

Predicting Employee Attrition Using Decision Tree, KNN, and Logistic Regression

University of California, Berkeley | School of Information

DATASCI 207 Applied Machine Learning - Spring 2023

Final Project | Baseline Presentation

Team: Ivy Chan, John Gibbons, Mark Herrera, Maria Manna

February 28, 2023

AGENDA

1. Research Question
2. Data
3. Summary Statistics
4. Correlation Heatmap
5. Prediction Algorithms
6. Decision Tree for Info Gain
7. Baseline & Evaluation Metrics

Research Question

- What is the likelihood of an active employee leaving the company?
- What factors are most predictive of an employee's attrition?



Understanding these questions allows businesses to better identify at-risk employees. It can also assist in the hiring process if there are known factors that influence employee churn. In the past two years, employers have been significantly impacted by events such as the “Great Resignation,” losing employees to other employment opportunities. The knowledge gained by our research has the potential to lessen the impact of future periods of high employee turnover for businesses who value a low employee turnover rate.

Data

We will use an employee attrition dataset from Kaggle containing the employee data from IBM HR Employee Attrition and Performance. The data from this dataset is structured.

The dataset size: 1,470 observations and 35 features.

The target variable: Attrition.

Some main features we will be utilizing from this dataset are:

- Age
- DistanceFromHome
- Education
- EnvironmentSatisfaction
- Gender
- HourlyRate
- JobInvolvement
- JobLevel
- JobSatisfaction
- MonthlyIncome
- MonthlyRate
- NumCompaniesWorked
- OverTime
- PercentSalaryHike
- PerformanceRating
- StockOptionLevel
- TotalWorkingYears
- WorkLifeBalance
- YearsAtCompany
- YearsInCurrentRole
- YearsSinceLastPromotion
- YearsWithCurrManager

The Kaggle logo, featuring the word "kaggle" in a light blue, lowercase, sans-serif font.

Data Source: <https://www.kaggle.com/code/hamzaben/employee-churn-model-w-strategic-retention-plan/data>

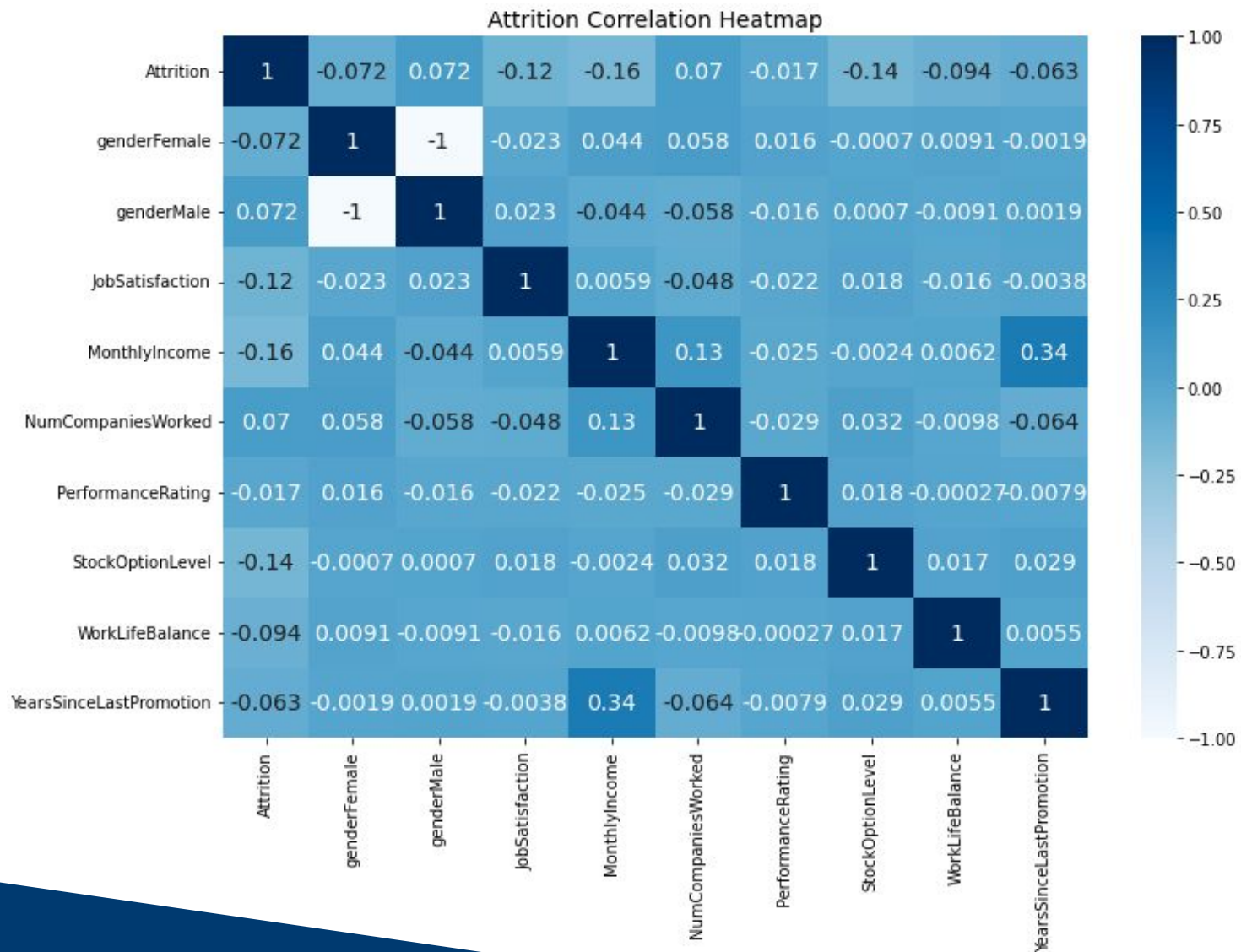


Summary Statistics

	Attrition	Age	JobSatisfaction	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	PerformanceRating	WorkLifeBalance	YearsSinceLastPromotion
count	1029.000	1029.000	1029.000	1029.000	1029.000	1029.000	1029.000	1029.000	1029.000
mean	0.161	36.985	2.756	6517.126	2.723	15.231	3.158	2.737	2.188
std	0.368	9.194	1.095	4658.337	2.532	3.668	0.365	0.710	3.215
min	0.000	18.000	1.000	1052.000	0.000	11.000	3.000	1.000	0.000
25%	0.000	30.000	2.000	2936.000	1.000	12.000	3.000	2.000	0.000
50%	0.000	36.000	3.000	4969.000	2.000	14.000	3.000	3.000	1.000
75%	0.000	43.000	4.000	8381.000	4.000	18.000	3.000	3.000	3.000
max	1.000	60.000	4.000	19999.000	9.000	25.000	4.000	4.000	15.000

df = training data subsetting for key variables of interest (data not yet standardized)

Correlation Heatmap



Prediction Algorithms

1. Decision tree
 - a. Uses: feature selection & prediction
 - b. Measure with information gain
 - c. Potential Ensemble Method: Random Forest (bagging/bootstrapping)
2. K-Nearest Neighbors (KNN)
 - a. Use: identify employees with higher risk of attrition by comparing to profiles of known former employees
 - b. Measure with Euclidean distance
3. Multivariate Logistic Regression
 - a. Use: binary classification problems (will an employee turnover or not?)
 - b. Will utilize Stochastic Gradient Descent (SGD) and logistic loss



Decision Tree - info gain algorithm to assist in feature selection

- Utilized information gain algorithm to help with feature selection. Identifying parameters with most information gained (partitioning the data).

```
0 0.011 Age
1 0.008 BusinessTravel
2 0.004 DistanceFromHome
3 0.000 Education
4 0.006 Environmentsatisfaction
5 0.006 JobLevel
6 0.005 JobSatisfaction
7 0.009 MonthlyIncome
8 0.001 NumCompaniesWorked
9 0.028 OverTime
10 0.000 PercentSalaryHike
11 0.000 PerformanceRating
12 0.033 StockOptionLevel
13 0.012 TotalWorkingYears
14 0.006 WorkLifeBalance
15 0.013 YearsAtCompany
16 0.014 YearsInCurrentRole
17 0.003 YearsSinceLastPromotion
18 0.010 YearsWithCurrManager
```

Feature Selection

1. Stock Option Level
2. Overtime
3. Years in current role
4. Years at the company
5. Total working years
6. Age
7. Years with current manager
8. Monthly income
9. Business Travel
10. Job level/ Environment Satisfaction

Baseline & Evaluation Metrics



- We use Log Loss to calculate the baseline error.
- The target variable “Attrition” is binary
 - Yes means leaving the company
 - No means staying
- The attrition rate is 16.1% (percentage of people leaving the company).
- Our baseline prediction is to always predict the majority class, i.e., No.
- The Log Loss of the baseline prediction is 5.57



- Matthew’s Correlation Coefficient (MCC)
- F1 Score - useful for unbalanced classes
- Precision

Thank You

References

References

Content

- https://scikit-learn.org/stable/supervised_learning.html
- <https://medium.com/@mikeusru/common-metrics-for-evaluating-natural-language-processing-nlp-models-e84190063b5f>
- https://xgboost.readthedocs.io/en/stable/get_started.html
- <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62>
- <https://www.v7labs.com/blog/f1-score-guide#:~:text=F1%20score%20is%20a%20machine%20learning%20evaluation%20metric%20that%20measures,prediction%20across%20the%20entire%20dataset.>

References

Images

- <https://www.voxco.com/blog/employee-turnover-a-guide/>
- <https://en.wikipedia.org/wiki/Kaggle>
- https://d3njcjbhbojbot.cloudfront.net/api/utilities/v1/imageproxy/https://coursera-course-photos.s3.amazonaws.com/4e/2b9450fd5011e88a28fd978cb69b7d/Public-Health-Biostatistic_Logo5_Multiple-Regression-Methods-04.png?auto=format%2Ccompress&dpr=1&w=175&h=175&fit=fill&bg=FFF
- <https://commons.wikimedia.org/wiki/File:Performance-Evaluation-Process-z.jpg>
- <https://freesvg.org/colorful-gague>
- https://www.ibm.com/brand/experience-guides/developer/b1db1ae501d522a1a4b49613fe07c9f1/01_8-bar-positive.svg