Yelp Dataset and Review Analysis

By: Amr Attyah, Ashutosh Deoghare, Nidhi Mehrotra, Radhika Mittal, Saiteja Nakka, Sanya Seth

Introduction:

Thousands of users leave a review for places they visit, give honest feedback about what they liked and what might need improvement. For businesses to pay attention to these reviews manually is a task which is why some technique that will help them draw conclusions is the need of the hour. We have applied data mining and modeling techniques on yelp dataset to analyze the customer reviews and extract meaningful results out of them.

Goal:

The goal is to classify reviews as good or bad by applying some NLP techniques for sentiment analysis and to perform predictive analytics on the dataset and explore the attributes resulting in good (star rating greater than or equal to 3) or bad rating (star rating less than 3) of US restaurants..

Data:

The data is acquired from Kaggle it had 7 csv files including yelp_business, yelp_reviews, yelp_checkins, yelp_tips. The review file had 5 million rows with review id's, business_id's, restaurant names, locations, reviews and so on that had 1.7 million businesses.

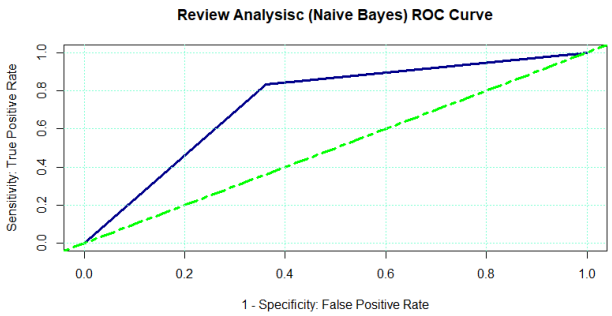
Sentiment Analysis using Text Mining package (Naïve Bayes):

For this only ten thousand reviews were considered from review file. Text and review stars columns are used from the review file. We want to use supervised learning approach for our sentimental analysis, so we are annotating the reviews based on review stars and making a new column review. This review column contains "Good" or "Bad" for review star from 0 to 2.5 we are annotating review as "Bad" and from 3 to 5 we are annotating review as "Good".

First, we converted text column into documents using VectorSource function. Then used corpus function to convert our document into corpus. After this we used text mining to clean our corpus using function tm_map by removing stop words, stemming, removing punctuations and removing white space. After cleaning the corpus, we convert our corpus into Document Term Matrix using DocumentTermMatrix function. The Document Term Matrix is the Matrix where for each review the frequency of each word will be calculated. After this data was separated for training and testing, the words with frequency less than five were removed. Then training data was given to Naïve Bayes Classifier along with review column as ground truth to train and after training we are we giving testing data to predict the reviews whether it is Good or Bad.

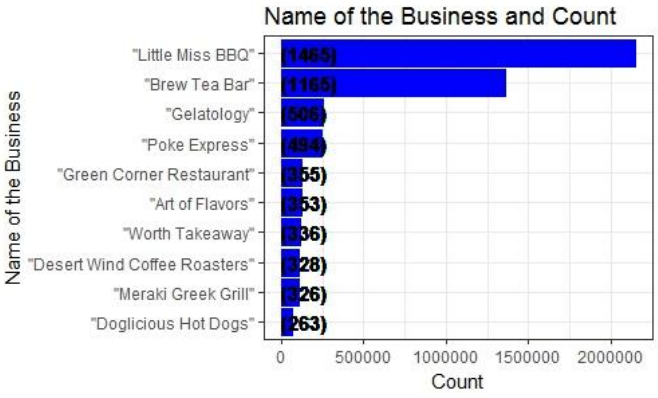
On testing set accuracy was 73.90%.

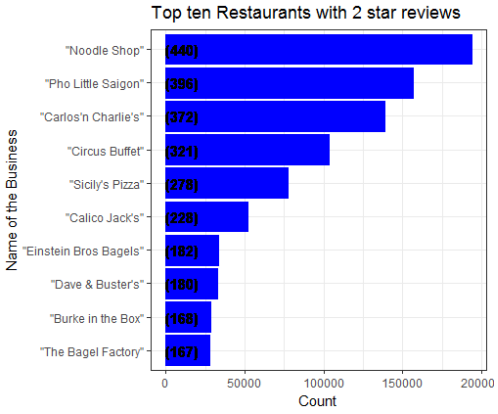
```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad   918  263
      Good  520 1299

               Accuracy : 0.739
                 95% CI : (0.7229, 0.7546)
    No Information Rate : 0.5207
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4734
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8316
            Specificity : 0.6384
         Pos Pred Value : 0.7141
         Neg Pred Value : 0.7773
             Prevalence : 0.5207
         Detection Rate : 0.4330
   Detection Prevalence : 0.6063
      Balanced Accuracy : 0.7350

       'Positive' Class : Good
```



