Heliverse

Assignment: AI & ML Role - Python Programming

Objective:

This assignment aims to evaluate your proficiency in Python programming and your understanding of key concepts in artificial intelligence (AI) and machine learning (ML).

Problem Statement: Predicting Employee Attrition

Dataset: IBM HR Analytics Employee Attrition & Performance

Attrition- Company losing its customer base

Attrition is a process in which the workforce dwindles at a company, following a period in which a number of people retire or resign, and are not replaced.

- A reduction in staff due to attrition is often called a hiring freeze and is seen as a less disruptive way to trim the workforce and reduce payroll than layoffs
- Our Aim will be to analyse the datasets completely with respect to each and feature and find the reason behind Attrition of Employees.
- And what the top factors which lead to employee attrition?

Description about the data

- Age: A period of employee life, measured by years from birth.
- Attrition: The departure of employees from the organization.
- BusinessTravel: Did the employee travel on a business trip or not.
- Daily Rate: Employee salary for the period is divided by the amount of calendar days in the period.
- Department: In which department the Employee working.
- DistanceFromHome: How far the Employee live from the office location.
- Education: In education 1 means 'Below College', 2 means 'College', 3 means 'Bachelor', 4 means 'Master', 5 means 'Doctor'
- EducationField: In which field Employee complete his education.
- EmployeeCount: How many employee working in a department
- EmployeeNumber: An Employee Number is a unique number that has been assigned to each current and former State employee and elected official in the Position and Personnel DataBase (PPDB).

- Job involvement: Is the degree to which an employee identifies with their work and actively participates in it where 1 means 'Low', 2 means 'Medium', 3 means 'High', 4 means 'Very High'
- JobLevel: Job levels, also known as job grades and classifications, set the responsibility level and expectations of roles at your organization. They may be further defined by impact, seniority, knowledge, skills, or job title, and are often associated with a pay band. The way you structure your job levels should be dictated by the needs of your unique organization and teams.
- JobRole: What is the jobrole of an employee.
- JobSatisfaction: Employee job satisfaction rate where, 1 means 'Low', 2 means 'Medium', 3 means 'High', 4 means 'Very High'
- MaritalStatus: Marital status of the employee.
- MonthlyIncome: total monetary value paid by the organization to an employee.
- MonthlyRate: The per-day wage of the employee.
- NumCompaniesWorked: Before joining this organization how many organizations employee worked.
- Over18: Is the employee age over than 18 or not.
- OverTime: A Employee works more than 9 hours in any day or for more than 48 hours in any week.
- PercentSalaryHike:
- PerformanceRating 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'
- EnvironmentSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
- RelationshipSatisfaction 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
- StandardHours: Is the number of hours of production time that should have been used during an working period.
- StockOptionLevel: Employee stock options, also known as ESOs, are stock options in the company’s stock granted by an employer to certain employees. Typically they are granted to those in management or officer-level positions. Stock options give the employee the right to buy a certain amount of stock at a specific price, during a specific period of time. Options typically have expiration dates as well, by which the options must have been exercised, otherwise they will become worthless.
- TotalWorkingYears: Total years the employee working in any organization
- TrainingTimesLastYear: Last year how many times employee took training session.
- WorkLifeBalance 1 'Bad' 2 'Good' 3 'Better' 4 'Best'
- YearsAtCompany: How many years the employee working in the current organization
- YearsInCurrentRole: How many years the employee working in the current position
- YearsSinceLastPromotion: How many years the employee working in the current position after promotion
- YearsWithCurrManager: How many years the employee working under the current manager

Roadmap

1. Data Collection
2. Data pre-processing
3. Model Building
4. Model Training
5. Model Evaluation

1. Data Collection

- Data was been taken from kaggle in the form of Microsoft Excel file.
- The data was in the tabular form i.e. in the form of rows and column.
- Library files and then the data file are imported in the code.

2. Data Pre-processing

- In Data Processing data exploration is been performed.
- Importing Libraries: In the first place, Let import libraries to help us in the manipulation the data set, such as `pandas`, `numpy`, `matplotlib`, `seaborn`.
- **Basic Data Exploration:-**

  - This is an Important Step in Data Science and Machine Learning to ensure about the columns, and rows present.
  - First, we will check the shape of the dataset
  - Second, we will check the head, tail, and sample of the datasets
  - Third, we will check the Data Description
  - Then, we will check the Data Types of the columns present in the data.

