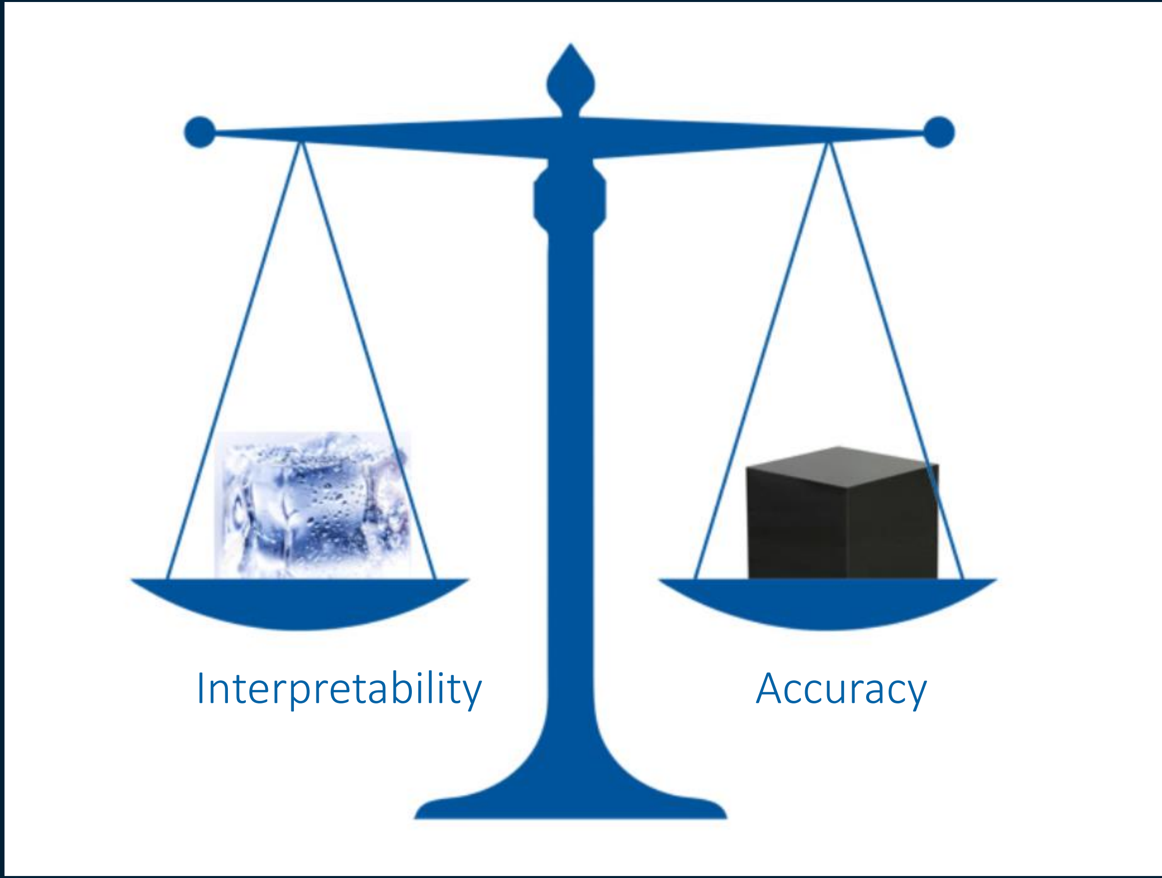
Model Interpretability

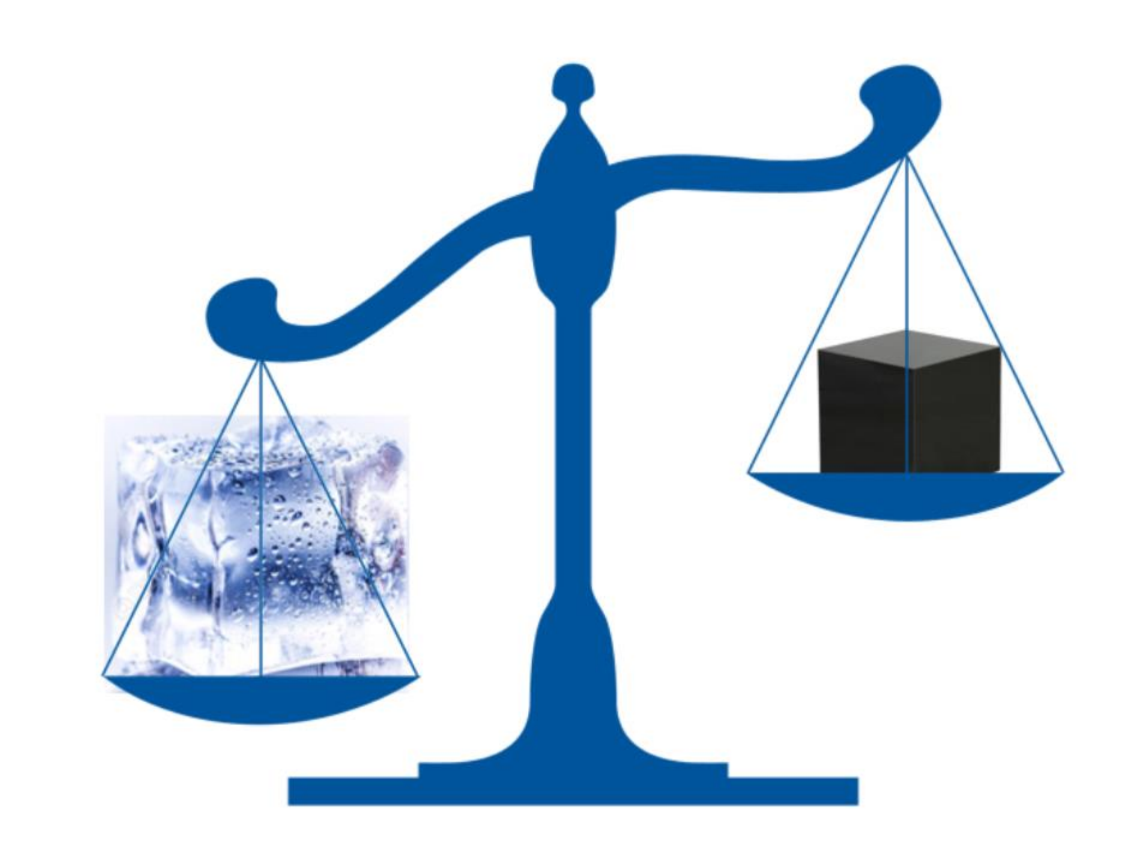
FANS Network Meeting | Data Science |June 2, 2021

