ROBERT H. SMITH
SCHOOL OF BUSINESS
Leaders for the Digital Economy

Employee Demographics:
Exploring similarities and differences
in the private and public sector

Abrar Al-Hasan, Jorge Christian, Hideki Kakuma, Li Wei Chen, Yavor Nikolov

Course BUDT733

Professor Galit Shmuéli

May 12th 2008

Executive Summary

Acting as a Human Resources consulting firm, our team used a dataset from the 1994 U.S. Census to build an explanatory model that explains which demographic factors influence a person's placement within the public or private sector. Our data consisted of 13 variables -Age, Education, Marital Status, Occupation, Relationship, Sex, Race, Native Country, Capital Gain, Capital Loss, Hours worked per Week, Less than or Greater than $50K, and Work class (see **Exhibit A** for a detailed data description). The goal was to find the right strategy for targeting desired employees for our clients, and to help prospective employees target the right sector for them (Public or Private). We also hoped to find which sector(s) have deficiencies in a demographic and find which demographics are heavy in a particular industry so that we could make some inferences as to why the gaps exist, and how our clients can correct for it. Our clients are composed of both public and private organizations, and we also publish our research materials for prospective employees.

Overall, our model yielded general results which have implications for our clients depending on their needs but the more actionable results we found were specifically for the Public sector. Since our study was very general our clients can derive conclusions from the study depending on their goals and objectives. The following table summarizes a few findings:

Private	**Public**
• Higher percent of white people (including Hispanics)	• Older workforce (median 41)
• More sales professionals	• More flexible work hours (greater number of people working less than 40 hrs/week)
• Greater variance in education (less than high school to Ph.D)	• More females
• More people from "other" minority groups (excluding Black, Asian)	• More highly educated professionals

Out results give both our job-seeking clients and HR departments in both sectors important takeaways based on their goals. For the private sector, we did not find many clear actionable items from our high level analysis due to the wide variety of companies with various needs resulting in a wide variety of demographics. From our analysis, we find that if an HR firm needs actionable results for the private sector then it must do so focusing on a narrower domain in the private sector. However, for example one implication for private firms needing highly educated people is that they could find a strong supply in the Public sector.

In the Public sector, having an older work force can become a cause for concern. As employees get older the likelihood of a 'brain drain' from retirement, illness, etc increases. Depending on the organizations needs and goals, this could be a very important issue to address immediately. Creating an environment friendly towards younger employees may become a must (e.g. emphasizing quality of life and recruiting directly from universities). Another interesting note is that the Public sector has a very strong need for highly educated people. Attracting and retaining these people can become expensive especially if faced with Private sector pressure offering higher salaries. Also, our analysis showed a stronger female presence in the Public sector so this can be positive or negative depending on the agencies goals (e.g. strong female presence can attract more women).

Lastly, for our job-seeking clients, depending on their profile and their needs we can give them a sense of where they might be most needed and feel most comfortable. For example, highly educated, females over 40 might find many similar colleagues in the Public sector. Also, the Public sectors flexible work hours can be very attractive for those needing more time outside of work.

In general, our study could have a variety of implications for our clients depending on their needs and this model will help our firm work with their individual cases to tailor an appropriate strategy.

Technical Summary

Exploratory data analysis process

1. Data reduction

Having a considerable large dataset (more than 30,000 records with 15 variables), we started with a data reduction process to eliminate records that would not add value to our analysis. We looked closely into each variable and found that some had too specific information that would be irrelevant, or seriously impair the parsimony in our analysis.

Firstly, we decided to drop redundant variables, such as *Education* (captured in the variable *Education number*) and *Relationship* (sufficiently described for our purposes by *Marital status*)

Then, we reduced the categories in the different variables by aggregating them into larger ones. For example, the variable *Native-Country* shows every single country where each record (i.e. person) was born, a total of 42 categories. However, as an HR consultancy operating in the U.S., we concentrated on aggregated regional background information and grouped the countries into six groups: *Asia-Developed*, *Asia-Developing*, *European*, *Hispanic*, and *North America* and *Other*. For this purpose we used our domain knowledge of the social and cultural traits of the different countries and we looked at the counts for different countries.

