Decreasing Employee Attrition: One Prediction At A Time

By: Emily Phillips

Introduction to the Business Problem(s)

What is Employee Attrition?

Definition : “When an employee leaves the company through any method, including voluntary resignations, layoffs, failure to return from a leave of absence, or even illness or death”

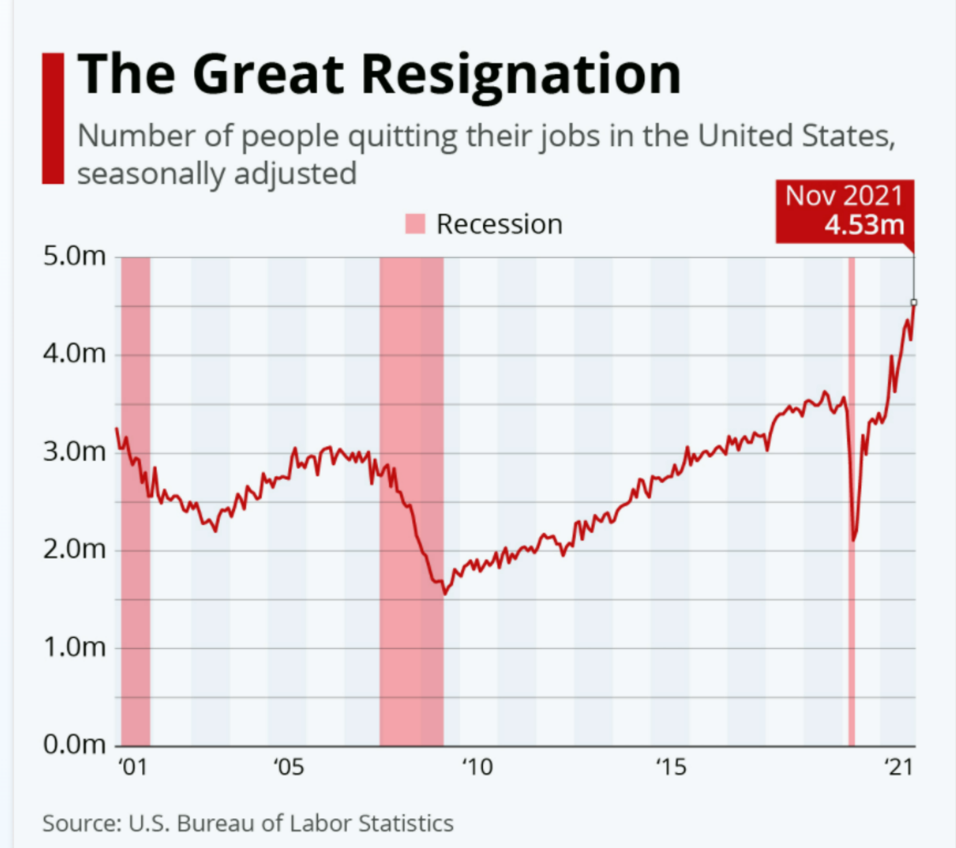
Two types of employee attrition:

1. Voluntary
   a. Employee deliberately chooses to leave their company
   b. Company has the decision to not replace the employee
2. Involuntary
   a. The company chooses to let go of (fire) an employee
   b. EX) Reorganization, layoffs
   c. In this case, the company either eliminates the employee’s position altogether or chooses not to replace it

https://www.digitalhrtech.com/employee-attrition/

Issues with Voluntary Employee Attrition

1. Lack of continuity and consistency within organization and within teams
2. Training gaps
3. Lack of institutional knowledge
4. Higher recruiting efforts, longer recruiting time
5. Empty positions can be left for some time; leaves teams understaffed
6. Possible burnout for remaining employees -- overloaded and overworked
7. Increase in recruiting, hiring and training costs



The United States is at an all-time high for number of people quitting their jobs!

[Great Resignation Statistics]

Statistical Question(s): What factors play a significant role in voluntary employee attrition, and how can we predict it?

Dataset Explanation

The Dataset: IBM HR Analytics Employee Attrition & Performance

- Employee survey from IBM, a technology corporation
- Gathered information around employee satisfaction, income, job level and some personal demographic data such as age and gender
- Dataset structure:
  - 1470 rows of data
  - 35 columns

- Each record in the dataset pertains to an employee that is either currently at IBM or voluntarily left
  - Attrition = ‘Yes’ or ‘No’

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Key Factors in Employee Attrition

- New survey by consulting firm McKinsey & Company released on March 9, 2022
- Surveyed nearly 600 employees between December 2020 and December 2021
    - The Great Resignation was at a peak in November 2021
- Found four top factors contributing to this historical event:
    - Unsustainable workloads
    - Uncaring managers
    - Inadequate compensation
    - Lack of career advancement potential
- "About 35% of respondents said unsustainable work performance expectations were the reason they left a job without another in hand"

https://www.forbes.com/sites/emmylucas/2022/03/09/employees-say-unsustainable-workloads-and-expectations-are-driving-them-to-quit/?sh=5880b2d57c34

	JobSatisfaction	YearsWithCurrManager	YearlyIncome	YearsSinceLastPromotion
count	1273.000000	1273.000000	1273.000000	1273.000000
mean	2.722702	4.018853	79330.812255	2.165750
std	1.102140	3.590863	57494.432802	3.251031
min	1.000000	0.000000	12108.000000	0.000000
25%	2.000000	2.000000	34932.000000	0.000000
50%	3.000000	3.000000	60804.000000	1.000000
75%	4.000000	7.000000	103452.000000	2.000000
max	4.000000	17.000000	239676.000000	15.000000

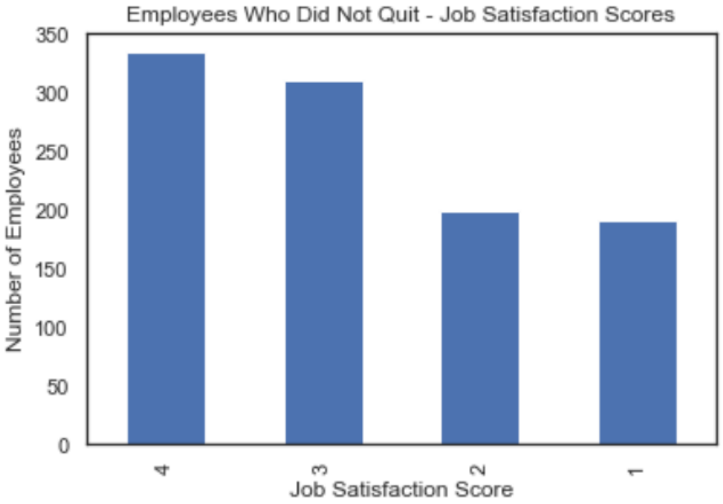
Description of Aligning Factors from the Kaggle Dataset

Data Visualizations – Exploration with Target Variable ‘Attrition’

Target Variable – Class Distribution

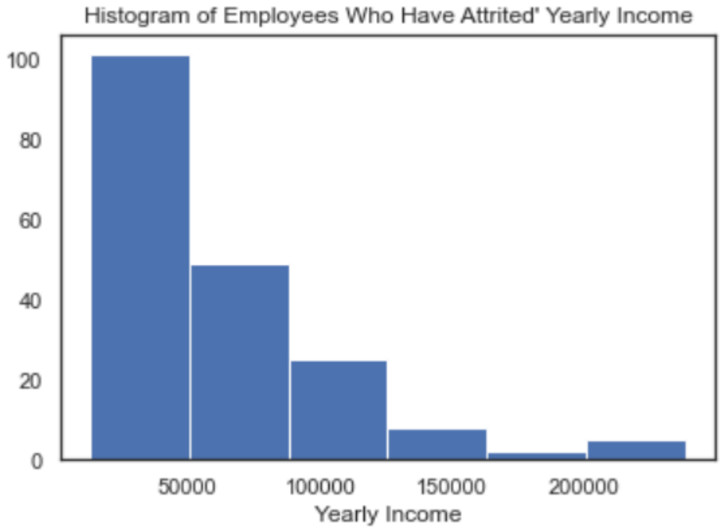
- Major imbalance between the two classes of ‘Yes’ & ‘No’
- ‘No’: 1,059 values
- ‘Yes’: 214 values
- CHALLENGE: Need to handle class imbalance before performing predictive modeling

