Big Data Approaches to Modeling the Labor Market

Gerunov, Anton

Sofia University "St. Kliment Ohridski"

2014

Big Data Approaches to Modeling the Labor Market

Anton Gerunov
Sofia University "St. Kliment Ohridski", Faculty of Economics and Business Administration
125 Tsarigradsko Shosse Blvd., 1115 Sofia, Bulgaria
gerunov@uni-sofia.bg

Abstract: The research paper leverages a big dataset from the field of social sciences – the combined World Values Survey 1981-2014 data – to investigate what determines an individual's employment status. We propose an approach to model this by first reducing data dimensionality at a small informational loss and then fitting a Random Forest algorithm. Variable importance is then investigated to glean insight into what determines employment status. Employment is explained through traditional demographic and work attitude variables but unemployment is not, meaning that the latter is likely driven by other factors. The main contribution of this paper is to outline a new approach for doing big data-driven research in labor economics and apply it to a dataset that was not previously investigated in its entirety.

Keywords: Labor market, Unemployment, Big data, WVS

1 Introduction

Traditionally econometric modeling has perused relatively small datasets to answer questions of substantive economic interest. A typical approach would be to formulate a scientific hypothesis, collect a limited number of theoretically-informed variables and subject them to statistical testing using largely linear models under the assumption of normality of the underlying data distribution [1]. This methodological framework has provided many fruitful insights and deepened our understanding of the underlying economic processes.

However, it suffers from a number of potential pitfalls –small samples raise questions about bias, estimation precision, and generalizability. Further, the number of observations imposes constraints on the maximum feasible number of independent model variables so that a researcher often has to make a judgment call on what to include. The availability of large datasets containing hundreds of millions or more data points ("big data") can now help overcome those limitations and provide an additional perspective on economics research [2].

This paper focuses on modeling the labor market and proposes a way in which a well-known machine-learning algorithm can be applied to glean novel conclusions from a large-scale dataset. We first review traditional theories and methods for modeling the labor market and outline how big data approaches can supplement and enrich the existing paradigms.

Then we fit an ensemble Random Forest model and outline the model qualities and main results. The paper concludes with directions for further research and possible applications of this very new approach in economics.

2 Labor Market Theories and Approaches

Labor market theories can be broadly subdivided into two main groups: microeconomic theories, and macroeconomic theories. Micro-level approaches emphasize the supply of labor as resulting from optimizing decisions by households and the demand for labor as resulting from optimizing decisions by firms. At the resulting equilibrium employment, the firms pay for their workers a wage equal to their marginal productivity. A newer strand of theories – the "search and matching" theories – have focused much more on the process by which workers are matched to certain positions, given their useful economic characteristics like productivity, and employers' needs [3].

Macroeconomic approaches have tended to emphasize the connection between unemployment and the level of economic activity, thus intimately connecting labor market developments with economic growth and recessions [3]. The logic behind this is lucid – greater output is usually produced by an increase in inputs, and labor is one of the most important ones. Those two angles to understanding employment are of crucial importance – they emphasize firm-level needs, labor market processes, and macroeconomic structural needs. Some authors also point at the importance of psychological characteristics, values, and perceptions at the individual level for a given adult's employment prospects [4], [5]. The availability of large-scale data on individual characteristics allows us to see what values and attitudes determine employment in addition to a worker's productivity and job vacancies.

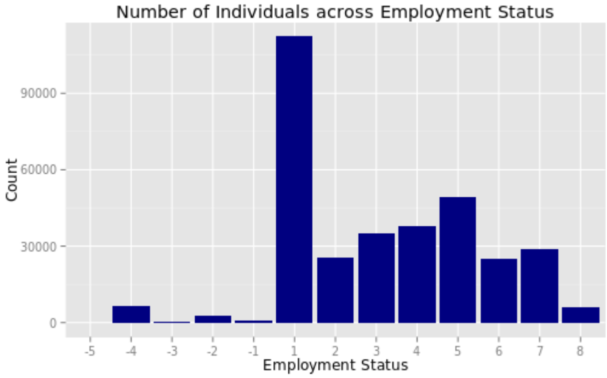
3 Data and Methods

To add the individual level dimension to modeling the labor market situation, we will explore sociological data coming from one of the largest social science primary data collection initiatives – the World Values Survey [6], and will outline a possible approach to gleaning insight from such large data for the purposes of economics research.

