Daniel González Muela
Antoni Carbonell Muñoz

Predicting house prices in Brooklyn

For this project we studied and predicted the public house prices of sales in Brooklyn, New York from 2003 to 2017. The original Kaggle dataset [1] has 390.883 rows and 110 features, described in [2] and [3], that are the 2 public datasets used to construct the Kaggle dataset. The target variable Y is house price.

Data Cleaning & Transformation

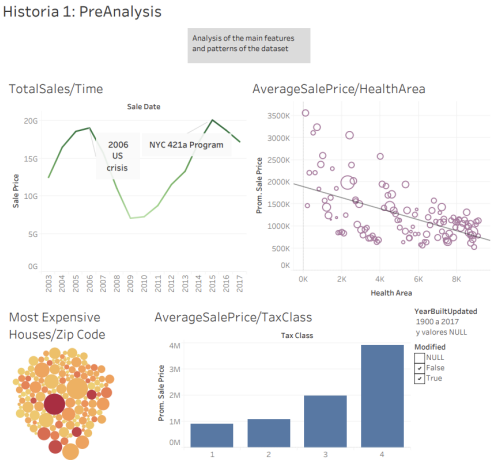
Of the 110 features, 94 of them contain null values. Our first decision for preprocessing the data was to eliminate columns with more than 25% of nulls, and then eliminate rows containing any null values. We splitted the sale date in day, month and year, and transformed the house prices to their present value, assuming a yearly risk free rate of 3% (due to inflation prices increase steadily and raw prices of 2003 can not be directly compared to 2017 prices).

Finally we removed rows with 0 sale price (about 100K samples), normalized the dataset and transformed the objects (which are categorical variables) to dummy variables using one-hot encoding but, to avoid having a huge dataset we removed objects that have more than 80 unique values. We ended up with 2 cleaned datasets of 192.205 rows and 336 columns.

Data Visualization

Before predicting the house prices a first visualization of the dataset was done with Tableau. [Figure 1: TotalSales/Time] Overall the 2003 to 2017 Brooklyn sample resembles the economic situation of the US house market: the 2006 and 2017 sales peaks are followed by the 2006 US housing crisis, and the 2017 renewal of the NYC 421-a. From [Figure 2: AverageSalePrice/HealthArea] we can infer the importance of Health Care in the US, observing a direct relationship between average sale price and the health status of the citizens. Then in [Figure 3: MostExpensiveHouses/ZipCode] we studied the most expensive houses sold by zip code, which correspond to Brooklyn Heights, Downtown Fulton and Williamsburg South, three of the most expensive areas in Brooklyn.

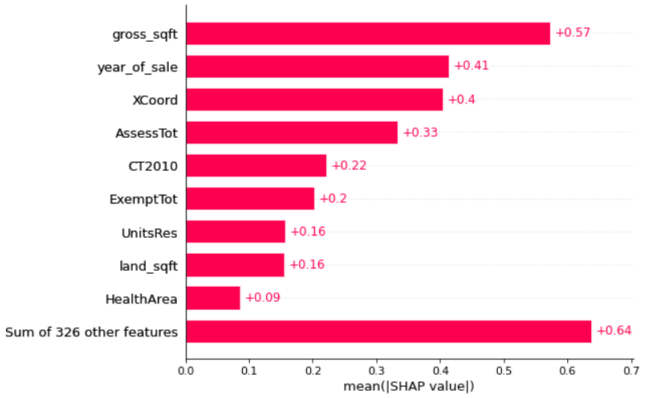
At the end, we plotted the average sale price per tax class [Figure 4: AverageSalePrIce/TaxClass], and created a Dashboard [Dashboard 1:PreAnalysis] in which you can play using this last Figure as a filter, as well as also filtering by the years in which the buildings were built and if they had any modification along the years.



Data prediction
Our first attempt consisted of a simple linear regression model that resulted in a poor regression score of 16%. A PCA plot of the 2 principal components verified it.

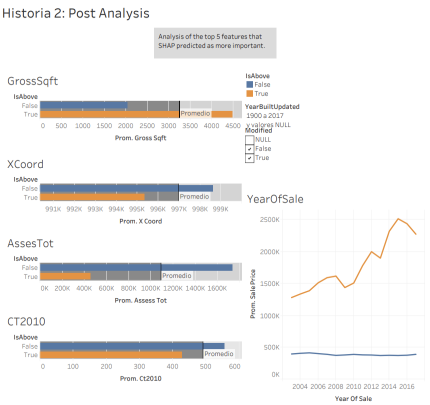
Classification (I)
Our second attempt consisted of classifying if a price was above or below the median price. We tried 7 different models: Linear regression, support vector machine, logistic regression, XGB Classifier, Random Forest, Gradient Boosting, Naive Bayes and a majority voting.
The better-working model was the XGB Classifier, which is the one analyzed in the XAI section. It gave us an accuracy of 79.66%, an average precision of 80% and an average recall of 79.5%. (the results for the other models can be found in the notebook).

