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

**Problem Statement:**

Every year a lot of companies hire a number of 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. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

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 in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

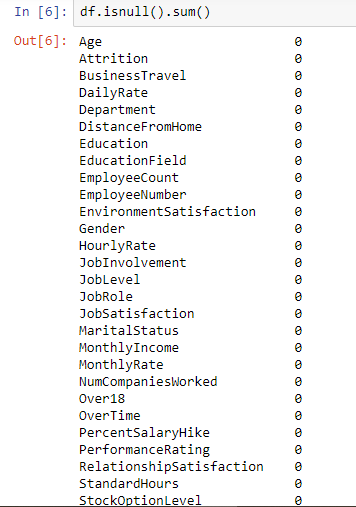
How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**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.

**Analysis Done on the Projects (Insight Gathering):-**

We have checked for the missing values in the dataset as it affects our model prediction to some extent and in some cases it leads to invalid conclusions in the case of small dataset , however on the contrary if the dataset is large enough we can remove the row completely containing the null values but that should be our last approach as any form of data loss whether big or small is not a good practice for a Data Scientist . We can find out null values by using inbuilt libraries in python i.e, isnull() / isna() or we can even visualize it using seaborn library which includes heatmap , histogram etc.



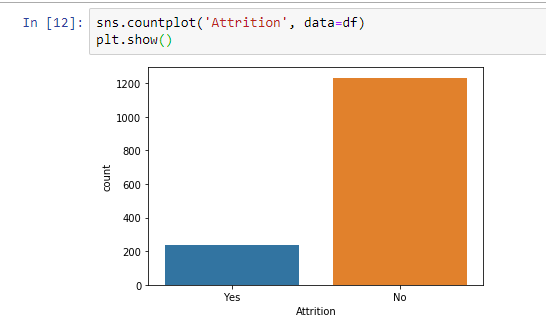
Their were no Null Values in our Dataset. So we do not have to worry about it in case if there is any missing data is present in our dataset we can treat it using any of the techniques available in machine learning . Details of which are mentioned in the link below:

<https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/>

Always try take references from the Web in case you are stuck anywhere.

Now we will check if there is any type of class Imbalance or not class imbalance can be referred to as the number of Observations belonging to one class is significantly lower than the other class in case of classification problems where the output is basically 0 & 1 , True or False , Yes or No.

These type of class imbalance problem causes the predictive model that is developed using machine learning algorithms could be biased and inaccurate.



Here our dataset has class imbalance which can be treated by various approaches:-

1. Data Level Approach : Resampling Techniques

I ) Random Under-Sampling

II ) Random Over-Sampling

III ) Cluster Based Over-Sampling

IV ) Informed Over-Sampling : Synthetic Minority Over-Sampling Technique (SMOTE)

V ) Modified Synthetic Over-Sampling Techniques (MSMOTE)

1. Algorithmic Ensemble Techniques

I ) Bagging Based techniques for Imbalanced Data

II ) Boosting Based Techniques for Imbalanced Data

i )Adaptive Boosting-Ada Boost techniques for Imbalanced Data.

ii ) Gradient Tree Boosting Techniques for Imbalanced Data.

Iii ) XG Boost Techniques for Imbalanced Data.

For Detailed Explanation of the following approaches you can refer to :

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/>

Now we will check for the data type of the target variable whether the data is continuous.

We in this dataset use SMOTE for treating our class imbalance following is the code attatched for your understanding on how to perform SMOTE on the dataset.

This technique is followed to avoid overfitting which occurs when exact replicas of minority instances are added to the main dataset. A subset of data is taken from the minority class of an example and then new synthetic similar instances are created. These synthetic instances are then added to the created. These synthetic instances are added to the original dataset.

The new dataset is now used as a sample to train the classification model.

Ex:- Total Observations=1000

Fraudulent Observation=20

Non Fraudulent Observation=980

Event Rate=2%

A sample of 15 instances is taken from the minority class and similar synthetic instances are generated 20 times

Post generation of synthetic instances, the following data set is created

Minority Class (Fraudulent Observation)=300

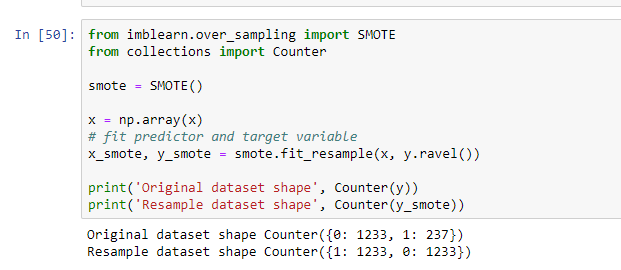
Majority Class (Non Fraudulent Observation)=980

Event Rate=300/1280=23.4%

Now the minority instances are increased by replicating the dataset . Now the classification model development is more accurate and reduces invalid Output.

