**Predicting US Domestic Migratory Patterns, Post-Pandemic**

**White Paper**

Donald Price

**Business Problem & Background**

The COVID pandemic of 2020 changed people’s lives. Lockdowns occurred in most of the Unites States forcing many employees and students to work from home. This arrangement persisted for months. Companies are allowing people to perform jobs remotely or in a hybrid environment, and employees are using this benefit to consider changes to their living situation. With fewer employees tied to a specific geographic location, many are opting to relocate for family, enjoyment, financial, or just a change in scenery.

The hypothesis for our project is to predict migration of the US population by county and explain the factors that influence migration patterns.

**Stakeholders & Goals**

There are two groups who could benefit from a model predicting migration: government officials and corporations.

Understanding population changes is crucial to understand the locality’s revenue base. Governments need to model local population changes to predict tax collections and understand effects on government spending budgets. Government officials would also seek data on factors that influence population change so they can alter budgets to improve the attractiveness of their location to prospective citizens.

The business problem is also useful to the private sector. Corporations in the private sector seek predicted demographic changes by location, which could be used to help home builders, retail developers, and other firms maximize store expansion by focusing on growing areas.

**Data Explanation**

The United States Federal Government (USFG) publishes statistics on domestic migration. Domestic migration data from April 2020 through July 2021 will be used as our target variable as it most closely represents the period immediately following the pandemic and shutdown. The study will be conducted at the county level.

Many factors may contribute to a person’s decision to move. The Indiana Business Research Center (IBRC) at Indiana University publishes health, economic, and quality of life statistics for each county. The IBRC data set will be joined to the census data at the county level and will form over 100 features for each county. See Appendix for full list of features and the Data Preparation section for information on transformations and variable reductions.

**Modeling/Analytical Approach & Methods**

The business problem is a regression problem. The model needs to predict the percentage change in migration, positive or negative, for each county. Models will be evaluated using the R2 value. Multiple linear regression, ridge regression, and the random forest regressor will be used in modeling.

A good model should produce an R2 of 0.30 or better. Ozili suggests an R2 exceeding 0.10 is acceptable for social science uses assuming no multicollinearity of variables (Ozili 2022). Correlated variables were removed during data preparation to avoid issues with multicollinearity. Dean Abbott further supports our R2 goal; he suggests an R2 of 0.3 may be deemed “excellent” for social sciences (Abbott 2014, p.301). The random forest regressor outperformed other algorithms and produced a R2 of 0.55, exceeding our goal.

**Analysis**

The percentage of a population with social security income is highly correlated with migration (0.41 correlation as shown in the line below). This variable identifies retired individuals. Retired individuals are unlikely to move, and presumably individuals soon to retire are likely to move to county with other retirees.

**Exhibit 1: Census Migration Rates, by Social Security Recipients**

Chart, scatter chart

Description automatically generated

The IBRC data contains other features that are independently correlated to migration. Variables with a strong positive or strong negative correlation are noted in the exhibit 2 below. These correlations are independently run and may not be included in the final model when all variable interactions are included.

**Exhibit 2: Top Features Correlated to Migration**



Temperature is positively correlated to migration, which suggests people prefer to live in warmer conditions. This may or may not relate to an older population that appears to migrate more often than younger populations. Income also appears to be correlated. Higher income, as denoted by tax returns that contain business income and charitable deductions, see higher migration. Conversely, counties with a higher poverty rate and SNAP benefits see reductions in population over time. Modeling will explore the contributions of these variables, both independently and collectively. These specific features will be incorporated as one of our modeling approaches.

**Conclusion**

Our tuned random regressor model generated an R2 of 0.55, which significantly exceeded our goal. The chart in exhibit 3 identifies the most significant factors influencing domestic migration.

**Exhibit 3: Top Features in the Random Forest Model**

Chart

Description automatically generated

Our stakeholders expect explanations behind the predicted migration. The SHAP summary chart in exhibit 4 explains which variables are influential for migration. SHAP charts can be produced for each county to explain each factor and the influence each feature has for a particular county. Local factors will be explained in the implementation section.

