County AGI Prediction

Carl Barbee
NYU Courant Institute of Mathematical Sciences
New York City, USA
crb616@nyu.edu

Mike Urciuoli
NYU Courant Institute of Mathematical Sciences
New York City, USA
mlu216@nyu.edu

Alex Spence
NYU Center for Data Science
New York City, USA
ajs811@nyu.edu

***Abstract*—Our objective was to employ a linear regression model that predicts the average Annual Gross Income (AGI) of a county for the year 2017 using features gained from three US government data-sets—demographics, educational attainment, and types of businesses by NAICS—for the years 2011 to 2016. We evaluated the accuracy of the model by comparing it with the Internal Revenue Service's (IRS) Annual Gross Income (AGI) data-set for 2017. 91% of the model's predictions come within 30% of the expected value.**

***Index Terms*—analytics, big data, annual gross income, demographics**

I. Introduction

There are multiple use cases for knowing income level of an area. Investors use it to drive business decisions, policy makers and social workers use it to track gentrification and as a factor in calculating socioeconomic status. The IRS releases very detailed information on the income of each U.S. county, however they lag by a few years, so current income level can only be approximated by analyzing correlated features.

Machine learning has a natural fit for this problem as long as the feature set is robust enough for the given areas. For some this is easy to do if the local government prioritizes data availability; an example of this would be NYC Open Data. When trying to develop a national model, however, where information is required for every U.S. county, the datasets that can be used are limited and often proprietary. The U.S. Census is a valuable resource when it comes to data on a national scale, and although it lags in releasing data like the IRS its data can be used to form a historical model. Having a historical model would simplify the problem of predicting present day income levels to just gathering up to date information that describes the necessary features of an area, e.g., using Yelp to approximate number of businesses in an area [1]) that are important to the model.

In this study we aim to predict the average Adjusted Gross Income (AGI) of a county based on demographics, educational attainment, and business counts by type. AGI is just gross income with specific tax deductions taken out. Average AGI specifically is regularized by population, since total AGI can be increased simply by adding more tax payers, so we view average AGI as a good indication of the affluence of an area.

The machine learning technique we chose to deploy for our prediction model is Linear Regression since it has a number of useful properties. First, it is capable of making predictions where the target output is continuous. Second, the model is easily explainable by analyzing its feature coefficients. Finally, Linear Regression with elastic net regularization will filter out features that do not contribute to accurate predictions. We found these properties helped in examining the validity of the model and provides quantification of social patterns. We developed this model at a national scale with individual predictions being made at the county level. We utilized Core Spark and Spark ML, running on NYU's DUMBO cluster that output a visualization of results in Tableau.

II. Motivation

Predicting Annual Gross Income (AGI) is useful for a variety of stakeholders: policymakers, investors, and citizens. The Internal Revenue Service (IRS) does produce a report, but it lags by a few years, so a predictive model for current day AGI would be useful. We decided to create a machine learning model which could predict, within a range, the average AGI of a county for stakeholders to use for planning.

Our model outputs average AGI predictions per county and the most/least important features contributing to these predictions. These features can be used to help policymakers decide what factors will contribute to growth in their area. For investors, it's an opportunity to investigate new investments in a community such as expanding their business to a particular county or finding out which counties have benefited from a particular business category. For researchers investigating equality, it's helpful to understand how different factors influence income which is a strong indicator of prosperity.

III. Related Work

In "Predicting Twitter User Socioeconomic Attributes with Network and Language Information" the authors investigated whether an individual's Facebook (FB) data could be used to determine their income [2]. The researchers conducted a survey to collect FB user data on 2,623 people including their likes, profile info, personality, etc. The survey contained general information questions about the participants age, gender, zip code, income, etc. The author's then created two matrices based on the user profiles–likes and status updates. They then created train/validation/test split and trained with a LASSO Model with 10-fold cross-validation to handle outliers and ensure robustness.

