NUS
National University of Singapore

IS5152

DECISION MAKING TECHNOLOGIES

Group Project: Text Mining and Performance Tuning

Prepared for:

ASSOC PROF Huang Ke Wei

Prepared by:

CHEN Wanli Desiree

Shifang, ZOU

WEI, Yulai

Yu-Chang, TSAO

Date:

26th April 2017

Table of Contents

Section 1: Overview of Problem and Analytics Research Objectives

Introduction

The Economic Development Agency (called the "agency") of Tallmadge in the state of Ohio, situated in the United States, is the authority overseeing the economic growth of the city of Tallmadge. In the 2010 census, Tallmadge had a population of 17,537 [1].

Background

To promote the economic growth of restaurants in Tallmadge, the agency should attach importance to the ratings (also known as stars) on internet review forums such as Yelp, and engage restaurants to improve the restaurants' relatively weaker aspects. Such ratings impact customers' purchasing decisions to a large extent. According to a research [2] conducted by Harvard University, when the rating of a restaurant increases by one star, the revenue of that restaurant will increase by 5% to 9%. Therefore, if the agency can help restaurants enhance their performance on Yelp, it will greatly increase economic growth in Tallmadge.

Industry players

As part of industry development efforts, the agency would like to analyse potential factors contributing to a restaurant's rating, and in turn seek to ensure sustainable growth and vibrant business opportunities for the city. This research focuses on customer reviews of 25 food and beverages (F&B) businesses in Tallmadge and proposes areas of improvement for the F&B industry in Tallmadge.

Section 2: Data

Based on customer reviews, this study would like to develop a model to predict the ratings of F&B businesses in Tallmadge. The agency issues F&B licences to both new (wishing to set up a business in Tallmadge) and existing (currently running a business in Tallmadge) establishments. For new establishments, the model can assist the agency to gauge the performance of the potential establishment in Tallmadge. In an industry where there is a limit on retail space allocated to F&B businesses, it is imperative that new establishments complement, rather than cannibalise, the F&B scene in Tallmadge.

Variables created by text mining methods

From customer reviews, a **sentiment score** for each business is obtained. This sentiment score indicates if the business evokes a positive or negative sentiment. In addition, **three social indexes** are obtained. These indexes represent if the particular establishment attracts family-oriented crowds or business-oriented crowds. The establishment can employ different strategies if it were attracting a family-oriented crowd as compared to a business-oriented crowd. For example, establishments catering more towards families may choose to incorporate kid's meals in their menu, so that every member of the family, from the young to the old, gets to enjoy the experience at the establishment.

Next, the model considers whether the review is made by a customer who had previously visited the establishment (**repeat index** = 1) or who had visited the establishment for the first time (repeat index = 0). For establishments where there is a higher proportion of repeat customers, the establishment could create loyalty rewards programme for customers whereby customers earn points on spending which could be redeemed in subsequent visits. It is hoped that such loyalty rewards programme would help entrench customers and build up a pool of loyal customers of the establishments.

Lastly, the model uses the **five topic scores** obtained from the probability distribution of topics in its prediction of the rating for F&B businesses. Details of the topics can be found in Section 3.

Section 3: Methods

Summary statistics

To remove any potential selection bias, stratified sampling is applied to the data. Train Data is made up of 70% of the reviews from *each* business. Test Data is made up of 30% of the reviews from *each* business. The allocation of reviews from each business to either dataset is random.

Total number of reviews	373
Number of reviews in Train Data (~70% of data)	263
Number of reviews in Test Data (~30% of data)	110
Mean number of reviews per business (Base: All Data)	15
Median number of reviews per business (Base: All Data)	9
Maximum number of reviews per business (Base: All Data)	60
Minimum number of reviews per business (Base: All Data)	3

Word Cloud

To obtain a general sense of the reviews, a word cloud was created of the Train Data.

The top five most frequent words in the reviews are good (occurs 167 times), order (occurs 124 times), time (occurs 124 times), great (occurs 112 times) and service (occurs 107 times).

Dictionary Approach

Sentiment Score

The positive and negative dictionaries were used to calculate a Sentiment Score for each review.

$$Sentiment\ Score_{review} = Number\ of\ Positive\ Words - Number\ of\ Negative\ Words$$

The Sentiment Score for each business is the average of the Sentiment Score for the reviews relating to the business.

$Sentiment\ Score_{business} = \frac{1}{i}\sum_{1}^{i} Sentiment\ Score_{review}$, where i is the number of reviews for the business.

