Impact of Covid-19 on Working Professionals

IST 687 : Intro to Data Science : Final Project

Team 3: Matthew Woods, Avi Kazassi, Dan Myers, Sandra Suresh, Ty Hall

Purpose

- Covid-19 was an unprecedented event that forced most of the world to shut down and shelter in place.
- This led to many jobs utilizing work from home technologies in order to maintain employment and create a semblance of normality.
- In this project we wanted to understand the various impacts that Covid-19 had on different sectors in the work force and used a simulated data set from Kaggle to do this.

Goal

- Business related questions regarding reported work changes due to the COVID-19 pandemic. Specifically :
  - **Did any sectors have to work significantly more hours due to COVID-19**
  - **Did changes in working hours due to the pandemic affect stress levels?**
  - **Do changes in productivity due to the pandemic affect stress levels?**
  - **Could the features presented here be used to predict an employee’s stress level accurately?**

Why is this important?

Real World Applicability

Identify stressed indivimitigate the effects of chronic stress and prevent burnout.

Impact

- Increased productivity
- Increased retention rates
- Increased job satisfaction and engagement of employees
- Improved mental health of employees.

Understanding an employees stress levels is the first step in unleashing creativity, enabling potential and supporting sustainable productivity!

ABOUT THE DATASET

- The data set we used was found on Kaggle and linked below.
- *https://www.kaggle.com/datasets/gcreatives/impact-of-covid-19-on-working-professionals*
- The data is not a real-world data set and has been simulated to contain noise to mimic data that we would find in real life.

Step 1 : Data Modification & Cleaning

Quick Exploration of the data

- In the column "Meetings Per Day", some of the data points were negative values, so to fix this, those points were changed to 0
- 2 continuous variables
- 2 variables with more than two classes
- Remaining variables binary

1A. Descriptive Statistics of the continuous variables

	Avg Hours Per Worked per Day	Avg Number of Meetings Per Day
Mean	8.007	3.98
Median	8.011	3.98
Range	0.29-16.9	0-9.59
Standard Deviation	1.978	1.515
Skewness	0.014	0.036
Kurtosis	2.968	2.88

1B. Histograms of the Continuous Variables

Then we plotted the continuous variables graphically so that we can see whether or not they are normally distributed.

