1. Motivation & Research question

1.1 Motivation

The COVID-19 pandemic has a profound effect on all countries over the world. Almost everything around us has been changed. Companies have allowed employees to work from home to have minimum contact. Schools have closed the campuses to ensure health and safety. People are living a massive shift in their lives and the global business is slowing down. Since the outbreak has begun, many long-established businesses have filed for bankruptcy protection. Small businesses are affected the most by the shutdown. According to Yelp data, permanent closures have reached 97,966, representing 60% of closed businesses that won't be reopening. Consumer behaviors have been changing rapidly, therefore, all businesses need to adopt a speed-up transformation strategy.

In this paper, we scale down to the restaurant businesses to see what kinds of indicators may correlate with restaurants' survival during the pandemic, and New York City is selected to conduct the research. We are using the Yelp dataset to analyze a time period, from May to November in 2020, of restaurant information. Pulling out the outlook of restaurants that are possible closing doors in the future. To provide suggestions for restaurant businesses what strategies should take during this unexpected situation.

1.2 Research question

What constitutes a good restaurant? What are the most important concerns for the customers' satisfaction for a restaurant?What enables restaurants to survive during a pandemic like COVID-19. Common sense might attribute the success of a restaurant to delicious food, great service, location etc. We are going to explore the topic by proposing the following research questions in this paper.

Research questions 1: What features are the vital factors that can affect the survival state of restaurants located in New York City when facing disasters like COVID-19? Which features have an important effect?

Research question 2: Whether the extracted risk assessment features can be applied to restaurants located in other comparable areas and predict their survival state during 2020?

2. Background and related work

2.1 Background

Yelp is an American Technology company that is focusing on helping customers to target local business based on social network functionality. As one of the most popular and largest crowd-sourcing websites in the United States, Yelp has millions of users writing comments and along with providing a star-rating system. The review data in yelp is reliable and up-to-date and has wide coverage of all businesses. Nowadays, Yelp data has affected millions of customers' food choice decision-making and a large number of customers rely on Yelp for food hunting. Therefore, the review data in Yelp is valuable and has become an important proxy for the success of restaurants especially during the pandemic like COVID-19. In this paper, we are going to collect the data from Yelp and conduct sentiment analysis on them.

Sentiment analysis is a method of determining whether the content of text is positive and negative, which has a variety of applications ranging from detecting feedback from predicting the overall performance of the business. The recent years in the field of Natural Language Process has seen an increasing number of researches in text mining and sentiment analysis especially when it comes to business review analysis. In the meantime, a variety of classification algorithms such as Naive Bayeusan, Logistic Regression, Random Forest, Support Vector Machine, Genetic Algorithm have been employed in the field of text mining and sentiment analysis.

2.2 Related work

Preventing business closure is a critical topic of all time. Exploring business properties is essential for outlining the prediction model. In the project of Las Vegas Business Closure, five different machine learning models are used to predict the chance of business closure.[1] Besides, text content reviews provide consumers with information about experience goods, which have the quality that is observed only after consumption.[2] Text classification like Latent Dirichlet Allocation (LDA) is used as a topic model to discover the underlying topics that are covered by a text document. Hoffman et al. present a variation to Expectation Maximization algorithm which is described as an Online Learning algorithm for LDA.[3] Alternatively, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) is an algorithm which models the document space in a more discriminatory manner using nearest neighbors. To detect linguistic strategies corresponding to particular hypotheses, standard methods from computational linguistics and sentiment analysis that measure characteristics of words and sentences are being used widely in text reviews.[4] Still, there are extensive approaches to be applied regarding text analysis.

In 2013, workers in Google, Mikolov et al, put forward word2vector. This model makes up of the shortcomings of TF-IDF, since the word vector can predict the text meaning in context based on the inner liner relationship of encoding. Word2vec contains two novel models: Continuous Bag-of-Words Model (CBOW), which can predict a word based on the context; and Skip-gram model, which can predict the following text according to the current words [5]. Scholars also applied this model in sentiment tasks [6]. As an efficient approach, word2vec can not only process larger dataset within less time and can be applied in most NLP tasks, but also provides start-of-the-art-performance on the task for measuring syntactic and semantic word similarities at that time [5]. Yu et al. built refining word embeddings for sentiment analysis [7].

