Big Data Group Project Report
People Analytics with Human Resources Dataset
Team 9: Xiao Han, Jent LaPalm, Aoran Wang, Yushan Yang

Executive Summary

This report on our company's HR data answers broad business questions: *who are we? how are we doing?* and *how can we do better?* Key insights: online recruiting leads in landing new & diverse hires. We have gender disparities to address. The majority of our staff meets performance goals; we have identified attrition reasons and potential redress. We've drilled down on manager performance and quantified manager-team impact. Finally, we've modeled employee attrition, enabling proactive retention.

Introduction

Many companies lack data maturity in HR — so-called 'people analytics.' (Mckinsey). Addressing this is multiplicative in value, from questions such as 'which interview questions predict top performers' to complex questions like 'should we use an internal team or 3rd-party vendor for a project?' This adds value, for example, by increasing worker-effectiveness, driving recruiting strategy, and lowering attrition costs.

Data Wrangling

For our EDA and summary statistics, we used SQL to query and Tableau to visualize (*exhibits 1-7*). For modeling and manager-performance, data-changes were needed: employees with no manager were given an id, and dates were converted to DateTime. This was challenging for 2-digit years, as Python converts some to future dates (e.g. 2067), requiring additional correction. We created new features: employee age, salary/age ratio, and a below-average-position-salary flag. For predictive modeling, we created binary variables representing recruiting source (LinkedIn, Google Search, etc.). This was worthwhile, (*exhibit 11*); our best model used both salary/age ratio and a recruiting variable. The data was randomly shuffled and split into the train, validation, and test sets.

Challenges and Limitations

One challenge in HR data is its private, personal nature. As such, there is a dearth of public HR

datasets. Most are synthetic, created by professors or data professionals. This causes limitations: our dataset is of a single company, preventing industry analysis. The data contains engagement and satisfaction scores, however, these only appear once, so we cannot analyze trends. The data's smaller size (>1000 rows) means overfitting is more likely when modeling.

Insights & Analytics: *Model-free data insights*

In our company, there are more females than males (*exhibit* 3); for balance, we suggest HR hire more males (*exhibit 4*). Based on the gender-distribution by department, we calculated the ratio of females over males. The admin office has twice the number of females than males, however, the gender ratio is inverted in the IT/IS department. Knowing this, hiring managers can improve departmental gender balance. Hiring more males would improve the balance for the company as a whole.

Secondly, we focused on staff with good performance (*exhibit 7*). The top reasons for 'good' employees to leave are 'another position', 'unhappy' and 'more money'. To increase retention, HR should consider their career development, mental health, and provide raises if possible.

Looking at 'unhappy' employees, their average number of late days over the last month is 1, which is higher than the average of others (0.3). This may be because people are unhappy, because they are not satisfied with their schedule, or perhaps the punishment for lateness. Or, they may be late because they are unhappy about their work. In general, looking at late-days over rolling 30-day periods could help HR identify employees who are more likely to have problems.

Our company evaluates employee-effectiveness by performance score: 1 for 'PIP', 2 for 'needs improvement', 3 for 'fully meets', and 4 for 'exceeds'. We created a manager-group performance overview (*exhibit 9*). Some managers led employees (EmpID) that all had a score ≥ 3, while a few led disparate performers, like manager 12, who had groups of outstanding employees and people under PIP.

Insights & Analytics: *Model-driven data insights*

One goal was to model employee attrition. This can not only proactively address about-to-churn employees, but also highlight what predicts churn. To establish baseline model performance, we ran a linear regression with forward-step feature selection, logistic regression, random forest, XGBoost, and

neural net. Surprisingly, the linear regression was our best initial model as measured by AUC score. Next, we used cross-validated randomized search to tune XGBoost, grid-search cross-validation to tune the logistic regression, and tested various neural net depths. We were able to improve each model's results; our best performer was the tuned XGBoost (*exhibit 8*). The initial superior performance of linear regression may be due to the dataset's small size, which can lead to overfitting. The linear regression's most-predictive features were marital status, performance score, zip code, absences, age, and recruitment via Google Search. These insights can be used as the basis for future investigations: *does advertising with Google increase high-value applicants?, can we better raise performance score?* etc.