Next, we also binned *Marital-Status* into fewer groups for the same reason as the previous example. Since we are HR consultancy firm, marital status was important because it gives clues as to their lifestyle like for instance that they may have children. We decided that whether the person is married is an important factor and created three groups. The first one was *Married* (*Married-AF-spouse*, *Married-civ-spouse*, *Separated*, and *Married-spouse-absent*); the second one was *DivWid* (*Divorced* and *Widowed*); and the third one was *Never-married*.

2. Findings from boxplots

Seeking further information on relationships from our data, we created boxplots which gave us some interesting insights. First of all, **Exhibit B** shows that *Public* had great flexibility regarding *Hours-per-Week*. That is, about 44% of *Public* are outliers whereas the working hours in *Private* are more concentrated around the median, although both *Public* and *Private* had the same median value, 40 hours. We hypothesized this may be due to several factors like workers in *Public* could have important responsibilities outside of work such as being parents or caring for elderly relatives and more time-flexible jobs would be tremendously valuable for them. Also, in terms of *Age*, *Public* had about a 5 year older median than *Private*. This might be attributable to the high responsibility and experience requirements for many jobs in the public sector, as well as the more flexible working hours, better suited for older employees.

Another aspect we found from the boxplots was the difference in education. **Exhibit B** shows *Education-num* (a number assigned to level of education) and it shows that people in *Public* tend to have a higher degree of education. This was at first somewhat surprising. However, this turned to be reasonable, because *Private* includes a number of jobs that do not require higher education, such as janitors or checkout clerks. To sum up, we found that *Public* sector had a higher degree of education, older age, and more flexibility regarding working hours.

3. Data relationships

We further examined the strength of the relationships between the categorical variables in our dataset by using Chi-squared tests. The *Occupation* variable has the highest Chi-squared number (4026.73), followed by *Race* (173.59) and *Income class* (90.75). From a detailed visualization, it proves that most occupations are much more represented in the private sector.

We decided to delete two sub-categories in “occupation”, “Armed forces” and “Priv-house-serv”, since these are, understandably, exclusively to the public and private sector. Hence, they do not provide any information in separating the two work classes.

4. Pivot Tables

In order to compare the categories within the variables based on percentage ratio rather than counts (necessary due to the much larger number of records in the private sector) we created dummies for all categorical

variables. We then ran pivot tables and found that people with a *Sales* occupation and *Hispanic* race (derived from country of origin) are much less likely to work in the Public sector. Moreover, *Race-white* and *Sex-male* are the other two variables having a higher proportion in the private sector (see **Exhibit C**)

Model Estimation and Interpretation

1. Classification Tree

Since XLMiner only takes 10,000 rows when modeling, we created a random sample from our approximately 30,000 rows, and used this random sample to run the classification tree on. We ran the classification tree on our 29 variables, and since our goal is explanatory, we did not prune the tree. The following variables turned out to be the most important predictors in differentiating the profile of a public job versus a private job, ranked according to the level at which they appear in the tree (see **Exhibit D**): *Occupation-Professional-Specialty, Occupation-Sales, Education Number, Age, Sex, Hours-per-week.*

The classification tree gave us a low percentage error on the training data of 14.02% (see **Exhibit D**). Also, the significant predictors shown in the tree are consistent with our findings from the boxplots and pivot tables. In general, we can note from the tree that there is a higher percentage of professional specialty occupations (e.g. lawyers, doctors, etc.) in the public sector, more females in the public sector, more variety in occupations in the private sector, and lastly more flexibility in hours per week in the public sector.

2. Logistic Regression

Using the knowledge obtained from the visualizations and classification trees, we then performed numerous logistic regression models in order to better understand the public and private profiles.

The final logistic regression was determined iteratively. We started with a model that included all of our 29 variables. Again, we used our random sample data. The logistic regression model gave us a Multiple R-squared of 17.08% and a total percentage error of 13.63% on the training data, which was lower than the classification tree's overall error. However, a lot of the variables had high p-values, and some were contradicting the tree's predictors. Therefore we decided to run a best subset model.

The best subset model gave us a total of 18 predictors that were all significant at a 5% significance level (see **Exhibit D**). These predictors were also consistent with the tree's predictors. Overall, the logistic regression shows that demographic features such as age, education, race, type of occupation, and the number of hours worked per week are important predictors to profile public sectors versus private sectors. Also, this 18-predictor model has a Multiple R-squared that was very close to the original 29-predictor model of 16.93% and an overall error rate of 13.64%.