Base: All 25 businesses	**Sentiment Score**	
	Train	**Test**
Mean	2.8	2.2
Median	2.8	2.6
Maximum	5.8	7.0
Minimum	-1.3	-3.3
Standard Deviation	1.9	2.6

We compare the statistics of the scores for both Train and Test Data to check for robustness of both datasets. As there is a smaller number of reviews in the Test Data, the standard deviation of the Test Data is expected to be larger than that of the Train Data.

Social Indexes using three dictionaries

The three dictionaries are 1) Affpt containing affection words valuing love and friendship, 2) Kin containing words denoting kinship, and 3) HU containing words referencing social categories of humans.

Name of Dictionary	**Number of words in the dictionary**	**Examples of words in this dictionary**
Affpt	55	Cousin, Dad, Friend, Husband, Mother, Nephew, Sister, Son, Wife
Kin	50	Aunt, Brother, Cousin, Family, Father, Marriage, Parent, Relative, Wedding
HU	795	Adult, Colleague, Friend, Grandchildren, Husband, Lover, Manager, Sister, Wife

Base: All 25 businesses	**Affpt score**		**Kin score**		**HU score**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
Mean	0.16	0.09	0.24	0.15	1.44	1.06
Median	0.10	0.00	0.20	0.00	1.09	0.88
Maximum	0.75	1.00	1.00	1.00	6.33	3.75
Minimum	0.00	0.00	0.00	0.00	0.20	0.00
Standard Deviation	0.20	0.23	0.25	0.30	1.30	1.01

Text Classification

Using a supervised learning approach, Text Classification was done based on the manually-coded repeat index for each of the 373 reviews. This is a binary index with value 1 when the

review is made by a customer who had previously visited the establishment and value 0 when the review is made by a customer who had visited the establishment for the first time.

Four extracts of reviews manually coded with repeat index = 1

Quote My family and I have been going to Wally's for years. *Unquote*

Quote Love Sammies, we come here anytime we are out in Tallmadge… *Unquote*

Quote The food was better this time… *Unquote*

Quote … I used to come religiously every Friday but the last few times were such a bad experience that I won't be back. *Unquote*

Two extracts of reviews manually coded with repeat index = 0

Quote My wife and I came here for the first time after a recommendation… *Unquote*

Quote … We'll probably go again and try something else on the menu. *Unquote*

The table below shows the proportion of Repeat index manual coding for the Train and Test data.

Manual code **Base: All reviews**	**Train Data**	**Test Data**
Repeat index = 1	28%	36%
Repeat index = 0	82%	64%

Comparing performances of algorithms

The agency used **eight algorithms** to train models that were used subsequently for prediction. With a sparsity threshold of 0.998, terms that appear in more than 372 (this is 0.998 multiplied by 373) of the reviews were removed. The document term matrix has 3,471 terms.

Training of the model was done on the Train Data which contains 263 documents. The trained model was then used to predict on the Test Data which contains 110 documents. The classification accuracy shows how accurate the algorithm is in labelling the repeat index of each review in the Test Data.

$$Classification\ Accuracy = \frac{True\ Positives}{Number\ of\ Reviews}$$

Algorithm	**F-Score**	**Classification Accuracy**
SVM	0.690	0.755
SLDA	0.550	0.609
Logitboost	0.690	0.736
Bagging	**0.755**	**0.791**
Random Forest	0.560	0.700
GLMNET	0.665	0.745
Tree	0.650	0.709
Max Entropy	0.720	0.755

Based on the F-Score and Classification Accuracy, the best algorithm for Text Classification is Bagging.

Manual coding

The repeat index for each business is calculated as follows.

$Repeat\ Index_{business} = \frac{1}{i}\sum_{1}^{i} Repeat\ Index_{review}$, where i is the number of reviews for the business.

Even though the agency could obtain the predictions from each of the eight algorithms used in the previous sub-section, for prediction purposes in the main model, the agency used the

manually-coded repeat index of each review. This is to ensure greater accuracy of the data points used in the main model. The Classification Accuracy from the previous sub-section shows that even the best algorithm, which was Bagging in this case, predicted correctly about 79% of the time. The manually-coded repeat index is the most accurate, as that was coded by personnel at the agency after having read through each of the review.

