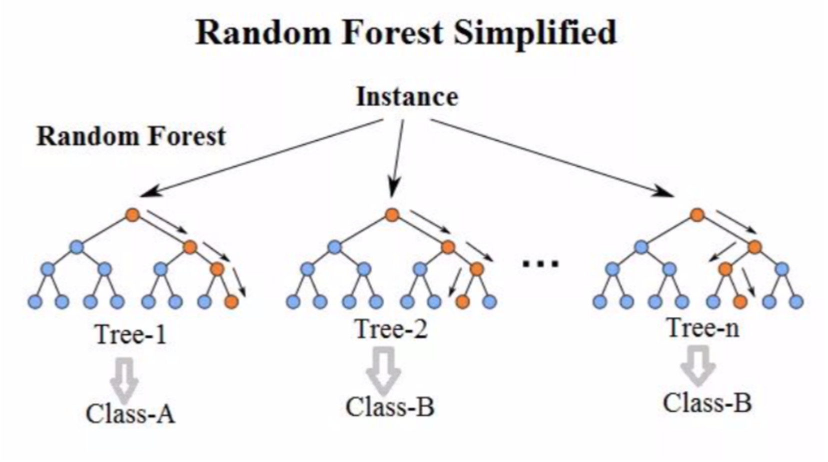
Model Agnostic Methods for Interpretability

Analytics Vidhya
Learn everything about analytics

Some Models are hard to interpret

Ensemble models (random forest, boosting, etc…)

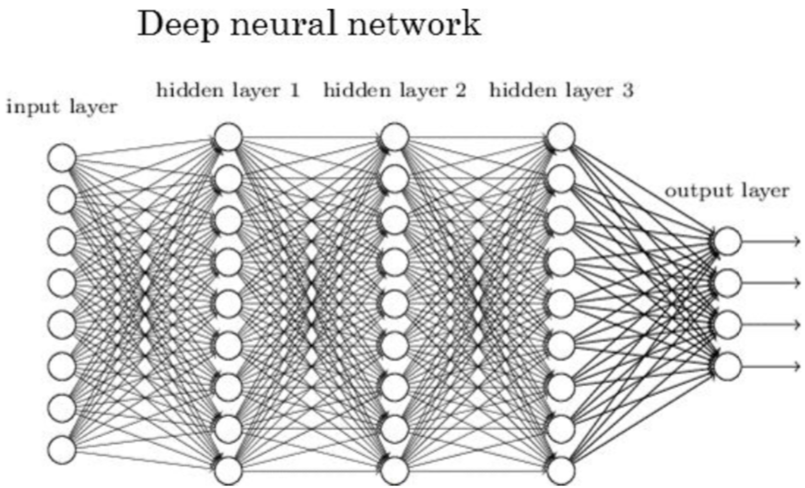
- Hard to understand the role of each feature
- Usually comes with **feature importance**
- Doesn’t tell us if feature affects decision positively or negatively



Some Models are really hard to interpret

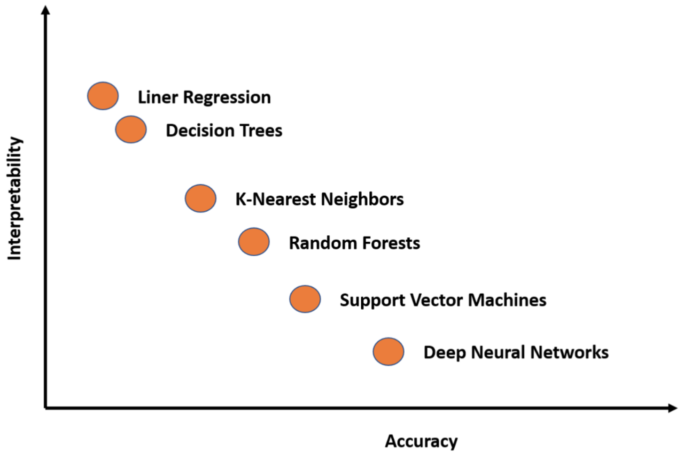
Deep Neural Networks

- No straightforward way to relate input to output layer
- Millions of parameters
- “Black-Box”



Use only simple models?

