**FINAL REPORT**

**STRESS PREDICTION**

**Team: Anwesha Sinha, Gandharva Dua, Armando Miranda**

**BUSINESS UNDERSTANDING**

As identified by our client, there has been a decrease in performance at the different operational levels in their company. They have been struggling to find the reason for this as understanding such an issue would require the analysis of an array of factors that come into play. Throughout eliminatory analysis, they ruled out any macro-operational or organizational flaws and have come to the early conclusion that such performance decrease could be due to a more personal driver. As such, they have hired us to understand whether stress, one of the suspected culprits, is a key factor of the low productivity.

There are many factors that could hinder operational performance in companies. Such factors may include things such as career realization, personal circumstances, organizational culture, and structure, pay, and more broadly, stress. It is well known that stress is a key factor that could negatively impact an individual’s performance at the workplace as suggested by (Yaseer Saied, 2009, p.470).

In this regard, we will be making use of secondary data from a survey performed to predict depression, anxiety, and stress (DAS). Since our task is to look for drivers for stress specifically, we have ruled out depression and anxiety metrics. By using this data, we analyzed and identified the key factors that drive stress and ran a series of models to test our model’s ability to predict stress. Our client expects us to receive a finalized model that could allow him to predict the likelihood of any of his workers to experience what is defined as “extremely severe” stress levels. This, with the goal of addressing the issues faced by such workers and only assigning to key projects those workers who are deemed to be within normal levels of stress. This with the goal of making their processes more effective by allocating the right resources at the right time.

**DATA UNDERSTANDING**

Understanding that we are using secondary data for this endeavor, this brings multiple issues for which we need to settle and/or try to improve. As data analysts, we must know how to work around such issues and that is why we shaped the data in the way we saw fit. This included getting rid of unnecessary variables, renaming variables, creating new columns with custom values, among other things. Variables that we decided to keep in our model are those which proved to have significant correlation to stress prediction. These include marital status, education level, gender, age, among others.

It is worth noting that in our data, we have a column that defines the stress levels in an ordinal manner ranging from normal, mild, moderate, severe, and extremely severe. We are particularly interested in targeting the cases that showed extremely severe stress levels as this would help us predict and hopefully avoid this situation with new data. Data showed to be organized and only required us to perform standardization of features before running our models to ensure all features had equal chance to be relevant in our models.

**DATA PREPARATION**

During our data cleaning process, we assumed that those respondents who did not respond to the degree question had no degree and thus, imputed the “No degree” value for those answers. Concerning the Ten Item Personality Inventory (TIPI) we named each of the columns from TIPI1, TIPI2, and so on to what those values meant for ease of analysis and for clearer visualizations. For the different ages of the respondents, we created a new column with age ranges that would allow us to better segment our sample. Similarly, we created another column grouping the different stress conditions using the score from the original column “Total\_Count”. Lastly, for the family size column, we limited the family size to 9, putting all those families of 9 and above in the same group understanding that most of those families larger than 9 were outliers.

**EXPLORATORY DATA ANALYSIS**

In this phase we performed exploratory data analysis on variables that we presume to be important for describing stress levels. We created a correlation matrix between the dependent and independent variables to check their collinearity. Additionally, we created charts to visualize the distribution of values. This helped us set the base for understanding which values we should take into consideration when performing further analysis such as random forest, decision tree and logistic regression trying to predict and explain stress levels.

Figure 1.1 Figure 1.2

A picture containing histogram

Description automatically generated Chart, bar chart

Description automatically generated

Here is a visualization of all the respondents of the survey and their stress levels

1. **Correlation Matrix:**

‘Married’ and ‘Age Groups’ have multicollinearity between them, and so do ‘Education’ and ‘Age Groups’, hence, the relation might not be considered. Also, it can be observed from both the plots and the matrix, that ‘Married’, ‘Education’ and ‘Age group’ have higher collinear relation with ‘Conditions’. This makes sense too, since a person in a certain age group might be too stressed, the reason being either marriage, or when going a level higher in their education or appearing for exams, etc. Marriage, and Education are the areas where the stress level of a person considerably increases, especially when they are in a particular age group.

