Predictive Analytics for Workforce Stability: Identifying Attrition Factors

Michelle Zheng
Spring 2026 8W1
MSDS692

What is employee attrition?

- Employee leaves & position not immediately replaced w/ new hire
  - Voluntary
    - Quitting
    - Retirement
  - Involuntary
    - Layoffs
    - Termination

↑ employee attrition = ↑ hiring cost + ↓ team productivity

- Motivation? Current role in department conducting new hire 5 wk basic training course, typically 20% stayed after 6 mths

Employee Data & Ethical Concerns

1,470 records
35 features

- Dataset: synthetic HR Employee Attrition & Performance dataset from IBM

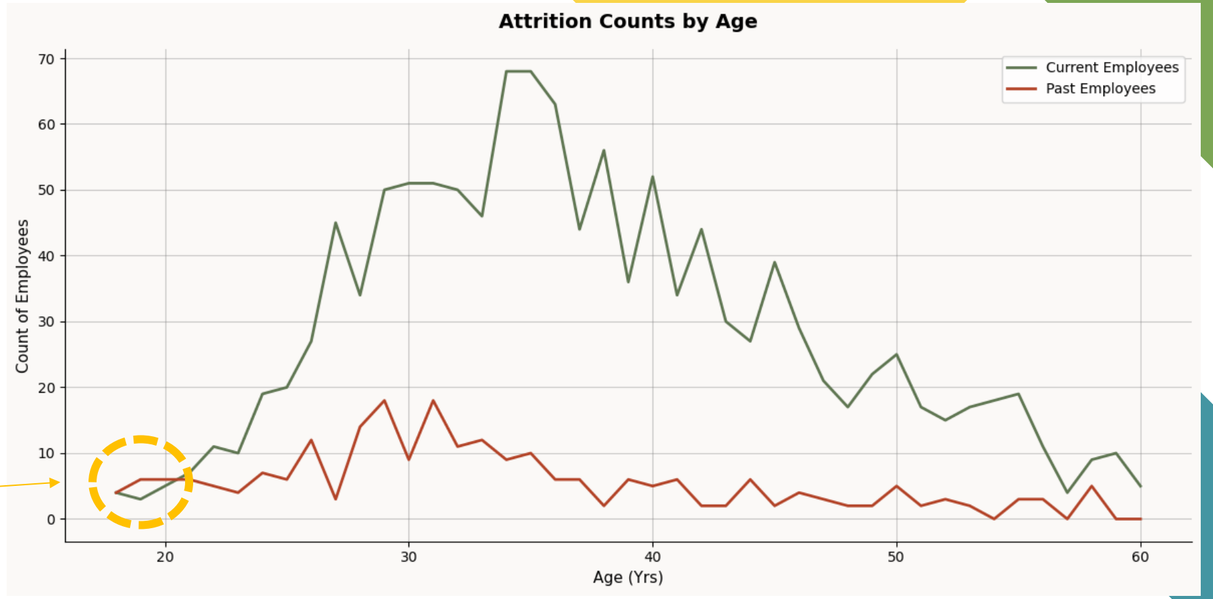
- Difficult to obtain real employee data
  - companies do not make personnel records public
  - accessing non-synthetic demographic data requires NDA making it inaccessible for open research

- Anonymized personal data is vulnerable to re-identification
  - Netflix public dataset challenge: researchers were able to reverse-engineer anonymized customer records to identify individuals

Current Employees: 83.88%

Past Employees: 16.12%

EDA & Preparation



(EDA) – Who had higher attrition rates?

- Department: 1. Sales 2. HR 3. R&D
- Monthly Income Group <$1,000-$4,000
- Years w/ Current Manager: 0-2 years
- 18-20 Age Employees

Data Preparation:

- Dropped constant/non-informative columns (EmployeeNumber, EmployeeCount)
- Rating based columns → categorical types (JobSatisfaction)
- Numerical column skewness >0.75 → log transformed

Independent t-test:

- H_0: Average income by month of past employees is the same as employees still present.
- $p < 0.0001$ → Average income by month of past employees **is different** as employees still present

