**Hopper – DA5030 – Spring 2018**

**CRISP-DM notes for Gentrification Project**

**Phase 1 - Business Understanding**

Background and business objectives

Gentrification is a process in which households with relatively higher incomes move into an area that is generally lower-income. These higher-income households are able to spend more on housing than the original residents, and they can out-compete and out-spend lower-income households. This drives up prices (home values and rent). It can also change the nature of businesses in the neighborhood to favor establishments that cater to a higher-income clientele. As the area becomes less affordable, lower-income households are often driven out by the high prices and increased cost of living. This raises important equity questions about who is able to live in the neighborhood. If we are striving for inclusive, mixed-income communities, and preserving the right of community members to stay in the neighborhood they have helped to build and create, then gentrification and its associated displacement are signs that those goals have not been met. Are there indicators that might help researchers identify which areas are likely to see gentrification in the coming years? Are there economic, social, infrastructural, and demographic characteristics that can help us to predict which locations will see significant levels of displacement? In addition to hyperlocal factors, are there broader regional dynamics that increase the likelihood of gentrification within a metro area? My project will attempt to build a predictive model that will use census-tract level data to classify which areas are most likely to see significant gentrification and displacement by 2020.

Assess Situation

There have been a number of research pieces that attempt to provide a quantitative method for identifying which census tracts have gentrified. Governing .com released a [report](http://www.governing.com/gov-data/gentrification-in-cities-governing-report.html) in 2015 using 2000 and 2010 decennial census information to identify which urban census tracts across the country had experienced gentrification. The analysis was done in two steps. First, the researchers set criteria for which census tracts were eligible for gentrification:

1. The tract had a population of at least 500 residents at the beginning and end of a decade and was located within a central city.

2. The tract’s median household income was in the bottom 40th percentile when compared to all tracts within its metro area at the beginning of the decade.

3. The tract’s median home value was in the bottom 40th percentile when compared to all tracts within its metro area at the beginning of the decade.

Second, the researchers used a very simple formula to create a simple binary outcome field for gentrification based on the following criteria:

1. An increase in a tract's educational attainment, as measured by the percentage of residents age 25 and over holding bachelor’s degrees, was in the top third percentile of all tracts within a metro area.

2. A tract’s median home value increased when adjusted for inflation.

3. The percentage increase in a tract’s inflation-adjusted median home value was in the top third percentile of all tracts within a metro area.

This is not the first project to attempt to create a leading indicator of gentrification.

Determine data mining goals

This is a prediction data mining goal. Using the 2000 to 2010 period for model training, the goal is to create a list of the census tracts in the fastest growing large metros that are most at risk for gentrification by 2020. The goal of the analysis is to achieve high accuracy (at least higher than random chance), while minimizing false negatives. In this particular research problem, the goal is to target assistance and programmatic energy to census tracts that need it. Leaving the fewest at-risk tracts behind is the ultimate goal, even if it means incorrectly classifying a larger number of tracts that ultimately never face gentrification pressures.

Produce project plan

The bulk of the time spent on this project will be on data exploration and preparation. Gentrification is a complex process, and analysis requires multiple data sources and data points that are likely to have high correlations and meaningful interactions. There will likely be a lot of trial and error in choosing the appropriate data points, and experimentation in creating interaction variables. Having a good initial data exploration process will hopefully serve to provide a path forward on these choices.

Because there are multiple data sources, integration may also pose complications. The data sets being used are at various geographic levels and time periods, and ensuring that these integrations are occurring correctly will be essential to a functioning model.

One important step of the data preparation phase will be to create a binary outcome variable for gentrification. Instead of attempting to measure the scale of gentrification in a census tract, this project will instead aim to

Once the data has been thoroughly explored and prepared, the

**Phase 2 - Data Understanding**

Collect initial data & Describe data

Data was acquired from multiple sources:

* 1990, 2000, and 2010 decennial census data downloaded from Brown University’s “Diversity and Disparities” project, which provides data and analysis around changing demographics and equity impacts across the United States. <https://s4.ad.brown.edu/Projects/Diversity/Researcher/Bridging.htm>

This data set was chosen because it has already resolved the geographic complexity of working with census tracts over time. Boundaries of census tracts shift/divide/combine as population changes, so researchers need to use a crosswalk to compare data from one decennial census to another. The “Diversity and Disparities” project offers the “Longitudinal Tract Data Base,” or LTDB, which crosswalks earlier decennial census data sets to the 2010 census tract geographies. Since this project involves tracking and predicting gentrification over time, this data set helps save the time and headache of having to perform that geographic reconciliation.

