**Employee Attrition Prediction using Machine Learning**

**Project Proposal**

The major objective of this project is to create and implement a machine learning-driven solution to forecast employee attrition. Employee attrition carries severe consequences for organizations, such as increased recruitment, training, and productivity expenses. It is a comprehensive account of the capstone project beginning as a conceptual proposal and evolving into an end-to-end data science solution including exploratory data analysis, feature engineering, model comparison, and business strategy recommendations. The outcome of this project will provide actionable insights that can aid human resource departments in taking preventive steps towards employee turnover, leading to enhanced talent retention and organizational stability.

The predictive modeling approach used in this project considers various employee characteristics such as demographic information, job experience, geographic location, educational attainment, and measures of engagement such as whether the employee was ever benched. Using the strengths of a combination of machine learning algorithms and cross-validation techniques, the report will present an integrated comparison of model performances, their strengths, and practical implications. Furthermore, project management practices have been touched upon to guarantee timely completion, ethicality, and professional presentation of findings in a final report and poster.

**Introduction**

With today's rapid, rapid growth and extremely competitive business landscape, human capital is probably one of the most prized assets a company has. Workers have direct impact on productivity, culture, and overall success of a company. Despite that, most companies are unable to retain employees. When one leaves, not only is workflow dynamics affected, but it also incurs enormous costs of replacing, training, and hiring them. Due to that, forecasting employee turnover that is, the likelihood an employee will leave the company has become a priority for companies of all sizes.

Historically, companies have employed qualitative methods like exit interviews and questionnaires to ascertain why people are leaving. Although they can provide some information, they are already too late to prevent turnover from occurring and furthermore limited in scope. One silver lining to this issue is that the advent of data science and machine learning offers an alternative solution to this issue. By analyzing past information about employees, we can search for patterns to forecast whether a current employee will leave or not. Therefore, organizations can go the extra mile to get employees happier, more engaged, and retained.

This project entails using machine learning models to examine a dataset that contains information regarding employees. The data contains columns such as the educational level, gender, age, city, years in current field, and whether and when the employee was ever benched (i.e., never assigned to a project for a long period of time). All of those factors help us build a better knowledge of what could make someone quit.

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

By effectively projecting employee attrition, businesses can save money, demystify turnover, and create a well-balanced workplace. Further, the determination of the reasons for employee attrition helps the management make informed decisions on workplace policies, motivational issues, and career development opportunities. Through this, data-driven decisions may not only be beneficial to individual employees but may also lead to organizational long-term success.

**Objectives**

The key objectives of this capstone project are as follows:

* Develop a predictive model to identify employees at risk of leaving the organization using historical data.
* Compare and evaluate multiple machine learning algorithms for model accuracy and generalization performance.
* Investigate and identify the most significant factors influencing employee turnover.
* Provide business-oriented recommendations based on data insights to enhance employee retention strategies.
* Demonstrate the application of project management methodologies in the execution of a real-world data science project.

By achieving these objectives, the project seeks to demonstrate how advanced analytics can contribute to organizational performance and sustainability.

**Problem Definition**

The fundamental problem addressed in this project is employee attrition—i.e., the capability to detect employees who are about to leave. Employee turnover has significant far-reaching effects: direct costs such as replacing employees, indirect costs such as lost productivity, and non-monetary effects such as reduced morale among the team. Being aware of the causes and having the capability to predict attrition allows organizations to proactively respond to mitigate these effects.

The predictive challenge is formulated as a binary classification problem with the target variable, LeaveOrNot, being 1 for the employees who left and 0 for the employees who stayed. The solution is to create and train machine learning models on a data set that has numerous attributes (e.g., education level, age, city, experience level, gender, and whether the employee has been benched or not). The focus is not so much on accuracy as on interpretability and business relevance of the models.

**Scope and Methodology**

**Scope**

This project centres around a publicly available dataset of anonymized employees. The information includes variables providing insight into employees' demographic, educational, and professional attributes. The scope of analysis includes:

• Preprocessing and data cleaning

• Exploratory Data Analysis (EDA) with a view to exposing trends and patterns

• Feature engineering and feature selection

• Creation, training, and testing a batch of machine learning models

• Comparison of models to uniform measures

• Creation of insights and business recommendations

The project yields an end-to-end report, poster presentation at expert level, and a working predictive model. The provided data set, while incomplete, is a good mix and range of attributes from which to create worthwhile predictive models.

