Identifying Trends and Market Opportunities in the Service Industry

Vikaasa Ramdas Thandu Venkat Kumar

***Abstract*—Online user reviews have evolved into a key aspect of the service industry today. Crowd-sourced review websites have become a major source of attracting customers to businesses, and the role of ratings and customer reviews can be vital in predicting the success and sustainability of businesses. A recent survey revealed that over 90% of the respondents said that online reviews influenced their buying decisions. In this project, the author proposes an approach to identify trends and market opportunities in the service industry from customer reviews over a period of time from the Yelp dataset.**

The proposed approach involves generating sets of features through simple manipulation of the given data, including weighting reviews in terms of recentness and user popularity. The businesses are also clustered geographically which will help us to identify popular business categories by location. The future business attention is then predicted using a regression model, and rules are mined to identify connections between various business categories. Finally, potentially lucrative business locations are identified by finding clusters with a large number of complementary businesses and low number of competing businesses.

***Index Terms*— machine learning, unsupervised learning, hierarchical clustering, random forests regression, association rule mining**

I. Introduction

YELP is a popular crowd-sourced reviews platform for businesses primarily in the service industry. Users can submit a review for each individual business on their products or services, using a one to five star rating system. After ten years of operation, Yelp has accumulated an enormous amount of data on information, reviews, and ratings of businesses. Sifting through this information can reveal many interesting trends such as cultural, seasonal, and regional trends, as well as trends that show the popularity of a business within a specific demographic. It would also be possible to recommend potential lucrative locations to establish new businesses, by identifying and building an index of complementary businesses and comparing regional trends, for example.

The underlying assumption we make to determine the popularity of a business is the number of 'positive' reviews it gets in a particular time period. However, while the rating and number of reviews for a business gives us a fair idea of its popularity, it ignores a lot of other information, such as the various attributes of the business, the reliability of individual users' ratings, the geographical location of the business and other businesses in its vicinity, and the recentness of the reviews themselves. The Yelp dataset provides a rich repository of information about its users as well, which allows weighting of individual reviews based on user information, such that it is possible to distinguish between experienced, highly up-voted users with a large number of fans, and new users, for example. Similarly, reviews are weighted in terms of recentness using a weighting function. This is done because recent reviews are much more relevant than reviews from several years before.

After the feature extraction and selection process, temporal prediction of the popularity of businesses can be performed by entering the final feature-set to the learning algorithm. In order to identify potential lucrative locations for establishing a new business, clustering techniques can be used to group the businesses in clusters of complementors and competitors. The preferred location for the business would be in a clusters that contain a high density of complementary businesses and low density of competitors.

II. Related Work

Michael Susplugas et al. proposed a novel method of identifying business opportunities that was based on identifying complementing businesses and complimentary businesses using the DBSCAN clustering technique. [1] Eric Wang et al. explored this topic as well, and proposed a method to exploit geographical data and apply self-organizing maps to create business neighborhoods, which are clusters of businesses that are in close proximity to each other. [2] In this paper, the authors generated a heat map to analyze emergent patterns and arranged the clusters by generating a dendrogram (i.e., hierarchical clustering) of the clusters. They perform k-medoids clustering to cluster business neighborhoods, and examine the usability of the identified business clusters by using support vector regression to predict the average star ratings of the businesses in these clusters. Bryan Hood et al. attempt to solve the problem of inferring future business attention by building a feature set of time dependent features and review text features, user clustering, followed by Principal Component Analysis for dimensionality reduction. [3] The authors then used Support Vector Regression (SVR) as the regression model. This project aims to identify trends and market opportunities in the service industry from the Yelp dataset, and the papers discussed above

Author is with the Computer and Information Science and Engineering Department, University of Florida, USA (e-mail: vikaasa@ufl.edu).

form an excellent basis for formulating an approach for this problem, since there are a lot of parallels that can be drawn from these related works.

III. Description

A. *Data Extraction*

The dataset will be procured from the Yelp Dataset Challenge [4], which is an open source database. The dataset contains data for more than 77,000 businesses, with more than 2.2M reviews. The data is in JSON format, with separate files for different object types. There is 1 JSON object per line in each file.

The different object types are: business, check-in, review, tip, user, and photos. The JSON data can be parsed into Python using the JSON library.

Here, a subset of the Yelp dataset has been chosen for the purpose of this project. This subset chosen is limited to the state of Pennsylvania. However, the methodology and the code written in this project can be extended to other states and cities from the Yelp dataset.

