Decreasing Employee Attrition: One Prediction at a Time

Emily Phillips

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catherine Williams

April 3,2022

Background

Starting from the year of 2021, the Great Resignation began its rise to impact. Employees began to quit at higher rates than ever before, causing great turnover in the organizations they left behind. According to an article titled 'The Number of Americans Quitting Their Jobs in 2021 - And Will It Continue to Grow' written by Andrew Lisa from yahoo!, the pivotal event in history started in April 2021 "when a record-high 3.8 million workers quit their jobs in a single month" (Lisa, 2021). This record-high amount of movement led to a "national quit rate of 3.1%" which was "the highest since [the Bureau of Labor Statistics] BLS started keeping records in 2000" (Lisa, 2021). Now moving into 2022, the Great Resignation has continued to pose a problem in organizations' ability to hold onto their employees, and has led to more jobs remaining unfilled. This problem in society has all come back to the basic concept of employee attrition which is "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" (S. Lucas, 2022). For the sake of this project, the focus will be on voluntary employee attrition.

Business Problem

With voluntary resignation, it can have a negative impact on the resigning employee's company due to lack of continuity within the organization, training gaps, lack of institutional knowledge, possible burnout for remaining employees, etc. (Markovich, 2019). Self-sufficient and high-performing teams are pivotal to a company's success, and with increased turnover at a company, it can pose a great danger to the company's ability to perform to their expectations. Therefore, by being able to predict the employees that are at risk for attrition and identify the critical factors in the employee's decision, it can put the given companies in a more proactive vs. reactive position.

Data Explanation

The dataset chosen for this project was sourced from Kaggle. The data is completely made-up, but it is meant to imitate 1,470 employees working at International Business Machines Corporation (IBM), which is an "American multinational technology corporation" (Wikimedia Foundation, 2022). Each record in the dataset pertains to a distinct employee, and contains information on the employee's

demographical and occupational information. There is also a clearly defined target variable, 'Attrition', which is a binary categorical variable that specifies whether the given employee resigned from IBM. The following are examples of some of the thirty-five total features in the dataset:

- Age
- Gender
- Education Level
- Environment Satisfaction Score
- Job Satisfaction Score
- Distance From Home (in miles)
- Job Department
- Monthly Income
- Total Working Years
- Number of Companies Worked

In terms of data cleansing, there were no missing values or duplicates in the dataset. There also were not any variables which were highly correlated with one another. The non-numerical attributes were also converted to numerical types using one-hot encoding before modeling was performed.

Once the variables were explored at length, multiple transformations occurred to derive more meaning from the attributes. In the corporate world, most of the discussion around employee pay occurs at the salary level. Therefore, since the original dataset contained a 'MonthlyIncome' variable, I decided to derive a 'YearlyIncome' variable by multiplying the monthly values by twelve to determine the employees' salaries. For the variable 'NumberOfCompaniesWorked', there were employees who had an assigned value of zero; I subsetted the original dataset to remove these employees with zero companies worked, since I wanted to focus on the employees who were full-time and also had prior career experience. Along with this subsetting of the dataset, multiple columns were removed throughout the data exploration process since they were deemed as not being relevant. Some of these columns included: 'DailyRate', 'HourlyRate', 'MonthlyRate', 'EducationField', 'Over18', 'OverTime', 'EmployeeCount', and 'StandardHours'.

Methods

Once the dataset was cleansed and transformed, various forms of exploration were conducted to gain a better understanding of the data especially in relation to the target variable. In assessing the 'Attrition' variable's class distribution, it became apparent that the classes were highly imbalanced in their representation of the employees, as visualized in Appendix A. The value 'No' had ten times as many records represented in the dataset compared to the value 'Yes', and since 'Yes' was our main value of interest, this imbalance represented a concern for the modeling portion of the project. Therefore, to handle this before modeling, I used a concept called Random Over-Sampling which entails adding more records to the minority class. The strategy helped to ensure that bias was not introduced towards the majority class during the predictive modeling steps.

