Employee Performance Prediction

Taruna Aggarwal
Master of Business Administration
Galgotias University, Greater Noida, Uttar Pradesh, India
Email: tarunaaggarwal54@gmail.com

ABSTRACT

In an increasingly data-driven corporate environment, employee performance has emerged as a central determinant of organizational success, influencing not only operational efficiency but also long-term strategic positioning. As traditional performance management systems face criticism for being subjective, reactive, and inconsistent, organizations are shifting towards predictive analytics to gain proactive, evidence-based insights into workforce behavior and outcomes. This study explores the use of machine learning techniques to predict employee performance based on publicly available review data. Specifically, it utilizes a structured dataset comprising 20,995 anonymized employee reviews from Capgemini, sourced from Kaggle. The dataset captures employee sentiment across multiple workplace dimensions, including career growth, skill development, work satisfaction, salary and benefits, job security, and work-life balance. These self-reported ratings serve as predictors for the overall performance rating, which is used as the target variable. The research employs a predictive research design, with a focused application of a single supervised learning model—Decision Tree Classifier. The decision to use this model is based on its interpretability, computational efficiency, and strong predictive performance. Prior to modeling, the dataset underwent data cleaning, transformation, class labeling (High vs. Low performers), and exploratory data analysis. The Decision Tree model was trained on 70% of the dataset and validated on the remaining 30%. Model performance was evaluated using standard classification metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. The model achieved an accuracy of 84.9% and an F1-score of 88.8%, indicating a robust ability to classify employee performance levels. Feature importance rankings revealed that career growth was the most influential predictor, followed by work satisfaction, salary and benefits, and skill development. Factors such as job type and job security were found to have relatively minimal influence. The findings have strong managerial implications. HR teams can leverage such predictive insights to identify high-potential talent, design targeted upskilling programs, and address early signs of disengagement. Importantly, the explainability of the Decision Tree model supports its integration into organizational decision-making processes, offering transparent justifications for employee-level classifications. While the study is limited by the absence of demographic and internal performance metrics, it demonstrates the feasibility and value of using open-source review data for predictive HR analytics. Future research could enhance the model by incorporating textual sentiment analysis and expanding the dataset to include multi-company comparisons. In conclusion, this thesis presents a practical and scalable framework for forecasting employee performance using machine learning. It reinforces the growing role of predictive analytics in human resource management and contributes actionable insights that can help organizations build and sustain high-performing workforces..

Keywords: employee performance, decision tree model

I. INTRODUCTION

In today's intensely competitive and digitally interconnected marketplace, organizations are increasingly judged not only by the quality of the products and services they deliver but also by the caliber and engagement of the people who deliver them. Employee performance—encompassing productivity, innovation, adaptability, and discretionary effort—has therefore moved to the center of strategic discourse in boardrooms across the globe. Human-resource (HR) leaders who once relied primarily on annual performance reviews and subjective managerial judgment are now turning to data analytics for sharper, real-time insights. The accelerating adoption of cloud-based HR information systems, applicant-tracking tools, collaboration platforms, and social-feedback channels has created unprecedented volumes of structured and unstructured data. Harnessing these data streams to predict how employees will perform, and why, promises to transform talent management from a reactive, intuition-driven function into a proactive, evidence-based discipline.

Against this backdrop, the present thesis—“Employee Performance Prediction Using Data Analytics”—investigates whether ratings drawn from public employee-review portals can be modeled to forecast overall performance sentiment. Capgemini, a global leader in consulting, technology services, and digital transformation, provides an ideal context for such an inquiry. The firm operates in more than 50 countries, employs over 300,000 professionals, and is consistently ranked among the most attractive employers in the IT services industry. The Capgemini Employee Reviews dataset sourced from Kaggle offers 20,995 real-world observations, each containing an overall rating together with granular assessments of work-life balance, skill development, salary and benefits, job security, career growth, and work satisfaction. Although the reviews are external and self-reported, their scale and richness make them a valuable proxy for gauging organizational climate and individual engagement levels.

The Strategic Centrality of Human Capital

Employee performance is universally recognized as a cornerstone of organizational success. It directly influences innovation, service quality, customer satisfaction, and financial outcomes. As global markets become more competitive and talent shortages intensify—such as the projected shortfall of 85 million workers by 2030 reported by Korn

Ferry—organizations are compelled to maximize the value derived from their human capital. In this context, understanding the predictors of employee performance becomes not only a human resource concern but a strategic imperative.

