Are coffee shops a good predictor of gentrification?

Isgandar Asgarov

July 5th, 2021

1 Introduction

1.1 Background

Gentrification is a process in which a poor area (as of a city) experiences an influx of middle-class or wealthy people who renovate and rebuild homes and businesses and which often results in an increase in property values and the displacement of earlier, usually poorer residents[1]. This process is extremely controversial, because gentrification changes the social and economic environment of an area, frequently resulting in the displacement of poorer citizens, which furthermore often belong to various social minorities. The phenomenon of gentrification is often measured through changes in neighbourhood features, in housing, and in the composition of residents over a period of time[2], however there are also less data-driven and more anecdotal indicators, such as the influx of coffee shops[3]. This study aims to find a link between the number of certain venue types in the neighborhood and gentrification.

2 Data acquisition

The data necessary for this analysis included the names and coordinates of the neighborhoods, the number of venues of various types within the neighborhood, and the status of the neighborhood with regards to gentrification.

The first component, the names of the neighborhoods were scraped off of the Wikipedia page for the postal codes of Canada[4]. After processing the string data into a pandas dataframe, the names were sent to the Google Geocoding API. Returned coordinates were then passed to the FourSquares API in a script provided by IBM to acquire the venues in the surrounding area. The only information about venues to be used were their type. The data on the types of venues was then one-hot encoded to be fit for use in a regression. Since the FourSquares API returned multiple venues per neighborhood, to facilitate analysis dummy indicators for venue types were averaged and summed in two different data frames.

Determining the gentrification status of a neighborhood was one of the main challenges at this stage in the study. Due to the limited scope of the analysis, to avoid excessive time wastage the most direct method available was chosen. First, the coordinates of the neighborhoods obtained in the early stages of the study were visualised on maps using the Folium library in Python. The maps were cross-referenced with the GENUINE tool created by Statistics Canada. The gentrified neighborhoods of Vancouver and Toronto

[1] Definition of “Gentrification”, the Merriam-Webster dictionary

[2] “Gentrification, Urban Interventions and Equity (GENUINE): A map-based gentrification tool for Canadian metropolitan areas”, Statistics Canada

[3] “A brief history of the coffee shop as a symbol for gentrification”, Pacific Standard

[4] “List of postal codes of Canada: V”, Wikipedia

were then manually selected.

