Interpretable Machine Learning

Patrick Hall
Dec. 11, 2016

H2O.ai

- I have said before that machine learning is uninterpretable – but I was wrong
- Nonlinear, nonpolynomial models
- ML models capture high degree interactions
- Allow a variables impact on the model predictions and interactions with other variables to change in complex ways over the variable's domain
- Interpretability is about trust and understanding – ways to increase trust and understanding

- This talk needs a white board

Contents

Part 1: Seeing all of your data
- Correlation graphs
- 2-D projections
- Glyphs

Part 2: Using machine learning in regulated industry
- OLS regression alternatives
- Build toward ML model benchmarks
- ML in traditional analytics processes
- Small, interpretable ensembles

Part 3: Understanding complex ML models
- Surrogate models
- LIME
- Maximum activation analysis
- Constrained neural networks
- Variable importance measures
- Partial dependence plots
- TreeInterpreter
- Residual analysis

H2O.ai

Part 1: Seeing all of your data

Need 2-D - even 3-D or VR is too cumbersome for quick analysis

Correlation Graphs

The nodes of this graph are the variables in a data set. The weights between the nodes are defined by the absolute value of their pairwise Pearson correlation.

How does it increase understanding? By visually displaying relationships between columns
How does it increase trust? Trust is increased if known relationships are displayed and/or correct modeling results are reflected in the graph - also if patterns are stable or change predictably over time

To create:
- calculate Pearson correlation between columns/variables
- build undirected graph where each node is a column/variable
- connection weights between nodes are defined by Pearson correlation absolute values; weights below a certain threshold are not displayed
- node size is determined by number of connections (node degree)
- node color is determined by a graph communities calculation
- node position is defined by a graph force field algorithm

Free graph software: https://gephi.org/

2-D projections

786 dimensions to 2 dimensions
with PCA

786 dimensions to 2 dimensions
with autoencoder network

H_2O.ai

Source: http://www.cs.toronto.edu/~hinton/absps/science_som.pdf

How does it increase understanding? If possible all records are shown in a single 2-D plot
How does it increase trust? Trust is increased if known or expected structures (i.e. clusters, outliers, hierarchy) are preserved and displayed in 2-D plots - also if patterns are stable or change predictably over time

There are numerous types of useful projections (or “embeddings”):
- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)
- t-SNE (t-distributed Stochastic Neighbor Embedding)
- Autoencoder networks

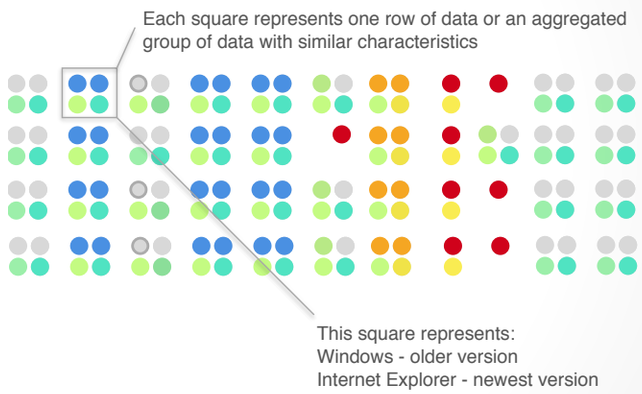
Here PCA and autoencoders are shown - better scalability than many other methods

Autoencoder projections can be augmented by training clusters in the original high dimensional data before projecting into lower dimensional space - look for clusters to be preserved in 2-D projections and confirm cluster relationships are reasonable on 2-D plots. For instance older, richer customers should be relatively far from younger, less affluent customers.

Glyphs

Variables and their values can be represented by small pictures with certain attributes, called “glyphs”.

Here the four variables are represented by their position in a square and their values are represented by a color.