**Methodology**

In this project, a structured data science workflow was followed, adapted to address a binary classification problem with structured HR data. The methodology consisted of the following principal steps, based on pragmatic machine learning development cycles and facilitated by tools like Jupyter Notebook, Python (Pandas, Scikit-learn, XGBoost), and Matplotlib for visualization.

**1. Problem Definition:** The main task was to predict employee attrition. This was done by casting the problem as a supervised binary classification problem, with the label LeaveOrNot designating whether an employee stayed (0) or left (1).

**2. Data Acquisition and Understanding:** A public, structured HR dataset was used. Data understanding consisted of:

* Verifying column data types and ranges
* Identifying null or outlier values
* Understanding the distribution of key demographic and experience-based features

**3. Data Preprocessing:** Data preprocessing involved a number of transformation steps to prepare the data for modeling:

**Missing Value Handling**: Median imputation for numeric variables and mode imputation for categorical variables

**Categorical Encoding:** Label encoding and one-hot encoding as required.

**Feature Scaling:** Standardization of numeric variables for use in algorithms like SVM and KNN

**4. Exploratory Data Analysis (EDA):** EDA provided insights into how features relate to attrition:

* Univariate and bivariate plots
* Correlation heatmaps
* Target variable imbalance analysis

**5. Feature Engineering and Selection:** Feature engineering involved:

* Creating new indicators from domain expertise
* Elimination of irrelevant or redundant columns
* Feature ranking by tree-based importance and correlation with the target

**6. Model Development:** The following algorithms were attempted:

* Logistic Regression
* K-Nearest Neighbors (KNN)
* Decision Tree (CART)
* Support Vector Machine (SVM)
* XGBoost Classifier

Each model was trained on an 80% training set and evaluated on a 20% test set. Model selection was based on F1 score, AUC, and stability across folds.

**7. Cross-Validation and Hyperparameter Tuning**:

* Parameters like depth, learning rate, and estimators were tuned
* Balanced class weights were used where relevant

**8. Model Evaluation:** Key evaluation metrics were:

* Accuracy, Precision, Recall, F1 Score
* ROC-AUC
* Confusion Matrix Analysis

**9. Interpretation and Explainability:** Values and feature importance rankings interpreted XGBoost's decisions. Feature insights informed business-oriented recommendations.

**10. Deployment Planning (Conceptual):** Though not deployed to production, deployment considerations were:

* Integration into HR dashboards
* Model retraining schedules
* Ethical use policies

This was an iterative process with feedback cycles at each phase ensuring improvement and mapping against the goals of the project.

# **VI. Ethical Considerations**

As with any project involving data, ethical considerations are of the utmost importance, especially when working with human subjects, even in anonymized form. Predictive modeling for human resource management raises issues of fairness, transparency, and biases that can be introduced by the data or modeling methods employed. A number of measures were taken to ensure that the project was following ethical best practices:

**1. Data Privacy and Anonymization:** The data set used did not contain personally identifiable information (PII), thus complying with data protection regulations such as GDPR. There was no attempt to re-identify individuals.

**2. Bias Detection and Mitigation:** As features like gender and city may be bias sources, their influence on model performance was carefully monitored. Models were evaluated for disparate impact, and techniques like balanced class weighting and fairness-aware preprocessing were explored.

**3. Ethical Decision-Making:** The ultimate application of the model's predictions will not be utilized in punitive or discriminatory actions. Instead, they are to be utilized to inform positive retention efforts, such as by offering mentorship, flexible work schedules, or tailored engagement initiatives.

In following these ethical guidelines, the project ensures responsible machine learning application in human resources.

# **VII. Data Source Overview**

The foundation of this predictive modeling project is established on the basis of a well-structured dataset that provides a detailed description of employee profiles. The dataset, available in the public domain, contains a range of features that span demographic and professional aspects. The main intention behind utilizing this dataset was to utilize pertinent attributes to make predictions about employee attrition.

The dataset includes the following major features:

**• Education:** Categorical variable indicating the highest educational qualification obtained (e.g., Bachelor's, Master's, PhD).

**• City:** Indicates geographical regions of employment (e.g., Bangalore, Pune, New Delhi).

**• Gender:** Binary or categorical data denoting the employee's gender.

**• Age:** Numerical variable denoting the age of the employee.

**• ExperienceInCurrentDomain:** Numerical value indicating the number of years the employee has been in the current line of business.