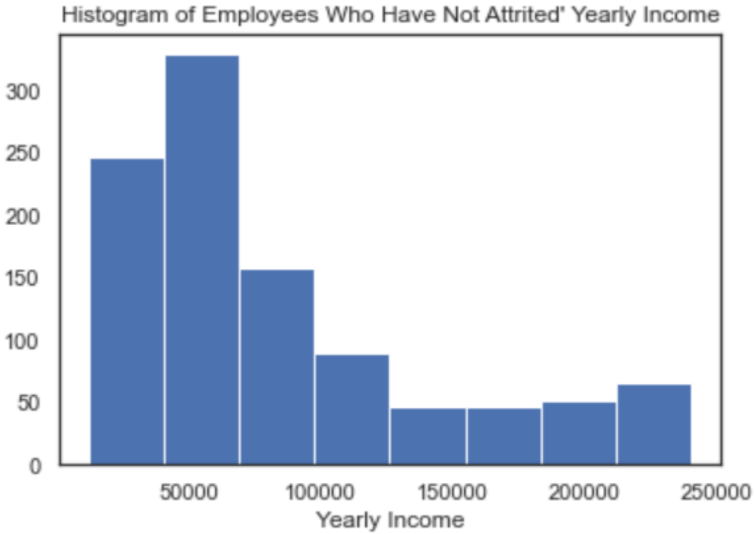
Bar Charts of Job Satisfaction Scores



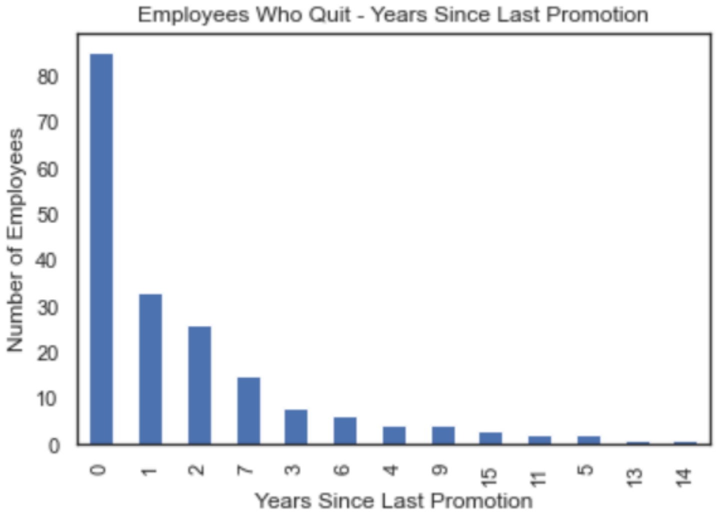
Bar Charts of Years With Current Manager

Histograms of Yearly Income (Salary)





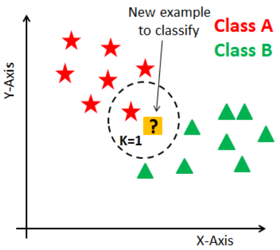
Bar Charts of Years Since Last Promotion



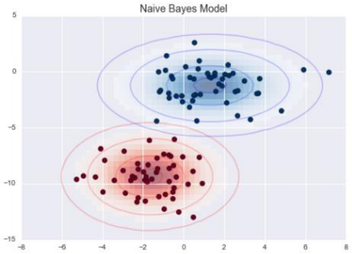
Predictive Modeling – Methods & Results

Modeling Deployment

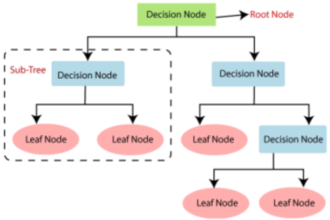
Following my data preparation, I began to evaluate algorithms and methods to help use on the prediction problem. I focused on models appropriate for classification(predictor of ‘Yes’ or ‘No’).



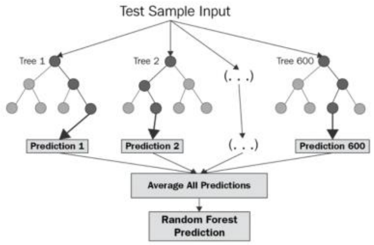
K-Nearest Neighbor - Straight forward pattern recognition model which allows the testing of several k values and leaf sizes to determine the best performance



Naive Bayes - Calculates the possibility of whether a data point belongs within a certain category



Decision Tree Classifier - A decision tree is a supervised learning algorithm that performs strong in classification problems



Random Forest Classifier - Expands beyond a decision tree by constructing multiple decision trees to remediate forcing a binary decision

Modeling Results

- Three Sets of Models:

1. Twenty-one features (A)
2. Four features (B)
    1. Job Satisfaction
    2. Years With Current Manager
    3. Yearly Income
    4. Years Since Last Promotion
3. Two features (C)
    1. Job Satisfaction
    2. Yearly Income

- Models were evaluated based on the following measures:
    - Accuracy
    - Precision
    - Recall
- Confusion matrices were also generated for each model in each set of data
- Best model differed among the sets

A. Random Forest – 84.5% accuracy, high precision and recall for 'Yes' class (58.3% and 16.7% respectively)

B. Naïve Bayes – 83.2% accuracy, Decision Tree Classifier has better precision and recall (64.3% and 21.4% respectively)

C. Naïve Bayes – 83.2% accuracy, Random Forest has overall best precision and recall (25% and 26.2% respectively)

Future Uses & Recommendations

Proactivity vs. Reactivity

Companies need to stay ahead of employee resignations!

- Build models based on organization’s human resources data
  - Other techniques can be taken to gauge feature importance in order to better narrow down the features that are critical for accurate model predictions
- Deploy them throughout departments, one at a time, to start testing models on employees
- Gauge trends and design initiatives to check-in with employees on their workloads, job satisfaction, career goals, etc.
  - Suitable team sizes
  - Bi-weekly manager 1:1s
  - Awareness of job opportunities → within and outside of teams
  - Mental health days
- Proactive plans can be rolled out in phases to ensure proper controls and management are in place for an ethical process

Conclusion and Q & A

1. Why were you interested in this business problem of employee attrition?

2. How did you account for the imbalance in the target classes?

3. How often do you think these models should be tested on employees?

4. Should the dataset consider outside-of-work features that can play into employee attrition such as mental health, changing career paths, etc.?

5. Is there a certain department of organizations that is experiencing higher attrition rates than others?

6. Why was employee job level not considered in the modeling, i.e. leadership positions vs. technical positions? Do you think this would make a difference?

7. Could you have employed other techniques for feature reduction?

8. Did you consider other forms of visualizations for the exploration data analysis (EDA)?

9. Would you suggest that companies be transparent around this work if they choose to implement it for employees?

10. How does predicting whether an employee will resign (Yes/No) compare with assigning them risk scores as also recommended?