Possible reason why establishments go out of business in Chicago

Rubin Daija
rubin.daija@epfl.ch

Rastislav Kováč
rastislav.kovac@epfl.ch

Sena Necla Çetin
sena.cetin@epfl.ch

Berk Mandıracıoğlu
berk.mandiracioglu@epfl.ch

Abstract

This project aims to analyze the Food inspections dataset of the city of Chicago and gain insight on reasons why businesses might be ceasing their activities. Different analysis is done based on the lifetime of the businesses, their violation types and their location. We further our investigation by utilizing other complementary datasets. We were able to conclude that the position of an establishment, the high number of restaurants clustered together, as well as their lifetime are indicating factors of their possible failure. Furthermore, we observed that violations, contrary to our belief, do not represent a decisive factor.

1 Introduction

Establishments continuously go out of business due to several factors. We used the provided data sets to gain some insight into the possible correlation between the results of different inspections done on establishments that have gone out of business, and the social-economic factors surrounding them. The city of Chicago is separated into 77 regions. The regions have various socioeconomic standings, as such we tried to use this fact to further investigate the reason of establishments going out of business.

2 Data Description

The open-source data set is provided by the city of Chicago at Kaggle [5]. It contains information resulting from inspections performed at different establishments in Chicago, starting from January 1st, 2010. The relevant features of this data set and their brief descriptions can be found in Table 1. Even though the time period is overlapping with our main data set only in two year, we were interested only in the ratios between the regions, which we assumed stays similar over the years.

Feature	Brief Description
DBA	'Doing business as', legal name of the establishment
Type of facility	e.g. restaurant, coffee shop, grocery store, etc.
Risk	1: High, 2: Medium, 3: Low
Latitude,Longitude	Precise location of the establishment
Inspection Type	Canvas, License, Complaint
Results	Pass, pass with conditions or fail
Violations	The violations observed during the inspection

Table 1: Description of the features of the data set.

To develop our analysis, we tried to use other supporting data sets, such as the Per Capita Income and the Affordable Rental Housing provided by the Chicago Data Portal [4, 6, 3].

The Per capita income data set provides various socioeconomic information about each region of Chicago. As we will later see, the most interesting features were the per capita income, percent of occupied housing units with more than one person per room (i.e. crowded housing) and hardship index, which that incorporates each of the six selected socioeconomic indicators.

3 Preprocessing

In this section we describe different thorough approaches to data preprocessing, namely data cleaning and how we identified each establishment.

3.1 Data Cleaning

The data cleaning process was an iterative one, as after each iterations other parts of the data had to be cleaned. Since there were different inspectors for different inspections, there were discrepancies on the data. The DBA name for big brand chains was written in different ways, e.g. KFC was written as 'KFC' or 'Kentucky Fried Chicken' or in different letter casing. Thus the first step that we took was to try and make all these points have the same value. Then we went back and put all the places that were in the same location together so that we would be able to identify any misspelling or words missing from the name.

The next feature that we cleaned was the facility type, where there would be facility type and the respective identifying value, just the value or in some cases a descriptive name which would be unique to that establishment. We looked at those facility types and tried to unify them in coherent groups.

The next feature that we tried to clean was the Inspection Type, as some of the inspectors would use either an inspection code or a more descriptive title value. Hence we went through the specification of the data set and tried to clean the data such that they would be only the inspection types mentioned in the specification. The ones that we could not fit to one group we simply dropped.

One more step that we tried to fill was some of the missing data such as the facility type. To do that, we looked at establishments with the same DBA name and if there would be one that had a value missing but had it mentioned in another point we would just fill it in.

3.2 Unique Establishment Identification

Our main goal was to investigate the reasons why an establishment goes out of business. During our investigation we encountered a hurdle in identifying business as one unique business could have multiple inspections and license numbers. Below we explain on how we dealt with this challenge.

We disregarded license numbers as one business could have multiple license numbers, either because they renewed it or because of inspection type. We decided that to identify a unique establishment we would use the DBA name, and the latitude and longitude. The DBA was used because that represents how the establishment is identified by the city of Chicago, thus if that stays the same it would mean that it is the same establishment. The latitude and longitude were used because they represent precise locations and as such ambiguities that can arise from street names were avoided. Furthermore, the location made it possible to differentiate between different establishment that are registered under the same brand name.

4 Analysis

In this section, we will dive into the analysis of the information we have gathered about our data sets. We will start with general exploratory analysis, the methodology will be examined and in the end we will take a deeper look into our findings.

4.1 Data Exploration

To have a better understanding of our data set, we start by doing exploratory data analysis. We first explore the number of unique facilities in the data set with respect to the procedure we explain in section 3.2. We find that there are 27,068 unique facilities.