**• EverBenched:** A dummy variable indicating if the employee was ever on the bench (i.e., briefly without a work assignment).

**• LeaveOrNot:** The response variable indicating if the employee stayed (0) or left (1) the company.

Although not exhaustive with regard to all possible employment factors, the dataset provides a diverse and representative representation that allows for predictive analytics and data-driven HR interventions.

Initial data was provided in CSV format, which was imported into a Jupyter Notebook platform for exploratory data analysis and preprocessing. Minimal integrity checks for data preparation were carried out to ensure that the data was ready and consistent for model build. Missing values, data type issues, and category consistency issues were handled in the preprocessing step.

# **VII. Project Plan**

The project plan adhered to a logical sequence as per CRISP-DM, with checkpoints and iterative evaluation in between. The following steps were formulated and executed:

**1. Requirement Gathering and Problem Definition:** Established business goals and translated them into data mining objectives.

**2. Data Understanding and Preprocessing:** Performed intensive exploratory analysis, identified missing or inconsistent data, and used appropriate transformation techniques.

**3. Feature Engineering:** Derived new features and carried out feature selection based on correlation and significance.

**4. Model Building:** Several machine learning models were tried, compared, and assessed.

**5. Model Tuning:** Hyperparameters were optimized for improved model performance.

**6. Interpretation of Results:** Visualized results, outlined model results, and explained the impact of features.

**7. Reporting and Documentation:** Presented findings in a report and summarized results graphically for poster presentation. This structured workflow helped ensure that all aspects of the project were thoroughly addressed, from data wrangling to strategic insights.

# **IX. Data Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) plays a foundational role in understanding the characteristics of the dataset, detecting anomalies, recognizing trends, and informing model development decisions. For this project, EDA was conducted using Python’s powerful data visualization and manipulation libraries including Pandas, Matplotlib, and Seaborn.

## **1. Summary Statistics**

We began with descriptive statistics in order to have a broad idea of each feature:

**Mean (average):** Helps us understand the central value (e.g., average age of an employee).

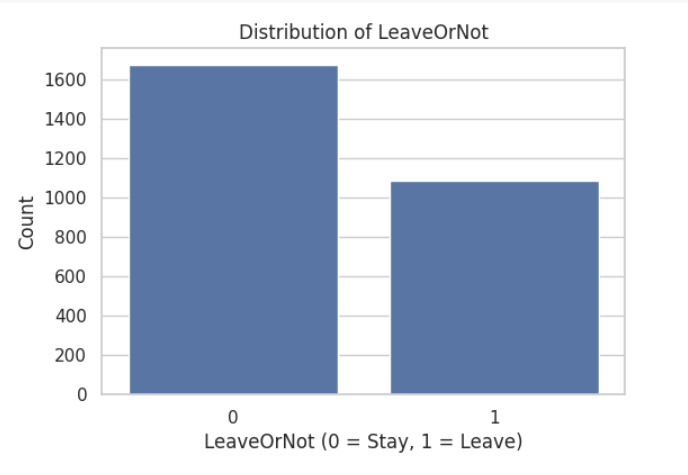
**Median:** The middle value, giving a clearer picture than the mean in the presence of outliers.

**Minimum and Maximum values:** Help to identify outliers.

**Standard Deviation:** It tells us about how dispersed the data is.

These figures give a quick but good overview of the data set and allow us to identify any abnormalities, such as very high or low values that may affect our models.

**# Fig 1:** Plot distribution of target variable



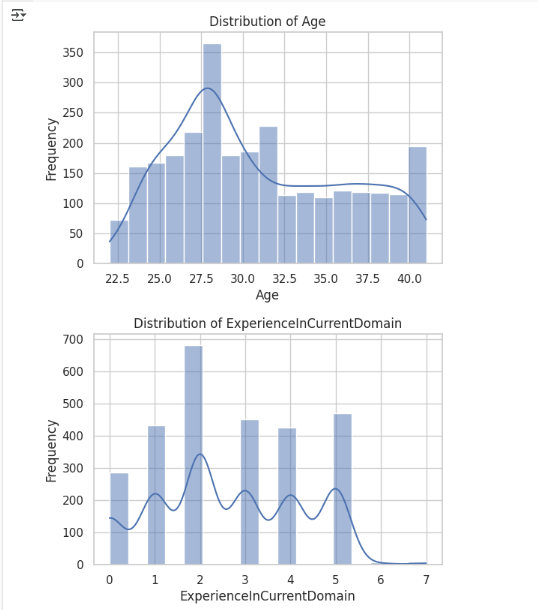
Interpretation: This graph provides a visual summary of employee attrition, and more precisely, the count of workers who chose to depart from the company versus the count of workers who chose to stay. The plotting variable is known as LeaveOrNot, which has two possible values:

**0:** Indicates that the worker stayed with the company.

**1:** Indicates that the worker departed from the company.

* On the x-axis, it is easy to identify these two groups. On the y-axis, it is easy to recognize the count of employees within each group as simply counts.
* The chart clearly allows us to note that: '0' (Stayed) bar is clearly higher than the '1' bar.
* We can understand from this that in the data set, there are more employees staying with the organization than the one who leaves.
* Overall, this chart tells us that the majority of the employees stayed, but there is still a large chunk who left, and that provides a good foundation to build a predictive model on employee attrition.

## **Fig 2:**



Interpretation:

**a. Age Distribution**

This chart indicates the distribution of employee ages within the dataset. The x-axis presents the age of employees, which goes from approximately 22 to 40 years. The y-axis represents frequency, or the number of employees in each age category.

From the chart, we can see:

The most common age is 27 to 30 years, with a peak at 28 years, where frequency is highest. There is a consistent decline in frequency as age exceeds 30, indicating fewer older employees are represented in the data. The curve drawn over the bars (a KDE curve) provides the smoothed sense of the distribution, emphasizing that the majority of employees are in their late 20s to early 30s.

The distribution is right-skewed, suggesting a younger employee population, with decreasing numbers of employees in the older age brackets (35+). This information is useful as age is usually a deciding factor in most employee decisions, such as staying or leaving an organization.

**b. Distribution of Experience in Current Domain**

This chart shows how many years of experience employees have in their current work domain. The x-axis shows the number of years of experience from 0 to 7 years. The y-axis shows the number of employees that fall under each experience level.

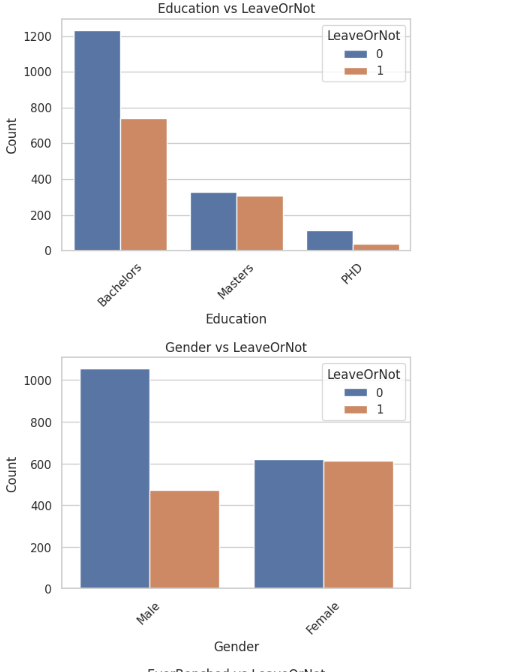
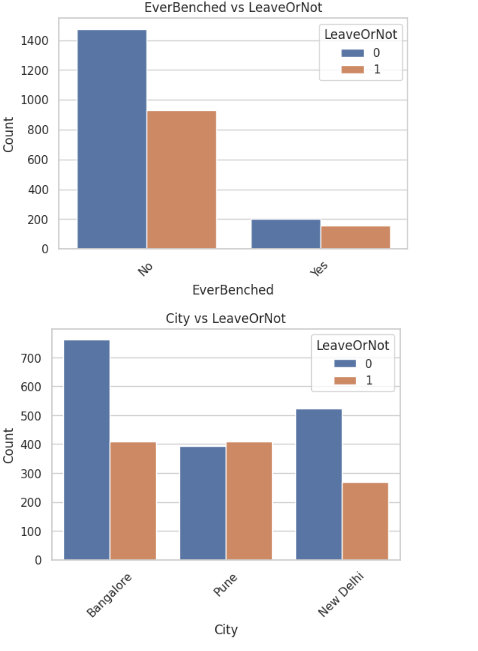
**Key findings:**

There are many employees in the 2 years of experience, which is the highest bar in the chart. Some employees also have 1, 3, 4, and 5 years of experience, showing a diverse distribution of early- to mid-level experience. There is a fall-off in employees with 6 or more years of experience.

