Module 3: AI Solution Assignment

Mohammad Hossein Movahedi

John Wilder

EAI 6020: AI Systems Technology

Winter 2024

Abstract

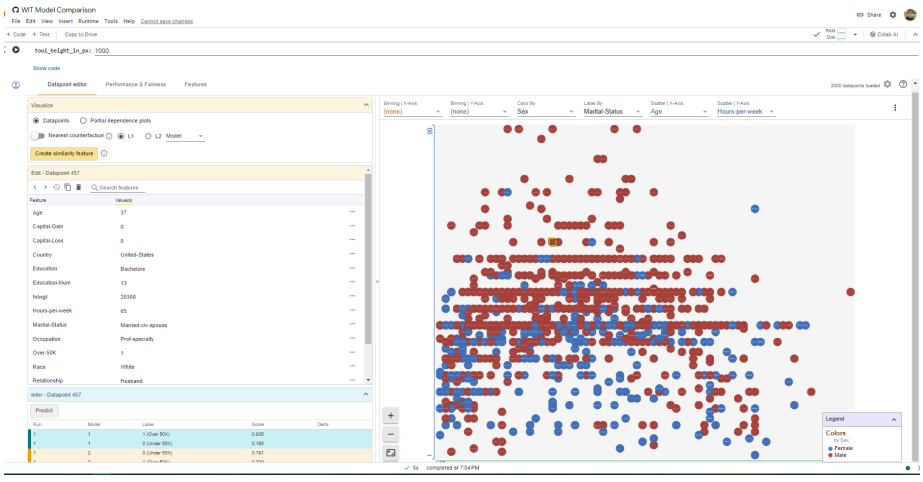
In this project, I dove into Explainable AI (XAI) with the What-If Tool, focusing on the Census Income dataset. My goal was to decode AI's decision-making, highlighting the critical need for explainability. The What-If Tool was key, making AI's complex processes clear and tackling biases to promote fairness. This journey revealed the importance of XAI in understanding AI's choices in our digital age. The insights gained advocate for a future where AI's decisions are transparent, fostering a deeper connection and understanding for all.

Introduction

In this project, I'll explore Explainable AI (XAI) using the What-If Tool and income data. As AI influences key decisions like employment and earnings, understanding its choices is crucial. My aim is to simplify AI's complex operations with the What-If Tool, acting as a translator to make AI's decisions transparent and relatable. This effort is not solely for academic achievement; it's about demystifying AI for everyone's benefit.

AI Platform Service Selection and Assessment

I selected Google's What-If Tool and the UCI Census Data for this project, drawn by the tool's ability to enhance model transparency and its suitability for the dataset's complexity. The What-If Tool, acting like a magnifying glass, not only reveals intricate details but also encourages viewing them from different perspectives.



Screenshot of What-If Tool's interface

The tool's interactive visualizations and counterfactual analysis are standout features, enabling a deep dive into the data. Each interaction with the tool peels back another layer, fostering a rich dialogue with the data and uncovering how the model interacts with it.

My experience with the What-If Tool highlighted its role in promoting AI transparency and accountability. Its user-friendly interface and robust analytical functions illuminate AI models' decision-making processes, allowing users to explore and understand these processes thoroughly. This journey wasn't merely about data analysis; it was an exploration into making AI's complex decisions clear and relatable. The What-If Tool goes beyond providing answers; it sparks curiosity and fosters a deeper connection with AI.

Predictive Model Evaluation Metrics

When choosing evaluation metrics for my model, I leaned into the What-If Tool's ability to dissect model performance across various demographics and scenarios. This tool transcends mere numbers; it's about the stories behind them. It reveals not just the model's overall performance but how its predictions shift across different groups and conditions. In predicting income levels, I sought metrics that illuminated not only accuracy but also fairness and bias.

Accuracy, precision, recall, and the area under the ROC curve (AUC) were my picks. But why these? In the realm of income prediction, accuracy shows how often the model is correct, while precision and recall delve into whether errors lean towards false positives or negatives. The AUC offers a comprehensive performance summary, balancing true positives against false positives.