The researchers found that including FB profile information and personality data increased the accuracy of a baseline sociological-demographic model (from r = 0.42 to r = 0.49). Finally, they examined the word choice of lower and higher income individuals. The researchers determined that higher income individuals employed positive/future-oriented word choice, whereas lower income individuals were more likely to be negative/self-focused. The paper concludes with concerns around the privacy implications of utilizing prediction models to determine someone's income and hopefulness for how it could be used positively. We found that different demographic profiles would be useful for our model, so we included gender, race, and educational attainment to enhance our model's predictive capabilities. Additionally, we used a Linear Regression model and were influenced by this paper's design approach for our model.

Agencies like the Census Bureau provide information that adds insights into gentrification, but often after a lag. Being able to predict things like gentrification in real-time could be useful towards analyzing such topics. One study used were Yelp business information, American Community Survey (ACS), Census Zip Code Tabulation Area (ZCTA), and Federal Housing Finance Agency (FHFA) [1]. They also incorporated a StreetScore, which is a measure of how humans perceive the safety of a neighborhood based on a Google Streetview image. Their paper then goes on to show the correlation between Starbucks in a neighborhood and Housing Price Index (HPI) change. They tried multiple methods of regression, since they were concerned that one method might just be capturing Starbuck's expansion strategy (and thus that the existence of a Starbucks is because of the affluence of a neighborhood, not the other way around). They also generalized to all cafes, not just Starbucks, in other regressions. They then found that local businesses had a correlation to how educated an area was. They find similar, but weaker, correlations between businesses and racial components of demographics. Finally, they were able to find correlations between Yelp reviews and average age of an area. We found this information to be directly related to our goal of predicting average AGI.

Our final background source is from researchers at the Northern Policy Institute. Researchers there performed an economic impact analysis report to determine how a change in the demand for goods and services affects economic activity in Northern Ontario [3]. The report finds that income multipliers are greatest in retail trade, professional, scientific, and technical services, and arts, entertainment, and recreation. The report finds that employment multiplies are greatest in mining, quarrying, oil and gas extraction, utilities, professional, scientific, and technical services. Multipliers attempt to quantify the amount by which a change in cash investment is magnified or multiplied to find the total final change in expenditures and income in the region in question. We considered these report's findings compelling when considering the inclusion of the U.S. County Business Patterns data-set into our model.

IV. Design and Implementation

A. Design Details

Fig. 1 shows the design diagram. The first step of our application is to clean/profile each dataset, and then join on county and year. The end result is that each row of this combined table which we split into training and testing sets. The model is fit to the training set, and evaluated on the test set. Our application outputs the model, a summary of the performance of the model, the model's coefficients, and the predictions made on the test set. This output is then visualized with Tableau as seen on Fig. 2.

V. Datasets

A. U.S. Census County Business Patterns (CBP)

Counts of businesses at the county level, organized by year, NAICS code and number of employees [4]. The data-set size is in the GB range [5].

B. IRS SOI Tax Stats

Income data aggregated at the county level. The data-set size is in the GB range [6].

C. U.S. Census American Community Survey (ACS) Demographics

The US Census collects demographic data organized at the county level and is found on the US Census website [7]. The dataset size is in the MB range and was collected once for the years 2011-2016. The demographics data

D. U.S. Census ACS Educational Attainment

Level of education attained by specified age groups, organized by county. The data-set size is in the MB range [8].

See Appendix for dataset schemas.

VI. Results

Nationally, the mean value for AGI is $22,750 with a standard deviation of $9,190. We split our data into training and test sets, the training set consisting of all data points gathered between 2011 - 2016 (inclusive) and the test set consisting of all data points gathered in 2017. Our model is fitted with a root mean squared error (RMSE) of 5.47 and an R2 value of 0.63, and when evaluated on the test set achieves an accuracy of 91% of predictions falling within 30% of their expected values. We interpret this as a good start with room for improvement.

