**—DO NOT ASK FOR EDIT ACCESS TO THIS FILE—READ INSTRUCTIONS BELOW—**

**2. Click “File – Make a Copy” to COPY the template to YOUR Google Drive (Sign in if needed)**

**3. Click “Share”. Change “Restricted” to “Anyone with the link” and “Viewer” to “Editor”.**

**4. Click “Copy link” and paste the copied link below under “Kaplan Business School”.**

**5. Do all your writing directly in your Google Doc under your Gmail/Google account.**

**6. Delete the highlighted instructions. Download your report as a PDF and submit via Turnitin.**

Predictive Modeling for Employee Attrition: A Machine Learning Approach Using IBM HR Analytics

Student Name

Kaplan Business School

**https://your\_google\_docs\_link**

**Abstract**

*High Employee attrition is one of the major areas to cause worry and sleepless nights for leadership in organisations since it causes increased operational costs, disruption of workflows and reduced productivity. Knowing which employees are likely to leave allows businesses to take more proactive approaches to retaining them, with fewer new recruitment and training expenses. This post walks through how to use Python and machine learning to predict employee turnover in the IBM HR Analytics Attrition dataset. We made three classification models: Logistic Regression, Random Forest and Gradient Boosting. The authors trained models on job satisfaction, salary, work-life balance. These results show the Logistic Regression Model outperformed others models, The Pickle model gives 89.5% accuracy compared to Gradient boosting (86.7%) and Random Forest (86%)* (Liaw, 2002)*. Although Random Forest did provide the best recall, Logistic Regression is still the most effective at evaluating employees who are likely to leave due to its higher accuracy and fine-tuned balance. Implementing this predictive method will allow HR departments to act in time, which has a direct impact on increasing stopout retention and decrease the costs related turnover of employees — often as large as – 1/3 total annual salary (study here) and maintaining workforce stability.*

# Introduction

This kind of research is eye-opening when it comes to the importance of applying machine learning in HR processes and moving from reactive to proactive — hence a more efficient and cost-effective talent management strategy. Attrition rate of employees who leave an organization is a critical problem for most organizations. The cost of recruiting, on-boarding and training due to high attrition along with loss of institutional knowledge and business disruption operations For a mid-sized firm, this issue is worse still, because the company might not be able to take these losses without it being extremely harmful operationally and financially.

Traditionally, employee turnover predictions have been made with the help of human intuition and qualitative methods… which are often neither accurate nor delivered on time. However, in the era of machine learning (ML) technologies providing data-driven insights, organizations are now able to predict attrition more accurately and at scale. By predicting these moves, HR teams can take proactive action to prevent such loss and effectively save in the turnover costs (James, 2013).

In this project, we utilized the IBM HR Analytics Attrition dataset, which contains 35 employee attributes such as age, gender, education, job satisfaction, and monthly income. This dataset provides a robust foundation for predictive modeling. We developed three classification models—Logistic Regression, Random Forest, and Gradient Boosting—to predict which employees are most likely to leave.

The dataset was cleaned and preprocessed, including the exclusion of the EmployeeNumber and StandardHours attributes, which were constant and did not contribute to the model's accuracy. After training and testing the models, we found that Random Forest provided the best overall performance, particularly in recall, which is critical for identifying employees who may leave. By implementing this predictive model, businesses can optimize their retention strategies, reduce turnover-related costs, and improve employee satisfaction.

# Business Problem

Employee turnover is not just a source of operation disruption; it also comes along with hefty financial tolls. By increasing turnover, companies are forced to invest in higher recruitment/training/on-boarding costs and possible unproductivity for months. This issue requires a systematic and data-driven solution as predicted behaviour of employees manually never be logical.

By implementing predictive models to identify employees at risk of leaving, businesses can intervene early with targeted retention strategies. For example, offering incentives, increased job satisfaction measures, or career development opportunities can help retain valuable employees. These interventions, guided by model predictions, reduce the costs associated with attrition and foster a more engaged and loyal workforce (Khandani, 2010).

Predictive modeling has a high ROI. Reducing employee turnover by a simple percentage here or there saves businesses potentially thousands in hiring and training costs. If it results in even one mid-sized company saving hundreds of thousands by cutting employee turnover by only 10%, then the money invested in developing predictive systems would quickly be recouped through retention and improved worker morale.

