Predicting Employee Productivity in Remote Work Environments: A Machine Learning Approach

Jihyeon LEE (20646185)
Juhyeon PARK (20581452)
Yunso KIM (20553699)

Abstract. After the onset of COVID-19, work from home (WFH) policies became a viable option for many companies. This study aims to identify reliable models for predicting productivity in remote work scenarios using conventional machine learning, deep learning, and feature selection techniques. The results show that the conventional models achieved the highest accuracy while deep learning models achieved the lowest mean squared error (MSE) values. Additionally, we identified ten survey questions that strongly influence the predictive performance of the models. This research provides insights into the factors that can enhance remote work productivity, benefiting both employers and employees.

Keywords: work-from-home, productivity, machine learning, deep learning, feature selection

1. Introduction

The Covid-19 pandemic has caused unprecedented disruptions in the global workforce, leading to widespread adoption of work-from-home policies. While remote work offers numerous benefits such as flexibility and reduced commuting time, concerns have been raised over its potential impact on productivity. As a result, there is a growing need for reliable models that can predict productivity in remote work scenarios. In this study, we aim to identify such models using conventional machine learning, deep learning, and feature selection techniques. Through our research, employers can gain insights into the factors that contribute to productivity and develop strategies to improve remote work performance. Also, employees can benefit by understanding the factors that affect their productivity and adjusting their work habits accordingly.

2. Literature Review

Prior to conducting the study, several works of literature were reviewed to gain insights into the overview of work from home and existing studies on remote work. The review was divided into three major topics: trends of remote work, the satisfaction of people with remote work, and productivity when working remotely. By examining these topics, this literature review provides a comprehensive understanding of the current state of remote work and its impact on employees' productivity.

To gain insights into the trend of work-from-home, two research were reviewed. One study is written by Smite et al., which explores the practical implications of the new trends emerging from the pandemic in tech companies (2023). The study examines employee preferences for WFH and post-pandemic work policies of 17 companies across 12 countries using ANOVA and post hoc tests. Another study by Islam examines work-from-home through the lenses of home-based work and precarious work, using a long-term case study of Prachi who worked in a small Indian e-commerce start-up (2022). The studies indicate that many companies are adopting remote work policies with limited work hours to promote better work-life balance and flexibility. This trend towards remote work is on the rise.

In addition, one of the insights from the literature reviews is people's perceptions of work-from-home policies. Two studies conducted by Kenny Roz et al. and Ritu Rani Sarker and Fariy Tabassum were analyzed to explore the relationship between work from home policies and work-life balance, work stress, and job satisfaction. The study by Kenny Roz et al. surveyed 472 participants and found a positive correlation between work from home policies and work-life balance, as well as a negative correlation between work from home policies and work stress (2021). The study by Ritu Rani Sarker and Fariy Tabassum surveyed 68 participants and found a significant positive correlation between work from home policies and job satisfaction (2021). Overall, these studies suggest that people have a positive attitude towards work-from-home policies and that such policies can improve work-life balance, reduce work stress, and increase job satisfaction.

Also, the impact of remote work on productivity and ways to improve it were investigated. Two studies on this topic were reviewed. Bucurean Mirela's study examined the benefits and challenges of remote work with employees of large companies in Bihor County and found that remote work has a negative impact on productivity (2020). Galanti et al.'s study investigated the hindrances and facilitators of remote work productivity, engagement, and stress among employees (2021). These studies found that family-work conflict and social isolation negatively influence remote work productivity and engagement while increasing stress levels. Therefore, a suitable strategy for managing conflicts and addressing social isolation is important to improve productivity.

3. Data

The data for this study was sourced from a survey conducted by the Treasury of New South Wales, a department of the New South Wales Government, in August-September 2020. The survey included 1,500 remote workers from the Australian state of New South Wales and aimed to identify the changes in remote work experiences and attitudes that occurred during various stages of the COVID-19 pandemic and to provide insights into the potential long-term implications of these changes. In this study, the dataset is utilized to predict the productivity of employees when they work from home. The survey comprised multiple types of 72 variables, and all the categorical variables were one-hot encoded to facilitate machine learning. Also, as the length of each survey question was long, they are modified into shortened names (Appendix 1).

