Paper 1784-2014

Dining With the Data: The Case of New York City and its Restaurants

Pruthvi Bhupathiraju Venkata, Dr.Goutam Chakraborty ,Oklahoma State University, OK, US

ABSTRACT

New York City boasts a wide variety of cuisine owing to the rich tourism and a large immigrant population. The quality of food and hygiene maintained at the restaurants, serving different cuisines, has a direct impact on the people dining in them. The objective of this paper is to build a model that predicts the grade of the restaurants in New York City. Our research also provides insights into the distribution of restaurants, cuisine categories, grades, criticality of violations etc. and concludes with a sequence analysis performed on the complete set of violations recorded for the restaurants at different time periods over the years 2012 and 2013. The data for 2012 is used to build the model, while the data for 2013 is used to score the model.

The data set consists of 15 variables that capture restaurant background and violation details. The target is an ordinal variable with three levels, A, B, and C, in the descending order of the quality representation. Various SAS EM® models, Logistic Regression, Decision Trees, Neural Networks, Ensemble models are built and compared using the validation misclassification rate. Stepwise Regression Model appears to be the best model with a prediction accuracy of 75.33%.

INTRODUCTION

The New York City department of Health and Mental Hygiene has established inspection procedures that monitor restaurants' compliance with food safety regulations. All restaurants are graded based on the violation scores received for two categories – critical and non-critical violations. The food inspectors use their discretion in allotting the violation scores to each category of violations recorded for the restaurant. Sometimes, restaurants, which receive lower grades request for revaluation of their violations scores and their grade. Restaurants continuously strive to improve the grade by curbing the violations. A predictive model for restaurant grades based on relevant factors will help the restaurants to take corrective actions and receive better inspection grades. Also, as the repeated non critical violations lead to several other critical violations, the food inspectors can impose strict regulations around certain types of non critical violations, helping the restaurants to maintain the food quality and safety.

DATA DICTIONARY

Variable Name	Level	Description
CAMIS	Nominal	This is an unique identifier for the entity (restaurant)
DBA	Nominal	This field represents the name (doing business as) of the entity (restaurant)
BORO	Nominal	Borough in which the entity (restaurant) is located. 1 = MANHATTAN 2 = THE BRONX 3 = BROOKLYN 4 = QUEENS 5 = STATEN ISLAND
BUILDING	Nominal	This field represents the building number for the entity (restaurant)
STREET	Nominal	This field represents the street name at which the entity (restaurant) is located.
ZIPCODE	Nominal	Zipcode as per the address of the entity (restaurant)
PHONE	Nominal	Phone Number
CUISINECODE	Nominal	This field is used to group restaurants by Cuisine Category. Full description can be obtained by lookup dataset (cuisine).
INSPDATE	Nominal	This field represents the date of inspection

GRADEDATE	Nominal	The date when the current grade was issued to the entity (restaurant)
CUISINE GROUP	Nominal	This field shows the full description related to the cuisine code
VIOLATIONCODE	Nominal	This field corresponds to VIOLCODE from WEBEXTRACT
CRITICALFLAG	Nominal	This indicates if Violation is critical or not. Y = Critical N = Not Critical.

Table1. Data Dictionary for Restaurant Data

DATA PREPARATION:

The overall data set consisted of over 560,000 records that spanned over the years of 2009-2013. For the analysis and modeling purposes, we sampled the data for 2013 and 2012. Data analysis and modeling is performed on 2012 data and scoring is done on 2013 data . We also created a combined data set for 2012/2013 for the sequence analysis. A flag variable, "changecritical" , is created that indicates whether there is an increase in the critical violations when compared to the previous year is created and coded as Y=yes, N=no. Since there are many cuisines, we aggregated the cuisines into the broad categories to be more specific in our analysis. Below are the broad categories used:

Cuisine Group	Cuisine Code
African	02,31,17,3
American	03, 10, 07, 15, 12, 16 , 24, 25, 29 , 39 , 41, 42, 43, 60, 65, 69, 70, 76, 78, 81, 40, 62,73
European	04, 30 ,11, 35, 37, 38, 26, 32, 47,57,64,66,71,77,83,80
Italian	48,63
Asian	05, 52, 09, 28,34,46,59,50,49, 45, 56,01, 06
Mexican	55
South American	53,13,19,61
Chinese	20, 21,22
Indian	44
Thai	82
Mediterranean	54,23
Seafood	72
Others	99,08,14,18,27,36,43,51,58,74,75,84