H2O.ai

How does it increase understanding? Glyphs are typically much easier to digest than just staring at plain rows of data

Part 2: Using ML in regulated industry

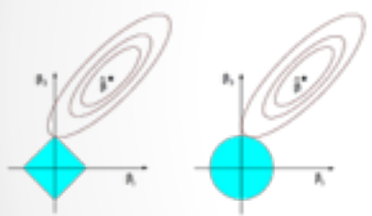
The inherent trade-off between accuracy and interpretability in predictive modeling can be a catch-22 for analysts and data scientists working in regulated industries. Professionals in the regulated verticals of banking and insurance often feel locked into using traditional, linear modeling techniques to create their predictive models. This is mainly due to strenuous regulatory and documentation requirements. As machine learning becomes more mainstream, the forces of innovation and competition often drive these same analysts and data scientists to break out of the mold and try new algorithms with more predictive capacity. Such algorithms for machine learning include gradient boosted ensembles, neural networks, and random forests, among many others. These algorithms are typically more accurate for predicting nonlinear, faint, or rare phenomena. Unfortunately, more accuracy almost always means less interpretability, and interpretability is crucial for documentation and regulation processes.

Due to their inscrutable inner-workings, many machine learning algorithms have been labeled “black box” models. What makes these models accurate is what makes their predictions difficult to understand: they are very complex. This is a fundamental trade-off. So how can you improve the accuracy of more traditional linear models while still retaining some degree of interpretability?

https://www.oreilly.com/ideas/predictive-modeling-striking-a-balance-between-accuracy-and-interpretability

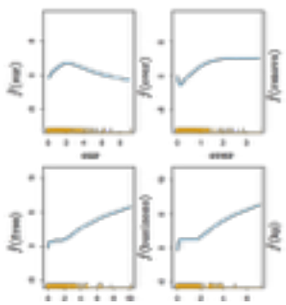
OLS regression alternatives

Penalized Regression



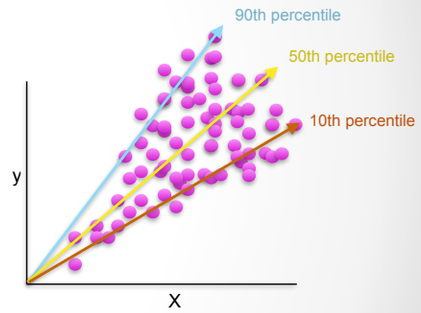
- Fewer assumptions
- Well suited for **N << p**
- No multiple comparison issues during variable selection
- Preserves interpretability by selecting a small number of variables (L1 penalty)

Generalized Additive Models



- Fit linear terms to certain variables
- Fit nonlinear splines to other variables
- Hand-tune a trade-off between interpretability and accuracy

Quantile Regression



- Fit an interpretable linear model to different percentiles of training data
- Find different sets of drivers across percentiles of an entire customer market or portfolio of accounts

H2O.ai

Source: https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

How does it increase trust and understanding?
It is the same understandable, trust worthy models used in different ways

Penalized regression:
- Penalized regression techniques are particularly well-suited for wide data.
- Avoid the multiple comparison problem that can arise with stepwise variable selection.
- They can be trained on datasets with more columns than rows.
- They preserve interpretability by selecting a small number of original variables for the final model using L1 regularization
- Nearly always predictive

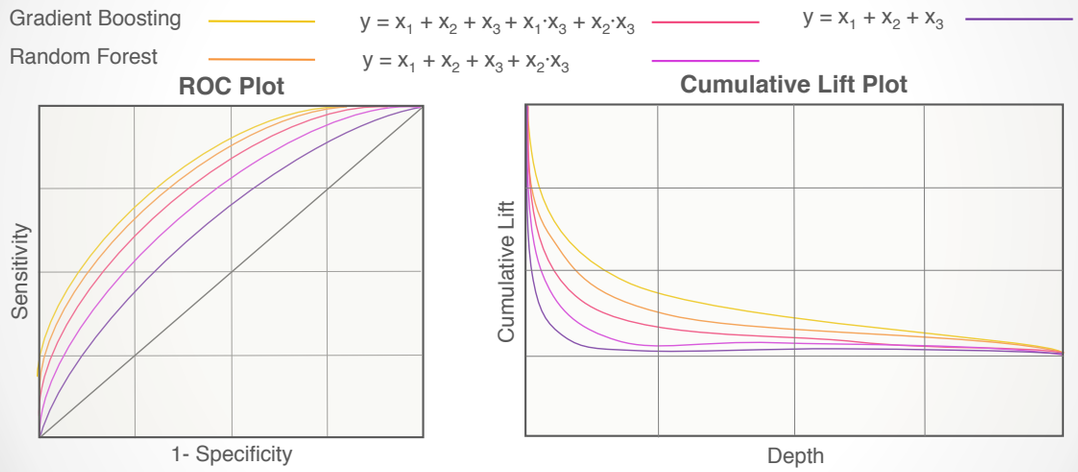
(L1 also works to increase interpretability across many different types of models.)

