Data Mining Methods for Describing Federal Government Career Trajectories and Predicting Employee Separation

Kimberly Healy, Dan Lucas and Cherilyn Miller

Abstract Data mining methods can be applied to human resources datasets to discover insights into how employees manage their careers. We examine two elements of career trajectories in federal government HR data. First, we apply association rule mining and sequential pattern mining to understand the prevalence and direction of interdepartmental transfers. Then we apply logistic regression and decision tree induction to understand and predict employee separation. In this specific application, we find that interdepartmental transfers are uncommon, except between branches of the armed services and out of these branches to the Department of Defence. We also find that demographics, compensation, and political transitions are significant factors for retention, but they account for only a small portion of the probability of a federal employee leaving service. We expect these methods would perform better in industry with a small amount of additional data gathered upon hiring and exit interviews.

1 Introduction

Prior to the emergence of service science as a distinct discipline [1], service and operations research were conducted without consideration of human resource management. However, many problems afflicting a service have human issues as their root causes [2]. Some work has been done since to model this interaction, including policy models and mathematical/statistical models such as the Markov model [3]. This history is summarized well in [4]. These models require significant business understanding to set parameters. For this reason, students of service management may be interested in data mining approaches requiring less initial configuration to help them discover under what circumstances employees will leave a department, either for other departments within the same employer, for employment elsewhere, or for retirement. These insights will allow service managers to consider turnover risks as they develop their service models. A recent release of data by the Office of Per-

K. Healy · D. Lucas (✉) · C. Miller
Engineering Division (Great Valley), Pennsylvania State University, Malvern, PA 19355, USA
e-mail: dj1252@psu.edu

H. Yang and R. Qiu (eds.), *Advances in Service Science*, Springer Proceedings in Business and Economics, https://doi.org/10.1007/978-3-030-04726-9_9

sonnel Management within the United States federal government under the Freedom of Information Act provides an excellent dataset for demonstrating these techniques, including association rule and sequential pattern mining, logistic regression, and decision tree induction.

1.1 Problem Definition

Over 40 years of United States federal government employment data was released to the public for the first time in downloadable format in May 2017. The data was made available through three Freedom of Information Act inquiries by BuzzFeed News. This data can now be retrieved from the Internet Archive at http://www.archive.org [5]. The records span from 1973 to March 2017. They include federal employees' employment details, such as occupation title, salary, and supervisory status, along with employee demographic details such as age, education level, and location.

1.2 Objective of Report

- Objective 1—Describe career trajectories of federal employees
- Objective 2—Predict employee separation over the course of a calendar year using data available in the third quarter of the previous year.

2 Data Understanding

The Federal Employment Dataset comes from the U.S. Office of Personnel Management via the Freedom of Information Act (FOIA). The data files contain four decades of the United States federal payroll spanning the years 1973–2017. The data files within the three dated chunks are partitioned by Department of Defense (DoD) data and Non-Department of Defense (Non-DoD). The data includes quarterly snapshots.

Status files give static data about employees during the quarter of the dated file. Attributes of the Status data files include ID, Employee Name, Date of Filing, Agency and Sub Agency, Station, Age, Education Level, Pay Plan, Pay Grade, Length of Service (LOS), Occupation, Occupation Category, Adjusted Basic Pay, Supervisory Status, Type of Appointment, Work Schedule, and Non-Seasonal or Full-Time Permanent Indicator.

Dynamic files give activity data about employee turnover. The files indicate whether an employee moved into (Accession) or out of (Separation) a position during the quarter of the dated file. Attributes of the Dynamic data files include ID, Employee Name, Agency and Sub Agency, Accession or Separation Indicator, Effective Date (of Accession or Separation), Age, Pay Plan, Pay Grade, Length of Service, Station,

Occupation, Occupation Category, Adjusted Basic Pay, Type of Appointment, and Work Schedule.

There are several limitations to this dataset, including:

- The dataset does not include detailed salary data, including bonuses or additional compensation.
- Thousands of employees' data are withheld by the U.S. Office of Personnel Management, including:
  - Name and duty stations of employees from the Department of Defense agencies, FBI, Secret Service, DEA, IRS, U.S. Mint, Bureau of Alcohol, Tobacco, Firearms, and Explosives, law enforcement officers, nuclear engineers, and some investigators.
  - No data is provided for employees from the White House, Congress, Judicial branch, CIA, NSA, the Department of State's Foreign Service, the Postal Service, Congressional Budget Office, Library of Congress, Panama Canal Commission, and among others.
- The data obtained from the last two FOIA inquires does not include the primary identification attribute, the pseudo-employee ID. This precludes tracking individual careers from September 2014 to March 2017.

3 Analysis I—Career Trajectories

The Career Trajectory analysis used stratified samples of the data to track the movement of employees from 1973–2012. We used sequential pattern mining and association rule mining techniques. We focused on tracking the movement of employees from one department to another. We relied on the employee ID number attribute to track an employee's movement from quarter to quarter.

3.1 Sample Description

The sampling of employees was done in two parts. In the first phase, strata were created based on government agencies. Then, systematic random stratified sampling was used to select 0.3% of government employees from each stratum. In order to remove the sampling bias for employees retained for longer periods, the sampling is done from a list of distinct employees. In the second phase, the stratified sample of randomly selected employees were joined with complete history (from 1973 to 2012) of every selected employee.

Table 1 Association rules for non-department of defense staff

Con	Ant	Support	Support (%)	Rule conf (%)	Lift (%)
HE	SZ	211	1.0715	42.5	4.0901
HS	TD	335	1.7012	38.7	4.4391
HS	EM	90	0.45704	33.7	3.8637
TD	HS	335	1.7012	19.5	4.4391
AG	IN	277	1.4067	16.7	1.1755
HS	DJ	187	0.94962	13.5	1.5521
DJ	HS	187	0.94962	10.9	1.5521
SZ	HE	211	1.0715	10.3	4.0901

3.2 Modeling and Analysis

The first technique we used was association rule mining. This technique can be used to identify frequent patterns among antecedent and consequent observations. Reference [6] provides an introduction to this method. Criteria such as lift, support, and confidence were used to determine the importance and prevalence of each rule. Association rule mining was primarily used in order to determine how likely it was that an employee transferred to a particular agency, given their previous employment at another agency. The support of an association rule is the percentage ratio of the records that contain both the antecedent and consequent to the total number of observations in the data set.

The confidence is the percentage ratio of the number of records that contain both the antecedent and the consequent to the number of records that contain just the Antecedent: $\alpha(A \rightarrow C) = P(C|A) = P(A \cap C)/P(A)$. The lift is the percentage of the records that contain both the Antecedent and the Consequent—to the percentage of records that contain the Consequent and to the percentage of records that contain the Antecedent. The formula for the lift is as follows: $l(A \rightarrow C) = s(A \cup C)/s(A) \cdot s(C)$. The lift tells us how much better the association rule is at predicting the result rather than simply assuming the result on its own.

Table 1 provides a summary of the association rules worth noting from the non-dod data sets and Table 2 provides a summary for the dod data sets. Note that agency codes are provided here for brevity. Meanings of agency codes can be found at the Office of Personnel Management's website [7]. We began with the rules that had the greatest support. Then, we evaluated the confidence of the rules. We were most interested in the rules with the higher confidence percentages. Note that while there are some agency transitions that have a substantial confidence and support, we should also focus on the lift. If the lift score is below 1, this indicates that the antecedent and the consequent appear less often together than expected. We filtered out any rules that had a lift significantly less than 1. If the lift was near 1, then we evaluated the support of the rule. The larger the support, the more actionable the rule.

Table 2 Association rules for department of defense staff

Con	Ant	Support	Support (%)	Rule conf (%)	Lift (%)
AR	AR	4151	42.129	42.1	1
NV	NV	2776	28.174	28.2	1
AF	AF	2434	24.703	24.7	1
DD	DD	2150	21.821	21.8	1
DD	[AF, AR]	74	0.75104	24.2	1.1083
DD	[NV, AR]	70	0.71044	23.3	1.0693

Table 3 Sequential rules for department of defense staff

Pattern	Support
<AR, AR, AR, AR, AR>	726
<AF, AF, AF>	725
<NV, NV, NV, NV, NV>	669
<AR, AR, AR, AR, AR, AR, AR>	639
<AF, AF, AF, AF>	630
<NV, NV, NV, NV, NV, NV>	603
<AR, AR, AR, AR, AR, AR, AR, AR>	564
<NV, NV, NV, NV, NV, NV, NV>	555
<DD, DD, DD>	553
<AF, AF, AF, AF, AF>	549
<NV, NV, NV, NV, NV, NV, NV, NV>	514
<AR, AR, AR, AR, AR, AR, AR, AR, AR>	501

The second technique used in our analysis was sequential pattern mining which focuses on the discovery of rules in sequences. Reference [6] provides an introduction to this method. The sequence of an employee's transition from one agency to the next is very helpful to know for making career trajectory predictions. The rules presented in Tables 1 and 2 were found using a method which did not take time into account; sequential pattern mining considers the sequence of agencies in an employee's career. Once again, we evaluated the results with special attention given to the confidence and the support of the rules. In this part of our analysis, the sequential pattern mining rules were created by utilizing the GSP algorithm in the SPMF software.

First, the data was condensed to a year-to-year basis before importing into the SPMF program which allowed for a summarized view of the patterns in the data. The rules generated did not describe any switching between agencies. Therefore, we were able to conclude that on a year to year basis, employees remained in their current agencies at a rate higher than would be expected by chance. The Army, Navy, and the Veterans Administration were associated with very long rules showing continued employment. These rules can be seen in Tables 3 and 4.

Table 4 Sequential rules for non-department of defense staff

Pattern	Support
<VA>	62
<VA, VA>	45
<VA, VA, VA>	40
<VA, VA, VA, VA>	32
<VA, VA, VA, VA, VA>	27
<AG>	25
<VA, VA, VA, VA, VA, VA>	24
<HS>	23
<IN>	22
<VA, VA, VA, VA, VA, VA, VA>	21
<DJ>	20
<HE>	19
<HS, HS>	19
<IN, IN>	19

Table 5 Sequential rules for inter-departmental transitions among DoD staff

Pattern	Support
<AR, DD>	125
<NV, DD>	87
<AF, DD>	82
<DD, AR>	79
<NV, AR>	73
<AR, NV>	72
<AR, AF>	69
<AF, AR>	57
<DD, AF>	50

Table 6 Sequential rules for inter-departmental transitions among non-DoD staff

Pattern	Support
<DJ, HS>	3
<HE, SZ>	3

Next, we ran another sequential pattern mining analysis, however we only considered the instances when an employee switched agencies. The results in Table 4 show some of the most common transitions and their support. These results provide further evidence that interdepartmental transfers are rare. Most of the rules are associated with transfers between the branches of the armed forces or between one of the branches and the Department of Defense (Tables 5 and 6).

4 Analysis II—Identifying Factors Contributing to Separation and Predicting Separation

We examined data from 1973–2012 to see what factors predict that a non-seasonal full time permanent federal employee will leave employment (separate) within a year. We relied on the ability to use employee Pseudo-IDs to track an employee across multiple quarters. Since the data from 2013 onward did not have Pseudo-IDs, the data was removed from this analysis.

In addition, we sought to understand whether presidential elections influence employees' retention. We added indicators to the data indicating whether that year was an election year, whether control of the White House transitioned from one political party to another, and, if so, which party assumed control.

