Predicting Population Growth in the United States

Georgia Institute of Technology – MGT 6203 Fall 2022

Team 15

Andrew Akers Alec Martin Dongsuk Cisco Kim Naveen Ram

CONTENTS

1. INTRODUCTION & BACKGROUND INFORMATION

The COVID pandemic has changed the way people live globally, including America in many ways. One of the biggest changes that we have observed over the last two years in the US is increased population migration across states as people look to find better conditions for their personal and professional lives. Based on the Census Bureau state population and domestic migration estimation data from 2020 to 2021, the net population migration in certain states was quite significant [1]. During this period, California and New York were the two states with the highest population decline (New York: -319,020, California: -261,902). On the other hand, Texas added 310,288 new residents and Florida gained 211,196 at the same time.

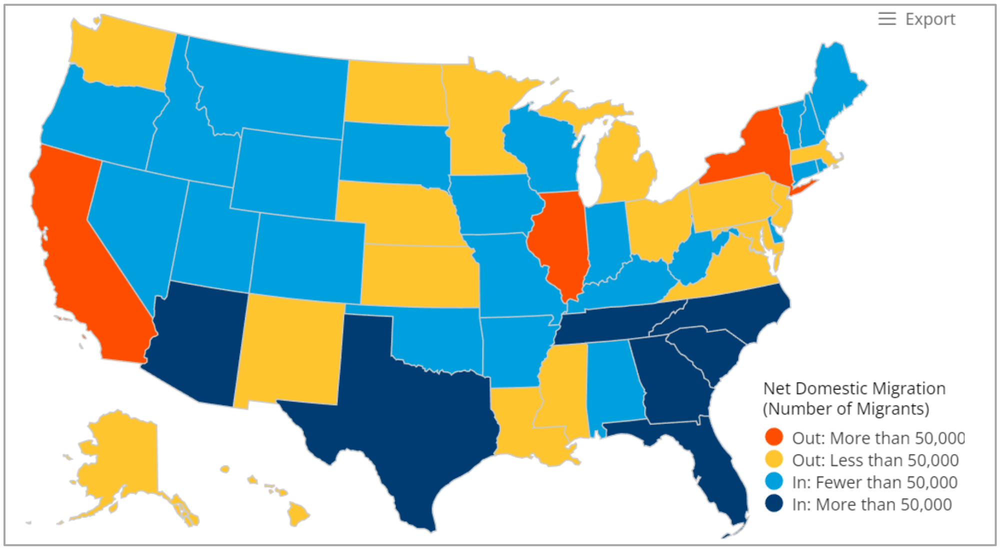

Figure 1 - Net migration by state in 2021

We cannot conclude that it is purely driven from new lifestyle preferences coming from lockdown. There are lots of other factors that could be influencing this migration such as tax increase/decrease, home prices, average salary, etc. At the beginning of the pandemic, people argued that this trend would be a short-term phenomenon, but this has been proven incorrect. Some of the changes have become a new norm, which has accelerated population migration with new dynamics.

Changing migration pattern will impact not just individuals, but corporate level strategy as well—understanding these patterns and their potential drivers is becoming more important for companies to adjust their go-to-market strategy and manage their cost by reopening or relocating their physical offices. Securing stable talent pools and having a strong customer base is one of the biggest and fundamental success factors for companies. COVID-19 has created a significant demographic shift globally and understanding this change and building a digital workplace strategy has become a critical corporate initiative. To be a winner in the new era, companies should be able to accurately anticipate major locations that will have high population growth, leading to a strong potential customer base and high-quality labor pool that can support their future investment accordingly.

The results of this analysis could impact a broad set of decisions across many different businesses with a focus on planning. We have identified three key areas of business decision-making that could benefit:

1) Companies that are planning on expanding office space to a new geographic region.
2) Retail-focused businesses determining where to build new store locations.
3) Real estate developers that are deciding on the best locations for new projects.

All three of these decisions should be influenced by the expected population growth in the area. For example, a company expanding its office space to a new location should be concerned with the size of the available talent pool in which to hire from (e.g., Amazon thinking about where to set up its next shipping center). The rationale for the last two areas is straight-forward: future population growth is essential in determining demand and revenue forecasts. Both retail-focused businesses and real estate developers will want to target areas with high expected population growth, which will have a positive impact on expected demand for their products.

