Compare Perspectives: Yelp Restaurants & Users

Ji Won Chung, Wenqin Chen, Khadidja Fares

Introduction

The potential target audience are Yelp and all restaurant owners interested in increasing user activity. We analyzed the business dataset and the user dataset with its corresponding reviews.

Our goal is to explore the restaurants business dataset and discover which features define a "successful" restaurant business from both the restaurants' perspectives and users' perspectives. We focused on restaurant information from the business dataset, because we believe that Yelp reviews are generally written for restaurants. We hypothesized that the features that restaurants value may differ from those that users value.

Data & Methods

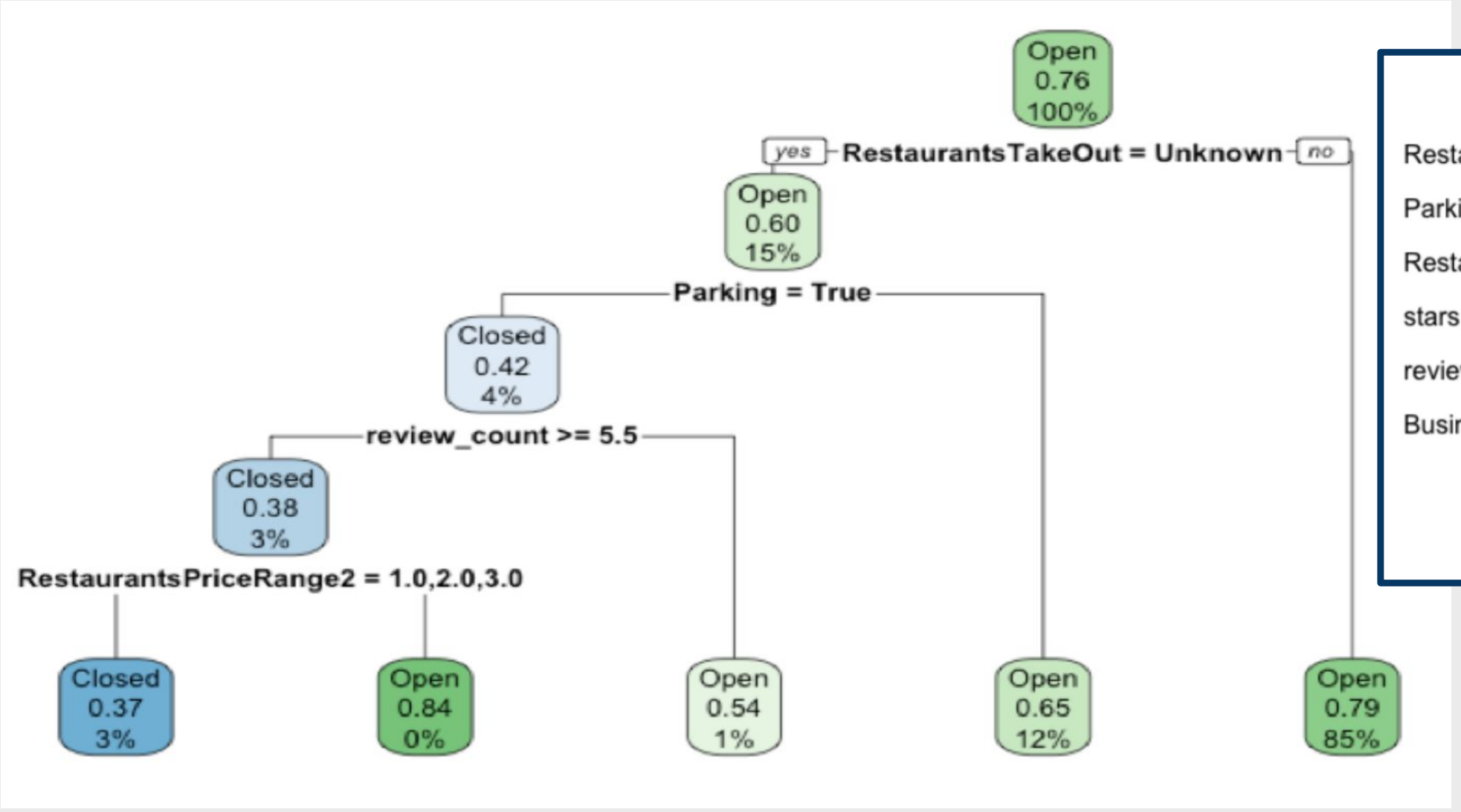
We used subsets of data from user.json, review.json, and business.json from the Yelp dataset challenge.