Table 1. Cuisine Groups

The target variable, "Final Grade", is an ordinal variable with three levels (A,B and C), with A indicating a good quality restaurant and C indicating a bad quality restaurant. The grades, based on the literature made available from the New York City government , were computed based on inspection scores as follows:

Grade	Score
A	0-13
B	14-27
C	28 and above

Table3. Restaurant Grade and Score Ranges

DATA ANALYSIS:

The Borough of Manhattan, with great commercial investments and tourist rate, has the highest number of restaurants, followed by the Boroughs of Brooklyn, Queens, Bronx and Staten Island respectively.

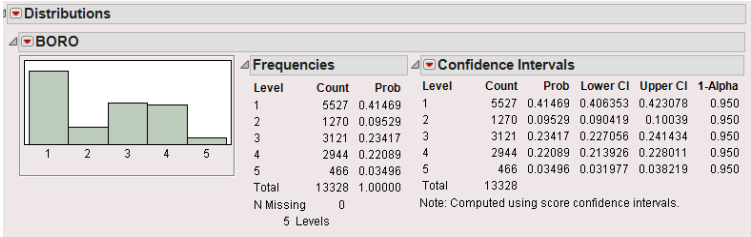

Figure 2. Distribution of Restaurants in each Borough of New York City

In New York City, based on the quality score, a vast majority of the restaurants are of A grade, followed by B grade restaurants and C grade restaurants. This emphasizes the point that New York City is a great place to dine for the locals as well as for the tourists.

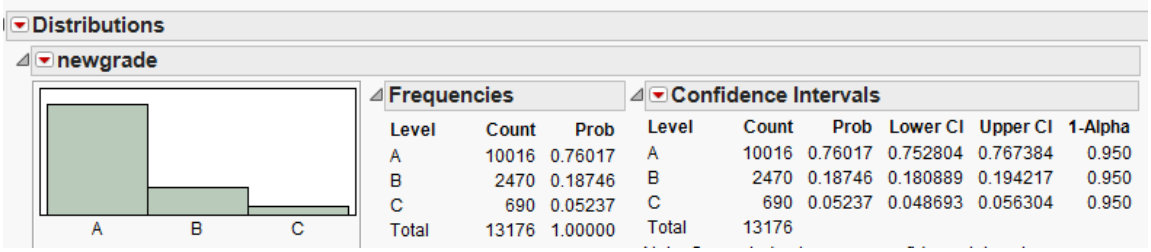

Figure 3. Distribution of Restaurant Grade

As shown below, among A grade Restaurants, for 67.35% of them there is no increase in the number of critical violations from the previous year (2011). However, for 63.91 % of C grade restaurants, there is an increase in the number of critical violations. This explains that the number of critical violations in a restaurant play an important role in determining the grade of the restaurant.

The FREQ Procedure

Table of changecritical by newgrade

		newgrade			
		A	B	C	Total
changecritical					
N	Frequency	6746	1192	249	8187
	Col Pct	67.35	48.26	36.09	
Y	Frequency	3270	1278	441	4989
	Col Pct	32.65	51.74	63.91	
Total	Frequency	10016	2470	690	13176

Figure 4. Cross Tab of Change in Critical Violations and Restaurant Grade

MODELING:

The business understanding and the Initial analysis of data helped in selecting the input variables and setting up their roles. The missing values are minimal and they are addressed with the help of base SAS code. The skewness and kurtosis of the interval values are within the optimum range. Data is partitioned into Training and Validation data of 50% each using the data partition node. Various Models, along with different variations within each model, are Implemented as follows:

Model Built	Variations
Regression	1. Stepwise Input selection 2. Forward Input Selection. 3. Backward Input Selection. 4. Stepwise with 2 factorial input combination
Neural Network	1. MLP Architecture 2. RBF Architecture (Ordinary Radial with Equal Width)
Decision Tree	1. Gini Index Splitting 2. Entropy Measure Splitting
Ensemble Model	Combinations of models above

Table 4. Summary of Different Models Built

Because the target is an ordinal variable, logistic regression is used for model building. For different variations of logistic regression that include stepwise input selection, forward input selection and backward input selection the stay level significance and the entry level significance are both set at the default values of 0.05. Also, a regression model (stepwise input selection) with a two factorial input combinations was used to model the data.