Following is the Code Snippet from the Project.

Advantage of SMOTE is that there is no loss of useful information whereas on the opposite side while generating synthetic examples SMOTE does not take into consideration neighbouring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise and also it is not very efficient for high dimensional data.



Now after replication we use this dataset as a training dataset.

**Exploratory Data Analysis:-**

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data Scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals.

**EDA Tools:-**

Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.

Univariate visualization of each field in the raw dataset, with summary statistics.

Bivariate visualization and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you’re looking at.

Multivariate visualizations , for mapping and understanding interactions between different fields in the data.

K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K-Groups.

Predictive models , such as linear regression use statistics and data to predict outcomes.

**Types of EDA:-**

1. Univariate non-graphical.
2. Univariate graphical. (Stem leaf plots , histograms , boxplots)
3. Multivariate non-graphical.
4. Multivariate graphical. (Scatter plot , Run chart , Bubble chart , Heatmap)

In this project we have used all types of EDA :-

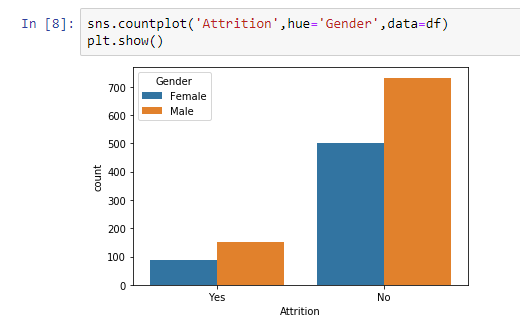
Heatmap and scatterplots are used to check for the null values in every feature and find out the relation between two features of the dataset this is a type of Multivariate Graphical Method.

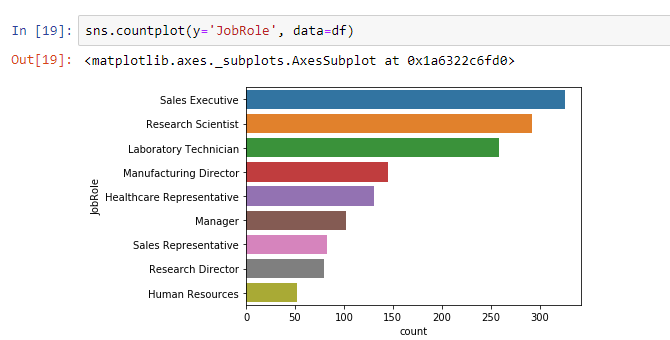
Countplot is used to analyse various aspects like gender wise attrition , age group of people in attrition , Job Satisfaction to attrition , Department wise attrition . We also use countplot for visualizing all the columns in a efficient way and see more clearly and get the info we need this is a type of Univariate Graphical Method.

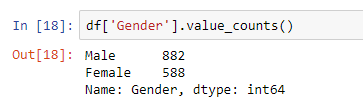
We also use value\_counts() function inbuilt in python to count the number of unique elements in the feature this is a type of Univariate Non-Graphical Method.

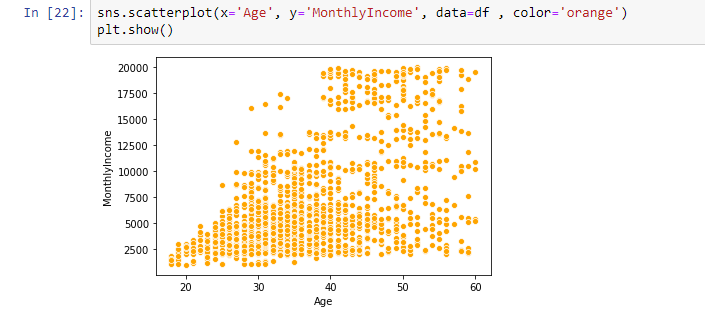
We also used crosstabs to visualize two features together and get their relation this is a type of Multivariate Non Graphical Method.