The data was downloaded in two parts for each year. Each of the two files includes demographic and housing information that I want to use in my model, and they have a common key field that allows for joining the two data sets (a geogrphic ID number). Unfortunately, some of that key field was incomplete in the data set, so I needed to download a crosswalk from the Census Bureau and use county, state, and tract ids to reconstruct the composite geographic ID used as the common key for joining the data sets.

* The next data point I needed was a binary yes/no variable for whether a census tract gentrified over the 2000 to 2010 time period. No direct data download was available, but I adopted a variation of the approach that Governing.com took in their gentrification analysis:

<http://www.governing.com/gov-data/gentrification-in-cities-governing-report.html>

Then, I applied this methodology to the 2000-2010 LTDB data to create a binary did/did not gentrify target variable for the model training phase.

* Where my modeling approach differs a bit from other efforts to predict or measure gentrification is in the inclusion of regional economic and housing production measurements as predictors of a census tract’s likelihood of gentrifying. In order to incorporate these regional data elements, I needed to download a number of data points from the U.S. Census Bureau on the metropolitan level. I also needed to use a crosswalk to join metro-level data to the census tracts that comprise those metros.
* Lastly, I wanted my model to include data points on transit accessibility and general mobility and connectedness, hypothesizing that locations that provide residents with the best or most broad access to as many locations as possible would be the most likely to gentrify. In order to do this, I downloaded census-tract level information for 2010 from the EPA’s Smart Location Database:

<https://edg.epa.gov/data/PUBLIC/OP/SLD>

This data set includes data points such as walking distance to a transit stop, number of locations accessible by transit, and number of locations accessible by car. Since these data points are on the tract level, joining them to my data set is straight-forward. One drawback of this data, however, is that it was only available for 2010, but I want to incorporate these predictor fields into my 2000-2010 model. An assumption was made that transit infrastructure in 2000 was not incredibly different in 2000 as it was in 2010, even though in some cases there might have been major upgrades and expansions of service.

Describe data – explored in attached RMD document

Explore data-

Data was explored for two reasons: technical joining of the data and relationships between data points

Data relationships explored mostly through correlation plots (via pairs.panels) and basic charts that sought to provide information about relationships between variables.

Data exploration was also heavily focused on feature selection. By studying the pairs panels and scatterplots, I was able to reduce the number of features in my data to the ones that seemed most impactful or potentially interesting.

explored in attached RMD document. Focused on pairs panels and comparative density curves for gentrified/non-gentrified tracts.

Verify data quality - explored in attached RMD document. Data imputation and outlier detection was necessary. Some reconstruction of geographic identifiers was required for successful joining of the different data sets.

**Phase 3 - Data preparation**

I did attempt a principal component analysis, which ran fine, but I did not think the principal components were strong enough to replace the individual features. It took around 20 principal components (based on 22 features) to get to about 90% of the variance. I prefer to have 22 straightforward features than 20 complex, difficult to understand features. The results of this Principal Component Analysis were an initial sign that perhaps the features in my dataset might not have very strong predictive power.

There were a lot of transformations and adjustments that needed to be made in order to join and normalize the data. I won’t share all of them, but they are all documented within my code if you’re interested. A few of the more significant transformations included turning all of the totals from the census tract file into proportions of the whole. Since I am modeling census tracts of different sizes and densities, measuring variables as proportions of the whole is essential to get an apples to apples comparison within the model.

A lot of work also went into the geographic crosswalks between metro areas, census tracts, and census tract block groups. All of these have complicated one-to-many relationships, and then these geographic definitions change over time. I spent a lot of the data preparation phase ensuring that my primary keys were perfect for the joins.