Base: All 25 businesses	**Repeat Index obtained from Manual Coding**	
	Train	**Test**
Mean	0.24	0.34
Median	0.25	0.38
Maximum	0.55	1.00
Minimum	0.00	0.00
Standard Deviation	0.18	0.32

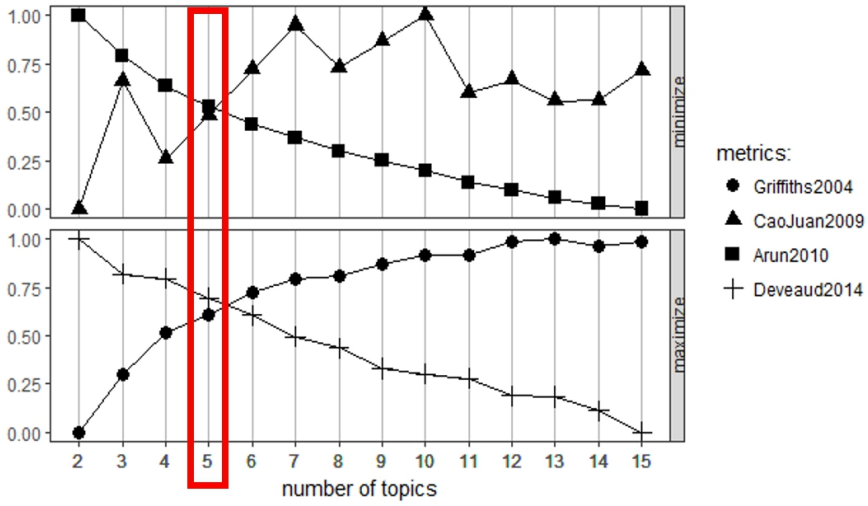
Topic Modeling

For topic modeling, the agency used a sparsity threshold of 0.995. Terms that appear in more than 371 (this is 0.995 multiplied by 373) of the reviews were removed. The document term matrix has 1,671 terms. Due to the difficulty in deriving the individual topics from the frequent terms, the sparsity threshold used for Topic Modeling is less than that for Text Classification (sparsity threshold of 0.998). This is so that the number of terms is reduced and the algorithm can train more effectively on a smaller set of terms. It was also easier to deduce the topics from the lists of terms generated.

Choosing optimal number of topics

In choosing the number of topics for this unsupervised learning approach, the agency employed metrics developed in previous research. These metrics are calculated based on the

number of chosen topics. Both Juan et. al [3] and Arun et. al [4] developed metrics that are to be minimised. Both Griffiths et. al [5] and Deveaud et. al [6] developed metrics that are to be maximised. The ldatuning package in R is used to plot the metrics relating to the number of topics for Topic Modeling. Based on the metrics, we chose five as the optimal number of topics for Topic Modeling by Latent Dirichlet allocation (LDA).

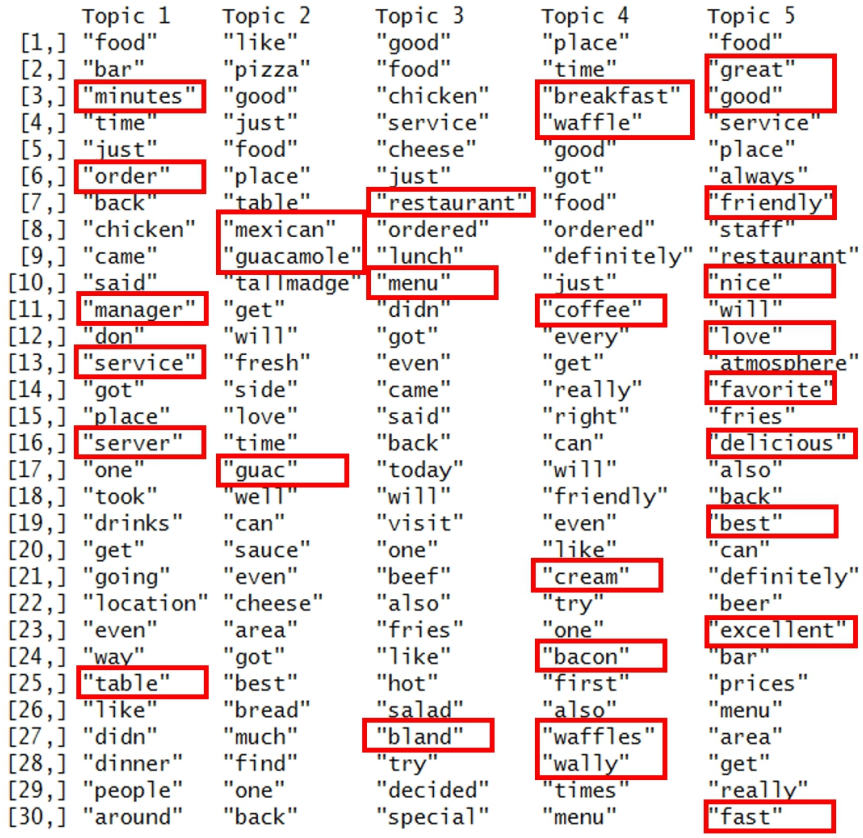


Subject of each topic

The topic allocation is as follows.

