The Effect of Pro Sports Teams on a City's Economy

By Jae Hyun Lee, Jordan Bales, and Ben Rogers

Abstract:

There are 90 total NFL, NBA, and MLB teams in the US. These teams are valued in the billions of dollars with cities paying hundreds of millions in tax breaks to keep many of those teams in their city. We wanted to know, is this a smart idea? Our goal is to look at the effect these teams have upon their respective city's GDP, or in other words whether it is an economic benefit to have a professional sports team for a city. We separated this into three different questions. Does having a sports team increase a GDP over cities with no teams? Does market size matter for the teams in terms of economic impact? Does ownership change for a sports team impact the GDP? From our research, we determined that professional sports have a positive impact on GDP of only the largest of cities and adding professional sports to cities without a team or changing ownership does not positively impact the cities GDP. Future possible research to expound upon our findings would be to look at the impact of college teams on their local economies and the impact sports teams have on the happiness of the citizens in their local cities.

Data and Methods:

For our data, we focused on three primary groups of data to gather. They are city economic data, sports financial, and sports non-financial. Our goals for obtaining data was to maintain city definition through our economic data and to gather from as few sources as possible to make our initial data cleaning easier.

We chose 24 cities from which to obtain data. They are separated into 4 different groups: large market, medium market, small market, and no teams. Since we are focusing on if market

size for a sports team affects its economic impact on the city, we chose cities that fit into each of these categories. Unfortunately, there is no commonly accepted belief on which cities fit into which markets beside the largest cities being Large Market teams. We can be assured that there are both large and small market teams as many of the sports world agree to this. Large market teams tend to attract larger stars in free agency and generally have larger fan bases even without regular team success. On the other hand, small market teams are always afraid of losing their best players for these bigger market cities is the generally agreed upon logic. We chose to solve this problem by separating the cities into different market sizes based on market size with six cities in each market. The large market cities are Atlanta, Chicago, Los Angeles, New York City, San Francisco, and Washington DC. The medium market cities are Boston, Denver, Houston, Minneapolis, Seattle, and Tampa Bay. The small market cities are Charlotte, Cleveland, Green Bay, Milwaukee, Nashville, and Las Vegas. The cities with no teams we selected are Austin, Boise, Louisville, Providence, Tucson, and Virginia Beach. This poses one problem with our research. We tried to control for population size among the cities with no teams to make them equal with the other cities, but there are few of the major cities that do not have sports teams. In addition to this, New York and Los Angeles both have multiple sports teams in all three of the sports we looked at since they are by far the largest cities in the U.S. We decided that having as much diversity in the data as possible was good so we wanted to include these two cities. We also tried to mix in multiple cities that only have sports teams from one or two leagues. For example, Green Bay only has the Green Bay Packers and Nashville only has the Tennessee Titans. For the large markets, Washington D.C. obtained the Nationals in this data set. While some sports teams may have their stadium outside of the city limits, such as the New England

Patriots, we still attributed this team to that major city as we believed this is a widely accepted view.

There is not much available economic data for specific cities available in easy to access forms. We checked through several common government databases such as FRED (Federal Reserve Economic Data) and the Bureau of Labor Statistics with no luck. The problem was that for most Government organizations, they do not need the data for most cities but instead just the largest ones. Ideally, we would be selective for what the city geographically covers, for example hypothetically choosing New York City just to be the island of Manhattan. However, we could not control for this given the lack of data sources for all the cities. Our data came from the Bureau of Economic Analysis and is based upon the Metropolitan Statistical Area definition for a city. This means that these numbers may be different from other sources just because of the area selected of the city but should have no major impact on the accuracy of our models. The city economic data for GDP and personal income we gained included the city Industry Total, which we used as GDP, and then different subcategories with the data being in thousands of current dollars. There were 34 total columns of data from 2001 to 2018. This was the best data set we could find for City GDP so we decided that only having 19 years of data for 24 cities would be enough given the number of potential variables for this data. We also used the BEA's Personal Income Data for the years 2001 to 2018. Some of these variables of interest are Population, Per capita personal income, and employment. These variables are either in dollars or persons. In total, there are 65 columns of economic data we can use. Both of these sets of data could be downloaded by the city for the selected years.

