Ilina Mitra
DATA 1030 - Final Paper

For my capstone project, I am using machine learning techniques to study gentrification using the Kaggle data set "Gentrification and Demographic Analysis." This data has been collated from three sources; the US Census Bureau, censusreporter.org, and Logan et al's Longitudinal Tract DataBase. To understand the descriptions of the columns below, it is important to understand some key terminology. The data set measures features within census tracts, which are small entities within counties. These boundaries are updated every ten years, based on demographic shifts. Each tract has about between 2500 and 8000 residents. The data set has 8281 and 17 columns, described below [1].

1. geoid - Census tract ID
2. name - The name of the census tract
3. total_population - The total population in the tract
4. total_population_over_25 - The total population in the tract who are over the age of 25
5. median_income - Median income of the tract
6. median_home_value - Median home value in the tract
7. educational_attainment - The number of people who are 25 years or older with a 4 year degree in the tract
8. white_alone - The number of people in the tract whose race is white alone
9. black_alone - The number of people in the tract whose race is black alone
10. native_alone - The number of people in the tract who are American Indian and Alaska Native alone
11. asian_alone - The number of people in the tract who are Asian alone
12. native_hawaiian_pacific_islander - The number of people in the tract who are Native Hawaiian and Other Pacific Islander alone
13. some_other_race - The number of people in the tract who are some other race alone
14. two_or_more - The number of people in the tract who are two or more races
15. hispanic_or_latino - The number of people who are Hispanic or Latino
16. city - The name of the city
17. metro_area - The name of the metro area

I am hoping to study the effects of the demographic shifts, educational attainment, and median income on median house prices in a city. This is a regression problem, with the target variable being median house prices. I believe this problem is particularly relevant, as it is a data centered look into the gentrification of cities. As median home prices in cities increase, original residents, most often residents of color, are pushed out of their homes and forced to relocate.

This data set was used by Buzzfeed News earlier this year to run an analysis of the five cities (Washington DC, Atlanta, Oakland, New York, and Baltimore) in question, highlighting the demographic changes between 2000 and 2017 and exploring which census tracts have gentrified in that time period. While the article did not make any predictive statements, it created interactive maps that help visualize how communities of color are being pushed out by white populations. [2]

EDA

During time spent performing exploratory data analysis, I wanted to see how the different features in the set related to one another. I constructed a number of graphs and below I have included some of the ones I found most relevant.

The first graph is a scatter plot showing the relationship between educational attainment and median income. The plot shows educational_attainment, or the number of people over the age of 25 within a tract that have a four year degree, on the x axis and graphs median income of said tract on the y axis. I had preconceived notions about the positive relationship between these two variables. Though the plot does not disprove this idea, it certainly is not the strong positive correlation I was expecting to see between the two variables.

