ACENET Microcredential in Advanced Computing ISP Report

Project title: Predicting Community Growth Using Census Data and Road Network Characteristics

Participant name: Dane Sheppard

Date: July 31, 2024

Abstract:

This project was an investigation into using machine learning to predict community population change between census periods. A random forest regression model predicted well and showed R2 values as high as 0.99 in some cases, although prediction capability underperformed for higher population communities.

1. Introduction

Population forecasting is a complex and time consuming task that is typically done through a Request for Proposal[1] process. Applying and calibrating the typical models used for this process requires considerable expertise, and as a result is not accessible to most of the population. This project is an attempt to provide a method for predicting population change that would be faster and easier to use than the current state-of-the-art approach, while maintaining reasonable predictive capability.

The work focused on three research questions (RQs):

RQ1: Can census data from different census years be used to train a model to predict population change over time within a limited geographical scope (eg. provincial scale)?

RQ2: Does the inclusion of geospatial characteristics (road network characteristics, building density, population density, etc.) improve the prediction capability of the model from research question 1?

RQ3: Do prediction capabilities persist when geographical scope is expanded to several provinces, or to a Canada-wide analysis?

2. Background

Forecasting how a community will change over time is a critical part of community planning[2]. Infrastructure changes, development of major services and amenities (schools, hospitals, etc), and the scaling of hydro and electrical grids are all projects that can take years to come to fruition, and without proactive investment, a community may find it difficult to meet the needs of a changing population. This project is a step towards a tool to empower communities to engage in proactive, data-driven decision making with minimal cost and time investment required.

3. Analysis

In order to investigate the research questions, datasets were required for census data across multiple years. The original goal was to programmatically download data across multiple years; 2011, 2016, and 2021. Statistics Canada provides APIs for census data for 2016 and 2021, but the syntax is different for each, which makes it cumbersome to retrieve the data for both, and the 2011 data had no API provided. It proved more expedient to manually download data for each. For RQ2 in particular, a dataset of geospatial characteristics was also required. Initially, several different dimensions of geospatial characteristics were considered, but investigation of the available open geospatial data (primarily openstreetmap.org, as the most readily available data source) indicated that only road network data would be suitable for any province-wide analysis for a project of this scope.

3.1 Census Data Preprocessing

The census data was collected by downloading the collection of census subdivisions from Statistics Canada for each province. The data had hundreds of features for each community organized in a stacked data structure, and data across years had mismatched characteristic names, miscellaneous special characters that python had trouble parsing, and a different syntax for naming communities. Preprocessing of the census data focused on pivoting the data to a record format so each community would have a list of features, and ensuring that the datasets were compatible.

3.2 Road network preprocessing

The road network data for New Brunswick was retrieved and processed using the python library OSMNX. The library was able to search the OpenStreetMap database by community name and extract the desired features when a match was found. Some additional encoding was required for some of the road network measures to provide a format suitable for machine learning. See file ./dataScraping.ipynb in the Github repo for details.

3.3 Exploratory Data Analysis

After data preprocessing, an exploratory data analysis revealed many columns with very limited data, including hundreds of columns about language that had only a few valid entries in the entire dataset. This prompted some feature elimination, using keywords thought to be more relevant to population change. The remaining features in the census data and extracted road network characteristics followed expected trends, showing near-normal distributions across much of the dataset. See file ./dataScraping.ipynb in the Github repo for further details.

3.4 Machine Learning Model

A model needed to be selected to predict a numerical output, given a large number of features. The two standout candidates for machine learning algorithms were Random Forest Regression and Histogram Gradient Boosting (HGBT). Both offer reasonable performance in terms of prediction capabilities and computation time for this type of model, but Random Forest Regression was ultimately chosen because it is considered to be more resilient to overfitting, particularly when there are a large number of features[3]. Models were trained on NB census data, NB road networks, and Canada-wide census data, to predict community population in 2021, in various configurations. See files ./growthPredictionNB.ipynb, generalizedPopGrowth.ipynb, and ./stressTest/stressTesting.ipynb in the Github repo for details.