After extracting the businesses from the state 'PA', and the corresponding reviews and users' information, the data is split into two such that all reviews before the date June 1, 2015 are placed in the training dataset. Thus, metadata about the businesses in the state of 'PA' is collected by aggregating its review information through the years until June 1, 2015, and is stored as the training dataset.

B. *Feature Extraction*

The process begins by extracting features with simple manipulation of the data to create time-dependent features, and falls under two categories. The first category contains features which describe a subset of reviews. These features include: number of reviews in the set, average number of stars across reviews in the set, recentness of the review, number of the fans of the user who posted the review, number of votes received, minimum number of stars, number of unique users logging these reviews every six months, and so on.

The second category contains features which describe metadata about the business, such as location, number of businesses within 1km and number of businesses sharing a category. These features are obtained by performing clustering based on geographical location (latitude and longitude). The clustering technique used is discussed in detail in the next section.

A list of weighted time-dependent features are created, that are defined below:

a) Weighted recency score

The reviews in the dataset range from the year 2005 to 2015. The reviews are thus weighted by year to obtain a recency score, such that a review posted in 2005 has a score of 0.5, and a review posted in 2015 has a score of 1. The weighting function is given below:

$$recency\ score = 0.5 + \left(\frac{\left(10 - (2015 - YEAR)\right)^2}{10 * 2}\right)$$

b) Weighted user review count

The review count of each user who makes a review is weighted as follows:

$$weighted\ review\ count\ (wrc) = FLOOR(1 + 100 * \frac{(user\ review\ count)}{maximum\ count}$$

c) Weighted fan count

The fan count of each user who makes a review is weighted as follows:

$$weighted\ fan\ count\ (wfc) = FLOOR(1 + 100 * \frac{(fan\ count)}{maximum\ count}$$

d) Weighted vote count

The number of votes received by the user for a review is weighted as follows:

$$weighted\ vote\ count\ (wvc) = FLOOR(1 + 100 * \frac{(number\ of\ votes)}{maximum\ votes}$$

e) Final weighted score

The final weighted score for every review is calculated as:

$$final\ weighted\ score = recency\ score * (\frac{wrc + wfc + wvc}{3})$$

f) Weighted user sum

The weighted user sum is defined as the sum of all the final weighted scores for all the reviews for a particular business.

g) Weighted ratings sum

The weighted ratings sum is defined as the sum of all the weighted ratings for all the reviews for a particular business.

h) Weighted average rating

The weighted average rating is calculated as follows:

$$weighted\ average\ rating = \left(\frac{weighted\ ratings\ sum}{weighted\ user\ sum}\right)$$

i) Regression value

The regression value is the value that our regression model needs to predict. This value is defined as the product of the review count of a business and its average rating.

C. *Feature Selection*

There might often be redundant or irrelevant features that can hinder the performance of a learning algorithm, and the best features are identified by those that make data easily separable or predictable. A naive approach for feature selection uses an exhaustive search over all possible subsets of features to pick the subset with the smallest test error, which is not feasible for large datasets with many features. Other approaches include using dimensionality reduction techniques such as Principal Component Analysis (PCA), or feature selection techniques such as Greedy Feature Removal, Scaling, or Univariate Feature Analysis. The feature selection method used in this project uses Random Forests. This method is selected primarily because of compatibility, as the Random Forests regression model is used to solve the regression problem as well. As a precursor to that step, Random Forests offers two straightforward methods for feature selection: mean decrease impurity and mean decrease accuracy, in order to find important variables for interpretation and to try and design a good prediction model. Robin Genuer et al. examines these two methods in detail. [5]

In general, for feature selection, a scoring function and a search method to optimize the scoring function is needed. Random Forests is used to rank feature based on an importance score. Random Forests will select features randomly with replacement and group every subset in a separate subspace (called random subspace). One scoring function of importance could be based on assigning the accuracy of every tree for every feature in that random subspace, which is then done subsequently for every separate tree. A threshold for computing the importance score may be used. In random forests, the most extensively used score of importance of a given variable is the increase in the mean of the error of a tree (which is MSE for regression and misclassification rate for classification) in the forest, when the observed values of this variable are randomly permuted among the OOB samples.

Dıaz-Uriarte, Alvarez de Andres proposed a strategy based on recursive elimination of variables, where they first compute RF variable importance. [6] Subsequently, at every step, 20% of the variables having the smallest importance are eliminated and a new forest is built with the remaining variables. Finally, the set of variables that leads to the smallest OOB error rate is selected. This is the strategy followed in this project as well, where the size of the initial feature list with 21 features was reduced to 8 features.