In this research, we descoped to prove co-relationship between population growth rate and potential business revenue increases assuming there is a positive relationship between those two. Instead, we will focus on estimating the population growth based on three major categories—Affordability, Opportunity, and Desirability—to provide the best location for companies to expand their future investment. We expect that migration patterns within the United States should be predictable by capturing data that influences people's decision making and rationale for moving. Building a regression model and validating it on out-of-sample data will allow us to test whether this hypothesis is true.

2. LITERATURE SURVEY

Migration patterns as a field study have been steadily growing over the past decades. The technological advances of the 21st century have allowed researchers access to vast amounts of data which they have used to analyze who, where, when, and why people move. With hundreds of analytical and theoretical papers available on the subject we have found three which provide a basis for the question we seek to answer in this project.

A first critical finding in our literature survey was a contribution from the United States Census Bureau in conjunction with researchers from Harvard University who provided a thorough empirical analysis using census data to determine the migration patterns of young adults in the United States. One of the main contributions of this work was a resulting publicly available dataset which outlined the number of young adults who moved from one "Commuting Zone" (CZ) to another. The result is a 2-dimensional matrix of ratios whose index is the origin commuting zone and destination and whose value is the percentage of young adults from the origin who live in the destination [2]. This data provides a baseline which our project can use for validation and adds precedence for the use of zoning definitions in migratory analysis like we will do with MSAs.

Another aspect of modern migratory analysis is "big data". There have been surprisingly few efforts to utilize the power of social media and GPS data while studying migration patterns, and unsurprisingly these approaches have yet to be refined and validated. The effort it takes to scrape these large data sources and clean them is no simple task and would take a considerable amount of time we have to provide our analysis. Not only has big data's use in migratory pattern research proven to be difficult but it also poses some legal and ethical dilemmas [3] which we would like to steer clear of in the beginning phases of this project.

In most of the studies which we have read the primary factors which were being estimated for causation was economic, namely wage data. However, the research on this topic often proved that wage data had significant but small impact on migratory patterns [2, 4]. In both approaches the analysis involved creating novel models for labor market strength and migration likelihood based on the census data for income, age, and location as well as immigration policy. The conclusion of this analysis was that while wages did influence the migratory patterns of young adults this effect was small, and that a more indicative predictor of a young adult's likelihood of moving CZs was their birth family's household income [2].

3. DATA

3.1 Sources

For this project, we needed to collect data that was relevant for predicting population migration in the US. As such, we started by defining three high level categories of data, which included Affordability, Opportunity, Desirability. Each of these also had subcategories. For Affordability, we looked at the average cost of living and housing costs. For Opportunity, we looked at wages, the job market, and education. Lastly for desirability, we looked at crime and weather. We collected data for these subcategories from various online sources and merged this into one feature set. The sources we used included the Bureau of Economic Analysis, Data.gov, the Bureau of Labor Statistics, the OpenWeather API, the FBI, and the Census Bureau.

3.2 Cleaning

Most of the data required light cleaning and restructuring, including pivoting from wide to long format. For example, many of the files have columns for each year, which is not suitable to pass into a model in most statistical software, including R or python. In addition, there was data that needed to be preprocessed before it could be imported properly to be cleaned.

We aggregated historical weather data by month using the OpenWeather API and the latitude and longitude values for each MSA, retrieved using the geocoders API. This data was converted from JSON API responses to data tables organized by MSA code and month. In total, there are 53 features related to temperature, humidity, precipitation, etc. with 12 observations per MSA. For the organization of the data, we included this data for each year from 2008 to 2020. We also gathered data on employment by industry and grouped these data points by MSA and year going from 2008 to 2020. This initial dataset has over 100 features due to the various categories of industry that are converted to factor variables. Through feature selection, we will reduce these factors to the most important industries.

