**Name of the Project**

HR Analytic Project

**Submitted By: - Prakhar Dixit**

**Problem Definition: -**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**Data Analysis**

Data set contain the 35 Columns and 1470 rows including the target columns.

After analyzing the data set, we are happy to see there is no null value in the data set. After more analyzing we observed that there are so many differences in min and max value which concluded we need scaling in future.

There are following categorical columns in this data set:

**Data Analysis**

|  |  |  |
| --- | --- | --- |
| Column | Data Type | Null Values |
|  |  |  |
| Age | Int 64 | 0 |
| Attrition | Object | 0 |
| Business Travel | Object | 0 |
| Daily Rate  Department  Distance From Home  Education  Education Field  Employee Count  Employee Number  Environment Satisfaction  Gender  Hourly Rate  Job Evolvement  Job Role  Job Satisfaction  Marital Status  Monthly income  Monthly Rate  Over 18  Over Time  Percent Salary Hike  Performance Rating  Relationship Satisfaction  Standard Hours  Stock Option Level  Total Working Years  Training Times Last Year  Work Life Balance  Years At Company  Years In Current Role  YearsSinceLastPromotion  Years With Curr Manager | Int64  Object  Int64  Int64  Object  Int64  Int64  Int64  Gender  Int64  Int64  Int64  Object  Object  Int64  Int64  Object  Object  Int64  Int64  Int64  Int64  Int64  Int64  Int64  Int64  Int64  Int64  Int64  Int64 | 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |

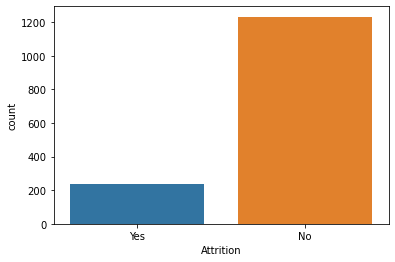
From the table we can observed that the there are two type of columns object and nominal. We are happy to see there is no null value in the data set.

From the data set we observed that there is difference in min and max value of data so we can understand we need to scaling of data in future.

**EDA (Exploratory Data Analysis)**

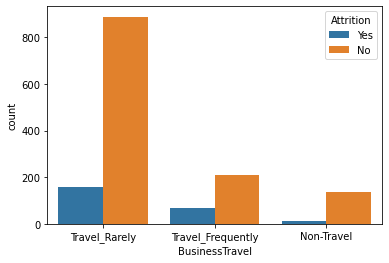
1.-

After observing the attrition column, we conclude that This is highly imbalance data set, here out of 7 employee 1 is facing Attrition.



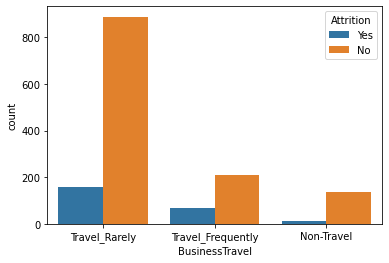
2.-

**After observing the business travel column. We found that mostly people travel rarely**



3.-

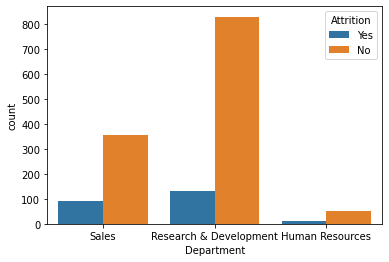
**Let us see the count of each category of business travel where employee attrition is true.**



based on the graph we conclude that portion of attrition is higher where travel is frequent

4.-

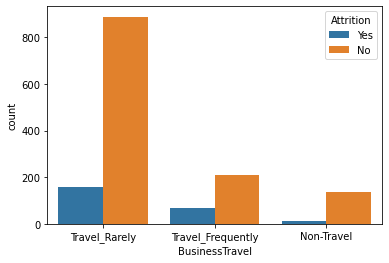
**Let us check in which department attrition is high**



here as we compare between sales and research department 1 out of 4 and 1 out of 9 ppl attrition ratio is coming

5.-

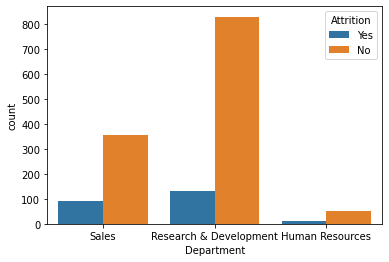
**Let us see the count of each category of business travel where employee attrition is true**



based on the graph we conclude that portion of attrition is higher where travel is frequent.

6.-

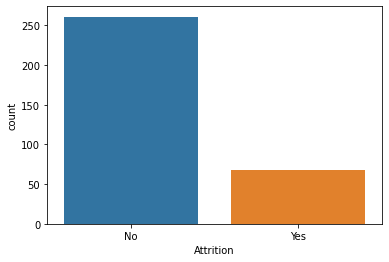
**Let us check in which department attrition is high**



here as we compare between sales and research department 1 out of 4 and 1 out of 9 ppl attrition ratio is coming.

7.-

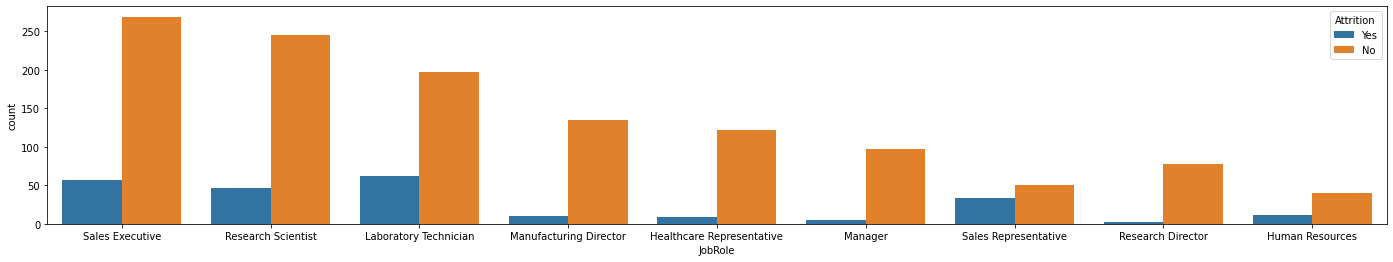
Let us check when distance from home is above 15 km what is the attrition



Not too many people effected due to the distance above the 15 km.

8.-

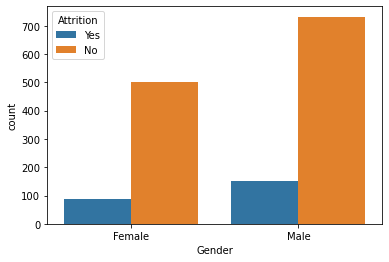
**Let us check the Attrition in each job role**



We observed from the graph Sales executive job profile person is largely attrition.

9.-

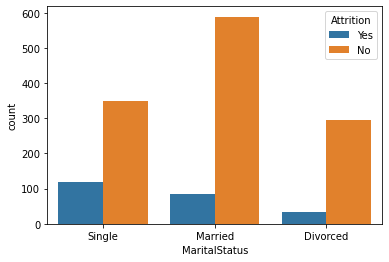
**Let us check the Attrition Ratio based on Gander**



Male employee is largely Attrition as compare to female gender employee.

10.-

**Let us Check the Attrition in Marital Status**



Married employee is more attrition

**Pre-Processing Pipeline: -**

1.-

There is column employee number in the data set its does not help us , so we drop it.

2.-

Let us see over 18 columns this column has only one value throughout the dataset, so it will not help us any way, so let us drop it.

**Building Machine Learning Model**

We observed that are target column have two option se need classification model.

We used Logistic Regression to find our best Random state and accuracy.

So, we conclude that at random state 123 our model gives the 91-percentile accuracy.

We Used following models for check accuracy.

1. Random Forest Classifier
2. Decision Tree Classifier
3. Logistic Regression

By using the following models, we find out Logistic Regression is best model.

**Concluding Remarks**

Logistic regression work as best model with 91 percentile accuracy.