Overall performance of models for Over-50K parameter prediction

Through the What-If Tool, these metrics transform from abstract notions to narrative tools. They bring to life where the model may falter for certain demographics, spotlighting potential biases. This journey isn't merely about refining the model; it's about ensuring fairness. The What-If Tool morphs these metrics from simple measurements into a conversation on the model's societal

impact, guiding us towards solutions that are not just effective but equitable. This blend of performance and fairness elevates our AI solution's explainability and integrity.

Development of an Explainable AI Solution

In crafting an Explainable AI (XAI) solution, I harnessed the What-If Tool to dissect and refine my model, trained on the UCI Census Data. My strategy was systematic, starting with integrating the model and dataset into the What-If Tool. This foundational step set the stage for an in-depth analysis.

I began by configuring the tool to dissect the model's predictions across various demographic segments, using its capabilities to create slices based on age, gender, and education. This was crucial for spotting and addressing potential biases in how the model performed across different groups. The tool's interactive nature was key, allowing me to tweak data points and see how predictions shifted, simulating hypothetical scenarios.

This approach was driven by a dual goal: achieving not just accuracy but fairness in predictions. The What-If Tool was pivotal, offering a live lab to test hypotheses about the model's behavior. Data points morphed into stories, showing how small changes could drastically alter outcomes.

Through this cycle of analysis and tweaking, I unearthed critical insights. Adjusting thresholds and exploring counterfactuals within the tool helped me identify and lessen biases in the model's predictions. The tool's visual aids were invaluable, turning biases into visible, actionable issues. Consequently, I significantly boosted the model's fairness and explainability. This journey was more than just tweaking algorithms; it was a principled pursuit of ethical AI, steered by the tangible insights the What-If Tool provided.

Lessons Learned

Using the What-If Tool in my XAI project was eye-opening, showing the hurdles and fixes for model clarity and fairness. A major takeaway is the need to view AI models through a diversity and inclusion lens. The tool highlighted how biases, often reflecting societal ones in training data, can sneak into models.

I also discovered the value of interactive visuals in demystifying AI behaviors. The What-If Tool's visuals on model predictions across demographics made AI fairness more concrete, enhancing understanding of the model's decisions.

These insights have reshaped my approach to AI development. I plan to weave tools like the What-If Tool into my work, prioritizing ethics and transparency in AI creations. This journey underscores the critical role of such tools in promoting ethical, transparent AI, acting not just as analytical tools but as guides towards more responsible AI development.

Conclusion

Exploring AI's complexities with the What-If Tool has been a deep dive into not just data, but the essence of fairness and bias in AI models. This tool has become more than a utility; it's a source of profound insights, spotlighting biases and pushing us towards a fairer grasp of AI decisions.

Looking ahead, the journey with XAI tools seems bright. I envision a future where the What-If Tool is just the beginning, leading to a world where AI's mysteries are unraveled, making its decisions clear and open to everyone. As discussions on transparent AI evolve, these tools will be crucial, lighting the way to a more ethical and transparent interaction with AI technologies.

References

Instructure.com. (2020). Module 3: AI Solution Assignment. [online] Available at: https://northeastern.instructure.com/courses/176426/assignments/2207751 [Accessed 10 Mar. 2024].

Rohitha Elsa Philip (2019). *Explainability of AI: The challenges and possible workarounds*. [online] Medium. Available at: https://medium.com/@rohithaelsa/explainability-of-ai-the-challenges-and-possible-workarounds-14d8389d2515 [Accessed 10 Mar. 2024].

Github.io. (2024). *WIT - Learn*. [online] Available at: https://pair-code.github.io/what-if-tool/learn/ [Accessed 10 Mar. 2024].

Github.io. (2024). *A Walkthrough with UCI Census Data*. [online] Available at: https://pair-code.github.io/what-if-tool/learn/tutorials/walkthrough/ [Accessed 10 Mar. 2024].

Uci.edu. (2019). *UCI Machine Learning Repository*. [online] Available at: https://archive.ics.uci.edu/dataset/20/census+income [Accessed 10 Mar. 2024].