Post-Pandemic Working Modalities: Working from home or working in office?

By Linh Phung (692428)

I. Introduction

Covid 19 has made a dramatic change to our normal working life. In response to the imperative of social distancing regulations, a global workforce has undergone a paradigm shift toward remote working. Demonstrated as an effective solution to prevent the transmission of diseases within communities, working from home has facilitated the daily operation of business during this challenging time. However, whether it will persist as the new norm post-pandemic remains a controversial matter for organizations globally. The subject elicits conflicting perspectives, emphasizing the necessity to tailor strategies to accommodate individual preferences, leading to the formulation of the following research question: *Who would prefer working from home post pandemic? What are motivations of choosing working from home?*

Using survey data collected from employees about their opinions and backgrounds, this research aims to apply supervised machine learning methods to assist companies anticipating their employees' preferred working concepts, thereby fostering an adaptive and resilient work environment after the pandemic.

II. Data

Data for this research, which is available on Kaggle, was originally collected through surveying workers by S. A. D. D. Abesiri (S. A. D. D. Abesiri, 2022). It recorded answers of 325 employees about 4 types of information: Demographic, Working information, Technology and Experience/Opinion of online working (as described in Table 1).

At the end of the survey, each respondent would answer our target question: Do you want to work remotely post-pandemic? Responding to this question, 38.8% people said "No" and 61.2% people said "Yes", indicating a relatively balanced result.

Section	Variable	Data type
Demographic	Age group	Ordinal
	Gender	Nominal
	Marital status	Nominal
	Status of having children (Yes/ No)	Nominal
Working information	Type of working place	Nominal
	Employment category	Nominal
	Working experience (divided by groups)	Ordinal
	Commuting distance to work (divided by groups)	Ordinal
	Mode of work transportation	Nominal
	Monthly salary (divided by groups)	Ordinal
Technology	Better facilities for working online (Yes/ No)	Nominal
	Better internet access (Yes/ No)	Nominal

	Internet coverage	Ordinal 5-scale rating from 1 (lowest value) to 5 (highest value)
	Data charge for online working	
Experience/ Opinions of online working	Ease of completing tasks while working online	
	Work-life balance while working online	
	Computer skill for online working	
	English skill to handle computer	
	Preference of online working during Covid-19	
	Number of working days at office per week during Covid-19	Numeric
	Online working hours per day (divided by groups)	Ordinal
	Working online experience before Covid-19 (Yes/ No)	Nominal
	Working online experience during Covid-19 (Yes/ No)	Nominal

Table 1: Variable Properties

In the survey, question about monthly salary was designated as optional, resulting in 15.2% of respondents opting not to disclose this information. Given the potentially crucial nature of salary data, an imputation method was employed. Missing responses were filled in by assigning the most frequent salary group within each distinct employment category and working experience group, on the basis of close interaction between monthly salary and working experience and employment category indicated by low p-value through Chi-square test (Table 2). Subsequently, during the data preparation phase, categorical variables were transformed into factor, and the natural order of ordinal variables was preserved through ordered factors. Additionally, for Support Vector Machine, nominal variables were encoded into dummy variables, while ordinal variables were numerically represented based on their ordinal values. Finally, feature scaling through standardization was executed particularly for Support Vector Machine model to ensure all variables have the same scale.

III. Methods

In this research, I started with a simple, interpretable model – Logistic Regression. It uses Logistic function, which could be written as $p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$ where X = $(X_1 \ldots X_p)$ are p predictors. By taking logarithm on both sides, we can achieve a formula of log odds, presented as $log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$. To fit the model for estimating coefficients, we would use maximum likelihood method which finds the best estimates for coefficients such that it could maximize the likelihood of getting the observed data. Despite its ease of implementation, this method is based on certain assumptions: linear relationship between predictors and log odds of outcome variable, absence of multicollinearity, independence of residuals. These assumptions are difficult to meet in real life, possibly resulting in poor prediction and bias interpretation.