GAMs:
- Generalized additive models fit linear terms to certain variables and nonlinear splines to other variables
- Allowing you to hand-tune a trade-off between interpretability and accuracy
- Can be predictive based on the application

Quantile regression:
- Fit a traditional, interpretable linear model to different percentiles of your training data
- Allowing you to find different sets of variables for modeling different behaviors across a customer market or portfolio of accounts
- more inferential than predictive

https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

Build toward ML model benchmarks



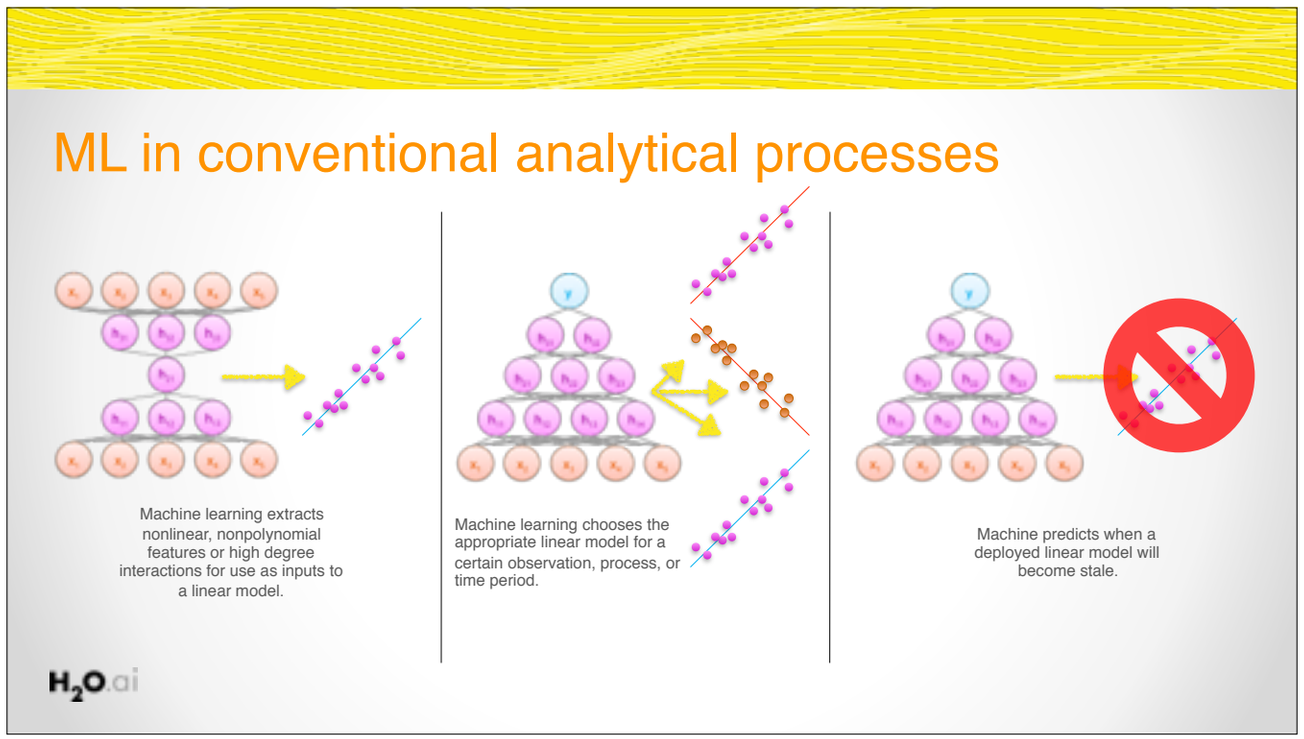
Machine learning models typically incorporate a large number of implicit variable interactions and easily fit nonlinear, non-polynomial patterns in data. If a traditional regression model is much less accurate than a machine learning model, the traditional regression model may be missing important interactions or a piecewise modeling approach maybe necessary.

