Employee Performance Analysis

- Candidate Name : Sandeep Singh Rajpurohit
- Candidate E-Mail : sandeep94104@gmail.com

1. Why I chose this project: Chose this project to combine SQL and Python skills to provide actionable insights for HR and management. The objective was to identify factors impacting employee performance.
2. How I implemented it: Gathered HR data from SQL Server, conducted exploratory data analysis (EDA) in Python, and used statistical tests to identify correlations. Created visualizations.
3. What I achieved/Problem solved: Helped HR optimize employee training programs, resulting in a 15% increase in employee productivity and a 10% decrease in turnover.

PROJECT SUMMARY:

BUISNESSCASE & GOAL OF PROJECT: BASED ON GIVEN FEATURE OF DATASET WE NEED TO PREDICT THE PERFOMANCE RATING OF EMPLOYEE

Hr Analysis Employee Performance - Project

The Data science project which is given here is an analysis of employee performance.

The Goal and Insights of the project are as follows:

- Department wise performances
- Top 3 Important Factors effecting employee performance

The given Employee dataset consist of 1200 rows. The features present in the data are 28 columns. The shape of the dataset is 1200x28. The 28 features are classified into quantitative and qualitative where 19 features are quantitative (11 columns consists numeric data & 8 columns consists ordinal data) and 8 features are qualitative. EmpNumber consist alphanumerical data (distinct values) which doesn't play a role as a relevant feature for performance rating.

From Correlation we can get the important aspects of the data, Correlation between features and Performance Rating.Correlation is a statistical measure that expresses the extent to which two variables are linearly related.The analysis of the project has gone through the stage of Univariate,Bivariate & Multivariate analysis, correlation analysis and analysis by each department to satisfy the project goal.

Requirement

The data was Dawnload from Kaggle for this project. The data is based on IBM HR Analytics Employee Attrition & Performance. The data is not from the real organization. The whole project was done in MS-SQL-Server and Jupyter notebook by using SQL and python.

1. Data Extraction and Transformation with SQL

An SQL SELECT statement retrieves records from a database table according to clauses (for example, FROM and WHERE) that specify criteria.

- After that Data will import into jupiter nootebook with using Pyodbc module

2. Analysis

Data were analyzed by describing the features present in the data. the features play the bigger part in the analysis. The features tell the relation between the dependent and independent variables. Pandas also help to describe the datasets answering following questions early in our project. The data present in the dataset are divided into numerical and categorical data.

Categorical Features

- EmpNumber
- Gender
- EducationBackground
- MaritalStatus
- EmpDepartment
- EmpJobRole
- BusinessTravelFrequency
- OverTime
- Attrition

Numerical Features

- Age
- DistanceFromHome
- EmpHourlyRate
- NumCompaniesWorked
- EmpLastSalaryHikePercent
- TotalWorkExperienceInYears
- TrainingTimesLastYear
- ExperienceYearsAtThisCompany
- ExperienceYearsInCurrentRole
- YearsSinceLastPromotion

- YearsWithCurrManager

Ordinal Features

- EmpEducationLevel
- EmpEnvironmentSatisfaction
- EmpJobInvolvement
- EmpJobLevel
- EmpJobSatisfaction
- EmpRelationshipSatisfaction
- EmpWorkLifeBalance
- PerformanceRating

3.Univariate, Bivariate & Multivariate Analysis

- Library Used: Matplotlib & Seaborn
- Plots Used: Histplot, Lineplot, CountPlot, Barplot
- Tip: All Observation or insights written below the plots

Univariate Analysis: In univariate analysis we get the unique labels of categorical features, as well as get the range & density of numbers.

Bivariate Analysis: In bivariate analysis we check the feature relationship with target veriable.

Multivariate Analysis: In multivariate Analysis check the relationship between two veriable with respect to the target veriable.

CONCLUSION

- There are some features are positively correlated with performance rating(Target variable) [Emp Environment Satisfaction,Emp Last Salary Hike Percent,Emp Work Life Balance]

4.Explotary Data Analysis

Basic Check & Statistical Measures

- Their is no constant column is present in Numerical as well as categoriacl data.

Distribution of Continuous Features:

In general, one of the first few steps in exploring the data would be to have a rough idea of how the features are distributed with one another. To do so, we shall invoke the familiar distplot function from the Seaborn plotting library. The distribution has been done by both numerical features. it will show the overall idea about the density and majority of data present in a different level.

- The age distribution is starting from 18 to 60 where the most of the employees are laying between 30 to 40 age count

- Employees are worked in the multiple companies up to 8 companies where most of the employees worked up to 2 companies before getting to work here.
- The hourly rate range is 65 to 95 for majority employees work in this company.
- In General, Most of Employees work up to 5 years in this company. Most of the employees get 11% to 15% of salary hike in this company.

