CS 9223 Foundation of Data Science

Project Report

Predicting the volume of Yelp reviews during NFL Season for sports bar of USA cities and compare the change from previous year

Submitted By
Bhavin Vijay Mehta
Shyam Rajendra Joshi

1) What did you propose to do? What is the motivation/background?

Our proposal was to identify the change in volume of reviews for Sports bar in USA cities during the NFL season and predict this change.

Businesses around the world are paying more and more attention to Yelp, arguably the Internet's largest social-local platform, and with good reason. Touting its average 135 million monthly users, dominance in local search engine results, and ability to shut down businesses overnight, Yelp is force to be reckoned with. As per BrightLocal, 92% of consumers now read online reviews for local business. Hence every bit of analysis done on Yelp Dataset is valuable for business. We propose to analyze the Yelp Dataset and focus our analysis on Sports Bar in USA. We aim to identify if NFL season has any change in volume of reviews for Sports bar. NFL season is of importance for Sports Bars since it has been the most famous sports for 30 consecutive years in USA, as stated by Sports Illustrated.

On checking the reviews for sport bars on Yelp, we find that people pay attention to the bar's environment while watching the game. People's review affect the bar's business and even more for a Sports bar during a major event like NFL. If we see that the NFL does have some impact on the volume of reviews for Sports bar, we can than further do sentimental analysis to identify how positive or negative are these reviews. People tend to go to Sports Bar during NFL season, where they can enjoy socializing, get all the important updates related to matches and in between timeout entertainment from such businesses. People post reviews to share their experience, suggest improvement, support their favorite team and read the reviews to select a bar with similar taste of people. On top of this, Elite users contributed 10% of Yelp reviews. All these reviews decide the credibility of a Sport Bar. Therefore, it is important for Sports Bar to analysis reviews during NFL season. Most of the NFL matches are hosted over the weekends and people love football because the teams are valued more equally than in baseball, reports columnist Michael Fitzgerald at "Bleacher Report."

The below paper discusses how people watch sports at Sports bar and how customer reviews make an impact on specific business.
http://plei-plei.info/wp-content/uploads/2012/03/tv-watching-at-sports-bars-as-social-interaction.pdf
http://www.hbs.edu/faculty/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-8f951238dba2.pdf
http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business

Below are the link where we understood how users put reviews on Yelp and how authenticity of reviews are important for all the user.
https://biz.yelp.com/support/responding_to_reviews
https://www.fundera.com/blog/2015/05/28/yelp-reviews-does-anybody-really-care

2) Explain the data you used and model in detail.

We are using the following datasets

1. Yelp Dataset
   Obtained from Yelp Dataset Challenge https://www.yelp.com/dataset_challenge
2. NFL Schedule
   Obtained by web scrapping from
   http://static.pfref.com/years/2015/games.htm#games::none

Yelp Dataset

The yelp dataset contains

- 2.7M reviews and 649K tips by 687K users for 86K businesses 566K business attributes, e.g., hours, parking availability, ambience.
- Social network of 687K users for a total of 4.2M social edges.
- Aggregated check-ins over time for each of the 86K businesses
- 200,000 pictures from the included businesses

Below is the schema for the yelp dataset,
We have removed some of the features that we are not using in this project. The removed features are marked in Red.

business
{
'business_id': (encrypted business id), Type: String
'name': (business name), Type: String
'city': (city), Type: String
'state': (state), Type: String
'stars': (star rating, rounded to half-stars), Type: Numeric
'review_count': review count, Type: Numeric
'categories': [(localized category names)] Type: String
'type': 'business', Type: String
'neighborhoods': [(hood names)], Type: String
'full_address': (localized address), Type: String
'latitude': latitude, Type: String
'longitude': longitude, Type: String
'open': True / False (corresponds to closed, not business hours), Type: String
'hours': {(day_of_week): {'open': (HH:MM),'close': (HH:MM)},...}, Type: String
'attributes': {(attribute_name): (attribute_value),...}, Type: String
}

review
{
'business_id': (encrypted business id), Type: String
'user_id': (encrypted user id), Type: String
'stars': (star rating, rounded to half-stars), Type: Numeric
'date': (date, formatted like '2012-03-14'), Type: Date, Temporal
'type': 'review', Type: String
'text': (review text), Type: String
'votes': {(vote type): (count)}, Type: Numeric
}

User

{
'user_id': (encrypted user id), Type: String
'review_count': (review count), Type: Numeric
'elite': [(years_elite)], Type: String
'type': 'user', Type: String
'name': (first name), Type: String
'average_stars': (floating point average, like 4.31), Type: Numeric
'votes': {(vote type): (count)}, Type: Numeric
'friends': [(friend user_ids)], Type: String
'yelping_since': (date, formatted like '2012-03'), Type: Date
'compliments': {(compliment_type): (num_compliments_of_this_type), ... }, Type: String
'fans': (num_fans), Type: Numeric
}

NFL Dataset

The NFL data is obtained using web scrapping from PRO FOOTBALL REFERENCE website.
The data is obtained for 2012,2013,2014 and 2015 Seasons.
Each season has 17 weeks of play before the playoff starts. Playoff starts with Wildcard, Division, ConfChamp and ends with SuperBowl. The weeks of playoff are label as Week 18,Week 19, Week 20 and Week 21 respectively.
Below is the schema for NFL Dataset.

NFL Schedule
{
Week : Number Type: Numeric
Day : String (Mon,Tue) Type: String
Date: Date Type: String
Time : TimeStamp Type: Date
Winner/Tie: Team Name Type: String
Loser/Tie : Team Name Type: String
PTSW : Points Scored By Winning Team Type: Numeric

PtsL -- Points Scored by the losing team (second one listed) Type: Numeric
YdsW -- Yards Gained by the winning team (first one listed) Type: Numeric
TOW -- Turnovers by the winning team (first one listed) Type: Numeric
YdsL -- Yards Gained by the losing team (second one listed) Type: Numeric
TOL -- Turnovers by the losing team (second one listed) Type: Numeric
}

- All the unwanted columns were removed and data was cleaned before processing.
- Yelp dataset was provided in JSON format. All the strings were converted to lower case. Duplicate records were removed as needed. All variables are mapped to the corresponding R datatypes. Date datatype is used for the Date variables.
- After cleaning the datasets, we prepare the data for the model used.
- We take only the business that have category Sports Bar. Then we filter the reviews for these bars. Finally, we take only the reviews that fall in the NFL season week. To identify these reviews, we take into consideration the next day of the NFL match also. Week refers to the NFL game week.
- We then aggregate these over cities and take the count of reviews for each city for each week in the NFL schedule. We then hand label the review counts that we obtain. Review counts are divided into class labels each of length 5. The Mean of the aggregated data is 7 and the median is 3. We wanted the labels to be not so limited and as well as not too sensitive, hence we are taking 5 reviews as our class interval. Using this we label the aggregate data into classes. For example, Las Vegas with review count 4 in week 1 will be assigned class 1. Similarly, Charlotte with review count 6 in week 1 will fall in class 2.
- After hand labelling the data we divide this labelled data into training set and testing set. The labelled data corresponding to the 2012, 2013 and 2014 NFL Seasons is taken as the training set and those belonging to the 2015 NFL Season is taken as the test data.

Models used

Initially, we planned to apply the linear regression on the data set without the labelled data. We find that linear regression is not feasible since the data is skewed and sparse. Also, there is no linear relation among variables that we found. Hence linear regression is not used as a feasible model.

