Employee Burnout prediction

Linear Regression:

- power statistical technique to model the relationship between a dependent variable and one or more independent variables.

What is Employee Burnout prediction?

- Predicting the stress level of a employee

→ For an organization it is required to go through the Nature & stress levels of an Employee So it is going to contribute the whole organization's growth.

→ So, it is important to address the employees mental health, & stress levels.

What is Supervised learning:

→ Basically we will be feeding the machine / computer both the input & output so it can generate data & apply it to future datasets.

→ This model uses Labelled data.

→ In this project (Employee Burnout prediction) we will be using Linear Regression Technique.

Dependent and independent Variables

Dependent variable – like an outcome that we want to predict or explain

Independent variable – factors that influence the Dependent variable.

eg: • Final examination score is DV

• All our struggle, efforts, internal marks, class Assignment, activies that influence our final score is IV.

Defining Employee Burnout:

- Employee Burnout is a state of emotional, physical & mental exhaustion caused by prolonged or excessive stress

1) Emotional exhaustion
2) De-personalization
3) Physical symptoms
4) Reduced personal Accomplishment.

Data Collection & Pre-processing

- online resources : Kaggle.

1) Data Acquisition
2) Data cleaning
3) Feature Engineering.

- All these machine learning models are based on iterations / data
- ML is impossible without data (quality data leads to optimized result)

Work-related Factors (contributing to Employee Burnout)

- Workload, Job control, Role clarity, Work-life Balance.

Personal Factors

- Age, Gender, Personality Traits, Emotional intelligence

Organizational Factors

- Organizational culture, leadership style, support systems.

Linear Regression:

General Formula:-

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

where,

y - dependent variable (outcome of employee Burnout)

$x_1, x_2 \dots x_n$ - independent variables (factors which influence dependent variables)

$\beta_1, \beta_2, \beta_n$ - coefficients that give you co-relation b/w Dependent & independent variable.

β_0 - intercept

ϵ - error term (difference b/w observed & predicted values of y)

Google Co-Lab: online platform where you can implement Machine Learning & Data Science related

The dataset consist of 9 columns.

- Employee ID
- Date of Joining
- Gender
- Company Type
- WFH set-up available
- Designation
- Resource allocation
- Mental Fatigue Score

} input column.

- Burn Rate. [output column]

Libraries:

numpy: numerical python (Mathematical & Array related computations) to support

why numpy ? → 'n' dimensions (like Matrices, Arrays)

Pandas: data Manipulation (dataset forom different Platform)

Matplotlib.pyplot, Seaborn } used for visualization, graphs, animated graphs, charts

relation blw data with visualization (pictorial rep)

sklearn.model_selection : In the ML model we divide data for testing & training.

sklearn.preprocessing : Converting statistical data into unit variance with the help of standard scalar.

sklearn.linear_model : Linear Regression Model. (created using training data-set)

sklearn.metrices : to evaluate the performance of our model, metrices i.e we have; mean_absolute_error, mean_squared_error, r2_score

pickle
os } Storing transmitted data from one format into another
↓
opening a file, reading a file etc.

Data Overview:

1) data.head() - in default case it gives the first '5' entries.
Top '5' entries of the data-set

2) data.tail(3) - displays the last 3 entries of a data set.

3) data.describe() – it gives the statistical information of your data set. (Min, Max, count values... , mean, SD)

4) data.columns.tolist() – List of column values in the data set

5) data.nunique() – No. of unique values under each category / columns it will show you count

6) data.info() – Gives the information of your dataset.

7) data.isnull().sum() – Count of Null values in each column.
   ↓
   returns boolean values (whether given column has Null value or not)

8) data.isnull().sum().values.sum() – Count of total No. of Null values in whole data set

9) data.corr(numeric_only=True)['Burn Rate'][:-1]

– Finds the correlation b/w Burn Rate & all other columns having only numerical values.

→ '-1' → it excludes to find correlation among Burn Rate itself

10.) sns.pairplot(data)
plt.show()
- visualization of correlation of data in terms of the graphical representation.

(correlation between burnout rate & remaining 3 columns), mental fatigue, resource allocation, designation

Observation: to check / infer the co-relation b|w variables.

11) data = data.dropna() - Drop off all Null values of our dataframe.

Why?: because, due to the missing values we have some less correlation b|w the models/ variables thats why we drop all Null values.

12.) data.shape - to verify whether values are dropped off or Not.

13) data.dtypes - Analyze data type of Each column.

14.) data = data.drop('Employee ID', axis = 1)
- dropping the entire employee ID column as it consists of inconsistent data.

To check Employee seniority can contribute for the data set or Not.

Checking the correlation of Date of Joining with Target Variable.

→ The answer is; Date of joining to No. of years of exp we can find the seniority.

One-hot Encoding for categorical values

- Based on Number of categories it will be converted into 0 & 1's.

(Converting categorical features into numerical features)

Preprocessing

- dividing the data into input & output.
  - y - output
  - x - input
- split the data into training & testing
  - training size: 70%
  - testing size: 30%

We create a Linear Regression Model where, we feed both the input & output.

Then, we evaluate the model based on error metrics.

- For calculating the errors we will be feeding the machine both predicted & original outcome so it can generate the error.

Final Observation:

- Based on the evaluation metrics, the Linear Regression model appears to be the best model for predicting burnout analysis.
- It has the lowest mean squared error, root mean squared error, and mean absolute error, indicating better accuracy and precision in its predictions.
- Additionally, it has highest R-squared score indicating a good fit to the data and explaining a higher & explaining a higher ~~propagation~~ proportion of the variance in the target variable.
- So, we are choosing this model for deployment.

Output:

Linear Regression Model Performance Metrics :

Mean Squared Error : 0.0031569779113610717

Root Mean Squared Error: 0.05618690588 2831

Mean Absolute Error: 0.045950326447 73

R-squared score : 0.9188226742 47248.