Figure 1.3

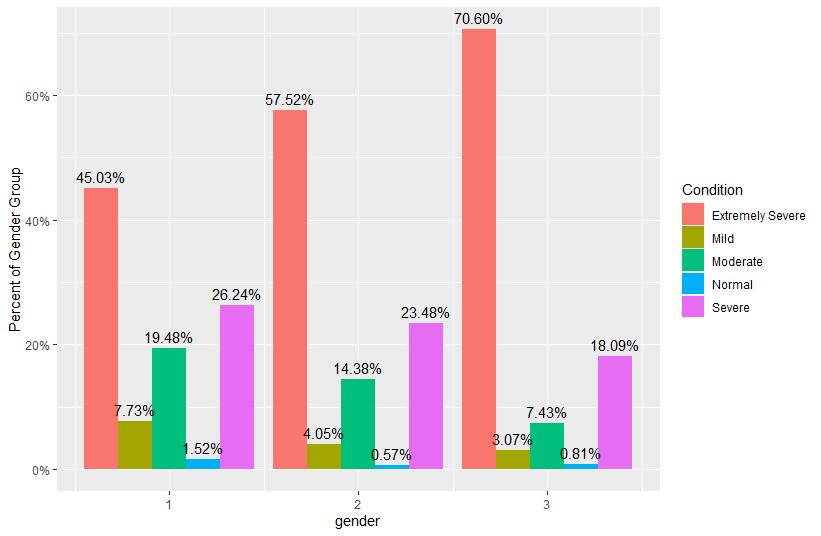
Chart, scatter chart, bubble chart

Description automatically generated

1. **Gender vs Condition:**

Here, gender ‘1’ corresponds to ‘Males’, gender ‘2’ to ‘Females’, and gender ‘3’ to ‘Others’. It can be interpreted from the visualized and textual results, that gender is a significant factor for stress prediction since the P value is less than 0.05 as per the Chi-square test performed. Contrary to what the end-user might believe, gender 3 does not have the highest relative extremely severe percentage, it is gender 2. This is explained by the very small sample size of gender 3 compared to gender 1 and gender 2.

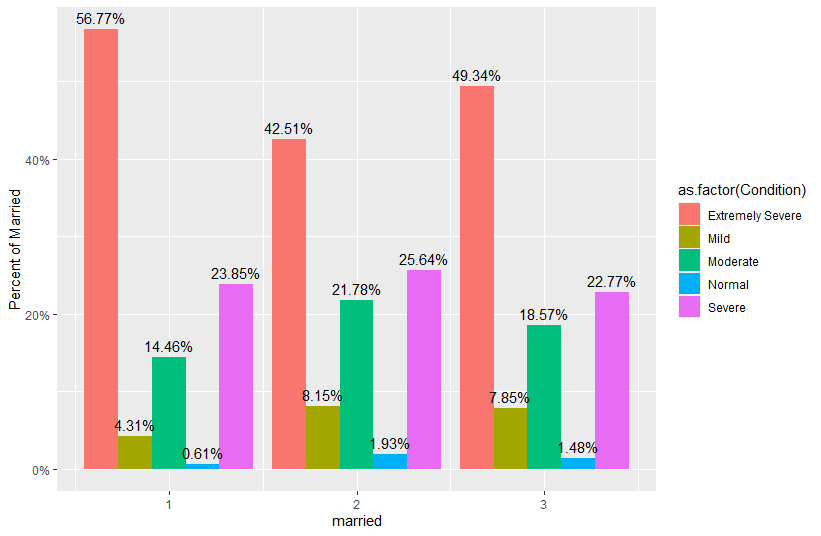
Figure 1.4



1. **Married vs Condition**

Here, ‘1’ has been considered as ‘Never Married’, ‘2’ as ‘Currently Married’ and ‘Previously Married’ has been assigned to ‘3’. According to the chart, individuals who never married face higher levels of extremely severe stress levels, but lower levels of other types of stress.  We have a chi-squared value of 536.01. Since the p-Value is less than the significance level of 0.05, we reject the null hypothesis and conclude that the two variables are in fact dependent, and that ‘Marriage’ is an important factor for predicting stress.

Figure 1.5

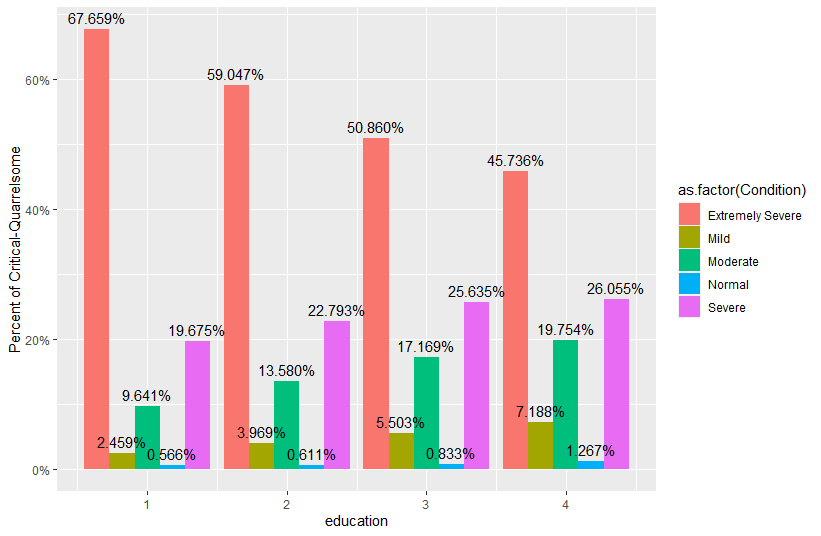


1. **Education vs Condition:**

Numbers have been assigned to the four levels of education that were specified in the survey. ‘1’ is to be interpreted as education less than high school, education till a high school degree is ‘2’, university degree is being assigned the number ‘3’, and a graduate degree is ‘4’.