We aggregated the data into a single multi-dimensional table, which required us to join the various datasets using a common index, namely the Metropolitan Statistical Area (MSA) code. Most of the datasets used in this project contained MSA codes, which allowed us to easily join them. Datasets that did not contain MSA codes had some geographic indicator (city name, state name, latitude and longitude, etc.) which we performed preprocessing on to enable proper unions that follow the MSA code indexing convention. Our final data matrix contained a row for each MSA and time period (e.g., year), a column for the population change (dependent variable), and 23 columns containing the final set of independent variables.

3.3 Exploratory Data Analysis

First, we analyzed the properties of the dependent variable: population change. The distribution of the variable is roughly normal, but it has fat tails (i.e., positive kurtosis), especially the right tail. The density plot below displays this observation visually. This means that it is more likely for an MSA to experience abnormally high population growth than abnormally high contraction. This has the potential to cause issues in a linear regression by making the error terms non-normal, which we will monitor during the modeling stage.

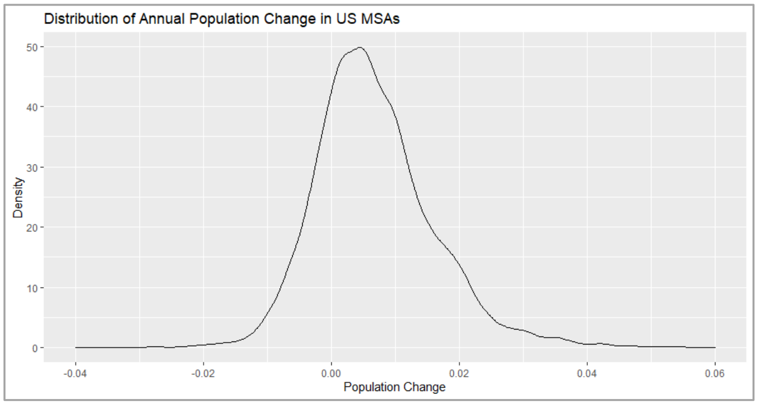

Figure 2 - Distribution of Annual Population change

Another important characteristic of population change we discovered is that it is serially correlated from year to year. The correlation between population changes across MSAs in year $t-1$ and year t in our dataset is 0.78. While this observation will be helpful in a predictive model, it will not be useful in a descriptive model that has a goal of determining the important factors driving population change.

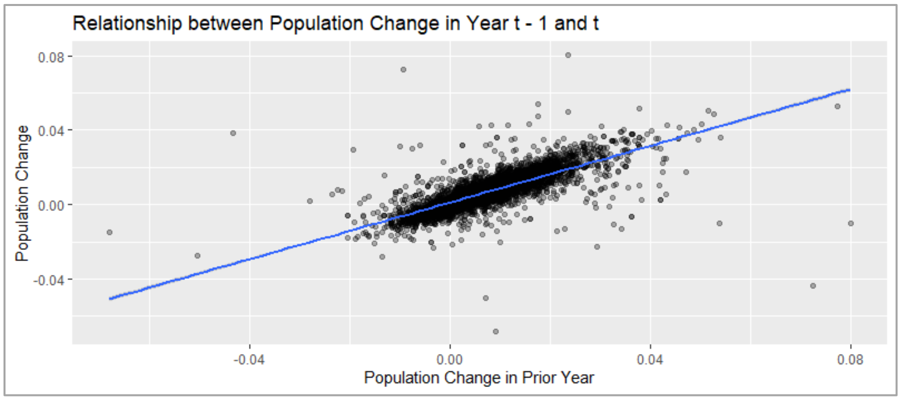

Figure 3 - Relationship between population change in year t-1 & t

Next, we tested the linear relationship between population change and the cleaned independent variables as part of the exploratory data analysis stage. A correlation matrix revealed that several of the independent variables had a moderately strong linear relationship, including the change in employed persons, the change in aggregate wages, the change in economic output (GDP), and a cost-of-living index. Interestingly, the cost-of-living index had a positive correlation with population change, suggesting that higher prices do not inhibit population growth at least in the short-term. We also noted that many of the independent variables are correlated, which will be an important consideration when implementing a model.