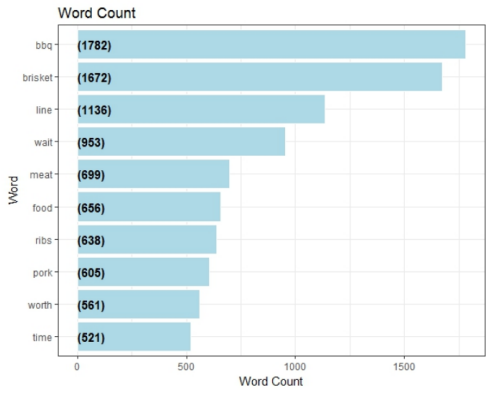
Sentiment Analysis using tidytext:

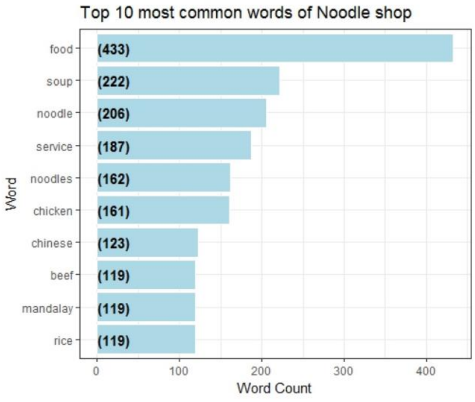
Tidytext is a library in R that has implementations of the NLP methods. These were applied on the dataset to retrieve the most preferred restaurant and least preferred.



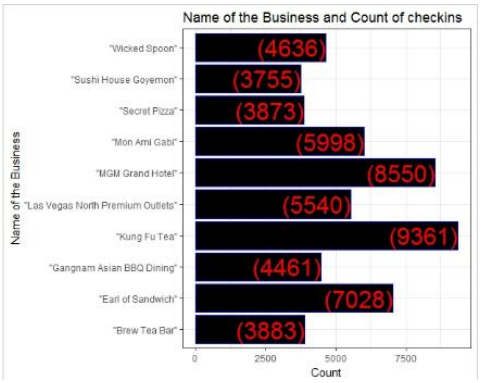


These are the top 10 restaurants with 5 star reviews on the right and 2 star reviews on the left. The locations were Nevada, Arizona and Pennsylvania. The most common/frequently occurring words in the reviews of these restaurants are shown in the graph below.



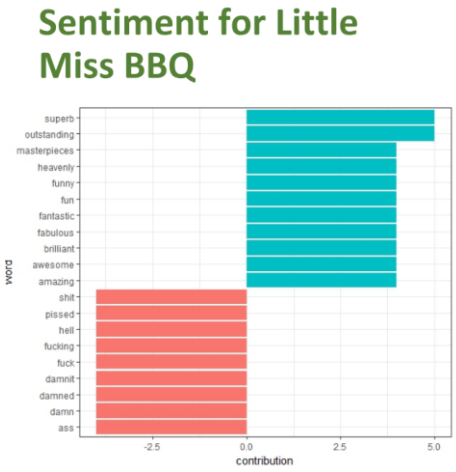


The left graph shows words for Little Miss BBQ and right one shows The Noodle Shop



Kung Fu Tea had the maximum check-ins and the location of all the above restaurants was Nevada, Las Vegas.

The wordclouds represent the words and their frequency and are ideal for deducing topics, thus we can see the left wordcloud is for the top ten restaurants with 5 star ratings and the right one is with 2 star ratings. We can compare the reviews by just looking at the cloud, the 5 star review cloud has comparatively positive words.



We can see the positive words in blue and negative ones in light red with their respective scores on a scale of -5 to +5.

