BUS 41204: Machine Learning | Final Project

Business Patterns and Development

Team: Manuel ***Aragonés****, Anastasiya* ***Yarygina****, Andrés* ***Ponce de León Rosas***

Introduction

Industrial growth can be seen as a trigger of economic development and social welfare. It is the sources of employment, higher wages and investment what generates positive externalities on welfare. However, the aforementioned effects can vary depending on the type of industry that the economy is based on. As policymakers we have the power and responsibility to reduce poverty levels and increase general wellbeing of society. The objective of the following project is to analyse the link between industrial composition of regional economies and their socio economic indicators. Using the latest Machine Learning (ML) algorithms[1], in Sections I and II we analyse the industrial mix across the United States counties. We then use this information in Section III to explain and predict two socio economic indicators: (i) poverty index and (ii) income level.

Data Source

Industrial Data: The United States Census Bureau periodically measures the business patterns across counties in the United States. The 2012 Economic Census data contains information such as employment, annual payroll or number of establishments by industry using the North American Industry Classification System (NAICS) and by region up to a county level. NAICS codes are available at different levels of disaggregation. We focused on NAICS level 4, from which we have been able to decompose the US economy into 312 different types of industries for our data set of 3 thousand 143 American counties.

Using this information, we first build a matrix with the total number of establishments (E) for each industry (i) in county (c). Then, using a measure of concentration known as Location Quotient (LQ) defined in Equation 1, we calculate for each county (c) the measure of county's comparative advantage (rca) in industry (i) defined in Equation 2. Finally, we build the

[1] The R code for Sections I and II can be found in the on-line Appendix 1. The R code for Sections III.1 and III.2 can be found in on-line Appendix 2.1 and Appendix 2.2.

competitive advantage matrix where rows corresponds to counties, columns corresponds to industries and cells take the *rca* index. If a county has comparative advantage in the given industry, then the cell in the intersection between the county's row and the industry's column has value "1" and "0" otherwise. The value is "NA" if there are no establishments of this industry in a particular county. By way of example, in Table 1 we show the extract from the matrix for 5 industries in Baldwin, Alabama.

$$LQ_{c,i} = \left(\frac{E_{c,i}}{\sum_i E_{c,i}}\right) \Big/ \left(\frac{\sum_c E_{c,i}}{\sum_c \sum_i E_{c,i}}\right) \qquad (1)$$

$E_{c,i}$ = Total number of establishments of industry i in county c

$$rca_{c,i} = \begin{cases} 1 & if \quad LQ_{c,i} \geq 1 \\ 0 & if \quad LQ_{c,i} < 1 \\ NA & if \quad E_{c,i} = 0 \end{cases} \qquad (2)$$

Table1. **Competitive Advantage of Baldwin, Alabama in 5 industries**

industry index / county name	rca_1131	rca_1132	rca_1133	rca_1141	rca_1142
Baldwin, Alabama	1	NA	1	1	NA

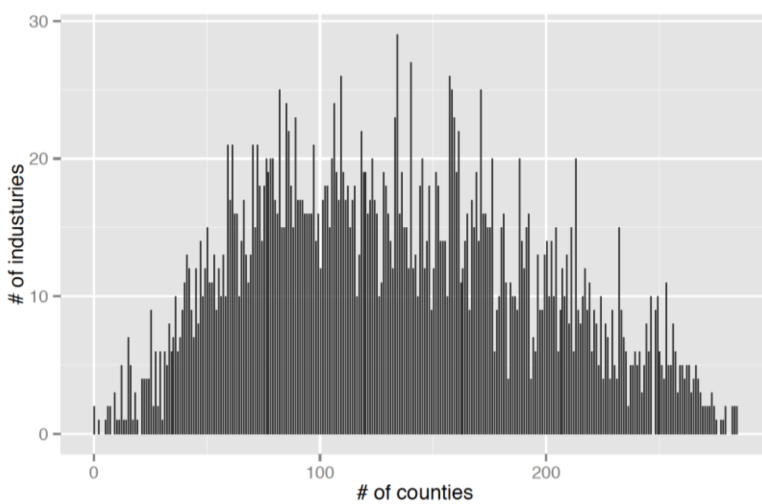
The competitive advantage matrix shows the most promising sectors in each county. We use this matrix as a raw data in this project.