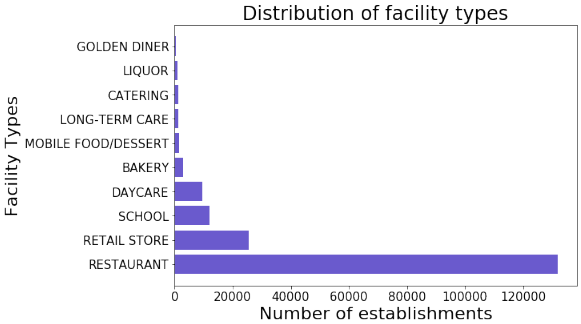

Figure 1: Distribution of facility types.

When we plotted the facility type distributions (Figure 1), we discovered that the distribution is heavily skewed, i.e. most of the establishments have an average risk score of 1. Likewise, we discovered the average risk score of the establishments to be skewed as well (Figure 2), where most of the establishments carry a high level of risk.

TODO Add general information about the dataset, distribution, outliers, continous, discrete data.

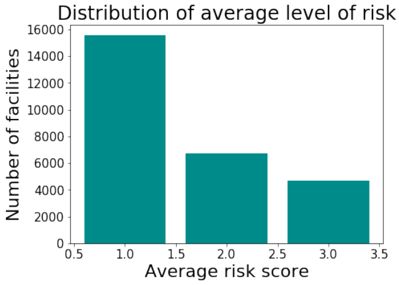

Figure 2: Distribution of average level of risk of the establishments.

4.2 Methodology

The aim of our methodology was to interpret the data to give an answer to our main question. We formulated a hypothesis that the main indicator of a failing establishment is its position and possibly the type of committed violations. One of the math tools we applied was the Pearson correlation coefficient which was used as a measure for linear relationships between two variables x and y, defined as follow:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (1)$$

To cross-check this correlation, we have also applied Spearman's correlation coefficient which can be used for variables that are not normally distributed and have a non-linear relationship. In order to get a deeper insight about the location, we have applied k-means++ algorithm [7], an augmented version of the standard clustering algorithm, which selects initial cluster centers for k-means clustering in a smart way to speed up convergence. In order to find good value for k centers, we have applied the Elbow method. However, since most of the regions did not have a clear elbow shape, we have decided to use Silhouette Score which reaches its global maximum at the optimal k.

4.3 Location Analysis

We augmented our original data set with appropriate region for each inspection we look at the ratio of the out of business establishments and total number of establishments in a given region.

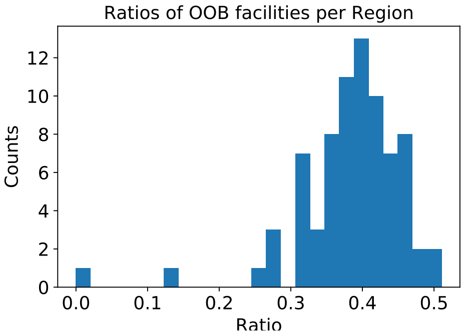

Figure 3: Distribution of OOB rations in regions

We observe that the distribution in Figure 3 is fairly similar to normal distribution, except the two outliers which could be disregarded give the low amount of inspections, 7 and 8 respectively. Together with the Per income database, we could define each location more precisely. Correlations are presented as follow:

	Crowded Housing	Per Capita Income	Hardship Index
Pearson	0.23	-0.11	0.17
Spearman	0.36	-0.33	0.33

Table 2: Socieconomic correlations.

Besides the anticipated negative correlation with income, the more money people earn, the less likely is it to go out of business, a correlation with Crowded housing has been found. Too many people in one house might be a good indicator of consumers who do not have enough resources to spend on food establishments, therefore we can observe a higher rate of closing down.

To further examine the locations, we try to test the hypothesis [2] that if there are too many restaurants at one place, they will start to fight over the customers and eventually someone will be forced to close down. To examine this, we have applied k++ means clustering algorithm on the establishments location within the region. K-Means starts by randomly defining k centroids. From there, it iteratively performs two tasks: Assign each data point to the closest corresponding centroid, using the standard Euclidean distance. For each centroid, calculate the mean of the values of all the points belonging to it. The mean value becomes the new value of the centroid. We looked at the clusters with the highest number of facilities and compared them to the OOB ration of a given region. We have found out that out of the 77 community areas, 45 highest density clusters had larger out of business ratio than its region, which is just above 50%. Therefore, too much competition can result in premature closure. One of the possible clustering of a region is shown in the Figure 4.

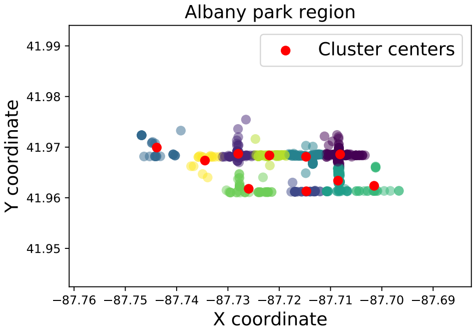

Figure 4: K++ clustering of establishments in Albany park region with 10 clusters

4.4 Violation type analysis

In the description of the data set in Kaggle, they provide the numbers of the violation types that are critical (1-14) and serious (14-29). We hypothesize that violation types 1-29 could have an influence on putting establishments out of business. To check this hypothesis, we plot the grouped histogram in Figure 5, where we can observe the difference in the distributions of the ratios of critical and serious violation types between the facilities that have gone out of business and the ones that have not.