3.1 Descriptive Statistics Analysis

3.1.1 Demographics

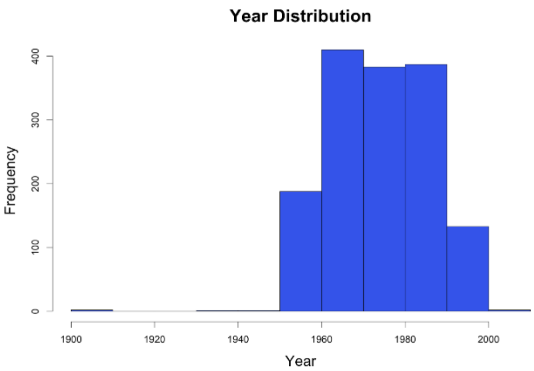

Figure 1. Histogram for age

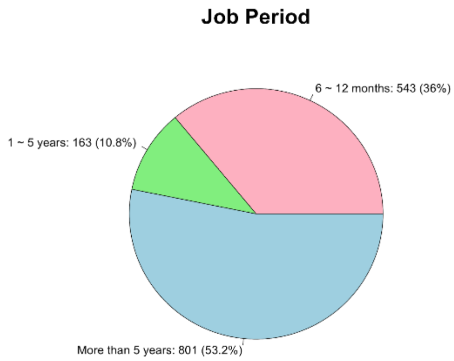

Figure 2. Pie chart for job period

The study participants were predominantly born between 1960 and 1990, with a mean and median year of birth of 1975, indicating a mature and experienced workforce. Over half of the respondents had worked for more than 5 years in their current company, suggesting a high level of job tenure. These demographic characteristics suggest that the participants were likely professionals with substantial knowledge and experience, capable of providing reliable and valuable responses to the survey.

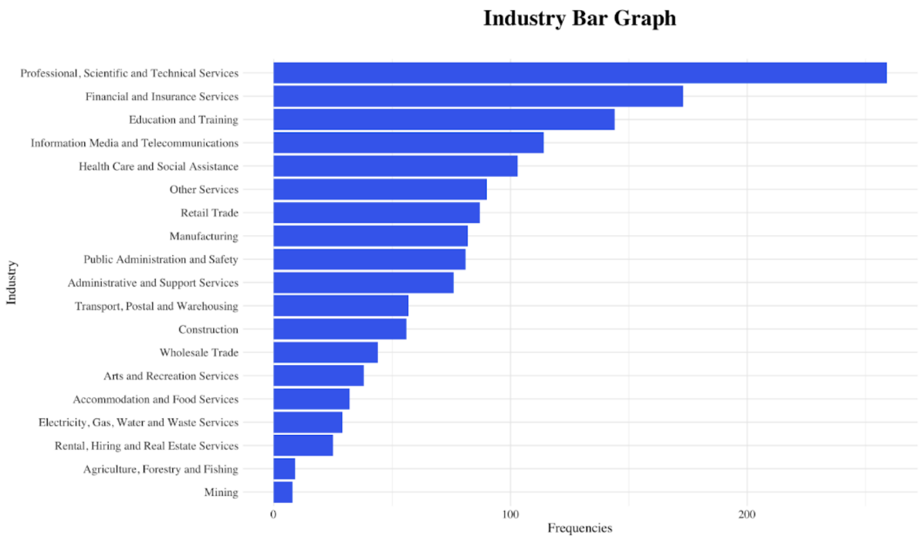

Figure 3. Bar graph for respondents' industries

Figure 3 is the distribution of respondents across various industries. The distribution shows that the majority of participants were employed in professional, scientific, and technical services, followed by financial and insurance services. The IT industry also had a high number of respondents, while non-IT industries had a lower number of respondents. These results suggest that the increasing trend in the IT industry may have contributed to the adoption of work from home policies as a means of improving employees' work-life balance and well-being.

