HR Analytics: Does having 'Experience" in profession necessarily give "Exposure" also?

1. Problem Definition.

Today IT industry has reached its peak when it comes to the use of different skills and an immense amount of projects, they handle both onshore and offshore. With this comes the need for a huge amount of workforce to handle.

The introduction of HR Analytics has put the workforce at the center of the success factors for any company whether it's a small-time or an MNC. We see how quickly HR analytics is growing day by day in most of the companies as initially HR or HR-related jobs and tasks were categorized as support roles and not as business revenue departments unlike Marketing, Finance, etc.

Does attrition happen due to lack of work? Or is it because of the incompetence of an employee only? Does the performance indicator apply only to an employee? Or does it also indicate the factors that are directly or indirectly promoted by the company?

Every year a lot of companies hire several employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well.

For any Organization, high employee attrition is an additional cost and a major problem. Some of the common expenses are hiring processes, Job postings, and new or existing hire training that end up losing employees and replacing them. The knowledge base or experience over time reduces if employee turnover. It prohibits the organizations from increasing its collective base. If your business is client-facing, then it could be a problem as clients often prefer to interact with regular and familiar people and this may also increase issues and errors as new workers are not accustomed to business practices.

My project speaks about how employee attrition affects a company and why it's important for employers to look out at new approaches to improving their workforce.

2. Data Analysis.

The objective of my project is what factors majorly influence the attrition or retention of an employee at an organization.

I have considered a generic dataset that is used by one of the HR Departments and has no missing values and it's a classification problem.

Since the majority of the variables including the target variable are categorical, I used LabelEncoder to convert it into its corresponding numerical values. Also, because these features are not in an orderly manner and it's not required for it be as well

The columns 'Attrition', 'BusinessTravel', 'Department','EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18' and 'OverTime' have been converted into numerical values.

Also, the dataset has some outliers or skewness as the distribution is not normal. Hence I used Z-Score to remove outliers if any as follows:

Only 5 continuous data features seemed less normal compared to others hence this method was applied only to these.

```
In [123]: from scipy.stats import zscore

          z_score = zscore(df[['DailyRate','HourlyRate','MonthlyIncome', 'MonthlyRate','PercentSalaryHike']])
          abs_zscore = np.abs(z_score)

          filtering_entry = (abs_zscore < 3).all(axis=1)

          df = df[filtering_entry]
```

There was no significant change in the data loss and I followed up with splitting the dataset into X and Y as follows before scaling the data using StandardScaler.

```
In [6]: x = df.drop(columns = ["Attrition"], axis=1)
        y = df["Attrition"]
```

```
In [7]: from sklearn.preprocessing import StandardScaler

        scaler = StandardScaler()
        x_scaled = scaler.fit_transform(x)
        x_scaled

Out[7]: array([[ 0.4463504 ,  0.59004834,  0.74252653, ..., -0.0632959 ,
                -0.67914568,  0.24583399],
               [ 1.32236521, -0.91319439, -1.2977746 , ...,  0.76499762,
                -0.36871529,  0.80654148],
               [ 0.008343  ,  0.59004834,  1.41436324, ..., -1.16768726,
                -0.67914568, -1.15593471],
               ...,
               [-1.08667552,  0.59004834, -1.60518328, ..., -0.61549158,
                -0.67914568, -0.31487349],
               [ 1.32236521, -0.91319439,  0.54667746, ...,  0.48889978,
                -0.67914568,  1.08689522],
               [-0.32016256,  0.59004834, -0.43256792, ..., -0.33939374,
                -0.36871529, -0.59522723]])
```

Since there was a possibility of multicollinearity, I used Variation Inflation Factor to identify such features and dropped "monthly income" as it had high collinearity and its correlation with target feature “Attrition” was very less compared to other features.

```
In [130]: from statsmodels.stats.outliers_influence import variance_inflation_factor

          vif = pd.DataFrame()
          vif["vif"] = [variance_inflation_factor( x_scaled, i) for i in range(x_scaled.shape[1])]
          vif["Features"] = x.columns
          vif
```

3. EDA Concluding Remark.

Based on the given information in the dataset, I observed that the Attrition rate is very less compared to the retention rate. We have roughly about 220 that fall under the "Yes" category and over 1200 under the "No" category. Now attrition could be for various reasons. An employee doesn't need to stick to the same company even if they are provided with a high salary.

