Analysis on the Survivability of NYC Restaurants During COVID-19

Ling Ao, Lei Zhang, Szu-Yu Chen
Instructor: Rong Liu

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics
May 28th, 2021

Introduction

- **Background:** The COVID-19 pandemic has a profound effect on all countries over the world. Food industry is one the businesses in the U.S. that got hit by the pandemic the hardest.
- According to Yelp data, permanent closures have reached 97,966, representing 60% of closed businesses that won't be reopening.
- **Research Question:** 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?
- **Significance:** 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.

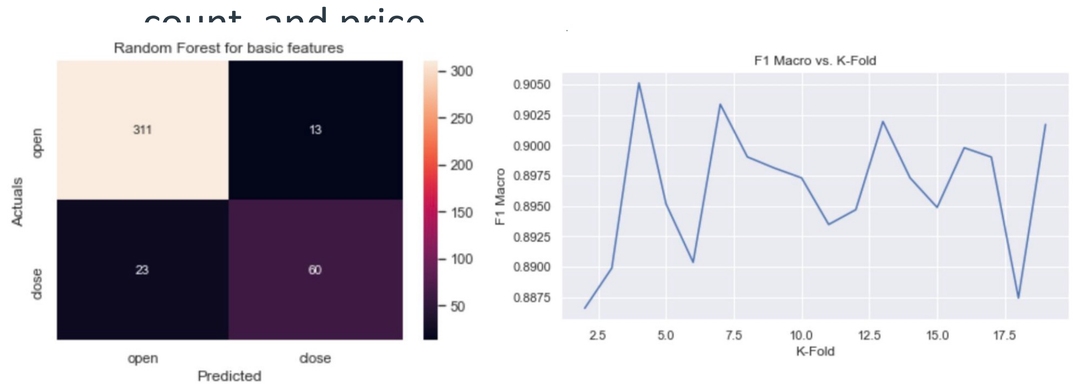
Data Collection & Exploration

- **Data collection:** Mined permanent close restaurants due to COVID-19 from May 8th to October 31th from Easter New York by using python, total 302 restaurants. Collected 1000 survived restaurants' information by using Yelp API.
- **Features:** Category, Location (5 arears in New York), Rank (from 1 to 5), Is_closed (0: closed, 1: open), Price (1 for $, 2 for $$, 3 for $$$, 4 for $$$$), Review (10 reviews for each restaurant)

Methodology & Results

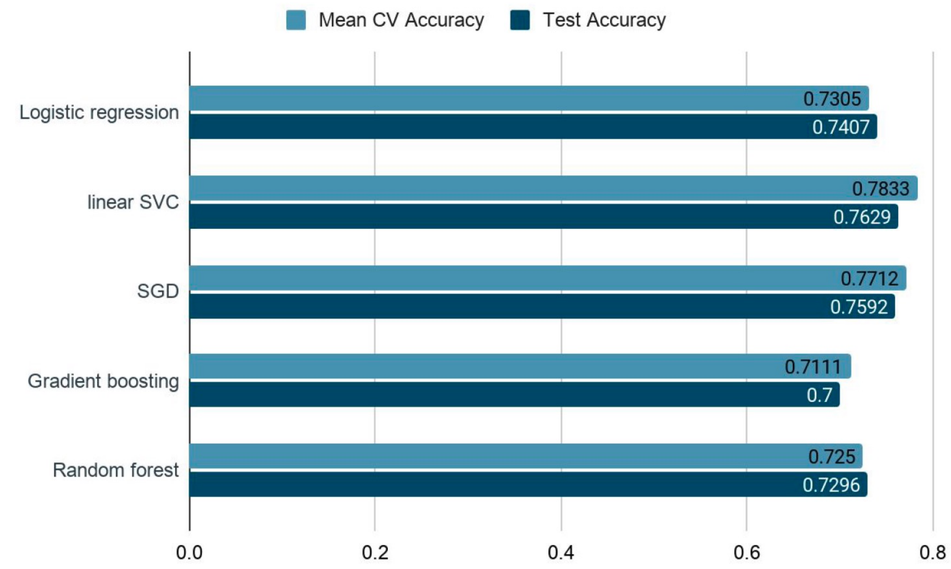
Method 1: Basic features classification model.