D. *Geographical Clustering of Businesses*

The objective of this phase of the project is to group businesses in a 1-mile radius into clusters. This data is later used as a feature for the regression model used in the subsequent phase of this project, which is later used to determine the most popular and emerging categories of businesses in a particular cluster.

The hierarchical clustering approach was chosen in this project, because of the following reasons. [7] Since the primary motivation here is to find clusters of businesses in a 1-mile radius, algorithms such as k-means, which requires a value of ‘k’ clusters to be specified before-hand, are not ideal. DBSCAN, which forms clusters on the basis of density, is also not applicable in this scenario as the objective is to strictly find clusters of businesses in a 1-mile radius.

Hierarchical clustering is an unsupervised classification method that is used in exploratory data analysis mainly to identify subgroups within a dataset. It is an alternative approach to other clustering algorithms as it builds a hierarchy from the bottom-up, and does not require the number of clusters to be specified beforehand. Instead, hierarchical clustering, when combined with the Haversine distance function, can determine clusters that are within a specified geographical distance. In this case, we can specify the geographical distance as 1 mile (1.60934 km).

The hierarchical clustering algorithm works as follows:

i. First, every data point is put in its own cluster.

ii. Then, the closest two clusters are identified and combined into one new cluster.

iii. The above step is repeated till all the data points are combined into a single cluster.

Hierarchical clustering uses some form of dissimilarity (such as Euclidean distance) to determine differences between observations. The method produces a dendrogram which provides an excellent visual method for determining the ideal number of clusters or categories within the data.

In hierarchical clustering, there are a few ways to determine how close two clusters are. The approach followed here involves complete linkage clustering, where the distance between two clusters is defined as the longest distance between two points in each cluster. Complete linkage clustering avoids the ‘chaining’ phenomenon, which is a significant drawback of the alternative single linkage method, where clusters formed via single linkage clustering may be forced together due to individual elements being located close to each other, even though several of the elements in each cluster may actually be very distant from each other. Complete linkage tends to find compact clusters of approximately equal diameters.

Though complete linkage leads to outliers being assigned a cluster of their own, this property actually meets the requirements of this project, where it is required to assign every business to a cluster with radius of one mile.

Since the distance between points on the map are distances on a sphere, the shortest distance between two points on the map (i.e., the ’great-circle-distance’) is calculated according to the Haversine method.

For any two points on a sphere, the haversine of the central angle between them is given by

$$\operatorname{hav}\left(\frac{d}{r}\right) = \operatorname{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1)\cos(\varphi_2)\operatorname{hav}(\lambda_2 - \lambda_1)$$

where,

hav is the haversine function:

$$\operatorname{hav}(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1-\cos(\theta)}{2}$$

d is the distance between the two points (along a great circle of the sphere; see spherical distance),

r is the radius of the sphere,

φ_1, φ_2: latitude of point 1 and latitude of point 2, in radians

λ_1, λ_2: longitude of point 1 and longitude of point 2, in radians

E. *Random Forests Regression Model*

The second phase of this project implements Breiman’s random forest algorithm for regression, to predict the product of review count and average rating for businesses. [8]

This phase of the project aims to predict the growth of the businesses in the dataset. The product of the review count and average rating for businesses is chosen as the dependent variable which is to be predicted by the regression model, as it is able to better depict the growth of a business than using either the star ratings or review count independently.

Here, some of the reasons why this approach was chosen to solve the regression problem over other methods are listed below:

i. Random Forests runs efficiently on big data bases.
ii. It consists of an ensemble of different regression trees that are aggregated over to determine a better prediction.
iii. It can handle a large number of input variables without variable deletion, and does not suffer from overfitting problems.
iv. It gives good estimates of what variables are important to the regression model.
v. It generates an internal unbiased estimate of the generalization error as the forest creation progresses.
vi. It has an efficient process for estimating missing data and preserves accuracy even when a large proportion of the data are missing.
vii. Another advantage is that generated forests may be stored for future use on other data.