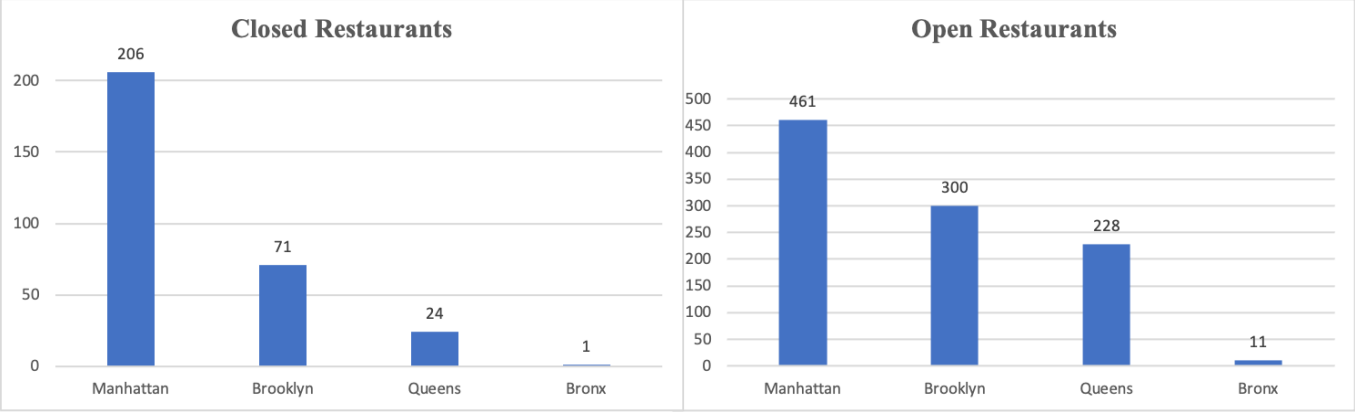
3. Data overview

3.1 Data collection

The current study sample covers restaurants in New York City. Given that there is no official city or state organization keeping track of closings in real-time, it is hard to get accurate restaurant closure data in the city during the pandemic. However, we obtained data from Eater, a food and dining online guide, which has been documenting the city's permanent restaurant closure since the start of the pandemic, tracking by local reporters and bloggers across the city. To establish confirmed closed data, we collected the recorded sample period from May 8th to October 31st, 2020. To access business details, we matched the recorded name to its site on Yelp then scrape the business ID. After collecting business IDs, we used Yelp API to return restaurant details. For characteristics analyses, this study used the city's restaurants that are still open in November 2020, collecting by sorting search from Yelp API. After deleting unmatched restaurant observations, this study retrieved 302 closed restaurants and 1,000 open restaurants in New York City.

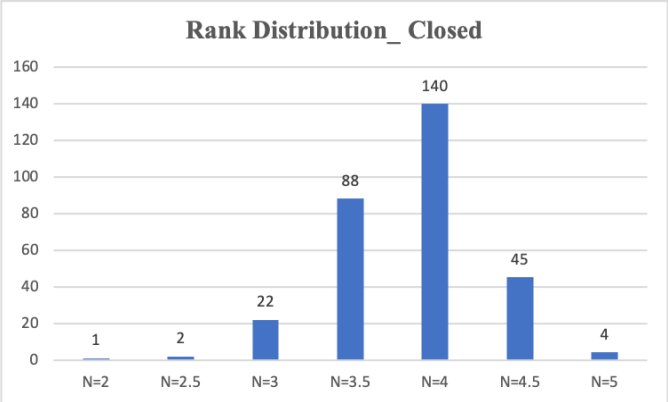
3.2 Data exploration

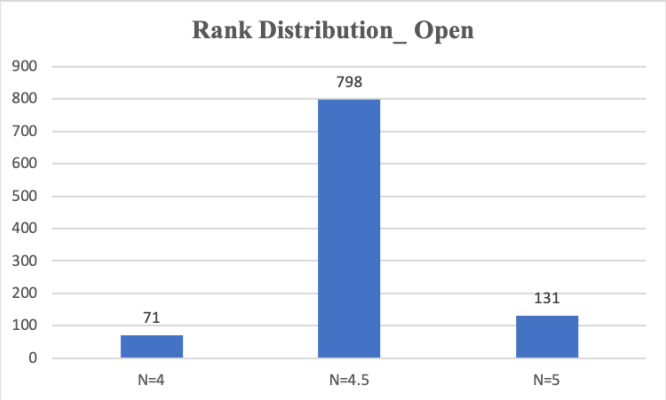
Area Distribution: Restaurants are divided into five sub-area: Manhattan, Brooklyn, Queens, Bronx, Staten Island by using zip-code. Here are the distribution tables:Closed restaurant: Manhattan': 206, 'Brooklyn': 71, 'Queens': 24, 'Bronx': 1. Open restaurant: Manhattan': 461, 'Queens': 300, 'Brooklyn': 228, 'Bronx': 11



