Backstory:

The proposal outlined in this project centers around an imaginary individual who is a county economic and development planner in an unnamed county in an unnamed state. This individual (myself) is approached by his superior who requests that the I come up with an answer of what are some important factors for improving tax revenue in the area. Upon review of county tax records it was discovered by this individual that the largest tax revenue is income tax revenue in this area. The county receives a % of income tax revenue that is collected from individuals in this town. Furthermore, it was discovered by this individual that the tax code is written such that areas that contribute more to the tax bucket get a higher percentage of the tax returns from state government. When faced with this revelation, this individual reasoned that a model that depicted factors that contributed to a higher than average median household income relative to the state that the county resided in. Determined to meet my audience wherever their attention level/interest level is at to gain buy in I created this technical report, the high level narrated presentation and a 2-3 minute summary video.

Business Understanding:

The county in question derives the majority of its funding through tax rebates returned to the county by the government. The most direct way that the county can increase it’s share of the tax rebate pool is to increase the income taxes collected from within the county. The county would like to commission a preliminary study with its resident “data-guy” to gain a better understanding of what factors are involved with determining if a county has higher that average household median income for its state. Thus the county can then determine if there are policies that it may push that would promote the higher median incomes in the county and perhaps increase income tax revenue for the county.

I will seek to develop first a data model that can be then reduced down to contributing factors by weight, analyzed and then presented before the county commissioner and board in an engaging way to show if there are policies that maybe undertaken to promote higher incomes within the county. It is my understanding that the board is less technical in capabilities thus, I will seek to ensure that the model is intuitive to understand. Furthermore, the emphasis in this project will be on ensuring that accuracy and overall classification error metrics are maximized; the board would like to ensure that they know which policy strategies help and hurt personal economic growth. The county board knows that there will be some factors that will remain outside of their control, however, if there are some macro strategical changes that may incentivize growth, they would like to know.

Data Understanding:

The county is seeking to increase funding, thus there was no budget available for data acquisition and surveys would take too long to conduct, thus data for this project consisted of publicly sourced data; findings associated with this phase of the study will be used by myself to lobby for additional data sources for follow up studies.

The data for this study was aggregated from 3 data.gov data sets. The first dataset consisted of Education Data, the second data set consisted of Unemployment data and the third consisted of Population Estimates. Each of the data sets go back to 2005 with reliable data and have aggregated data points going through the 1970’s. Also, the data sets have a common unique identifier system for county, a “FIPS Code” and each have roughly the same counties. This data is semi reliable as it is from a public government source, however some of the data is in unusable formats and there are several data points that are missing. The highlight of the entire data set for my purposes are a unique feature to the Unemployment data set which has “County Household Median Income as a percent of the State Total Median Household Income, 2017. This is aggregated data based on county that should show the percent the relative wealth of the county based on the rest of the state that the county belongs too. Studying this data point it is apparent that there are more counties below the 100 in this data set than above (meaning that the wealth distribution between counties is skewed; thus wealth distribution is not random and there should be some distinguishing features of these counties that make them more attractive to wealthy individuals. It should also be noted that even in some major metropolitan areas, neighboring counties can have vastly differentiated Median Household Income.

The raw data file overall has over 3400 rows of data (each representative of a county or state aggregated record) and has over 120 features. The US Government sites the bureau of labor statistics, the USDA economic Research Service, the Census Buereau, and the US Department of Education for the data gathered in this data set.

Data Preparation.

The first thing noted about the data set was that the data was in 3 separate downloaded files. Since the data sets had a common UID the data was merged using the following python merge code:

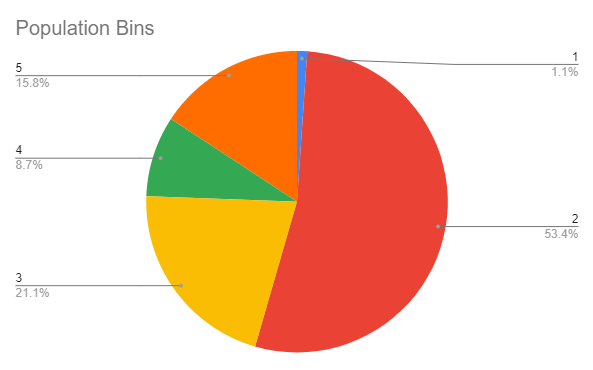


The resulting “mergedf.csv” file was then opened in excel for additional processing.