	Proportion of reviews allocated (%)	
	Train	**Test**
Topic 1	15	23
Topic 2	16	14
Topic 3	15	15
Topic 4	17	13
Topic 5	37	35

The top 30 most frequent words of each of the five topics are shown in the figure below. Based on the highlighted words, the agency derived the related topic for each grouping of words. Topic 1 is related to service standards. Topic 2 is related to Mexican cuisine. Topic 3 consists of a list of words describing a normal experience at the establishment. Topic 4 is related to breakfast and waffles. Topic 5 consists of a list of words describing a positive experience at the establishment.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[1,]	"food"	"like"	"good"	"place"	"food"
[2,]	"bar"	"pizza"	"food"	"time"	"great"
[3,]	"minutes"	"good"	"chicken"	"breakfast"	"good"
[4,]	"time"	"just"	"service"	"waffle"	"service"
[5,]	"just"	"food"	"cheese"	"good"	"place"
[6,]	"order"	"place"	"just"	"got"	"always"
[7,]	"back"	"table"	"restaurant"	"food"	"friendly"
[8,]	"chicken"	"mexican"	"ordered"	"ordered"	"staff"
[9,]	"came"	"guacamole"	"lunch"	"definitely"	"restaurant"
[10,]	"said"	"tallmadge"	"menu"	"just"	"nice"
[11,]	"manager"	"get"	"didn"	"coffee"	"will"
[12,]	"don"	"will"	"got"	"every"	"love"
[13,]	"service"	"fresh"	"even"	"get"	"atmosphere"
[14,]	"got"	"side"	"came"	"really"	"favorite"
[15,]	"place"	"love"	"said"	"right"	"fries"
[16,]	"server"	"time"	"back"	"can"	"delicious"
[17,]	"one"	"guac"	"today"	"will"	"also"
[18,]	"took"	"well"	"will"	"friendly"	"back"
[19,]	"drinks"	"can"	"visit"	"even"	"best"
[20,]	"get"	"sauce"	"one"	"like"	"can"
[21,]	"going"	"even"	"beef"	"cream"	"definitely"
[22,]	"location"	"cheese"	"also"	"try"	"beer"
[23,]	"even"	"area"	"fries"	"one"	"excellent"
[24,]	"way"	"got"	"like"	"bacon"	"bar"
[25,]	"table"	"best"	"hot"	"first"	"prices"
[26,]	"like"	"bread"	"salad"	"also"	"menu"
[27,]	"didn"	"much"	"bland"	"waffles"	"area"
[28,]	"dinner"	"find"	"try"	"wally"	"get"
[29,]	"people"	"one"	"decided"	"times"	"really"
[30,]	"around"	"back"	"special"	"menu"	"fast"

Below are extracts of reviews allocated to the various topics.

Topic 1 relating to service standards

Quote Rude service!! … Delivery takes an hour. Won't be going again. *Unquote*

Topic 2 relating to Mexican cuisine

Quote Very authentic, small, no wait (7 pm Saturday) enjoying the golden pitcher. Very good guacamole and tacos. *Unquote*

Topic 3 describing a normal experience at the establishment

Quote Wasn't super impressed the first time but that was early after the opened so decided to give it another try. Unfortunately I'm still not real impressed. *Unquote*

Topic 4 relating to breakfast and waffles

Quote These are the best waffles I've ever had. They were light, not heavy as waffles can sometimes be, but generously portioned. They were slightly crispy on the outside, chewy on the inside and perfectly flavored. *Unquote*

Topic 5 describing a positive experience at the establishment

Quote Beautifully decorated restaurant, great service, and amazing salsa. The food was impressive and very tasty. Definitely recommend this place! *Unquote*

Topic scores for each business

The topic scores for each business are calculated by the equation below. As this topic score is an *average* of the probability distributions of the reviews for this business and for this topic, the topic scores do not add up to 1.

$$Score_{business,topic} = \frac{1}{i}\sum_{1}^{i} Probability\ distribution_{review,i,\,topic}$$

The tables on the next page show the probability distributions of each of the five topics for the Train Data and Test Data.