The following shows the sentiment score for little miss bbq and the noodle shop:

Sentiment score for little miss bbq

```
# A tibble: 6 x 3
      X sentiment words
  <int>     <dbl> <int>
1     1  0.800000    20
2     2  1.750000    12
3     3  1.615385    13
4     4  1.214286    14
5     6  2.555556     9
6     9  1.400000     5
```

Average sentiment score for little miss bbq

```
> head(sentiment_lines)
  sentiment words
1  1.615463 10800
```

Sentiment score for the noodle shop

```
> head(sentiment_lines)
# A tibble: 6 x 3
      X sentiment words
  <int>     <dbl> <int>
1 513174  2.142857     7
2 513175 -1.666667     9
3 513177  2.857143     7
4 513179  2.444444     9
5 513180  2.700000    10
6 513182  0.000000     8
```

Average sentiment score for noodle shop

```
> head(sentiment_lines)
  sentiment words
1 0.6930052  2316
```

Negative reviews of Little Miss BBQ

	sentiment	text
7	-0.65	big line already when I got here, but I was gonna do it! I recommend parking else around the area and walk there cause you'll need that walk after you eat! My buddy and I decided to go all out and ordered as much as we could; Pork ribs, beef brisket, sausages, and beef ribs. Needless to say, the meat was tender and if slide off the bones and as you pulled them off. The aroma itself made your mouth watery! Be ready to eat with your hands and napkins, napkins, napkins is a must for your hands and face! On all the tables came with 3 homemade BBQ sauces, which were phenomenal and was based on the amount of heat you can handle. The meats were all definitely slow cooked in large vats of steam broilers. Man I could still smell it and you I swear I started to salivate as I stood in line! Overall, one of the best spots in Phoenix for sure and must go. Yes if you must play hooky from work or school and go for it. I give you permission! - Enjoy!
8	-0.666666666666667	You can't go to Phoenix and miss this place. I haven't had BBQ this good in a very long time. The brisket is to die for. If you go to long after noon or 1 you may miss out. You have to try Little Miss BBQ. It is going to be your new favorite.
9	-0.833333333333333	It's a little pricey . The staff and crew were so attentive and hospitable. We came in with the last serving portion. And they threw the (4) sides for free . The pecan pie is yucky ! There's too many Problems - the worse of them all is I should have bought a dozen !!! I end up licking the last morsel as my verdict by just buying one ... the pie in itself it's worth the line (very long)
10	-1	Sometimes the beaten path is beaten for good reason. Such is the case for Little Miss BBQ. Go at opening as they often sell out of some items, especially wondrous fatty brisket and pork ribs. I waited 90 minutes in line and did not regret it.

Previous
1
2
Next

Negative reviews of The Noodle Shop

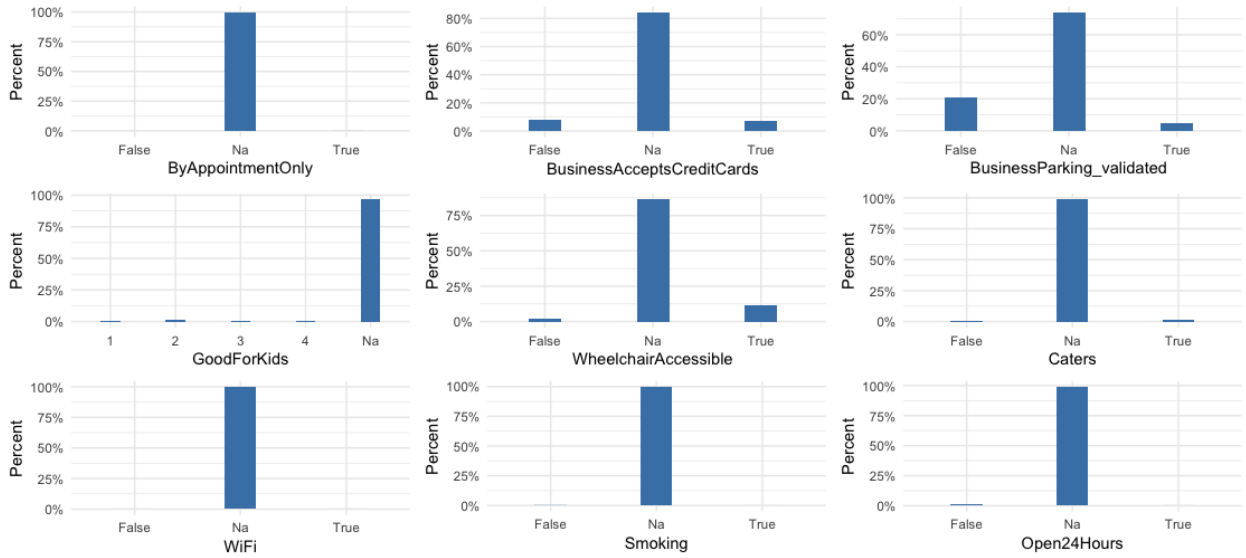
	sentiment	text
		Do not eat here.
7	-2	Yuck, Yuck, Yuck! The name is Noodle Shop, yet their noodle soup is probably THE worst on the strip or in the entire Vegas. AND probably the most expensive too! If you are stuck in Mandalay and you must have a bowl of noodle or you will die - still, don't go there. Just pick up a hamburger and go to bed. Yes, it is that bad!
8	-2.08333333333333	Me and my girlfriend walked in right behind a different couple....they got served along with ice water offered....we sat there for over 15 min without anyone asking us if we wanted anything to drink. The restaurant wasn't busy, maybe 6 tables total. WTF Finally received our water....straight up tap water where I can taste the treatment chemical...yuck. WTF The noodles soup came out very quick....alert...if you are asian...dont come here. The soup is tasteless...yuck.....one of the worst, if not the worst noodles soup ever.....and it cost $22 a bowl...WTF I think they use the can soup to cook the broth because there is no way anyone can cook such bad broth. To sum up... Bad service The worst food Super expensive
9	-2.33333333333333	If I could give this place no stars, I would. The food is bad and the service is worse. We told the waiter our pork was raw, he took it, left and never came back. And they charged us for it. They also brought the wrong main dish. I blame myself for eating here
10	-2.83333333333333	What the.... worst Asian restaurant I have ever been in. Terrible food, terrible time. They poured something odd in the water. Terrible noodles, terrible service. First and last time.