Another goal was to explore the relationship between employee performance and their manager. Having observed the difference in each group (*exhibit 9*), we ran a logistic regression and calculated odds ratio (OR) to evaluate how a manager affected the employee's performance: OR > 1 if the effect is positive, OR < 1 if negative and OR ≈ 1 if no effect ([0.99999, 1.00001])

In the result (*exhibit 10*), the OR of managers aligned with the manager-employee performance chart and identified how managers affected employee performance. Managers (by ID) who positively affected the employee's performance are: 2, 4, 5, 6, 10, 12, 14, 16, 18, 19, 40. Managers 7, 11, 15, 17, 20, 21, 22, 39 negatively affected employee performance — they had quite a few employees whose performance score was 1 or 2. Managers who had no obvious impact (1, 3, 9, 13, 30) led a group with a performance score of 3. Our analysis can answer business-related questions, like *how does the manager affect an employee's performance? How can we use metrics to evaluate our managers' leadership?* etc.

Insights & Analytics: *Reasons for Selected Approach*

As an exploratory problem, the model of performance versus manager explored the association between ordinal and nominal variables, making metrics like correlation meaningless. Instead, we used ordered multinomial logistic regression. We also chose OR as our metric because it quantifies the strength of association (whether it exists, positive or negative).

To predict attrition, we iteratively followed the crisp-dm framework. After initial modeling, we

refined the most-promising candidates. As an example of this iterative approach, initially, we did not use recruitment source as a variable, however, after doing so, we saw marked improvements. The final selection of XGBoost is appropriate not only for its superior performance, but also because it can grow with our business over time. The algorithm performs well over large datasets, allowing for fast computations as more categorical features are recorded.

Discussion: *Generalizability of the insights*

Our framework of EDA, demographic statistics, performance-quantification, and attrition-prediction is broadly applicable to any company's HR data. The drivers of attrition in our model *may* be predictive at other companies (particularly absences), however, this requires further study. Our other insights apply only to our company. This is one of the difficulties in people analytics: because the data is private, it is difficult to obtain, and difficult to compare between companies or industries.

In the short and medium term, our company can continue to store HR information on a relational database, however, as the department grows, we recommend investigating a switch to a NoSQL DB like MongoDB, and storing data in JSON files.

Discussion: *SWOT Analysis of Insights*

Strength: 1. Clear overlook at general performance by staff 2. Solid understanding of employee structure 3. Identify group and individual level problems like gender disparity and bad performance 4. Exhibit performance on recruiting strategy 5. Predict attrition

Weaknesses: 1. limited data size 2. limited data source 3. potential data leakage

Opportunities: 1. Expansion into more effective recruiting methods 2.Discover potential cause of structural and individual problems and make improvement, thus increase organizational effectiveness 3. Reduce retention cost by acting proactively

Threats: 1. Lacking ability to apply model industrial-wise 2. Potential overfitting when applying a model to a specific organization 3. Emerging new organizational structure and performance measurements