Two decision tree models with ordinal target criterion as Gini and Entropy are used. Both Gini impurity Index and Entropy use impurity reduction as goodness for a split. For both the splitting criteria, splits depend only on the ordering of the levels, making tree models robust to outliers in the input space. The goal during the split is to always maximize Gini Index and minimize Entropy.

Neural Network Models with Multilayer perception architecture (MLP) and ordinary radial basis function (with equal width) architecture are built. For both the Neural Net models, significant variables from the stepwise regression are used as input. The MLP architecture model is run with the default settings, but the radial architecture model is implemented with six hidden units.

The predictions from all the above models are combined to create a single consensus prediction through the ensemble model. The commonly believed advantage of the ensemble models is that it is better than the individual models that compose it. However, in this data, the regression model (stepwise) seems to be the better model even in comparison to the ensemble model.

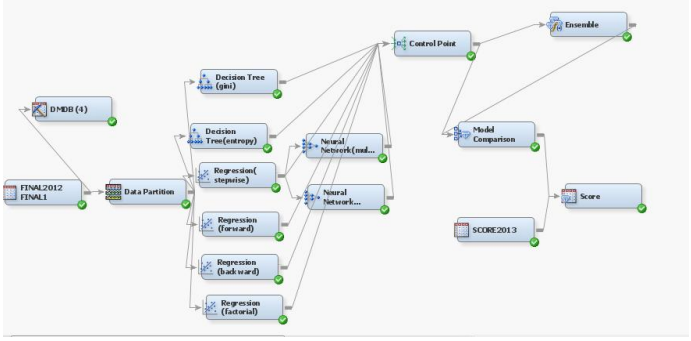

Figure 5: Process Flow Diagram

EXPLANATION OF THE BEST MODEL:

The best model, selected based on validation misclassification rate is the Logistic regression model with stepwise input selection and prediction accuracy of 75.33%

Selected Model	Model Description	Selection Criterion: Valid: Misclassification Rate
Y	Regression(stepwise)	0.244674
	Regression (forward)	0.244674
	Regression (backward)	0.244974
	Regression (factorial)	0.245575
	Neural Network(multi layer perception)	0.245725
	Ensemble	0.246625
	Decision Tree (gini)	0.247075
	Neural Network (ordinary radial-equal ...	0.247825
	Decision Tree(entropy)	0.248575

Figure 6. Model Selection Based on Validation Misclassification rate

Based on the validation misclassification rate, the step wise regression model at step 3 is optimal.

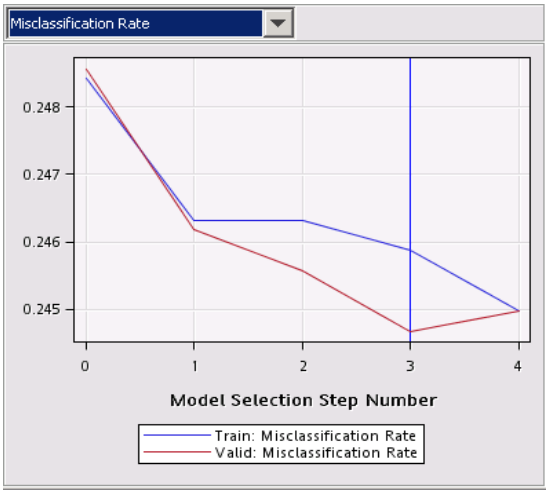

Figure 7. Model Optimization Based on Validation Misclassification Rate

From the odds ratio it can be observed that for one unit increase in the ' critical' (no. of critical violations) the odds of the restaurant grade being downgraded is 30% more.

Odds Ratio Estimates

Effect	Point Estimate
Critical	1.300
NonCritical	0.928
cricnt	1.017

Figure 8. Variables Odds Ratio.