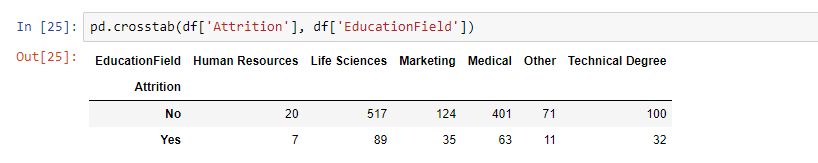












**Preprocessing Pipelines:-**

Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. **Incomplete, noisy, and inconsistent data** are the inherent nature of real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

* **Incomplete data** can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and misfunctions.
* **Noisy data** can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.

There are many stages involved in data preprocessing.

* **Data cleaning** attempts to impute missing values, removing outliers.
* **Data integration** integrates data from a multitude if source into single data warehouse.
* **Data transformation** such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement.
* **Data reduction** can reduce the data size by dropping out redundant features. Feature selection and feature extraction technique can be used.

Pipelines are a simple way to keep your data preprocessing and modelling code organzed. Specifically, a pipeline bundles preprocessing and modelling steps so you can use the whole bundle as if it were a single step.

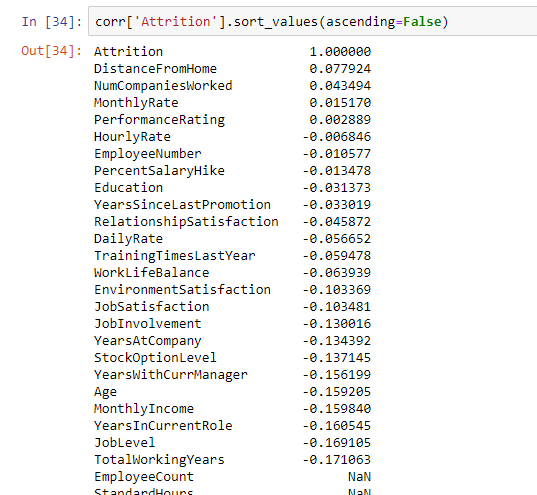
Pipelines have some important benefits:-

1. Cleaner Code
2. Fewer Bugs
3. Easier to Productionize
4. More Options for Model Validation

We will now check our dataset for any noise , incompleteness or inconsistent data for that we will check the correlation which shows us the relation of any feature with the target variable.



Now we will check which of the features have very less correlation with the target variables.

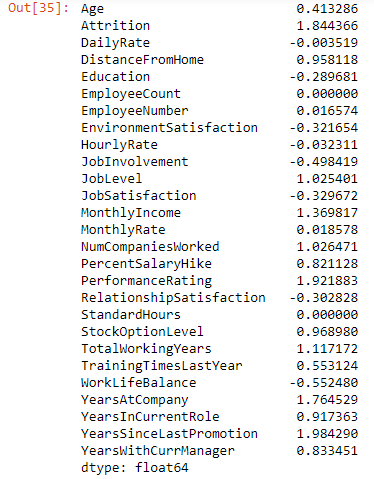


Now we will check for the skewness in the dataset.

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to left or right it is skewed.

If the **skewness** is between -0.5 and 0.5, the data are fairly symmetrical. If the **skewness** is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed. If the **skewness** is less than -1 or greater than 1, the data are highly skewed.

Here in our dataset we can observe that there is some skewness in some of the features.



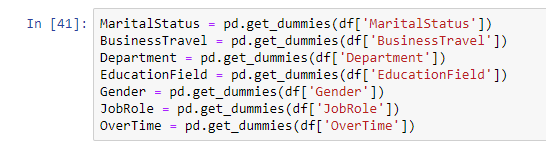
We can treat the skewness in our dataset by using 3 techniques mentioned below:-

1. Log Transform
2. Square/Cube Root Transform
3. Box Cox Transform

Detailed Explanation of which you can read at:-

<https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45>

Now to build our model for prediction we have to make our dataset understandable to the machine learning algorithms. So we will convert all the object data types into machine language i.e, 0 & 1 using One Hot Encoding Technique.



Class Imbalance Problem also comes under preprocessing details of which are mentioned above .