As observed in this chart, there is a negative relationship between extremely severe stress and education levels. As education levels increase, the likelihood of experiencing extremely severe stress decreases. However, for the other kinds of stress levels, there is a positive relationship. For stress prediction, Education should be considered since it seems to be a significant factor according to the Chi-square test.

Figure 1.6



For thorough EDA, it can be referred to in the Appendix.

**MODELING**

**Random Forest:**

Random forest is a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree. Random Forest benefits in solving the problem of overfitting (model learns the training data by heart instead of learning the patterns which prevent it from being able to be generalized to the test data).

For Random Forest we did a split ratio of 80-20 training/test dataset. We were able to get an accuracy of 93.32% which is a higher percentage of our prediction. With this accuracy our business model would be able to predict if stress is a major factor in the decline of the company’s productivity.

Some pros and cons we figured while performing Random Forest are as follow:

Pros-

* It performs well on classification model (i.e., categorical target variable).
* There are very few assumptions related to it, so it makes data preparation less challenging and gives results which helps save time.

Cons-

* Random Forests are not very good at generalizing cases with completely new data.
* Random Forest can cost some loss of interpretation.

Error vs no. of tress plot of the model:

Figure 2.1

Chart, line chart

Description automatically generated

Minimal depth:

The distribution of the mean minimal depth allows us to appreciate the variable’s role in the random forest’s structure and prediction. The smaller the mean minimal depth, the more important the variable is and the higher up the y-axis the variable will be.

Plot for Minimal depth:

Figure 2.2

Chart, bar chart

Description automatically generated

Plot for Multi-way importance scores: Plot for Confusion Matrix:

Figure 2.3 Figure 2.4

Chart, scatter chart

Description automatically generated Chart

Description automatically generated

**Multinomial Logistic regression:**

The multinomial logistic regression algorithm is an extension to the logistic regression model that involves changing the loss function to cross-entropy loss and predict probability distribution to a multinomial probability distribution to natively support multi-class classification problems.

For multinomial logistic regression model, we did a split ratio of 80-20 training/test dataset. We were able to get an accuracy of 87.13%.

Some pros and cons we figured while performing Multinomial Logistic regression are as follow:

Pros:

* Logistic regression is easier to implement, interpret, and very efficient to train.
* It makes no assumptions about distribution of classes in feature space.

Cons:

* If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.
* The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

Plot for Confusion Matrix:

Figure 2.5

Chart

Description automatically generated

**Decision Trees:**

Decision Tree solves the problem of machine learning by transforming the data into a tree representation. Each internal node of the tree representation denotes an attribute, and each leaf node denotes a class label.

For Decision Tree we did a split ratio of 80-20 training/test dataset. We were able to get an accuracy of 84.63%.

Some pros and cons we figured while performing Decision Trees are as follow:

Pros:

* Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
* A decision tree does not require normalization of data.

Cons:

* A small change in the data can cause a large change in the structure of the decision tree causing instability.
* For a Decision tree sometimes, the calculation can go far more complex compared to other algorithms.

Tree plot

Figure 2.6

Timeline

Description automatically generated

Plot for Confusion Matrix:

Figure 2.7

Chart

Description automatically generated

**How does these models solve our business problem?**

The problem that the company is facing about the decline of its productivity could be due to a huge issue ‘stress’. Since our dataset is categorical, we analyzed that the models that will fit best in this situation are random forest, logistic regression and decision tree. We chose Random Forest as our best fit model as it gave us higher accuracy of 93.32% for predicting whether the employee’s condition falls into any of the five categories of stress levels. With the help of this model and the accuracy of our prediction, the company can take necessary measures to provide its employees with the required help they need which in return can boost the company’s productivity.

**EVALUATION**

An evaluation of the performance of our model in allowing management to assign team members to any given project would be easily measurable. This could be done by obtaining a Key Performance Indicator (KPI) that would tell us the success of ‘x’ or ‘y’ project. We could treat project ‘x’ as a treatment group and project ‘y’ as a control group. If the performance of the treatment group (i.e., those who were assigned to the project using our model) is significantly higher, then that would be an indicator that our model is indeed working. Moreover, the company could evaluate the success of our endeavor by measuring their return on investment (ROI) and compare it to previous cycles where the stress prediction was not used when assigning employees to any given task.

Although, not a measurable metric, the company would be perceived as a more socially aware organization that care for its employees. This in turn, would increase its current and prospective clients’ trust in the company brand as suggested by (Enio Marcos Babireski Barcelos et al, 2015, p.434). It is not easy to measure something as subjective as a customers’ perception of a company, but our client could measure other factors that could be influenced by this such as churn rate and acquisition of new customers. In the same lines, the company’s employees will be happier if their psychological state is valued and taken into consideration when assigning new tasks. This might prove to increase employee retention rates which is detrimental to a company’s success and should be seen as a long-term investment (Journal of Leadership, 2015, p.125). All in all, we expect our model to allow our client to evaluate the benefits in multiple fronts as it will have effects throughout the entire organization.

**Business Case handled with the help of prediction models-**

In our dataset we analyzed that the worst case of error we can come across is Type 2 Error i.e., False Negative.

We found sensitivity is a major metric of the confusion matrix. Sensitivity explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where False Negative is of higher concern than False Positive. It is important in medical cases where it doesn’t matter whether we raise a false alarm, but the actual positive cases should not go undetected.

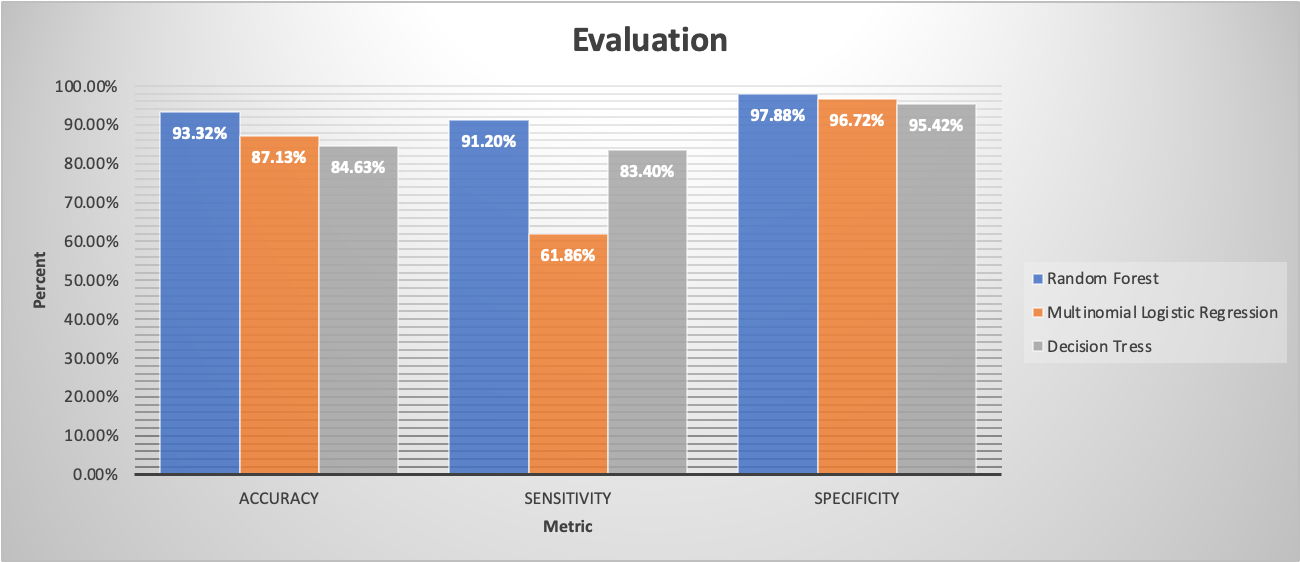
In our business problem, if an employee who is suffering from stress but remains undetected for it then it is considered as a False Negative. If this issue arises and the employee experiencing a severe stress level is assigned to a given task/project could lead to catastrophic consequences for the company. These consequences could go from damaged client relations due to irritable character, security breach by the employee not being attentive to its environment, or simply sub-par performance causing economic losses. Moreover, the company won’t be able to benefit from doing a survey of its employees’ mental health as the return on investment will never see positive figures if this issue is persistent. Based on the models we ran; we were able to predict the person’s sensitivity percentage.

The appendix could be referred to for the model statistics of every model used.

Table 1.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | Preferred |  |
|  | Model | Accuracy | Sensitivity | Specificity |
| Top Model | Random Forest | 93.32% | 91.20% | 97.88% |
|  | Multinomial Logistic Regression | 87.13% | 61.86% | 96.72% |
|  | Decision Tress | 84.63% | 83.40% | 95.42% |

Figure 3.1



**ROI-**

By definition, ROI is a type of business evaluation that calculates the monetary returns from a financial investment. In this case, the venture of choice is wellness benefits in the workplace. Usually, by implementing such programs, companies typically see a return on investment in health care cost savings and increased productivity due to reduced employee absenteeism. In other words, supporting workers’ mental and physical well-being is a smart business investment.

Stress has negative consequences for individuals but also for your organization. The more stress, the less productive your employees' engagement is. Of a total of 39775 employees surveyed over a time period of 2017-2019, 54.96% experienced severe stress problems which means more than half of the workplace dealt with stress which in turn caused decline of the company’s productivity.

**How is ROI achieved by this prediction?**

ROI in this case is conservative and does not consider the many other benefits that investment in workplace mental health programs can generate.

Higher retention rates-

Programs that support employees along the continuum of care, from mental health promotion to treatment, can decrease voluntary turnover related to mental health issues by supporting employees in need, so that they see alternatives to leaving the organization. Decreased turnover, in turn, reduces costs related to hiring and training new employees.

The methods of converting data involved a variety of approaches, including tabulating direct costs, using standard values, using external data, and securing estimates from a variety of target audiences. The cost categories represented fully loaded costs for the program. Expected intangible benefits from the program were based on the experience of other organizations and other stress reduction programs. The communication target audience consisted of six key groups ranging from corporate and business unit managers to participants and their immediate supervisors.

Table 1.2

Table

Description automatically generated

**Why invest in Mental Health?**

Cost and Cost-Effectiveness-

* There is a substantial return on investment in mental health.
* Every US$ 1 invested in scaling up treatment for depression and anxiety leads to a return of US$ 4 in better health and ability to work.

Equitable Access and Financial Protection-

* Investment improves equitable access and fairness in financial contribution to essential mental health services.
* This helps us to move closer to Universal Health Coverage