Base: 25 businesses in Train Data	**Probability Distribution of Topics**				
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Mean	0.20	0.16	0.15	0.20	0.29
Median	0.13	0.05	0.14	0.14	0.29
Maximum	0.63	0.76	0.39	0.81	0.77
Minimum	0.0024	0.0011	0.0021	0.0029	0.0011
Standard Deviation	0.19	0.21	0.11	0.19	0.21

Base: 25 businesses in Test Data	**Probability Distribution of Topics**				
	Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Mean	0.29	0.15	0.13	0.14	0.29
Median	0.19	0.04	0.06	0.01	0.25
Maximum	0.91	0.74	0.64	0.74	0.98
Minimum	0.0018	0.0013	0.0011	0.0011	0.0013
Standard Deviation	0.28	0.22	0.17	0.22	0.29

Building the models

The 10 new variables created from text mining are Sentiment Score, Affpt Score, Kin Score, HU Score, Repeat Index, Topic 1 Score, Topic 2 Score, Topic 3 Score, Topic 4 Score and Topic 5 Score.

Two variables from external data and missing values handling

External data on Price and Quality are sourced from Yelp and TripAdvisor respectively. Yelp represents the price level of a restaurant on a scale of $ to $$$$. The four levels and their corresponding price range are listed in the diagram below.

Level	Price range as classified by Yelp	Category as classified by Yelp
$	≤ $10	Inexpensive
$$	Between $11 and $30	Moderate
$$$	Between $31 and $60	Pricey
$$$$	> $60	Ultra high-end

Quality represents how the restaurant compares to other restaurants serving the same type of cuisine, be it fast-food, dessert places or general restaurant types. It is calculated by taking the rank of the restaurant divided by the total number of restaurants of that same type. For example, Nuevo Acapulco is ranked 5th out of 22 restaurants in Tallmadge city. Thus, its Quality index (used interchangeably with Rank index) is $5/22$ which is 0.227. The smaller the Rank Index, the better the quality of the restaurant. As the restaurants are compared across restaurants serving the same type of cuisine, the Rank Index calculated for each business is comparable across the 25 businesses in this research. Not all the Rank Indexes can be found from information obtained in Trip Advisor. For missing Rank Indexes, the agency used the Ordinary Least Squares (OLS) method to predict missing values.

Features transformation

The variable Repeat Index on its own only indicates that the customer had previously visited the establishment, but this does not affect the Business rating which the model is built for. The agency conducted feature transformation by **multiplying Sentiment Score with Repeat**

Index to form a **new variable**. As the Repeat Index for the business is the average of the Repeat Index for the reviews relating to the particular business (and is not a binary value), the segmentation is done whereby businesses with Repeat Index greater than the Business Repeat Index median value of 0.25 are considered as businesses for which reviews tend to be made by customers who had previously visited the establishment. Businesses with Repeat Index less than or equal to 0.25 are considered as businesses for which reviews tend to be made by customers who had visited the establishment for the first time. The various combinations of Sentiment Score and Repeat Index are shown and explained below.

		Sentiment Score	
		Positive	**Negative**
Repeat Index	$> \mathbf{0.25}$	Business has a positive sentiment and relatively many of the reviews are made by repeat customers.	Business has a negative sentiment and relatively many of the reviews are made by repeat customers.
	$\leq \mathbf{0.25}$	Business has a positive sentiment and relatively few of the reviews are made by repeat customers.	Business has a negative sentiment and relatively few of the reviews are made by repeat customers.

Another aspect of Features transformation is in **Features scaling**. The variables are scaled in the range of 0 and 1. Features scaling is done so that the machine learning algorithm can work well and is able to converge more quickly.

$Scaled\ value, x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, where x is the original value.

Features selection

Several Linear Regression models were built with Business Ratings (known as Stars in the Business dataset) being the dependent variable and the relevant variables such as Price, Rank Index, Sentiment Score, Repeat Index, Sentiment Score * Repeat Index, Affpt Score, Kin Score, HU Score, Topic 1 Score, Topic 2 Score, Topic 3 Score, Topic 4 Score and Topic 5 Score being the independent variables. Features selection is done by the **AIC method**.

Ensemble Method and Parameters Tuning

Using the variables selected by the AIC method, the agency applied the **SVM algorithm** to the dataset. Using the Caret package, 10-fold cross validation was conducted to obtain preliminary parameter values for sigma and C. The parameter values were tuned by trying other parameter values for the classifier.

For the Ensemble Method, the agency applied a **combination of Bagging and SVM**. The data was sampled 10 times to form 10 bags. SVM was applied to each bag and the best parameters for gamma and cost were obtained. The best parameters were then used in the model built by the SVM algorithm. Four kernel types (linear, polynomial, radial and sigmoid) were used and predictions were made based on each of these four kernel types. For each business, the 10 bags provide 10 predictions in total. The average of the 10 predictions for each business is used for calculating the performance metrics of the models.