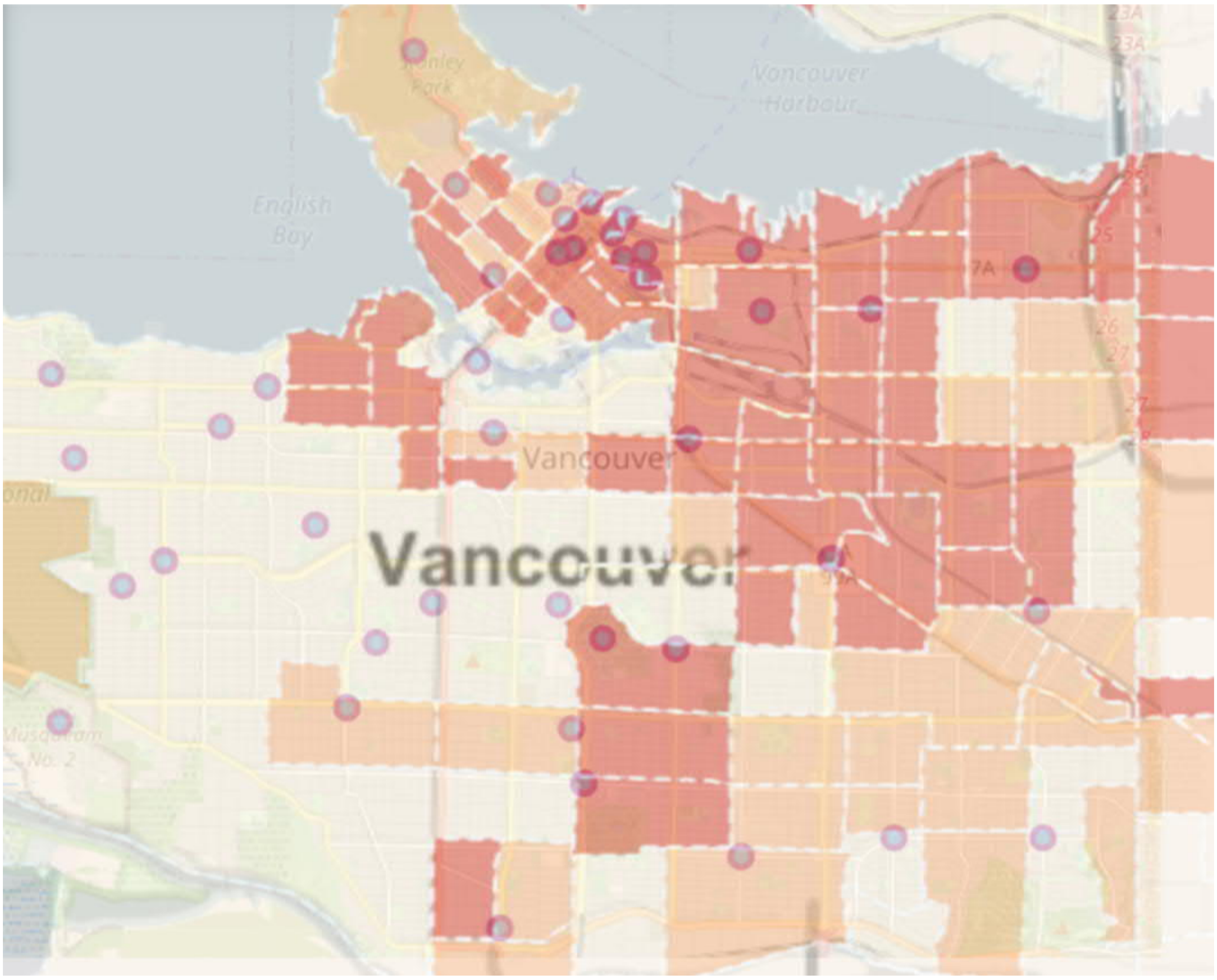


Figure 1: Overlapping maps of the neighborhoods in the dataset and the GENUINE tool for Vancouver

