Yelp
Predicting Restaurant Success

Neel Vasa
Aditya Vaidya
Shreya Kamani
Manan Upadhyay
Mark Thomas

IOM 528 SPRING 2014

Executive Summary

In this day and age of social media, with its instant reviews and referrals, many businesses seek to find the best way to improve their online reviews in order to attract and keep clientele or they seek to establish a new business that is mindful of what drives good reviews. This is particularly true for restaurants and food establishments. Several studies have been done that have looked at the correlation between a restaurants star rating on Yelp and their success related to the number of reservations and the revenue increase. We have obtained a dataset from Yelp, which lists businesses in and around Phoenix AZ. The dataset has each business's number of reviews, including their ratings, and other common attributes. Through this info we plan to mine the data for the aspects that make a particular business type successful.

We decided to narrow our current models on restaurants business types as this is what Yelp is primarily used for. We started by determining what successful exactly means given the data we have obtained. We decided to use the number of reviews and the number of star ratings, and then chose a minimum threshold that the business must obtain to be considered successful. Using this threshold we assigned a binary number to each business with a 1 denoting success. This allows us to run various prediction models, such as decision tree, logical regression, and neural networks in order to find what independent variables are most likely to lead to the success of the business.

The data set was quite large to begin with, so we grouped many of the attributes together, and aimed our focus on these grouped categories. We then narrowed the number of businesses we included to those with zip codes that contained more than 20 businesses in that particular zip code. So each of our business IDs had the following independent variables to help us determine the success outcome: Zip Code, Mexican, Asian, American, European, Latin American, Fast Food, Cafes, # Meals, Meal Options, # Features and Amenities.

Using these independent variables we ran several models in SAS to identify those variables that most accurately predict the successfulness of a business on Yelp. We ultimately arrived at a model that used the most significant variable of success, and found that approximately 50% of success, as determined for our model, could be explained by the variables. The independent variables that had the greatest effect on the success were Fastfood, European Dishes, Latin American Dishes, Number of Meals and Number of Features and Amenities.

This information can assist businesses in improving their success through ratings and number of reviews. It can also provide valuable information to those who want to establish a new restaurant.

As businesses try to cater to more and more customers through social media, and as they develop strategies and methods for attracting those customers, social media can help to supply valuable info on customers' preferences. Through better mining of the data that is available through sites like Yelp and others, businesses can gain the info they need to market better to their customers.

Problem

We set out to discover if business attributes given by Yelp can help to predict whether a restaurant is successful. Business attributes that are collected by Yelp are things such as location (address), Food Type, Number of Meals, Amenities (TV, WiFi, etc.), Late Night Hours, Delivery, etc. We want to use this information to run models that help point to those attributes required to gain success.

We will run prediction analysis using decision trees, multiple regression, and neural networks using both JMP and SAS to find the model that has the best prediction rate. The main application of our result will be in helping restaurateurs decide the factors to consider for improving an existing restaurant or while setting up a new restaurant. On the basis of this idea we have defined the following hypotheses.

Hypothesis

Ho: Food type or categories would not determine the success of a restaurant
Ha: Food type or categories would determine the success of a restaurant

Ho: Having a large number of features and amenities would not increase the chances of a restaurant being successful
Ha: Having a large number of features and amenities would increase the chances of a restaurant being successful

Ho: Location of the restaurant does not play an important role
Ha: Location of the restaurant plays an important role

Data

Our project sources data from the Yelp Academic Dataset available at:
https://www.yelp.com/academic_dataset

Yelp is providing a generous sample of data from the greater Phoenix, AZ metropolitan area including:

15,585 businesses
111,561 business attributes
11,434 check-in sets
70,817 users
151,516 edge social graph
113,993 tips
335,022 reviews

Dataset Description: The dataset is a single compressed file, composed of one json-object per line. Every object contains a 'type' field, which tells whether it is a business, a user, or a review. The data consists of Business objects that contain basic information about local businesses. The 'business_id' field can be used with the Yelp API to fetch even more information for visualizations. The fields are as follows:

```
{   'type': 'business',
  'business_id': (encrypted business id),
  'name': (business name),
  'neighborhoods': [(hood names)],
  'full_address': (localized address),
  'city': (city),
  'state': (state),
  'latitude': latitude,
  'longitude': longitude,
  'stars': (star rating, rounded to half-stars),
  'review_count': review count,
  'categories': [(localized category names)]
  'open': True / False (corresponds to closed, not business hours),
  'hours': {        (day_of_week):
                    { 'open': (HH:MM), 'close': (HH:MM) },    ... },
  'attributes': {   (attribute_name): (attribute_value),       ... },       }
```

These business objects also contain restaurants which of primary interest and these are extracted to enrich the raw dataset. The restaurant business objects contain various other fields specific to restaurants fields in json representation that are then converted to more amenable formats. The restaurant data fields also contain additional fields for the type of content served and also amenities available eg. Wi-Fi, TV screens etc.