3. Evaluation of Model Performance: Goodness of Fit

Since we are seeking an explanatory model, model performance is based on measures of overall fit, such as the Multiple R-squared, p-values and training data classification error, and other factors such as interpretability. Both models were useful and easily interpretable. In the end, we chose the logistic regression as our final model since it has a lower training data error and a straightforward (parsimonious) interpretation using odds. However, we also used the tree to complement the profile of public and private employed people, and to cross-check the results from the logistic regression.

Overall, our findings from the exploratory data analysis, backed up by the results from the logistic regression and the classification tree, led to the example of profiles for people employed in the public and private sector outlined in the executive summary. These high-level, general profiles can be used by our company to draw tailored conclusions and better fulfill the individual needs of our clients.

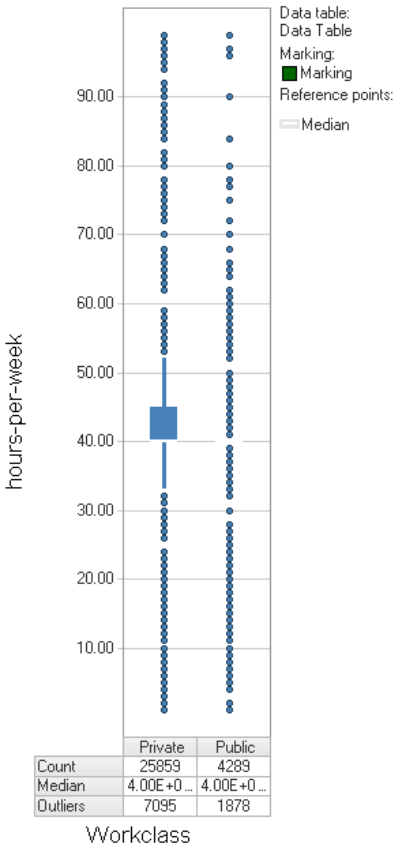
Exhibit A: Data Description

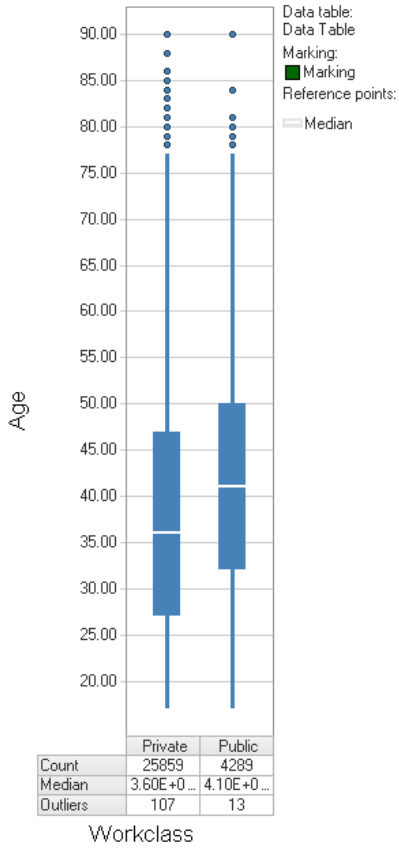
Data Source: Census bureau database: http://www.census.gov/ftp/pub/DES/www/welcome.html (1994)

	Variable	Type	Measurement
1	Age	Numerical	Continuous number indicating a person's age.
2	Education number	Numerical	Continuous number indicating the level of education.
3	Marital status	Categorical	Married, Divorced, Never-married, Separated, Widowed.
4	Occupation	Categorical	Occupations (ex. Tech-support, Craft-repair, Other-service, Sales, Exec-managerial).
5	Relationship	Categorical	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
6	Race	Categorical	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
7	Sex	Binary	0= Female, 1=Male.
8	Hours worked per week	Numerical	Hours worked per week.
9	Native country	Categorical	Native countries.
10	Capital gains	Numerical	Continuous number measuring the capital gains.
11	Capital loss	Numerical	Continuous number measuring the capital loss.
12	Income class	Categorical	Indicating whether income is more than $50K per year or not.
	Y-Variable: Work Class	Categorical	Private, Public.

