**Assignment 2**

-Yixi Rao u6826541

Total words: 2073

**1 Introduction**

From the beginning of the 20th century, with the invention of the webcam, a new method that is different to the traditional face-to-face interview was born, namely video interview. At the same time, many companies were born using this technology, such as the HireVue company. The HireVue company developed an AI product based on a machine learning algorithm that helps the hiring managers quickly select the desirable candidates based on their diction, tone of voice, speaking patterns, and facial movements and body language. HireVue claims that this kind of technique can eliminate the unconscious bias in the hiring process and capture some cues overlooked by human. However, this company has been criticized for perpetuating bias on this algorithm.

This paper will analyze the controversy and identify the significant ethical issues and reasons that could explain this critique in section 2. Next, in section 3, a comprehensive plan, for undertaking the project in a way that addresses ethical concerns and problems mentioned in section 2, is presented. In section 4, some unavoidable trade-offs and difficult tensions are present, and justification is provided in this section.

**2 Reasons**

The source of the bias may come from the dataset. Most of this kind of automated decision is built based on machine learning, which works in the following way, it uses some specific dataset to train the model and predict the result based on the trained model. So, what happens if the datasets used to train machine learning model are not representative of objective reality? As a result, the algorithmic decisions are unavoidably unfair, and the results of this algorithm are often with discrimination and bias against some specific groups. An Example supporting this argument is the “Amazon’s sexist AI recruiting tool in 2018 [2]”, in [2], it said amazon using a secret AI recruiting tool that showed bias against women because it trained the model using the historical hiring dataset while Amazon has hired more men in the past. ML model learns this feature of this dataset, making it easier to ignore female candidates in its decision making.

Another reason why people criticize HireVue AI product is that this ML model does not provide sufficient interpretability. After reviewing the HireVue official website, I found that HireVue does not have a function that provides explanations to the rejected applicants, and HireVue uses the employability score to prioritize and rank the applicants while not providing the score and any explanation for that score. The explanation is very important for the applicants because the applicants might want to understand why they are rejected to prepare for the next application, or they might want to argue that the decision is unreasonable (such as they inferred from the interpretation that the decision is discriminatory) in the hope of withdrawing the decision.

Now assumed that HireVue has trained the AI product with a fair, balanced, and unbiased dataset, now the model can provide proper and fair prediction, however, another issue appears. Does the model really can capture human true nature or many important intangible qualities like emotional intelligence? Can the model really comprehend the facial movements? Such criticisms have been presented in [1] and [3].

In [3], Barrett evaluates more than 1000 research papers studying whether the human face shows universal expressions of emotion and how well algorithms can understand them. Barrett found that “this kind of system has become quite astute at detecting facial movements, such as the difference between a smile and a frown, but it is very difficult for the system to understand what those facial movements exactly means, and especially people communicate anger, disgust, fear, happiness, sadness, and surprise varies substantially across cultures, situations, and even across people within a single situation [3]”. For example, frowning during an interview does not mean a person is in a negative attitude, it is also possible that he or she is immersed in thought. Therefore, it is very skeptical that whether the HireVue AI product has the ability to predict an applicant’s employability.

**3 Plan**

To continue undertaking this HireVue AI project, a comprehensive plan must be presented to address the ethical concerns shown in section 2. The basic idea is to make the ML algorithm transparent and improve the AI or ML algorithm with high interpretability. If the HireVue company integrates the algorithm with high interpretability, it will give an explanation to the applicants who are rejected. Even though the training dataset is biased, the explanation will reflect the discrimination as long as the explanation properly explains how discrimination arises from prediction, then the applicants can contest this discrimination in the hope of having a second chance of a face-to-face interview.