In this research, an additional methodology considered is tree-based methods. In contrast to logistic model, it offers a non-parametric approach, eliminating the need to adhere to diverse assumptions. Decision tree is known as the simplest, intuitive algorithm that works by recursively partitioning data based on the most significant features (determined by

impurity measures), but it commonly produces lower accuracy due to its tendency to overfit data. Hence, instead of building a single tree, I would use ensemble methods to combine a set of base classifiers (decision tree) to reduce variance. Specifically, I would use Random Forest which utilizes the fundamental principle of bagging technique that creates bootstrapped samples (random sampling with replacement) from training data. Furthermore, it decorrelates these trees by forcing each split performed in decision tree to consider only a random subset of predictors, thus, every variable has a better chance of being used in prediction. To decide how many variables should be used at any given split (m_{try}), cross-validation technique will be conducted.

Besides bagging, another ensemble method rooted in decision tree algorithm is boosting. Its core concept is growing trees sequentially so that each tree is constructed by learning from its predecessors. Numerous variants of the boosting technique exist, and for this study, I would use CatBoost because it is designed to handle categorical variable which is the dominant data type in our dataset. The algorithm could transform categorical variables to numerical features automatically using the ordered target statistics strategy so that we do not need to pre-process them beforehand. At the same time, it applies the ordered boosting method on training data by generating permutations and obtaining gradients on its basis to avoid overfitting on small dataset. For each permutation, we train n different supporting models $M_1,\ldots, M_n$ such that M_i uses the first i examples in the permutation and in each step, the residual of j-th sample is calculated using the model M_{j-1}. Based on model $M_{j-r,j}(i)$, we compute the corresponding gradients using the formula $\frac{\partial \mathrm{L}(\mathrm{y_i},\mathrm{s})}{\partial \mathrm{s}}|_{\mathrm{s}=\mathrm{M_{r,j}(i)}}$ and then, we approximate the gradient using cosine similarity. Similar to other gradient boosting methods, hyperparameter tuning through cross-validation will be performed to improve performance. Currently, the algorithm in R allows us to tune these parameters: tree depth, learning rate, number of trees, L2 regularization coefficient, the percentage of features used in each iteration, the number of splits for numerical features.

Support Vector Machine (SVM) will also be employed since it generally performs well with small dataset. The primary goal is to find a hyperplane that maximizes the margin between classes. However, practical scenarios often involve an inability to perfectly segregate classes using a hyperplane. Consequently, a concept known as "soft margin" is introduced, allowing some instances on the incorrect side of the margin. More importantly, SVM transforms data into higher dimensional space by using kernels to handle non-linearly separable data. One popular choice of kernel function is radial kernel, expressed as $K(x_i, x_{i\prime}) = exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{p}(x_{ij} - x_{i\prime j})^2\right)$, measure the similarity between pairs of data points in the feature space. There are hyperparameters for tuning using cross-validation technique: Cost (regularization parameter, or C, which balances between maximizing the margin and minimizing the training error) and σ (parameter defines how far the influence of a single training example reaches).

Finally, it is vital to interpret the model to understand motivations of working from home. However, except for Logistic model, other algorithms are considered “black-box” models because of their complexity. Hence, I would apply Global Surrogate method which trains an interpretable model to approximate the predictions of the complicated model. Subsequently, for a more granular examination of specific instances, LIME will be implemented. LIME creates a novel dataset using perturbed samples and predictions derived from the complex model. An interpretable model shall be selected to train on this new dataset, weighting based on the similarity of the sampled instance to the instance of interest. In the formula $\mathrm{explanation(x)} = \arg\min_{\mathrm{g}\in\mathrm{G}} \mathrm{L(f, g, \pi_x)} + \Omega(\mathrm{g})$, LIME would search for the best explanation to minimize the loss function.

IV. Analyses & Results

1. Implementation

The initial phase involved partitioning the data into training (70%) and testing (30%) sets. To preserve the original ratio of outcome variables, stratified partitioning would also be employed during the splitting process.

The first method to be implemented is Logistic Regression model, which resulted in 68.75% Accuracy. Given the predominance of categorical variables in our dataset, Chi-square test was employed to assess variable interactions. This analysis revealed significant correlations among several variables, particularly those within the same information group. For instance, as can be seen in Table 2, most p-value from Chi-square test for variables in the Employment information group are smaller than 0.05 (red text), signifying dependence among these variables.

