**Machine Learning to Understand & Predict HR Attrition**

Do you want to know what is that one single thing for grand success of any business Enterprise, startup in current competitive era, where razor edge like fine technological upgradation take place continuously along with continuous evolving SOP’s, business model? Fundamental answer to this question trace back to core of business which is need & demand of product.

***So next Super Fundamental question comes here: from where product comes? Product comes from innovative ideas & solution from talented minds and effort of domain expertise people making idea into reality, a useful product.***

The key to success in an organisation is the ability to attract and retain top talents. But where Machine learning comes here? For that we need to go in background of HR analytics.

HR department hire lot of employees every year. 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. HR department often play significant role in designing company compensation programs, creating ambiance in work environment with various team activities, imprinting work culture habits on mindset of peoples and training skill systems that help the organization retain smart minds & talented brain. Most of organisation run different HR analytics vertical to gather data, analyse data to improve process within organisation and to make major decision about human resources.

***The gradual loss of employees’ overtime refers as HR attrition. Attrition can happen due to retirement, involuntary (employee is fired or terminated) or voluntary (resigns from an organization).***

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. For Tech companies’ loss of talented brain means loss of domain expertise, knowledge base as for these company’s Technological capability exists in domain expertise, intellectual property rights of employee. On marketing or sales side customers often prefer to interact with familiar people.

Just like several factors contribute towards building a reliable team and organization, similarly, numerous factors contribute to the attrition rate: Improper work-life balance, better job opportunities, salary hike, Lack of growth or work recognition, unhealthy relations with managers. We can gather data about these factors and utilise Machine learning classification techniques to predict whether employee like to leave organisation or stay in organisation. Here I will show you what leads employee’s attrition and Predication of attrition using case study on IBM HR analytics Dataset.

**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. 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:

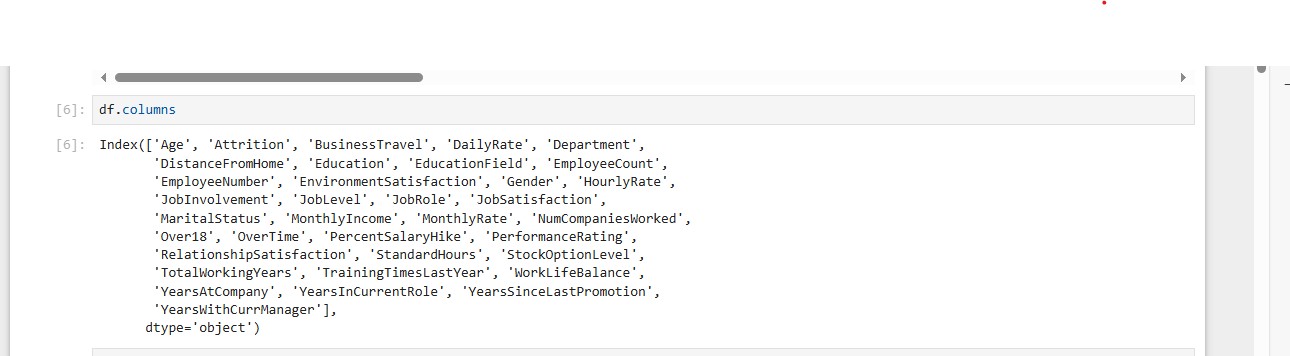
1. Which key factors result in employee attrition?
2. Building ML Model for predicting attrition.

