Overcoming the Barriers to Production-Ready Machine Learning Workflows

Josh Bloom    Henrik Brink
Lorica’s “Data Science Workflow”
Real-World Data Science = Optimization over this full Workflow
Data Science Optimization Space

Interpretability

Accuracy

Dimensionality
3 large + ~8 compact

Implementability
Our Background ...
“Data-Driven Scientists”

- Built & Deployed Real-time ML framework, discovering >10,000 events in > 10 TB of imaging → 50+ journal articles
- Built Probabilistic Event classification catalogs with innovative active learning
- Collective over 350 refereed journal articles including ML & timeseries analysis
Accuracy

Scalar proxies

- RMSE
- RMSLE
- [adjusted] $R^2$
- ...

cf. sklearn.metrics

Evaluation Metric: What’s the essence of what I care about?

$R^2 = 0.91$
$\text{RMSE} = 692.3$
$\text{Pearson R} = 0.96$
Accuracy

Evaluation Metric: What’s the essence of what I care about?

- FPR = 0.011
- FNR = 0.043

AUC = 0.996

Prob. = 0.5
Evaluation Metric: *What’s the essence of what I care about?*

which classifier is best? depends...
Evaluation Metric: What’s the essence of what I care about?

42-dimensional feature space

Some ML algorithms just do better
Accuracy

More Data (Dimensions) is better, but Protect Against Curse of Dimensionality

Performance Improvement

- RB2 all features
- RB2 optimal features
- RB1 features, RB2 training set
- RB1 features, old training set
More Data (Dimensions) is better, but Protect Against Curse of Dimensionality

"More data beats clever algorithms but better data beats more data."
- Peter Norvig
Accuracy

model 1 building + validation on historical data

Testing Set & Continuous (Streaming)
Testing & Model Updates

Model # in production

Date

actual value  good prediction  “bad” prediction
ML Algorithmic Trade-Off

Interpretability

High

Low

Accuracy

High

Low

Lasso
Linear/Logistic Regression
Decision Trees
Naive Bayes
Bagging
Random Forest®

Deep Learning

Nearest Neighbors
Splines
Nearest Neighbors
Gaussian/Dirichlet Processes
Neural Nets
SVMs

Warning
Unscientific & opinionated!

* on real-world data sets

Random Forest is a trademark of Salford Systems, Inc.
Interpretability
Consider a nonlinear system of equations:

\[
\begin{cases}
3x_1 - \cos(x_2 x_3) - \frac{3}{2} = 0 \\
4x_1^2 - 625x_2^2 + 2x_2 - 1 = 0 \\
\exp(-x_1 x_2) + 20x_3 + \frac{10\pi - 3}{3} = 0
\end{cases}
\]

suppose we have the function

\[
G(x) = \begin{bmatrix}
3x_1 - \cos(x_2 x_3) \\
4x_1^2 - 625x_2^2 + 2x_2 \\
\exp(-x_1 x_2) + 20x_3 + \frac{10\pi - 3}{3}
\end{bmatrix}
\]

where

\[
x = \begin{bmatrix}
x_1 \\
x_2 \\
x_3
\end{bmatrix}
\]

and the objective function

\[
F(x) = \frac{1}{2} G^T(x)G(x)
\]

\[
= \frac{1}{2} ((3x_1 - \cos(x_2 x_3) - \frac{3}{2})^2 + \ldots)
\]

With initial guess

\[
x^{(0)} = \begin{bmatrix}
x_1^{(0)} \\
x_2^{(0)} \\
x_3^{(0)}
\end{bmatrix} = \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

We know that

\[
x^{(1)} = x^{(0)} - \gamma_0 \nabla F(x^{(0)})
\]

where

\[
\nabla F(x^{(0)}) = J_G(x^{(0)})^T G(x^{(0)})
\]

The Jacobian matrix \(J_G(x^{(0)})\)

\[
J_G = \begin{bmatrix}
3 & \sin(x_2 x_3) & \sin(x_2 x_3) \\
8x_1 & -1250x_2 + 2 & 0 \\
-x_2 \exp(-x_1 x_2) & -x_1 \exp(-x_1 x_2) & 20
\end{bmatrix}
\]

Then evaluating these terms at \(x^{(0)}\)

\[
J_G (x^{(0)}) = \begin{bmatrix}
3 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 20
\end{bmatrix}
\]

and

\[
G(x^{(0)}) = \begin{bmatrix}
-2.5 \\
-1 \\
10.472
\end{bmatrix}
\]

So that

\[
x^{(1)} = 0 - \gamma_0 \begin{bmatrix}
-7.5 \\
-2 \\
209.44
\end{bmatrix}
\]

and

\[
F (x^{(0)}) = 0.5((-2.5)^2 + (-1)^2 + (10.472)^2) = 58.456
\]

Now a suitable \(\gamma_0\) must be found such that \(F(x^{(1)}) \leq F(x^{(0)})\). This can be done with algorithms. One might also simply guess \(\gamma_0 = 0.001\) which gives

\[
x^{(1)} = \begin{bmatrix}
0.0075 \\
0.002 \\
-0.20944
\end{bmatrix}
\]

evaluating at this value,

\[
F (x^{(1)}) = 0.5((-2.48)^2 + (-1.00)^2 + (6.28)^2) = 23.306
\]
How does the model work?

