Fairness Explorer: An Interactive Machine Learning Tool for Bias Assesment in Recruitment

Simon Iyamu Perisanidis Tom Wartmann Vito Pagone Jannis Widmer

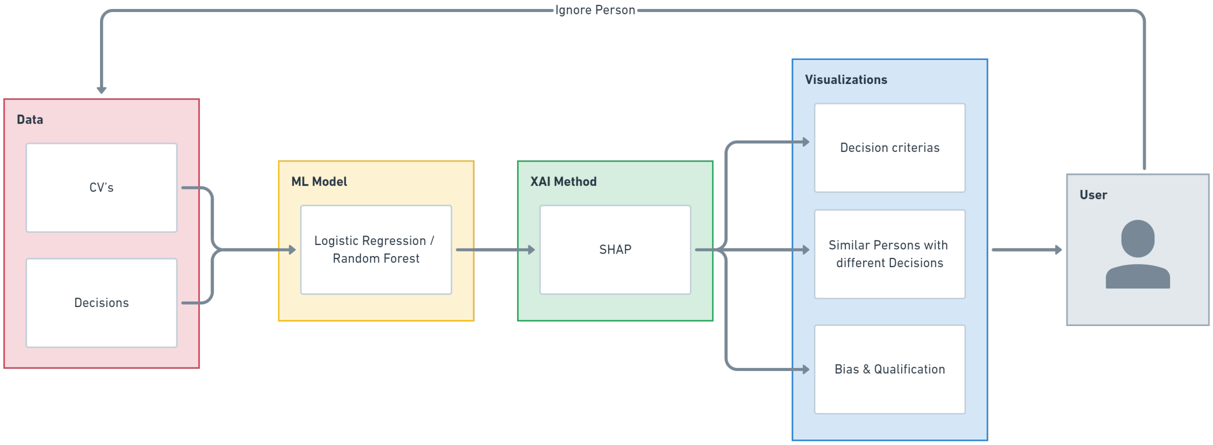


Figure 1: XAI Pipeline for our Application. Our system comprises four main components at a high level. Firstly, the data is inputted into a machine learning model, specifically a logistic regression or a random forest classifier in this case. Subsequently, the model undergoes analysis using SHAP (SHapley Additive exPlanations). Finally, the results are visualized to the user, who has the capability to interact with the system by removing specific data entries.

1 Introduction

In recent years, the advancement of machine learning has remarkably transformed our lives. Simultaneously, the study areas of human-computer interaction have gained popularity, in order to address the increased need for seamless user experience. In response to this evolving landscapes, the course Interactive Machine Learning was conceived, with the primary objective to explore the potential of creating user-centric applications that effectively employ artificial intelligence.

For the project of this course, our team developed a web application that aims to enable recruiters to explore any potential unconscious biases that they might have during applicant selection. We designed a platform for users, particularly recruiters, to explore and analyze their own unconscious biases based on their past hiring decisions. Our project focuses heavily on the development of the interactive user interface, and less on the methodology used for measuring human bias.

The core methodology employed in our project is contrastive learning. By presenting the user with counterexamples that highlight potential biases, we enable the user to reflect on their past decisions. As we recognise that the model should not act as an authority and confront the user with its own beliefs, we designed a system where the model and the user work together in collaboration to measure the fairness of the past hiring decisions. This is done by allowing the user to accept or reject the counterexample that was provided, and thus provide feedback to the model. This interactive feedback loop enables the model to be fine-tuned by the user's input, and creates a more nuanced assessment of one's biases.

Through the development and implementation of this application, we had the opportunity to put the course's theoretical concepts into practice and actively experiment with its various concepts that we learned. We learned not only how to put ourselves into the shoes of the user and design a user-centric system, but also the complexity of implementing a web application from scratch.

