Profiling the Different Types of Data Scientists: Which One is Right for You?

Sanaz Bahargam
Boston University
bahargam@bu.edu

Theodoros Lappas
Stevens Institute of Technology
tlappas@stevens.edu

The growing popularity of data-driven decision making, coupled with the increased availability of affordable infrastructure for storing large volumes of information, have established data as a firm's most valuable asset. To take advantage and monetize from this asset, a firm needs talented employees with the skills required to analyze large data repositories, identify opportunities to improve the various aspects of a firm's operations, and inform strategies that can be used to fully take advantage of such opportunities. While such individuals have always existed and held various titles within firms, the era of Big Data and Analytics has brought them together to create a new cast of "data scientists". Famously dubbed as "the sexiest job of the 21st century"[1], this new role has become a hiring priority for firms across industries. In addition, the ever-increasing demand and highly competitive salaries have established data science positions as highly desirable targets for fresh graduates and experienced workers who feel that they have the skills required to transition to data scientist roles. For all interested parties, the key question is the same: ***what is it that makes a good data scientist?***

A simple exploration of the data science openings on a website like Indeed.com or Linkedin.com verifies that firms can have dramatically different definitions of the role. Consider the two following examples:

Ad1[2]. Required skills:

- Experience with designing and implementing machine learning, data mining, statistics, or graph algorithms
- Experience in programming with object-oriented languages, such as Java, C++, C#, or Python
- Experience with development in Hadoop, MapReduce, and HDFS
- Experience in working with NoSQL or column-oriented distributed database

Ad2[3]. Required skills:

- Very strong stats skills & knowledge of HLM, ANOVA, IRT, & Structural Equation Models, Decision Trees, GLM, Clustering, Bayesian methods, SVM, linear/non-linear programming, Multi-level models, Random Forests, Choice Models
- Highly proficient in SPSS, R, Excel, and Access
- Comfortable with documentation of approaches for reference and reproducibility

These two firms are clearly looking for two significantly different individuals. Therefore, it is safe to assume that the two openings will attract applicants with different educational and professional backgrounds. ***Motivated by this diversity and the ever-growing need for talented workers in this domain, our work focuses on identifying and comparing different castes of data scientists, in terms of both their background and behavioral patterns.***

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
[2] http://www.jobs.net/jobs/booz-allen-hamilton/en-us/job/United-States/Data-Scientist/J3K1KT5X7MF9BGVTJ2B/
[3] http://www.indeed.com/cmp/All-In-Analytics/jobs/Big-Data-Scientist-efd15655120dfa43

2. Methodology

To conduct this study, we collected the profiles of 14960 LinkedIn users who have held at least one data science role. At the time of collection (July 2015), this was a complete dataset of all such individuals on the platform. An analysis of the data reveals that data scientists can be grouped into the following 6 castes, according to their undergraduate major: Computer Science (CS), Business, including Information Systems and all other majors offered by business schools (BSN), Humanities (HUM), Engineering (ENG), Math/Physics/Statistics (MPS), and Biology (BIO). ***In this work, we compare the six castes based on (i) their most prevalent skills and (ii) their commitment to the data science role.***

2.1 Different Backgrounds, Different Skills

Given a skill *S* and a pair of castes, we first count the number of worker profiles from each caste that include the skill. We then use Fisher's exact test [Fisher 1925] to determine if *S* is significantly more likely to occur in one of the two castes, at a confidence level α=0.01. The process is repeated for all skills and all possible caste pairs. For each pair (C1, C2) , we report the 5 most frequent skills in C1 that are also significantly less likely to occur in C2 (and vice versa). The results are shown in Table 1.