4.1 Sample Description

We randomly selected employees employed in the third quarter of each year. We then counted the quarters in which each was employed during the next calendar year. A count of 4 indicated that the employee was retained (1). A count of 0–3 indicated that the employee was not retained (0). This is an imprecise method as some agencies were not subject to the Freedom of Information Act request, so transfers into these agencies would be indicated as a non-retention outcome. We then removed from the sample any employees exhibiting rare values (fewer than 30 observations) for the variables agency, appointment type, or pay. We adjusted for inflation by assuming a constant rate of 3% annually since 1980.

4.2 Modelling and Analysis

We trained a logistic regression model to determine which attributes have a statistically significant impact on retention when considered together. This model outputs a probability that a record belongs to the target class. Reference [8] provides an introduction to the algorithm and a tutorial for training a model using R. In order to improve our understanding of the minority class, we undersampled from the majority class at a rate of 15%. We found the variables listed in Table 7 to be statistically significant predictors at the given coefficients.

This model has a McFadden R^2 of only 0.081, suggesting that most factors contributing to separation are not represented in this dataset. We do find a statistically significant result indicating that employees are less likely to be retained in the year following the transition from Democratic control of the White House to Republican control, all other factors held constant. It is important to note that there is also a statistically significant result indicating that transitions in party-control are correlated with

Table 7 Coefficients of logistic regression model

Coefficients	Estimate	Std. error	z value	Pr(>|z|)
(Intercept)	−1.94E+01	2.50E+00	−7.765	8.19E−15
Grade	1.24E−02	4.96E−03	2.493	0.012656
appt_type15	−2.80E−01	4.46E−02	−6.27	3.60E−10
appt_type30	−4.05E−01	9.81E−02	−4.127	3.68E−05
appt_type32	−6.41E−01	1.27E−01	−5.042	4.60E−07
appt_type38	−6.63E−01	4.30E−02	−15.43	<2e−16
appt_type40	−5.63E−01	2.92E−01	−1.927	0.054039
appt_type50	−3.87E−01	2.47E−01	−1.569	0.116739
appt_type55	−3.55E+00	1.05E+00	−3.37	0.000751
Year	8.22E−03	1.27E−03	6.47	9.78E−11
Election year	−1.64E−01	3.67E−02	−4.474	7.68E−06
Transition of white house to other party	2.66E−01	5.69E−02	4.677	2.92E−06
Transition of white house to republican control	−2.05E−01	7.15E−02	−2.87	0.004111
Education level	9.44E−02	2.64E−02	3.582	3.41E−04
Pay	3.54E−05	4.97E−06	7.12	1.08E−12
Age	5.84E−02	5.96E−03	9.806	<2e−16
Length of service	1.94E−01	1.68E−02	11.576	<2e−16
Age:Length of service	−3.96E−03	3.17E−04	−12.49	<2e−16
Pay:Age	−4.87E−07	1.00E−07	−4.862	1.16E−06
Pay:Length of service	7.59E−08	2.07E−07	0.366	0.714046
Education level:Age	−2.03E−03	5.81E−04	−3.489	0.000485
Education level:Length of service	1.69E−03	4.09E−04	4.128	3.66E−05
Education level:Pay	−1.79E−06	3.60E−07	−4.957	7.15E−07
Pay:Age:los.numeric	−7.92E−09	3.89E−09	−2.038	0.041536
Education level:Pay:Age	3.44E−08	7.10E−09	4.845	1.26E−06

an increase in the likelihood that an employee is retained. This transition coefficient is greater in magnitude than the Republican coefficient. As such, the reasonable interpretation would be that transition to a Democratic White House increases employee retention more than transition to a Republican White House. Neither has a negative impact on employee retention.

Education level appears in several interaction terms in the logistic regression. Some of these terms have negative coefficients while others have positive coefficients. Given the expected magnitude of the Education variable (approximately 10^1), the magnitude of its coefficient (approximately 10^{-1}), the magnitude of the Education:Length of Service interaction variable (approximately 10^2) and its coefficient (10^{-3}), a unit increase in the education variable would be expected to increase the log likelihood of retention by a factor of 1.1 before considering the negative coefficients on interaction effects including education. The expected magnitude of Education level:Age is 10^2 with a coefficient of 10^{-3} and the magnitude of Education level:Pay is 10^6 with a coefficient of 10^{-6}. In the case of a unit increase in education with all other factors held constant, these factors would decrease the log likelihood of retention by a factor of approximately 1.1. This means the expected impact of a unit increase in education on retention is approximately zero. We are left with a weak conclusion that the effect of education on retention is idiosyncratic based on the employee's age, length of service, and pay.

