**1. Introduction**

In today's fast-paced, rapidly growing, and highly competitive business world, human capital is perhaps one of the most valuable assets a company might have. Workers directly influence productivity, culture, and overall success of a company. Even with that, numerous businesses find it challenging to retain their workers. When a worker leaves, not only is workflow dynamics disrupted, but it also results in substantial costs associated with replacing them, training, and bringing them onboard. Due to this, predicting employee turnover that is, the likelihood an employee will depart from the company has emerged as a leading priority for firms of all sizes.

Previously, companies have relied on manual methods like exit interviews and surveys to establish why employees leave. Though these can provide some information, they are too late to prevent turnover and are limited in scope. A silver lining to this issue is that the advent of data science and machine learning offers a new way to solve this issue. By analyzing past data about employees, we can search for patterns to forecast whether an existing employee will be prone to leave. Therefore, organizations can actively try to improve the happiness of employees, their engagement, and retention.

This project revolves around using machine learning models to research a dataset which contains information regarding employees. The dataset contains fields such as the level of education, gender, age, city, number of years in current field, and whether or not the employee was ever benched (i.e., not placed on a project for a stretch of time). All of that information helps us build a more solid concept of what could potentially make someone leave.

The general goal of this project is to create accurate and reliable predictive models from this employee data. These models will enable us to estimate the value of a target variable called LeaveOrNot. This column in the data indicates whether an employee has left the company (value 1) or not (value 0). We want to create models that can analyze all the other features of an employee and predict this outcome as accurately as possible.

By properly predicting employee attrition, firms can cut costs, ease turnover, and create a healthful working environment. Moreover, the identification of reasons underlying employee attrition helps management in decision-making for workplace policies, employee engagement, and career advancement opportunities. With this, evidence-based decisions may not only prove to be advantageous to individual employees but can also aid towards organizational long-term success.

Over the course of this report, I will walk you through each phase of the machine learning process, from understanding datasets to model creation and comparison. By the final page, we will have a good sense of how well machine learning can help predict employee attrition and how this can be applied in the real world to improve HR decision-making.

**Dataset Overview**

The data used in this project is a data set of employee data with various attributes relating to his or her professional and demographic background. These attributes give details regarding the background of the employee, work experience, and potential factors influencing his or her intention to stay at or leave the company. The next is a detailed description of each attribute

**Key Features:**

**Education:**

**Description:** It represents the employee's highest level of education. It can be a categorical variable with levels such as 'High School', 'Bachelor's', 'Master's', or 'Ph.D.'.

**Importance:** Educational level usually corresponds to one's career orientation, decision-making, and expectations in an organization. Employees holding higher levels of education can have different job expectations or career orientation than employees holding lower levels of education.

**City:**

**Description:** The geographic location where the employee works or lives. It can be represented as a categorical variable with values such as 'City A', 'City B', etc.

**Importance**: City may be a factor in employee turnover. Employees working in cities with greater living expenses, for instance, may have varying levels of job satisfaction or career advancement opportunities, which could affect their probability of staying or leaving.

**Gender:**

**Description:** Refers to the gender of the employee, usually as 'Female' or 'Male', although in certain information it may also have 'Other' or 'Non-binary'.

**Importance:** Occasionally gender can be tied to employee history in the workforce, though the latter is controversial and sensitive as a topic. It's helpful to record whether there are gendered trends among those leaving employment.

**Age:**

**Description:** Employee age.

**Importance**: Age can affect an employee's choice to quit. Younger workers might have greater turnover, looking for additional opportunities or advancement, while older employees might be looking for stability or retirement options. Age is a significant driver to predict career satisfaction and probability of leaving.

**ExperienceInCurrentDomain:**

**Description:** The number of years the employee has worked in their current job or field.

**Importance**: More experienced staff might be less likely to leave, as they are invested in the business and possess domain expertise. However, high-experience workers can leave when they perceive that they are being undervalued and are not challenged anymore.

**EverBenched:**

**Description:** Refers to whether the employee ever has been on the bench (i.e., not working on any project).

**Importance:** Being benched is likely to be among the major reasons why an employee quits a firm. Benched workers are most likely to feel neglected, under-valued, or demotivated and therefore more likely to quit the firm.

**Target Variable:**

**LeaveOrNot:**

**Description:** This is our target variable to predict. It indicates whether a worker left the company (1) or stayed (0).

**Importance:** The general goal of this project is to predict worker turnover, which has significant implications for HR management. Understanding why workers are leaving can allow HR departments to act pre-emptively in reducing turnover, improving retention, and increasing job satisfaction.

**Data Preprocessing**

The raw data must be cleaned and transformed into model-understandable format before we can train machine learning models. This is called data preprocessing. It ensures the dataset is stable, complete, and structured for analysis. Below is a description of the main preprocessing steps that have been used in this project:

**1. Handling Missing Values**

Sometimes datasets hold missing data—e.g., an employee's age or education level may never have been recorded. Missing pieces here can negatively impact the accuracy of our models.

* Carefully checked for missing values in each column.
* Where columns were numerical like Age or Experience, were substituted with the average or median value.
* For categorical columns like Gender or City, replaced missing entries using the most occurring category.

**2. Encoding: Converting Categorical Data into Numbers**

Some columns in the data, like City and Gender, contain text values. But machine learning processes work with numbers, not with words—so categorical variables had to be converted to numerical numbers.

For columns like Gender (with only two categories), we used Label Encoding. This is giving each category a number (for example, Male = 0, Female = 1).

For columns with over two categories, like City, One-Hot Encoding was used, which creates a separate column for each category.

By doing this, made sure the models could interpret and process all columns correctly.

**3. Feature Scaling**

The dataset has values which are in a different range or scale. For example, Age can be anything between 20 and 60, but Experience can range between 0 to 10. These scales will affect how the models can be learned.

An approach called Standard Scaling that made all the numeric features' scale have an average of 0 and standard deviation of 1.

This guarantees that every numerical feature contributes equally to the model's learning process.

Scaling guarantees improved performance and stability of certain algorithms, especially those involving distance or gradient computation (e.g., Logistic Regression).

**4. Selecting the Most Significant Features**

Not all the features are created equal when it comes to predicting whether an employee will leave. Some may be of little or even detrimental effect.

Correlations between each feature and the target were checked (LeaveOrNot).

Tools like feature importance scores from models like Random Forest or XGBoost were used to see which features were most significant.

Deleting unnecessary features can make the model faster, easier, and more understandable.

**5. Preparing the Final Dataset**

Once the dataset was converted and cleaned, it was divided into:

**Features (X):** All the input variables that are used to predict things (Age, Gender, City).

**Target (y):** The variable that is to be predicted — in this case, if the employee quit (1) or not (0).

This division is very important when we are going to train and test the model.

**6. Separation of Data into Training and Testing Sets**

To verify how well models perform, dataset was divided into two:

**Training Set (80%):** Used to train the model by exposing it to instances of employee traits and outcomes.

**Testing Set (20%):** Used to test the performance of the model on unseen data.

train\_test\_split function from Scikit-learn was used to accomplish this. Separating the data this way ensures that model is not just memorizing the training set but can actually generalize and make predictions on real cases.

After preprocessing, a clean, consistent, and well-structured dataset was accomplished:

* All missing values had been handled.
* All features were numeric and scaled properly.
* Irrelevant or extreme values had been handled.

The dataset had been split into training and testing datasets for model development.