Data Usage in Project

The raw data format is JSON and is not supported by natively by JMP and SAS Miner. Also the signal to noise ratio in the raw data is too large and thus we clean the data using a JAVA code and convert it to a .CSV format that is supported by various applications.

Of the approximately 15000 business objects, about 5000 restaurant business objects were extracted yielding a Signal to Noise Ratio (SNR) = 0.33. This apprised us of the fact that enrich of the data is an extremely important part of the project and we further concentrated more efforts on enriching the data to improve the SNR. Also, these 5000 rows were partitioned into training and testing datasets. K-Fold partitioning was done to negate the effects of potential bias and randomization factors. This partitioning approach helped us achieve stable performance that was resilient and robust.

Procedure

We learned numerous ways in class with respect to analysis of large datasets and their usage in deriving business insights. Of them, we chose the MAGIC framework of analysis where we proceed through the steps of Modeling, Analysis, Grilling, and Improving Continuously (MAGIC) in-order to create an accurate model to address the problem that we tackle.

The following steps were followed in line with the MAGIC framework learnt in class:

1. Research Data sources

    a. We thoroughly researched data sources at various websites for ways and means to get data and potential ways to enrich it so as it is pertinent to our project problem statement. The library session held in class about getting data was especially helpful as it opened up more avenues to direct our research and ultimately helped us reach Yelp and finalize it as a primary source of our data.

2. Enrich Data and Format Conversion

    a. The data obtained from Yelp as explained above had a low SNR(signal to noise ratio). This adversely affected our preliminary models as we were not able to achieve the desirable accuracy in prediction results.

    b. We thus, enriched the raw data, using various approaches and also converted the json format data into a more universal .csv format data.

    c. The format conversion helped us in using various models and analytic tools on our data to compare on and we were able to achieve the best results in SAS Miner.

3. Model Building

    a. This was the crux of the project and we considered various models from Decision Trees to Logistic Regression to Neural Networks.

    b. Each approach had its own advantages and fallbacks for eg where neural networks provided a tremendous accuracy in training data set, the failed miserably on the testing dataset.

4. Improving Continuously

    a. We implemented various models constantly tweaking and changing the parameters in order to achieve the desired accuracy.

    b. We also followed a novel way of K-Fold partitioning in the data set to obtain better results and robust performance.

5. Final Model: A final model was thus reached which provided us with a satisfying accuracy, robust performance and was consistent with our business sense.

Extract, Transform, and Load Data

The original data set was a 300 MB JSON file containing a large number of extraneous attributes. We refined this in a first pass of the extraction to restrict only business objects.

In the second pass, the business objects were further refined using various JSON parsing libraries to remove various attributes, like location and category data and then these were stored in a data format (.csv) which can be opened using JMP and SAS Miner. This data set still had numerous independent variables, and therefore grouping some of these variables needed to be considered to improve our model outputs.

In a third refinement of the data, the attributes were counted and accumulated, replacing individual binary 0/1 values of attributes with a count of the number of attributes and amenities the restaurant provides. In this pass over of the data set, outliers in zip codes were eliminated as well, retaining only zip codes with 20 or more restaurants. Data attributes accumulated include meals served, open times (late-night etc.), parking, Wi-Fi etc. Additionally, food categories were grouped into broader categories, for example various categories like Pizza, Burgers, and Sandwiches were grouped under Fast Food.

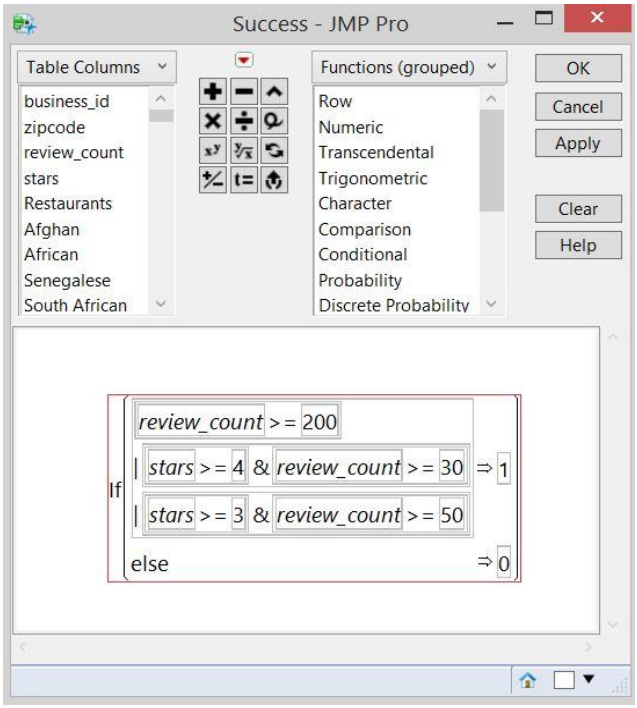
Appendix 1 shows a sample tuple of the refined data set.