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

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



Checking different datatypes in dataset: -

****

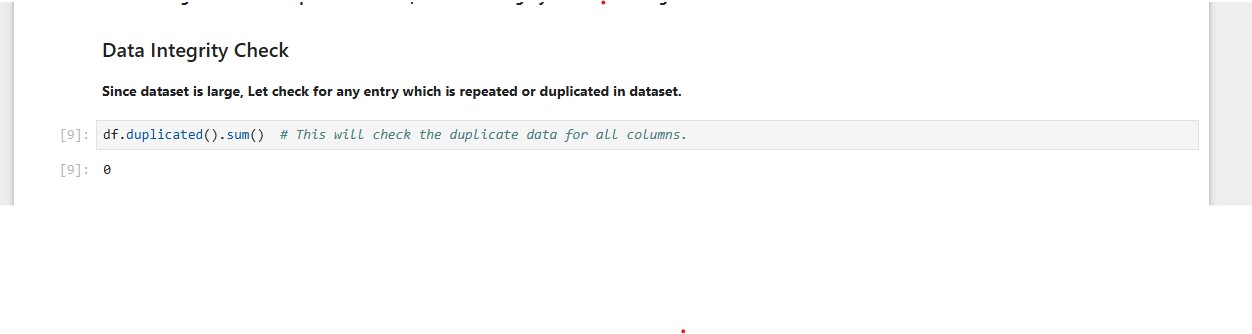
We have 9 features with object datatypes and rest are Numeric feature with int64. Out of all numeric features Education, Environment-Satisfaction, Job-Involvement, Job-Satisfaction, Relationship-Satisfaction, Performance Rating, Work Life Balance are ordinal variable. These ordinal features have unique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

**So, Yes Good to Go Further!!!**

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

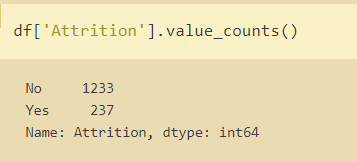
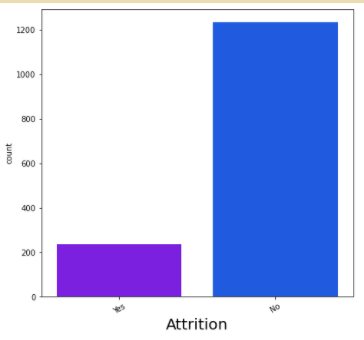
*Few key observations from this statistical matrix are listed below: -*

* 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.

**Exploratory data analysis**

**Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover PATTERNS, TO spot ANOMALIES, TO test hypothesis and to check assumptions with the help of summary statistics and graphical representations.**

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

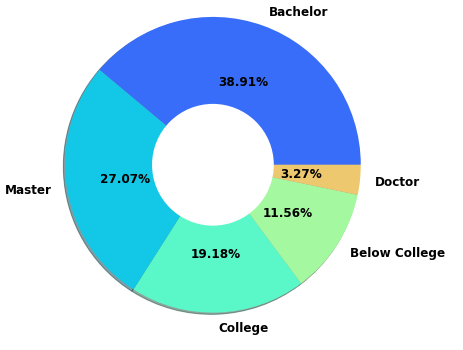
83.88% (1237 employees) Employees did not leave the organization while 16.12% (237 employees) did leave the organization ***making our dataset to be consider as imbalanced***since more people stay in the organization than they actually leave.

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.

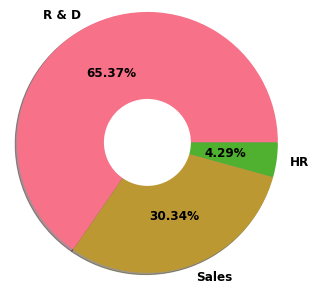
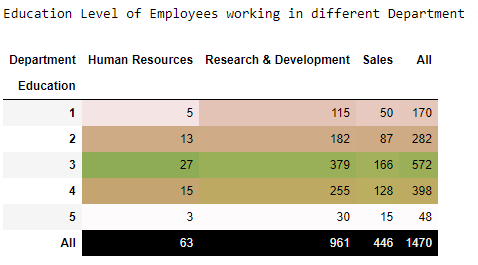
Education level of Man power available:

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.

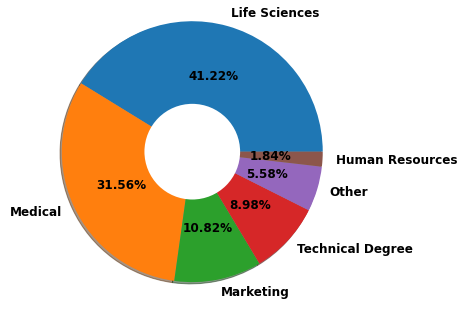
****

Department wise Distribution of Man power:

****

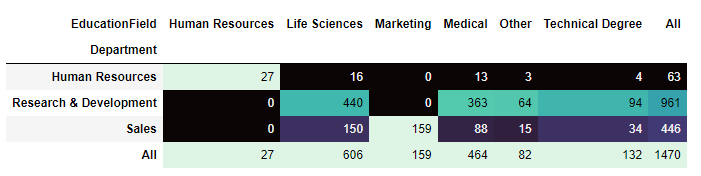
Key Insights on Department and Education Level of employee in each Department

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:**

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.

