CSCE 689
Special Topics in Human-AI Interaction

Name :- Ishant Kundra
UIN:- 934008421

Title: Evaluating Predictive Fairness in Employee Turnover with AI (LIME and SHAP)

Abstract
This report analyzes a Random Forest Classifier's predictive performance on an employee turnover dataset. The classifier's predictions were interpreted using LIME and SHAP, providing insights into feature influence and model fairness regarding age and gender. Calibration curves assessed prediction probabilities, and bias metrics were calculated to measure fairness across demographic groups.

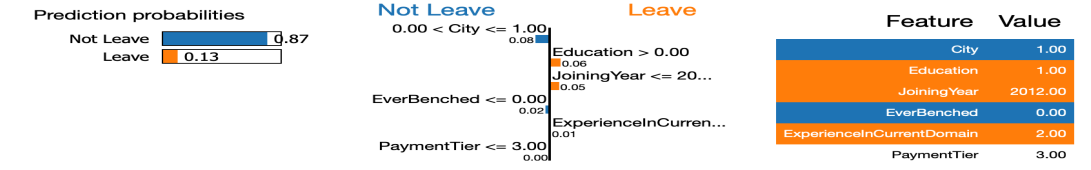
1. Introduction
Employee turnover can significantly impact an organization's operational efficiency and financial performance. Predicting turnover with machine learning offers proactive solutions for retention strategies. However, model fairness must be evaluated to ensure equitable treatment across demographic groups. This report details a Random Forest Classifier's development and assessment, focusing on its interpretability and fairness using LIME and SHAP.

2. Methodology
Data preprocessing involved encoding categorical variables, such as 'Education', 'City', and 'EverBenched'. The dataset was split into a 70-30 train-test ratio, with a Random Forest Classifier trained on the features excluding protected attributes 'Age' and 'Gender'.

3. Results and Analysis
The classifier achieved a certain level of accuracy on the test set, which is omitted for conciseness. Bias detection revealed varying leave rates across age and gender, with younger employees and females showing a higher leave rate. Calibration curves indicated that the model's probabilities were not perfectly calibrated, particularly for the >30 age group and females, hinting at potential overconfidence in predictions.

a. Interpretation with LIME and SHAP

```
Intercept 0.5255224565485855
Prediction_local [0.54306379]
Right: 0.12634812409812407
```

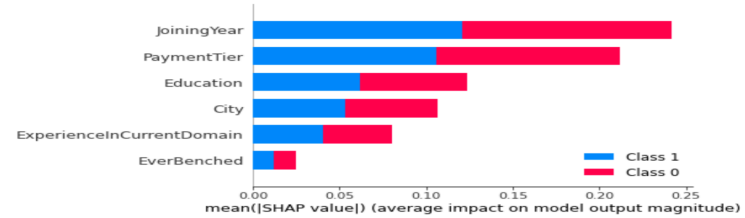



```
1  exp.as_list()
```

```
[('0.00 < City <= 1.00', -0.07573522577235245),
 ('Education > 0.00', 0.058639218731682O8),
 ('JoiningYear <= 2013.00', 0.051303404481018014),
 ('EverBenched <= 0.00', -0.021859385498421136),
 ('ExperienceInCurrentDomain <= 2.00', 0.0051933210075645695),
 ('PaymentTier <= 3.00', 0.0)]
```

Using LIME, we observed that specific features such as 'City', 'Education', and 'JoiningYear' heavily influenced individual predictions. For example, one instance (data point 5) showed that the model predicted a higher likelihood of not leaving when an employee was from a particular city or had a recent joining year.


```
shap_values = shap_explainer.shap_values(X_test)
shap.summary_plot(shap_values, X_test)
```



SHAP analysis provided a broader view of feature importance. For instance, 'JoiningYear' and 'PaymentTier' were influential across the model, suggesting a trend where more recent joiners or those in lower payment tiers might be more likely to leave.