3.5 HPC and Scaleup

HPC was used to scale up the census-related analysis done in NB to a Canada-wide analysis. The data processing and analysis were automated using three workhorse scripts that were run on ACENET's Siku cluster. The file ./pythonScripts/censusProcessing.py processes the provincial census for use with Random Forest Regression. The file ./pythonScripts/dataAssembly.py assembles the processed provincial datasets into a single file for all of Canada. The file ./pythonScripts/popGrowthModelTraining.py imports that data in a Pandas dataframe, then runs training, hyperparameter tuning, and analysis using Scikit Learn. See the Github repository for details.

4. Results

To answer RQ1, several tests were conducted on the NB-only model. Goodness of fit was evaluated primarily using R^2 and Root-mean-squared error (RMSE). The original prediction against the validation set showed an R^2 of 0.89, indicating a reasonable explanation of the variance in the dependent variable, shown in figure 1 (left) below. However, the RMSE was 3130, quite high when compared to the mean of 2562. Interestingly, the NB-trained model scored better when predicting using a 2016 dataset for NL, shown in figure 1 (right). Several other experiments with other provinces showed a good-to-excellent prediction capability for training and predicting with provincial datasets. See the folder ./growthPredictionNB and the visualizations in ./figures for further details about this process.

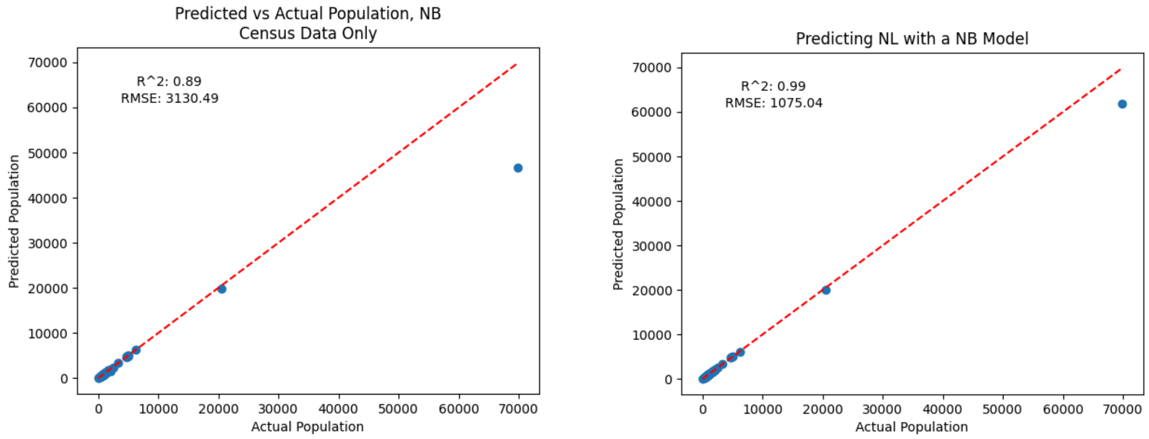

Figure 1: Predicted vs Actual for NB (left) and NL (right) using NB-trained model

A second analysis was run using road network characteristics in an attempt to answer RQ2. Training a model on the network data only, however, had an R^2 of -26, indicating that the model did worse than a random distribution around the mean. This is not surprising, as the road network data only had only three features for each community, and a small number of data points. Combining the road network data with the census data set and retraining gave better prediction results, with an R^2 of 0.98 and a RMSE of 255. See ./growthPredictionNB.ipynb for further details.

The third research was evaluated in several steps. Training, hyperparameter tuning, and validation done on Siku gave an R^2 of 0.96 and an RMSE of 5346. The variance is well explained, but once again the RMSE is quite high compared to the mean of 5137. This is likely due to the presence of a small number of extreme outliers (Toronto, Vancouver, etc.). The model tends to underestimate these high population cities, and a few errors in the order of 10^6 skew the mean and the RMSE considerably. To counteract this, additional training and tuning was done for models using subsets of the data; one model using only populations less than 100k, and another using populations less than 30k. After all models were trained, each was used to predict population change for every province in Canada as a subset of the larger dataset. Results for Ontario are shown in figure 2 below as an example. Discussion of these results will be done in the following section.