Models: Logistic Regression vs Random Forest

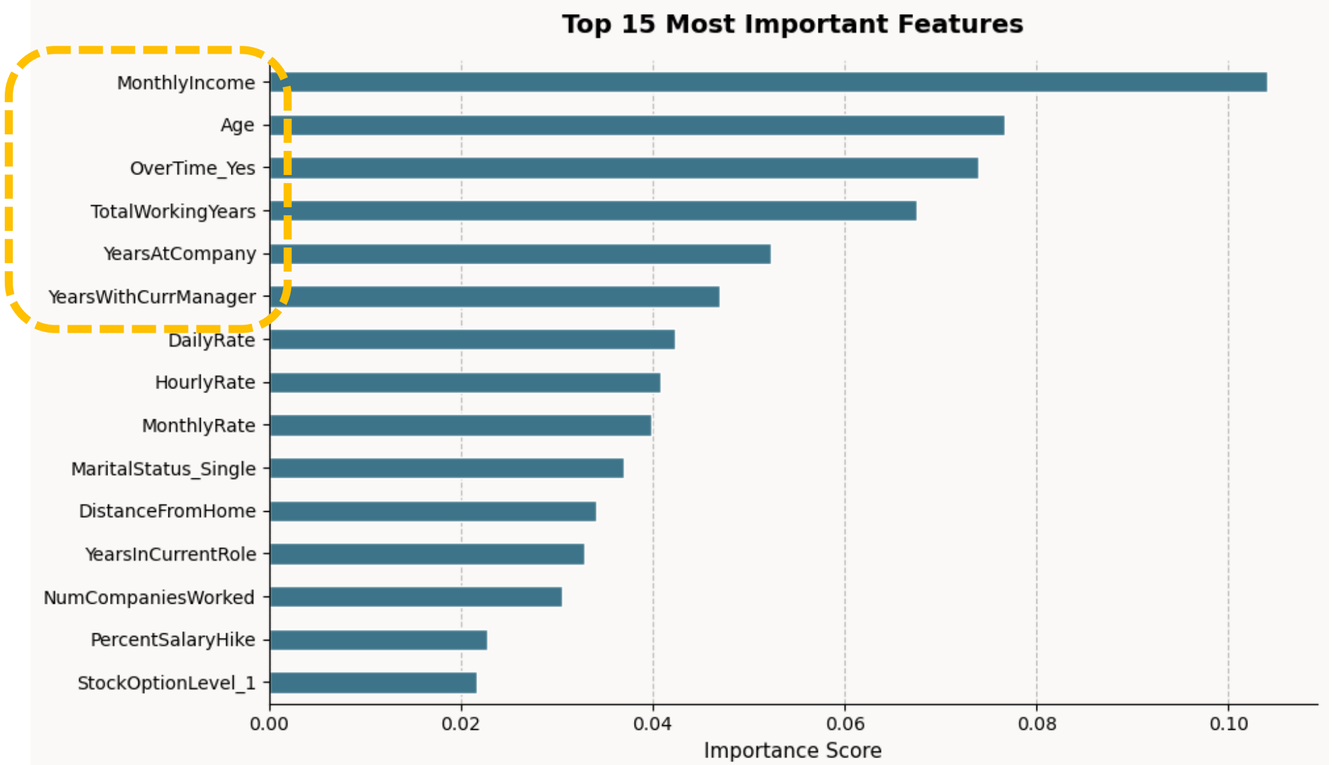
Correctly identified how many employees left in the test set

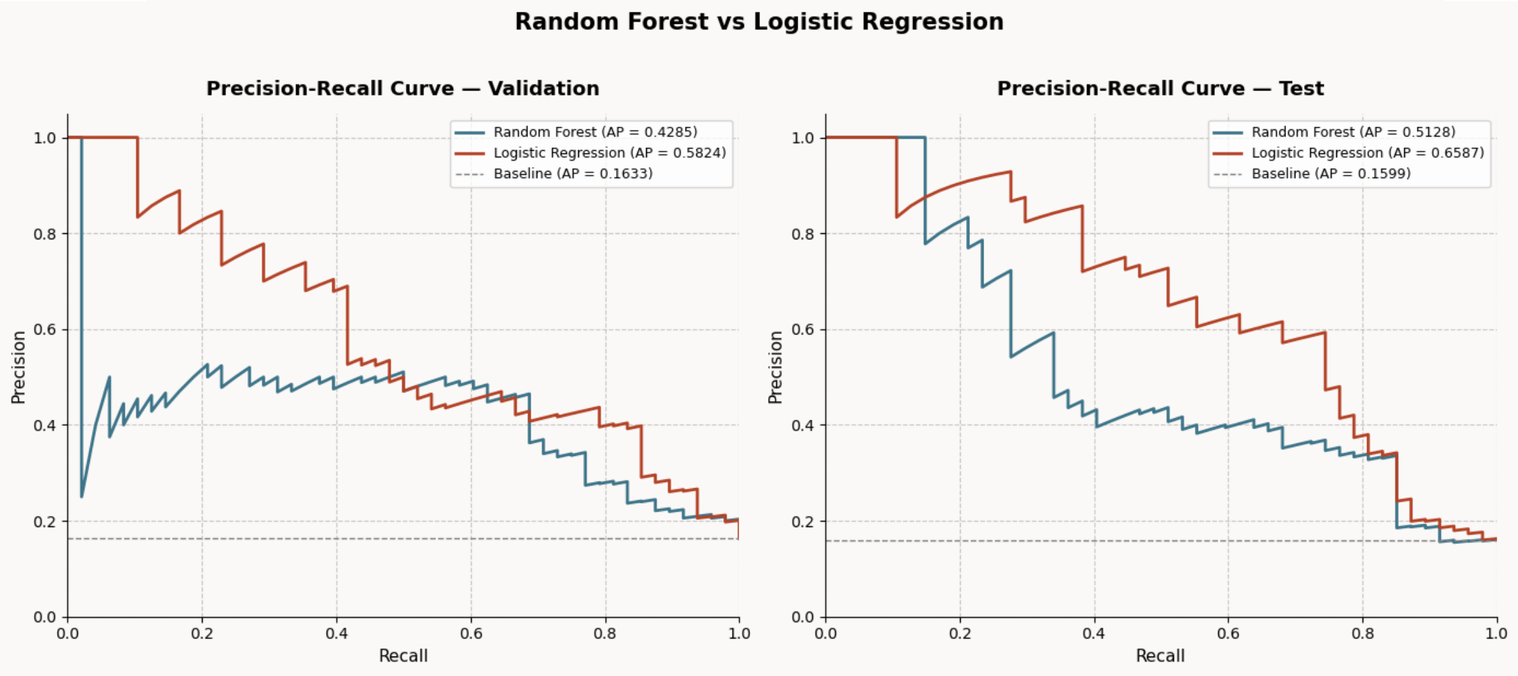
TABLE I
MODEL COMPARISON — METRIC SUMMARY

