Fernando Chapa, Chang Li, Monica Martinez, June Yuan
MIS 381N Final Project
Dr. Chakrabarti
August 8th, 2022

Employee Burnout Analysis

Project Goals

This project investigates the "Burn Rate Dataset" to identify substantial factors that influence employee burnout and trains statistical models to help predict burnout rates. Findings and predictive power will help businesses understand their employees better and provide managers with information to adjust workplace settings and requirements in order to decrease employee dissatisfaction, turnover rate, and improve performance.

Exploratory Data Analysis

The dataset we are working with contains six predictors: gender (male or female), company type (service or product), work-from-home setup availability (yes or no), job level (intern, entry-level, associate, supervisor, executive), hours worked (from 20 to 70 hrs/wk), and mental fatigue score (from not at all to extremely). These factors are then used to predict the employee's burn rate, which is a continuous variable from 0.0 to 1.0.

Through exploratory data analysis, we used statistical graphs and other visualization techniques to find different patterns in the data and come up with our initial hypotheses. We began by performing a stratified train-test split of our data using a 70/30 split and making sure the class imbalance in the two sets was similar. By checking correlation statistics for the continuous variables, we found that the average job level is an entry-level position, the mean hours worked is about 45/week and the mean mental fatigue score is 5.7 out of 10. Next, we created some data visualization plots as shown in figure 1.

Predictive Model Analysis

To begin our modeling analysis, we performed multi-linear regression using the burn out rate as our dependent variable. We used Ordinary Least Squares to estimate the parameters in our model. After running the model, we observed all variables to be statistically significant, with gender and job level having the highest p-values of the predictors between $2.9e\text{-}10<p<8.8e\text{-}08$. Our R^2 was about 0.9210, which means that our six explanatory variables are able to explain 92% of the variation in burnout rate.

To simplify the model, we took out the variables with the highest p-values: gender and job level. The simplified model has an R^2 of 0.9203, and notice that it hasn't changed much from the previous model. This implies that the two variables we excluded from this model were not contributing significantly to the model. Now we can look at our p-values. Our p-values for number of hours worked and mental fatigue score are 0.000 and the p-value for the work from home variable is 2.0353e-43. These p-values are the strongest predictors for burn out rate. All predictors but work from home have a positive coefficient, which makes logical sense. If employees have more hours they work and a high mental fatigue score, their burnout rate will be higher. We set our test and training sets and calculated for the model's MSE. Looking at Figure 2, we can deduce that it is probably a good model for prediction, but we calculated the MSE to be certain. Our MSE is 0.0463, which confirms our inference that OSL for multi-linear regression is a good model and predictor.

Other Models

Naive Bayes

Firstly, in Naive Bayes, we made a classification: take Burn Rate as the observed value; those less than 0.49 are classified as not needing a lot of attention, and those greater than 0.49 are classified as fatigued employees and labeled as our target. After that, except Burn Rate, our dataset has one continuous variable: Mental Fatigue Score, and two variables that appear to have continuous values but are actually categories: Job Level and Hours Worked; we discretize these variables which take integers or floats as statistical results into bins, and create the design matrices based on producing new dummy columns for these binned variables and concatenate

them into the old dataframe. In this case, after setting up the classifier, we test this predictor on the training data and get 0.903 for the accuracy score, which is a good fit.

K-Nearest Neighbors

Secondly, in KNN, since this model is similar to Naive Bayes, which is suitable for classification conditions, we adopt a similar classification approach. To start with the building model, and because we need Burn Rate to be the basis for classification, we use the continuous variable Mental Fatigue Score and the category variable Hours Worked to be the features. Then, same as what we did in Naive Bayes, we classified greater than 0.49 for the employees who need more attention, and less than 0.49 for the employees who need less attention. Next, we create design matrices based on the two features and plot the data to have a more sensible conception. According to figure 3, we can deduce that the combination of Mental Fatigue Score and Hours Worked have a strong impact on Burn Rate. The increase on both of the variables will cause the Burn Rate to be closer to 1. We use the train set data that had been split before EDA, and get 0.919 for the accuracy score after setting up the classifier. Which is also a good fit, and has a more accurate prediction than Naive Bayes.

Random Forest

Finally, in the Random Forest model, we get the hyperparameter which is obtained from grid searches. However, using grid searches will make the model run very slowly. But a good point is that by using it, we can know which variables are more significant. According to figure 4, the ranking of variables from most to least important is: Mental Fatigue Score, Hours Worked, Job Level, and WFH Setup Availability. The result we get for the mean score is 0.924, a little bit higher than other models, which means it's not only a good fit, but also a best fit in all of our models.

