Identifying strongly correlated factors relevant to restaurants' operating statuses

World Count: 1486

Introduction

Many restaurants operate in a competitive environment across the United States. It is one of the popular industries in which quick-service restaurant franchise alone generate more than \$250 billion in 2018 to the US economy (Statista, 2020). However, some businesses failed in the operation, and motivated researchers focused on studying the financial position and failure rate (Parsa, Self, Njite, & King, 2005). Furthermore, there are other factors to consider, such as negative reviews received from customers and restaurants' attributes themselves. This study will analyse and identify factors strongly correlated to the business operating status under normal circumstances (before the COVID-19 pandemic). The factors were retrieved from the restaurants' attributes, and coefficient correlation was applied to find the association degree.

Data and Methodology

Firstly, the yelp dataset [1] was preprocessed in Python to take restaurants from business category and transform the attribute columns to a list of features. As a result of data transformation, the dataset contains 23,284 restaurants from 491 cities in 21 states consisting of 6,818 opening restaurants and 6,466 that had already closed. Also, the definition of 50 features used to identify strongly correlated factors is available in Appendix A. Restaurants operating statuses, whether it is opening or closed, were used as a target discrete variable to define the relationship with other factors. The coefficient correlation is the best measurement that yields the strength and direction of a linear relationship between two variables of factor (x) and operating status (y).

The coefficient correlation gives the degree of relationship with the range value between $[-1, 1]$ where 1 indicates a perfect positive correlation and both variables move in the same direction. Conversely, they will move in the opposite direction if the coefficient is negative, which is anti-correlated. The zero coefficient means no correlation between two variables, but this does not mean that they are not independent. The most well-known coefficient method is Pearson correlation, which takes covariance of continuous two variables (x and y) divided by the multiplication of each variable's standard deviation. However, this is not an applicable approach as the target variable is boolean with two possible values, True and False. Also, the factors (x) consist of quantitative, nominal, and ordinal variables. Thus, alternative coefficients such as Point Biserial, Rank-based correlation, and Thiel's U are required (Calkins, 2005).

Point Biserial coefficient measures the association degree between continuous variables, such as the total number of opening restaurants in the same area and the boolean variable of operating

[1] The dataset was downloaded on 27th January 2021 from https://www.yelp.com/dataset.

status (Lev, 1949). Formula 1 gives the value of the Point Biserial correlation [2] which will be equivalent to Pearson correlation (Scipy Contributors, n.d.).

$$r = \frac{\bar{Y_0} - \bar{Y_1}}{s_y} \sqrt{\frac{N_0 N_1}{N(N-1)}} \tag{1}$$

The rank correlation is required to find the correlation between ordinal variable (e.g., price range in ranking) and nominal variable. Kendall's Tau is one tool that treats observations as tied data. Its Tau-b (2) takes concordant and discordant pairs that measure how two variables have the same and opposite signs respectively[3] (Kendall, 1938). Another alternative rank correlation is Spearman's coefficient with simpler computation but its confidence intervals is less reliable and interpretable compared with Kendall's Tau coefficient (Kendall & Gibbons, 1990).

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{2}$$

The last coefficient is the Thiel's U or uncertainly coefficient (Equation 3), which measures the nominal relationship based on conditional entropy ($H(X|Y)$) indicates the expected quantity of information to describe the outcome given that the X value is already known (Press, Teukolsky, Vetterling, & Flannery, 2007). This conditional entropy also solves the symmetric issue found in the Cramer's V as the asymmetric information between two variables was lost, and the independent column was not defined (Cramer, 1946). Unlike other coefficients, this uncertainty coefficient will give the range values of [0, 1].

$$U(X|Y) = \frac{H(X) - H(X|Y)}{H(X)} \tag{3}$$

The last step is performing statistical testing by calculating the p-value, a 2-sided test on a null hypothesis such that there is no association between two variables. The p-value is statistically significant if it is less than the threshold of 0.05, which means less than a 5% chance that the null hypothesis is true. The p-value is calculated by using formula $p = 2 * CDF(-|r|)$ where CDF is cumulative distribution function of beta distribution and r is the coefficient value returned from the above association measurements (Scipy Contributors, 2020).

