**A Blog on Machine Learning Model HR Analytics Project- Understanding the Attrition in HR**

Author : Amey Prabhune

**INTRODUCTION:**

Machine Learning is method of data analysis which automates analytical model building. It is branch of Artificial Intelligence. Machine Learning algorithms automatically build a mathematical model using sample data (training data) to make decisions without being specifically programmed to make those decisions. Most industries working with large amounts of data have recognized value of machine learning technologies.

Industries use machine learning are:

* Health care
* Government
* Retail
* Financial services
* Transportations

**HISTORY:**

Machine Learning is based on model of brain of cell interaction. The model is created in 1949 by Donald Hebb in the book titled The Organization of Behavior. In 1967, the nearest neighbor algorithm was conceived, which was the beginning of basic pattern recognition. The algorithm was used for mapping routes and used in finding a solution to the travelling salesperson’s problem of finding most efficient route. Currently, much of speech recognition training is being done by a deep learning technique called Long Short-Term Memory (LSTM). A neural network model describe by Jurgen Schmidhuber and Sepp Hochreiter.

**Problem Definition:**

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.

**Attrition in HR**

Attrition in HR refers to the gradual loss in employee overtime. High attrition is problematic for companies.

**Attrition Affection Companies**

A major problem in high employee attrition is its cost to the organization.

**Problem Statement**

We have one dataset containing information of employee and we will try to predict the employee attrition.

**Data Analysis:**

In our dataset we have 35 columns and 1470 rows in total we have record of 1470 employees. With the help of this information we will train our model and predict the attrition. In our dataset there are 35 columns one of them is our target variable. Target variables result is depend upon other remaining columns. First we need to find out our target variable and in this dataset out target variable is Attrition column. In the Data Analysis phase we will try to analyze the data. First we will import our data, then we will check our record and shape and other information like null values, any missing values. Using (df.info) command we will check null values from the dataset. Using (df.describe) command we will check count, min, max values and standard deviation of every column. Also we need to drop columns from dataset which are not helping us to predict values like Employee Number and Hours.

Incomplete data can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and malfunctions.

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.

**EDA Concluding Remark:**

In statistics, Exploratory Data Analysis is an approach of analyzing data set to summarize their main characteristics, using statistical graphics and other visualization methods. Primarily EDA is seeing what the data can tell us beyond the modeling or hypothesis testing. EDA focuses on checking assumptions required for model fitting.

In EDA we will visualize our data in the graphical form with the help of graph means Univariate analysis (plotting a box plot for single column) then bivariate analysis (plotting a scatter plot for 2 variables) and third one is multivariate analysis (plotting pair plot, heatmap for all dataset) using count plot we will find out count of each categorical value from every column. With help of heat map we can check relation between target variable and other variable. Also we need to convert all non-numerical columns into numerical columns with the help of encoding. We will replace the missing values with mean. Also we will drew a graph for visualization. Like count plot, subplot, pairplot.

**Pre-Processing Pipeline:**

Preprocessing is very important step in Machine Learning to yield highly accurate and insightful result. Greater the quality of data, the greater is the reliability of the produced result. Data preprocessing helps in increasing the quality of data filling in missing data, smoothing noisy data.

In our dataset luckily we don’t have any null values, but we have some column which has object data types so we need to convert them into integer, with help of Encoding. Like Attrition, Business travel, Department, Education field, Gender etc. we will use label encoder only for those column who has only 2 category like male female, yes no etc. and for the others we will use ordinal encoder. After cleaning machine learning model we will check skewness and correlation but form my point of view it better it is a god practice to remove skewness after the splitting our dataset. Because I want my target variable untouched for better predication. We will check our columns are normally distributed or not with the help of histplot. It will give us clear vision of our data is skewed or not. We will remove skewness of those columns which are skewed (right or left).

Note : Categorical column doesn’t have any correlation.

**Building Machine Learning Models:**

Building machine learning model consist of algorithm that can automate analytical model building. Building machine learning models that have the ability to generalize well on future data requires thoughtful consideration of the data at hand and of assumptions about various training algorithm.

After dealing with missing values and cleaning our data set we will split out data set into feature columns and target variable. And then we will remove skewness from out data set. After removing skewness, we will build our model with train and test data. For building model we need to find our highest accuracy score on random state. In this model I got best accuracy score of 0.91 at random state 26. We will use random state 26 and will train our model. Then we will use other algorithms to check highest accuracy like. GaussianNB, Random Forest, Decision Tree, KNN, AD Boost, Support Vector Classifier. After that we will cross check all our models using cross validation. And we will compare our both score and check the difference between score. Any model give us best accuracy, we will save that model. Using Random Forest algorithm I got high accuracy score of 0.91. We will save that model. Before choosing our best algorithm accuracy we need to cross validate our model with help of cross validation function.

**Technology used:**

* Python 3
* Jupyter Notebook
* Numpy
* Pandas
* Seaborn
* Sklearn
* Matplotlib
* Different types of algorithm
* Min max scaler
* Zscore

**Concluding Remarks:**

After training and testing our data my models accuracy score is 0.91 and we can predict employee attrition with the help of this model.