Consider a nonlinear system of equations:
\[
\begin{align*}
3x_1 - \cos(x_2x_3) - \frac{3}{2} &= 0 \\
4x_1^2 - 625x_2^2 + 2x_2 - 1 &= 0 \\
\exp(-x_1x_2) + 20x_3 + \frac{10\pi - 3}{3} &= 0
\end{align*}
\]

suppose we have the function
\[
G(x) = 
\begin{bmatrix}
3x_1 - \cos(x_2x_3) \\
4x_1^2 - 625x_2^2 + 2x_2 \\
\exp(-x_1x_2) + 20x_3 + \frac{10\pi - 3}{3}
\end{bmatrix}
\]

where
\[
x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}
\]

and the objective function
\[
F(x) = \frac{1}{2} G^T(x)G(x) = \frac{1}{2} \left( (3x_1 - \cos(x_2x_3) - \frac{3}{2})^2 + \right.
\]

With initial guess
\[
x^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\]

We know that
\[
x^{(1)} = x^{(0)} - \gamma_0 \nabla F(x^{(0)})
\]

where
\[
\nabla F(x^{(0)}) = J_G(x^{(0)})^T G(x^{(0)})
\]

The Jacobian matrix
\[
J_G(x^{(0)}) = \begin{bmatrix}
3 & \sin(x_2x_3)x_3 & \sin(x_2x_3)x_2 \\
8x_1 & -1250x_2 + 2 & 0 \\
-x_2 \exp(-x_1x_2) & -x_1 \exp(-x_1x_2) & 20
\end{bmatrix}
\]

Then evaluating these terms at \(x^{(0)}\)
\[
J_G \left( x^{(0)} \right) = \begin{bmatrix}
3 & 0 & 0 \\
0 & 2 & 0 \\
0 & 0 & 20
\end{bmatrix}
\]

and
\[
G(x^{(0)}) = \begin{bmatrix}
-2.5 \\
-1 \\
10.472
\end{bmatrix}
\]

So that
\[
x^{(1)} = 0 - \gamma_0 \begin{bmatrix} -7.5 \\ -2 \\ 209.44 \end{bmatrix}
\]

and
\[
F \left( x^{(0)} \right) = 0.5((-2.5)^2 + (-1)^2 + (10.472)^2) = 58.456
\]

Now a suitable \(\gamma_0\) must be found such that \(F(x^{(1)}) \leq F(x^{(0)})\). This can be done with algorithms. One might also simply guess \(\gamma_0 = 0.001\) which gives
\[
x^{(1)} = \begin{bmatrix} 0.0075 \\ 0.002 \\ -0.20944 \end{bmatrix}
\]

evaluating at this value,
\[
F \left( x^{(1)} \right) = 0.5((-2.48)^2 + (-1.00)^2 + (6.28)^2) = 23.306
\]
Why do I get these answers?

Interpretability

e.g., Credit score

Sample FICO® Scoring Model

<table>
<thead>
<tr>
<th>Category</th>
<th>Characteristic</th>
<th>Attributes</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payment History</td>
<td>Number of months since the most recent derogatory public record</td>
<td>0 – 5</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 – 11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 – 23</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24+</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No public record</td>
<td>55</td>
</tr>
<tr>
<td>Outstanding Debt</td>
<td>Average balance on revolving trades</td>
<td>12 – 23</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 – 47</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 or more</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No revolving trades</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 – 99</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100 – 499</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500 – 749</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td></td>
<td>750 – 999</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000 or more</td>
<td>75</td>
</tr>
<tr>
<td>Credit History Length</td>
<td>Number of months in file</td>
<td>Below 12</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 – 23</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 – 47</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 or more</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Below 12</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 – 23</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>24 – 47</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48 or more</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4+</td>
<td></td>
</tr>
<tr>
<td>Pursuit of New Credit</td>
<td>Number of inquiries in last 6 mos.</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4+</td>
<td>50</td>
</tr>
<tr>
<td>Credit Mix</td>
<td>Number of bankcard trade lines</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4+</td>
<td>50</td>
</tr>
</tbody>
</table>
Peering Inside the Black Box

Random Forest® model-level feature importance
Peering Inside the Black Box

Individual-level prediction feature importance

Probability of Default in 1 year: 76% [deny loan]

Driving factors:
- Credit history: 10 months 14%
- Outstanding debt: $1200 5%
- Inquiries in 6 months: 2 1%

e.g. microcredit application scorecard
Implementability

How long does it take to put the model into production? At what cost?
Implementability

Leaderboard data from Kaggle & Netflix

Netflix

$>$50k Prize

$<$50k Prize

many teams get within
~few % of optimum

so which is easier to
put into production?
“We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.”

Xavier Amatriain and Justin Basilico (April 2012)
The divide between data science & production
Implementability

Treat Machine Learning Deployment as you would Software

- Continuous Deployment
- RESTful API
- Language bindings
- Security
- SLA
Integration

Connect data

Consume predictions
Scalability & Speed

Implementability

Micro-scaling
Fast, efficient use of memory hierarchy

Horizontally scalable data processing

Spark
Hadoop
Machine-Learning, Data Science Workflow is an Optimization in *many* dimensions
We are Hiring!

- Full-stack developers
  - Javascript, Python, Spark/Shark
- Front end developers
- DevOps engineers
- C++ engineers
  - C++ template metaprogramming
- Data scientists
  - Python, Deep NN, ML expertise

jobs@wise.io
http://wise.io/jobs/