# CLOUD COMPUTING

## Project Report

## Title: Employee Years at Company Prediction

## Team Members:

Sagar Varma Samanthapudi

Priyanka Surapaneni

Leena Guduru

## Introduction:

Recruitment is always long process which consume more time and energy of HR team. There is plethora of challenges faced by recruiters in selecting the ideal candidate for the job. Some of them are

- Finding the ideal applicant in the pool of talented people is limited and even if you identified the best person from the people who applied is not suitable for the job description. At end it all depends on the candidates who will apply to the job.
- Engaging the qualified candidates via emails or through phones is crucial.
- Due to vacant positions and delay operation the hiring team must hire faster to meet their deadlines.
- HR team must also ensure Employing a data-driven recruitment strategy, creating a powerful employer brand, providing a positive candidate experience, Fair recruitment, Creating a productive recruitment process.

What if after all this tiresome process, the candidate stays only for 6 months or left after training period?

## Problem Description:

Employee churn is a major problem for many firms these days. Great talent is scarce, in high demand and hard to keep if found. Hiring and retaining employee

are highly difficult activities that necessitate a lot of money, effort, and expertise. Some study tells that a company may pay 15% to 20% of an employee's income to hire a new employee, which is a significant sum, especially for large corporations with thousands of employees. The time it takes to get a new worker up to speed costs an average company between 1% and 2.5 percent of their entire revenue.

**Solution: Develop a model which will be highly effective in determining the number of years that an employee can stay at an organization.**

## Description of data:

**Data source:**

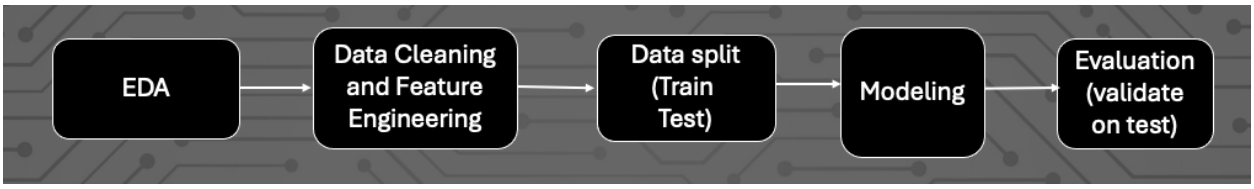
1. https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study
2. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Total number of records: 5882
Total number of features: 35
Output features: 1 (YearAtCompany)

```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

**Data Dictionary:** https://github.com/sonithapriya/CC_Employee-Years-at-Company-Prediction/blob/main/data/data_dictionary.xlsx

**Workflow:**



# Methodology:

### Data Cleaning and Preprocessing:

- Combined data from both URL's by matching the columns.
- Replaced the null values with median.
- Handled outliers with MAD method (Median Absolute Method).
- Dropped Columns like age, employee count etc.
- Dropped age group 10-20 with zero experience as outliers.

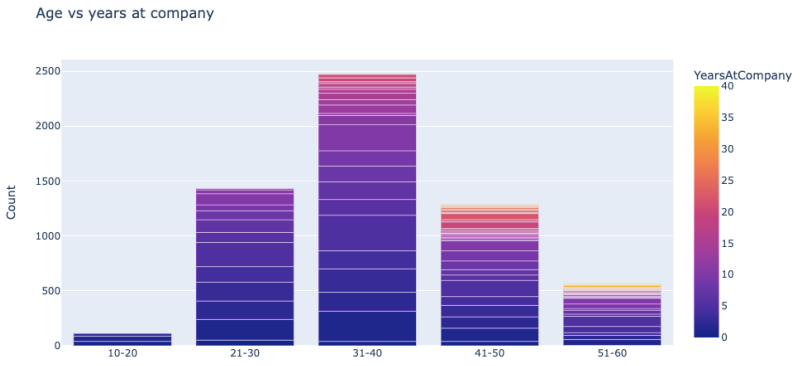
### Feature Engineering:

- Created age groups from age column by aggregating them to bins.
- Converting categorical into vectors for modeling and assembling the indexed data and numeric features in a vector.

### Exploratory Data Analysis:

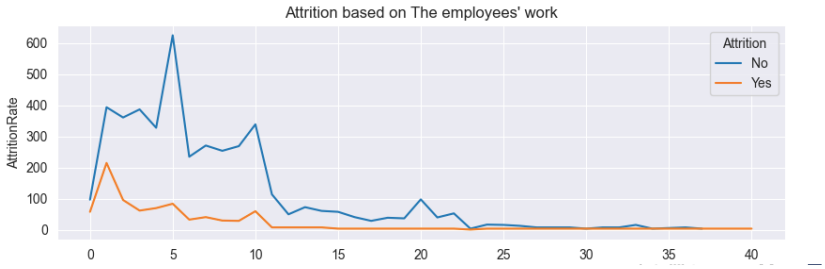
### Age vs Years at company:

The age group of 10-20 contain zero experience with row count zero which indicates as outliers. As age increases years at company increases according to their experience.



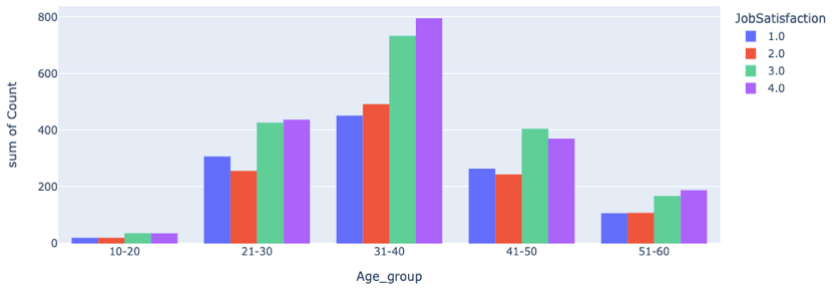

### Attrition vs Years at company:

Suppose 5 years at company tends to retain in company than to leave and as years at company increases attrition is insignificant.



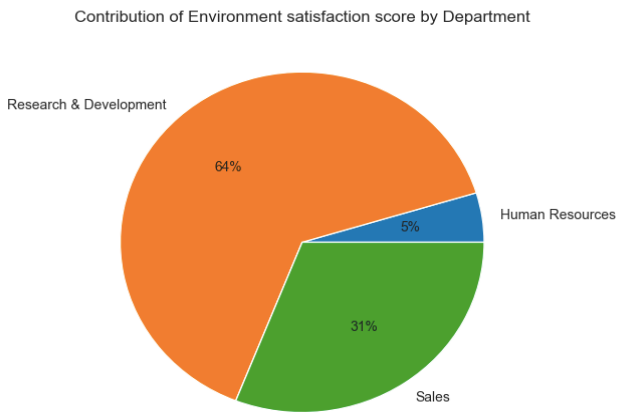

### Age group vs Job satisfaction:

Age group 31-40 are mostly satisfied with their jobs and other age groups are mostly 3 or same at job satisfaction.



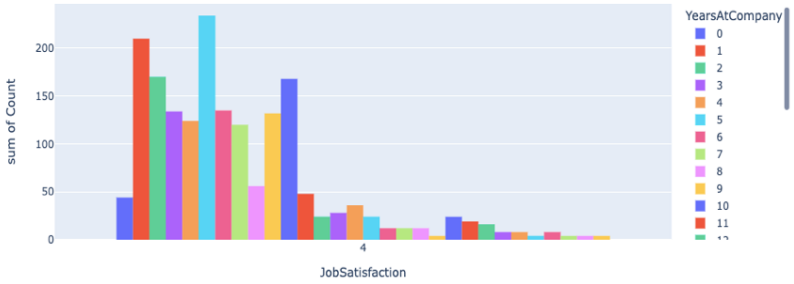
**Department vs Environment satisfaction:**

Research and development department has most environment satisfaction than other departments.




**Year at company vs Job satisfaction:**

5 years at company experience people has more job satisfaction and as experience increases job satisfaction is insignificant.




**Correlation plot:**

Age is correlated with total working years. Years at company is correlated with total working years etc.

**Modeling:**

- Split the data in to train and test datasets with 70-30% spilt.
- Implemented Linear Regression, Decision Tree Regressor, Random Forest models.
- Models are tested on test dataset.
- Tools: Pyspark, Jupyter notebook.

## Results:

| Models | RMSE | R2 |
|---|---|---|
| Linear Regression | 3.182349 | 0.729678 |
| Decision Tree Regressor | 2.12158 | 0.894548 |
| Random Forest | 2.08155 | 0.896293 |

Random Forest performed better and predictions are according to RF model

**Predictions:**