Key Summary Stats

Appendix 2 shows the distribution info between several of the independent variables from the dataset before we condensed them down into smaller groups. These were the most significant variables from the larger dataset. They were review count and stars, which are quantitative, and then Mexican, Caters, Price Range, Really Fast Food, and Alcohol, which are all qualitative. The greatest number of stars for the restaurant businesses was 3, which had 2538 instances; the second most number was 4 with 2015 instances. The review counts had a greater distribution, ranging from 0 to over 1000. The mean for reviews was at 44, with a standard deviation of 76, and the median was 16. So a majority of the restaurants had less than 100 reviews, with a small number receiving over 100, including some outliers that received over 500.

Appendix 3 shows the distribution from the refined data set. These are some of the key variables which include review count and stars, as well as categories Mexican, Really Fast Food, American Dishes, and the # Meals, and # Features and Amenities. Similar to the larger data set, the mean for review counts is 47 and the median is 18, so the majority of reviews are under 100 with a few outliers over 500. The stars also are largest at 3 with 2328, and second is 4 with 1857 at that level. Another variable we created is Features and Amenities, which has a fairly normal distribution with a mean around 4. We also found that there are some businesses that have some missing data, like number of meals per day listed is 0. So we need to consider the best way to handle this, for example defaulting all those with number of meals zero to a one.

Below is the calculation used for measuring success according to the data we obtained.



Models

Our next best model was run using Decision Tree. We used the optimal values and got to 7 splits. It separated first by the number of Features and Amenities and then split by a number of different zip codes as well as number of meals and then the type of food.

Appendix 4 shows the full tree. Appendix 5 shows the R Square for the model.

The R square for the testing and validation were very far off though, so it was over fit for training data. We decided to try Neural Network to test the model and see if we could improve on the R Square and the significance of the variables.

Our best model was run using Neural Network. We tried several variables and came up with the following parameters to give us the best R Square model for both Training and Validation.

Parameters:

Zipcode
Fastfood
European Dishes
Latin American Dishes
Meals
Features and Amenities

Rsq: 0.58 (Training) 0.51 (Testing)

I had first Fit Y by X for all our variables versus success. These 6 variables were among the ones that had maximum significance on success. The reason why these are statistically significant and make business sense:

1. Zip code: Geographical areas play a significant role in contributing towards success. Downtown areas and ASU campus areas in particular have a high success rate.

2. Fastfood: Quick bites. People in need of a quick snack or a cheap meal frequent fast food joints.

3. European Dishes: More often than not, fine dining restaurants where food quality and service is monitored and scrutinized strictly. Clients usually are big spenders, rich families or a venue for business meals.

4. Latin American: Frequency of Latin American restaurants is low compared to Mexican/Fast Food/Asian (less than a quarter of these places). But it is a popular cuisine because of lot of Latin Americans in Phoenix.

5. Meals: Restaurants serving 1 or 2 meals (lunch and dinner) are more successful than those serving 3 or more meals (like brunch). Maybe because of less popularity of brunch?

6. Features and Amenities: More the features and amenities (WiFi, Parking, Home Delivery, Alcohol etc) more likelihood of being a success.

Appendix 6 and Appendix 7 show the neural network.

Improvements

The project has a few shortcomings which can be overcome by utilizing various other data sources. The accuracy of Yelp data might be questionable, as competitors might deliberately give bad reviews or the owners themselves might start with very good reviews of their own business. To eliminate this, a few techniques could be used, such as grouping of reviews by time period (month-wise or quarterly).

Another improvement could be the addition of other regions' data. Including 5-6 cities could give a robust model, as this data might suffer from over fitting to the Phoenix area.

Lastly, including demographic data from census or other information sources could lead to a better indication of success/failure of a restaurant. For example, a restaurant in a prosperous neighborhood might have several 4/5 rated competitors, and would need a higher sustained rating to be deemed a success

Conclusion

The Yelp Phoenix data set proved to be a very rich and powerful data source for predicting restaurant success as defined in the project. Undertaking an extensive and rigorous ETL process helped significantly increase the accuracy rates for prediction in the models as well. The findings were clear and concise: A restaurant is highly likely to be a success if it caters to certain popular categories and provides a large number of services, regardless of its location. This particular finding is in contrast to the preliminary belief of location being the most crucial factor to success, and can probably be attributed to the fact that a better location also equals greater competition.

