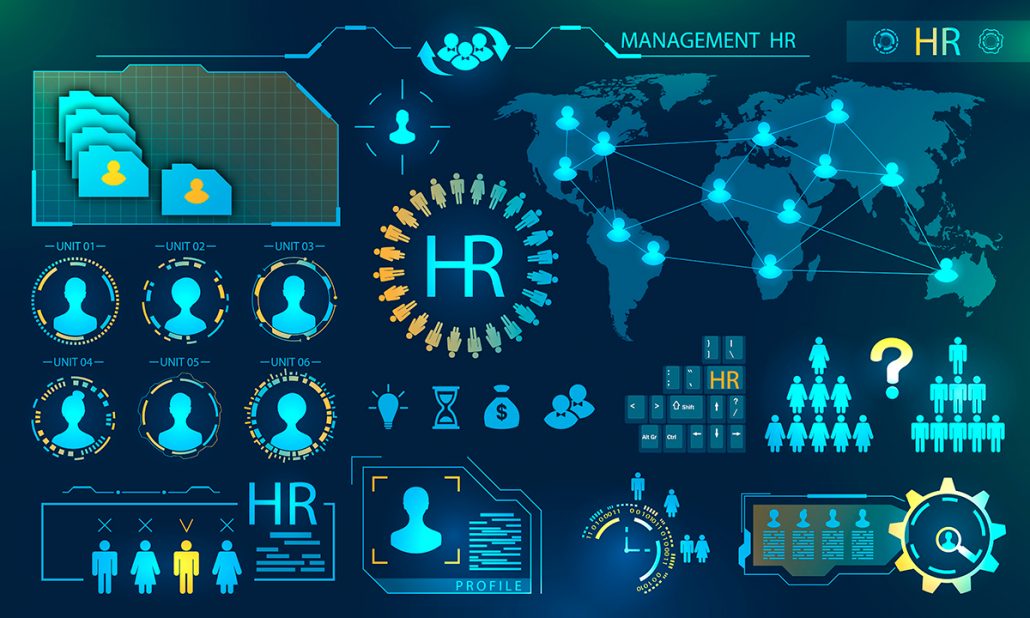
HR Analytics Project- Understanding the Attrition in HR

Written by Sanjay Sivakumar on November 21,2021



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 ofimproving 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 over time. 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 isespecially 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.

Hope the basics made sense. Let’s move on to coding and try finding out how HR Analytics help in understanding attrition.

**Introduction**

This dataset is given as an evaluation project during my course in Data Science at my institute*.*  The dataset includes features like Age, Employee Role, Daily Rate, Job Satisfaction, Years At Company, Years In Current Role etc. For this exercise, we will try to study the factors that lead to employee attrition. This is a fictional data set created by IBM data scientists.

Let’s get start with the work.

# Data Preparation: Load, Clean and Format

Data cleaning is the process of preparing data for analysis by weeding out information that is irrelevant or incorrect.

This is generally data that can have a negative impact on the model or algorithm it is fed into by reinforcing a wrong notion.

Data cleaning not only refers to removing chunks of unnecessary data, but it’s also often associated with fixing incorrect information within the dataset and reducing duplicates.

Without clean data, your models will deliver misleading results and seriously harm your decision-making processes. You'll end up frustrated (been there, done that!), and it's simply not worth it.

Data cleaning is therefore an important part of any machine learning pipeline, and one should not ignore it.

# 

# Firstly we can drop the unwanted univariant columns.

****

Checking for any null values in our dataset.



Thus there is no null values and also no need to format our dataset.

# Data Analysis

Data analysis is the process of collecting, modeling, and analyzing data to extract insights that support decision-making. There are several methods and techniques to perform analysis depending on the industry and the aim of the analysis.

All these various methods for data analysis are largely based on two core areas: *quantitative methods*and*qualitative methods in research.*

Now we can check for the datatypes of the columns in our dataset.

# 

# In order to take a peak into categorical/object values we have to bind them with a numeric variable and then you will be able to see their relevance to the dataset.