We also developed a decision tree model using the C5.0 algorithm in order to develop rules that might help understand these interaction effects. This model is described in [9]. The decision tree developed for all employees is provided in Fig. 2. As with the logistic regression, this model was trained with an undersampled set to improve our understanding of the minority case. It has an overall accuracy on unseen data of 65.8%, and its accuracy given that the actual class is 0 is 55.5%. Its accuracy given the actual class is 1 is 75.0%. This gives an average by class of 65.2%. This compares favorably to the null model in which we always predict that an employee will be retained, yielding an overall accuracy of 90.0% but an average by class of 50.0%. The pruned tree does not reference education level. It describes a complex interaction between age, length of service, and pay. As age and length of service increase, retention generally drops. For employees with non-permanent appointments paid less than $43,000 per year who are between the ages of 40–54, however, retention increases with length of service greater than 9.5 years. As pay increases, retention generally increases. However, for employees with permanent appointments paid more than $43,000 per year between the ages of 50 and 54, retention drops significantly above 30 years of service. The retention of employees under age 54 with non-permanent appointments is not affected by length of service. This suggests that employees with a permanent appointment tend to work until they can earn their pension, unless they began their service later in life. A plot of variable importance is shown in Fig. 1.

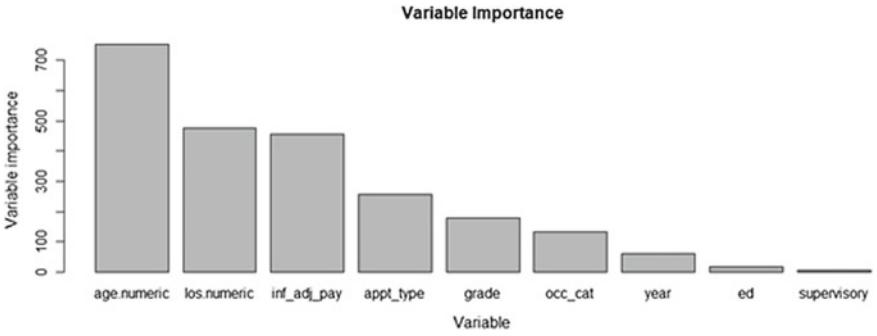

Fig. 1 Variable importance for decision tree

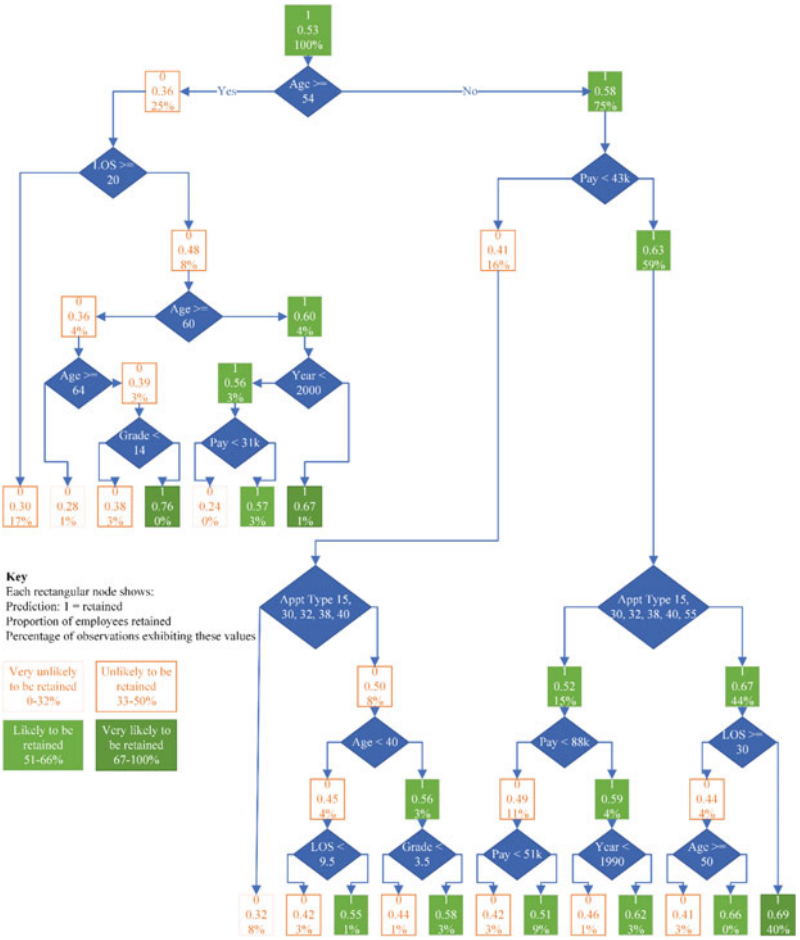

Fig. 2 Decision tree predicting separation

5 Summary

Through pattern mining, we discovered that interdepartmental transfers are rare in the federal government. They tend to be more common in the Department of Defense. Outside of the Department of Defense, we find only two significant sequential rules: employees from the Department of Justice are more likely than most to transition to the Department of Homeland Security, and employees from the Department of Health and Human Services are more likely to transition to the Social Security Administration.

