Yelp Data Exploration

Pradip Hayu, M.S.

Gaberial Campese, M.S.

Dr. Abdi Awl, Ph.D.

The George Washington University

Abstract

Entrepreneurs have historically been challenged with the assumption of risk especially when embarking on a new venture as this is what makes them an entrepreneur. Opening any type of business, more specifically a restaurant, requires careful consideration of all resources in order to maximize initial investment and reduce risk. With the research to be conducted in this report, entrepreneurs will have the ability to better offset their risk when opening a new restaurant. Python and several libraries will be utilized for the research into Yelp restaurant reviews. The research will involve the application of data science and machine learning through the use of several classification models and text analysis methodologies executed on the massive Yelp dataset containing over 6,600,000 reviews. The Yelp reviews are of scope in this research because they are created by customers and detail exactly what customers consider a good or bad restaurant. Among these millions of reviews, key success indicators will be revealed in regards to what makes a customer pleased when dining at a restaurant.

Keywords

Yelp dataset, classification, text analysis, restaurants, data science, machine learning, entrepreneur

Introduction

Yelp is a social networking site and app, centered on customer reviews of local businesses. Anyone can post a review on Yelp, and anyone searching the web can use Yelp to find and evaluate local businesses. With more and more people being introduced to this app, it also provides Yelp a huge amount of market information, which could be used for local businesses to improve their performance based on ratings and reviews and local governments to identify the most thriving business type in town. As Cui (2015) mentions, “Analyzing the real

world data from Yelp is valuable in acquiring the interests of users, which helps to improve the design of the next generation system". It is highly likely that this is one of the reasons that Yelp created a big data challenge for the public to analyze this enormous amount of data from different perspectives. This research utilizes numerous classification models and texting mining methodologies to provide insight on reducing risk and maximizing initial investments for restaurant entrepreneurs.

Literature Review

The majority of Yelp Dataset challenges and explorations work specifically with Natural Language Processing (NLP) in their studies, which is why this research focuses mainly on classification modeling, while dabbling in NLP with the text mining procedures. Many projects use Markov Chain to finish reviews based on inputted text and text classification to return predicted review category based on inputted text. Another popular technique involves the use of sentiment analysis to focus specifically on the context of text embedded in reviews by customers. A study using sentiment analysis discovered that "unexpectedly, the flavor of dishes does not rank first among all positive reviews. Instead, the service of restaurants seems to be the priority for most customers, since the word friendly ranks first and the word attentive is also in top 5" (Yu et al. 2017). This is interesting because the customer appears to value service over food quality.

Another interesting application of data science involving the Yelp Dataset involves the use of the Neo4j, known as py2neo, to conduct Cypher queries ultimately incorporating clustering on the dataset (Cui, 2015). In simpler terms, this study uses Neo4j to create relationships among users to detail how users themselves affect Yelp and how the reviews are aggregated. The research initially finds: "From the high level, we found that there are 2576179

knows directed friend relationships among 366715 users. So, on average, each user has roughly 7 friends (Cui, 2015). The study then uses this to move deeper into their investigation because this has to do with the "fans" feature on Yelp, fans referring to the reaction to a user's review by fellow users: "From the results, we know that only a small portion of users actively comment a lot of businesses, which makes them followed by others. Most of users only care about the businesses that they have ever been experienced, which are limited" (Cui, 2015). This is a different, yet interesting approach to the Yelp data exploration in that it analyzes the actual reviewing activities of the users on Yelp.

Research Methodology

It is evident, that there is notable value in analyzing restaurant data in the industry because of the insights offered in regards to problems faced by restaurants and entrepreneurs. This problem, repeatedly mentioned throughout the introduction and abstract, is recognized early on and it is where the research commences. The machine learning, including several classification models, and text mining analysis offers a deep dive into the complex Yelp dataset in efforts to unearth trends to ultimately help determine what makes a restaurant successful. These models will help predict star ratings, the core performance indicator of a Yelp review, to identify trends among the restaurants that hold these good reviews. The text mining will reveal key words categorized as either good or bad based on sentiment analysis. Finally, this research will help entrepreneurs make better decisions to not only help deter their assumption of risk, but also understand common patterns that make a restaurant both good and bad in the eyes of customers, which could also help later on down the road when they have an active restaurant.

