Employee Burnout Prediction Using Data Sceince

,Umaira[1], Ayshathul Sajeena[1],Nusaiba[1], Nabeesath Sunaina

[1]Department of Computer Science and Engineering, BIT, Mangaluru, Karnataka, India.

*Corresponding Author: Ayshathul Sajeena Email: sajeenaacm@gmail.com

Abstract:

Nowadays employees face a lot of stress due to the workload and lack of leisure time, so it is the need of hour for having to predict and analyze employee burnout. One of the most important new concerns that organizations are grappling with is employee or job burnout. Workers in the manufacturing and service sectors who are frequently exposed to demanding work environments may become more stressed out at work, burn out, or even quit their jobs. Our research identifies the main causes of burnout by gathering and analyzing data from a variety of sources, such as performance indicators, questionnaires, and HR records.

The study makes use of machine learning models and statistical methods to find patterns and correlations in the data. Our goal is to forecast employee burnout occurrences by using predictive models, which will allow employers to take prompt action. To understand their impact on burnout, common contributing elements are investigated, including workload, job demands, interpersonal connections, and job satisfaction.

To identify trends and correlations in the data, the study uses statistical techniques and machine learning algorithms. Our objective is to use predictive models to anticipate employee burnout so that companies can respond promptly. The effects of typical contributing factors, such as workload, job demands, interpersonal relationships, and job satisfaction, on burnout are examined.

Key Words: Machine Learning, Burnout, prediction, Linear Regression, Ridge Regression,Lasso Regressor, catboost regressor.

1. Introduction

A common problem in today's hectic and demanding work situations is employee burnout. Chronic occupational stress that has not been effectively controlled is what defines it. Numerous detrimental outcomes, such as a drop in general wellbeing, increased absenteeism, and decreased productivity, can result from burnout. Organizations are using data science more and more to evaluate and forecast burnout tendencies as they realize how important it is to combat employee burnout.

The goal of this research is to use data science methods to assess and forecast employee burnout in a

company. We may create models that shed light on potential burnout risks by looking at a variety of burnout-related parameters, including workload, work hours, interpersonal interactions, and individual traits. After that, these insights can be used to put proactive measuresin place, like workload adjustments.
Predicting employee burnout using data science is a crucial application in modern workplaces,
aiming to proactively identify and mitigate factors contributing to burnout among employees. Burnout is a state of physical, emotional, and mental exhaustion resulting from prolonged stressor overwork, and it can have significant negative impacts on individual well-being andorganizational productivity.
Data science techniques provide a strong method for analyzing different elements contributing toburnout and antedate which employees are more vulnerable. Here's an introduction to how this process typically works:

1. **Data Collection**: Obtaining diverse data from the organization is the initial stage. This mayinvolve employee demographics, roles at the job, working hours, the performance metrics, feedback from employee surveys, communication patterns, and even external elements such as business trends or the state of the economy.
2. **Feature Engineering:** Once the data is collected, feature engineering is performed to preprocess and in this process the raw data extracted only meaningful features that can be used for analyzing purpose.
3. **Model Selection**: Various machine learning algorithms that can be applied to foresee employee burnout based on the prepared dataset.
4. **Model Training and Evaluation:** The chosen model is trained on a subset of the data and assessed using performance metrics such as accuracy level, precision, recall, and F1-score. Cross-validation techniques are often employed to guarantee the robustness of the model and prevent over fitting.
5. **Prediction and Deployment:** after, the model is trained and evaluated, it can be deployedinto production to forecast burnout risk of new or existing employees

2. Methodology

2.1 Modules overview

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied, and the model istrained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset

2.) Selection of attributes3.) Data Pre-Processing 4.) Burn Out Prediction

1. Collection of Dataset

Initially, we collect a dataset for our Burn Out prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset consists of 9 attributes.

