2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) | 978-1-7281-9772-2/20/$31.00 ©2020 IEEE | DOI: 10.1109/RTEICT49044.2020.9315726

Predicting Employees under Stress for Pre-emptive Remediation using Machine learning Algorithm

Anusha Garlapati
Dept of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India.
garlapatianusha@am.students.amrita.edu

Doredla Radha Krishna
Dept of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India.
doredlakrishna4@am.students.amrita.edu

Kavya Garlapati
Dept of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India.
garlapatikavya@am.students.amrita.edu

Gayathri Narayanan
Dept of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India.
gayathrin@am.students.amrita.edu

***Abstract*—With the ongoing COVID-19 pandemic, businesses and organizations have acclimated to unconventional and different working ways and patterns, like working from home, working with limited employees at office premises. With the new normal here to stay for the recent future, employees have also adapted to different working environments and customs, which has also resulted in psychological stress and lethargy for many, as they adapt to the new normal and adjust their personal and professional lives. In this work, data visualization techniques and machine learning algorithms have been used to predict employees stress levels. Based on data, we can develop a model that will assist to predict if an employee is likely to be under stress or not. Here, the XGB classifier is used for the prediction process and the results are presented showing that the method facilitates getting a more reliable model performance. After performing interpretation utilizing XGB classifier it is determined that working hours, workload, age, and, role ambiguity have a significant and negative influence on employee performance. The additional factors do not hold much significance when associated to the above discussed. Therefore, It is concluded that concluded that increasing working hours, role ambiguity, the workload would diminish employee representation in all perspectives.**

Keywords -Data Visualization, Machine Learning Algorithms, people analytics, Employee stress.

I. INTRODUCTION

On March 11, 2020, the World Health Organization (WHO) reported coronavirus (COVID-19) a pandemic that signifies a global, epidemic disorder frightening the entire universe [3]. COVID-19 is a contagious disease affected by the coronavirus. 'Coronaviruses' are a huge family of viruses that cause ailments varying from the typical flu to other critical complications. According to WHO, on March 31, 2020, the virus had reached 202 countries. Due to this, stock markets and other sectors have experienced a severe downturn in growth. This, in turn, affects employees too, who feel stressed when they are unable to cope with prolonged uncertainty and pressure. The application of machine learning and artificial intelligence to the field of business is seeing a lot of promising growth. The pattern of employee behavior is analyzed in [11].

Vis-à-vis, they do not have any satisfaction due to long working hours in addition to having a heavy workload. Here, the foremost objective of this research is to analyze the consequence of stress on employee appearance. Moreover, this influences physical ailments and a lack of commitment to work. However, in the contemporary situation, COVID-19 has put the world population in an unprecedented position. Through this work, we intend to analyze the stress level that employees are subjected to owing to a phenomenon like the present pandemic. Here, machine learning algorithms are used to predict whether employees undergoing stress or not.

II. LITERATURE SURVEY

Workplace stress is a thriving concern for employees like human resource managers and so on. Although considerable scholarly and practical awareness has been dedicated to stress management over the years, the time has come for distinct perspectives and research. Extracting from the emerging field of organizational performance, this research proposes analysis conclusions including implications for combating occupational stress.

Specifically, data from a large sample of working employees beyond a variety of organizations and industries suggest that positive resources of efficacy, optimism, and resilience may be key to better understanding adaptation in perceived manifestations of stress. Numerous investigations and experimentation has been done in the last few years, most of the studies have been conveyed in countries that endeavor to promote to enhance advanced economically and socially. stress has become one of the most widespread 'occupational disorders', Aloft the past years, close to 3 billion employees" are experiencing stress at their workplace and it is influencing their overall job performances on regular basis.

III. DATA SET

Each row in the data set belongs to a unique individual and in detail. Here, data has parameters like Employee ID, Target, Age, Average Daily Hours, and so on. The data is in the format of CSV (Comma-Separated-Values). The whole data set is divided into two parts – first, is the training data set which is fed into our machine learning algorithm to tell our model about the data, and second is the test data set which can be used to test our model performance. The structure of data including sample rows is as follows:

TABLE1: Data set

Employee ID	Target	Age	Avg Daily Hours
100001	0	36.0	6.45
100002	0	24.0	8.48

TABLE 2: Data set

Department	Education	Gender	Job Role
Sales	Technical Degree	Male	Manufacturing Director
Sales	Technical Degree	Male	Sales Representative