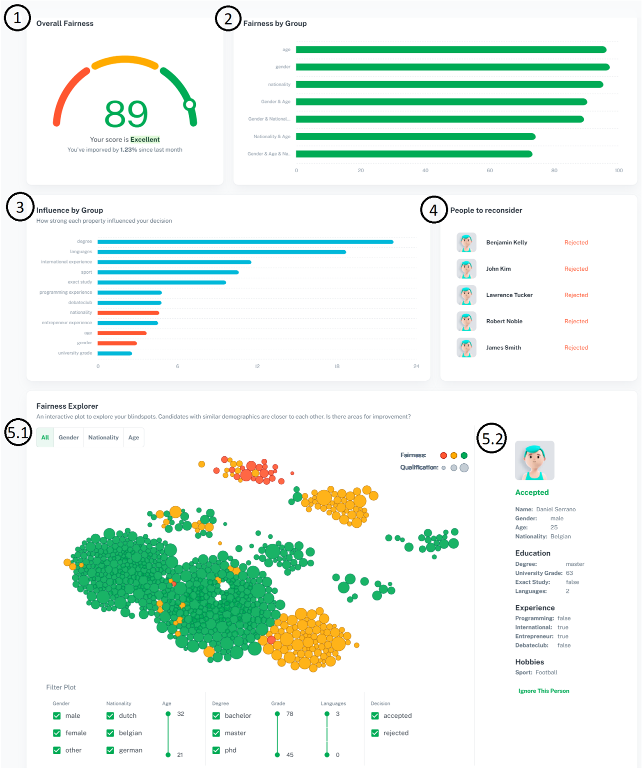
In this report we present an overview of our project, detailing the users, tasks, user interface and interactive machine learning techniques employed to develop our platform.

2 Users

The primary users of our web application are recruiters involved in the hiring process. Recruiters play an important role in evaluating job applicants and making decisions that can impact individuals' lives. Our application aims to empower recruiters by providing them with a tool to explore their own unconscious biases and make more fair and inclusive hiring decisions.

3 Tasks

The main task of recruiters using our web application is to explore and analyze potential unconscious biases in their past hiring decisions. The application provides a user-friendly interface that allows recruiters to interact with the machine learning model and gain insights into their biases. The tasks include looking at charts and plots about their hiring practices, reflecting on past hiring decisions, and exploring pairs of candidates that exhibit similar qualifications but different decision outcomes. Recruiters can reflect on these pairs and evaluate if any unconscious biases might have influenced their decision-making. The application also allows recruiters to provide feedback on the model's counterexamples, fine-tuning the analysis based on their expertise and knowledge of fairness in the recruitment process. Moreover, they are able to review a list of people that they might want to reconsider for a job role. Ultimately, the goal is to help recruiters identify and mitigate biases, leading to a more equitable and inclusive hiring process.



Overall Fairness

89

Your score is Excellent

You've improved by 1.23% since last month

Fairness by Group

Influence by Group

How strong each property influenced your decision

People to reconsider

Benjamin Kelly	Rejected
John Kim	Rejected
Lawrence Tucker	Rejected
Robert Noble	Rejected
James Smith	Rejected

Fairness Explorer

An interactive plot to explore your blindspots. Candidates with similar demographics are closer to each other. Is there areas for improvement?

Accepted

Name: Daniel Serrano
Gender: male
Age: 25
Nationality: Belgian

Education
Degree: master
University Grade: 63
Exact Study: false
Languages: 2

Experience
Programming: false
International: true
Entrepreneur: true
Debateclub: false

Hobbies
Sport: Football

Ignore This Person

Filter Plot

Gender: ☑ male ☑ female ☑ other
Nationality: ☑ dutch ☑ belgian ☑ german
Age: 32 – 21
Degree: ☑ bachelor ☑ master ☑ phd
Grade: 78 – 45
Languages: 3 – 0
Decision: ☑ accepted ☑ rejected

Figure 2: View arrangement of dashboard with panels numbered from 1 to 5.

4 User Interface

The view arrangement is structured as a dashboard with five panels. See Figure 2 for an overview of the design. The boards in our visualization have an implicit order from left to right and top to bottom, similarly to text.

The ”Overall Fairness” panel (Figure: 2, Panel: ①) shows a general fairness score, followed by the ”Fairness by Group” panel (Figure: 2, Panel: ②), that shows fairness with respect to common biased attributes such as gender nationality and age, including all their permutations. Those two panels are intentionally put at the first position - they are supposed to give users quickly and at first glance an overview of the situation.