The merged file contained 3400 rows of data and over 120 features. One problem with this data set is that it contained aggregated points. Since the project called for only comparing counties within a state to determine factors in wealthier counties, it was determined that these aggregated data points should be removed. To accomplish this, a column was added and the state names were flagged and promptly deleted form the data set. Additionally, the data set contained 51 rows of data that appeared to be indicative of incomplete county surveying. These counties did not contain the KPI that the project called for thus these were also flagged and removed. This left the data set with 3141 complete rows of data and still over 120 features. However, many of these features utilized data points that were from decades ago. Upon consulting with the board, it was determined that the board was interested in only the most recent data points, the most recent median household income point was from 2017; thus I eliminated any feature that did not contain 2017 data. Further, many of the features were simply rehashing of other features which appeared to be indicative of multiple agencies holding similar data and then merging this data. So duplicate or nearly duplicative features were removed. Additional investigation of the remaining 60 features showed that a number of the features did not appear to have much to do with economic output and were instead unknown internal KPIs, while understanding these KPI’s may have helped the dataset, it was unclear from the readme files associated with the datasets that these KPI’s came from what the underlying datapoints from the KPIs were. In the interest of eliminating any symbolance of bleedthrough into the training/testing data and preventing unnecessary correlation/bias of features these KPI’s were removed leaving 19 features.

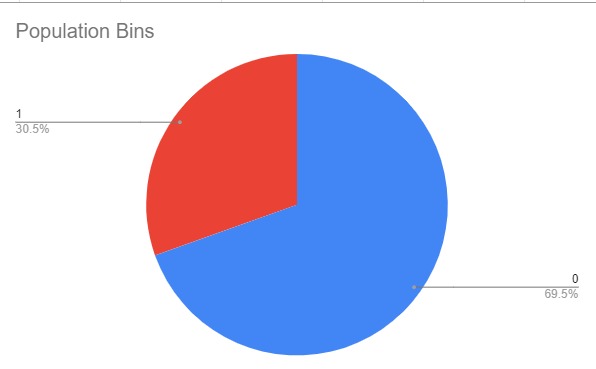
Review of the remaining 19 features revealed some possible issues within the data set. First many of the metrics were based off of total population of the county for instance the unemployment numbers were based off of total unemployment and births within the county where in a real number form as well. This appeared to overemphasize the effect of population on the county in my opinion, and as the median average income is an average and not an overall sum, it did not make since to me at the time to give such a large bearing on the number of individuals within a county. Thus these numbers were all re-engineered to be percentage of population instead of raw numbers. Further, there were vast disparities in population of county estimates. The range was from 123 to over 1,000,000 per county. I was immediately worried about the effect that such a large disparity in county population may have on a model that oveall has numbers between -1 and 50. Thus these populations were recategorized and binned into micro [1] counties (0-1000), Rural [2] counties (1001-30000), suburban [3] counties (30001-70000), urban [4] (70001-250000) and Metropolis [5] counties (250001+). The breakdown was as follows:

|  |  |
| --- | --- |
| Pop\_Estimate\_Bin | count Pop\_Estimate\_Bin |
| 1 | 34 |
| 2 | 1677 |
| 3 | 663 |
| 4 | 272 |
| 5 | 495 |



Another interesting feature was added by investigating the percent differences in individuals who did not graduate highschool and individuals that have atleast an undergraduate degree completed. Thus another metric was added to indicate if more individuals in that particular county had a bachelors degree or higher (1) vs less than a highschool diploma (1). This is broken out as follows:

|  |  |
| --- | --- |
| nohighschool>college? | count nohighschool>college? |
| 0 | 2184 |
| 1 | 957 |



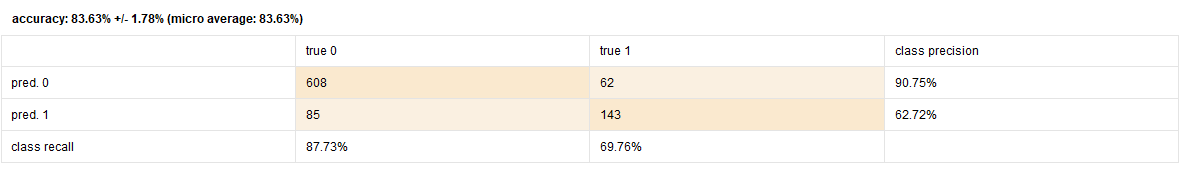
The features considered in the modeling phase of this project were as follows:

|  |  |
| --- | --- |
| FIPSCODE, Area Name | UID |
| State\_x | State |
| Percent of adults with less than a high school diploma, 2013-17 | Percent of adults with less than a high school diploma, 2013-17 |
| Percent of adults with a high school diploma only, 2013-17 | Percent of adults with a high school diploma only, 2013-17 |
| Percent of adults completing some college or associate's degree, 2013-17 | Percent of adults completing some college or associate's degree, 2013-17 |
| Percent of adults with a bachelor's degree or higher, 2013-17 | Percent of adults with a bachelor's degree or higher, 2013-17 |
| nohighschool>college | Are more individuals highschool dropouts than college graduates in this county? |
| Pop\_Estimate\_Bin | Population bin (2017) |
| N\_POP\_CHG\_2017 | Population change in 2017 |
| Births\_2017 | Births in 2017 |
| Deaths\_2017 | Deaths in 2017 |
| NATURAL\_INC\_2017 | Natural pop increase in 2017 |
| INTERNATIONAL\_MIG\_2017 | Migration from international 2017 |
| DOMESTIC\_MIG\_2017 | Migration from domestic 2017 |
| NET\_MIG\_2017 | Migration net 2017 |
| Civilian\_labor\_force\_2017 | % of population that is civilian labor force in 2017 |
| Employed\_2017 | employed % in 2017 |
| Unemployed\_2017 | unemployed % in 2017 (filing for unemployed benefits) |
| Higher than us average income | Higher than 50% of Average US Income indicator (Key KPI) |

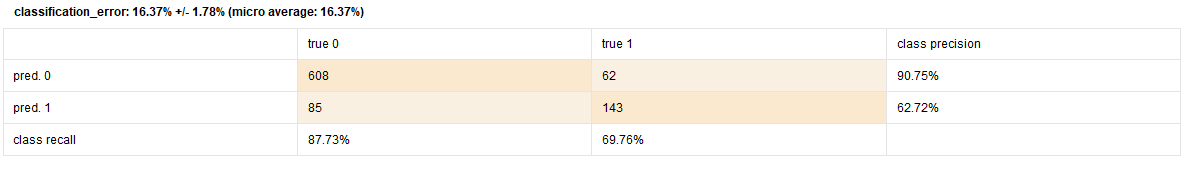
Modelling phase and evaluation:

After the data was selected, prepped, cleaned and features selected, engineered and reengineered, the project moved into the modelling phase. For this I was given access to the counties Rapid Miner trial license. For this phase I considered 3 types of models that I believed would yield strong results for the data set. [evaluation of each model was conducted by cross validation within the data set. The first was Deep Learning:

Deep learning gave 84% accuracy,



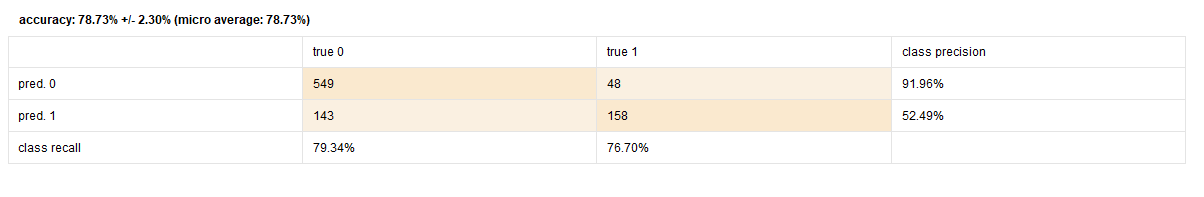
And 16.37% classification error



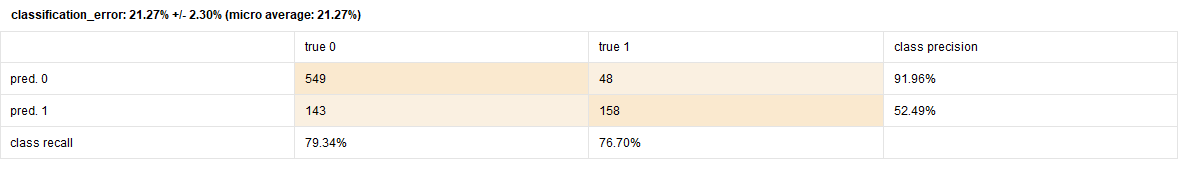
While this was a strong showing out of an initial run at a model, I was concerned with explaining deep learning to the board as a means of acquiring additional funds for my initiative.

