Final Report Team 30

James Domingo, Michael Jones, Amandeep Singh, Kyle Lindteigen,
Paul Horton, Yogesh Raparia

April 24 2021

I. Background/Introduction

The demographic profile of the US is rapidly changing. Current projections have the US population reaching 438 million in 2050 [1]. With this growth, key features of US states and/or communities such as population, cost of living, crime rate, and unemployment are expected to drastically change. It is therefore important to understand the current relationship between communities with respect to these features and to understand how these relationships are predicted to change in the future. Significant work has been done to predict individual features such as population, racial makeup, and demographics; however, there has been little work that attempts to combine a wide range of features into a singular tool [2][3][4]. The methods involved in current predictions range from assumptions based on current patterns to machine learning methods such as SVM [5] and Regression [6], as well as statistical models such as ARIMA [7][8].

The goal of this project is to provide users with an easy-to-use and interactive tool that allows them to visualize key trends in various communities, from past and current behavior to forecasted outcomes. Local governments can use the tool to identify areas that need more attention in their community – for example, which areas in their community are most accident-prone and what are the factors contributing the most to these accidents. The general population can also benefit greatly from this tool – for example, a family seeking to move may want to compare the average rent prices in Georgia to that in Texas over the next 5 years.

II. Problem Definition

Knowing how key features of the US community such as population, income, cost of living, crime rate, and safety change is critical in decision making for both local government units and the common population but current efforts to collect this information into a comprehensive tool have been limited. This project aims to: (1) forecast these key features through the appropriate statistical and machine learning (ML) models, and (2) consolidate these forecasts, as well as the most important factors affecting these forecasts, into an interactive and visual tool to aid our stakeholders in decision making.

III. Literature Survey

Demographic information and corresponding change tracking are important across different domains such as commercial and industrial applications, real estate development, and urban planning. As such, demand for tools that visualize these data while maximizing human-centered design elements and include interactive features [9]. Cities such as Dublin [10] and Atlanta [11] have found dashboards displaying city performance to be useful. Such information included population, housing prices, crime, income, power consumption, and resources such as apps for services like transportation. Pu et al also list features important to tourism and settlement such as image, culture, traffic, and economic development [12].

Forecasting is difficult because demographics are dynamic and can be influenced by external factors which obfuscate models developed on an assumption that the system is static. Population trends, for instance, can shift due to increases in life expectancy or prevalence of diseases [13]. Additionally, time-series data tends to be aggregations of more volatile elements which masks inherent short-term trends [4]. Some methods have developed which use dynamic models weighting variables differently throughout time [14].

Numerous ML algorithms have been adapted to predict demographics over the years. Aside from typical regression models, the application of SVM (such as in housing prices [2] and temperature [4]), as well as ensemble methods such as Gradient Boosting [13] have been shown to improve accuracy. Nonetheless, Makridakis et al have shown that these ML methods are still dominated by statistical time-series forecasting in terms of performance [7]. Auto-regressive methods have proved to be suitable for trends such as in population [8] and crime [15], and for cyclical systems such as the weather [4][16].

Several works have seen positive results in combining these statistical forecasting methods with ML algorithms. Rapach and Strauss have shown that these models in combination with other predictors have worked well in forecasting housing prices [2]. Wang et al [3], as well as Chi and Voss [6], have also noted how network analysis or considering events that have happened nearby can boost prediction results. Qiokata and Khan have shown that artificial neural networks are capable of outperforming auto-regressive methods

[17]. Dubin shows the impact of a "neighborhood effect" in which how correlations with correlated data can be incorporated in modeling, such as pricing of neighboring houses can affect predictions for housing prices [18]. However, one challenge of working with demographic data is the availability of similarly relevant data. Additionally, smaller geographic regions do not necessarily record the same metrics as others.

IV. Proposed Method

Given our audience and the weight of decision making we aim to cater to (e.g. relocation, local government decisions), it is imperative that our tool uses methods that are better than current state-of-the-art methods. To achieve our objectives, this project focused on developing novel methods on 2 key aspects: modeling and visualization. These are explained in depth in this section.