	Type of working place	Employment Category	Working experience	Commuting Distance	Transpotation	Monthly Salary
Type of working place	-	9.884e-12	6.906e-09	2.768e-06	0.098663	9.685e-05
Employment Category	9.884e-12	-	0.163003	0.000507	7.837e-10	<2.2e-16
Working experience	6.906e-09	0.163003	-	0.077578	0.000169	<2.2e-16
Commuting Distance	2.768e-06	0.000507	0.077578	-	1.1e-05	0.239219
Transpotation	0.098663	7.837e-10	0.000169	1.1e-05	-	0.10833
Monthly Salary	9.685e-05	<2.2e-16	<2.2e-16	0.239219	0.10833	-

Table 2: P-value from Chi-square test

Because one of the principal assumptions was not fulfilled, and taking into account the limited predictive power indicated by low Accuracy, I would try other advanced techniques to get better prediction.

As explained in the Method section previously, tree-based methods shall be deployed. Firstly, Random Forest would be applied using caret package. The package made it easy to tune hyperparameter (m_{try}) through 5-fold cross-validation.

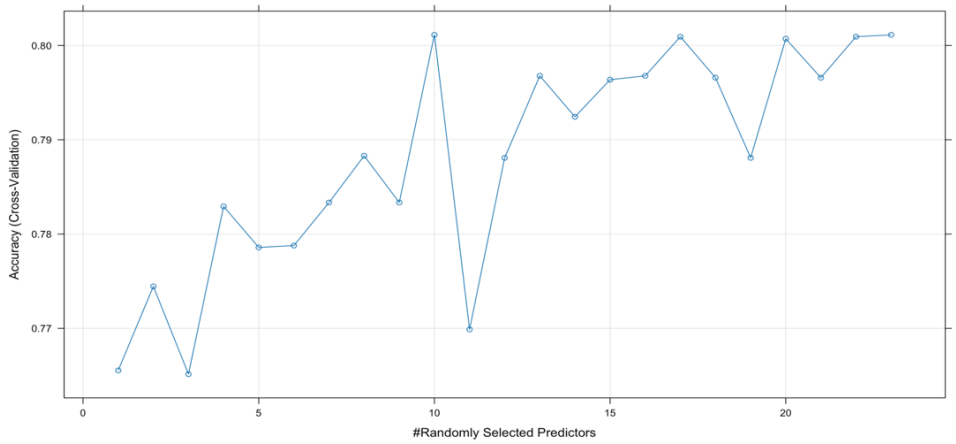

Figure 1: Hyperparameter tuning vs Accuracy for Random Forest

Figure 1 illustrated the variation in Accuracy with the increment of features in the model. Commencing at a lower value, it then reached the peak of approximately 81% when m_{try} = 10, then dropping sharply before rising again, but the Accuracy could not be as high as 81%. Therefore, the optimal choice for m_{try} in Random Forest model would be determined as 10. With the selected hyperparameter, applying Random Forest model on the training set resulted in an average of 80.97% Accuracy across 5-folds.

Aside from bagging technique used in Random Forest, another popular advanced technique to improve tree-based method performance is boosting. From various boosting alternatives, CatBoost was selected for its excellent handling of categorical variables. To optimize the model, a 5-fold cross-validation approach would be employed for the fine-tuning of multiple hyperparameters: tree depth, learning rate, number of trees, L2 regularization coefficient, the percentage of features used in each iteration, the number of splits for numerical features. Note that during the tuning process, I set up early stopping round equals to 100 to reduce the running time as well as avoid overfitting by stopping the algorithm if the evaluation metric did not improve after 100 rounds. The optimal combination of parameters yielded the highest Accuracy of 83.63% on training set.

Finally, Support Vector Machine would be deployed as another potential, powerful prediction method. Again, 5-fold cross-validation was executed using caret package to tune hyperparameters: Cost and σ.