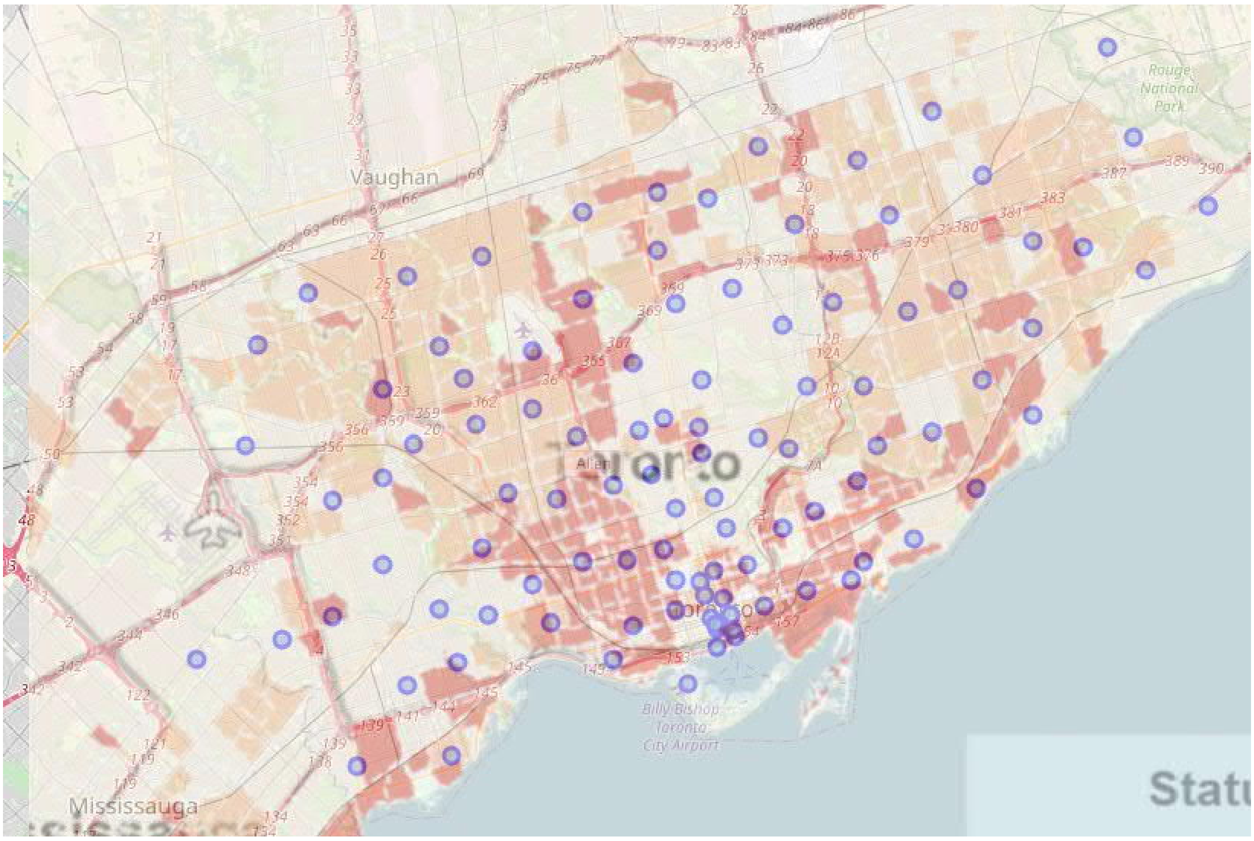


Figure 2: Overlapping maps of the neighborhoods in the dataset and the GENUINE tool for Toronto

3 Analysis

It was determined that the most appropriate statistical tool for classification as in the presented research question is a logistic regression. To create the various regressions, scikit-learn machine learning library for Python was utilized. In all models, a 75/25 train/test split was applied to the Vancouver data.

3.1 Approach 1

The first approached was the simplest: an attempt to establish a causational link between coffee shops, the anecdotal indicators, and gentrification. The model was fitted using the Newton's method of cost minimization.

Despite the relatively promising Jaccard score of 0.83, the model, in fact, utterly failed to identify a single gentrified neighborhood, returning 3 false negatives and 0 true positives.