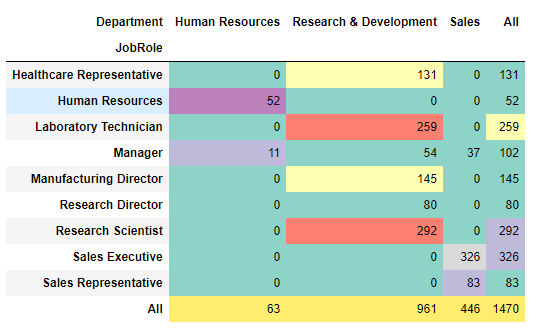


The probability of Employees Retention is more when there working domain is in alignment with education background. Let check this with crosstab of department against education field.

Key Insights from 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.

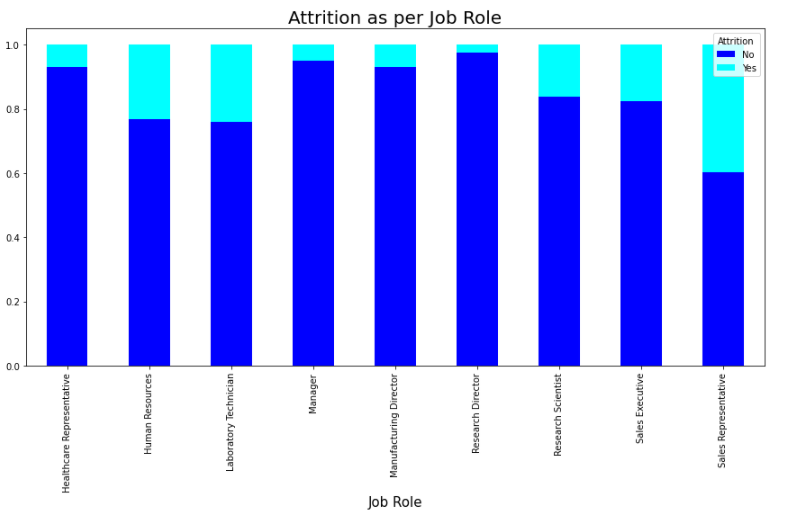
We will Analyse Attrition in department according to education background based on above insight further but before that explore Job role in order to include it in further attrition analyse. First build matrix of department vs job role which will give us idea about number of employees of different job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



Key Insights from above Cross Tab:

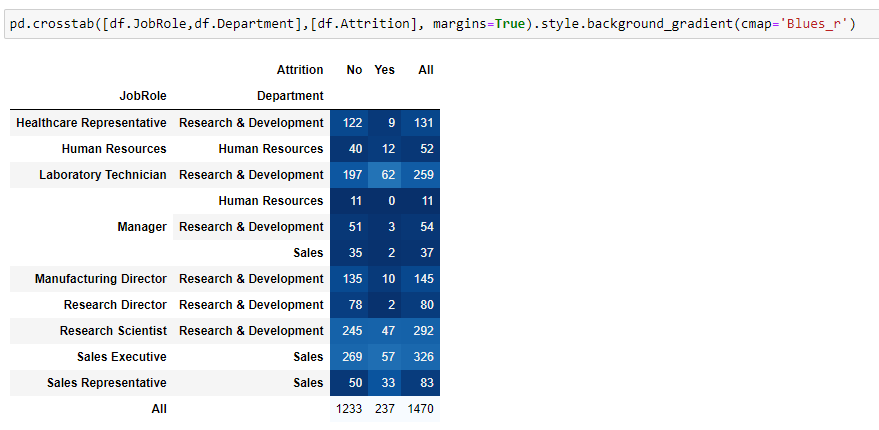
* There are 3 job roles in HR Department, maximum of which are sales Executive with 446 Total Employees.
* Human Resources department has 2 Job role i.e., HR & Manager.
* There 6 different Job role in R&D department with total 961 employees and until now we know that all of them belong to their respective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