Data

The Yelp dataset was obtained directly from Yelp in both json and csv format composed of over 6,600,000 reviews. The dataset is large in nature containing 6 separate datasets named Business, Review, Check-in, Tip, User, and Photo. For the purposes of this research, the Check-in, Tip, User, and Photo datasets will be disregarded leaving the Business and Review datasets remaining. The Business dataset is in both json and csv format, while the Review dataset is in csv format. These datasets were then read into Jupyter Notebook to perform data preprocessing, exploratory data analysis, modeling, visualizations, and further analysis. While analyzing the reviews dataset, the restaurants had to be merged with the restaurant Business data so we would only focus on the restaurants reviews in accordance with the purpose of this research.

Data Analysis

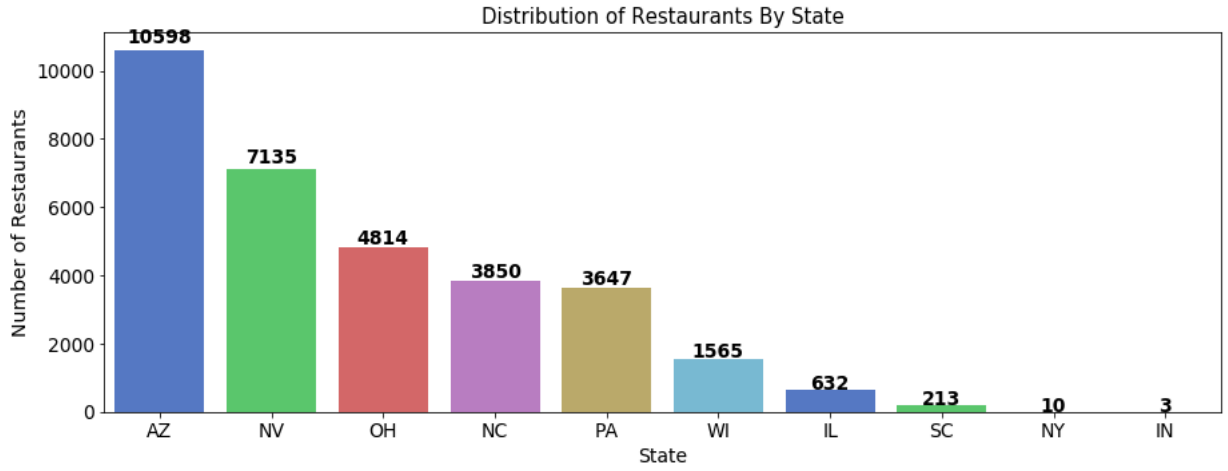
The data analysis is executed primarily using Python libraries in Jupyter Notebook. The Business datasets, in json and csv format, and Review dataset are used in the analysis. Specific libraries utilized include pandas, numpy, scikit-learn, seaborn, matplotlib, pydotplus, Natural Language Toolkit (nltk). Scikit-learn is mainly used for the machine learning classification modeling and Nltk is utilized for the text mining analysis. There are five classification models used for the purposes of predicting Yelp review star ratings including Random Forest, Decision Tree, KNN, Logistic Regression, and Gaussian Naïve Bayes. The Business dataset in json format is used to create attributes for the classification modeling. The overall business reviews are in Yelp-provided increments of 0.5, while the individual user reviews are in increments of 1. The overall reviews are benchmarked at 4 and above being classified as positive (1), with 3.5 and below being negative (0). The individual reviews are classified similarly except 4 or more is

