**HR Analytics: An Employee Attrition Predictive Analysis**

*August 1st, 2024*

Hunter Berberich

# Executive Summary

In today's competitive business environment, retaining valuable employees is crucial for success. Losing a good employee can significantly impact a company's performance and morale. In his book on predictive analytics, Eric Siegel discusses a model Hewlett-Packard (HP) developed to predict employee attrition, assigning a "flight risk" score to each employee. Inspired by this, the current case study aims to develop a similar predictive model using a fictional dataset created by IBM. This dataset includes 1470 employees with 35 attributes, such as job satisfaction, work-life balance, and whether they left the company. The best-performing model was a logistic regression classification model. The final model accurately predicted 77% of the employees who left, which led to an estimated cost savings of over $4 million.

# Summary of Analysis

Talent retention is crucial to organizational success. The cost of employee attrition is approximately 1 to 1.2 times employee salary on an individual basis, costing the U.S. economy an estimated $27 billion in December 2016 alone (Marsden, 2016). Considering the unemployment rate in the United States is only 4.1%, vacancies may not always be filled promptly, even if a company is willing to endure the financial costs.

Predicting which employees quit in advance would be extremely valuable to any organization. To this end, two predictive models were developed using data from 1470 employees to assist human resources in identifying employees who are at the highest risk of leaving and target retention efforts accordingly.

The two model types used were a logistic regression model (Model 1) and a decision tree classification model (Model 2). Model 1 was refined using stepwise variable selection. Model 2 was refined using decision tree pruning techniques.

The objective for each model was to minimize false negatives because failing to predict an employee will quit (i.e., false negative) is far more costly than incorrectly predicting an employee will stay (i.e., false positive). Each model was tuned by adjusting the cut-off value to maximize the balanced accuracy.

The logistic regression model had an overall accuracy of 72.01% and accurately predicted 85.11% of the employees who left. The decision tree model had an overall accuracy of 73.37% and accurately predicted 61.70% of the employees who left.

The logistic regression model performed best. The final model performance and net cost savings are summarized in Table 1.

Table 1. Final Model Performance

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Type | Balanced Accuracy | Net Savings |
| 1 | Logistic Regression | 0.77 | $4,096,838 |
| 2 | Decision Tree | 0.69 | $2,926,312 |

Although it had a lower/similar sensitivity to the decision tree model, it predicted 85.11% of high-risk employees. The average employee salary for all employees was $78,053. Assuming an attrition cost of 150% of salary and a retention cost of 10%, the final model would have saved this hypothetical company approximately $4.1 million.

# Description of Analysis

## Data Preparation

The models were developed using a fictional dataset created by IBM that contains information (35 attributes) on 1,470 employees, including whether they left their positions (Attrition = Yes if they left, No otherwise). The first step was to import the data, encode variables, and drop variables with no predictive value. Once data transformation was complete, the data was then split, 80/20, into a training and validation set. Due to the class imbalance, special care was taken to ensure sufficient “Yes” values were present in both sets for proper analysis.

## Descriptive Statistics & Visualizations

This section provides an overview of the dataset and explores various features that may influence employee attrition. Key variables include relationship satisfaction, job satisfaction, years in the current role, monthly income, job role, and overtime. The visualizations help understand the factors contributing to employee attrition and identify potential areas for further investigation by human resources and upper management. By analyzing these variables, a more comprehensive predictive model for employee attrition can be developed.