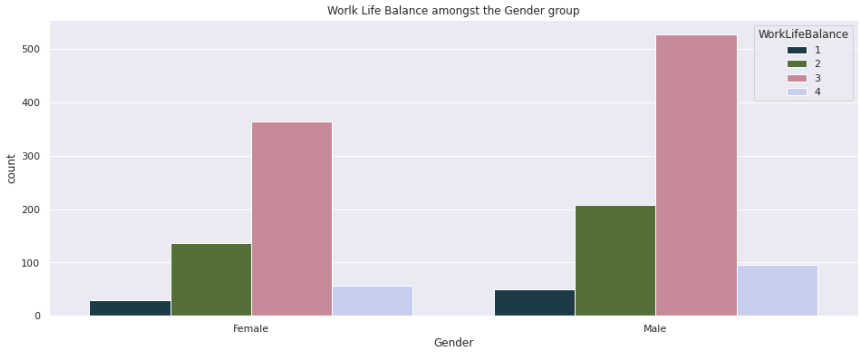
Since the dataset is imbalanced concerning gender distribution, it appears attrition of Female employees is less compared to Male employees in this scenario.

It's a common notion one would have how often a person changes the company or how long a person prefers to stay in the same company. From my observation, I found employees that who have high experience tend to stay in the company much longer than the ones without experience or very little experience.

Every person prefers stable growth and opportunities down the line as they would have explored all the possible fields or roles. But a junior or a new employee still has a lot to grow and may not prefer sticking to the same company in the early stages of the career progression.

Also, as and when a person ages, job opportunities reduce gradually. It wasn't the case a couple of years ago but now given the mindset of people, the demand for technologies, the urge to learn and succeed have all made most of the companies look out for employees from outside the market also, thus introducing fresh blood in the team.

From the above graph, we can observe that as and when the age increases, changing of jobs decreases. Hence in this dataset, we can observe a gradual change in job seems to be between 18 to 30 years and also up to 40 years to some extent. But after that, it's balanced and employees may even prefer getting retired in the same company.

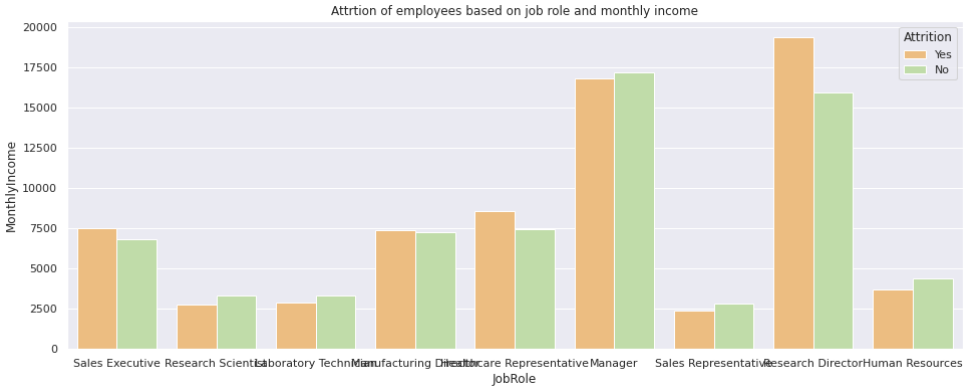


Work-life balance and job rotation are other factors that may or may not influence a person's decision to leave the company. Being a recruiter myself, I have observed these to be key issues where a company needs to focus on. Hence even in the analysis, I could find the same inferences.

As mentioned, due to the availability of different types of jobs, roles, responsibilities a person is naturally inclined to go for such positions.