3.1.2 Company's perspective on remote work

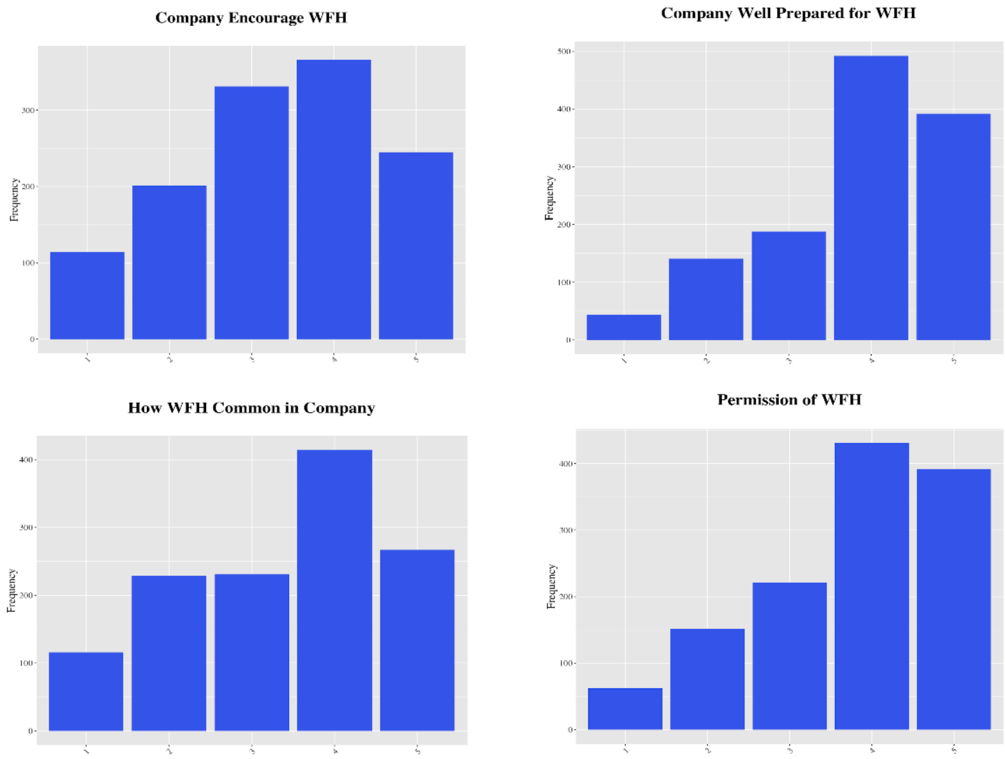

Figure 4. Barr graphs showing companies' perspective on WFH

The dataset includes information on how companies perceive remote work, which was analyzed through four variables: the extent to which companies encourage remote work, their level of preparedness for remote work, the ease of permitting remote work, and the prevalence of remote work within the company. The scores ranged from 1 (strongly disagree) to 5 (strongly agree). The results indicate that the most common response across all four variables was "somewhat agree" (with a score of 4), and the frequency of agreement was significantly higher than that of disagreement in all cases. These findings suggest that companies view remote work policies as beneficial and are already providing support for such policies to their employees.

3.1.3 Worker's perspective on work-from-home policy

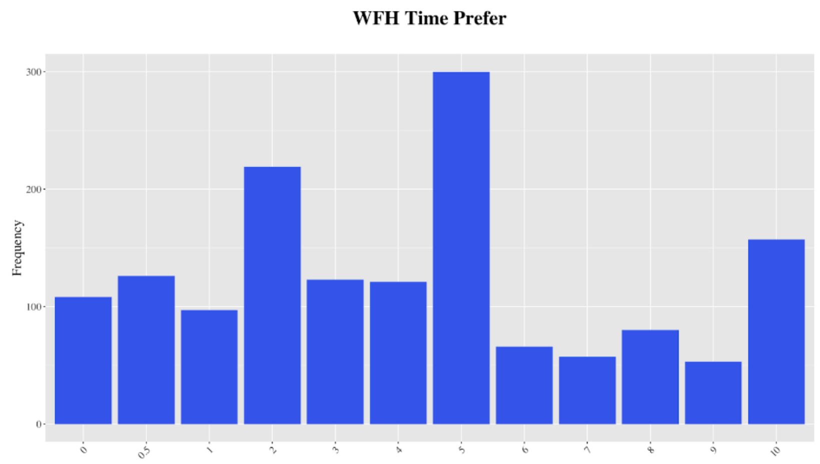

Figure 5. Bar graph for preferred working time for WFH

Also, several variables were analyzed to gain insight into how workers perceive work-from-home policies. One important variable, "time_prefer" showed the respondents' attitudes towards remote work. When asked how much they would like to work from home if given the opportunity, the highest frequency of workers indicated that they would like to work remotely more than 50% of the time. Additionally, the majority of respondents expressed a desire to work from home at least partially. The figure shows that very few workers (represented by the small number in the 0 category) do not want to work remotely at all. Overall, these findings suggest that most workers are supportive of remote work policies.

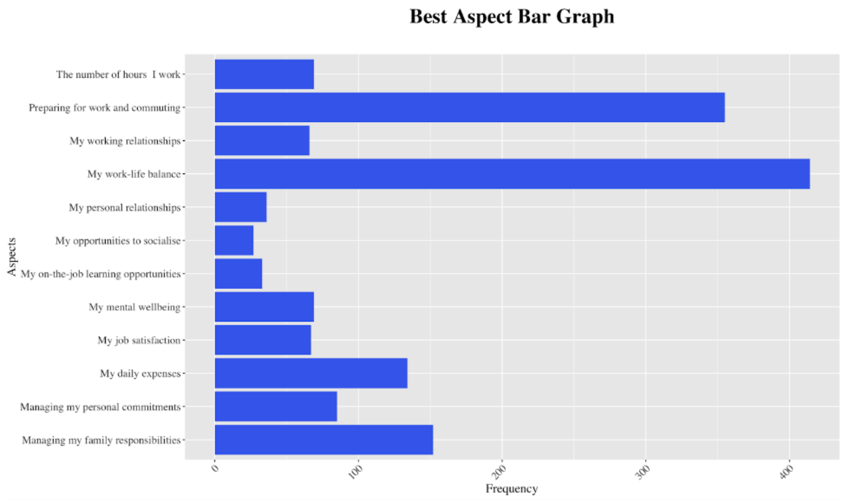

Figure 6. Bar graphs for respondent's favorite aspects on WFH

The presented bar graph sheds light on the reasons why workers have a positive attitude towards remote work. The graph indicates that the primary reason for choosing to work from home is to improve work-life balance, followed by the desire to reduce preparation time and commuting. These findings suggest that remote work policies can offer employees greater flexibility and control over their work schedule.

