Employee Attrition: What makes an employee leave?

Christopher Boomhower1, Stacey Fabricant 2, Alex Frye1, David Mumford2, Michael Smith1, Lindsay Vitovsky1

1 Southern Methodist University, Dallas, TX, US

2 Penn Mutual Life Insurance Co, Horsham PA

**Abstract.**  In this paper, we present a model for employee attrition prediction and discuss the ethical impacts of using such a model within private and public sector organizations. As it is in Human Resource personnel’s best interest to improve retention, implementing statistical and machine learning techniques is the most viable means to attrition abatement. To this end, we examine Office of Personnel Management public sector, Bureau of Labor Statistics public sector, and IBM anonymized private sector employee separation data. Three classification models (Methodologies include Logistic Regression, Random Forest, and K Nearest Neighbor) are trained and tested on these data before selecting our best fit model for attrition prediction. We finally use metrics such as Gini and Permutation Importance to identify the most impactful variables in determining prediction outcome before presenting the ethical ramifications of using such outputs in HR planning. [WILL ADD SENTENCE FOR MAIN RESULT AND SENTENCE FOR MAIN CONCLUSION ONCE THESE DEVELOPMENTS ARE COMPLETE].

1 Introduction

How much does it really cost to lose an employee? Studies such as the Center’s for American Progress analysis (November, 2012) indicate a separated employee may cost anywhere between 16 percent to 213 percent depending on the position [1]. Precisely quantifying this may seem out of reach depending on the complexity of a particular role, but areas of impact that one may foresee at many organizations are: 1) determining if the employee’s vacancy should be replaced or duties handed off to others; 2) posting the job opportunity to various outlets; 3) interviewing, hiring, and training a replacement; 4) enduring lowered employee morale / possible lower productivity from remaining employees; and 5) tolerating a lower skill set from an underdeveloped replacement [2].

Corporations are keenly aware of the downsides to losing employees and exert great effort to maintain retention levels. In their efforts to not only attract talented workers but retain them as well, businesses provide substantial benefits [3]. With industry competitiveness the norm, many employers may still face retention challenges as their employees have alternate employment options. To become even more proactive in attrition prevention, companies must gain a solid understanding for the reasons their employees separate. Foresight into attrition development and contributing factors empowers Human Resource departments to improve retention efforts through improved planning and intervention. While such insights are available to organizations that store employee data, these understandings are not within reach without sufficient analysis.

The first step in gaining foresight into employee attrition is obtaining pertinent data. Companies are understandably reluctant to release the methods, proprietary or purchased, that use even anonymous data to help them in their management of human resources. Various articles allude to this challenge [4, 5, 6]. However, we identified three valid sources of Human Resources data in the forms of Office of Personnel Management data, Bureau of Labor Statistics data, and the “IBM HR Analytics Employee Attrition” dataset. All three forms were analyzed in unison to complement one another in insight and model validity.

During analysis, dimensionality reduction was performed on the datasets. This was essential to reduce numerous correlating and covariant relationships present between dataset variables. Only after these relationships were addressed and the datasets simplified were the data prepared for modelling.

[ADD PARAGRAPH ABOUT MODELLING PHASE AS PROJECT DEVELOPMENT PROGRESSES]

[ADD PARAGRAPH SUMMARIZING MAIN RESULTS AS PROJECT DEVELOPMENT PROGRESSES]

[ADD PARAGRAPH SUMMARIZING MAIN CONCLUSIONS AS PROJECT DEVELOPMENT PROGRESSES]

[ADD PAPER OVERVIEW PARAGRAPH ONCE REMAINING PAPER CONTENT FILLS IN]

2 Attrition as Seen in Civil Service Workers

The U.S. Office of Personnel Management (OPM) serves as the central Human Resources department for all Federal agencies, including the management of federal agency health insurance and retirement benefits. Their oversight of policy implementation as well as being a general resource for all agency Human Resource departments makes their employment data of particular interest for this paper. OPM regularly captures a wide range of data on the millions of federal civil service workers, and we used several of their data sources that focus on attrition. These data [7], include variables such as job level, locality, salary, length of service, and basic reason for termination of employment. Important for understanding our findings in the correct context are the following points:

* **October 2014 – September 2015:** The date range used to filter the final dataset (one full calendar year).
* **General Schedule Pay Scale (GS)**: This is the centralized pay scale used by many civil service agencies which covers the majority of white collar professionals. Whether or not a person is paid under the GS, the OPM converts the level he or she is paid under to the GS for data collection purposes. The scale numbers 1-15 and there are ten “steps” within each of those levels.
* **Locality adjustments:** To remain competitive with industry salaries, the GS operates a locality adjustment scale that adds a particular percentage to a person’s salary, based on city of occupation alone.
* **Federal Employees Retirement System (FERS):** Beginning on January 1, 1987, new civil employees are paid under FERS. This mix of Social Security, a Basic Benefit Plan, and a Thrift Savings Plan helps make civil service positions stand out from many current private sector jobs. We account for this factor in our research, taking into consideration the confounding effect this has on length of service (LOS). At the time of this paper, the pension is based on salary and length of service.
* **Reasons for separation:** OPM captures termination information as "Transfer Out – Individual," "Transfer Out – Mass," "Quit," "Retirement – Voluntary,” and "Retirement – Early Out."