Using the models for prediction

The models are trained on the Train Data and predictions are done on the Test Data. The root mean squared error on the Test Data is used to assess the performance of the models.

A summary of the modeling is shown on the next page.

Steps taken	**Variables in training model**
Build a Baseline Model	Star ~ 10 text mining variables + Price variable obtained from external data
Conduct Features transformation to create a new variable that is the product of Sentiment Score and Repeat Index (for ease of reference, we call the new variable "Impression"). Build a model and apply StepAIC to the model.	Star ~ 10 text mining variables + Price + Impression variable
Use OLS to predict missing Rank Indexes. To prevent multicollinearity in the later part of our modeling, the variables in this model are the five variables that were removed by StepAIC previously.	Rank index ~ Sentiment Score + Affpt Score + Kin Score + HU Score + Topic 5 Score
Build a model on the seven variables from the previous StepAIC model and include the Rank Indexes which now have no missing values. Apply StepAIC to the final model.	Star ~ Repeat Index + Topic 1 Score + Topic 2 Score + Topic 3 Score + Topic 4 Score + Price + Impression + Rank
Conduct Features scaling on the variables in the Final Model.	Same as above.
Use SVM on the variables in the Final Model. Tune SVM.	Same as above.
Use Ensemble Method of Bagging and SVM on the variables in the Final Model. Tune SVM.	Same as above.

Section 4: Baseline Results

The results are shown for the businesses that have the top 10 highest number of reviews. Due to limited resources, the agency would like to engage businesses with relatively higher number of reviews. Our industry development efforts can subsequently be applied to relatively smaller businesses in the future. While our initial strategy focuses on the larger businesses, for our predictive model described in Section 5, we used the variables relating to *all* 25 businesses.

Dictionary Approach

Sentiment Score

The more popular (characterised by relatively high review count) businesses have above average sentiment scores. This shows that the F&B industry in Tallmadge is doing relatively well.

	Name of Business	Review Count	Sentiment Score	
			Train	Test
1	Sammies Bar & Grill	60	4.5	3.7
2	Nuevo Acapulco	47	4.7	4.9
3	La Mexicana Cantina and Grill	28	3.8	3.8
4	Wally Waffle	27	4.4	2.6
5	El Tren Grill	26	4.8	4.0
6	Delanie's Neighborhood Grille	24	4.3	4.3
7	Firehouse Grill & Pub	23	2.6	-1.3
8	Cortabella's Italian Eatery	15	5.8	0.8
9	Gionino's Pizzeria	14	1.9	4.3
10	Seven Grains Natural Market	14	3.9	0.8

Social Indexes

The most popular business, Sammies Bar & Grill, tends to attract family-oriented crowds as seen from the relatively high social indexes. On the other hand, La Mexicana Cantina and Grill, which is the third most popular business, has relatively low social indexes. This means

that the business tends to attract business-oriented crowds. The next table shows the Social Indexes of the top 10 most popular businesses.

Name of Business		Affpt score		Kin score		HU score	
		Train	**Test**	**Train**	**Test**	**Train**	**Test**
1	Sammies Bar & Grill	0.40	0.11	0.60	0.22	1.57	1.67
2	Nuevo Acapulco	0.24	0.14	0.24	0.21	1.52	0.64
3	La Mexicana Cantina and Grill	0.05	0.00	0.30	0.00	0.90	0.88
4	Wally Waffle	0.16	0.00	0.26	0.00	1.05	0.50
5	El Tren Grill	0.11	0.13	0.11	0.13	0.33	0.50
6	Delanie's Neighborhood Grille	0.06	0.14	0.12	0.29	1.12	1.43
7	Firehouse Grill & Pub	0.13	0.57	0.25	0.71	1.69	3.00
8	Cortabella's Italian Eatery	0.45	0.00	0.55	0.00	1.09	1.25
9	Gionino's Pizzeria	0.20	0.00	0.30	0.00	1.00	0.75
10	Seven Grains Natural Market	0.10	0.25	0.20	0.25	1.70	3.75

Text Classification

Even among the popular businesses, there is a mixture of those with a repeat crowd and those with a first-time crowd. Despite the popularity of Sammies Bar & Grill, its repeat index is less than the median Repeat Index. This indicates that many people visit Sammies Bar & Grill perhaps for first-time special occasions. There may be reasons why people tend not to revisit Sammies Bar & Grill, but the scope of this research does not delve into the reasons why. The second most popular business, Nuevo Acapulco, has a relatively high Repeat Index and this is consistent for both the Train Data and Test Data. This shows that Nuevo Acapulco tends to attract a repeat crowd. The Repeat Indexes of the top 10 most popular businesses are shown on the next page.

