# EMPLOYEE BURN RATE ANALYSIS

PREMIER PROJECT BY TEAM GITHUB

HDSC’22

**PP22/J611**

* **Scope: Jobs and Career**
* **Topic: Are Your Employees Burning Out?**
* **Project Description:**

**ABSTRACT**

Globally, World Mental Health Day is celebrated on October 10 each year. The objective of this day is to raise awareness about mental health issues around the world and mobilize efforts in support of mental health. According to an anonymous survey, about 450 million people live with mental disorders that can be one of the primary causes of poor health and disability worldwide. These days when the world is suffering from a pandemic situation, it becomes really hard to maintain mental fitness.

In today’s fast-paced world, we frequently find ourselves running here and there in order to get pile-ups done, for any undone task of the day will meet tomorrow’s myriad of activities. Sometimes, we get too busy and forget to take a step back and rest, that is when burnout can occur. According to **WebMD Editorial Contributors**, burnout is a form of exhaustion caused by constantly feeling swamped. It’s a result of excessive and prolonged emotional, physical, and mental stress, in many cases, burnout is job-related. It happens when one is overwhelmed, emotionally drained, and unable to keep up with life’s incessant demands. It is a tricky condition, although not medically diagnosed, burnout can affect your physical and mental health if not acknowledged or treated.

What is more, burnout keeps you from being productive, it reduces your energy, making you feel hopeless, cynical, and resentful. The effects of burnout can hurt your home, work, and social life. Long-term burnout can make you more vulnerable to colds and flu. It can often be confused with stress or escalate into depression. Signs to look out for if you or someone close to you is experiencing burnout: exhaustion, frustration, reduced work performance, lack of concentration etc. The major reasons for burnout include: unmanageable workloads, unfair treatment at work, confusing work responsibilities, lack of communication and support from manager, immense deadline pressure. There are ways to avoid a breakdown for anyone experiencing burnout, namely : getting enough sleep, find support, exercise mindfulness, talk to your supervisor, try a relaxing activity. Thus, this project aims to predict the burnout of employees in a particular organization using machine learning.

**DATA SOURCE AND DESCRIPTION**

We got our Dataset from kaggle, Link - <https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out>

The dataset comprises of 22750 instances with 9 columns representing 9 features in our data-set, below are brief description of what our features represent.

* Employee ID: The unique ID allocated for each employee (example: fffe390032003000)
* Date of Joining: The date-time when the employee has joined the organization (example: 2008-12-30)
* Gender: The gender of the employee (Male/Female)
* Company Type: The type of company where the employee is working (Service/Product)
* WFH Setup Available: Is the work from home facility available for the employee (Yes/No)
* Designation: The designation of the employee of work in the organization. In the range of [0.0, 5.0] bigger is higher designation.
* Resource Allocation: The amount of resource allocated to the employee to work, ie. number of working hours. In the range of [1.0, 10.0] (higher means more resource)
* Mental Fatigue Score: The level of fatigue mentally the employee is facing. In the range of [0.0, 10.0] where 0.0 means no fatigue and 10.0 means completely fatigue.
* Burn Rate: The value we need to predict for each employee telling the rate of Burn out while working. **In the range of [0.0, 1.0] where the higher the value is more is the burn out.**

**DATA CLEANING / PREPROCESSING**

The train set has three columns with missing values.The column with most missing values has a missing values percentage of approximately nine percent and the least with approximately 4 percent missing values.

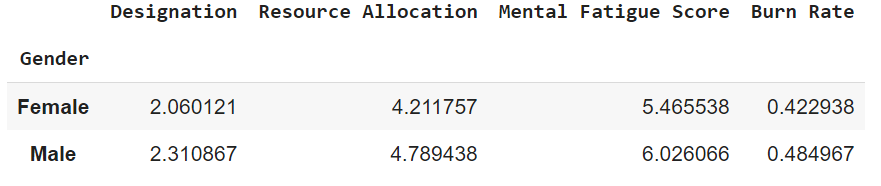
Here we had to drop all instance with missing values and reset the index of the dataframe, the reasons for dropping and not filling in those values are

* The Burn rate is our target variable and the missing values for a target variable should not be filled.
* We worked with the rule of having a tolerance mark of less than 5% missing values, and since the remaining two columns with the missing values have them greater than the 5% mark, we dropped it to avoid getting our model biased.