Addition of demographic data, additional cities' data and improvement in Yelp data as described in the Improvements section above has the potential to lead to a robust and highly accurate model for future use.

Bibliography:

The following resources were used for the development of this project:

http://www.yelp.com/dataset_challenge/

http://www.jmp.com/support/downloads/documentation.shtml

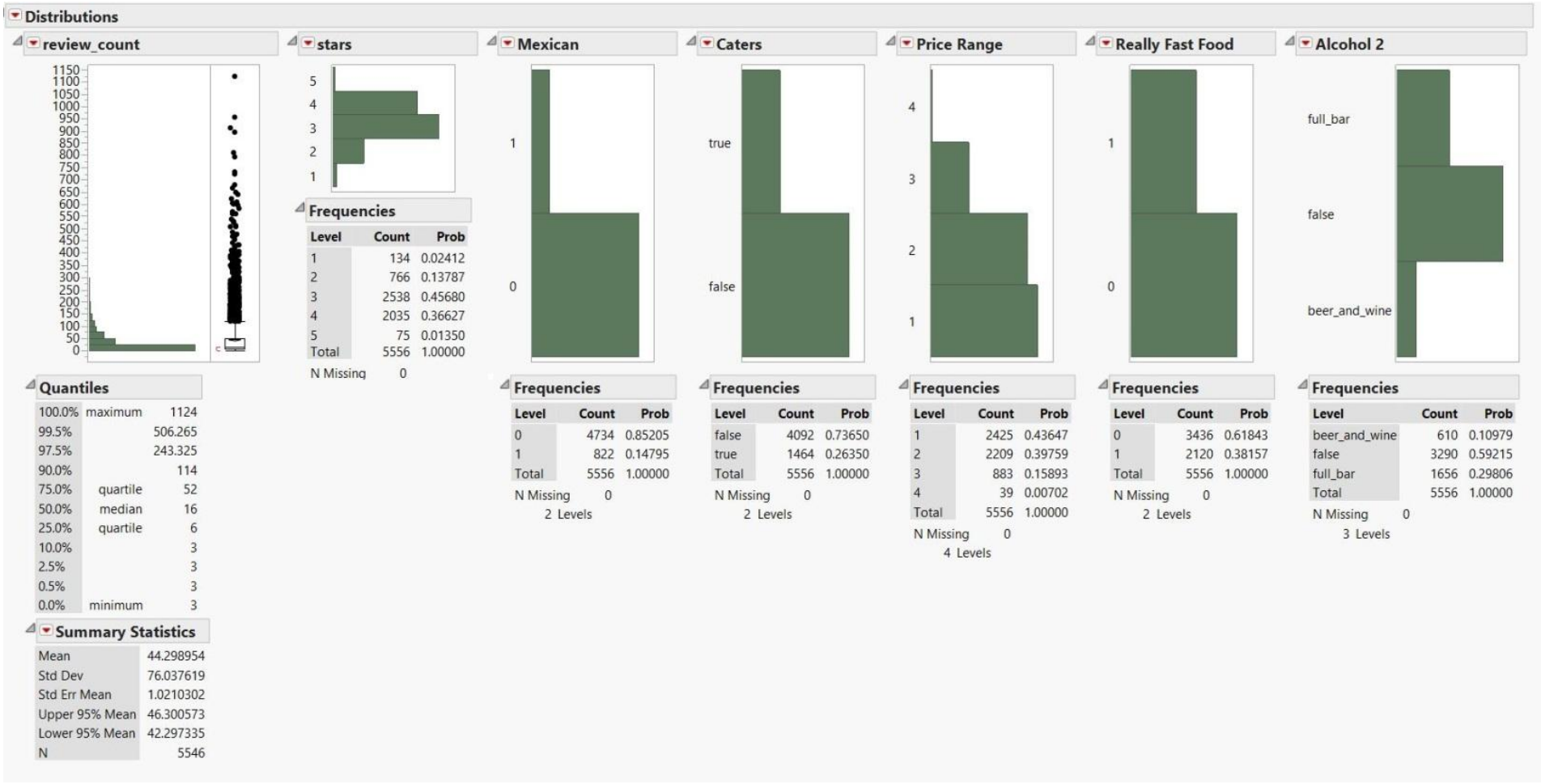
http://support.sas.com/documentation/onlinedoc/miner/