Socio-economic indicators: The United States Census Bureau (USCB) conducts the Current Population Survey every 5 years. It provides social and demographic information for each county. We employ two variables from the USCB database as outcome variables in our analysis: (1) poverty index, which takes value of 1 if the precent of population living below poverty level in a county is larger than the national average and zero otherwise and (2) median household income. We predict these outcomes using other 49 socio-economic indicators from USCB and the competitive advantage matrix.

I Industrial mix in the United Sates Counties

The average county in the U.S. has 133 industries (see Figure 1). The 3 counties with the highest competitive advantage (their industrial mix has the largest number of industries where they have competitive advantage) are: Harris (Texas) with 284 industries, Los Angeles, California with 284 industries and Cook (Illinois) with 283 industries. The 2 most spread industries (industries present in most countries) are: "Gasoline Stations" and "Restaurants and Other Eating Places". These industries are present in 3118 counties.

Figure 1. Distribution of industries. Counts by counties



The average competitive advantage index across counties, computed as the ratio of industries where county has competitive advantage to the total number of industries available in the county, is 0.58.

II Similarity analysis and industrial mix recommendation

Similarity analysis helps understand distribution of industrial mix patterns across counties in the United States. In addition, it can provide advice on industrial composition that gives regional economy higher competitive advantage. In this project we analyse similarities in industrial composition of a selected set of counties. Analysis of this kind, however, can be done for any arbitrary county or region. This provides a useful tool to practitioners who design and implement industrial policy.

Similarity

The similarity analysis was conducted using the "recommenderlab" package from Hashler (2015). Using this package, we identified counties that have similar industrial mix.

For example, **Cook County, Illinois**, has a similar industrial composition to the following counties (in decreasing order of similarity): [1] "Los Angeles, California"; [2] "Douglas, Illinois"; [3] "Crawford, Ohio"; [4] "Atlantic, New Jersey"; [5] "Grant, Minnesota"; [6] "Newaygo, Michigan"; [7] "San Francisco, California"; [8]" Montgomery, New York"; [9] "Fulton, Georgia"; [10] "Plymouth, Massachusetts".

At the same time, **Autauga, Alabama**, 5 most similar counties are: [1] "Loudon, Tennessee" , [2] "Mason, Texas"; [3] "Murray, Oklahoma"; [4] "Henderson, Tennessee"; [5] "Martin, Kentucky" .

Recommendations

Recommender package algorithms can also be used to advice industries to counties, so that they can build more competitive industrial mix. This is useful when policymakers want to explore which industries can spur growth in their region. We focus our attention on Gary's, Lake County, Indiana. We believe that this kind of analysis can be interesting to Gary's practitioners, since the industrial decline in this city caused abandonment of approximately one third of homes.

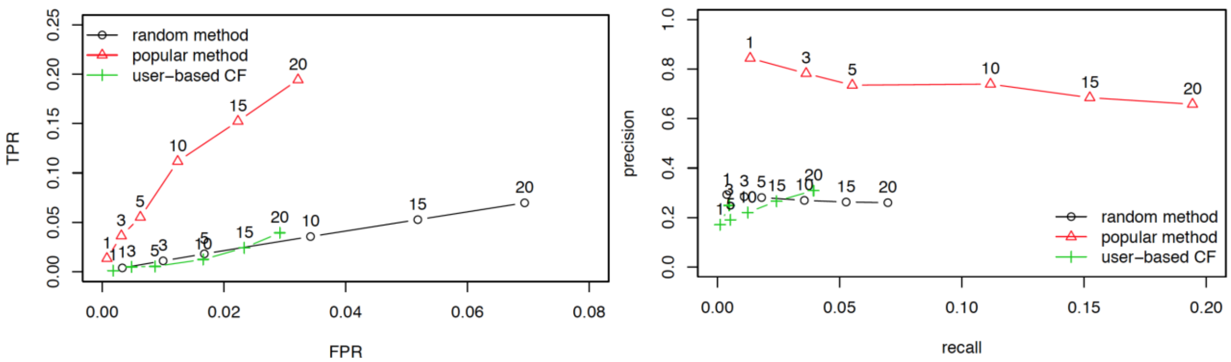
The "popular method" from recommender package suggests the following industries, in order of importance, to Lake County:

[1] "Sawmills and Wood Preservation"; [2] "Museums, Historical Sites, and Similar Institutions"; [3] RV (Recreational Vehicle) Parks and Recreational Camps; [4] Logging; [5] "Support Activities for Forestry"; [6] "Boiler, Tank, and Shipping Container Manufacturing".