In a regression tree, since the target variable is a real valued number, the regression model is fitted to the target variable using each of the independent variables. Then, for each independent variable, the data is split at several split points. The sum of squared error (SSE) is calculated at each split point between the predicted value and the actual values. Then, the variable resulting in minimum SSE is selected for the node. This process is recursively continued till the entire data is covered. [10] The “mean of squared residuals” is computed as

$$\mathrm{MSE}_{\mathrm{OOB}} = n^{-1}\sum_{1}^{n}\{y_i - \hat{y}_i^{\mathrm{OOB}}\}^2,$$

where $\hat{y}_i^{\mathrm{OOB}}$ is the average of the OOB predictions for the ith observation. The “percent variance explained” is computed as

$$1 - \frac{\mathrm{MSE}_{\mathrm{OOB}}}{\hat{\sigma}_y^2},$$

where σ^2_y is computed with n as divisor (rather than n − 1). We can compare the result with the actual data, as well as fitted values from a linear model.

Random forests can be used for regression analysis, and are otherwise known as Regression Forests. They are considered an ensemble of multiple regression trees, and are used for nonlinear multiple regression. [9] Ensemble methods use multiple variant models and aggregate over them to determine a better prediction. Random forests use the concept of bagging and produces k trees based on a random sampling of features in the data set. Each tree is based on a few features (instead of all features) of the given data, and all the trees together create a forest. Random Forests try to de-correlate trees to reduce autocorrelation. The fundamental objective of Random Forests is to make each tree as independent as possible.

Breiman proposed that random forests that can be used for regression can be constructed by building trees that depend on some random vector Θ in such a way that the tree predictor $h(x,\Theta)$ takes on numerical values instead of class labels. [8] When the output values are numerical, it can be said that the training set is drawn independently from the distribution of the random vector Y, X. The mean-squared generalization error for any numerical predictor h(x) is

$$E_{X,Y}\,(Y-h(X))^2$$

The random forest predictor is formed by taking the average over k of the trees { h(x,Θk)}. Similar to the classification case, the following holds:

As the number of trees in the forest goes to infinity, almost certainly, we can say that

$$E_{X,Y}\,(Y-av_k h(X,\Theta k\,))^2 \rightarrow E_{X,Y}\,(Y-E_\Theta h(X,\Theta))^2$$

Assume that for all Θ, $EY = EXh(X,\Theta)$. Then,

$$PE^*(\mathrm{forest}) \le \rho PE^*(\mathrm{tree})$$

where ρ is the weighted correlation between the residuals $Y-h(X,\Theta)$ and $Y-h(X,\Theta')$, and where Θ,Θ' are independent. This theorem states the constraints for precise regression forests, which are low correlation between residuals and low error trees. The random forest reduces the average error of the trees used by the factor ρ. The randomization that is used must aim at low correlation.

F. Association Rule Mining

The third and final phase of this project focuses on identifying potential business opportunities for establishing new businesses of different categories. This part of the project is built upon the results from the previous two phases of the project.

The results of the regression model are used to predict the most popular businesses categories in a cluster. From the predicted results of the regression model, we consider all businesses in a cluster with an average rating of greater than or equal to 3.5 as popular, and with an average rating of lesser than 3.5 as unpopular. The predicted variable for each of these businesses (which is the product of their star rating and review count) is assigned as +ve for popular businesses (where the average rating >= 3.5) and -ve for the other businesses (average rating < 3.5), and is then summed up, and finally sorted to get a list of the most popular categories in a cluster.

This list represents popular categories of businesses in a cluster that could be regarded as complementing each other. Association rule mining is performed after converting this list to a transactional data set in order to mine rules that identify such complementing business categories. These rules are generated with a constraint that they should have support greater than 0.015 and confidence greater than 0.75.

From the dataset, a set of most frequently occurring categories is identified for every cluster. Then, this is compared with the list of mined rules to see if there are any common elements. If one or more categories from the mined rules are missing in the set of frequently occurring categories for that cluster, then it can be said that there is a potential opportunity to establish a business of that category in that cluster's geographical locality. If there are no missing categories from the rule in that cluster, the algorithm skips the cluster and moves on iteratively to the next one. This method of identifying potential business opportunities also accounts for competing businesses, since the algorithm skips clusters with high concentration of competing businesses.

IV. Evaluation

A. *Geographical Clustering of Businesses*

In this section, the results for hierarchical clustering algorithm used to cluster businesses in a 1-mile radius are described. A subset of the Yelp dataset that consist of businesses only from the state 'PA' is first extracted for the purpose of this project. This had a total of 3624 businesses. This data was then geographically clustered using hierarchical clustering with complete linkage and using Haversine distance, and the distance constraint set to 1 mile (1.60934 km), which resulted in the 3624 businesses being clustered into 135 clusters. A map plotting the clusters against the map of Pittsburgh and its surrounding region is shown in Figure 1.

