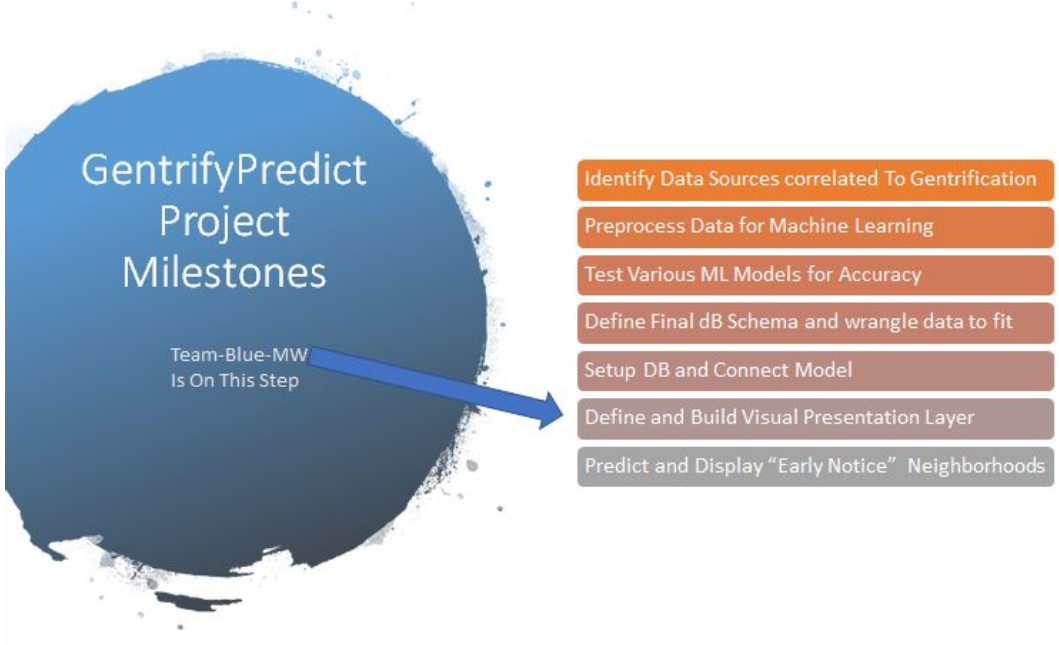
Predicting Gentrification Using Machine Learning

Team-Blue MW Project Milestones



Presentation Deliverables Week 2

The following Week 2 Presentation Deliverables are complete and shown in this presentation.

✓ Selected topic

✓ Reason how and why this topic was selected

✓ Question we hope to answer with the data

✓ Description of our data sources

✓ Description of our data exploration phase of the project

✓ Description of the analysis phase of the project

Our Selected Topic is Predicting Gentrification

So what is Gentrification?

Gentrification is the process of investing in a neighborhood - using public/private monies - to repair and rebuild homes, businesses, and infrastructure in a deteriorating area (such as an urban neighborhood).

The investment sets off a chain reaction by attracting more middle-class or upper income home buyers and business owners which help support the financial investment in the neighborhood, but which may have the unintended consequence of displacing many lower-income, longtime diverse residents of color.

Team-Process Used For Selecting This Topic

To arrive at this Team selection:

- Each team member submitted 4 topics as candidates.
- Topics were summarized in a slide set (available in the repository) and a meeting was held to discuss each topic and available data sources.
- Our team then conducted a multi-vote with each team member selecting their favorite 3 topics.
- Votes were summarized and topics narrowed to 3 based on the most votes.
- The Team then discussed the 3 remaining candidates and by consensus selected Gentrification Predication.
- The Team felt that this topic was an important social issue and wanted to see if the use of Data Science could lead to better overall outcomes in minimizing displacement.

Data Questions

- Can Machine Learning provide early notice of Gentrification?
- Would early warning enable Policy Makers to minimize Displacement while still enabling neighborhood revitalization?

Data Exploration Phase

Data Exploration was done iteratively - *and with some challenges* - as follows:

- Research was done to find past studies on Gentrification and factors leading to Gentrification which might be used as factors.
- A target list of Features was derived from this research.
- Python code was written to generate “fake data” to initially test ML algorithms.
- We then used multiple sources to try and “piece together” a real dataset covering 2 snapshots of the Features across a 10 year time horizon for each geographic zip code. Working from 3 initial data sources, this proved very challenging to build sufficient rows of data without “Null values.”
- Further research from a team member led us to “USA.com” where features were available for the years we needed across all zip codes. We further limited the project to California only.
- From the separate CSV files and Python/Pandas, a dataframe was built and tested.

Analysis Phase Description

The Analysis has been following the steps below:

- Our initial Feature List was derived from research papers from previous government studies on Gentrification. These studies called out socio-economic Features correlated to Gentrified Neighborhoods.
- We then constructed Fake Data for some of the Features to test multiple Machine Learning Models, looking at Accuracy and the Confusion Matrix to see best fit model.
- In the fake data, we labeled known Gentrified Neighborhoods as Y=1, then ran it through the training and testing.
- In assembling final dataset from USA.com CSV files (2 columns per feature and 1 column to calculate % Change between years on the Feature) we set again, known zip codes which had Gentrified to 1. These rows train our model.
- We have iterated testing models and different Features to move Accuracy from 53% to >90%.

Analysis Phase - Gentrification Stages

Stages of Gentrification		
Early Stage	**Transitional Stage**	**Late Stage**
Artists, writers, musicians, affluent college students, homosexuals, hipsters and political activists move in to a neighborhood for its affordability and tolerance.	Upper-middle-class professionals, often politically liberal-progressive (e.g. teachers, journalists, librarians), are attracted by the vibrancy created by the first arrivals.	Wealthier people (e.g. private sector managers) move in and real estate prices increase significantly. By this stage, high prices have excluded traditional residents and most of the types of people who arrived in stage 1 & 2.
Retail gentrification: Throughout the process, local businesses change to serve the higher incomes and different tastes of the gentrifying population.		
Source: Caulfield (1996)[*pages needed*]; Ley as cited in Boyd (2008)[*pages needed*]; Rose (1996)[*pages needed*]; and Lees, Slater & Wyly (2010)[*pages needed*] as cited in Kasman (2015) [*pages needed*].		

Neighborhood Characteristics Before Gentrification

- Older Housing Stock
- Long-time Residents
- High Racial Diversity
- Low Income
- High Percent of Rentals
- Use of Public Transportation

Gentrification In Progress

- Investment Comes Into Neighborhood
  - New Housing
  - New Upscale Businesses
  - Investment in new public transportation (i.e. Light Rail Systems)
- Area Becomes More Desirable To Higher Income Buyers
- Rents Increase – Low Income Residents Get Displaced

Machine Learning (X) Features

X Features (Timeframe 2000>2010)

- X1 - Percent Change in Rental Price
- X2 - Percent Change in Caucasian Resident
- X3 - Percent Change in Median Housing Prices
- X4 - Percent Change in Median Income

Additional Features Under Consideration

- Percent Change in Use of Public Transportation
- Percent Change in Education Level

Analysis Phase - Machine Learning

- Run a supervised machine learning algorithm to determine the factors that signify potential gentrification of a neighborhood.
- Use cross-verification to find goodness of each model's fit
    - Will implement ensemble learners
    - If none of the ensemble learners show a good result on testing data, different algorithms will be used

From early tests of Random Forests and Gradient Boost prediction from our targeted Features is quite possible.

Next Challenge Week 3: High level of accuracy suggests model may be overfitting. We are experimenting with different parameters of the model to determine if this is the case. Model will then be adjusted.