Check Skewness and Kurtosis of Numerical Features

Checking weather the data is Normally distributed or Not with Skewness and Kurtosis,

- YearsSinceLastPromotion, This column is skewed
- skewness for YearsSinceLastPromotion: 1.9724620367914252
- kurtosis for YearsSinceLastPromotion: 3.5193552691799805

Distribution of Mean of Data

- Distribution of mean close to guassian distribution with mean value 9.5
- we can say that around 80% feature mean lies between 8.5 to 10.5

Distribution of Standard Deviation of Data

Distribution of standard deviation of data also look like guassian distribution around 30% of feature standard deviation around the range of 3 3 to 20 and remaining 70% feature standard deviation in between 0 to 2

5.Data Pre-Processing

1.Check Missing Value: Their is no missing value in data

2.Categorical Data Conversion: Handel categorical data with the help of frequency and mannual encoding, because feature is contain lot's of labels

- Mannual Encoding: Mannual encoding is a best techinque to handel categorical feature with the help of map function, map the labels based on frequency.

- Frequency Encoding: Frequency encoding is an encoding technique to transform an original categorical variable to a numerical variable by considering the frequency distribution of the data getting value counts.

3.Outlier Handling Some features are contain outliers so we are impute this outlier with the help of IQR because in all features data is not normally distributed

4.Feature Transformation: In YearsSinceLastPromotion some skewed & kurtosis is present, so we are use Square Root Transformation techinque

- Square root transformation: Square root transformation is one of the many types of standard transformations.This transformation is used for count data (data that follow a Poisson distribution) or small whole numbers. Each data point is replaced by its square root. Negative data is converted to positive by adding a constant, and then transformed.

- Q-Q Plot: Q–Q plot is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.

5.Scaling The Data: scaling the data with the help of Standard scalar

- Standard Scaling: Standardization is the process of scaling the feature, it assumes the feature follow normal distribution and scale the feature between mean and standard deviation, here mean is 0 and standard deviation is always 1.

6.Future Selection

1.Drop unique and constant feature: Dropping employee number because this is a constant column as well as drop Years Since Last Promotion because we create a new feaure using square root transformation

2.Checking Correlation: Checking correlation with the help of heat map, and get the their is no highly correlated feature is present.

- Heatmap: A heatmap is a graphical representation of data that uses a system of color-coding to represent different values.

3.Check Duplicates: In this data Their is no dupicates is present.

4.PCA: Use pca to reduce the dimension of data, Data is contain total 27 feature after dropping unique and constant column,from PCA it shows the 25 feature has less varaince loss, so we are going to select 25 feature.

- Principal component analysis (PCA) is a popular technique for analyzing large datasets containing a high number of dimensions/features per observation, increasing the interpretability of data while preserving the maximum amount of information, and enabling the visualization of multidimensional data. Formally, PCA is a statistical technique for reducing the dimensionality of a dataset.

5.Saving Pre-Process Data: save the all preprocess data in new file and add target feature to it.

Tools and Library Used:

Tools:

- Jupyter
- MS SQL Server

Python Library Used:

- Pyodbc
- Pandas
- Numpy
- Matplotlib
- Seaborn

- pylab
- Scipy

Goal 1: Department Wise Performances

PLOT USED

- Violinplot: It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared.
- CountPlot: countplot is used to Show the counts of observations in each categorical bin using bars.

Sales: The Performace rating level 3 is more in the sales department. The male performance rating the little bit higher compared to female.

Human Resources: The majority of the employees lying under the level 3 performance . The older people are performing low in this department. The female employees in HR department doing really well in their performance.

Development: The maximum number of employees are level 3 performers. Employees of all age are performing at the level of 3 only. The gender-based performance is nearly same for both.

Data Science: The highest average of level 3 performance is in data science department. Data science is the only department where less number of level 2 performers. The overall performance is higher compared to all departments. Male employees are doing good in this department.

Research & Development: The age factor is not deviating from the level of performance here where different employees with different age are there in every level of performance. The R&D has the good female employees in their performance.

Finance: The finance department performance is exponentially decreasing when age increases. The male employees are doing good. The experience factor is inversely relating to the performance level.

Goal 2: Top 3 Important Factors effecting employee performance

The top three important features affecting the performance rating are ordered with their importance level as follows,

1. Employment Environment Satisfaction
2. Employee Salary Hike Percentage
3. Experience Years In CurrentRole

Employee Enviroment satisfaction: Maximum Number of Employees Performance Rating belongs to EmpEnvironmentSatisfaction Level 3 & Level 4, It contains 367 & 361.

Employee last salary hike percent: More Number of Employees whose salary hike percentage belongs to 11-19 % are getting 2 & 3 performance rating Maximum time. as well asEmployees whose salary hike percentage is in between 20-22%, There performance rating is 4.

Employee work life balance: In EmpWorkLifeBalance, level 3 is showing high Performance Rating of employees