The final model produced several interesting insights. Counties with a high percentage of social security recipients tend to attract the most people to the county. These counties may contain retirement communities where people seek to live post-employment. People are also moving to areas with higher average temperatures. Having a county offering strong amenities to citizens (natural amenity scale features) increases the county’s ability to attract migrants. Migration wasn’t only for pleasure; the fourth-most significant feature (exhibit 3) was employment growth. Counties attracting workers were much more likely to see sustained increases in people moving to the county.

Lastly, what is most interesting is what did not come up as significant. Although the news often reported people moving to more rural areas, we did not see much evidence of this in the model. Our population feature did not come up as significant in any model, and the only slight indication that people may be leaving cities is the 8th most significant feature stating those counties that have a higher percentage of people walking to work experienced some attrition of citizens. This would not seem to be enough to support a flight to rural areas.

**Exhibit 4: SHAP Explanations of Features**

**A picture containing text

Description automatically generated**

**Assumptions**

The model assumes current causes of migration will continue. Our migration period and model are based on a significant period of change in the United States. If pandemic risks abate or companies reduce remote roles, forecasts from this model may become less accurate and may require the model to be retrained. Another factor may be changes in the percentage of population that is at or near retirement, which is our largest variable.

**Limitations & Challenges**

Many factors influence what would cause someone to migrate, and the complications of the many different drivers of migration make identification of significant factors difficult. As we saw in the SHAP values for exhibit 4, counties have existing population bases that differ and cause counties to factors to vary by county. Although our model is deemed strong by social standards, there is still quite a bit of unknown influence on migration not captured in our model.

**Implementation, Future Uses, & Recommendations**

County government leaders would best gain access to this model through a published dashboard that identifies key attributes and influences of those attributes on the model. Describing a key influence and a county’s statistic against the country average would help a government office to decide how best to act on the model’s insights. Exhibit 5 shows key influences for Jefferson County, NE. Migration predicted to be negative, mostly because the county’s amenities are below national averages. Temperature is also a negative in Nebraska, but that is not something the county can alter.

**Exhibit 5: SHAP Explanations for Jefferson County, NE**

**A picture containing graphical user interface

Description automatically generated**

Businesses would probably prefer an API where they can access county-level predictions of migration. This would enable a business to assess which growth counties to target. Factors affecting migration are also likely desirable from the API as it contains information about the demographic changes of the population that may further enable market segmentation.

**Ethical Assessment**

Few ethical issues are present with this project. The census data and IBRC data is anonymized and made available for public consumption. No data disclosure risks are present with this information.

Uses of this information are mostly ethical. One consideration might be lower income areas where investment may suffer. There are features in the data set that identify income and education levels, and if the model suggests investment here is less profitable for corporations, these areas may suffer further economic difficulties. Limitations on providing these predictions to companies should be weighed against creating economic opportunity Alternatively, this may necessitate national government officials using this data to identify areas needing public funding if private funds were reduced.

**Appendix**

**References**

Abbott, Dean, Applied Predictive Analytics, 2014.

Axios, (July 2022). Census Data Reveals COVID Population Shift.

https://www.axios.com/2022/07/01/population-shift-west-sunbelt-census.

Cambon, Sarah Chaney, June 28, 2022. Rural counties are booming, but can it last?. *Wall Street*

*Journal.* https://stocks.apple.com/AuplFwjoBTk-wJRaOc7XjZw.

Indiana Business Research Center (2020). Metrics for Development. *StatsAmerica.*

https://www.statsamerica.org/downloads/default.aspx#research.

Ozili, Peterson K, The Acceptable R-Square in Empirical Modelling for Social Science Research

(June 5, 2022). Available at

SSRN: <https://ssrn.com/abstract=4128165> or [http://dx.doi.org/10.2139/ssrn.4128165](https://dx.doi.org/10.2139/ssrn.4128165)

United States Census Bureau. County Population Totals: 2020-2021.

https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-total.html.