HR Analytics:

Job Change Rates of Data Scientists

DATS 6311

Bayesian Computing

Final Project Report

Harish Ram and Mariko McDougall

12-08-2021

Final Project Report

1. Introduction

Problem and Motivation

For many non-government companies promotions are rare, with much greater rates of external hiring than internal promotion. These new hires are often compensated much at a higher rate than existing employees; external hires earn 18 percent to 20 percent more than existing employees promoted to similar positions on average.

These facts combined have led employees to embrace “job-hopping”, where employees frequently move between companies to gain compensation and seniority, rather than remain with a company in hopes of receiving a promotion.

For companies, replacing employees who leave represents a significant cost in time spent seeking new candidates and in productivity loss during training. If companies were instead able to accurately predict potential turnover candidates they may be able to intervene and negotiate an offer to retain them, rather than needing to find a new employee.

In specific, we examined the rate of attrition for Data Scientists, for both their increasing role in many industries, and for personal relevance as an exploration of our future field.

Dataset Overview

To investigate the possibility of turnover prediction, we used the kaggle dataset “HR Analytics: Job Change of Data Scientists”, which documents professional demographics of Data Scientists in a variety of companies, and if they are currently seeking to leave their current position.

2. Results

Methods

Preliminary data exploration showed that while the dataset contained 13 features, and that relatively few of these features were likely to be informative. For example, nearly all candidates in the data had relevant experience, which rendered the binary “relevant experience” column irrelevant. As a result, several features were removed from the dataset. The remaining features to be used for the analysis included: education level, total years of experience, company size, company type, and tenure at their current position.

To reduce the total number of responses in each feature, we binned the possible responses as appropriate for each category. For example, company size was split into three bins, “Small”, “Medium” and “Large” representing companies with <100, 101-1000 and 1000+ employees respectively. Reducing the number of categorical responses will be important later in the processing pipeline, as we will be performing one hot encoding, and unbinned features would create thousands of uninformative columns.

The original data contained many missing values, with over half of all entries containing at least one missing value. To resolve this, we imputed the missing values for each feature using the median response for that feature.

With no missing values, we then performed One Hot Encoding for all categorical features. This process separates out categorical features into binary columns for each possible response in that feature.

To calculate the probability of a given candidate leaving within each category, we then isolated the true values for each of the one hot encoding values, and for each of those responses we imported the target variable (i.e. if they were seeking new employment).

Finally, we upsampled each one hot encoding feature column within an original feature category so that all columns contained the same number of total observations. For example, within the “company size” original category, we created Company_Size_Small, Company_Size_Medium, and Company_Size_Large. All three of these columns were upsampled to have an identical number of total observations to make them compatible for future analysis. These groups of columns were then saved together as a csv.

To analyze each group, we performed MCMC (Markov chain Monte Carlo) analysis within each original category group. MCMC is an extremely useful method of approximating a posterior distribution which allows us to compare the likelihood to a prior that we found through research. The prior, in this case, is the real world statistic about the attrition rate of a specific company type; while, likelihood is what the data is conveying. By comparing these two values, we can look at the probability of an event that happens exclusively in our dataset to the probability of an event that happens in real life. For example, we can compare the probability that a data scientist would leave a startup compared to the actual attrition rate of an average employee in a startup in the United States.

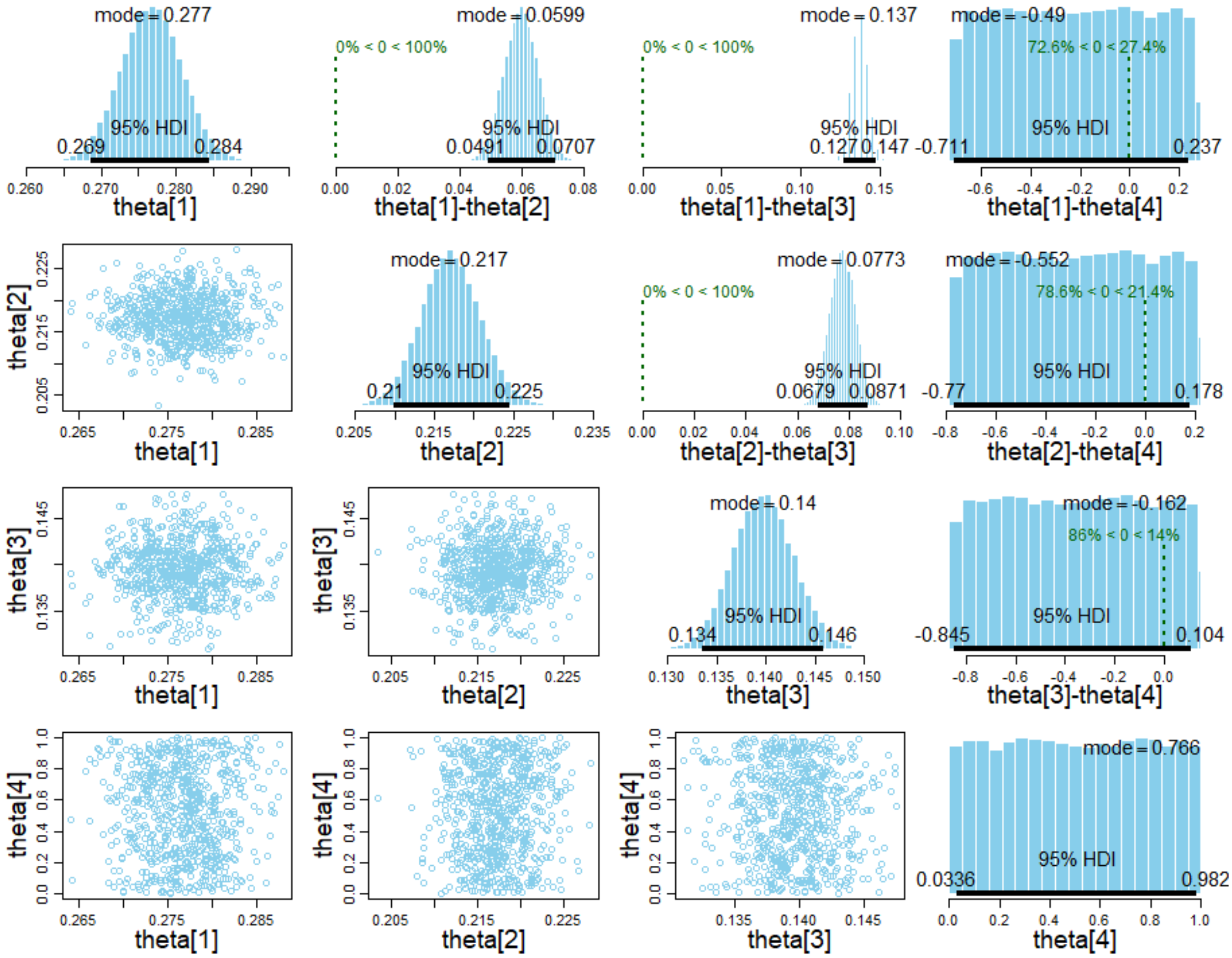
Results

Our analysis revealed many insights into the potential attrition of data scientists across many categories. For features in which the prior was unavailable through research, we focused primarily on the mode of the distribution to tell us meaningful information about the dataset and used prior knowledge of an average business's attrition rate (18%) to put it into context.

Education Level

Education was divided into two sub-categories, highest level of education achieved and total education level. In the highest level of education achieved, only the most advanced degree was considered for each observation. For total education, each degree earned for an observation was considered, such that a candidate with a Masters degree would also be considered to have completed public education and a bachelor's degree.

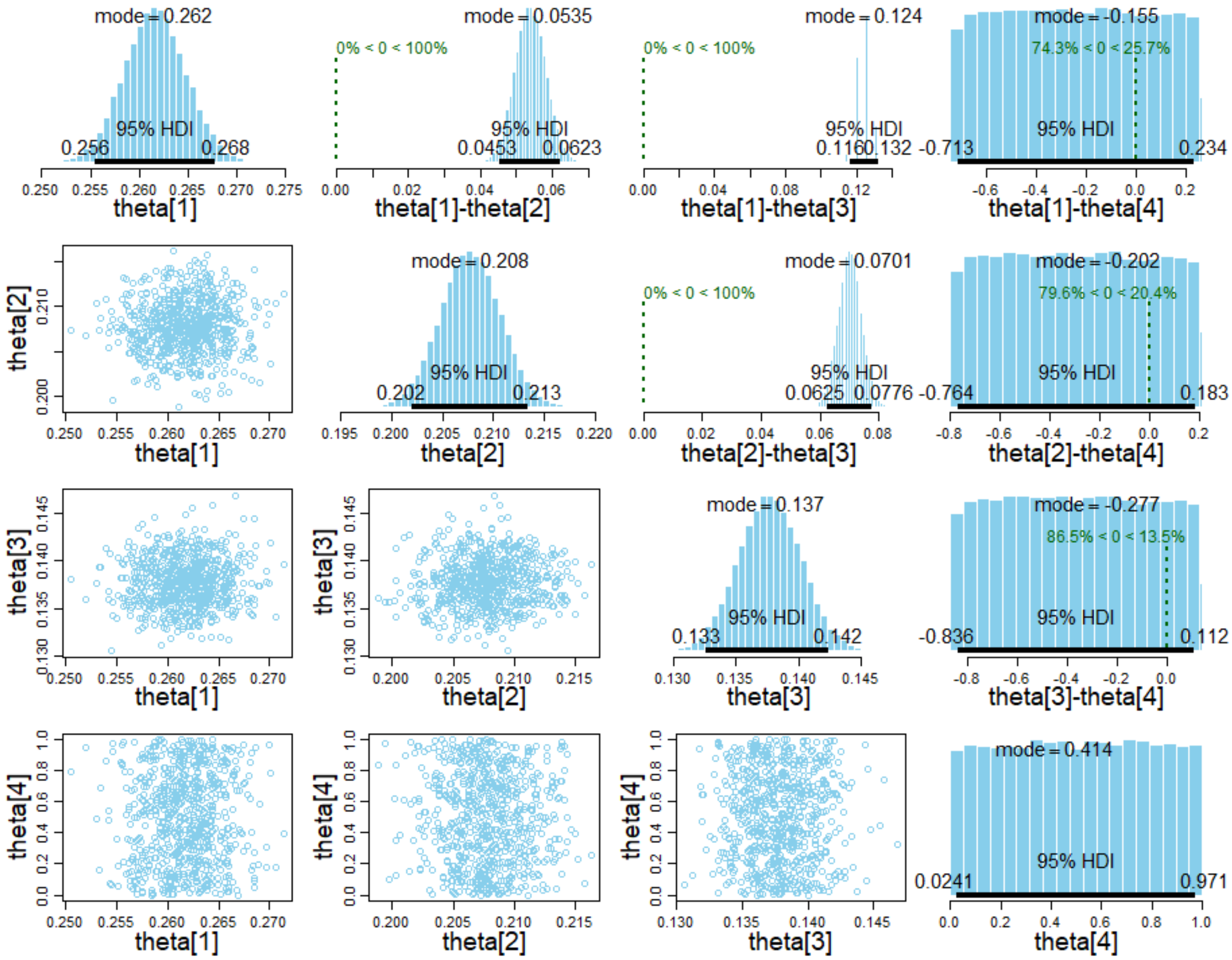
Highest level of Education Achieved



Theta	Highest Education	Mode
Theta 1	Bachelors	27.7%
Theta 2	Masters	21.7%
Theta 3	PhD	14.0%
Theta 4	Public Education	76.6%

When compared to the overall industry standard rate of attrition (18%), we see that most Data Scientists with degrees lower than PhD look for a new position at a greater rate than expected. This is particularly true of those without a post-secondary degree at 76.6%. It is possible that the discrepancy may be due to the different position types afforded to each degree level - it is entirely possible that PhD data scientists are in academia, or are at the peak of their careers, whereas those without a secondary degree are early in their career. This line of reasoning will be explored further in the “Total Years of Experience” section.

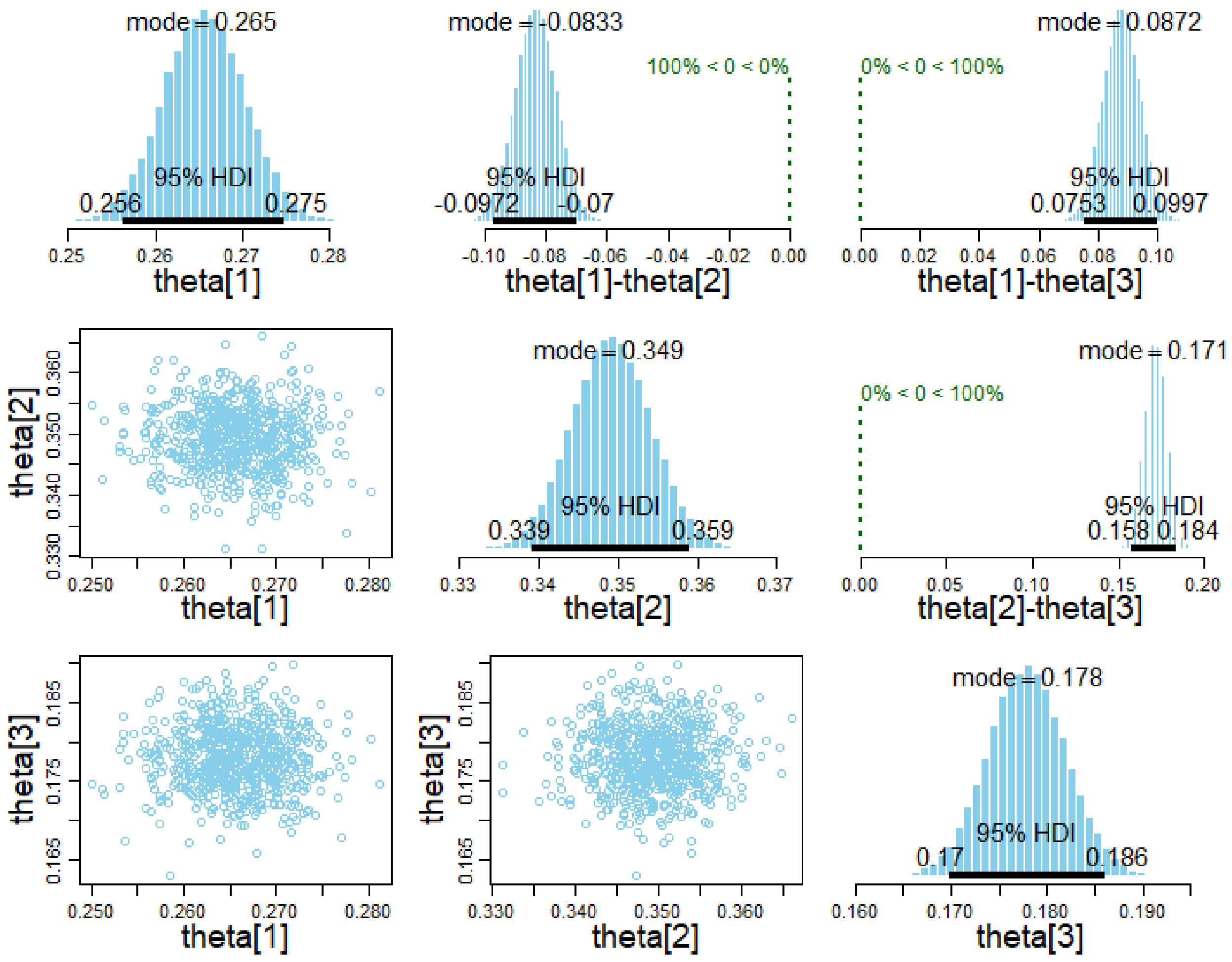
Total Education



Theta	Highest Education	Mode
Theta 1	Bachelors	26.2%
Theta 2	Masters	20.8%
Theta 3	PhD	13.7%
Theta 4	Public Education	41.1%

While conceptually including total education achieved was interesting, the overall effect was a dilution of lower levels of education, without additional information of interest. Both PhD and Masters remain relatively unchanged, but the means of both Bachelors and Public education are shifted towards the means of the higher levels. This is especially true of Public education, which dramatically shifts from 76.6% to 41.1%. The effect is pronounced in this category as each candidate with a degree, which is every candidate that was not originally in the category, is now also represented in the public education category.

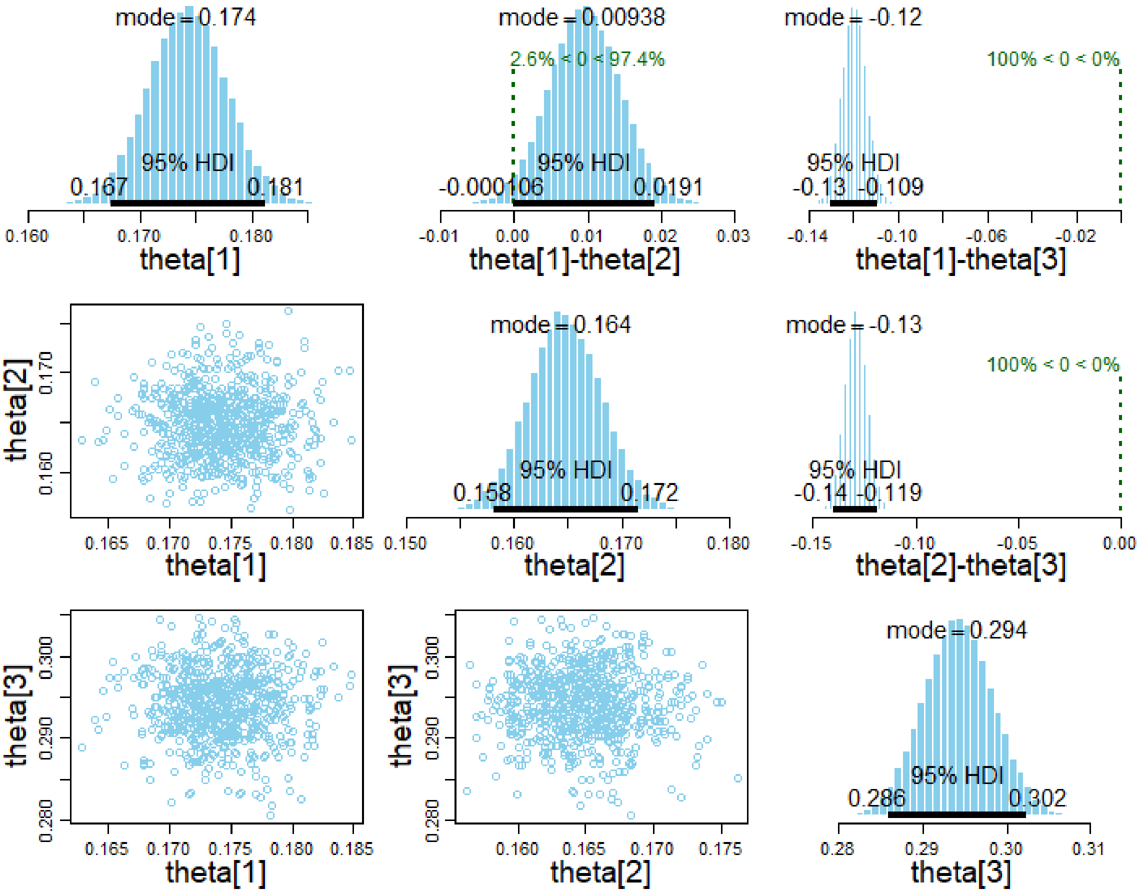
Total Years of Experience



Theta	Years Experience	Mode
Theta 1	5-10	26.5%
Theta 2	<5	34.9%
Theta 3	>10	17.8%

As supposed in the education category, Data Scientists that are earlier in their career are more inclined to seek new employment. As Data Scientists become mature in their careers, with over 10 years of experience, their probability to seek new employment decreases to the average rate for most industries. Prior to that however, it is much higher than average.

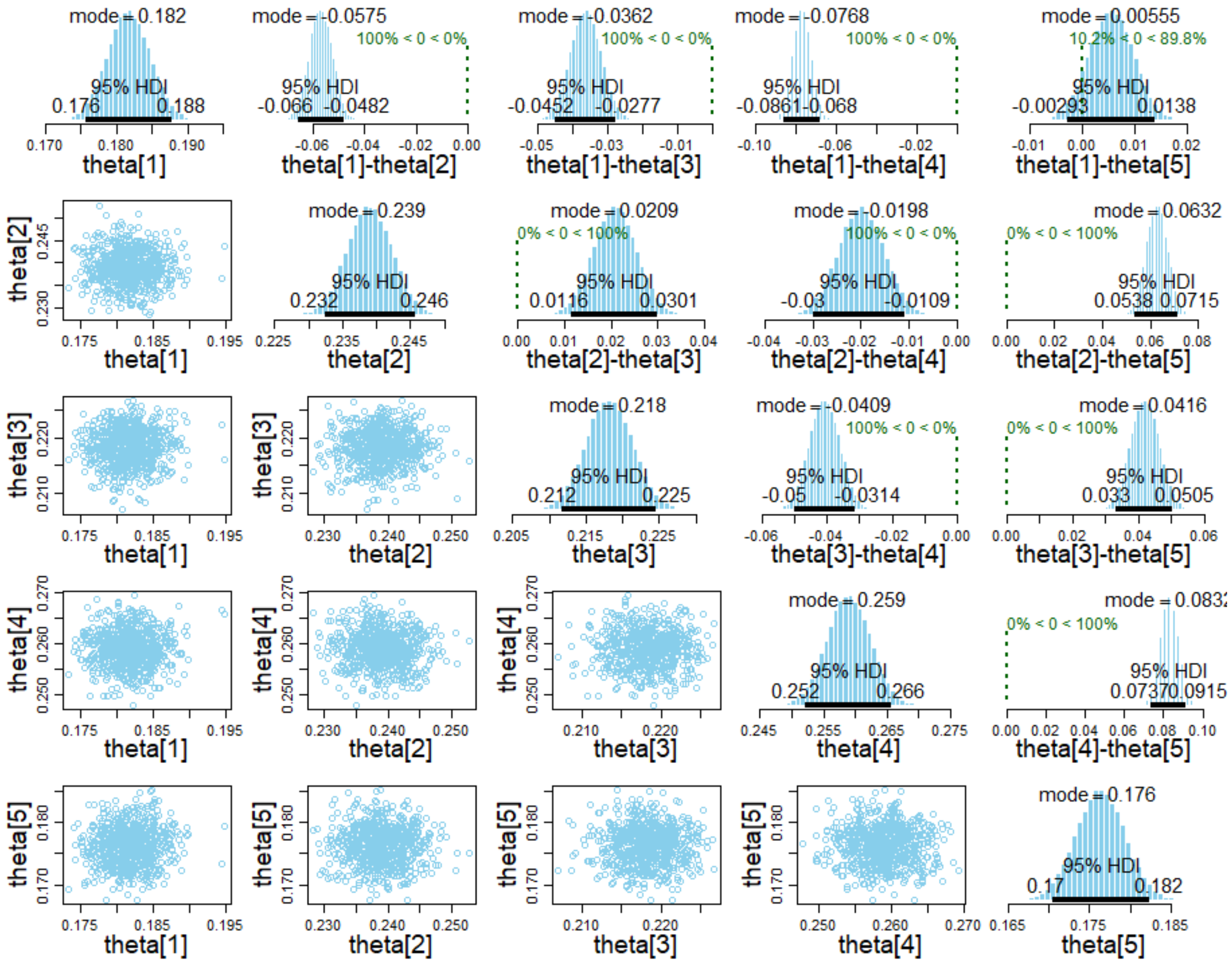
Company Size



Theta	Company Size	Mode
Theta 1	Large	17.4%
Theta 2	Medium	16.4%
Theta 3	Small	29.4%

Company size appears to play a role in the predilection of its employees to seek new employment. Data Scientists that are in smaller companies look for new positions at a greater rate than both data scientists at large or medium companies, and more likely than industry standard in general. This may be due to a perceived or real lack of stability in smaller start-ups or firms, or a lack of possible promotions.

Company Type



Theta	Company Type	Prior	Mode
Theta 1	NGO	19%	18.2%
Theta 2	Other	25%	23.9%
Theta 3	Public Sector	6.2%	21.8%
Theta 4	Private	4.1%	25.9%
Theta 5	Startup	25%	17.6%

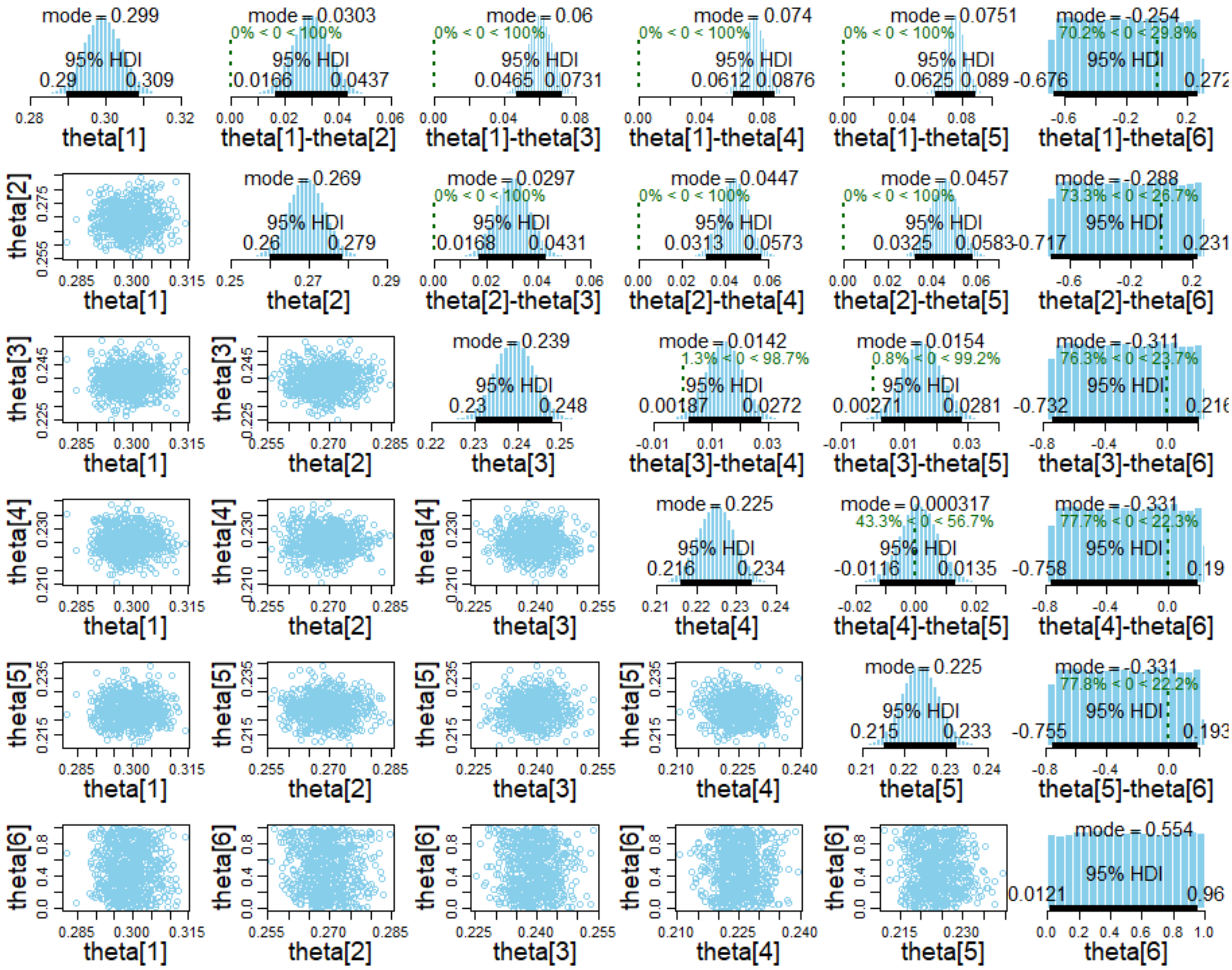
The type of company that one decides to work at can prove one of the most influential factors in whether they decide to leave the company or not. In the table above, we can see that the attrition rate for NGOs or Non-Government Organizations for employees in all fields is 19%. This indicates that in a given year, 19% of the staff will leave the company or be fired. Within that statistic, 18.2% of employees who leave will be data scientists according to our dataset. The "Other" category encompasses types of companies that are not already listed and the prior we researched was purely voluntary turnover which is around 25%. Compared to the attrition rate calculated from our dataset of 23.9%, data scientists tend to leave other types of jobs relatively more often than NGOs.