In exploring the weather data, we found that the most correlated data to the population change by MSA was when we took the mean of the monthly data on a yearly level for record min/max temperature, average temperature, and information on average cloud cover. However, even with these top values, the correlation between the values was not strong. For example, displayed below is the plot of average cloud cover and population change. We will need to transform or find meaningful interaction terms to find features that are more correlated with the population change attribute.

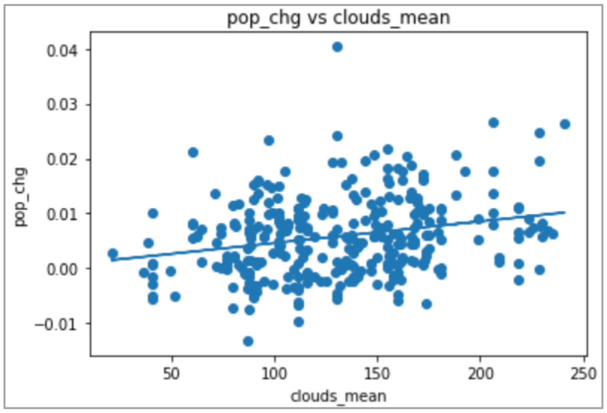

Figure 4 - Relationship between population change & clouds

3.4 Feature Engineering and Transformation

There are several data transformations we needed to apply to the features. In general, there were three broad types of possible features: levels, normalized levels, and changes. Levels are typically raw data measures and were often not useful because the data was not comparable across MSAs. For example, the number of employed persons in an MSA was heavily dependent on population size. Since the dependent variable, change in population, is not explicitly dependent on population size, we needed to make appropriate adjustments. A normalized level is a simple solution that involves dividing the feature value by the population size. Continuing the previous example, number of employed persons becomes the percentage of the population that is employed, which has the same interpretation regardless of MSA size. Lastly, a change variable involves calculating the percentage difference between a feature value and its one-year lagged value. For example, instead of the total number of employed persons or the percentage of the population is employed, we can look at the growth (or contraction) in employed persons over the past year.

Since the data is time dependent, a key step in the data preparation process was appropriately lagging the feature values. This ensured that we used past data to predict future population changes as opposed to concurrent changes. Therefore, we used feature values associated with year $t-1$ to predict the population change in year t.

Finally, because we considered a variety of independent variables in our analysis, many of which vary in not only magnitude but units, we will have to perform scaling and normalization to ensure that our regression models are not skewed towards any one feature, particularly if regularization is applied (see modelling section below). Features such as average temperature and household income, for example, needed to be normalized to achieve meaningful results. Additionally, there was an opportunity to normalize features over time and across MSAs. The motivation for this normalization is that population change resembles a "zero sum game." That is, people leaving one MSA typically go to another (ignoring net immigration at the country level). For example, in our initial data analysis, we found that employment growth is positively correlated with future population change. In poor economic times, it is possible that employment growth is negative in the majority of MSAs. However, it is unrealistic that population growth would also be below average in the majority of MSAs.

See the Appendix for the list of features, where the data came from, and the transformation applied, if any.

4. MODELING

4.1 Models

The first type of model we employed was a simple linear regression of the form $\text{Population change}_{t+1} = \alpha + \mathbf{x}_t\beta$ fit via ordinary least squares where α is a constant intercept, $\mathbf{x}_t$ is the vector of feature predictors captured at time t, and β is the vector of model coefficients. Given its simplicity, linear regression will serve as a baseline model to compare other more complex models regarding performance. Linear regression models are also easy to interpret, which will be a consideration when selecting the final model.

We also tried several variations of the linear regression model, including a linear mixed model and regularized models. A linear mixed model is useful for modeling hierarchical data (also called longitudinal data) where there are multiple observations across different groups. For example, our data is grouped by MSA since we have 13 (annual data from 2008 to 2020) observations for each MSA. The key innovation of the linear mixed model over classic linear regression is that it allows the model coefficients to vary from the population coefficients for each group. The basic form of the linear mixed model that we used allows varying intercept terms, making the form of the regression $\text{Population change}_{i,t+1} = \alpha_\text{i} + \mathbf{x}_t\beta$ where α_i represents the intercept for the i^{th} MSA.