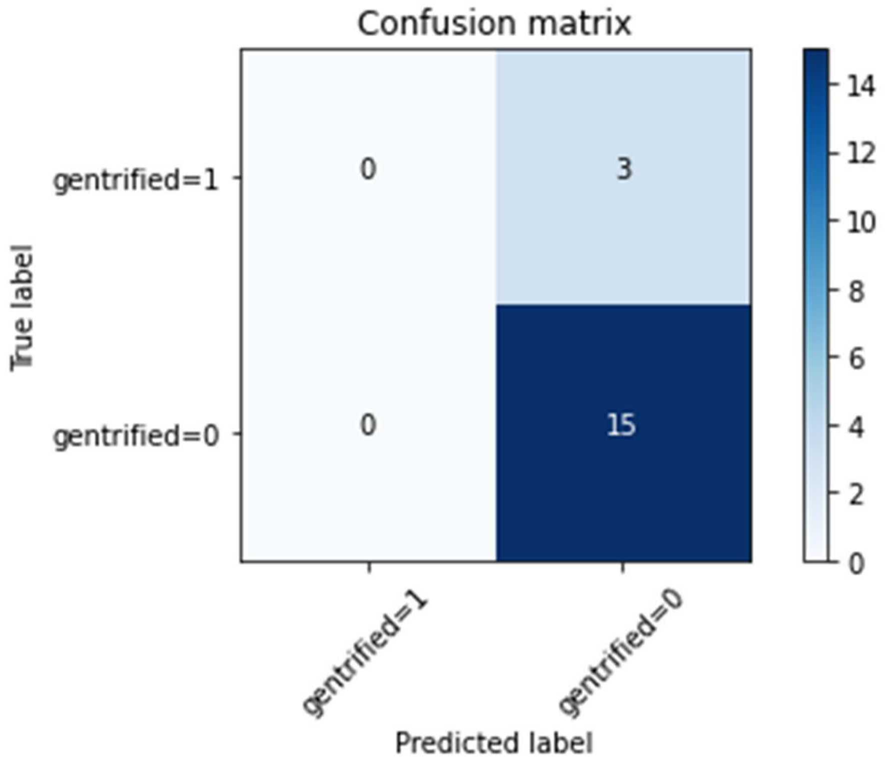


Figure 3: The confusion matrix for the test set of the logistic regression fitted in approach 1

When the model fitted to the Vancouver testing set was applied to the Toronto data set, the results were similarly disappointing. None of the 26 identified gentrified neighborhoods were picked up by the regression.

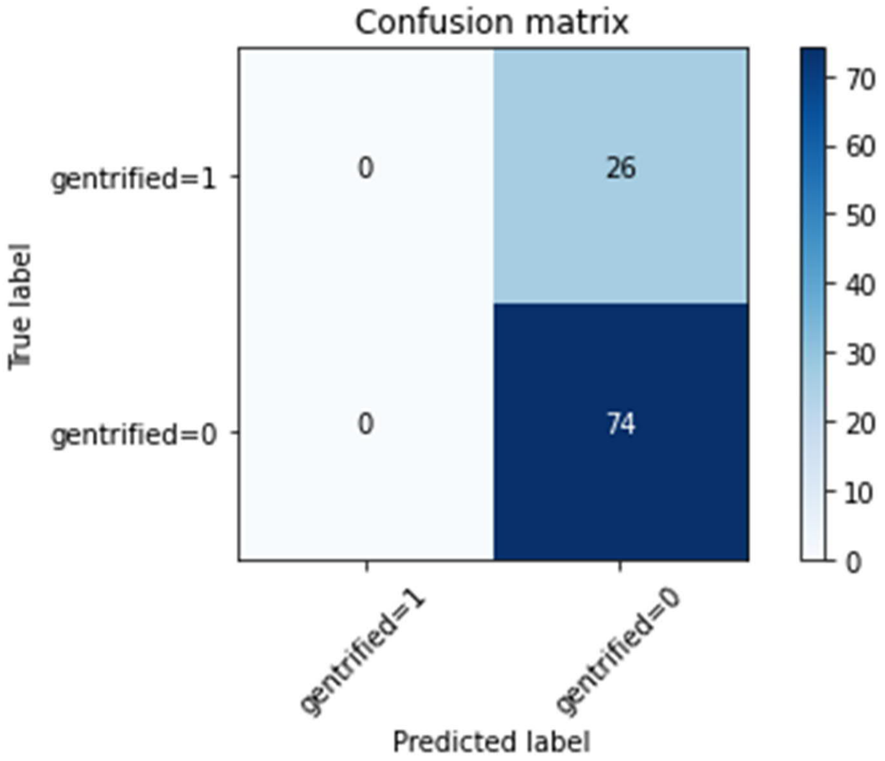


Figure 4: The confusion matrix for the Toronto set of the logistic regression fitted in approach 1

3.2 Approach 2

In the second approach, all venues types were used in the regression. To compensate for a larger volume of data, the Library for Large Linear Classification was used as the cost minimizing function.

The results of this approach were similarly inauspicious. With the Jaccard score of 0.78 and logarithmic loss of 0.66, the model nonetheless again falsely pegged the gentrified neighborhoods as non-gentrified.