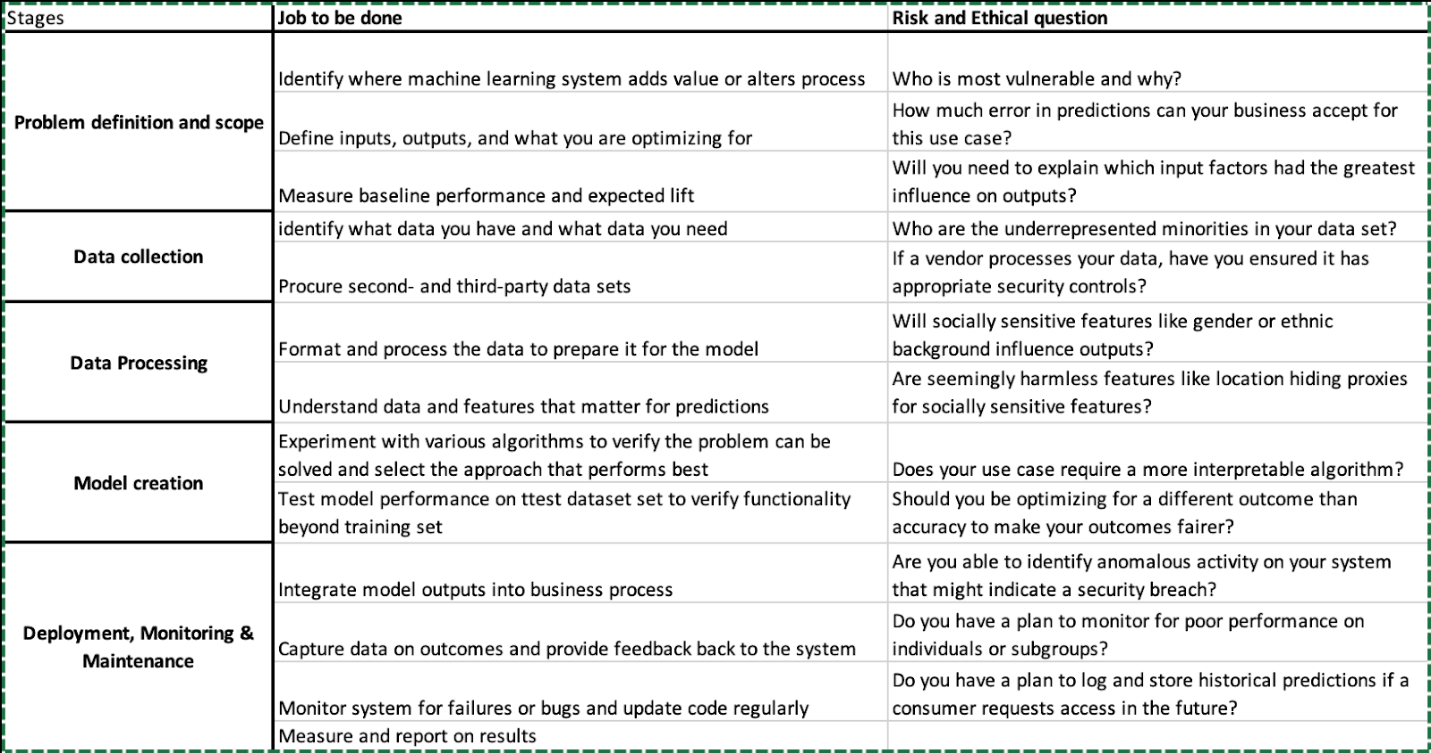
**DEPLOYMENT**

The deployment of this new management-aiding tool should be performed in a slow and controlled manner. Trainings should be provided to management and executives first as they are the ones who will be employing this stress predicting tool when making decisions. Further training and exposure to this new tool could be provided to the rest of employees so that they are aware of its implementation. During this stakeholder training, it should also be mentioned that the metrics that the model takes into consideration such as family size, marital status, housing location among others should be periodically updated to that the model can be adjusted to employee’s ever-changing conditions and predict accordingly.

The data used to feed to the model is sensitive data that the company will have to be obtaining from their employees regularly in order to maintain it relevant to the current circumstances. Baring this in mind, the company should ensure the data is kept safe and encoded as to ensure their employees identity concealed. With this, the company should ensure a proper data governance structure is in place.

A few important ethical considerations need to be made during each stage of machine learning. A table is provided below which shows the stages and some ethical questions that should be answered so that it does not lead to a complete loss of trust, which can have a cataclysmic impact on any business

Table 1.3



**APPENDIX**

Below given are the detailed analysis for our other factors which were not included in the main report due to space constraints.

1. **Age Groups vs Condition:**

**Text

Description automatically generated with medium confidenceText

Description automatically generated**

Age groups do play a significant role in predicting stress levels in a human being. It is proven generally, as well as per the Chi-squared test performed. According to the chart, in the group ‘Adults’, more than half of the group were found to have extremely sever stress levels. That goes down to 42% in Elder Adults, and in the group of Older People, almost 33% have found to have extremely sever stress levels. Overall, every age group seems to have less than 2% of normal stress levels and are extremely affected by stress at high levels.

**Chart, bar chart

Description automatically generated**

1. **English As a Native Tongue vs Condition:**

**Text

Description automatically generated with low confidence Text

Description automatically generated**

In general, a person who doesn’t have English as a native language might go through stress when living/working in an English-speaking country/workplace. This is shown through the textual and graphical results here. According to our analysis, this factor does play an important role in predicting stress, and that, in the group of people not having English as a native language, more than half of the surveyed people go through extreme stress levels. For both the groups, a normal condition is extremely rare, since the percentage showing for that is no more than 1%.

**Chart, bar chart

Description automatically generated**

1. **Area of Living vs Condition:**

**A picture containing graphical user interface

Description automatically generated Text

Description automatically generated**

Stress conditions might vary according to living conditions. People in the survey were asked whether they come from 1=Rural Areas (Villages), 2=Suburban Areas (Countryside), or 3=Urban Areas (Towns/Cities). Living through the extreme rush in huge urban cities might prove to be extremely stressful for an individual. Our analysis shows that too. Since the p-value according to the Chi-Square test is less than 0.05, it is considered that stress levels can be predicted using the area an individual hails from. In the group of Urban people, stress levels are found to be extremely severe in almost 60% of the population. The same is lesser for people belonging to the group of individuals coming from either rural or suburban areas.

**Chart, bar chart

Description automatically generated**

1. **Family Size vs Condition:**

Text

Description automatically generated with medium confidence Text

Description automatically generated

In this case, the values for family sizes ranged from 1 to more than 100. Thus, for a concise analysis, the numbers were narrowed down, and the final range was from 1 to 9, where family sizes more than 10 were considered in 9 itself. As per our analysis, in almost every group, the percentage of people having extreme stress levels is quite high, followed by severely, and moderately stressed. As per the chi-squared test, family size plays a significant role in determining stress.

Chart, bar chart

Description automatically generated

Below given are a few other factors which we thought can contribute to an individual’s stress levels. The questions had responses divided into the given sections:

1-Did not apply at all

2- Applied to some degree, sometimes

3- Applied to a considerable degree, a good part of the time

4- Applied very much, most of the time

**Getting Upset vs Condition:**

**Text

Description automatically generated with medium confidence Text

Description automatically generated with medium confidence**

This was one of the first questions in the survey. It is quite certain, that generally, as well as per our analysis; getting upset at trivial things can contribute to the stress levels of an individual. The responses were divided into numbers 1 to 4 as per the indicator given above. It can be observed that among people in group 4, who get upset at trivial things almost all the time, more than 95% of the population go through extremely severe stress levels, whereas normal stress condition is almost non-existent. In group 3, the percentage of people having extremely severe stress levels is considerably lower than group 4, and Group 1 seems to have the lowest of them all. In group 1, the percentage of people who are moderately stressed is higher than the others.