The panels on the next row consist of the ”Influence by Group” panel (Figure: 2, Panel: ③), which displays an estimate of how strongly each property of the applicants influenced the decision regarding them, and the ”People to Reconsider” panel (Figure: 2, Panel: ④) showing 5 people which were supposedly mostly influenced by bias. Clicking on one of these people opens a pop-up window and the dashboard moves to the background so that the user's entire focus lies on the pop-up. It contains detailed information about the individual, as well as a second person who possesses highly similar characteristics, yet the recruiter has reached a contrasting decision. Next to each person's information, there is a ”ignore this person” button, which triggers a process further described in section 7.

Allowing users to quickly interact with the most biased decisions already should be sufficient to quickly explore edge cases of potentially biased recruiters decisions. It also allows to quickly check people whose decision might have been biased and remove them if this is found to be wrongly marked by the machine learning model or to check further people.

The last row contains the ”Fairness Explorer” which is a more engaging and itneractive view of the dataset. The plot (Figure: 2, Panel: (5.1)) contains decisions as points clustered by demographic groups. With the use of PCA we were able to encode the position of the candidates in the plot as their demographics, allowing for candidates of similar demographics to be closeby. The size of the points indicates the estimated ”quality” of an applicant and it's color how strongly this applicants decision was influenced by biases. Whereas a large circle stands for a high qualification and the red color for a large bias. For more detailed views we allow clustering the groups by different subsets of demographic information, namely by Gender, Nationality or Age. Users have the flexibility to change the view of these clusters, triggering smooth animations that dynamically reposition the data points. This dynamic visualization creates a sense of liveliness and interactivity, allowing users to explore and analyze the data from different perspectives. By seamlessly transitioning between cluster views, users can gain deeper insights into the relationships and patters within the dataset. For example, if they view the clusters by ”Gender” and see that the females cluster is more red than the male cluster (Figure: 3), that would be an indicator that they might have gender bias, and they can further investigate why these points are red.

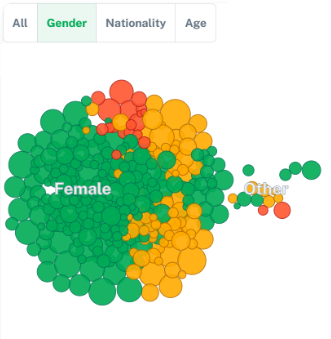


Figure 3: View of clusters grouped by ”Gender”.

Hovering over the points immediately shows a persons detailed information on the side of the plot (Figure: 2, Panel: (5.2)). We also allow filtering out certain people (e.g. only display people who where accepted). This plot is designed to provide a set of tools to the user which enable it to search effectively for biases and explore the application data. Actually understanding this plot and searching in it is the most complex task in our tool, which probably would not be done that often, which is why we decided to put it as the last panel. However it is also the most interesting one, as this actually allows drawing conclusions about prevalent biases. In our example data set this view on the data allows to actually understand the bias indicated above - the recruiters tend to discriminate against young Belgian people while being positively biased against older Dutch people.

5 Data

The data we used is from a synthetic dataset, called *Utrecht Fairness Recruitment Dataset*[0] which was published under the *CC BY-SA 4.0*[1] Licence. This dataset consists of decisions made by 4 companies about 4000 people and 14 attributes per person. The attributes contain information on demographics such as age, gender and nationality, information on education such as university grade, how many additional languages the applicant speaks and whether he/she has programming experience and what hobby is practiced and finally if the applicant was accepted or rejected. We extended the attributes

[0] https://www.kaggle.com/datasets/ictinstitute/utrecht-fairness-recruitment-dataset

[1] https://creativecommons.org/licenses/by-sa/4.0/

with an imaginary name for each person to make the user experience more realistic. Since for our use case we were only interested in the data of one company, we restricted the data to a random subset of 800 records from company C.

6 Machine Learning

In our collaborative endeavor, we developed an exhaustive, systematic pipeline that transforms raw candidate data into valuable insights, with the aim of informing fair, transparent, and efficient hiring decisions. By deploying machine learning techniques, specifically logistic regression and random forests, we have showcased the transformative potential of these algorithms for hiring analytics and beyond.

