Pitfalls of AutoML

ANALYTICS FRONTIERS CONFERENCE 2019
CLIFF WEAVER
Speaker Introduction

• **Cliff Weaver**
  • Let’s just say decades of work in technology
  • Data obsessed
  • R programming enthusiast and evangelist
  • Limitless passion to solve problems with data
  • Current work includes:
    • Local large company
    • RisMyHammer!
    • Local startups
Agenda

• AutoML – What is it?
• AutoML Options
• What are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
Why This Topic?

Why is this topic important to you?

To improve your business by helping you make informed decisions when evaluating AutoML solutions.

• Dodge the avoidable risk
• Plan and implement AutoML thoughtfully
Agenda

• AutoML – What is it?
• AutoML Options
• What are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
Before defining AutoML, distinguish machine learning from data science – it is more than a matter of semantics.

• Machine learning includes data modeling (selecting the best algorithm, tuning its parameters, etc.) which is part of a larger data science toolkit including data preparation, descriptive analytics, feature engineering, operationalization, etc.

• AutoML is **NOT** automated data science.
AutoML – What is it?

• AutoML (Automatic Machine Learning) refers to automated methods for model selection and/or hyperparameter optimization.

• The goal of AutoML software is two-fold:
  • Enable non-experts to train high quality machine learning models
  • Improve the efficiency to find optimal solutions to machine learning problems

• AutoML provides:
  • High-level machine learning algorithms to detect the best model or combination of models for a specific problem
  • Methods that automate hyperparameter optimization (relieving the data scientist from repetitive and tedious tasks)

  the automation of automating automation¹
Hyper-Parameter Tuning

model <- h2o.deeplearning(
x=predictors,
y=response,
data=train.hex,
classification=T,
activation="RectifierWithDropout",
hidden=c(1024,512,256),
hidden_dropout_ratio=c(0.5,0.5,0.5),
input_dropout_ratio=0.05,
epochs=50,
l1=1e-5,
l2=1e-5,
rho=0.99,
epsilon=1e-8,
train_samples_per_iteration=2000,
max_w2=10,
seed=1)

Nearly unlimited options!
What did we learn?

- AutoML is not data science – it helps with just a piece of the overall process
- AutoML saves time by automating model optimization
- Model optimization looks (and is) painful
Agenda

• AutoML – What is it?
• **AutoML Options**
• Where are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
AutoML – Options

• Almost every provider in the ML space provides some sort of AutoML solution
  • Some options are **FREE**, others $$$
    • Free solutions are typically provided as packages to be used in R or Python
      • As powerful as the $$$ options without the interface
      • Require data scientists to properly leverage the AutoML packages
  
• Paid option typically provide
  • Deployment options
  • Inviting interfaces that guide users developing models
  • A platform for monitoring deployed models
<table>
<thead>
<tr>
<th>AutoML Options</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoML R - free</td>
<td></td>
<td><a href="https://cran.r-project.org/web/packages/AutoML/vignettes/howto_AutoML.pdf">https://cran.r-project.org/web/packages/AutoML/vignettes/howto_AutoML.pdf</a></td>
</tr>
<tr>
<td>caretEnsemble- free</td>
<td></td>
<td><a href="https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html">https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html</a></td>
</tr>
<tr>
<td>Autoxgboost – free</td>
<td></td>
<td><a href="https://github.com/ja-thomas/autoxgboost">https://github.com/ja-thomas/autoxgboost</a></td>
</tr>
<tr>
<td>H2O – free and $$$</td>
<td></td>
<td><a href="https://www.h2o.ai/products/h2o/">https://www.h2o.ai/products/h2o/</a></td>
</tr>
<tr>
<td>BigML - $$$</td>
<td></td>
<td><a href="https://bigml.com/">https://bigml.com/</a></td>
</tr>
<tr>
<td>MLJar - $$$</td>
<td></td>
<td><a href="https://mljar.com/">https://mljar.com/</a></td>
</tr>
<tr>
<td>AutoKeras - free</td>
<td></td>
<td><a href="https://autokeras.com/">https://autokeras.com/</a></td>
</tr>
<tr>
<td>missinglink.ai - $$$</td>
<td></td>
<td><a href="https://missinglink.ai/">https://missinglink.ai/</a></td>
</tr>
<tr>
<td>Uber’s Ludwig – free</td>
<td></td>
<td><a href="https://uber.github.io/ludwig/">https://uber.github.io/ludwig/</a></td>
</tr>
<tr>
<td>TPOT - free</td>
<td></td>
<td><a href="https://github.com/EpistasisLab/tpot">https://github.com/EpistasisLab/tpot</a></td>
</tr>
<tr>
<td>DataRobot - $$$</td>
<td></td>
<td><a href="https://www.datarobot.com/">https://www.datarobot.com/</a></td>
</tr>
<tr>
<td>Cloud AutoML - $$$</td>
<td></td>
<td><a href="https://cloud.google.com/AutoML/">https://cloud.google.com/AutoML/</a></td>
</tr>
<tr>
<td>MLBox - free</td>
<td></td>
<td><a href="https://github.com/AxeldeRomblay/MLBox">https://github.com/AxeldeRomblay/MLBox</a></td>
</tr>
<tr>
<td>Binah.NOW - $$$</td>
<td></td>
<td><a href="https://www.binah.ai/">https://www.binah.ai/</a></td>
</tr>
<tr>
<td>Microsoft Neural Network Intelligence - free</td>
<td></td>
<td><a href="https://github.com/Microsoft/nni">https://github.com/Microsoft/nni</a></td>
</tr>
<tr>
<td>DATAIKU - $$$</td>
<td></td>
<td><a href="https://www.dataiku.com/">https://www.dataiku.com/</a></td>
</tr>
</tbody>
</table>
What did we learn?