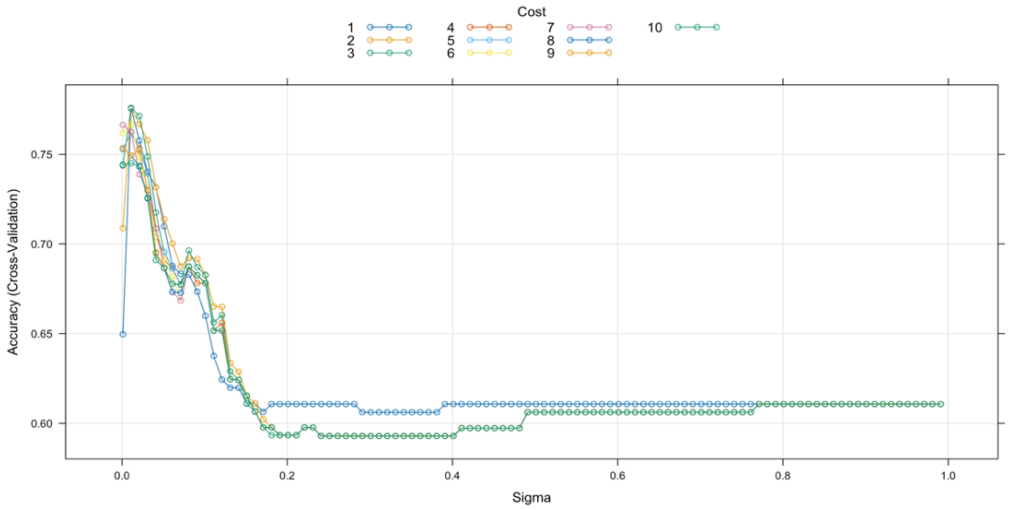

Figure 2: Hyperparameter tuning vs Accuracy for Support Vector Machine

As presented in Figure 2, the maximum Accuracy on the training set was attained at a low value of σ, specifically 0.011. Beyond this point, a pronounced decline in the metric was observed. Conversely, the Accuracy rates for Cost within the range of 1 to 10 did not exhibit substantial distinctions. Notably, with σ set at 0.011, the optimal Cost was determined to be 2, generating the highest average Accuracy of 80.97%.

2. Comparison and Results

Subsequent to the training of various models on training set, it is important to apply them on unseen data (testing set) for comparison purpose. In the context of a relatively balanced dataset, Accuracy serves as the most common metric. In addition, given that the research question's focused on those who express a preference for working from home (“Yes”), I

would use the Sensitivity (measure of model's ability to predict true positive) as the second metric for evaluation. Besides, model's running time would also be taken into consideration.

	Logistic Regression	Random Forest	CatBoost	Support Vector Machine
Accuracy	68.75%	83.33%	84.34%	79.2%
Sensitivity	69.49%	88.14%	88.14%	84.75%
Time	0.043 seconds	41.37 seconds	897 seconds	51.11 seconds

Table 3: Comparison of models

Apart from Logistic model which clearly performed the worst and violated some assumptions as mentioned before, the remaining models demonstrated good Accuracy and Sensitivity. Out of the three models, CatBoost gave the most favorable outcomes in both Accuracy and Sensitivity, even though it required more time for tuning process. Considering the relatively modest sample size, the runtime of 897 seconds (approximately 15 minutes) is acceptable in light of the enhanced performance. Notably, CatBoost exhibited the smallest disparity in Accuracy between the training (83.63%) and testing (84.34%) sets, indicative of a more consistent predictive capability. For these reasons, the selected model is CatBoost. To get a comprehensive understanding of the model, ROC Curve was developed.