As pointed out earlier, job rotation is very much important in any company. This benefits both the company and the employees. A company ends up training the person on a cross-functional basis so he/she contributes across all the modules and a company ends up saving the cost of hiring new employees.

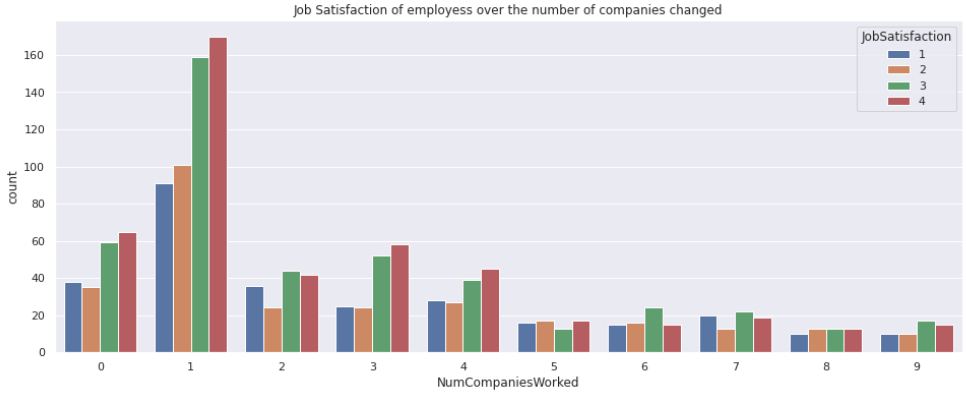
But without a proper career growth plan for an employee, things are likely to fall apart. This is where HR / Workforce analytics comes to place as it helps collect all the relevant information on available job positions, how long that specific position is open from as it's a liability for any company, how to provide constructive training to employees at a general level and also at a specific or higher level to convert them from possible liabilities to potential assets of the company.



From the above graph, we can see "Manager" and "Research Director" are highly paid compared to other employees. They could also be belonging to a specific department as salaries differ from position/role/department. But on the contrary, Attrition is also high among the same categories.

This show's how having a good salary alone doesn't indicate if a person would stay for a long time as there is a good chance of work being monotonous and no much innovation. Sometimes one's job title may be fancy but the work may not be that critical that would fetch a high salary.

Sometimes employees also seek non - financial benefits like the flexibility to some extent, better shifts, which may be of more relevance as per individual as an employees loo at how their future should and would look like even before getting into companies.

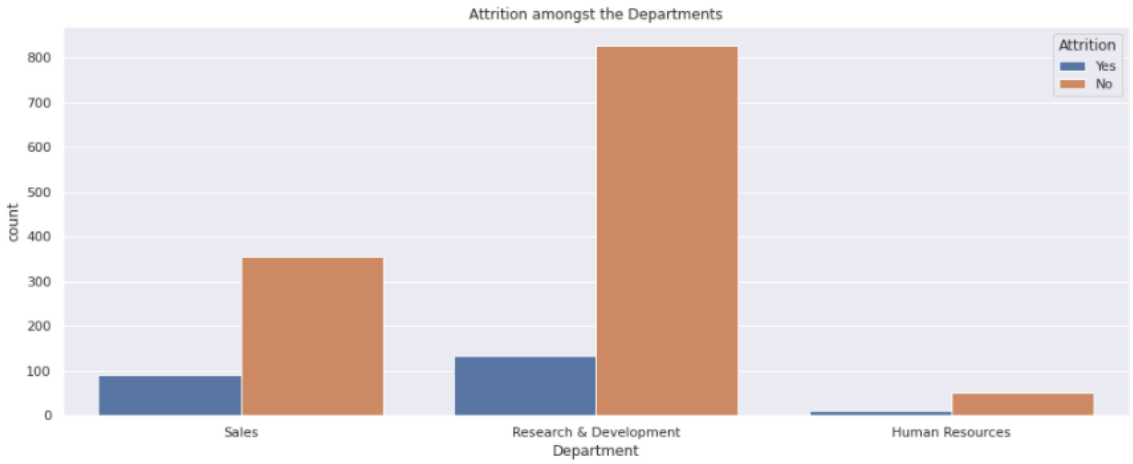


Let’s observe the relationship between job satisfaction and the number of times companies changed as shown in the above graph.