IOM 528 class notes and minicases

Appendix 1

business_id	zipcode	review_count	stars	Mexican	Really Fast Food	Asian Dishes	American Dishes	European Dishes	Latin American Dishes	Cafes	Meals	Meal Options	Features and Amenities	Success
OYlq2UoKPQO_ElqhC3a8	85003	125	3	0	0	0	0	0	0	0	3	1	6	1

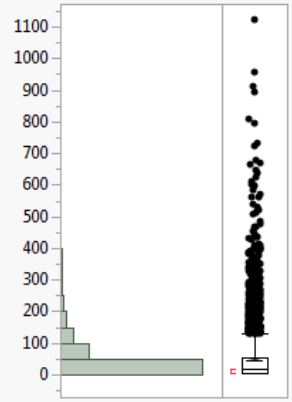
Appendix 2



Appendix 3

Distributions

review_count



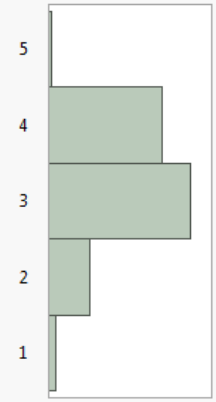
Quantiles

100.0%	maximum	1124
99.5%		521.7
97.5%		253.7
90.0%		120
75.0%	quartile	56
50.0%	median	18
25.0%	quartile	6
10.0%		4
2.5%		3
0.5%		3
0.0%	minimum	3

Summary Statistics

Mean	47.13592
Std Dev	78.890513
Std Err Mean	1.1129015
Upper 95% Mean	49.317693
Lower 95% Mean	44.954148

stars



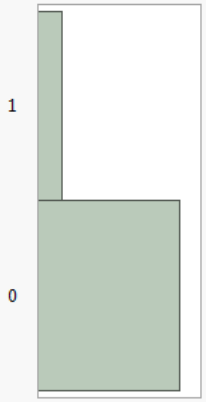
Frequencies

Level	Count	Prob
1	115	0.02289
2	667	0.13274
3	2328	0.46328
4	1857	0.36955
5	58	0.01154
Total	5025	1.00000

N Missing 0

5 Levels

Mexican



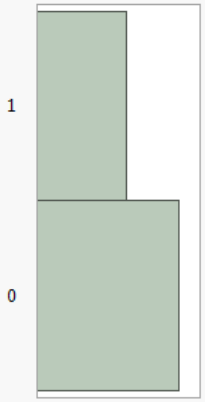
Frequencies

Level	Count	Prob
0	4291	0.85393
1	734	0.14607
Total	5025	1.00000

N Missing 0

2 Levels

Really Fast Food



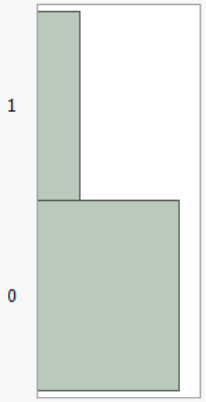
Frequencies

Level	Count	Prob
0	3091	0.61512
1	1934	0.38488
Total	5025	1.00000

N Missing 0

2 Levels

American Dishes



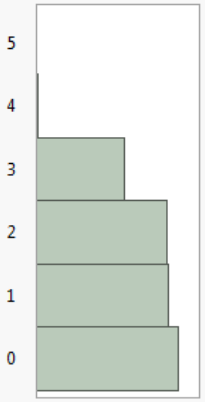
Frequencies

Level	Count	Prob
0	3877	0.77154
1	1148	0.22846
Total	5025	1.00000

N Missing 0

2 Levels

Meals



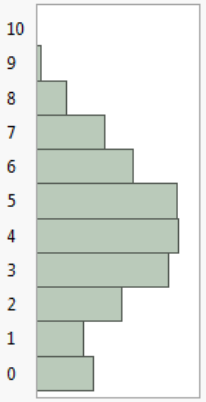
Frequencies

Level	Count	Prob
0	1442	0.28697
1	1334	0.26547
2	1327	0.26408
3	898	0.17871
4	20	0.00398
5	4	0.00080
Total	5025	1.00000