TABLE 3: Data set

Working Hours	Flexible Timings	Workload level
8.0	No	Low
1.0	No	High

TABLE 4: Data set

Work satisfaction	Years at company	Monthly Income
Medium	8.0	175000
Very High	0.0	16667

TABLE 5: Data set

Performance Rating	Percent salary hike	companies worked
1	21	4
3	11	1

TABLE 6: Data set

Worklife balance	Training Times last year	Micromanaged at work
2	2	3.0
2	5	4.0

TABLE 7: Data set

Is Individual Contributor	Years with Current manager	Marital status
Yes	8.0	Divorced
Yes	0.0	Married

TABLE 8: Data set

Education Field	Years since the last promotion	Self -Motivation level
5	1.0	3
5	0.0	2

This is some sample data. Before implementing any model or algorithm on any dataset we need to interpret the data by using exploratory data analysis using Python or R. Here, obtained results can conjecture data better and can analyze the correspondence between variables in data. It can be examined by using diverse representations of plots like pair plot, box plot, violin plot, heat map, scatter plot, histogram, pie charts, and so on.

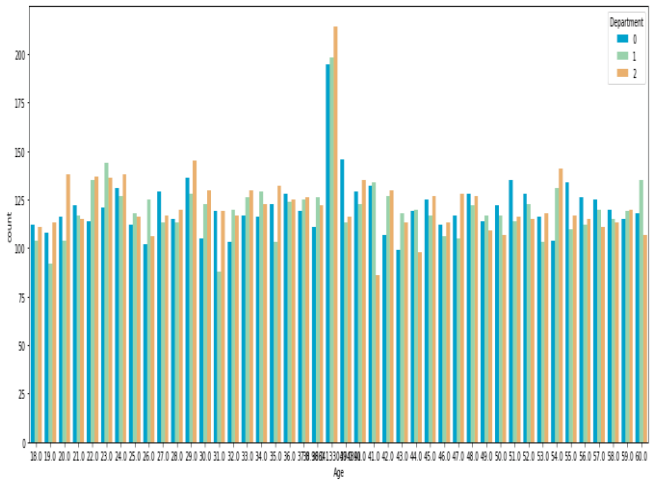

Figure1– Frequency distribution of employees with age

The X-axis is the Age and Y-axis is a count for Age. Here, the numbers are coded as: 0- Sales Department, 1- Research and development, and 2- Human Resources. The obtained result shows that Human Resources department employees record maximum age with starting value – 18, ending value – 60 and step-size: 1

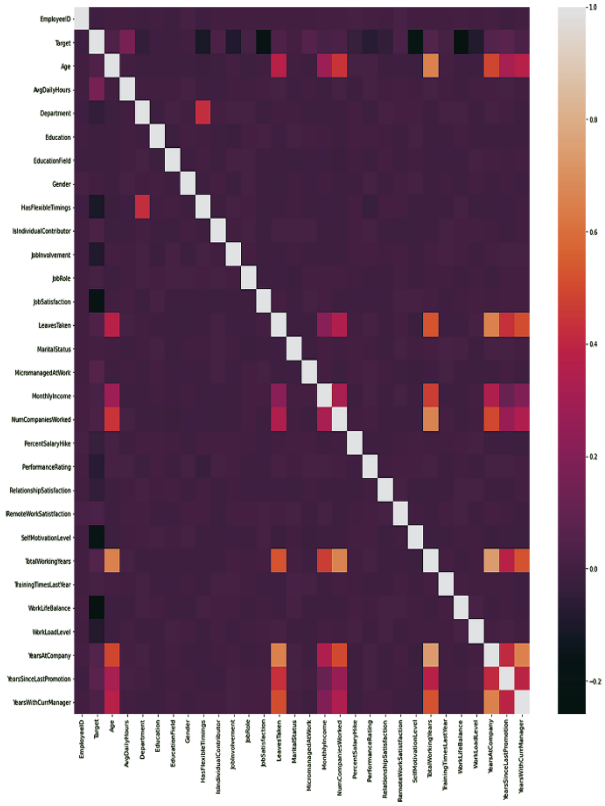

Figure2 – Correlation plot for the obtained data

The above figure (Figure2) represents the correlation plot for the overall data. It shows the relationship among all variables whereby every individual is alternated with the remaining variables. It will examine which are the most significant variables in our data and can progress for distant steps. Here, variables in the X and Y axis are column names in the dataset and those column names are quotes at the inception.