- Sticking to simpler models is the best way to be confident about interpretation
- However, more complex models such as ensembles and neural network can provide better performance



Analytics Vidhya
Learn everything about analytics

Use only simple models?

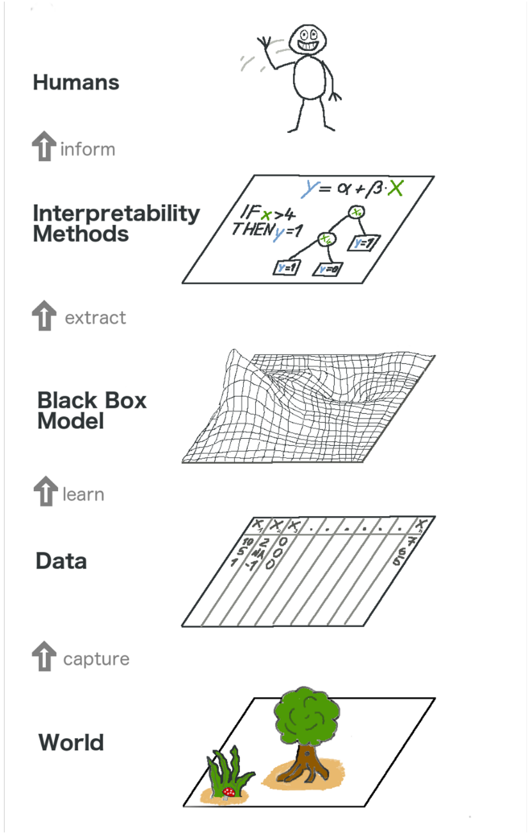
Model agnostic techniques allows usage of more complex models without losing all interpretability power

Black Box → Model Agnostic Methods

White Box → Model Specific Methods

Analytics Vidhya
Learn everything about analytics

Model Agnostic Interpretability



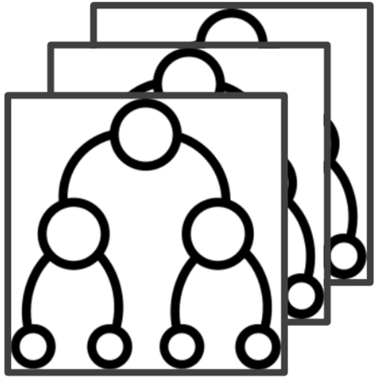
Analytics Vidhya
Learn everything about analytics

Global Surrogate methods

Idea:

Approximate complex model output with simpler model

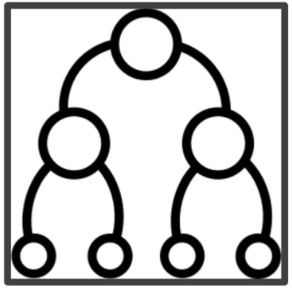
Complex Model



Random Forest Classifier

Predictions: [0,1,0,1,1,0]

Simpler Model



Decision Tree Classifier

Predictions: [0,0,0,1,1,0]

