ISM 6136 Data Mining

Final Project

Review-Based Restaurant Performance Analysis

Darya Gahramanova

Anirudh Govindraj

Apeksha Bhise

Nikhitha Sama

Executive Summary

In this project, Yelp's challenge dataset on different businesses in US and Canada, their attributes and user reviews was used to analyze and predict factors affecting restaurants' performance and customer satisfaction. The paper covers following topics:

- Introduction, data set overview
- Data cleaning and preprocessing, R
- Predicting restaurant ratings, Multiclass decision trees, SAS Miner
- Exploring associations, SAS Miner
- Predicting restaurant closure, Two-class Bayes Point machine, Boosted decision tree, Azure ML
- Cluster analysis, K-Means, Azure ML, R
- Sentiment analysis (text mining), R

Resulting models provide valuable insights on the importance of business attributes in user ratings, finding similar restaurants based on customer visits and restaurant attributes, predicting business closure based of features, and words associated with good or bad reviews. The models can be applied to predict different scenarios in future, and businesses explored in the project can benefit from new findings affecting their performance.

Introduction

Customer satisfaction continues to be a crucial determinant in businesses' performance. With increasing diversity and competition, restaurants especially value customer feedback, as even small details that the businesses might overlook can significantly affect performance. The main purpose of this project is to gain insights and make predictions on marketing-related issues in food business. A large and reliable dataset was required to have enough records and variables to make significant predictions and discover new findings.

Yelp is one of the most popular business directory services – it has over 135 million business and restaurant reviews worldwide. The company released its records of hundreds of thousands businesses across US and Canada with their features, reviews, ratings, and user attributes for its data analytics challenge on their official page: yelp.com/dataset.

This dataset satisfied our requirements in terms of reliability, number of records and features. Although it is a public dataset and many people performed analysis on it, the scope of work that can be done on it is boundless. Since dataset is very large, some people performed very specific analysis on it, chose a narrow target, used only certain tools, or only some of the variables. This project is different in a way that it is narrowing down to restaurant businesses only, using extensive data preprocessing but saving as many variables as possible, using different tools for both supervised and unsupervised learning. Most importantly, this analysis aims to provide real-life business implications that should be applicable to real customers.

The main challenge was to make the enormous dataset readable for machine learning, as well as significant for the analysis without getting confused. This will be discussed in the following section.

Preprocessing

It consists of 6 .json files, 4 of which were used for the analysis:

business.json: *business_id, name, address, city, state, postal_code, latitude, longitude, stars, review_count, is_open, attributes (nested), categories (nested), hours*

checkin.json: *business_id, weekday, hour, checkins*

review.json: *review_id, user_id, business_id, stars, useful, funny, cool, text*

user.json: *user_id, name, review_count, yelping_since, useful, funny, cool, elite, friends, fans, average_stars, compliment_hot, compliment_more, compliment_profile, compliment_cute, compliment_list, compliment_note, compliment_plain, compliment_cool, compliment_funny, compliment_writer, compliment_photos*

Photos posted and tips left by the reviewers were left out of the scope of this analysis.

At first, none of the data mining software that was aimed to be utilized in the analysis would open these files. The problem turned out to be in inconsistencies in json formatting convention used by Yelp. As a result, readily converted .csv files of Yelp dataset were obtained from Kaggle, but errors in data were spotted later during analysis.

Those files have been terminated and replaced by original json files from Yelp again. R was used to convert these datasets into more readable format. While *rjson* library was helpful in reading in most of the data, it failed to recognize Yelp's non-standard usage of nested loops, particularly in *business* dataset. These problematic columns contained large amount of business attributes and categories that were crucial for the analysis, so manual approach was used to break this data into columns.

First, *data.table* library was used to spot only restaurant-related businesses in *categories* column, which reduced the business dataset from ~200,000 to ~60,000 instances.

```
yelp_business_small <- yelp_json[yelp_json$categories %like% "Restaurants" |
yelp_json$categories %like% "Coffee & Tea" | yelp_json$categories %like% "Fast Food",]
```

Furthermore, the unreadable *attributes* column was broken down into separate indicator variables with binary records. Discretion was used to merge some attribute categories, but none of the attributes were omitted. An example:

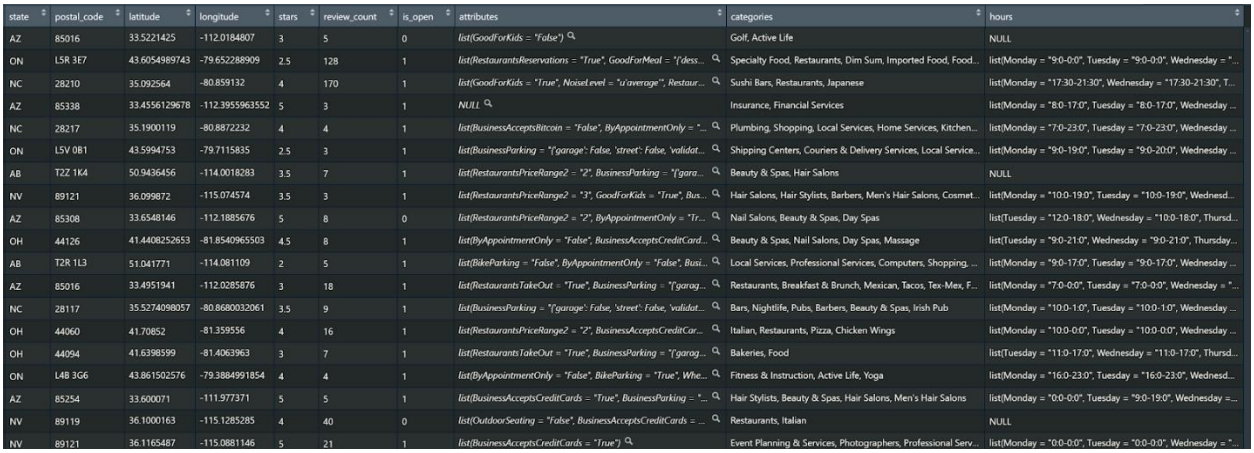
```
for (i in 1:nrow(at)) {
  if ((grepl("'dessert': False", at$GoodForMeal[i]) == TRUE)){
  at$good.for.dessert[i] = 0}
  else if ((grepl("'dessert': True", at$GoodForMeal[i]) == TRUE)){
  at$good.for.dessert[i] = 1}
  else{
```

```
    at$good.for.dessert[i] = ""}
}
```