Previous
1
Next

Preprocessing and transformation:

Missing Data:

The original dataset contained around 80 binary predictors. One of the main challenges is dealing with the amount of missing data. There are many predictors like Wifi, Smoking, and Open 24 hours, that has missing values for each row. Most predictors had about 4% worth of data. The predictors that had the most information are Business Parking attributes with only 37% of the data known. Simply omitting the NA values is not an option because it would remove all the observations, so preprocessing the missing data was undeniably important.



Missing Data Imputation:

We have used mice function in R to impute NA values. Method used was pmm (predictive mean matching method) with following parameter setting: m (multiple imputations) = 5 and maxit (maximum iterations) = 10. MICE assume that the missing data are Missing at Random (MAR), meaning that the probability that a value is missing depends on the observed values in the dataset and mice uses these observed values to predict missing values.

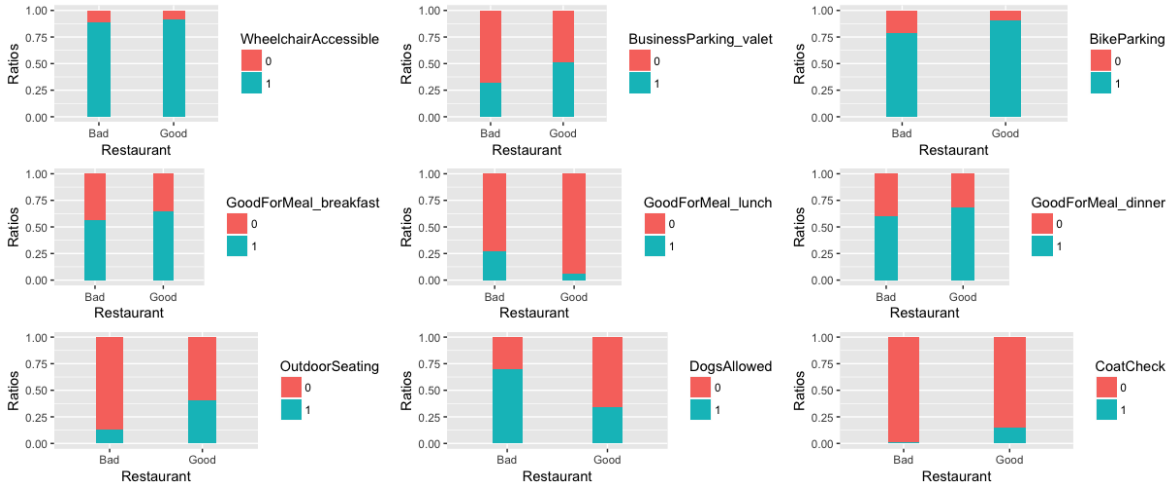
For all those columns where, missing data was less than equal to 96%; we used imputation to impute valid values.