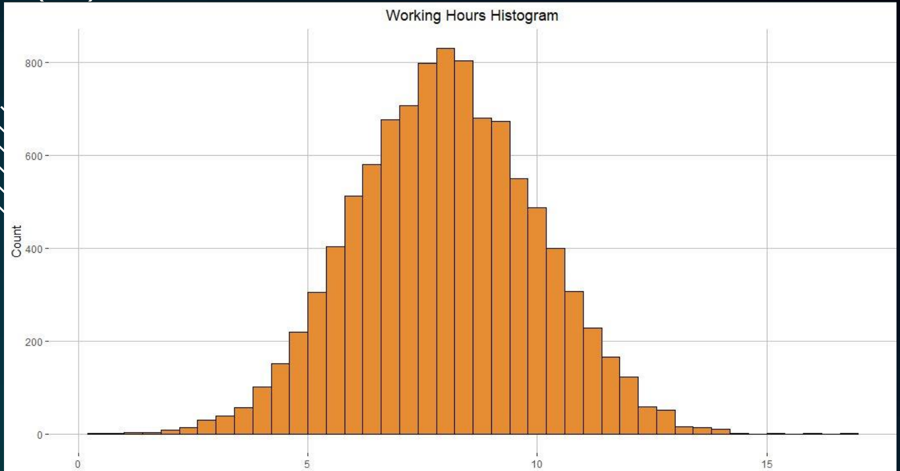


Figure 1. Histogram Average Hours worked per Day

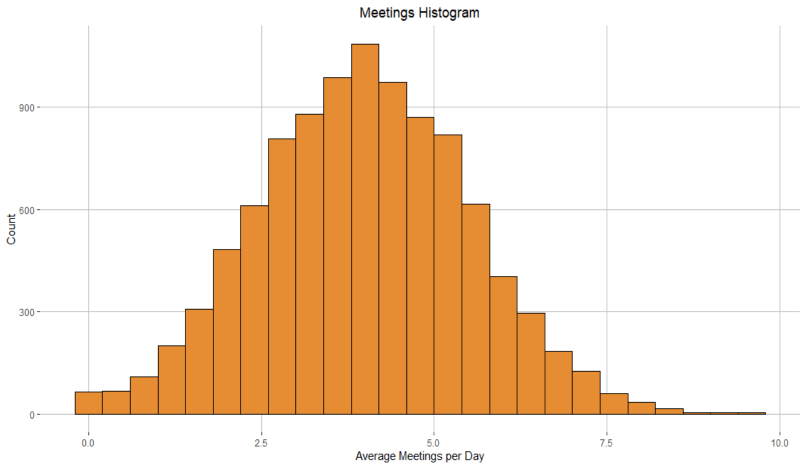
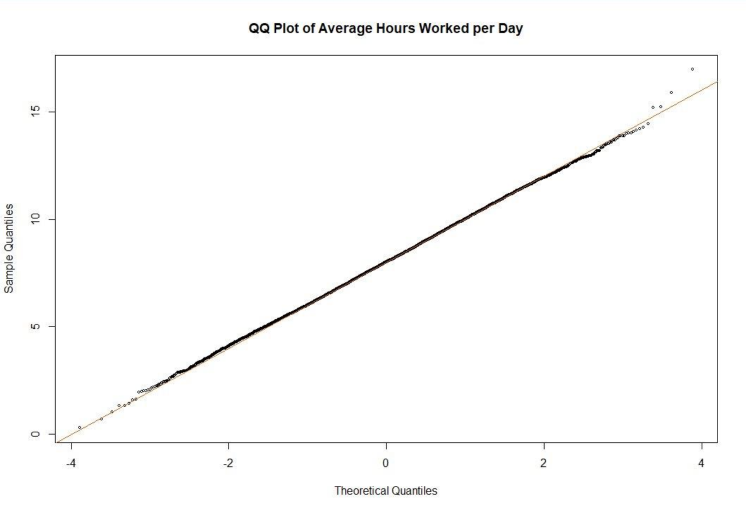


FIgure 2. HISTOGRAM AVERAGE MEETINGS PER DAY

1C. QQ Plots of the Continuous Variables

Next we ran QQ Plots of the continuous variables in order to check that they are normally distributed.



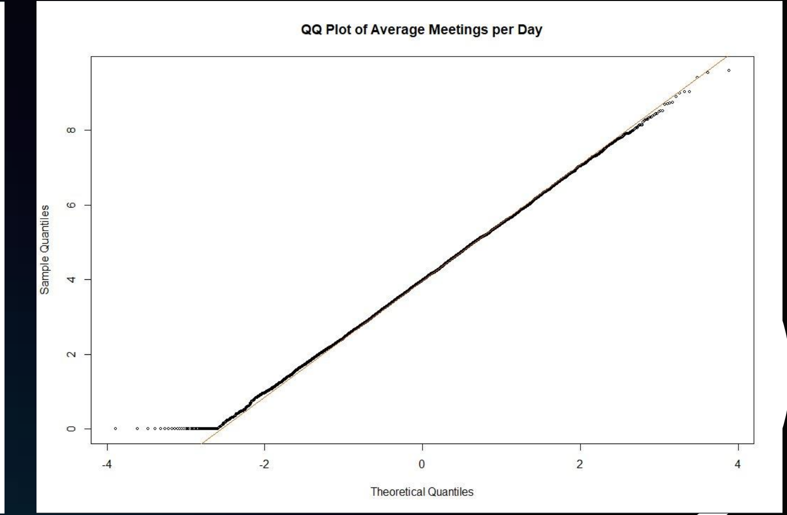


Figure 1. QQ Plot of Average Hours Worked per day

Figure 2. QQ Plot of Average Meetings per day

<u>Step 2 </u>: Feature Engineering

Problem
Many features binary when they could contain more information

Example
Productivity_Change – response 1 for change, 0 for no change

<u>Solution</u>: Engineer/Impute some variables

Feature engineering helps us transform a dataset and make it more suitable for a data scientists needs.

Step 2 : Feature Engineering (Continued..)