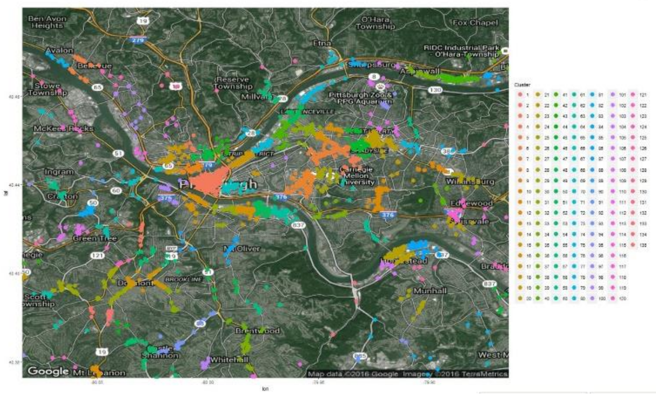


Fig. 1. Plot showing geographical clusters of businesses in 1-mile radius formed using Hierarchical clustering algorithm with complete linkage using Haversine distance.

The cluster information was added as a feature to the dataset, and was passed to the Random Forests regression model for the second phase of the project.

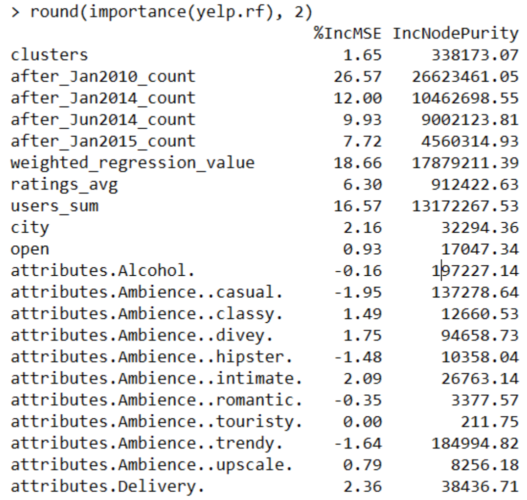
B. *Temporal Prediction using Random Forests Regressor*

In this section, the results of the second phase of the project using the Random Forests Regression model to predict the expected product of a business's average rating and review count, given the attention received so far, is described. Given metadata about a business and all reviews in the Yelp database logged before a target date, which is taken as June 1st 2015, the Random Forests Regression model predicts the product of the average rating and review count for a business 6 months after the target date.

First, a feature vector of 21 features highlighting metadata about each business was generated for every business in the dataset. The data set itself was split into an 80:20 split by randomly sampling with setting a seed, such that 80% of the data records was used to train the Random Forests Regression model and 20% of the data records was used for testing.

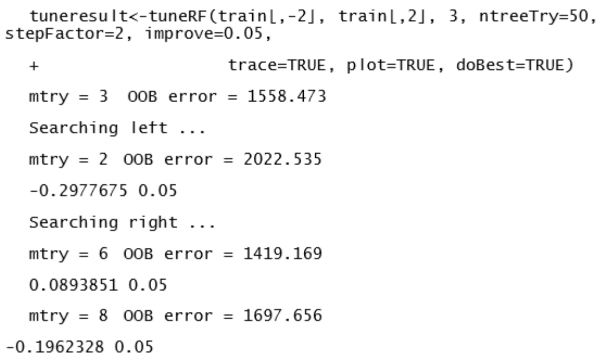
Then, feature selection was performed using Random Forests by computing the RF variable importance. Then, at each step, 20% of the variables that had the smallest importance were removed, and a new forest was built with the remaining variables. Finally, the set of variables that led to the smallest OOB error rate was selected, and the m-try value was found to be 6. The final feature set had 8 features.

The tuneRF function in the randomForest package was used to select the "optimal" mtry value, using the out-of-bag error estimate as the criterion. The feature set at the beginning of the feature selection process is shown in Fig. 2, the tuneRF function is shown in Fig. 3, and the feature set after the feature selection process was completed is shown in Fig. 4.