3 OPM Data Consolidation, Sampling, & New Attributes

Worth mentioning are a few measures taken to consolidate, sample, and enhance the OPM dataset from its original form.

* **Non-US jobs** – we are not interested in jobs outside of the US given the wide fluctuations in standard of living costs among other countries.
* **Missing Occupation Type** – all records without and OCCTYP was dropped since one of our main points of interest was the connection of someone’s industry to their length of service.
* **Missing Specified Salary** – we assume that salary is of particular importance in attrition, so we remove the instances with missing salaries.
* **Missing Length of Service** – in order to create a model with a response variable of LOS, we removed those records that did not have this recorded.
* **Missing Length of Service Level** -
* **Younger individuals / Missing Age Level** – given the turnover of young workers and that they are not of interest to this paper, we removed all records for employees under age 20. We also removed those who did not have their age level recorded.
* **< GSGRD of 7** – The fifteen payscale levels of the General Schedule include ten steps within each level. This equates to 150 possible job levels. In order to keep our research to mid-level professions and higher, which are more expensive to train and replace, we only analyzed job levels that began at a level 7.

It is worth noting, that by dropping records under a level 7, all blue collar jobs were, thus, deleted.

Firstly, observations for locations outside the United States were removed as our intent is to model domestic jobs only. In addition, records with no specified occupation, no specified salary, no specified length of service level, and no specified age level were removed from the dataset. Age level *A* employees were also removed as this category includes employees age 20 to as young as age 14, and, therefore, may not reflect relevant work schedules, share common benefits, or separate for valid reasons as compared to the remaining age level categories. These actions reduced the dataset from 8,423,336 to 8,223,193 observations. In addition to observation removal, the second and third measures include new attribute derivation and proper sample design as described in sections 3.1 through 3.3 below.  
  
**3.1 OPM Computed Attributes**

Within the original OPM data, six new attributes were created through aggregation or calculation amongst various attributes: 1) SEP Count by Date & Occupation – Total number of separations (of any type) for a given Date and Occupation; 2) SEP Count by Date & Location – Total number of separations (of any type) for a given Date and Location; 3) Industry Average Salary – Average salary amongst non-separated employees, grouped by quarter, occupation, pay grade, and work schedule; 4) Lower Limit Age – Youngest age within each age level category; 5) Years to Retirement – Based on FERS retirement eligibility baseline of 57 years of age [8]; and 6) Salary Over/Under Industry Average – Difference between computed average salary of non-separated employees and actual salary for each observation. Note also another 1,293 observations were removed after calculating industry average salary as they had no matching non-separation observations (matched on quarter, occupation type, pay plan /grade, and work schedule), which were utilized to ensure realistic salary averages.  
  
**3.2. Bureau of Labor Statistics Derived Attributes**

In addition to the OPM data, we merged 10 attributes from the Bureau of Labor Statistics (BLS). Data were sourced from Federal Government industry codes across all regions. Although assumed to be highly correlated, we sourced both Level (Total number) and Rate (Percentage of Level to total employment and / or job openings) for the following statistics: 1) Job Openings, 2) Layoffs, 3) Quits, 4) Total Separations, and 5) Other Separations. While Rate paints an aggregated, holistic picture for job market trends, Level provides a raw count for total separations alone. Both these statistics were captured by a monthly aggregate and merged to the OPM data by their respective months.  
  
**3.3 Sample Design**

Of the 8,221,900 observations present in the reduced OPM dataset, only 214,282 contained actual separation data and all other observations were considered non-separation. This state of data was inoperable for analysis; therefore, a sample design was determined to mitigate sample size constraints and high variance amongst frequencies of separation types.

Data were divided into groups based on separation type, allowing a maximum of 50,000 observations per type to persist forward during analysis. In so doing, the following retirement separation types were combined: 1) SD Retirement – Voluntary, 2) SE Retirement – Early Out, 3) SF Retirement – Disability, and 4) SG Retirement – Other. Next, the following separation types were dropped completely: 1) SB Transfer Out – Mass Transfer, 2) SK Death, 3) SL Other Separation, and 4) SJ Termination (Expired Appt/Other). Within each separation group (including non-separation), proportional allocation was performed on a combination of date and age level strata to ensure a sample demographic which, as closely as possible, represents that of the original strata-level populations. After sampling was complete, we were left with 229,826 observations.  
  