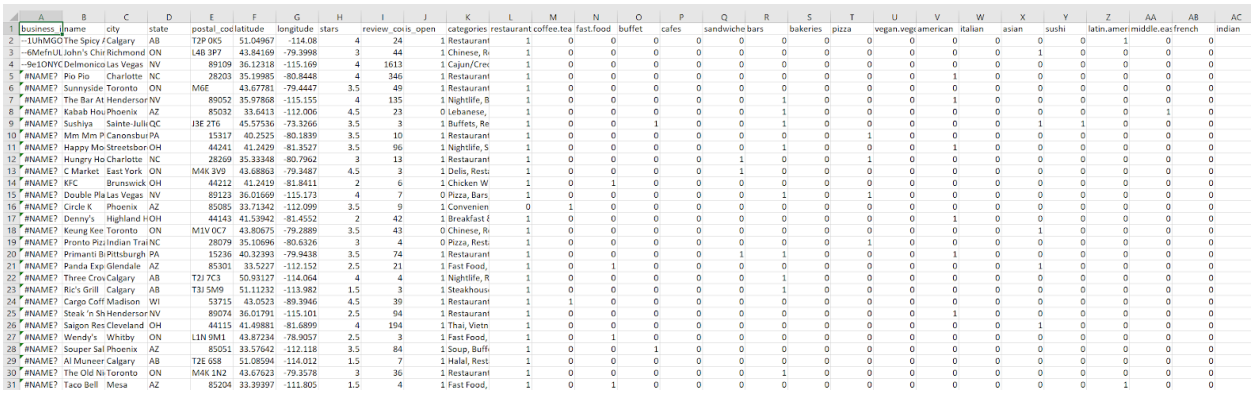
Resulting dataset had the following business attributes: *price.range, noise.level, reservations, reservations.only, table.service, delivery, caters, takeout, drive.thru, wifi, has.tv, attire, coat.check, counter.service, credit.cards, dogs.allowed, happy.hour, open.24.hours, alcohol, smoking, corkage, accepts.bitcoin, wheelchair.access, outdoor.seating, bike.parking, parking.garage, parking.lot, parking.street, parking.valet, parking.validated, ages.adults.only, ages.all.ages, good.for.dancing, good.for.kids, good.for.groups, good.for.breakfast, good.for.brunch, good.for.lunch, good.for.dinner, good.for.latenight, good.for.dessert, ambience.casual, ambience.classy, ambience.divey, ambience.hipster, ambience.intimate, ambience.romantic, ambience.touristy, ambience.trendy, ambience.upscale, music.background, music.dj, music.jukebox, music.karaoke, music.live, music.video, no.music, bestnight.monday, bestnight.tuesday, bestnight.wednesday, bestnight.thursday, bestnight.friday, bestnight.saturday, bestnight.sunday, diet.dairy.free, diet.gluten.free, diet.halal, diet.kosher, diet.soy.free, diet.vegan, diet.vegetarian, open.monday, open.tuesday, open.wednesday, open.thrursday, open.friday, open.saturday, open.sunday.*

A similar technique was used to extract separate business categories as indicator variables from the *categories* column. In this case, only those categories that deemed to be valuable for the analysis were extracted: *restaurants, coffee.tea, fast.food, buffet, cafes, sandwiches, bars, bakeries, pizza, vegan.vegetarian, american, italian, asian, sushi, latin.american, middle.eastern, french, indian, canadian.*

Resulting final *business* dataset had 108 columns, most of which were binary. Before and after file snapshots:



state	postal_code	latitude	longitude	stars	review_count	is_open	attributes	categories	hours
AZ	85016	33.5221425	-112.0184807	3	5	0	list(GoodForKids = "False")	Golf, Active Life	NULL
ON	L5R 3E7	43.6054989743	-79.652288909	2.5	128	1	list(RestaurantsReservations = "True", GoodForMeal = "{'dess...	Specialty Food, Restaurants, Dim Sum, Imported Food, Food...	list(Monday = "9:0-0:0", Tuesday = "9:0-0:0", Wednesday = "...
NC	28210	35.092564	-80.859132	4	170	1	list(GoodForKids = "True", NoiseLevel = "u'average'", Restaur...	Sushi Bars, Restaurants, Japanese	list(Monday = "17:30-21:30", Wednesday = "17:30-21:30", T...
AZ	85338	33.4556129678	-112.3955963552	5	3	1	NULL	Insurance, Financial Services	list(Monday = "8:0-17:0", Tuesday = "8:0-17:0", Wednesday ...
NC	28217	35.1900119	-80.8872232	4	4	1	list(BusinessAcceptsBitcoin = "False", ByAppointmentOnly = "...	Plumbing, Shopping, Local Services, Home Services, Kitchen...	list(Monday = "7:0-23:0", Tuesday = "7:0-23:0", Wednesday ...
ON	L5V 0B1	43.5994753	-79.7115835	2.5	3	1	list(BusinessParking = "{'garage': False, 'street': False, 'validat...	Shipping Centers, Couriers & Delivery Services, Local Service...	list(Monday = "9:0-19:0", Tuesday = "9:0-20:0", Wednesday ...
AB	T2Z 1K4	50.9436456	-114.0018283	3.5	7	1	list(RestaurantsPriceRange2 = "2", BusinessParking = "{'gara...	Beauty & Spas, Hair Salons	NULL
NV	89121	36.099872	-115.074574	3.5	3	1	list(RestaurantsPriceRange2 = "3", GoodForKids = "True", Bus...	Hair Salons, Hair Stylists, Barbers, Men's Hair Salons, Cosmet...	list(Monday = "10:0-19:0", Tuesday = "10:0-19:0", Wednesd...
AZ	85308	33.6548146	-112.1885676	5	8	0	list(RestaurantsPriceRange2 = "2", ByAppointmentOnly = "Tr...	Nail Salons, Beauty & Spas, Day Spas	list(Tuesday = "12:0-18:0", Wednesday = "10:0-18:0", Thursd...
OH	44126	41.4408252653	-81.8540965503	4.5	8	1	list(ByAppointmentOnly = "False", BusinessAcceptsCreditCard...	Beauty & Spas, Nail Salons, Day Spas, Massage	list(Tuesday = "9:0-21:0", Wednesday = "9:0-21:0", Thursday...
AB	T2R 1L3	51.041771	-114.081109	2	5	1	list(BikeParking = "False", ByAppointmentOnly = "False", Busi...	Local Services, Professional Services, Computers, Shopping, ...	list(Monday = "9:0-17:0", Tuesday = "9:0-17:0", Wednesday ...
AZ	85016	33.4951941	-112.0285876	3	18	1	list(RestaurantsTakeOut = "True", BusinessParking = "{'garag...	Restaurants, Breakfast & Brunch, Mexican, Tacos, Tex-Mex, F...	list(Monday = "7:0-0:0", Tuesday = "7:0-0:0", Wednesday = "...
NC	28117	35.5274098057	-80.8680032061	3.5	9	1	list(BusinessParking = "{'garage': False, 'street': False, 'validat...	Bars, Nightlife, Pubs, Barbers, Beauty & Spas, Irish Pub	list(Monday = "10:0-1:0", Tuesday = "10:0-1:0", Wednesday ...
OH	44060	41.70852	-81.359556	4	16	1	list(RestaurantsPriceRange2 = "2", BusinessAcceptsCreditCar...	Italian, Restaurants, Pizza, Chicken Wings	list(Monday = "10:0-0:0", Tuesday = "10:0-0:0", Wednesday ...
OH	44094	41.6398599	-81.4063963	3	7	1	list(RestaurantsTakeOut = "True", BusinessParking = "{'garag...	Bakeries, Food	list(Tuesday = "11:0-17:0", Wednesday = "11:0-17:0", Thursd...
ON	L4B 3G6	43.861502576	-79.3884991854	4	4	1	list(ByAppointmentOnly = "False", BikeParking = "True", Whe...	Fitness & Instruction, Active Life, Yoga	list(Monday = "16:0-23:0", Tuesday = "16:0-23:0", Wednesd...
AZ	85254	33.600071	-111.977371	5	5	1	list(BusinessAcceptsCreditCards = "True", BusinessParking = "...	Hair Stylists, Beauty & Spas, Hair Salons, Men's Hair Salons	list(Monday = "0:0-0:0", Tuesday = "9:0-19:0", Wednesday =...
NV	89119	36.1000163	-115.1285285	4	40	0	list(OutdoorSeating = "False", BusinessAcceptsCreditCards = ...	Restaurants, Italian	NULL
NV	89121	36.1165487	-115.0881146	5	21	1	list(BusinessAcceptsCreditCards = "True")	Event Planning & Services, Photographers, Professional Serv...	list(Monday = "0:0-0:0", Tuesday = "0:0-0:0", Wednesday = "...