H_2O.ai

How does it increase trust and understanding?
It helps us make our understandable, trust worthy models more accurate

- Machine learning models often take into consideration a large number of implicit variable interactions
- If your regression model is much less accurate than your ML model, you've probably missed some important interaction(s)
- Decision trees area great way to see the potential interactions
- Important interactions may only be occurring at certain values of certain variables

ML models intrinsically allow:
- high degree interactions between input variables - include 2nd, 3rd degree interactions to approximate
- nonlinear, nonpolynomial behavior across the domain of a single input variable - use piecewise models to approximate



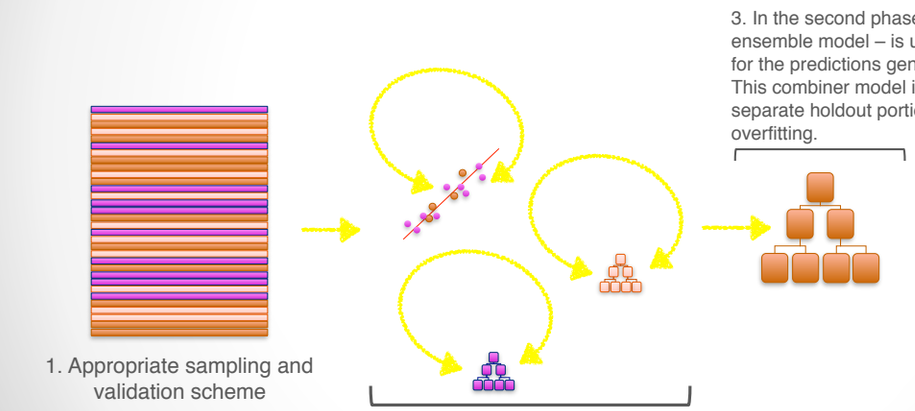
How does it increase trust and understanding?
It helps us make our understandable, trust worthy models more accurate and helps us use them more efficiently

- Introduce nonlinear predictors into the model
- (Predictors that capture more complex, nonlinear, nonpolynomial relationships)

Based on past model performance data:
- Use an ML model as a gate to pick which linear model to use
- Use a machine learning model to predict when traditional deployed models need to be retrained or replaced before their predictive power lessen

Small interpretable ensembles

Ensembles can optimally combine the results of multiple interpretable models



1. Appropriate sampling and validation scheme

2. In the first phase of training, several different models are trained, validated and/or cross-validated on a portion of the training data.

3. In the second phase of training, another model – often an ensemble model – is used to determine optimal weightings for the predictions generated in the first phase of training. This combiner model is often trained and validated using a separate holdout portion of the training data to prevent overfitting.

H2O.ai

How does it increase trust and understanding?
It allows us to boost the accuracy of traditional trustable models without sacrificing too much interpretability
It increases trust if models compliment each other in expected ways, e.g.
A logistic regression model that is good at rare events slightly increases a good decision tree model that is not good at rare events in the presence of rare events

- Probably the most important recent breakthrough in machine learning aside from deep learning
- Combining predictions between a handful of good, but different, models often results in better predictions
- Ex: train an interpretable regression model and an interpretable decision tree and average their predictions
- Different over sampling in each model very useful for rare events
- Hill climbing
- Stacking:
    - A linear model is often used to optimally weight the predictions of several different models that are then assembled together
    - Cross validation

Van der Laan 2007 reference: http://biostats.bepress.com/ucbbiostat/paper222/

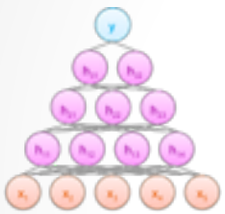
Part 3: Understanding complex ML models

H2O.ai

For those who can use any type of ML model, this is how can they explain their behavior.

Probably a combination of these techniques works best.

Surrogate models

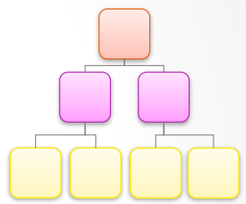


y	x_1	x_2	x_3
1	2.54	1.65	0.02
0	1.14	0.70	0.82
1	0.99	0.51	2.11
⋮	⋮	⋮	⋮

Target vector Input vectors

First fit a complex machine learning model to your training data.

Then train a single decision tree on the original training data, but instead of using the actual target in the training data, use the predictions of the more complex algorithm as the target for this single decision tree.*



y	$\hat{y}$	x_1	x_2	x_3
1	0.98	2.54	1.65	0.02
0	0.32	1.14	0.70	0.82
1	0.85	0.99	0.51	2.11
⋮	⋮	⋮	⋮	⋮

Predicted target vector

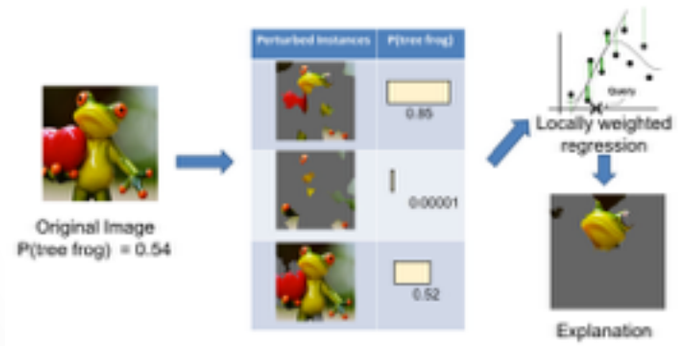
H2O.ai

How does it increase trust and understanding?
It helps us understand the inner workings of a complex system
It increases trust if we can see the logic in the surrogate model matches our domain experience or expectation
It increases trust if the logic is stable under mild perturbations of the data

- Interpretable models used as a proxy to explain complex models
- For example:
    - Fit a complex machine learning model to your training data.
    - Then train a single decision tree on the original training data, but use the predictions of the more complex algorithm as the target for this single decision tree
    - This single decision tree will likely* be a more interpretable proxy you can use to explain the more complex machine learning model

* Few (possibly no?) theoretical guarantees that the surrogate model is highly representative of the more complex model

Local Interpretable Model-Agnostic Explanations (LIME)



Source: https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime

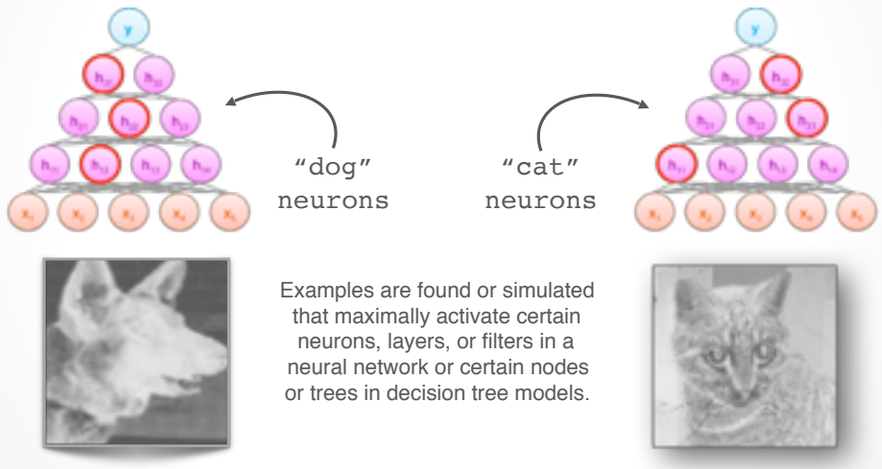
How does it increase trust and understanding?
It helps us understand the predictions made for key observations
It helps us understand the behavior of the model at local, important places no matter how complex the global model is
It increases trust because we can see how the model makes decisions for key observations
It increases trust if we see decisions being made about similar observations being made in similar ways