Appendix *(Data Source, Exhibits)*

Data Source

Prof. Huebner, Rich. *Human Resources Data Set*. *Kaggle*, 4/29/2021. Dataset

https://www.kaggle.com/rhuebner/human-resources-data-set, accessed 11/22/2021

Visualizations

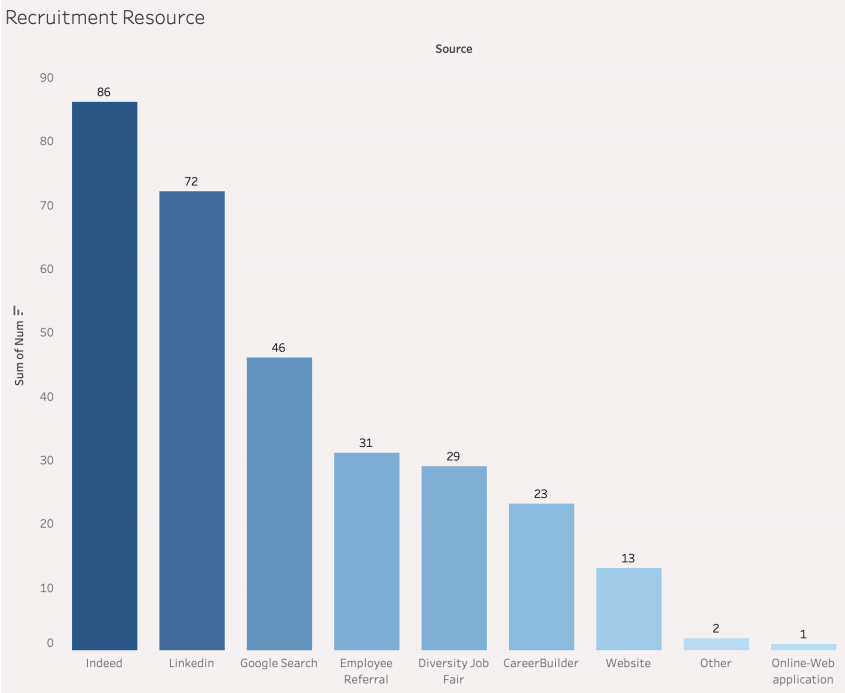

Exhibit 1

Number of employees from each of recruiting source

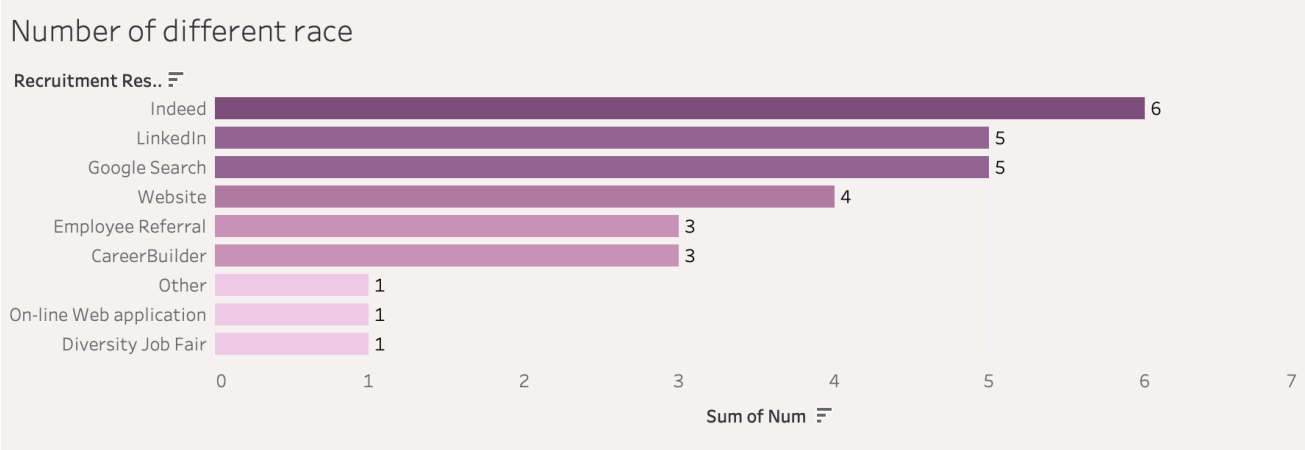

Exhibit 2

Different number of races of employees from each of recruiting source

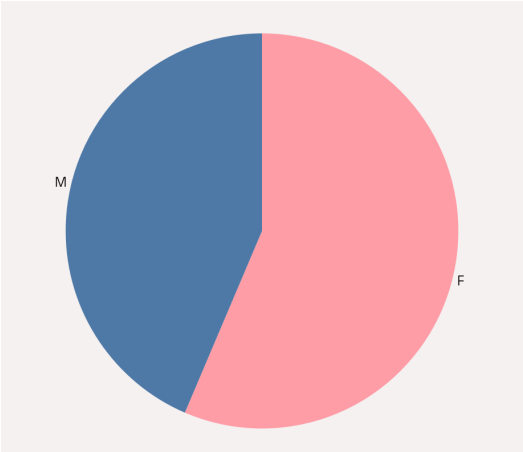

Exhibit 3

Male and Female in this company

There are more males(132) than females(171) in this company.

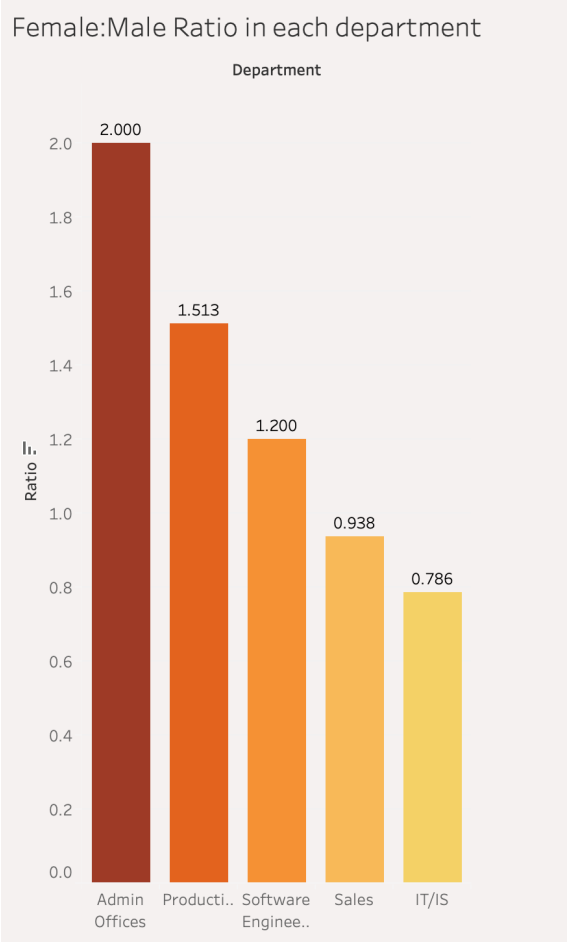

Exhibit 4

Female : Male ratio in each department

Number of people in each department

Department	Num	
Production	201	Abc
IT/IS	50	Abc
Sales	31	Abc
Software Engineering	11	Abc
Admin Offices	9	Abc
Executive Office	1	Abc