We then apply the supervised learning model for classification. Below classification models were used

1. SVM Multiclass Classification
2. Random Forest
3. Multinomial logit model

The following are the predictors that we considered

Predictors

City
NFL Season
NFL Week
Number of Reviews with 5 stars rating
Number of Reviews with 4 stars rating
Number of Reviews with 3 stars rating
Number of Reviews with 2 stars rating
Number of Reviews with 1 stars rating

Number of Elite user reviews
Number of Non-Elite user reviews

Response variable
Label – Hand labelled for Number of reviews, divided into interval of 5.

The data from 2012 to 2014 NFL season was used for training and the data of 2015 NFL season was used for testing. This is turns out to be around 75% data for training and 25% data for testing.

SVM Multiclass Classification
The SVM classification model was applied on the label data. We used confusion matrix for evaluation

The SVM confusion matrix is displayed below,

```
> svm_confusion_matrix
Confusion Matrix and Statistics

          Reference
Prediction  0  1  2  3  4  5  6
         0  0  5  0  0  0  0  0
         1  0 81  0  0  0  0  0
         2  0  4 42  2  0  0  0
         3  0  0  1 24  0  0  0
         4  0  0  0  2  4  1  0
         5  0  0  0  0  1  4  0
         6  0  0  0  0  0  2  0

Overall Statistics

               Accuracy : 0.896
                 95% CI : (0.8406, 0.9372)
    No Information Rate : 0.5202
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8428
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity                NA   0.9000   0.9767   0.8571  0.80000  0.57143       NA
Specificity            0.9711   1.0000   0.9538   0.9931  0.98214  0.99398  0.98844
Pos Pred Value             NA   1.0000   0.8750   0.9600  0.57143  0.80000       NA
Neg Pred Value             NA   0.9022   0.9920   0.9730  0.99398  0.98214       NA
Prevalence             0.0000   0.5202   0.2486   0.1618  0.02890  0.04046  0.00000
Detection Rate         0.0000   0.4682   0.2428   0.1387  0.02312  0.02312  0.00000
Detection Prevalence   0.0289   0.4682   0.2775   0.1445  0.04046  0.02890  0.01156
Balanced Accuracy          NA   0.9500   0.9653   0.9251  0.89107  0.78270       NA
```

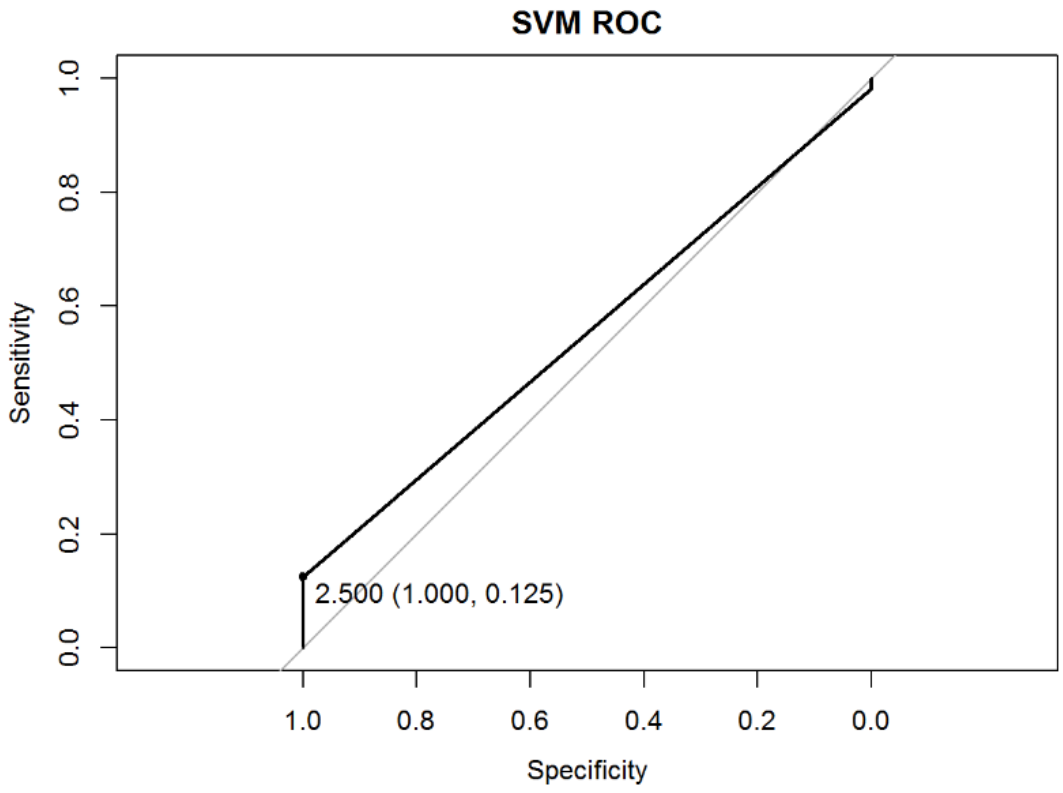
From the confusion matrix, we can see that SVM gives a good accuracy of 0.896 (89.6%). The sensitivity for class 1, class 2, class 3 and class 4 is very high, indicating that the model retrieved most of the relevant information for these classes, but we also see that specificity is high as well. Hence SVM prediction includes a lot of junk values. Also, the boundary class, class 1 and class 6, contains NA values and are not predicted. SVM is not able to predict these boundary classes.

From the ROC curve, we see that the SVM falls in the worthless area when we compare to the ideal ROC curve as mentioned in the lecture slides. The AUC is 0.5 indicating that SVM classifier gives a bad performance.

To improve the performance of the SVM, we penalize the model. By penalizing a model we avoid the repeated classification to a same class. We see the NA values that were obtained earlier for the boundary classes were corrected when a penalizing cost was included. Below is the confusion matrix for SVM including a penalizing cost. The accuracy came out to be 0.8435(84.35%).