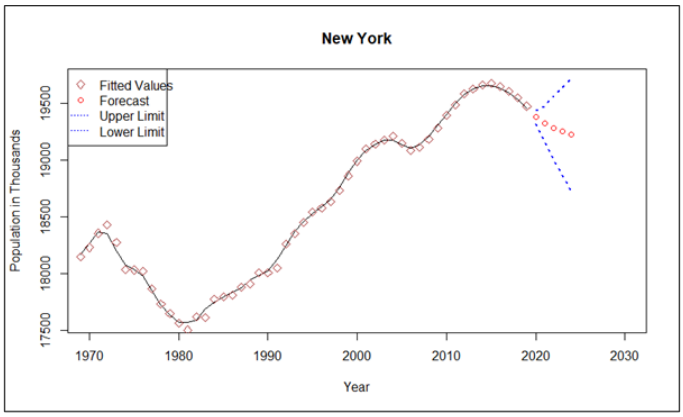
Modeling: Time-Series Data

The goal of this aspect of the project is to fit predictive models to states' historical data so that they can be used to forecast future values of interest that will be displayed by our visualization tool. The type of model we decided to use is an ARIMA model because our research found that ARIMA models are can successfully forecast different trends [8] and because of their general ease of use since we will be building many models.

ARIMA stands for auto regressive integrated moving average and is used to explain the future value of a variable based on its past values. An ARIMA model is characterized by its order (p, d, q) where p represents the autoregressive term, d represents the number of differencing applied, and q represents the moving average term. The autoregressive component uses previous observations as input to predict the next observation and the selected order indicates how many previous observations or lags to consider as input. The moving average component uses the error from previous forecasted values as input and the selected order again indicates how many lags to consider. Finally, time series data is more successfully modeled when the data is stationary, that is the mean and variance of the data do not depend on time. In order to convert a non-stationary time series to a stationary time series, one must take the lagged difference of the series. The number of differences necessary to convert the process to stationary is the difference order in the ARIMA model. The equation below shows the final ARIMA model where β represents the AR coefficients for the Y variable lagged to order p and ϕ represents the MA coefficients for the error term ϵ lagged to order q.

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q}$$

For our project, we fit ARIMA models to our datasets for population, temperature, income, and housing price indices. We used an iterative process for model selection. First a model was fit with order (1, 0, 0) and the resulting AIC was stored. A new model was then fit for each order up to (6, 1, 6) for a total of 37 unique models. The parameters that gave the lowest AIC value were determined to be the optimal order. This process was performed on the data for each state giving a total of 50 different ARIMA models for each feature. An example forecast including prediction error is shown below for the population of New York for the next 5 years.

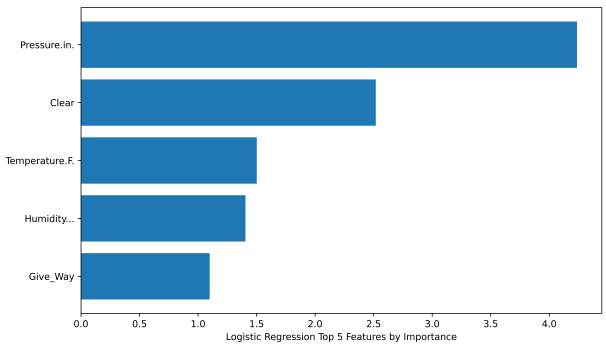


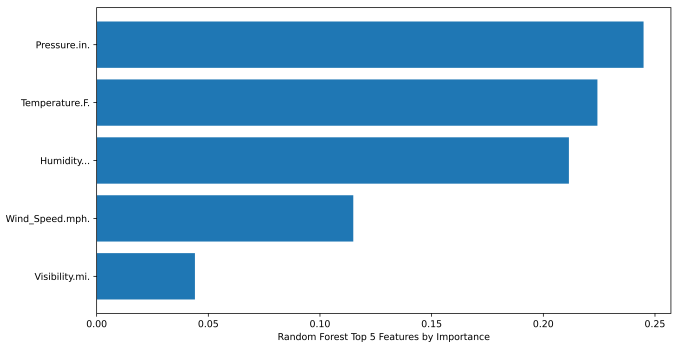
Limitations to this type of modeling is that the prediction error significantly increases the further out the prediction is. This is seen by the upper and lower limits increasing in magnitude. Additionally, because this model is only based on the previous observations, it does not respond well to shocks or quick changes in trends caused by external factors such as a pandemic.