Exhibit 5

Number of people in each department

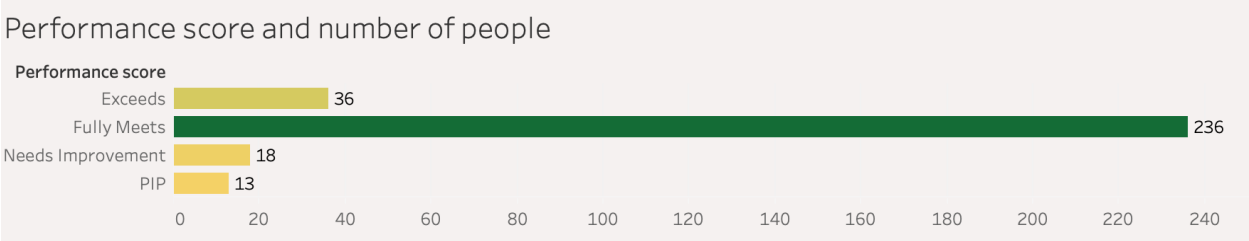

Exhibit 6

Number of people for each performance score

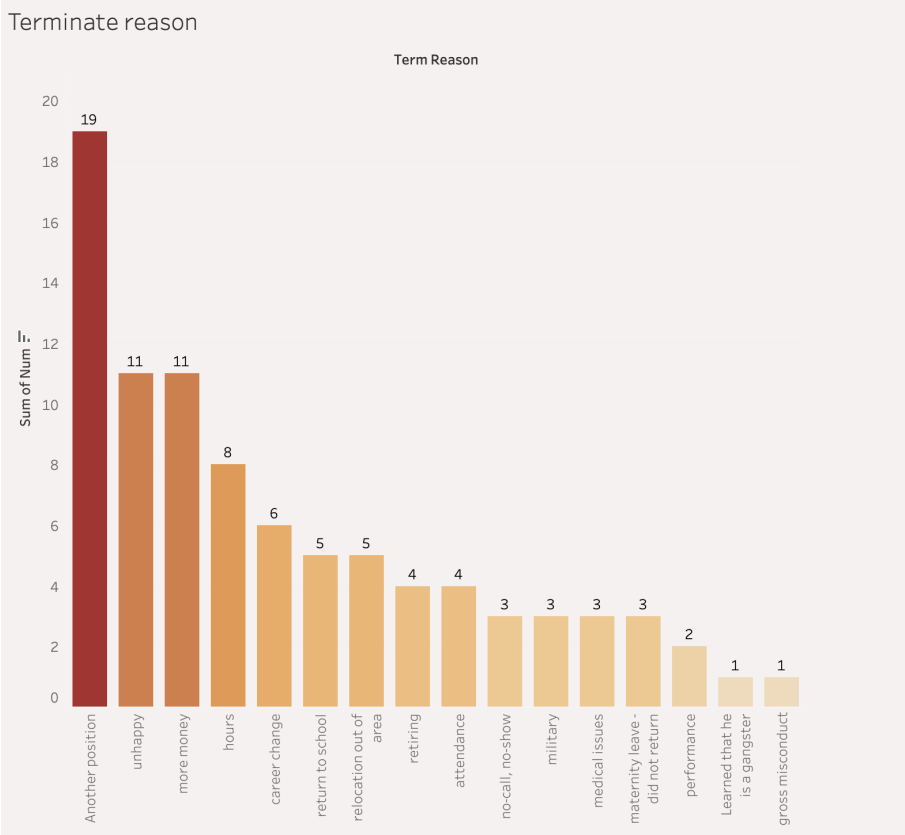

Exhibit 7

Terminate reason collection

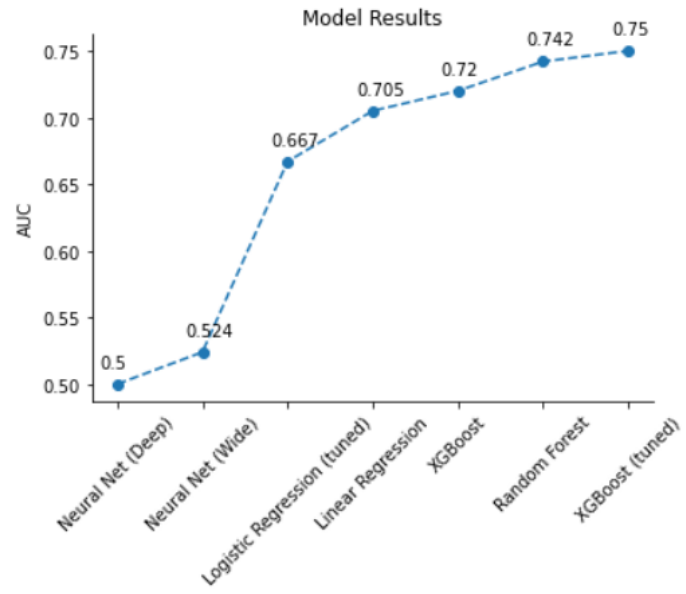

Exhibit 8: *Attrition Model Results*

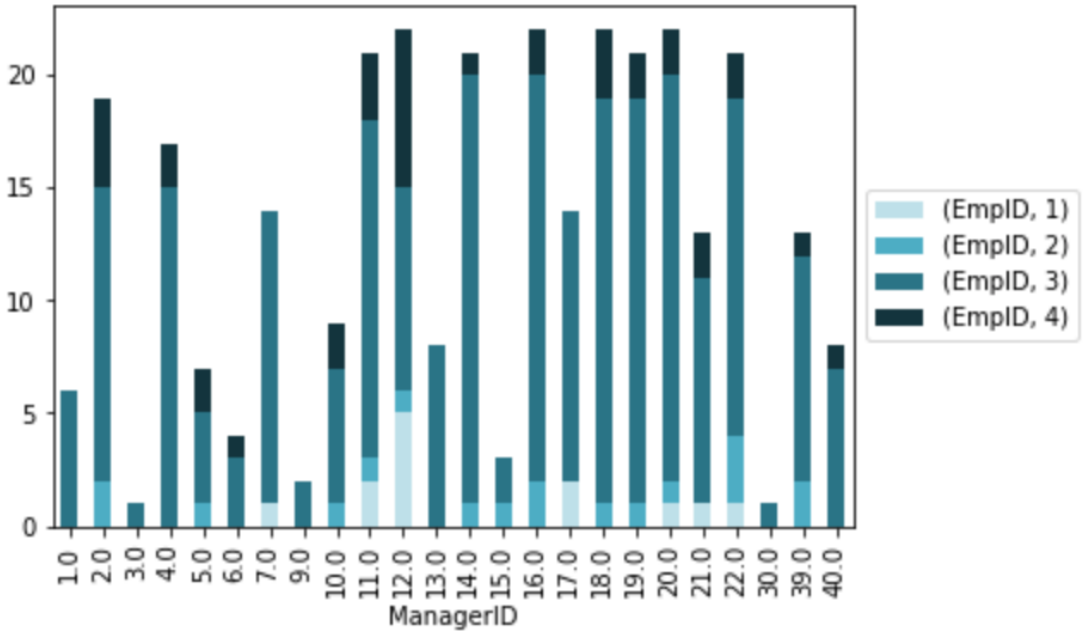

Exhibit 9: *How each group led by managers perform?*

```
##                    OR      2.5 %    97.5 %
## ManagerID2  1.8328956 0.22078044 15.132163
## ManagerID3  1.0000738 0.01024337 97.623620
## ManagerID4  1.8065533 0.21619686 14.980673
## ManagerID5  2.3281327 0.17253114 28.241924
## ManagerID6  3.3470503 0.18229595 50.501038
## ManagerID7  0.6761920 0.07574898  6.126039
## ManagerID9  1.0000357 0.02905644 34.415729
## ManagerID10 1.9017390 0.16625576 20.449244
## ManagerID11 0.9763544 0.11789337  8.090488
## ManagerID12 1.7406275 0.20491592 14.597648
## ManagerID13 0.9999381 0.08844520 11.306359
## ManagerID14 1.0112437 0.12671161  8.077790
## ManagerID15 0.2590817 0.01625942  5.517091
## ManagerID16 1.0227739 0.12851206  8.147887
## ManagerID17 0.4512219 0.05127248  4.056224
## ManagerID18 1.6260696 0.20643928 