Friends don’t let friends deploy Black-Box models
The importance of transparency in Machine Learning

Rich Caruana
Microsoft Research
Friends Don’t Let Friends Deploy Black-Box Models
The Importance of Transparency in Machine Learning

Rich Caruana
Microsoft Research

Joint Work with
Ran Gilad-Bachrach, Yin Lou, Sarah Tan, Johannes Gehrke
Paul Koch, Marc Sturm, Noemie Elhadad

Thanks to
Greg Cooper MD PhD, Mike Fine MD MPH, Eric Horvitz MD PhD
Nick Craswell, Tom Mitchell, Jacob Bien, Giles Hooker, Noah Snavely
When is it Safe to Use Machine Learning?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

is it safe to deploy this model and use on real patients?
is high accuracy on test data enough to trust a model?
When is it Safe to Use Machine Learning?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

is it safe to deploy this model and use on real patients?

is high accuracy on test data enough to trust a model?
When is it Safe to Use Machine Learning?

- data for 1M patients
- 1000’s great clinical features
- train state-of-the-art machine learning model on data
- accuracy looks great on test set: AUC = 0.95

- is it safe to deploy this model and use on real patients?
- is high accuracy on test data enough to trust a model?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital (≈10% of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
  - ...  

Most accurate ML method: **multitask neural nets** (shallow MTL nets)

Safe to use neural nets on patients?

No — we used logistic regression instead…
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital (≈10% of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
- ...

Most accurate ML method: **multitask neural nets** (shallow MTL nets)

Safe to use neural nets on patients?
- No — we used logistic regression instead...
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- **LOW Risk:** outpatient: antibiotics, call if not feeling better
- **HIGH Risk:** admit to hospital ($\approx 10\%$ of pneumonia patients die)

One goal was to compare various ML methods:
- logistic regression
- rule-based learning
- k-nearest neighbor
- neural nets
- Bayesian methods
- hierarchical mixtures of experts
- ...

Most accurate ML method: **multitask neural nets** (shallow MTL nets)

Safe to use neural nets on patients?

No — we used logistic regression instead...
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: \( \text{HasAsthma}(x) \implies \text{LessRisk}(x) \)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering \( \text{HasAsthma}(x) \)… was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: \texttt{HasAsthma}(x) \implies \texttt{LessRisk}(x)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering \texttt{HasAsthma}(x)... was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: HasAsthma(x) => LessRisk(x)

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering HasAsthma(x)... was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Motivation: Predicting Pneumonia Risk Study (mid-90’s)

- RBL learned rule: $\text{HasAsthma}(x) \implies \text{LessRisk}(x)$

- True pattern in data:
  - asthmatics presenting with pneumonia considered very high risk
  - receive aggressive treatment and often admitted to ICU
  - history of asthma also means they often go to healthcare sooner
  - treatment lowers risk of death compared to general population

- If RBL learned asthma is good for you, NN probably did, too
  - if we use NN for admission decision, could hurt asthmatics

- Key to discovering $\text{HasAsthma}(x)$... was intelligibility of rules
  - even if we can remove asthma problem from neural net, what other "bad patterns" don’t we know about that RBL missed?
Lessons

- Risky to use data for purposes it was not designed for
- Most data has unexpected landmines
- Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
- Our approach is to make the learned models as intelligible as possible
- Must be able to understand models used in healthcare
- Also true for race and gender bias where the bias is in the training data
All we need is an accurate, intelligible model
Problem: The Accuracy vs. Intelligibility Tradeoff
Problem: The Accuracy vs. Intelligibility Tradeoff

![Diagram showing the tradeoff between Accuracy and Intelligibility for various machine learning models.](image)

Models:
- Boosted Trees
- Random Forests
- Neural Nets
- Logistic Regression
- Naive Bayes
- Single Decision Tree
- Decision Lists

Rich Caruana (Microsoft Research)
Faculty Summit: Intelligible Models
July 18, 2017
Model Space from Simple to Complex

- **Linear Model**: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)
- **Additive Model**: \( y = f_1(x_1) + \ldots + f_n(x_n) \)
- **Additive Model with Interactions**: \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \)
- **Full Complexity Model**: \( y = f(x_1, \ldots, x_n) \)
Model Space from Simple to Complex

- **Linear Model:** $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$

- **Additive Model:** $y = f_1(x_1) + ... + f_n(x_n)$

- **Additive Model with Interactions:** $y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + ...$

- **Full Complexity Model:** $y = f(x_1, ..., x_n)$
Model Space from Simple to Complex

- Linear Model: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)
- Additive Model: \( y = f_1(x_1) + \ldots + f_n(x_n) \)
- Additive Model with Interactions: \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \)
- Full Complexity Model: \( y = f(x_1, \ldots, x_n) \)
Model Space from Simple to Complex

- **Linear Model**: \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n \)

- **Additive Model**: \( y = f_1(x_1) + \ldots + f_n(x_n) \)

- **Additive Model with Interactions**: \( y = \sum_i f_i(x_i) + \sum_{ij} f_{ij}(x_i, x_j) + \sum_{ijk} f_{ijk}(x_i, x_j, x_k) + \ldots \)

- **Full Complexity Model**: \( y = f(x_1, \ldots, x_n) \)
Add ML-Steroids to old Stats Method: GAMs → GA2Ms

- Generalized Additive Models (GAMs)
  - Developed at Stanford by Hastie and Tibshirani in late 80’s
  - Regression: \( y = f_1(x_1) + ... + f_n(x_n) \)
  - Classification: \( \logit(y) = f_1(x_1) + ... + f_n(x_n) \)
  - Each feature is “shaped” by shape function \( f_i \)