Besides investigating how to predict whether an employee will leave their organization, I also wanted to determine the factors which were important in the employees' decisions. According to an article from Forbes titled *Employees Say Unsustainable Workloads and Expectations Are Driving Them To Quit* by Emmy Lucas, "Unsustainable workloads are one of the top factors contributing to the Great Resignation, along with uncaring managers, inadequate compensation and lack of career advancement potential, according to a new survey by consulting firm McKinsey & Company" (E. Lucas, 2022). In the dataset, there were four variables which pertained to these mentioned factors: 'JobSatisfaction', 'YearsWithCurrentManager', 'YearlyIncome' and 'YearsSinceLastPromotion'. I split the dataset based on the target values of 'Yes' and 'No'; then, I created bar charts for the groups with these variables to compare the differences and determine if any were in alignment with the Forbes' article findings. These bar charts are shown in Appendix B.

The variables where I noticed a significant difference between the two groupings was for job satisfaction and yearly income. For employees who did quit, their job satisfaction scores mostly leaned towards falling into the 1 and 3 levels. Whereas for the employees who did not quit, they scored their job satisfaction at the highest scores of 3 and 4. With the lowest-valued score having the second-highest value count for the resigned employees, it did raise a flag to the validity of the concern from the Forbes article. This data showed that the employees in the 'Yes' group were either dissatisfied or relatively satisfied

which are uneasy places to stand. For yearly income, the distributions of the variable were left-skewed for the two groups of employees; however, for the employees who did resign, they were mainly represented at the lower end of the yearly income scale and barely had any representation past ~ $125,000. The employees who did not quit had representation across the entire scale from $25,000 to $230,000, and had much higher counts at the higher values than the other group. Pay is the most important factor to an employee's lifestyle, and if an employee is not being paid enough to take care of their means, then it would make sense that these lower-paying employees would leave to pursue more optimal financial situations.

For model selection, several models which are appropriate for a classification model were used to evaluate performance accuracy.The models evaluated included the following:

- K-Nearest Neighbor - Straight forward pattern recognition model which allows the testing of several k values and leaf sizes to determine the best performance	- Random Forest Classifier - Expands beyond a decision tree by constructing multiple decision trees to remediate forcing a binary decision
- 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

Analysis

At first, the models were performed on all of the twenty-one features in the dataset, post-data exploration. The model performance was then evaluated based on the models' accuracies and confusion matrices. Precision and recall can be derived from a confusion matrix; recall helps to determine the ability to find all relevant instances of a class in a data set, whereas precision is used to determine the proportion of data points that the model says exists within the relevant class were indeed relevant. All accuracy, precision and recall results can be viewed in Appendix C.

The k-Nearest Neighbor model had the strongest performance with 84.48% accuracy; it also had the second-highest recall score for the 'No' class of Attrition at 97.7%. For the 'Yes' class of Attrition, it

also had the highest precision score of 58.3% which was relatively higher than the scoring of the other models. This measure showed that 58.3% of the employees who were predicted as resigning were correctly predicted. Given the count imbalance in the target classes, I think these scores are relatively good and that the models were able to compensate for the lack of original data for the resigned employees.

The second set of models only included the features mentioned above from the Forbes article: Job Satisfaction, Yearly Income, Years With Current Manager and Years Since Last Promotion. The scores for the four models differed quite extensively in their performance; in terms of accuracy, the Naive Bayes model had the highest measure at 83.19%. However, when it came to precision and recall for the 'Yes' target variable, the Decision Tree Classifier model was the best performer with precision and recall scores of 64.3% and 21.4% respectively which is slightly lower than the precision and recall values from the original model with all features.

I also generated models only for the two features which seemed to have significant differences between the two groups: Job Satisfaction and Yearly Income. All models performed at a lower ability in terms of precision and recall in comparison to the model with four features, giving some indication that job satisfaction and yearly income have more predictive influence when combined with other features.

Assumptions, Limitations & Challenges

Throughout the steps conducted for this project, I assumed that the employees recorded in the dataset were full-time employees who worked a standard of nine hours. Full-time employees are usually the main focus of an organization, since they are the consistent workers in the environment. I also assumed that although the data was made-up, it still replicated the employees at an organization in a relatively close manner. Therefore, with this assumption, I made choices in the project that would pertain to why an employee would leave an organization, which hopefully is what the dataset set out to accomplish as well.

The two major limitations in this project were the small dataset size and the major class imbalance of the target variable. These points were both limitations and challenges, in that they caused

problems in truly being able to compare the distinct groups of employees. While some differences were observed, it was difficult to know if they were true differences or just biased data, and this uncertainty showed in the model performances as well since the results were quite inconsistent and difficult to rank. I hoped for the ability to make clear conclusions from the dataset, but in the end, it was a challenge to drive honesty from the data.

Future Uses & Recommendations

The power of this predictive modeling for the Great Resignation has the potential to make great strides for this organizational issue; it will allow companies to stay ahead of their employees and ensure that they are being helped. Organizations can start to utilize their human resources data to build models and deploy them throughout departments to gauge trends and put initiatives in place that will put employees at the top of companies' minds. These initiatives could entail weekly manager check-ins, mental health days, consistent job opportunities, etc. The models can also entail tagging employees with a "risk score" which will put a numerical measure onto how likely it is for an employee to quit their career. An implementation plan can also be found in Appendix D to understand how these future uses and recommendations could be rolled out at an organization.

Ethical Assessment

Overall, I think utilizing employee data and tagging employees that are likely to quit their careers has major ethical implications. First, employee data can contain personally-identifiable information which is private to the employee and can pose a high risk to the company if the data were to be leaked. The data being used for the modeling portion of the project needs to be well-managed and well-transformed to ensure that no critical risk data is in the clear for anybody to access. Also, in terms of false positives that may arise from the models, it could be incredibly disadvantageous if an employee is approached about wanting to resign when they don't actually want to. The results of the models need to be taken with a grain of salt, and they may require extra evaluation to ensure that all employees identified are actually at risk. With employees, and people in general, the data on them needs to be carefully assessed to ensure that it is not crossing any ethical boundaries that would put an organization at danger of breaking any laws.

Conclusion

While it may be grimm, according to an article from the Gettysburg College titled *One third of your life is spent at work*, "the average person will spend 90,000 hours at work over a lifetime" (Gettysburg College, n.d.). If people are going to spend a majority of their lifetime as an employee, then it makes sense that they would want that time to be well-spent and somewhat-enjoyed.

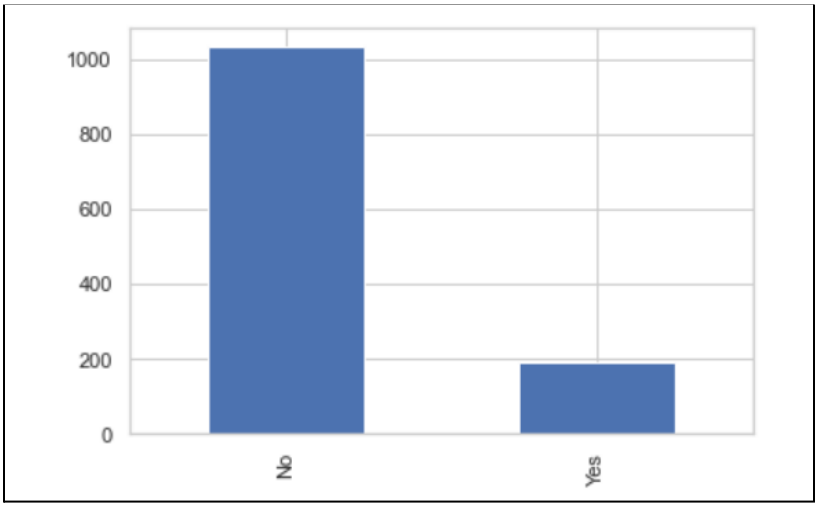
The dataset for this project conveyed that there are many features that factor into an employee's decision to leave their job; two of the important ones being job satisfaction level and yearly income. Employees want to enjoy their jobs, and they also want to be paid enough to compensate for the time, energy and effort that they put into their work. If these needs are not being met, then it is highly likely that an employee will choose to leave their organization rather than stay and feel dissatisfied.

With the use of the Random Forest model or the K-Nearest Neighbor model, the features mentioned above (along with others) can even be utilized from a predictive perspective to identify at-risk employees ahead of time and potentially prevent them from resigning. The results from these predictions could help organizations save money on staff replacement and productivity loss, and also hopefully decrease turnover rates. Employees will be happier; teams will be happier, and organizations will be happier that they have satisfied and motivated employees who want to show up to work everyday and get things done!

References

1. Lisa, A. (2021, December 7). *The number of Americans quitting their jobs in 2021 - and will it continue to grow?* Yahoo! Retrieved March 30, 2022, from https://www.yahoo.com/video/number-americans-quitting-jobs-2021-160334757.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAAHj25VF2fhDSsUBmlAlVGcIXkL4YAMD3eIn9yXHKWs3xKFQPEScQKnEQOujUv9oKgCTeq1fu8hRe2ns3eoQqsHioe_ImiR1OTKbwn2T7fSov0oUTP4yQ9F0sAd2RIA1M1A2vAdr3XCdnMp1mBq2XkZa9LIjodlXmubfZWMI5oOQ#:~:text=The%20Great%20Resignation%20started%20in,started%20keeping%20records%20in%202000.
2. Lucas, S. (2022, March 1). *Employee attrition: All you need to know*. AIHR. Retrieved March 30, 2022, from https://www.digitalhrtech.com/employee-attrition/.
3. Markovich, M. (2019, February 4). *The negative impacts of a high turnover rate*. Small Business - Chron.com. Retrieved March 30, 2022, from https://smallbusiness.chron.com/negative-impacts-high-turnover-rate-20269.html
4. Wikimedia Foundation. (2022, March 27). *IBM*. Wikipedia. Retrieved March 30, 2022, from https://en.wikipedia.org/wiki/IBM
5. Lucas, E. (2022, March 9). *Employees say unsustainable workloads and expectations are driving them to quit*. Forbes. Retrieved March 30, 2022, from https://www.forbes.com/sites/emmylucas/2022/03/09/employees-say-unsustainable-workloads-and-expectations-are-driving-them-to-quit/?sh=5880b2d57c34
6. Gettysburg College. (n.d.). *One third of your life is spent at work*. Gettysburg College. Retrieved March 30, 2022, from https://www.gettysburg.edu/news/stories?id=79db7b34-630c-4f49-ad32-4ab9ea48e72b

Appendix A - Target Class Distribution




```
No     1032
Yes     190
Name: Attrition, dtype: int64
```

Figure 1: Distribution of the Target Variable ‘Attrition’

Appendix B - Feature Importance Visualizations

Figure 2: Bar Chart of Job Satisfaction Scores For Resigned Employees

Figure 3: Bar Chart of Job Satisfaction Scores For Non-Resigned Employees

A. Years With Current Manager

Figure 4: Bar Chart of Years With Current Manager for Resigned Employees

Figure 5: Bar Chart of Years With Current Manager for Non-Resigned Employees

3. Yearly Income

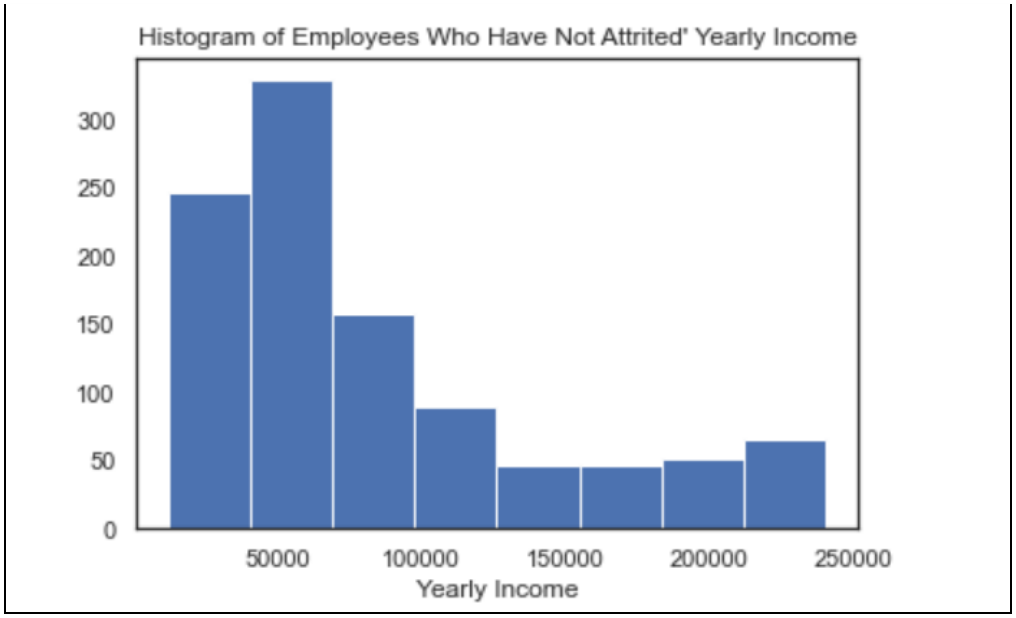


Figure 6: Histogram of Yearly Income For Non-Resigned Employees

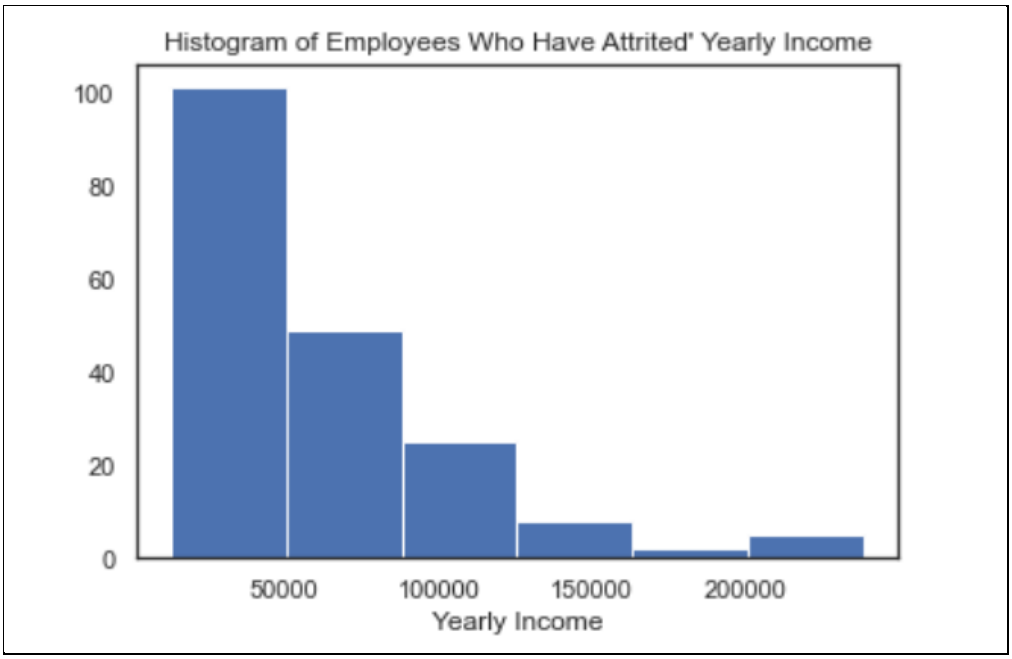


Figure 7: Histogram of Yearly Income for Resigned Employees

4. Years Since Last Promotion

Figure 8: Bar Chart of Years Since Last Promotion for Resigned Employees

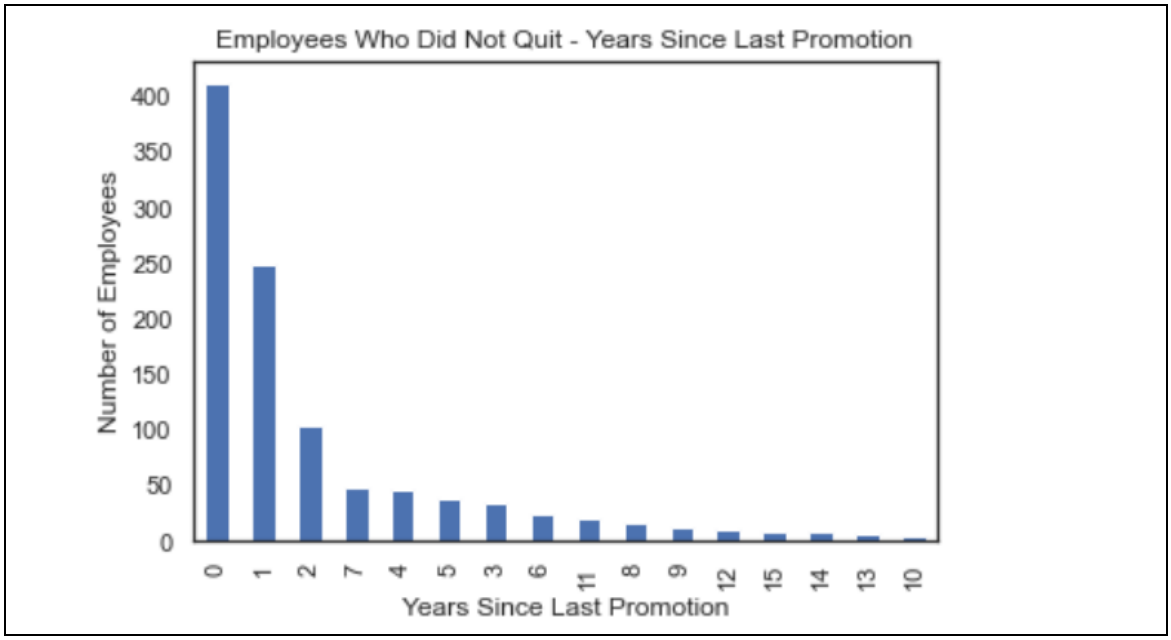


Figure 9: Bar Chart of Years Since Last Promotion for Non-Resigned Employees

Appendix C: Modeling Evaluation Results

```
---------------------------------------------------------------------------------------------------
KNN Results
Best Accuracy : 0.8310231023102309
Best Parameters : {'weights': 'uniform', 'n_neighbors': 21, 'leaf_size': 25, 'algorithm': 'kd_tree'}
Best Estimator : KNeighborsClassifier(algorithm='kd_tree', leaf_size=25, n_neighbors=21)
---------------------------------------------------------------------------------------------------
Naive Bayes Results
Mean Accuracy : 0.7386915162104446
---------------------------------------------------------------------------------------------------
Decision Tree Classifier Results
Best Accuracy : 0.8408755581440497
Best Parameters : {'max_features': 'sqrt', 'min_samples_leaf': 18, 'min_samples_split': 13}
Best Estimator : DecisionTreeClassifier(max_features='sqrt', min_samples_leaf=18,
                       min_samples_split=13, random_state=42)
---------------------------------------------------------------------------------------------------
Random Forest Classification Results
Best Accuracy : 0.8448130976528543
Best Parameters : {'bootstrap': True, 'criterion': 'gini', 'max_depth': None, 'max_features': 17, 'min_samples_leaf': 1, 'min_s
amples_split': 3}
Best Estimator : RandomForestClassifier(max_features=17, min_samples_split=3, random_state=42)
---------------------------------------------------------------------------------------------------
```

Figure 10: Four Models & Twenty-One Features - Evaluation Results (Accuracy)

