CASE STUDY REPORT

Group No.: Group 17

Student Name: Maitreya babar

I. Background and Introduction

Whether on a vacation with the family or on a business trip — or simply at home wanting to try something new, “Yelp” has been a great way to find good restaurants. One problem we encounter sometimes is that there can be a lot of restaurants of the same cuisine with similar ratings. With so many restaurants having similar ratings, it can be challenging to figure out which place to try. Is there a way to “cut through the noise” and extract more information so that a clearer choice emerges? So, whenever we are hungry not just for food but even a good place to eat out some authentic cuisine, we tend to go through reviews posted on yelp. But there are hundreds of restaurants offering various types of cuisines. Each of these restaurants has hundreds of reviews and ratings given on it. One usually tends to go for a place which tends to be of higher rating, but again the paradox is that, higher rating does not always make the food authentic. Therefore, our approach to this problem include analysis of yelp dataset using classification method. Food connoisseurs devour the best quality of food served by relying on “yelp” based reviews and ratings to select their choice of cuisines. However, sometimes it is unclear to the customers which businesses offer the quality authentic cuisines which simply doesn’t only depend on ratings. Therefore, through meticulous analysis of ‘yelp’ parameters , we can predict authenticity of cuisines by classification model.

Problem:

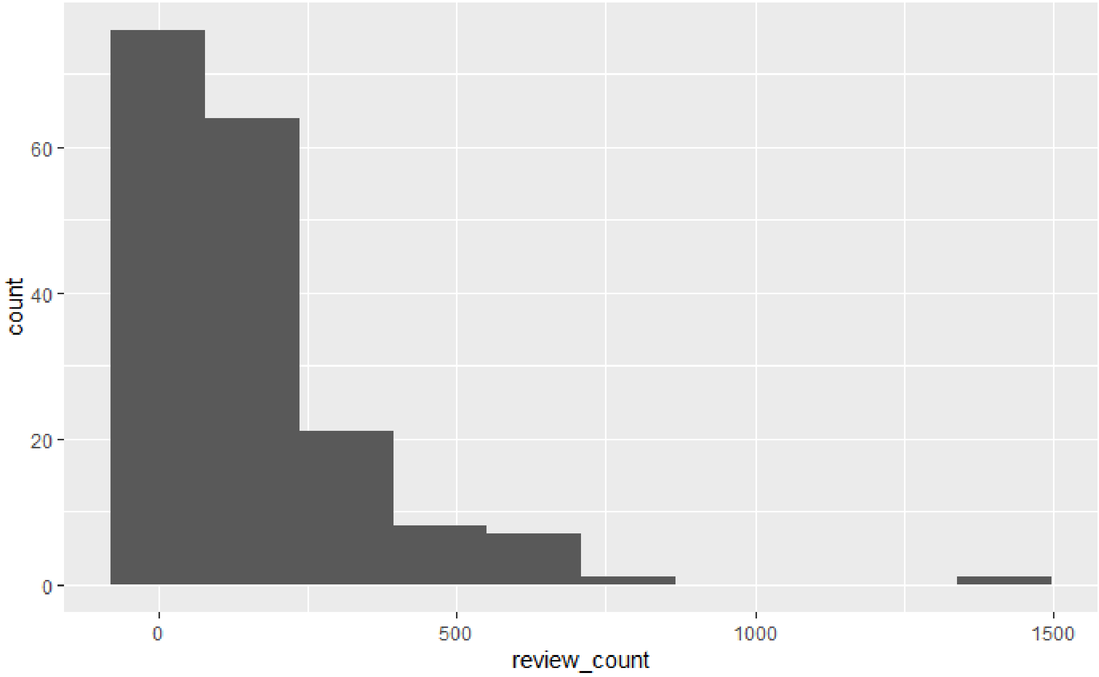
Classification of parameters necessary to deem the food cuisine as authentic or not by cleaning and wrangling of original yelp data set consisting of million rows and then shortening to 100,000 rows. Reviews, ratings or even pictures do not tell the whole story, to a user who is deciding to visit the restaurant & user won’t always get authentic food, even if it is highly rated. Identifying the parameters which are responsible for indicating how authentic food it is. Constraining the datasets to few hundred thousand rows from millions, as original yelp dataset consists information from more than 5 million users.

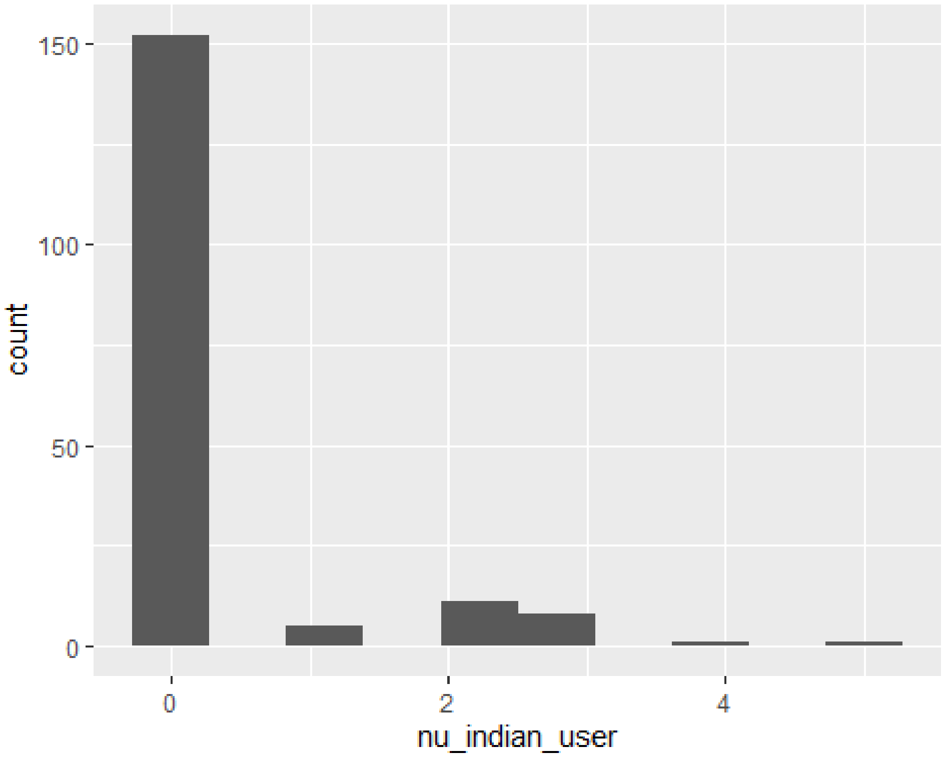
Possible solution:

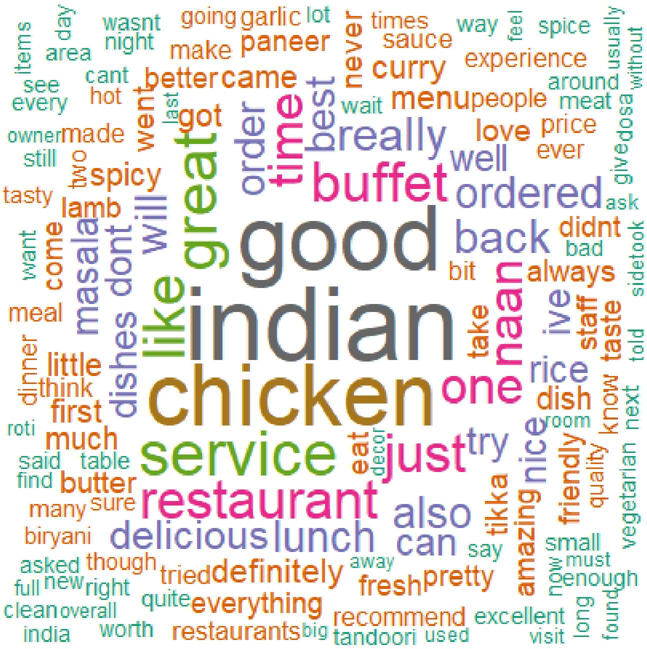
The goal was to classify the food cuisines based on no of reviews, no of users, cuisines, ratings to be authentic. Things like number of reviews & ratings by same person on restaurants serving same cuisine, name of the person (which can help identify his/ her ethnicity), age of the person, age of the account & many such parameters can be considered to understand whether a restaurant’s food can be classified as authentic or not. Using reviews to understand the sentiments of the person and categorizing it as a good or bad, can give more information about the food & help segregate those users who tend to be dissatisfied with the service instead of food. It would be interesting to explore improving the Authenticity rating by using other parts of the Yelp database — such as using modeling to extract an authenticity rating from the text of the reviews, or giving more weight to those reviewers who have also reviewedClassifying the restaurants on above mentioned parameters and using supervised machine learning techniques on training data & use it on test dataset to make sure it identifies food as authentic or not.we went on to try a method of tweaking Yelp's rating system. This approach was used on Indian restaurants and seeing how many Indian people were eating inside them (example-lots of Indian restaurants, such as Devon Avenue in Chicago). The first step was to extract all of the Indian names. After analyzing all of the names of reviewers in the "Yelp" dataset we found names ranging from "Aayush" to "Yuvraj". The next step was to extract all reviews of Indian restaurants with reviewers having those names. This turned out to be pretty simple using R's "%in%" construct and "subset" command. From there, generating the new "Indian name rating" was a simple matter of using the group_by and summarize commands in "dplyr" again.Using text mining, natural language processing libraries and sentiment analysis, we found out that the sentiment attached to the reviews containing the word "authentic" either to be positive or negative. Therefore, we can now classify the review to be authentic or not.

II. Data Exploration and Visualization

The "Yelp" data set consisted of three JSON format files of business ID's, reviews ,user ID's and there were less number of 'NA' values which were removed from the final data frame. There were 700000 rows of reviews,5000000 rows of businesses and Data quality seems good. three histograms were plotted and also the word plot.




150
100
50
0
count
0
2
4
nu_indian_user


items
day
wasnt
going
garlic
lot
never
times
way
feel
spice
usually
area
night
make
paneer
sauce
experience
without
see
cant
better
came
curry
around
every
hot
went
last
best
wait
menu
people
meat
dosa
owner
made
got
order
time
really
love
price
still
two
great
well
ever
give
tasty
spicy
buffet
ordered
ask
lamb
will
good
back
didnt
want
come
masala
dont
bad
took
bit
always
side
meal
like
indian
naan
ive
staff
taste
told
dinner
little
take
think
chicken
rice
first
dishes
one
dish
roti
much
room
know
next
service
eat
decor
just
try
nice
said
table
friendly
quality
find
butter
vegetarian
many
sure
restaurant
also
tikka
biryani
delicious
lunch
can
amazing
small
asked
though
say
away
must
full
new
right
tried
definitely
now
enough
fresh
pretty
clean
overall
quite
everything
long
india
worth
restaurants
big
recommend
excellent
found
tandoori
used
visit

III. Data Preparation and Preprocessing

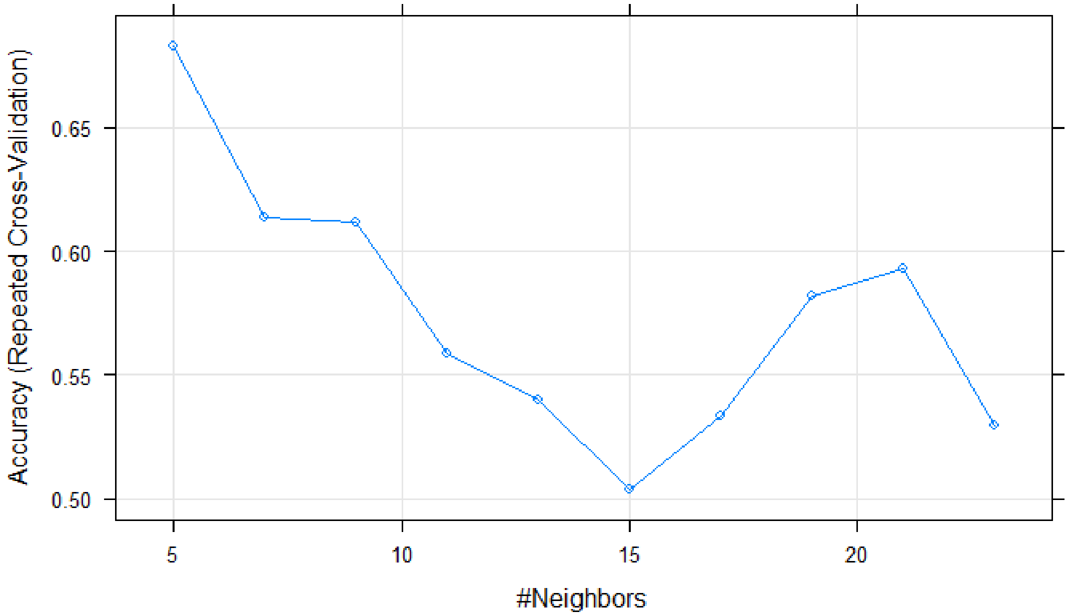
The file for reviews contained all the business using yelp and those connections were opened in chunk of 100000 rows for convenience. This was done to search for Indian reviews as we are interested to find authenticity for Indian cuisines and then they were combined into one single data frame. Unessential information such as geographical information , open 24 hrs and other categories were also removed. So finally variables selected such as business_id, city ,state, stars, review_count were selected.

IV. Data Mining Techniques and Implementation

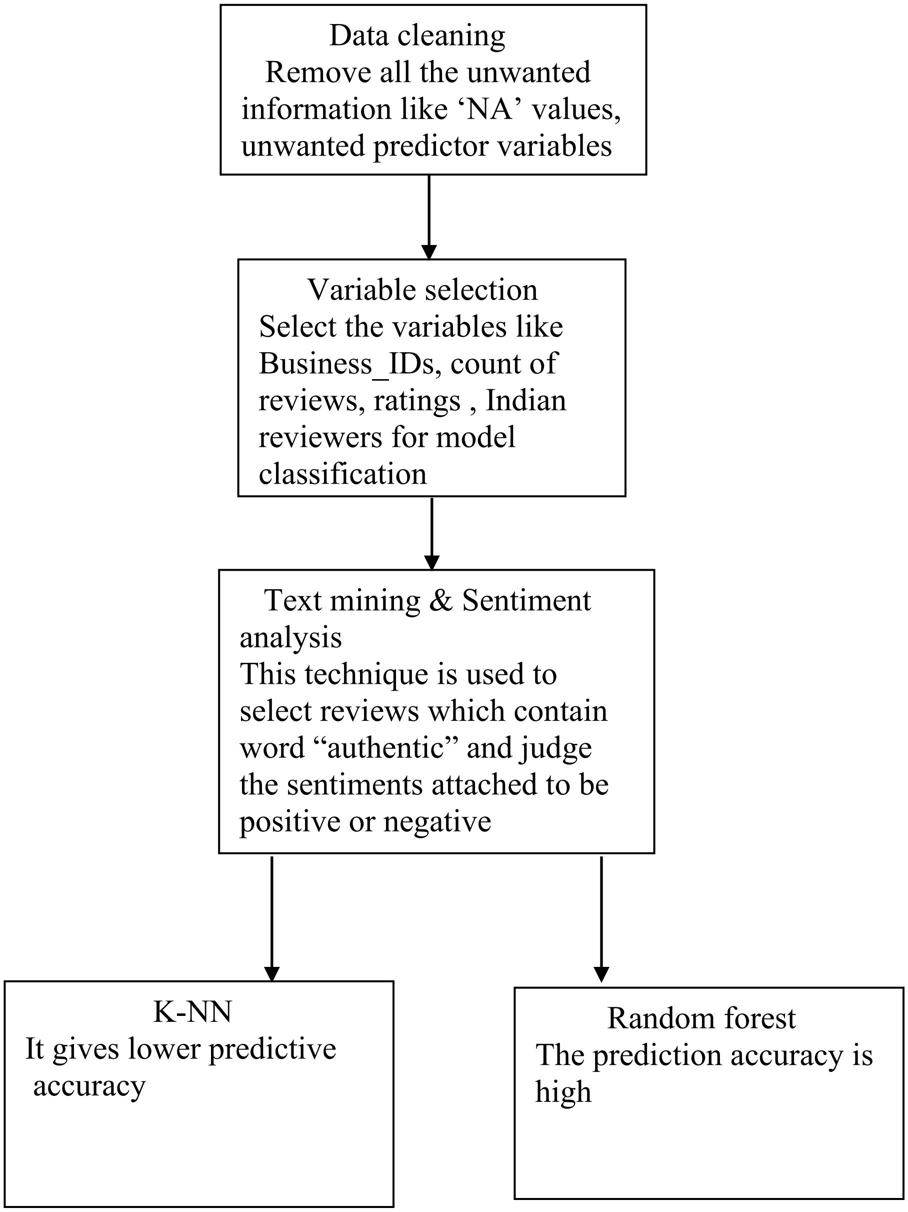
Supervised learning is used to classify the cuisines to be authentic or not. We are using the k-NN classification for classifying the reviews to be authentic or not and for that we have normalized the variables. Summary of the variables is given as follows :-

```
  Min.     1st Qu.    Median    Mean      3rd Qu.     Max.
0.00000  0.00000   0.00000   0.06742   0.00000   1.00000
```

Now , after using the k-NN classification, we get graph with certain prediction accuracy and as the number of neighbors increases we see there is a decrease in the accuracy.



So, using K-NN we found out that the classes to be authentic or not.
Now, we are using another algorithm of random forest to improve upon the prediction accuracy so that we get best results.In random forest , we get the good accuracy and lower out of bag error. Misclassification rate also comes to low.



Data cleaning
Remove all the unwanted information like 'NA' values, unwanted predictor variables

Variable selection
Select the variables like Business_IDs, count of reviews, ratings , Indian reviewers for model classification

Text mining & Sentiment analysis
This technique is used to select reviews which contain word "authentic" and judge the sentiments attached to be positive or negative

K-NN
It gives lower predictive accuracy

Random forest
The prediction accuracy is high

V. Performance Evaluation

By using K-NN we found the accuracy to be 52% where as the performance of random forest was better with 62% and all other parametrs of confusion matrix was

VI. Discussion and Recommendation

Provide discussion of the overall approach, including advantages and shortcomings. Based on your result

VII. Summarize

It was overall good project as we have made the classification for Indian cuisines to be authentic or not.

Appendix: R Code for use case study

.