Mathias Lanner SAS Institute

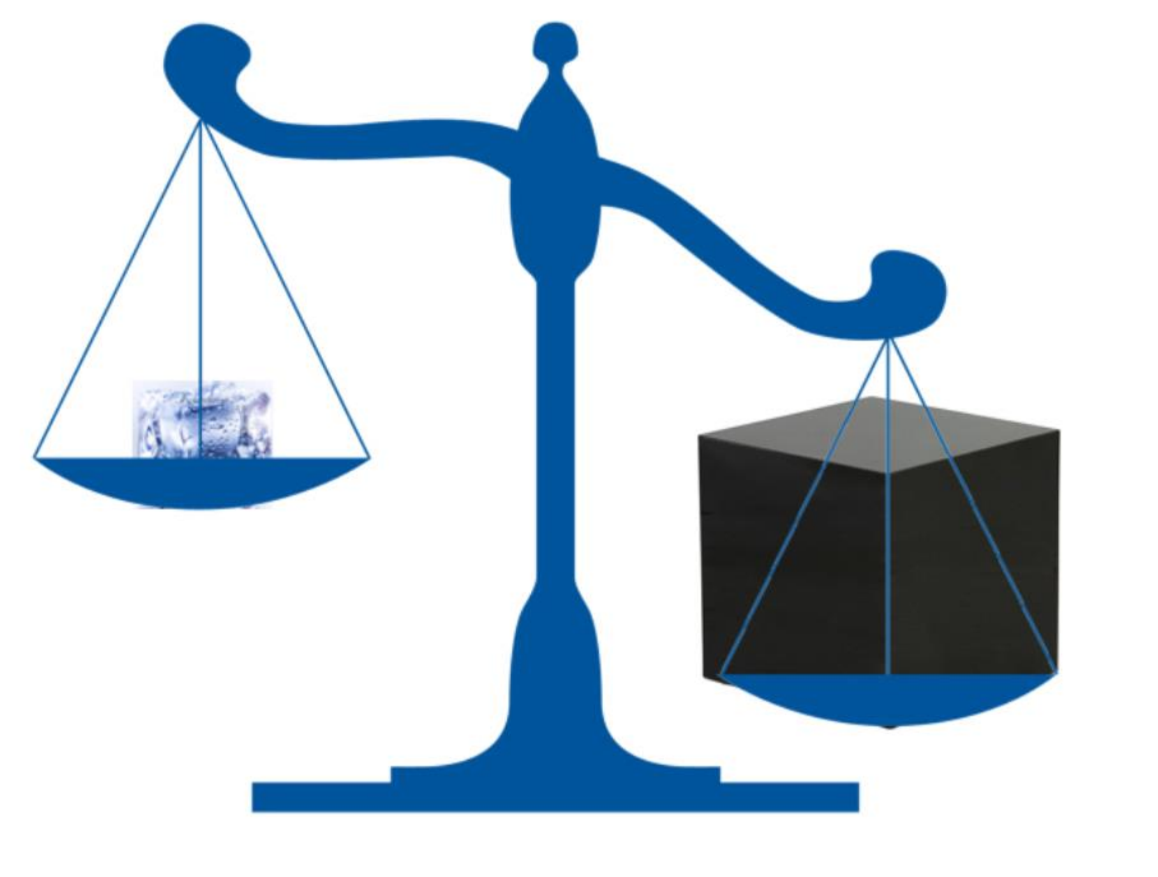