Accuracy: 83.33 % accuracy

Analytics Vidhya

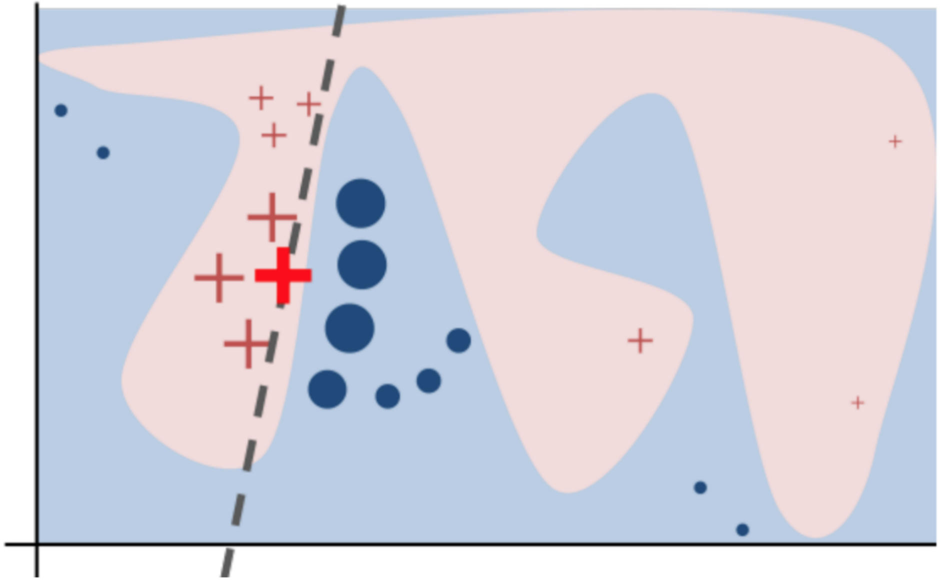
Learn everything about analytics

Global Surrogate methods: Steps

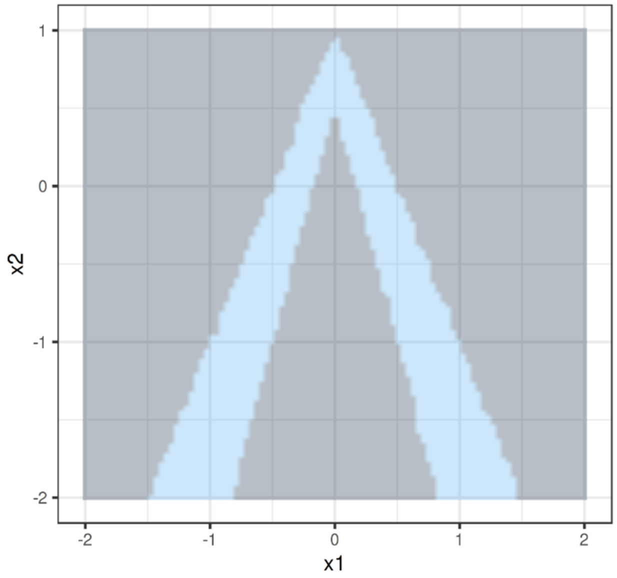
- Interpretable surrogate model that is trained to approximate the predictions of a black box model
- Steps:
  - Get predictions from black box model
  - Select an interpretable model (Linear, DT….)
  - Train interpretable model on original dataset and black box predictions as target
  - Measure performance of surrogate model
  - Interpret the surrogate model

LIME (Local Interpretable Model Agnostic Explanations)

Local interpretation of each prediction for a Black Box Model

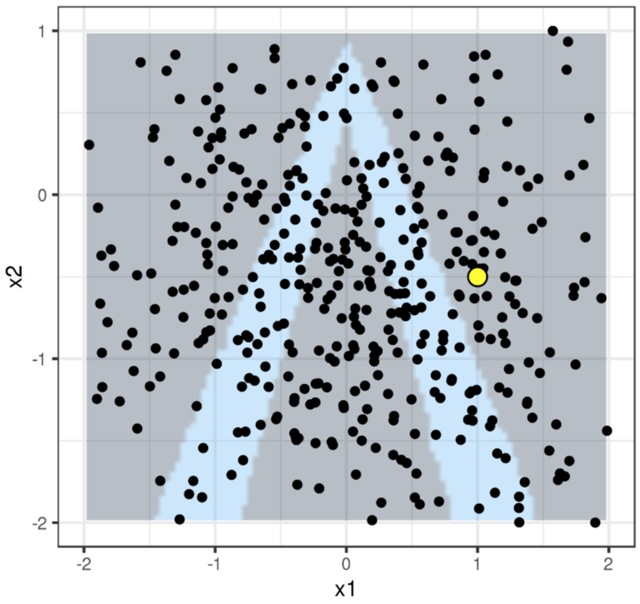


LIME - How does it work?