	Name of Business	**Repeat Index**	
		Train	**Test**
1	Sammies Bar & Grill	0.21	0.28
2	Nuevo Acapulco	0.55	0.57
3	La Mexicana Cantina and Grill	0.10	0.63
4	Wally Waffle	0.42	0.38
5	El Tren Grill	0.22	0.25
6	Delanie's Neighborhood Grille	0.35	0.43
7	Firehouse Grill & Pub	0.19	0.14
8	Cortabella's Italian Eatery	0.27	0.00
9	Gionino's Pizzeria	0.20	0.50
10	Seven Grains Natural Market	0.30	0.50

Topic Modeling

Topic Modeling does not reflect much accurate allocation of topics to reviews. It was expected that reviews from businesses such as La Mexicana Cantina and Grill and El Tren Grill serving Mexican food would reflect the most probable topic as Topic 2 which was related to Mexican cuisine. However, both businesses reflected Topic 4 relating to waffles and breakfast. It was also expected that reviews about Wally Waffle would reflect Topic 4 as its most probable topic, however those reviews resulted in Topic 1 relating to service standards as the most probable topic. Topic 5 relating to a positive experience had the highest proportion of reviews allocated to it. However, none of the top 10 most popular businesses reflected it as the most probable topic. Despite the inaccuracy of Topic Modeling, the approach is robust as both Train Data and Test Data reflected similar topics as the most probable for the business.

Name of Business		**Most probable topic**	
		Train	**Test**
1	Sammies Bar & Grill	4	4
2	Nuevo Acapulco	4	4
3	La Mexicana Cantina and Grill	4	4
4	Wally Waffle	1	1
5	El Tren Grill	4	4
6	Delanie's Neighborhood Grille	4	4
7	Firehouse Grill & Pub	3	3
8	Cortabella's Italian Eatery	3	3
9	Gionino's Pizzeria	1	1
10	Seven Grains Natural Market	2	2

Industry segmentation

The new variable formed by Features Transformation, that is, multiplying Sentiment Score and Repeat Index, allows the agency to segment the F&B businesses and provide customised strategies for these businesses. Details are found in the SWOT strategy analysis below. The examples below are a non-exhaustive list of the F&B businesses in Tallmadge.

		Sentiment Score	
		Positive	**Negative**
Repeat Index	> 0.25	**Strength:** Business to continue maintaining what they are good at. Examples are Nuevo Acapulco, Wally Waffle, Delanie's Neighborhood Grille, Cortabella's Italian Eatery, Seven Grains Natural Market.	**Threat:** Customers are visiting again (assuming they are rational and would visit again only if previous visits were positive experiences), but the reviews now are negative. Business seems to be declining in its utility to customer. An example is White House Chicken.
	≤ 0.25	**Opportunity:** Business to work at entrenching their new customers. Examples are Sammies Bar & Grill, La Mexicana Cantina and Grill, El Tren Grill. Firehouse Grill & Pub, Gionino's Pizzeria.	**Weakness:** Customers visit once and form a negative impression of the business. Business to find out what they are weak at and seek to improve in those aspects. An example is Taco Bell.

Variables in final model

The independent variables, selected by the AIC method, in the Final model are Repeat Index, Topic 1 Score, Topic 2 Score, Topic 3 Score, Topic 4 Score, Price, Impression, and Rank. The variables Sentiment Score, Affpt Score, Kin Score, HU Score and Topic 5 Score were insignificant and removed from the model after stepAIC was applied.

From the linear regression models, the variables Repeat Index, Topic 1 Score, Topic 2 Score, Topic 3 Score, Topic 4 Score, Price, and Rank have a negative correlation to Business Rating. Only the Impression variable has a positive correlation to Business Rating.

When the Repeat Index is low, which means that the customer had visited the establishment for the first time, the Business Rating increases. When the Repeat Index is high, which means that the customer had previously visited the establishment, the Business rating decreases. This may seem to suggest that restaurants need to provide new improved experiences in subsequent visits to continue to provide utility to the customer. The customer may get bored of the current offerings of the restaurants, if no changes are made at the restaurant.

Out of all the four topic scores in the final model, only Topic 1 Score relating to service standards is significant. This shows that it is service standards, and perhaps less so the cuisine, that has a greater impact on Business rating.