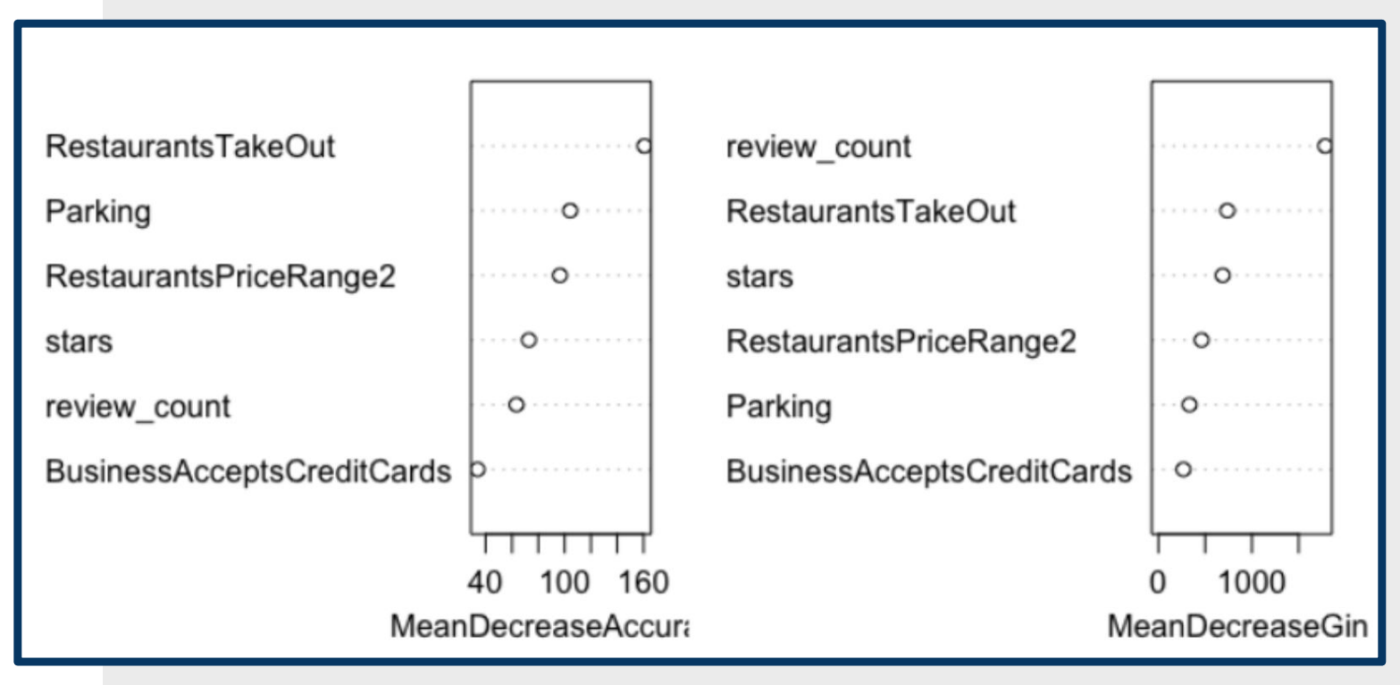
We explored user and review data through a Naive Bayes Classifier and a Support Vector Machine. We defined inactive users as those with 'review_count' equaling to 1 and active users as those with 'review_count' >= 90. By using Term Frequency times Inverse Document Frequency (tf-idf), we tokenized texts, counted word frequency, and downscaled less informative words such as 'is', 'a', 'are' for all active users and inactive users with at least 80% of reviews available in review.json. We then ran the models on the tf-idf matrix to predict an active user with the user review.

We explored the business data by creating a Decision Tree and Random Forest for each response variable: 'is_open', 'stars', and 'review_count'. Our assumption was that a successful restaurant is most likely still open, has good star ratings, and a lot of review counts. We used the following predictors, but excluded the response variable: 'BusinessAcceptsCreditCards', 'RestaurantsPriceRange2', 'RestaurantsTakeOut', 'Parking', 'is_open', 'stars', and 'review_count'. If a decision tree splits on a feature and the top ranked feature on a variable importance diagram overlaps, we determined that those factors could potentially define what makes a restaurant business successful.