• AutoML is not data science – it helps with just a piece of the overall process
• AutoML saves time by automating model optimization – saves time
• Model optimization looks painful!
• Lots of AutoML choices including open source and paid solutions
Agenda

• AutoML – What is it?
• AutoML Options
• What are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
AutoML Pitfalls

AutoML is just one part of the data science process

• View the business as a machine
• Understand the drivers
• Measure the drivers
• Uncover problems and opportunities
• Encode algorithms - AutoML
• Measure the results
• Report financial impact
AutoML Pitfalls

• AutoML assists the data scientist only in one step of the overall process.
• While AutoML can save time for the data scientist by running many experiments automagically, it does not address the rest of the data science process tasks.
  • This is often overlooked by organizations overly-eager to chase the promise of predictive analytics.

AutoML is NOT automated data science
AutoML Pitfalls

• AutoML acceptance is growing rapidly and an ever-growing number of industries rely on it.

• This success crucially relies on data science experts to perform the following tasks:
  • Preprocess and clean the data
  • Select and construct appropriate features
  • Select an appropriate model family
  • Optimize model hyperparameters
  • Postprocess machine learning models
  • Critically analyze the results obtained
  • Determine if the model needs to be re-trained
AutoML Pitfalls

• Businesses adopting AutoML platforms as a solution and not just as a piece of the data science process workflow expose themselves to avoidable financial risk.

• At one extreme, the AutoML algorithm may under-perform others. Model performance is important.

• Skilled interpretation of model results directly impact financial performance
AutoML Pitfalls – Results Matter

AutoML performance varies greatly across solutions

How do you pick the right one?

<table>
<thead>
<tr>
<th>Dataset Id</th>
<th>AutoML 1</th>
<th>AutoML 2</th>
<th>AutoML 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>0.497789992</td>
<td>0.315297615</td>
<td>0.304903671</td>
</tr>
<tr>
<td>38</td>
<td>0.165474774</td>
<td>0.045179262</td>
<td>0.038953435</td>
</tr>
<tr>
<td>44</td>
<td>0.383845093</td>
<td>0.136005382</td>
<td>0.125079404</td>
</tr>
<tr>
<td>722</td>
<td>0.4829143</td>
<td>0.314242093</td>
<td>0.225437935</td>
</tr>
<tr>
<td>737</td>
<td>0.314242093</td>
<td>0.225437935</td>
<td>0.524730876</td>
</tr>
<tr>
<td>740</td>
<td>0.572952603</td>
<td>0.567157133</td>
<td>0.295373383</td>
</tr>
<tr>
<td>741</td>
<td>0.484707493</td>
<td>0.295373383</td>
<td>0.254556068</td>
</tr>
<tr>
<td>742</td>
<td>0.560163406</td>
<td>0.254556068</td>
<td>0.232634802</td>
</tr>
<tr>
<td>819</td>
<td>0.517901444</td>
<td>0.232634802</td>
<td>0.23753078</td>
</tr>
<tr>
<td>821</td>
<td>0.420366818</td>
<td>0.23753078</td>
<td>0.23753078</td>
</tr>
<tr>
<td>822</td>
<td>0.442981634</td>
<td>0.23753078</td>
<td>0.23753078</td>
</tr>
<tr>
<td>823</td>
<td>0.342793913</td>
<td>0.394451891</td>
<td>0.206475294</td>
</tr>
<tr>
<td>833</td>
<td>0.436367941</td>
<td>0.243808173</td>
<td>0.25310065</td>
</tr>
<tr>
<td>837</td>
<td>0.455984028</td>
<td>0.267071446</td>
<td>0.255029891</td>
</tr>
<tr>
<td>843</td>
<td>0.545753158</td>
<td>0.314863141</td>
<td>0.246755019</td>
</tr>
<tr>
<td>845</td>
<td>0.304903671</td>
<td>0.246755019</td>
<td>0.240192515</td>
</tr>
<tr>
<td>846</td>
<td>0.390824138</td>
<td>0.390824138</td>
<td>0.390824138</td>
</tr>
</tbody>
</table>