# Thus we can encode our dataset using Label Encoder.

# 

# Now we can check for any skewness present in our dataset.

# 

# Thus skewness present in most of our columns.So we can use Z-score method,Box-Cox method to remove it.

# *Z SCORE METHOD*

# A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.A z-score describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units. The z-score is positive if the value lies above the mean, and negative if it lies below the mean.

# *BOX-COX TRANSFORMATION*

# A Box Cox transformation is a [transformation](https://calculushowto.com/transformations/) of non-normal [dependent variables](https://www.statisticshowto.com/dependent-variable-definition/) into a [normal shape](https://www.statisticshowto.com/probability-and-statistics/normal-distributions/). [Normality](https://www.statisticshowto.com/assumption-of-normality-test/)is an important assumption for many statistical techniques; if your data isn’t normal, applying a Box-Cox means that you are able to run a broader number of tests.

# 

Our aim is to predict the employee attrition and it is important to see which variables are contributing the most in attrition. But before that we need to know if the variables are correlated if they are, we might want to avoid those in model building process.

There are many continuous variables, we can have a look at their distribution and create a grid of pair plots but that would be too much code to see the correlation as there are a lot variables. Rather, we can create a seaborn heatmap of numeric variables and see the correlation. The variables which are not poorly correlated(i.e correlation value tend towards 0), we will pick those variables and move forward with them and will leave the ones which are strongly correlated(i.e correlation value tend towards be 1).

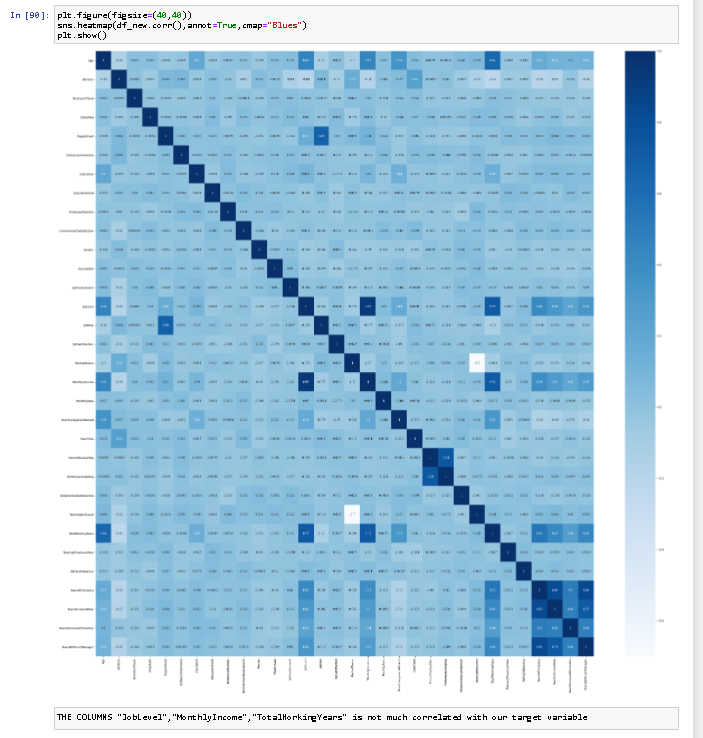
***CORRELATION TEST***

Correlation is a term used to represent statistical measurement of linear relationship between two variables. It can also be defined as the measure of dependence between two different variables. If there are multiple variables and the goal is to find correlation between all of these variables and store them using appropriate data structure, the **matrix data structure**is used. Such matrix is called as **correlation matrix.**

Dependence between two variables, also termed as correlation, can be measured using the following:

**Correlation coefficient / Pearson correlation coefficient** which measures how the value of two different variables vary with respect to each other.

Rank correlation coefficient metric such as **Spearman correlation coefficient**is used to measure the extent to which one variable increases / decreases as the other variable increases / decreases.