Arguably, some of suggested industries could help foster development in this region that much needs it.

We chose "popular method" from the recommender package to make our recommendations because it outperforms "user-based-CF" and the "random-method" in terms of true-positive rate and false positive rate, as shown on ROC curves and Precision-Recall curves on Figure 2.

Figure 2. Precision and Recall Curves for Similarity Algorithms



Note: TPR: true positive rate; FPR: false positive rate.

III Industrial mix and Socio-Economic Indicators

In general, policymakers want to foster growth and development to reduce poverty and improve general wellbeing of society. One of the most challenging problems they face is how to identify specific features that might be causing stagnation, or in contrast, promote growth. In this section we analyse how industrial composition of the US counties can explain and predict poverty index and income level.

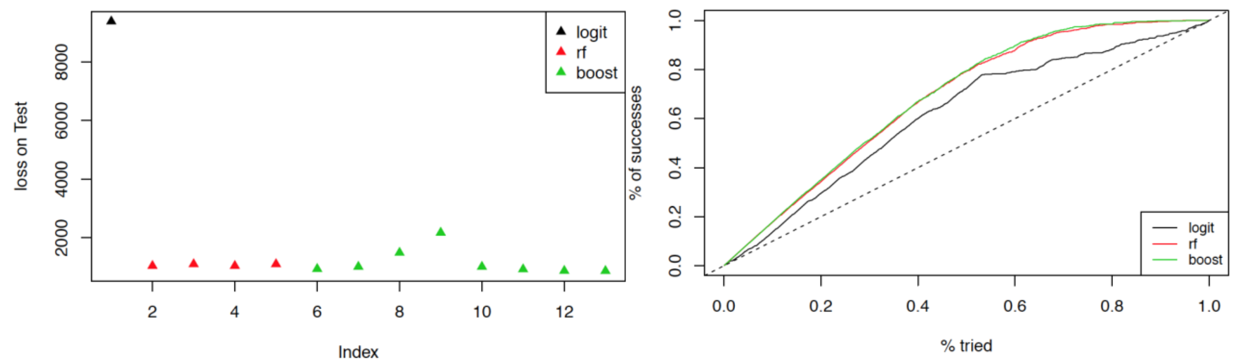
III.1 Poverty Index

As mentioned above, it is difficult to identify variables linked to poverty. However, there are certain ML algorithms that can uncover underlying relationships not identifiable at the first glance. In particular, when the relationship between the outcome variable and predictors is not clear, tree-based methods such as Random Forests (RF) or Boosting (GBM) can be helpful. Depending on the data at hand, one method can yield better performance than another. In this project, we try Logistic, RM and GBM methods to predict poverty index. The objective of our analysis is twofold: (i) identification of the best model, (ii) identification of the best predictors.

Identification of the best model

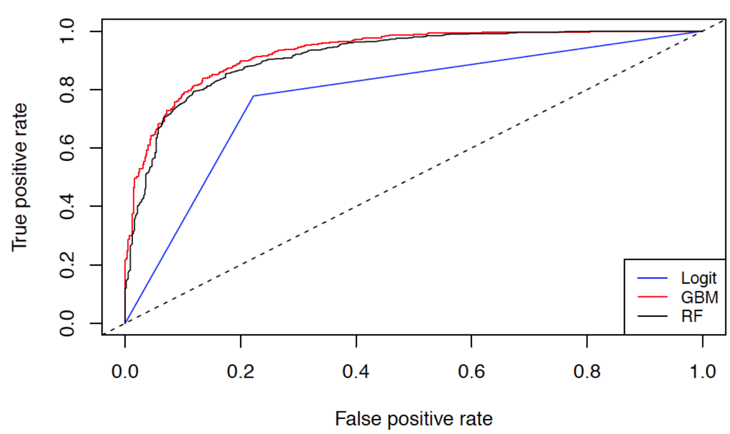
To predict poverty index, we build model using 49 socio-economic predictors from USCB and the competitive index matrix. We train model on train dataset (60% of observations) and test predictive capacity on test dataset (40% of observations). We select across three methods (Logit, RF, GBM) and different specifications of RF and GBM the best predictive model using such criteria as loss, lift and ROC curves. As seen on Figures 3 and 4, GBM (boosting) model with specification depth=4, trees=6000, shrinkage=0.01 has the best predictive capacity.

