents customers from various background that like to enjoy Indian food.

IV. DATA PREPARATION

Since some data elements have binary output (for example: Delivery – Yes/No, Outdoor seating – Ye/No), this data is separately used to generate a separate decision tree. Since binary output heavily influences the decision tree and results in large pruning, such attributes are used separately for creating a separate decision tree. The data from reviews is entered in Excel Spreadsheet and the file is saved in '.CSV' format so it can be uploaded to Weka. Total of 180 records were used to generate the decision tree.

Following attributes have binary output (Attribute list A)

Attribute	Values
Delivery	Yes, No
Take reservations	Yes, No
Accept Credit Card	Yes, No
Parking	Parking lot, Street
Good for	Lunch, Diner
Good for Kids	Yes, No
Good for Groups	Yes, No
Noise	Average, Quiet
Outdoor Seating	Yes, No
Wi-Fi	Free, No
Has TV	Yes, No
Waiter Service	Yes, No

Following attributes have more than two outputs (for example: Portion size – Big/Average/ Small, Ambience – Casual/Classy/Upscale) – (Attribute list B)

Attribute	Values
Portion size	Big, Small, Good
Quality/Flavor	Excellent, Good, Poor
Price	Under $10, $11 to $30, Above $82
Service	Friendly, Unfriendly, Neutral
Menu options	Good, Average, Limited
Location	Near College, Downtown, State highway
Ambience	Casual, Whole in the wall, Upscale
Noise Level	Quiet, Average, Loud

V. DATA MINING ALGORITHM

Decision Tree algorithm is used for data mining because it is a classification technique which can be used on existing data as a test set and the resulting decision tree could be leveraged to create a strategy for new restaurant and streamlining operations of existing restaurants in New Brunswick and Somerset cities.

Decision tree visually represents a decision situation and hence aids in communication. The branches of a tree explicitly show all the factors within the analysis that are considered relevant to the decision. For example, we can see the price factor was not used in any of the branches because people are not very sensitive to money when it comes to better dining experience.

The Decision tree technique can be used to identify the impact of change on result, if one of the underlying attribute's value is changed. Hence it allows businesses to identify the factors that are more sensitive and less sensitive. This kind of sensitivity analysis is difficult to do in other modeling environments.

Decision tree allows for forward and backward calculation paths to happen and hence the choice of the correct decision to take is made automatically. Since the moving parts / risk and success factors in restaurant business are limited and business is small scale, decision tree is quick, easy and accurate technique for decision making.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

Weka Decision Tree algorithm is used for generating the decision tree, with 15 cross-validation Folds. The total number of attributes are divided into two parts

- Attribute list A: Static Factors, not directly linked with food (less controllable by restaurant owner)
    - These attributes have binary values
    - Example: Parking, Wi-Fi, Has TV, Take reservations, etc.
- Attribute list B: Behavioral factors, directly linked with dining experience (sensitive factors, which are essential factors for a restaurant)
    - These attributes have more than two values
    - Example: Service, Food quality, Portion size, etc.

Result Analysis: As per the below Decision Tree in figure 1.3, there are three possible outcomes

- Excellent rating
- Good rating
- Average rating

The essential qualities for excellent rating are listed below:

a) Noise level = Quiet → Good for = Dinner
b) Noise level = Average → Alcohol = Yes → Delivery = No
c) Noise level = Average → Alcohol = No → Good for = Dinner → Parking = Private Lot → Delivery = Yes