XAI



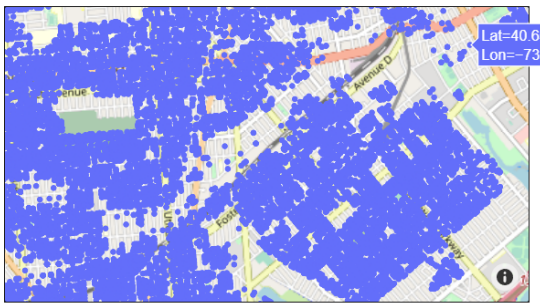
From the Shap values, the top features that have more importance were : gross square feet, year of sale (new buildings tend to be more expensive), XCoord (expensive areas), Assessed total value and Census Tract 2010. Once again, we see the relevance of XAI to explain the models in a more visual and intuitive fashion.

Returning again to Tableau, we checked that these features have a significant average difference depending on the housing classification ,summarized in [Dashboard 2: PostAnalysis]. In the case of year of sale, the relation was not that obvious but while the average price of the houses above the median increased with time, the average price of the houses below it kept constant at about 400K. You can filter by year of edification, if_modified or not, and click on specific years of the YearOfSale figure to use it as a filter.



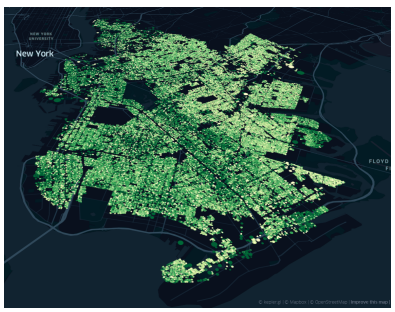
Classification (II)
Finally we used the two best algorithms of the previous section (XGB and Gradient Boosting) to classify in 4 classes (the quartiles). Again, the model that worked better was XGB Classifier with a 59.99% accuracy, 60.25% average precision and 60% of average recall.

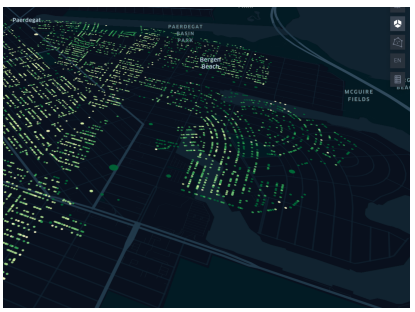
Geospatial representation of our data
The strong part of our data was not its geospatial information alone. However, we did have a dataset of 200,000 brooklyn houses with their location so we wanted to plot and see them. In addition, we wanted to take advantage of some column information for better visualization.
The main issue was that we did not have *Longitude* and *Latitude,* but rather *XCoord* and *YCoord*. Moreover, it was not explained in which system those coordinates were. Hence, in spite of trying to convert it, we could not make any sense of *XCoord* and *YCoord*. But because we had the addresses of every house, we took a small subset of data (500 rows) and used a function to look up the *Longitude* and *Latitude* for each of these rows. With those rows, we trained two Linear Regression Models to predict *Longitude* from *XCoord* and *Latitude* from *YCoord.* The accuracy for both models was 0,86 and 0,87. So the coordinates we did get were not 100% accurate on the map as we will see on the images below. However, the distance between points was respected.
Once with the *Longitude* and *Latitude* for our 200,000 points, we proceed to represent it using 'plotly express' scatter_mapbox:



We can see that the points are slightly misplaced but because the distance between points and the shape is perfect, we can easily recognise where it should have been.

Then, we did the same on KeplerGL online (keeplerGL notebook) but using 40,000 rows as with more it did not work.Here are some shots: we used 'sale_price' as color and 'land_sqft' as radius.





Finally, we filtered by 'neighborhood' and 'year_built' to produce a little video on the evolution of the neighborhood 'Brooklyn Heights' from the 1850's on. The video can be accessed through link [5]

External Links

[1] https://www.kaggle.com/datasets/tianhwu/brooklynhomes2003to2017?select=brooklyn_sales_map.csv

[2] https://www.nyc.gov/site/finance/taxes/glossary-property-sales.page

[3] https://www.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page

[4] https://www.nytimes.com/2022/03/31/nyregion/nyc-tax-credit-housing-crisis.html

[5] https://drive.google.com/file/d/1PPRcVOmmnZC0zP66K0bKgX04oQSjlByy/view?usp=sharing