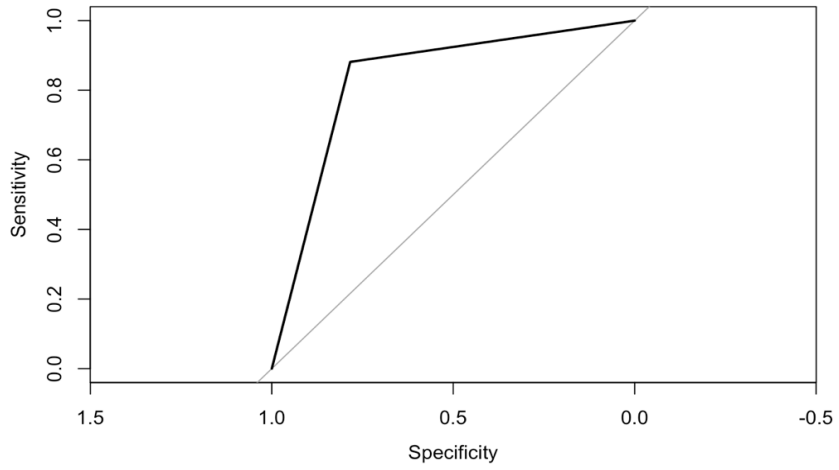

Figure 3: ROC Curve

Figure 3 expressed the trade-off between predicting correctly positive case (Sensitivity) and negative case (Specificity). The plot showed higher Sensitivity (88.14%), and slightly lower yet good Specificity (78.38%), suggesting that the model performed effectively in predicting both classes.

3. Interpretation

To address the second question regarding the motivations behind a preference for working from home, it is essential to interpret CatBoost model thoroughly. Because of its complexity, the model is commonly known as a "black-box" model and thus, model-agnostic methods are recommended for interpretation of such a complicated model.

As demonstrated during the diagnostic step for the Logistic model (depicted in Table 2), interactions between variables were observed. Consequently, Partial Dependence Plot and

Permutation Feature Importance could not be used since it may result in bias interpretation. For the same reason, Logistic model would not be a suitable choice as a surrogate model. Instead, Decision tree would be selected as the interpretable model. By running Decision tree model on the training set and replacing the original target variable with the predicted values from CatBoost model, Accuracy of 85.84% was achieved, meaning that the surrogate model effectively approximates the underlying model CatBoost. Hence, instead of CatBoost, I could interpret Decision tree. First, we would look at the learned tree.

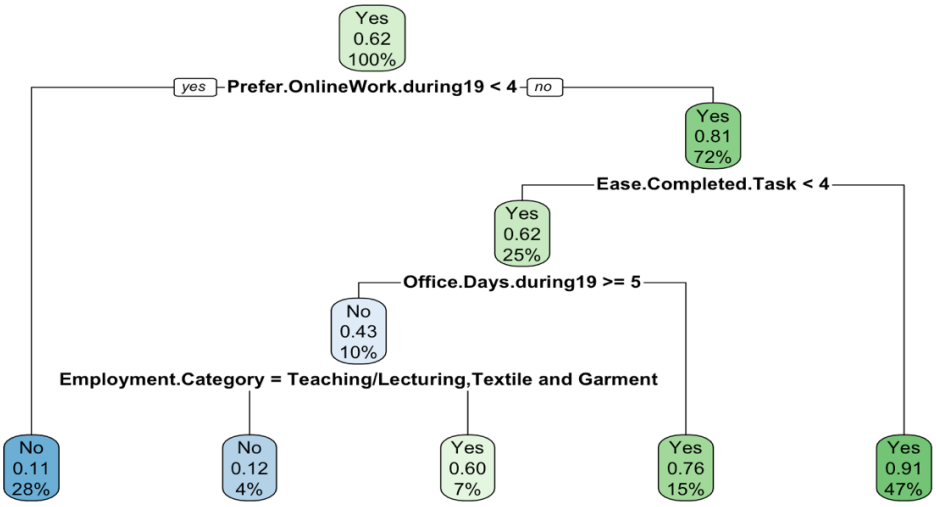

Figure 4: Decision Tree plot

Decision Tree model utilized Impurity measure (GINI) to determine the optimal node splitting and the most important feature would be used in the first split, which is the preference of online working during Covid-19. It is revealed that 72% individuals who highly favored remote work during the pandemic expressed a likelihood to continue working remotely in the future. Moreover, the likelihood of individuals preferring remote work increased with the ease of task completion online, especially for those who rated the ease of task completion equal or higher than 4. Besides, the visualized tree identified the number of office working days during Covid-19 and the employment category of participants as the subsequent pivotal variables for node splitting.

To further understand the importance of each feature in the model, I would draw a feature importance plot for the Decision Tree model based on its impurity measure.