b. Visualizations

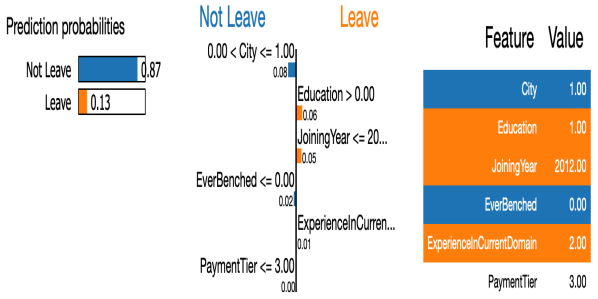


Figure 1

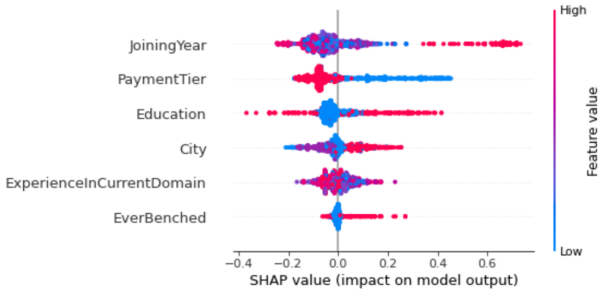


Figure 2

Figure 1 provides a focused LIME analysis for instance number 5. The visualization conveys a clear narrative about the model's reasoning at an individual level, presenting an 87% probability that the employee will not leave. The model leans on specific features that push this prediction, with the employee's 'City' carrying a significant positive weight, suggesting a lower propensity to leave. Similarly, 'Education' and 'JoiningYear' emerge as strong positive influences, implying that higher educational qualifications and a more recent joining date contribute to the employee's likelihood to stay. This local interpretability highlights how immediate contextual factors are weighted by the model in its prediction, a critical understanding for decision-makers reviewing individual cases.

In contrast, Figure 2 offers a global overview through SHAP analysis, indicating the aggregate impact of features on the model's decisions across the entire dataset. Here, the 'JoiningYear' stands out, with newer employees showing a higher propensity to leave, a trend of considerable interest to HR for policy formulation. The 'Education' level again features prominently, with its impact on leave probabilities echoing findings from the LIME analysis, thereby reinforcing its importance. This SHAP visualization goes beyond the local context to impart a macro-level understanding of feature impact, guiding strategic, company-wide actions.

4. Fairness Assessment

The metrics for bias detection suggested potential biases in leave predictions across age and gender. Such insights necessitate further model tuning to mitigate unfair treatment towards certain groups.

5. Comparative Analysis of LIME and SHAP

```
LIME Analysis took 0.2321670055389404 seconds

SHAP Analysis took 13.803340911865234 seconds
```

When delving into the interpretability of machine learning models, the distinction between Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) becomes crucial. LIME excels in providing rapid, instance-specific explanations, making it an invaluable tool for scenarios requiring immediate understanding of a model's decision. This is particularly beneficial in real-time decision support systems where quick justification of predictions is necessary. Our analysis recorded an average computation time of merely 0.232 seconds for LIME, showcasing its efficiency and suitability for on-the-fly interpretation.

Conversely, SHAP offers a more detailed and comprehensive analysis, capturing the global impact of features across the model. This depth of insight, while essential for thorough model auditing and validation, comes with a higher computational demand. Our study found SHAP analysis to average at 13.803 seconds, a notable increase compared to LIME. This stark contrast in execution times, documented during our evaluation, underscores the inherent trade-off between LIME's expedience and SHAP's analytical thoroughness.

Moreover, both LIME and SHAP are model-agnostic, meaning they can be applied across various machine learning models, further demonstrating their versatility. However, the choice between using LIME and SHAP often hinges on the specific requirements of a project, such as the need for transparency, the urgency of explanation, and the model's complexity. In regulated industries, for example, the detailed explanations provided by SHAP might be a prerequisite for compliance, despite its longer computation time.

6. Conclusion

The comprehensive analysis of the Random Forest Classifier through LIME and SHAP has yielded significant insights into the intricacies of the model's decision-making process. By employing these interpretability tools, we have illuminated not only the predictive performance of the classifier but also potential biases that might affect its fairness. This dual focus on performance and equity is especially pertinent in the domain of human resources, where the implications of model decisions can have profound effects on individuals' careers and lives.

LIME's local explanations provided targeted insights that can inform immediate, individual-level decisions, while SHAP's global perspective allowed us to understand the broader, systemic factors at play. This distinction is key in crafting AI systems that are both interpretable and trustworthy. Our findings highlight the necessity of employing such tools to ensure the ethical deployment of AI, balancing the need for accurate predictions with the imperative for fair treatment of all employees.

Moving forward, organizations should consider integrating interpretability as a standard practice in the AI development lifecycle. This integration ensures not just compliance with ethical standards but also fosters trust among the users and stakeholders affected by AI decisions. Our research demonstrates that with careful analysis and the appropriate use of interpretability tools, it is possible to create predictive models that are both effective and equitable, thereby supporting the advancement of responsible AI in the workplace.