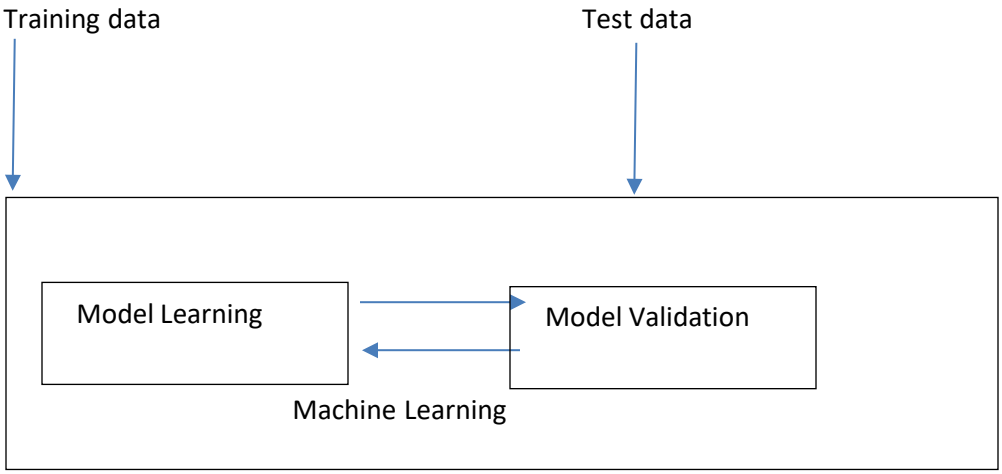
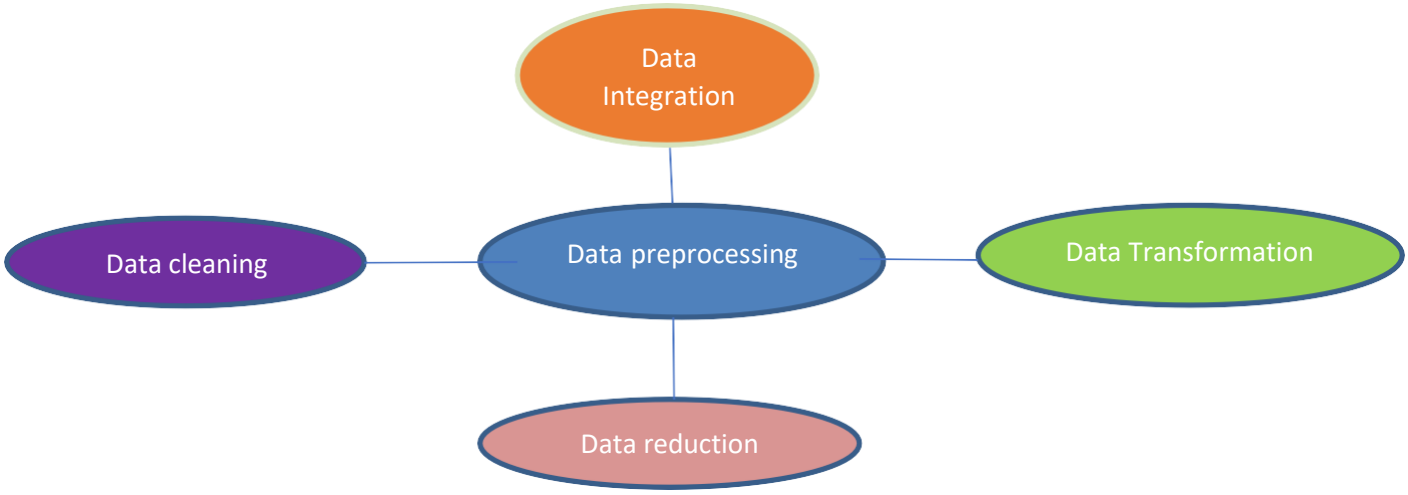


Figure 2.1.1: Collection of Dataset

2. Selection of attributes



Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the employee like gender, Company type, WFH setup, Designation, Resource Allocation, Mental Fatigue Score etc are selected for the prediction.

3. Pre- Processing of The Data

4. Prediction of Burnout

Machine learning algorithms like Linear Regression, Ridge, Lasso Regressor and catboost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for Burnout prediction.

2.2 machine learning methods

1. KNeighborsRegressor:

KNeighborsRegressor identifies the 'k' nearest neighbors of that data point from the training set based on the chosen distance metric. Then, it computes the average (or sometimes weighted average) of the target values of these neighbors and uses that as the predicted value for the new data point.

