Fairness Debugging of Tree-based Models using Machine Unlearning

**Motivation and Introduction**
- Concern continues to mount that ML systems reinforce systemic biases and discrimination often reflected in their training data.
- There has not been much work on understanding and debugging tree-based classifiers in the context of fairness.
- Tree-based ML models, such as decision trees and random forests, are one of the most widely used ML models.
- Following machine unlearning advances are utilized to find fairness-based bias inducing subsets in the training dataset:
  1. **DaRE – RF**: Data Removal-Enabled Random Forests [1]
  2. **SISA**: Sharded, Isolated, Sliced, and Aggregated [2]

**Experimental Setup**
- Fairness Metric – Predictive Parity [4].

<table>
<thead>
<tr>
<th>Subsets</th>
<th>Support</th>
<th>Bias Reduction</th>
<th>Accuracy Reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>housing = rent</td>
<td>18.12 %</td>
<td>88.87 %</td>
<td>5.40 %</td>
</tr>
<tr>
<td>property = real estate</td>
<td>28.25 %</td>
<td>84.25 %</td>
<td>0.60 %</td>
</tr>
<tr>
<td>cred_amt = high, debtors = none</td>
<td>21.87 %</td>
<td>79.78 %</td>
<td>3.37 %</td>
</tr>
<tr>
<td>age = young, employment = &lt; 1 year</td>
<td>15.12 %</td>
<td>83.77 %</td>
<td>4.70 %</td>
</tr>
<tr>
<td>duration = high, existing_creds = high</td>
<td>17.25 %</td>
<td>84.60 %</td>
<td>4.05 %</td>
</tr>
<tr>
<td>foreign_worker = yes, install_plans = bank</td>
<td>13.75 %</td>
<td>76.19 %</td>
<td>4.72 %</td>
</tr>
</tbody>
</table>

**Conclusion and Future Works**
- Fairness-based bias inducing subsets of training data (removing which has negligible accuracy loss) can be optimally found using techniques from machine unlearning.
- We plan to expand this study to other standard datasets and fairness metrics used in the fairness literature [4].

**References**

**Acknowledgements**
- This research was partially funded by a Google Research Scholar Award.