**Attrition by Job role :**

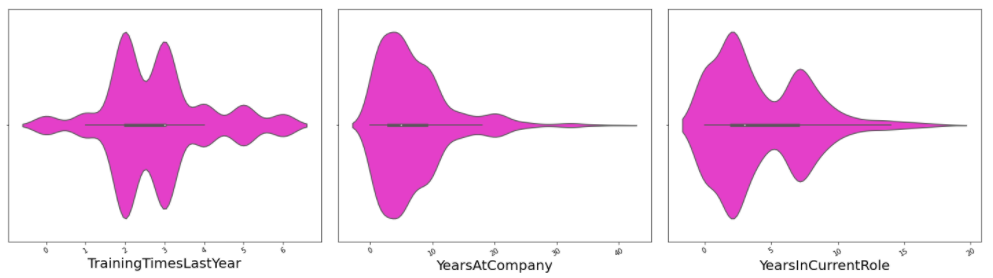
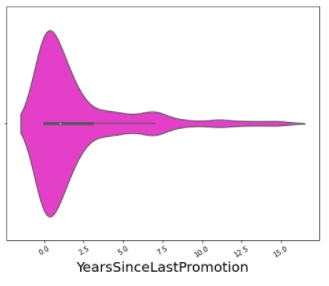
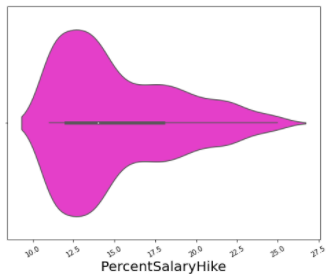
****

Key Insights on Department & Education level of from Pie Plot

Let’s check absolute number matrix of attrition according job role, again this time using crosstab.

Key Insights from above Cross Tab:

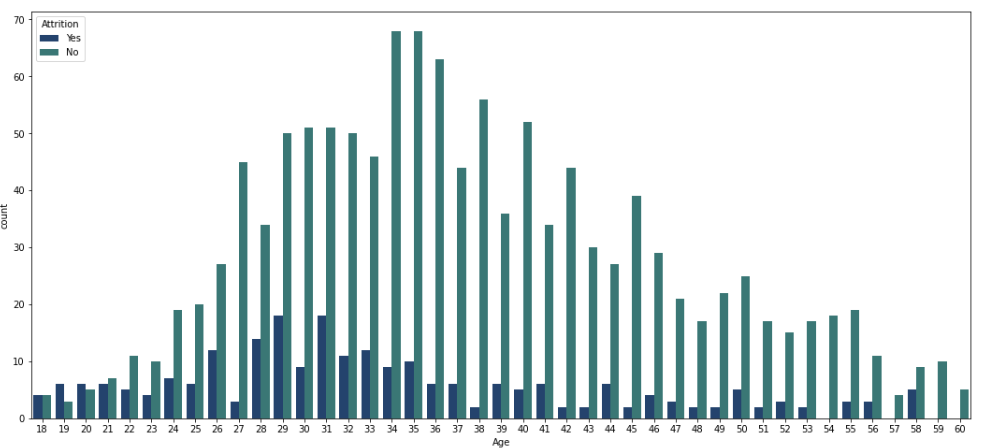
* Percentage of attrition is high in Sales Representative, Laboratory Technician, Human Resources.
* At the Top chart 62 Laboratory Technician has resign from job, followed by 57 sales executive and 47 Research Scientist.
* 16 % attrition rate for Research Scientist, which involve huge investment from company. Company not only loses employee but its knowledge base, expertise & Intellectual property rights in some cases.

Let’s check violin plot of some numerical features to gain more insight.

Key Insights from above Cross Tab:

* For Majority of people have spent 3 to 10 years at company.
* Most of people staying company up to 2 years after promotion.
* Majority of people are train 2-3 times in last year
* Majority of people stay in same role for maximum 4 yrs.
* Majority of Employees have salary hike of 10 to 15%.