From the above tables: Manhattan and Brooklyn cover the most restaurants. Thus, we decided to research these two areas in detail.

Rank Distribution: Restaurants' rank ranges from 1 to 5. Here are the distribution of rank: Closed restaurants: 4.0: 140, 3.5: 88, 4.5: 45, 3.0: 22, 5.0: 4, 2.5: 2, 2.0: 1. Open restaurants: 4.5: 798, 5.0: 131, 4.0: 71





From the above tables: 62.79% closed restaurants have rank over 4.0, 100% open restaurants have rank over 4.0. Thus, rank can be a feature to measure a restaurant's bankruptcy risk.

Price column: 1 stands for $, 2 stands for $$, 3 stands for $$$, and 4 stands for $$$$.1 means very cheap and 4 means very expensive. Price is also considered as a feature to measure a restaurant's bankruptcy risk.

Is_closed column: After collecting data, we label our dataset into open and close. 0 stands for closed and 1 stands for open

Review column: We scripe 10 reviews from the official website of the restaurant.

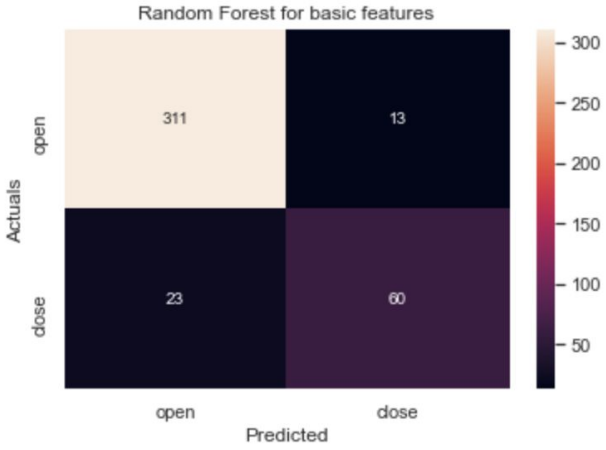
4. Methodology & Results

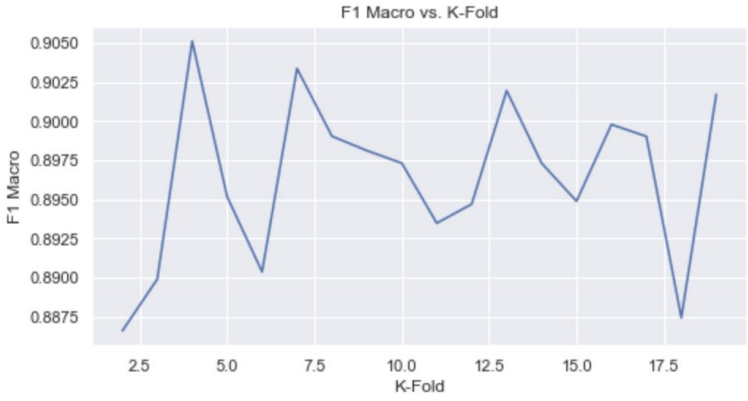
There are four parts in our methodology. In part 1, we use basic features that are displayed on Yelp, such as review count, rating, price, categories, to conduct preliminary classification. In part 2, we only use reviews to conduct sentiment analysis by using word to vector embedding approach, and six classification models are selected: SVC, Logistic Regression, LinearSVC, SGD, GradientBoosting, and RandomForest. In part 3, we enforce our feature engineering by TextBlob and other text analytics methods by combining the tf-idf matrix, document vector, and Vader for classification. In part 4, we use LDA topic modeling approach to analyze the negative review dataset.

