Final Report

Kanav Bhalla

Ceni Babaoglu

501145702

July 25th, 2022

An emerging industry in numerous markets throughout the world has been food delivery apps and the use of technology to make the process of ordering deliveries more efficient. This market has expanded exponentially during the pandemic as more consumers grew reliant on delivery services rather than having to leave their homes to get their goods. Due to this increase in demand, there has been an influx of new startup apps that have been joining the market to meet the new demand, among these startups is a new startup pizza delivery app GETPIZZA. With a growing market there have been more apps being developed to compete for dominance within the industry “the sustainability of the FDA is dependent on two aspects, from a trade point of view: first, the FDAs must be able to serve the needs and expectations of its existing consumers. Second, the increasing number of FDA service-providers is also increasing competition among them (Ray et al., p. 2022, 2018). Being able to filter out potential partners will greatly help expand the reach of GETPIZZA and ensure that the company is able to differentiate from competition. Food delivery apps have changed the overall landscape of the food and service industry increasing the profitability and opportunities for new technology companies, restaurants, drivers and suppliers, growing the overall economy exponentially. To meet the demand GETPIZZA has by their customers and to ensure their app is run efficient the business is requiring a quantifiable list of potential pizza stores that they can partner with, to do so they are requesting a list of the top pizza restaurants to work with. Having the right partnerships is extremely important as it ensures the business stays profitable and grows, we add restaurants that customer would want to shop from and to avoid complaints from customers ordering through the delivery app ensuring this satisfaction will also ensure the GETPIZZA app stays profitable and retains the customers they have.

The Yelp Dataset was an extremely valuable dataset to use as it had over 150,346 business in 11 of the largest metropolitan areas with numerous businesses that fall under various categories and focuses. The Yelp Dataset was the primary data used for this project, however as there were numerous businesses that did not include pizza within the dataset as this is a broader dataset thus there were many techniques that are going to have to be used to ensure we are focusing our mining and recommendations based on the most applicable parts of the data set, at first my goal was to extract only stores that solely indicated that they sold pizza, however to ensure we had a more broad pool and to not limit more opportunities for the app any store that indicated that they had pizza as an option was included in the analysis.

**Research Questions:**

* What are the top ranked pizza restaurants on Yelp? Graphical user interface

  Description automatically generated with low confidence

Within the quantitative analysis after clearing out the broader yelp dataset, a list was used to identify all of the pizza stores with the highest rank of 5 stars, this concluded that out of the 6067 overall pizza stores within the YELP dataset there was only 123 pizza stores with the ranking of 5 stars, With satisfaction being such a key indicator of customer retention ensuring GETPIZZA’s partners with these stores would be a great start in ensuring GETPIZZA is efficient in terms of selecting its partners.

* What are the lowest ranked pizza restaurants on Yelp? (This list will benefit GETPIZZA on shops we recommend avoiding partnership with).

Graphical user interface, table

Description automatically generated

* Which states have the most pizza stores on yelp? (Based on this info the app can know which areas to target for max partnership opportunities).

Table

Description automatically generated with medium confidence

* The major goal of this research and project was to determine which pizza stores will be the most suitable for the pizza delivery app to target as potential partners?

A screenshot of a computer

Description automatically generated with medium confidence

The Yelp Dataset was an extremely valuable dataset to use as it had over 150,346 business in 11 of the largest metropolitan areas with numerous businesses that fall under various categories and focuses. The Yelp Dataset was the primary data used for this project, however as there were numerous businesses that did not include pizza within the dataset as this is a broader dataset thus there were many techniques that are going to have to be used to ensure we are focusing our mining and recommendations based on the most applicable parts of the data set, at first my goal was to extract only stores that solely indicated that they sold pizza, however to ensure we had a more broad pool and to not limit more opportunities for the app any store that indicated that they had pizza as an option was included in the analysis. Using Classification analysis once the dataset was divided into only pizza stores, I was able to begin the cleaning of the dataset and begin the preliminary classification analysis.