White Box- vs Black Box-models



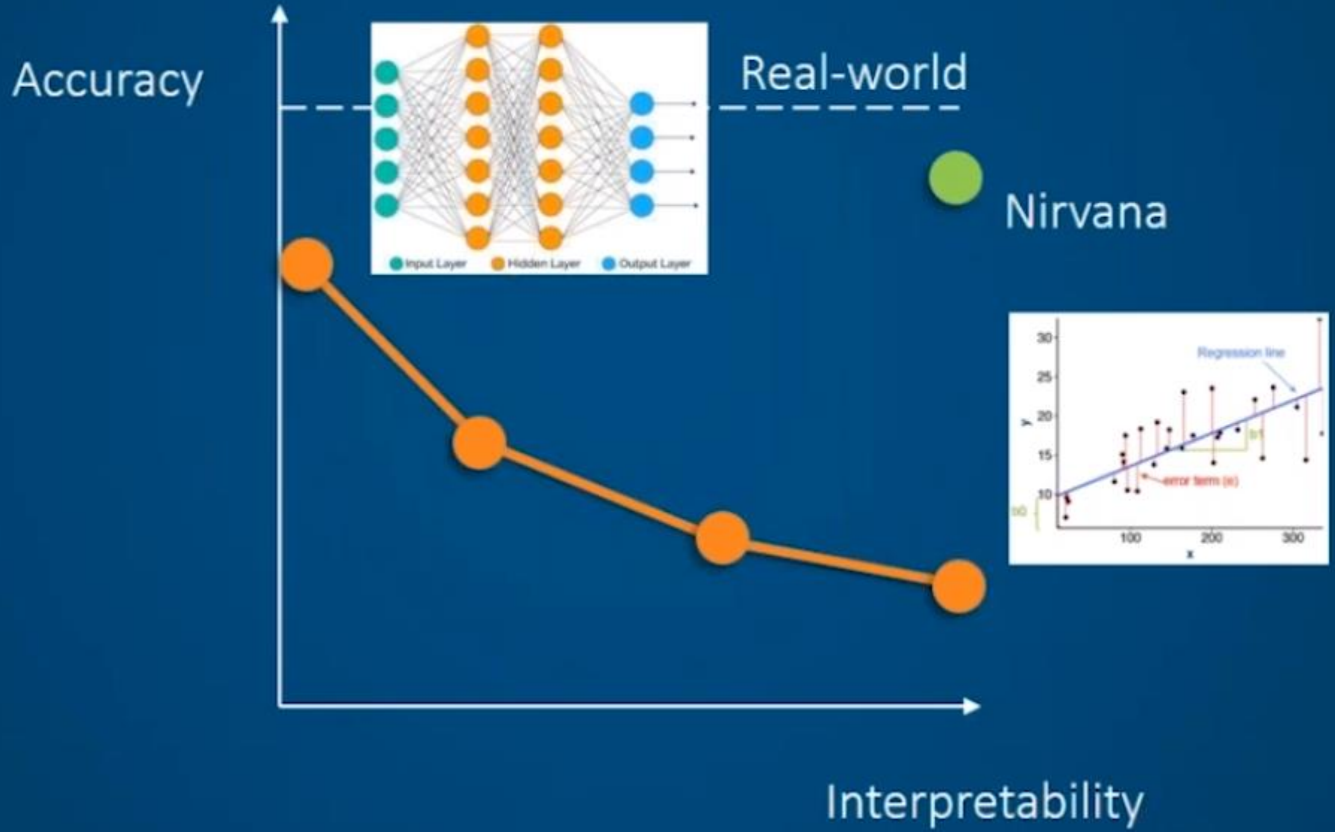
What do you need?



