CASE STUDY REPORT

Group No.: Group 17

Student Names: Bhavesh Thakkar and 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 this restaurant 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, through which we classify authenticity of the cuisine offered by restaurants. Thus, in a way improving user experience by introducing an authentic class of choices when a user demand's one.

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. 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.

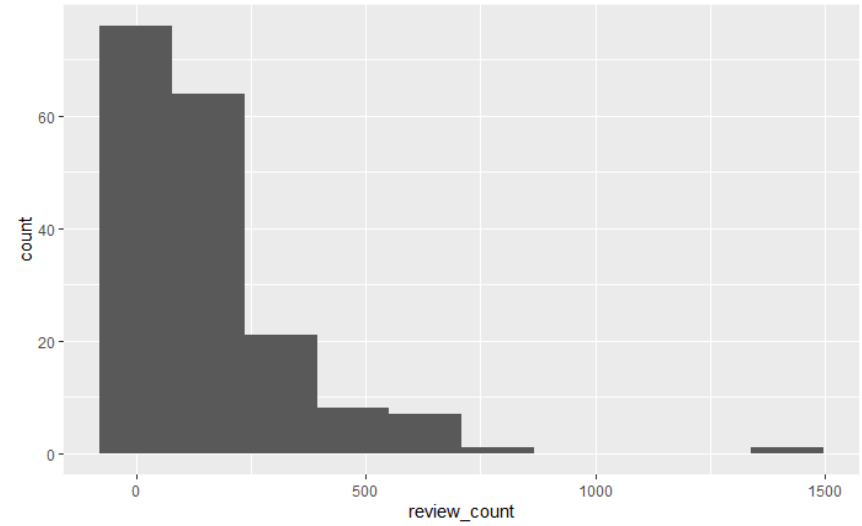
Possible solution:

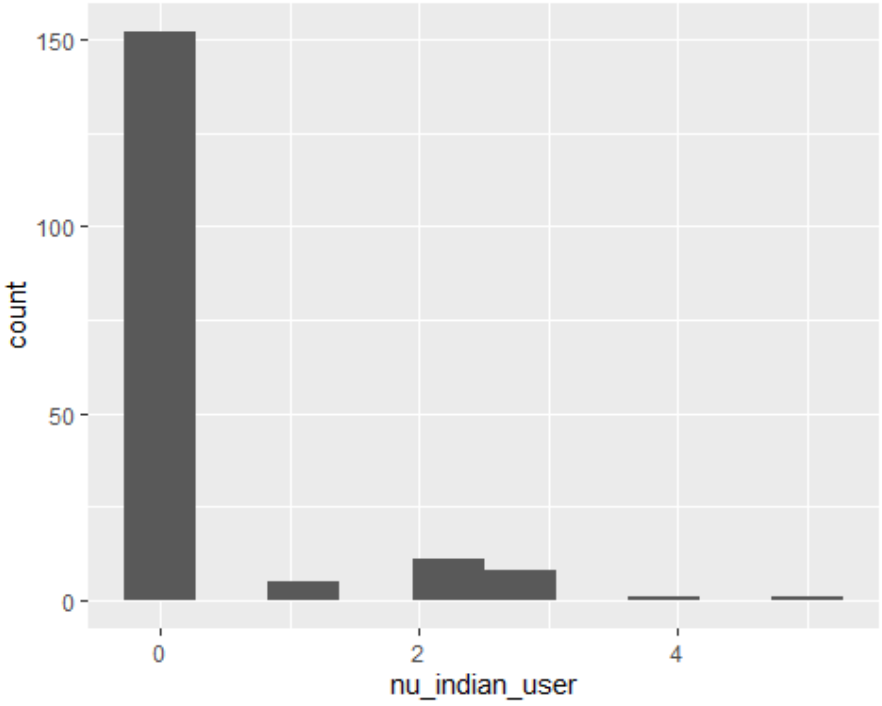
The goal is to classify the food cuisines based on no of reviews, no of users, cuisines, ratings to be authentic. Things like number of reviews, name of the person (which can help identify his/ her ethnicity) and the ratings of the restaurant can help us identify whether the cuisine offered is authentic or not. Adding authentic badges to those restaurants can make user experience significantly better, as it gives users one more parameter to select the restaurant to dine out.