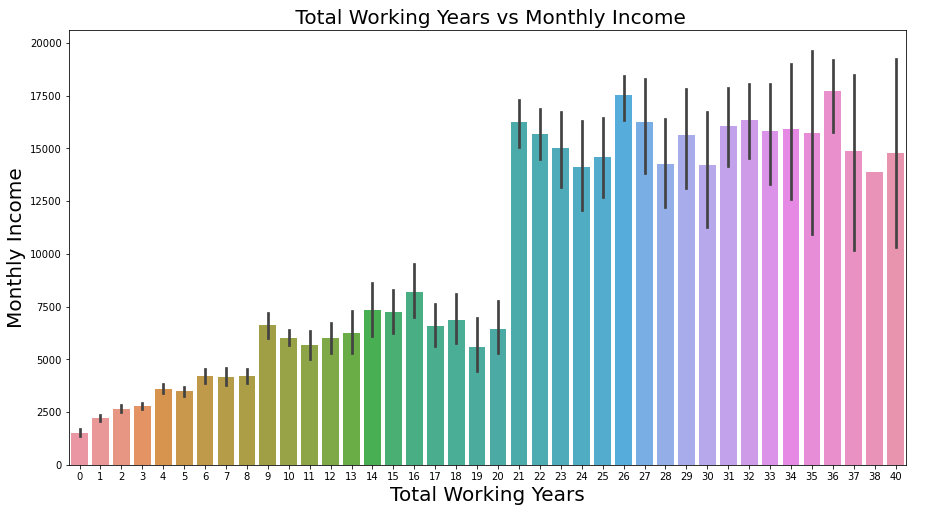
Q. In which age group attrition rate is high?



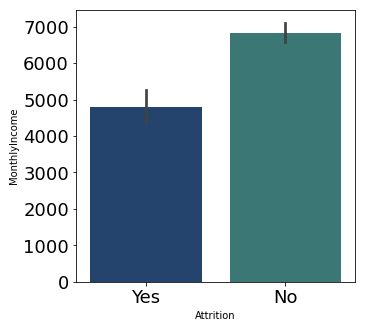
Key Insights from count plot of Age Vs Attrition:

1. The Attrition rate is minimum between the Age years of 34 and 45.
2. The Attrition rate is maximum between the Age years of 29 and 33.

Q. What is variation in monthly income as Total working year increases.

****

Monthly Income is higher for the employees with 21 or more number of Total working years. For first 9 years monthly income is less than 5000$. **But what about attrition, let’s bar chart of Monthly income so we can come across some benchmark of average monthly income in both attrition categories.**

****

*Key Insights on Average Monthly Income as per attrition*

* We can see that Average monthly income is less in employees who choose to resign compare to rest. Less Monthly Income is major reason behind attrition.
* To prevent attrition average monthly to be greater than 6900$ is recommended.

**Feature Engineering: Data Pre-processing**

*Feature engineering 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.

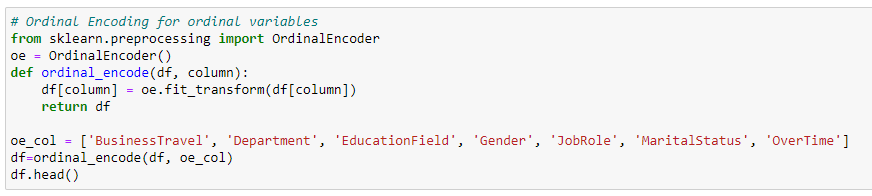
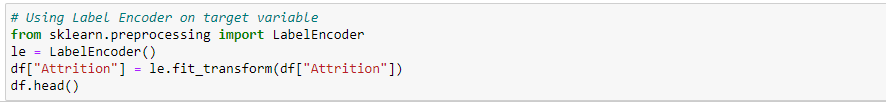
1. **Dropping unnecessary features**

Feature like ‘Over18’, ‘StandardHours’ contain single unique value. Features like EmployeeCount, EmployeeNumber 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.