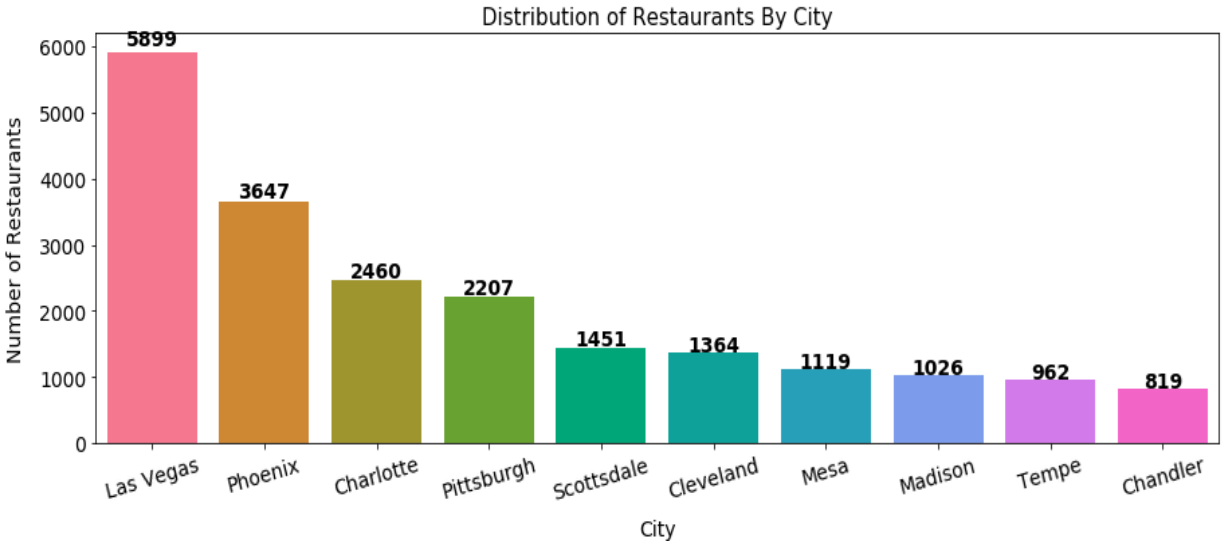
considered positive and 3 and below is negative. Lastly, hyperparameter tuning is used to optimize the parameters for each classification model, resulting in improved accuracy.

The Review and Business datasets in csv format are used for the text mining efforts. The reason why the Business dataset in csv is required for text mining is because it is merged with the Reviews dataset to correctly associate restaurants from the Business dataset with the reviews from the Reviews dataset. Arbitrarily, 7 popular cuisines are selected among the restaurant data for text mining analysis including French, Chinese Italian, Japanese, Korean, Thai, and Vietnamese. The Nltk library helps to calculate a polarity score to identify positive and negative words embedded in individual restaurant reviews. The polarity score is a sentiment analysis, which determines where the word's status is based on its context in the review.