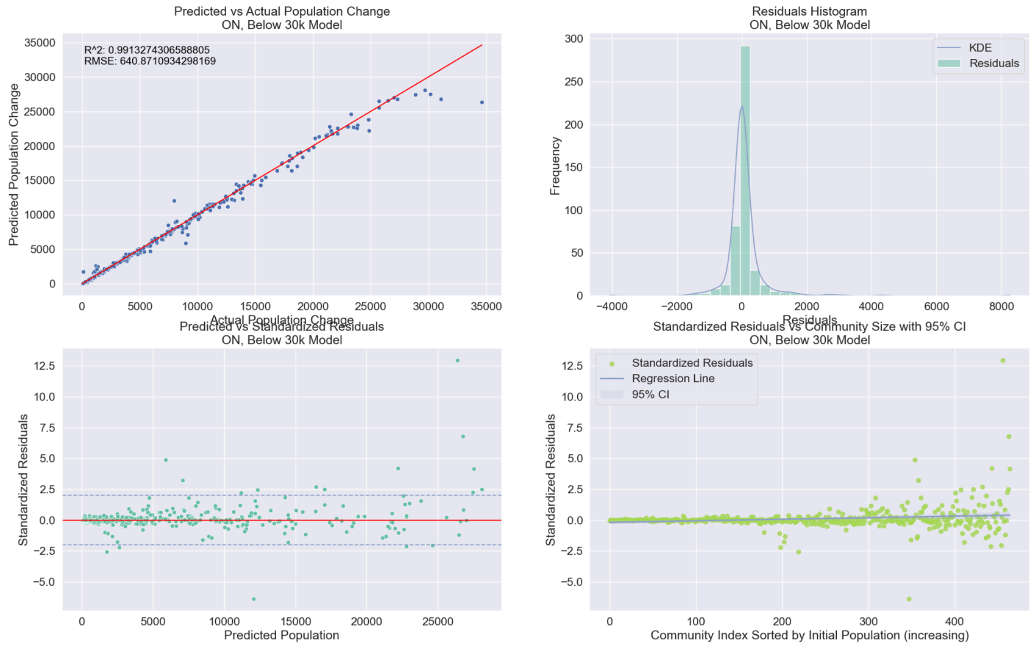

Figure 2: Prediction results for ON trained on populations <30k

The feature importances for the various models were also recorded in script outputs and visualized using bubble plots similar to the one shown in figure 3.

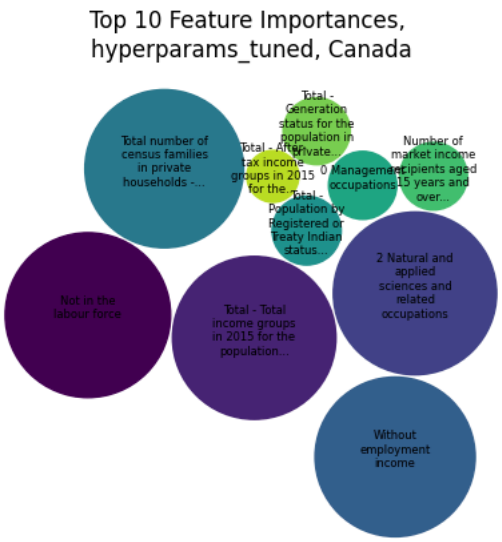

Figure 3: Bubble plot of feature importances for full data model

These plots show feature importances in a bubble sized for their level of importance, with larger bubbles indicating a more important feature. Note that some feature names are quite long and get truncated by the visualization. The trends in feature importances will be discussed in the following section. For more bubble plots in the various models, see folders ./figures, ./sikuOutput, and ./stressTest/figures for files of the format feature_importances_bubble_*.png. For the prediction outputs for each province in the final stress testing, see ./stressTest/figures.