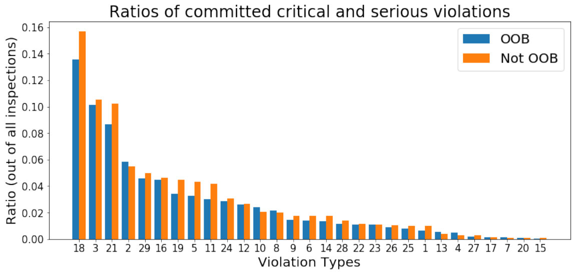

Figure 5: Ratios of critical and serious violation types (OOB vs. not OOB)

To gain further insight into the violation types that cause facilities to go out of business, we check the descriptions of these violations. To our surprise, the violations that have the biggest ratio differences between the two groups are the non-critical (i.e. serious) violations. We discover that critical violation types do not have an influence on making establishments go out of business.

4.5 Timeline analysis

Another area on where we looked in to data was on a time progress manner, as we believed that longer standing establishments can have a stronger foundation and customer basis. Thus we calculated the period of activity of all establishments that went out of business based on the first and last inspection. The start date will be around the time when the business would have started, since a license would need to be obtained prior to the opening, received during an inspection. Our results can be observed from Figure 6.

We can clearly observe that businesses that have a shorter lifetime, have a higher tendency to go out of business. The peak is at one which would mean that establishments stop functioning the most around the time on when they have one year since the opening. After that we see a decline as we move further. This result directly supports our belief that the shorter a business operating lifetime is the higher its chances are to for it to stop functioning.

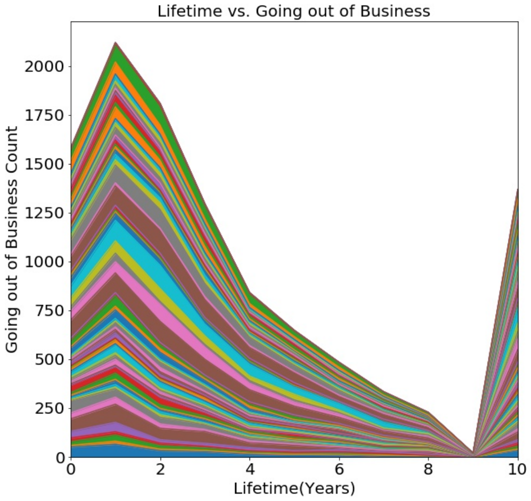

Figure 6: Stack area chart of lifetime versus going out of business count. Each stack level represents a region.

5 Conclusion

We were able to observe that the age of the establishments and location and play an important role to stay in business, in the city of Chicago. However, the correlation was not decisive, always below 0.5. Moreover, we were not able to show any significant correlation between the number, type of violations or risk type. This can be perhaps an incentive to create new set of indicators to properly reflect the seriousness of the current state of an establishment.

6 Future work and critique of our work

As mentioned before, social aspects have an impact on the success of a restaurant. Closely connected to this area is migration of citizens to and from a specific area. However, we could not find a well suited data about the number of people in each region during a specific time. Moreover, the proper answer to our question would explain the causal relationship, not just correlation. However, due to the hidden socioeconomic nature of a seemingly simple question is beyond data analysis of relatively small data. This dataset has been analyzed before [8, 5], even used to train a model to forecast future inspections. In the future we would like to try data prediction model such as Krishna et al [1].

References

[1] Alok Choaoudhry Amar Krishna Ankit Agarwal. “Predicting the outcome of startups:Less failure ,More Success”. In: *IEEE* ().

[2] Dalton Barker. *Does Chicago have too many restaurants?* URL: https://www.chicagobusiness.com/restaurants/does-chicago-have-too-many-restaurants.

[3] City of Chicago. *Affordable Rental Housing Developments*. URL: https://data.cityofchicago.org/Community-Economic-Development/Affordable-Rental-Housing-Developments/s6ha-ppgi.

[4] City of Chicago. *Food Inspection*. URL: https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5.

[5] City of Chicago. *Food Inspection Forecasting*. URL: https://chicago.github.io/food-inspections-evaluation/.

[6] City of Chicago. *Selected socioeconomic indicators in Chicago, 2008 – 2012*. URL: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2.

[7] Sergei Vassilvitskii David Arthur. “k-means++: The Advantages of Careful Seeding”. In: *IEEE* ().

[8] andMustafa Bilgic Vinesh Kannan Matthew A. Shapiro. “Hindsight Analysis of the Chicago Food Inspection Forecasting Model”. In: *IEEE* ().