We argue that the level of the model's accuracy can be the basis for analysing its coefficients to quantify social patterns. Out of 251 coefficients only 35 are non-zero, and again this is due to linear regression with elastic net regularization "pruning out" features that do not contribute to model accuracy. Also

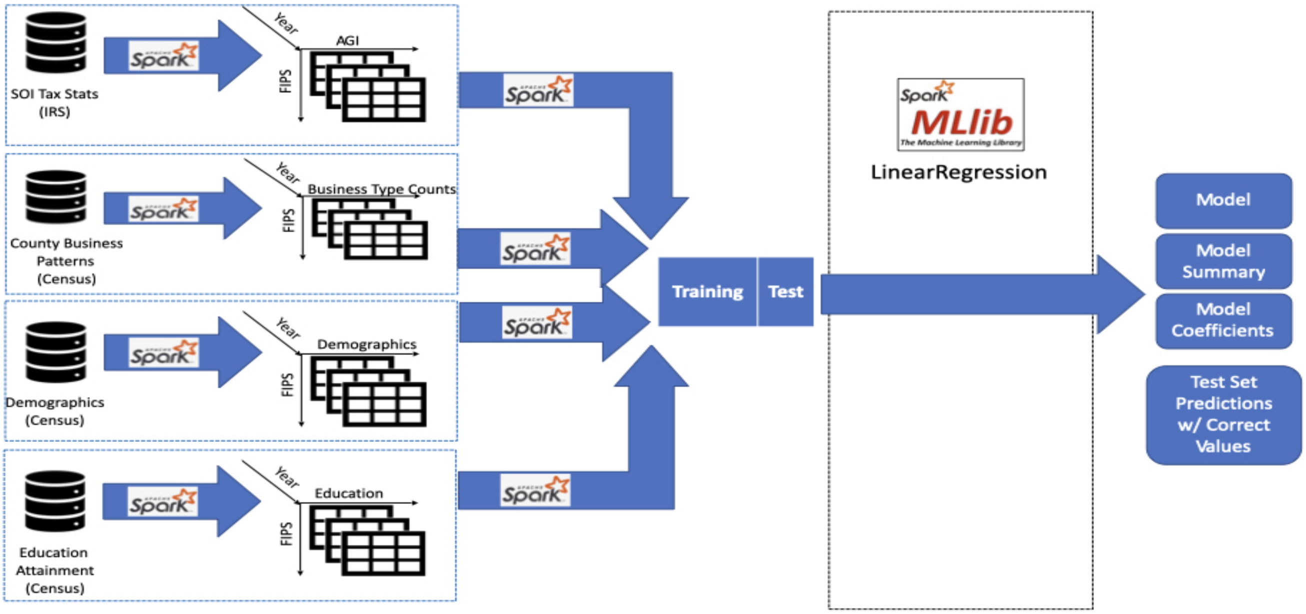


Fig. 1. Design Diagram

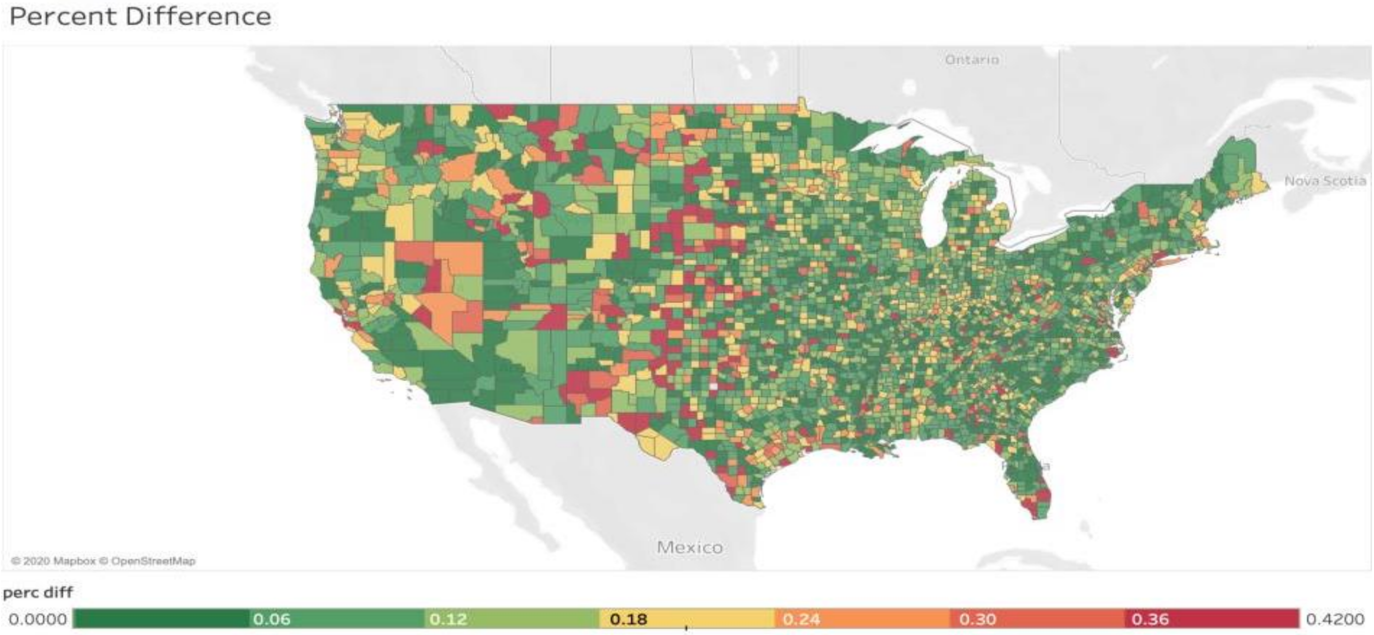
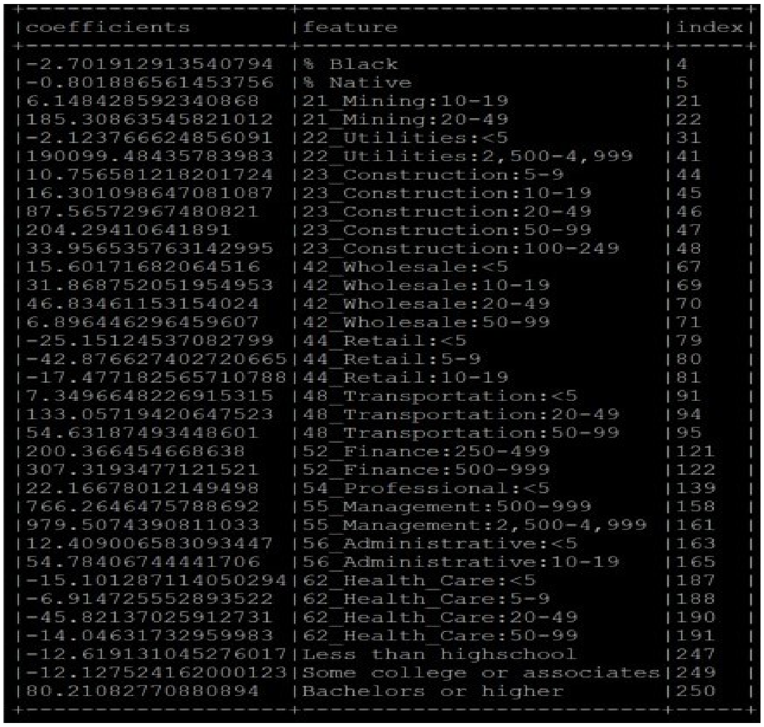


Fig. 2. AGI Predictions

these coefficients are not random, but are produced deterministically based on the algorithm and the datasets we extract features from. This means that an individual coefficient shows a correlation between the feature and an effect on the average AGI of a county. This warrants analysis from domain experts, but we will make a few speculations.

In figure 3 we see three coefficients tied to educational attainment features at the bottom of the table, and they line up with what we would expect. Attaining at least a bachelors degree correlates with higher average AGI, which makes sense since those with that level of education have access to higher paying jobs. Similarly, not graduating from high school (or having an equivalent degree) correlates with lower average AGI. Having achieved some level of college education, to include achieving an associates degree, is also correlates with lower average AGI, and we hypothesize that this population is mostly students that (for the most part) aren't employed in high paying full-time jobs.