The distribution is multi-peaked, which can mean there are several entry points or career plateaus in the company hierarchy. Overall, this chart reveals that the dataset is dominated by employees with low to moderate domain experience, which may influence their job stability and likelihood of leaving. Less experienced workers may be in the early stages of career exploration or job hopping.

Collectively, these charts provide a baseline understanding of the employee profile — young individuals with comparatively fewer years of domain-related experience — which are both relevant factors when exploring trends related to employee attrition.

## **Fig 3**

**a. EverBenched vs. LeaveOrNot**

This bar graph shows the relationship between whether or not an employee has ever been benched (i.e., taken out of active projects briefly) and whether they tend to leave the company or not. The "No" column shows a significantly greater number of employees staying than leaving. Of those employees who were benched at some point in time, the proportion of employees leaving is much more similar to that of those who remained, but still lower. This would mean benching would be associated with an increased risk of attrition, possibly for job dissatisfaction or perceived job insecurity.

**b. City vs. LeaveOrNot**

The chart displays the distribution of employee retention and attrition in three cities: Bangalore, Pune, and New Delhi. The largest overall employee number is in Bangalore, with a huge difference between the employees retained and those lost. Pune has very close numbers for both categories, while New Delhi has a larger retention than attrition but with a larger difference than Pune. This could imply that geographical location is a factor in employee turnover to a certain degree, possibly because of differences in job markets, cost of living, or working culture among cities.

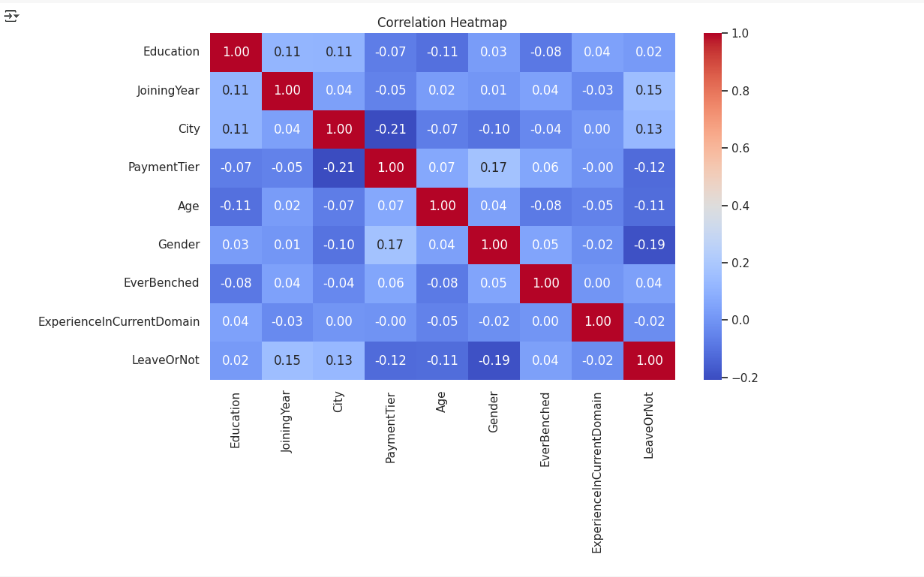
**c. Education vs. LeaveOrNot**

This graph analyzes the relationship between employee turnover and education level. Employees with Bachelor's degrees comprise the majority of the dataset and have the highest number of both staying and leaving. The category who possess Master's degrees have relatively balanced proportions, while PhD holders have the lowest number in total and a much lower attrition. This may suggest that more advanced degrees like a PhD are associated with better retention, perhaps from more senior roles or job satisfaction, whereas those holding a Bachelor's degree might have more career mobility or constricted development, so that they leave.

**d. Gender vs. LeaveOrNot**

This bar graph contrasts gender with employee turnover. For male workers, significantly more stayed than left, whereas that for women is nearly equal. This pattern may be explained by variations in job roles, support networks, or other external factors affecting women's retention more significantly, e.g., work-life balance or organizational culture. These results can guide organizations to design gender-specific retention strategies.