**EXPLORATORY DATA ANALYSIS**

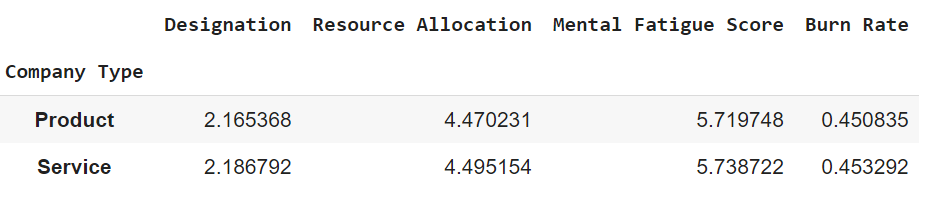
* **DEDUCTIONS**

We were able to segment the employees into different groups and they show to exhibit like characteristics, we were also able to make some deductions with the Analysis done.



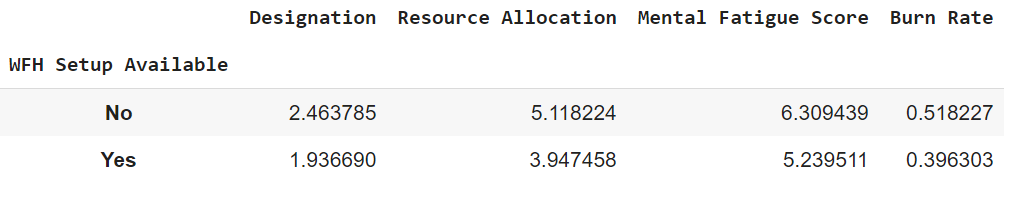
From the data-frame above, we could deduce that

* The males seems to have a more resource allocated to them,
* The males also have an average higher mental fatigue score than the females
* And the males also have an average burn out range greater than the average burn out range of the females



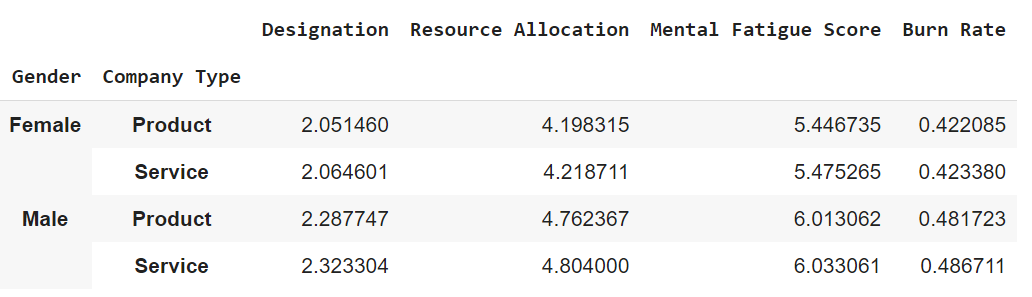
From the data-frame above, we could deduce that

* Those that work in a product company have lesser working hours( Resource Allocation) than those working on with a service company.
* Those with the working in a service company have an average mental fatigue score and burn out range greater than those working with a product company.