Next, we fit two types of regularized linear models. Regularization encourages smaller absolute model coefficients and provides two important potential benefits: 1) it can prevent overfitting and 2) it can be an efficient form of feature selection. Regularization works by adding a penalty term λ to the model's objective function, giving it the form $\sum_i^n (y_i - \hat{y}_i)^2 + \lambda|\beta|$ where $|\beta|$ is either the L1- or L2-norm of the coefficient vector (excluding the intercept). Using the L1- and L2-norm results in the LASSO and ridge regression models, respectively.

The main drawback of the models discussed above is they assume the relationship between population and each feature is linear. Therefore, in addition to the linear models, we fit a more flexible machine learning model: the random forest regression. Random forests are an ensemble technique that builds many individual decision trees using only a subset of the features in each tree. An individual decision tree splits the data multiple times into subgroups based on the feature values and then takes the average response value of all the observations in a final subgroup (called a leaf) as the predicted value. The final predicted value is the average prediction across all the individual regression trees in the forest.

4.2 Validation and tuning

Since all the data has a time component, we implemented a time series validation process for model selection, often called step ahead validation. The basic idea is that we started with an initial training set using the first five years of data (2008-2012) to train the model as well as tune any hyperparameters via 5-fold cross validation, if applicable. Then, we used the trained and tuned models to make out-of-sample population change predictions for the following year (2013). We then looped through the rest of the data, expanding the training set by one year at each iteration. For example, during the second iteration, the models were trained and tuned on data from 2008-2013 and out-of-sample predictions were made for 2014. The key benefit of this validation method is that it matches how the model would be used in practice. Additionally, it allows us to assess the stability of model performance over time.

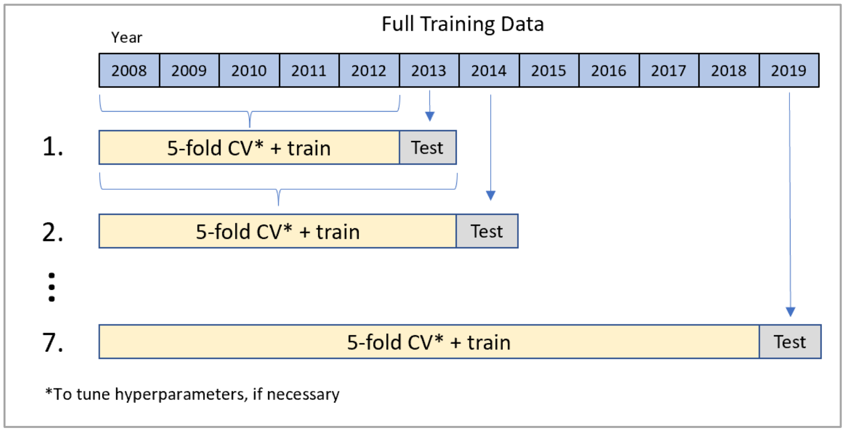

Figure 5 - Visual depiction of the step-ahead validation process.

4.3 Performance and model selection

The tables below summarize the performance (R^2 and mean squared error) of each model for the validation data according to the process outlined above.

	2015	2016	2017	2018	2019	**Average**
OLS	0.73	0.72	0.79	0.75	0.59	**0.71**
Linear mixed effects	0.59	0.57	0.65	0.65	0.48	**0.59**
LASSO	0.72	0.71	0.80	0.75	0.59	**0.71**
Ridge	0.72	0.71	0.79	0.75	0.58	**0.71**
Random forest	0.77	0.78	0.80	0.78	0.58	**0.74**

Table 1 - Validation R^2 by model and year

	2015	2016	2017	2018	2019	**Average**
OLS	0.27	0.30	0.19	0.21	0.44	**0.28**
Linear mixed effects	0.40	0.45	0.32	0.29	0.55	**0.40**
LASSO	0.28	0.30	0.19	0.21	0.43	**0.28**
Ridge	0.28	0.30	0.20	0.21	0.44	**0.29**
Random forest	0.23	0.24	0.18	0.18	0.44	**0.25**

Table 2 - Validation mean squared error by model and year

The linear models all perform similarly, outside of the linear mixed effects model, which has substantially inferior performance. The random forest model has the edge in performance in terms of both R^2 and mean squared error in every validation year. The average R^2 and mean squared error was 0.74 and 0.25, respectively. For context, a null regression model (intercept only) has a mean squared error of 0.88. Given its validation performance advantage, we selected the random forest as the best model. Lastly, we calculated performance using the selected model on the holdout test set with data from 2020. The R^2 and mean squared error were 0.82 and 0.15, respectively.

