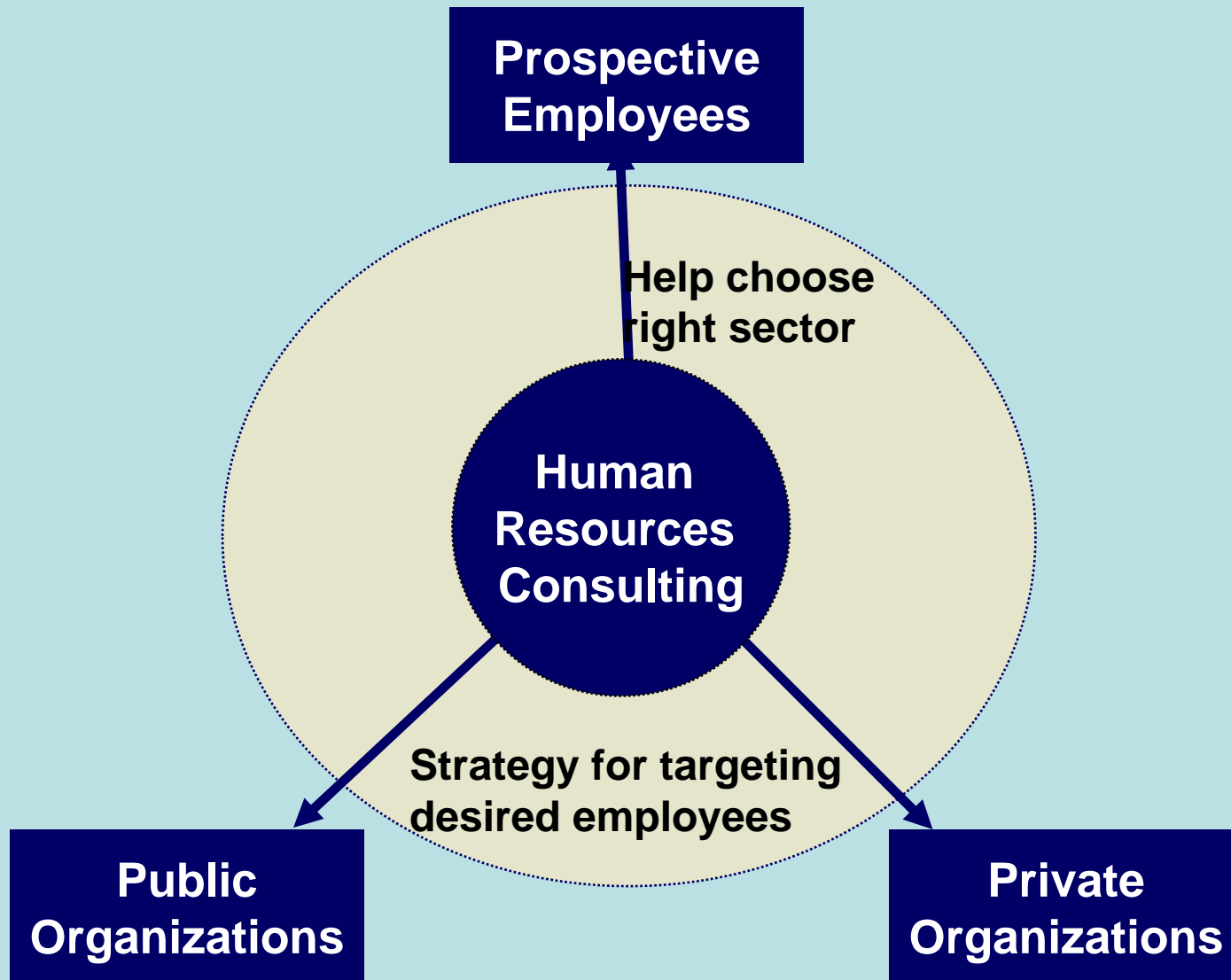


Employee Demographics: Exploring similarities and differences in the private and public sector