Understandably, price has a negative correlation to Business rating. When price increases, customers tend to experience a decrease in utility and hence, Business rating decreases.

As explained previously in Section 3, the smaller the Rank Index, the better the quality of the restaurant. The model also shows that when the Rank decreases, which means that the restaurant is of a better quality, Business rating increases.

The Impression variable is the product of Sentiment Score and Repeat Index. When the Impression variable increases, Business rating increases. This is explained by, when a rational customer gives a positive review (which means that the Sentiment Score is positive), the customer will give a higher Business rating.

Results of SVM models and Ensemble Method models

Using the Baseline Model, the Root Mean Squared Error (RMSE) of the Test Data is 1.02. After taking steps to improve prediction performance, the RMSE of Test Data using SVM models and Ensemble Method models are lower.

Model	**RMSE of Test Data**
Baseline Model	1.02
After Missing Value handling, Features selection using stepAIC, Features Transformation, that is, creating a new variable and Features scaling.	
SVM model without parameters tuning sigma = 0.05701049, C = 1	0.56
SVM model with parameters tuning sigma = 0.05701049, C = 5	0.65
Ensemble method using Bagging and SVM linear kernel	0.76
Ensemble method using Bagging and SVM polynomial kernel	0.97
Ensemble method using Bagging and SVM radial kernel	0.67
Ensemble method using Bagging and SVM sigmoid kernel	0.63

Section 5: Performance Improvement

Root Mean Squared Error of Test Data

Features scaling resulted in performance improvement. The RMSEs of Test Data predicted on scaled variables are lower than that predicted on unscaled variables.

Model	**RMSE of Test Data**	
	Before features scaling	**After features scaling**
Baseline Model	1.02	Not applicable
SVM model without parameters tuning sigma = 0.05701049, C = 1	0.59	0.56
SVM model with parameters tuning sigma = 0.05701049, C = 5	0.73	0.65
Ensemble method using Bagging and SVM linear kernel	1.08	0.76
Ensemble method using Bagging and SVM polynomial kernel	0.95	0.97
Ensemble method using Bagging and SVM radial kernel	0.78	0.67
Ensemble method using Bagging and SVM sigmoid kernel	0.73	0.63

Root Mean Squared Error of Train Data

Missing Value handling, Features selection using stepAIC, Features Transformation, that is, creating a new variable and Features scaling resulted in a decrease in the RMSE of Train Data.

Model	RMSE of Train Data
Baseline Model	0.5221
Conduct Features transformation to create a new variable. Build a model.	0.5216
Conduct Features transformation to create a new variable. Build a model and **apply StepAIC to the model**.	0.4693
Use OLS to predict missing Rank Indexes. Build a model on the seven variables from the previous StepAIC model and include the Rank Indexes which now have no missing values. **Apply StepAIC to the final model**.	0.4352

Section 6: Conclusion

This research revealed the potential factors contributing to a restaurant's Business rating. Based on these factors, industry segmentation was done, so that the Economic Development Agency of Tallmadge can better partner establishments to grow their businesses.

The model can also be used for prediction of Business ratings. This is especially crucial, as the licensing department in the agency receives many applications from F&B establishments wishing to enter the F&B scene in Tallmadge and resources are limited. In an industry where there is a limit on retail space allocated to F&B businesses, the agency can utilise the model to gauge the performance of an establishment and choose to issue licences to more viable

establishments. In turn, this will help to ensure sustainable economic growth and vibrant business opportunities for the city of Tallmadge.

In this research, the focus was on F&B businesses in Tallmadge. The methodology can be applied to other cities or to other industries such as the Beauty and Spa industry.

References

[1] *Tallmadge, Ohio.* Retrieved April 20, 2017, from https://en.wikipedia.org/wiki/Tallmadge,_Ohio

[2] Anderson, M., Magruder, J. (2012). Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database. *The Economic Journal, 122*(563), 957-989.

[3] Juan, C., Tian, X., Jintao, L., Yongdong, Z., Sheng, T. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(7-9), 1775-1781.

[4] Arun, R. Suresh, V., Veni Madhavan C.E., Narasimha Murthy M.N. (2010, June 21-24). *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations.* Paper presented at Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. DOI:10.1007/978-3-642-13657-3_43

[5] Griffiths, T.L., Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(supplement 1), 5228-5235.

[6] Deveaud, R., Sanjuan, E., Bellot, P. (2014). *Accurate and effective latent concept modeling for adhoc information retrieval.* DOI:10.3166/DN.17.1.61-84