	Prevalent In	Skills
BIO vs CS	BIO	Data analysis, R, Statistics, Bioinformatics, Matlab
	CS	Machine learning, Java, Python, Data mining, Algorithms
BIO vs BSN	BIO	R, Python, Statistics, Machine learning, Bioinformatics
	BSN	SQL, Analytics, Analysis, Business intelligence, SAS
BIO vs HUM	BIO	Bioinformatics, Matlab, Molecular Biology, Java, computational Biology
	HUM	Analytics, SPSS, Analysis, Quantitative research, Qualitative research
BIO vs ENG	BIO	Data analysis, R, Statistics, Bioinformatics, Molecular Biology
	ENG	Machine learning, Matlab, Data mining, Algorithm, C++
BIO vs MPS	BIO	Bioinformatics, Molecular Biology, Computational Biology, Genome, Genetics
	MPS	Statistics, SQL, Statistical model, Analytics, SAS
CS vs BSN	CS	Machine learning, Java, Python, Data mining, Algorithms
	BSN	Data analysis, SQL, R, Analytics, Statistics
CS vs HUM	CS	Machine learning, Java, Python, Data mining, Algorithms
	HUM	Data analysis, Statistics, R, Analytics, SPSS
CS vs ENG	CS	Machine learning, Java, Python, Algorithms, C++
	ENG	Matlab, R, Data analysis, Statistics, Analytics
CS vs MPS	CS	Machine learning, Java, Python, Data mining, Algorithms
	MPS	Data analysis, R, Statistics, Matlab, Statistical model
BUS vs HUM	BSN	SQL, Java, Business intelligence, Database, Microsoft excel
	HUM	Data analysis, Statistics, SPSS, visualization, Quantitative research
BUS vs ENG	BSN	Data analysis, SQL, Analytics, Analysis, Business intelligence
	ENG	Machine learning, Python, Matlab, Data mining, Algorithms

BUS vs MPS	**BSN**	SQL, Analytics, Java, Analysis, business intelligence
	MPS	Data analysis, R, Statistics, Python, Machine learning
HUM vs ENG	**HUM**	Data analysis, Statistics, SPSS, Analysis, SAS
	ENG	Machine learning, Python, Matlab, Data mining, Algorithms
HUM vs MPS	**HUM**	SPSS, visualization, Javascript, Quantitative research, Qualitative research
	MPS	R, Statistics, Python, Machine learning, Data mining
Eng vs MPS	**ENG**	Machine learning, Matlab, Algorithms, C++, Java
	MPS	Data analysis, R, Statistics, SQL, Statistical models

Table 1: Comparing prevalent features across data science castes.

The table reveals the association of the CS caste with machine learning/data mining and programming languages such as Python, Java. For BSN, skills such as databases, business intelligence, and analytics tend to be more prevalent. Skills like SPSS, quantitative research, and statistics tend to occur more frequently in the HUM caste. ENG demonstrates the highest diversity, displaying skills such as Matlab, machine Learning, data mining, algorithms, and a variety of programming languages. For MPS, stand-out tools include statistics and tools such as Matlab, R, SPSS and SAS. Finally, the Biology caste stands out for its domain-specific skills, such as bioinformatics, molecular biology and computational biology.

2.2 Commitment to Data Science

We hypothesize that the explosive popularity of data science can attract opportunistic candidates with little in the domain. Such individuals are more likely to abandon the role and transition to different jobs. On the other hand, committed workers who consider data science a critical career stage are more likely to persist. In order to evaluate the commitment of individuals from each caste, we perform a survival regression via an Accelerated Failure Time (AFT) model [Wei 1992]. In this setting, the lifetime of each worker begins when she first becomes a data scientist. A death event is observed when and if the worker abandons the role. An AFT regression is used to model the time-to-death with multiple control features from each worker's profile, including (among others) the gender, age, academic rank of undergraduate institution, master and PhD degrees, and the number of different jobs prior to the data science position. The regression is completed independently for each caste, in order to identify features that are correlated with the time-to-death. The workers are then projected on the space of the selected features and clustered via the DBSCAN method [Ester 1996] to identify cohesive subgroups within each caste. Finally, we plot the survival curve for each cluster in Figures 1-6. The x-axis hold the time, while the y-axis holds the survival probability (i.e. the probability of maintaining a data science role).

The figures reveal a number of interesting findings. First, all 6 castes include 2-3 clusters that display significantly different survival curves. A deeper study of these clusters reveals notable differences in their composition. For instance, for the CS and BSN castes, the workers in the two identified clusters differ significantly in terms of their mobility, as captured by the number of jobs that they held prior to the data scientist role. In the HUM caste, the workers with the highest commitment to data science were those that received their Bachelor's from a top-10 university. As we discuss next, our future plans a detailed analysis of the clusters and their differences, both within and across castes.

3. Future work

Further analysis is required to fully understand the different types of data scientists. In addition to studying the clusters that emerged from our commitment analysis, we will conduct a secondary survival study to compare the time that it takes for workers in each caste to enter the data science domain following their graduation. In addition, our study will identify the different types of firms that tend to employ graduates from each caste, thus presenting a further comparison of their career paths.

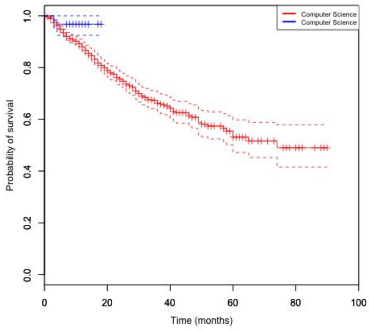

Figure 1: CS Clusters

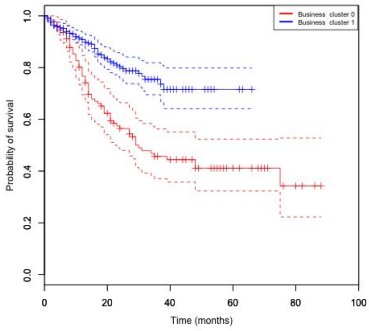

Figure 2: BSN Clusters

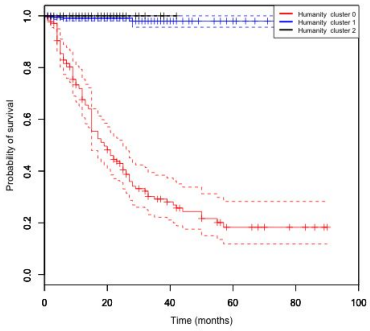

Figure 3: HUM Clusters

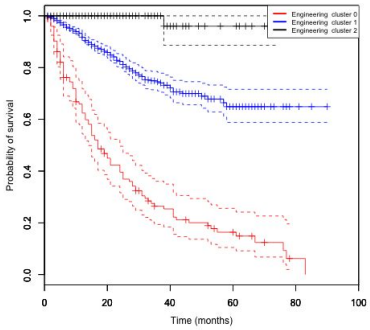

Figure 4: ENG Clusters

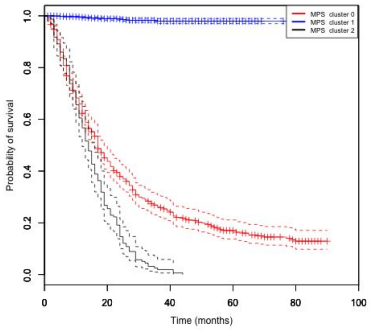

Figure 5: MPS Clusters

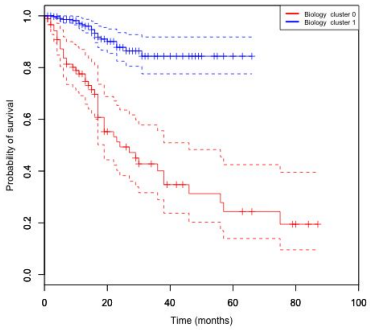

Figure 6: BIO Clusters

REFERENCES

[Fisher 1925] Fisher, Ronald Aylmer. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[Wei 1992] Wei, L. J. "The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis." *Statistics in medicine* 11.14-15 (1992): 1871-1879.

[Ester 1996] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).