Through logistic regression and a C5.0 decision tree, we determined that the most important factors for predicting departure of employees of the federal government are length of service and age. Among employees who are started their careers later in life, pay becomes an important factor.

These methods could be applied within a company using a richer dataset. For the pattern mining approach, significant insights could be added by including data about the previous employer and the next employer for departing staff. There are some attributes which could enrich the logistic regression and decision tree induction, including previous employer(s), disciplinary actions, employee performance review scores, and supervisor performance review scores.

Acknowledgements This paper is a summary of the results of our winning submission to Penn State's university-wide Data Analytics Challenge, which was chaired by Dr. Robin Qiu with the support of the following committee members from Penn State's Smeal College of Business, College of Engineering, College of Information Sciences and Technology, and the Great Valley Engineering Division. We are grateful for their support. Thank you, Jason Acimovic, Saurabh Bansal, Adrian Barb, Guoray Cai, Terry Harrison, Ashkan Negahban, Robin Qiu, Kathleen Riley, Chris Solo, Satish Srinivasan, Hui Yang, and Tao Yao.

References

1. Moussa S, Touzani M. A literature review of service research since 1993. J Serv Sci. 2010;2(2):173–212.
2. Boudreau J. On the interface between operations and human resources management. Manuf Oper Manag. 2003;5(3):179–202.
3. Lagard M, Cairns J. Modelling human resources policies with Markov models: an illustration with the South African nursing labour market. Health Care Manag Sci. 2012;15(3):270–82.
4. Hafeez K, Aburawi I. Planning human resource requirements to meet target customer service levels. Int J Qual Serv Sci. 2013;5(2):230–52.
5. Internet Archive. Federal employment data from the offices of personnel management. https://archive.org/details/opm-federal-employment-data. Accessed 20 Feb 2018.
6. Penn State. SWENG 545: Data Mining—7.2 Discovering Frequent Sequential Patterns on a Computer, Online Course, Accessed May 2018.

7. Office of Personnel Management. Federal Agencies List. https://www.opm.gov/about-us/open-government/Data/Apps/Agencies/. Accessed 20 Feb 2018.
8. Forte R. Logistic regression. In: Mastering predictive analytics with R. Packt Publishing, Birmingham;2015. p. 93–109.
9. Forte R. Tree-based methods. In: Mastering predictive analytics with R. Packt Publishing, Birmingham;2015. p. 201–8.