- Pick or simulate ‘marker’ records/examples
- Score them for probability or value of target with complex ML model
- Choose a ‘query’ record/example with a prediction to be explained
- Weight ‘marker’ records/examples closest to the query record/example
- Train an L1 regularized linear model on the data set of ‘marker’ records/examples
- The parameters of the linear model will help explain the prediction for the ‘query’ record/example

- Local surrogate model + more structured type of activation analysis
- You can include ‘marker’ records/examples in your training data
- For traditional analytics data, explanatory data samples could potentially be simulated – e.g. customers with highest, lowest, and median credit scores

https://www.oreilly.com/learning/introduction-to-local-interpretable-model-agnostic-explanations-lime
https://arxiv.org/pdf/1606.05386.pdf

Maximum activation analysis



Examples are found or simulated that maximally activate certain neurons, layers, or filters in a neural network or certain nodes or trees in decision tree models.

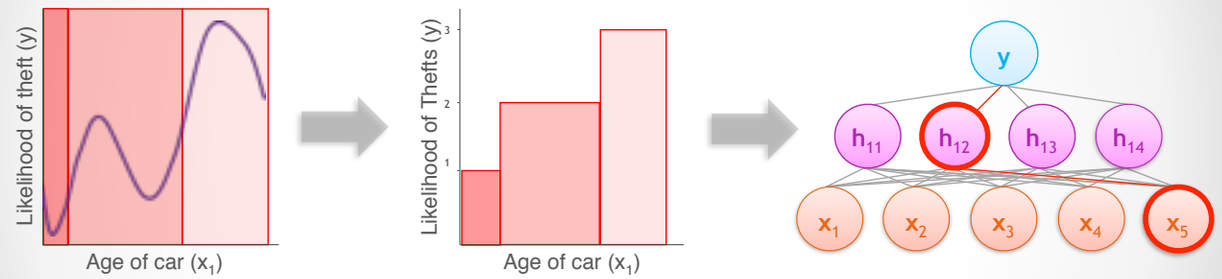
H2O.ai

How does it increase trust and understanding?
It increases understanding because it elucidates the structure of the model
(If we have dogs and cats in our data we would expect certain neurons to maximally learn certain visually features, i.e. dog nose neuron is activated for all dog picks, but not in cat pictures)
It increases understanding because see interactions when input units activate the same hidden unit consistently
It increases trust if we see stability in what units are activated for similar inputs
It increases trust if similar data points proceed through the model in the same way
It increases trust if interactions and structure match

- Which data creates the maximum output from certain neurons
- Which neurons create the maximum output for some archetypal data example
- You can include 'marker' records/examples in your training data

http://yosinski.com/deepvis
http://yosinski.com/media/papers/Yosinski__2015__ICML_DL__Understanding_Neural_Networks_Through_Deep_Visualization__.pdf

Constrained neural networks



Age of car is a nonnegative quantity, but is not monotonic wrt to car theft, the target in this example.

Through a binning scheme, age of car can be transformed to be monotonically increasing with the target.

When all inputs are nonnegative, monotonic wrt to the target, and model weights are constrained to be nonnegative it's easier to parse extra information from machine learning models.