3.1 World Values Survey Dataset

The World Values Survey is an ongoing project since 1981 whereby respondents from almost 100 countries, representative for approximately 90% of world population, are questioned regarding their values, attitudes, beliefs, life conditions, demographic characteristics and evaluations. Currently about 330,000 respondents have been interviewed over the Survey's six waves and an additional seventh wave is presently under way. Separate variables from different waves have been used extensively by economists, psychologists and other social scientists to study in depth questions about political participation, economic development, culture and psychological issues. This wealth of data has never been analyzed in its entirety as the Survey Committee made publicly available the beta version of integrated and compatible data throughout all waves only very recently. This dataset contains 330,354 observations of 1377 variables for a total of nearly 455 million data points. While more common in machine learning contexts, such volumes of data are rarely studied in the social sciences. This paper will present a feasible approach to analyzing it in view of possible computational resource constraints. We will look in particular into the variable Employment.status, which codes whether the respondent is employed full-time (coded 1), part-time (2), self-employed (3), retired (4), housewife (5), student (6), unemployed (7) or in other position / not asked (codes 8 and negative). The WVS data will allow us to see what individual-level characteristics are important for classifying individuals in either of those positions, thus providing a useful exploratory analysis which can spur additional research at the boundary between economics, psychology and sociology.

Fig. 1: WVS Dependent Variable Distribution



3.2 Data Processing and Dimension Reduction

A dataset of such dimensionality – with a lot of variables per observation (so-called "fat data") – provides a clear computational challenge, which calls for optimization of its size. As with many big data sets, the World Values Survey is also a sparse dataset – some of the observations have either missing values or very low variance. Those are unlikely to be useful as classification and regression algorithms leverage variable variances. We remove the variables where the variance is either zero or they have few unique values with the ratio of the most common to the second most common is at least 95 to 5. A key insight from social science research is that a lot of the collected variables exhibit high multi-collinearity due to the complex feedback loops in social and economic systems. This means that some correlates can be dropped at the expense of very small information loss, so we check all pairs of variables with correlation of above r=0.9 and remove the one with the largest mean absolute correlation. We should note that rank variables need to be processed through rank correlation (e.g. Spearman), while continuous – through continuous correlation (e.g. Pearson). In the current dataset results from both were practically the same.

Those two simple steps lead to a dramatic reduction of the number of variables per observation – these now count 233, or a total of 77 million data points. Such a dataset can be usefully analyzed with relatively fewer computational difficulties. For the analysis presented here we use the R Studio IDE on a server of 24 GB RAM and an eight-core Intel i7-4770 processor at 3.5 GHz, and will report calculation times for both sequential and parallel computation. There are many other possible approaches for dimension reduction that economists may fruitfully apply in their big data modeling. For example, experimentation with Principal Component Analysis (or Singular Value Decomposition) proved suboptimal for the WVS data but may be useful in other contexts. Large volumes of information also make it possible to select only a subsample from the data thus decreasing computation time.

3.3 Random Forest Ensemble

After pre-processing we can model the data by using a scalable machine-learning algorithm. Classification and Regression Trees (CART) are familiar to economists and are certainly one of the most useful and easier to interpret tools in the big data toolbox. There are also many standard and useful implementations but trees suffer from a number of problems, most notably high susceptibility to noise in data. Ensembling a large number of trees ameliorates this and effectively decreases the prediction or classification variance. Those ideas lead to the combination of trees into forests, so-called Random Forests, which can leverage a large number of trees for prediction purposes [7], [8]. Research in the machine learning literature has revealed that Random Forests show very good performance on a large number of problems under a wide range of specifications, making them a widely applicable algorithm.

For this reason we will pursue modeling with them but naturally there are many other viable approaches [2].

The Random Forest algorithm proceeds as follows [7]: for ***b=1*** to ***B***, it selects a bootstrapped sample from the training data ***Z***. For our research, the size of this sample equals the whole dataset. Then, ***m*** variables are selected at random from the ***p*** predictors, so that ***m = sqrt(p)***. Out of those 15 variables the best variable/split-point is picked and the node is split into two daughter nodes, thus growing the tree T_b. We have limited the number of terminal nodes to a maximum of 10,000 for computational convenience. After all trees are grown (with ***B = 504***), those are combined in an ensembled Random Forest $\{T_b\}_1^B$. The prediction for a new point **x** in case of a regression Random Forest is then:

$$f_{rf}{}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x).$$

In case of classification, the trees in the ensemble generate a classification $C_b(x)$ and then “vote”, and the majority vote assigns the class, or:

$$C_{rf}{}^{B}(x) = majority.vote\{C_b(x)\}_1{}^{B}.$$

We fit a classification random forest to the WVS data to investigate what variables are most important for classifying a respondent as Employed, Non-employed, Retired, Part-time Employed, Self-employed, Student, Housewife, etc. The model was sequentially calculated for 72.9 minutes. Since a Random Forest is particularly suitable for parallel computation, we also split the 504 trees into 8 processes of 63 trees. The latter model was computed for 36.6 minutes, or an improvement of about 50%. Big data analytics of such scale reaps the benefits from parallelization and the algorithm selection needs to be performed with this possibility in mind. Reported results are for classification Random Forest, which is the more theoretically sound to apply. If one experiments with a regression Random Forest, the computation time increases to more than 20 hours sequentially and 8.3 hours in parallel.