**4 Data Visualization**

5 Modeling and Evaluation   
5.1 Dimensionality Reduction  
  
5.2 Classification Model Training  
  
5.3 Classification Model Comparison  
  
6 External Validation

7 Ethical Considerations

Perhaps at the crux of any activity that collects and interprets data on human is where the value lies, with the gatherer or the subject. Would our model prompt a company to strive for increased employee tenure, or would it reduce compensation and other benefits once it is aware of an employee’s “shelf life”? We touch upon several ethical dilemmas below.

**Size of the Interested Party**

One of the benefits of using third party data is that a smaller organization does not risk breaking anonymity. For example, maintaining anonymity in a smaller size company, with only one employee in accounting, would be quite outside the realm of possibility should the organization gather such data on their own employees. However, by using data from such a large entity as the US Government, a Human Resources professional would be able to gain insight about the departments in their company without risking important employee relationships.

That being said, the data we analyzed is applicable to employee groups as a whole, not individually. There are many factors that could impact an employee choosing to leave, and when big data is applied to small departments or single individuals, an HR professional might incorrectly assume that one of their employees will act similarly to the behaviors of general groups. In reality, an employee’s family life, manager, financial situation, and perception of self-worth may drastically affect the outcome of applying large-scale behavioral models to small populations.

**Data Collector**

There is no denying that, in the case where employee disclosure is required as part of the data gathering process (as in the case of the performance data from the IBM data set), an employee may not feel comfortable being completely honest about certain attributes. For example, if an employee is to rate their job satisfaction, would said person be concerned that their results would be seen by a manager and used against them? Would the realization that he or she is not satisfied prompt them to then begin a job search?

We contend that surveying employees (rather than analyzing known facts that are collected by default) are more prone to ethical issues. Certainly, the collection of employee data, such as length of service, years to retirement, etc. are features that are not very attributable to individual respondents. But, data collection that reflects personal feelings (i.e. job satisfaction) or other types of individual metrics (i.e. performance reviews) risk being attributable to specific employees.

For these reasons, if a company wishes to gain insight into the attrition rate of their own employees, it would be advisable to either 1) form an internal team or committee to monitor and advise on the collection of employee data and 2) perhaps bring in a third party researcher that will commit to maintaining anonymity and autonomy.

**Improving the Life of an Employee**

How do results from employee data collection and analysis improve the life of an employee? Does a company only care about saving money in their management and development of their ranks? Is the concern to actually improve the experience of an employee?

After the collection and analysis of such employee data, how our results of our research, or anyone else’s for that matter, actually improve the life of an employee? Does a company only care about saving money in their management and development of their ranks? Is a concern to actually improve the experience of an employee?

The reasons for staying at a company vary from person to person, of course. A company that is wanting to simply reduce attrition cost, and nothing else, will most likely not receive the full benefits of understanding attrition. Ethically, making inferences of what prompts an employee to leave should be balanced on how to improve the life of employees.

For example, of gender is collected as part of a survey, perhaps an employee attrition model finds that women have a higher attrition rate than their male peers. If we build our model around that feature, we then are at risk of making the statement that gender is just as important as, say, salary when determining attrition rate, even though gender may have societal associations that would result in attrition, but would be considered morally reprehensible if those associations were taken into consideration for hiring practices.

Also, consider the ever elusive “work/home life balance.” If personal events are found to affect employee attrition, then the question becomes, is that something you, a person in Human Resources or Management, should involve yourself wit when considering hiring and attrition? If you catch wind that an employee is going through a difficult time at home, is that a red flag for the company? There's a possibility that such data can impact the model, and while there is likely quite a bit of improvement that may come from collecting such data, the risk of bias that could come of it is ever present.

**Systematic Propagation of Unfair Practices**

A glaring ethical challenge from employing an attrition model is the chance that once trends are seen, a company actually instills processes that create systematic disadvantages for its employees. For example, say a company learns that 30-45 year olds working in marketing average 3 years at their employer and are often underpaid in relation to their industry when they quit. In response, the company updates its profit sharing plan to only vest for its employees once they have reached 5 years of service. Rather than keep more employees around, the company inadvertently created more discontent, profiting by keeping unvested profit sharing contributions and continuing to pay their workers less than the industry average.

Just because there is a risk, doesn't mean you should act on that risk. This falls into the Worst-Case Fallacy and the Zero Tolerance Fallacy - both of which many business practices are already susceptible of being considered.