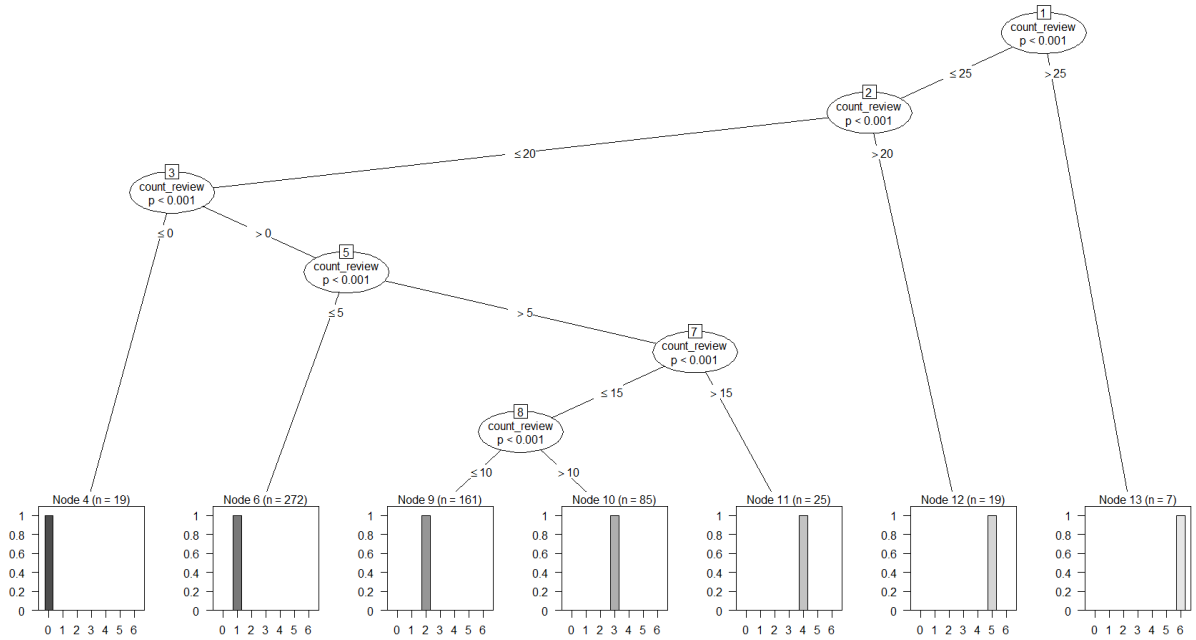
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1  2  3  4  5  6
##          0  0  1  0  0  0  0  0
##          1  2 48  0  0  0  0  0
##          2  0  7 40  1  0  0  0
##          3  0  0  3 18  4  0  0
##          4  0  0  0  0  8  1  0
##          5  0  0  0  0  0  9  4
##          6  0  0  0  0  0  0  1
##
## Overall Statistics
##
##                Accuracy : 0.8435
##                  95% CI : (0.7745, 0.8982)
##     No Information Rate : 0.381
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.789
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity          0.000000   0.8571   0.9302   0.9474  0.66667  0.90000
## Specificity          0.993103   0.9780   0.9231   0.9453  0.99259  0.97080
## Pos Pred Value       0.000000   0.9600   0.8333   0.7200  0.88889  0.69231
## Neg Pred Value       0.986301   0.9175   0.9697   0.9918  0.97101  0.99254
## Prevalence           0.013605   0.3810   0.2925   0.1293  0.08163  0.06803
## Detection Rate       0.000000   0.3265   0.2721   0.1224  0.05442  0.06122
## Detection Prevalence 0.006803   0.3401   0.3265   0.1701  0.06122  0.08844
## Balanced Accuracy    0.496552   0.9176   0.9267   0.9463  0.82963  0.93540
##                      Class: 6
## Sensitivity          0.200000
## Specificity          1.000000
## Pos Pred Value       1.000000
## Neg Pred Value       0.972603
## Prevalence           0.034014
## Detection Rate       0.006803
## Detection Prevalence 0.006803
## Balanced Accuracy    0.600000
```

Below is the ROC curve for SVM.



Random Forest

After SVM we applied the Random Forest model on the labelled data. Before applying the random forest, we create a decision tree for the labelled data, to identify probabilities for each class label. Decision tree shows all the factors that are considered relevant to the decision. Below is the decision tree for the labelled data.



From the decision tree, we see that every class has a high probability for prediction.
Random Decision Forest is very effective in eliminating noise in the model input data. Since our labelled data input for the model is imputed, decision tree helps to control overfitting and identifies decisions that would generalize well enough to future input it has not seen.

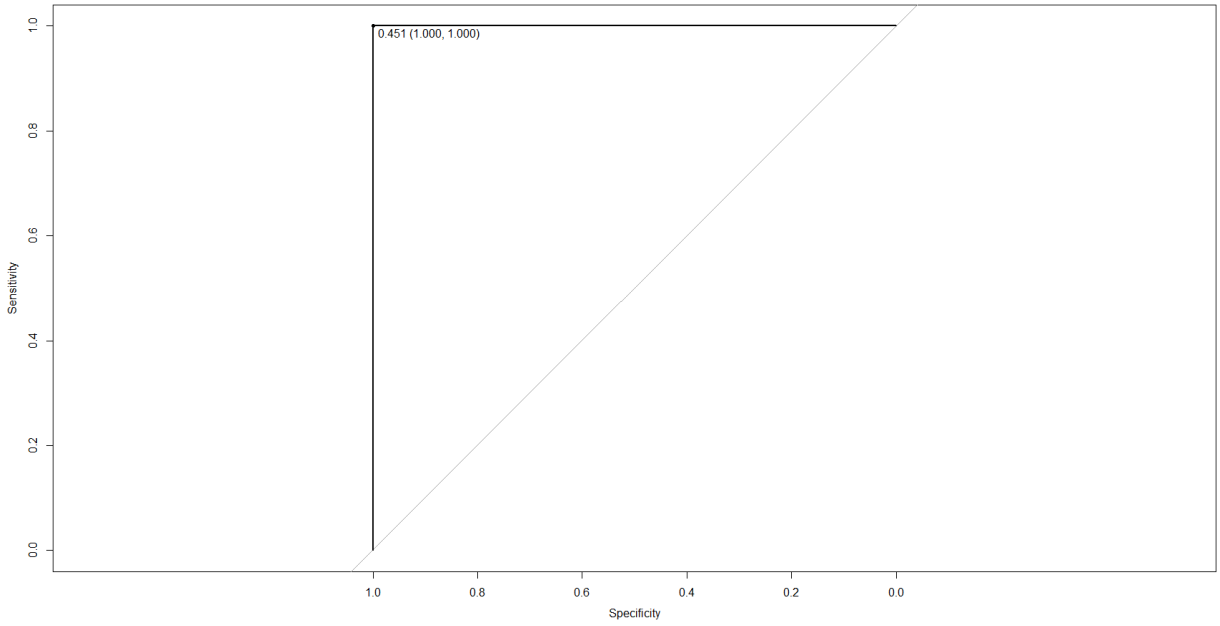
Random Forest builds many trees using a subset of the available features. It builds some underlying decision trees that omit the noise generating features. In the end, when it is time to generate a prediction a vote among all the underlying trees takes place and the majority prediction value wins.

We applied Random Forest Classifier on the labelled data. The results of Random Forest are better than SVM and Multinomial Model. Below is the confusion matrix for Random Forest.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1  2  3  4  5  6
##          0  2  0  0  0  0  0  0
##          1  0 55  0  0  0  0  0
##          2  0  1 40  0  0  0  0
##          3  0  0  3 19  2  0  0
##          4  0  0  0  0 10  0  0
##          5  0  0  0  0  0 10  3
##          6  0  0  0  0  0  0  2
##
## Overall Statistics
##
##                Accuracy : 0.9388
##                  95% CI : (0.887, 0.9716)
##     No Information Rate : 0.381
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9175
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity           1.00000   0.9821   0.9302   1.0000  0.83333  1.00000
## Specificity           1.00000   1.0000   0.9904   0.9609  1.00000  0.97810
## Pos Pred Value        1.00000   1.0000   0.9756   0.7917  1.00000  0.76923
## Neg Pred Value        1.00000   0.9891   0.9717   1.0000  0.98540  1.00000
## Prevalence            0.01361   0.3810   0.2925   0.1293  0.08163  0.06803
## Detection Rate        0.01361   0.3741   0.2721   0.1293  0.06803  0.06803
## Detection Prevalence  0.01361   0.3741   0.2789   0.1633  0.06803  0.08844
## Balanced Accuracy     1.00000   0.9911   0.9603   0.9805  0.91667  0.98905
##                      Class: 6
## Sensitivity           0.40000
## Specificity           1.00000
## Pos Pred Value        1.00000
## Neg Pred Value        0.97931
## Prevalence            0.03401
## Detection Rate        0.01361
## Detection Prevalence  0.01361
## Balanced Accuracy     0.70000
```

The accuracy for random forest is 0.9388 (93.88%).
The Sensitivity and specificity are not too high and lies between 0.5 and 1.00. This indicates that the Random Forest has good performance. The ROC curve is display below, and we see that it falls in between the good and the excellent region.