For the sports economic data, we mainly used the Forbes Team Value yearly estimation. Forbes publishes their estimates of the team value for all of the major sports teams and is widely

accepted as the closest to true estimation for team value. It is impossible to know the true value given most teams do not trade on the Stock Exchange and their goal is not to make a profit. The team revenue is also available through Forbes as well, but this is not estimation as the teams can provide this data. We also included the payroll for Major League Baseball teams as they are not restricted by a salary cap. For cities with multiple sports teams, we added the data together for the year of the two teams so the team value for the city is the combined value of its individual teams. Two additional financial columns included of note are Team Purchased and Team Relocated. These were categorical variables marking the years for which cities had a team changed ownership and/or a team move to the city.

The sports non-financial data was obtained from sports reference affiliated sites such as Pro Football reference. We obtained the total attendance numbers for each sports team, home wins, wins, titles, number of teams in the specific league to list some.

To make the data set, we combined all the data from each separate spreadsheet onto one spreadsheet. We chose to do this over excel because there were 162 spreadsheets of data used, and it was easier to maintain data accuracy in excel. In excel, we combined the data by setting the columns to be variable names and then inputting the data by year adding a year column to identify each data point. Since many of the data points from the excel sheets had teams with different names than the city the team was in, it was simplest to copy data from the selected cities onto the spreadsheet.

Methods:

After preprocessing the data, we came up with three questions to answer:

1. Does a city having a sports team affect its GDP?
2. Does market size matter to economic impact?

3. Does a city purchasing or relocating a sports team affect its GDP?

We approached the collected data with regression analysis, a conventional form of supervised learning, to answer these questions using pyspark.ml.regression package. We built and compared three different regression models: linear regression, decision tree regression, and gradient-boosted tree regression. The goal was to predict a dependent variable, GDP (Gross Domestic Product), from a set of independent variables related to sports teams (features) and uncover scalar relationship between them.

We imported the preprocessed data, saved in CSV format, as a spark dataframe. We utilized a scatter matrix with pandas to explore correlation between independent variables and determine potential linear relationships. However, the visualized image of the scatter matrix was difficult to comprehend given the high volume of independent variables (total of 77). We narrowed down the options and studied correlation of variables against GDP only. Based on the result, we decided to only include both sports financial and sports non-financial variables on top of a few select city economic variables, such as “total_employment” and “per_capita_personal_income”, for features. These columns from the initial dataframe were then combined and transformed into a single vector column using VectorAssembler. We created another spark dataframe that contains the transformed features and GDP, which was split into train set (70%) and test set (30%).

We initialized a linear regression object first with the following parameters: maxIter=10, regParam=0.3, elasticNetParam=0.8. The measures we looked for are coefficient values, R-squared, and RMSE (root mean square error), which measures the differences between values predicted by a model and the values observed. We used adjusted R-squared when possible, but the parameter metricName of the function RegressionEvaluator does not support adjusted R-

squared so we would have to use R-squared. We dropped three variables with coefficient value of zero after the first evaluation of linear regression. The process of selecting different sets of variables and evaluating repeated multiple times until we came down to 24 variables.

Next, we built decision tree regression and evaluated R-squared, RMSE, and feature importance. We noticed a pattern that when we removed a variable from a model due to its significantly larger value (0.7~0.9) than any other variables, the new model without the said variable would still have another variable with outstanding feature importance value rolled over to it. This could be explained by considering the nature of decision tree algorithms, but we repeated a similar process as with linear regression in an attempt to optimize the model and produce better RMSE value. We then built a gradient-boosted tree regression and compared R-squared and RMSE of the three regressions.