```
## predictions vs Original data
predictions.select("prediction", "YearsAtCompany").show()
```

```
+----------+--------------+
|prediction|YearsAtCompany|
+----------+--------------+
|     16.57|            18|
|      4.45|             3|
|      4.04|             3|
|      2.93|             4|
|      2.93|             4|
|      3.97|             5|
|      9.98|            10|
|      4.66|             9|
|      4.47|             5|
|      0.98|             0|
|      3.09|             3|
|      4.42|             5|
|      4.42|             5|
|     10.49|            11|
|     12.14|            13|
|     12.14|            13|
|      4.07|             5|
|      4.89|             4|
|      4.34|             5|
|      4.97|             5|
+----------+--------------+
only showing top 20 rows
```

## References:


1. https://resources.workable.com/stories-and-insights/common-recruiting-challenges
2. https://towardsdatascience.com/employee-retention-using-machine-learning-e7193e84bec4
3. https://spark.apache.org/docs/latest/ml-classification-regression.html

# Appendix 1:

**Contribution from each member:**

| Tasks | Team Members |
|---|---|
| Data Collection | Sagar, Priyanka |
| Data Cleaning and Preprocessing | Priyanka, Sagar, Leena |
| EDA | Priyanka |
| Modeling, Feature Engineering, Predictions | Priyanka, Sagar, Leena |
| Presentation and Report | Priyanka, Sagar, Leena |

# Appendix 2:

**Link to GitHub repo with project code:**
https://github.com/sonithapriya/CC_Employee-Years-at-Company-Prediction