**Predicting and Enhancing Employee Performance Through HR Analytics**

Cali Hugelen, Edith Tache, Jacob Brooks, Remington Webb

June 30, 2024

**Table of Contents**

# Introduction

# Business Problem/Hypothesis

# Methods/Analysis

# Results

# Recommendations/Ethical Considerations

# Conclusion

# References

# Appendix

# **Introduction**

In today's competitive business landscape, organizations strive to enhance productivity and retain top talent. Employee performance is a critical factor influencing organizational success, and optimizing this performance is a primary objective for Human Resources (HR) departments. Leveraging data analytics within HR, commonly referred to as HR analytics, provides a powerful means to achieve this objective. By systematically analyzing employee data, organizations can uncover insights into key performance drivers, optimize recruitment strategies, and evaluate the effectiveness of training programs.

The dataset includes various features that influence employee performance:

* **Employee ID:** Unique identifier for tracking performance.
* **Department:** The department in which the employee works.
* **Region:** Geographical location of the employee.
* **Education:** Employee's educational background.
* **Gender:** Gender of the employee.
* **Recruitment Channel:** Source of recruitment.
* **Number of Trainings:** Number of training programs attended.
* **Age:** Age of the employee.
* **Previous Year Rating:** The performance rating of the prior year.
* **Length of Service:** Number of years the employee has worked in the company.
* **KPIs Met More Than 80:** Number of key performance indicators met.
* **Awards Won:** Number of awards won.
* **Average Training Score:** Average score from training programs.

**Source and Preprocessing:** The dataset was sourced from Kaggle and titled "Employee's Performance for HR Analytics." Before analysis, the dataset underwent preprocessing steps to handle missing values. Summary statistics, unique value counts, and a review of each unique value in the columns were conducted to better understand the dataset and check for any data value errors.

**Specific Goals of the Project**

The specific goals of this project are to:

1. Identify the key predictors of employee performance.
2. Assess how different recruitment channels impact employee performance.
3. Evaluate the effectiveness of training programs on employee performance.
4. Develop a predictive model to forecast employee performance based on identified features.
5. Provide actionable recommendations for HR strategies to improve employee performance and retention.

By focusing on these goals, the project aims to help organizations improve their hiring processes, enhance employee development, and ultimately boost overall workforce productivity and retention.

# **Business Problem/Hypothesis**

The core of our project revolves around several fundamental research questions aimed at addressing the business problem of optimizing employee performance through HR analytics. Specifically, we seek to understand:

1. What are the key predictors of employee performance?
2. How does the recruitment channel affect employee performance?
3. What is the impact of training programs on employee performance?

By focusing on these questions, we aim to improve hiring processes, enhance employee development, and ultimately boost overall workforce productivity and retention. Additionally, we seek to build a model that will help the organization effectively predict employee performance.

**Hypothesis**

Based on our preliminary research and analysis of the dataset, we hypothesize that:

1. Key indicators such as KPIs met, previous year ratings and average training scores are significant predictors of employee performance.
2. The recruitment channel through which an employee is hired has a measurable impact on their performance.
3. Participation in training programs positively influences employee performance.

These hypotheses will guide our analysis and model development as we work to identify the critical factors that influence employee performance and provide actionable recommendations for HR strategies.

# **Methods/Analysis**

**Data Loading and Preparation**

The data was loaded into the Google Colab environment for analysis. Initial data exploration involved handling missing values, with categorical variables filled using the mode and numerical variables filled using the median. Summary statistics, unique value counts, and a review of each unique value in the columns were conducted to better understand the dataset and check for any data value errors.

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis was conducted to comprehend the distributions and relationships within the data. Key visualizations included:

* **Outliers:** Outliers were identified and analyzed using box plots for each numerical variable. Upon further investigation, it was determined that the outliers present in the dataset were within the realm of normal variation and did not negatively impact the data's integrity or reliability. Therefore, these outliers were retained for further analysis.
* **Employee Performance Level by Recruitment Channel:** A bar plot revealed how employee performance varied across different recruitment channels, indicating the channels through which high-performing employees were hired.
* **Employee Performance Level by Number of Trainings:** A bar plot illustrated the relationship between the number of training programs attended and employee performance, highlighting the impact of training on performance.

**Statistical Tests**

ANOVA tests were conducted to assess the impact of recruitment channels and training programs on employee performance. The results indicate significant differences among the groups (p < 0.001), suggesting that both recruitment channels and training programs influence performance outcomes.



