PREDICTING EMPLOYEE BURNOUT

Meghavi Singhaniya, Akash Goyal, Bindu Raghu Naga, Alfredo Sandoval, Meenakshi Sundaram

DESCRIPTION OF PROJECT GOALS

Importance of the Problem

"To win in the marketplace, you must first win in the workplace" - Doug Conant. In today's fast-paced and competitive business landscape, organizations are increasingly recognizing the significance of their most valuable asset: their employees. The well-being and productivity of the workforce directly impact an organization's success and overall performance. However, in recent years, a concerning trend has emerged – the rise of employee burnout.

Employee burn rate refers to the rate at which employees experience burnout, stress, and disengagement in the workplace. It is a complex phenomenon that affects individuals across industries and job roles, transcending geographical boundaries. According to Deloitte's 2015 external workplace well-being survey, three-fourths of employees have experienced burnout at specific periods. This alarming statistic highlights the pervasive nature of the issue and the urgent need to address it.

This report aims to shed light on the critical issue of the "Employee Burn Rate" and its implications on both employees and organizations. By understanding the factors contributing to burnout and implementing strategies to mitigate its effects, organizations can foster a healthier and more engaged workforce, leading to increased productivity, reduced turnover, and ultimately, a competitive advantage in the business landscape.

Description of Problem

In this project, data associated with employees of a company is analyzed, and predictions over which employees are at a high risk of experiencing 'burnout' were made. The primary objective is to leverage data-driven insights from the dataset, which includes various factors that may contribute to burnout, such as resource allocation, mental fatigue rate, and WFH setup. This project will enable us to identify individuals who may be at risk of burnout before it becomes a critical issue.

DATA CLEANING AND PREPROCESSING

Handling Missing Values: Three variables were identified with missing values i.e., Resource Allocation, Mental Fatigue Score, and Burn Rate. The rows containing these values were dropped reducing the total number of rows to 18,590 rows.

Mapping Categorical Variables: Three categorical variables, i.e., Gender, Company Type, and WFH

Dataset: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out

Setup Available, were converted into binary values.

OneHot Encoding: The Designation column is hierarchical in nature, with no measurable unit. To capture any non-linearity associated with the variable, one-hot encoding was required. This approach treats each designation independently, enabling us to represent the categorical information accurately.

Feature Engineering: DOJ column in the dataset was converted to a new variable called Days_count. This variable represents the number of days between the date of joining of the most recent employee and the date of joining of the oldest employee.

Dropping Unimportant Columns: To streamline the dataset, the Employee ID and Date of Joining columns were dropped, as they do not provide significant value for the analysis and modeling.

Train-Test Split: To prepare for model training and evaluation, a 70 – 30 % train-test split was made.

By completing these data cleaning and preprocessing steps, the foundation for the predictive model to identify employees at risk of burnout was set.

EXPLORATORY DATA ANALYSIS

This next phase involved gaining valuable insights into the dataset, which helped understand the characteristics and relationships between different features. Here are the key findings:

Count plots of features:

- 52.4% of the employees in the dataset are female.
- Approximately 54% of the employees had the option to work in a hybrid (WFH) setup.
- A majority of employees (around 33%) belong to designation level 3, while only a small percentage (1.7%) are in designation level 5.
- The "days_count" variable, which was binned into groups of 90 days, shows a relatively even distribution of employees across the brackets, with around 4,600 employees in each bracket.

Burn Rate Profiling:

- Employees experiencing a high burn rate (in the 4th quartile range) tend to work long hours, approximately between 6.8 to 10 hours.
- Employees in higher designation levels (4 and 5) experience higher burnout rates compared to those in lower designation levels.
- Notably, no employees from designation levels 0 and 1 fall into the high burn rate group, whereas all 317 employees categorized as high burn rate fall into higher designation levels. This suggests that designation level and resource allocation may have a significant impact on the burn rate of employees.

Correlation Analysis:

- The correlation heatmap reveals that resource allocation and mental fatigue score have a strong positive correlation with burn rate. This indicates that employees with higher resource allocation and mental fatigue scores are more likely to experience burnout.

- Designation levels 3, 4, and 5 are also positively correlated with burn rate, suggesting that employees with higher job levels may be more susceptible to burnout.
- On the other hand, the availability of a WFH setup and lower designation levels (0, 1, and 2) show negative correlations with burn rate. This suggests that employees with WFH options and lower job levels may experience lower burnout rates.
- Gender, company type, and days_count variables show no significant correlation with burn rate, implying that these factors do not directly influence burnout levels.

These findings provide crucial insights into the factors influencing employee burnout within the organization. By understanding the relationships between burn rate and various features, better strategies and interventions can be made to mitigate burnout and promote employee well-being.