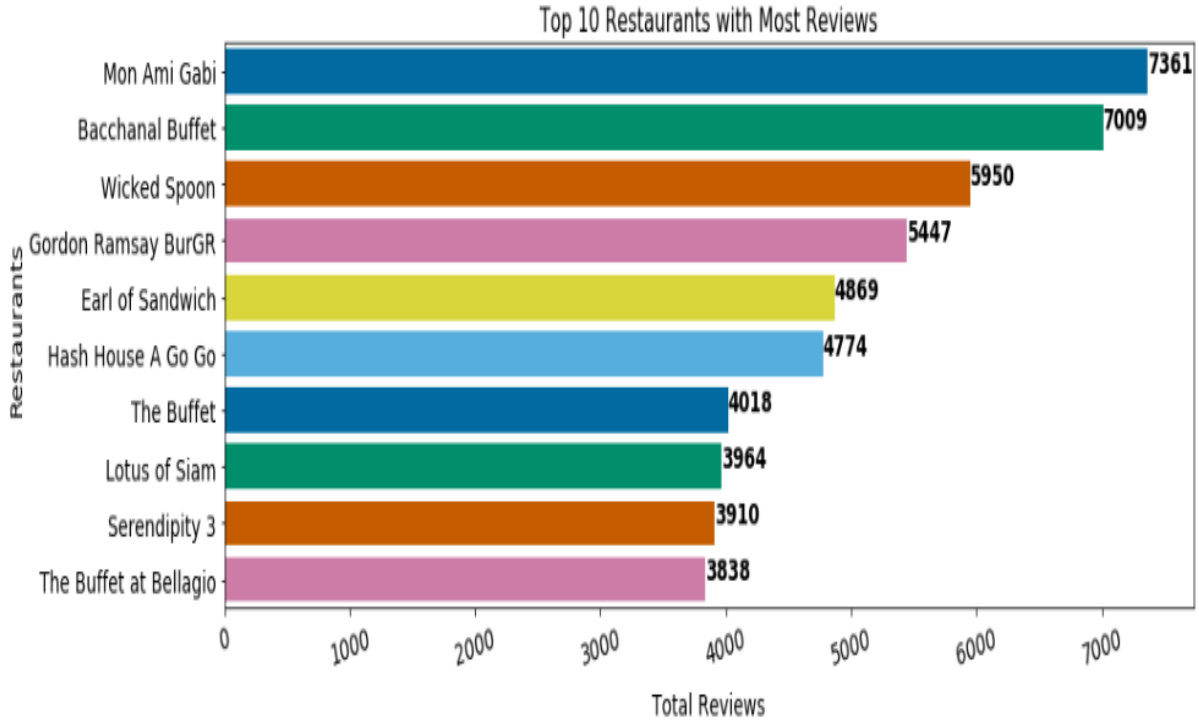
Key Findings

In providing key insights for the Yelp dataset, the following is conducted: Exploratory Data Analysis (EDA), construction of various models, hyperparameter tuning, model comparison, model selection, fitting the selected models, and finally text mining the reviews. For EDA, the datasets were preprocessed and narrowed down to strictly analyze United States restaurants.



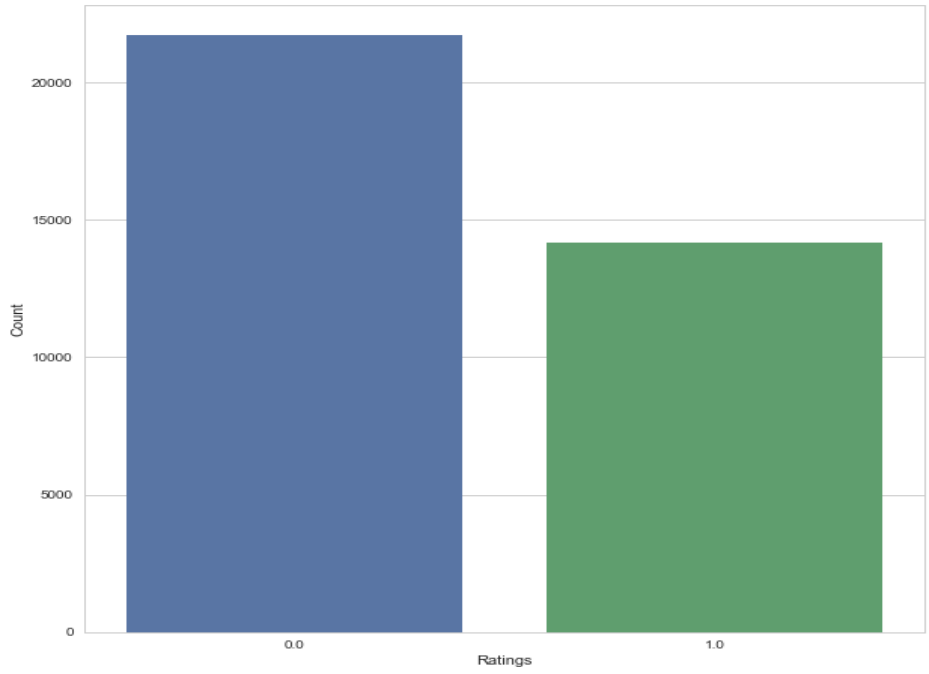


It is discovered that the majority of restaurants in the Yelp dataset reside in Arizona. However, when looking to the lense by city, it is seen that most restaurants are in Las Vegas by a considerable margin. The restaurants with the most reviews are then extracted and visualized.