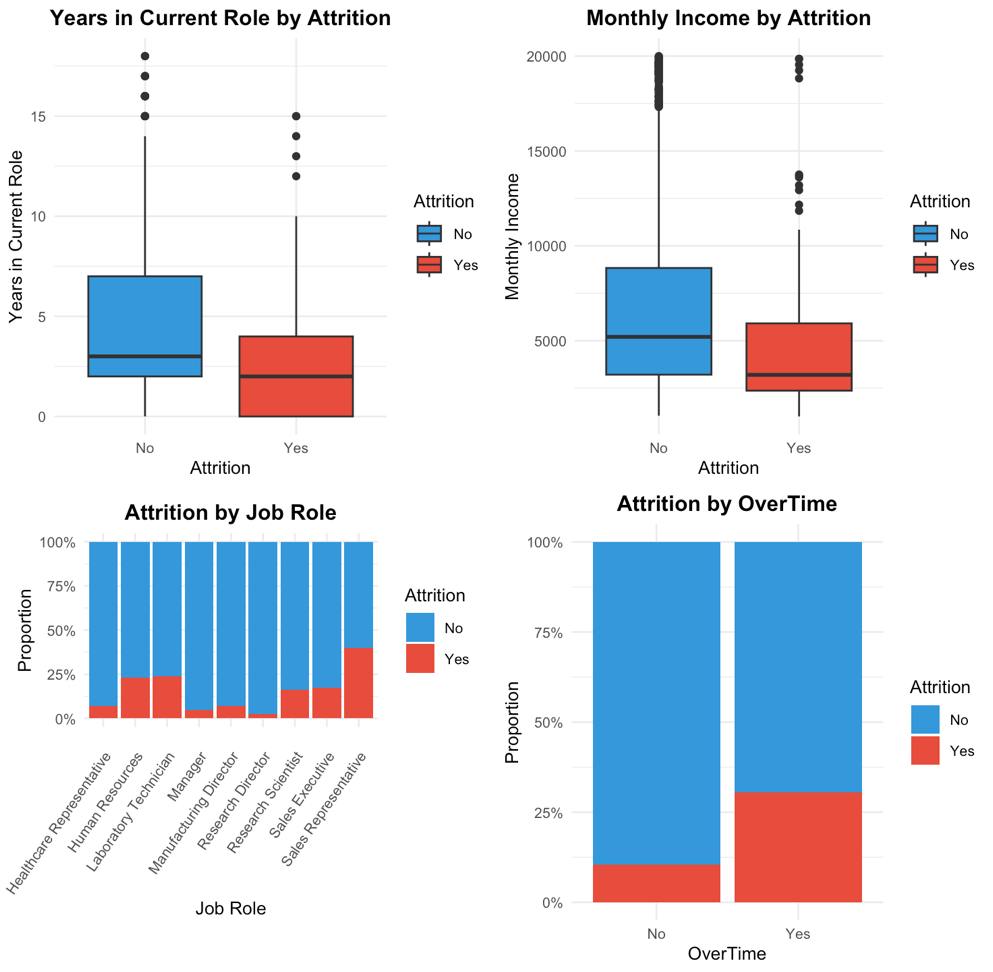


Figure 1: Attrition by Key Factors

Figure 1 includes box plots and bar plots showing the distribution of years in the current role and monthly income for employees who left the company (Yes) versus those who stayed (No). It also illustrates the proportion of employees who left versus those who stayed by their job roles and whether they worked overtime. Employees who left had shorter tenures, lower monthly incomes, and certain job roles and overtime work correlated with higher attrition rates.

A group of blue and red bars

Description automatically generated

Figure 2: Attrition by Key Factors Continued

Figure 2 includes bar plots showing the proportion of employees who left the company (Yes) versus those who stayed (No) based on their work-life balance, business travel frequency, job involvement, and job satisfaction. Employees who left generally had lower work-life balance, more frequent business travel, lower job involvement, and lower job satisfaction rates.

## Model 1 Development: Logistic Regression Model

Model 1 utilized logistic regression, a type of generalized linear model designed for classification tasks. This model outputs a probability ranging from 0 to 1, where 1 indicates a positive outcome, and 0 indicates a negative outcome. In the context of this case study, a probability of 1 represents an employee leaving the organization, while a probability of 0 represents an employee staying.

All variables were used to fit the initial model. Then, stepwise variable selection techniques were employed to remove variables based on the default AIC metric to avoid overfitting the model to the training data. The refined model used 39 coefficients compared to the initial model, which used

58 coefficients. The refined model performed well on the training set. The initial model AIC was 770.44 compared to the final model AIC of 748.64.

The default cut-off value for the logistic regression model used is 0.5. However, these default models suffered from a very low sensitivity, which is the objective metric for this application. The cut-off values were tuned to maximize the balanced accuracy, which is the average of sensitivity and specificity. The effect of the cut-off score on accuracy and balanced accuracy is shown in Figure 3

A graph with a red line and blue line

Description automatically generated

Figure 3: Effect of Cut-Off Values on Model Performance

As shown in Figure 3, the overall accuracy peaks near a cut-off of 0.55, but the balanced accuracy peaks near a cut-off of 0.1. Ultimately, the final cut-off score for Model 1 was 0.1.

The overall accuracy of the model on the test set was 71.01%. The sensitivity was 85.11%, meaning it predicted 40 of the 47 employees that left the organization. The full model performance metrics are discussed in more detail in the Model Performance and Selection section.