****

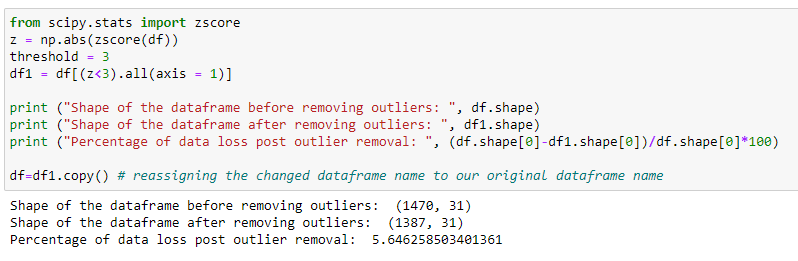
Since now encoding is done we will move towards outliers’ detection and removal.

1. **Outliers’ detection and removal**

Identifying outliers and bad data in your dataset is probably one of the most difficult parts of data clean-up, and it takes time to get right. Even if you have a deep understanding of statistics and how outliers might affect your data, it’s always a topic to explore cautiously.

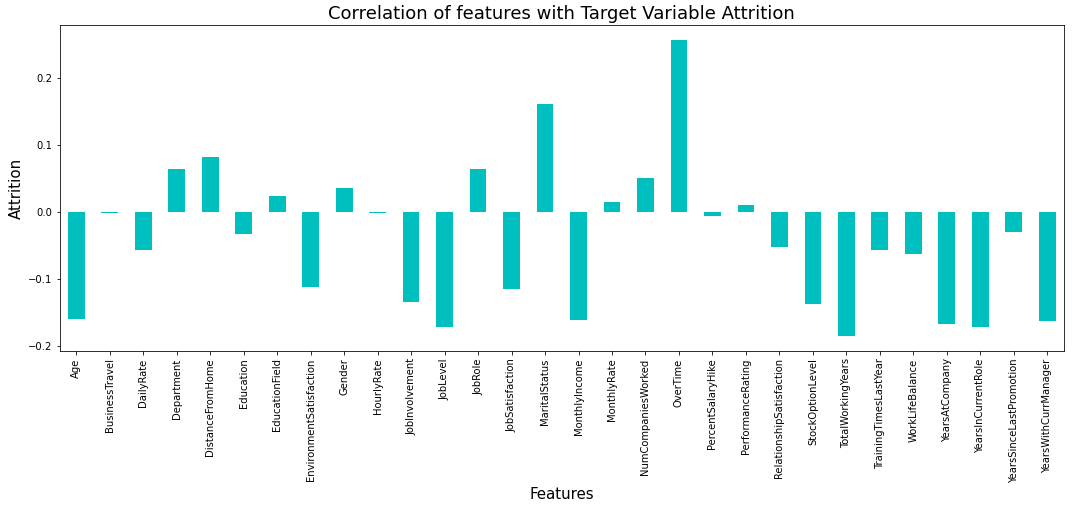
* Page 167, Data Wrangling with Python, 2016

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.

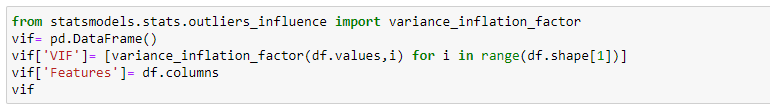


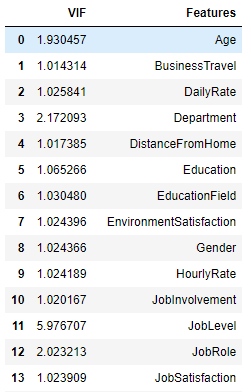
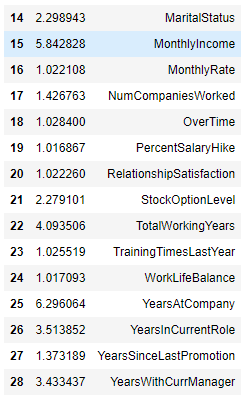
1. **Correlation Heatmap**

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**

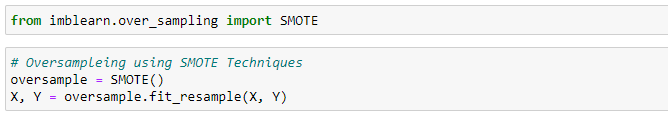
Variance Inflation factor imported from statsmodels.stats.outliers\_influence to check multicollinearity between features.