Results and Discussion

After applying correlation coefficient methods and retrieving the p-value, some features needed to be removed from the analysis; thus, any variable with a p-value of correlation coefficient less than 0.05 was discarded. It did not reject the null hypothesis that there is no relationship between the restaurant attribute and the target variable, i.e., operating status. The coefficient of these non-significant variables is available in Appendix B. As a result, the number of features was reduced down to nine features from 50 in total as showed in table 1.

The number of reviews received from customers had the highest positive correlation coefficient value of 0.153 with a p-value of 1.138e-121. This coefficient was from the Point Biserial approach and can be interpreted that once the number of reviews increased, it is associated with opening

[2] The coefficient uses with n-1 degree of freedom where $\bar{Y_0}$ and $\bar{Y_1}$ are the mean of observations labelled by 0 and 1 respectively, N_1 and N_2 are the number of observations labelled by 0 and 1 and s_y is standard deviation (Scipy Contributors, n.d.).

[3] n_c is number of concordant pairs, n_d is number of discordant pairs, $n_0 = n(n-1)/2$, $n_1 = \sum t_i(t_i - 1)/2$ where t is number of tied values for the first quantity and $n_2 = \sum u_i(u_i - 1)/2$ where u is number of tied values for the second quantity (Kendall, 1938).

Feature Name	Correlation Coefficient	P-value
ReviewCount	0.153	1.138e-121
Dinner	0.052	1.605e-15
Classy	0.047	7.581e-13
Lunch	0.046	2.353e-12
HasTV	0.029	7.809e-06
DriveThru	0.026	6.966e-05
Breakfast	0.019	0.004
RestaurantsPriceRange	-0.056	1.230e-18
OpenCount	-0.086	3.592e-39

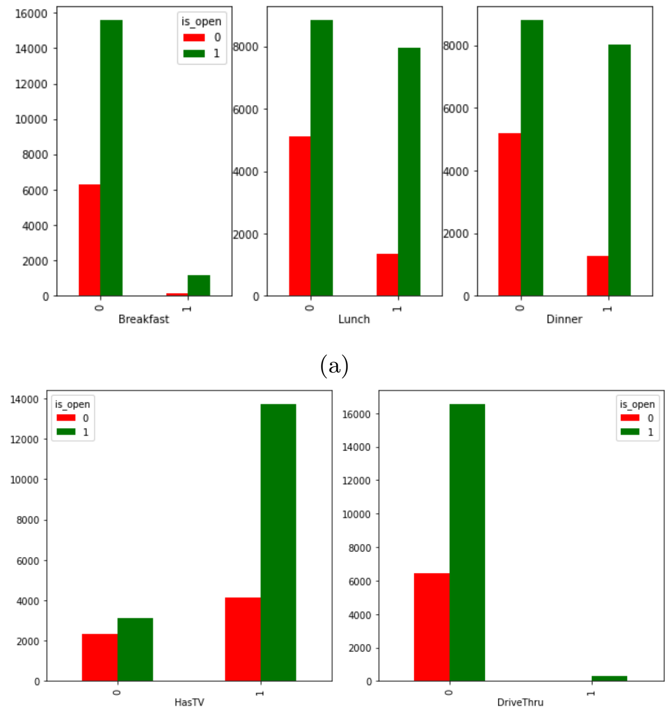
Table 1: The correlation coefficient of nine features with the p-value less than 0.05 and sorted from the highest to the lowest coefficient value by 4 significant figures.

status (i.e., the value of one). Also, the average review received for operating restaurants was 156.5, which is higher than the average of closed businesses with a value of 66.5. Figure 2 shows the frequency of reviews received from customers by the operating status. However, the business ratings presented different results.

The frequency analysis revealed that 5,340 restaurants received high ratings larger than 3.5. Meanwhile, there are low rating restaurants (i.e., lower than 2.5) still operating. It can be implied that the business ratings did not necessarily cause the business to shut down. Importantly, this review rating factor was discarded due to a low coefficient value of 0.0002, and a p-value greater than 0.05 explained no correlation with the target variable.