business_i	name	city	state	postal_cod	latitude	longitude	stars	review_co	is_open	categories	restaurant	coffee,tea	fast,food	buffet	cafes	sandwiche	bars	bakeries	pizza	vegan,vege	american	italian	asian	sushi	latin,amer	middle,eas	french	indian
--1UhMGO	The Spicy /	Calgary	AB	T2P 0K5	51.04967	-114.08	4	24	1	Restaurant	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
--6MefnU	John's Chir	Richmond	ON	L4B 3P7	43.84169	-79.3998	3	44	1	Chinese, R	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
--9e1ONY	Delmonico	Las Vegas	NV	89109	36.12318	-115.169	4	1613	1	Cajun/Cre	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Pio Pio	Charlotte	NC	28203	35.19985	-80.8448	4	346	1	Restaurant	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
#NAME?	Sunnyside	Toronto	ON	M6E	43.67781	-79.4447	3.5	49	1	Restaurant	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	The Bar At	Henderson	NV	89052	35.97868	-115.155	4	135	1	Nightlife, B	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
#NAME?	Kabab Hou	Phoenix	AZ	85032	33.6413	-112.006	4.5	23	0	Lebanese,	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	
#NAME?	Sushiya	Sainte-Julie	QC	J3E 2T6	45.57536	-73.3266	3.5	3	1	Buffets, Re	1	0	0	1	0	0	1	0	0	0	0	0	1	1	0	0	
#NAME?	Mm Mm P	Canonsbur	PA	15317	40.2525	-80.1839	3.5	10	1	Restaurant	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
#NAME?	Happy Mo	Streetsbor	OH	44241	41.2429	-81.3527	3.5	96	1	Nightlife, S	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
#NAME?	Hungry Ho	Charlotte	NC	28269	35.33348	-80.7962	3	13	1	Restaurant	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	
#NAME?	C Market	East York	ON	M4K 3V9	43.68863	-79.3487	4.5	3	1	Delis, Rest	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
#NAME?	KFC	Brunswick	OH	44212	41.2419	-81.8411	2	6	1	Chicken W	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Double Pla	Las Vegas	NV	89123	36.01669	-115.173	4	7	0	Pizza, Bars	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	
#NAME?	Circle K	Phoenix	AZ	85085	33.71342	-112.099	3.5	9	1	Convenien	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Denny's	Highland H	OH	44143	41.53942	-81.4552	2	42	1	Breakfast &	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
#NAME?	Keung Kee	Toronto	ON	M1V 0C7	43.80675	-79.2889	3.5	43	0	Chinese, R	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
#NAME?	Pronto Pizz	Indian Trai	NC	28079	35.10696	-80.6326	3	4	0	Pizza, Rest	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
#NAME?	Primanti Bi	Pittsburgh	PA	15236	40.32393	-79.9438	3.5	74	1	Restaurant	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	
#NAME?	Panda Exp	Glendale	AZ	85301	33.5227	-112.152	2.5	21	1	Fast Food,	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	
#NAME?	Three Crov	Calgary	AB	T2J 7C3	50.93127	-114.064	4	4	1	Nightlife, R	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
#NAME?	Ric's Grill	Calgary	AB	T3J 5M9	51.11232	-113.982	1.5	3	1	Steakhous	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
#NAME?	Cargo Coff	Madison	WI	53715	43.0523	-89.3946	4.5	39	1	Restaurant	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Steak 'n Sh	Henderson	NV	89074	36.01791	-115.101	2.5	94	1	Restaurant	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
#NAME?	Saigon Res	Cleveland	OH	44115	41.49881	-81.6899	4	194	1	Thai, Vietn	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
#NAME?	Wendy's	Whitby	ON	L1N 9M1	43.87234	-78.9057	2.5	3	1	Fast Food,	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Souper Sal	Phoenix	AZ	85051	33.57642	-112.118	3.5	84	1	Soup, Buff	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	Al Muneer	Calgary	AB	T2E 6S8	51.08594	-114.012	1.5	7	1	Halal, Rest	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
#NAME?	The Old Ni	Toronto	ON	M4K 1N2	43.67623	-79.3578	3	36	1	Restaurant	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	
#NAME?	Taco Bell	Mesa	AZ	85204	33.39397	-111.805	1.5	4	1	Fast Food,	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	

