Examining Factors with the Most Impact on the Per Capita Income of Tulsa City Tracts

TDS

Steve Green, Ph.D & Ben Harris
June 17th 2017

Tulsa Data Science, Inc

Highlights

- We found that education level had the greatest impact on predicting income level of Tulsa city tracts.
- With high school drop outs having the strongest impact on the per capita of city tracts, we recommend the city pursue policies which increase the number of students graduating from high school.

Executive Summary

The Mayor has set a primary goal of increasing the per capita income of Tulsa. To select policies necessary to meet this goal, the city needs to understand the factors correlated to per capita income. Using Census Bureau Data, we performed feature selection to identify factors with the most impact on city tract income levels. Results showed the most impactful factors were related to education level with high school drop-outs having the most impact on a tract's income. For this reason, Tulsa Data Science, Inc, formally recommends that the City work with school districts to increase the high-school graduation rates. By reducing high school drop-outs, our model suggests that the city of Tulsa should see an increase income level.

Introduction

For the first year in office, Mayor Bynum has set increasing per capita income for the city of Tulsa as a key goal. Succeeding in this goal will translate into a real tangible effect of putting more money into the pockets of the city's citizens. In addition, it will make Tulsa a more attractive destination for top talent allowing the city to develop a variety of industries and create jobs. Finally, by increasing income of poorer tracts, the city can reduce income disparity across all of Tulsa.

While improving income level can provide benefits, identifying the right policies to achieve this goal can be difficult. This is because a variety of factors determine income level and it's unclear which ones have the most impact. For example, education level, availability of natural resources, and demographic composition can all have an impact on the overall income of the city. With limited resources, the Mayor's office needs to know which factors have the most impact in order to select those policies that will most efficiently increase income levels.

The current analyses sought to identify the factors with the most impact on determining income level. Using census bureau data, we examined the effect of over an hundred different factors on the income level of all Tulsa city tracts. We employed several machine learning algorithms, to select the features with the most impact in relationship to income level. This allowed us to build a model that more accurately predicted the income level of each of the city's tracts. Using this approach, we hypothesized that both education level and social factors would have the most impact on income levels.

Results

Average ranking of all selected features revealed four features that were consistently at the top. We found that the per capita income of a tract was most strongly associated with the number of individuals holding a professional degree (medical or law). This was followed by the number of doctoral degree holders. The third and fourth highest ranked features were the number of white people holding a bachelor's degree and the number of people educated at the 11th grade level.

Since the Mayor's office is interested in understanding how these features can be leveraged to increase the overall wealth of Tulsa, we examined each of these four factors more closely. We wanted to better understand the relationship between individual features and per capita income.

Amount of Professional Degree Holders

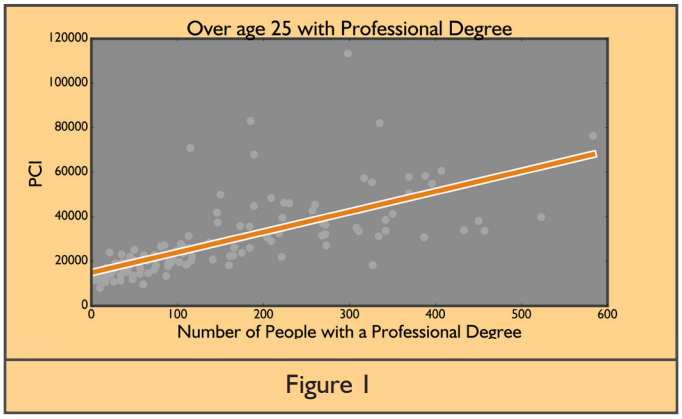

Figure 1

Examining the figure above (Figure 1) we can see that as a city tract contains more professional degree holders the higher the per capita income. Statistical analysis using univariate linear regression confirmed this result ($F(106) = 118$, $p < .001$). Those holding professional degrees typically are either doctors or lawyers so we expect that areas of city that contain many of these kinds of professionals would have higher per capita income.