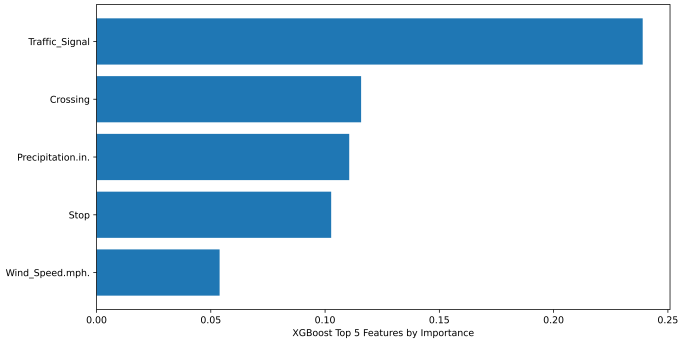
Modeling: Feature Data

We have developed initial models for the US road accidents dataset [19][20], which is a large dataset containing 4.2M observations from 2016-2019 and 49 variables. Each observation shows information on the location of the accident, its severity (with 4 levels), and several factors describing the accident. For this data set, instead of forecasting the number of future accidents or predicting the likelihood of an accident, we focused on identifying the most important features for determining the severity of an accident for each state. This will help state governments get a better idea of which factors influence accidents the most and thus work on improving road safety.

We tried Logistic Regression, Random Forest, and XGBoost algorithms to determine the importance of each feature. While these methods are already known to work well in classification tasks, we believe that we have brought bring innovation in its application on road accident prediction – most previous studies have only considered modeling accident data in a specific city or state. Our project scope and purpose require us to model the dataset for the entire US, which brings additional complexity not just due to the size of the dataset, but also due to the differences in data across states.