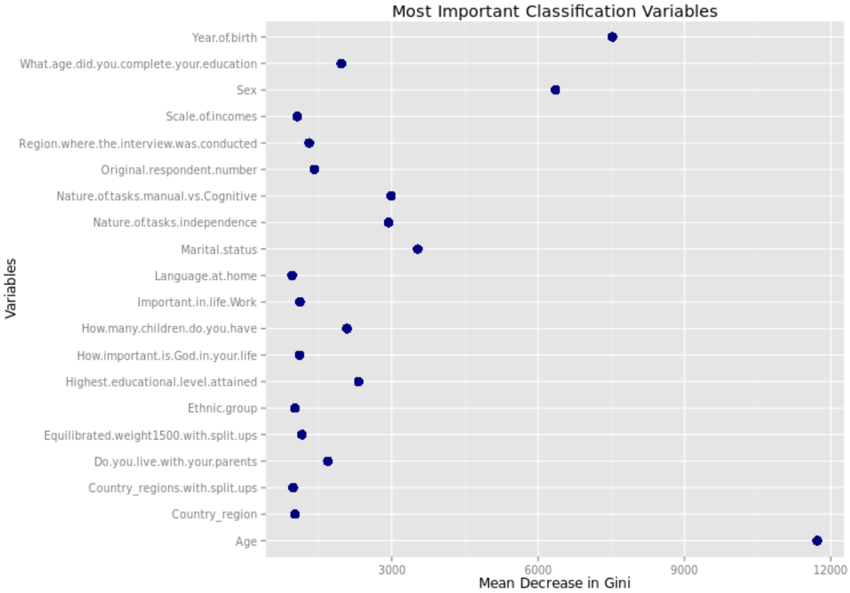
4 Results and Discussion

The calculated model provides very good results, with an estimated out-of-bag (OOB) error rate of 39.29%, meaning correct classification of 60.71% of cases under scrutiny. Looking at the most important classification variables in Figure 2, we observe few surprises. Age, sex, marital status, ethnic group, education, and region top the list. Work attitudes which have captured the imagination of organizational theorists such as preference towards specific types of tasks (manual or cognitive) and preference for workplace autonomy are also present.

The variables we observe are standard for the labor market literature and emphasize the importance of individual-level demographics well above attitudes in the job allocation process. In that sense the current exploratory study confirms results from previous research on the labor market.

One exception is the importance of God in one’s life, which is slightly negatively correlated with better employment prospects. Atheists seem to be more likely to be employed full-time that the devout. A possible explanation could be their more pragmatic and less metaphysical focus. Such results naturally need to be interpreted with care and further investigated.

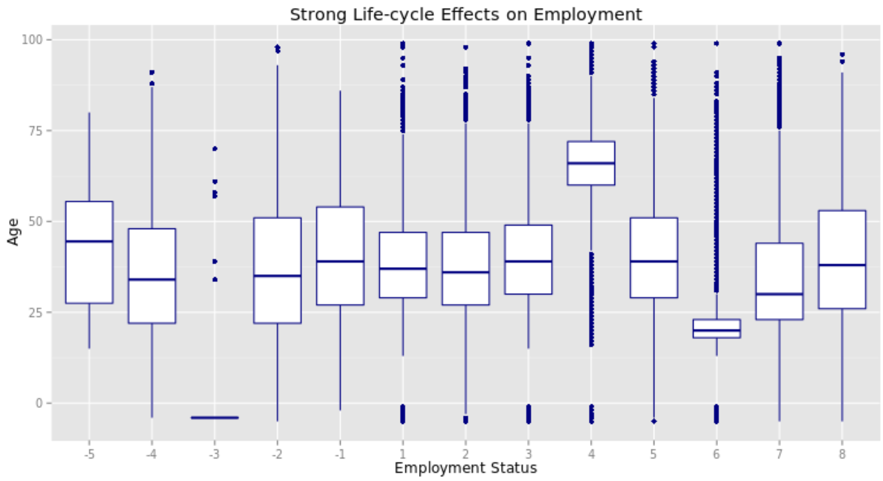
Fig. 2: Random Forest Classification Variables with Highest Mean Decrease in Gini



The variable with greatest predictive power by far is Age, underlining the strong life cycle effects on individual employment prospects (Fig. 3). Full-time and part-time employed individuals (codes 1 and 2) tend to be of the same age, whereas self-employed (code 3) are slightly older. Retired (code 4) and students (code 6) are also easy to recognize. The unemployed (code 7) tend to be on average younger than those in employment. This piece of statistics is likely driven by the fact that unemployment has larger prevalence among young adults than in the general population.