- Used three numerical variables and one categorical variable to build the primary classification model, and chose random forest classification algorithm.
- Evaluating the performance by the score of f1 macro and AUC, with accuracy 0.8573 and 0.9064 respectively. The order from highest: rating, review count, and price



Method 2: Use sentiment score to enforce feature engineering

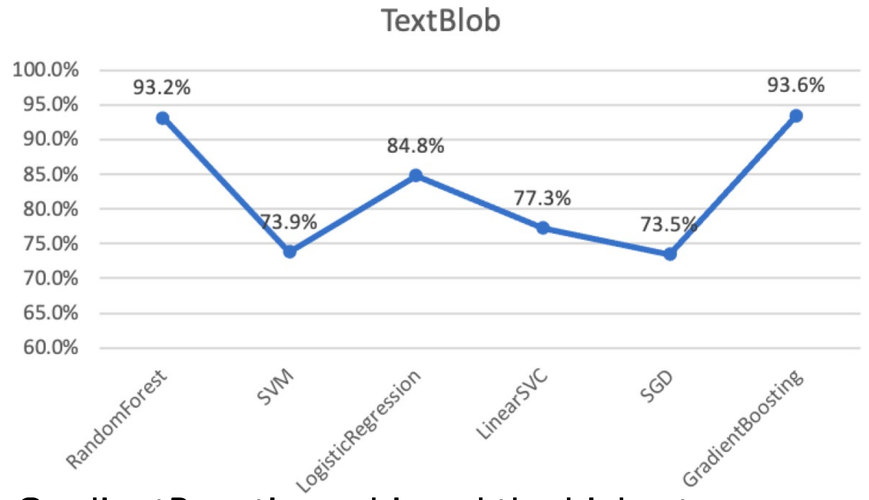
- Cleaned the reviews first, and then used the TF-IDF matrix and document vector as feature selection.
- Chose five classification algorithms for model venturing : logistic regression, linear SVC, SGD, Gradient Boosting, and random forest
- Used four sentiment scores from Vader: negative, positive, neutral, and compound
- Manhattan dataset isused to do training and Brooklyn dataset is used to do testing



- Linear SVC model has the highest test accuracy of 0.7379. After adding sentiment scores as, score increased to 0.7629.

Method 3: Use TestBlob to enforce feature engineering

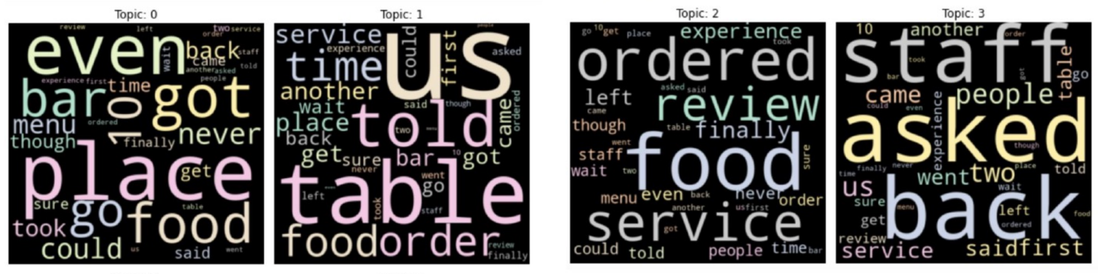
- Adopting the TextBlob approach, using polarity and subjectivity to represent reviews.
- Dependent variable: is_closed, Independent



- GradientBoosting achieved the highest accuracy at 93.6%

Method 4: Negative reviews clustering and LDA topic modeling

- The result shows that the generated topics are regarding place, table, food, staff



Conclusion

- Categories belong to traditional American, bars, and tea are more likely affected by unpredictable changes
- Bars are more obvious because our model extracts the important keywords like bar and bartender
- Feature Importance : rating > review count > price > category
- Adding review as a feature can enforce the feature engineering. Compared with sentiment score approach, TestBlob approach performanced better, and achieved the accuracy at 93.6%