Published works

Awada M. Becerik Gerber B, L. G. (n.d.). Stress appraisal in the workbplace and its associations with productivity and mood: Insights from a multimodal machine learning analysis. *PLoS ONE*. doi:https://doi.org/10.1371/journal.pone.0296468

Pencavel, J. (2018). Diminishing Returns at Work: The consequences of Long Working Hours. *Oxford Academic, Online edition*. doi:https://doi.org/10.1093/oso/9780190876166.001.0001

Yerkes, R. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *J. Comp. Neorol. Psychol., 18*, 459-182. doi:https://doi.org/10.1002/cne.920180503

Impute Productivity_Change_Direction
(1, 0, -1)

Sampled BLS 2020 salary data where productivity increased and salary changed
Sampled BLS 2019 salary data everyewhere else

Log Normal distribution

2A. Histogram of simulated salaries

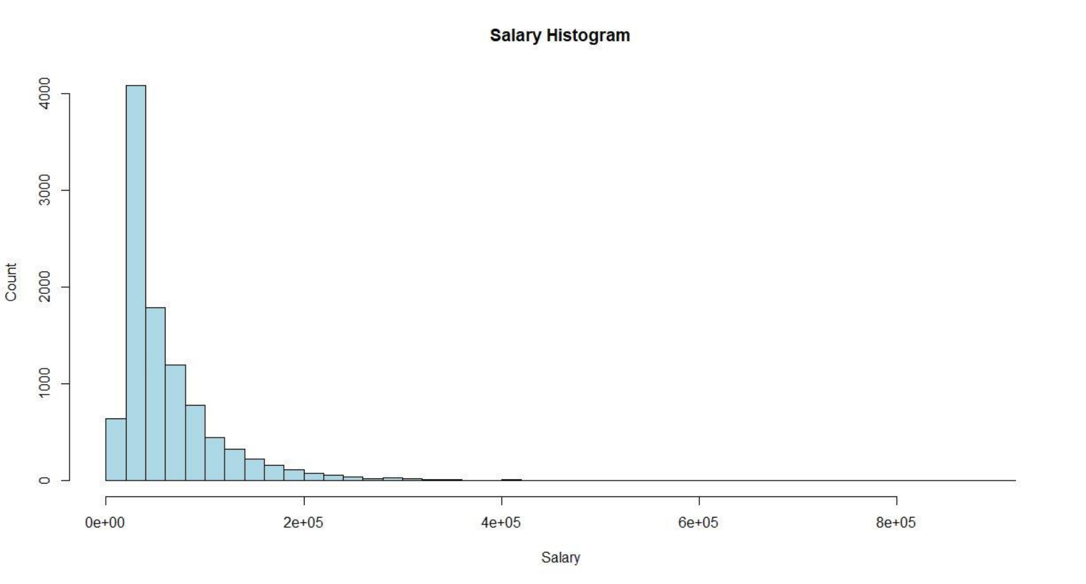


Figure : Histogram of simulated salaries showing a log normal distribution.

Look at that! Our simulated data was a success!

Step 3 : Exploratory Data Analysis

After cleansing and modifying the data, we can begin to understand some of the relationships between the variables by using Exploratory Data Analysis (EDA).

A. Hours Worked By Sector

Did any sectors work significantly more hours?

B. Stress Levels By Increased Working Hours

Were stress levels affected by increased working hours?

C. Productivity by stress levels

Were stress levels affected by changes in productivity?

st

3A. Hours Worked By Sector

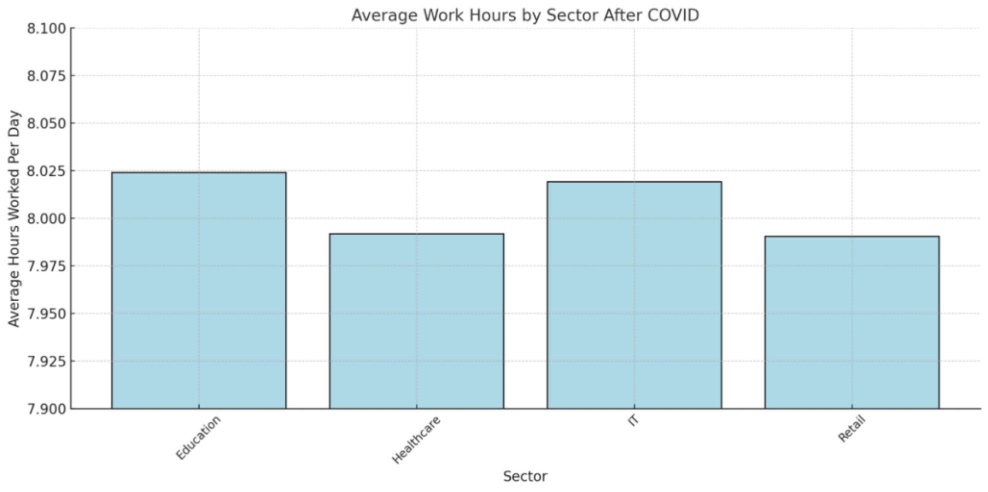


Figure 3A. Bar Plot of Average Working Hours by Sector

Results

Varying levels of average working hours

Analysis of variance $p = 0.897$.