What do you need?



Accuracy vs Interpretability



Can you taste the difference?

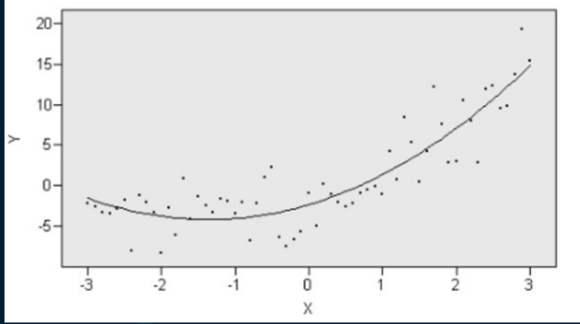
How does the ingrediencies impact the taste experience?

- Gin & Tonic

- Boeuf Bourguignon

Transparent "White Box"	Opaque "Black Box"
Regressions	Neural Networks
Decision Trees	Random Forests
Rules-Based	Gradient Boosting

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$



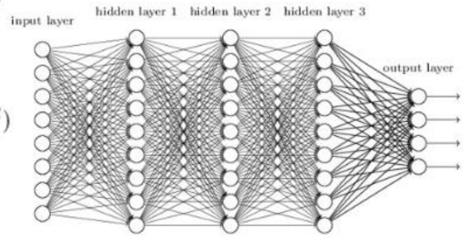
$$a^l_j = \sigma\left(\sum_k w^l_{jk} a^{l-1}_k + b^l_j\right)$$

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

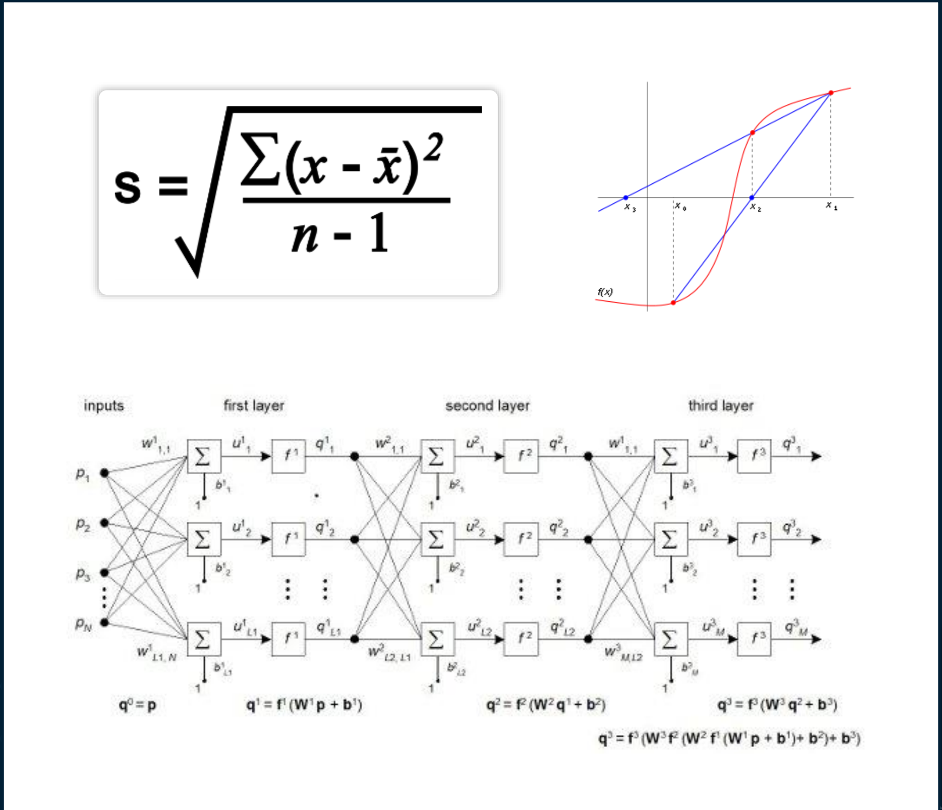
$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j$$