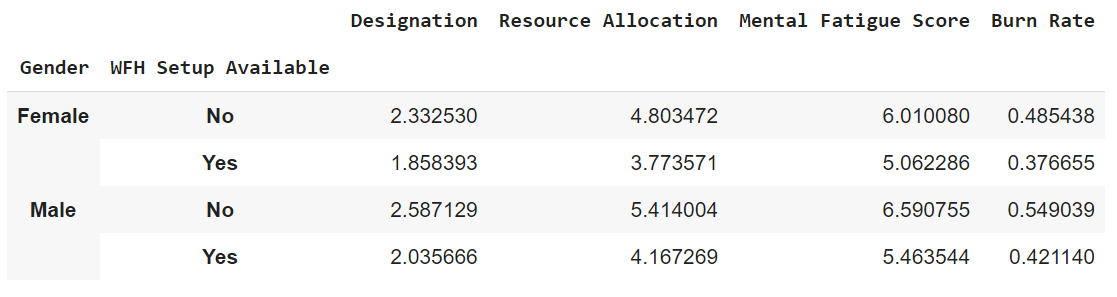
From the data-frame above, we could deduce that

* There is a very significant gap to the working hours (resource allocation) of those who work from home (WFH) and those who don't, with those working from home with a lesser average working time than those who don't.
* The gap is also noticed the mental fatigue score and burn rate of, where those who work from home have a lesser mental fatigue score and burn rate than those who don't.



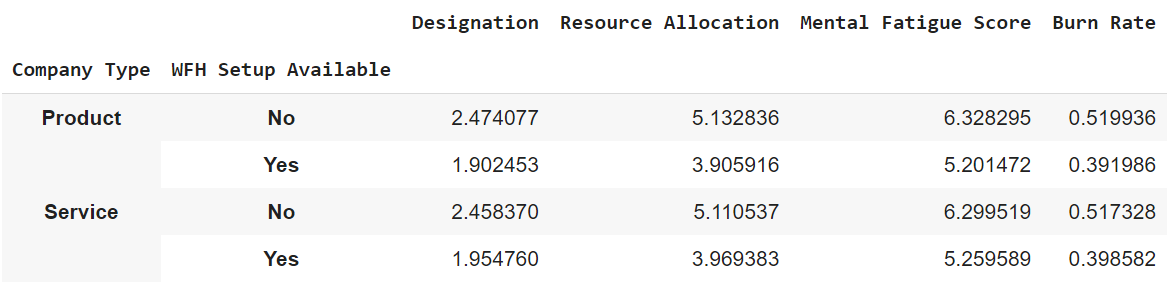
From the data-frame above, we could deduce that

* Females who work in a service company tends to have an average resource allocation (working hours), mental fatigue score and burn rate greater than females that work in a product company.
* Males who work in a service company tends to have an average resource allocation (working hours), mental fatigue score and burn rate greater than males that work in a product company.



From the data-frame above, we could deduce that

* Females who have the work from home setup available tends to have an average resource allocation (working hours), mental fatigue score and burn rate lesser than females that don't.
* Males who have the work from home setup available tends to have an average resource allocation (working hours), mental fatigue score and burn rate lesser than males that don't.



From the data-frame above, we could deduce that

* Employees under a product company, who have the work from home setup available tends to have an average resource allocation(working hours), mental fatigue score and burn rate lesser than that work in a service company and don't have the work from home facilities available.
* Employees under a service company, who have the work from home setup available tends to have an average resource allocation(working hours), mental fatigue score and burn rate lesser than that employees who work in a service company and don't have the work from home facilities available

We did our EDA and visualization mainly based on three grounds

* The Univariate Analysis : The distribution and characteristics of unique variable in each feature
* The Bivariate Analysis. : The relationship between two features, especially with the target feature.
* The Multi-Variate Analysis. : The relationship between more than two features.