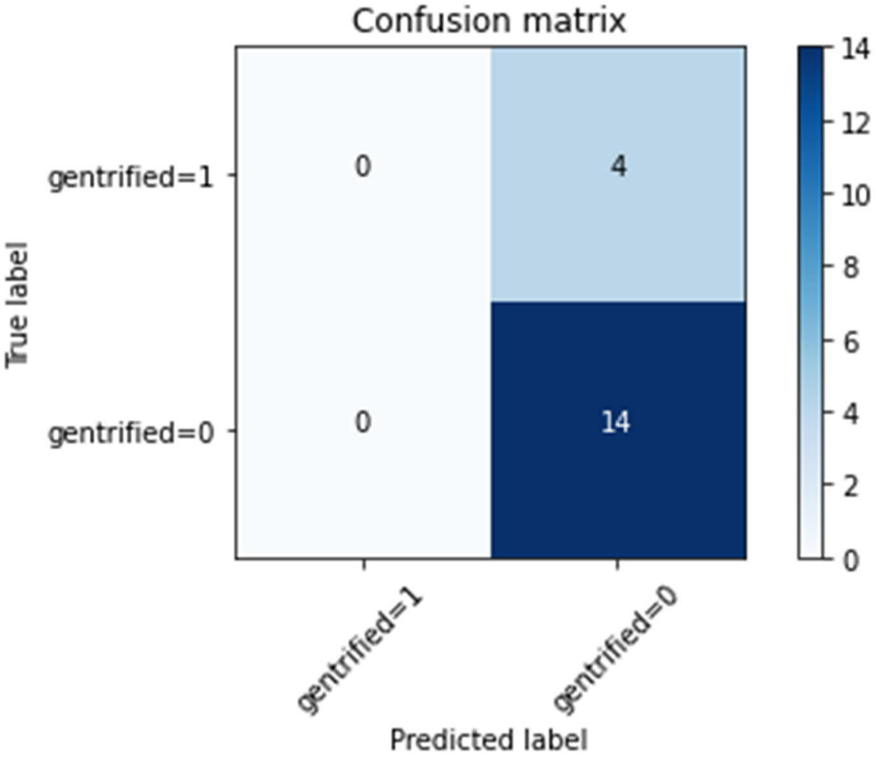


Figure 5: The confusion matrix for the test set of the logistic regression fitted in approach 2

Due to the differences in venue types, this model could not be fitted to the Toronto dataset.

3.3 Approach 3

According to a paper titled paper "Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change", "the opening of a Starbucks — and cafes more generally — is a leading indicator of gentrification, and is associated with an increase in local housing prices of .5%. Gentrifying neighborhoods tend to spawn a growing number of grocery stores, cafes, restaurants, and bars"[5]. Based on this assessment, in the third approach the venue types used were limited to restaurants, bars, cafes and similar. The exclusion of other venue types did not, however, produce a difference.

[5] "Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change", By Edward L. Glaeser, Hyunjin Kim, and Michael Luca