2. Lasso Regressor:

The Lasso (Least Absolute Shrinkage and Selection Operator) regression, commonly known as Lasso regression or Lasso regularization, is a linear regression technique used for feature selection and regularization. It extends ordinary least squares regression by adding a penalty term to the loss function, which encourages sparsity in the coefficients of the regression model. The objective of Lasso regression is to minimize the sum of squared residuals between the observed target values and the predicted values, while simultaneously penalizing the absolute values of the regression coefficients. Mathematically, the objective function for Lasso regression

$$\text{minimize} \left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right)$$

3. Ridge Regression:

The ridge regression reduces standard errors by adding a degree of bias to the regression estimates. It is hoped that the net effect will be to provide more reliable estimates. Ridge regression is modifying the least squares method which allows to have biased estimators of the regression coefficients in the regression model. Ridge regression puts a particular form of constraints on parameters.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

4. Linear Regression:

For finding a relationship between two continuous variables, Linear regression is useful. One variable is predictor or independent, and the other variable is variable response or dependent. It looks for a relationship that is statistical but not deterministic. It is said that the relationship between two variables is deterministic if the other can express one variable accurately.

where Y is the dependent variable, X is the independent variable. Theta is the coefficient factor.

$$Y = \theta_1 * X + \theta_0$$

5. Catboost Regressor

CatBoost Regressor is a machine learning algorithm specifically designed for regression tasks. It belongs to the family of gradient boosting algorithms and is particularly known for its efficiency, robustness, and ability to handle categorical features without the need for pre- processing.

3. Results and Discussions

Prediction is conducted using different machine learning algorithms such as linear regression lasso regression, ridge regression, K-NN regression and adaBoost repressor. After analysis weconclude that catboost regression holds good for burnout prediction.

	Employee ID	Date of Joining	Gender	Company Type	WFH Setup Available	Designation	Resource Allocation	Mental Fatigue Score	Burn Rate
0	fffe32003000360033003200	2008-09-30	Female	Service	No	2.0	3.0	3.8	0.16
1	fffe3700360033003500	2008-11-30	Male	Service	Yes	1.0	2.0	5.0	0.36
2	fffe31003003200370039 00	2008-03-10	Female	Product	Yes	2.0	NaN	5.8	0.49
3	fffe32003400380032003900	2008-11-03	Male	Service	Yes	1.0	1.0	2.6	0.20
4	fffe31003900340031003600	2008-07-24	Female	Service	No	3.0	7.0	6.9	0.52
...	...	...	...	...	...	...	...	...	...
22745	fffe31003500370039003100	2008-12-30	Female	Service	No	1.0	3.0	NaN	0.41
22746	fffe33003000350031003800	2008-01-19	Female	Product	Yes	3.0	6.0	6.7	0.59
22747	fffe390032003000	2008-11-05	Male	Service	Yes	3.0	7.0	NaN	0.72
22748	fffe33003003200360039 00	2008-01-10	Female	Service	No	2.0	5.0	5.9	0.52
22749	fffe3400350031003800	2008-01-06	Male	Product	No	3.0	6.0	7.8	0.61

Figure 3.1: data set used in project

```python
def preprocess_inputs(df):
    df = df.copy()

    # Drop Employee ID column
    df = df.drop('Employee ID', axis=1)

    # Drop rows with missing target values
    missing_target_rows = df.loc[df['Burn Rate'].isna(), :].index
    df = df.drop(missing_target_rows, axis=0).reset_index(drop=True)
    # Fill remaining missing values with column means
    for column in ['Resource Allocation', 'Mental Fatigue Score']:
        df[column] = df[column].fillna(df[column].mean())
    # Extract date features
    df['Date of Joining'] = pd.to_datetime(df['Date of Joining'])
    df['Join Month'] = df['Date of Joining'].apply(lambda x: x.month)
    df['Join Day'] = df['Date of Joining'].apply(lambda x: x.day)
    df = df.drop('Date of Joining', axis=1)

    # Binary encoding
    df['Gender'] = df['Gender'].replace({'Female': 0, 'Male': 1})
    df['Company Type'] = df['Company Type'].replace({'Product': 0, 'Service': 1})
    df['WFH Setup Available'] = df['WFH Setup Available'].replace({'No': 0, 'Yes': 1})
    # Split df into X and y
    y = df['Burn Rate']
    X = df.drop('Burn Rate', axis=1)

    # Train-test split
    X_train, X_test, y_train, y_test = train_test_split(X, y, train_size=0.7, shuffle=True, random_state=1)
     # Scale X
    scaler = StandardScaler()
    scaler.fit(X_train)
    X_train = pd.DataFrame(scaler.transform(X_train), index=X_train.index, columns=X_train.columns)
```