Predictors and response:

business_id	review_count	BusinessParking_street	BusinessParking_validated	BusinessParking_lot	BusinessParking_valet	BikeParking	Caters	WheelchairAccessible	OutdoorSeating	GoodForMeal_latenight	GoodForMeal_lunch	GoodForMeal_dinner	GoodForMeal_breakfast	GoodForMeal_brunch	CoatCheck	DogsAllowed	NumberOfCheckins	BusinessType
--9e1ONYQ	1451	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	2568	Good
-01XupAWZ	77	1	0	0	1	1	0	0	1	0	0	1	0	1	0	1	295	Good
-05uZNVbb8	3	0	0	0	0	0	1	1	0	0	0	1	1	0	0	0	23	Bad
-092wE7j5H	83	0	0	0	1	1	1	1	1	0	0	1	1	0	0	0	72	Good
-0aIra_B6iA	32	0	0	0	1	1	1	1	0	0	0	1	1	0	0	0	73	Good
-0Sgh0QlUK	44	0	0	0	0	1	1	1	1	0	0	1	1	0	0	0	23	Bad
-0tgMGl7D9	120	1	0	0	0	1	1	1	0	0	0	1	1	0	0	0	772	Good
-0WegMt6C	77	1	0	0	0	1	1	1	1	1	0	1	1	0	0	0	340	Bad
-1ea69SVW	4	0	0	0	0	1	0	1	1	0	1	1	1	1	0	0	18	Bad
-1HW4ALuB	4	0	0	0	0	1	1	1	0	0	0	1	1	0	0	1	17	Bad
-1m9o3vGR	38	0	0	0	0	0	1	1	1	0	0	1	1	0	0	1	16	Good
-1UMR00eX	325	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	750	Good
-1VaIJza42I	176	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0	416	Good
-2bYV9zVtn	4	0	0	0	0	1	1	1	1	0	0	1	1	0	0	1	4	Good

Predictors: Business_Id, Review_Count, BusinessParking_Street, BusinessParking_Validated, BusinessParking_Lot, BusinessParking_Valet, BikeParking, Caters, WheelchairAccessible, OutdoorSeating GoodForMeal_latenight, GoodForMeal_lunch, GoodForMeal_dinner, GoodForMeal_breakfast, GoodForMeal_brunch, CoatCheck, DogsAllowed, NumberOfCheckins

Visualization after handling NA's:



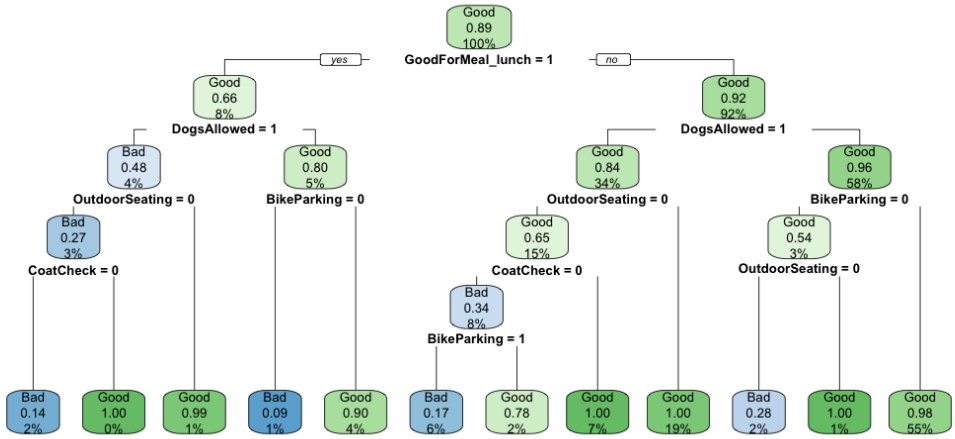
Visualizing the relationship between the predictors and the response before modeling can help us determine which predictors are expected to be informative. "Weelchairaccessible" and "GoodForMeal_Breakfast" are examples of predictors that might be uninformative because they seem to be almost the same when comparing the good and the bad restaurants. However, "DogsAllowed" seems to an important predictor, since bad restaurants are likely to allow dogs, whereas good restaurants are less likely to allow dogs.

Predictive Analytics:

Classification Tree:

cp	Accuracy	Kappa
0.04216216	0.9465772	0.6916447
0.05675676	0.9391730	0.6218817
0.06283784	0.9162769	0.3241376

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.04216216.