The abridged model confusion matrix is presented in Table 1. Individuals in full-time employment are largely correctly classified, showing that there are some very distinct characteristic within this group. Other groups in the labor force such as self-employed, or part-time employed are largely classified as Employed, meaning that these are only little different in attitudes and perceptions from those actually having a full-time job.

Fig. 3: Median Age of Individuals Across Different Employment Statuses



Unemployed individuals are difficult to classify correctly. One possible interpretation of this result is that there is hardly a stable set of psychological attitudes, social beliefs or evaluations that clearly distinguish the unemployed from the other respondents. These results underlie the fact that reasons for unemployment will need to be found in individual labor productivity, motivation, and overall job market conditions, rather than in a set of psychological beliefs and attitudes.

Table 1: Truncated Confusion Matrix for Random Forest Classification

Class / True	*Full-time*	*Part-time*	*Self-employed*	*Retired*	*Housewife*	*Student*	*Unemployed*	*Other*	*Error Rate*
Full-time	102604	30	1152	2791	3352	1899	260	58	8.6%
Part-time	19807	187	630	1381	2128	1352	114	84	99.3%
Self-employed	23079	9	6070	2041	2618	634	208	129	82.6%
Retired	6493	4	145	29224	1646	57	202	28	22.8%
Housewife	12032	24	294	3040	32893	603	313	63	33.4%
Student	5987	0	21	48	576	17745	482	2	28.7%
Unemployed	15874	2	526	1418	3450	3698	3826	57	86.7%
Other	3067	12	165	689	583	518	236	571	90.2%

The results so far underline an interesting conclusion – we are very well aware what makes a person employed – active age, good education, being in the right ethnic group and the right region of the country, and possessing pro-work attitudes. What remains elusive is what makes a person unemployed – even those with favorable characteristics might end up without a job, and we seem to be unable to statistically distinguish between the former and the latter using individual demographics and attitudes. In that sense the current paper opens interesting venues for labor market research.

Firstly, employment and unemployment do not seem to be the flipsides of the same coin, as is commonly assumed in labor economics, but rather two distinct conditions that need to be studied separately. Secondly, demographics, psychological attributes, and social perceptions seem unable to explain unemployment and other explanatory factors need to be investigated further. An obvious determinant of unemployment is individual labor productivity which probably plays a role. Another viable contender is chance. If there is a structural labor market need for downsizing the labor force, some individuals may lose their jobs purely by chance, irrespective of their objective qualities. While this interpretation substitutes randomness for causality, it might be worth exploring further.

Thirdly, such results can be uniquely gleaned only through leveraging a combination between big data and advanced machine learning algorithms. Under the standard econometric inference testing approach one could utilize a version of the Generalized Linear Model to interpret regression coefficients and their significance levels. This will only show that some regressors reach statistical significance, and are therefore important for predicting the dependent variable. We will not be able to see the subtle differences and discern that employment and unemployment are two very different conditions that may need to be studied within distinct theoretical and analytical frameworks.

5 Concluding Remarks

The current exploratory study leverages a new and previously unutilized dataset – the complete integrated comparable World Values Survey data spanning 1981-2014 – to investigate if individual level employment can be explained by a combination of demographics, psychological attributes, and social attitudes. Using a Random Forest model we classified respondents, with over 60% out-of-bag correct classification. Employed individuals were largely correctly classified, but the unemployed ones were more challenging. A possible reading of this result is that unemployment is hardly defined by traditional individual level attributes (age, education, region, work attitudes) but could be attributed to either individual labor productivity or structural labor market characteristics and randomness of outcomes. Such results can serve to refocus the research agenda in labor economics and steer it towards a better understanding of the determinants of individual employment status.

References

[1] Greene, W. (2011). *Econometric Analysis, 7th Edition*. US: Prentice Hall.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The Elements of Statistical Learning*. NY: Springer.

[3] Romer, D. (2012). *Advanced Macroeconomics*. US: McGraw-Hill.

[4] Kalil, A., Schweingruber, H. & Seefeldt, K. (2001).Correlates of Employment among Welfare Recipients: Do Psychological Characteristics and Attitudes Matter? *American Journal of Community Psychology*, Volume 29, Issue 5, 701-723.

[5] Kessler, R. C., Turner, J. B., & House, J. S. (1987). Intervening processes in the relationship between unemployment and health. *Psychological Medicine*, *17,* 949–961.

[6] World Values Survey, Wave 1-6 1981-2014. (2014). World Values Survey Association (www.worldvaluessurvey.org).

[7] Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5-32.

[8] Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2-3, 18-22.