SCORING:

The separate sampling technique is implemented to adjust the prior probabilities. The prior probabilities are adjusted as shown in Figure 9 . The best model, the stepwise regression model is used to score the 2013 data and the results are as shown in Table 5:

Level	Count	Prior
C	690	0.0518
B	2622	0.1967
A	10016	0.7515

Figure 9. Adjusting The Prior Probabilities For Scoring

Grade	Score	Training	Validation
A	97.91	92.02	91.59
B	1.85	6.9	7.1
C	0.23	1.03	1.2

Table 5. Scoring Results

SEQUENCE ANALYSIS

Sequence Analysis was performed on the entire 2012 and 2013 data. The data contains an ID field that identifies the restaurant number. There are a number of violations recorded per restaurant number which became the target of analysis. The sequence variable is taken as the inspection date, which contained entries from the year 2012 and 2013.

There are repeated violations for 10F, “Equipment not easily movable or sealed to floor, adjoining equipment, adjacent walls or ceiling. Aisle or workspace inadequate". Also, there are several critical and non-critical violations that occur along with the non-critical violation of 10F. These are:

- Critical violation 2G – cold storage violation, “Cold food item held above 41º F (smoked fish and reduced oxygen packaged foods above 38 ºF) except during necessary preparation”.
- Critical violation 4l – “Evidence of rats or live rats present in facility's food and/or non-food areas”.
- Critical violation 6D – “Food contact surface not properly washed, rinsed and sanitized after each use and following any activity when contamination may have occurred”.
- Critical violation 6C – “Food not protected from potential source of contamination during storage, preparation, transportation, display or service”.
- Non critical 8G violation, “ Facility not vermin proof. Harborage or conditions conducive to vermin exist”.

This clearly implies that 10 F is one violation, although non-critical, that the food inspectors and the administrators should look into and make sure that this violation is corrected across all the restaurants. Also, this violation should be included in the critical category because of its recurrent nature; and because it might be the root cause for many other critical violations.

Rule	Support (%)	Confidence (%)	Rule
1	43.55	55.31	10F ==> 10F
2	34.87	59.3	02G ==> 10F
3	34.47	60.74	08A ==> 10F
4	31.37	55.27	08A ==> 08A
5	29.73	50.85	02G ==> 02G
6	27.52	53.5	10B ==> 10F
7	27.44	60.37	04L ==> 10F
8	26.62	45.52	02G ==> 08A
9	25.5	56.09	04L ==> 04L
10	24.87	57.59	06C ==> 10F

Figure 8 . Sequence Analysis Results

CONCLUSION:

The statistical insights drawn from the data analysis will aid people in choosing better restaurants serving different cuisines across the five boroughs. The predictive model helps in forecasting the grade in future years. The results of sequence analysis give insights into the most frequently occurring critical and non-critical violations. These results can be used by both Departments of Health and Mental Hygiene to implement strict guidelines to curb the violations as well as take necessary corrective actions for every individual restaurants. Curbing one such non-critical violation, '10F' may resolve many other critical violations associated with it.

REFERENCES:

1. Source of Data: https://nycopendata.socrata.com

2. NYC DOHMH – A Guide for Food Service Operators: http://www.nyc.gov/html/doh/downloads/pdf/rii/blue-book.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Pruthvi Bhupathiraju Venkata
Organization: Oklahoma State University
Address: 74 S, university pl apt#1
City, State ZIP: Stillwater, OK- 74075
Work Phone: 201-218-5741
Email: pruthvi@okstate.edu

Pruthvi Bhupathiraju Venkata is a graduate student in Management Information Systems at Oklahoma State University. He has two years experience with SAS Data Mining and Analytical tools. He is a Base SAS certified programmer and SAS Certified Predictive Modeler using SAS Enterprise Miner 7.1. In 2013 he received SAS and OSU Data Mining Certificate.

Name: Dr. Goutam Chakraborty
Enterprise: Oklahoma State University
Address: Department of Marketing, Oklahoma State University
City, State ZIP: Stillwater, OK - 74074
Work Phone: (405)744-7644
E-mail: goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is a professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU business analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.