## **Fig 4:**



The correlation heatmap gives a graphical representation of the correlation between various numerical and encoded categorical features of the dataset. The correlation is between -1 and 1, with numbers close to 1 indicating strong positive correlation, numbers close to -1 indicating strong negative correlation, and numbers close to 0 indicating little or no linear relationship between variables.

**Key Observations:**

Low Overall Correlation with Target (LeaveOrNot): None of the features have a high correlation with the target variable LeaveOrNot.

**Gender (-0.19 correlation with LeaveOrNot):**

A moderately negative correlation suggests that gender could be an issue with attrition, and one gender is slightly more likely to depart than the other. The correlation is not strong enough, however, to make anything conclusive without further analysis.

**Joining Year (0.15 correlation with LeaveOrNot):**

Moderate positive correlation indicates that the date an employee started working for the company may have an influence on whether or not they are going to leave, perhaps due to changes in organizational policies and expectations from the workforce over time.

**City (0.13 correlation with LeaveOrNot):**

A correlation by location, where employees in certain cities might have more job mobility or other opportunities, which would influence their stay or leave decision.

**Payment Tier (-0.12 correlation with LeaveOrNot):**

There is a weak negative correlation, suggesting that staff in lower-paid brackets are maybe more likely to leave, perhaps due to dissatisfaction with pay or career advancement opportunities.

**Age (-0.11 correlation with LeaveOrNot):**

Older employees might be that little bit more settled or committed, and this leads to a slightly reduced level of attrition.

**Education, EverBenched, and ExperienceInCurrentDomain:**

All have extremely close-to-zero correlation with the target variable, and thus no direct linear effect on attrition in themselves.

Some traits have low correlations with each other, which are:

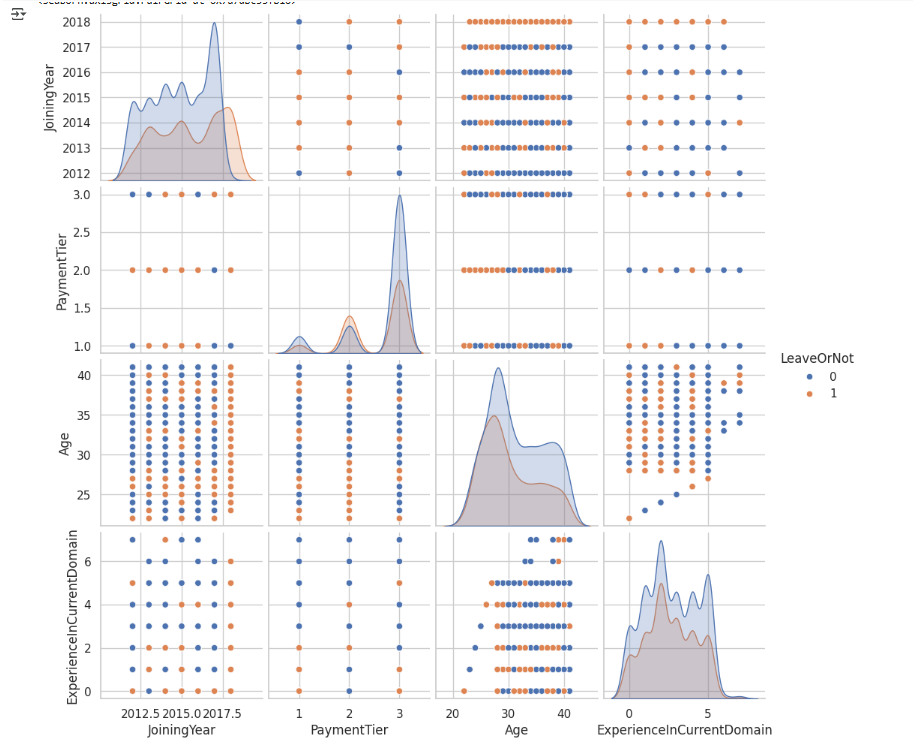
**PaymentTier and Gender (0.17)**

**Education and City (0.11)**

These can indicate structural in the data set, such as the genders being spread out over pay levels.

This heatmap is a first diagnostic tool to understand the dataset. While it shows broad trends, it does not expose non-linear interactions or variable interactions. Hence, more advanced tools like feature importance analysis and SHAP values can expose more about how each variable contributes to model predictions.