Sample size, n= 32561; Number of variables, k= 13

Exhibit B: Boxplots





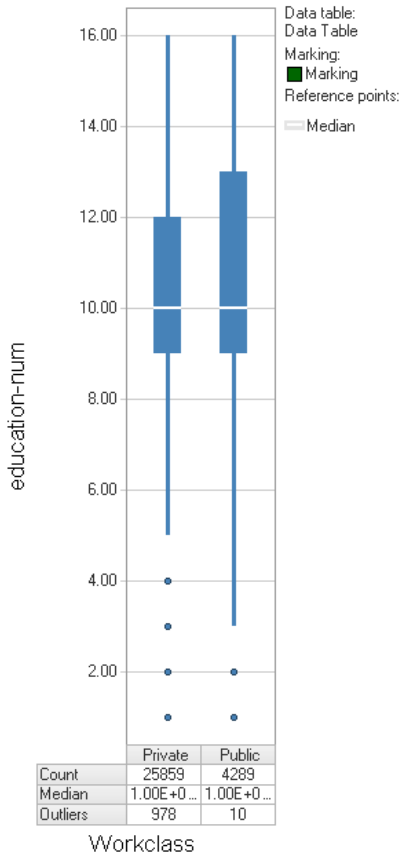


Exhibit C: Pivot Tables

Count of Workclass_ Public	Workclass_ Public		
Hispanic	0	1	Grand Total
0	85.45%	14.55%	100.00%
1	93.98%	6.02%	100.00%
Grand Total	85.79%	14.21%	100.00%

Count of Workclass_ Public	Workclass_ Public		
sex_ Male	0	1	Grand Total
0	83.42%	16.58%	100.00%
1	86.92%	13.08%	100.00%
Grand Total	85.79%	14.21%	100.00%

Count of Workclass_ Public	Workclass_ Public		
race_ White	0	1	Grand Total
0	80.49%	19.51%	100.00%
1	86.65%	13.35%	100.00%
Grand Total	85.79%	14.21%	100.00%

Count of Workclass_ Public	Workclass_ Public		
Sales	0	1	Grand Total
0	83.99%	16.01%	100.00%
1	99.11%	0.89%	100.00%
Grand Total	85.79%	14.21%	100.00%

Exhibit D: Logistic Regression and Classification Tree outputs

Classification Tree

Level 1	Prof-Special								
Level 2	Sales	Education_Num							
Level 3	Age	Age	Education_Num	Education_Num					
Level 4	Education_Num	hrs_per_wk	Age	Age	Age	Sex_male	Sex_male	Education_Num	
Level 5	Age	Age	Education_Num	Education_Num	Age	Age	Age		
Level 6	hrs_per_wk	Craft_Repair	hrs_per_wk	hrs_per_wk	Machine_Op	Craft_Repair	class_>50K	Education_Num	hrs_per_wk

Error Report			
Class	# Cases	# Errors	% Error
1	8560	37	0.43
0	1440	1365	94.79
Overall	10000	1402	14.02

Classification Confusion Matrix		
	Predicted Class	
Actual	1	0
1	8523	37
0	1365	75

Logistic Regression

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	2.22655797	0.21810414	0	*
Age	-0.02170395	0.00234693	0	0.97852987
education-num	-0.12294044	0.01451445	0	0.88431633
Craft-repair	1.40744495	0.12747101	0	4.0855031
Exec-managerial	0.75133455	0.09279788	0	2.11982703
Farming-fishing	1.8333509	0.29099211	0	6.25481081
Handlers-cleaners	1.11341715	0.18750669	0	3.04474497
Machine-op-inspct	2.48784757	0.26867896	0	12.03534222
Other-service	0.74952132	0.1145769	0	2.11598682
Protective-serv	-1.94934726	0.15952255	0	0.14236698
Sales	3.78213954	0.33811864	0	43.90988922
Tech-support	0.46796659	0.15849172	0.00315085	1.59674406
Transport-moving	0.78580767	0.15214679	0.00000024	2.19417834
race_ Asian-Pac-Islander	0.72615516	0.18768629	0.00010929	2.06711745
race_ Other	1.31176102	0.50166774	0.00892796	3.71270609
race_ White	0.7085942	0.09101161	0	2.03113389
capital-gain	0.00002729	0.00000767	0.00037649	1.0000273
hours-per-w eek	0.00970774	0.00276339	0.0004431	1.00975502
country_Other	0.81870979	0.3976253	0.03949441	2.26757216

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	8474	86
0	1278	162

Error Report			
Class	# Cases	# Errors	% Error
1	8560	86	1.00
0	1440	1278	88.75
Overall	10000	1364	13.64