Yelp Business Dataset

Technical University of Denmark

42578 Advanced Business Analytics

June 12, 2020

Kia Kafaei Yahyavi

Table of contents

1 Business

In the yelp Business data set a lot of insight can be gained about different businesses in different cities. In this notebook the focus was set on the city of Charlotte in the state of North Carolina, US and what potential investors may want to look for when investing in a business in Charlotte.

1.1 What makes a business successful in Charlotte?

As an investor looking to invest in a business in Charlotte several factors might play a role in the consideration of whether to buy a business or not. It is in the interest of the investor to invest in a business that will be successful and stay open. Since success can be measured in many ways and due to the limitations of the data set, the criteria of success will in this project be focused on whether the business is open or closed. It was determined that this factor is the best proxy for success since closed businesses - by definition - are unsuccessful businesses.

A comprehensive analysis of both open and closed businesses is performed in this project with the goal of understanding what factors play a role in whether a business stays open or closes.

One factor that played a role in the success of a business is the review count of the business. This is a KPI describing how many reviews the business has gotten. An intuition is that the good businesses has a lot of reviews while bad businesses don't have many reviews. The analysis showed the following.

Openness	review mean	review median
Open	37.87	10.00
Closed	25.82	8.50

Table 1: Mean and median review count for open and closed businesses in Charlotte.

The median review count is 15% smaller for closed businesses compared to open businesses. The mean review count is around 32% smaller for closed business compared to open businesses. This suggests that the number of reviews could have a significant influence on whether a business stays open or closes.

When looking at the categories of businesses that stays open compared to those who closes, one particular category strikes out, namely the category of *home services*. This can be seen from the histograms below, where home services is the 2'nd

most popular category among open businesses while being the 9'th most popular business among closed businesses. This large discrepancy suggests that from the 10 most popular business categories – all else being equal – the best business category for a potential investor to invest in would be home services [1].

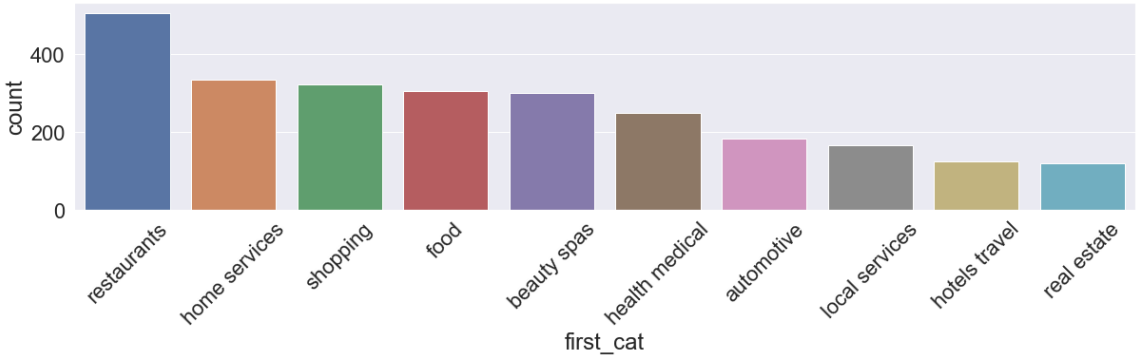

Figure 1: Histogram of open business categories

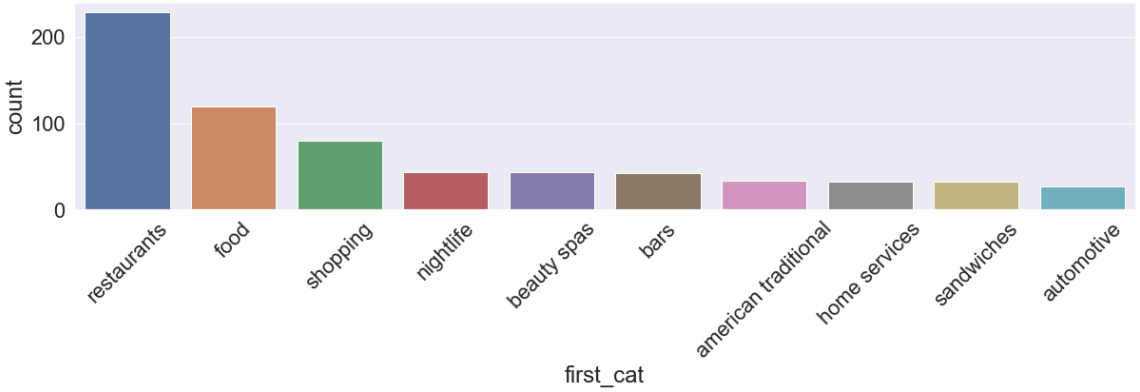

Figure 2: Histogram of closed business categories

The project further investigates the role of location in whether a business is open or closed. A statistical significance test (Chi squared test with a 95% significance level) is performed to check the correlation between these two variables, and on the basis of this test it was concluded that location in fact do play a role in how a business fares.

The last variable that was looked into was rating. Here the intuition was very straightforward: That good businesses get better ratings than bad businesses.
The analysis showed that there is a difference of 0.1 (from 3.51 for open businesses to

3.41 for closed businesses) between the average rating of open and closed businesses. This is a drop of around 3%. This might suggest that the 'rating' of a business, has somewhat of an influence in whether a business closes or stays open; though probably not a very large one. This challenges our initial intuition that "good" business has significantly better ratings than "bad" businesses; this simply doesn't seem to be the case.

1.2 Predictive analysis

In this notebook predictive analysis of two KPI's were performed, namely: review count and star rating. Although not as important as whether a business is actually open or closed, these two KPI's are still interesting to investigate and predict from a business perspective. Since they will be attractive KPI's to increase for any business owner.

The focus here was on restaurant related businesses, since many of the attributes given in the data was related to restaurants. Furthermore restaurants was the most frequent business category in Charlotte, making it the most relevant category to investigate.

Below are the error tables for 3 different classifiers/regressors for the star rating problem and review count problem respectively.

Models	(MSE) Test error
Baseline (mean pred)	0.91
Linear regression	0.69
Random Forest regression	0.66

Table 2: Review count: Test error for three regression models

Models	Accuracy	F1 score
Baseline (mean pred)	0.23	0.09
Linear regression	0.22	0.18
Random Forest regressor	0.21	0.23

Table 3: Rating classifier: Accuracy and F1 score for 3 classifiers

The feature importance analysis showed that by far the most significant feature for predicting the rating of a restaurant is the review count.
For predicting the review count the importance of features were more even, with a slight edge given to the star rating. Indicating that these two variables are somewhat correlated; increasing one KPI should lead to an increase in the other.

1.3 Conclusion and recommendations

1.3.1 Descriptive analysis

The main takeaways and recommendations - on the basis of the descriptive analysis -for investors looking to invest in businesses in the city of Charlotte, are the following:

- Look at the review count of the business, businesses with a review count less than 10 have a higher likelihood of closing compared to businesses with a higher review count.
- It seems that businesses that are related to home services have a higher chance of staying open, compared to other business categories such as restaurants.
- The rating of the business should not play as significant a role when looking to invest, especially if the review count for the business is high.

1.3.2 Predictive analysis

The predictive analysis showed that it is difficult to predict either the rating or the review count of a restaurant given the features included in the data set.

Neither of the models performed significantly better than their respective baseline models (the mean prediction). This makes giving recommendations to restaurant owners on the basis of these predictive models, futile.

References

[1] Stanislav Borysov. *Lecture 3: Text Analytics*.