Results & Model Visualizations

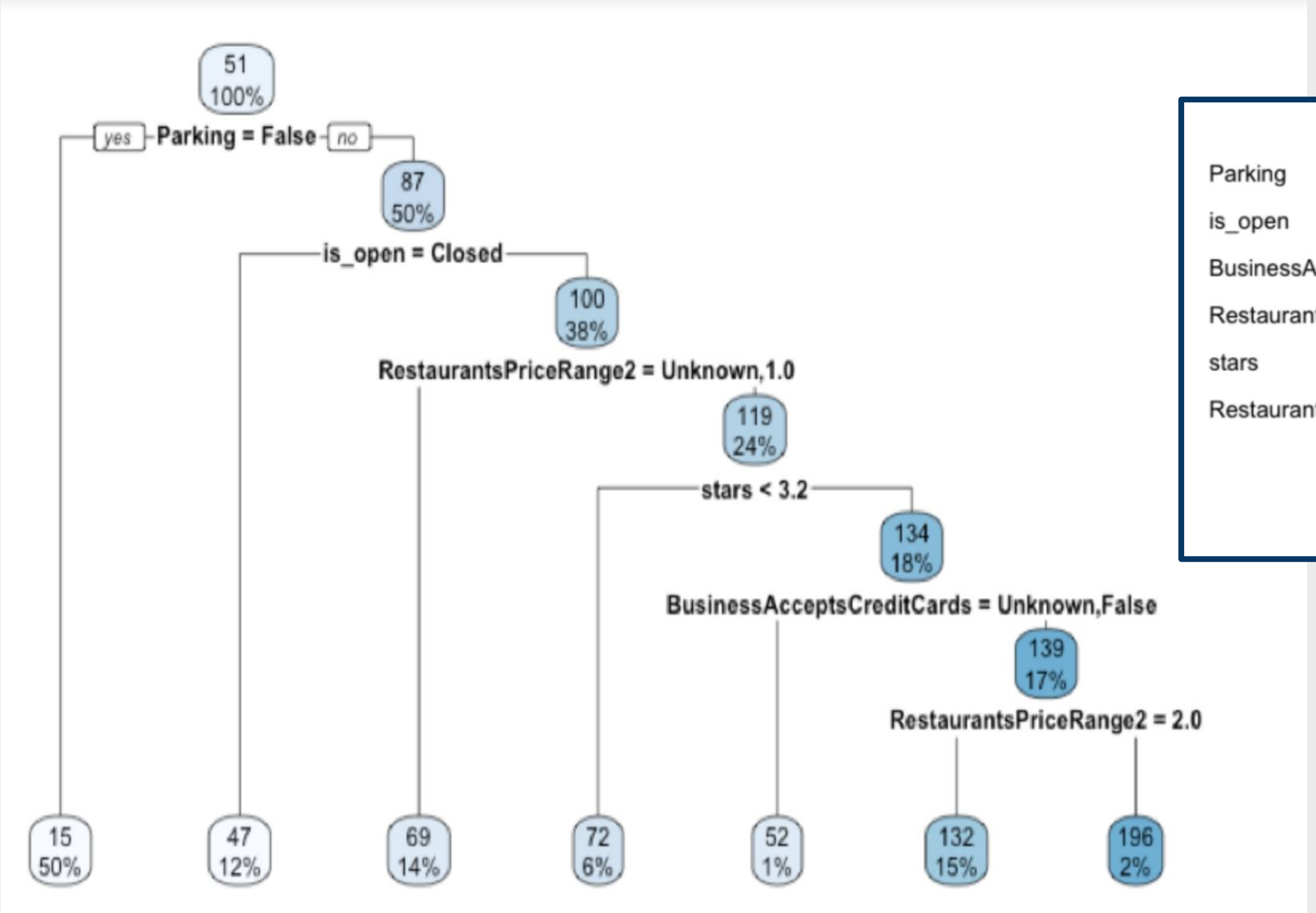
Classification Decision Tree and Random Forest Variable Importance: Predicting Open and Closed Restaurants

