A MIXED-METHOD TO CHECK DATA OUTPUTS FROM RESEARCH WORKS

Remy Marquier - Titouan Rigaud
CONTEXT OF OUR STUDY

Welcome to the world of output checking!
OUR SEMI-AUTOMATED SYSTEM

Warning: it looks complicated, but it's actually simple.
FEATURE ENGINEERING

A magic spell capable of transforming a zip containing files to ... variables?
TRAINING MODELS ON THESE FEATURES

A gentle introduction to parallel architectures
DO YOU KUBERNETES?
TRAINING MODELS

Kubernetes Master

Node 1

Node 2

Node 3

Node 4

mlflow

PostgreSQL
WE CAN COMPARE THE MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>Learn Rate</th>
<th>Depth</th>
<th>Leaves</th>
<th>Estimators</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
<th>TP</th>
<th>ROC-AUC</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.07</td>
<td>4</td>
<td>10</td>
<td>200</td>
<td>113</td>
<td>994</td>
<td>2182</td>
<td>309</td>
<td>0.771</td>
<td>0.69</td>
</tr>
<tr>
<td>2</td>
<td>0.07</td>
<td>4</td>
<td>10</td>
<td>250</td>
<td>118</td>
<td>964</td>
<td>2212</td>
<td>304</td>
<td>0.770</td>
<td>0.70</td>
</tr>
</tbody>
</table>
PERFORMANCE OF THE MODEL

Does it actually work?

Sort of ... I guess
PERFORMANCE

Normalized confusion matrix:

- True label: 0
  - Predicted label: 0: 0.69
  - Predicted label: 1: 0.27
- True label: 1
  - Predicted label: 0: 0.31
  - Predicted label: 1: 0.73

ROC curve:
- AUC = 0.771

False Positive Rate (Positive label: 1) vs True Positive Rate (Positive label: 1)
IN OTHER WORDS:

Controlling 1 export out of 3 allows to obtain 75% of the total exports that needs to be rejected.
WHAT VARIABLES ARE IMPORTANT TO PREDICT?

That's a secret, but we know how to compute them.
FUTURE WORK

"I need to keep my job, please !"

(My boss is in the room)
CONTINUOUS TRAINING

- Detecting data shifting
- Labelling the data correctly
OTHER IDEAS

- Training a model file by file (350k files labelled available)
- Using more data from automatic controls
THANK YOU FOR YOUR ATTENTION

Feel free to ask any question!