Amount of Doctoral Degree Holders

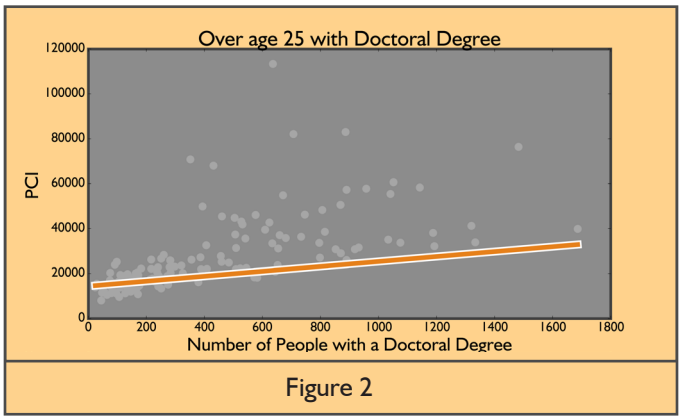

Figure 2

Similar to the professional degree holder, the figure above (Figure 2) reveals that per capita income increases as the number of people in a city tract hold a doctoral degree. We confirmed this result with a univariate linear regression model ($F(106) = 89$, $p < .001$). An expected result, we speculate that areas of the town where advance degree holders are concentrated, like around universities, have a higher per capita income than other areas of Tulsa.

White Bachelor Degree Holders

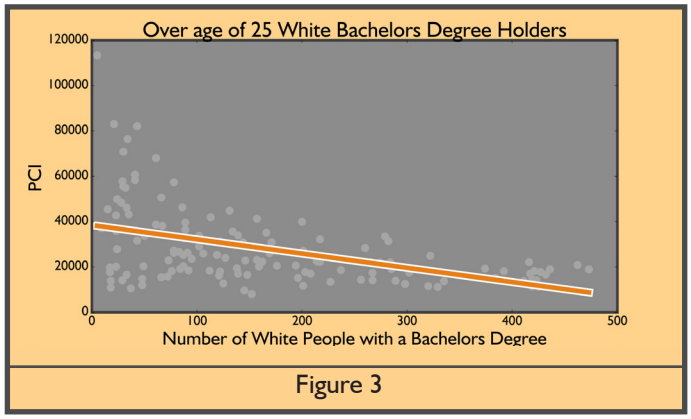

Figure 3

Visual inspection of figure 3 indicates that as the amount of white bachelor degree holders increases in an areas of the city, the per capita income decreases. Although weaker than the previous two features, univariate regression confirms the finding ($F(106) = 36$, $p < .001$). One possible explanation may be due to white bachelor degree holders pushing out more professional and advance degree holders in a region, thereby driving down the per capita income of the tract.

Those With at Most an 11th Grade Education

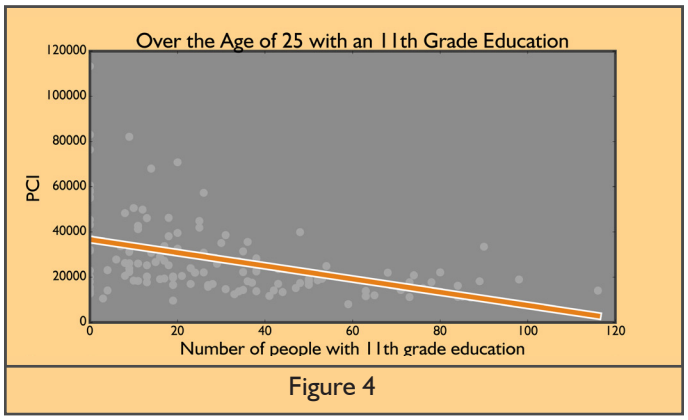

Figure 4

As the above figure confirms (Figure 4), tracts where there are high numbers of people that failed to complete high school, have substantial decreases in per capita income. This finding was confirmed with univariate regression ($F(106) = 32$, $p < .001$). Even though the result is expected, it underscores the importance for the city to work with the Tulsa high schools to find ways to reduce drop out rates as a way to increase the per capita of the city.

Comparing the Effects of each Feature on Per Capita Income

Rank	Feature	Slope Estimate
1	Professional Degree Holders	$91
2	Doctorate Degree Holders	$31
3	White Bachelor Degree Holders	-$62
4	Completed 11th Grade	-$291

Table 1

The above table (Table 1) shows the four top features-ranked according to their impact on per capita income. Results show that areas with many professional and doctoral degree holders have a positive impact on per capita income. In contrast, the next two features – white bachelor degree holders and those that did not finish high school – reduced per capita income.