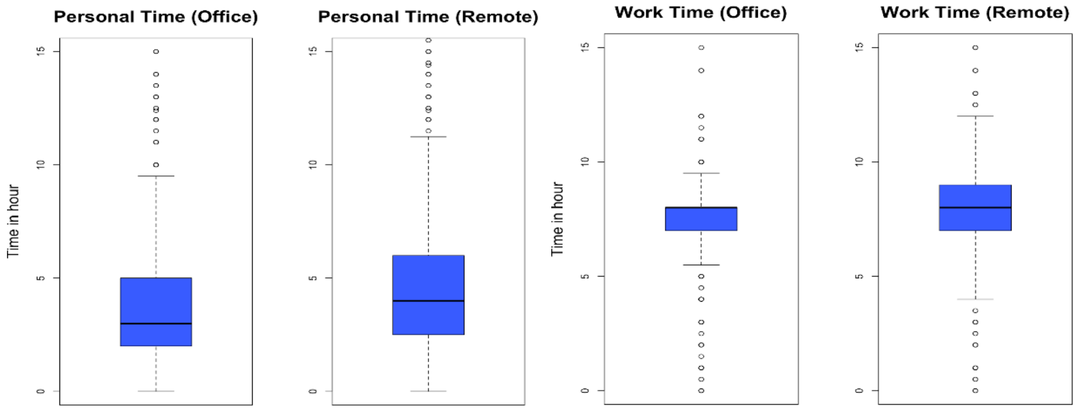

Figure 7. Box plots comparing changes in working and personal time

To assess the impact of remote work on work-life balance, the amount of time spent working and engaging in personal activities was compared between remote work and office work. Figure 7 illustrates that remote work has led to a significant improvement in both areas. Specifically, employees who worked remotely had a mean work time that was 0.3 hours longer than those who worked in the office, while also having an average of 0.7 hours more personal time. These findings suggest that remote work can provide benefits to both employees and companies, particularly in terms of promoting a better work-life balance.

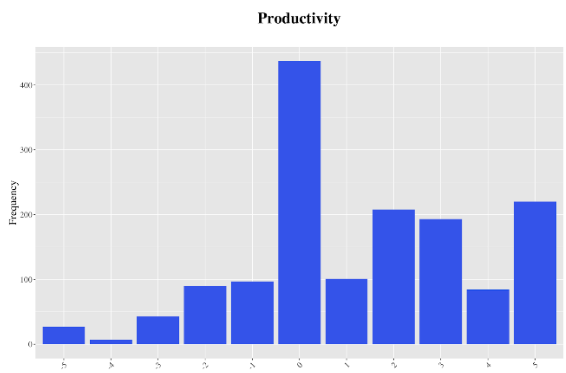

Figure 8. Bar graph for productivity for WFH

Remote work policies have been found to be beneficial for both work-life balance and flexibility, as well as for the productivity of workers. Figure 8 depicting the distribution of productivity, measured on a scale from -5 to 5 where a score of 5 indicates a 50% increase in productivity when working from home and -5 represents the opposite, reveals that there is no significant difference in productivity between working from home and working in an office. Additionally, the positive side of the distribution is slightly larger than the negative side, indicating a slight improvement in productivity when working remotely.

4. Research methods

This study investigated the effectiveness of five conventional machine learning models, namely Logistic Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting, and Multilayer Perceptron (MLP) with four different activations for predicting the productivity of employees based on a remote working dataset. The following sections provide a justification for why these models were chosen.

4.1 Conventional machine learning models

Multinomial logistic regression is a statistical method that models the probability of both nominal and binary outcome variables as a function of predictor variables. It is a simple and interpretable model that can handle both continuous and categorical predictor variables. As the outcome variable in this study was nominal and then transformed to binary, logistic regression was a natural choice.

Decision trees are a non-parametric method that can handle both categorical and continuous independent variables. They are easy to interpret and can capture non-linear and complex relationships between predictor variables and the outcome variable, so they were included in this study.

Random forests are an ensemble learning method that combines multiple decision trees to improve the accuracy and generalization performance of the model. They can handle both categorical and continuous predictor variables and can capture complex non-linear relationships as Decision trees do. Also, they are known to perform well on a wide range of machine learning tasks and can handle high-dimensional feature spaces.

Gradient boosting is an ensemble learning method that combines multiple weak models to improve the accuracy and generalization performance of the model. It can handle both regression and classification tasks and is known for its ability to capture complex non-linear relationships between predictor variables and the outcome variable.

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm that can handle both regression and classification tasks. It is known for its speed and scalability, making it suitable for large datasets and has been shown to perform well on a wide range of machine learning tasks.

All the models were implemented in default settings, as it allows for the reproducibility of results and facilitates comparisons between different models. However, hyperparameter tuning was performed in this study but only applied to the best models with high accuracy scores. This approach ensured that the best-performing models were further optimized while maintaining consistency in the experimental setup for all models.