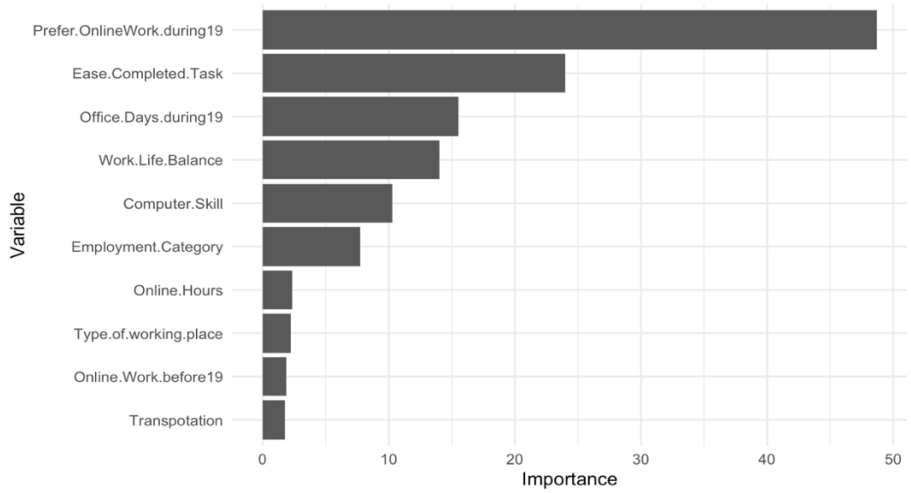

Figure 5: Feature Importance Plot

Figure 5 presented top 10 variables with the greatest impact on the model. As clearly seen from the plot, an individual's preference for an online working arrangement during the pandemic prominently influences their inclination to choose remote work post-pandemic, followed by the ease of completing task online and the number of working days at the office during Covid-19. These features collectively reflected an individual's experience with remote work, where a positive experience during the pandemic correlated with a heightened likelihood of seeking a similar arrangement in the future. Second in terms of importance were features related to the individual's working setting. Work-life balance played a significant role in the model, meaning that employees who prioritized their personal lives were more inclined towards remote work. Stay-home working option affords greater flexibility in working hours, minimizes commuting time, and facilitates an easier transition between working and non-working modes, thereby enhancing work-life balance. Also, employment category, type of working place and transportation appeared in the figure, emphasizing that people working in certain careers and working condition would be likely to opt for working from home option compared to their counterparts in other fields.

In the next stage, I would dig deeper into how the algorithm predicted whether a person like working from home or not. Understanding why certain predictions were made for specific instances aids in discerning potential factors contributing to prediction inaccuracies. Hence, I would use LIME to gain local interpretation of prediction for specific observations.