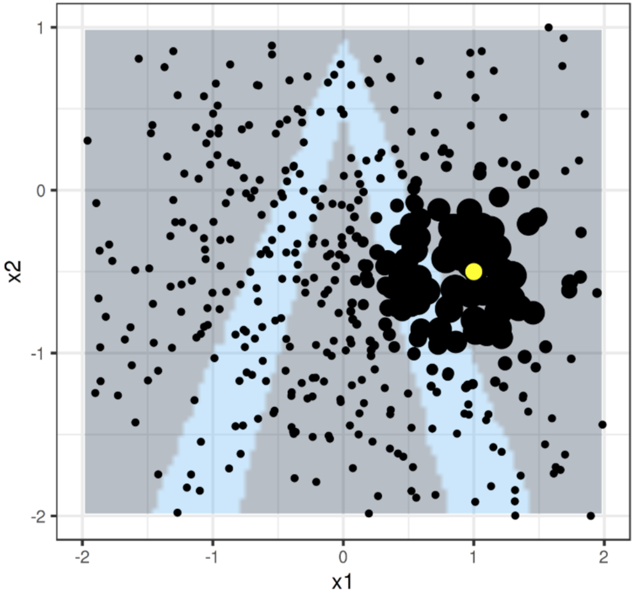
Decision Boundary for a black box model with features x1 and x2

LIME - How does it work?



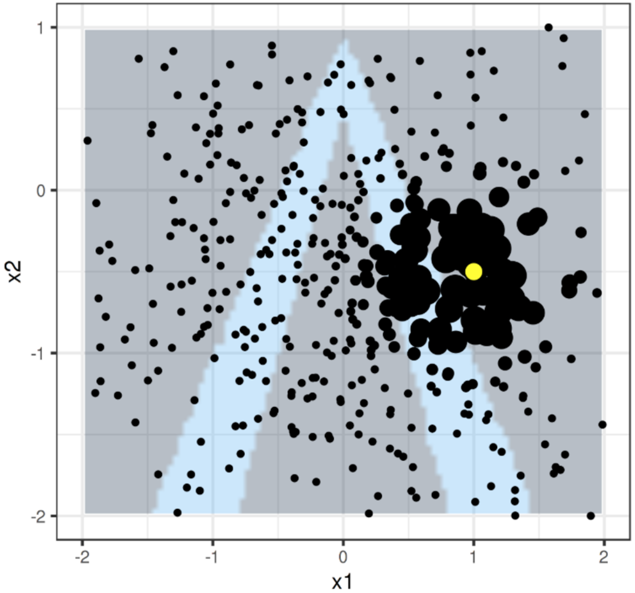
Selected observation (yellow) and data sampled from a normal distribution (black dots)

LIME - How does it work?



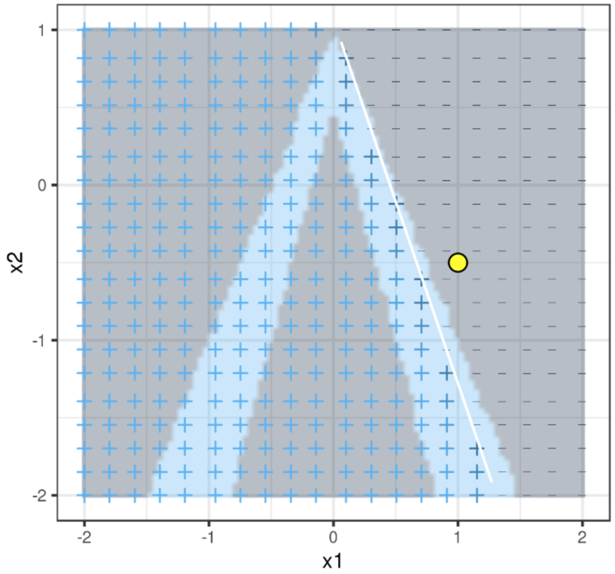
Assign higher weight to points near the our observation

LIME - How does it work?



Train an interpretable model over the fake data generated from the distribution

LIME - How does it work?



The white line marks the new decision boundary for locally learned model

LIME - Let’s Summarise

- Select your instance of interest for which you want to have an explanation of its black box prediction.
- Perturb your dataset and get the black box predictions for these new fake data points.
- Weight the new samples according to their proximity to the instance of interest.
- Train a weighted, interpretable model on the dataset with the variations
- Explain the prediction by interpreting the local model

Thank You!

Analytics Vidhya
Learn everything about analytics