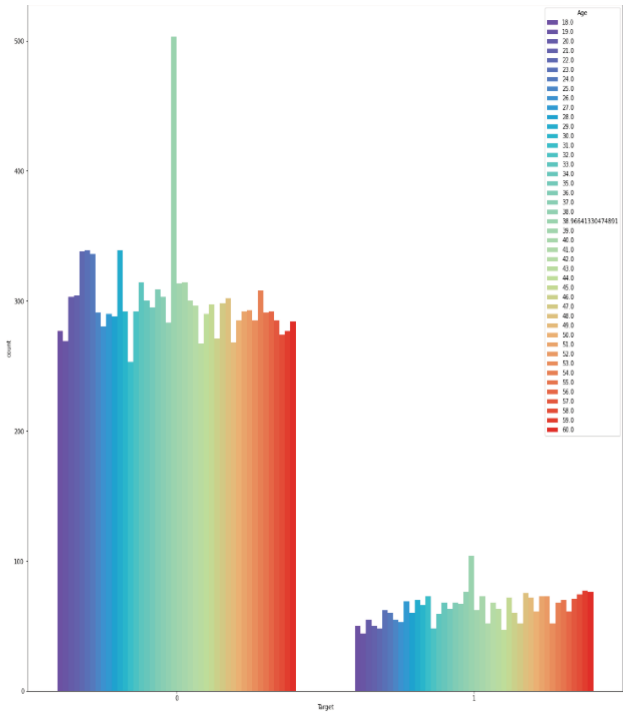

Figure3 – Histogram plot of employees under stress

The Figure 3 is an indication of the age group of employees who are undergoing greater stress. Therefore, the obtained results show that the age group 35-38 is subject to more stress than others.

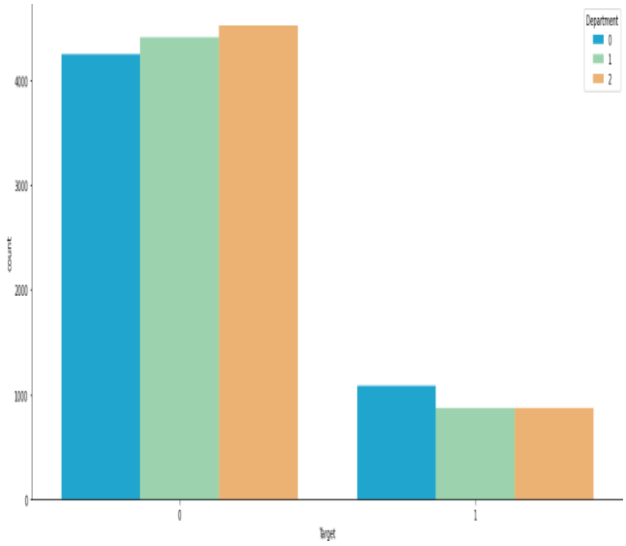

Figure4 – Department wise distribution of employee under stress

In Figure 4, the X-axis is its Age, Y-axis represents the count and the hue is 'Department'. The acquired results project which department employees are enduring more stress. we infer that the employees who are in the sales department are undergoing more stress when compared to others.

These are some plots used to analyze the data. Prior to data analysis, data cleaning techniques were implemented with regulations like handling null values and retaliating them. This will help to get more performance and accuracy of the model that is going to assist the stress levels of employees.

IV. METHODOLOGY

One of the best important steps while dealing with data is cleaning the data. Without doing that, if we go for model execution, we can not obtain better performance in model execution. Therefore, it ought to deal with null values, zeros, NAN values.

Here, data has 3895 null values. The categorical null values are fixed by using mode, and numeric data is fixed by using mean, median, floor techniques. Here, altogether, it can drop null values also, but by doing that, it may dissipate some data. So that is preferred for good model execution.

TABLE9: Target symmetry

Class 0	13180
Class 1	2820

This is the target variable count. By this, it shows that data is imbalanced and it can be balanced by using various techniques.

It is also possible to implement resampling data, oversampling, under-sampling data, or either it can be done by using 'smote', hyperparameter techniques. It can also be overcome by using better algorithms that can give the most reliable performance model or can be done by using bagging and boosting techniques, algorithms like XGB Classifier (XS Boost), PCA (Principal Component Analysis), right evaluation metrics, changing performances metrics, ROC curve.