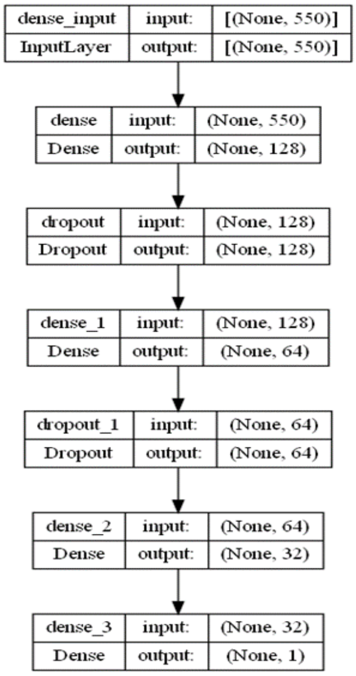

Figure 9 Architecture of MLP model

4.2 Deep Learning model

MLP is the most suitable deep learning model for the categorical survey dataset which can handle non-linear relationships between input features and output variables and can learn complex decision boundaries that may be difficult to model using linear methods. The performance of MLP depends on the number of layers used in the model, and three dense layers and two dropout layers were used in this study as displayed in figure 9. The three dense layers were chosen to allow the model to capture complex and non-linear relationships between the predictor variables and the outcome variable, and two dropout layers were used to reduce overfitting and improve the generalization performance of the model. With these five multiple layers, four different activations were also applied to MLP, including ReLU, sigmoid, tanh, and softmax. Sigmoid is an activation function that maps the output of a neuron to a value between 0 and 1. It is useful for binary classification tasks. ReLU (Rectified Linear Unit) is an activation function that sets negative values to zero and leaves positive values unchanged which is simple and computationally efficient. Tanh is a hyperbolic tangent function that maps the output of a neuron to a value between -1 and 1 which is useful for classification tasks and can introduce non-linearities into the model. Softmax is an activation function that is commonly used in the output layer of a neural network for multi-class classification tasks and maps the output of each neuron to a probability distribution over the classes. In the compilation process of the MLP model, the Adam optimizer, which computes individual adaptive learning rates for each parameter based on the first and second moments of the gradients, was used to optimize the learning rate.

4.3 Hyperparameter Tuning

In the field of machine learning, hyperparameters refer to the settings that determine the behavior and performance of a model during training, and their selection and tuning can have a significant impact on the model's ability to generalize to new data and avoid overfitting. To optimize the model's performance, the gridsearchCV library was employed, which exhaustively searches over a diverse range of hyperparameters to find the optimal combination. The selected hyperparameters are based on the best model and will be fine-tuned to achieve the most accurate results.

5. Results

5.1. Evaluations of Predictive Models

The target variable consists of 11 distinct categories that describe the level of productivity. Each category shows a percentage change in productivity, ranging from 50% less productive to 50% more productive. To use this variable in machine learning models, categories were encoded with numerical values from -5 to 5, where 0 represents no change in productivity.

The accuracy and mean squared error (MSE) scores for each model are listed in Table 1. Accordingly, the random forest model achieves the highest accuracy score of 0.3344, followed by the gradient boosting model at 0.2947 and the XGBoost model at 0.2914. These models appear to make more correct predictions than the logistic regression, decision tree, and multilayer perceptron models, which have lower accuracy scores between 0.1258 and 0.2781. However, in terms of the MSE, the multilayer perceptron with the ReLU activation function is the best-performing model with the lowest MSE of 4.6867, while the decision tree model has the highest MSE of 8.5397.

Table 1. Accuracy and MSE with 11 Prediction Categories

	Accuracy	**MSE**
Logistic Regression	0.2781	7.2417
Decision Tree	0.1788	8.5397
Random Forest	0.3344	6.0265
Gradient Boosting	0.2947	7.0298
XGBoost	0.2914	5.9768
Multilayer Perceptron_ ReLU	0.1291	4.6867
Multilayer Perceptron_ Sigmoid	0.1258	5.8091
Multilayer Perceptron_ Tanh	0.1457	5.8605
Multilayer Perceptron_ SoftMax	0.1623	4.8794

Due to the weak performance of previous models with 11 prediction classes, the number of categories in the target variable was reduced from 11 to 3, aiming to simplify the analysis and improve the interpretability of our results. The target variable was encoded using 3 numeric values: -1 for decreased productivity, 0 for no change in productivity, and 1 for increased productivity.

According to Table 2, the Gradient Boosting model achieved the highest accuracy of 0.6225, followed closely by XGBoost and Random Forest with accuracy scores of 0.6060 and 0.6026, respectively. On the other hand, the multilayer perceptron with the ReLU activation function has the lowest MSE of 4.6867. In terms of these two metrics, the decision tree model has the worst predictive performance with the lowest accuracy score of 0.4040 and the highest MSE score of 0.9735.