logloss

Not all AutoML solutions are equal
An organization evaluates their overtime policy to determine which of several policies will benefit the organization best.

A new data scientist exclaims, *That’s easy!*

The data scientist inputs his data into an AutoML solution and behold – an answer! The data scientist reports to management that the Max F1 score is 54%.

The answer is wrong. (Missed by ~$340k)
What did we learn?

• AutoML is not data science – it helps with just a piece of the overall process
• AutoML saves time by automating model optimization – saves time
• Model optimization looks painful!
• Lots of AutoML options both open source and paid solutions
• There is no easy button for data science
• There can be significant differences between AutoML solutions
• Data Science training is required to interpret AutoML results with business domain expertise
Agenda

• AutoML – What is it?
• AutoML Options
• What are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
Future of AutoML

• The future of AutoML is coupled with the future of data science.
  • The future of data science will likely follow the path of past innovations:
    • A calculator was once a person
    • Webmaster was once a hot career
    • Microsoft Office was once a typing pool (does anyone but me recall those days?)

• The data science role will (is?) changing and AutoML is the first visible evidence
  (data scientists will still have a job, it simply changes).

• Machine learning will become industry-specific, so perhaps will AutoML.
Future of AutoML

- Data scientists will have different enterprise roles once tools democratize machine learning.
  - Entry level data scientists will interpret data and make it usable and focus on educating end users.
  - Advanced data scientists will improve the performance of algorithms and designing new generalized approaches.
  - Domain expertise with data science skills will become a key role in the future. These data scientists will apply data science techniques and tools in specific verticals like manufacturing, healthcare and finance.
  - The evolution of the data science practice manager with proven abilities to build effective teams and the processes that drive the business to greater success.
### TODAY
Currently, selecting the “best” algorithm to use per dataset requires a level of intuition or expertise about the data. Data scientists leverage their experience to experiment with different combinations of models and hyperparameter values to achieve the highest accuracy.

### TOMORROW
AutoML will lessen the data scientist dependency on intuition by iteratively trying out an algorithm, scoring its performance, and choosing and refining other models. AutoML solutions will evolve to platforms automating more of the data science workflow.
Future of AutoML

• AutoML systems will become mainstream in the machine learning world.
  • Will include data prep, industry specific feature engineering, API deployment, monitoring and retraining
  • Future AutoML will be interactive where the user and AutoML system will work together
  • AutoML system will learn in real-time from the user's experience and adapt its optimization process

• Because of AutoML, the data science role will fade and become a business science role
What did we learn?

• AutoML is not data science – it helps with just a piece of the overall process
• AutoML saves time by automating model optimization – saves time
• Model optimization looks painful!
• Lots of AutoML options both open source and paid solutions
• There is no easy button for data science
• There can be significant differences between AutoML solutions
• Training is required to interpret AutoML results with business domain expertise
• AutoML will evolve dramatically and become mainstream tools – the future is bright
• The data scientist role will evolve in the coming years – the future is bright!
Agenda

• AutoML – What is it?
• AutoML Options
• What are the pitfalls?
• The Future of AutoML
• Recommendations
• Q&A
Recommendations

• AutoML should be viewed as a tool; it is not a solution or a substitute for data scientists

• Data Scientists must provide guidance and oversight to AutoML implementations

• AutoML solutions will advance quickly – time your investments wisely. (It's going to be exciting)

• Contrast and compare AutoML solutions – there are differences that matter. Shop around, become informed!

• Start small – your organization is not ready for a big-bang deployment (no one is). Requires process!
Questions?

Cliff Weaver

www.rismyhammer.com

https://www.linkedin.com/in/cliffordweaver/