**Domain Background**

Employee attrition is a huge and costly problem for most of the companies. The cost for replacing an employee with new employee involves huge cost.

Recent studies shows companies spends huge amount of money to replace an employee .The huge cost may occur due to below reasons

* Training cost
* Hiring cost
* Loss of productivity since new employee takes time to get accustomed to new role etc.,

**Problem Statement**

Employer invests a huge amount of money in finding a new employee in place of the employee who left the organisation. Since the cost of replacing employees for most employers is very high, if we predict the likelihood of an employee leaving the company, this will lead to actions to improve employee retention as well as possibly planning new hiring in advance. By using the predictive model, we will predict likelihood of an active employee leaving the company it would save a lot of time and money to the employers.

**Datasets and Inputs**

The dataset used is an open source data set created by IBM which contains the data about 5000 employees. Since the number of employees leaving will be less when compared to number of active employees, the data set is imbalanced with an attrition event rate of 16%.

The dataset contains several numerical and categorical features providing various information on employee’s details (not sensitive) and employment details. The target variable is ‘Attrition’ which is a binary variable, 0 (active employee), 1 (former employee) .Some of the categorical features include Department, Education level, Education Field, Job Role etc. Some of the continuous features includes Age, Number of companies worked, Hike percentage, performance rating etc.

**Benchmark Model**

Since the problem is a standard **supervised classification problem**, Logistic regression can be used as a starting step. The minimum accuracy would be 80% and minimum ACU-ROC will be 0.5 .The above benchmark metrics would be helpful in comparing the results of the final solution.

**Solution**

The employee churn detection will be addressed analytically through any of the below classification models.

* Regression based models (Logistic, other GLM model)
* Tree based models (Decision tree, Random Forest )
* XGBoost and other Boosting Algorithms

**Evaluation Metrics**

Below evaluation metrics will be used to evaluate the model performance

Since the dataset is imbalanced, ROC-AUC will be the best evaluation metric to measure the performance of the model

* **Accuracy**:  Overall, how often is the classifier correct
* **False positives (FP):** We predicted yes, but actual is No. (Type I error)
* **False negatives (FN)**: We predicted No, but actual is Yes (Type II error)
* **Precision**: True Positives/(True Positives False Positive)
* **Recall**: True Positives/(True Positives False Negatives)
* **ROC-AUC**: Area under the curve

**Project Design**

The workflow for approaching a solution the problem statement is

* **Exploratory Data Analysis :**Understating the patterns in the data .Uni variate and bivariate analysis
* **Data Pre-processing :** Standardising, Missing values and outliers treatment
* **Model Building:** Training the model to find the solution of problem statement
  + Since the dataset is imbalanced stratified sampling is used in order to maintain same event rate in train and test splits
  + Grid search is used to select the optimal hyper-parameters for the model
* **Model Selection:** Selecting the appropriate model
  + The final metrics will be compared with the benchmark model metrics
* **Model Evaluation:** Evaluating the Model on testing Data