SOLUTION AND INSIGHTS

The next steps in the analysis involve building a predictive model using these insights to identify employees at high risk of burnout and implement targeted measures to address the issue effectively.

Model Selection: In this, various machine learning algorithms to determine the most suitable model were explored for our prediction task.

Linear Regression considering the p-value: Workplace factors' impact on employee performance was analyzed initially using p-values (<0.05). The model gave a training RMSE of 0.0552, and a test RMSE of 0.0548.

Notably, gender, the availability of a Work-From-Home (WFH) setup, resource allocation, and mental fatigue score were found to have a statistically significant impact on employee performance.

PCR: The analysis aimed to find the optimal number of principal components (M) and identify significant features for prediction. Cross-validation was used to determine the best M, with the optimal value found to be 10. Using this optimal M, a PCR model was trained and evaluated on the test set, achieving an RMSE of 0.0548, indicating good prediction accuracy.

Lasso Regression: Features were standardized and optimal alpha (0.01) was chosen via 5-fold cross-validation. The model predicted an RMSE of 0.057 and R-squared of 0.917. The model focused on 'Resource Allocation' and 'Mental Fatigue Score,' simplifying for interpretability and shedding light on key factors impacting employee well-being and productivity.

Ridge Regression: Features were standardized using Standard-Scaler and optimal alpha (0.01) chosen via 5-fold cross-validation. The model exhibited a strong fit with explaining 92.2% of "Burn Rate" variance. RMSE was low at 0.055, affirming accurate predictions. 'Mental Fatigue' and 'Resource Allocation' were most impactful, while 'Company Type,' 'designation 4.0,' and 'days count' had minimal influence on predictions.

K-Nearest Neighbors: Using 5 fold cross-validation K=9 was chosen and model was fitted on the entire training data. The training RMSE (0.062) and test RMSE (0.066) were obtained with the fit. Variable importance was determined using permutation importance function. 'Resource Allocation' and 'Mental Fatigue Score' emerged as most significant variables.

Decision Tree: This optimal decision tree, having a maximum depth of 8 levels, only splits groups with at least 10 observations, and each terminal node contains a minimum of 4 observations.

Using this model, an RMSE of 0.055 was obtained, which translates to a 12.16% error.

Random Forest: Employing 5-fold cross-validation on training data, the optimized parameters for RF obtained were n estimators (100), max depth (8), max features (12), min samples split (10), min samples leaf (2). The resulting RMSE of 0.053 reflects an 11.79% error compared to the mean value of the sample burnout rate.

The model indicates that the most influential feature determining burnout rate is the mental fatigue score, followed by resource allocation and the number of days an employee has been working.

CONCLUSION

As various predictive models were explored [Table 1], Random Forest emerged with superior predictive prowess.

Unraveling the dynamics of burnout, two pivotal determinants i.e., Mental Fatigue Score and Resource Allocation (Billable Hours) wielded the most substantial influence in predicting Burn Rate, underpinning the critical importance of holistic employee well-being and balanced workloads. Conversely, other variables such as days count and gender exert minimal to negligible effects.

In conclusion, our comprehensive analysis not only identifies the best-performing predictive model but also unveils actionable insights that hold the potential to transform workplaces into thriving, sustainable environments. The synergy between employee well-being and business success is undeniable, and by heeding these recommendations, organizations can pave the way to a brighter, healthier future for both employees and the bottom line.

Dataset: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out

APPENDIX

Table 1

Model	**Training RMSE**	**Test RMSE**	**Error %**	**Parameters**
Linear Regression (p-values)	0.0552	0.0547	12.10%	*3 features dropped*
Lasso	0.057	0.056	12.44%	*2 features selected*
Ridge	0.0552	0.0548	12.17%	*All features selected*
KNN	0.062	0.066	14.6 %	*Optimal K = 9*
Decision Tree	0.052	0.055	12.16%	Max Depth=8 N estimators=100 Min samples leaf=4 Min samples split=10
Random Forest	0.051	0.053	11.79%	Max Depth=8 Max Features=12 N estimators=100 Min samples leaf=2 Min samples split=2