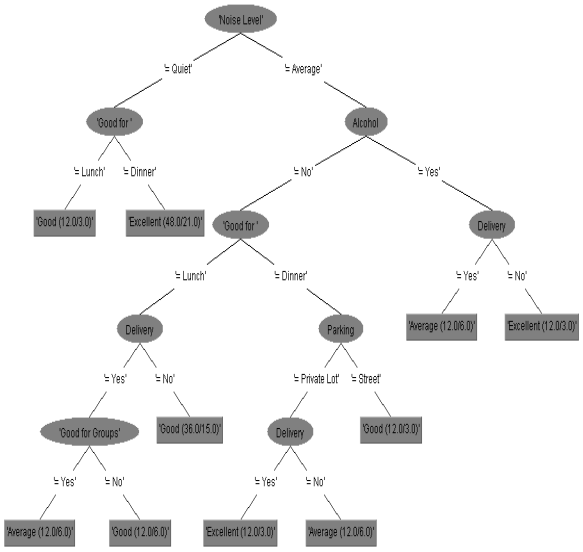

Figure 1.3: Decision Tree A

Conclusion: People in New Brunswick and Somerset cities like to dine in restaurant that are Good for Dinner, have a Private parking lot and deliver the food and Noise level is Quiet or Average.
This indicates that the majority of customers in this area go to restaurant at Diner time, so restaurant should be fully staffed in PM hours. Lunch buffet traffic is comparatively less than Diner traffic. Also the neighboring community likes home delivery so delivery service should be considered to gain edge over competitors.

```
Correctly Classified Instances         102               56.6667 %
Incorrectly Classified Instances        78               43.3333 %
Kappa statistic                          0.4028
Mean absolute error                      0.2742
Root mean squared error                  0.3839
Relative absolute error                 75.05   %
Root relative squared error             89.8482 %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.2       0.625     1         0.769      0.912    Good
                 0.683     0.274     0.573     0.683     0.623      0.706    Excellent
                 0         0.015     0         0         0          0.6      Poor
                 0.519     0.111     0.452     0.519     0.483      0.814    Average
Weighted Avg.    0.567     0.166     0.425     0.567     0.483      0.747

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 45  0  0  0 |  a = Good
  9 43  2  9 |  b = Excellent
  9 28  0  8 |  c = Poor
  9  4  0 14 |  d = Average
```

Figure 1.4: Statistics for Decision Tree A

Analysis of result: The overall confidence level is 56.67%, so it is not very insightful, however the ROC Area of restaurants with 'Good' rating is 0.912 and for Average rating it is 0.814. This means that the decision tree model is highly accurate for Good and Average ratings compared to 'Excellent' and 'Poor' ratings.

Decision Tree B: As seen in figure 1.5 below, the Decision Tree B accounts for the attributes that are directly linked with food and the values of such attributes cannot be expressed in binary fashion. It is essential to analyze the path for 'Excellent' and 'Poor' rating path for a successful strategic decision making.

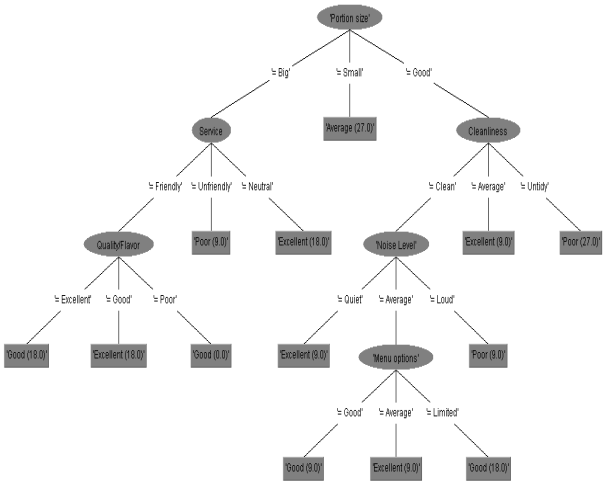

Figure 1.5: Decision Tree B

Conclusion based on Decision Tree B:

- **a)** A BIG portion size with Friendly or Neutral customer service can influence the rating positively and very quickly
- **b)** Portion size is the key factor that can make or break a restaurant. If a restaurant offers excellent service, pricing, ambiance, menu options, it will still remain average, if the portion size is SMALL
- **c)** Untidy and Loud / Noisy restaurants are highly disliked by residents of Somerset and New Brunswick. Such occurrence can immediately lead to failure / poor rating

d) A small or limited menu is not a cause of worry. So instead of focusing on offering variety of food items or being creative with offerings, one should focus more on portion size, cleanliness and noise level, to make the BIGGEST POSITIVE IMPACT.

```
Correctly Classified Instances         180               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              180

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Good
                 1         0          1         1         1          1        Excellent
                 1         0          1         1         1          1        Poor
                 1         0          1         1         1          1        Average
Weighted Avg.    1         0          1         1         1          1

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 45  0  0  0 |  a = Good
  0 63  0  0 |  b = Excellent
  0  0 45  0 |  c = Poor
  0  0  0 27 |  d = Average
```

Figure 1.6: Statistics for Decision Tree B

Analysis of result: The confidence level is 100% for all the results which means this model is full proof and one can blindly follow the conclusion listed above for **Decision Tree B.** The factors that are said to cause failure will certainly lead to failure, so those should be the first priority and plans should be made to never lag behind on such factors.

VII. CONCLUSION

As we can infer from section 8 and 9, following are the specific factors that drive a customer's experience. The key influencing factors are indicated 'Yes' in table 1.1 below. The Attributes which are indicated in bold orange (Noise and Portion Size) are of utmost importance.

Table 1.1: Key Influencing Factors for a dining experience

Attribute	Values	Key Influencing Factor
Delivery	Yes, No	Yes
Take reservations	Yes, No	No
Accept Credit Card	Yes, No	No
Parking	Parking lot, Street	No
Good for	Lunch, Diner	No
Good for Kids	Yes, No	No
Good for Groups	Yes, No	Yes
Noise	Average, Quiet	Yes
Outdoor Seating	Yes, No	No
Wi-Fi	Free, No	No
Has TV	Yes, No	No
Waiter Service	Yes, No	No
Portion size	Big, Small, Good	Yes
Quality/Flavor	Excellent, Good, Poor	No
Price	Under $10, $11 to $30, Above $82	No
Service	Friendly, Unfriendly, Neutral	Yes
Menu options	Good, Average, Limited	No
Location	Near College, Downtown, State highway	No
Ambience	Casual, Whole in the wall, Upscale	No
Cleanliness	Clean, Average, Untidy	Yes

VIII. FUTURE WORK

Detailed analysis on the favorite food items can be done with the help of sentiment analysis. Best rated reviews of a restaurant can be combined to form one chunk of text. This text can be imported to AlchemyAPI to find the Positive sentiment and its relevance ratio. As we can see below, in figure 1.7, the AlchemyAPI software is used to identify the most popular menu items among

the diners of one of the best rated restaurants in New Brunswick and Somerset cities.

Keyword	Relevance	Sentiment
New Brunswick/College Ave	0.986324	neutral
Butter Chicken	0.803121	positive
go-to meal	0.79506	neutral
Indian cuisine	0.716714	neutral
best places	0.6395	positive
naan	0.482491	positive
crispy	0.461074	positive
portions	0.408398	positive
opinion	0.395998	neutral
rice	0.388683	neutral
experience	0.354368	neutral
area	0.352884	positive
entrée	0.347351	neutral
food	0.346491	positive

Figure 1.7: Sentiment Analysis – Most Popular Item Analysis

Figure 1.8 is outcome of negative reviews, to find the least popular item in that particular restaurant. The least popular item is 'Paneer Tikka'. Similarly, reviews of group of restaurants in an area can be analyzed to find the likes and dislikes of people in a particular neighborhood.

Keyword	Relevance	Sentiment
paneer tikka	0.995466	negative
paneer platter	0.8344	neutral
paneer achari	0.818748	neutral
paneer fan	0.785458	positive
New Brunswick	0.700035	positive
Great Indian place	0.677115	positive
kati rolls	0.581091	positive

Figure 1.8: Sentiment Analysis – Least Popular Item Analysis

REFERENCES

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Boston: Pearson Addison Wesley, 2005.

Risk and Decision Analysis Blog: Dr. Michael Rees (*http://blog.palisade.com/2008/08/18/why-use-decision-tree-analysis/)*