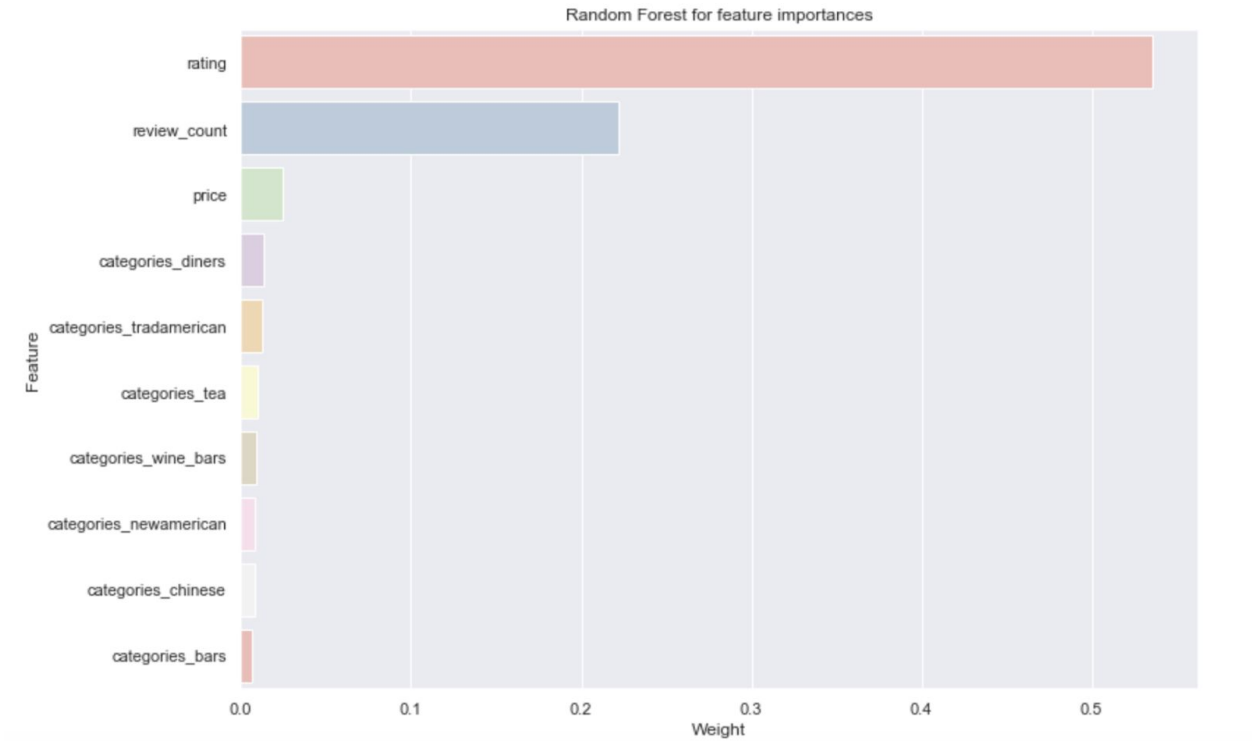
4.1 Basic features

In this section, we use three numerical variables and one categorical variable to build the primary classification model. Since each restaurant has at most three categories, we take the first as the representative. The classification algorithm chosen is Random Forest, which uses an ensemble of models to learn and predict outcomes. We then use feature importance for model interpretation.

We split our dataset into 70% of training and 30% of testing, evaluating the performance by the score of f1 macro and AUC. Both obtained well at 0.8573 and 0.9064, respectively. Also, we use k-fold validation to find that using 4-fold will obtain the maximum of average f1 macro, 0.9051. As expected, the three numerical variables have the top feature importance. The order from highest: rating, review count, and price. For categories, we find diners, traditional American, tea, wine bars, new American, Chinese, and bars are top survival features.