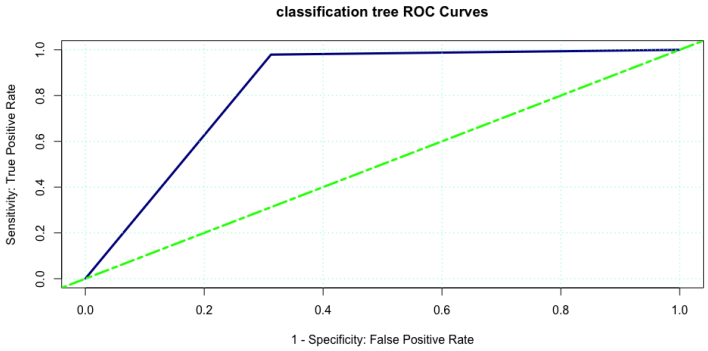


Tuning the model on the training dataset using the complexity parameter (cp) outputs different values for accuracy and Kappa. The best cp value is 0.042, which gives an accuracy of 94.65% and a kappa

of 69.16%. The tree plot above is color coded with green and blue, depending if the prediction is a good or a bad restaurant. On the training dataset it shows that only 11% of the time it will predict a restaurant to be bad, and 89% predicting it to be good.

Result

```
          Reference
Prediction  Bad Good
      Bad   350   82
      Good  159 3799

               Accuracy : 0.9451
                 95% CI : (0.9379, 0.9517)
    No Information Rate : 0.8841
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7134
 Mcnemar's Test P-Value : 9.801e-07

            Sensitivity : 0.9789
            Specificity : 0.6876
```



Evaluating the classification tree of cp = 0.042 on the testing dataset yielded good results, with an accuracy of 94.51%, kappa of 71.3%, and AUC value of 0.83.

Random Forest:

Random Forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

On testing dataset following results were observed:

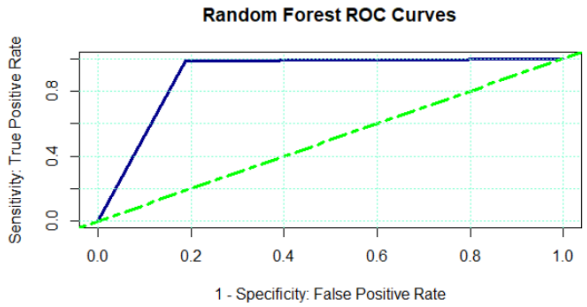
```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad   413   40
      Good   96 3841
```

```
               Accuracy : 0.969
                 95% CI : (0.9635, 0.9739)
    No Information Rate : 0.8841
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8413
 Mcnemar's Test P-Value : 2.403e-06

            Sensitivity : 0.9897
            Specificity : 0.8114
         Pos Pred Value : 0.9756
         Neg Pred Value : 0.9117
             Prevalence : 0.8841
         Detection Rate : 0.8749
   Detection Prevalence : 0.8968
      Balanced Accuracy : 0.9005

       'Positive' Class : Good
```

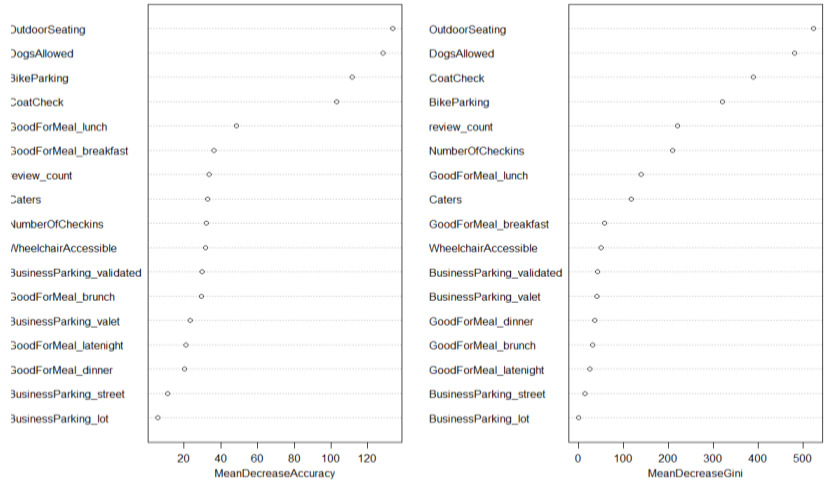
Random Forest turned out to be the most accurate model with 96.9% accuracy on test dataset among all other models. 98.9% of time model was able to correctly predict whether a restaurant was a good restaurant. Running this model with different turning parameters produced following results:

Tuning Parameter: ntree	Accuracy	Kappa
500	0.9685649	0.8386752
2000	0.9690205	0.8412983
8000	0.9690205	0.8412983

ROC Curve for Random Forest:



Importance of the Features per Random Forest:



KNN:

KNN for classification predicts a new sample using the K-closest samples from the training set. To allow each predictor to contribute equally to the distance calculation, centering and scaling of all predictors was done. 10-fold cross-validation was used to evaluate predictive model by partitioning the original sample into a training set to train the model, and a test set to evaluate it.