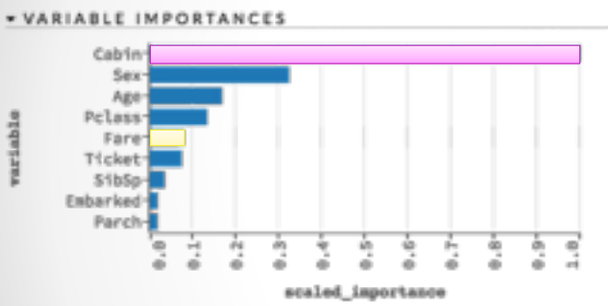
H2O.ai

How does it increase trust and understanding?
It increases understanding because it we can learn interactions, important variables, and the direction in which a input effects the predicted outcome
It increases trust if these are parsimonious with domain expertise or expectations
It increases trust if similar data points proceed through the model in the same way
It increases trust if interactions, important features, and direction of an input are consistent for similar data sets

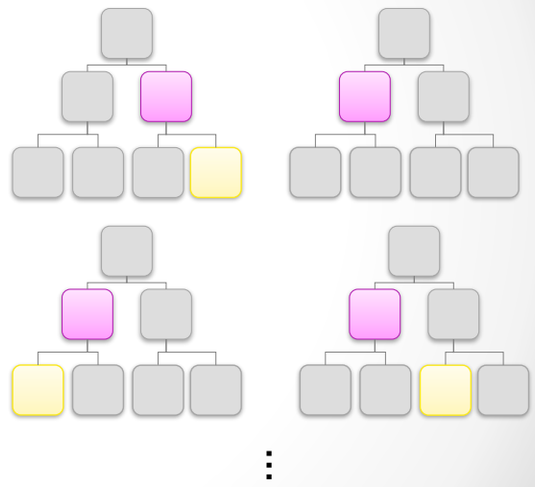
- Scale inputs to be non-negative
- Transform inputs such that their relationship with the target is monotonically increasing or decreasing
- Enables the human user to parse extra information from machine learning models:
    - In a neural network with only positive weights
    - For a binary classification task where the target value 1 indicates an event and the target value 0 indicates a non-event
    - All predictor variables are non-negative and monotonically increasing with respect to the target
    - Higher values of that predictor lead to increased occurrences of the target event
    - By following the maximum activation of neurons through the network it may even be possible to determine high-order interactions

- Binning does reduce the resolution of the information presented to the model during training
- But it can lead to better generalization (intricate patterns in the training data can be noise)
- Allows for elegant handling of outliers

Variable importance measures



Variable importance in tree-based models is often based on the amount splitting on a given variable increases predictive accuracy or node purity.



H2O.ai

How does it increase trust and understanding?
It increases understanding because we can learn important variables and their relative rank
It increases trust if these rankings match domain expertise or expectations
It increases trust if these ranks are repeatable in similar data

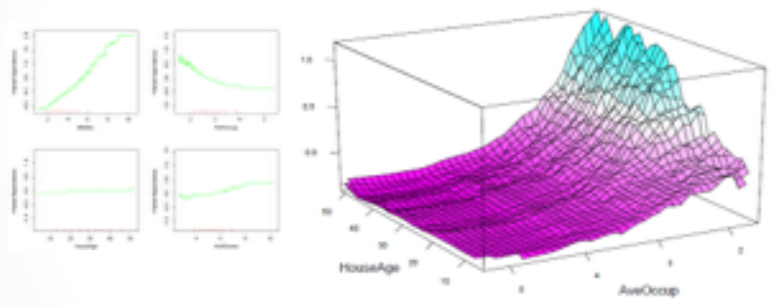
In Tree:
Split criterion change caused by an input for each node

In RF:
Split criterion change caused by an input for each node
Difference in OOB predictive accuracy when the predictor of interest is shuffled
(shuffling is seen as ‘zeroing out’ the effect of the variable in the trained model, because other variables are not shuffled)

In GBM:
Split criterion change caused by an input for each node

Simplistic variable importance measures can be biased toward larger scale variables or variables with a large number of categories