**Chart, bar chart

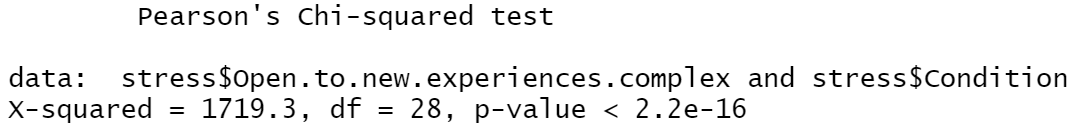
Description automatically generated**

1. **Open to New Experiences vs Condition:**

**Chart, bar chart

Description automatically generated**

Table

Description automatically generated ****

This condition refers to how the respondent feels about being exposed to new experiences. One (1) is the lowest score meaning the respondent does not identify with this trait and eight (8) is the highest score meaning the respondent very much identifies with this trait. Looking at the chi-squared test, we can see that its p-value with regard to their stress condition is significant.

As seen in the chart displayed above, there is an inverse relationship between being open and the likelihood of experiencing severe stress periods.

1. **Over-Reacting vs Condition:**

**Text

Description automatically generated** Text

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated

One of the questions in the survey was whether people tend to overreact to situations. The responses were recorded according to the indicator given above. This is a significant factor when measuring stress levels as per our analysis. People who tend to overreact all the time (group 4) have the highest number of individuals who are extremely stressed, with the percentage going up to 94.5%. It gradually reduces through group 3,2 and 1. So, it can be deduced that people who almost never overreact tend to have normal stress levels, and less levels of extreme stress compared to the other groups.

Along with those questions, the survey incorporated other factors called the Ten Item Personality Inventory (TIPI). The responses for these ten questions were divided into 7 responses:

1-Disagree Strongly

2-Disagree Moderately

3-Disagree a Little

4-Neither Agree nor Disagree

5-Agree a little

6-Agree Moderately

7-Agree Strongly

1. **Critical and Quarrelsome vs Condition:**

Text

Description automatically generated

**Text

Description automatically generated**

Chart, bar chart

Description automatically generated

One of the TIPI questions was about a person being Critical and Quarrelsome. It’s p-value was lower than 0.05, thus making it a significant factor to predict stress levels. As per the chart, among the people who

1. **Disorganized and Careless vs Condition:**

**Text

Description automatically generated**

**Text

Description automatically generated**

Chart, bar chart

Description automatically generated

1. **Dependable and Self-Disciplined vs Condition:**

Text

Description automatically generated

**Text

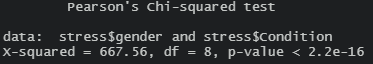
Description automatically generated**

Chart, bar chart

Description automatically generated

1. **Gender vs Condition:**

Figure 1.4 Figure 1.5

 Text

Description automatically generated with low confidence

1. **Married vs Condition**

Figure 1.7 Figure 1.8

A picture containing text

Description automatically generatedText

Description automatically generated

1. **Education vs Condition**

Figure 2.1 Figure 2.2

**Text

Description automatically generated** **Text

Description automatically generated**

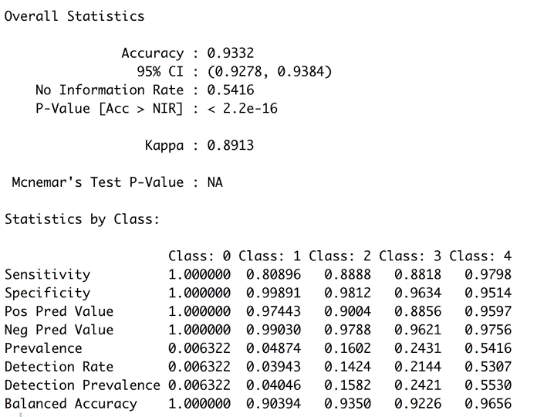
Output using Training data for Random Forest:

Table

Description automatically generated with low confidence

Random Forest Statistics:

Figure 4.1



Multinomial Logistic regression Statistics:

Figure 4.2

Table

Description automatically generated

Decision Tree Statistics:

Figure 4.3

Table

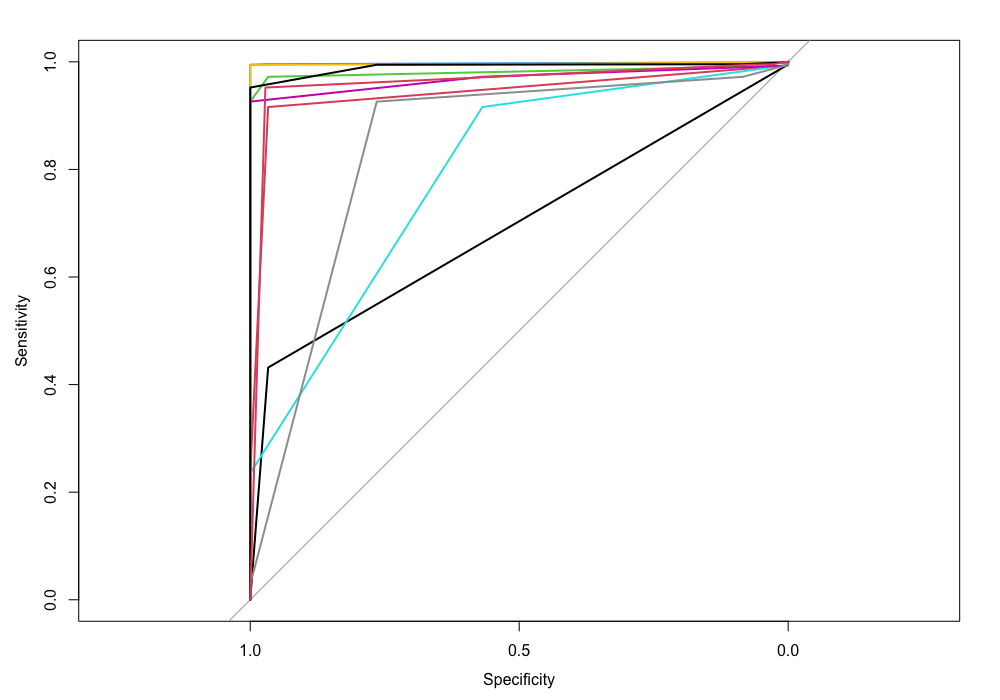
Description automatically generated

Plot for ROC and AUC: Random Forest

Chart

Description automatically generated

Plot for ROC and AUC: logistic regression



**SOURCES**

SaiEd, Y. S. (2014, May 22). *Impact of stress on employee productivity, performance and turnover; an important managerial issue*. Academia.edu. Retrieved December 1, 2022, from https://www.academia.edu/2400475/Impact\_Of\_Stress\_On\_Employee\_Productivity\_Performance\_And\_Turnover\_An\_Important\_Managerial\_Issue

Babireski Barcelos, E. M., Baptista, P. de P., Maffezzolli, E. C. F., Silva, W. V., Marchetti, R. Z., & da Veiga, C. P. (2015, April). *(PDF) relationship between an organization evaluated as being socially ...* ResearchGate. Retrieved December 1, 2022, from https://www.researchgate.net/publication/277132360\_Relationship\_Between\_an\_Organization\_Evaluated\_as\_Being\_Socially\_Responsible\_and\_the\_Satisfaction\_Trust\_and\_Loyalty\_of\_its\_Clients

Cloutier, O., Felusiak, L., Hill, C., & Pemberton-Jones, E. J. (2015). *The importance of developing strategies for employee retention*. na-businesspress. Retrieved December 1, 2022, from http://www.m.www.na-businesspress.com/JLAE/Pemberton-JonesEJ\_Web12\_2\_.pdf

Chiappelli, J., Kvarta, M., Bruce, H., Chen, S., Kochunov, P., & Hong, L. E. (2021). Stressful life events and openness to experience: Relevance to depression. *Journal of affective disorders*, *295*, 711–716. <https://doi.org/10.1016/j.jad.2021.08.112>

Bhandari, P. (2022, November 11). *Type I & type II errors: Differences, examples, visualizations*. Scribbr. Retrieved December 7, 2022, from <https://www.scribbr.com/statistics/type-i-and-type-ii-errors/>

Integrate.ai. (2018, October 18). *How to incorporate Ethics and risk  into your machine learning development process*. Medium. Retrieved December 7, 2022, from <https://medium.com/the-official-integrate-ai-blog/how-to-incorporate-ethics-and-risk-into-your-machine-learning-development-process-4b8e9bc78ce0>

Abrams, Z. (2022, June 1). *High stress levels during pandemic are making even everyday choices difficult*. Monitor on Psychology. Retrieved December 4, 2022, from <https://www.apa.org/monitor/2022/06/news-pandemic-stress-decision-making#:~:text=About%20half%20of%20U.S.%20adults,wear%20or%20what%20to%20eat>.

*Take a personality test - open source psychometrics project*. Take a personality test - Open Source Psychometrics Project. (n.d.). Retrieved December 8, 2022, from https://openpsychometrics.org/