**The Main Contribution of the Work Compared to Past Research:**

As this industry is just growing there is not many papers and projects that have focused on potential markets and finding ideal partnerships “The statistical results and discussions show that satisfaction is the most significant factor, and perceived task-technology fit, trust, performance expectancy, social influence and confirmation have direct or indirect positive impacts on users’ continuance usage” (Zhao and Bacao 2020, p. 1). With satisfaction being such a key indicator of customer retention finding ideal partners would ensure that GETPIZZA remains profitable and retains the customers using their platform as they continue to grow. Many papers have been done on the YELP dataset however no paper has focused on specifically restaurants under the context of food delivery apps rather majority of the work done under this dataset have been sentimental analysis to predict reviews of restaurants although this work is extremely valuable this project did not focus on the sentiment rather focusing on capitalization of the reviews and analysis of the dataset to ensure an emerging app could be optimal. “The findings show performance expectancy, social influence, and facilitating conditions positively influence behavioral intention to use food order-delivery features in ride-hailing applications” (Surya et al,. 2021, p.1363) finding the optimal stores to partner with to ensure performance and meet customer expectancy is what could set an emerging food delivery application apart which is why this project was selected.

**Description of Methods and Conducted Analysis**

import pandas as pd

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

yelp= pd.read\_csv("pizza.csv")

yelp

To conduct the analysis the first step was importing the necessary packages and reading the dataset into python, through excel I was able to filter the dataset into only containing businesses that had pizza as an option, as the yelp dataset was extremely broad with multiple types of businesses to ensure my analysis was focused this was a key step in ensuring my analysis would be valuable for GETPIZZA.

yelp.describe()

I then began by getting a grasp of the dataset I was working with using, this was valuable as within the dataset I was able to determine the mean of the stars within the dataset which equaled 3.34 and the average review count which was 49.12. Another valuable insight was learning the min and max of the key indicators of a valuable business including the review count which was a max of 3741.

yelp.columns

yelp.dtypes

yelp = yelp.replace({'\$':''}, regex = True)

I then displayed the columns and datatypes and removed any characters that could pose issue with future analysis and techniques.

plt.scatter(yelp['stars'], yelp['review\_count'])

I plotted the relationship between review count and stars to see how review count affects the stars and noticed that majority of the highly reviewed restaurants were among the higher starred.

star.sort\_values(by='review\_count', ascending=False)

I then began the process of sorting my data to begin my analysis of the best restaurants to partner with and avoid as partners, I began by sorting the dataset from highest star rating to highest and lowest, and then created a side dataset for only the 5-star restaurants after doing this I sorted this to the highest reviewed restaurants with 5 stars. This indication was extremely valuable as through this I was able to find out my prime restaurants to recommend to GETPIZZA including "Akropolis Gyro & Pizza", "In Forno Pizza" and also identify some 5 star restaurants that may not be the best partners based on the low amount of reviews they had which may be skewing their overall rating including "Blaze Fast-Fire'd Pizza" which although having 5 star review this store only had 3 review which is a indicator that the sample size is not big enough compared to a restaurant such as “Akropolis Gyro & Pizza" which has 214 reviews.

yelp.sort\_values(by='review\_count', ascending=False)

The next sort I had used was sorting based on review count, this became an extremely valuable sort for my overall analysis as after the primary results it indicated that although 5 star restaurants is the max in terms of rating, the review count may be a better indicator of restaurant success and customer loyalty this was also proven by the top 5 restaurants in terms of review count which the lowest star rating was 4.0 an extremely high rating and as mentioned the review count for these top 5 was over 1700 reviews this indicated that these restaurants had a loyal following that approved their service, this sort also demonstrated that the top 5 lowest rated restaurants did not indicate correlation to star as "Mama C's" only had 3 reviews but had a star rating of 4.5 stars. This was a key breakthrough in my analysis as at this point I had shifted my focus from believing that recommending based on star count would be most appropriate for GETPIZZA to looking into the overall factor of how many reviews the restaurant had and the corresponding star rating.

lowstar = yelp.loc[yelp['stars'] == 1]