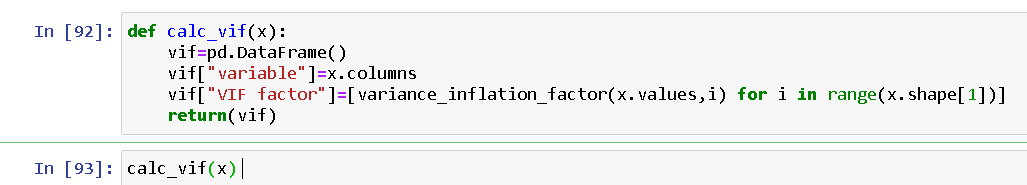
From the heatmap we can infer that the columns “Job level”,”MonthlyIncome”,”TotalWorkingHours” is not much correlated with our target variable.To take further decision to remove any column we can further check for any multicollinearity present in our input columns through VIF factor.

***VIF FACTOR***

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model.

Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

A large variance inflation factor (VIF) on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.

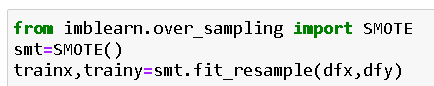


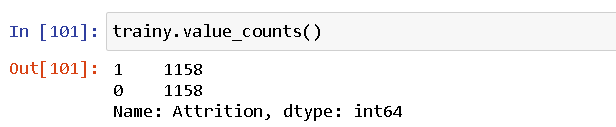
The multicollinearity is high and also the correlation with our target variable also low in the column “MonthlyIncome”.Thus it can be dropped.

As our target column’s data is unbalanced, We can balance using SMOTE technique.

#### **SMOTE: Synthetic Minority Oversampling Technique**

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.



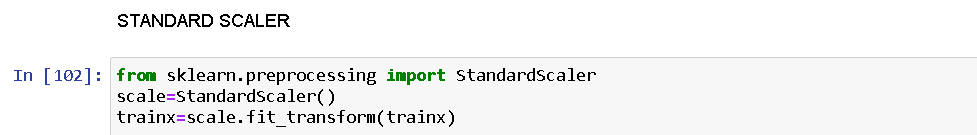


Thus our target column is balanced.Now we can use Standard Scaler to our dataset and proceed to model building.

***STANDARD SCALER***

Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g. Gaussian with 0 mean and unit variance).

For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.



The standard scaler technique should be only used for the input variable.

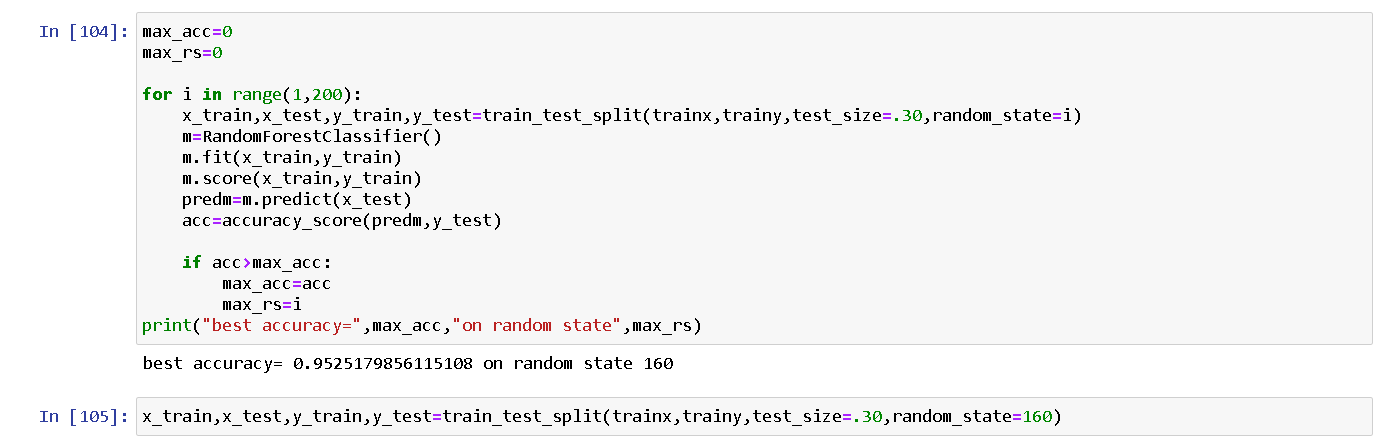
# Modelling the data

We have our final dataset. We now have to start modelling- Predicting the Attrition. If we have the true values we could be able to know that our predictions are correct. Now you will realize that, how important the training data phase is. We train the model in a way that it can predict(almost) correct results.