It is recognized that Mon Ami Gabi and Bacchanal Buffet are the leading restaurants in terms of number of reviews on Yelp. Classification model methodologies for predicting the Yelp star rating are then conducted.

For classification models, there must be features (inputs) and targets (star rating). For this research, the inputs are numerous attributes of restaurants: BikeParking, BusinessAcceptsCreditCards, Cater, GoodForKids, HasTV, OutdoorSeating, RestaurantsDelivery, and more. Each of these attributes are one hot encoded to 1 for yes, and 0 for no based on whether or not the restaurant provides that feature. The star rating is then predicted as either good (1) or bad (0) based on these inputted attributes. The following are visualizations to detail the Yelp data in the training and test sets for the models to be crafted.



Most of the reviews are found to be negative (0) in the training and test sets. Finally, the models are trained on the training set, and evaluated based on their accuracy in classification on the test set.

	Model	Score
3	Logistic Regression	61.54
2	KNN	61.15
0	Random Forest	60.11
1	Decision Tree	59.00
4	Naive Bayes	43.30

Initially, Logistic Regression is found to be the best model in terms of accurately classifying the restaurants as positive or negative based on the attributes of the restaurants at 61.54%. This is a slim margin from KNN at a 0.39% difference.

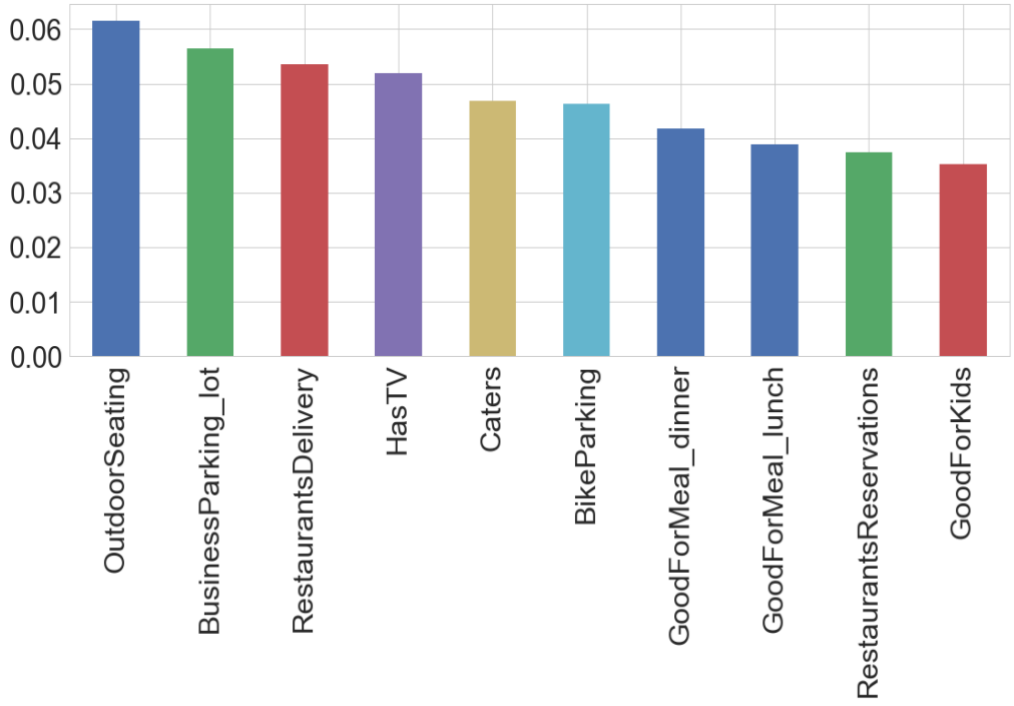
The accuracy percentages are not the best, so hyperparameter tuning is utilized to provide better results for accuracy scores.