Second, the sixth coefficient is tied to the number of utility companies in an area that have between 2,500 and 4,999



```
+-------------------+--------------------------+-----+
|coefficients       |feature                   |index|
+-------------------+--------------------------+-----+
|-2.701912913540794 |% Black                   |4    |
|-0.801886561453756 |% Native                  |5    |
|6.148428592340868  |21_Mining:10-19           |21   |
|185.30863545821012 |21_Mining:20-49           |22   |
|-2.123766624856091 |22_Utilities:<5           |31   |
|190099.48435783983 |22_Utilities:2,500-4,999  |41   |
|10.756581218201724 |23_Construction:5-9       |44   |
|16.301098647081087 |23_Construction:10-19     |45   |
|87.56572967480821  |23_Construction:20-49     |46   |
|204.29410641891    |23_Construction:50-99     |47   |
|33.956535763142995 |23_Construction:100-249   |48   |
|15.60171682064516  |42_Wholesale:<5           |67   |
|31.868752051954953 |42_Wholesale:10-19        |69   |
|46.83461153154024  |42_Wholesale:20-49        |70   |
|6.896446296459607  |42_Wholesale:50-99        |71   |
|-25.15124537082799 |44_Retail:<5              |79   |
|-42.876627402720665|44_Retail:5-9             |80   |
|-17.477182565710788|44_Retail:10-19           |81   |
|7.3496648226915315 |48_Transportation:<5      |91   |
|133.05719420647523 |48_Transportation:20-49   |94   |
|54.63187493448601  |48_Transportation:50-99   |95   |
|200.366454668638   |52_Finance:250-499        |121  |
|307.3193477121521  |52_Finance:500-999        |122  |
|22.16678012149498  |54_Professional:<5        |139  |
|766.2646475788692  |55_Management:500-999     |158  |
|979.5074390811033  |55_Management:2,500-4,999 |161  |
|12.409006583093447 |56_Administrative:<5      |163  |
|54.78406744441706  |56_Administrative:10-19   |165  |
|-15.101287114050294|62_Health_Care:<5         |187  |
|-6.914725552893522 |62_Health_Care:5-9        |188  |
|-45.82137025912731 |62_Health_Care:20-49      |190  |
|-14.04631732959983 |62_Health_Care:50-99      |191  |
|-12.619131045276017|Less than highschool      |247  |
|-12.127524162000123|Some college or associates|249  |
|80.21082770880894  |Bachelors or higher       |250  |
+-------------------+--------------------------+-----+
```

Fig. 3. Model Coefficients

employees (inclusive). The coefficient is very large, so at a glance it might appear that this feature correlates with very high average AGI. However, before inputting the feature into the model (either during training or evaluation on the test set) we regularize each business count by dividing by the total number of businesses in that county for the given year. It would make sense that a large utility company would be serving a high population area, and that there would be many businesses in that county. So the value that gets passed into the model for this feature is at most a very small decimal. This would would explain the coefficient value as compensation for this small decimal.

Finally, at the top we notice two demographic features that correlate with lower average AGI. What we think what is happening here is that there exists some feature(s) that would explain lower average AGI that correlate highly with these populations, and in future works we would need to find these features and incorporate them to increase the robustness and accuracy of the model. However, these populations are historically disenfranchised, so we would monitor these coefficients in the expansion of the model to see if they persist.

VII. Future Work

One addition we would like to experiment with would be automatically training a second model that only includes features that contribute to the accuracy of the original model above a specified threshold. The theory behind this is that in order to make predictions on current average AGI levels, one would need to perform information gathering (or approximation) to build the feature set the model needs to make a prediction. If this feature set is smaller, one would need to spend less time (or resources) to perform this information gathering, since most of the features may not be important. So a smaller model with a comparable level of accuracy would be more beneficial to the user of this application.

Another expansion to this project would be to build a more robust, user-friendly front end that allows the selection and analyzing of input features for model development. This would allow domain experts to perform these tasks in their own experiments without requiring our assistance.

Additionally we believe that collecting features at a more fine-grained level (i.e. by city or zip code) would allow for a more detailed analysis of the results. This could be useful towards developing policies that target the root of problems identified through our application.

Finally, it was out of scope of our project to incorporate standard ML optimization techniques in model development. Doing so automatically would increase the accuracy of the model and thus give more validation to the coefficient values.

VIII. Conclusion

A linear regression model was developed using Spark ML to predict the average AGI of a county using counts of businesses by NAICS type, demographics, and educational attainment as features. The IRS lags with releasing their information, so this model could be useful to those who need an approximation now. We achieve 91 percent accuracy of predictions being within 30 percent of their target. Interactive AGI maps, and feature importance maps were developed into an application using Tableau to illustrate the model results. The coefficients that linear regression quantify the importance of each feature towards model accuracy, and could be used to quantify social patterns.