5. DISCUSSION AND KEY TAKEAWAYS

5.1 Evaluation of Results

The resulting model described above returned results which were in line with our hypothesis regarding Affordability and Opportunity. The most unaffordable MSAs were predicted to have the most population decrease with Los Angeles, CA and New York, New York having the largest percentage decrease in population for Large Population MSAs (Figure 6). Even for lesser known MSAs the MSAs which our model predicts to have the most growth are MSAs which have relatively low housing costs and higher wages (Figure 6).

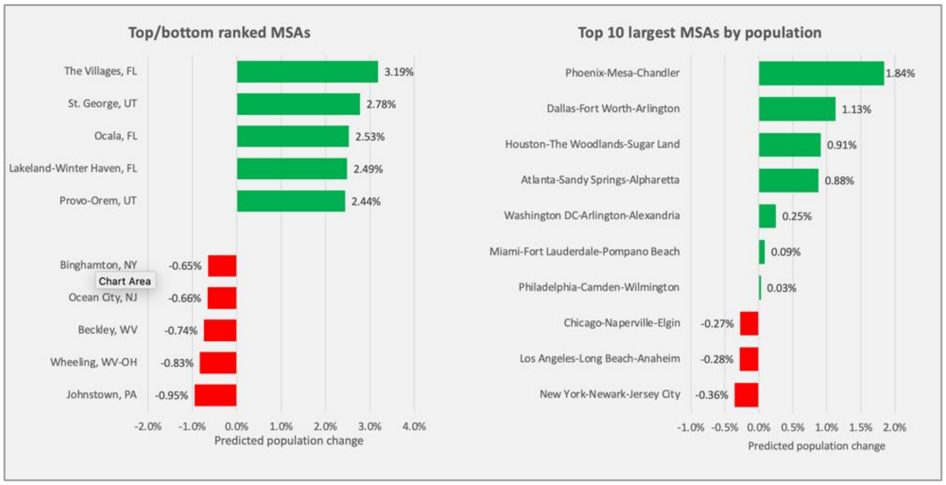

Figure 6 - Model predictions for population change

5.2 Key Drivers for Migration

As mentioned in the evaluation of results the main drivers for migration all fell into the categories of affordability and opportunity. When plotting a components and residuals plot for our resulting model, it is observed that there exists a clear positive correlation between opportunity features such as change in employment rates and wages. Meanwhile, affordability or lack thereof can also be seen to be strongly correlated with a decrease in population, features such as Regional Price Parities and Housing costs were large factors in why people chose to leave their current location in favor of another.

Surprisingly, the single most important feature for predicting population fluctuations was the prior year's population change. This was something which was not thought of in our initial hypothesis and should be studied more closely. Even more surprising was that desirability metrics such as crime and weather looked to have no predictive power and were almost completely uncorrelated to the rise and falls of population.