N Missing 0

6 Levels

Features and Amenities



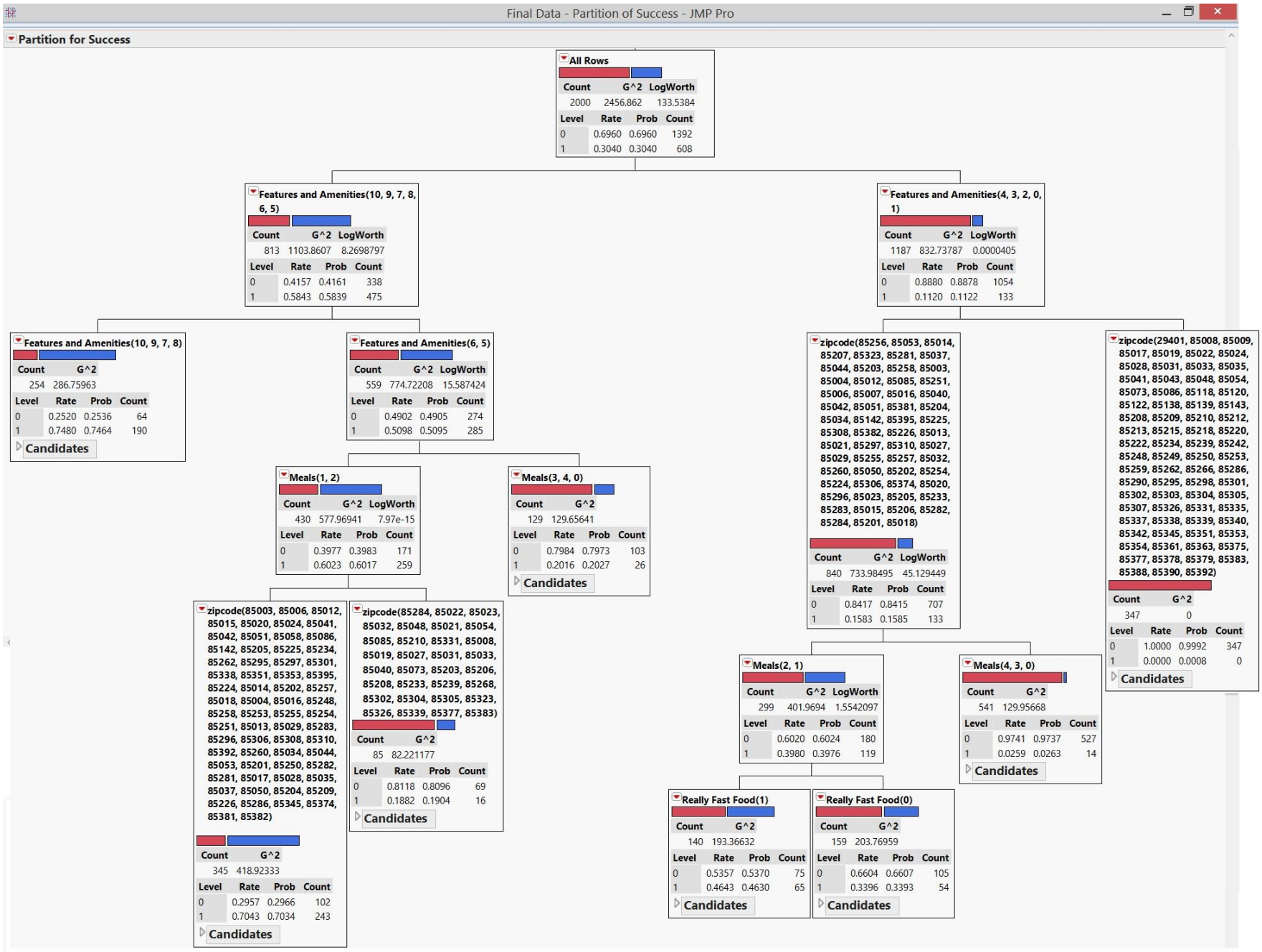
Frequencies

Level	Count	Prob
0	358	0.07124
1	296	0.05891
2	531	0.10567
3	827	0.16458
4	889	0.17692
5	878	0.17473
6	608	0.12100
7	423	0.08418
8	188	0.03741
9	24	0.00478
10	3	0.00060
Total	5025	1.00000

