SCB Business Analytics Competition – Final Report

Fardin Ahsan (Team Leader) - [fxa8225@g.rit.edu](mailto:fxa8225@g.rit.edu)

Dren Gara - [dg9901@rit.edu](mailto:dg9901@rit.edu)

Harshit Kumar - [hxk8955@rit.edu](mailto:hxk8955@rit.edu)

**Executive summary**

**Data preparation**

Our data preparation process can be broken down into 3 main steps; Data cleaning, Web-scraping Scraping and Data Augmentation. We used Python for all the data processing/cleaning.

**Data Cleaning and processing**

Before doing any sort of analysis we had to clean up the data in a way would allow us to treat the entire dataset as one and format it such that we would segment off parts of it to further analyze. For example, keeping NULL values in the dataset will raise an error when we try to apply any sort of function to the fields in the dataset. (Appendix, Data cleaning)

Here are the following steps we did to clean up the two given datasets (‘Pharmacies in ROC’ and ‘Pharmacy Sales’).

* Drop the ['telephone','address1','address2','state','website','zipcode4'] from ‘Pharmacies in ROC’ columns as we decided they wouldn’t meaningfully help with any analysis.
* Drop all stores in ‘Pharmacies in ROC’ that did not have a store number. Because the Store number is the primary key of both datasets, without a store number there is no way to connect the two datasets.
* Added new fields: [‘total\_cost’,’profit’].
* Merged the two datasets (outwards merge) on ‘Store Number’ as the primary key.
* Dropped all stores that did not have any sales in 2021, assuming that those shops closed down.
* Dropped all sales before 2018 because some shops only had sales in specific time periods, in order to do any aggregate analysis, we had to normalize the entries in some way, normalizing it to a specific time period (3 years chosen to not lose too much of the dataset) was the trade off that made most sense to us.
* Coerced the correct datatype. Everything in both the datasets were in the string data format. In order to do analysis, we had to change some fields to numeric and datetime formats. This was done purely to make the dataset work with the software and aid in our analysis, for example changing zip codes to numeric was not necessary.

**Web – Scraping**

We scraped demographic data from ‘www.zipdatamaps.com’. We did this to find out which of demographic factors have a strong correlation with the average profitability of a zip code.

Some of the categorical data had to be encoded into numbers, such as school grades, {Good = 3, Average = 2, Poor = 1}, this allows the software to find correlations in categorical data with numerical data.

We wrote a python script to scrape the data from the website. (Appendix, Web-Scraping)

**Data Augmentation**

After aggregating the merged dataset (Two initial datasets) and grouping by zip code. We merged it with the data we scraped from the web, with zip code as the primary key. This allowed us to generate a correlation matrix of demographic factors and a zip code’s average profitability. ­­

**Data analysis**

**Why did we normalize the data on time?**

On our initial attempt at exploratory data analysis, we noticed that not all shops had sales in all periods of times, for example shop A might have had sales from 2015 to 2017, whilst shop B might have had sales from 2014 to 2020, this means, any sort of aggregate function would be worthless, without some sort of normalization. Normalizing based on number of sales doesn’t remove shops that might have closed down. We thought of applying a normalizing method such as a linear mixed effects model, but we realized that we don’t have enough data for the model to actually produce an accurate normalization. Thus, we decided to normalize on time as that it is the method that requires the least tinkering with the data, and thus the least possibility of removing any data that we otherwise should have kept. Moreover, working with recent years might have the added benefit of capturing more recent market trends better. 2018 seemed to be the best cutoff as it required sacrificing the least number of shops whilst including enough years to offset the erratic trends that might have shown up during the pandemic, we didn’t want the data to be influenced too much by the pandemic as that might make our analysis and recommendations dated once the pandemic is over.

**Methodology and philosophy**

The target of all the analysis was to find the better (We decided better means most profitable, as the fundamental goal of a business is to make a profit and that is the safest assumption.) option of either building a pharmacy or buying a struggling pharmacy. And to find out what predicts the success of a pharmacy. The data in the dataset was not sufficient to answer either the latter or the former. The solutions to those problems were making some educated guesses about the cost of running a business in Monroe County (To compare profitability of shops), and scraping demographic data from ‘www.zipdatamaps.com’ to find correlations between demographic markers and profitability. Most of the analysis we did was using a ‘group by’ method. Which means that when we wanted to find something with respect to something else (for example average profit per zip code), we grouped the merged dataset based on that field (zip code), then applied an aggregation function (such as count, sum, mean, cumulative sum), then accessed the field (average profit) we need in the grouped and aggregated table.

To answer the first question as to which factors, drive success of a business and predicts future success. We decided that the best level of analysis was analyzing on a zip code basis. City wide analysis would produce trends that are not precise enough and street wise analysis would result in statistical “overfitting” where our models might capture patterns that are not necessarily there. Thus, going by zip codes meant the ideal tradeoff between too much and too little granularity given the size of our dataset. To find what makes a specific zip code profitable, we looked at the correlations of various demographic data with the average profitability of that zip code, which is nothing more than the total profits in that zip code divided by the number of shops in that zip code. Coming up with a product portfolio simply consisted of finding the top 500 overall profitable products and sorting them in ascending order (set A). Doing the same thing for the recommend shop (set B). Then replacing products that appeared in set A but not set B, set A minus set B are the recommended new products. We used the same method for selecting vendors.

As for recommending a between building a new store or buying a currently existing struggling store, which store in either case is informed by the data analysis. For building it would have to be a new store in a profitable zip code or buying an unprofitable store in an otherwise profitable zip code. The answer to both those questions is informed by secondary research on the cost of running a business in that zip code and buying a store in that zip code. Whichever combination is more profitable in the long term is the better option.

**Results**

**What drives the success of a business?**

In order to find determine what drives a success of a business other than its internal business practices, we looked at how demographics factors correlated with the average profitability of the of the zip code.

We found that the strongest driver for zip code profitability was its ‘Average Adjusted Gross Income’ which had a correlation of 61% with average profitability, followed by ‘School Test Performance’ (correlation = 35%). It is also worth noting that profitability was also anti-correlated (-26%) with unemployment rate. (Appendix, Correlation matrix)

Therefore, the best location to open a new business would be a business in a zip code with a high average adjusted income, high school performance, and low unemployment rate. ‘14618’ was the most profitable zip code by a large amount (12 times the average). And also, was consistently within the top 5 of all of the positively correlated metrics. (Appendix, Zip code Choropleth map)

**Buying a struggling pharmacy vs building a new one.**