However, interpretability is not a monolithic concept, but several distinct ideas that must be disentangled before any progress can be made. According to the description of Z. C. Lipton [4], the motivations of interpretability can be decomposed as trust, Causality, Transferability, Informativeness, Fair and ethical decision making. And the properties of the model that makes it interpretable fall into two different categories. Z. C. Lipton suggests, the first category is transparency which helps people to understand how the model works, the second is post hoc interpretability [4], which means the model gives the explanation without illustrating how the model works. Our plan is to build an interpretable model contains all of these two categories of properties. The implementation of transparency properties will be discussed in section 3.1, and section 3.2 will introduce what kinds of post hoc explanation will be implemented in this plan.

**3.1 transparency**

Transparency is the opposite of the term opacity; a model is opaque means that it is difficult for humans to understand why an algorithmic outcome is obtained. Therefore, to help people easily understand how the outcome is obtained, the transparency property needs to be further decomposed. It can contain three levels, as Z. C. Lipton suggested in [4], which is “the entire model level, the individual components level, and the training algorithm level [4]”.

The first level is the level of the entire model (simulatability), which means human beings can calculate every computation needed for each step to get the prediction result by using the provided input and all the parameters in a reasonable time period. In short, this person can simulate the whole process of the model at once. For the plan, a new webpage of algorithm’s introduction should be created under the HireVue website, this page is designed to descript the workflow of the model using a flow diagram. This diagram should contain different nodes to denote the different layers of the model and the diagram should contain all the weighted matrices that connect any two connected nodes; besides, the diagram should mark any of the specific algorithms or arithmetical operations it used. At the same time, this diagram should be intuitive and comprehensible for human to read. The challenge of this plan is that it is very hard to design an intuitive flow diagram for people to fully understand the whole process. This is because people’s cognition is limited.

The second level is the level of the individual components (decomposability). At this level, a model is decomposed into several explainable parts or steps, which means it should generate explanations for each part of the model, such as input and parameter. If HireVue using a decision tree model, then each layer of the tree can be tracked to see which features or thresholds are used to classify the child nodes, such as the model classifies samples with tone feature greater than a specific threshold to a specific node with score 5. The challenge of this level is that it requires the input to be interpretable as a prerequisite, which means it disqualifies some models with highly engineered or anonymous features [4]. On the other hand, in the pre-training and feature section state, the dataset should be selected appropriately since the biased in the dataset will influence the association between the features and labels.

The last level is the training algorithm level (algorithmic transparency), which will decide whether this algorithm can provide some guarantee, for example, proving that this model will converge to a solution or performing well on unseen dataset. The drawback of this level is Deep learning or neural networks is a little harder to explain from this perspective [4], for example, there is no guarantee of a convergence of a unique parametric setting since different settings are likely to achieve similar predictions. Therefore, our plan will not consider this level.

**3.2 post hoc explanations**

The post hoc explanation is giving some useful information without explaining how the model works, some common approaches contain Text explanations, Visualization explanations, Local explanations, example-based explanations [4]. our plan only considers implementing the text explanation and example-based explanation because these kinds of explanations are intuitive to the applicants.

The text explanation means using the natural language to explain why the model is producing this result. Krening et al using this kind of text explanation [5], their paper discusses a reinforcement learning agent chooses actions to maximize the cumulative reward based on the object-focused text explanation. For our plan, HireVue can develop one model to generate the prediction, and develop another model using the RNN (recurrent neural network) to generate the text explanation. The challenge of this approach is that the correctness of the explanation is not guaranteed, and there are no standard metrics to quantify how good an interpretation is.

An example-based explanation is proposed by Caruana et al [6], shows how to find the most similar case in the dataset to the case that needs to be explained by using the trained model as the distance metric. For our plan, the activations of the hidden layers of HireVue model can be used to find the most relevant sample by using the KNN algorithm. So, one of the explanations would be like this “Esther, who has a very similar speaking pattern to Nora but her diction is more polite, was given the job”.