The analysis reveals a significant improvement in the performance upon categorizing the dependent variable into 3 classes. Specifically, an approximately two- to three-fold increase in accuracy and a ten-fold reduction in MSE score were observed in comparison to the previous models.

Table 2. Accuracy and MSE with 3 Prediction Categories

	Accuracy	**MSE**
Logistic Regression	0.5530	0.7947
Decision Tree	0.4040	0.9735
Random Forest	0.6026	0.6656
Gradient Boosting	0.6225	0.6457
XGBoost	0.6060	0.6821
Multilayer Perceptron_ ReLU	0.4967	0.5376
Multilayer Perceptron_ Sigmoid	0.4801	0.6556
Multilayer Perceptron_ Tanh	0.4801	0.6188
Multilayer Perceptron_ SoftMax	0.5430	0.5467

To further improve the prediction performance, the number of prediction classes is reduced to 2 values: 0 for decreased productivity, and 1 for no change in productivity or increased productivity.

The XGBoost model has the highest accuracy score of 0.8642, followed by Random Forest and Gradient Boosting, with identical accuracy scores of 0.8477. In terms of MSE, the multilayer perceptron model with SoftMax activation function shows the lowest score of 0.1231. Once again, the decision tree model has the lowest accuracy score of 0.7550 and the highest MSE score of 0.2450.

By categorizing the target variable into two classes, all predictive models except for the decision tree achieved accuracy scores of at least 0.82 and MSE scores of no more than 0.17, demonstrating significant improvements in both metrics.

Table 3. Accuracy and MSE with 2 Prediction Categories

	Accuracy	**MSE**
Logistic Regression	0.8212	0.1788
Decision Tree	0.7550	0.2450
Random Forest	0.8477	0.1523
Gradient Boosting	0.8477	0.1523
XGBoost	0.8642	0.1358
Multilayer Perceptron_ ReLU	0.8444	0.1310
Multilayer Perceptron_ Sigmoid	0.8278	0.1500
Multilayer Perceptron_ Tanh	0.8212	0.1613
Multilayer Perceptron_ SoftMax	0.8311	0.1231

5.2. Feature Importance

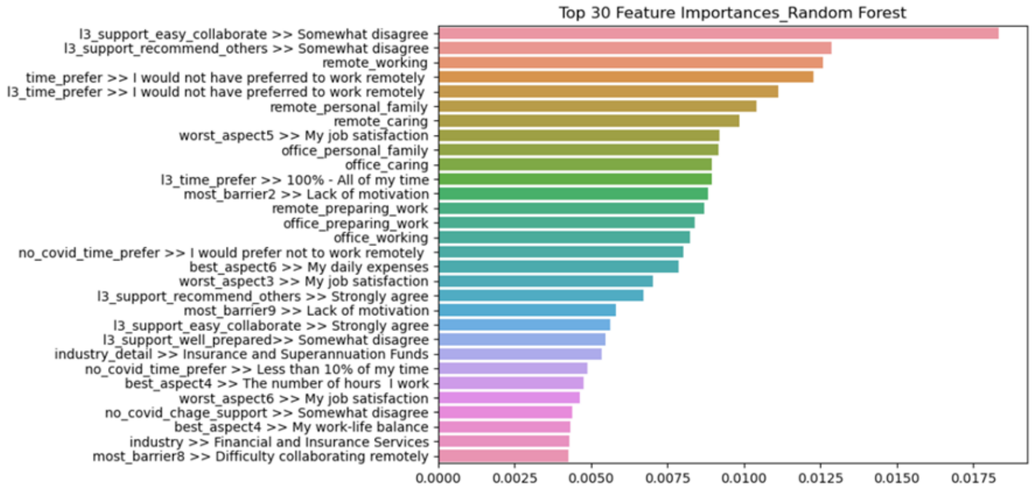

Figure 10. Top 30 Variables with Highest Feature Importance in Random Forest Model