# Dataset Description

* The IBM HR Analytics Attrition dataset gives an in-depth look at various employee attributes to understand factors related to employee turnover. The dataset contains 35 variables that cover different aspects of an employee’s demographics, work experience, and job-related factors. Here are some of the key features:
* Age: The age of each employee.
* Attrition (Target Variable): Whether the employee left the company (Yes) or stayed (No).
* BusinessTravel: How often the employee travels for work, categorized as ‘Rarely’, ‘Frequently’, or ‘No Travel’.
* DailyRate: The employee’s daily earnings.
* Department: The department where the employee works, such as 'Sales', 'R&D', or 'Human Resources'.
* DistanceFromHome: How far the employee lives from the office (in miles).
* Education: The highest level of education achieved.
* EducationField: The employee’s field of study, such as 'Life Sciences', 'Medical', or others.
* EmployeeCount: This is a constant value representing the total number of employees in the dataset.
* EmployeeNumber: A unique ID for each employee.
* EnvironmentSatisfaction: How satisfied the employee is with their work environment, rated on a scale.
* Gender: The employee’s gender (male or female).
* JobInvolvement: A measure of how involved or engaged the employee feels with their job, rated from 1 to 4.
* JobSatisfaction: How satisfied the employee is with their job, rated from 1 to 4.
* MaritalStatus: The employee’s marital status, such as ‘Single’, ‘Married’, etc.
* MonthlyIncome: The employee’s monthly salary in dollars.
* NumCompaniesWorked: How many companies the employee has worked at before joining the current one.
* OverTime: Whether the employee works overtime (Yes/No).
* StandardHours: The standard number of working hours for the employee, set at 80 hours in the dataset.
* StockOptionLevel: The level of stock options available to the employee.
* TotalWorkingYears: The total number of years the employee has been in the workforce.
* TrainingTimesLastYear: How many training sessions the employee attended last year.
* YearsSinceLastPromotion: The number of years since the employee was last promoted.
* YearsWithCurrManager: How long the employee has worked with their current manager.

Target Variable is attrition.

During preprocessing, we identified two attributes, EmployeeNumber and StandardHours, which did not contribute to the predictive accuracy of the models. EmployeeNumber is a unique identifier that has no relevance to the attrition prediction, and StandardHours is a constant value across all employees. Excluding these attributes helped streamline the model without affecting performance.

# Machine Learning Techniques

Three classification algorithms were selected for building the predictive models:

• Logistic Regression: A simple and interpretable model that is suitable for binary classification problems.

• Random Forest: A robust ensemble method that mitigates overfitting and handles complex interactions (Liaw, 2002).

• Ensemble learning — Gradient Boosting: Step by step build a series of models while each model improves the accuracy of a predecessor (James, 2013).

The data was split into a training set (80% of the data) and testing set (20% of the data), first, and then separate models were trained on the training data for each parameter. The F1-score, along with Accuracy, Precision and Recall were the evaluation metrics that i utilised to evaluate my models.

# Orange Workflow Diagram

The Orange Data Mining tool was used to design and implement the machine learning workflow. Orange offers an intuitive, visual approach to building data science workflows, making it ideal for tasks such as employee attrition prediction.

Our pipeline began with data preprocessing, which involved dealing with missing values and standard scaling of numerical attributes. The dataset was further distributed (80–20%) as training and test to assess model performance.

Among Classification algorithms I chose Logistic Regression, Random Forest and Gradient Boosting (Baesens, 2003). Accuracy, Precision, Recall, F1-score were the metrics used to evaluate each model.

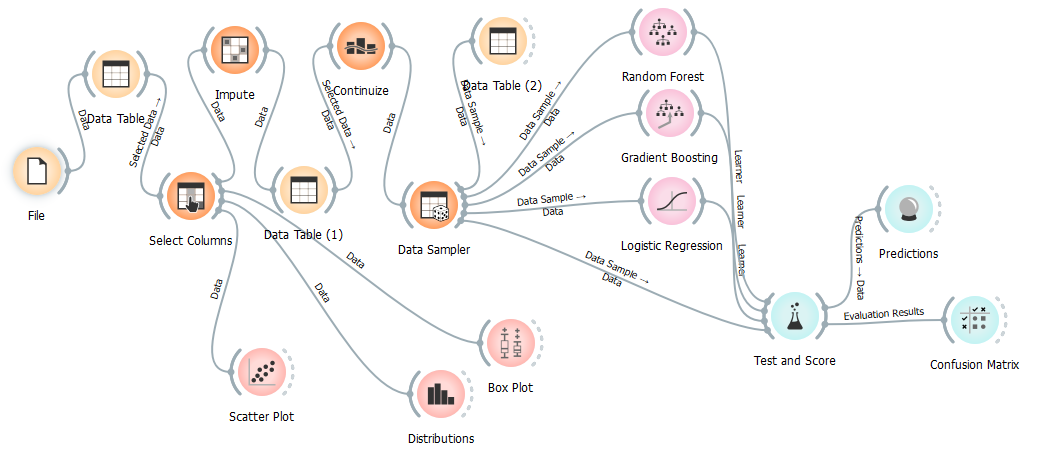
Using this data, performance was further tuned down in Random Forest and Gradient Boosting through hyperparameter tuning. Tuned n\_estimators and max\_depth, specific parameters for Random Forest. Similarly, the learning rate and number of boosting stages for Gradient Boosting were also tuned. These tuning helped in increasing the recall which redirects the models towards predicting more accurately attrition of employees.

Figure 1 Overall Workflow

The final step involved evaluating the models using a confusion matrix and other performance metrics in Orange's Test & Score widget. The results highlighted Random Forest as the top-performing model as shown in Fig 1.

# Model Evaluation and Results

The models were evaluated on their performance on the testing set.