Remaining datasets were also reduced in the number of instances by matching *business_id* and *user_id*, for instance:

```
yelp_review_small <- review[review$business_id %in% yelp_business_small$business_id,]
yelp_user_small <- user[user$business_id %in% yelp_review_small$user_id,]
```

Checkin dataset was broken down into separate binary columns for each day of the week and every hours of the day using *aggregate* and *spread* (*tidyr*) commands. *Checkin* was joined with *business* dataset on business_id as a separate file to reduce the number of files fed into SAS Miner.

```
b1=aggregate(checkins~business_id, data=check, drop=FALSE, simplify=FALSE, FUN=sum)
b2=aggregate(checkins~business_id+weekday, data=check, drop=FALSE, simplify=FALSE,
FUN=sum)
b3=aggregate(checkins~business_id+hour, data=check, drop=FALSE, simplify=FALSE,
FUN=sum)
sp1=spread(b2, weekday, checkins)
sp2=spread(b3, hour, checkins)
checkin = merge(b1,sp1,by="business_id", all.x=TRUE)
checkin = merge(checkin,sp2,by="business_id", all.x=TRUE)
check.final =
checkin[c("business_id","checkins","Mon","Tue","Wed","Thu","Fri","Sat","Sun","0:00","1
:00","2:00","3:00","4:00","5:00","6:00","7:00","8:00","9:00","10:00","11:00","12:00","
13:00","14:00","15:00","16:00","17:00","18:00","19:00","20:00","21:00","22:00","23:00"
)]
```

Two types of *review* datasets were generated: with ratings and full text reviews, and only with star ratings. The first one was to be used for text mining, and the second one for machine learning.

The number of columns in the *user* dataset was reduced by grouping fairly similar user attributes into *compliment* and *user.useful* columns:

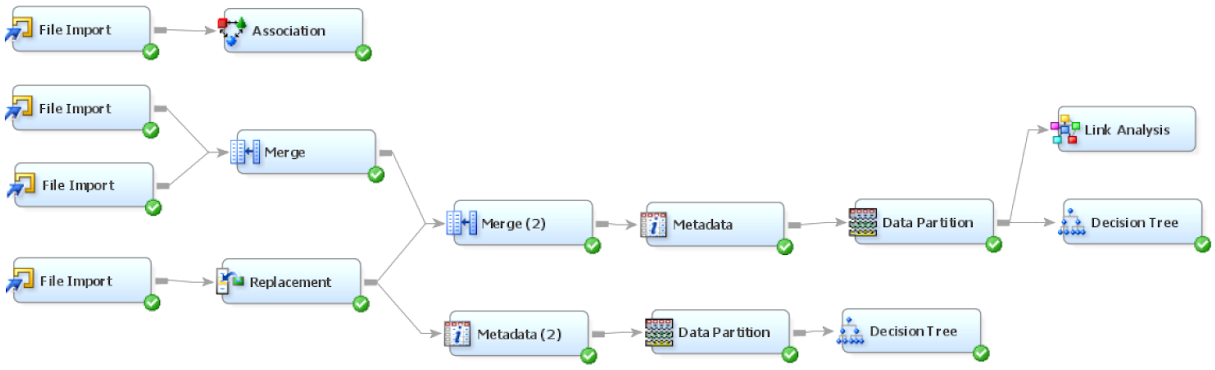
```
user$compliment = rowSums(user[,9:19],na.rm=TRUE)
user$user.helpful = rowSums(user[,4:7],na.rm=TRUE)
```

Additional cleaning involved replacing different types of missing values with consistent NAs, removing special characters, and converting variables.

Predicting Restaurant Ratings

One of the goals of this analysis was finding attributes that most strongly affect restaurant ratings. Knowing these attributes and being able to make predictions are crucial for any restaurant business.

Decision tree algorithms in SAS Enterprise Miner were used for filtering out these attributes. Both of the trees have the same parameters (depth 6, categories 5, leaf size 5, number of rules 5), but the first one used combined data with both business and user attributes, and the second one only used business attributes. Merged data was joined on respective *business_ids* and *user_ids*.



Data was partitioned into training (70%), validation (20%), and test (10%). Target variable was set as *stars*, all other variables were changed to binary, nominal, interval accordingly. Both decision trees demonstrated similarly good results on fit statistics and can be used for predictions:

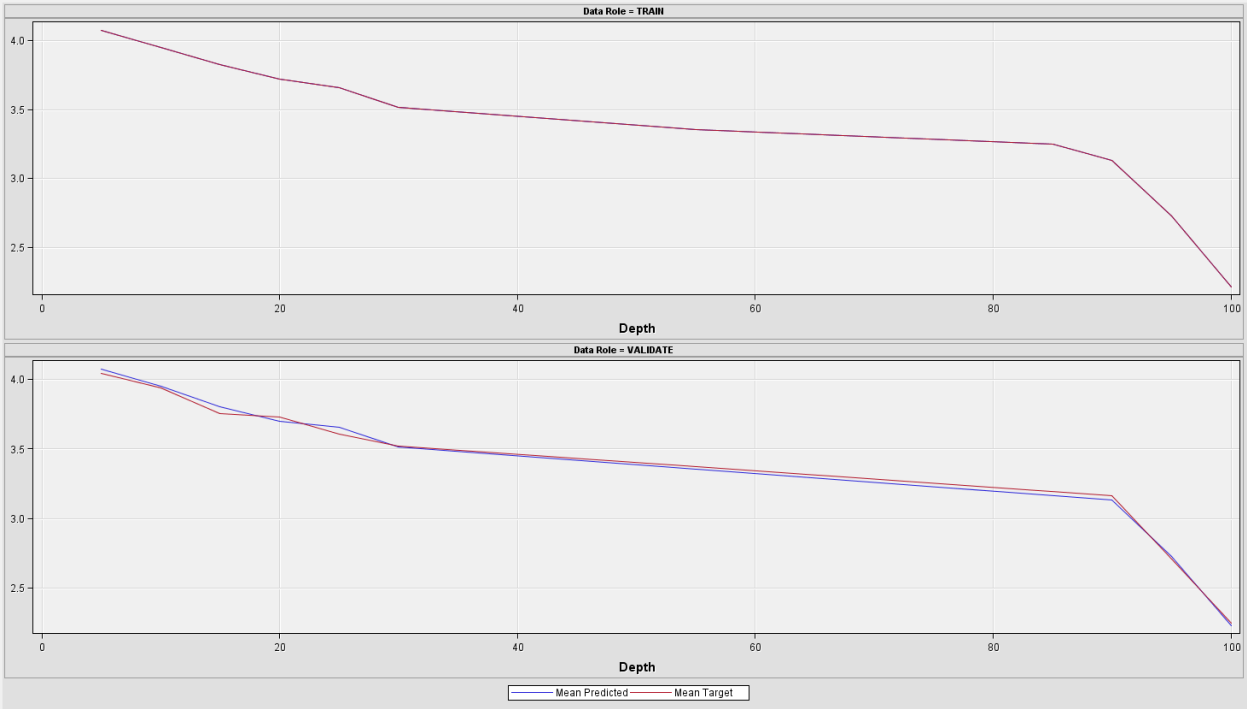
All attributes

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	43801.00	12735.00	6513.00
MAX	Maximum Absolute Error	3.05	3.06	2.58
SSE	Sum of Squared Errors	22742.65	6972.99	3576.10
ASE	Average Squared Error	0.52	0.55	0.55
RASE	Root Average Squared Error	0.72	0.74	0.74
DIV	Divisor for ASE	43801.00	12735.00	6513.00
DFT	Total Degrees of Freedom	43801.00	.	.

Business attributes

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	44134.00	12610.00	6305.00
MAX	Maximum Absolute Error	3.05	3.06	2.58
SSE	Sum of Squared Errors	23251.45	6984.19	3487.09
ASE	Average Squared Error	0.53	0.55	0.55
RASE	Root Average Squared Error	0.73	0.74	0.74
DIV	Divisor for ASE	44134.00	12610.00	6305.00

```
_DFT_       Total Degrees of Freedom      44134.00          .            .
```



Most importantly, decision trees identified attributes that affect restaurant rating the most. Summarizing both results, top attributes, ordered by descending importance, are:

1. Is not fast food
2. Is open on Sundays
3. Has many reviews
4. Is open on Mondays
5. Serves coffee/tea
6. Has low noise level
7. Has casual ambience
8. Has parking street
9. Does not have drive thru
10. Is good for kids
11. Is not in states: AB, ON, SC
12. Is in Eastern part of USA and Canada, according to longitude
13. Does not serve alcohol
14. Has pizza
15. Is not buffet
16. Has casual attire

These results provide interesting insights. While each attribute carries a very important implication for businesses, in sum, family restaurants which are not fast food or buffet, tend to have better ratings. It is important for restaurants to have parking and be open on Sundays and Mondays. One difference between the decision trees is that the one with all attributes identified

differences in ratings of customers that checked in at 6 pm. There is something important about dinner time or getting off work that food places should consider.

Exploring Associations

SAS Miner was also used for making associations between customer visits to restaurants. For this model, a new dataset was created in R that merged *user_ids* and *names* of the businesses they visited as was set as transaction data in SAS. Minimum confidence was set to 5% and minimum support was 3.5%.

Earl of Sandwich ==> Wicked Spoon
Wicked Spoon ==> Earl of Sandwich
Bacchanal Buffet ==> Wicked Spoon
Wicked Spoon ==> Bacchanal Buffet
Starbucks ==> Panera Bread
Panera Bread ==> Starbucks
Starbucks ==> Dunkin' Donuts
Dunkin' Donuts ==> Starbucks
Subway ==> Starbucks
Starbucks ==> Subway
Chipotle Mexican Grill ==> Starbucks
Starbucks ==> Chipotle Mexican Grill
McDonald's ==> Starbucks
Starbucks ==> McDonald's
Starbucks ==> In-N-Out Burger
In-N-Out Burger ==> Starbucks
Starbucks ==> Buffalo Wild Wings
Buffalo Wild Wings ==> Starbucks

These findings might be very useful for the companies, because resulting connections between them were not that obvious: for instance, people visiting coffee shops tend to visit fast food chains.