Model	Hyperparameters Used	Time Taken For Execution(sec)	Accuracy(%)
Random Forest(rf)	n_estimators, min_samples_split and min_samples_leaf.	84.54	68.03
Decision Tree(dt)	min_samples_split and min_samples_leaf.	14.86	66.29
KNN(knn)	n_neighbors	5166.92	64.27
Logistic Regression(lr)	multi_class, solver and C	550.60	61.83
Gaussian Naive Bayes(gnb)	var_smoothing	2.46	53.21

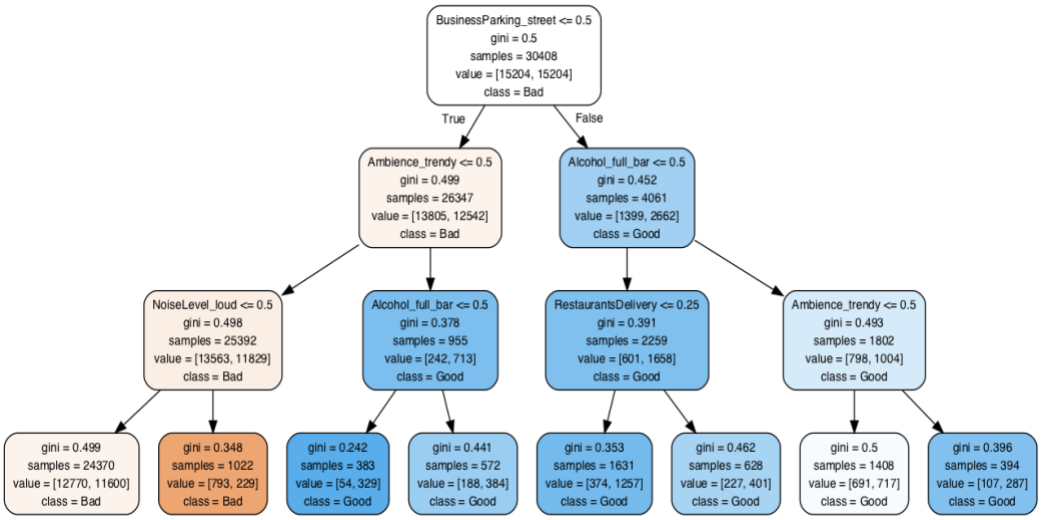
In order to perform hyperparameter tuning on each model, a parameter grid is created for each model to identify the parameters needed to be tuned. This is seen under the Hyperparameters Used column above. The execution time and accuracy are also detailed above. After hyperparameter tuning, Logistic Regression is no longer the most accurate model with Random Forest taking its place at over 68%. The Random Forest model saw an impressive boost of about

8% in accuracy after tuning. Decision Tree follows closely at 66.29%, leading to Random Forest and Decision Tree to be selected in model selection.

The following visualizations are important products of model selection, the first being from Random Forest Classification.