$$\frac{\partial C}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j$$



The black box



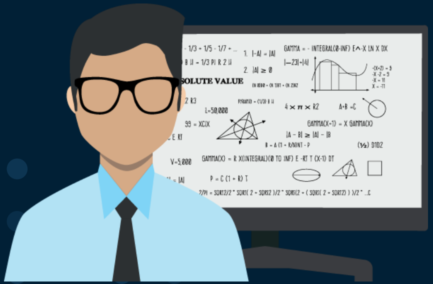
How can we generate models which are not only accurate, but

Fair
Accountable
Transparent
Trustworthy
Explainable
?

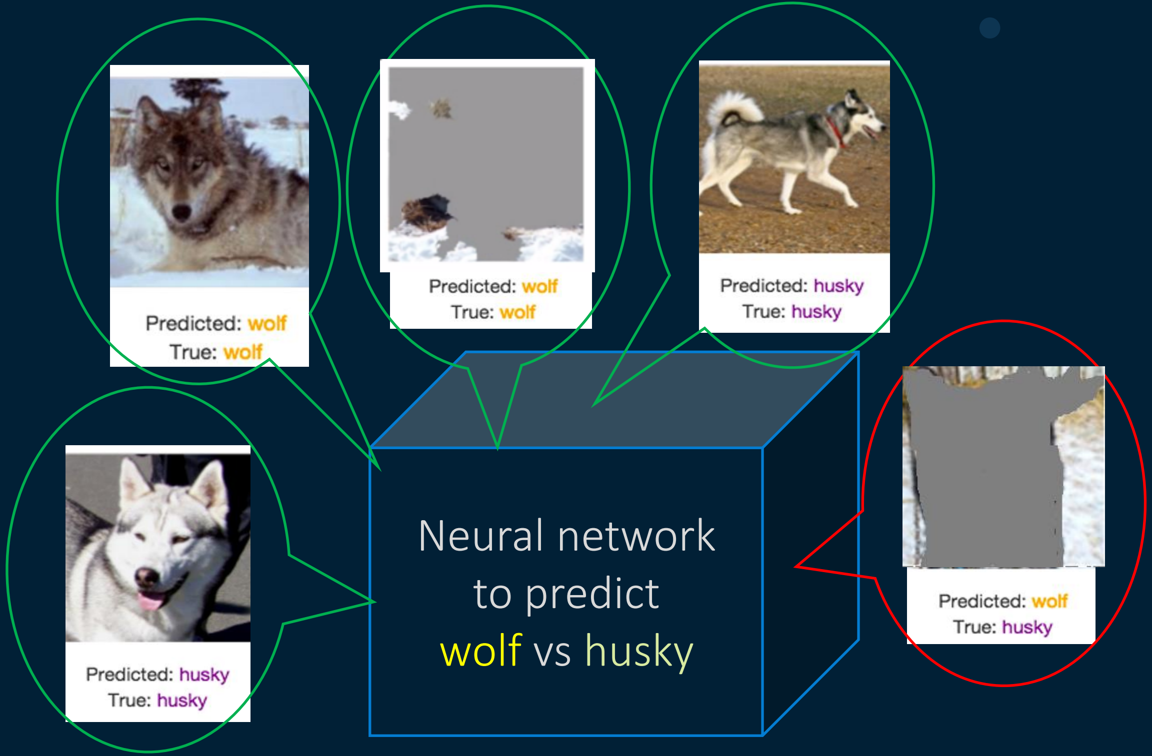
Figure out when NOT to trust a model

Prediction accuracy is very high. It is time to put this system online.