T. Hastie and R. Tibshirani. 
*Generalized additive models.* 
Skip all algorithmic details and jump to one result
What GA2Ms Learn About Pneumonia Risk (POD) as a Function of Age

![Graph showing Pneumonia Risk Score and Density as a function of age.](image-url)
Intelligible model also learned:
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used
- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Intelligible model also learned:
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used
- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Intelligible model also learned:
- Has_Asthma $\Rightarrow$ lower risk
- History of chest pain $\Rightarrow$ lower risk
- History of heart disease $\Rightarrow$ lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used
- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Intelligible model also learned:

- Has_Asthma => lower risk
- History of chest pain => lower risk
- History of heart disease => lower risk

Good we didn’t deploy neural net back in 1995
But can understand, edit and safely deploy intelligible GA2M model
Intelligible/transparent model is like having a magic pair of glasses

Model correctness depends on how model will be used

- this is a good model for health insurance providers
- but needs to be repaired to use for hospital admissions

Important: Must keep potentially offending features in model!
Interpretable GAM model class is a good match for homomorphic encryption

Interpretable models may help preserve data privacy

Potential issue with transparency vs. encryption
Interpretable GAM model class is a good match for homomorphic encryption

Interpretable models may help preserve data privacy

Potential issue with transparency vs. encryption
Why GAMs Are Good For Homomorphic Encryption

[Diagram showing a neural network with inputs at the bottom and outputs at the top.]
Why GAMs Are Good For Homomorphic Encryption

Original 2nd-degree Polynomial Fit

Rich Caruana (Microsoft Research)
Faculty Summit: Intelligible Models
July 18, 2017 19 / 28
### Why GAMs Are Good For Homomorphic Encryption

Poly-GAMs are competitive models

<table>
<thead>
<tr>
<th>Model</th>
<th>Pneumonia Test AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.8432</td>
</tr>
<tr>
<td>Random Forests</td>
<td>0.8460</td>
</tr>
<tr>
<td>LogitBoost</td>
<td>0.8493</td>
</tr>
<tr>
<td>Intelligible GAM</td>
<td>0.8542</td>
</tr>
<tr>
<td>Intelligible GA²M</td>
<td>0.8576</td>
</tr>
<tr>
<td>SEAL Poly-GAM</td>
<td>0.8502</td>
</tr>
</tbody>
</table>
- Interpretable GAM model class is a good match for homomorphic encryption
- Interpretable models may help preserve data privacy
- Potential issue with transparency vs. encryption
Why the Simplicity of GAM Models Might Be Good For Preserving Privacy

Complex Black-Box Deep Net
Why the Simplicity of GAM Models Might Be Good For Preserving Privacy

Complex Black-Box Deep Net

Transparent GAM Model
Interpretable GAM model class is a good match for homomorphic encryption

Interpretable models may help preserve data privacy

Potential issue with transparency vs. encryption
Potential Problem with Encryption if Model Remains Hidden
Potential Problem with Encryption if Model Remains Hidden
- ML trained on data will learn the biases in that data
  - ML for resume processing will learn gender bias
  - ML for recidivism prediction will learn race bias
  - ...
- Remember, the bias is in the data!
- How to deal with bias using intelligible models:
  - keep bias features in data when model is trained
  - remove what was learned from bias features after training
- If offending variables are eliminated prior to training
  - often can’t tell you have a problem
  - makes it harder to correct the problem
- EU General Data Protection Regulation (goes into effect 2018):
  - Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
ML trained on data will learn the biases in that data
- ML for resume processing will learn gender bias
- ML for recidivism prediction will learn race bias
- ...

Remember, the bias is in the data!

How to deal with bias using intelligible models:
- keep bias features in data when model is trained
- remove what was learned from bias features after training

If offending variables are eliminated prior to training
- often can’t tell you have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
ML trained on data will learn the biases in that data
  - ML for resume processing will learn gender bias
  - ML for recidivism prediction will learn race bias
  - ...

Remember, the bias is in the data!

How to deal with bias using intelligible models:
  - keep bias features in data when model is trained
  - remove what was learned from bias features after training

If offending variables are eliminated prior to training
  - often can’t tell you have a problem
  - makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
  - Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
ML trained on data will learn the biases in that data
- ML for resume processing will learn gender bias
- ML for recidivism prediction will learn race bias
- ...

Remember, the bias is in the data!

How to deal with bias using intelligible models:
- keep bias features in data when model is trained
- remove what was learned from bias features after training

If offending variables are eliminated prior to training
- often can’t tell you have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
ML trained on data will learn the biases in that data:
- ML for resume processing will learn gender bias
- ML for recidivism prediction will learn race bias
- ...

Remember, the bias is in the data!

How to deal with bias using intelligible models:
- keep bias features in data when model is trained
- remove what was learned from bias features after training

If offending variables are eliminated prior to training:
- often can’t tell you have a problem
- makes it harder to correct the problem

EU General Data Protection Regulation (goes into effect 2018):
- Article 9 makes it more difficult to use personal data revealing racial or ethnic origin and other “special categories”
Summary

- High accuracy on test set is not always enough — can be very misleading
- There are land mines hidden in most real data — need magic glasses to see landmines
- In some domains (e.g., healthcare) it’s critical to understand model before deploying it
- Correctness depends on how model will be used — data/model not inherently right/wrong
- GA2Ms give us accuracy and intelligibility at same time
- Important to keep potentially offending variables in model so bias can be detected and then removed after training
- Deep Learning is great — but sometimes we have to understand what’s in the black box
- GA2Ms can help insure privacy protection because models are so simple
- Poly-GAMs can be good for encryption, but the model needs to be visible to someone
Thank you