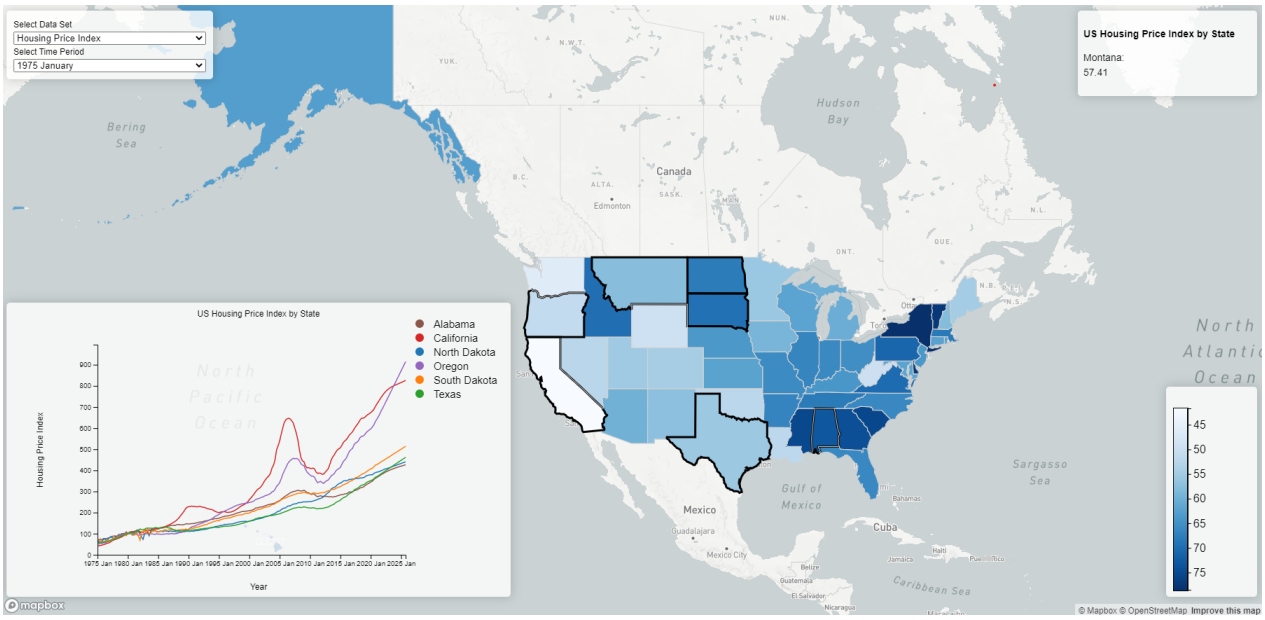


Nonetheless, training a model on such a large dataset takes up too much time and memory so we first selected the best model by training on a subset of the data (1M observations). Model validation results point to the Random Forest model being the most effective, with an OOB error estimate of only 28.7%. Using this algorithm, we then trained an individual Random Forest model for each state and measuring performance on test sets containing data for each state. This approach introduces several benefits: model development has become more manageable, and each model is better fit to the true conditions of each state, becoming more informative to their respective audience. Additionally, being an ensemble model, a Random Forest works reasonably well even in states with the fewest available data.

Visualization

For the data visualization aspect of the project, we focused on using an HTML page with d3 to create graphs and used an api called mapbox in order to build an appealing map. Our goal with the visualizations were to make something that allows for the user to interact with the data in meaningful ways such that they can gain insights into each of the states. The two primary components of the html page that allowed for this were the mapbox api and the d3 library. The d3 library was primarily used for the creation of the graph that allows for users to directly compare the different states and their trends over time for each of the provided data sets. It was also used to create the scale and color choices so users can easily compare lines on the graph and data on the map.

The bulk of the work involved transferring the choropleth from d3 to mapbox. The objective was to keep the same functionality as the original d3 version while improving the visuals and interactivity of the map making for a better user experience. Mapbox provided useful tools that allowed for the creation of the general map and use of a geo json in order to layer the choropleth onto the world map. There are also useful tools to help with interactivity that allowed us to keep the functionality so that when you hover over a state you get the current value of that state. It also allows for on click functionality so we could create the graph and highlight the state on click. All this is possible while being able to zoom in on the map, zoom out, and move around it while keeping the integrity of the different layers, presenting the data, and allowing for easy switching between data sets.



Our method is novel because it includes three combined features missing in current alternatives: aggregation of data, predictions, and a visualization. Without our tool, if an individual needs to understand how states differ, they must comb through multiple government organizations. Even if they aggregate the data, it is not easily digestible as the amount of information is overwhelming when analyzed in tabular form. We created a visual with intuitive features to aid the user in their search. Additionally, historical data only tells part of the story. We used predictive modeling to enable users to consider future implications in a dynamic environment.

V. Design of Experiments/Evaluation

The key analytical aspects of our project are modeling and visualization. Thus, we evaluated the success of our project in these two aspects through accuracy and end-user satisfaction, respectively. The following discusses the experiments and tests we performed for each metric:

Model Accuracy

For the prediction accuracy, we have divided our data into train and test sets. The ARIMA models use the last 5 years as the test set with all other years as the training set. Our table with results for Root Mean Squared Error on all predicted variables is below. The Random Forest model indicated through the variable importance output that the presence of a traffic signal was the most significant variable for accident severity in most states. Overall, we have gotten considerably accurate results which are comparable, and some even superior, to similar models in our literature survey. It is to be noted that state trends for 2020 differed markedly from previous years due to the pandemic and this understandably is where most of the error comes from for our models.

Metric	Population	Income	Temperature	Housing Price Index	Jobs Added
Mean	6,366,000	59,600	64.1	418.6	183.9
RMSE	1,323,000	6,230	1.4	5.7	0.7
RMSE as % of Mean	20.8%	10.5%	2.2%	1.4%	0.3%

End-user Satisfaction

With the final output of the project being an interactive visual tool, we wanted to get the feedback of users regarding quality of information and usability. We had respondents test out a prototype of the tool and answer a short survey, with the context that they are looking to move into another state. This survey

included questions on three main areas: (1) quality of modeling / forecasting, (2) features included, and (3) visualization and usability. Respondents have provided positive scores and have given insightful feedback on possible future improvements on the tool, as well as other useful features to consider. These include adding tooltips for other statistics for easier visualization, providing a detailed breakdown for some features such as population by age, and adding other features such as crime rate, cost of living, salaries, and taxes. All of these can be easily implemented as we spend more time developing the tool in the future. The table below shows a summary of questions included and average scores received.