N Missing 0

11 Levels

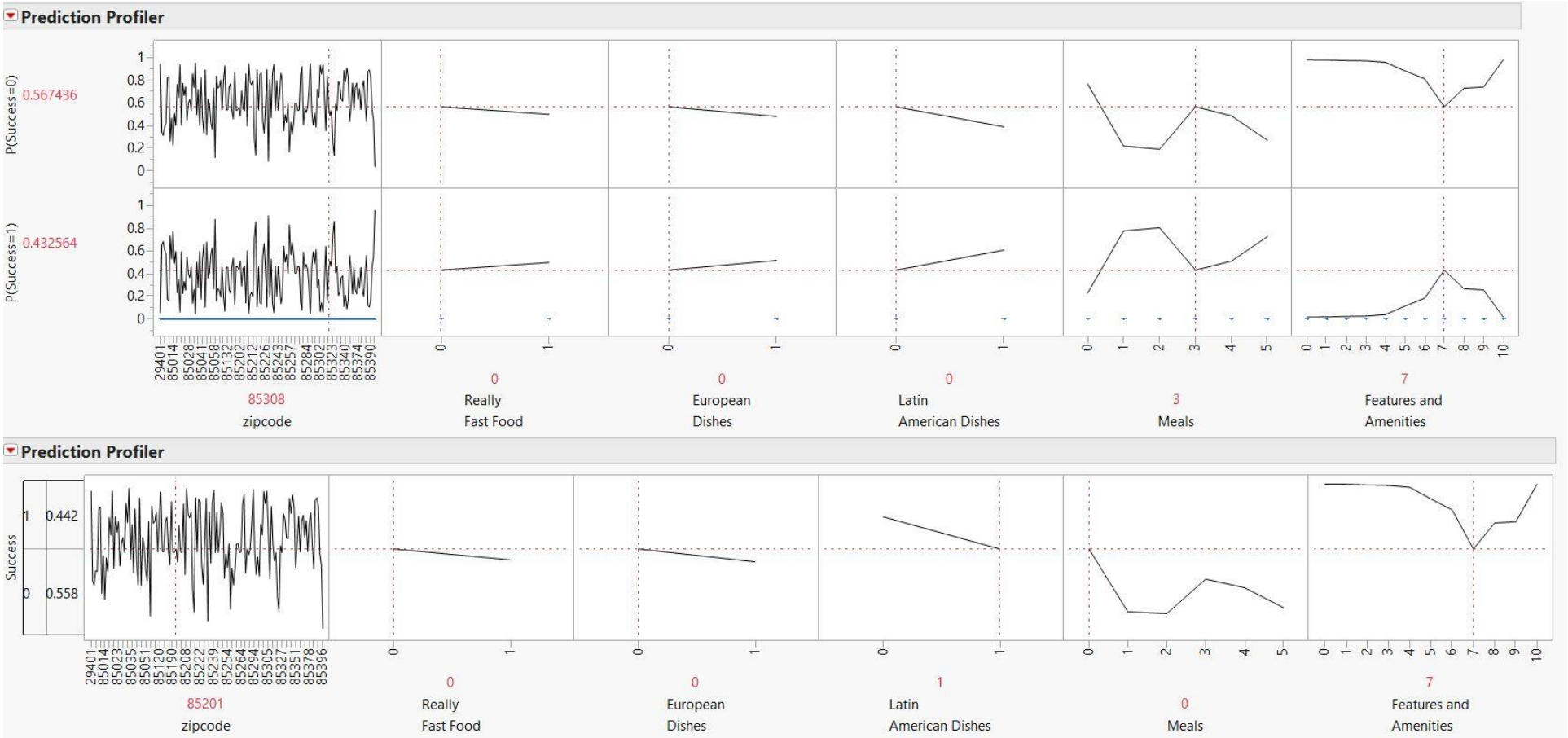
Appendix 4


Final Data - Partition of Success - JMP Pro
Partition for Success
All Rows
Count G^2 LogWorth
2000 2456.862 133.5384
Level Rate Prob Count
0 0.6960 0.6960 1392
1 0.3040 0.3040 608
Features and Amenities(10, 9, 7, 8, 6, 5)
Count G^2 LogWorth
813 1103.8607 8.2698797
Level Rate Prob Count
0 0.4157 0.4161 338
1 0.5843 0.5839 475
Features and Amenities(4, 3, 2, 0, 1)
Count G^2 LogWorth
1187 832.73787 0.0000405
Level Rate Prob Count
0 0.8880 0.8878 1054
1 0.1120 0.1122 133
Features and Amenities(10, 9, 7, 8)
Count G^2
254 286.75963
Level Rate Prob Count
0 0.2520 0.2536 64
1 0.7480 0.7464 190
Candidates
Features and Amenities(6, 5)
Count G^2 LogWorth
559 774.72208 15.587424
Level Rate Prob Count
0 0.4902 0.4905 274
1 0.5098 0.5095 285
zipcode(85256, 85053, 85014, 85207, 85323, 85281, 85037, 85044, 85203, 85258, 85003, 85004, 85012, 85085, 85251, 85006, 85007, 85016, 85040, 85042, 85051, 85381, 85204, 85034, 85142, 85395, 85225, 85308, 85382, 85226, 85013, 85021, 85297, 85310, 85027, 85029, 85255, 85257, 85032, 85260, 85050, 85202, 85254, 85224, 85306, 85374, 85020, 85296, 85023, 85205, 85233, 85283, 85015, 85206, 85282, 85284, 85201, 85018)
Count G^2 LogWorth
840 733.98495 45.129449
Level Rate Prob Count
0 0.8417 0.8415 707
1 0.1583 0.1585 133
zipcode(29401, 85008, 85009, 85017, 85019, 85022, 85024, 85028, 85031, 85033, 85035, 85041, 85043, 85048, 85054, 85073, 85086, 85118, 85120, 85122, 85138, 85139, 85143, 85208, 85209, 85210, 85212, 85213, 85215, 85218, 85220, 85222, 85234, 85239, 85242, 85248, 85249, 85250, 85253, 85259, 85262, 85266, 85286, 85290, 85295, 85298, 85301, 85302, 85303, 85304, 85305, 85307, 85326, 85331, 85335, 85337, 85338, 85339, 85340, 85342, 85345, 85351, 85353, 85354, 85361, 85363, 85375, 85377, 85378, 85379, 85383, 85388, 85390, 85392)
Count G^2
347 0
Level Rate Prob Count
0 1.0000 0.9992 347
1 0.0000 0.0008 0
Candidates
Meals(1, 2)
Count G^2 LogWorth
430 577.96941 7.97e-15
Level Rate Prob Count
0 0.3977 0.3983 171
1 0.6023 0.6017 259
Meals(3, 4, 0)
Count G^2
129 129.65641
Level Rate Prob Count
0 0.7984 0.7973 103
1 0.2016 0.2027 26
