Predicting Gentrification in Mexico City using Neural Networks

1st Leon Palafox
Facultad de Ingeniería.
Universidad Panamericana
Mexico City, Mexico
lpalafox@up.edu.mx

2nd Pedro Ortiz-Monasterio
Facultad de Ingeniería
Universidad Panamericana
Mexico City, Mexico
portizm@up.edu.mx

***Abstract*—Gentrification is a process that affects millions of people every year. In this process, high-income residents replace low-income residents in neighborhoods near city centers and business centers. The gentrification process drives low-income people to find unfamiliar places to live, which in and of itself brings multiple social problems about housing, transportation, and schooling.**

Many studies in social geography have looked at the onset of gentrification in a neighborhood. Yet, studying gentrification in a single neighborhood is a slow process that requires expertise about the different elements that can cause it, like housing prices, businesses in the neighborhood, and other social elements. Each of these elements is idiosyncratic to each country and area.

In this work, we mix the predictive power of Neural Networks with an Interpretability Method called LIME, which helps understand which factors are driving the classification given the data and a trained classifier. With this, we expect to have an overall model that gains a deeper understanding of which effects drive gentrification in different settings.

***Index Terms*—Gentrification, Neural Networks, Interpretability**

I. Introduction

In recent years, from social networks to e-commerce sites, we have plenty of new data to analyze and explore the consumers from different points of view: psychological, sociological, and geographical. However, it is still a complex task to match all these data points to individual people.

Nowadays, we have access to geographical data from the different people inhabiting a city. Like census data, commuting data from Uber [1], living data from AirBnb [2] and the cost by square meter in some of the largest cities in the world [3].

However, given the plethora of geographical data available, not enough work has been done to address social issues pertaining housing. One of the most pervasive problems in our present society is the problem of gentrification. Gentrification is defined as the process in which people with high-income geographically displace people with lower income. This displacement has many different causes, from economic ones to political ones. Gentrification comes along with the re-purposing of old neighborhoods to make them more alluring to the new tenants in the nearby residences. New businesses open, new facilities get added, and the extra tax revenue helps increase the real state value of neighborhoods [4].

Yet Gentrification brings with it the displacement of people, which puts pressure in other areas of the city, and, as consequence, to the infrastructure systems of the city, by having to move larger numbers of people from their new living places to the business centers where they usually work [5], [6].

Mexico City is an example where the effects of gentrification are being felt [7]. Mexico City is a bustling business center in Latin America, and it has a constant influx of affluent people looking to live in the richer areas of the city, near the downtown and the entertainment centers. Yet, this has had as a result the displacement of people, who, in turn, are forced to live in areas far away from the center of the city.

There have been many models to try and predict whether a neighborhood will become gentrified [8], the main problem with these models is that they require expert knowledge to label the data. Furthermore, models trained in one city cannot be used in modeling gentrification in other cities. Most studies in gentrification seldom use Machine Learning to predict the onset of the phenomenon. One of the main reasons is that social scientists lack the training to deploy these kinds of models in real data, and Machine Learning experts lack the experience to understand and tune the model to better mimic the reality.

There has been some work to try and use Machine Learning to predict gentrification trends in cities like London [8] and New Orleans [9],however, both London and New Orleans are vastly different from Mexico City, for one, Mexico City does not have a waterfront, and in London and New Orleans, waterfronts are areas primed for gentrification.

We already have presented an approach using classification trees to predict gentrification in Mexico City [10], however, in that work, many of the effects and expert knowledge were lacking, since we follow an approach where we did not account for many social effects.

In this work we use a Neural Network, and we are using a data-set that has been vetted by a social scientist, who classified multiple neighborhoods of Mexico City in gentrifiable or non-gentrifiable neighborhoods, which are better labels than a stark gentrified and non-gentrified.

To address the social issue of interpretability, we also use LIME [11], a model to help us interpret the different factors that affect a classifier when it does its classification.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Using LIME, we can use the predictive power of the Neural Network and try and analyze particular neighborhoods in Mexico City. And which factors have particularly helped those neighborhoods to get gentrified.