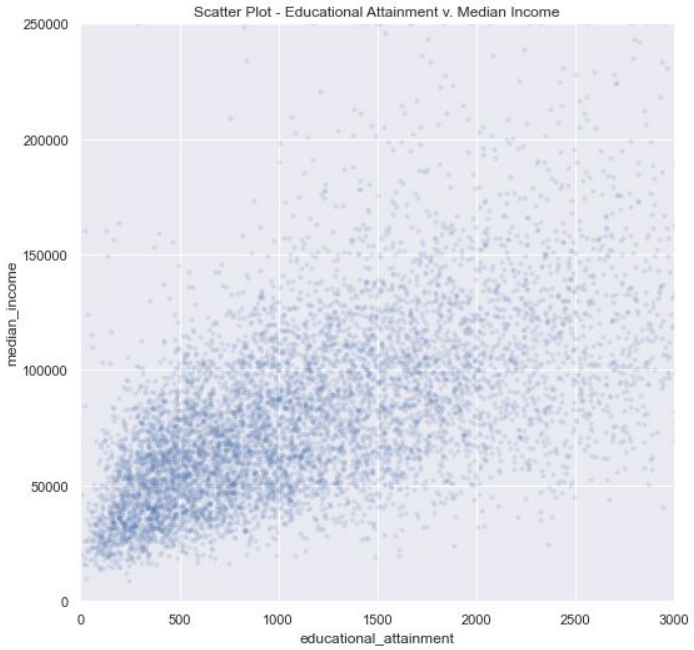

Fig 1: Education Attainment v. Median Income across census tracts

The second graph shows distribution of the target variable (median_home_value) for every city. I created both individual histograms as well as one category-specific one, shown below. It has the median_home_value per census tract on the x axis with the counts on the y axis. Though the colors overlap a little bit, there is a visual depiction of how the distribution of median home values differ across cities. Oakland median home prices extend well into the millions of dollars, while Atlanta and Baltimore are more clustered in between 100,000 and 200,000 dollars.

Fig 2: Median Home Value Distribution Across Cities

Given that the data is collected by city, my first intuition was that this data is not independent and identically distributed. I thought that there could be something intrinsic to specific cities that determined median home value and this motivated me to create such a graph. As shown in the graph above, there is significant overlap in the distributions of median home value among all the cities, despite some of the differences mentioned above. I feel that performing a basic split also makes sense, especially given that the distribution is not drastically different between cities, except for Atlanta.

The last plot I created was a heatmap representing the collinearity of the multiple feature variables in the data set. The darker squares signify a strong positive correlation between two variables while light free is a significant negative correlation. This information is useful when considering how to improve the model moving forward.

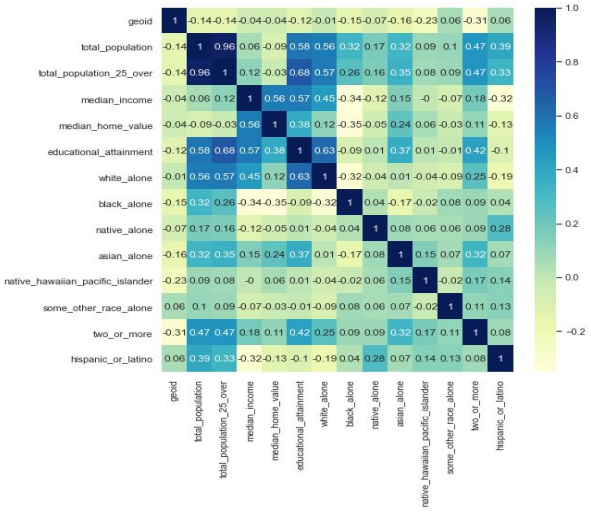

Fig 3: Collinearity Heatmap

Preprocessing and Splitting

Because the categorical features in the data set, city and metro_area, cannot be ranked, I preprocessed using OneHotEncoder. For the continuous variables, I selected the StandardScaler preprocessor, to best cover any outliers within the features.

The only two columns in my data that contained missing values were 'median_income' and 'median_home_value.' Upon further inspection, I found that about 4% of points were missing. Given that only a small fraction of points contain missing values, I dropped the respective values. After dropping these rows and preprocessing the data I was left with 7936 rows and 17 columns.

Methods/Results

As mentioned at the onset of the paper - I want to predict the median home value using the given feature variables and I felt that performing a regression analysis was the most prudent.

Given that I am answering a regression question, I considered two metrics. First, R^2 which is the proportion of variation in the outcome that is explained by the feature variables. As someone who studied economics in college, this metric made the most sense to me. Furthemore, I was interested in seeing how good the models I created were in predicting the mean median home value. I also considered using root mean squared error (RMSE) which measures the average error performed by the model in predicting the outcome of the observation, but ultimately decided to stick with R^2.