**FEATURE ENGINEERING**

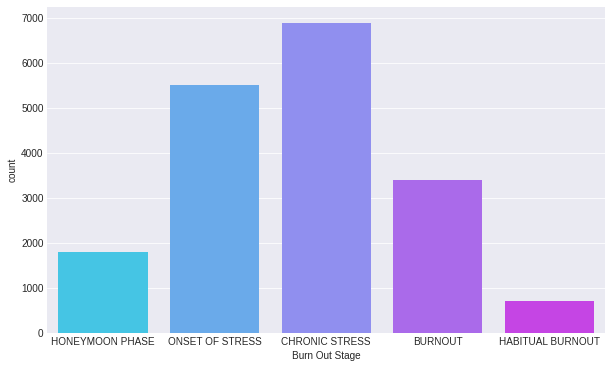
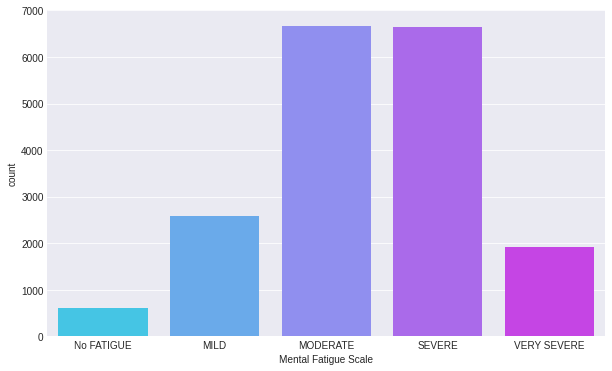
* We converted our date of joining to a date-time feature, separate the year, month and day into different separate columns, then we dropped the initial date of joining column
* To aid our analysis and visualization, we binned our continuous features (Mental fatigue score and Burn rate into segment of equal intervals.
* **HONEYMOON PHASE** are value range from 0.0 - 0.2 in our Burn out stage columns
* **ONSET OF STRESS** = 0.2 - 0.4
* **CHRONIC STRESS** = 0.4 - 0.6
* **BURNOUT** = 0.6 - 0.8
* **HABITUAL BURNOUT** = 0.8 - 1.0

They are also grouped in an ordinal manner with -'No FATIGUE' < 'MILD' < 'MODERATE' <

'SEVERE' < 'VERY SEVERE'

Same applies with the Mental fatigue **Score** that was binned into Mental fatigue **Scale**

**Note : These categories added features/ columns and the parent feature it was obtained from still remains in our data.**