```
---------------------------------------------------------------------------------------------------
KNN Confusion Matrix
[[210   3]
 [ 39   3]]
Precision:  [0.84337349 0.5       ]
Recall:  [0.98591549 0.07142857]
---------------------------------------------------------------------------------------------------
Gaussian Naive Bayes Confusion Matrix
[[162  51]
 [ 21  21]]
Precision:  [0.8852459  0.29166667]
Recall:  [0.76056338 0.5       ]
---------------------------------------------------------------------------------------------------
Decision Tree Confusion Matrix
[[199  14]
 [ 36   6]]
Precision:  [0.84680851 0.3       ]
Recall:  [0.9342723  0.14285714]
---------------------------------------------------------------------------------------------------
Random Forest Confusion Matrix
[[208   5]
 [ 35   7]]
Precision:  [0.85596708 0.58333333]
Recall:  [0.97652582 0.16666667]
---------------------------------------------------------------------------------------------------
```

Figure 11: Four Models & Twenty-One Features - Evaluation Results (Precision & Recall)

```
---------------------------------------------------------------------------------------------------
KNN Results - Reduced Features
Best Accuracy : 0.8300524170064065
Best Parameters : {'weights': 'uniform', 'n_neighbors': 23, 'leaf_size': 1, 'algorithm': 'ball_tree'}
Best Estimator : KNeighborsClassifier(algorithm='ball_tree', leaf_size=1, n_neighbors=23)
---------------------------------------------------------------------------------------------------
Naive Bayes Results - Reduced Features
Mean Accuracy : 0.8319020669291339
---------------------------------------------------------------------------------------------------
Decision Tree Classifier Results - Reduced Features
Best Accuracy : 0.83003300330033
Best Parameters : {'max_features': 'log2', 'min_samples_leaf': 9, 'min_samples_split': 7}
Best Estimator : DecisionTreeClassifier(max_features='log2', min_samples_leaf=9,
                       min_samples_split=7, random_state=42)
---------------------------------------------------------------------------------------------------
Random Forest Classification Results
Accuracy of Random Forest: 80.00%
---------------------------------------------------------------------------------------------------
```