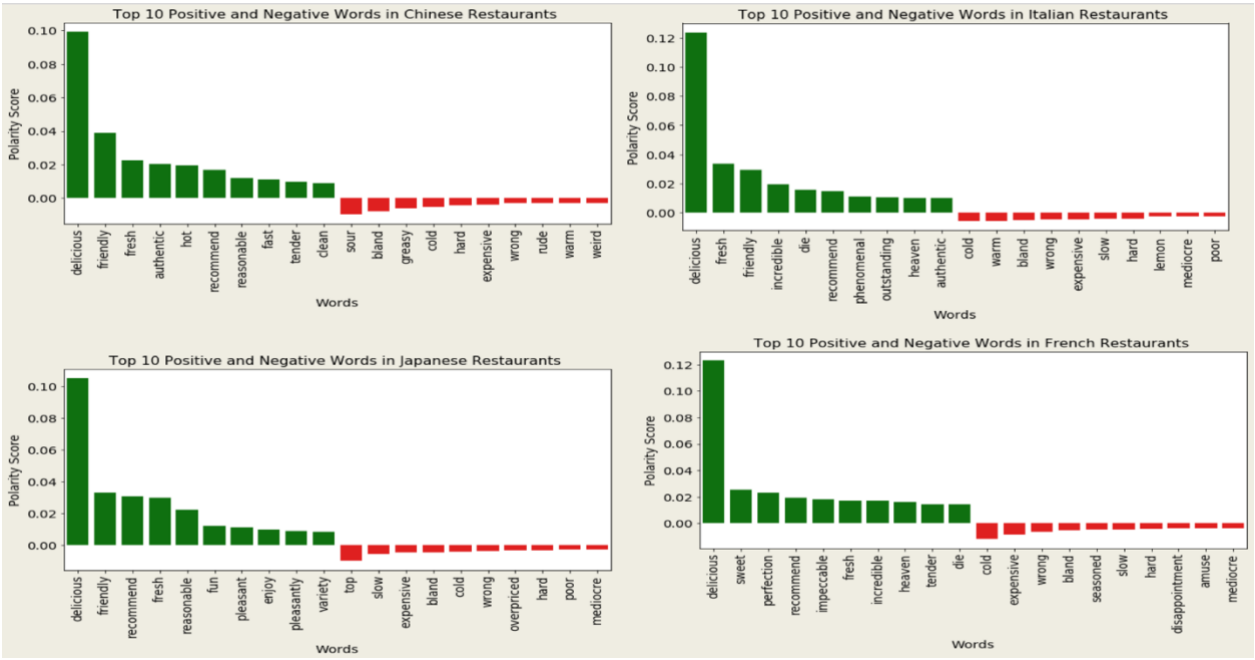


The graphic above is feature importance in the Random Forest Model. These features are ordered by their importance in classifying the Yelp restaurant star ratings as either good or bad. It is seen that specifically OutdoorSeating, BusinessParking_lot, and RestaurantsDelivery lead the way in feature importance when classifying the reviews with Random Forest. Entrepreneurs should pay close attention to these features.



This visualization is the Decision Tree Classifier in action. This decision tree is easily navigable by determining whether a certain attribute exists at the restaurant eventually resulting in whether is placed in the class of either Good or Bad Review.

Moving on, the final methodology used in the Yelp Data Exploration research revolves around text mining reviews among 7 different cuisines. Text mining is used to identify key words utilizing sentiment analysis (polarity score) to determine good and bad words embedded in reviews. The features are a “bag of words”, while the targets are the polarity scores.



The graphics depicted above lay out the top 10 positive and negative reviews accompanied by their polarity scores in each of the 7 selected cuisines. Notice that each cuisine's top positive word existing in the individual reviews is "delicious". Negative words including "cold" and "slow" are commonly seen in negative reviews. The following graphics provide further insight of the positive and negative words.

cuisines										
Chinese	delicious	friendly	fresh	authentic	hot	recommend	reasonable	fast	tender	clean
French	delicious	sweet	perfection	recommend	impeccable	fresh	incredible	heaven	tender	die
Italian	delicious	fresh	friendly	incredible	die	recommend	phenomenal	outstanding	heaven	authentic
Japanese	delicious	friendly	recommend	fresh	reasonable	fun	pleasant	enjoy	pleasantly	variety
Korean	delicious	friendly	reasonable	attentive	tender	recommend	generous	fresh	cute	variety
Thai	delicious	friendly	fresh	recommend	reasonable	perfection	fast	hot	bomb	outstanding
Vietnamese	delicious	fresh	friendly	clean	fast	attentive	reasonable	enjoy	recommend	helpful

This table provides the top 10 positive words by cuisine.

cuisines										
Chinese	sour	bland	greasy	cold	hard	expensive	wrong	rude	warm	weird
French	cold	expensive	wrong	bland	seasoned	slow	hard	disappointment	amuse	mediocre
Italian	cold	warm	bland	wrong	expensive	slow	hard	lemon	mediocre	poor
Japanese	top	slow	expensive	bland	cold	wrong	overpriced	hard	poor	mediocre
Korean	bland	work	slow	warm	cold	expensive	mediocre	sweet	fried	lack
Thai	sweet	bland	slow	cold	wrong	sour	expensive	mediocre	greasy	weird
Vietnamese	bland	wrong	slow	cold	greasy	warm	rude	expensive	dirty	sad

This table provides the top 10 negative words for each cuisine.