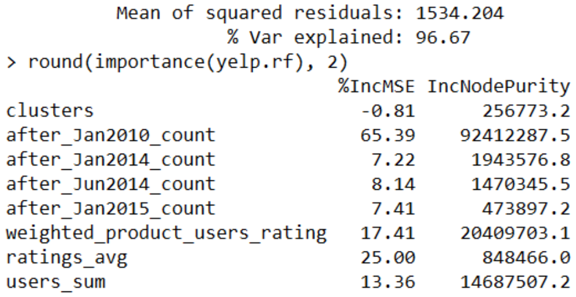
```
> round(importance(yelp.rf), 2)
                              %IncMSE IncNodePurity
clusters                         1.65     338173.07
after_Jan2010_count             26.57   26623461.05
after_Jan2014_count             12.00   10462698.55
after_Jun2014_count              9.93    9002123.81
after_Jan2015_count              7.72    4560314.93
weighted_regression_value       18.66   17879211.39
ratings_avg                      6.30     912422.63
users_sum                       16.57   13172267.53
city                             2.16      32294.36
open                             0.93      17047.34
attributes.Alcohol.             -0.16     197227.14
attributes.Ambience..casual.    -1.95     137278.64
attributes.Ambience..classy.     1.49      12660.53
attributes.Ambience..divey.      1.75      94658.73
attributes.Ambience..hipster.   -1.48      10358.04
attributes.Ambience..intimate.   2.09      26763.14
attributes.Ambience..romantic.  -0.35       3377.57
attributes.Ambience..touristy.   0.00        211.75
attributes.Ambience..trendy.    -1.64     184994.82
attributes.Ambience..upscale.    0.79       8256.18
attributes.Delivery.             2.36      38436.71
```

Fig. 2. Feature set at the beginning of the feature selection process with 21 features.



```
  tuneresult<-tuneRF(train[,-2], train[,2], 3, ntreeTry=50,
stepFactor=2, improve=0.05,

  +                   trace=TRUE, plot=TRUE, doBest=TRUE)

  mtry = 3  OOB error = 1558.473

  Searching left ...

  mtry = 2  OOB error = 2022.535

  -0.2977675 0.05

  Searching right ...

  mtry = 6  OOB error = 1419.169

  0.0893851 0.05

  mtry = 8  OOB error = 1697.656

-0.1962328 0.05
```

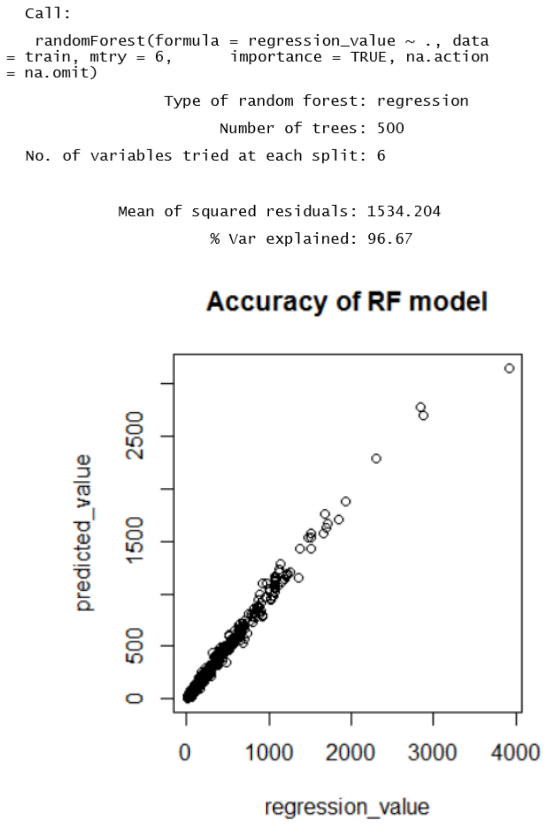
Fig. 3. Result of TuneRF function, that leads to finding the optimal m-try value to be 6.



```
          Mean of squared residuals: 1534.204
                    % Var explained: 96.67
> round(importance(yelp.rf), 2)
                              %IncMSE  IncNodePurity
clusters                        -0.81       256773.2
after_Jan2010_count             65.39     92412287.5
after_Jan2014_count              7.22      1943576.8
after_Jun2014_count              8.14      1470345.5
after_Jan2015_count              7.41       473897.2
weighted_product_users_rating   17.41     20409703.1
ratings_avg                     25.00       848466.0
users_sum                       13.36     14687507.2
```

Fig. 4. Final feature set after the completion of the feature selection process with 8 features.

After running the Random Forests Linear regression model with mtry=6, the mean of squared residuals was found to be 1534.204, and the %Var explained was found to be 96.67. A graph plotting the output of the RF model is shown in Fig. 5.