ROC for Random Forest



We also applied the random forest model with different combination of the predictors, to identify which combination of these predictors are strongest.
The following predictors gave high accuracy for the random forest
City, NFL Week, Star 5, Elite count, Non Elite Count

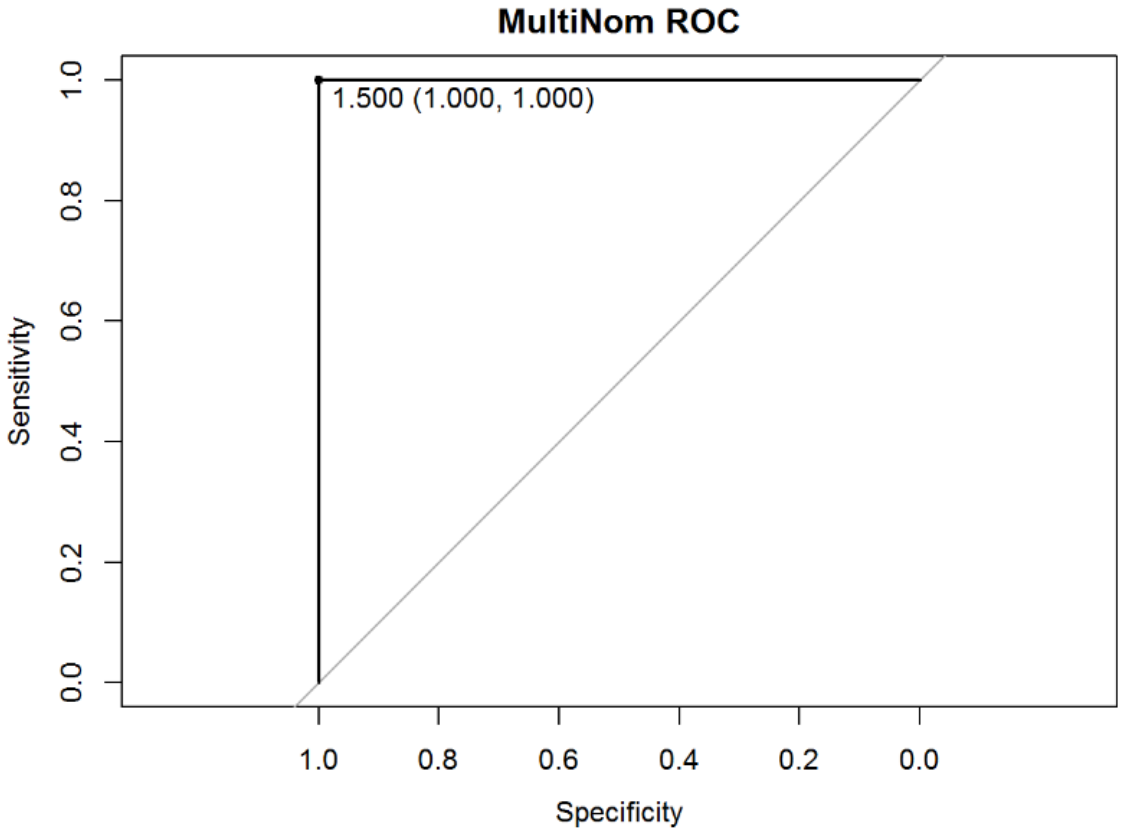
Multinomial Logit Model

We also apply the Multinominal logit model. It is a probabilistic model and estimates maximum likelihood depending upon the probabilities of the response categories by nominating one class as the baseline or reference. The multinomial logit models have a dependent variable that is a categorical, unordered variable which suits our response variable. It works on non-linear relationships between features and generates co-efficients with marginal effects from the reference label.