3B. Stress Levels By Increased Working Hours

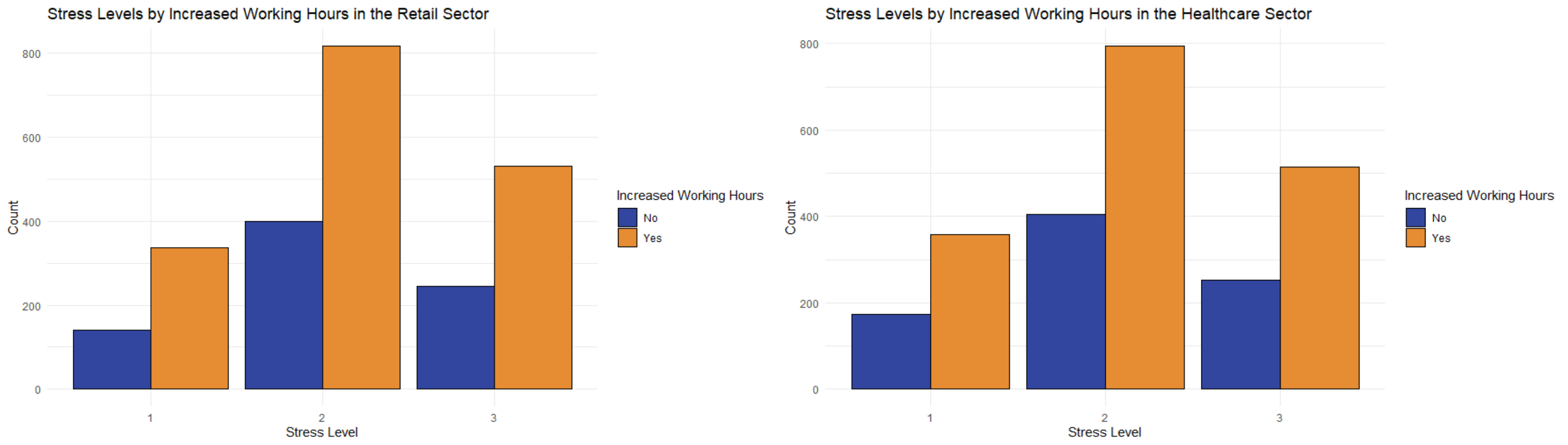


Figure 3B. Stress Levels by Increased Working Hours in each Sector (Retail & Healthcare)

3B. Stress Levels By Increased Working Hours (Cont...)