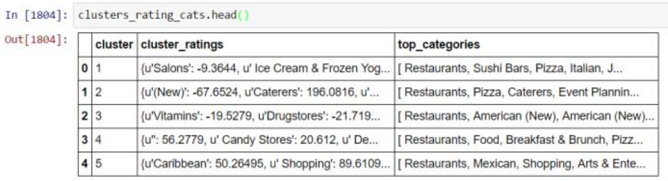
```
 Call:
  randomForest(formula = regression_value ~ ., data
= train, mtry = 6,      importance = TRUE, na.action
= na.omit)
                Type of random forest: regression
                      Number of trees: 500
 No. of variables tried at each split: 6

            Mean of squared residuals: 1534.204
                      % Var explained: 96.67
```

Fig. 5. Output graph plotting the predicted value of the RF model against the actual data (regression_value) from the dataset.

C. Identfying Potential Business Opportunities using Association Rule Mining

In the third and final phase of this project, the results of identifying potential business opportunities for establishing new businesses of different categories are described.

The predicted variable for each of these businesses (which is the product of their star rating and review count), was assigned as +ve for popular businesses (where the average rating >= 3.5) and -ve for the other businesses (average rating < 3.5), and was then summed up, and finally sorted to get a list of the most popular categories in a cluster, as shown in Table 1.



In [1804]: clusters_rating_cats.head()

Out[1804]:

	cluster	cluster_ratings	top_categories
0	1	{u'Salons': -9.3644, u' Ice Cream & Frozen Yog...	[Restaurants, Sushi Bars, Pizza, Italian, J...
1	2	{u'(New)': -67.6524, u'Caterers': 196.0816, u'...	[Restaurants, Pizza, Caterers, Event Plannin...
2	3	{u'Vitamins': -19.5279, u'Drugstores': -21.719...	[Restaurants, American (New), American (New)...
3	4	{u'': 56.2779, u' Candy Stores': 20.612, u' De...	[Restaurants, Food, Breakfast & Brunch, Pizz...
4	5	{u'Caribbean': 50.26495, u' Shopping': 89.6109...	[Restaurants, Mexican, Shopping, Arts & Ente...

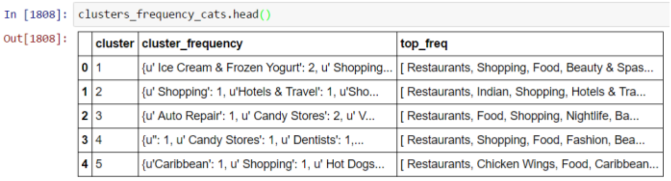
Table 1. Tabulation of top categories per cluster, calculated from the predicted value of the RF Regression model output.

This list represented complementary and popular categories of businesses in a cluster. After converting this list to a transactional data set, association rule mining was performed in order to mine rules that identify such complementing business categories. [12] These rules were generated with a constraint that they should have support greater than 0.015 and confidence greater than 0.75, and top few rules sorted by support are shown in Table 2. The first rule, “Bars” => “Nightlife” indicates that “Bars” and “Nightlife” are complementary business categories.

	rules	support	confidence	lift
86	{Bars} => {Nightlife}	0.19852941	0.8709677	2.575035
85	{Pizza} => {Restaurants}	0.19117647	0.8666667	1.403175
84	{American (New)} => {Restaurants}	0.17647059	0.8571429	1.387755
321	{Restaurants,Bars} => {Nightlife}	0.14705882	0.9090909	2.687747

Table 2. Tabulation of top 4 mined association rules sorted by support.

From the dataset, a set of most frequently occurring categories was identified for every cluster, as shown in Table 3.



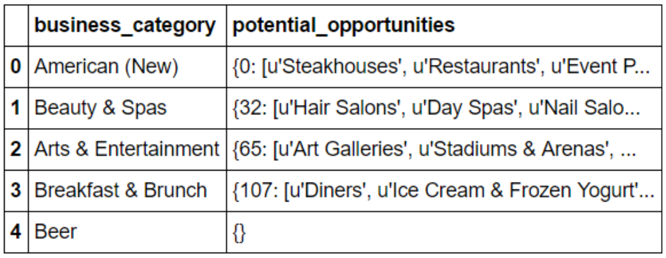
In [1808]: clusters_frequency_cats.head()

Out[1808]:

	cluster	cluster_frequency	top_freq
0	1	{u' Ice Cream & Frozen Yogurt': 2, u' Shopping...	[Restaurants, Shopping, Food, Beauty & Spas...
1	2	{u' Shopping': 1, u'Hotels & Travel': 1, u'Sho...	[Restaurants, Indian, Shopping, Hotels & Tra...
2	3	{u' Auto Repair': 1, u' Candy Stores': 2, u' V...	[Restaurants, Food, Shopping, Nightlife, Ba...
3	4	{u'': 1, u' Candy Stores': 1, u' Dentists': 1,...	[Restaurants, Shopping, Food, Fashion, Bea...
4	5	{u'Caribbean': 1, u' Shopping': 1, u' Hot Dogs...	[Restaurants, Chicken Wings, Food, Caribbean...