**Feature Engineering**

To prepare the data for machine learning models, categorical variables were transformed into numerical values using label encoding for binary variables, frequency encoding for high cardinality variables, and one-hot encoding for other categorical variables.

**Correlation Analysis**

Correlation analysis was conducted to identify the relationships between various features and the target variable, "KPIs met more than 80". This analysis helped in understanding which features were most strongly associated with meeting KPIs more than 80%. The top three features with the highest correlation were previous\_year\_rating, awards\_won, and department\_operations. A bar plot and a heatmap were also used to visualize these correlations.

A screenshot of a computer program

Description automatically generated

**Model Training and Evaluation**

**Model Selection**

Several machine learning models were chosen for this analysis based on their strengths and suitability for the task:

* **Logistic Regression:** Chosen for its simplicity and interpretability, logistic regression is effective for binary classification tasks and provides insights into the importance of different features.
* **Random Forest:** Selected for its ability to handle complex interactions between features and its robustness to overfitting. Random forests can provide feature importance scores, which are valuable for understanding the impact of different variables.
* **Gradient Boosting:** Chosen for its high performance in predictive tasks, gradient boosting combines multiple weak learners to create a strong predictive model. It is particularly effective in handling non-linear relationships and interactions between features.
* **Support Vector Machine (SVM):** Selected for its effectiveness in high-dimensional spaces and its ability to handle both linear and non-linear classification tasks.

**Hyperparameter Tuning**

The Gradient Boosting Model underwent hyperparameter tuning to optimize its performance. The tuning process involved using GridSearchCV to explore a range of hyperparameters and identify the best combination for the model. The following hyperparameters were tuned:

* **learning\_rate:** Controls the contribution of each tree to the final model.
* **max\_depth:** Limits the depth of the individual trees to prevent overfitting.
* **min\_samples\_leaf:** Specifies the minimum number of samples required to be at a leaf node.
* **min\_samples\_split:** Specifies the minimum number of samples required to split an internal node.
* **n\_estimators:** The number of trees in the ensemble.
* **subsample:** The fraction of samples used for fitting the individual base learners.

The best parameters found were:

* **learning\_rate:** 0.038573365384388155
* **max\_depth:** 5
* **min\_samples\_leaf:** 2
* **min\_samples\_split:** 6
* **n\_estimators:** 101
* **subsample:** 0.944399754453365

This tuning process involved fitting 3 folds for each of 50 candidates, totaling 150 fits, to ensure robust performance and avoid overfitting. The tuned Gradient Boosting model achieved an accuracy of 0.718, indicating its effectiveness in predicting employee performance.

# **Results**

**Model Performance**

The performance of several machine learning models was evaluated to predict employee performance. The models included Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Each model's performance was assessed using accuracy, precision, recall, and F1 score.

**Interpretation of Model Performance Metrics**

* **Accuracy:** The proportion of true results (both true positives and true negatives) among the total number of cases examined. It indicates the overall effectiveness of the model in correctly classifying employee performance. However, accuracy alone can be misleading if the classes are imbalanced.
* **Precision:** The ratio of true positive predictions to the total predicted positives. High precision indicates a low false positive rate, which is important for ensuring that employees predicted to meet performance standards actually do.
* **Recall (Sensitivity):** The ratio of true positive predictions to the total actual positives. High recall is crucial for capturing as many true positive cases as possible, which helps identify all employees who meet performance standards.
* **F1 Score:** The harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when the classes are imbalanced, as it considers both false positives and false negatives.

**Comparative Analysis Models**

1. **Logistic Regression:**
   * **Accuracy:** 0.715
   * **Class 0 (Not Met KPIs):** Precision: 0.74, Recall: 0.87, F1 Score: 0.80
   * **Class 1 (Met KPIs):** Precision: 0.64, Recall: 0.43, F1 Score: 0.52
   * **Macro Average:** Precision: 0.69, Recall: 0.65, F1 Score: 0.66
   * **Weighted Average:** Precision: 0.70, Recall: 0.71, F1 Score: 0.70
2. **Random Forest:**
   * **Accuracy:** 0.707
   * **Class 0:** Precision: 0.73, Recall: 0.86, F1 Score: 0.79
   * **Class 1:** Precision: 0.63, Recall: 0.42, F1 Score: 0.50
   * **Macro Average:** Precision: 0.68, Recall: 0.64, F1 Score: 0.65
   * **Weighted Average:** Precision: 0.70, Recall: 0.71, F1 Score: 0.69
3. **Gradient Boosting:**
   * **Accuracy:** 0.716 (improved to 0.718 after hyperparameter tuning)
   * **Class 0:** Precision: 0.73, Recall: 0.89, F1 Score: 0.80
   * **Class 1:** Precision: 0.66, Recall: 0.41, F1 Score: 0.50
   * **Macro Average:** Precision: 0.70, Recall: 0.64, F1 Score: 0.65
   * **Weighted Average:** Precision: 0.71, Recall: 0.72, F1 Score: 0.70
4. **Support Vector Machine (SVM):**
   * **Accuracy:** 0.708
   * **Class 0:** Precision: 0.72, Recall: 0.91, F1 Score: 0.80
   * **Class 1:** Precision: 0.67, Recall: 0.34, F1 Score: 0.45
   * **Macro Average:** Precision: 0.70, Recall: 0.62, F1 Score: 0.63
   * **Weighted Average:** Precision: 0.70, Recall: 0.71, F1 Score: 0.68

**Justification for Selecting Gradient Boosting Model**

The Gradient Boosting model was selected as the best performer due to its higher accuracy and balanced performance across precision, recall, and F1 score compared to the other models. Specifically:

* **Accuracy Improvement:** The accuracy of the Gradient Boosting model (0.718 after tuning) was the highest among all models, indicating it correctly classified the most cases overall.
* **Precision and Recall Balance:** The precision and recall for class 0 (employees not meeting KPIs) were notably high, suggesting the model correctly identified underperforming employees. For class 1 (employees meeting KPIs), although the recall was lower, the precision was relatively better, indicating fewer false positives.
* **F1 Score:** The F1 score for both classes was higher in the Gradient Boosting model than in the other models, indicating a better balance between precision and recall, which is crucial for maintaining both the identification of high performers and minimizing misclassification of low performers.

Overall, the Gradient Boosting model's superior performance in these metrics makes it the most reliable choice for predicting employee performance, helping the organization make informed decisions based on accurate and balanced predictions.

**Feature Importance**

The feature importance analysis from the Tuned Gradient Boosting model provides insights into the key predictors of employee performance. These findings help answer our research question: "What are the key predictors of employee performance?" The analysis shows that previous year rating, average training score, and length of service are the top 3 predictors of employee performance, providing valuable insights for future performance improvement strategies.

A graph with blue and white text

Description automatically generated

Figure 1 Feature Importance from Tuned Gradient Boosting

**Measure of Model Performance Accuracy – ROC Curve**

The ROC (Receiver Operating Characteristic) curve for the Gradient Boosting model provides insight into the trade-offs between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various threshold settings. Here is what we can interpret from the ROC curve.

* The ROC curve is above the diagonal (red dashed line representing random chance), indicating that the model performs better than random guessing.
* The ROC curve and AUC value of 0.74 suggest that the Gradient Boosting model has a decent performance in predicting the target variable.

This analysis demonstrates that the Gradient Boosting model is effective in distinguishing between positive and negative classes, making it a reliable model for predicting employee performance.

A graph of a curve

Description automatically generated

**Figure 2: ROC Curve**

**Employee Performance by Recruitment Channel**

This analysis addresses our research question: How does the recruitment channel affect employee performance?

**Performance Distribution**:

* The **'Other'** recruitment channel has the highest number of employees and shows a substantial number of both high and low performers.
* The **'Sourcing'** channel shows a higher number of low performers compared to high performers.
* The **'Referred'** channel has the smallest group of employees, with an almost equal number of high and low performers.

These observations highlight the impact of different recruitment channels on employee performance. The 'Other' channel appears to be the most common, encompassing a broad range of performance outcomes. The 'Sourcing' channel tends to have more low performers, while the 'Referred' channel maintains a balanced distribution of high and low performers.

A graph of a bar graph

Description automatically generated with medium confidence

Figure 2: Employee Performance By Recruitment Channel

**Employee Performance by Training Programs**

This analysis addresses our third research question: What is the impact of training programs on employee performance?

**General Trend:**

* Most employees attended only one training session.
* As the number of training programs increases, the total number of employees decreases significantly.
* In all training categories, there are consistently more employees who did not meet their KPIs compared to those who did.