Solution and Insights

As indicated by model results, mental fatigue score and number of hours worked are the most influential factors to predict an employee's burnout rate. Businesses should consider adjusting employee's working hours and increase importance of employee's mental health by organizing

more team events, granting access to counselors, and more. If applicable, businesses could also set up the option of working from home, since adding this option would decrease employee's burnout rate.

Stakeholders can also make use of these statistical models to keep track of how burned out employees are, and make adjustments in a timely manner.

One limitation of this study is that certain variables are not clearly defined. For instance, Hours Worked is a relative score, but it is not a quantitative measure of hours. Similarly, a burnout rate of 0.8 for example, is up for interpretation by the reader.

Following this analysis, business stakeholders could further look into how to decrease employee burn rate and dissatisfaction by training or improving models with discrete variables, which are more measurable.

References

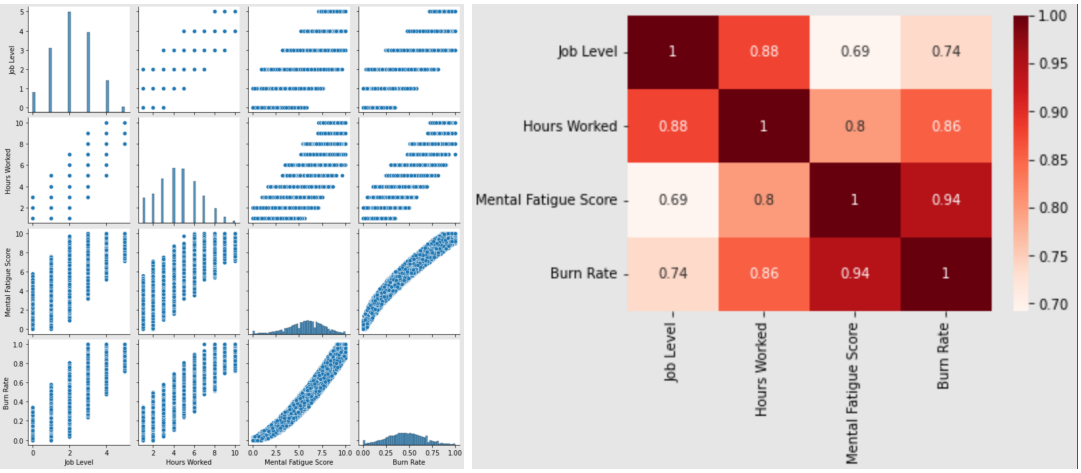

Figure 1. - Pair plots and heat map of job level, hours worked, mental fatigue score, and burn rate

	Real_Values	Predicted_Values
0	0.26	0.313433
1	0.49	0.402237
2	0.49	0.474014
3	0.66	0.689059
4	0.49	0.536699
...	...	...
3713	0.49	0.492297
3714	0.63	0.588115
3715	0.25	0.258969
3716	0.92	0.854962
3717	0.63	0.673823

Figure 2. - Actual Values vs. Forecast Values

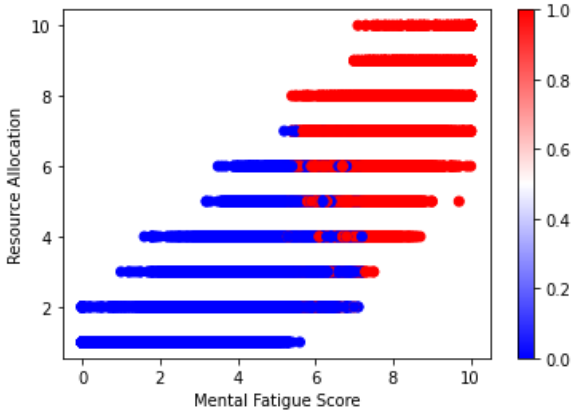

Figure 3. The influence degree of the combination of two variables on Burn Rate

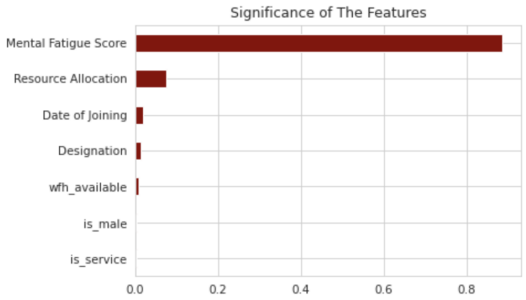

Figure 4:Significance of Feature