Candidates
Meals(2, 1)
Count G^2 LogWorth
299 401.9694 1.5542097
Level Rate Prob Count
0 0.6020 0.6024 180
1 0.3980 0.3976 119
Meals(4, 3, 0)
Count G^2
541 129.95668
Level Rate Prob Count
0 0.9741 0.9737 527
1 0.0259 0.0263 14
Candidates
zipcode(85003, 85006, 85012, 85015, 85020, 85024, 85041, 85042, 85051, 85058, 85086, 85142, 85205, 85225, 85234, 85262, 85295, 85297, 85301, 85338, 85351, 85353, 85395, 85224, 85014, 85202, 85257, 85018, 85004, 85016, 85248, 85258, 85253, 85255, 85254, 85251, 85013, 85029, 85283, 85296, 85306, 85308, 85310, 85392, 85260, 85034, 85044, 85053, 85201, 85250, 85282, 85281, 85017, 85028, 85035, 85037, 85050, 85204, 85209, 85226, 85286, 85345, 85374, 85381, 85382)
Count G^2
345 418.92333
Level Rate Prob Count
0 0.2957 0.2966 102
1 0.7043 0.7034 243
Candidates
zipcode(85284, 85022, 85023, 85032, 85048, 85021, 85054, 85085, 85210, 85331, 85008, 85019, 85027, 85031, 85033, 85040, 85073, 85203, 85206, 85208, 85233, 85239, 85268, 85302, 85304, 85305, 85323, 85326, 85339, 85377, 85383)
Count G^2
85 82.221177
Level Rate Prob Count
0 0.8118 0.8096 69
1 0.1882 0.1904 16
Candidates
Really Fast Food(1)
Count G^2
140 193.36632
Level Rate Prob Count
0 0.5357 0.5370 75
1 0.4643 0.4630 65
Candidates
Really Fast Food(0)
Count G^2
159 203.76959
Level Rate Prob Count
0 0.6604 0.6607 105
1 0.3396 0.3393 54
Candidates

Appendix 5

	RSquare	N	Number of Splits
Training	0.412	2000	7
Validation	0.280	3546	

Appendix 6



Appendix 7

Model NTanH(5)NTanH2(7)

Training

Success

Measures	Value
Generalized RSquare	0.5850783
Entropy RSquare	0.4347717
RMSE	0.3344498
Mean Abs Dev	0.2279717
Misclassification Rate	0.16
-LogLikelihood	694.34398
Sum Freq	2000

Confusion Matrix

Actual \ Predicted	0	1
0	1217	175
1	145	463

Confusion Rates

Actual \ Predicted	0	1
0	0.87428	0.12572
1	0.23849	0.76151

Validation

Success

Measures	Value
Generalized RSquare	0.5146876
Entropy RSquare	0.3694344
RMSE	0.3570678
Mean Abs Dev	0.244131
Misclassification Rate	0.1934574
-LogLikelihood	1362.7708
Sum Freq	3546

Confusion Matrix

Actual \ Predicted	0	1
0	2093	395
1	291	767

Confusion Rates

Actual \ Predicted	0	1
0	0.84124	0.15876
1	0.27505	0.72495