- **Statistical Observation:-**

  - We only have int and string data types features. There is no feature with float. 26 features are numerical and 9 features are categorical
  - Attrition in out target value which has no missing value. But, the quantity of data of employee having Attrition is less compared to employees which do not have Attrition.
  - It's very good that we are having a complete data-set, there is no any missing values in dataset.

- Check Duplicates
- Checking missing value
- **Target Variable** -Over here we noticed that the Target column is Highly Imbalanced, we need to balance the data by using some Statistical Methods.

- **Exploratory Data Analysis:-**

  OBSERVATIONS:-

- Employees working in R&D department are more, but employees from sales department or at position like sales executive, sale Representative leaves the job early.
- Males are more under Attrition then Females.
- Age column is very well normalised, most of employees are age between 25 to 40.
- We are having some of the numerical columns which are label encoded for us, they are ordinal labels, so let's have a look at them first.
- Employees from Bachelor are more, then from Masters background. Attrition with respect to bachelor can be seem more because they have more and more expectation from companies and it will be interesting to see the reason behind this in this database.

- **Label Encoding:-**

- In machine learning, we usually deal with datasets that contain multiple labels in one or more than one columns.
- These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words.

3. Model Building

- For Model building firstly the data is been described which states the statistics of the data. Information is been extracted which tells number of features in the dataset.
- Feature Scaling is done to normalize the data.
- Choose appropriate machine learning algorithms for training the model
- Machine Learning: Splitting the data into Training and Testing sample
- We don't use the full data for creating the model. Some data is randomly selected and kept aside for checking how good the model is. This is known as Testing Data and the remaining data is called Training data on which the model is built.
- Typically 70% of data is used as Training data and the rest 30% is used as Testing data.
- **Resampling:-**

- Resampling is the method that consists of drawing repeated samples from the original data samples. The method of Resampling is a nonparametric method of statistical inference Oversampling and under sampling in data analysis are techniques used to adjust the class distribution of a data set. These terms are used both in statistical sampling, survey design methodology and in machine learning. Oversampling and under sampling are opposite and roughly equivalent techniques
- We are going to use Over Sampling.

- We will not use Under Sampling to avoid data loss.

Algorithms used:

1. **Logistic Regression in Machine Learning**

- Logistic Regression is used for predicting a category, specially the Binary categories (Yes/No , 0/1).
- For example, whether to approve a loan or not (Yes/No)? Which group does this customer belong to (Silver/Gold/Platinum)? etc.
- When there are only two outcomes in Target Variable it is known as Binomial Logistic Regression.
- If there are more than two outcomes in Target Variable it is known as Multinomial Logistic Regression.
- If the outcomes in Target Variable are ordinal and there is a natural ordering in the values (eg. Small< Medium< Large) then it is known as Ordinal Logistic Regression.
- Logistic regression is based on logit function $\text{logit}(x) = \log(x / (1 - x))$
- The output is a value between 0 to 1. It is the probability of an event's occurrence.
- E.g. There is an 80% chance that the loan application is good, approve it.
- The coefficients β0, β1, β2, β3… are found using Maximum Likelihood Estimation Technique. Basically, if the Target Variable's value (y) is 1, then the probability of one "P(1)" should be as close to 1 as possible and the probability of zero "P(0)" should be as close to 0 as possible. Find those coefficients which satisfy both the conditions.

2. **Random Forests**

- Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.
- For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned Random decision forests correct for decision trees' habit of overfitting to their training set.
- Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.
- Random Forest often produces highly accurate predictions.The ensemble nature of Random Forest helps mitigate overfitting, especially when compared to individual decision trees.
- It can provide insights into feature importance, helping with feature selection. Random Forests are widely used in various domains such as finance, healthcare, and natural language processing. However, like any model, they have their limitations,

and the choice of the right algorithm depends on the specific characteristics of the data and the problem at hand.

4. Model training:

- Several models are been imported in the code from sklearn.
- The training data i.e. X_train and Y_train is been fit into the model.
- The model is been trained through the data provided and then it is ready to predict the target value for test data.

5. Model Evaluation:

- The predicted model and the test dataset are used to predict the accuracy score of the model.
- The model is been evaluated by using accuracy score, precision score, recall score and f1 score is used.
- More the accuracy score, more accurate is the model.

Conclusion

❑ The most accurate model is Random Forest Classifier with an accuracy of 95.7.

❑ The model is saved and then trained on whole dataset.

❑ The model is ready to predict the outcome for the new input of data