On the training dataset tuning the parameter 'K' gives the following output:

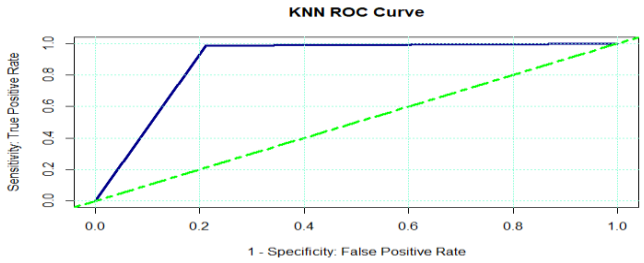
k	Accuracy	Kappa
5	0.9628091	0.7951258
7	0.9627635	0.7931175
9	0.9623533	0.7898504
11	0.9622850	0.7892763
13	0.9620459	0.7869370
15	0.9621941	0.7869768
17	0.9616358	0.7840041
19	0.9612258	0.7815432
21	0.9616245	0.7838562
23	0.9618410	0.7851602
25	0.9618182	0.7848244
27	0.9615107	0.7826602
29	0.9611462	0.7802128
31	0.9610322	0.7792957
33	0.9604627	0.7755647
35	0.9596767	0.7701756
37	0.9589249	0.7653446
39	0.9582985	0.7605693
41	0.9575923	0.7562288
43	0.9576264	0.7562045

Accuracy was used to select the optimal model using the largest value. The final value used for the model was K = 5, Accuracy = 96.28% and Kappa = 79.51%.

Result

On testing dataset following results were observed:

```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad   402   57
      Good  107 3824

               Accuracy : 0.9626
                 95% CI : (0.9566, 0.9681)
    No Information Rate : 0.8841
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8096
 Mcnemar's Test P-Value : 0.0001301

            Sensitivity : 0.9853
            Specificity : 0.7898
```

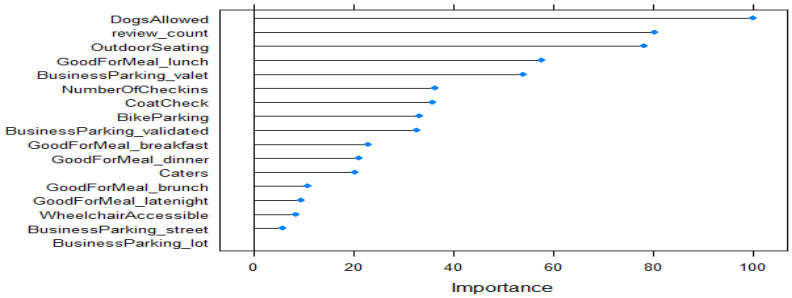


i. Accuracy = 96.26% , AUC = 0.8875

ii. Sensitivity = 0.9853 i.e. 98.53% of times the model will be able to detect a 'Good' restaurant, when it is actually 'Good'.

iii. 1 – Specificity = 0.2102 i.e. 21.02% of times the model will not be able to detect a 'Bad' restaurant, when it is actually 'Bad'.

Key features based on KNN:



SVM:

For Support Vector Machine two types of kernel were used:

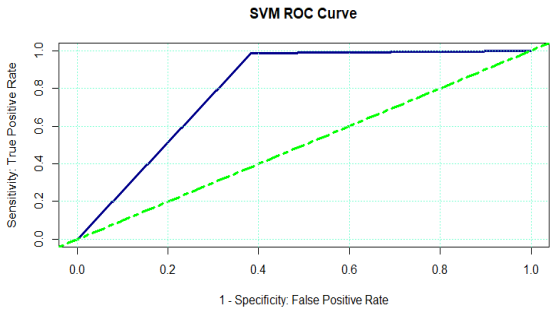
1. svmLinear: This kernel is from kernlab package.
2. svmLinear2: This kernel is from e1071 package.