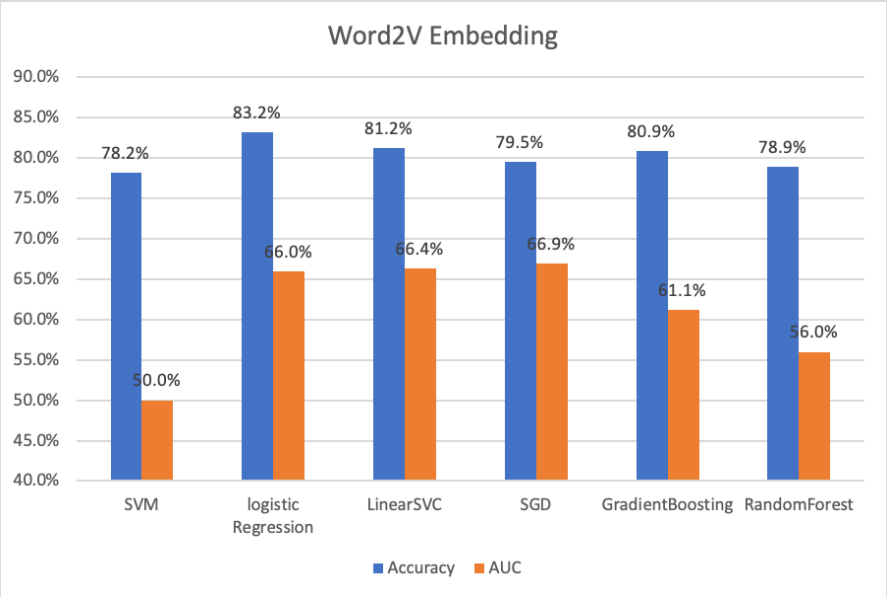


4.2 Word to Vector Model

First step is data selection & process. The Manhattan dataset, which contains open restaurants and closed restaurants in the Manhattan area, is the training set. The Brooklyn dataset, which contains open restaurants and close restaurants in the Brooklyn area is the testing set. Only the review column was chosen. Delete all rows that contain null value. Convert all characters into lower case, and split the reviews as a single word. Second step is Embedding and training. Import word2vec from gensim models. Training size to testing size is 7:3. The parameters are size=100, window=5, iter=10, workers=1, seed=2018, min_count=1. Third step is choosing models and testing accuracy, AUC. We choose six models: SVC, Logistic Regression, LinearSVC, SGD, GradientBoosting, and RandomForest. The table shows the results:

Embedding Method	Model	Accuracy	AUC
Word to Vector	SVC	78.19%	50.00%
Word to Vector	Logistic Regression	83.22%	65.98%
Word to Vector	LinearSVC	81.21%	66.35%

Word to Vector	SGD	79.53%	66.94%
Word to Vector	GradientBoosting	80.87%	61.15%
Word to Vector	RandomForest	78.86%	55.98%

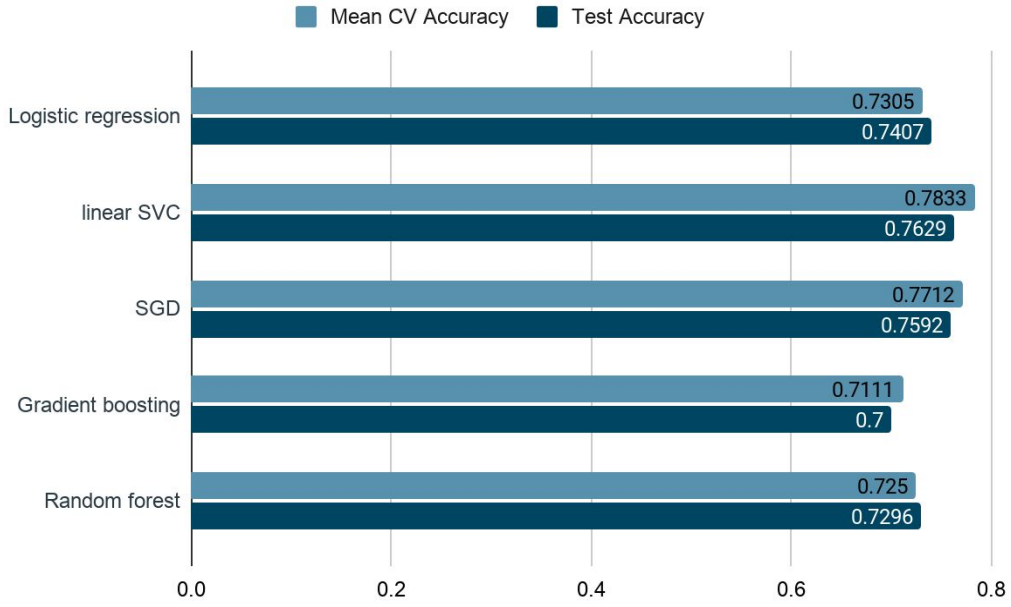


From the above table and chart, it is obvious that logistic regression has the highest accuracy of 83.22%, and SGD model has the highest AUC of 66.94%.