## Model 2 Development: Decision Tree Model

We developed a basic decision tree using all available variables with the default complexity parameter (cp), which served as our baseline for further analysis. This initial step was crucial for gaining a preliminary understanding of the data's structure and the potential influence of various factors on employee attrition.

We then expanded our analysis by fully developing the tree, allowing it to grow to its maximum depth without restrictions (cp=0). This method was instrumental in uncovering the full range of variables and interactions present in the dataset, although with an increased risk of overfitting. The complexity parameter table of the full tree was examined to determine the optimal cp value that minimizes cross-validated error, leading to the selection of the best-pruned tree model. The best cp value used for the final model was 0.01052.

The pruning process, critical for enhancing the model's generalizability, involves reducing the size of the full tree to avoid overfitting. The pruned tree was finalized by applying the optimal cp value obtained from the full tree's complexity parameter table. This model strikes a balance between complexity and predictive accuracy.

A comparison was also made with a tree pruned using cost complexity (with a predefined cp=0.01), showcasing an alternative approach to controlling the tree's growth and complexity.

All variables were used to fit the initial model. Then, stepwise variable selection techniques were employed to remove variables based on the default AIC metric to avoid overfitting the model to the training data. The final pruned tree had 23 nodes compared to the fully expanded tree, which used 43 nodes. The refined model performed well on the training set, with a balanced accuracy of 0.7104 compared to the fully expanded tree model, with a balanced accuracy of 0.7434 compared to the final model.

**Model Comparison and Performance Evaluation**

The effectiveness of both the basic and refined (pruned) models was measured against a testing dataset. This evaluation focused on various key metrics, such as accuracy, sensitivity, and specificity, which were instrumental in evaluating the models' performance.

A graph of a tree model performance

Description automatically generated

Figure 4. Decision Tree Model Performance Metrics

Figure 4 showcases the pruned decision tree model's performance in terms of overall accuracy and balanced accuracy. The overall accuracy peaks at the 0.15 threshold, while the balanced accuracy peaks at the 0.1 threshold, suggesting that the optimal threshold for the model is 0.1.

**Performance Insights**

The overall accuracy of the pruned decision tree model on the test set was 73.37%, with a sensitivity of 61.70%, meaning it accurately predicted 29 out of the 47 employees who left the organization. The model's performance metrics highlight its utility in accurately predicting employee attrition. These findings underscore the model's predictive strength and potential to serve as a reliable tool for identifying at-risk employees. Further details and specific figures are presented in Table XX, which outlines the model's performance across different metrics.

# Model Performance and Selection

As previously discussed, both models had similar overall accuracy ratings. Key performance metrics for both models are shown in Figure 3 for comparison.

A graph of different colored bars

Description automatically generated

Figure 5. Model Performance Comparison

As seen in Figure 5, Model 2’s accuracy and specificity performed marginally better than Model 1, but Model 1’s balanced accuracy was higher due to its superior sensitivity. To ensure the cut-off scores were optimized as intended, the cost savings for each model were estimated using the employee salary data and an assumed attrition-to-retention cost ratio of 15:1. The results from the net savings analysis for Model 1 are shown in Figure 4.

A graph showing different types of thresholds

Description automatically generated

Figure 6. Net Cost Savings vs. Cut-off Scores for Model 1.

As shown in Figure 6, the cut-off score used in the final model, 0.10, maximized the net cost savings. The net estimated cost savings from Model 1 was $4,096,838. For comparison, the results from the net savings analysis for Model 2 are shown in Figure 5.

A graph with a line graph

Description automatically generated with medium confidence

Figure 7. Net Cost Savings vs. Cut-off Scores for Model 2.

As shown in Figure 7, the cut-off score used in the final model, 0.10, maximized the net cost savings. The net estimated cost savings for Model 2 was $2,926,312.

# Conclusion

As demonstrated by this analysis, predicting employees who are leaving before they leave can lead to substantial cost savings for a company. Although implementing either model would save this company millions of dollars, the logistic regression model was the best-performing one. This model accurately predicted 40 of the 47 employees who left the company, leading to potential cost savings of $4.1 million. The company should adopt this model to help target its incentives for its retention program.

**References**

Bureau of Labor Statistics. (2024, July 5). *Employment Situation Summary—2024 Q02 Results*. Bureau of Labor Statistics. [https://www.bls.gov](https://www.bls.gov/news.release/empsit.nr0.htm)

Marsden, T. (2016). What is the true cost of attrition? *Strategic HR Review*, *15*(4), 189–190. <https://doi.org/10.1108/SHR-05-2016-0039>