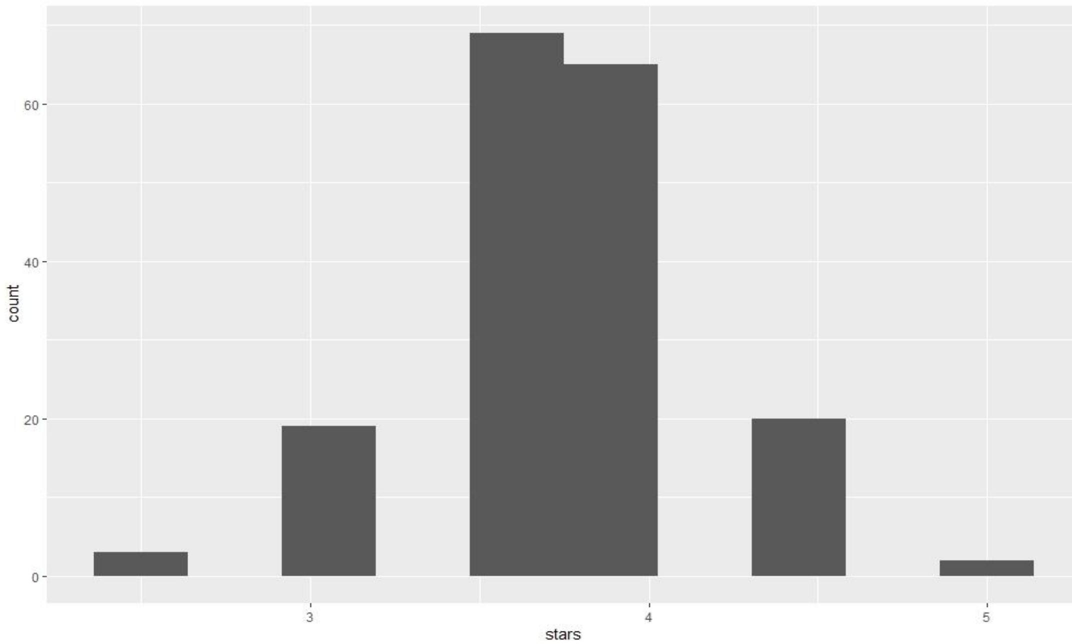
II. Data Exploration and Visualization

The "Yelp" data set consisted of three JSON format files of business ID's, reviews, user ID's and there were a smaller number of 'NA' values for the features we selected for the final data set. The original dataset contained more than 5 million observation. For the final data set, we narrowed it down to less than 200 observation. To the understand the distribution of three-predictor variable we plotted a histogram.

The first one is of total review-count, second one number of Indian users who reviewed the restaurant and the third average rating given by all the users. The fourth plot is a word plot for understanding the reviews.






count
60
40
20
0
3
4
5
stars


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

III. Data Preparation and Preprocessing

All the files for the preparation were opened in chunks as the file sizes were huge. Libraries like jsonlite were used to open JSON files. For filtering out the reviews related to Indian restaurants we created an index out of the business offering Indian cuisine and then we selected the review files on basis of that.

After the final review file was obtained, we identified the reviews which consisted of word authentic in the sentences and then performed sentiment analysis on them. The sentiment analysis we performed gave us a score between -1 to +1, with that we labeled the data set. And the final data frame consisting of business id, review count, stars or ratings, and a column computed on basis of the sentiment score classifying each restaurant as authentic or not was created.