4.3.1 Use sentiment score to enforce feature engineering

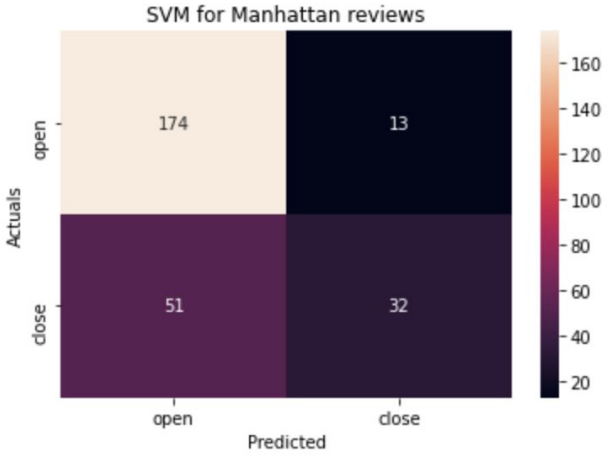
We clean the reviews by removing redundant symbols and lemmatizing the words. After cleaning, we use the tf-idf matrix and document vector as feature selection. We choose five classification algorithms for model venturing, e.g., logistic regression, linear SVC, SGD, Gradient Boosting, and random forest. We evaluate the performance by test accuracy and mean accuracy of five cross-validations. In the second trial, we will add four sentiment scores from Vader: negative, positive, neutral, and compound, to see how models perform.

We use tf-idf and document vectors in five classification algorithms, obtaining that the linear SVC model has the highest test accuracy of 0.7379. We then continue to add sentiment scores as our additional features, resulting in the score increased to 0.7629.



We try to generalize important survival features from reviews, therefore, we use random forest algorithms to extract feature importance to check. From the result, we find negative scores and some words, e.g., food, bar, bartender, have the top feature importance.

Finally, we use this model to predict Brooklyn district. We obtain an overall model accuracy of 0.69 on the test data.

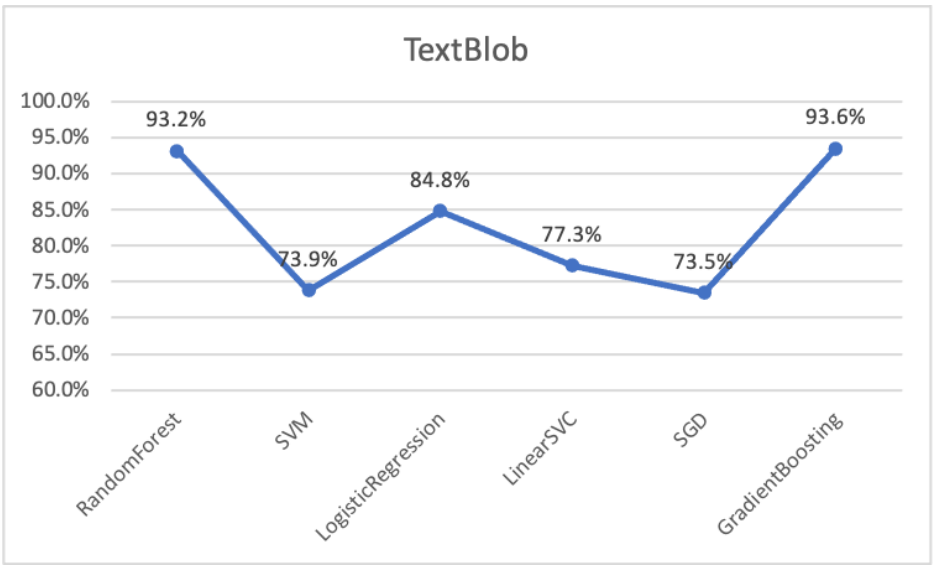


	precision	recall	f1-score	support
open	0.77	0.93	0.84	187
close	0.71	0.39	0.50	83
accuracy			0.76	270
macro avg	0.74	0.66	0.67	270
weighted avg	0.75	0.76	0.74	270

4.3.2 Use TextBlob to enforce feature engineering

First step is data selection & process. Same with Word to Vector Model, Manhattan dataset is used to do training and Brooklyn dataset is used to do testing. Five columns are selected: review_count, rating, price, reviews, is_closed. Second step is adopting the TextBlob approach, using two measurements: polarity and subjectivity to represent reviews. Third step is build regression, in which " is_closed" is the dependent variable, and "review_count", "rating", "price", "polarity", "subjectivity" are the independent variables. Next step is to choose models and test accuracy. The table shows the results:

Model	Accuracy
RandomForest	93.18%
SVC	73.86%
LogisticRegression	84.85%
LinearSVC	77.27%
SGD	73.48%
GradientBoosting	93.56%



From the above table and chart, it is obvious that GradientBoosting has the highest accuracy of 93.56%.