**4 Trade-off and tensions**

When the project is going to run based on our plan, there are still some difficult tensions and unavoidable trade-offs. The first difficult tension is the requirement that the algorithm is interpretable conflicts strongly with the interests of the enterprise. In some cases, giving away too much information about how an algorithm works could allow malicious people to attack the system, which means that publishing information about the model’s inner workings may actually make it less secure or expose the company to more liability. In 2020, D. Slack et al showed that variants of LIME and SHAP, the two main technologies used to explain black-box algorithms, could be fooled to generate innocuous explanations which do not reflect the underlying biases of the classifier [10], meaning that explanations made by AI could be deliberately tampered with, leading to a loss of trust not just in the model but in its explanations too.

Another unavoidable trade-off is the trade-off between interpretability and predictive accuracy. Breiman claimed in [7] that complex machine learning models (especially random forests and neural networks) are almost unexplainable. So, in choosing a model, he said we always need to strike a balance between interpretability and accuracy. However, significant progress has been made in interpreting model’s prediction in the past few years, in particular, SHAP and LIME approaches can be applied to any model from random forest to neural network, and provide a reasonable explanation for the prediction of each model. But the argument of Breiman is still valid because algorithms with high accuracy can be developed much faster than interpretability. This is reasonable as we need to make sure the algorithm is accurate before trying to interpret them, and it is useless to explain the predictions of inaccurate models. Recently, Yi Luo et al also suggested that “no single interpretable (IP) and non-interpretable (NIP) approach is located at a Pareto optimum, where it enjoys both the highest accuracy and the highest interpretability, but it rather exists as a comprise between them [8]”. Another data scientist Sharayu proposes the relationship between the accuracy and summarizes the characteristics of a model with high interpretability. She said a model is highly interpretable if it has one of the following characteristics: linear & smooth, well-defined relationships, or it is easy to compute [9]. Thus, we have to find a balance between them, Yi Luo et al introduce some approaches to balance the accuracy and interpretability for some common ML approaches in [8].

**5 Reference:**

[1] A. Smith and M. Anderson, “Americans’ attitudes toward hiring algorithms” *PEW RESEARCH CENTER*, page 5, October 4, 2017. [Online], Available: <https://www.pewresearch.org/internet/2017/10/04/americans-attitudes-toward-hiring-algorithms/>

[2] J. Dastin “Amazon scraps secret AI recruiting tool that showed bias against women” *PEW RESEARCH CENTER*, page 5, October 4, 2017. [Online], Available:

[3] L. Barrett et al. “Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements.” *Psychological Science in the Public Interest*, vol. 20, no. 1, July 2019, pp. 1–68, doi:10.1177/1529100619832930.

[4] Z. C. Lipton. 2018. “The mythos of model interpretability”. Commun, 10, 36–43. October 2018.

[5] V. d. Maaten, L., Hinton, G. “Visualizing data using t-SNE”. *Journal of Machine Learning Research 9*, 2579-2605, 2008.

[5] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl and A. Thomaz, "Learning From Explanations Using Sentiment and Advice in RL," in IEEE Transactions on Cognitive and Developmental Systems, vol. 9, no. 1, pp. 44-55, March 2017, doi: 10.1109/TCDS.2016.2628365.

[6] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson. “Case-based explanation of non-case-based learning methods.” Proceedings. AMIA Symposium: 212-5, 1999.

[7] Breiman L. “Statistical Modeling: The Two Cultures[J]”. *Statistical Science,* 16(3): 199-215, 2001.

[8] Luo Y, Tseng HH, Cui S, Wei L, Ten Haken RK, El Naqa I. “Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling”. *BJR Open*. 2019 ;1(1):20190021. DOI: 10.1259/bjro.20190021. PMID: 33178948; PMCID: PMC7592485.

[9] S. Rane “The balance: Accuracy vs. Interpretability” *Towards Data Science*, December 3, 2018. [Online], Available: <https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062>

[10] D Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, “Fooling lime and shap: Adversarial attacks on post hoc explanation methods”. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180-186) February 2020.