Below is the confusion matrix for multinomial model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1  2  3  4  5  6
##          0  2  0  0  0  0  0  0
##          1  0 51  0  0  0  0  0
##          2  0  5 40  0  0  0  0
##          3  0  0  1 16  0  0  0
##          4  0  0  0  0  8  0  0
##          5  0  0  1  0  1  9  0
##          6  0  0  1  3  3  1  5
##
## Overall Statistics
##
##                Accuracy : 0.8912
##                  95% CI : (0.8293, 0.9365)
##     No Information Rate : 0.381
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.855
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
## Sensitivity           1.00000   0.9107   0.9302   0.8421  0.66667  0.90000
## Specificity           1.00000   1.0000   0.9519   0.9922  1.00000  0.98540
## Pos Pred Value        1.00000   1.0000   0.8889   0.9412  1.00000  0.81818
## Neg Pred Value        1.00000   0.9479   0.9706   0.9769  0.97122  0.99265
## Prevalence            0.01361   0.3810   0.2925   0.1293  0.08163  0.06803
## Detection Rate        0.01361   0.3469   0.2721   0.1088  0.05442  0.06122
## Detection Prevalence  0.01361   0.3469   0.3061   0.1156  0.05442  0.07483
## Balanced Accuracy     1.00000   0.9554   0.9411   0.9171  0.83333  0.94270
##                      Class: 6
## Sensitivity           1.00000
## Specificity           0.94366
## Pos Pred Value        0.38462
## Neg Pred Value        1.00000
## Prevalence            0.03401
## Detection Rate        0.03401
## Detection Prevalence  0.08844
## Balanced Accuracy     0.97183
```

We can see that the accuracy of multinomial is 0.8912(89.12%). It has higher accuracy than SVM. Below is the ROC curve for multinomial model



3) What did you end up doing?

- We thought of linear regression initially, but as stated above, linear regression cannot be used in this case.
- So used supervised learning classification approach.
- We performed other model application, penalizing model behavior, avoiding over-fitting data and comparison of modelling accuracy.
- In terms of hand labelling the features and response the approach and processing remains the same.
- Furthermore, we used text mining to find the NFL sports related review texts to differentiate the data from NBA,NHL and MLB
- We first cleaned the data, prepared the data for the model and finally applied 3 classification models on the hand labelled data. Among these models, RandomForest gave the best results.

4) What if anything did you change about your approach and why?

A few of the changes that we made along the way are the following:

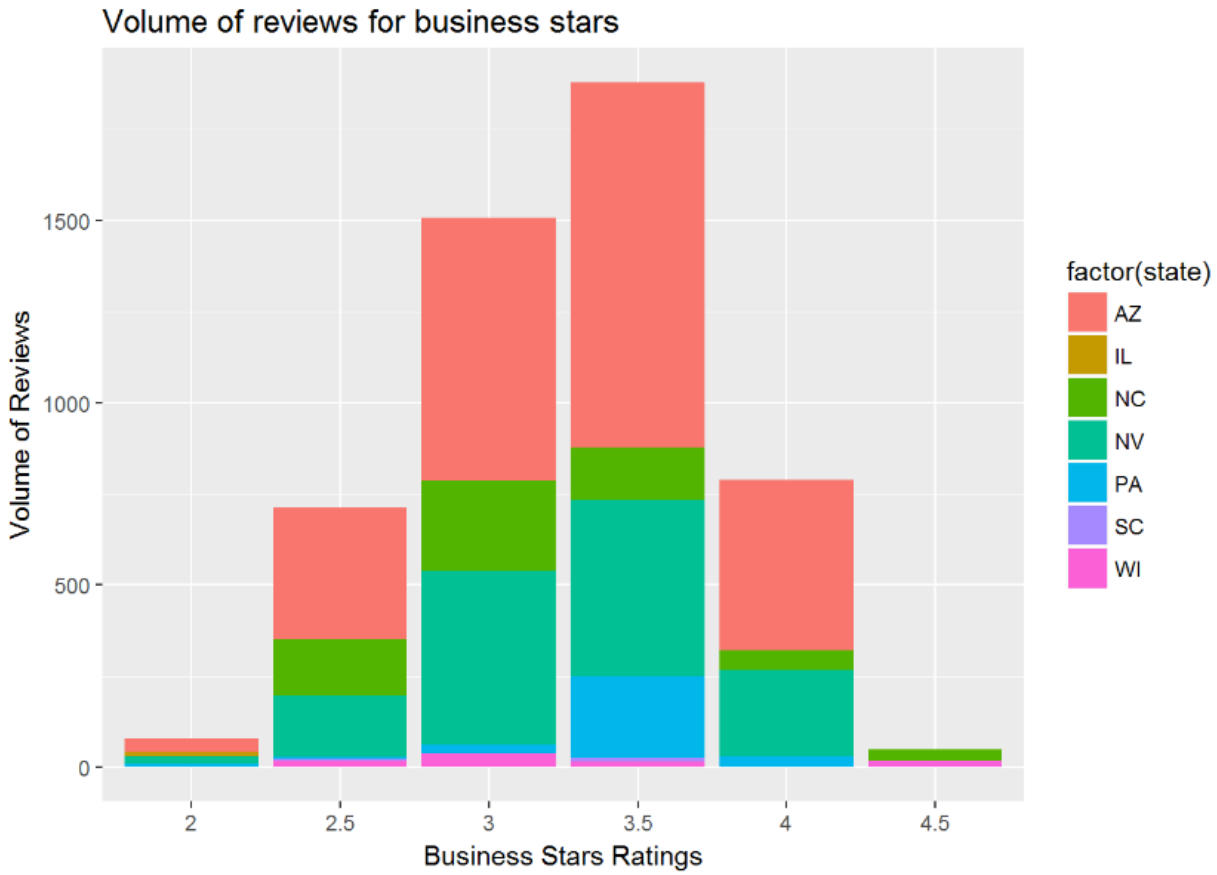
- For each Review start rating, we added the number of reviews falling in each of the stars for each city as predictor. This was done to get increase the modelling performance. Furthermore, added the number of reviews by elite user and non-elite user as predictor.