Acknowledgment

Thank you to Prof. McIntosh for answering our analysis questions. Thank you United States Census, IRS, for providing our datasets. Thank you specifically to the Census clerk that answered the phone immediately and answered our questions without putting us on hold! Thank you to HPC for providing the technology to make this project possible. Thank you Naren Chittar of J.P. Morgan for checking the validity of our model Thank you Danielle Branton of Columbia University for insights that went into model formation and analysis

References

[1] Edward L. Glaeser, Hyunjin Kim, and Michael Luca. Nowcasting gentrification: Using yelp data to quantify neighborhood change. *AEA Papers and Proceedings*, 108:77–82, May 2018.

[2] Nikolaos Aletras and Benjamin Paul Chamberlain. Predicting twitter user socioeconomic attributes with network and language information. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, page 20–24, New York, NY, USA, 2018. Association for Computing Machinery.

[3] Bahktiar Moazzami. *Income and Employment Multipliers for 20 Industries in 11 Census Divisions in Northern Ontario*. Northern Policy Institute, 2019.

[4] Naics sic identification tools.

[5] US Census Bureau. County business patterns (cbp) apis, Jul 2017.

[6] US Internal Revenue Service. Irs soi tax statistics, Jul 2017.

[7] US Census Bureau. Us census demographics, Jul 2017.

[8] US Census Bureau. Us census educational attainment, Jul 2017.

Appendix

U.S. Demographics Data Schema

Data schema is broken down into: field, data type, and description.

1) county_id - long - the county ID which is a numerical representation of the state and county.
2) state - string - the name of state.
3) county - string - the name of the county.
4) total_population - long - the number of people living in the county.
5) men - long - the number of men living in the county.
6) women - long - the number of women living in the county.
7) hispanic - long - the number of hispanic people living in the county.
8) white - long - the number of white people living in the county.
9) black - long - the number of black people living in the county.
10) native - long - the number of Native American people living in the county.

U.S. Business Patterns

1) fipstate - integer - the state ID which is a numerical representation of the state.
2) fipscty - integer - the county ID which is a numerical representation of the county.
3) naics - string - indicates the NAICS code subcategory, or total if "——".
4) est - integer - total number of establishments.
5) n1_4 - integer - number of establishments: 1-4 employee size class.
6) n5_9 - integer - number of establishments: 5-9 employee size class.
7) n10_19 - integer - number of establishments: 10-19 employee size class.
8) n20_49 - integer - number of establishments: 20-49 employee size class.
9) n50_99 - integer - number of establishments: 50-99 employee size class.
10) n100_249 - integer - number of establishments: 100-249 employee size class.
11) n250_499 - integer - number of establishments: 250-499 employee size class.
12) n500_999 - integer - number of establishments: 500-999 employee size class.
13) n1000_1 - integer - number of establishments: 1,000-1,499 employee size class.
14) n1000_2 - integer - number of establishments: 1,500-2,499 employee size class.
15) n1000_3 - integer - number of establishments: 2,500-4,499 employee size class.
16) n1000_4 - integer - number of establishments: 5,000+ size class.
17) year - integer - year data was recorded

U.S. Educational Attainment

1) fips_combined - integer - ID that identifies a county.
2) less_than_high school - integer - population count of those that have not achieved a high school or equivalent education
3) highschool - integer - population count of those that have achieved a high school or equivalent education but nothing more
4) some_college - integer - population count of those that are working on college or have achieved an associates
5) bachelors - integer - population count of those who have achieved a bachelors or more
6) year - integer - year data was recorded

SOI Tax Statistics Data Schema

1) fipstate - integer - the state ID which is a numerical representation of the state.
2) STATE - string - U.S. State
3) fipscty - integer - the county ID which is a numerical representation of the county.
4) COUNTYNAME - string - US County name
5) agi_stub - integer - a tax bracket
6) agi - float - Annual Gross Income
7) fipstate_fipscty - integer - combined score
8) year - integer - Year of collection