Technical Report: Influences of National Geographic Migration

Prepared by Kholiswa Tsotetsi, Michael Striffler and Azzy Caceres

Rutgers University Data Science Bootcamp

January 26th, 2021

Abstract: In this report, data from the U.S. Census Bureau ACS was obtained by extracting Excel files of state to state migration, median income, and unemployment rate over the past decade (from 2010-2019). In addition to this, state median home prices were obtained from Zillow Research hub. From here, data was cleaned by dropping unnecessary columns and adding in columns for the other data files. These files were then read through a Python jupyter notebook in order to append the different yearly datasets and combine into one comprehensive dataset.

Using this file, we tested different types of machine learning techniques to hypothesize if our model would have the ability to predict the migration from one state to another using the different variables listed above. Visualizations were then created using Tableau. Ultimately, a dashboard was created where the user could input certain state criteria to predict migration as well to display our predictive model and visualizations.

Introduction

Geographic migration is commonplace. For years, people have moved from one area to another, whether it be to a different county, state or country for a myriad of reasons. Looking at the extracted Census data, it is clear that there are trends in the movement of people year by year. Therefore, the question we then asked ourselves was, what influences someone to move from one place to another.

The idea of migration peaked our interest as there have been many changes throughout the recent years that could have had an impact on this. In 2008, the market crash and subsequent economic downturn led to years of a cheaper housing market. Starting in 2018, certain state and local taxes were capped which may have influenced people to migrate to areas where this tax break was in effect. Most recently, due to travel restrictions enforced due to the current Covid-19 pandemic, a larger part of the workforce now works remotely which may influence people's decision to live in a particular state as they are not as limited. These major events made us question what factors matter to people:

1. Are people looking at the home prices in the states they are looking to migrate to?
2. Do the median income and unemployment rate of a state sway someone to move (or stay away from a state)?
3. Did the cap on property taxes cause people to move somewhere where the property taxes are lower?
4. What were the states that people decided to migrate to?

To answer these questions, we searched for data pertaining to each of the factors listed. All migration, income and unemployment data was obtained via the U.S. Census American Community Survey yearly estimates. We also incorporated home prices, which we found via Zillow Research data to obtain home prices by month and year and calculated the average. There are many other factors that we could have included in this model, but due to time constraints, we felt these would be the most valuable.

Data Acquisition

In order to have all the information in one place, we compiled all of our collected data into one universal file. We downloaded each excel file by year and manually cleaned up each

file and dropped the columns that we did not need. We then made sure our other datasets for the average home price, unemployment rate and median income matched with our state migration files and added these columns to the data set. Using an 'If' statement on Excel, we were able to create a binary option for positive and negative migration. Once we had each year file cleaned, we transferred this into a Jupyter Notebook Python file. We then used Pandas to read the files and append them together as a new dataframe to create a dataset that subsequently lists all of the years. This newly created file was then saved as an Excel file and used for our machine learning model.

Data Analysis & Visualizations

Machine Learning

Using our combined data, we created a few different models to see which would be the best fit. We used K means, Logistic Regression, Random Forest and Linear Regression. When we first started looking at the data, our dependent variable was ‘Average Home Price’. In the Limitations section we will discuss why we changed our dependent variable. The machine learning models we started with were all based on our ‘Average Home Price’ variable as the dependent variable.

With the K means algorithm, we looked at where clusters formed based on home price and net migration. We realized this illustrated areas of similar home prices and migration, but didn’t give us much insight on the impact one area had over another. With Logistic Regression, our model returned with ‘0’ accuracy and precision. Out of the previous two models we attempted, the random forest model better suited our model. Here, we used ‘Unemployment Rate’ as the dependent variable, which we were not satisfied with as this model didn’t answer the questions we initially had. Ultimately, we decided to use a Linear Regression model. With the Linear Regression we tried many combinations of dependent and independent variables. However, nothing was giving a significant R-squared which led us to believe there was no correlation.