II. Methodology

In this work, we extracted data from various sources of information, once we had the data, we trained a Neural Network to build the classifier, and we use LIME to evaluate the importance of the features for different neighborhoods in Mexico City.

We describe the system in Figure 1:

- First we merge the different data sources according to the indexes by neighborhood given by the Mexican National Institute of Statistics
- We clean the data and drop columns that we do not consider important for the classification, such as the name of the neighborhood.
- We train a Neural Network with the data and have a black box classifier capable of predicting our different labels.
- We use LIME to analyze which variables are indicators of the gentrification process.

A. LIME

Interpretability is one of the key features of a gentrification-prediction algorithm. Since it is particularly important for public servants to understand the specific reasons that onset the gentrification phenomenon. Is impossible to have interpretability with a traditional Neural Network since the network's classification power depends of the networks' capability of abstracting and combining the features of a dataset.

For the sake of interpretability, we use LIME. In LIME, we can analyze neighborhood by neighborhood and investigate the risk of said areas for getting gentrified. Thus, we can be sure that we can supply better policy insights when it comes to neighborhoods and data.

LIME is an algorithm that perturbs data points and analyzes the effects that the perturbations have in the resulting label. The perturbation allows LIME to infer for each data point which feature explains better the variations in the labels. Once the process is done, we get factors which show us how important was each feature to achieve the label that we are giving to it.

Generally, the perturbations are regressed to the true labels via a linear model, which can be linear regression or another stochastic gradient descent based model.

In this case we use our architecture of NN+Lime to have a system that can help us analyze which factors drive gentrification in a city like Mexico City.

B. Data Sources

As we mentioned before, we limit the geographical coverage of this work to Mexico City, the most populated city in Mexico. Mexico City has enough studies done in gentrification and plenty of data associated with commerce and the social fabric. This availability of data will allow us to model the city in the most exact way.

TABLE I
Gentrified neighborhoods

1.Alamos	9. Irrigacion
2.Centro	10. Juarez
3.Condesa	11.Obrera
4.Cuauhtemoc	12. Roma Norte
5.Doctores	13. Roma Sur
6.Escandon	14. San Rafael
7.Hipdromo	15. Santa Maria la Ribera
8.Hipodromo de la Condesa	16. Tabacalera

The figure 2 shows the 1,436 neighborhoods of the CDMX. For this work, we focused in 251 central neighborhoods which are prime candidates for gentrification.

The data that we use was collected in 2000, 2010 and 2016, which mostly comes from the Population and Housing Censuses conducted by INEGI, Mexico's geographical institute.

Table I shows the neighborhoods that are defined as gentrified in the present work ([12], [7]).

Figure 3 shows the location of the gentrified neighborhoods on a map. The gentrified neighborhoods are in the center of Mexico City.

C. Data

The different data sources we used are referenced in the Figure 4.

- Inventario Nacional de Viviendas 2016: National inventory of living quarters
- Censo de Población y Vivienda: national census in Mexico
- SCINCE: System to query the census information
- DENUE: National directory of economic activities
- Softec: Private database with information regarding the cost of the quare meter.
- Shapefiles: Different shapefiles for the geographical units in Mexico City

D. Data Processing

Since the data that we used comes from a variety of sources, we had to undergo a heavy process of pre-processing to be able to match the different tables in terms of having the same locations coincide in the various sources.

In some cases, same neighborhoods had different names, and we did not have a unification code for them. In some other, for temporal data, neighborhoods ceased to exist, or altogether new neighborhoods were created in the years between the different census.