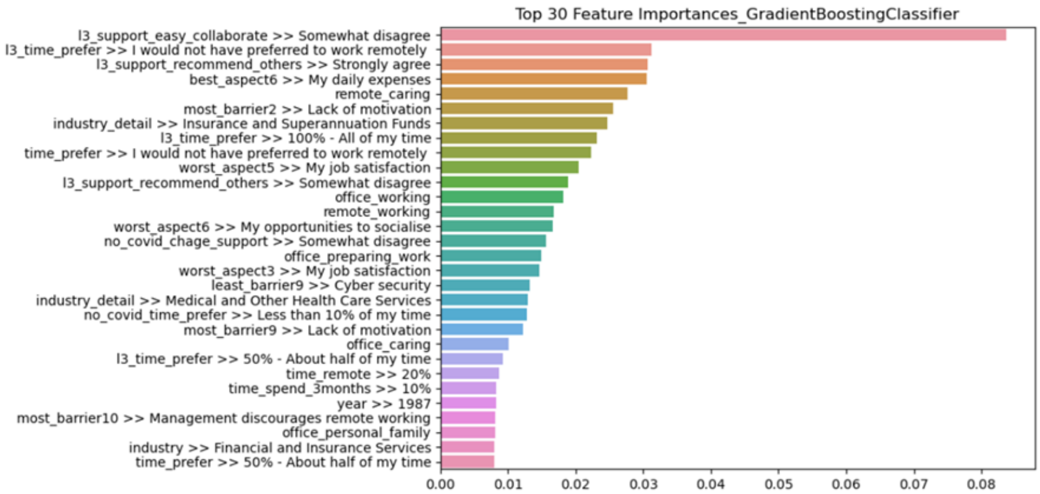

Figure 11. Top 30 Variables with Highest Feature Importance in Gradient Boosting Model

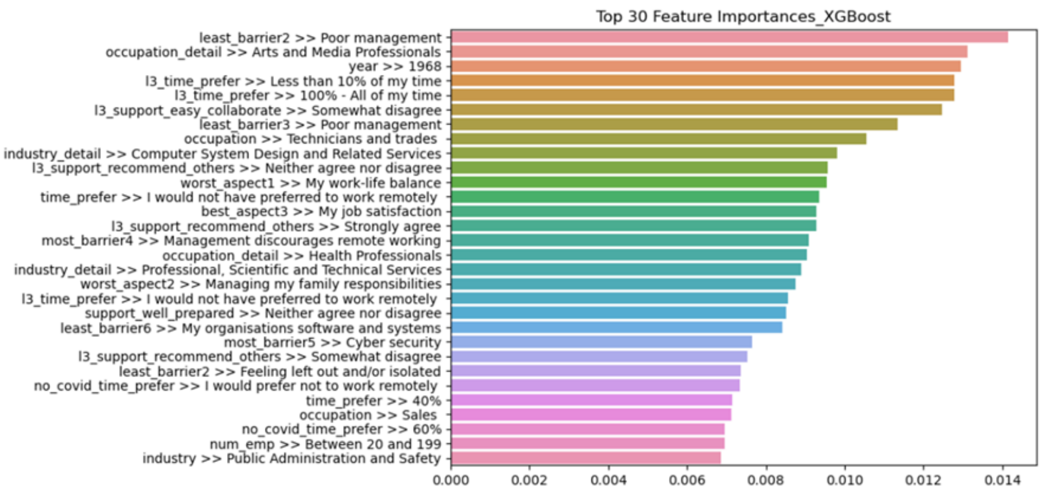

Figure 12. Top 30 Variables with Highest Feature Importance in XGBoost Model

Figures 10, 11, and 12 display the top 30 variables of the best-performing machine learning models, namely Random Forest, XGBoost, and Gradient Boosting. These variables are selected based on their feature importance values, which indicate the degree of contribution to the prediction of the target variable. The symbol '>>' is included in these figures to differentiate between the question and answer sections of the survey data, while continuous variables that do not have one-hot encoded categories do not include this symbol. This analysis enables researchers to identify the best survey questions that can be used to predict the productivity of remote workers. Each of the three models has a different set of important variables, and the best feature in each model has a distinct level of importance: 0.01834, 0.014163, and 0.083729 for the Random Forest, XGBoost, and Gradient Boosting models, respectively. The following are 10 questions that are commonly included in the top 30 features across all three models:

1. Which of the following best describes your industry?

2. Which of the following best describes your industry? (Detailed)

3. Compare remote working to working at your employer's workplace. Select the best aspect of remote working for you

4. Compare remote working to working at your employer's workplace. Select the worst aspect of remote working for you

5. From the following, please select the most significant barrier to doing your work remotely

6. Imagine that COVID-19 is cured or eradicated. Going forward, how much of your time would you prefer to work remotely?

7. How much of your time would you have preferred to work remotely last year?

8. How much of your time would you have preferred to work remotely in the last 3 months?

9. Thinking about remote working in the last 3 months, how strongly do you agree or disagree with the following statements? - I would recommend remote working to others

10. Thinking about remote working in the last 3 months, how strongly do you agree or disagree with the following statements? - I could easily collaborate with colleagues when working remotely

5.3. Hyperparameter Tuning

To optimize the prediction performance of the XGBoost model, which showed the highest accuracy on the given dataset, the 5-fold cross validation was employed to identify the best combination of hyperparameters using the GridSearchCV method provided by the scikit-learn library in Python. The hyperparameters that were tuned during the optimization process include the number of trees (n_estimators), the step size for each training iteration (learning_rate), the maximum depth of each tree (max_depth), the fraction of training data (subsample) and features (colsample_bytree) to be used for each tree, minimum loss reduction required to make a further partition on a leaf node of the tree (gamma), L1 regularization (reg_alpha) and L2 regularization (reg_lambda) on the leaf weights.