Breakfast, Lunch, and Dinner features were extracted from the GoodForMeal attribute, which defines whether the restaurant is suitable for a particular meal type. These three attributes all had weak positive correlations with values close to zero, 0.019, 0.046, and 0.052, respectively. Similar to the Ambience attribute (i.e., the atmosphere of a place), only Classy was significant with a coefficient value of 0.047 and p-value at 7.581e-13. The availability of television and drive-through service can also affect the operating status, but it was not considered a strong impact due to coefficient values close to zero (0.029 and 0.026). These coefficients from Thiel's U also followed a similar interpretation above. Once the category value moves from zero to one, it also corresponds to opening restaurant status. The frequency plot of these nominal variables against the operating statuses is available in the figure 1.

In contrast, two variables had a negative correlation, RestaurantsPriceRange, and OpenCount. The price range is an ordinal variable, and Kendall's Tau gave a negative coefficient with a value of -0.056. The result suggested that increasing the rank of price range tends to decrease the category or rank of the variable to zero, i.e., closing down. There are 85 restaurants that offer high price service at rank four had already shut down, and 155 restaurants that charge the same price range are still opening. In comparison, 5,624 restaurants with low-priced food are still operating while 1,857 of them were closed. Overall, the proportion of closing down restaurants with more expensive services is higher than restaurants offered food at a lower price. Next is the competitiveness, determined by the total of operating restaurants operated in the same area (by city and state) can also increase the chance of business closing down with the coefficient value of -0.086 given by point biserial. The shutdown restaurants faced a higher number of competitors with a mean of 1,996.014, and opening business has a lower average of competitors at 1,695.720 (Figure 2). This number of competitors is the second correlated attribute to operating status following the number of reviews. Nevertheless, it is still considered as a weak association as the coefficient value is closed to zero.



(a)

(b)

Figure 1: The frequency plot of boolean restaurant attribute against operating status.

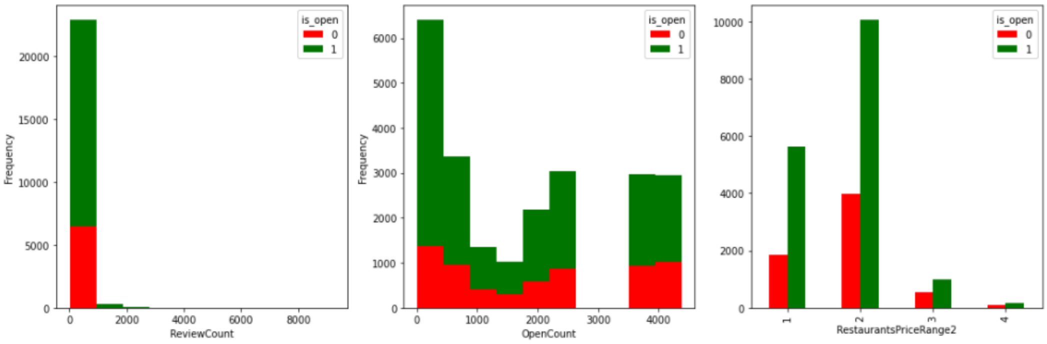


Figure 2: The frequency plots of continuous variables (ReviewCount and OpenCount), and ordinal feature of restaurant price range against operating status.

Conclusion

Despite many available features for analysis, most of the nominal restaurant attributes were discarded due to the low significance value of the correlation coefficient determined by the p-value, which left only nine features. The number of reviews was the most strongly correlated to restaurants' operating status with the coefficient of 0.153 and p-value at 1.138e-121. It could be said that the restaurant owner should pay attention to the number of reviews that reflect the number of customers who share their experience. Meanwhile, they should also consider offering a meal at different times of the day, including breakfast, lunch, and dinner, and ensuring a classy atmosphere. Additionally, the restaurant may have television as entertainment for customers or provide Drive-Through service for customer convenience to receive the order. The negative correlation of price range suggests that the price range is to keep reasonable price to persist competitive position especially with a higher number of restaurants operating in the same area to remain open. However, all correlation coefficients were low due to a large number of samples (i.e., 23,284 restaurants) used in this study. The coefficient can be improved by reducing the population and the scope to consider a particular cuisine. The study could also be extended to consider economics and financial statements (e.g., profits and losses) and investigate whether they gave a strong correlation against restaurant attributes.