Predicting Restaurant Closure

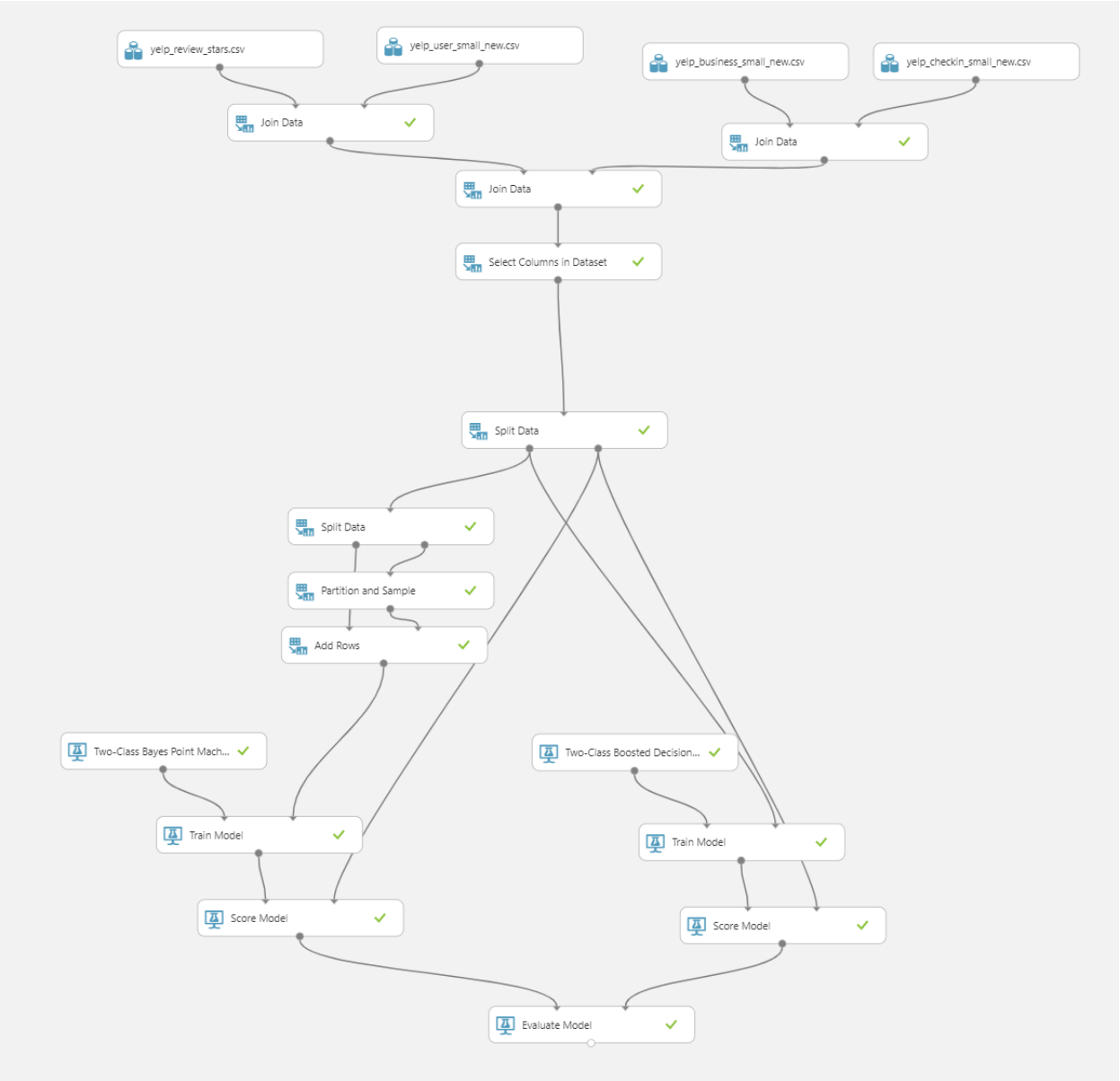
One of the variables in *business* dataset was whether the restaurant was currently open. Interestingly, about 30% of records were businesses that were no longer functioning. The question is: can we predict, based on our variables, closure of a restaurant business?

The model was created in Azure ML and, like in previous case, datasets were merged on respective keys. Target variable is binary: *is_open.* Data was partitioned into 80% train and 20% test.

The two models compared were Two-Class Bayes Point Machine and Two class Boosted Decision Tree. The results of running these algorithms demonstrated slightly better results for Bayes Point Machine. Furthermore, because the data was skewed (70% open businesses vs 30% closed), a bias was introduced to the model. Split data, partition and sample, and add rows nodes were used to increase the number of low instances and decrease the number of high instances. This change decreased the number of false negative and false positive predictions. Overall precision of the model is 99%.

On successful completion of the model, it was deployed as a web service so that it could be used to test real-time parameters.

The diagram and results of the model are shown below.

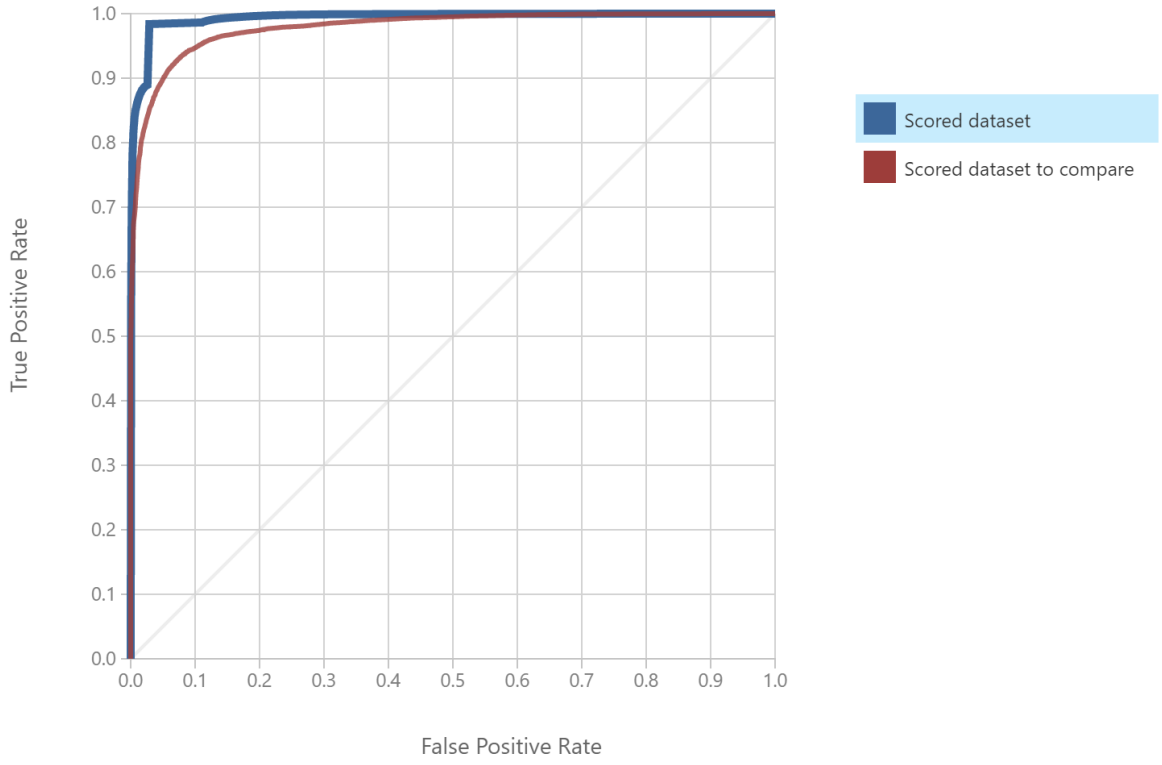


True Positive	False Negative	Accuracy	Precision	Threshold	AUC
230314	3800	0.981	0.993	0.5	0.993