4.4 Negative reviews clustering and LDA topic modeling

To understand possible factors of closure, we use sentiment analysis to extract negative reviews from closed restaurants. We convert our negative reviews into the tf-idf matrix using bi-gram for better interpreting and removing stop words. Finally, we will try to generalize reasons from clustering results of the Gaussian mixture model. Meanwhile, we will use LDA modeling to generalize possible topics of closure based on negative reviews.

First, we separate the negative reviews with sentiment scores. We find nearly 10% of reviews are negative. We then check the optimal number of clusters with silhouette scores. Finally, we use GMM model with 4 clusters for clustering.

```
======= GMM =======


0 20 minutes understaffed hasty delayed finally order misunderstood good lot finally swapped food delayed food excell
ent hasty order lot miscommunication miscommunication server misunderstood food rushing understaffed lot rushing serv
ice good server lot excellent service food great ignored said definitely good

1 food isn 30 minutes left hungry menu price lunch menu noodles known modify website gotten noodles noodles saw terri
ble fresh known modify receipt charged paid noticed regular price ordered noodles noodles terrible website picked not
iced receipt website noodles price 17

2 sat bar probably microwaved food lukewarm servers said ignore took problem food right problem short ribs nearby got
bother send didn bother proceeded ignore place quiet night place green curry sitting nearby lukewarm didn got special
said hello asparagus cooked

3 mozzarella balls don know know place ricotta jalapeño mediterranean restaurant assumes anorexic toddler left choose
eat smallest portions payed 10 left nut ordered payed balls obviously balls size portions imaginable size toddler sel
fish places places ordered place owner anorexic choose
```

```
#miscommunication, understaffed, hasty misunderstood food order
#misinformation, Website menu price is incorrect
#microwaved food, servers ignored
#smallest portions, Not satisfied with the size of the meal
```

We draw the possible negative effects, including miscommunication between servers and customers, misinformation between menu price and website, microwaved quality of food, and small portions of food.

In the LDA topic model, we generate four topics for negative reviews. We use words at least five times showing and remove stop words and non-negative words. The result shows that the generated topics are regarding place, table, food, staff.

5. Analysis of Experiment results

5.1 Random Forest for Basic features

From the basic feature model, our model performance is really good, AUC obtained 90.64%. We attribute to the data not having much noise and the amount is large enough. We find that the rating is the key factor for closing. Indeed, rating represents the foundation of how we distinguish a restaurant. If a restaurant has a good reputation, its customer return will be higher, and it will be more resistant to unpredictable changes. We also find some types of restaurants are prone to close, compared with our exploratory data. For example, traditional American, bars, wine bars, tea, new Americans, etc. Bars were less likely to open with takeout or deliveries, resulting in a lot of closures. Another worth discussing category is traditional Americans. Possible reasons are that traditional restaurants are less with flexibility of adjustments, such as changing the menu for takeout or adopting delivery.

5.2 Word to Vector

Overall speaking, this method performs not well because the highest AUC is 66.94%. Even though the logistic regression model has accuracy of 83.22%, the AUC is just 65.98%. The main reason may be the dataset size is not large enough. Our first attempt uses the dataset that only contains the top 20% closed and open restaurants in the Manhattan and Brooklyn area in the beginning, while the results are bad: AUC is around 50%. Then we changed into larger datasets that contain all closed and open restaurants in the Manhattan and Brooklyn area. However, the results are still not good, but better than before.

5.3 TextBlob & Tf-idf, document vector, Vader

In this research, the Textblob approach performs better than word to vector approach. The highest accuracy of TextBlob approach is 93.56% under GradientBoosting model. It seems that using the TextBlob approach to represent reveiws is better than using the word vector model.