- To improve the SVM model prediction, we increased the penalization cost.

- We take the cities that have reviews over all the 4 NFL season, from 2012 to 2015.
- We performed text mining and taken into consideration only those reviews which are related to NFL games. For these we take reviews having NFL related words in their text. These reduced the number for records to 524. These data size was small and hence we did not apply our models on this text mined data. Since this will improve the model efficiency, this can be done in future by taking more review over a longer time period.

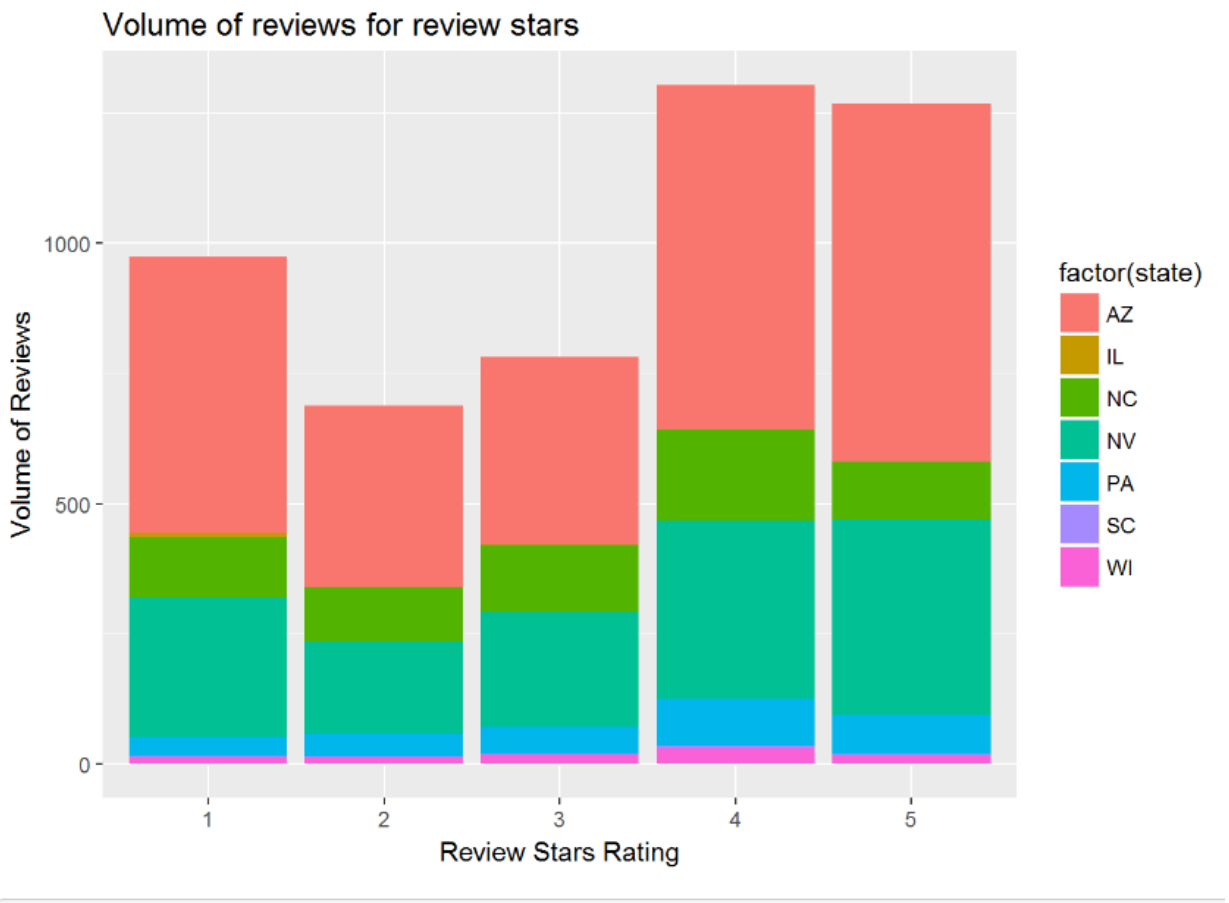
5) What visualization(s) have you included? Explain what is conveyed in the visualization and why.

For visualization, we have used bar graphs, bubble chart , word cloud and area graph.

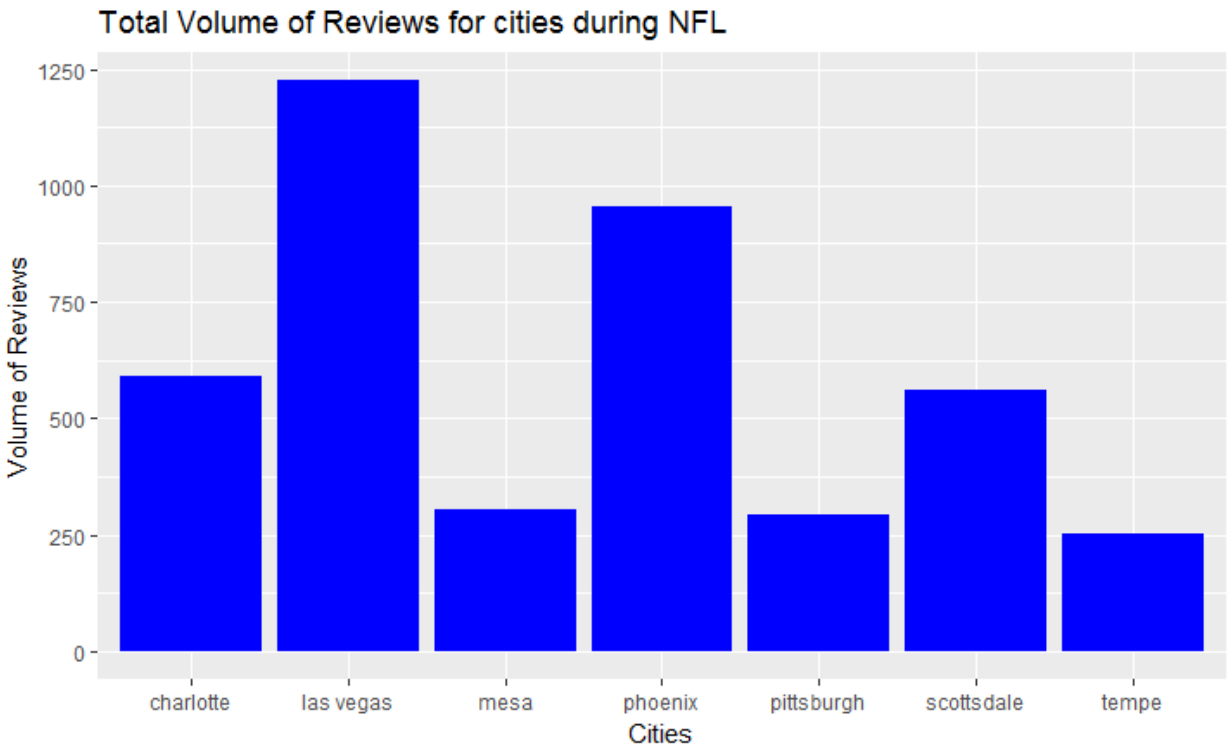
- We have used bar graphs for comparison of relative point values.
- Bubble chart is used to check the frequency of number of reviews over a week for each city.
- Area graph is used to check the change in volume of reviews per week.



The above bar graph depicts the comparison of volume of reviews with respect to business star ratings for various states. We can see that the maximum business star rating across state is 3.5 and state of Wisconsin has the least business stars rating.

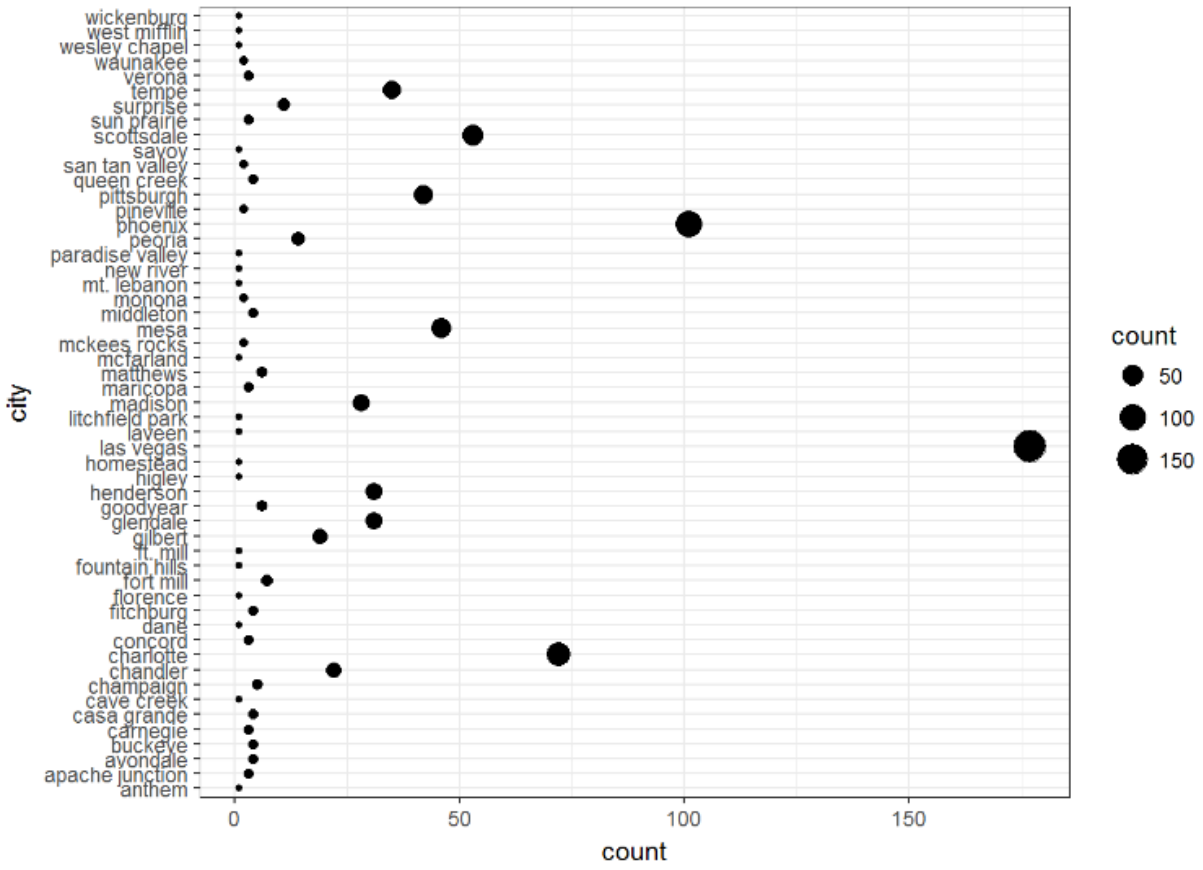


The above bar graph depicts the comparison of volume of reviews with respect to review star ratings for various states. We can see that the maximum review star rating across state is 4 and state of Wisconsin has the least business stars.

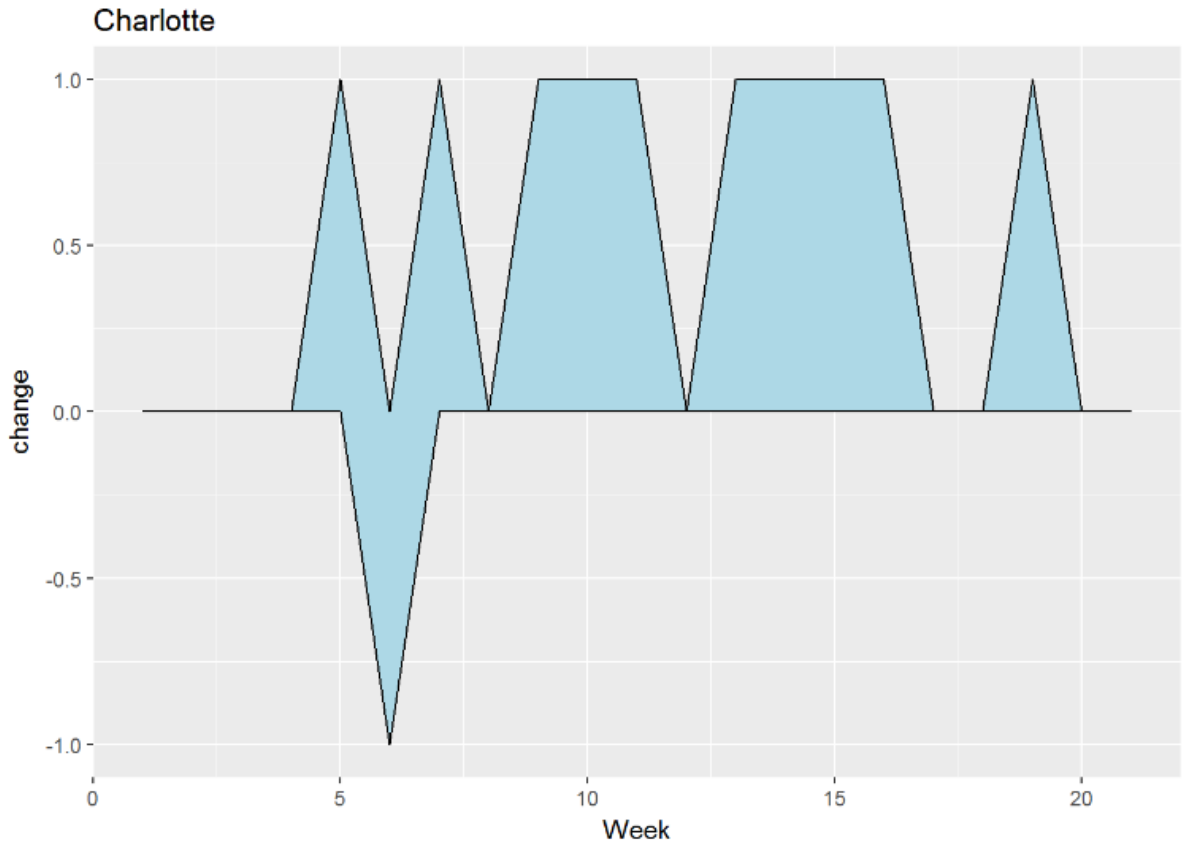


The above graph depicts the number of reviews for major cities in USA during NFL Season. Las vegas has the maximum number of reviews.

Word cloud for the reviews posted during NFL season.



The above chart depicts the frequency of number of reviews over a week for each city.

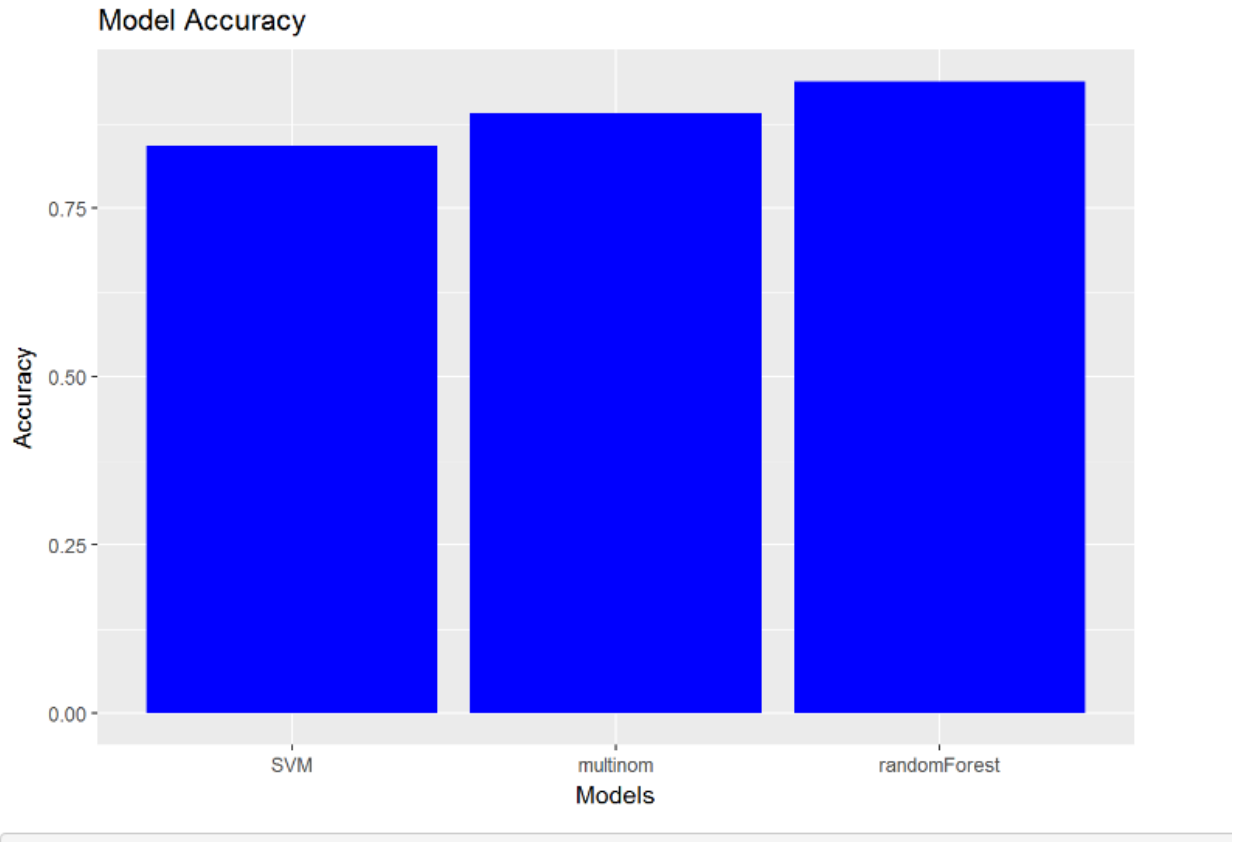


The above graph depicts the change in volume of reviews per week for Charlotte.

6) What evaluation method did you propose?

- We propose to use confusion matrix and ROC curve for evaluation. The accuracy of the models were calculated using the confusion matrix.
- We get the Sensitivity and Specificity from the confusion matrix. ROC curve was plotted to represent the relation between sensitivity and specificity. AUC is obtained from the ROC curve plot.

7) How did your model perform according to this evaluation?

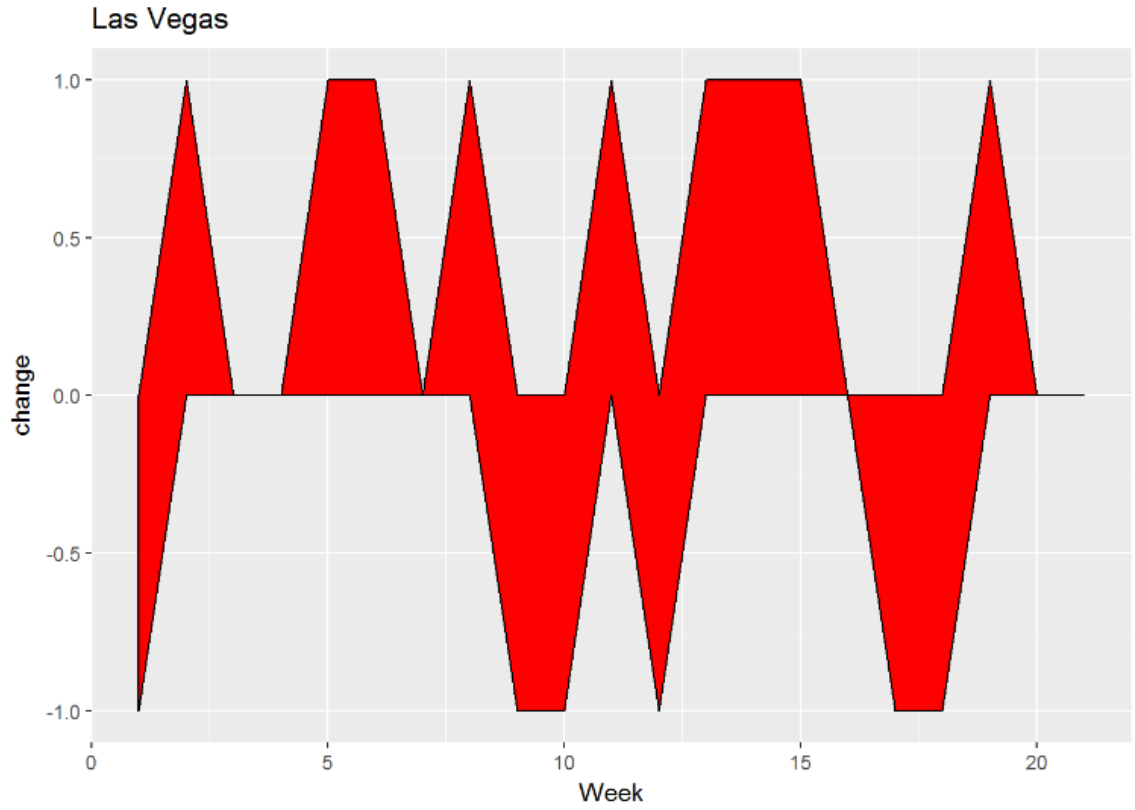


The above bar graph depicts the comparison of accuracy across the 3 models that we used. We see that Random forest has the highest accuracy.