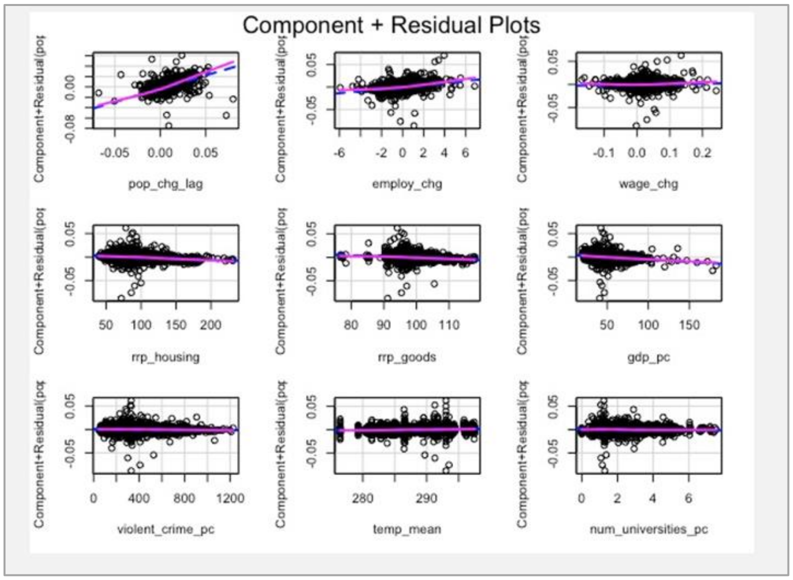

Figure 7 - Components + Residuals Plot of Model

5.3 Challenges

There are two main challenges in building a predictive model for this problem, both of which centered around the nature of the data and our chosen hypothesis. The most difficult challenge we discovered quickly is that desirability is a hard metric to quantify. In the end we only had two broad categories for desirability which were crime and weather. Other factors such as recreational areas, public schooling, and cultural destinations were not included and all and I think would play a large role in a person's choice to move to a less affordable destination like New York for example. Another challenge this model faces is handling yearly outliers. As you can see here 2019 was a huge outlier in terms of our model's accuracy with a 20-point hit in mean square error for that year. Outliers such as this would hopefully be smoothed out as more historical data becomes available.

Affordability and Opportunity data points such as regional price parities and income per capita were all found to be highly correlated, which also reduced the number of unique features we could use in our model.