The second was Naïve Bayes:

Naive bayes gave 79% accuracy,



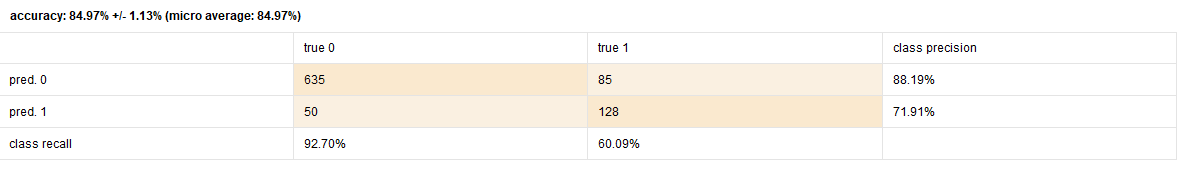
And 21.27% classification error



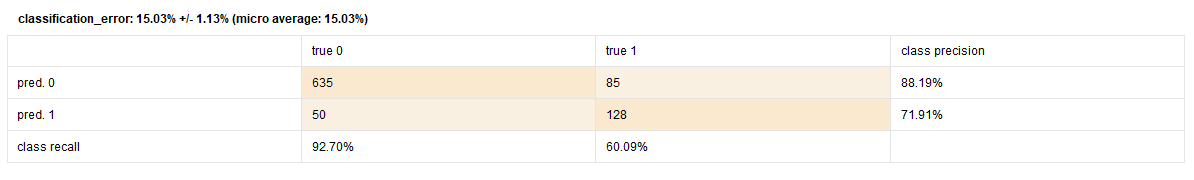
While this method gave a strong accuracy and is simple enough to explain in non technical terms, this model had a little too high of classification error than I was prepared to work with.

I was able to quickly determine that logistic regression had the best accuracy and the least classification error. As seen here:

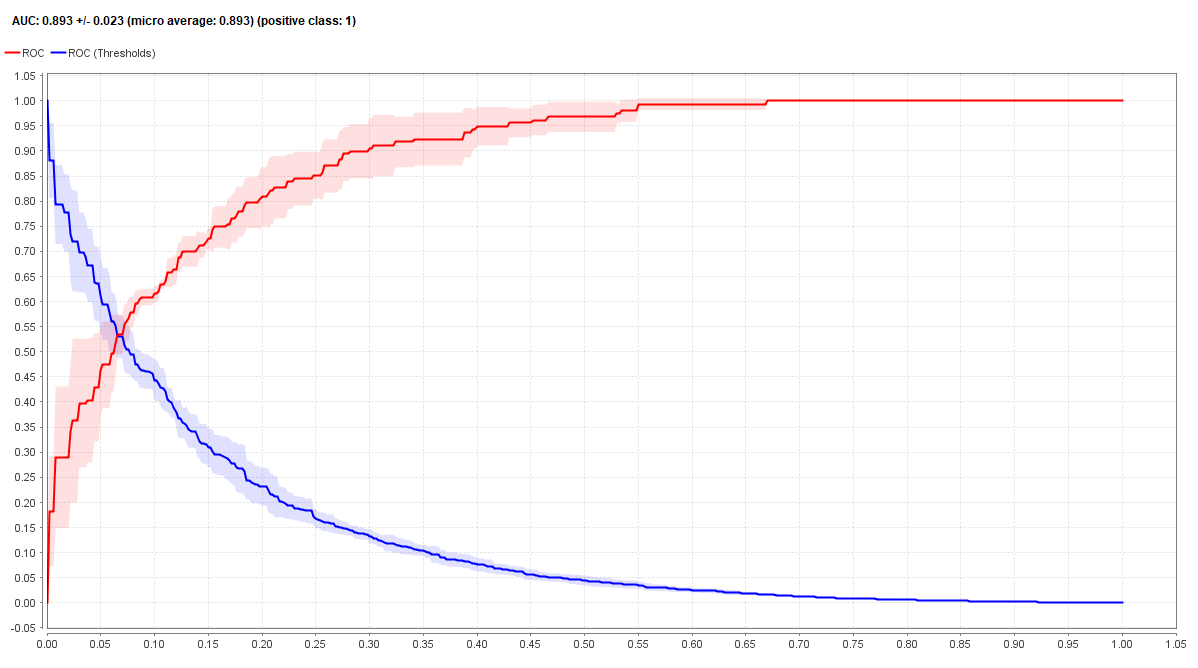
Logistic regression gave the highest accuracy at 85%