E. Labels

One of the essential elements for an effective gentrification algorithm is the correct assignments of the labels for the classifier. For this work, we assigned three different labels:

- Non Gentrifiable: Means the neighborhood cannot be gentrified. This can happen due to several issues. Once

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

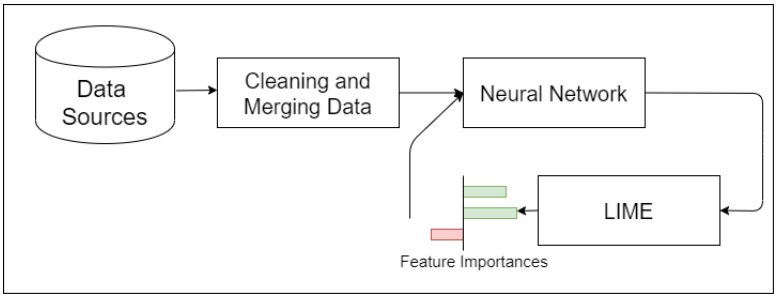

Fig. 1. Flow diagram of the implemented system

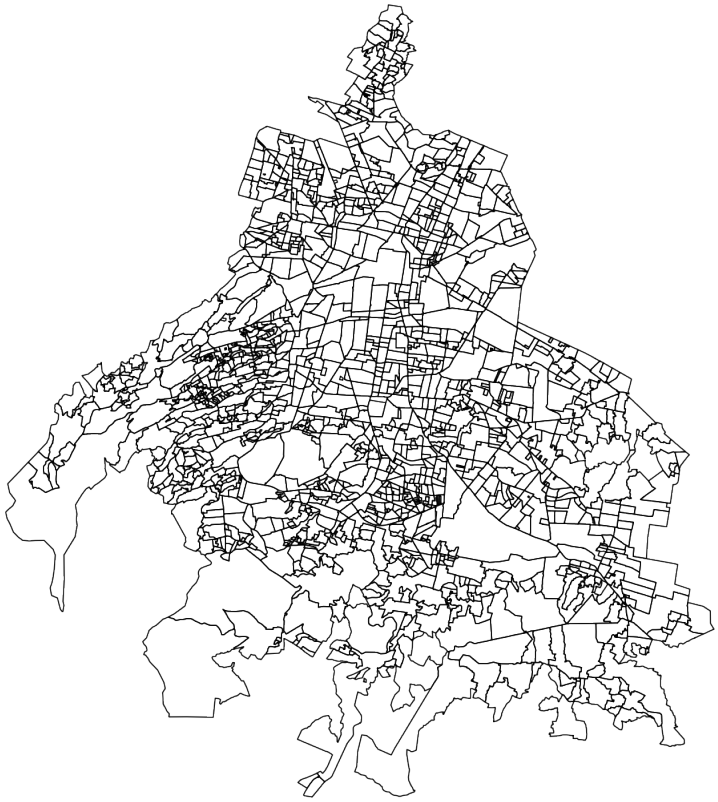

Fig. 2. CDMX neighborhoods

being that the neighborhood is already affluent, thus the social displacement that comes with gentrification is impossible.

- Non-Gentrifying: Means the neighborhood can be gentrified but is not currently undergoing the process of gentrification. These neighborhoods are usually near the center of the city with low-income people.
- Gentrifying: It means that the neighborhood is in gentrification. This is the most important label since is the one that may be used to create affordable housing policies to prevent people from being displaced.

To set the labels, we had the help of experts in gentrification, which helped us identify the current condition of the neighborhoods, and in doing so, set the labels in the most precise way.

III. Experiments

To test the algorithms, we tried different NNs architectures, after multiple iterations with different architectures, we got to the network defined in Figure 5. The activation functions for the first two layers are sigmoid and the activation for the output layers is a softmax function since the problem is a multilabel problem.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