Score Scale: 0 = lowest; 5 = highest

Key Aspect	Question Topics	Average Score
Modeling / Forecasting	Sufficiency of Forecast Coverage, Accuracy / Trustworthiness of Forecasts, Quality of Information Included	4.0
Features	Scope and Relevance of Features Included, Feature Importance to Decision Making	4.0
Visualization	Ease of Use, Display Quality on Screen, Intuitiveness, Usefulness	4.4

VI. Conclusion

Our completed tool has the features to enable a user to build a deeper understanding of regional demographics. The accuracy of our predictive models is satisfactory despite significant impacts in 2020 due to the pandemic. The initial user feedback was positive indicating that we have included features and functions that are relevant and helpful.

To go further in this project, we would evaluate the more specific applications to see which would be more promising for this tool. An employee in the Department of Housing and Urban Development may be interested in additional metrics for unemployment or the number of homeless people. This tool could be refined for government officials at a state level where these metrics are shown for each county within their respective state. We could also expand this tool for more public use with additional metrics for individuals like employment by sector, crime rate, or tax rates.

VII. Distribution of Work

All team members contributed equally to the project.

Bibliography

[1] Passel, J. S., & D'Vera Cohn, D. (2008). US population projections, 2005-2050 (p. 20). Washington, DC: Pew Research Center. link

[2] Rapach, D. E., & Strauss, J. K. (2007). Forecasting real housing price growth in the eighth district states. Federal Reserve Bank of St. Louis. Regional Economic Development, 3(2), 33-42. link

[3] Wang, H., Kifer, D., Graif, C., & Li, Z. (2016, August). Crime rate inference with big data. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 635-644). link

[4] Aghelpour, P., Mohammadi, B., & Biazar, S. M. (2019). Long-term monthly average temperature forecasting in some climate types of Iran, using the models SARIMA, SVR, and SVR-FA. Theoretical and Applied Climatology, 138(3), 1471-1480. link

[5] Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the US real house price index. Economic Modelling, 45, 259-267. link

[6] Chi, G., & Voss, P. R. (2011). Small-area population forecasting: Borrowing strength across space and time. Population, Space and Place, 17(5), 505-520. link

[7] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS one, 13(3), e0194889. link

[8] Dai, J., & Chen, S. (2019, October). The application of ARIMA model in forecasting population data. In Journal of Physics: Conference Series (Vol. 1324, No. 1, p. 012100). IOP Publishing. link

[9] Lock, O., Bednarz, T., Leao, S. Z., & Pettit, C. (2020). A review and reframing of participatory urban dashboards. City, Culture and Society, 20, 100294. link

[10] McArdle, G., & Kitchin, R. (2016). The Dublin Dashboard: Design and development of a real-time analytical urban dashboard. link

[11] Edwards, D., & Thomas, J. C. (2005). Developing a municipal performance-measurement system: Reflections on the Atlanta Dashboard. Public administration review, 65(3), 369-376. link

[12] Pu, Z., Du, H., Yu, S., & Feng, D. (2020, February). Improved Tourism Recommendation System. In Proceedings of the 2020 12th International Conference on Machine Learning and Computing (pp. 121-126). link

[13] Şahinarslan, F. V., Tekin, A. T., & Çebi, F. MACHINE LEARNING ALGORITHMS TO FORECAST POPULATION: TURKEY EXAMPLE. link

[14] Bork, L., & Møller, S. V. (2015). Forecasting house prices in the 50 states using Dynamic Model Averaging and Dynamic Model Selection. International Journal of Forecasting, 31(1), 63-78. link

[15] Rattner, A. (1990). Social indicators and crime rate forecasting. Social Indicators Research, 22(1), 83-95. link

[16] Eni, D. (2015). Seasonal ARIMA modeling and forecasting of rainfall in Warri Town, Nigeria. Journal of Geoscience and Environment Protection, 3(06), 91. link

[17] Qiokata, V., & Khan, M. G. (2015, December). Modeling emigration of Fiji's population using Artificial Neural Network. In 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE) (pp. 1-8). IEEE. link

[18] Dubin, R. A. (1998). Predicting house prices using multiple listings data. The Journal of Real Estate Finance and Economics, 17(1), 35-59. link

[19]] Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409

[20] Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019, November). Accident risk prediction based on heterogeneous sparse data: New dataset and insights. In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 33-42)