To allow each predictor to contribute equally to the distance calculation, centering and scaling of all predictors was done. After partitioning data in training and testing set we are doing 10-fold cross validation on our training data. The tuning parameters to tune our SVM model on our testing data is given below:

Tuning Parameters	Accuracy (Cost = 0.01)	Accuracy (Cost = 0.1)	Accuracy (Cost = 1)
svmLinear	89.57%	91.58%	93.94%
svmLinear2	91.08%	93%	94.58%

From the above parameters, we get the best result for svmLinear2 model and the ROC curve for that is given below:

```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad   290   57
      Good  181 3860

               Accuracy : 0.9458
                 95% CI : (0.9386, 0.9523)
    No Information Rate : 0.8927
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6799
 Mcnemar's Test P-Value : 1.55e-15

            Sensitivity : 0.9854
            Specificity : 0.6157
         Pos Pred Value : 0.9552
         Neg Pred Value : 0.8357
             Prevalence : 0.8927
         Detection Rate : 0.8797
   Detection Prevalence : 0.9209
      Balanced Accuracy : 0.8006

       'Positive' Class : Good
```

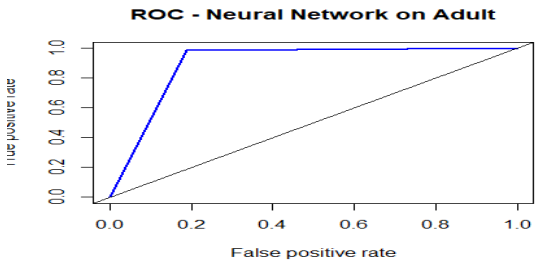


Neural Network:

We have used nnet package in R, which allows us to limit the number of hidden layers to one, we decided to model it with 40 nodes in that single hidden layer, since it is commonly believed that the number of nodes should be close to double the number of predictors. The activation function between the nodes in different layer was default chosen to softmax function.

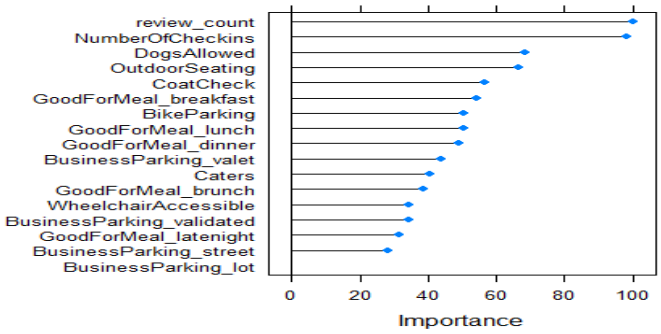
The loss function i.e., error between the predicted values and the test set is default chosen to least squares, and this function is optimized using the stochastic gradient descent approach with learning rate of 0.05, but the accuracy using 0.01 learning rate was little more better, but it is computationally time taking. Since, the model has a lot of training samples and predictors, it needs a lot number of iteration for the efficient model, for which we set them to 1000. Neural network being the most complex model, trains the model repeatedly until it achieves the least error. Training data was scaled and normalized for better performance.

```
Confusion Matrix and Statistics

          Reference
Prediction  Bad Good
      Bad   417   53
      Good   92 3828

               Accuracy : 0.967
                 95% CI : (0.9612, 0.9721)
    No Information Rate : 0.8841
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8333
 Mcnemar's Test P-Value : 0.001601

            Sensitivity : 0.9863
            Specificity : 0.8193
         Pos Pred Value : 0.9765
         Neg Pred Value : 0.8872
             Prevalence : 0.8841
         Detection Rate : 0.8720
   Detection Prevalence : 0.8929
      Balanced Accuracy : 0.9028

       'Positive' Class : Good
```



The ROC curve with an area of 0.97 under shows us how well our model performed.

Below are the predictors that contributed most in building our model in descending order:



So, the neural network model with 40 number of nodes and 1000 iterations has evaluated the test set with an accuracy of 96.7%.

Conclusion:

Of all the models Random Forest performed the best on our data based on accuracy. With outdoor seating, dogs allowed, bike parking, coat check as top four predictors. So, if you want a 'Good' rating based on our analysis do not allow dogs, have outdoor seating, bike parking and coat check service available. It is interesting to note that a restaurant can be classified as 'Good' i.e. can have 3 or more stars on scale of 1-5 without considering food as a factor. As people consider the whole experience while dinning out. If you want to expand on the project and be able to classify restaurants based on stars, then we suggest finding more data related to quality of food. As we believe it is one of the most crucial factors in deciding restaurant rating.