7 Conclusion

# References

1. H. Boushey and S J. Glenn. "There are Significant Costs to Replacing Employees." Center for American Progress. https://www.americanprogress.org/wp-content/uploads/2012/11/CostofTurnover.pdf. November 2012,
2. J. Kantor. “High Turnover Costs More Than You Think.” Huffington Post. http://www.huffingtonpost.com/julie-kantor/high-turnover-costs-way-more-than-you-think\_b\_9197238.html. February 2016.
3. Glassdoor Team. “Top 20 Employee Benefits and Perks.” Glassdoor. https://www.glassdoor.com/blog/top-20-employee-benefits-perks/. February 2016.
4. J. Walker. “Do New Job Tests Foster Bias?” Wall Street Journal. https://www.wsj.com/articles/SB10000872396390443890304578006283936708970. September 2012.
5. Abellin. “Woman Sues Over Personality Test Job Rejection.” ABC News. http://abcnews.go.com/Business/personality-tests-workplace-bogus/story?id=17349051. October 2012.
6. D. Richter. “Why Does HT Need to Collect Data? An Employer’s Guide.” CascadeGo HRMoz Blog. http://www2.octopus-hr.co.uk/hrmoz/article/why-does-hr-need-to-collect-data.aspx September 2014.
7. Office of Personnel Management. "Separations Data" from "Data, Analysis and Documentation." OPM. https://www.opm.gov/data/. 2106.
8. Bureau of Labor Statistics. "Help & Tutorials: Data Descriptions." Bureau of Labor Statistics. https://www.bls.gov/help/def/jl.htm#rate/level. November 2002.
9. Z. Lipton. "The Foundations of Algorithmic Bias." Approximately Correct. http://approximatelycorrect.com/2016/11/07/the-foundations-of-algorithmic-bias/. November 2016.
10. J. Angwin, J. Larson, M. Surya, and L. Krichner. "Machine Bias." in "Machine Bias: Learning the Algorithms that Control Our Lives." ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. May 2016.
11. A. Akhtar. "Is Pokémon Go Racist? How the app by be redlining communities of color" in "Inequity in Silicon Valley." USA Today. https://www.usatoday.com/story/tech/news/2016/08/09/pokemon-go-racist-app-redlining-communities-color-racist-pokestops-gyms/87732734/. August 2016.
12. H. Reese. "Why Microsoft’s ‘Tay’ AI Bot Went Wrong" in "Innovation." TechRepublic. http://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/. March 2016.
13. N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos. (2016) "Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective." PeerJ Computer Science 2:e93. https://doi.org/10.7717/peerj-cs.93. October 2016.
14. IBM Watson Analytics. "Sample Data: HR Employee Attrition and Performance." IBM Watson Analytics Blog. September 2015.
15. M. Miller. "This Algorithm is Better at Predicting Human Behavior Than Humans Are" in "Co.Design: Evidence." Fast Company. October 2015.
16. J. Burn-Murdoch. "The Problem with Algorithms: Magnifying Misbehaviour" in "Big Data." The Guardian. https://www.theguardian.com/news/datablog/2013/aug/14/problem-with-algorithms-magnifying-misbehaviour. August 2013.
17. C. Chu. "Machine Learning Done Wrong" from "ML in the Valley: ML Lessons and Insights Learned from Industry Practice." Posthaven. http://ml.posthaven.com/machine-learning-done-wrong. June 2014.
18. E. Volini, and B. Dussert. "The Informed Executive: Improving Organizational Agility Through Workforce Analytics." Deloitte Touch Tohmatsu, Ltd. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/human-capital/us-cons-workforce-analytics.pdf. 2016.
19. C. Nielsen. "Collect Your Employee's Data Without Invading Their Privacy." Harvard Business Review. https://hbr.org/2014/09/collect-your-employees-data-without-invading-their-privacy. September 2014.
20. J. Walker. "Do New Job Tests Foster Bias?" Wall Street Journal. https://www.wsj.com/articles/SB10000872396390443890304578006283936708970. September 2012.
21. Bureau of Labor Statistics. "Help & Tutorials: Data Descriptions." Bureau of Labor Statistics. https://www.bls.gov/help/def/jl.htm#rate/level. November 2002.
22. Office of Personnel Management. "FERS Information." Office of Personnel Management. https://www.opm.gov/retirement-services/fers-information/eligibility. 2017.
23. C. Merhar. "Employee Retention – The Real Cost of Losing an Employee." Zane Benefits HR Blog. https://www.zanebenefits.com/blog/bid/312123/employee-retention-the-real-cost-of-losing-an-employee. February 2017.
24. J Fieldsend and R Everson. "Visualisation of multi-class ROC surfaces." University of Exeter, Department of Computer Science. http://users.dsic.upv.es/~flip/ROCML2005/papers/fieldsend2CRC.pdf. 2005.