Figure 12: Four Models & Four Features - Evaluation Results (Accuracy)

```
---------------------------------------------------------------------------------------------------
KNN Confusion Matrix
[[210   3]
 [ 39   3]]
Precision:  [0.84337349 0.5       ]
Recall:  [0.98591549 0.07142857]
---------------------------------------------------------------------------------------------------
Gaussian Naive Bayes Confusion Matrix
[[213   0]
 [ 42   0]]
Precision:  [0.83529412        nan]
Recall:  [1. 0.]
---------------------------------------------------------------------------------------------------
Decision Tree Confusion Matrix
[[208   5]
 [ 33   9]]
Precision:  [0.86307054 0.64285714]
Recall:  [0.97652582 0.21428571]
---------------------------------------------------------------------------------------------------
Random Forest Confusion Matrix
[[196  17]
 [ 34   8]]
Precision:  [0.85217391 0.32      ]
Recall:  [0.92018779 0.19047619]
---------------------------------------------------------------------------------------------------
```

Figure 13: Four Models & Four Features - Evaluation Results (Precision & Recall)

```
---------------------------------------------------------------------------------------------------
KNN Results - Reduced Features
Best Accuracy : 0.8300524170064065
Best Parameters : {'weights': 'uniform', 'n_neighbors': 23, 'leaf_size': 1, 'algorithm': 'ball_tree'}
Best Estimator : KNeighborsClassifier(algorithm='ball_tree', leaf_size=1, n_neighbors=23)
---------------------------------------------------------------------------------------------------
Naive Bayes Results - Reduced Features
Mean Accuracy : 0.8319020669291339
---------------------------------------------------------------------------------------------------
Decision Tree Classifier Results - Reduced Features
Best Accuracy : 0.8241409435061152
Best Parameters : {'max_features': 'log2', 'min_samples_leaf': 9, 'min_samples_split': 7}
Best Estimator : DecisionTreeClassifier(max_features='log2', min_samples_leaf=9,
                       min_samples_split=7, random_state=42)
---------------------------------------------------------------------------------------------------
Random Forest Classification Results
Accuracy of Random Forest: 74.90%
---------------------------------------------------------------------------------------------------
```

Figure 14: Four Models & Two Features - Evaluation Results (Accuracy)

```
--------------------------------------------------------------------------------------------------------
KNN Confusion Matrix
[[210   3]
 [ 39   3]]
Precision:  [0.84337349 0.5       ]
Recall:  [0.98591549 0.07142857]
--------------------------------------------------------------------------------------------------------
Gaussian Naive Bayes Confusion Matrix
[[213   0]
 [ 42   0]]
Precision:  [0.83529412        nan]
Recall:  [1. 0.]
--------------------------------------------------------------------------------------------------------
Decision Tree Confusion Matrix
[[206   7]
 [ 37   5]]
Precision:  [0.84773663 0.41666667]
Recall:  [0.96713615 0.11904762]
--------------------------------------------------------------------------------------------------------
Random Forest Confusion Matrix
[[180  33]
 [ 31  11]]
Precision:  [0.85308057 0.25      ]
Recall:  [0.84507042 0.26190476]
--------------------------------------------------------------------------------------------------------
```

Figure 15: Four Models & Two Features - Evaluation Results (Precision & Recall)

Appendix D: Implementation Plan

Step	**Who**	**What**
1.1	Stakeholders	Identify the business problem and the problem's impact on the organization
2.1	Managers & Product Owners	Develop a vision for the project and OKRs to set goals for the work
3.1	Data Scientists	Gather and prepare the employee data
3.2	Data Scientists	Define the target variable of the dataset for modeling
3.3	Data Scientists	Perform predictive modeling of the data on training set & identify important features
3.4	Data Scientists & Stakeholders	Assess modeling results and determine if any rework needs to be done in alignment with the business problem
4	Data Scientists & Stakeholders	Deploy model on test set of

		employees and validate findings
5	Human Resources	Based on model results, brainstorm and plan initiatives that could help with findings
6	Data Scientists	Continuously re-train the model to ensure it is up to standards and using new employee data
7	Organization	Roll out models on all employees and monitor its effect