We have used 3 models

1) SVM Multiclassification Model
   SVM had low accuracy among the other model. Its accuracy was improved by adding a penalization cost. Without the penalization cost the performance was even lower.
2) Multinomial Logit Model
   Multinomial Logit Model performed better than SVM. It performed better without any penalization cost. It is a probabilities model due to which there is overfitting of data. From the ROC curve we can see that it curve is close to excellent area.
3) Random Forest Model
   Random Forest performed better than the rest of the models. It uses decision trees which controls the over fitting. From the ROC curve, we see that it falls in between the good and the excellent region, hence Random forest has good performance.

8) Based on your results what conclusions do you draw?

- Our results show that the 7 selected cities can expect an rise of 42.85% in reviews for Sports bar during the NFL season.



The above graph shows that there was an increase in number of reviews for 9 out to 21 weeks, which is 42 % approx.

- From the text mining, we found frequent terms like food, order, service etc. This shows that people are interested in the food quality and experience at the Sports bar.
- We found that it is not feasible to differentiate NFL/NBA/NHL/MLB simply from mining review text, since we find that the most frequent words are loosely related to this sporting events.
- Also we find that to analyze the reviews for a specific Sports bar during the NFL season, there must be data available for most number of NFL seasons.

9) Based on your results what further studies would you do or are warranted?

- We focused on NFL schedule game days and its effect on the volume of review of sport bar. Our results show that there is an increase in volume of reviews for Sports Bar during NFL season. Further sentimental analysis can be done to find out how positive and negative do people talk about Sport Bar during a NFL game.
- Since NFL, NBA, NHL and MLB schedule are overlapped. We can further improve the model by doing monthly analysis, taking specific months that have only one of these games and compare across different seasons. This overlapping schedule is depicted below, month wise.

Rank	1	2	3	4
Month	NFL	MLB	NBA	NHL
Jan	NFL	NA	NBA	NHL
Feb	NA	NA	NBA	NHL
March	NA	NA	NBA	NHL
April	NA	MLB	NBA	NHL
May	NA	MLB	NBA	NHL
June	NA	MLB	NBA	NHL
July	NA	MLB	NA	NA
Aug	NA	MLB	NA	NA
Sept	NFL	MLB	NA	NA
Oct	NFL	MLB	NBA	NHL
Nov	NFL	NA	NBA	NHL
Dec	NFL	NA	NBA	NHL

Thank You Note
We would like to thank Qi Wang and Professor Rumi Chunara for their valuable inputs.

Reference:-

Lecture Slides.

https://www.yelp.com/dataset_challenge

http://static.pfref.com/years/2015/games.htm#games::none