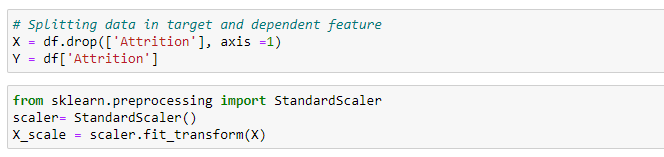
We can see that for all features Variance inflation factor in within permissible limit of 10. Multicollinearity do not pose any threat here.

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.

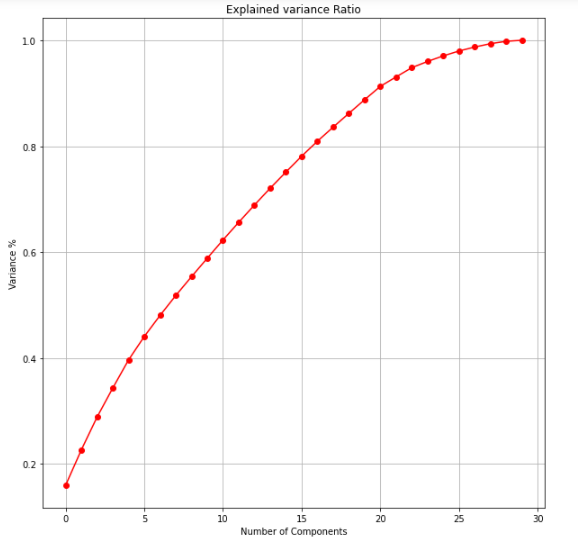


1. **Scaling of data using Standard Scalar**

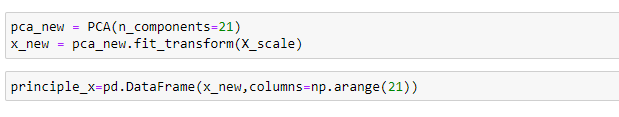


1. **Dimensionality Reduction Using PCA**

PCA used find patterns and extract the latent features from our dataset.

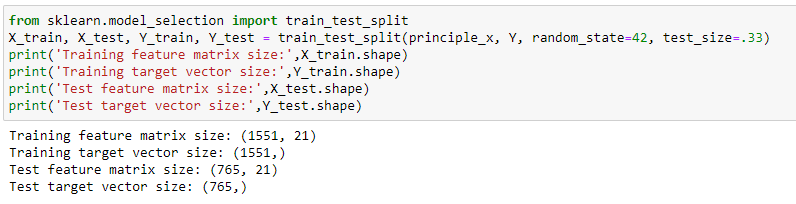


We can see that 21 principal components attribute for 90% of variation in the data. PCA applied for 21 components.

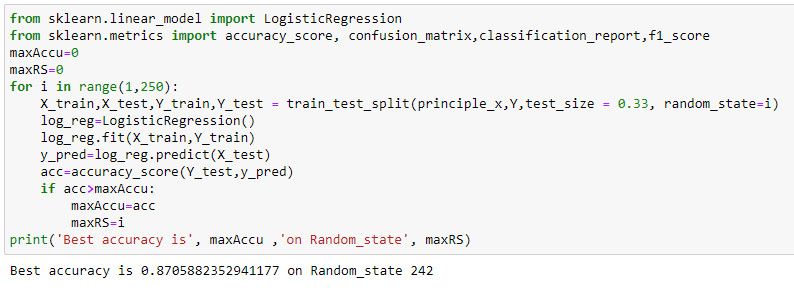


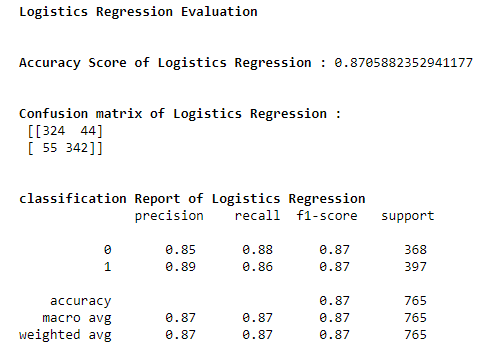
**Machine Learning Model Building:**

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).