Based on the slope estimate – which models the per capita change as people are added to a tract – we can see some substantial differences between the features. For example, as one more professional degree holder is added to a tract, we would predict the per capita income to increase by $91 dollars. In contrast, with a coefficient value of negative $291, adding those who have only completed the 11th grade will reduce the tract's per capita by by almost $300. If the city aims to increase the per capita of the city, they will likely have the most impact by reducing drop-outs, as this will increase the per capita of that tract by the most amount of money per person.

Model Validation

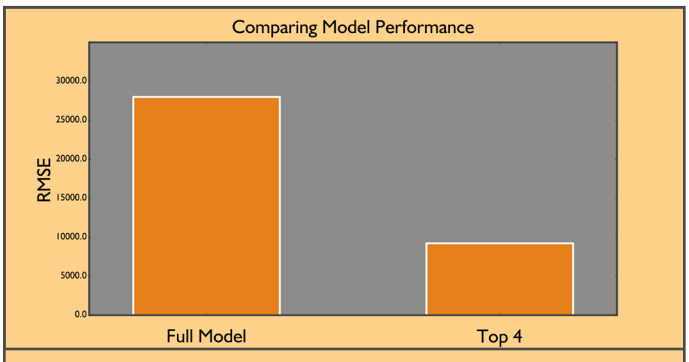

Figure 5

To determine whether our four features were effective at predicting per capita income we used a five-fold cross-validation to test against a model with all features. As can be seen in the figure above (FIgure 5), results from this analysis showed that the top-four model had significantly less error in predicting per capita outcome than the full model. We conclude that these four features have the most impact on the per capita income.

Discussion

With the Mayor's Office focused on increasing the per capita of the city of Tulsa we performed feature selection to reduce a large set of census bureau features into a subset with the most impact. Our analysis revealed four stable features in predicting per capita income in Tulsa City tracts. We found that advanced and professional degree holders increased per capita while white bachelor degree holders and those that did not complete high school decrease it. Examining each feature's coefficient, we found that those not completing high school had the greatest impact on per capita value, while professional degree holders had the second highest impact. These results indicate that overall improvement of per capita income is best directed at improving the education level of the citizens of the city.

Recommendations for Raising Per Capita Income

Results from our analysis suggest that one of the best ways to improve the city per capita income is by reducing the amount of high school drop-outs. According to the model individuals with only an eleventh grade education had the most impact on per capita income through substantially reducing it. Increasing the high school graduation rate across the region would not only have the largest impact on improving the per capita of Tulsa, but it will also provide these residents with greater financial outlook over their entire lifetime.

Another approach the city could take is to either produce and retain, or attract more professional degree holders. We recommend focusing on professional degree holders as the model showed they added more income to a region that those with an advanced degree. Since most professional degree holders are either lawyers or doctors, the city could implement programs to attract more of these professionals to the area. Alternatively, collaborating with local universities to de-

velop programs to encourage local residents to attain professional degrees might also increase the per capita income of the city.

Recent work at Tulsa Public Schools, in the office of Data Strategy and Analytics[1], has shown that chronic absenteeism is the strongest predictor of a student dropping out of school. Using this insight, the Mayor's Office could collaborate with school districts across the city to develop programs encouraging students to attend school. For example, the city might consider providing support for programs that fund counselors with the explicit task of increasing the attendance percentage of chronically absent students. Increasing attendance in these students will decrease the drop-out rate, and based on our modeling, would have the highest impact on improving the per capita income of the city. For this reason, we formally recommend the Mayor's office pursue policies to reduce high school drop-outs as this approach will most likely increase the city's per capita income.

Recommended Visualization

With our current model we could compute positive deviance for each city tract. Using these results, the city could examine over performing tracts to possibly identify features that may be increasing their per capita income. These insights could be used to develop policies that will increase per capita income for tracts that are under performing. To help facilitate this, Tulsa Data Science Inc. could develop an online visualization that would show how each tract is performing against its predicted outcome. Being able to see how tracts are performing will allow city officials to see trends across tracts which might provide greater insight into policies that are affecting income level across the city.