Table 3. Tabulation of mostly frequently occuring categories per cluster.

Then, this set of most frequently occurring categories wass compared with the list of mined rules to see if there are any common elements. If any one or more categories from the mined rules were missing in the set of frequently occurring categories for that cluster, these were labelled as potential opportunities to establish a new business of that category in that cluster’s geographical locality. This is depicted in Table 4. Here, the category “Beer” has no potential opportunities listed. This is because there are no missing categories from the rule in any of the clusters. On the other hand, for business category “Beauty and Spas”, for example, there are potential business opportunities in cluster 32 to open a hair salon, day spa, and nail salon.



	business_category	potential_opportunities
0	American (New)	{0: [u'Steakhouses', u'Restaurants', u'Event P...
1	Beauty & Spas	{32: [u'Hair Salons', u'Day Spas', u'Nail Salo...
2	Arts & Entertainment	{65: [u'Art Galleries', u'Stadiums & Arenas', ...
3	Breakfast & Brunch	{107: [u'Diners', u'Ice Cream & Frozen Yogurt'...
4	Beer	{}

Table 4. Table depicting potential business opportunities for different business categories by cluster.

V. Summary and Conclusion

This project has introduced a novel approach to identify business trends and market opportunities from the Yelp dataset. The project successfully incorporated geographic clustering of businesses using hierarchical clustering and predicting future business attention using the Random Forest Regression model. The Random Forest Regression model was tested and found to have a high accuracy with a mean of squared residuals error (MSE) of 1534.204, and with a high % of Var explained = 96.67.

The predicted data from the Random Forests Regression model was subsequently used to detect complementary businesses by applying Association rule mining, and finally identifying opportunities to open new businesses within specific geographic clusters. This data can be very useful for entrepreneurs who are looking for a good location to establish their business. The information gleaned from this project has wider implications, as it can also help current businesses stay relevant by proactively observing trends in their geographical vicinity, and tweaking their business model to stay competitive.

VI. Future Scope

Semantic analysis of text reviews is something that can be incorporated into this project in the future, which would add a new dimension into analyzing and rating the categories of businesses. [13][14][15] By parsing the review text and identifying keywords and associating sentiment ratings to these keywords, it would be possible to identify and explore more connections between different businesses and business categories. It would also be interesting to cluster the users based on the locations they frequent, to identify trends in their travel patterns to businesses of different categories, which may offer potential information on identifying further business opportunities.

References

[1] Susplugas, Michael, Radhika Garg, and Burkhard Stiller. "Recommending Ideal Location for New Business based on Yelp Dataset."

[2] Wang, Eric, and Charles Zhang. "Application of Unsupervised Learning Techniques to Business Meta-Data, using Yelp Data."

[3] Hood, Bryan, Victor Hwang, and Jennifer King. "Inferring future business attention." Yelp Challenge, Carnegie Mellon University (2013).

[4] Yelp: Yelp Dataset Challenge - Round five; http://www.yelp.com/datasetchallenge/

[5] Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests." Pattern Recognition Letters 31.14 (2010): 2225-2236.

[6] Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 1.

[7] Berkhin, Pavel. "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.

[8] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[9] https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[10] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[11] Scikit Learn. http://scikit-learn.org/stable/index.html.

[12] https://cran.r-project.org/web/packages/arules/arules.pdf

[13] Huang, James, Stephanie Rogers, and Eunkwang Joo. "Improving restaurants by extracting subtopics from yelp reviews." iConference 2014 (Social Media Expo) (2014).

[14] Wang, Hongning, Yue Lu, and Chengxiang Zhai. "Latent aspect rating analysis on review text data: a rating regression approach." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[15] Elkouri, Andrew. "Predicting the Sentiment Polarity and Rating of Yelp Reviews." arXiv preprint arXiv:1512.06303 (2015).