As a result, the best combination of hyperparameters for the XGBoost model was identified as: 'colsample_bytree': 1.0, 'gamma': 0.1, 'learning_rate': 0.2, 'max_depth': 7, 'n_estimators': 50, 'reg_alpha': 0.1, 'reg_lambda': 0, 'subsample': 0.75, resulted in an accuracy score of 0.8477 and MSE score of 0.1523 on the test set. Despite the extensive search conducted during hyperparameter tuning, the XGBoost model with tuned hyperparameters was found to offer the worse prediction performance on the test set, as measured by accuracy and mean squared error (MSE) scores. This suggests that the default settings of XGBoost were well-suited to the given dataset, and that further hyperparameter tuning could not yield significant improvements in performance.

6. Conclusion & Discussions

There are three main findings in this study. Firstly, the lowest MSE scores were obtained with the deep learning model, while the highest accuracy scores were obtained with conventional machine learning models. This is likely because deep learning models are better at capturing the complex, non-linear relationships between the input features and the output variable, resulting in lower MSE scores, while conventional machine learning models are better suited for classification tasks, resulting in higher accuracy scores. Secondly, the XGBoost model showed the highest accuracy score, while the Multilayer Perceptron with SoftMax activation function achieved the lowest MSE score, indicating that these models performed the best among all models evaluated in this study. Lastly, the analysis identified 10 questions that significantly contribute to predicting remote work productivity. These questions can be categorized into four groups based on their content: Industry of Occupation, Advantages/Disadvantages of Remote Work, Preference time for Remote Work, and Remote Work Experience.

Although the results of this study provide valuable insights into the effectiveness of the proposed approach, there are still several limitations to this study that must be acknowledged. Firstly, the use of a deep learning model for prediction limited the choice of available evaluation metrics. The attempt to use other evaluation methods like F1 score was made; however, because the deep learning model is based on regression rather than classification, it was really challenging to the researchers to find certain criteria to fit the values into confusion matrix. Thus, only accuracy and MSE are used in this study, as accuracy score is the same as micro-F1 score. Secondly, the data used in this study was survey data, which may be subject to response bias and may not fully capture the complexity of remote work experiences. The survey questions relied on the respondents' subjective answers, which may have been influenced by their personal biases or perception of remote work. Additionally, the sample size was limited, which may have impacted the generalizability of the findings to other populations. Thirdly, transforming the productivity target variable into a binary caused a skewness problem due to a high proportion of 'not changed' responses, potentially affecting the accuracy of the prediction. This imbalance may be due to central tendency bias in surveys. To deal with this issue, the data augmentation technique known as Synthetic Minority Over-sampling Technique (SMOTE) was employed to produce synthetic samples of the minority class, thereby balancing the dataset. However, this approach resulted in decreased accuracy scores of most models, ranging from 0.79 to 0.86, which ultimately reduced the overall predictive performance. Lastly, the data was collected during the COVID-19 pandemic, which may have influenced the remote work experiences of the respondents. The pandemic has brought about significant changes in the way people work, and the sudden shift to remote work may have affected the respondents' perceptions of their work experiences. As such, it is challenging to generalize the results of this study to remote work experiences outside of this context.

Despite the limitations of this study, the findings suggest competitive machine learning models to deal with remote working issues and the pathway that organizations may consider focusing on the industry of occupation, advantages/disadvantages of remote work, preference time for remote work, and remote work experience as key factors. With these suggestions, policymakers may update remote working policies to enhance employees' productivity in more efficient way.

Reference

Galanti, T., Guidetti, G., Mazzei, E., Zappalà, S., & Toscano, F. (2021). Work from home during the COVID-19 outbreak: The impact on employees' remote work productivity, engagement, and stress. *Journal of Occupational & Environmental Medicine*, *63*(7). , e426.

Irawanto, D., Novianti, K., & Roz, K. (2021). Work from home: Measuring satisfaction between work–life balance and work stress during the COVID-19 pandemic in Indonesia. Economies, 9(3), 96. https://doi.org/10.3390/economies9030096

Islam, A. (2022). Work-from/at/for-home: Covid-19 and the future of work – A critical review. Geoforum, 128, 33–36. https://doi.org/10.1016/j.geoforum.2021.11.018

Mirela, B. U. C. U. R. E. A. N. (2020). THE IMPACT OF WORKING FROM HOME ON PRODUCTIVITY. A STUDY ON THE PANDEMIC PERIOD. Annals of the University of Oradea, Economic Science Series, 29(2).

Smite, D., Moe, N. B., Hildrum, J., Gonzalez-Huerta, J., & Mendez, D. (2023). Work-from-home is here to stay: Call for flexibility in post-pandemic work policies. Journal of Systems and Software, 195, 111552. https://doi.org/10.1016/j.jss.2022.111552

Tabassum, F., & Sarker, R. R. (2021). Impact of work from home on employee satisfaction during COVID-19: An empirical study on bangladesh. Shanlax International Journal of Management, 9(2), 43–55. https://doi.org/10.34293/management.v9i2.4280