Partial dependence plots



HomeValue ~ MedInc + AveOccup + HouseAge + AveRooms

H2O.ai

Source: https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

How does it increase trust and understanding?
It increases understanding because we can see the behavior of individual inputs and their 2 way interactions
It increases trust if the displayed behavior is consistent with domain expertise and expectations
It increases trust if displayed behavior is repeatable

Images: Elements of Statistical Learning, https://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf, pg. 374

“Partial dependence tells us how the value of a variable influences the model predictions after we have averaged out the influence of all *other* variables. (For linear regression models, the resulting plots are simply straight lines whose slopes are equal to the model parameters.)”
- https://cran.r-project.org/web/packages/datarobot/vignettes/PartialDependence.html

Can be calculated efficiently for tree-based models, because of tree structure

Interesting variant: https://github.com/numeristical/introspective

TreeInterpreter

Tree interpreter decomposes decision tree and random forest predictions into bias (overall average) and component terms.

This slide portrays the decomposition of the decision path into bias and individual contributions for a simple decision tree.

For a random forest model, treeinterpreter simply prints a ranked list of the bias and individual contributions for a given prediction.

Prediction: 18.11 = 22.60 (trainset mean) - 2.64(loss from RM) - 5.04(loss from LSTAT) + 3.20(gain from NOX)

Source: http://blog.datadive.net/interpreting-random-forests/

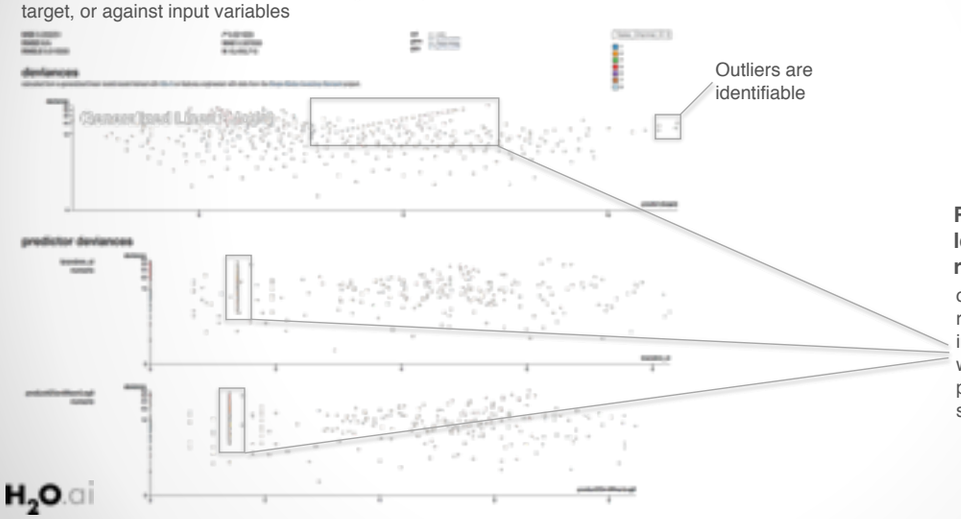
Currently only for sklearn decision tree and forest models.

How does it increase understanding? It allows for easy explanations of the internal mechanics of model
How does it increase trust? It increases trust if …
- internal mechanics represent known or expected phenomenon in the training data
- different decision paths lead to different results
- similar decision paths lead to similar results
- if model remains stable over time or over minor perturbations of training data

https://github.com/andosa/treeinterpreter

Residual analysis

Residuals can be plotted against the target, the predicted target, or against input variables



Residuals from a machine learning model should be randomly distributed

obvious patterns in residuals can indicate problems with data preparation or model specification

How does it increase understanding? Patterns in residuals can help elucidate patterns in the data that would otherwise be obscured by the curse of dimensionality, i.e. outliers, clusters, hierarchies, sparsity, etc.
How does it increase trust? If overall residuals are randomly distributed, this is a good indication that the ML model is fitting the data well. Obvious patterns in residuals could point to problems in model specification or data preparation that can be iteratively corrected by preprocessing data, building a model, and analyzing residuals

http://residuals.h2o.ai:8080/

Questions?