Next, looking at the Public Sector which references all government and government-related organizations, the attrition rate is extremely low at 6.2% for all fields. The takeaway from this statistic is that government positions are largely stable. Government jobs typically do not go out of business and come with benefits. Within the 6.2% of employees who leave the public sector, 21.8% of them are in the data science field.

The private sector has a similar, but slightly lower turnover rate of about 4.1% while 25.9% are those in data science jobs. Like in the public sector, private companies can also provide benefits and flexibility to the work-life schedule. However, with more competition in the private sector, the attrition rate for data science jobs also goes up. With STEM jobs becoming more and more prevalent with big data, there is more need to find employees with more recent knowledge or find a better job at any other company as most businesses tend to contain a data science department.

Finally, startup companies, being unreliable in profits and stability, have the highest turnover rate of a single company type at 25%. In other words, a ¼ of employees will leave a startup in a given year. Although not as high as the turnover rate in the public and private sector, 17.6% of data scientists leave startups most likely due to its highly volatile nature.

Tenure at Current Position



Theta	Most Recent Job	Mode
Theta 1	0	29.9%
Theta 2	1	26.9%
Theta 3	2	23.9%
Theta 4	3	22.5%
Theta 5	4	22.5%
Theta 6	5+	55.4%

The length of tenure at the current position is a factor in a data scientist’s decision to search for new employment. With a tenure of less than a year, there is a 29.9% chance that they are seeking a new position. This may be due to poor fits in their current position, or misrepresentation of their role in the initial job posting.

There is also a high rate of searching for new positions after the completion of year 1 at 26.9%, which would line up with our suppositions at the start that candidates are likely to job hop roughly every two years.

After years 2-4, the rate is still slightly above the average rate of attrition (~23% vs 18%), but is relatively steady. However, at years 5+, the rate spikes to 55.4%. This is likely due to the large tenure range that is encapsulated by such a large category. At an unknown time in the tenure after 5 years there is likely an inflection point where data scientists become progressively more likely to search for a new position, but we are unable to pinpoint this time with the given dataset.

3. Conclusions

As big data becomes more relevant and widespread, data science jobs in every industry at the moment are becoming more popular. Thus, we decided to focus on what types of companies data scientists are deciding to leave. Our analysis included Markov chain Monte Carlo methods to determine the probability that a data scientist would leave their position. This probability is exclusive to our data; thus, researching priors of that probability that are applicable to the job industry now was necessary for deeper understanding. However, this information was not readily available for many of the features we were looking at; thus, we decided to focus primarily on the context of those features to draw meaningful conclusions.

In our data analysis we have compared the factors of a company (size, number of employees, etc.) to the characteristics of the employee (years of experience, education level, etc.) to find which of these factors are the most influential in an employees decision to leave a company and seek a new job. In terms of the factors that are most influential in the company itself, the data has demonstrated that data scientists have a higher probability of staying within the public and private sector for stability and benefits and tend to leave startups more often because of its volatile nature. The size of the company was a very influential factor as well in that a more populated company was more inviting while smaller companies had a higher attrition rate.

Looking at the characteristics of the type of employee, the highest and total education level they achieved correlates to the attrition rate as well with a higher education indicating long-term commitments to a company. Years of experience is a major factor as well in that those with more than 10 years of experience tend not to look for new jobs while below that equates to probabilities above the average of 18%. Finally, the length of tenure of an employee can also influence an employee’s desire to leave the company. Our dataset shows that with 0 to 1 years of tenure can describe a higher probability, 2-4 years slightly lower and more than 5 years there is a massive increase as employees will want a more senior position at a different company.