False Positive	True Negative	Recall	F1 Score
1515	50784	0.984	0.989

Positive Label	Negative Label
1	0

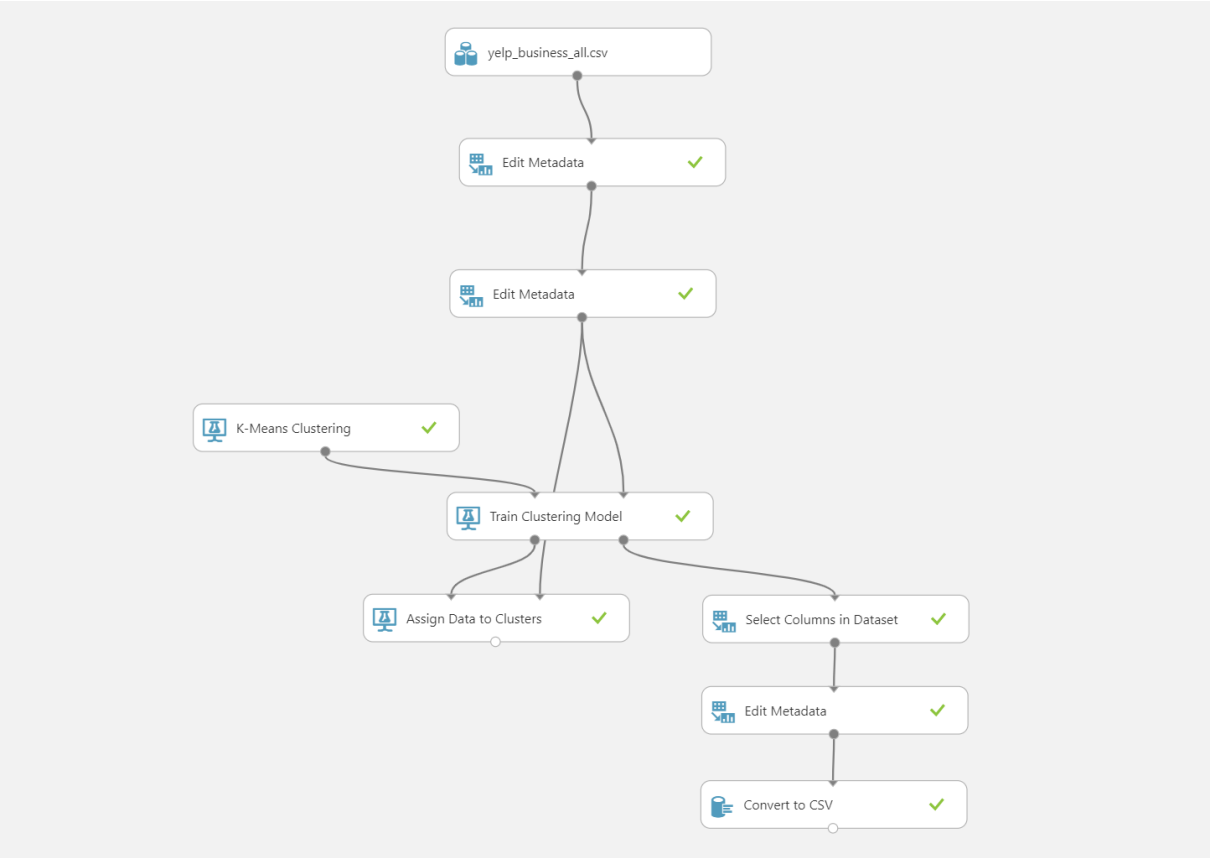
ROC PRECISION/RECALL LIFT

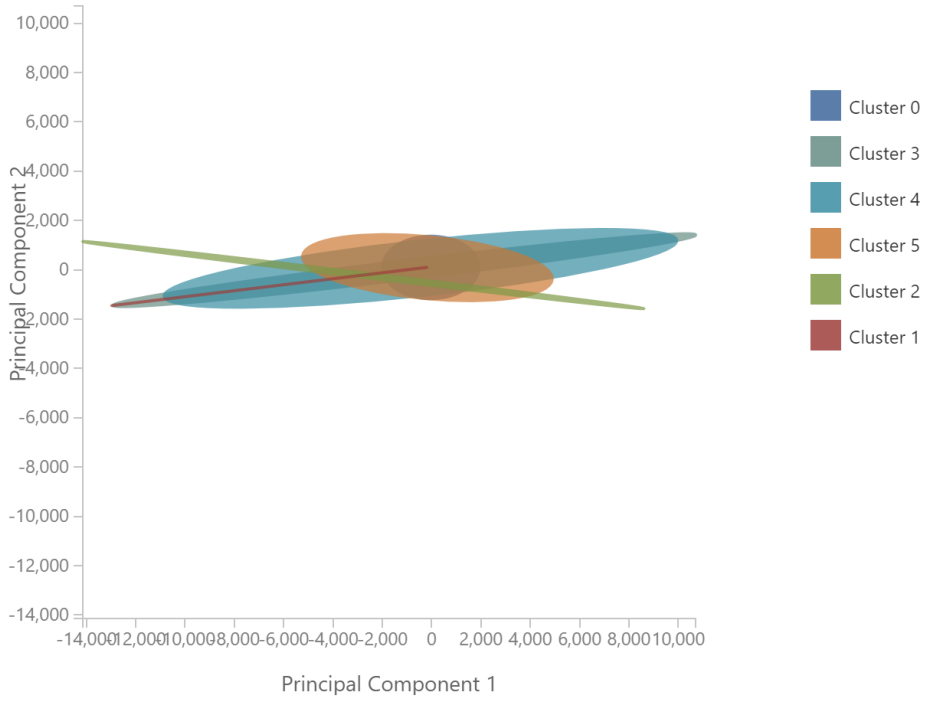


Cluster Analysis

Another objective with this dataset was performing exploratory analysis using clustering. Because we are not given enough data on customers, businesses themselves were grouped into clusters. The purpose of this analysis was to cluster businesses based on different attributes except for star ratings, and then compare the result to those clusters' average ratings to find out whether certain types of businesses that have something in common also tend to have similar ratings.