****

***Fig1.1 – Binned feature (Burn out Stage) visualized fig1.2 – Binned feature (Mental fatigue scale).***

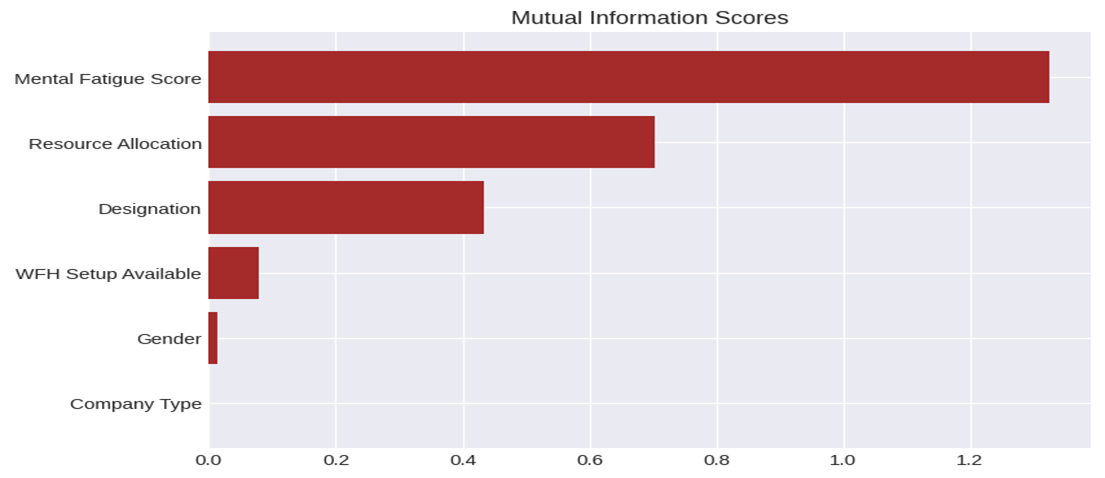
***With a countplot***

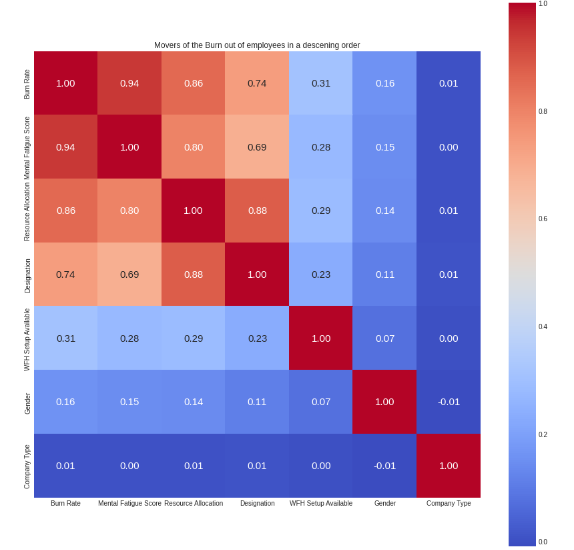
* The added features were dropped and our data was rescaled and standardize to attain normality (Gaussian-Distribution) for the sake of modeling.

**CHECKING FOR DEPENDENCIES FEATURE CORRELETIONS**:

We found the correlation coefficients of each feature with the target and every feature and made use of the sea-born library tool heat-map( plasma soup) to visualize this, the correlation ranges from -1 to 1, with 1 being the most positive correlation and -1 being the most negative correlation.

We also made use of the mutual\_info\_regression module to check for the dependencies, that is, it is calculated between two variables and measures the reduction in uncertainty for one variable given a known value of the other variable. and we visualized it using a barh plot.





**MODELING AND EVALUATIONS AND DEPLOYMENT**

We tried out a total of six models, with linear regression giving us the best results

**Linear Regression Model**

* The model was created with the default hyper-parameters, fitted into the train set and was evaluated using r2-score and RMSE as metrics. **r2\_Score - 0.928,** **RMSE - 0.055**
* **Linear Regression with Polynomial Features**

The model was done by creating a pipeline to find the polynomial features of our data before it’s application, we record approximately 10% improvement with the r2\_score accuracy using this model. **r2\_Score - 0.927,** **RMSE - 0.053**

**Other models used are :**

**Random Forest with Polynomial Features. : r2\_Score - 0.927, RMSE - 0.053**

**XGBoost Regressor. : r2\_Score - 0.928,** **RMSE - 0.053**

**Light GBM Regressor : r2\_Score - 0.923,** **RMSE - 0.055**

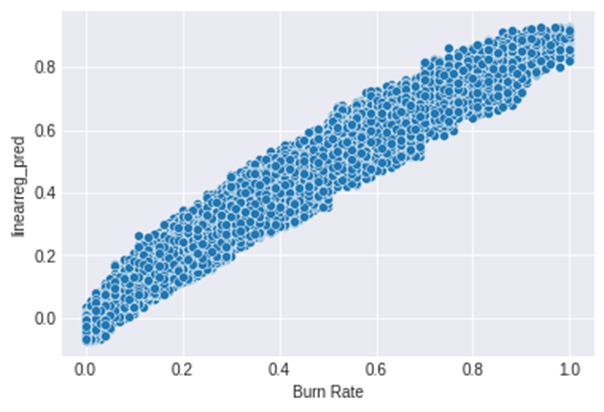
**GradientBoostingRegressor : r2\_Score - 0.922,** **RMSE - 0.055**

**Neural Net Model – RMSE - 0.053**

The model was deployed with Flask for end users to enable them determine the stage of their burnout.

GIthub Link - <https://github.com/HamoyeTeamGithub/Hamoye-Premier-Project--22---PP22-J612---Team-Github-->

1. **A scatterplot showing the distribution of the actual Burn Rate to the predicted values by the linear regression model.**



**2. A scatterplot showing the distribution of the actual Burn Rate to the predicted values by the linear regression model with polynomial features.**