Capgemini as a Case Context

Capgemini is a global leader in consulting, digital transformation, and technology services. With more than 300,000 employees spread across 50+ countries and annual revenues exceeding €22 billion, Capgemini operates in a high-performance, project-based environment. This decentralized structure creates unique challenges in assessing and managing employee performance. The reliance on cross-functional, geographically dispersed teams increases the need for scalable, data-driven tools that can monitor and predict workforce effectiveness in real-time.

The Capgemini Employee Review dataset—sourced from the Kaggle open data platform—provides a valuable foundation for performance analysis. It consists of nearly 21,000 anonymized reviews, each evaluating various facets of the employee experience such as work-life balance, skill development, salary and benefits, job security, career growth, and overall job satisfaction. These reviews offer a rare glimpse into employee perceptions and allow for empirical modeling of performance sentiment.

II. LITERATURE REVIEW

The literature on employee performance prediction intersects multiple disciplines, including human resource management, data science, and behavioral analytics. As businesses transition from intuition-based decision-making to data-driven strategies, a growing body of research has emerged to explore how predictive models can be used to evaluate and enhance employee performance. This section synthesizes key theoretical and empirical contributions that form the foundation of this study

Evolution of HR Analytics

Human Resource (HR) Analytics has evolved significantly over the last two decades. Initially, HR decisions were based on subjective appraisals, annual performance reviews, and managerial intuition. However, the introduction of cloud-based HR information systems and advanced analytics tools has enabled organizations to shift from reactive talent management to predictive and prescriptive HR practices (Minbaeva, 2018).

Predictive Modeling in Employee Performance

Machine learning algorithms have been widely used to classify and forecast employee outcomes such as attrition, promotion readiness, and performance levels. Research by Sharma & Singh (2020) and Jain & Kaur (2021) confirms that models such as Decision Trees, Logistic Regression, Random Forests, and Support Vector Machines (SVM) can accurately predict performance-related outcomes using employee-level data. These models analyze historical data to recognize patterns that can anticipate future behavior.

- Key Drivers of Performance
- Various studies have identified critical factors influencing employee performance.

These include training frequency, career development opportunities, managerial feedback, work-life balance, and organizational culture (Campbell, 1990; Law, 2018). For example, continuous learning programs and real-time feedback mechanisms have been associated with increased productivity and engagement (Noe et al., 2017).

Furthermore, employee perception of career growth has been repeatedly cited as a strong motivator and performance enhancer, particularly in knowledge-intensive industries like IT and consulting. When employees see clear paths for advancement, they are more likely to exhibit commitment and high-quality output (Herzberg, 1966)

Review-Based and Sentiment Data in HR

While many traditional studies rely on internal HR metrics such as key performance indicators (KPIs), project delivery scores, or attendance records, recent research has begun to explore publicly available employee review data as a proxy for performance sentiment. Platforms such as Glassdoor, Indeed, and Comparably provide open access to employee feedback, which, when structured and analyzed, can offer insights into performance perception.

Pang and Lee (2008) pioneered sentiment analysis techniques that have since been adapted for HR contexts, including performance prediction. However, the academic use of structured review ratings (rather than free-text sentiment) remains underexplored. This study addresses this gap by applying machine learning on quantitative review-based features to predict overall performance classification.

Theoretical Frameworks Guiding Performance Prediction

This study is anchored in three well-established theories:

- Human Capital Theory (Becker, 1964): Suggests that employees are valuable assets whose performance improves with investment in skills and knowledge.
- Resource-Based View (RBV) (Barney, 1991): Proposes that high-performing talent is a strategic resource that can offer sustained competitive advantage if nurtured properly.
- Social Exchange Theory (Blau, 1964): Highlights the reciprocal relationship between employer support (e.g., career growth, feedback) and employee performance.

Together, these theories support the notion that organizational inputs (learning, recognition, environment) are key drivers of employee effectiveness, which can be captured and predicted through data analytics.

III. RESEARCH METHODOLOGY

Research Strategy and Design

This study follows a predictive research design, which focuses on using existing data to forecast outcomes or classify categories. In the context of this project, the goal is to predict whether an employee is a high or low performer based on key workplace experience ratings. Predictive research is suitable when the objective is to generate actionable insights from historical data using statistical or machine learning models.

Unlike exploratory or causal research, which may seek to discover unknown patterns or prove theoretical relationships, predictive research aims to build a data-driven model that can estimate future or unobserved results. This approach is particularly relevant for HR analytics, where timely identification of performance trends can inform decision-making around hiring, training, and talent retention..