As our project is based on classification problem we can use upto 5 algorithms to train and test our dataset which is now ready for the model building.And we can use which can be the best suit for our model.

But firstly,we should identify our best random state.

Random\_state is used to set the seed for the random generator so that we can ensure that the results that we get can be reproduced. Because of the nature of splitting the data in train and test is randomised you would get different data assigned to the train and test data unless you can control for the random factor.



Thus we have found our best random state and used it in our train-test-split function for segregating our model for train and split.

The train-test split is a technique for evaluating the performance of a machine learning algorithm.

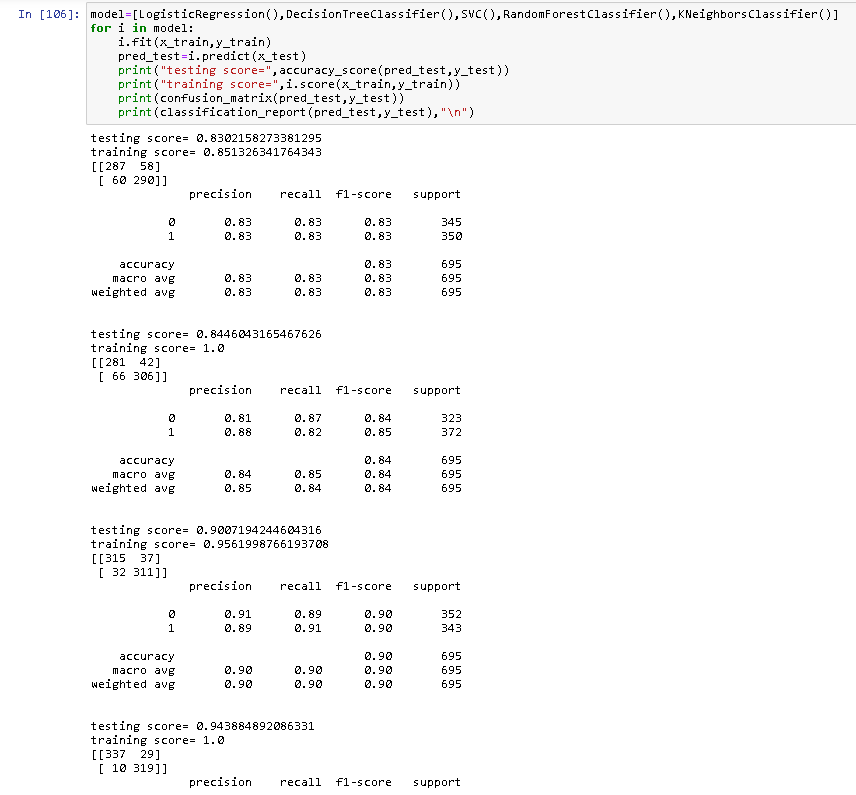
It can be used for classification or regression problems and can be used for any supervised learning algorithm.

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

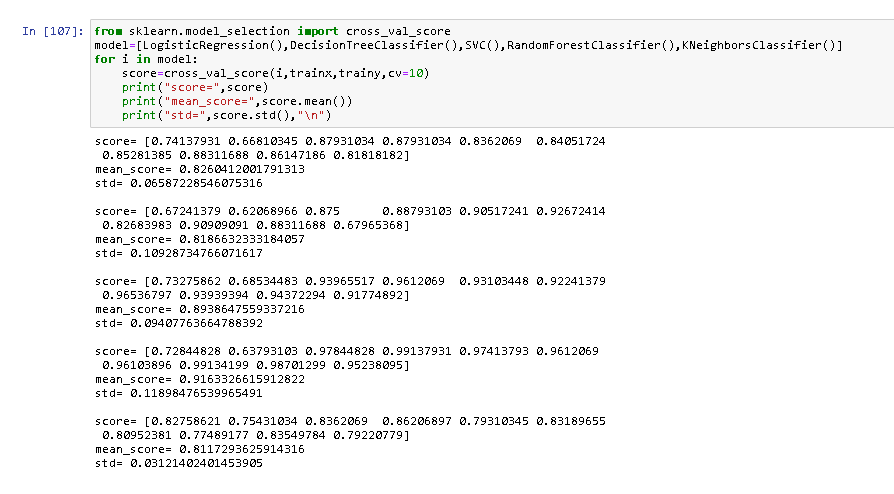
**Train Dataset**: Used to fit the machine learning model.