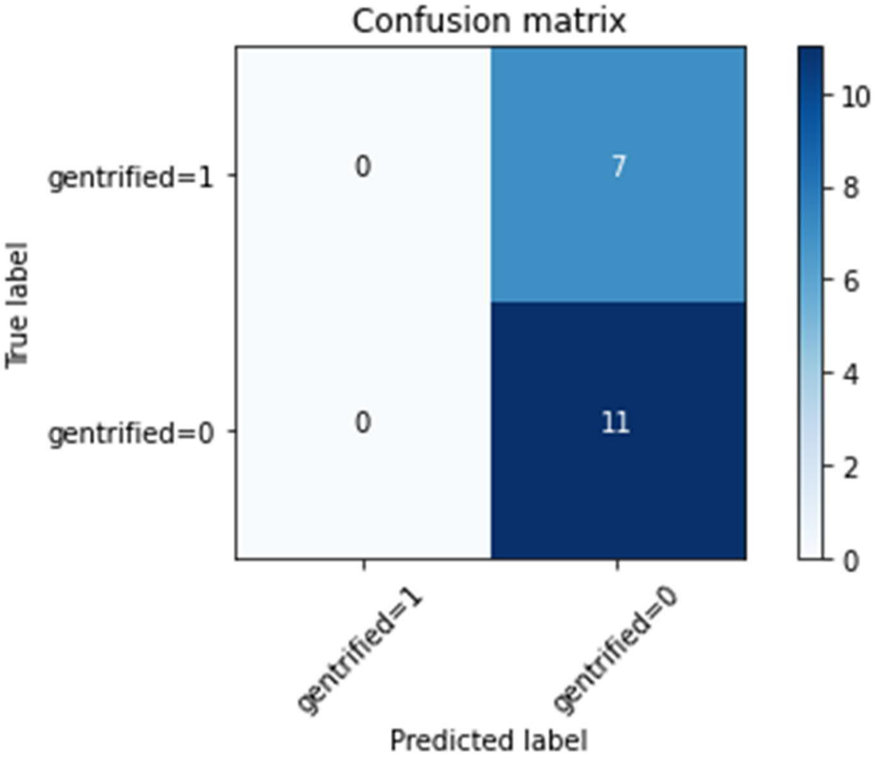


Figure 6: The confusion matrix for the test set of the logistic regression fitted in approach 3

The model once again returned not a single true positive. Retraining the model for the Toronto dataset (some of the venue types chosen for the analysis were not present in the Toronto dataset, rendering the usage of the already trained model impossible) similarly failed to produce results.

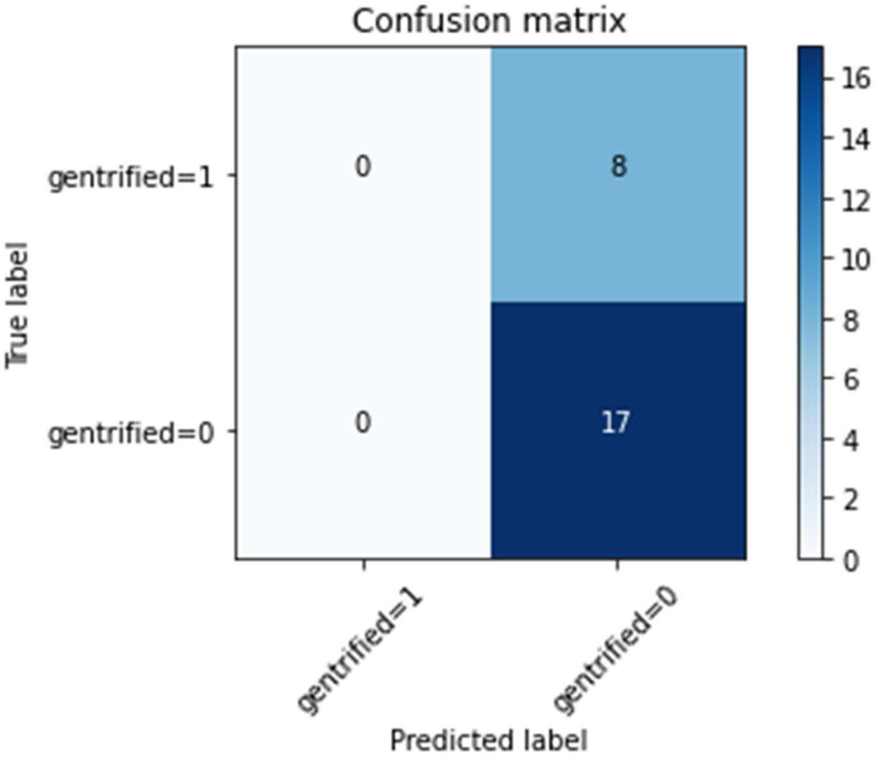
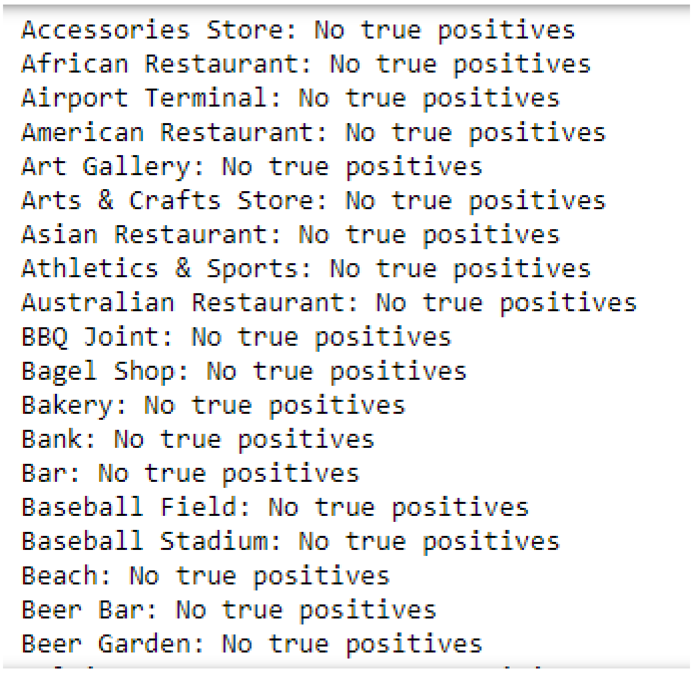


Figure 7: The confusion matrix for the Toronto set of the logistic regression fitted in approach 3

3.4 Approach 4

In this approach, a looping script was written to model a regression based on data on each individual venue type by neighborhood. After running the process several times, the outcome was the same: not a single regression based on any venue type produced any true positives.



```
Accessories Store: No true positives
African Restaurant: No true positives
Airport Terminal: No true positives
American Restaurant: No true positives
Art Gallery: No true positives
Arts & Crafts Store: No true positives
Asian Restaurant: No true positives
Athletics & Sports: No true positives
Australian Restaurant: No true positives
BBQ Joint: No true positives
Bagel Shop: No true positives
Bakery: No true positives
Bank: No true positives
Bar: No true positives
Baseball Field: No true positives
Baseball Stadium: No true positives
Beach: No true positives
Beer Bar: No true positives
Beer Garden: No true positives
```

Figure 8: An excerpt of the output of the looping script created in the approach 4

3.5 Approach 5

The fifth and final approach was based not simply on the number of coffee shops, but their relative popularity in the neighborhood. A script was written to generate a dataset with the 3 most popular venue types in each neighborhood. A dummy variable was then created for neighborhoods in which coffee shops were either in any of the 3 top rankings or the top 2 spots. For both cases, the models failed to identify gentrified neighborhoods.

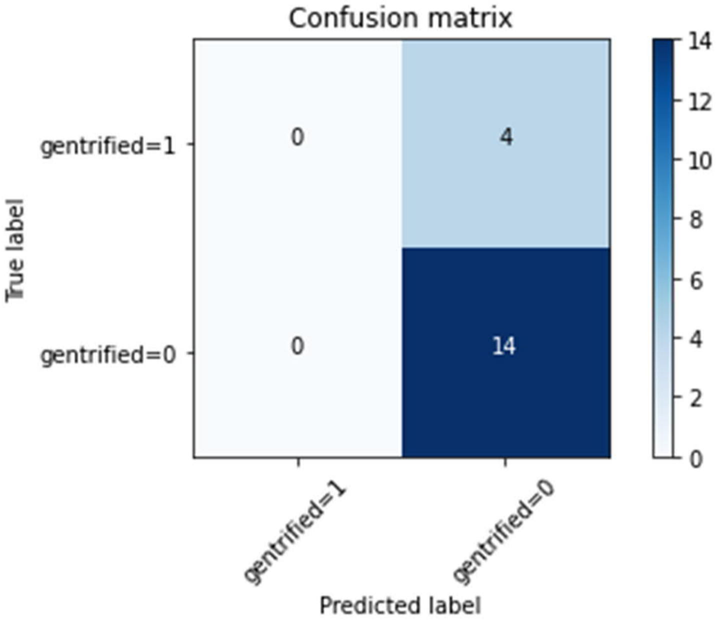


Figure 9: The confusion matrix for the test set of the logistic regression fitted in approach 5, for neighborhoods in which coffee shops are the most, the second most or the third most popular venues

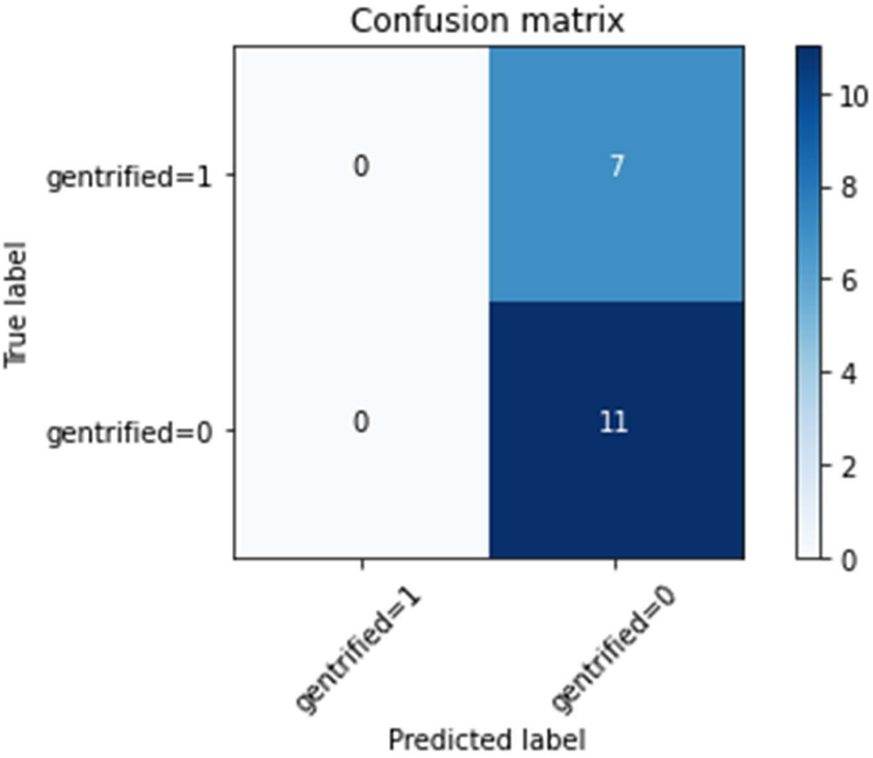


Figure 10: The confusion matrix for the test set of the logistic regression fitted in approach 5, for neighborhoods in which coffee shops are the most or the second most popular venues

4 Results

As is apparent from the above, this analysis failed to find a link between a number of certain venue types in the neighborhood and gentrification.

5 Discussion

While the results of this study imply no link between venue types and gentrification, it is important to keep in mind the limited scope of the research. First, it should be noted that the data on gentrified neighborhoods was taken as is based on one of many potential measurements of gentrification. Ideally, an analysis like this would look into the links between other indicators of gentrification (property prises, population composition, displacement of the poor citizens), but the resources necessary were not available for this study. Additionally, the small dataset is another hindrance to accurate analysis, which was further aggravated by the limited nature of the data (it being simply the number of venues, not separated by average prices of similar distinguishing factors). The fact that all models consistently flagged false negatives rather than a random number of false positives and false negatives might indicate that the model was overfitted or that the sample size or the nature of the data warped the model.

6 Conclusion

To conclude, while this study failed to identify a link between gentrification and recreational venues in an area, further research is necessary before any conclusive statements can be made.