We can observe that employees who newly joined without any experience don't seem very much satisfied. It could be because they are yet to experience the factors that the experienced employees may have. It also tells us in general, if it's your first company you cannot compare it with other work experience unless that person joins elsewhere.

We can see the employees who have changed jobs at least 1 time seem highly satisfied than the ones who did not change at all. It could be because when changing jobs, one always keeps in mind to get all benefits that one may not have gotten in the first company. The first company's experience becomes your benchmark and one tries to find something better or similar but not less than that.

We can see the other category employees have consistent job satisfaction and we can also notice it gradually increases as and when they change companies over the years. This could also mean down the line other factors like salary, role, skills, etc. could be more important than just focusing on job satisfaction.



Interestingly, I can see Human Resources department has the least number of attritions compared to other departments as shown above. This raises a question is it because the HR profession as a whole is dying and hence, they do not get opportunities elsewhere? No. HR field is changing day by day and it's one of the underrated departments in most organizations right from startups to MNC's.

Today, if you can find any opportunities, is because HR Department has well-planned and executed better strategies that ensure not only the creation of multiple opportunities but also directly influence the success of a company.

A company's success or strength is not only measured by its capital or seed funding but also by the strength of the workforce. It's all interconnected where if more positions are created, a greater number of employees are required. If there are a greater number of employees effective strategies need to be implemented to retain the best potential and identify the best career progression path for employees.

This ensures reduction of attrition to some extent if not completely eventually reducing the cost of hiring and creating unnecessary open positions leading to liabilities. Hence HR in today's world is a revenue generation department with better business models aiming at both employees and employers.

4. Pre-Processing Pipeline and Building Machine Learning Models.

Since it's an imbalanced dataset, I have considered SMOTE technique at first to balance the target variable column and the result is as shown.

<u>Before balancing the target variable:</u>

```
In [134]: df["Attrition"].value_counts()

Out[134]: 0    1233
          1     237
          Name: Attrition, dtype: int64
```

After balancing the target variable:

```
from imblearn.over_sampling import SMOTE
SM = SMOTE()
x_over, y_over = SM.fit_resample(x,y)
```

```
In [11]: y_over.value_counts()

Out[11]: 0    1233
         1    1233
         Name: Attrition, dtype: int64
```

Since it's a classification problem I have used 6 Algorithms i.e. LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, and ExtraTreesClassifier to build this model and select the best suitable model by applying hyperparameter tuning.

In this case, LogicticRegression gave better results with an accuracy of approximately 72.56% as shown below.

Also, the difference between the accuracy score and cross-validation score is very less for LogicticRegression compared to other algorithms.

```
In [138]: from sklearn.linear_model import LogisticRegression
log_reg = LogisticRegression()
log_reg.fit(x_train,y_train)

y_pred = log_reg.predict(x_test)

print(accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
0.7256756756756757
[[271  70]
 [133 266]]
              precision    recall  f1-score   support

           0       0.67      0.79      0.73       341
           1       0.79      0.67      0.72       399

    accuracy                           0.73       740
   macro avg       0.73      0.73      0.73       740
weighted avg       0.74      0.73      0.73       740
```

```
In [144]: from sklearn.model_selection import cross_val_score
```

```
In [145]: scr = cross_val_score(log_reg, x, y, cv=5)
print("Cross Validation score of LogisticRegression model is:", scr.mean())
```

```
Cross Validation score of LogisticRegression model is: 0.8408163265306122
```

I then used approximately 7 parameters for Hyperparameter tuning to check if the accuracy score could be improved to some extent if possible.

```
In [168]: parameters = {"penalty":["l1", "l2", "elasticnet", "none"],
                        "tol":[1e-4, 1e-2, 1e-3, 1e-1],
                        "intercept_scaling":[1, 2, 3, 4, 5],
                        "solver":["newton-cg", "lbfgs", "liblinear", "sag", "saga"],
                        "multi_class":["auto", "ovr", "multinomial"],
                        "max_iter":[50, 70, 100, 120, 130],
                        "intercept_scaling":[1, 2, 3, 4, 5]
                        }
```

The predictions are then re-run using GridSearchCV to find the best possible parameters by passing them to “best_params_” as follows:

```
In [169]: from sklearn.model_selection import GridSearchCV
          GCV = GridSearchCV(LogisticRegression(), parameters, cv=5)

In [170]: GCV.fit(x_train, y_train)

Out[170]: GridSearchCV(cv=5, error_score=nan,
                       estimator=LogisticRegression(C=1.0, class_weight=None, dual=False,
                                                    fit_intercept=True,
                                                    intercept_scaling=1, l1_ratio=None,
                                                    max_iter=100, multi_class='auto',
                                                    n_jobs=None, penalty='l2',
                                                    random_state=None, solver='lbfgs',
                                                    tol=0.0001, verbose=0,
                                                    warm_start=False),
                       iid='deprecated', n_jobs=None,
                       param_grid={'intercept_scaling': [1, 2, 3, 4, 5],
                                   'max_iter': [50, 70, 100, 120, 130],
                                   'multi_class': ['auto', 'ovr', 'multinomial'],
                                   'penalty': ['l1', 'l2', 'elasticnet', 'none'],
                                   'solver': ['newton-cg', 'lbfgs', 'liblinear', 'sag',
                                              'saga'],
                                   'tol': [0.0001, 0.01, 0.001, 0.1]},
                       pre_dispatch='2*n_jobs', refit=True, return_train_score=False,
                       scoring=None, verbose=0)

In [171]: GCV.best_params_

Out[171]: {'intercept_scaling': 5,
           'max_iter': 70,
           'multi_class': 'ovr',
           'penalty': 'l1',
           'solver': 'liblinear',
           'tol': 0.001}
```

Rebuild the model using the appropriate params we received from best_params_. It’s observed that the model accuracy was approximately 72.56 % earlier and post Hyper Parameter tuning it's now approximately 84.72 % better.

```
In [173]: mod_log_reg = LogisticRegression(intercept_scaling= 5, max_iter= 70, multi_class="ovr", penalty= "l1", s

          mod_log_reg.fit(x_train,y_train)
          pred = mod_log_reg.predict(x_test)
          print(accuracy_score(y_test,pred)*100)

          84.72972972972973
```

6. Concluding Remarks.

- There is some sort of inflation-like situation amongst the workforce. There once was a time where not everyone could afford education, hence not everyone got into white-collar jobs. Today availability of an abundant number of resources gives wide options for companies to choose best resources from a pool of candidates.
- My major understanding is as mentioned, "Exposure is far more important than the Experience alone". A person's salary is dependent on what he/she has learned so far, what are they willing to learn, and how flexible a person is when it comes to taking up initiatives.
- Most of the companies have opted out of hiring multiple people for multiple roles or projects and prefer and expect an individual to handle more tasks and responsibilities. This way the amount saved on hiring multiple people could be used in the increase of salary for the existing employees based on the specifications of work and their performance.
- This situation applies to HRs also. If you look at how startups work, they have a greater number of technical employees than HRs. This tells the same set of ideology is applied to every single department and people may be hired on need only basis when it comes to complex or filling serious roles.
- We see a lot of companies hiring contractors are approaching third-party consultants that find and take care of filling more generic positions. This is quite coming in most of the abroad countries and it's slowly catching up in Indian job markets also. The hiring time, screening efforts, negotiations, profile backups, etc. are all handled by these consultants to a greater extent and this saves a lot of time for companies which in turn increases productivity to some extent.
- Many companies have also tied up with learning centers that train employees on their behalf. This not only reduces company's burden on creating specific curriculum as it's already available customized a per company's needs but it also benefits an employee as they get to learn from industry professionals at a broader level.