After more consideration, the realization that the combined data we were running our model on was not picking up the states individually. It was then that we realized that using dummy variables would be necessary for our string values. Random Forest was used to see if that would have a good accuracy score, but using the Net Migration (binary positive/negative option) as the dependent variable and the other columns as the independent variables, we got a score of about 72%. We believed we could get better results with the Linear Regression model,

and attempted to run that model instead. As the dependent variable, we kept the ‘Net Migration’ and used the majority of other inputs as the independent variables. Running this model with many variables helped the accuracy of the model. Since there are multiple variables, we had to check for the adjusted R-squared. Ultimately we were able to obtain a model with an R-squared of 0.85. We then used this to implement predictions on our html page that was created on our local Flask server.

Tableau

Going back to our initial questions, we next wanted to explore how certain factors affected where people migrated to. Although our model says there is not a very strong relationship, we know that people are moving. We used Tableau to show these movements.

We read the combined data file on Tableau to pull through the information we were going to use. We created various dashboards. To begin, we showed a Timelapse map to see the migration changes year by year. We also used bar graphs to show the visualizations of the sum throughout the years of migration. We see the majority of people moved to Florida, Texas and Arizona, and moved out of New York, California, and Illinois.

To drill down into the possibilities in Tableau and to show information on a state level, we chose to look at California and Florida. There is no steady decrease or increase in migration but there was a steady decrease of unemployment and increase in home prices. Less people unemployed, gives the ability for people to buy homes, which increases demand and therefore the prices. For Florida there seems to be slightly steadier migration into Florida, but similar unemployment and home price situation as California.

Next we looked at the state and local tax deduction changes that went into effect in 2018. We see a significant increase of people moving out of states with higher home prices which leads us to believe those states would also have larger property taxes. As people lost this deduction, they may have been looking for areas with lower taxes as it was no longer considered as big of a benefit as it was before it was capped.

Lastly we reviewed the major regions between the states. As we had imagined, more people are moving south, with the majority of people moving away from the east coast. Our overall theory of people moving south was confirmed with the visualizations, although on a predictive outlook, we may have needed other factors.

Flask App

In order to run our html webpage, we utilized Flask to host our app. Here, we extracted our developed linear regression model by serializing the model and saving it by utilizing the pickle library. We then loaded our created pickle file into our deployment app python file. Here, we loaded our dependencies and created an app route to our initial html file (tableau.html). In order to utilize the data inputted by the user, we created a form and connected it to our deployment app. Here, we utilized requests to retrieve these values when the user clicks the 'Submit' button. From here, the values were inputted into a dataframe which was used to send these values into the model. The '.predict' function was utilized to predict the net migration depending on the values the user inputted. Here, the output was sent back to the html file to display the outcome to the user.

Limitations

Our main limitation was the type of data we collected. Prior to starting this project, we discussed factors we felt would influence our decision in deciding where to move. We collected all the data about these factors and found that the correlation was not as strong as we had imagined. Although we would have liked to look at demographics by state, property taxes at a deeper level, level of education or job types, there was a time constraint that did not allow us to expand our data further. Additionally, data at the state level for some of these factors simply didn't exist. Due to this, we shifted our strategy on building a working model with the data we could obtain. Our biggest relief was realizing the need to use dummy variables for our model to increase its accuracy. We had not realized how the data set was being read as one big state instead of 50 states.

Originally we started with three years worth of data. When our models were not showing correlation, we decided to add seven more years. We had been going in circles with the Machine Learning models and took much time in the data gathering process which pushed back our ability to exceed in other areas of the project.

We had also wanted to deploy to Heroku, and created the flask server to connect with a PostgresSQL, but again, due to time we decided to host on a local server to be able to provide a finished product.

Conclusion

From this project, we have gained a deeper understanding of migration, and in particular, that the factors we looked at do play a role to some extent. Using our site, users can input certain variables and have our model predict the net migration for that state in any given year. It also provides our users with the visualizations of the history of migrations. This can show what the history looks like and can consider how the factors we looked at have made an impact on people's decision to migrate to certain areas.

Ultimately, we created an HTML webpage that runs through our Flask server that holds all our findings. Despite our limitations, we successfully produced interactive visualizations of our extracted results.