Beyond that, there was outlier detection, data imputation of missing census tract level information (based on the average from the rest of the gentrification-eligible tracts within the same metro, and derived feature creation. One of the more interesting features that was derived is a measurement of whether a metro area has added as many housing units as households over the prior decade. The thinking here was that if a metro gained more households than housing units over the prior decade, then it would put additional pressure on the housing market, which could increase gentrification pressures.

Data set description – See RMD

Select data with rationale for inclusion/exclusion – Done via pairs panel analysis and correlation matrix. See attached RMD document

Clean data - explored in attached RMD document. Data imputation and outlier detection was necessary. Some reconstruction of geographic identifiers was required for successful joining of the different data sets.

Construct data – derived attributes – See RMD. Derived features include normalized rates as well as aggregated metro level statistics

Integrate data – merged data – See RMD – ultimately six different data sets were used and merged

Format data – reformatted data – See RMD. For certain models, features were normalized or made binary to properly fit into R package requirements.

**Phase 5 - Modeling**

Select modeling techniques

Because my outcome variable is a binary classification, but all my predictor features are continuous numeric, there are a few models that are obvious choices. For my models, I am using a logit regression, a support vector machine, a neural network, and a random forest.

I chose these models because I wanted to choose models that learn differently, and set different types of rules and decisions for sorting the data. The hope was that these different methods would have different strengths in classification, so that when they were combined into an ensemble model they would be able to make up for each other’s shortcomings.

When I could, I used k-fold cross-validation when training the models before running the validation data through the models.

Generate test design

Since the ultimate goal is a stacked ensemble learner, I’ve partitioned my 2000-2010 data to result in three data sets: a training set with 60% of the data, a validation set of 20% of the data (for use in testing the underlying models), and a test set of the remaining 20% of records, which I will use to test the stacked ensemble.

Build models – See attached RMD for code

Assess models – addressed within RMD

**Phase 6 - Evaluation**

Evaluate results

The biggest lesson learned is that gentrification is an incredibly hard phenomenon to predict with readily available data sources. This, of course, is not a surprise. However, this is a bit vexing because people with knowledge of a city can often guess which neighborhoods are changing and gentrifying. There are certainly other indicators that other researchers have thought of, including movement of a “creative class” into a neighborhood, or an increase in the number of certain businesses, such as dog groomers or coffee shops. Data points like these are harder to find, especially for a modeling process that includes locations around the country and over multiple periods of time.

Although the model was not particularly predictive, I still wanted to take the next decade’s data (that’s 2010 data) and make a prediction for which census tracts would gentrify in 2010. So, I ran the same data prep and selection processes as I did with the 2000 data, ran the data through the first layer models and then my averaging ensemble, and spit out a set of predictions. The model suggests that 217 census tracts across the large metros are at risk of gentrification by 2020.

Review process

The most obvious way to improve this project is to find new and less conventional input features. It would be wise to do a more thorough literature review to find new and more innovative ways of modeling and predicting gentrification.

Beyond that, additional algorithms could be explored for inclusion in the ensemble learner, such as kNN.

Determine next steps

Conduct thorough literature review, explore additional data sets, include additional modeling techniques. Iterate.

**Phase 7 - Deployment**

Plan development

The ultimate deployment of a successful model would include an interactive mapping interface that would allow users to select a metro area and view which census tracts are predicted to gentrify in the coming decade. This could be done using Shiny as a platform.

Plan monitoring and maintenance

Shiny would allow for integration with underlying R notebooks, which would make updating and maintenance easy. Additionally, if hosted on shinyapps.io, there are readily accessible metrics for the app’s usage from users.

Produce final report

This document in combination with annotated RMarkdown document and video presentation represent the final report.

Review project

Overall, despite the low accuracy and predictive power of my algorithms, this felt like a good start. The true complexity of the gentrification and displacement process was certainly apparent in the modeling process. If anything, this modeling exercise is a clear demonstration of the need for additional research and pursuit of less conventional data sets.

Peer review will also help to shed light on possible improvements within the existing model.