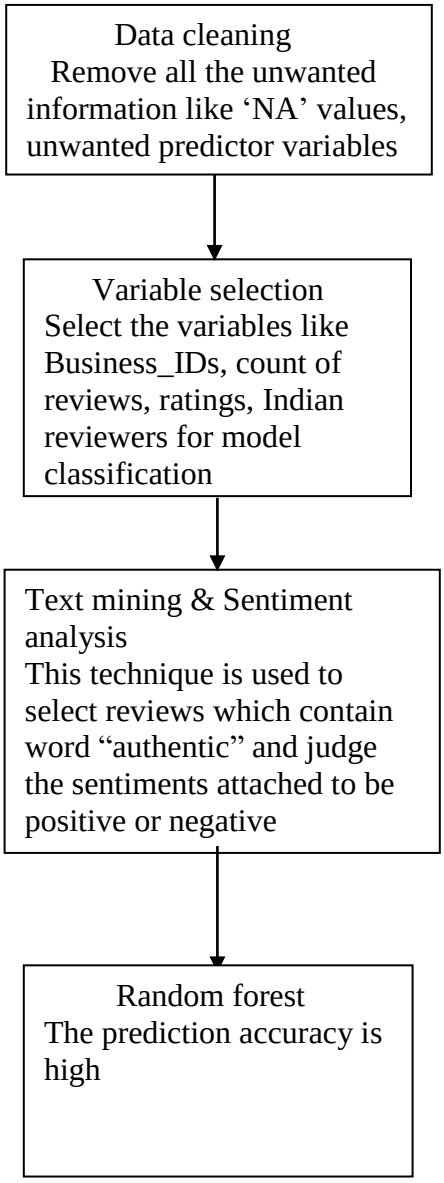
IV. Data Mining Techniques and Implementation

On the final data frame, we implemented Random Forest where we were able to obtain the accuracy of 70%.

Random Forest:

The data set was divided into two parts training and test dataset. Where the training part consisted of 40% and the testing part constituted of 60% of the total data. For the training dataset we oversampled the data using a method called SMOT or synthetic minority oversampling technique, with a help of library called DMwR. Which balanced our dataset into 50% authentic and 50% non-authentic

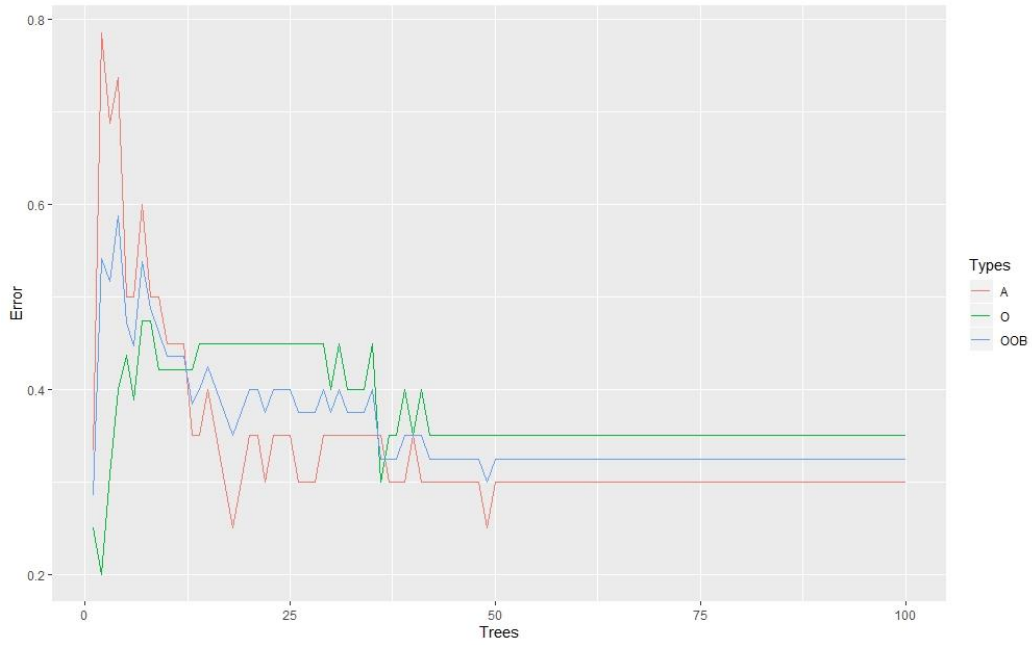
V. Flow Chart


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
Random forest
The prediction accuracy is high

VI. Performance Evaluation

Random Forest:

For finding how many numbers of tree to be selected we plotted OOB & misclassification error rate. Which gave us the number of trees to be used in Random Forest at 50.



Confusion Matrix:

	Reference	
Prediction	A	O
A	68	7
O	25	9

Accuracy: 0.7064
Sensitivity: 0.7312
Specificity: 0.5625
'Positive' Class: A