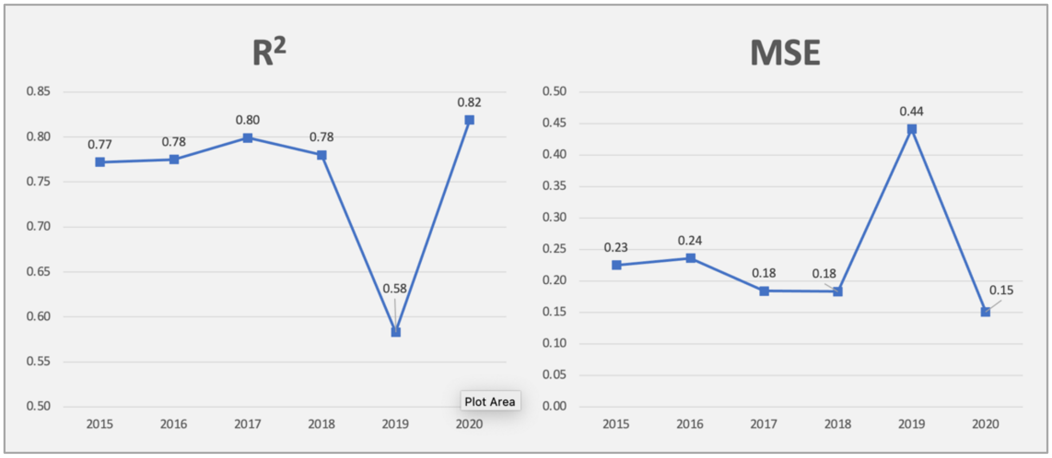

Figure 8 – Out-of-sample performance of the final model

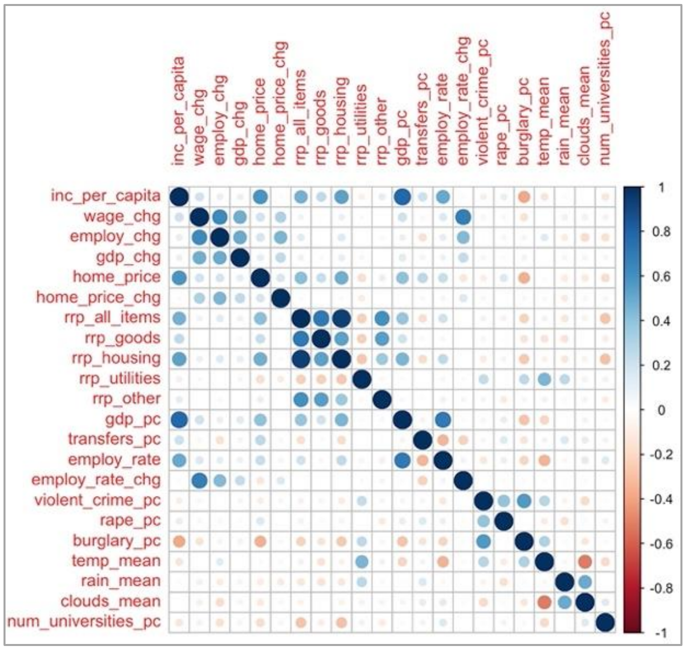

Figure 9 – Feature correlation matrix

5.4 Conclusion

Migratory patterns are a challenging analytical problem—the number of variables that goes into these decisions and the nature of human behavior make building accurate predictive models difficult. However, even from the extremely small set of predictive features we used, our random forest model proved to have a considerable amount of predictive power with R^2 values over 70% in five out of the six validation periods. The hypothesis of three composite drivers for migration which included Affordability, Opportunity, and Desirability proved only to be partially correct with Affordability and Opportunity having any feature importance. That said, even these broad driving factors seemed to be much less informative in predicting population changes than the simple year lag in population change. Something which we might seek to exclude in a further analysis into our hypothesis.

Going forward, more work needs to be done in collecting and quantifying desirability feature data for MSAs. The fact that our model saw no correlation between population changes and the desirability metrics we had in our data is unsatisfactory and we believe more data would uncover the true role these data points play. Additionally, more data on who is moving and not just where would be important information for businesses to have and would also play a role in how we define our model parameters and feature sets.

REFERENCES

1) DOMESTIC MIGRATION DROVE STATE AND LOCAL POPULATION CHANGE IN 2021 (Link)
2) Domestic Migration Trends June 2022 (Link)
3) United States Census Bureau – Net Domestic Migration Increased in Many U.S. Counties in 2021 (Link)
4) Where Are Americans Moving in 2021? (Link)
5) Mayda, Anna Maria. (2007). International Migration: A Panel Data Analysis of the Determinants of Bilateral Flows. Journal of Population Economics. 23. 1249-1274. 10.1007/s00148-009-0251-x. Link to Paper
6) Sîrbu, A., Andrienko, G., Andrienko, N. *et al.* Human migration: the big data perspective.*Int J Data Sci Anal* **11**, 341–360 (2021). Link To Paper
7) Sprung-Keyser, Ben, Nathaniel Hendren, and Sonya Porter. Working Paper. "The Radius of Economic Opportunity: Evidence from Migration and Local Labor Markets".

APPENDIX

Variable	Source	Category	Transformation
Population change	US Census Bureau	N/A	Lag
Income	Bureau of Economic Analysis	Opportunity	Pop. normalized
Wages	Bureau of Economic Analysis	Opportunity	Change
Employment	Bureau of Economic Analysis	Opportunity	Change
Employment	Bureau of Economic Analysis	Opportunity	Pop. normalized
Gross domestic product (GDP)	Bureau of Economic Analysis	Opportunity	Change
Gross domestic product (GDP)	Bureau of Economic Analysis	Opportunity	Pop. normalized
Federal transfers	Bureau of Economic Analysis	Opportunity	Pop. normalized
Home price	Federal Housing Finance Agency	Affordability	Level
Home price	Federal Housing Finance Agency	Affordability	Change
Regional price parity (all items)	Bureau of Economic Analysis	Affordability	Level
Regional price parity (goods)	Bureau of Economic Analysis	Affordability	Level
Regional price parity (housing)	Bureau of Economic Analysis	Affordability	Level
Regional price parity (utilities)	Bureau of Economic Analysis	Affordability	Level
Reported violent crimes	Federal Bureau of Investigation	Desirability	Pop. normalized
Reported rapes	Federal Bureau of Investigation	Desirability	Pop. normalized
Reported burglaries	Federal Bureau of Investigation	Desirability	Pop. normalized
Average temperature	OpenWeather API	Desirability	Level
Average rain	OpenWeather API	Desirability	Level
Average cloud cover	OpenWeather API	Desirability	Level
Number of universities	Data.gov	Desirability	Pop. normalized

Table 3 – List of features, data sources, and transformations used in the final model