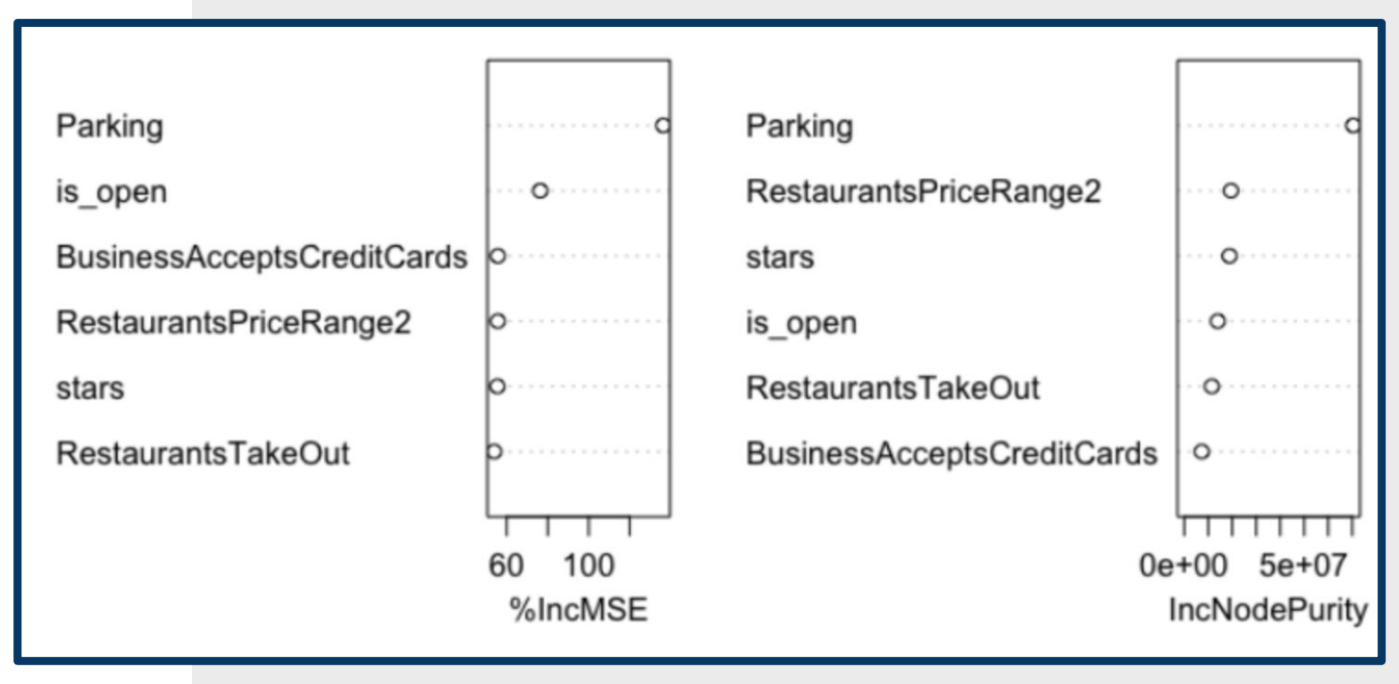




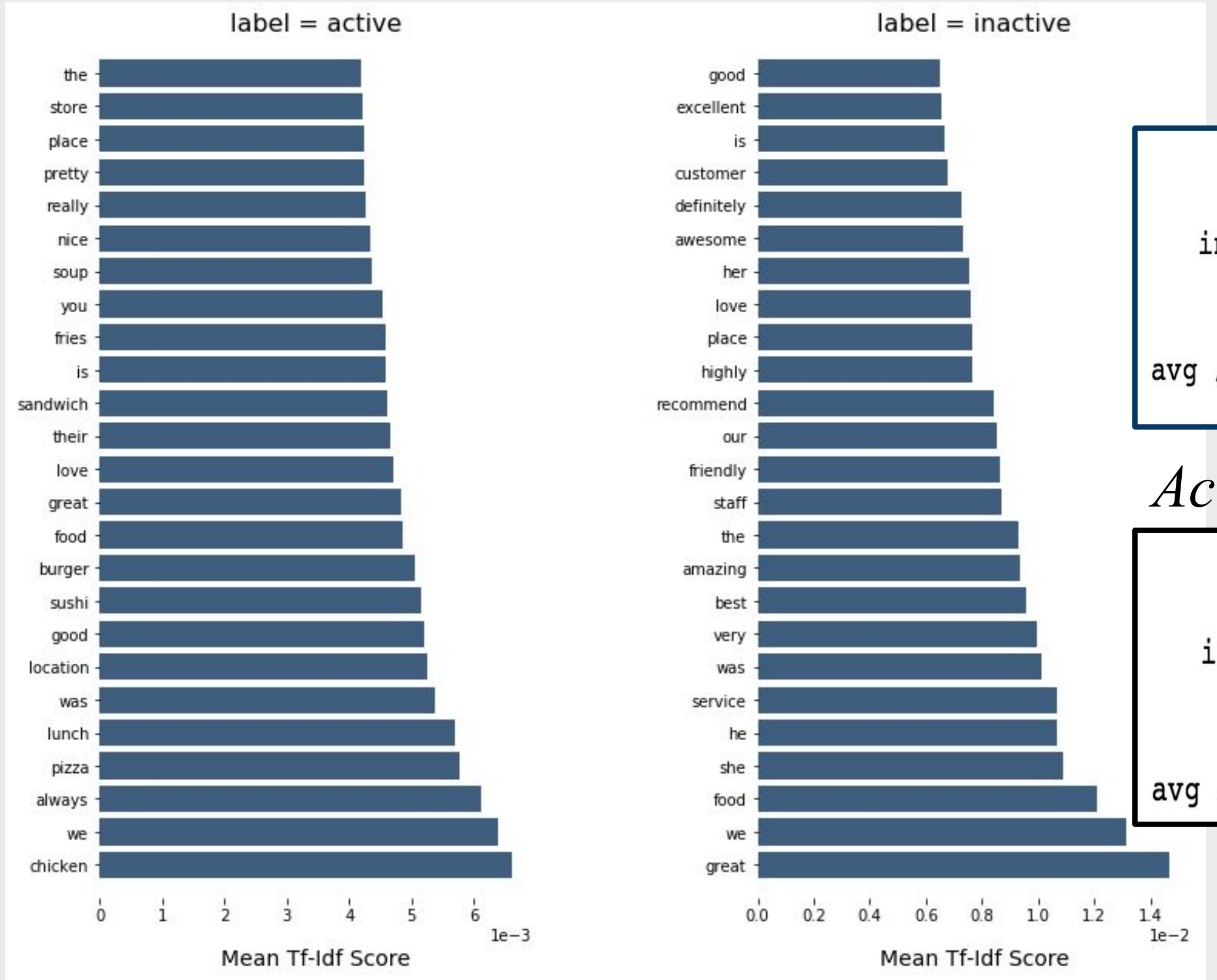
Classification Decision Tree Error Rate: 23.9%

Regression Decision Tree and Random Forest Variable