These findings suggest that while training is an essential aspect of employee development, merely increasing the number of training sessions does not necessarily lead to better performance outcomes. Further analysis may be needed to evaluate the quality and relevance of the training programs to improve their effectiveness. A graph with numbers and a number of trainings

Description automatically generated

Figure 3: Employee Performance by Number of Trainings

# Recommendations/Ethical Considerations

**Enhance Training Programs**

Based on the analysis, it was observed that the number of training sessions attended does not directly correlate with improved performance. Therefore, it is essential to focus on training programs' quality and relevance rather than quantity. Here are some concrete steps to enhance training programs:

1. **Tailored Training Programs:** Develop training programs that address specific skills and knowledge gaps identified through performance data. For example, if data analysis skills are crucial for high performance in a particular department, design targeted workshops to improve these skills.
2. **Interactive and Practical Sessions:** Incorporate interactive elements such as hands-on workshops, case studies, and real-world problem-solving sessions to make the training more engaging and practical.
3. **Continuous Learning Opportunities:** Provide ongoing learning opportunities through e-learning platforms, webinars, and online courses to enable employees to continuously develop their skills at their own pace.
4. **Mentorship and Coaching:** Implement mentorship and coaching programs where experienced employees can guide and support less experienced colleagues, fostering a culture of continuous improvement and knowledge sharing.
5. **Feedback Mechanisms:** Establish regular feedback mechanisms to gather input from employees on the effectiveness of training programs and make necessary adjustments based on their feedback.

**Optimize Recruitment Channels**

The analysis indicated varying performance outcomes associated with different recruitment channels. Here are strategies to optimize these channels:

1. **Strengthen 'Other' and 'Referred' Channels:** Since these channels show a balanced distribution of high and low performers, efforts should be made to strengthen them. This could include offering referral bonuses, leveraging employee networks, and enhancing employer branding to attract high-quality candidates.
2. **Improve 'Sourcing' Channel:** Given that the 'Sourcing' channel currently has a higher number of low performers, consider revising the sourcing strategies. This could involve more rigorous screening processes, better alignment of job descriptions with required skills, and using data-driven approaches to identify top talent.

**Leverage Key Performance Predictors**

Utilize the insights gained from the feature importance analysis to focus on the key predictors of employee performance. Here are some specific strategies:

1. **Enhance Performance Reviews:** Focus on improving the accuracy and effectiveness of performance reviews, as previous year ratings are a significant predictor. Implement structured performance review processes that provide clear, actionable feedback.
2. **Targeted Development Programs:** Develop targeted interventions and development programs to improve average training scores and length of service. This could include personalized development plans, career progression pathways, and recognition programs for long-serving employees.

**Performance Monitoring and Feedback**

Establish regular performance monitoring and feedback mechanisms to provide employees with timely and constructive feedback. This can help identify areas of improvement and implement corrective measures promptly. Here are some specific steps:

1. **Regular Check-ins:** Schedule regular one-on-one check-ins between managers and employees to discuss performance, set goals, and provide feedback.
2. **360-Degree Feedback:** Implement 360-degree feedback systems where employees receive feedback from peers, subordinates, and supervisors, providing a comprehensive view of their performance.
3. **Performance Dashboards:** Create performance dashboards that allow employees to track their progress and see how they are performing against set KPIs in real-time.

**Limitations and Future Research**

While the study provides valuable insights, there are potential limitations that should be addressed in future research:

1. **Data Quality and Completeness:** The analysis is based on the available dataset, which may have limitations in terms of data quality and completeness. Future research could focus on obtaining more comprehensive and high-quality data to improve the accuracy of the analysis.
2. **Generalizability:** The findings of this study are based on a specific dataset and may not be generalizable to all organizations. Future research could involve replicating the study with data from different industries and organizational contexts to validate the findings.
3. **Longitudinal Analysis:** This study provides a snapshot of employee performance at a given time. Future research could conduct longitudinal analyses to understand how employee performance evolves over time and the long-term impact of different HR strategies.

By addressing these limitations and implementing the recommended strategies, organizations can enhance employee performance, improve productivity, and foster a positive and equitable work environment.

# **Conclusion**

**Summary of Key Findings**

This study aimed to predict and enhance employee performance through HR analytics by leveraging a comprehensive dataset to uncover key performance drivers and evaluate the effectiveness of various HR strategies. The key findings of the study are as follows:

1. **Significant Predictors of Performance:** Previous year ratings, average training scores, and length of service were identified as the top three predictors of employee performance.
2. **Impact of Recruitment Channels:** Different recruitment channels showed varying impacts on performance, with the 'Other' and 'Referred' channels having a more balanced distribution of high and low performers, while the 'Sourcing' channel had more low performers.
3. **Effectiveness of Training Programs:** The number of training sessions attended did not directly correlate with improved performance. This highlights the importance of focusing on training programs' quality and relevance rather than quantity.