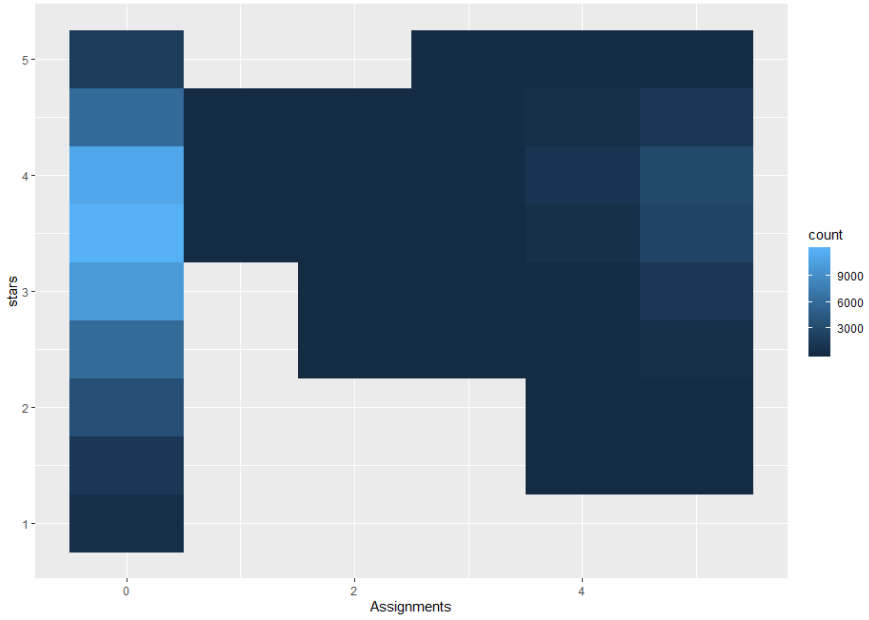
K-means clustering in Azure ML was used for this purpose. After several tries is was determined to use 6 k's, as they provided better and more distinct results. The metadata of trained model was then collected and analyzed using R. Resulting diagrams are presented below.





The algorithm was run several times, and each time the results were compared to the average star rating of the restaurants in that group. Below is the result of the final run, but even with other settings one cluster was constantly getting much lower rating.

Cluster	Stars
0	3.383514
1	4.000000
2	3.855072
3	3.984127
4	3.874765
5	3.706249



This analysis suggests also suggests that certain attributes that define a business also affect its rating. Clustering these businesses can help them find competitors with similar attributes and use this in their marketing strategy.

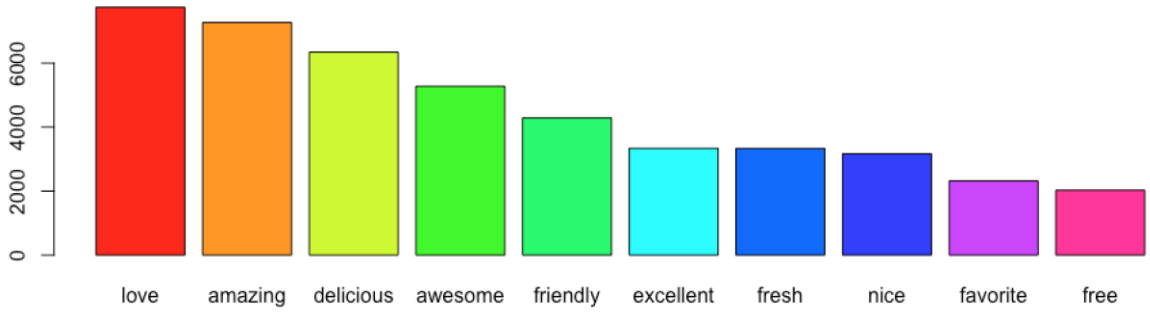
Sentiment Analysis (Text Mining):

The reviews of customers have been taken and text mining analysis is done to find the most frequent used words for positive and negative reviews. For this we used the libraries *dplyr* and *tidytext*. We created a data frame positive that had all the reviews of all restaurants that had a star rating greater than 4. We split the words in the reviews into individual tokens. The next thing that we did was removing the stop words. Then the inbuilt dictionary available in the *tidytext* library was used to compare and count the adjectives in the review. Similarly, we did this for negative reviews as well.

Results:

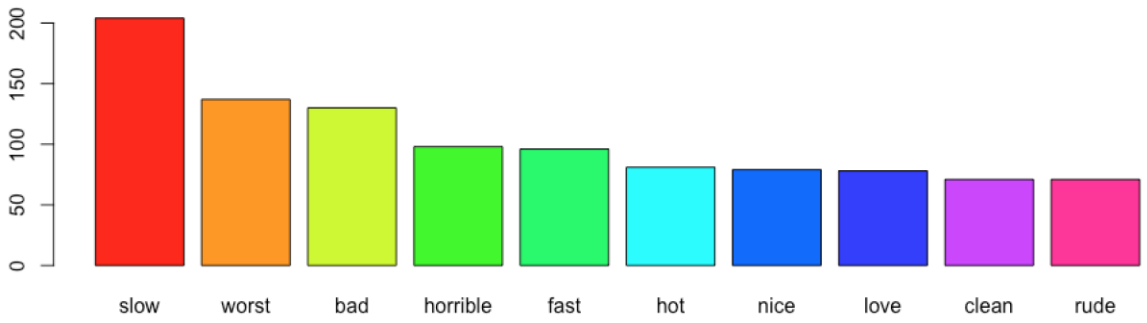
Positive

word	n
<chr>	<int>
1 love	7744
2 amazing	7272
3 delicious	6346
4 awesome	5277
5 friendly	4286
6 excellent	3337
7 fresh	3334
8 nice	3166
9 favorite	2318
10 free	2027



Negative

```
# A tibble: 668 x 2
  word        n
  <chr>    <int>
1 slow       204
2 worst      137
3 bad        130
4 horrible    98
5 fast        96
6 hot         81
7 nice        79
8 love        78
9 clean       71
10 rude       71
```



From this mapping of these frequencies we could easily know that over all a business has many positive reviews or negative reviews.

Conclusion

Yelp dataset is a valuable resource for analysts, as after proper cleaning it opens windows to new discoveries such as sentiment analysis, exploring and predicting business performance not in terms of financial statements, but in terms of customer feedback. Analyses performed in this paper provide new insights on what affects restaurant rating and closure, as well as help identify similar businesses. These results might be helpful for businesses to identify new customers, how to approach them, and discover their competitive strengths and weaknesses.

Although the models were performed on restaurant data, with different data preparation, similar techniques can be used for predicting performance and customer satisfaction for non-restaurant businesses or businesses in specific regions.