You are detecting snow, not wolves! I can't trust you



Data scientist



"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. ACM SIGKDD, 2016

Shine some light on the black box

Road to Trustworthy AI

Being able to interpret and explain machine learning models is key to trustworthy AI

Proxy Methods and Diagnostics

Proxy Methods:

- **Surrogate model approach**

  Fit a black box machine learning model (deep learning, gradient boosting, random forest, etc.) to your training data. Then use those predicted outcomes as the targets for a more interpretable model (decision tree, regression)

- **Machine learning as benchmark**

  Use a complex model to set the goal for potential accuracy metrics that could be achieved, then use that as the standard against which you compare the outputs of more interpretable model types

- **Machine learning for feature creation**

  Use ML/Deep Learning to extract the features, then use those features as inputs to a more explainable model type

Post-Modeling diagnostics:

- Variable importance(VI)
- Partial Dependence (PD)
- Individual Conditional Expectation (ICE)
- Local Interpretable Model-agnostic Explanations (LIME)
- **SHapley Additive exPlanations (SHAP)**

Post-Modeling Diagnostics

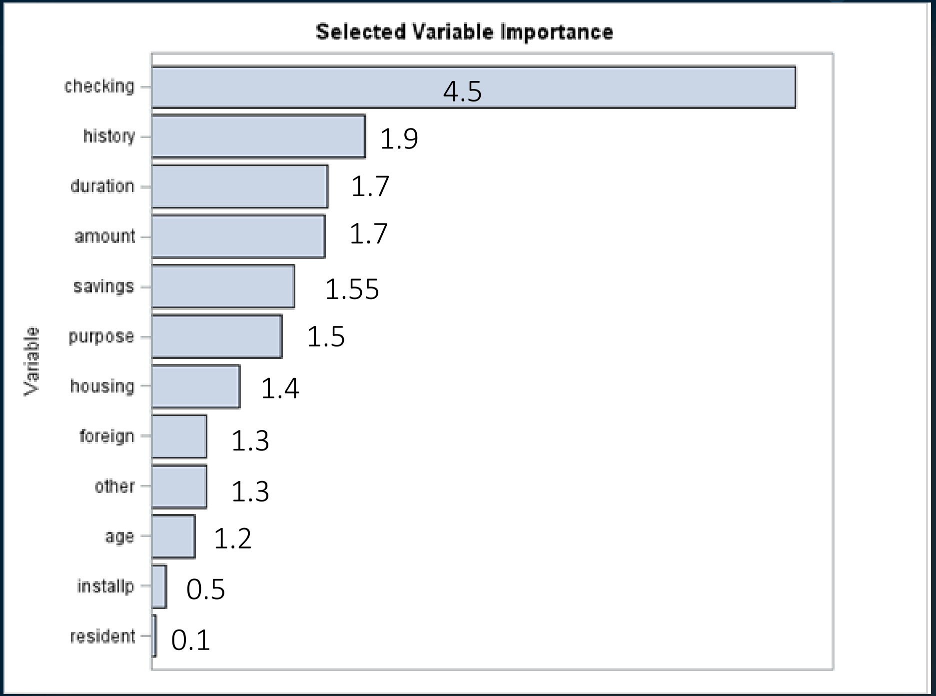
Input-Output Relationship

Question	Technique
What are the top inputs?	**Variable Importance(VI)/Relative VI**
How do the drivers work?	**Partial Dependence (PD)** **Individual Conditional Expectation (ICE)**
What is the explanation for a particular prediction?	**Local Interpretable Model-agnostic Explanations (LIME)** **SHapley Additive exPlanations (SHAP)**