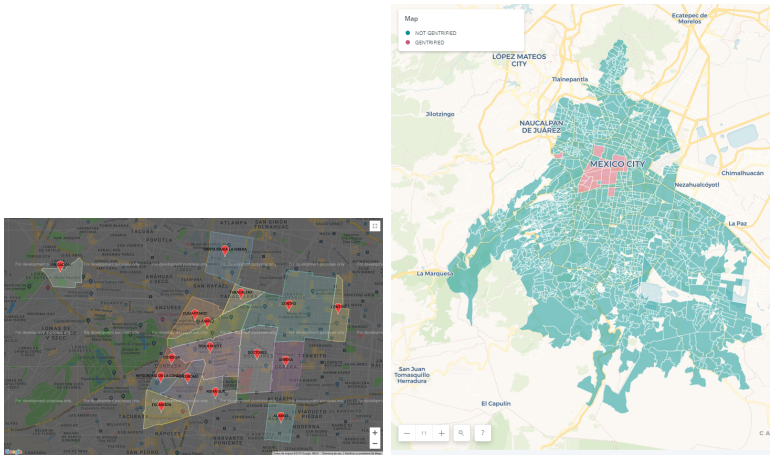

Fig. 3. Gentrified neighborhoods

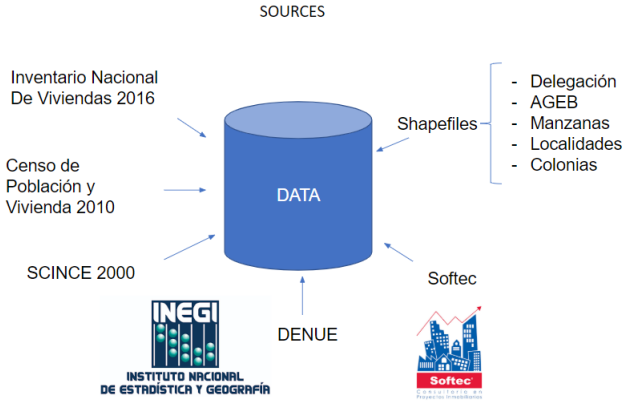

Fig. 4. Data Sources

We ran a 3-Fold Cross validation to obtain the best network architecture, and the experiments, accelerated with GPU take 3 seconds to run.

Finally, we chose the model with the highest accuracy and using that model, we used the trained classifier with LIME and then we tested multiple neighborhoods to analyze which variables are conducive to gentrification in any area.

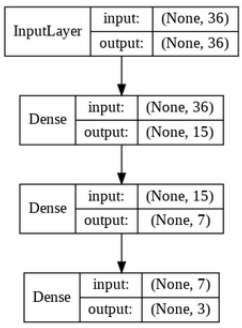

Fig. 5. Network Architecture

IV. Results

The results for the best neural network architecture were the following:

Name	Value
Training time	2 s
Accuracy	0.75
loss	0.5

In the Figure 6 we present the optimization of the NN, due to the number of parameters and the ambiguity of the different classes, we required 200 epochs with 50 sized train batches each. We tried increasing the epochs but did not find any noticeable results. We also see how the loss decreases.

In Figure 7 we present the output of the LIME result in a single example, after testing multiple instances we see that the model is overly sensitive to prices. Which is in line with the results that we presented before in [13]. The model is also somehow sensitive to the kind of businesses that are in the area, which shows that gentrification is also driven by the variety of business that start opening in the area.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