References

Calkins, K. G. (2005). *More correlation coefficients.* Retrieved from https://www.andrews.edu/~calkins/math/edrm611/edrm13.htm

Cramer, H. (1946). *Mathematical methods of statistics / by harald cramer* [Book]. Princeton University Press Princeton.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. Retrieved from http://www.jstor.org/stable/2332226

Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). A Charles Griffin Title.

Lev, J. (1949). The Point Biserial Coefficient of Correlation. *The Annals of Mathematical Statistics*, *20*(1), 125 – 126.

Parsa, H., Self, J., Njite, D., & King, T. (2005). Why restaurants fail. *Cornell Hotel and Restaurant Administration Quarterly - CORNELL HOTEL RESTAUR ADMIN Q*, *46*, 304-322.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2007). *Numerical recipes 3rd edition: The art of scientific computing* (3rd ed.). USA: Cambridge University Press.

Scipy Contributors. (n.d.). *scipy.stats.pointbiserialr.* Retrieved from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pointbiserialr.html

Scipy Contributors. (2020). Scipy stats.py [Computer software manual]. GitHub. Retrieved from https://github.com/scipy/scipy/blob/f2ef65dc7f00672496d7de6154744fee55ef95e9/scipy/stats/stats.py#L3781

Statista. (2020). *Output of the quick service restaurant (qsr) franchise industry in the united states from 2007 to 2020.* Retrieved from https://www.statista.com/statistics/252151/economic-output-of-the-us-quick-service-restaurant-franchise-industry/

Wikipedia. (2021). *List of cuisines.* Wikimedia Foundation. Retrieved from https://en.wikipedia.org/wiki/List_of_cuisines

Appendix

A Feature Engineering

This section describes 50 features based on Yelp business data with missing data handling by the highest frequency of each column. Any row with missing 'attributes' and 'hours' columns was previously removed. Then, the individual attribute was extracted from the 'attributes' column and convert to a list of features used in this study. The cuisine was extracted from the categories column by using the public list of cuisines and manually select it from existing the dataset itself (Wikipedia, 2021). Eventually, any categorical value of string or Boolean will be transformed to nominal numerical data.

Feature Name	**Description**	**Missing Data Handling**
ReviewCount	The number of reviewed received from customers	-
Dinner	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
Classy	Type of ambience which was extracted from 'Ambience' attribute.	Missing data were assigned to False value.
Lunch	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
HasTV	Boolean value whether the restaurant has television.	Missing data were assigned to True value.
DriveThru	Boolean value whether the restaurant has drive through service.	Missing data were assigned to False value.
Breakfast	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
RestaurantsPriceRange	The ranked variable with the value between 1 to 4. Value 4 means that the the restaurant offers a high-cost meal.	The row with missing data was dropped.
OpenCount	The number of opening restaurants operating in the same city and state.	-
Latenight	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
RestaurantsTableService	Boolean value whether the restaurant offers table service.	Missing data were assigned to False value.
Dessert	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
RestaurantsDelivery	Boolean value whether the restaurant offers delivery service.	Missing data were assigned to False value.
IsOpenWeekend	Boolean value whether the restaurant open either Saturday or Sunday or both.	-
NoiseLevel	The noise level of the restaurant with possible values of quiet, average, loud, and very loud.	Missing data were assigned to average value.
AgesAllowed	There are four groups of ages; allages, 18plus, 19plus and 21plus.	Missing data were assigned to allages value.
Brunch	The type of meal which was extracted from 'GoodForMeal' attribute.	Missing data were assigned to False value.
RestaurantsAttire	Type of outfit to wear at the restaurant such as casual, dressy, and formal.	Missing data were assigned to casual value.
GoodForKids	Boolean value whether the restaurant is good for kids.	Missing data were assigned to True value.
RestaurantsReservations	Boolean value whether the reservation is required.	Missing data were assigned to False value.
Casual	Type of ambience which was extracted from 'Ambience' attribute.	Missing data were assigned to False value.
DogsAllowed	Boolean value whether the restaurant allows dogs inside.	Missing data were assigned to False value.