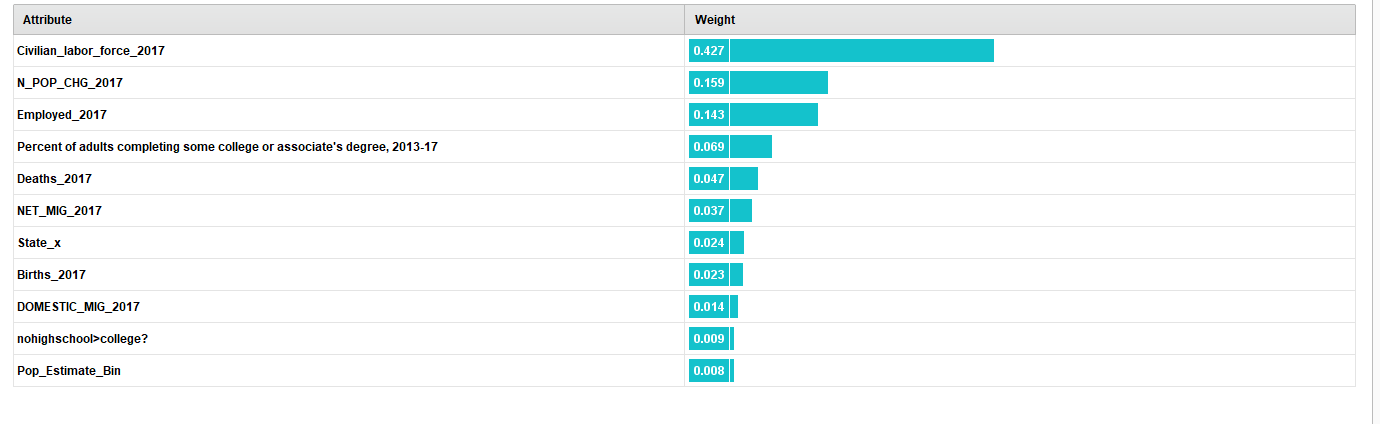
Gave the lowest classification error % at 15.03%



And had an AUC of .893



This model was used. The weights were as followed:



Here we see that the highest weights in the logistic model were given to the percentage of the population that was civilian labor force in 2017, the population change in 2017, employed % in 2017 and the percent of adults that have completed some college or associate’s degree between 2013-2017. Thus there appears to be a surprising trend here. It appears that these individuals are working class adults without over education (as this was not the highest education level included in the model) but willing to maintain as a member of the workforce. It appears that the wealthiest counties in America respective to the rest of the state in which they reside are attracting individuals to move there. This stands to reason as certain high paying jobs may follow were individuals are.

Recommendations based on Model:

It appears that this model suggests that perhaps a strong community college incentive system should be advertised both within the county and in surrounding counties for in-county residents. This may incentivize working individuals to migrate to the area to cash-in on reduced price education in competitive fields. Further, if there is a plethora of talent in highly competitive fields this may attract high-paying companies into the area and thus these individuals being residents within the county would be paying income tax within the county.

Further study:

Review of the data shows that there maybe additional studies that maybe undertaken to further enhance the counties appeal to this demographic. For instance, it maybe helpful to conduct a study on what degree these individuals are studying. If this is for instance a technical or computer based field, it may be in the counties best interest to attempt to attract partnerships with companies in that field to develop a community college training program within the county. Such a partnership could prove lucrative for both the company being partnered with and the county. Further, the county should look into the community college system within and consider the funding that it is receiving, the classes that are coming in and determine if infrastructure around the college is conducive to commuting students (i.e. is it difficult to both attend school and work a part-time or full-time job as parking, roads or other congestion prevents efficient transportation around the school. Lastly, the county should look at the population demographics within the county and promote programs that benefit working class individuals. For instance is there excessive subsidies given to retirement communities within the county; while these maybe good incentives to have, it maybe beneficial to attempt to partner with a neighboring county and [with risk of sounding callous] off load some of this community and thus increase the percentage of civilian labor force and employed labor force. (As retired individuals, including retired-early individuals do not have income coming in and thus do not contribute to the income tax incentives received by the county). A study showing the possible benefits and impact of this should be conducted before proceeding with these initiatives.