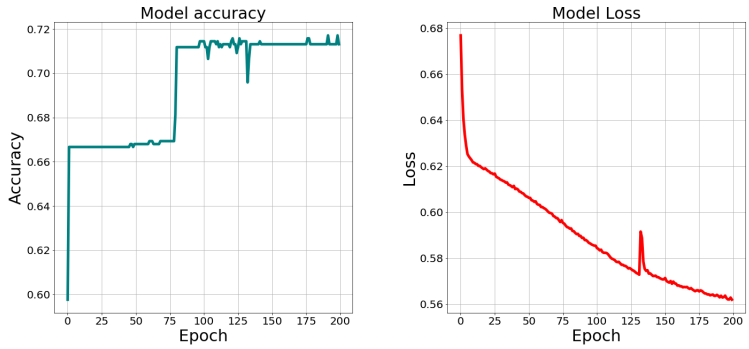

Fig. 6. Model Accuracy and Loss

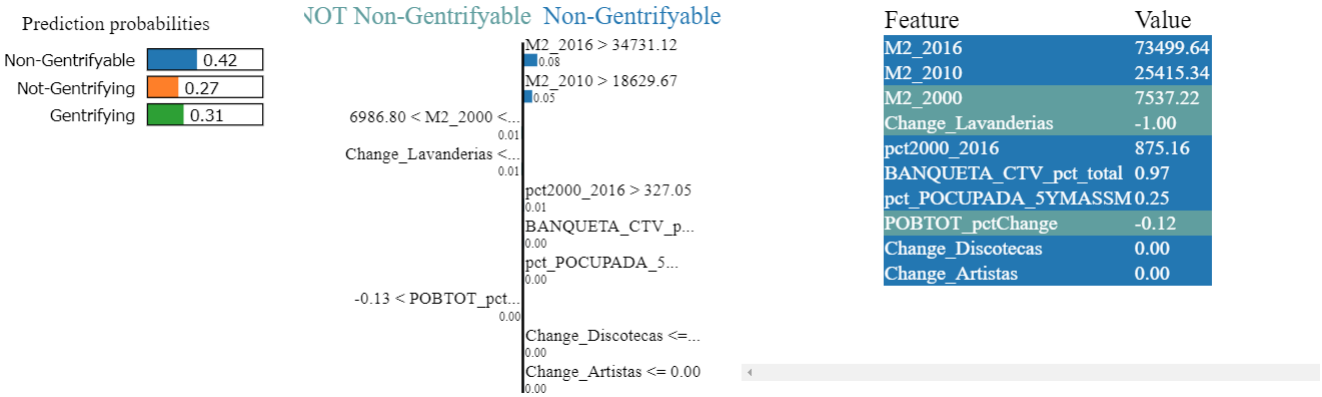

Fig. 7. Lime Output

V. Conclusions

In this work we have presented a new way to predict gentrification. We have used a Neural Network and while our accuracy is not as good as with the classification trees presented in our earlier work, we have used more labels to capture the complexities of the process of gentrification.

Adding the LIME analysis also helped to evaluate the process of gentrification for extremely specific neighborhoods, which is usually a problem with classifiers that in general apply a blanket approach when it comes to explainability of the features to perform a classification.

In our future work, we will use our model in other cities, however, there are extra degrees of complexity since we need information about the gentrification process in those cities. While other Mexico cities might be good candidates, we expect that our system can be used for other cities around the globe.

References

[1] L. K. Poulsen, D. Dekkers, N. Wagenaar, W. Snijders, B. Lewinsky, R. R. Mukkamala, and R. Vatrapu, "Green cabs vs. uber in new york city," in *2016 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2016, pp. 222–229.

[2] J. Oskam and A. Boswijk, "Airbnb: the future of networked hospitality businesses," *Journal of Tourism Futures*, 2016.

[3] F. R. B. of St. Louis. Average Square Feet of Floor Area for One-Family Units. (2020, January 30). [Online]. Available: https://fred.stlouisfed.org/series/HOUSTSFLAA1FQ, January 30, 2020.

[4] R. Murdie and C. Teixeira, "The impact of gentrification on ethnic neighbourhoods in toronto: A case study of little portugal," *Urban Studies*, vol. 48, no. 1, pp. 61–83, 2011.

[5] L. Lees, T. Slater, and E. Wyly, *Gentrification*. Routledge, 2013.

[6] T. W. Sanchez, R. Stolz, and J. S. Ma, "Inequitable effects of transportation policies on minorities," *Transportation research record*, vol. 1885, no. 1, pp. 104–110, 2004.

[7] *Gentrificación: las colonias de CDMX que se "aburguesan"*, 2017.

[8] J. Reades, J. De Souza, and P. Hubbard, "Understanding urban gentrification through machine learning," *Urban Studies*, vol. 56, no. 5, pp. 922–942, 2019.

[9] K. F. Gotham, "Tourism gentrification: The case of new orleans' vieux carre (french quarter)," *Urban studies*, vol. 42, no. 7, pp. 1099–1121, 2005.

[10] Y. Alejandro and L. Palafox, "Gentrification prediction using machine learning," in *Mexican International Conference on Artificial Intelligence*. Springer, 2019, pp. 187–199.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[12] *5 colonias con potencial en el DF*, 2014.

[13] L. Rokach and O. Maimon, *DATA MINING WITH DECISION TREES.Theory and Applications*, 2nd ed. World Scientific, 2015, vol. 81.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE