**HR Analytics Project- Understanding the Attrition in HR**

1. ***Problem Definition*** :-

Employee attrition, or turnover, is a significant challenge for many organizations. High attrition rates can lead to increased costs in recruiting and training new employees, along with reduced morale among remaining staff. In this project, our goal is to predict which employees are likely to leave the company using machine learning techniques.

**Objective** :-

Build a machine learning model to predict employee attrition based on features like job satisfaction, monthly income, years at the company, and work-life balance.  
Business Impact: With accurate predictions, businesses can proactively address potential retention issues, reducing turnover and improving employee satisfaction.

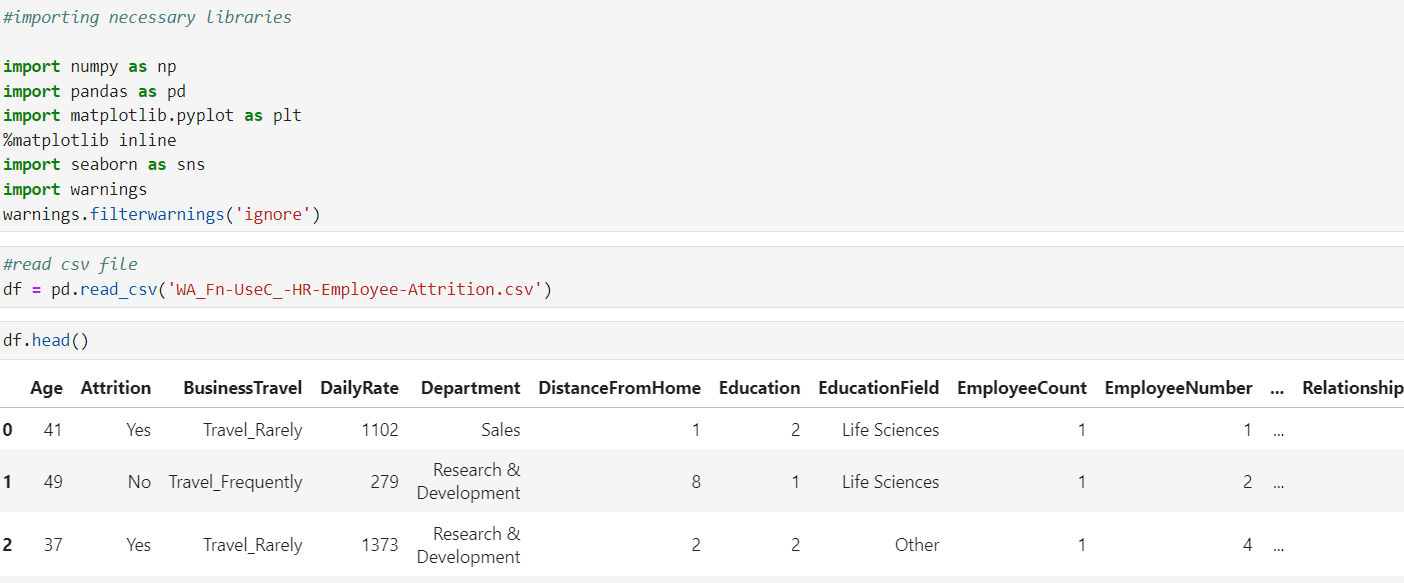
1. ***Data Analysis :-***

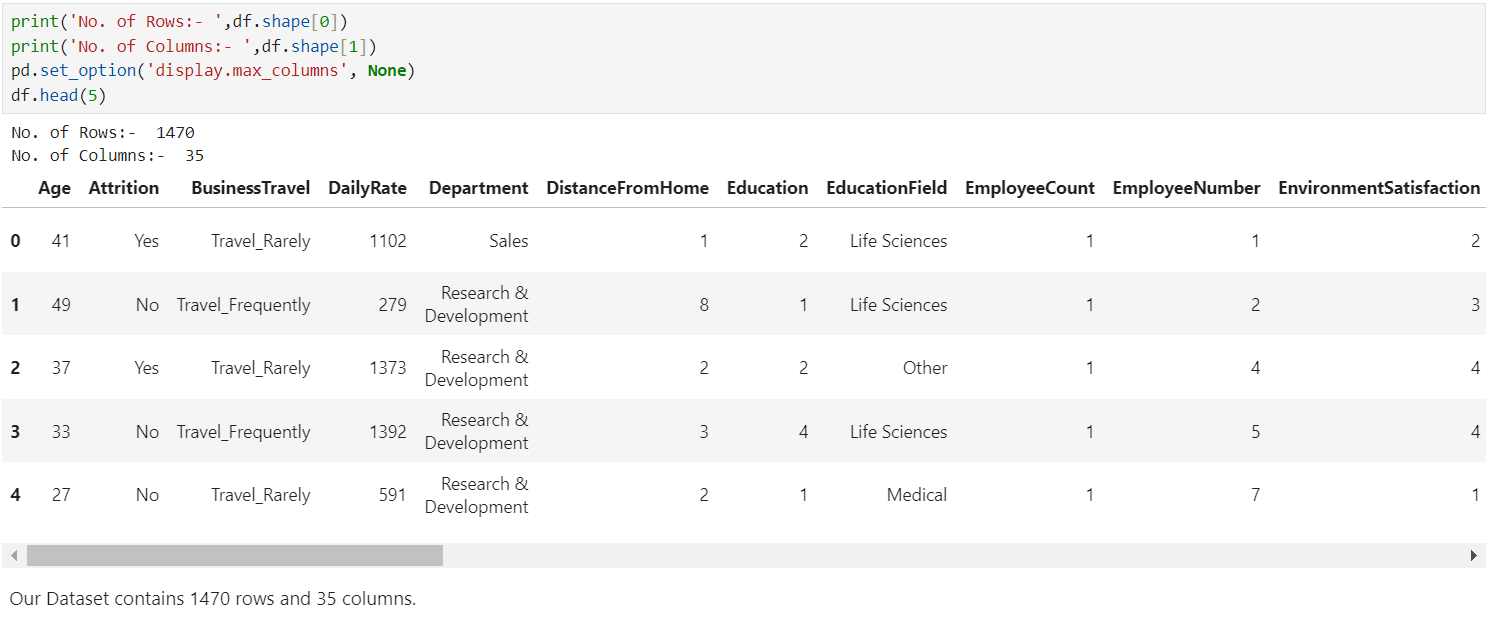
**IBM HR Analytics Employee Attrition & Performance Dataset**

In this case study we will use IBM HR Analytics database. This fictional dataset created by IBM employees and available to download from GitHub and Kaggle. You can also download dataset from my [GitHub profile here](https://github.com/Dishant-Bawankule/Internship/blob/main/Evaluation%20Project/HR%20Analytics%20project/HR%20Analytics%20Project.ipynb). This dataset consists of 1470 rows, 35 features describing each employee’s background and characteristics and target variable. Attrition is target variable to be predicted. As target variable is categorial in nature, this case study falls into classification machine learning problem. We have two objectives here:

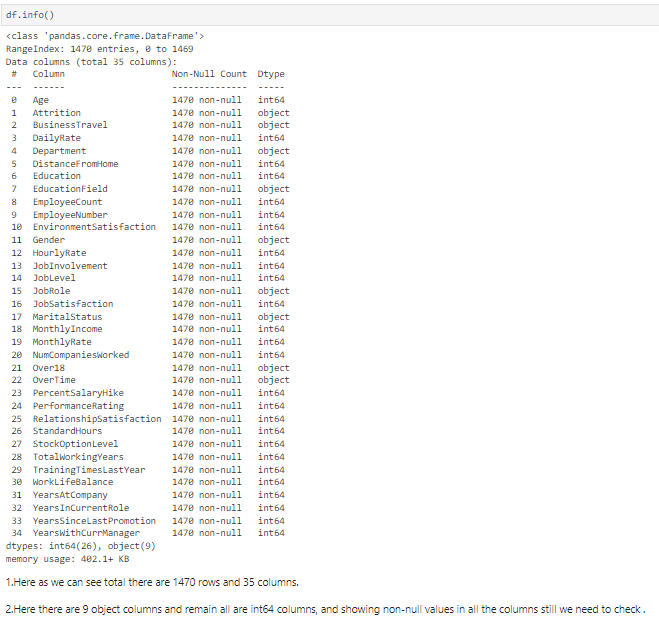
**Data Preparation: Load, Clean and Format :-**

Let’s begin with importing libraries for EDA and dataset itself

****



Checking different datatypes in dataset



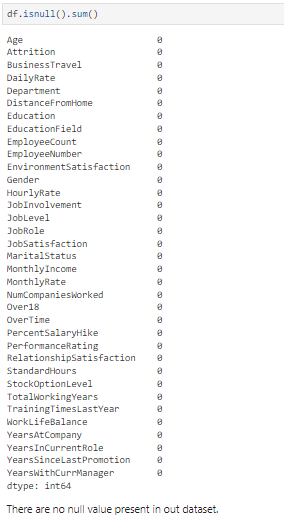
We have 9 features with object datatypes and rest are Numeric feature with int64. Out of all numeric features Education, Environment-Satisfaction, Job-Involvement, Job-Satisfaction, Relationship-Satisfaction, Performance Rating, Work Life Balance are ordinal variable. These ordinal features have unique label for each numeric value.

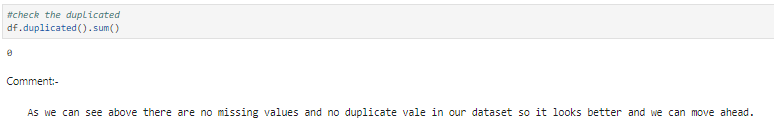
**These Ordinal features come with the following label encoding** :-

* **Education:** *1- 'Below College’, 2 -'College', 3 -'Bachelor', 4- 'Master' ,5 -'Doctor'*
* **Environment Satisfaction:** *1- 'Low', 2- 'Medium', 3 -'High', 4- 'Very High'*
* **Job Involvement:** *1 -'Low', 2- 'Medium', 3- 'High', 4- 'Very High'*
* **Job Satisfaction:** *1- 'Low', 2- 'Medium', 3- 'High', 4 -'Very High'*
* **Performance Rating:** *1- 'Low', 2- 'Average', 3 -'Good', 4- 'Excellent', 5- 'Outstanding'*
* **Relationship Satisfaction:** *1- 'Low', 2- 'Medium', 3- 'High', 4- 'Very High'*
* **Work Life Balance:** *1- 'Bad', 2- 'Good', 3- 'Better', 4- 'Best'*

Above nomenclature will help in better understanding of data when we perform EDA in this case study.

**Data Integrity Check :-**

Dataset can have missing values, duplicated entries and whitespaces. Now we will perform this integrity check of dataset.



Luckily for us, there is no missing data! this will make it easier to work with the dataset.

Dataset doesn’t contain Any duplicate entry, whitespace, ‘NA’, or ‘-’.

Statistical parameters like mean, median, quantile can give important details about database. Now is time to look at statistical Matrix of Dataset.

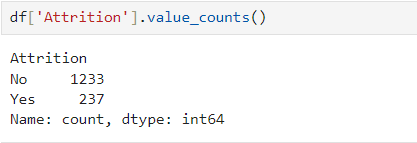
***Few key observations from this statistical matrix are listed below****: -*

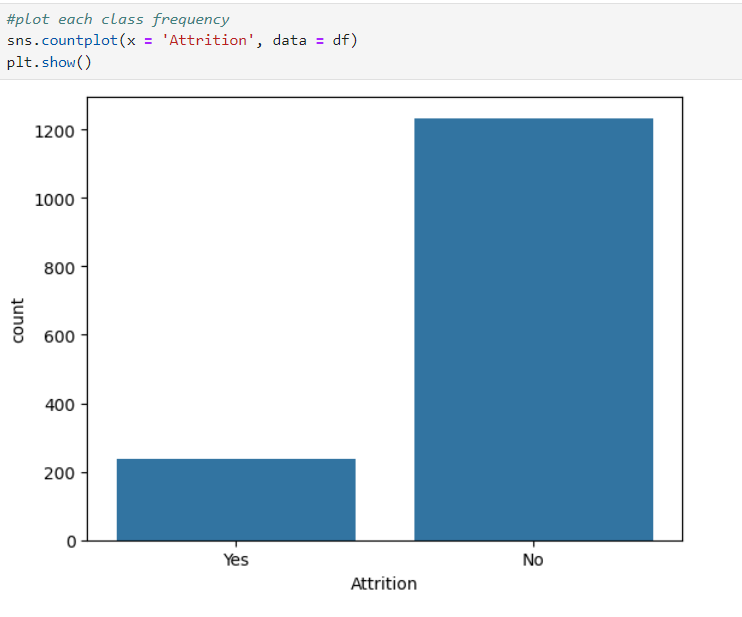
* **Minimum Employee Age is 18 and Maximum age of employee 60.**
* **Average distance from home is 9.1 KM. It means that most of employee travel at least 18 KM in day from home to office.**
* **Average performance Rating of employees is 3.163 with min value 3.0. This Means that performance of most of employee is 'Good’. This implies that Attrition of Employee with 'Outstanding' or 5 rating need to investigate.**
* **50% of Employees has worked at least 2 companies previously.**
* **For Monthly Income, Monthly Rate by looking at 50% and max column we can say outliers exist in this feature.**
* **By looking at Mean and Median we see that some of the features are skew in nature.**
* **For ordinal features statistical terminology like mean, median, std deviation are not applicable.**
* **Standard Hours and Employee Count contain same value for all statistical parameter. It means they contain one unique value.**

1. ***EDA Concluding Remarks :-***

**Exploratory Data Analysis (EDA) revealed key insights into the factors contributing to employee attrition**

Let’s begin data exploration of Target variable using count plot

****



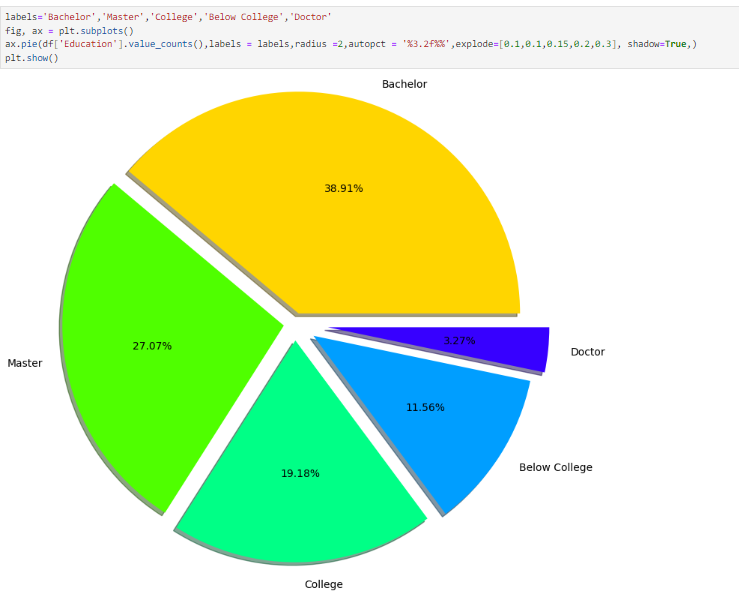
83.88% (1237 employees) Employees did not leave the organization while 16.12% (237 employees) did leave the organization *making our dataset to be consider as imbalanced*since more people stay in the organization than they actually leave.

In this dataset we have features like education, department, education field, job role, job satisfaction which are inter related with each other. Job role & job position not in alignment with educational background can lead attrition. Let investigate this by visualisation of these features one by one to gain more insights.

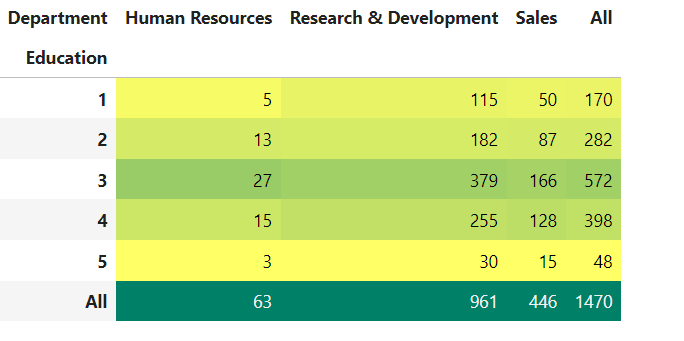
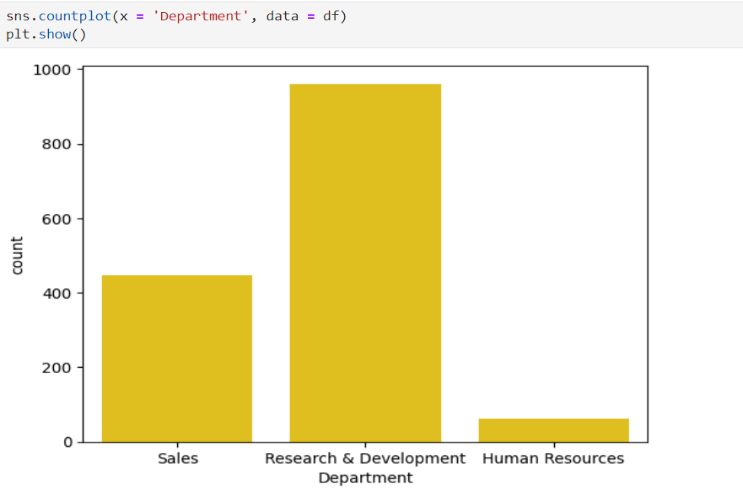
**Education level of Man power available:-**

Key Insights from Pie Plot

1. More than 38 % employees educated at Bachelor level.
2. 30 % of Employees are highly educated which involves master and doctor degree.
3. Almost 19% Employees are educated up to college & 12% are below college.

****

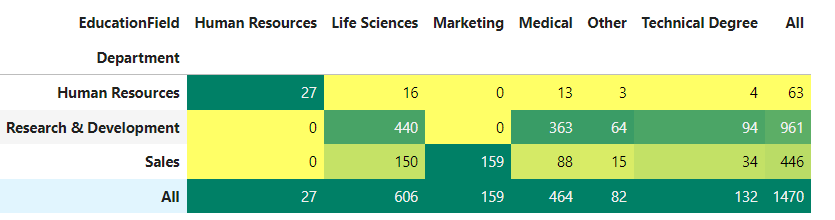
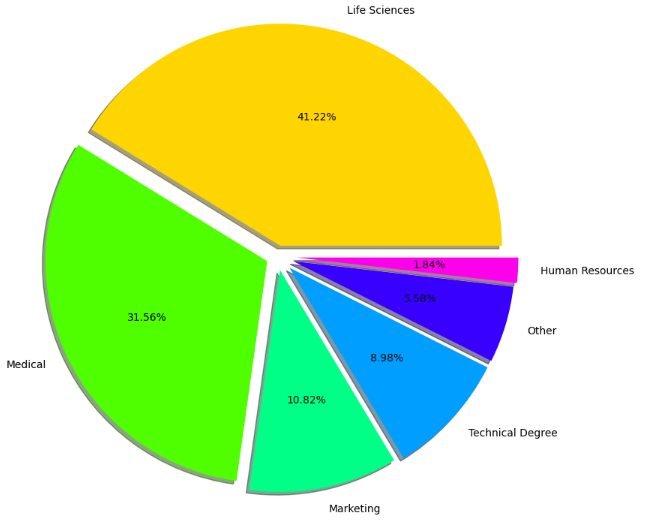
Department wise Distribution of Man power :-



Key Insights on Department and Education Level of employee in each Department

1. 65.37% of Employees work inside Research & Development Department. Out of Total 961 Employee number of employees with education level of Bachelors, Masters, Doctor are 379, 255 and 30 respectively.
2. Only 63 Employee work in HR department.

**Employee distribution as per education field :-**

** **

The probability of Employees Retention is more when there working domain is in alignment with education background. Let check this with crosstab of department against education field

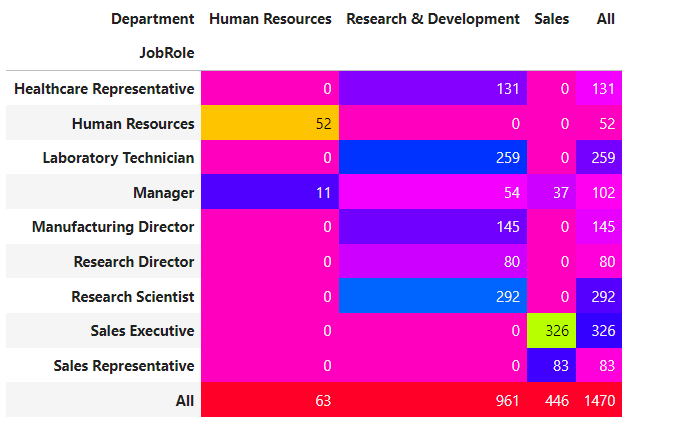
Key Insights from Pie Plot

1. Employees belongs to six different domains.
2. 41.22 % Employee comes from Life science background followed by medical profession with 31.56%
3. Least number of Employees comes from HR background.

Key Insights from above Cross Tab:

* There are only 27 people with HR background and We know that 63 people work in HR Department from previous result. This implies that at least half employee working in HR department do not have HR background.
* R&D department almost everyone comes from domain expertise or technical background except support staff. These employees usually have high salary, so it will be interesting to investigate attrition in this category.
* There are 159 Employee with Marketing background and all work in Sales Department.
* 50% Employees in sales department have background of Life sciences & Medical. We can clear see they are working in domain to which their educational background does not belong. So, it will be interesting to see attrition rate in these employees.

We will Analyse Attrition in department according to education background based on above insight further but before that explore Job role in order to include it in further attrition analyse. First build matrix of department vs job role which will give us idea about number of employees of different job role across department.



Key Insights from above Cross Tab:

* There are 3 job roles in HR Department, maximum of which are sales Executive with 446 Total Employees.
* Human Resources department has 2 Job role i.e., HR & Manager.
* There 6 different Job role in R&D department with total 961 employees and until now we know that all of them belong to their respective domain background.

**These findings highlight that compensation, work-life balance, and engagement play crucial roles in employee retention.**

1. ***Pre-processing Pipeline :-***

*Pre-processing is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.*

Feature Engineering is very important step in building Machine Learning model. Some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. In Feature engineering can be done for various reason. **Some of them are mention below:**

1. **Feature Importance**: An estimate of the usefulness of a feature
2. **Feature Extraction**: The automatic construction of new features from raw data (Dimensionality reduction Technique like PCA)
3. **Feature Selection**: From many features to a few that are useful
4. **Feature Construction**: The manual construction of new features from raw data (For example, construction of new column for month out date - mm/dd/yy)

There are Varity of techniques use to achieve above mention means as per need of dataset. Some of Techniques important are as below:

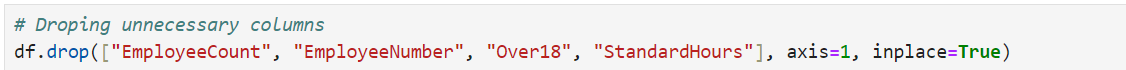
* Handling missing values
* Handling imbalanced data using SMOTE
* Outliers’ detection and removal using Z-score, IQR
* Scaling of data using Standard Scalar or Minmax Scalar
* Binning whenever needed
* Encoding categorical data using one hot encoding, label / ordinal encoding
* Skewness correction using Boxcox or yeo-Johnson method
* Handling Multicollinearity among feature using variance inflation factor
* Feature selection Techniques:
* Correlation Matrix with Heatmap
* Univariate Selection – SelectKBest

ExtraTreesClassifier method

**In this case study we will use some of the mention feature engineering Techniques one by one.**

1. **Dropping unnecessary features :-**

Feature like ‘Over18’, ‘Standard Hours’ contain single unique value. Features like Employee Count, Employee Number are irrelevant from ML model building perspective. We will drop these features.



1. **Encoding Categorical & Ordinal Features :-**

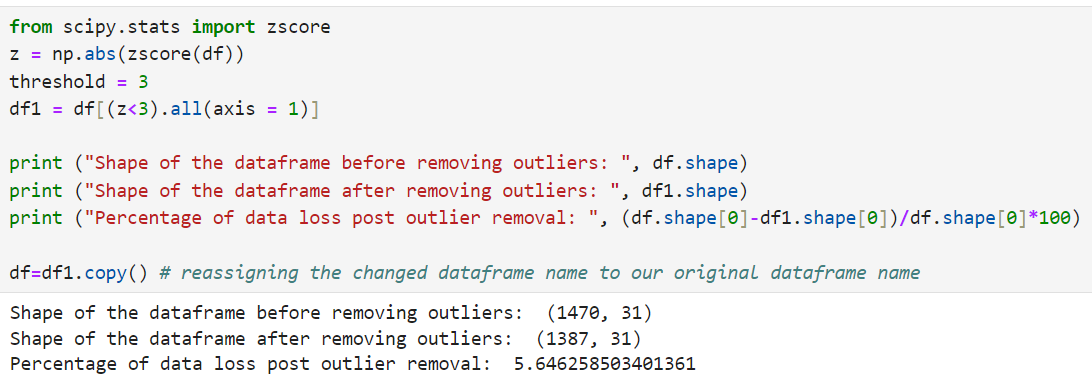
Label Encoding is employed over target variable ‘Attrition’ while Ordinal encoding employ for rest categorical features.

****

Since now encoding is done we will move towards outliers’ detection and removal

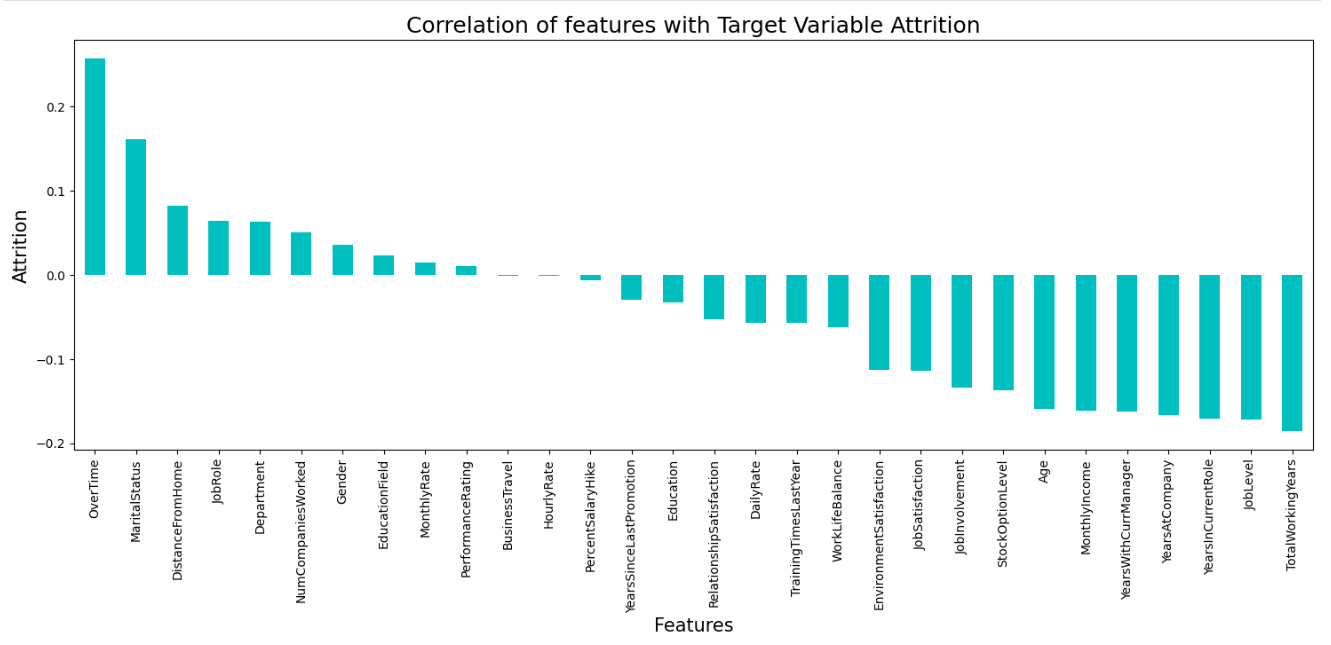
1. **Outliers’ detection and removal :-**

Machine learning algorithms are sensitive to the range and distribution of attribute values. Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models and ultimately poorer results. Outliers can be seen in boxplot of numerical feature. We did not added boxplot here as it will make this article length, I left it to reader to further investigate. Now we will use Z-score method for outliers’ detection.

****

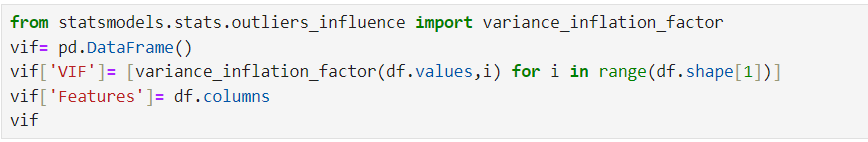
1. **Correlation Heatmap :-**

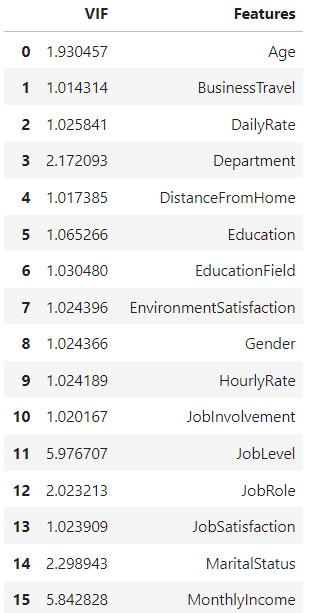
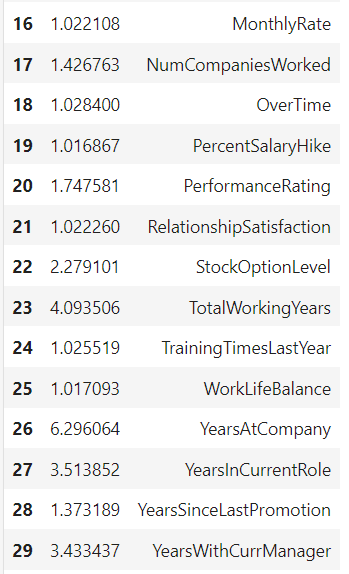
Correlation Heatmap show in a glance which variables are correlated, to what degree, in which direction, and alerts us to potential multicollinearity problems. The bar plot of correlation coefficient of target variable with independent features shown below

****

1. **Multicollinearity between features :-**

Variance Inflation factor imported from statsmodels.stats.outliers\_influence to check multicollinearity between features.

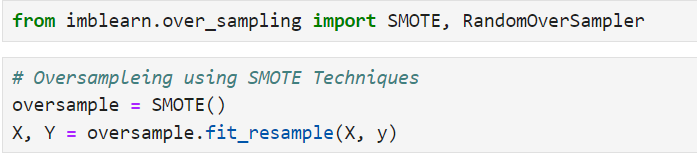


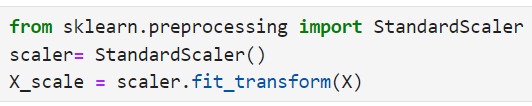
We can see that for all features Variance inflation factor in within permissible limit of 10. Multicollinearity do not pose any threat here.

1. **Handling imbalanced data using SMOTE :-**

This two-class dataset is imbalanced (84% vs 16%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented class. We can resolve this by Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class.

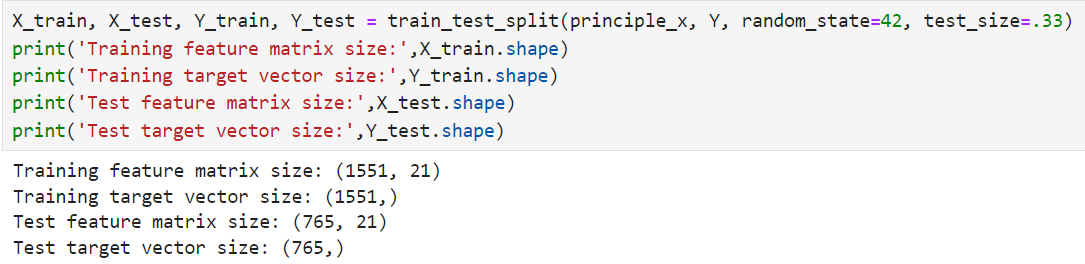


1. **Scaling of data using Standard Scalar :-**

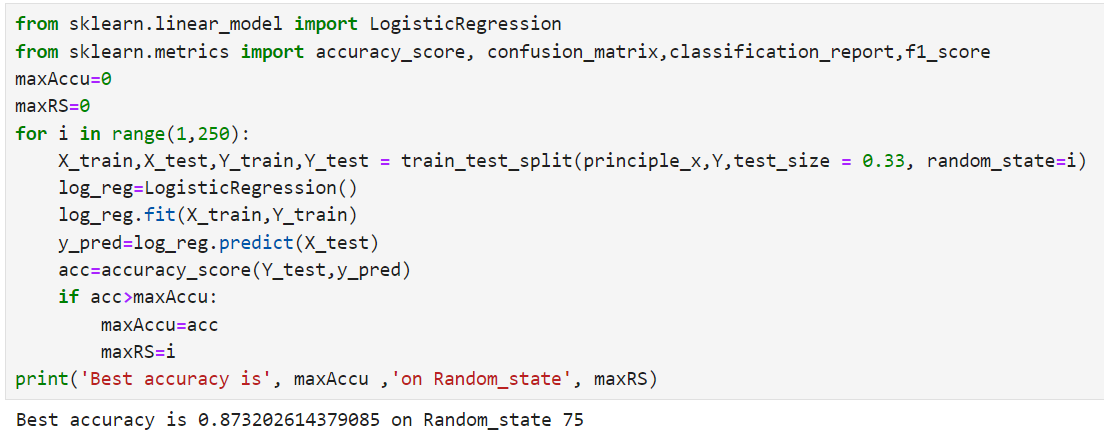


1. ***Building Machine Learning Models :-***

In this section we will build Supervised learning ML model-based classification algorithm. As objective is to predict attrition in ‘Yes’ or ‘No’ leads to fall problem in domain of classification algorithm. train\_test\_split used to split data with size of 0.33

******

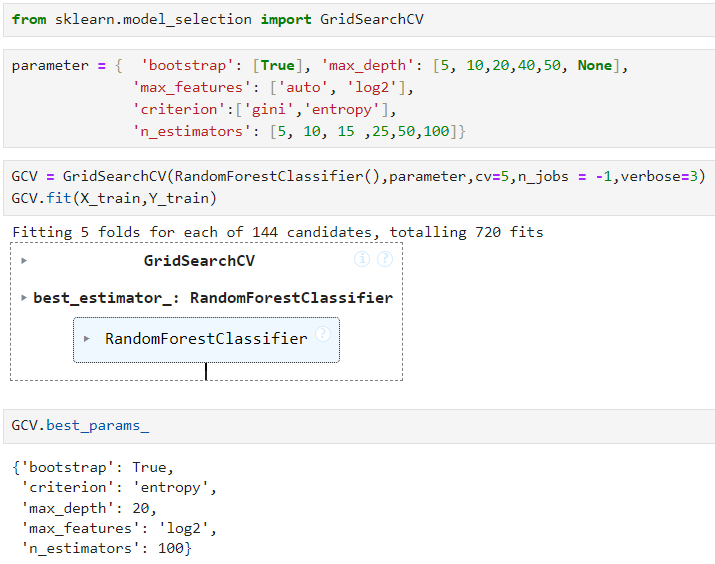
First we will build base model using logistic regression algorithim. Best random state is investigated using for loop for random state in range of (0,250).



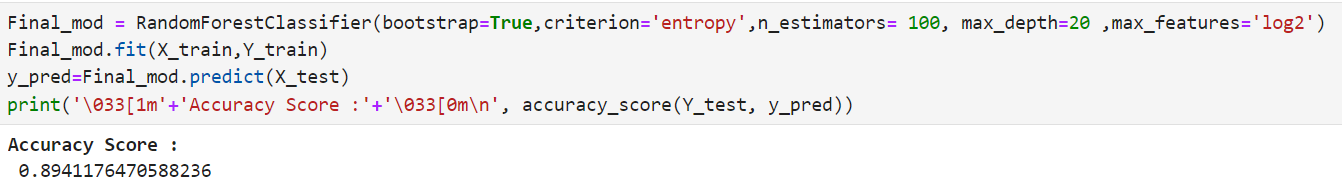
As Now base model is ready with f1-score of 0.86, we will train model with different classification algorithm along with k-5 fold cross validation. The final evaluation matrix different classification algorithm is as shown table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML Algorithm | Accuracy Score | f-1 Score | Recall | Precision |
| Logistics Regression | 0. 8614 | 0.86 | 0.86 | 0.85 |
| SVC | 0.9137 | 0.91 | 0.93 | 0.90 |
| GaussianNB | 0.8562 | 0.85 | 0.87 | 0.84 |
| DecisionTreeClassifier | 0.8261 | 0.82 | 0.82 | 0.82 |
| KNeighborsClassifier | 0.8405 | 0.82 | 0.74 | 0.92 |
| RandomForestClassifier | 0.8967 | 0.90 | 0.92 | 0.88 |
| AdaBoostClassifier | 0.8418 | 0.84 | 0.84 | 0.83 |
| GradientBoostingClassifier | 0.8758 | 0.87 | 0.88 | 0.86 |
| Bagging Classifier | 0.8509 | 0.85 | 0.88 | 0.8 |

**We can see that Random Forest Classifier gives us maximum f1-score & mean cross validation score. We will perform hyper parameter tuning on random forest classifier to build final ML Model**.

****

Next step is to build final machine learning model over best params in Hyper parameter tuning



We can see that Final model with hyper parameter tuning leads to slight decrease in accuracy score from 0.8980 in original model to 0.8915.

1. ***Concluding Remarks***

* Bench mark of 6900$ monthly income is recommended to Prevent attrition.
* Attrition rate is high in age group of 29 to 33. HR need to keep eye over need & expectation of this age group from company.
* Percentage of attrition is high in Sales Representative, Laboratory Technician
* 16 % attrition rate among Research Scientist and no company afford to lose them.
* Almost 50% employs in sales department from different education background. There is possibility of dissatisfaction among them as attrition high among these.
* Different feature engineering techniques like balancing data, outliers’ removal, label encoding, feature selection & PCA are perform on data.
* Random Forest Classifier model gives maximum Accuracy.

You can get code of this case study from my [GitHub Profile](https://github.com/Dishant-Bawankule/Internship/blob/main/Evaluation%20Project/HR%20Analytics%20project/HR%20Analytics%20Project.ipynb).