List of Figures

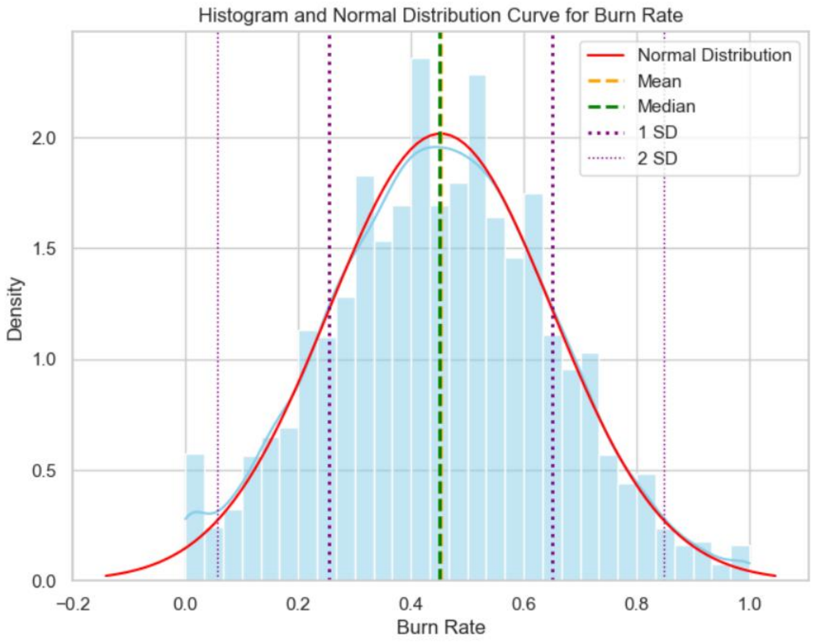
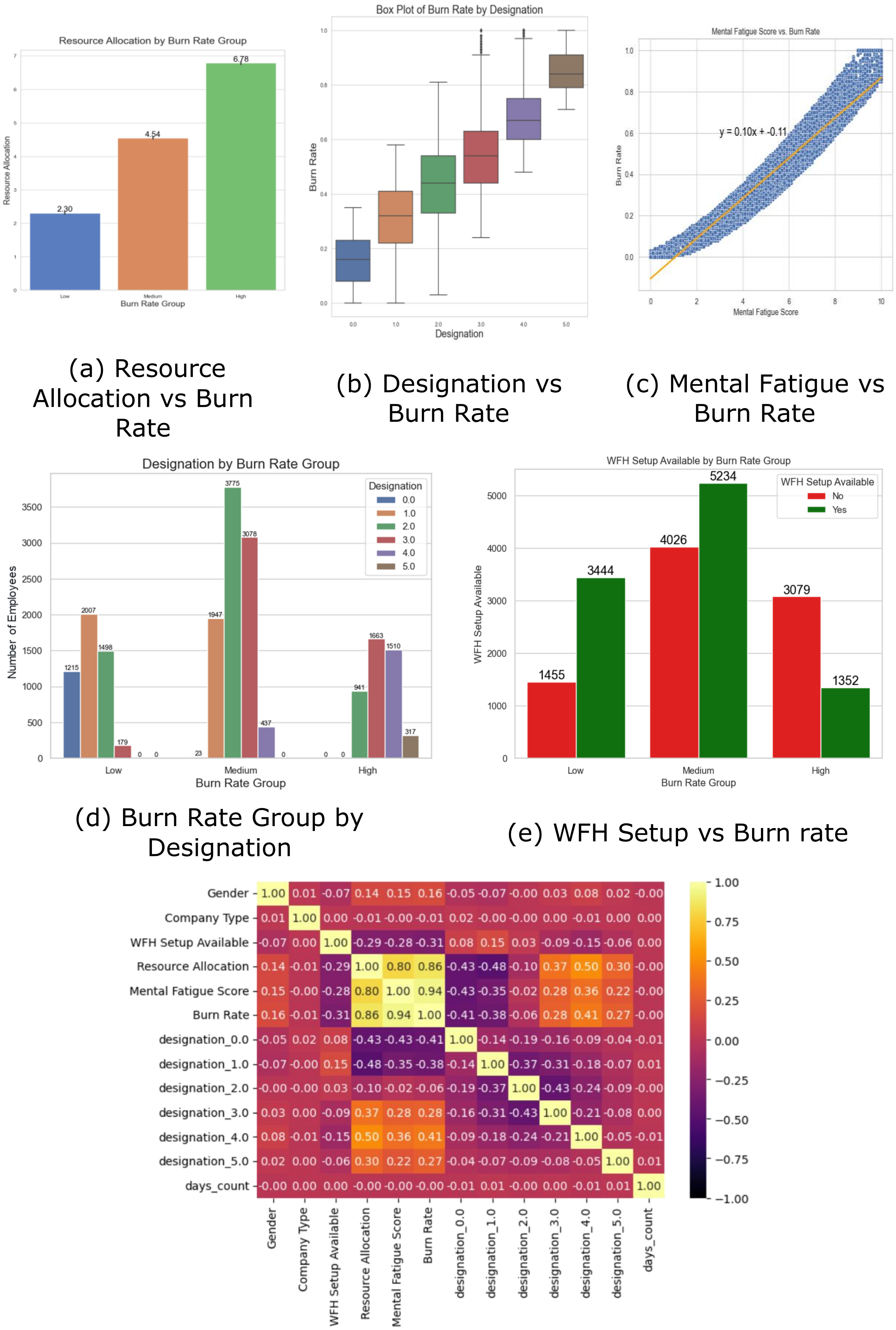