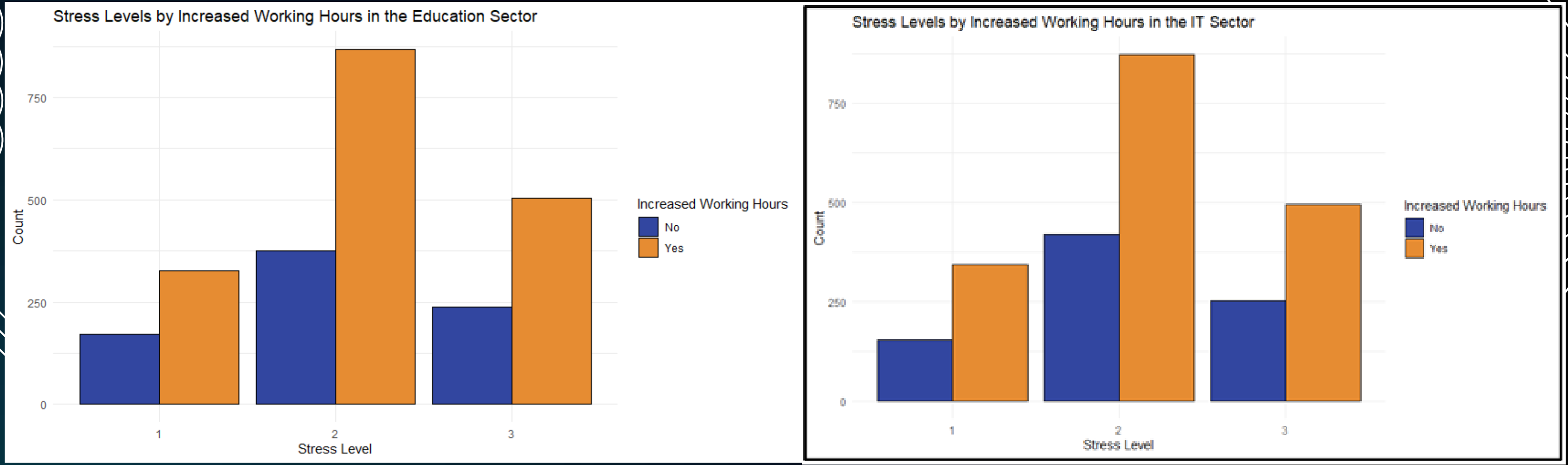


Figure 3B. Stress Levels by Increased Working Hours in each Sector (Education & IT)

3C. Stress Levels By Productivity Change

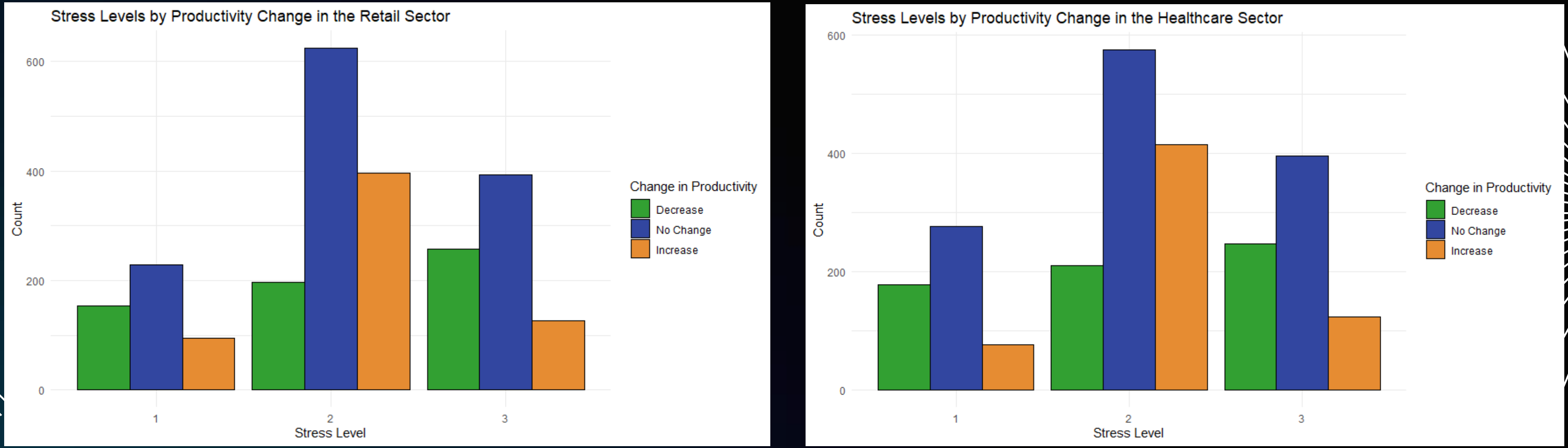
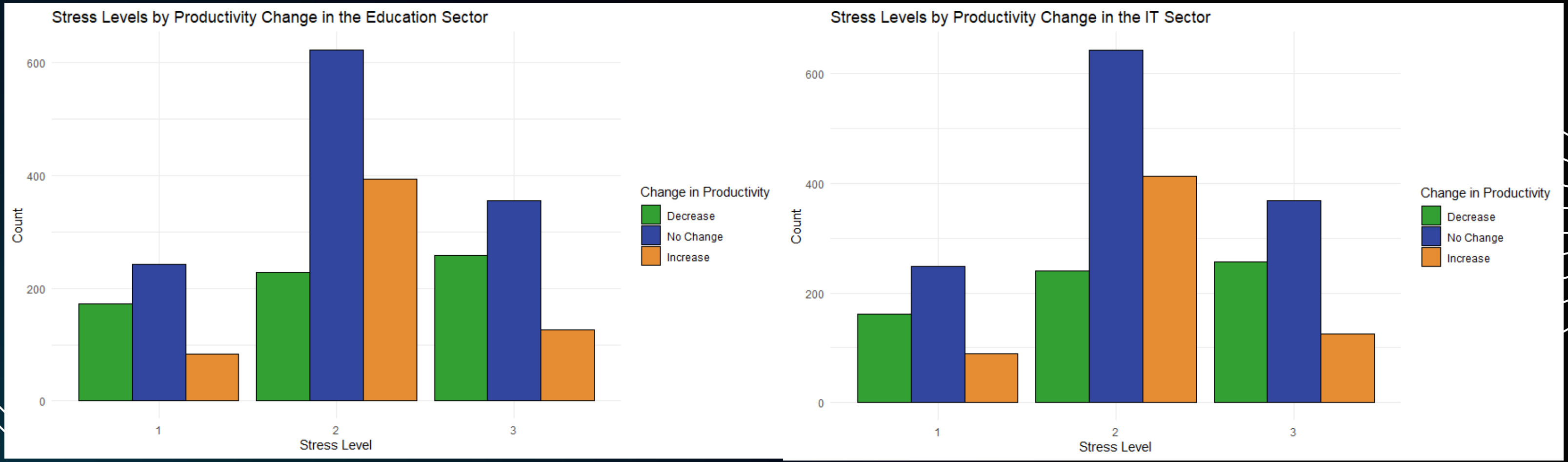