Limitations

Throughout the course of this research, several limitations are identified that may help in future research. Moreover, these limitations may serve as insight in regards to the shortcomings of the Yelp Data Exploration. The main limitations pertain to the overall nature of the Yelp data itself due to its complexity. The Yelp Dataset is composed of 6 datasets encompassing over 6.6 million reviews in a less than favorable format with missing data and NaNs. Additionally, this data exploration is limited to restaurants and the dataset offers various routes to go such as investigating movie theaters and social media posts. Other limitations are identified among the methodologies. For one, the scope of the research is limited to 5 classification models and more models could be assessed in the future. More focus on hyperparameter tuning could also help in this situation for both the existing models and models to come. For the text mining analysis, more advanced NLP techniques could also be applied such as BERT, yet the main scope of this research centers on classification modeling.

Recommendations & Future Research

It is apparent, that based on the Key Findings revealed in the Yelp restaurant data research, several recommendations can be made specifically to restaurant entrepreneurs. Based

on the classification models detailed previously, entrepreneurs must pay close attention to various restaurant attributes based on feature importance. Some key attributes that stood out are whether the restaurant has parking, delivery, TV, outdoor seating, and kid friendly. Other recommendations that can be made pertain to the modeling procedure itself. Random Forest and Decision Tree models appear to be the most successful classification models for the Yelp dataset. Hyperparameter tuning is also recommended based on the improvements in performance observed. The accuracy scores were not the greatest, so perhaps future research should focus more on effective data preprocessing and augmentation since the Yelp dataset is extremely complex.

As for the text mining research, several significant recommendations can additionally be set forth. Across the board, entrepreneurs must be keen on food temperature, flavor, price, and service because these are negative indicators among every cuisine. Moreover, specific cuisines need to be wary of their own, unique challenges. For instance, Italian cuisine entrepreneurs must pay close attention to lemon in their food, as this was commonly determined to have negative sentiment in customer reviews. Overall, the classification and text mining research described in this report are able to provide impressive recommendations.

Conclusion

All in all, entrepreneurs do not appear to be slowing down on opening new ventures including restaurants. Yelp remains the gold standard when it comes to reviews and putting the voices of customers in spotlights. These two facts further drive the high value of data in today's world. With the methodologies applied in this research, entrepreneurs are able to make better decisions to ultimately minimize risk because they are better equipped in their understanding of what customer's value when dining at restaurants. The accuracy scores could absolutely be

better, yet they are still good and the models provide an excellent framework to pursue future research. Furthermore, entrepreneurs are able to act on key recommendations and even apply the methodologies described throughout the course of the research.

Biographies

Gaberial Campese is a graduate student at George Washington University's Data Science program. Mr. Campese previously studied Information Systems at the University of Florida's Warrington College of Business. As of December, Gabe is finishing his Master's in Data Science at GWU, while working full time in the government technology consulting industry with nearly 2 years of experience.

Pradip Hayu is a graduate student in the Data Science Program at The George Washington University (GWU). Mr. Hayu received his undergraduate degree in Chemistry from Salem State University, Salem, MA and served in the United States Army for 5 years before joining the graduate program at GWU. Currently, he is working as a benefit coordinator at GWU's military and veteran office and a part-time academic tutor at Educational Connections, Inc.

Dr. Abdi Awl is a professor at the George Washington University's Data Science program. He has worked for over 16 years in the technology industry. Dr. Awl has been working as GWU's senior database administrator for the last 7 years and his courses include data warehousing and data science capstone.

References

Cui, Y. (2015). An Evaluation of Yelp Dataset. *Semantic Scholar*.

Yelp Dataset Challenge. (n.d.). Retrieved from https://www.yelp.com/dataset/challenge.

Yu, B., Zhou, J., Zhang, Y. and Cao, Y. (2017). Identifying Restaurant Features via Sentiment Analysis on Yelp Reviews. *Semantic Scholar*.