Figure 3.2.: Data processing

```
                    Linear Regression trained.
 Linear Regression (L2 Regularization) trained.
 Linear Regression (L1 Regularization) trained.
                 K-Nearest Neighbors trained.
                      Neural Network trained.
Support Vector Machine (Linear Kernel) trained.
   Support Vector Machine (RBF Kernel) trained.
                       Decision Tree trained.
                       Random Forest trained.
                   Gradient Boosting trained.
                             XGBoost trained.
                            LightGBM trained.
                            CatBoost trained.
```

Figure 3.3: Training the dataset using different algorithm

Comparison Analysis

```
                    Linear Regression R^2 Score: 0.87075
Linear Regression (L2 Regularization) R^2 Score: 0.87075
Linear Regression (L1 Regularization) R^2 Score: -0.00001
                K-Nearest Neighbors R^2 Score: 0.85605
                    Neural Network R^2 Score: 0.87242
Support Vector Machine (Linear Kernel) R^2 Score: 0.86897
   Support Vector Machine (RBF Kernel) R^2 Score: 0.88430
                    Decision Tree R^2 Score: 0.81606
                    Random Forest R^2 Score: 0.89753
                Gradient Boosting R^2 Score: 0.90257
                          XGBoost R^2 Score: 0.90310
                         LightGBM R^2 Score: 0.90912
                         CatBoost R^2 Score: 0.90842
```

Figure 3.4: Burnout prediction using different algorithm

Conclusion

Predicting employee burnout using data science techniques involves leveraging various data sources and machine learning algorithms to anticipate and prevent workplace stress. Here's a summary of key aspects in predicting employee burnout using data science

1. Data Collection
2. Feature Engineering
3. Model Selection
4. Evaluation Metrics

By leveraging data science techniques, organizations can proactively identify at-risk employees, implement targeted interventions, and foster a supportive work environment conducive to employee well-being and organizational success.

FUTURE SCOPE

Future work in predicting employee burnout using data science could focus on several areas to enhancethe effectiveness and applicability of predictive models.

Here are some potential directions for future research

1. Incorporating Real-Time Data Streams
2. Personalized Predictive Models
3. Robustness and Fairness
4. Validation and Deployment in Real-World Settings By addressing these future research directions, organizations can harness the power of data science to develop effective strategies for predicting and preventing

employee burnout, fostering a supportive work environment, and promoting employee well- being and organizational resilience.

References

[1] Małgorzata Grządzielewska, "Using Machine Learning in Burnout Prediction: A Survey", Child and Adolescent Social Work Journal, vol.38, no.2, pp.175, 2021.

[2] P. Zhernova, Y. Bodyanskiy, B. Yatsenko and I. Zavgorodnii, "Detection and prevention of professional burnout using machine learning methods", In 2020 IEEE 15th International Conference on Advanced Trends in Radioelectronics Telecommunications and Computer Engineering (TCSET), pp. 218-221, 2020, February.

[3] M. C. Staff, "Job burnout: How to spot it and take action", November 2018, [online] Available: https://www.mayoclinic.org/healthy-lifestyle/adult-health/in-depth/burnout/art-20046642.

[4] Alarcon, G., Eschleman, K. J., & Bowling, N. A. (2009). Relationships between personality variables and burnout: A meta-analysis. Work & Stress: An International Journal of Work, Health & Organizations,23(3), 244–263.

[5] Waheeda Almayyan, developing a Machine Learning Model for Detecting Job BurnoutDuring the COVID-19 Pandemic Among Front-line Workers in Kuwait, Vol. 19, No. 10, October 2021