Our other attempt is to use the tf-idf matrix and document vetor, then continue adding sentiment scores from Vader. The performance of the final model is improved, however, it still needs more refining. Since the number of features increases phenomenally with each document considering each distinct word becomes a feature, this leads to huge sparse word vectors for textual data. Thus, if we do not have enough data, we may end up getting poor models or even overfitting the data due to the curse of dimensionality.

5.4 Negative reviews clustering and LDA topic modeling

In addition to building classification models for closure, we also try to understand the negative impact behind those closing restaurants, to discover latent hidden structures and patterns in data. Therefore, we group similar words from the negative reviews using clustering and summarize documents based on topic models. We explore unsupervised methods in this section. Still, text data is unstructured and highly noisy and thus the results might be biased. The words from clustering include miscommunication, unstaffed, hasty, misinformation, microwaved food, servers ignored, small portions, etc. Clustering can give us a feel for the possible groups or categories that our data might consist of, based on similar patterns and attributes. And we conclude some negative impact behind those closing restaurants.

6. Conclusion and future work

This study works to find what factors can affect the survival state of restaurants when facing disasters like COVID-19, and the restaurants in New York city were selected to do this research. Our analysis contains four steps. In the beginning, we used the Random Forest algorithm for basic features analysis, obtaining 90.64% of AUC. And results show that rating is the most important in the features of review count, rating, price, and categories.

Next, we use word embedding to do the sentiment analysis of reviews with six different classification models. However, this method performs not well with the highest AUC of 66.94%. Then, we view reviews as a feature, and add it into the initial regression model to enforce the feature engineering. TextBlob method is used to present reviews, which improve the initial model with the accuracy of 93.56% under GradientBoosting model. Another approach is to use tf-idf matrix and document vector, then add vader scores, which obtained improved accuracy too.

In the end, we use topic modeling to analyze the negative reviews on purpose to find the common features of closed restaurants.

In sum, the important features affecting New York City restaurants closing are those categories belonging to traditional American, bars, or coffee and tea. Bars are more obvious because our model extracts the import keywords like bar and bartender. Meanwhile, we find that restaurants need to have good communication in all aspects, whether in service or information. Once building effective communication from the bottom, those restaurants communicate about the COVID-19 crisis can create clarity, build resilience, and catalyze positive change during uncertainty.

Contribution

The research adopts a machine learning approach to extract features to identify restaurants' risk tolerance ability and predict restaurants survival state during COVID-19 period, and can be applied to other unexpected pandemic or disaster in the future.

This work also can provide insights for restaurants' owners to take strategies in advance based on the features we propose, when they meet unexpected events in the future. Additionally, risk appraisers can give a more accurate rating, and investors can get more useful information.

Limitation & Recommendation for future work

The limitations of this research and our recommendations are about three aspects: data size, number of features, and models selection. Due to time and data limitation, our closure dataset is

small. We only script the closed restaurants from May 8th to October 31st, and 10 reviews of each restaurant. This is a main reason why the word embedding model doesn't perform well. If further work can use more information about closed restaurants and script more reviews, a better performance should be achieved. In addition, further work can try to use more specific features, e.g., delivery service, environment rating, service rating, for precise research study interpretation.

In the future, we may use more advanced feature engineering models, which leverage deep learning and neural network models, e.g., LSTMs, BERT, Fine-turning BERT, to consider the sequence of data (words, events and so on). These are more advanced models than regular fully connected deep networks and usually take more time to train.

Reference

[1] Chen, X. et al., (2018). “Las Vegas Business Closure Prediction Model.”

[2] Michael Luca, (2016). “Reviews, Reputation, and Revenue: The Case of Yelp.com” *Harvard Business School*

[3] M. Hoffman and D. Blei., (2010). “Online Learning for Latent Dirichlet Allocation.” *Neural Information Processing Systems*.

[4] Dan Jurafsky et al., (2014). “Narrative framing of consumer sentiment in online restaurant reviews” *First Monday*

[5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).

[6] Giatsoglou, Maria, et al. "Sentiment analysis leveraging emotions and word embeddings." *Expert Systems with Applications* 69 (2017): 214-224.

[7] Yu, Liang-Chih, et al. "Refining word embeddings for sentiment analysis." *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2017.

[8] Sarkar, D. (2019). Text Analytics with Python: *A Practitioner's Guide to Natural Language Processing.* Berkeley, CA: Apress.