Importance: Predicting Number of Review Counts

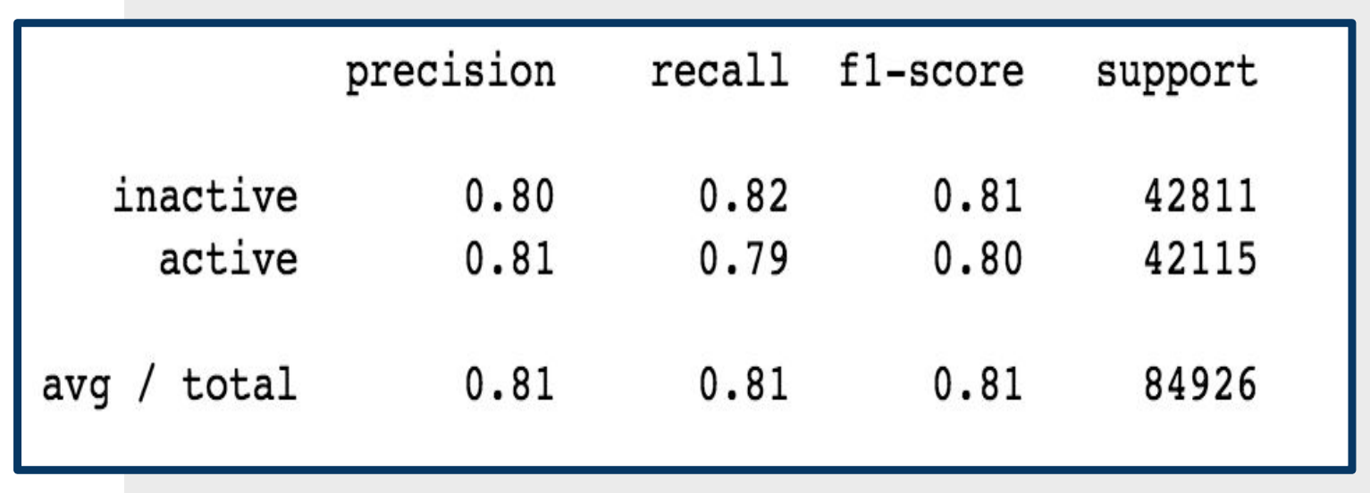




Performance of SVM and Naive Bayes Classifier with tf-idf: Predicting an Active User