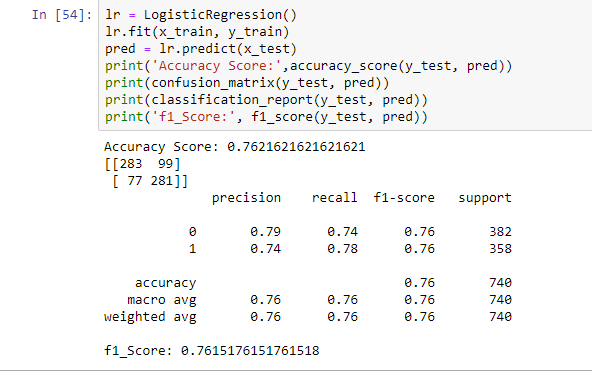
**Building Machine Learning Models:-**

Sklearn is the module that provide us with the machine learning models .Their are two types of problem i.e, classification and regression . In our problem we have to predict whether attrition is happening or not so it is a classification problem.

For building machine learning models we create train and test dataset on which we will train our models . In our case the train and test set will be the dataset output from the SMOTE .



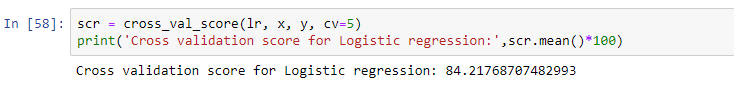
Now we will use different machine learning models and check which one gives us the better accuracy for the datasets provided which we will check by looking at the accuracy\_score.



In our problem logistic regression is selected because it is having the best accuracy and least difference with the Cross\_Val\_Score.

Cross\_Val\_Score is used as a accuracy score for classification problems and f2 score in regression models.

So the model which has the least difference in cross\_val\_score and their accuracy score is the best one.



After selecting the model we will do the Hyper Parameter Tuning of the Model for increasing its Accuracy for better prediction.

A hyperparameter tuning is a parameter whose value is set before the learning process begins.

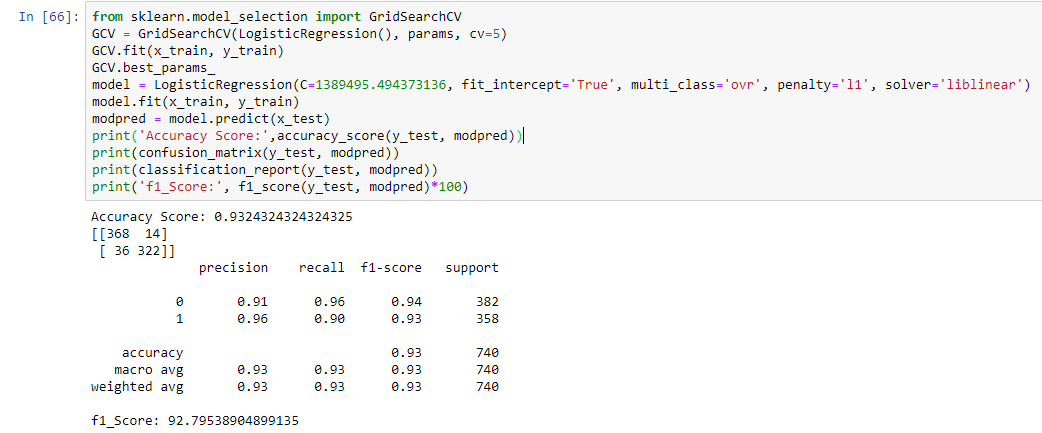
In sklearn, hyperparamters are passed in as arguments to the constructor of the model classes.

Tuning Strategies:-

We will explore two different methods for optimizing hyperparameters:-

* Grid Search
* Random Search

By using these we can improve the accuracy score of the model.



Our Model Accuracy has been up to almost 93% which is good and our model prediction is almost accurate 93 times out of 100.

**Plotting AUC-ROC Curve:-**

AUC – ROC Curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of seperability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0’s as 0’s and 1’s as 1’s. By analogy, the Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis.

**Conclusion:-**

This project has built a model to detect the HR Attrition in a Organization. By doing that we can save the Resources of the company by selecting the Employee correctly .

Various Models are used in our project but we have selected the best model.

We have got our best model i.e, Linear Regression with 93% accuracy . Here our model predicts 93 out of 100 correctly which helps the HR team to select their candidate with more precision keeping in mind the attrition problem.