Caveats

Although we recommend the Mayor's office pursue policy which effect the education level of citizen of Tulsa, we must mention such efforts may have unintended consequences[2]. In fact, results from our analysis indicate that increasing the number of white bachelor degree holders actually reduces the per capita income for those tracts. Our analysis implies a perverse result, where policies to increase college enrollment have the opposite effect of reducing per capita income. This may in part be due to the value of a degree being derived from the proportion of people in the city at that education level. As the proportion of bachelor degree holders increases in the city, the value of the degree decreases causing income levels to drop. We caution the Mayor's Office in policies affecting the amount of professional degree holders as it may have the unintended consequence of reducing the per capita income of the city.

Even though the analysis strongly suggests that education level has the strongest impact on the per capita income in different areas of the city, it's important to understand these results do not establish a cause and effect relationship between these two factors. Currently, the analysis indicates there is a relationship, but it may be spurious[3] or driven by an unknown third factor[4]. This is an important caveat because if the City chooses to pursue a policy on this relationship, as we have recommended, there is a possibility it will fail. In order to be confident of a causal relationship between these two factors the city would need to conduct an experiment manipulating the proportion of education level in a tract and observe the effect on that area's per capita income. Positive results from such a study would give the City confidence that their policies will have the intended effect of increasing per capita for Tulsa. Tulsa Data Science, Inc would be interested in providing support in designing and executing experiments testing these factors.

Methods

Data

The data was provided by the Mayor's Office and consisted of 130 observations, representing the individual census bureau tracts for the greater Tulsa area. The data also contained approximately 164 features measuring a variety of factors including education level, household types, and demographic information.

Preprocessing

Prior to the analysis, features that had the same values across all city tracts were removed. To reduce high levels of multicollinearity, columns with a measurement of error for one of the other factors in the data set were removed. After processing, there were 110 features remaining (36% of features were removed).

Parameter Tuning

For tuning the optimization parameters of each algorithm, we used five-fold cross validation and evaluated model performance using root mean squared error on the test sample. Prior to cross validation all features were centered and scaled by their standard deviation.

Lasso Regression – To identify a stable and optimal value of lambda we randomly sampled 500 times using 5-fold cross validation over a range value (.002 to .25). We selected a lambda value of .06 as this value produced the least amount of error in the test sample.

Ridge Regression – To select our values of lambda we randomly sampled 5000 times using 5-fold cross validation over a range of values (4 to 600). We selected a value of 68 as it produced the least amount of error in the test sample.

Random Forest- Three separate parameters were optimized by sampling 5000 times with 5-fold cross validation. The optimal number of trees was 125. For the maximum number features per tree a value of 60% was optimal. Finally, the optimal number of leaves was one.

Ranking the Impact of Features

To assess the reliability and impact of each feature on the income level, each feature was mean centered and scaled by their standard deviation. Next, Random Forest, Ridge, and Lasso Regression models were built to estimate the relative strengths of each factor in predicting income level. For each algorithm, we ranked order according to the strength of each feature. Finally, for each factor, their ranks were averaged across each algorithm to compute an average rank score.

Selecting Features Based on Statistical Reliability

To determine if the average ranking of the top features had a statistically reliable relationship with income level, an F-score and P-value were calculated after fitting each of the Top-5 features to a linear regression. Starting with the top rank feature, we included each feature with a p-value less than .05 and stopped when we came across a feature that did not meet this criteria.

Validating the Top-4 features

Finally, we wanted to test whether a model with only the top-ranked features would perform better at predicting income than a model which included all feature. Using five-fold cross-validation, we sampled 25 times and computed the average root mean squared error of each model. We evaluated model performance based on which model had the least amount of error when predicting income levels in the test data set.

Citations

1. Identifying Factors that Impact Drop Outs, Tusla Public Schools, Data Strategy and Analytics Office

2. https://en.wikipedia.org/wiki/Unintended_consequences

3. http://www.tylervigen.com/spurious-correlations

4. https://en.wikipedia.org/wiki/Confounding