Model	Acc.	AUC	Prec.	Recall	F1	Avg P
RF — Val	83.33%	0.7952	0.3333	0.0208	0.0392	0.4285
RF — Test	84.69%	0.7719	1.0000	0.0426	0.0816	0.5128
LR — Val	78.91%	0.8413	0.4125	0.6875	0.5156	0.5824
LR — Test	79.25%	0.8374	0.4186	0.7660	0.5414	0.6587

Metrics:

- *AUC & Recall > Acc.*





Findings

- Money (income) matters!!
- Younger people (18-20) tend to leave the company
- Logistic Regression > Random Forest

Future Directions

- Real anonymized employee data
- Real time system for continuous monitoring and flagging

References

Alqahtani, H., Almagrabi, H., & Alharbi, A. (2025). Dataset for predictive modelling and analysis of employee attrition and retention. *Data in Brief*, *63*, 112242. https://doi-org.dml.regis.edu/10.1016/j.dib.2025.112242

Cooper, A. K., & Coetzee, S. (2020). On the ethics of using publicly-available data. *Lecture Notes in Computer Science*, 159–171. https://doi.org/10.1007/978-3-030-45002-1_14

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random Forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1). https://doi-org.dml.regis.edu/10.1186/s12859-018-2264-5

Govindarajan, R., Kumar, N. K., P, S. R., E, S. P., B, D., & G, P. K. (2025). Predicting employee attrition: A comparative analysis of machine learning models using the IBM Human Resource Analytics Dataset. *Procedia Computer Science*, *258*, 4084–4093. https://doi-org.dml.regis.edu/10.1016/j.procs.2025.04.659

Hoffman, M., & Tadelis, S. (2021). People Management Skills, employee attrition, and manager rewards: An empirical analysis. *Journal of Political Economy*, *129*(1), 243–285. https://doi.org/10.1086/711409

Kim, J. (2025). The effect of mismanagement of poor performers on their coworkers' turnover intentions. *Public Personnel Management*, *55*(1), 118–144. https://doi-org.dml.regis.edu/10.1177/00910260251360823

Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous. *SMU Data Science Review*, *1*(3). https://scholar.smu.edu/datasciencereview/vol1/iss3/9

Lee, F., & Algarra, A. (2025). Leveraging topic modeling to predict and prevent employee attrition. *Information Systems Education Journal*, *23*(4), 59–84. https://doi.org/10.62273/mqnf1140

Manafi Varkiani, S., Pattarin, F., Fabbri, T., & Fantoni, G. (2025). Predicting employee attrition and explaining its determinants. *Expert Systems with Applications*, *272*, 126575. https://doi-org.dml.regis.edu/10.1016/j.eswa.2025.126575

Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the Netflix Prize dataset. arXiv. https://doi.org/10.48550/arXiv.cs/0610105

Srivastava, G. N., Sharma, H., Agarwal, R. N., & Jain, A. K. (2025). Analyzing employee attrition of research and development firms using mixed methods. *Cogent Business & Management*, *12*(1). https://doi-org.dml.regis.edu/10.1080/23311975.2025.2565439

Susser, D., Schiff, D. S., Gerke, S., Cabrera, L. Y., Cohen, I. G., Doerr, M., Harrod, J., Kostick-Quenet, K., McNealy, J., Meyer, M. N., Price, W. N., & Wagner, J. K. (2024). Synthetic Health Data: Real ethical promise and peril. *Hastings Center Report*, *54*(5), 8–13. https://doi.org/10.1002/hast.4911