Figure 3. Comparison of predictive capacity of Logit, RF and Boosting methods. Loss and lift curves



Note: for random forest and boosting methods we plot the results for the best model specification within each method.

Figure 4. Comparison of predictive capacity of Logit, RF and Boosting methods. ROC curves



Note: for random forest and boosting methods we plot the results for the best model specification within each method.

Identification of the best predictors

After having identified the model that yields the best predictive capacity, we use Variable Importance method to find features that are most relevant in this predictive model. As seen in Table 2, boosting model, which is the model that yields the best predictive capacity, relies on features from the industrial mix matrix.

Table 2. Variable importance for the best predictive model

BLACK	BLACK	4.869648
rca_2361	rca_2361	4.784303
FHH_CHILD	FHH_CHILD	3.867464
AMERI_ES	AMERI_ES	3.817313
rca_4529	rca_4529	3.775912
MED_AGE_M	MED_AGE_M	2.817762
VACANT	VACANT	2.722560
rca_2381	rca_2381	2.617724
rca_5222	rca_5222	2.505626
MARHH_CHD	MARHH_CHD	2.465957
rca_1133	rca_1133	2.340757
rca_4461	rca_4461	1.963703
rca_2383	rca_2383	1.898341
rca_6241	rca_6241	1.819618
rca_4451	rca_4451	1.757335
ASIAN	ASIAN	1.645390
PNTCNT_S	PNTCNT_S	1.551606
rca_6214	rca_6214	1.424837
AGE_85_UP	AGE_85_UP	1.301156
OTHER	OTHER	1.298451

Note: Variables starting with "rca_" prefix are indices of competitiveness in the industrial mix matrix. The rest of the variables are socio-economic indicators.

In contrast, as shown in Table 4, RF and Logit models, both build their predictive capacity on socio-economic indicators.

Table 4. Variable importance for Random Forest and Logit models

Logit		Random Forest	
POP2010	19655854	POP2010	8.533410
POP10_SQMI	26336341	POP10_SQMI	11.934628
POP2012	44093012	POP2012	8.528282
POP12_SQMI	26234559	POP12_SQMI	11.800887
WHITE	25475496	WHITE	12.934156
BLACK	35977124	BLACK	26.748412
AMERI_ES	61133435	AMERI_ES	13.667884
ASIAN	2861453	ASIAN	10.844283
HAWN_PI	58714781	HAWN_PI	8.149513
HISPANIC	40384000	HISPANIC	11.313124
OTHER	26465806	OTHER	11.740289
MALES	136529434	MULT_RACE	10.568144
AGE_UNDER5	23783434	MALES	8.752840
AGE_5_9	14516149	FEMALES	8.355474
AGE_10_14	10079067	AGE_UNDER5	9.058235
AGE_15_19	13107504	AGE_5_9	8.917370
AGE_20_24	43589069	AGE_10_14	9.148620
AGE_25_34	13792236	AGE_15_19	8.814471
AGE_35_44	16718007	AGE_20_24	12.397916
AGE_45_54	33068513	AGE_25_34	9.556742