Selected Technique: Decision Tree Classifier

To implement this predictive approach, the study uses a Decision Tree Classifier—a supervised learning algorithm ideal for classification tasks. It splits the dataset into branches based on the most informative input variables, enabling the model to predict whether an employee is likely to fall into a “High” or “Low” performance category.

Why Decision Tree?

- Interpretability: Decision Trees generate simple “if-then” rules that HR managers and non-technical stakeholders can easily understand.
- Accuracy: In this study, the Decision Tree achieved an accuracy of 84.9% and an F1-score of 88.8%—strong results for practical application.
- Feature Insight: The model identifies the most influential factors (e.g., career growth, work satisfaction), helping organizations prioritize interventions.
- Efficiency: The method is computationally inexpensive and does not require feature scaling or complex preprocessing.

IV. RECOMMENDATIONS

Based on the research findings and model outcomes, the following recommendations are proposed for HR leaders and managers:

- Enhance Career Growth Opportunities
  - The most influential factor, career growth, should be prioritized.
  - Establish clear promotion paths, regular performance feedback, and leadership grooming initiatives.

- Focus on Skill Development & Learning
  - Regular training and upskilling opportunities contribute directly to employee satisfaction and high performance.
  - Introduce personalized learning journeys aligned with individual and organizational goals.

- Improve Work Satisfaction through Culture
  - Focus on building a supportive work environment and strong team culture.
  - Conduct periodic engagement surveys and take action on employee feedback.

- Use Data-Driven Performance Management
  - Adopt predictive analytics in HR practices to proactively identify high-potential employees and those needing intervention.
  - Integrate such models with internal performance management systems.

- Target Low Performers Strategically
  - Use employee feedback to understand dissatisfaction among low performers.

Implement coaching, reskilling, or departmental transfers based on root-cause analysis.

- Tie Rewards with Development, Not Just Tenure
  - Move away from tenure-based rewards to merit and development-linked incentives.
  - This approach not only motivates but retains high-performing talent.

V. CONCLUSION

This study effectively demonstrates the power of data analytics in understanding and predicting employee performance patterns based on workplace experience ratings. By analysing 20,000+ employee reviews from Capgemini, the research identified key factors influencing overall performance perceptions.

Among the tested models—Logistic Regression, Decision Tree, Random Forest, and K- Nearest Neighbors—the Decision Tree model stood out with the highest accuracy (84.9%) and F1 score (88.8%), while also offering clear interpretability. The findings underscore the significant impact of Career Growth, Work Satisfaction, and Skill Development on how employees rate their performance and workplace experience.

Although constrained by limitations such as a lack of internal HR data, demographic details, and potential bias from public reviews, the project confirms that even with limited data, organisations can extract valuable insights using machine learning.

This project aimed to predict employee performance using data analytics techniques applied to employee review data. Through a systematic process involving data preprocessing, exploratory data analysis, model building, and evaluation, several insights and patterns were uncovered that are valuable for organizations seeking to improve workforce performance.

The distribution analysis of overall ratings revealed that a majority of employees rated their experiences positively (4 or 5 out of 5), suggesting general satisfaction but also indicating opportunities for targeted improvement among the lower-rated segments.

Several machine learning models—Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN)—were applied and evaluated based on their accuracy. All models demonstrated comparable performance, with Decision Tree slightly outperforming others, making it the most suitable model for this dataset due to its interpretability and performance.

The confusion matrix of the Decision Tree model indicated a high rate of correct classifications, especially for high-performing employees, though some misclassifications persisted. This suggests the model is effective but can be further improved with more diverse or granular features.

A key outcome of the project was identifying the most influential factors driving employee performance. The feature importance analysis revealed that:

- Career growth opportunities were the strongest predictor of employee performance,
- Followed by work satisfaction, salary and benefits, and skill development,
- While factors like job type, job security, and work-life balance had minimal predictive value in this model.

These findings underscore the importance of growth-oriented HR strategies and career development initiatives in enhancing employee performance.

VI. REFERENCES

Pustokhina, I. V. (2019). Predicting employee turnover: A study on decision trees and logistic regression models. Journal of Human Resource Management, 22(4), 112-124.

Sharma, S., & Singh, A. (2020). Predicting employee attrition using machine learning techniques. International Journal of Data Science and Machine Learning, 15(2), 35-48.

Law, B. G. (2018). The impact of organizational culture and job satisfaction on employee turnover. Journal of Business and Management Studies, 18(1), 56-63.

Jain, A. & Kaur, M. (2021). Predictive analytics in performance evaluation using decision trees. International Journal of Business Analytics.