Accuracy of SVM Model

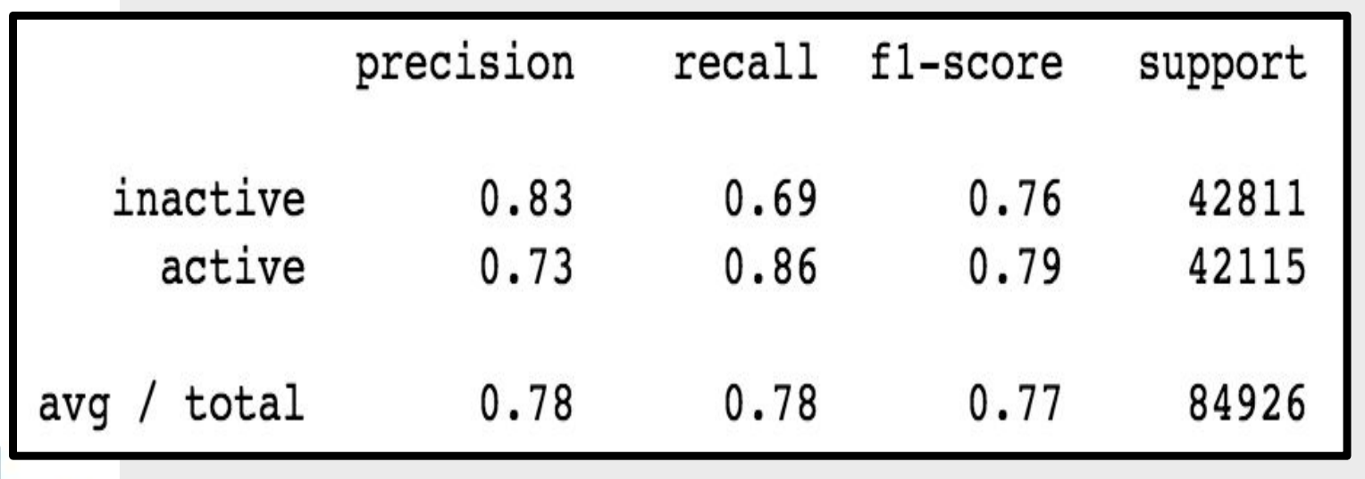


```
             precision    recall  f1-score   support

   inactive       0.80      0.82      0.81     42811
     active       0.81      0.79      0.80     42115

avg / total       0.81      0.81      0.81     84926
```

Accuracy of Naive Bayes Classifier



```
             precision    recall  f1-score   support

   inactive       0.83      0.69      0.76     42811
     active       0.73      0.86      0.79     42115

avg / total       0.78      0.78      0.77     84926
```

Conclusions & Future Work

In conclusion, SVM performs very well on text classification. Active Yelp users tend to give more specific feedback on food, and inactive users' reviews are more generic and are probably more related to business service and vibe. The decision trees and random forests predicting the number of a restaurant's review counts and if it is open demonstrates that information on parking and takeout are significant features that potentially measure the success of the restaurant.

For future work, we will sync active users' reviews and inactive users' reviews with the corresponding restaurants and identify significant features among the businesses. We could also add more columns to the restaurant dataset, instead of working with 6 predictors, to evaluate more thoroughly what denotes successful restaurants. If we had more time, we would explore the 'stars' feature of restaurants more and understand the regression tree better.

References

"Yelp Dataset Challenge." *Yelp Dataset*, 1 Sept. 2017, www.yelp.com/dataset/challenge.
"Working With Text Data." *Working With Text Data — Scikit-Learn 0.19.1 Documentation*,scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html.
Buhrmann, Thomas. "Analyzing Tf-Idf Results in Scikit-Learn." *Analyzing Tf-Idf Results*, 22 June 2015, buhrmann.github.io/tfidf-analysis.html.

Acknowledgements

This project was completed in partial fulfillment of the requirements of SDS293: Machine Learning. This course is offered by the Statistical and Data Sciences Program at Smith College, and was taught by R. Jordan Crouser during Fall 2017.