**Practical Applications**

The findings of this study provide actionable insights that can be implemented within organizations to enhance employee performance:

1. **Optimizing Recruitment Strategies:**
   * Strengthen and optimize recruitment channels that have shown better performance outcomes, such as 'Other' and 'Referred' channels.
   * Improve sourcing strategies by incorporating more rigorous screening processes and aligning job descriptions with required skills.
2. **Enhancing Training Programs:**
   * Develop tailored training programs that address specific skills and knowledge gaps identified through performance data.
   * Incorporate interactive and practical elements into training sessions to make them more engaging and effective.
   * Implement continuous learning opportunities through e-learning platforms, webinars, and online courses.
3. **Leveraging Key Performance Predictors:**
   * Focus on improving the accuracy and effectiveness of performance reviews, as previous year ratings are a significant predictor.
   * Develop targeted interventions and development programs to enhance average training scores and length of service.
4. **Performance Monitoring and Feedback:**
   * Establish regular performance monitoring and feedback mechanisms, including regular check-ins, 360-degree feedback systems, and performance dashboards.
5. **Implementing the Predictive Model:**

* Utilize the Gradient Boosting model to predict employee performance. This model can help HR professionals identify employees who are likely to perform well and those who may need additional support or training.
* Integrate the predictive model into the HR analytics system to provide real-time insights and forecasts. This can assist in making data-driven decisions regarding hiring, training, and employee development.
* Regularly update and retrain the model with new data to ensure its predictions remain accurate and relevant.

Organizations can enhance employee performance, improve productivity, and foster a positive and equitable work environment by implementing these strategies. This study demonstrates the value of HR analytics in guiding strategic decisions and highlights the importance of data-driven HR strategies in achieving organizational success and employee satisfaction. Future research should address identified limitations and explore additional features to further improve predictive accuracy and effectiveness.

# **References**

1. Kaggle. (n.d.). Employee's performance for HR analytics. Kaggle. [Employee’s Performance for HR Analytics (kaggle.com)](https://www.kaggle.com/datasets/sanjanchaudhari/employees-performance-for-hr-analytics)
2. Fitz-enz, J. (2010). The new HR analytics: Predicting the economic value of your company's human capital investments. AMACOM.
3. Pease, G., Byerly, B., & Fitz-enz, J. (2013). Human capital analytics: How to harness the potential of your organization's greatest asset. Wiley.
4. Davenport, T. H., Harris, J., & Shapiro, J. (2010). Competing on talent analytics. Harvard Business Review. <https://hbr.org/2010/10/competing-on-talent-analytics>
5. Bassi, L., & McMurrer, D. (2007). Maximizing your return on people. Harvard Business Review. <https://hbr.org/2007/03/maximizing-your-return-on-people>
6. Marr, B. (2018). Data-driven HR: How to use analytics and metrics to drive performance. Kogan Page. [1012042669.pdf (gbv.de)](https://www.gbv.de/dms/zbw/1012042669.pdf)
7. Lawler, E. E., Levenson, A., & Boudreau, J. W. (2004). HR metrics and analytics: Use and impact. Center for Effective Organizations. [HR Metrics and Analytics: Use and Impact - ProQuest](https://www.proquest.com/docview/224584705?sourcetype=Trade%20Journals)
8. Levenson, A. (2018). Using workforce analytics to improve strategy execution. IBM Center for The Business of Government. [Using workforce analytics to improve strategy execution - Levenson - 2018 - Human Resource Management - Wiley Online Library](https://onlinelibrary.wiley.com/doi/abs/10.1002/hrm.21850)
9. Rousseau, D. M., & Barends, E. (2011). Becoming an evidence-based HR practitioner. Human Resource Management Journal. [Becoming an evidencebased HR practitioner (cebma.org)](https://cebma.org/assets/Uploads/Rousseau-Barends-HRMJ-2011.pdf)
10. Siddiqi, M. A. (2024). Employee engagement and organizational performance: A human resource perspective. ResearchGate. [(PDF) Employee Engagement And Organizational Performance: A Human Resource Perspective (researchgate.net)](https://www.researchgate.net/publication/380284402_Employee_Engagement_And_Organizational_Performance_A_Human_Resource_Perspective)

# **Appendix**

# Anova TEst results

1. ANOVA result for recruitment channels: F\_onewayResult(statistic=21.5218030534083, p-value=4.620961379081536e-10)
2. ANOVA result for training programs: F\_onewayResult(statistic=5.446805540239299, p-value=6.98698897007903e-07)

CORRELATION MATRIX HEATMAP

A graph of a graph

Description automatically generated with medium confidence