## **Fig 5:**



This is a pair plot used to visually explore relationships between different employee features in a dataset, distinguishing between those who left the company (LeaveOrNot = 1) and those who remained (LeaveOrNot = 0). Both groups are color-coded: the one who stays are blue and who leaves are orange.

**Variables Examined**

JoiningYear, PaymentTier, Age, ExperienceInCurrentDomain, LeaveOrNot

**Diagonal Elements**

Every diagonal cell shows a distribution (density plot) of one variable divided by whether the employees left or not:

**JoiningYear**: More employees joined in 2016 and 2017. Those remaining are more likely to have joined recently.

**PaymentTier:** The majority of employees are Tier 1. Those in higher tiers (2 or 3) are slightly more likely to leave.

**Age:** Most employees are aged between 25 and 35. Density of stayees is higher in that group.

**ExperienceInCurrentDomain:** Most employees have 1–5 years of experience. More experience weakly corresponds to staying.

**Off-Diagonal Elements**

These show scatter plots between two variables, color-coded by leave or not. Some observations worth noting: There's a high concentration of those who stayed among newer hires (2015–2017), especially at younger ages and lower levels of experience. Higher Pay is more likely to have left (more orange dots in Tier 2 and 3). Age is strongly correlated with experience — older employees have more domain experience, and most of them stayed.

**Overall Insight is:**

Younger employees with low payment levels and fewer years of experience in the domain have higher chances of staying. On the other hand, older, more experienced, and higher-paid employees have a relatively higher chance of leaving.

# **X. Pre-processing**

Pre-processing is the process of preparing the dataset for algorithm ingestion through machine learning. It entails handling missing values, encoding categorical features, scaling numeric attributes, and selecting the most predictive features.

**Data Cleaning**

Missing values were encountered in fields Age and Education upon initial data exploration. The following was performed:

•**Numerical Imputation:** Missing values in fields Age and Experience were imputed with median values.

•**Categorical Imputation**: Entries for missing gender or city were imputed with the mode of the variable.

**•Outlier Handling:** Numerical feature outlier values were checked with IQR (Interquartile Range). Where required, values were clipped but not deleted in order to preserve data representativeness.

**Gender Pay Gap Calculation**

While the data lacked explicit salary, tier of payment and city were employed as proxies to approximate differences. Gender groupings uncovered little overrepresentation among males at greater tiers of payment. This was explored during model interpretation but did not dominate feature importance rankings.

**Encoding and Scaling**

**• label Encoding:** Applied to binary features like Gender and EverBenched.

**• One-Hot Encoding:** Applied to the variables with more than two categories (like City, Education).

**• Standard Scaling:** Performed on Age and ExperienceInCurrentDomain through Scikit-learn's StandardScaler such that all numeric features had a standard deviation of 1 and a mean of 0. This was particularly required for distance-based models like KNN and SVM.

# **XI. Feature Selection for Analysis**

**Feature importance was estimated on:**

• Correlation analysis

• Tree-based model feature importance (Random Forest, XGBoost)

Most important features which were consistently selected were:

1. ExperienceInCurrentDomain

2. EverBenched

3. Age

4. City (particularly Pune)

5. PaymentTier

These features were highlighted during training and model testing.

## **A)Modeling and Evaluation**

After data pre-processing, the next crucial step was building and evaluating the model. The goal was to identify the machine learning algorithm yielding the best predictive accuracy in employee attrition in terms of complexity, interpretability, and computational expense.

Five models were evaluated:

1. Linear Regression

2. K-Nearest Neighbors (KNN)

3. XGBoost (Extreme Gradient Boosting)

Each model was then trained on the preprocessed dataset and cross-validation used to evaluate generalization performance.

## **B)Data Split**

Splitting of the data was performed using training and test sets with an 80:20 ratio. This implied that models were trained on the majority of the data with sufficient samples left such that performance could be tested without prejudice.

Stratify parameter was used to maintain the target variable class distribution across splits, which was important since class imbalance was present.

## **C) Training and Test Data**

Training data were used to train models, and test data were used solely for final evaluation. Scikit-learn's default API was used to train models, and XGBoost was trained using the xgboost library. Accuracy was not sufficient, so other measures beyond that—precision, recall, F1-score, AUC-ROC—were used.