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

Problem

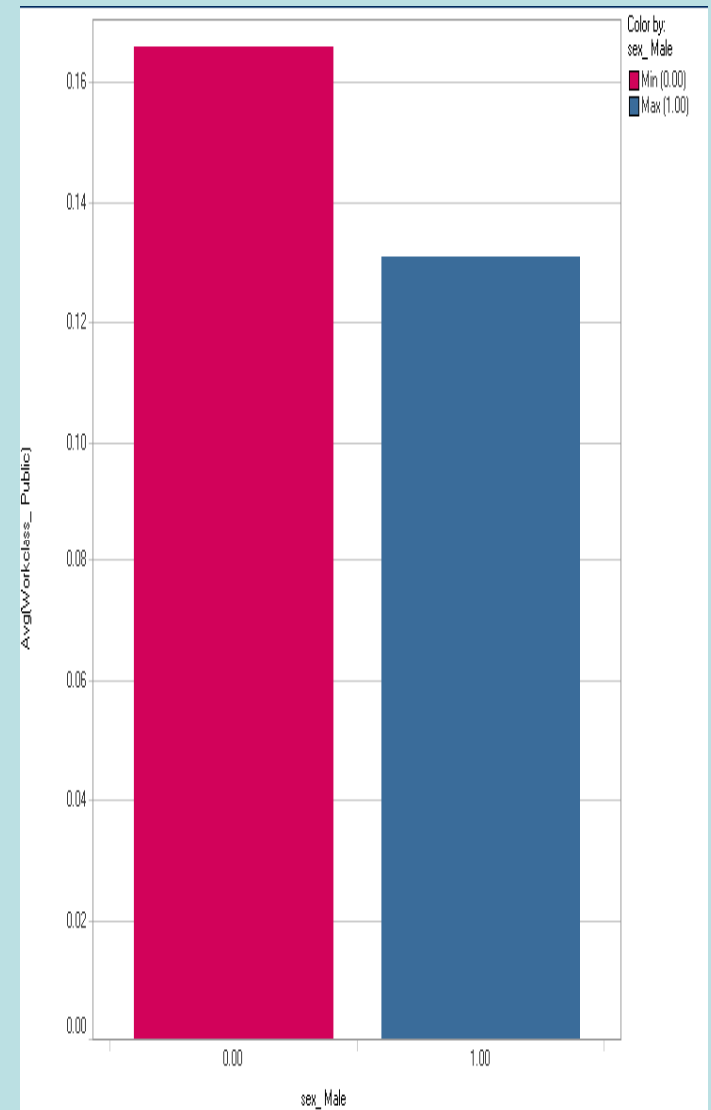
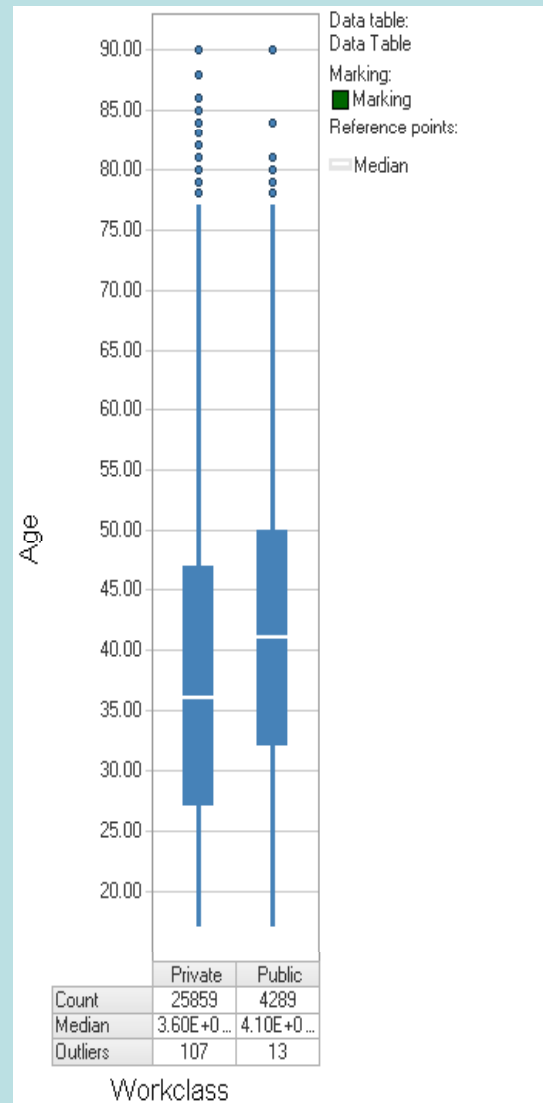
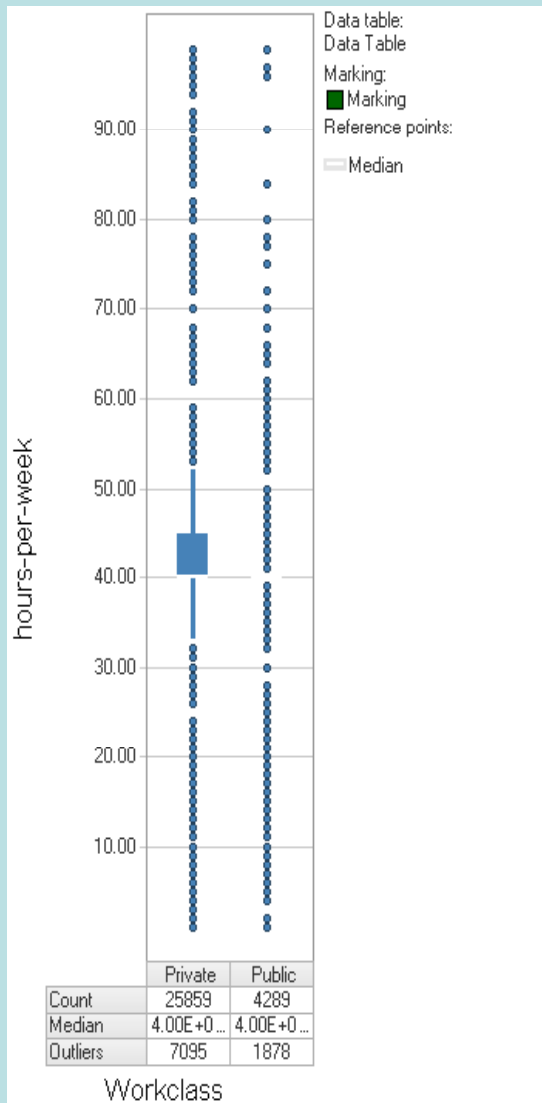


Data Source

- Data Source: UC Irvine Machine Learning Repository (US Census Bureau-1994)
- Y-Var Workclass: 1= Private, 0= Public
- Records: 32,561 → 30,148
 - Missing variables; irrelevant
- Variables: 15 → 13
 - Binning categorical variables; irrelevant

Sex
Marital status
Occupation
Relationship
Race
Native country
Income class
Age
Education number
Hours worked per week
Capital gains
Capital loss

Exploration



Models

The Regression Model

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_Asiac-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-week	0.00970774	0.00276339	0.0004431	1.00975502
country_Other	0.81870979	0.3976253	0.03949441	2.26757216

Residual df	9981
Residual Dev.	6847.646973
% Success in training data	85.6
# Iterations used	9
Multiple R-squared	0.16929489

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5
---	------------

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

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

Conclusion

- Model yielded general implications for our clients depending on their needs
 - Most actionable results for Public sector
 - Older workforce (median age 41)
 - More flexible work hours
 - Higher proportion of highly educated employees
 - Higher proportion of females
 - Private sector- no clear actionable results due to wide variety in demographics
 - Job seeking employees – fit depends on profile and individual needs