[Applied Analytics Project Week 11 Assignment]

Major: Applied Analytics

Name: Heng Bai, Yuang Guo

1. Identification of Important Features in Your Model. What Are the Top 5-10 Predictors?

Our model, which uses an XGBClassifier, calculates feature importance through SHAP values to understand each feature's contribution to the predictions. The SHAP summary plot indicates that key features like Feature_0, Feature_1, Feature_2, and Feature_3 play significant roles in the model's output. These top features drive the model's decision-making process, as they have the highest influence on prediction outcomes.

2. Select 5 Predictions at Random, Explain How the Model Generated Those Predictions

To understand the model's decision-making process, we examined five random predictions and analyzed the SHAP values for each one. In each case, the model's output was driven by a specific combination of high-impact features. For instance, when the model predicted a particular class, features like Feature_1 and Feature_3 played a dominant role, with positive SHAP values pushing the prediction toward that class. If these features were adjusted downward—say, by decreasing their values by 20%—the prediction could potentially flip to a different class. This sensitivity to feature adjustments shows that even small changes in key variables can significantly influence the model's predictions. This insight is valuable for assessing model robustness and understanding which features contribute most to the classification boundaries, allowing us to interpret the underlying data patterns that the model is capturing.

3. Discuss Whether Your Dataset Includes Protected Categories and Whether You Are Using Them in Your Model

Our dataset does not include protected categories, such as age, gender, or race. As a result, there are no direct risks of discrimination based on these attributes, and our model does not rely on any protected categories for making predictions. This simplifies our analysis by removing the need to address biases related to protected classes. However, we still need to remain cautious about any indirect biases that might exist within other features, as these could still affect fairness in the model's predictions. By focusing on indirect sources of bias, we can ensure that our model remains as fair and balanced as possible, even in the absence of explicit protected categories.

4. Discuss the Bias of the Model (With Respect to the Protected Categories If Our Dataset Includes Them)

Our model's performance metrics, as displayed in the classification report and confusion matrix, reveal differences in precision, recall, and F1-scores across classes. These disparities suggest that the model may not treat all classes equally, which could be indicative of bias if the classes are associated with protected categories. For example, lower recall for a particular class could mean that the model is less likely to correctly identify instances in that group, potentially disadvantaging those individuals. If protected categories are included in our dataset, examining these class-wise performance metrics can help us understand whether the model systematically favors or disadvantages any specific group. Addressing these biases is critical for creating a fair and balanced model, as biases not only impact the accuracy of predictions but also have ethical and societal implications.

5. Provide Bias Removal Strategies and Their Impact on the Predictions

To mitigate potential biases in our model, we can consider several strategies, each with unique implications for predictions. One approach is to re-weight samples during training to ensure that underrepresented groups receive sufficient representation, which can improve fairness without drastically impacting overall accuracy. Another strategy is to exclude protected features directly from the model to avoid any potential bias they might introduce. Alternatively, we could introduce fairness constraints during model training, which aims to equalize the treatment of different groups. Implementing these methods may alter the balance of predictions by promoting equitable outcomes across classes. While these changes might slightly reduce overall accuracy, they help ensure that the model's predictions are fair and inclusive, fostering trust among stakeholders who rely on its decisions.

6. [Optional] Clean Up/Treat the Input Data and Remove Bias, Retrain the Model, and Calculate Model Performance Metrics

To thoroughly evaluate and address bias, we could preprocess the dataset by applying bias treatment techniques and then retrain the model. After bias mitigation, we would calculate performance metrics—such as accuracy, precision, recall, and F1-scores—across training, validation, and test datasets. Comparing these post-mitigation metrics with the original values enables us to determine the effectiveness of the bias removal strategies. Ideally, the metrics should indicate that the model's performance across classes is more balanced, with less disparity in accuracy and other scores between groups. This approach allows us to confirm that bias treatment not only improves the model's fairness but also maintains predictive quality, ensuring that ethical considerations do not come at the expense of performance.

7. Discuss Other Risks of Using the Model on Stakeholders (Customers, Citizens, Environment, Government, etc.)

The deployment of our model carries several potential risks, particularly if biases are not addressed. Biased predictions can lead to unfair treatment of certain demographic groups, potentially harming individuals who rely on equitable outcomes. For instance, if the model is used in a customer-facing application, biased results could damage trust, leading to dissatisfaction and reputational harm. Moreover, stakeholders such as regulatory agencies may scrutinize models for bias, exposing the organization to legal and compliance risks. Beyond individual groups, biased outcomes could also exacerbate existing social inequalities, making it essential to ensure that the model is ethically sound and socially responsible. Addressing these risks by building fairness into the model from the outset not only benefits affected individuals but also aligns with broader societal goals, fostering transparency, accountability, and trust in AI-driven decisions.