A baseline model situates a more complex model and is able to provide context. By definition, the model must be simple and as such I thought it best to set up a simple linear model, with no hyperparameter tuning to operate as my baseline. The mean R^2 value associated with this is listed below.

```
test_mean: 0.5358468045470104
```

As is evident from the picture, the mean R^2 value suggests that the baseline model is fairly predictive. As I worked through this project I wanted to develop models that would increase the predictive power of the regression. My first logical step was to run a ridge regression, given that it shrinks the coefficients, helps reduce potential multicollinearity issues and impose constraints. In a ridge regression, the cost function is altered by adding a penalty equivalent to the square of the magnitude of the coefficients. In this model, I tuned the hyperparameter alpha, which measures the coefficient shrinkage. The mean R^2 value is again listed below, as well as the standard deviation.

```
test_mean: 0.539376374747736
test_std: 0.01268781687147004
```

I also thought it smart to try a Random Forest Regressor, thinking that a forest algorithm would improve accuracy and reduce overfitting. It is important to note that there are some

uncertainties introduced into the model when fitting tree algorithms. When running this model I tuned the parameters for max_depth and max_features. Max_features represents the number of features considered when looking for the best split while max_depth represents the depth of each tree. The mean R^2 value is again listed below, as well as the standard deviation. We can see there is a substantial improvement from the previous model.

```
test_mean: 0.6710416153151422
test_std: 0.01616417622695386
```

Lastly, given the success I had with the RF regressor, I also thought it would make sense to train an XGBoost regressor. XGBoost builds one tree at a time - in essence gradient boosting algorithms combine results along the way. For this model, I tuned the max_depth hyperparameter. Though it is not obvious from the python file, I considered max_depths 100 and 1000 and found that 1000 splits resulted in more information about the data being captured. The mean R^2 value is again listed below, as well as the standard deviation and we can see that this model gives us our best R^2 value.

```
test_mean: 0.6832196633032148
test_std: 0.017151838877981964
```

Results

With regards to the baseline scores, the model that performed the best was XGBoost Regressor, giving the highest R^2 results. It looks to be approximately eight standard deviations above the mean test score obtained in the baseline model. This mean test score tells us that approximately 68% of variance in median home value can be attributed to the feature variables in the model. This is a fairly substantial and non trivial amount.

To study feature importance, I calculated permutation scores for my best model - XGBoost. I was surprised to find that the top 3 most important features were asian_alone, city_Atlanta, and white_alone. Intuitively, city_Atlanta and white_alone do certainly explain variations in median home value, especially given the distribution for home values in Atlanta that we saw above. However I was not expecting the significant role asian_alone played in this model and I hope future models can elaborate on this more.

Outlook

There are four main things that come to mind when thinking about how to improve the model.

1. Change the metric - Adjusting the performance metric to RMSE would give us the ability to see how close our predictions for median home value are to the true values and this could give us greater insight into the model's accuracy.

2. Incorporate missing data into the XGBoost model - Unlike other scikit learn models, XGBoost can work with missing data. If given more time, I would not drop rows with missing values and implement an XGBoost regressor on the dataframe as is.
3. Multicollinearity - Figure 3 above tells an interesting story about how a number of feature variables are related to each other. A main goal of regression is to isolate the relationship between each independent variable and the dependent variable. The coefficients obtained from running a regression tell us how much the dependent variable changes for a one unit change in the independent variable, holding all other independent variables constant. When a correlation exists between independent variables, there is reason to believe that changes in one variable are associated with changes in other variables. If this is the case, it can be challenging for the model to interpret the individual relationships. In order to get accurate coefficient results, I would use Figure 3 as a starting point to determine if columns closely related to each other can be dropped to obtain more robust coefficient calculations.
4. Group Split by City - Given that city_Atlanta was determined to be of significant importance, I would consider implementing a group split by city.

In summary, I believe that the model developed is an interesting starting point to study what variables affect rising median home values across America. However, it is not the case that any concrete conclusions about gentrification can be made using this preliminary analysis.

References:

1. https://www.kaggle.com/mrmorj/gentrification-and-demographic-analysis, Gentrification and demographic analysis: census tract-level gentrification in five major cities
2. Lam Thuy Vo (2020) “If You Live In New York, Atlanta, Oakland, Baltimore or DC, You Need To See These Maps About Gentrification,” Buzzfeed News, https://www.buzzfeednews.com/article/lamvo/gentrification-maps-white-black-people-neighborhoods

GitHub repo - https://github.com/ilina-mitra/1030-project