12.818396
## ManagerID19 1.2965023 0.16186203 10.376640
## ManagerID20 0.9954016 0.12423075  7.976375
## ManagerID21 0.9397813 0.09403816  9.369710
## ManagerID22 0.6078240 0.07580432  4.894598
## ManagerID30 1.0000738 0.01024337 97.623620
## ManagerID39 0.6940828 0.07562086  6.528888
## ManagerID40 1.8743579 0.16153964 20.285553
```

From the OR table, we can find that all OR values are within the range of confidence interval.

Findings

Managers who lead improvement: 2, 4, 5, 6, 10, 12, 14, 16, 18, 19, 40.
Managers who may lead inhibition: 7, 11, 15, 17, 20, 21, 22, 39.
Managers who have no obvious association with performance: 3, 9, 13, 30.

Exhibit 10: *Odds Ratio for Manager Performance*

Where is manager 1: In our case, manager 1 is the reference for other managers, so analysis of the reference will be omitted by R.

Manager 1 leads a small group where everyone's performance score = 3. According to our findings, this manager should belong to those who have no obvious association with performance, although we cannot statistically prove that. It also makes manager 1 a good reference with no bias.

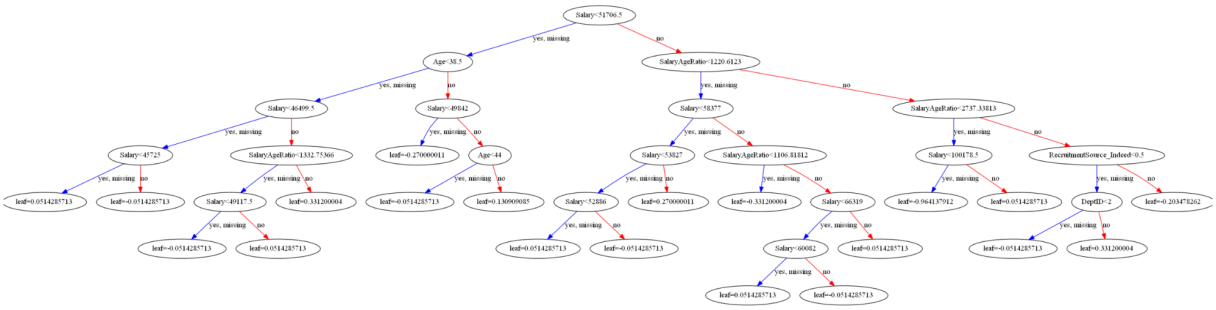

Exhibit 11: *XGBoost Tree*