http://www.businessinsider.com/most-watched-sporting-events-of-2015-2016-1

https://www.brightlocal.com/2015/08/20/92-of-consumers-now-read-online-reviews-for-local-businesses/

https://www.tutorialspoint.com/r/r_decision_tree.htm

https://www.r-bloggers.com/predicting-wine-quality-using-random-forests/

http://www.fftoday.com/nfl/schedule.php

http://www.espn.com/nfl/schedulegrid

http://www.sportsmediawatch.com/2016/07/halftime-most-watched-sporting-events-year-so-far-nflnba/

http://variety.com/2015/tv/news/nfl-record-ratings-for-opening-week-1201595991/

https://www.mathworks.com/help/stats/classificationlinear.predict.html

https://biz.yelp.com/support/responding_to_reviews

https://www.fundera.com/blog/2015/05/28/yelp-reviews-does-anybody-really-care

http://plei-plei.info/wp-content/uploads/2012/03/tv-watching-at-sports-bars-as-social-i
nteraction.pdf

http://www.hbs.edu/faculty/Publication%20Files/12-016_a7e4a5a2-03f9-490d-b093-
8f951238dba2.pdf

http://hbswk.hbs.edu/item/the-yelp-factor-are-consumer-reviews-good-for-business

http://localvox.com/blog/how-to-avoid-the-yelp-review-filter-and-get-more-positive-reviews/

http://stories.journalism.ku.edu/j415-sp15/2015/03/11/impact-yelp-reviews-have-on-local-businesses-
ishard-to-gauge-owners-managers-say/

http://www.kdnuggets.com/2015/05/3-things-about-data-science.html

https://www.yelp.com/search?find_desc=Sports+Bars&find_loc=New+York,+NY

http://www.mobilelifecentre.org/sites/default/files/gz_mobileHCI_final.pdf

http://www.si.com/extra-mustard/2015/12/03/fifa-scandal-arrests-hotel-baur-au-lac-yelp-review

https://en.wikipedia.org/wiki/Sports_in_the_United_States

https://en.wikipedia.org/wiki/2015_Copa_Am%C3%A9rica

http://www.skillsyouneed.com/num/percent-change.html

https://rstudio-pubs-static.s3.amazonaws.com/127992_a060e7d374d549998df02fc11ac8c334.html

https://www.cct.lsu.edu/~pkondi1/bare_jrnl

http://www.cs.ucsb.edu/~korpeoglu/cs290d/yelpbusy.pdf

http://www.galvanize.com/blog/bayesian-statistics-analyzing-yelp

http://www.topendsports.com/events/calendar-2016.htm

http://www.topendsports.com/events/

https://en.wikipedia.org/wiki/List_of_multi-sport_events

http://www.espn.com/nfl/schedule

http://www.sbnation.com/nfl/2016/4/14/11435628/2016-nfl-schedule-released-dates-times-regular-season

https://www.kaggle.com/maxhorowitz/nflplaybyplay2015

http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know#explore-frequent-terms-and-their-associations

https://www.quora.com/How-does-randomization-in-a-random-forest-work

http://www.statmethods.net/advgraphs/ggplot2.html

http://docs.ggplot2.org/dev/vignettes/qplot.html

https://sites.google.com/site/econometricsacademy/econometrics-models/multinomial-probit-and-logit-models
https://www.quora.com/How-does-randomization-in-a-random-forest-work