XG Boost: This is an optimized dispensed gradient boosting library. XG Boost is an accumulate learning method. And also Provides more dependable explications than other machine learning algorithms. XG Boost is more durable than other ensemble classifiers and confers more high-grade performance on a variety of machine learning data sets. Internally has parameters for cross-validation, regularization, user-defined objective functions, and missing values.

The proportion of the target variable is of symmetry 4.67:1. There are some techniques to balance this data. Xg boost, PCA, Gain charts, Smirnov charts. We need to standardize the data before applying algorithms as the data consists of categorical values also. Numeric data is done by using Standardization whereas categorical data is done by using Label Encoder. It ought to only standardize numeric data and should not standardize or normalize categorical data.

PCA: Used in Exploratory data analysis and for constructing predictive models. It is commonly utilized for dimensionality reduction by forecasting each data point onto only the first few principal components to capture lower-dimensional data while impersonating as much of the data modifications as possible. It is simply like the simplest true Eigen vector based multivariate analysis and closely related to factor analysis.

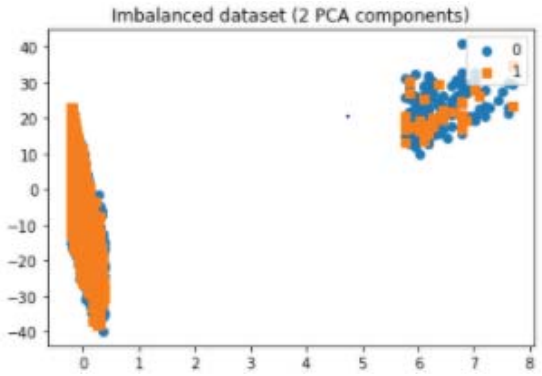

Figure 5– X-axis: train data for PCA, Y-axis: Count

when dealing with categorical data, the prerequisite to apply label encoder or One-hot encoding that depends on the data set. The below is sample standardized array data.

array([[-1.73194256, -0.24459007, -0.72185123,0.5931524,
-0.4682794, 2.47620414],
[-1.73172605, -1.23402768, 0.38921277, -0.7853377,
-1.06421259, -0.83526512],
[-1.73150955, 0.497448813, -0.45913659, 2.1439539,
1.31952015, 3.30407145]])

This is a sample of standardized data after persuading it to the data frame. Consolidated standardized numeric data and label encoded categorical data by using the merging technique and then cleave data into train and test. There is one more technique that OLS which is called an Ordinary least square method. It will give us some data which is used to estimate the model which helps to minimize the sum of squared errors between observed data and predict data. This is a statistical approach of analysis that estimates the relation of one or more independent variables. So far, by using various techniques the above data was obtained.

After applying algorithm to model the obtained results are:

XGBClassifier:

83.76

XGBClassifier Test Score:

81.68

Confusion matrix:

6524 60
1406 10

TABLE10: Classification Report

	Precision	Recall	F1-Score	Support
0	0.82	0.99	0.90	6584
1	0.14	0.01	0.01	1416
Accuracy	--	--	0.82	8000
Macro Avg	0.48	0.50	0.46	8000
Weighted avg	0.70	0.82	0.74	8000

The below mentioned are OLS Results in TABLE11 and TABLE12.

Here in OLS, we have R^2 which is termed as R-Squared which is a statistical measure that represents the goodness of fit of the model.

R-Squared=(TSS-RSS)/TSS, TSS=Total sum of squares

RSS=Residual sum of squares.

P-Value: It helps us to determine how likely it is to get a particular result when the null hypothesis is assumed to be true. It is probability of getting a sample like more extreme than others if null hypothesis is correct.

TABLE11: OLS Results

Dep. Variable	**Target**	**R-Squared**	**0.002**
Model	**OLS**	Adj. R-squared	**0.002**
Method	**Least square**	F-Statistic	**15.11**
Date	**29 Aug 2020**	Prob(F-Statistic)	**2.78e-07**
Time	**19:51:41**	Log-Likelihood	**-8801.1**
No.of observations	**16000**	AIC	**1.761e+04**
Df.Residuals	**15998**	BIC	**1.762+04**
DfModel	**2**		
Covariance type	**Nonrobust**		
Omnibus	**4141.356**	Durbin-waston	**1.672**
Prob(Omnibus)	**0.000**	Jarque-bera(JB)	**8169.219**
Skew	**1.694**	Prob(JB)	**0.00**
Kurtosis	**3.884**	Cond.No	**491e+04**

TABLE12: OLS RESULTS

	Coef	Std err	t	P>|t|	[0.025	0.975]
X1	1.51e-08	4.92e-09	2.237	0.019	1.8e-09	2.12e-08
X2	-0.0012	0.00	-4.96	0.00	-0.002	

RSS: Difference between actual and predicted value by the model.