**Test Dataset**: Used to evaluate the fit machine learning model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

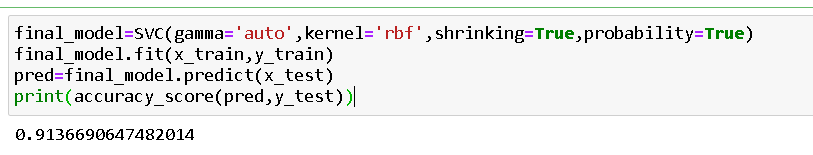


We can choose our best suit algorithm only with this accuracy score and metrics,we should also go for cross-validation to find our best algorithm for our model building.



From this we can able to find that SVC is our best fit algorithm for this dataset as our past accuracy score and our cross val score have less difference.

Now for this best fit model we can do Hyperparameter Tuning process with the help of GridSearchCV.

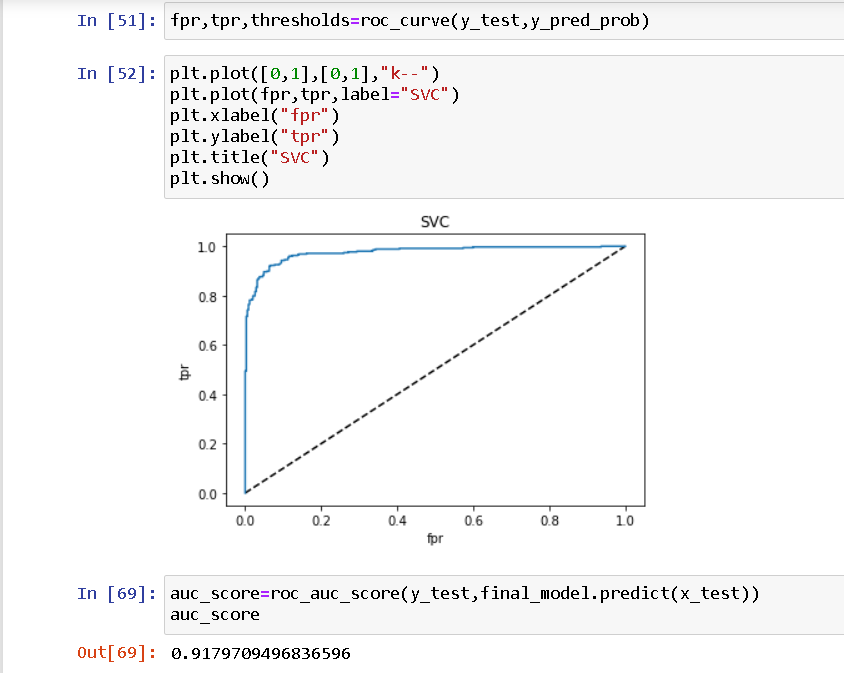


After this tuning process we have build our final model with the accuracy\_score of 0.91,which is quite good in this scenario.Now we can also see the AUC-ROC curve for this model.

***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 separability.

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



Thus this curve is also quite good as the curve is only slightly twisted.Hence we can use this model to predict our input testing columns.

# Summary

Throughout this post, we saw Data is important in Human Resource department(actually in most of places it is important). We saw how we can avoid using correlated values and why it is important not to use those while modelling.We used as much as 5 algorithms to train and test our dataset and identifies the advantages and disadvantages present in all of them. Most of all we found factors which are most important to employees and if are not fulfilled might lead to Attrition.