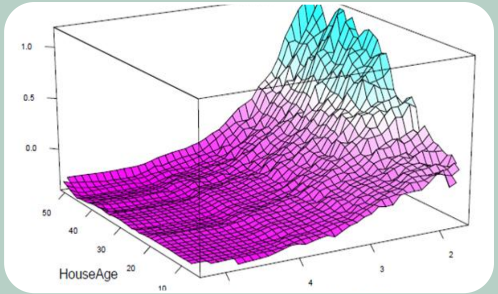
Variable Importance

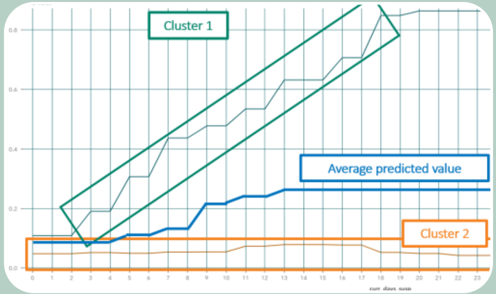
Variable Importance

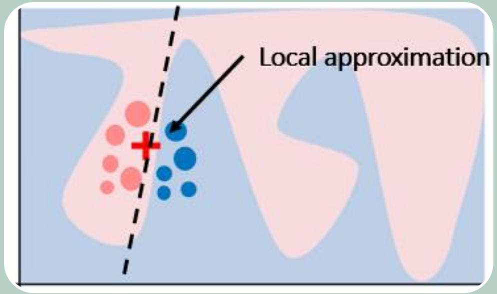
Variable Name	Train Importance
curr_days_susp	81.3108
handset_age_grp	46.5221
ever_days_over_plan	30.2300
pymts_late_ltd	24.8962
billing_cycle	16.2053
avg_days_susp	15.8511
calls_care_ltd	13.3913
call_category_1	13.1900



Train Importance is calculated as sum of the decrease in error when split by a variable.







Partial Dependence Plots

depicts relationship between the value of an input variable and the value of the model predictions after the influence of all other variables has been averaged out

Individual Conditional Expectation (ICE)

helps identify subgroups and interactions

Local Interpretable Model-agnostic Explanations (LIME)

builds an interpretable model of explanatory data samples at local areas in the analyzed data

LIME (Local Interpretable Model-agnostic Explanations)

Flu Prediction (Neural Network)

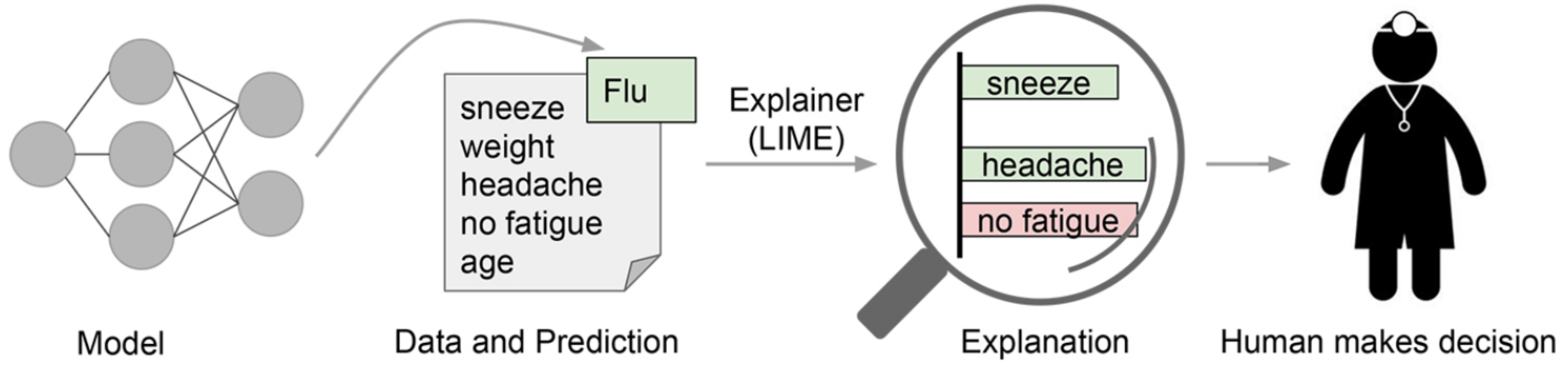


Figure 1. Explaining individual predictions to a human decision-maker. Source: KDD, Ribeiro, Singh, and Guestrin

Some examples…….

sas.com