Figure 3C. Stress levels by productivity change direction in each sector (Retail & Healthcare)

3C. Stress Levels By Productivity Change (Cont..)


Stress Levels by Productivity Change in the Education Sector

Stress Levels by Productivity Change in the IT Sector

Figure 3C. Stress levels by productivity change direction in each sector (Education & IT)

Step 4: Predicting Stress Levels

Is it feasible to use the features presented in this data set to predict the stress levels of an employee?

SVM Model

- Good with highl dimensional data
- Used all features in the data including engineered features except
  - Affected by Covid
- 66% Train - 34% Test split

Random Forest Model

- Feature importances
- Handful of features
- Robust model insensitive to overfitting
- 70% Train - 30% Test split

4A. SVM Model

Results

- Accuracy: 58%
- F1 Score: 0.61

- F1 Scores for Low Stress Level: 0.265
- F1 Scores for Medium Stress Level: 0.673
- F1 Scores for High Stress Level: 0.550

- Suggests imbalance – bias toward medium stress level class

	High	Low	Medium
High	622	90	335
Low	37	139	69
Medium	555	574	1578

Figure 4A. Confusion matrix for SVM prediction of stress level.

4B. Random Forest Model

Results

- Feature importance estimation
- Hours_Worked_Per_Day
- Meetings_Per_Day
- Technology_Adaptation
- Increased_Work_Hours
- Job_Security
- Sector

- Accuracy: 52%.
- F1 Score: 0.66
- F1 Scores for Low Stress : 0.045
- F1 Scores for Medium Stress 0.671
- F1 Scores for High Stress : 0.082

- Suggests imbalance – bias toward medium stress level class

	High	Low	Medium
High	91	0	0
Low	0	32	0
Medium	2035	1374	3470

Figure 4B. Confusion Matrix for Random Forest Prediction of Stress Level

Step 5: Interpretation

Now lets go back to our **Goal** and analyze the results we obtained from our analysis.

A. Business Questions

1. **Did any sectors have to work significantly more hours due to COVID-19?**
2. **Did changes in working hours due to the pandemic affect stress levels?**
3. **Do changes in productivity due to the pandemic affect stress levels?**
4. **Could the features presented here be used to predict an employee's stress level accurately?**

B. Interpretation

1. Bar chart and ANOVA indicate no evidence of significant differences
2. Side-by-side bar charts show some correlation but no causation
3. Side-by-side bar charts show some correlation but no causation
4. Models didn't prove accurate enough for predicting stress.

Step 6. Challenges/Improvements

Challenges

- Binary variables where more information could have been gathered
- Imbalanced stress level class
- Model accuracy

Potential solutions

- Gather more data
- Resample/Bootstrap minor classes
- Optimization, hyperparameter tuning, grid search
- Try other algorithms

Thank you!

Questions?