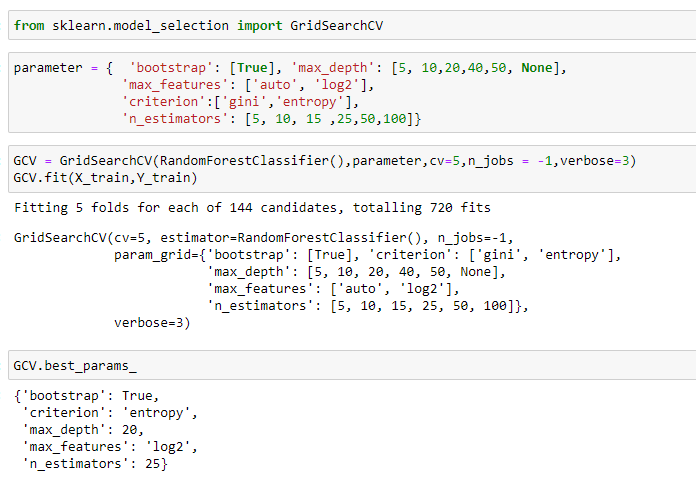
Logistics regression model is train with random state 242. The evalution matrix along with classification report is as below :

As Now base model is ready with f1-score of 0.87, 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:

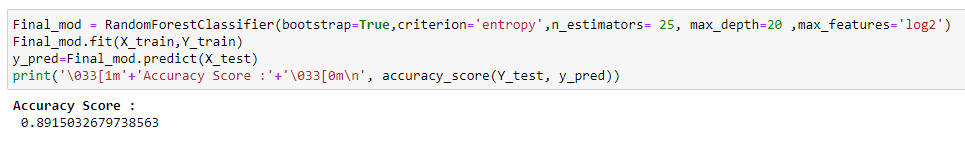
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ML Algorithm | Accuracy Score | CV Mean Score | f-1 Score | Recall | Precision |
| Logistics Regression | 0.8705 | 0.6869 | 0.87 | 0.87 | 0.87 |
| SVC | 0.9019 | 0.6075 | 0.90 | 0.90 | 0.90 |
| GaussianNB | 0.8470 | 0.7405 | 0.85 | 0.85 | 0.85 |
| DecisionTreeClassifier | 0.8039 | 0.8381 | 0.80 | 0.80 | 0.80 |
| KNeighborsClassifier | 0.8379 | 0.7374 | 0.84 | 0.84 | 0.85 |
| RandomForestClassifier | 0.8980 | 0.9171 | 0.90 | 0.87 | 0.93 |
| AdaBoostClassifier | 0.8457 | 0.8718 | 0.85 | 0.85 | 0.85 |
| GradientBoostingClassifier | 0.8560 | 0.8831 | 0.86 | 0.86 | 0.86 |
| Bagging Classifier | 0.8792 | 0.8856 | 0.87 | 0.87 | 0.88 |

(Min Value in column -Green, Max Value in column - Pink Colour )

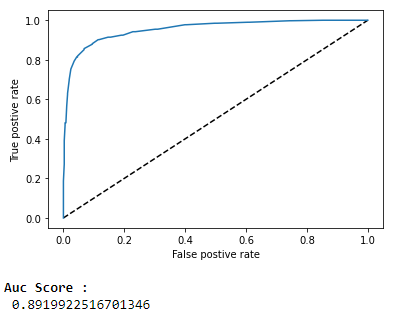
We can see that Random Forest Classifier gives us maximum f1-score & mean cross validation score. We will perform hyper parameter tuning on random forest classifier to build final ML Model.



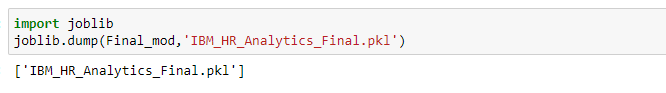
Next step is to build final machine learning model over best params in Hyper parameter tuning.



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. This complete possible We will use model with default values as our final model. AOC-ROC score of final random forest classifier model is shown below:



At last, we will save final model with joblib library, so it can be deploy on cloud platform.



**Concluding Remarks on EDA and ML Model**

* *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.*.