Fig.1: Histogram and Normal Distribution Curve for Burn Rate

Dataset: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out



(a) Resource Allocation vs Burn Rate

(b) Designation vs Burn Rate

(c) Mental Fatigue vs Burn Rate

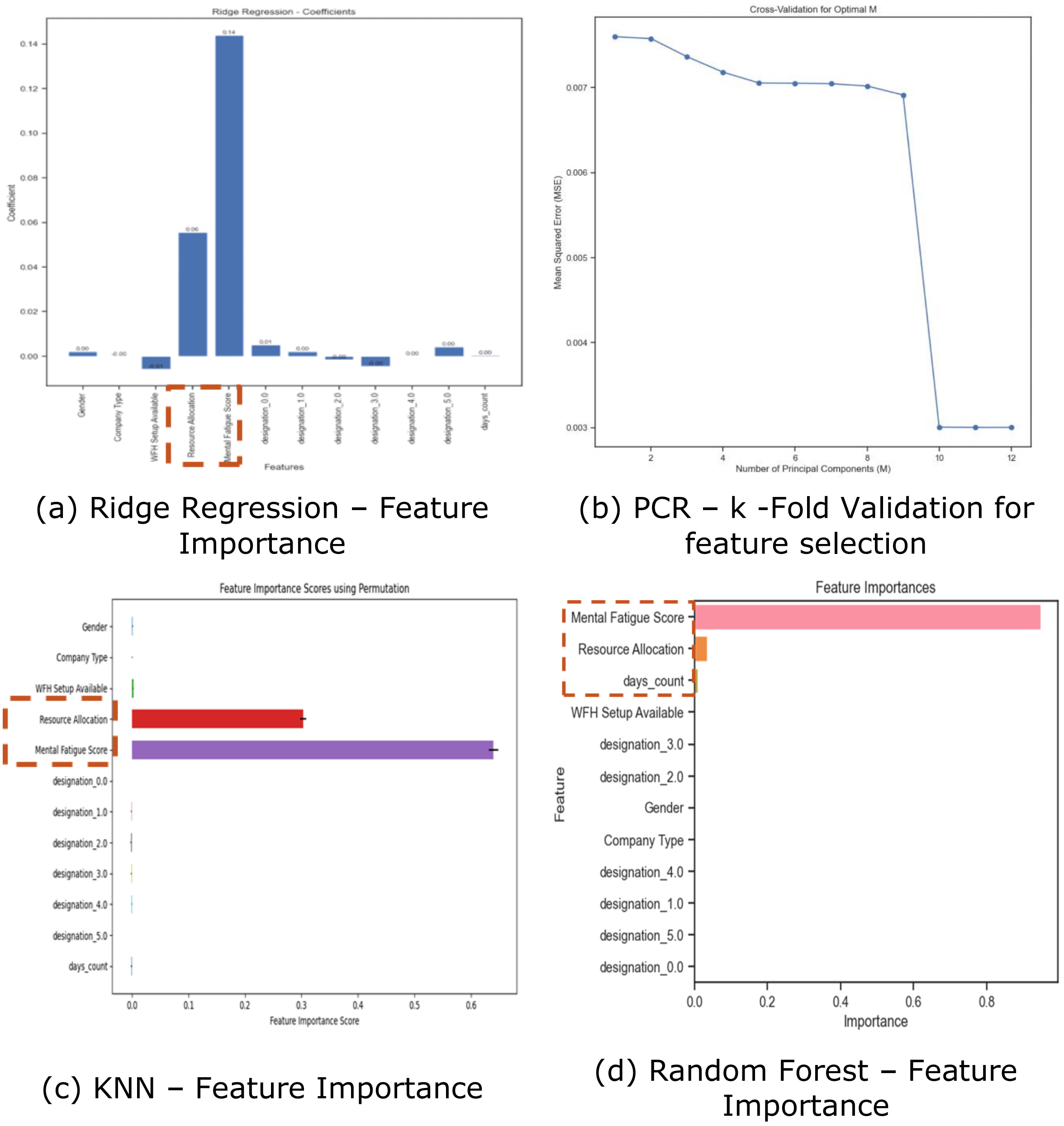
(d) Burn Rate Group by Designation

(e) WFH Setup vs Burn rate

(f) Heat Map for feature importance

Fig.2: EDA Visualization

Dataset: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out



(a) Ridge Regression – Feature Importance

(b) PCR – k -Fold Validation for feature selection

(c) KNN – Feature Importance

(d) Random Forest – Feature Importance

Fig.3: Model Analysis and Feature Selection

Dataset: https://www.kaggle.com/datasets/blurredmachine/are-your-employees-burning-out