I then shifted my focus to the restaurants that are less preferable for GETPIZZA to partner with separating the overall yelp dataset into a subset labeled lowstar this resulted to a total of 69 restaurants with only 1 star rating, this list was extremely valuable however a key observation made during this process was that most of the lowest rated pizza stores within the dataset had an average of 5.75 reviews this was much lower than the average review count of 5 star pizza restaurants which was 12.3. A key takeaway at this point and observation was that review count could directly be a more appropriate as an area focus in my recommendations for GETPIZZA to operate with. Another observation through this process was that there were much less restaurants reviewed with 1 star than higher ratings including the max rating of 5 star. With the clear observations from my initial analysis of this data by filtering the 5 star and 1 star restaurants I was able to determine that review count would be a major factor to utilize in terms of my recommendation, I then shifted my focus to finding out the restaurants with the highest amount of review counts with 1 star, the highest reviewed restaurants in this rating would be what I would highly recommend GETPIZZA not partner with these restaurants included surprising results as majority of them were highly identifiable pizza stores including PIZZA HUT, SBARRO, LITTLE CEASERS PIZZA, This was an interesting observation as although these specific locations were rated 1 star there were numerous other locations with higher ratings specifically when looking at Pizza Hut there were over 10 restaurants with higher than 4 stars however the clear median was of all of Pizza Hut’s ratings were 2 stars. Pizza Hut out of all of the major food chains rated the worst with 22 restaurants out of the overall 69 restaurants ranked 1 star were Pizza Huts. At this point in my analysis there was clear point to be made that although restaurants may have one stars the low review count especially for top brands was an indicator that their reviews may be skewed based on customers only leaving sentiments when they have bad experiences, in contrast the highest reviewed pizza restaurant in the entire yelp dataset was “Secret Pizza” with over 3741 reviews which in contrast had much higher ratings but only one location unlike the 304 Pizza Huts in the dataset.

lowstar.sort\_values(by='review\_count', ascending=False)

The next process was filtering the data set with 1 stars to the highest review count, to see which stores were most reviewed and had lowest rating. This result also showed that Pizza Hut had 3 of the top 5 lowest ranked pizza stores with the most review count.

pizzahut= yelp.loc[yelp['name'] == '"Pizza Hut"']

pizzahut.describe()

littleceasers = yelp.loc[yelp['name'] == '"Little Caesars Pizza"']

littleceasers.describe()

I then began the process of evaluating all of the top pizza brands that appear in more than one state and have over 50 pizza store locations, this found that Pizza Nova was the best pizza chain to partner with, followed by Domino’s and Little Caesars and Papa Johns lastly Pizza Hut. Although these stores had lower stars than the top-rated ranking because of the amount of stores they had any of these stores would be ideal to partner with to ensure that GETPIZZA could be utilized in more locations.

yelp.loc[yelp['review\_count'] > 1000]

The next observation I found was filtering the data set by restaurants with over 1000 reviews this was when I became certain with the value of the review count metric in terms of recommendations as although there were only 14 results the lowest star rating amongst them was 4.0 with a peak of 4.5 this indicated that as the review count grew the more support a restaurant had, this indicator is what I would base most of my recommendations on focusing on high review count restaurants with over 2.5 stars to ensure efficiency for GETPIZZA, however to ensure efficiency of partnerships I recommended that GETPIZZA avoid discrediting stores that were lower than 2.5 stars as long as they had 50+ locations as the value these stores would bring in terms of cross state platforming would be immense in the development of GETPIZZA’s awareness and growth.

yelp.loc[yelp['state'] == "NV"]

yelp['state'].value\_counts()

NV= yelp.loc[yelp['state'] == "NV"]

NV.describe()