We initiated our approach by carrying out an array of preprocessing steps on the collected candidate data. Recognizing that the structure and cleanliness of the data is of paramount importance in the successful application of machine learning algorithms, we diligently executed data cleaning, normalization, and one-hot encoding of categorical variables. These careful pre-processing steps served as the bedrock of our pipeline, enhancing the overall quality of the data and, by extension, the reliability of our subsequent analyses.

Normalization was a fundamental part of our pre-processing phase. We addressed the potential issue of disproportionate influence from larger-scaled features by adjusting all feature values to a unified scale. This important step established an even playing field for all candidate features, and made certain that no single characteristic unduly dominated the outcome.

For categorical variables such as 'gender', 'nationality', 'sport', and 'ind-degree', we adopted a method known as one-hot encoding. This technique transforms categorical data into a format that can be effectively interpreted by machine learning models. By converting these variables into a binary vector representation, we ensured a fair, accurate, and equitable assessment of all candidates, regardless of their backgrounds or personal attributes.

When it came to identifying similarities among job applicants, we relied on the mathematical concept of cosine similarity. Cosine similarity is a metric used to determine how similar two entities are, and it is computed as the cosine of the angle between two vectors. This concept is of paramount importance in high-dimensional spaces where traditional notions of distance may become less meaningful. The cosine similarity between two vectors, A and B, is given by the formula:

$$cosine\_similarity(A,B) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

In this equation, A_i and B_i represent the components of vectors A and B, respectively, and n is the dimensionality of the vectors (i.e., the number of features for each candidate). The dot product $A \cdot B$ in the numerator measures the alignment of the two vectors, while the denominator normalizes this value by the magnitudes of the vectors, ensuring that the cosine similarity lies between -1 (completely dissimilar) and 1 (identical). Using this approach, we calculated the cosine similarity for every pair of candidates, forming a robust picture of the landscape of applicants. Candidates with cosine similarity exceeding a predetermined threshold were deemed similar. This not only allowed us to spot potential redundancies but also enabled us to uncover intriguing patterns within the candidate pool. The application of cosine similarity thus significantly enhanced the depth and breadth of our analysis, enabling a more nuanced and detailed understanding of our applicants.

At the heart of our pipeline is the logistic regression model. As a machine learning algorithm, logistic regression has a unique advantage: it is inherently interpretable. By using logistic regression, we can understand and articulate the relationship between each candidate's features and the hiring decision. This interpretability promotes fairness and transparency, as it gives us a clear picture of why certain candidates are chosen over others.

Lastly, to further elucidate the internal logic of our logistic regression model and demystify its predictions, we calculated SHapley Additive exPlanations (SHAP) values. These values provide a unified measure of feature importance, enabling us to determine the impact of each candidate feature on the hiring decision for every individual application. By leveraging SHAP values, we not only illuminated the decision-making process but also ensured that all candidate features were given equitable consideration.

7 Interaction Design

This section takes a look on how we enabled the user to interact with the machine learning model in order to collaboratively assess the user's fairness. This is done by allowing the user to choose to "ignore" from the measurement candidates that they consider they treated fairly. Behind the scenes, since our model is relatively static, ignoring people technically leads to the machine learning model being trained again with a dataset that does not contain ignored people. The rationale behind this is, that while our model can take guesses of potential biases, it's perception of the world is very limited, e.g. due to it's simple design or missing data. As such we do not want to make a strict statement that the decision for a certain person was biased and rather give the option to the user to take a second look at an applicant and figuring out if there was indeed bias involved. If the model detects some pattern in the decisions as bias which are actually not, then removing the applicants who introduced the pattern will remove it's detection as bias.

8 Conclusion

In this project, we successfully explored the potential of interactive machine learning in addressing unconscious biases within the recruitment process. Our user interface allows our users to explore and reflect on their biases through self-analysis and contrastive learning. By allowing the collaboration between the user and the model, we were able to create a system that works with and not against recruiters. Our dynamic visualizations provide a seamless and engaging experience. Future work is needed to prioritize the potential development of a more reliable machine learning model. The knowledge we gained through this project have been invaluable, and it serves as a solid foundation for future endeavors.