Feature Name	**Description**	**Missing Data Handling**
City	The city of the restaurant operates. There are different 491 cities.	-
Touristy	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
Caters	Boolean value whether the restaurant offer catering service at a remote site.	Missing data were assigned to True value.
BikeParking	Boolean value whether the restaurant has bike parking.	Missing data were assigned to True value.
Smoking	There are three possible values - yes, no and outdoor.	Missing data were assigned to no value.
RestaurantsTakeOut	Boolean value whether the restaurant has take away service.	Missing data were assigned to True value.
WheelchairAccessible	Boolean value whether the restaurant is accessible for wheelchair.	Missing data were assigned to True value.
HappyHour	Boolean value whether the restaurant has happy hour.	Missing data were assigned to False value.
Divey	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
Cuisine	Type of cuisine extracted from ’category’ column. Each restaurant is filtered to have one cuisine. There are different 68 cuisines.	The row with missing data was dropped.
Alcohol	Type of alcohol available at the restaurant such as full bar, beer and wine or none.	Missing data were assigned to none value.
Upscale	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
OpenMusic	Boolean value whether the restaurant turn on the music or have musicians.	Missing data were assigned to False value.
BYOBCorkage	The restaurant may charge alcohol brought in by customers. There are three possible values: yes_free, yes_corkage, and no.	Missing data were assigned to no value.
WiFi	The type of WiFi service available in the restaurant with a possible value of none, free, and paid.	Missing data were assigned to no value.
State	The state of the restaurant operates. There are different 21 states.	-
Romantic	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
Intimate	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
Trendy	Type of ambience which was extracted from ’Ambience’ attribute.	Missing data were assigned to False value.
RestaurantsGoodForGroups	Boolean value whether the restaurant is good for groups.	Missing data were assigned to True value.
GoodForDancing	Boolean value whether the restaurant is good for dancing.	Missing data were assigned to False value.
Parking	Boolean value whether the restaurant has parking available for customers.	Missing data were assigned to False value.

Feature Name	Description	Missing Data Handling
CoatCheck	Boolean value whether the restaurant has coat check for customers.	Missing data were assigned to False value.
StarType	The type of business rating whether it is low or high based on value in 'stars' column. The business with neutral ratings of 3.0 was removed from the dataset.	-
BusinessAcceptsBitcoin	Boolean value whether the restaurant accepts Bitcoin for payment.	Missing data were assigned to False value.
OutdoorSeating	Boolean value whether the restaurant has outdoor seating space.	Missing data were assigned to False value.
Hipster	Type of ambience which was extracted from 'Ambience' attribute.	Missing data were assigned to False value.
BusinessAcceptsCreditCards	Boolean value whether the restaurant accepts credit card for payment.	Missing data were assigned to True value.

B Correlation Coefficient of insignificant features

The correlation coefficient and p-value of features classified as insignificant as p-value are greater than or equal to 0.05.

Feature Name	**Correlation Coefficient**	**P-value**
Latenight	0.013	0.055
RestaurantsTableService	0.011	0.087
Dessert	0.010	0.144
RestaurantsDelivery	0.006	0.192
IsOpenWeekend	0.008	0.222
NoiseLevel	0.008	0.234
AgesAllowed	0.008	0.247
Brunch	0.007	0.250
RestaurantsAttire	0.007	0.307
GoodForKids	0.006	0.316
RestaurantsReservations	0.006	0.342
Casual	0.006	0.358
DogsAllowed	0.005	0.445
City	0.005	0.476
Touristy	0.005	0.481
Caters	0.004	0.489
BikeParking	0.003	0.580
Smoking	0.003	0.673
RestaurantsTakeOut	0.003	0.681
WheelchairAccessible	0.002	0.705
HappyHour	0.002	0.708
Divey	0.002	0.724
Cuisine	0.002	0.759
Alcohol	0.002	0.802
Upscale	0.002	0.805
OpenMusic	0.002	0.811
BYOBCorkage	0.001	0.815
WiFi	0.001	0.872
State	0.001	0.892
Romantic	0.001	0.898
Intimate	0.001	0.901
Trendy	0.001	0.904
RestaurantsGoodForGroups	0.001	0.912
GoodForDancing	0.001	0.922
Parking	0.001	0.934
CoatCheck	0.0004	0.954
StarType	0.0003	0.967
BusinessAcceptsBitcoin	0.0002	0.969
OutdoorSeating	0.0002	0.971
Hipster	0.0001	0.987
BusinessAcceptsCreditCards	5.681e-07	0.999