5. Discussion

The prediction results for the NB-only dataset, particularly when used to predict NL population change with such high accuracy, was deemed sufficient to answer RQ1; training a model on census data and using it to predict population change was indeed possible on the provincial scale. The model tended to underpredict growth in larger communities, which persisted in the other areas of the project.

The combined dataset of NB census data and the road network characteristics gave an R^2 of 0.98 and a RMSE of 255. This is an excellent result, although it's not clear if the road network data caused this directly, or if it was a result of a change in randomization sequence through using the same seed, but a different sized dataset. The result for RQ2 was inconclusive. The inclusion of road network data may have been helpful, but was likely superfluous when using the full census dataset, as prediction capabilities were already quite good. This question requires further investigation.

Prediction using the Canada-wide dataset showed similar issues with large population communities, as with the provincial scale analysis. Estimates were skewed further due to multiple communities with populations greater than 1 million people. However, trimming the dataset to communities less than 100k people showed excellent predictive capabilities in 9/10 provinces, and ⅔ territories. The smallest model, less than 30k people, showed excellent prediction capabilities for all provinces and territories. This is a positive result for RQ3, and an indication that prediction capabilities scale quite well. This also implies that there may be similar leading indicators for population growth in the majority of Canadian communities, which bears further investigation in the future.

Feature importances showed a number of interesting trends. For example, the number of management jobs showing up as the most important feature across several iterations of the models. The most important features in general were consistently related to income levels, number of people of an age to be in the labour force and having a family, and employment industry. One hypothesis is that the employment and income features are serving as a proxy for the health of the local economy in a given community, which may be a good predictor of future growth.

Several challenges were faced during the exploratory data analysis in particular. The inconsistency in characteristic labels and lack of a reliable API resulted in a lot of labour cleaning, rearranging, and renaming the census datasets, and restricted the scope of the project to only 2016 and 2021 datasets. Collection of the road network characteristics was also quite time consuming, and the data was generally more sparse than the census datasets, although OSMNX was quite reliable overall. The consistent underperformance of the models for larger communities also presented some challenges, although this may simply be the limitation of the model for this type of prediction. One might expect larger communities to have more influences on population change that are not represented in census data.

Conclusion

The project provided strong evidence that accurate population forecasting is possible using a machine learning approach. This is an encouraging step towards a powerful, accessible tool for population forecasting and data-driven community planning.

Future work on this topic could proceed in several directions. It would be interesting to see how predictive capabilities changed after recursive feature elimination, to prune the census data down to some number of high-importance features. This may also allow the road network characteristics to be more influential as predictors. There is also an opportunity to pair ML-based population forecasting with machine-vision based geospatial change estimation, for a more comprehensive forecasting model. Finally, it would be worthwhile to expand the scope of this analysis when the 2026 census is released, to see if predictive capabilities change when 2016, 2021, and 2026 census datasets are used in tandem.

References

1. RFP for Population and Economic Forecasting for Tillsonburg, ON. https://pub-tillsonburg.escribemeetings.com/filestream.ashx?DocumentId=21723

2. Chi G, Wang D. Population projection accuracy: The impacts of sociodemographics, accessibility, land use, and neighbour characteristics. Popul Space Place. 2018 Jul;24(5):e2129. doi: 10.1002/psp.2129. Epub 2017 Dec 21. PMID: 30140176; PMCID: PMC6100728.

3. Comparing Random Forests and Histogram Gradient Boosting models. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_hist_grad_boosting_comparison.html

Supplementary Materials

The github repository for this project can be found here. The graphs for work related to RQ1 and RQ2 are available in the ./figures directory, HPC results are stored in the ./sikuOutputs directory, and the outputs for final testing of the model are available in ./stressTest/figures. The raw census datasets can be found on the Statistics Canada website:

- 2011 Data
- 2016 Data
- 2021 Data

Processed versions of each provincial dataset are available in the ./processedData directory of the github repository.

Road network data was scraped from https://www.openstreetmap.org/ using the OSMNX python library.