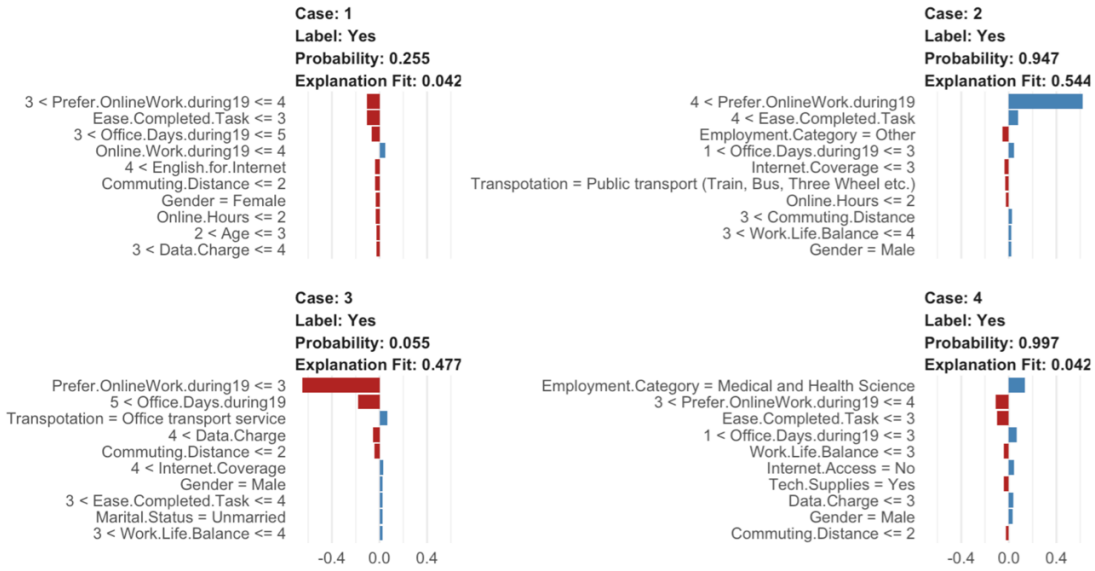

Figure 6: LIME visualization plot

Figure 6 comprised multiple graphs, each corresponding to an instance predicted with an incorrect label. It is noticeable that an individual's preference for online work during the pandemic, the ease of performing tasks online, and the number of working days at the office consistently ranked among the top 10 influential variables. This alignment reinforces confidence in the global feature importance interpretation of our model.

In case 1 and 4, the low explanation fit, accounting for less than 5% of the prediction outcome, may lead to incorrect prediction. For case 3, while the explanation fit was nearly 50%, the class probability of "Yes" was 55.5%, suggesting a relatively weak prediction that could easily lean towards class "No". In case 2, the features were able to explain 54.4% of the model with high probability (94.7%) associated with the label “Yes”. Surprisingly, this individual actually did not express a desire to continue working remotely post-pandemic. This anomaly underscores a rare scenario where a positive online working experience

during Covid-19 did not consistently translate into a preference for this working arrangement after the pandemic.

V. Conclusions

Covid-19 forced us to adapt to new digital working environment, but when social distancing is no longer a strict rule, the debate arises regarding the necessity for employees to return to physical workplaces. This research supports organizations to understand employees' preference for remote or on-site work, enabling tailored adjustments to working arrangements. Several methods have been implemented, and CatBoost emerges as the best model with a superior and stable predictive performance. Nonetheless, the method's drawback lies in the time-consuming hyperparameter tuning process. Therefore, it is suitable for small dataset such as the one used in this research, but I would suggest Random Forest for a larger dataset because of its robustness. In this study, I also figured out the major motivations of working from home. In general, individuals with positive experiences in remote work during the pandemic exhibit a preference to extend their remote work tenure, particularly when tasks can be efficiently completed digitally. Additionally, factors such as job nature, encompassing employment category, work-life balance, and type of working place, exert notable influence on their decisions.

Despites achieving good prediction using CatBoost, the research has certain flaws. The first concern lies in the small sample size, which may potentially lead to overfitting, diminishing the model's generalization ability. Secondly, current features in the model sometimes could not explain well the prediction of some instances (case 1 and 4 in Figure 6), revealing the room for an inclusion of more variables. Moreover, in spite of positive experience with remote working during Covid-19, some employees opposed the idea post-pandemic, highlighting the need to accommodate for challenges of working online (isolation, distraction, etc.). Finally, the survey only considered the limited binary option of working from home or not while in reality, hybrid working has become more famous. Thus, for future research, I would recommend diversifying the spectrum of working arrangement options, increasing sample size and exploring new variables. Such enhancements will contribute to a more comprehensive understanding of the dynamics surrounding modern work practices.

VI. References

S. A. D. D. Abesiri, R. A. H. M. Rupasingha, An Ensemble Learning Approach to Predict Employees' Choice for Continuing E-working Concept in the Post-Pandemic World, The IUP Journal of Information Technology, Vol. 18, Issue. 4, 2022.

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). christophm.github.io/interpretable-ml-book/

Tan, P.-N., Steinbach, M.„ Kumar, V. (2005). Introduction to Data Mining. Addison Wesley. ISBN: 0321321367

James, G., Witten, D., Hastie, T.„ Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R . Springer.

R Package: Usage Examples. (n.d.). Retrieved 12, 2023 from CatBoost: https://catboost.ai/en/docs/concepts/r-usages-examples

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: unbiased boosting with categorical features. http://arxiv.org/abs/1706.09516