I then shifted the focus to the premier locations that GETPIZZA should market and choose to focus on based on the prominence within the YELP Dataset, the top ranked states/provinces included Arizona, Ontario, Ohio, Nevada, and Pennsylvania. Based on these results the primary focus in the United States would be Arizona and in Canada the primary focus would be Ontario. This evaluation demonstrated a key limitation within the dataset as many states and provinces were not represented such as surprisingly one of the major markets for pizza which was New York, however even though this could be a limitation for this analysis with the knowledge GETPIZZA has on the individual states including the analysis on Nevada finding out the average review count and the average stars per state could give GETPIZZA a competitive advantage when entering the specific markets, for example the mean of the stars in Arizona was 3.36 higher than Nevada however the review count in Nevada was much higher with 109.50 for Nevada and 86 for Arizona.

yelp.isnull()

yelp.isnull().sum()

yelp.duplicated()

yelp.drop\_duplicates()

The next process was removing nulls which were mostly in the Neighbourhood column and dropping duplicates to clean datas.

yelp[['stars','review\_count']].corr()

sns.heatmap(yelp[['stars','review\_count']].corr(), annot=True, cmap = 'Reds')

plt.show()

I aimed to find the correlation between the stars and review count, through this I made a correlation matrix heatmap which was valuable as we were able to determine that stars had a high rating in terms of correlation.

X= yelp[['review\_count']]

y= yelp[['stars']]

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size =0.2, random\_state=10)

clf = LinearRegression()

clf.fit(X\_train,y\_train)

clf.predict(X\_test)

clf.score(X\_test, y\_test)

clf.score(X\_train, y\_train)

The next process was splitting the data set in 80/20 train and test data split, I was aiming to test the accuracy of the model I chose to make stars the X and the review count the y. I chose this because the stars were highly dependant on the review count as we determined in my classificational analysis. The accuracy was 36% for the test dataset and 29% for the training dataset.

**Summary of findings:**

* Arizona and Ontario are the top states for GETPIZZA to target first.
* Restaurants with over 1000 review count are ideal to partner with.
* Restaurants with over 50 restaurants are often starred lower than independent restaurants, Pizza Nova was the best franchise to partner with and the worst was Pizza Hut.
* Independent restaurants (lower than 50 restaurants) should be discredited if they have high review count and a lower rating of 2.5 stars or if the restaurant is under 2 stars with less than 10 reviews

**Shortcomings of the Work:**

The shortcomings of the work are mainly with the dataset present, as the Yelp dataset did not allow many opportunities to create a recommender system as there are not many numerical statistics to help predict with. Another limitation was that many states were not represented including key states and markets such as New York and British Columbia. Another limitation was not being able to create a system of predicting what the star rating would be based on which state they are located although we were able to classify which states would be best it would be valuable to predict which states would offer the highest star count for potential new Startup pizza restaurants. Moving forward with this topic finding a dataset of an existing delivery app including UBER EATS would be extremely valuable, however as this is an emerging market the analysis done with YELP dataset is still extremely valuable and valid for a Startup business and just as this is focused on pizza stores, we could duplicate this process with any other cuisine or any other area of focus for businesses trying to evaluate the overall market.

Sources:

Ade Permata Surya, I Made Sukresna, & Aris Mardiyono. (2021). FACTORS AFFECTING INTENTION TO USE FOOD ORDER-DELIVERY FEATURE OF RIDE-HAILING APPLICATIONS: THE UTAUT APPROACH. *International Journal of Business and Society*, *22*(3), 1363–1383. <https://doi.org/10.33736/ijbs.4306.2021>

Ray, Dhir, A., Bala, P. K., & Kaur, P. (2019). Why do people use food delivery apps (FDA)? A uses and gratification theory perspective. *Journal of Retailing and Consumer Services*, *51*, 221–230. <https://doi.org/10.1016/j.jretconser.2019.05.025>

Sharma, Dhir, A., Talwar, S., & Kaur, P. (2021). Over-ordering and food waste: The use of food delivery apps during a pandemic. International Journal of Hospitality Management, 96, 102977–. <https://doi.org/10.1016/j.ijhm.2021.102977>

Zhao, & Bacao, F. (2020). What factors determining customer continuingly using food delivery apps during 2019 novel coronavirus pandemic period? *International Journal of Hospitality Management*, *91*, 102683–102683. <https://doi.org/10.1016/j.ijhm.2020.102683>