Log-Likelihood: Used derive maximum likelihood estimator of parameters.

F-Statistic: Ratio of quantities that are expected to be roughly equal under the null hypothesis.

The confusion matrix:

6524 60
1406 10

This is a confusion matrix that shows diagonal values about predictions. Here, the algorithm predicts 6534 correct values and 1466 wrong predictions. The report obtained from XGB Classifier which are containing classification reports, precision, and recall.

Merits: This research study focuses on stress management levels of employee using various types of techniques and when compared to others the data set considered in this is related to present pandemic and has lot of parameters about employees stress levels.

V. CONCLUSION

To evaluate our model to achieve a better performance which is done by using XGB classifier. This is one of the best optimization technique and this is like a decision tree-based algorithm which adopts gradient boosting frame work technique for analysis and confusion matrix which tells us how many correct values are predicted by our model. XG Boost has tremendous predictive power and is about 10 times more durable than other gradient boosting techniques. It holds a variety of regularization which diminishes overfitting and enhances overall performance. Consequently, it is further recognized as the "regularized boosting" technique. Like it has true positive, true negative, false positive, false negative values. Used to evaluate the performance of the classification model.

REFERENCES:

[1] Shekhar Pandey, Supriya Muthuraman, Abhilash Shrivastava.The International Symposium on Intelligent Systems Technologies and Applications (2018), DOI: 10.1007/978-3-319-68385-0_10.

[2] Ramachandran, R; Rajeev, D.C; Krishnan, S.G; Subathra.P. International Journal of Applied Engineering Research (2015), Research India Publications, Volume 10,Number 10, p.25433-25448

[3] Ramin Zibaseresht: How to Respond to the Ongoing Pandemic Outbreak of the Coronavirus Disease (COVID-19) (WHO- World Health Organization) (2020), ISSN 2349- 8870.

[4] Chen, Tianqi; Guestrin, Carlos; "XG Boost: A scalable Tree Boosting System". Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, USA (2016). ACM. pp. 785-794.

[5] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002,XXIX,487p.28 illus. ISBN 978-0-387-95442-4.

[6] Manuela Aparicio and Carlos J. Costa "Data Visualization". Communication Design Quarterly (2014). DOI: 10.1145/2721882.2721883.

[7] Ali.M., Alqahtani.A., Jones.M.W., Xie.X ."Clustering and classification for Time Series Data in Visual Analytics: A Survey IEEE Access 7,8930535, pp. 181314-181338.

[8] Mouubayedd.A, Injadat.M., Nassif, Lutfuyya, H.Shami, A E-learning: Challenges and Research Opportunities Using Machine Learning and Data Analytics (2018) IEEE Access 6,8417405. pp. 39117-39138.

[9] Kim., Soyata, T., Behnagh, R.F. Towards Emotionally Aware AI Smart Classroom: Current Issues and Directions for Engineering and Education (2018) IEEE pp. 5308-5331.

[10] Priyonesi S.Madeh; EI-Diraby Tamer E. "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". Journal of Transprotation Engineering, Pavements (2020). DOI: 10.1061/JPEODX.0000175.

[11] Abhijeet Rawal, Sneha Mhatre. IOSR Journal of Bussiness and Management (IOSR-JBM), e-ISSN: 2278-487X, P-ISSN: 2319-7668, PP 15-23.

[12] Janne Skakon, Karina Nielsen, Vihelm Borg, Jaime Gazman. An international Journal of work, Health and organizations, Volume 24,2010-Issue 2.

[13] K. S. Santosh and S. H. Bharathi, "Non-negative matrix factorization algorithms for blind source sepertion in speech recognition," 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, 2017, pp. 2242-2246, doi: 10.1109/RTEICT.2017.8256999.

[14] Rajendra Prasad P, N. Narayan, S. Gayathri and S. Ganna, "An Efficient E-Health Monitoring with Smart Dispensing System for Remote Areas", 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 2018, pp. 2120-2124. doi: 10.1109/RTEICT42901.2018.9012480.