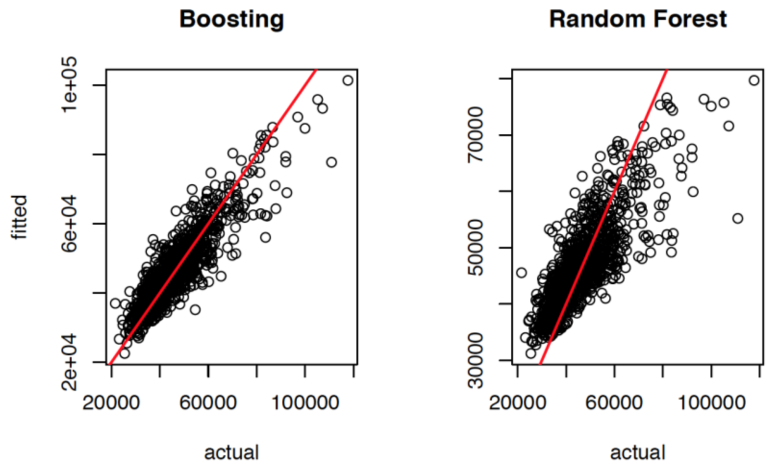
III.1 Household Income

A largely used proxy of welfare of society is the average or median household income. In this section we use tree-based methods to uncover the relationship between income level and the industrial mix in the US counties. The key difference with respect to the analysis performed in the previous section is that the outcome variable is continuous. Our objectives are the same: (i) identification of the best model, (ii) identification of the best predictors.

Identification of the best model

To predict median household income, we build predictive model using 49 socio-economic predictors from USCB and the competitive index matrix. We select the best model using as a creation the out of sample error. According to this criterion, boosting model with depth 4, 5000 trees and shrinkage=0.01 yields the best predictive capacity. As shown on Figure 5, the best specification of RF model tends to overestimate the outcome variable, when compared to the best specification of boosting model.

Figure 5. Model fit forecasts versus actual values



Identification of the best predictors

The variable importance results for the household income are qualitatively the same as those obtained for poverty index. Namely, the most important variables in the boosting model are drawn from the industrial mix matrix, while RF model draws its predictive capacity on socio-economic indicators (see Table 4).

Table 4. Variable importance for the best (Boosting) and Random Forest models

Boosting			Random Forest	
rca_5415	rca_5415	8.726229	POP2010	2077449787
rca_4529	rca_4529	5.472310	POP10_SQMI	5082578184
ASIAN	ASIAN	5.168479	POP2012	2639942405
VACANT	VACANT	3.834971	POP12_SQMI	6022587267
rca_4471	rca_4471	3.798647	WHITE	4073373880
rca_6116	rca_6116	3.214426	BLACK	3261925293
MARHH_CHD	MARHH_CHD	3.214350	AMERI_ES	1786009218
POP10_SQMI	POP10_SQMI	2.841777	ASIAN	6824962260
rca_5416	rca_5416	2.352813	HAWN_PI	2112224355
rca_4461	rca_4461	1.912648	HISPANIC	1848020293
rca_2381	rca_2381	1.852743	OTHER	1831589607
FHH_CHILD	FHH_CHILD	1.803999	MULT_RACE	2183968384
rca_6241	rca_6241	1.783765	MALES	2142423342
POP12_SQMI	POP12_SQMI	1.598140	FEMALES	2054066958
rca_2361	rca_2361	1.583794	AGE_UNDER5	2312553982
rca_1133	rca_1133	1.548054	AGE_5_9	3285406954
rca_7224	rca_7224	1.412911	AGE_10_14	3462768960
rca_4451	rca_4451	1.362849	AGE_15_19	1873626495
WHITE	WHITE	1.293358	AGE_20_24	1939506253
MED_AGE_M	MED_AGE_M	1.292223	AGE_25_34	1995387114

Conclusions

In this project we have explored industrial mix in the US counties. We also analysed how industrial diversity and competitiveness can be used to predict poverty index and income level. The key takeaway that we have make is that level and diversity of regional industrial mix is a relevant factor determining socioeconomic outcomes. When making recommendation on the industrial mix we have to take into account, though, that suggested changes that potentially foster welfare do not necessarily improve circumstances of the population at large. Consider the example of gasoline stations. Even though this industry can be recommend as fostering economic growth, exponentially increasing the number of gas stations would hardly improve welfare of average citizens. This paper is a good starting point to further discovery of new links between industrial development and welfare.