After exploring sports-related variables and its impact on GDP, we repeated the complete process with market sizes only (large market, medium market, small market, and no teams) and studied coefficients of respective variables.

Results:

As shown in the previous section we ended up with a few models we could use to answer our questions. We trained linear regression, decision tree regression, and gradient boosted tree regression models for each of our renditions. All coefficients and R-squared are drawn from the Linear Regression model. In only the first model with “total_employment” was linear regression not the strongest model.

The first model ran in our analysis was to get a barring on the data as well as utilize a variable, “total_employment”, that would most definitely have an effect on our dependent variable, GDP”. This yielded fantastic results with an adjusted R-squared of 0.970195 and the

decision tree attributed approximately 95% of variance to the variable “total_employment”. While this is a great model and logically the number of individuals employed would have a high correlation with GDP, we felt as though the aforementioned variable could have been weighed too heavily; therefore we decided to remove it altogether and repeat the analysis. Fortunately, running this initial model was not a complete loss. Upon analyzing the results it was shown that the coefficients for the factor variables “large_market” and “no_team” were positive with values 1.090789e+08 and 5.047166e+08 respectively while “medium_market” and “small_market” were negative. These variables were noted in order to see if the signs remained constant in the other models.

The next logical step was to run the analysis by removing the influential variable. This resulted in a model not as strong but with an adjusted R-squared of 0.837875 and an RMSE of 1.25e+09 which is still acceptable. When analyzing each rendition of the models two things held true one of which being the coefficient for “large_market” remained positive and when analyzing the decision tree the variability was spread evenly besides for one variable.

We ran the models again with only four variables, those being the three market size categorical variables and the no team categorical variable The model revealed that a large market (3.642708e+08) was the only variable with a positive coefficient while medium, small, and no team markets had negative coefficients (-1.872610e+07, -1.721790e+08, and -1.923107e+08, respectively).

Finally the question of team ownership changing alongside new teams joining a league were examined. This was accomplished by utilizing the factor variable ‘team_relocated’ in the model which gave us a negative coefficient of -2.486731e+08. The model itself had an adjusted R-squared of 0.86118 and a root mean squared error of 1.04525e+08.

Conclusion:

After running approximately six models, with variations, upon analysis the team arrived at the conclusion that a city having a sports team can impact GDP, but only when market size is large such as New York City or Los Angeles. In some cases the results even showed a city not having a sports team as having a positive effect on GDP. This was hypothesized in the beginning by the team based upon recent news releases showing the New York Knicks as the most valuable sports team in the nation regardless of highly criticized management and a thirty year losing streak. The results would have to be explored further, but we believe that in order for a team to have major economic impact on an area there would have to be three things a large population, the venue would need to be useful for hosting other events such as concerts, and finally the franchise has to operate more as a brand as opposed to a sports team.

The other question the team was able to explore with these models was does the year an organization joins a league or when a team changes ownership have an instant impact on GDP. It was discovered that a team relocating or joining a league during its first year actually had a negative impact on GDP. Theoretically this is due to the fact that the initial move requires the market/city to provide tax incentives along with a multitude of other investments to get the team to relocate to their city in hopes of the investment having a long term gain for a short term loss. The change in ownership still has a slight negative impact on GDP however it is much less than the initialization of a team or relocation. This is probably due to the fact the large upfront investment it takes from an owner to buy the team makes relocation out of the question, the tax incentives are long past, and the culture/branding of the organization is established.

Other questions that could be answered by taking this research further include does it make economic sense for a city to recruit a professional sports team or an additional sports

team to their market; is the return on investment in terms of tax incentives there? Does it have an impact on things that aren't as easy to measure such as happiness of the residents and social opportunities? Another possible avenue of future exploration is researching the localized effect of sports teams on restaurants and bars. It is also possible to see what impact college sports have on their localities as well.