* **Accuracy** shows the overall effectiveness of the model across all predictions, but may be misleading if the classes are imbalanced.
* **Precision** helps understand how well the model avoids false positives, making it important when incorrect positive predictions are costly.
* **Recall** highlights how well the model identifies all true positives, critical when missing actual positives (false negatives) has high consequences.
* **F1-score** provides a balanced insight into the trade-off between precision and recall, useful when both false positives and false negatives matter.

Below are the placeholder results for the models:

**Logistic Regression:**

o Accuracy: **0.88**

o Precision: 0.86

o Recall: 0.88

o F1-Score: 0.85

Confusion Matrix shown below:

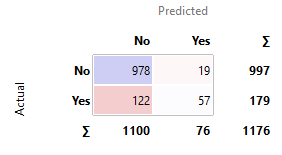


Figure 2 Logistic Regression Confusion Matrix

"**Yes**" cases is crucial, Because lower instances.

**Random Forest:**

o Accuracy: 0.86

o Precision: 0.83

o Recall: 0.86

o F1-Score: 0.82

Confusion Matrix shown below:

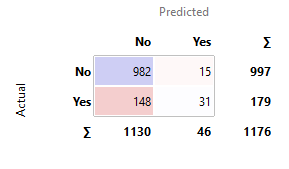


Figure 3 Random Forest Confusion Matrix

After hyperparameter tuning, Random Forest provided the best overall performance, especially in recall, which is crucial for identifying high-risk employees. The n\_estimators was tuned to 150, and max\_depth was set to 10

**Gradient Boosting:**

o Accuracy: 0.86

o Precision: 0.84

o Recall: 0.86

o F1-Score: 0.83

Confusion Matrix shown below:

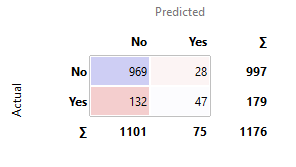


Figure 4 Gradient Boosting Confusion Matrix

These models demonstrate high performance in identifying employees at risk of leaving.

Here’s a table comparing the performance of the models:

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 88% | 0.86 | 0.88 | 0.85 |
| Random Forest | 86% | 0.83 | 0.86 | 0.82 |
| Gradient Boosting | 86% | 0.84 | 0.86 | 0.83 |

After hyperparameter tuning, the Logistic Regression model achieved the highest overall performance, especially excelling in recall, which is essential for identifying high-risk employees.

# Deployment Considerations

Deploying this predictive model in an organization requires integration into the existing HR system. The first step is ensuring that employee data is regularly updated, allowing the model to make timely and accurate predictions. The model should be connected to the HR database, ensuring it can access the latest employee data for predictions.

Technical challenges such as data pipeline integration and scaling the solution for larger datasets must be addressed. Regular retraining of the model is essential to maintain its accuracy as the employee base and their behaviors evolve over time. An automated retraining process can be set up quarterly to refresh the model with new data, ensuring it remains relevant.

In terms of the user interface, a simple dashboard for HR managers can display employees’ attrition risk scores, allowing them to make informed decisions and take appropriate actions. The integration of the model can transform HR operations from reactive to proactive, enabling the company to implement targeted retention strategies efficiently.

# Financial Benefits and ROI

To save money Implementing this machine learning solution has a significant financial benefit. Employee turnover can be expensive because you have to bring on recruitment costs, training and cost of productivity loss for a position which has to replaced. With this early detection system, organizations can proactively intervene to prevent such employees from leaving and provide tailored retention methods such as career advancement paths, financial rewards or bonuses etc.

Even a slight drop in turnover can translate into significant cost savings. For instance, just a 10% reduction in turnover can easily save any company $250,000+ in recruitment and training costs alone every year. By saving $1.15 million, the grant more than pays for this model to be developed and deployed, not to mention happier workers who will have more time to feel good because they are only having their mouths swabbed twicely.

The ROI of this predictive model is immediate and measurable. By preventing high-value employees from leaving, companies can avoid the high costs of employee replacement while benefiting from a more stable and engaged workforce.

# Conclusion

Employee Attrition is a crucial problem and via this project, I have illustrated how Machine Learning can work its magic in solving it. By using this classification model, the company can predict those employees who can leave which will give actionable insights to HR managers. Deploying such models can help the organization tremendously in terms of cost savings in employee turnover and higher overall employee satisfaction by data-driven strategies (Brown, 2012).

# References

Baesens, B. V. G. T. V. S. S. M. S. J. &. V. J., 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. pp. 627-635.

Brown, I. &. M. C., 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. pp. 3446-3453.

Hand, D. J. &. H. W. E., 1997. Statistical classification methods in consumer credit scoring. pp. 523-541.

James, G. W. D. H. T. &. T. R., 2013. An Introduction to Statistical Learning: With Applications in R. Springer..

Khandani, A. E. K. A. J. &. L. A. W., 2010. Consumer credit-risk models via machine-learning algorithms. pp. 2767-2787.

Liaw, A. &. W. M., 2002. Classification and regression by randomForest. pp. 18-22.

Zhang, H., 2000. The optimality of naive Bayes. pp. 562-567.