Machine Learning Explainability

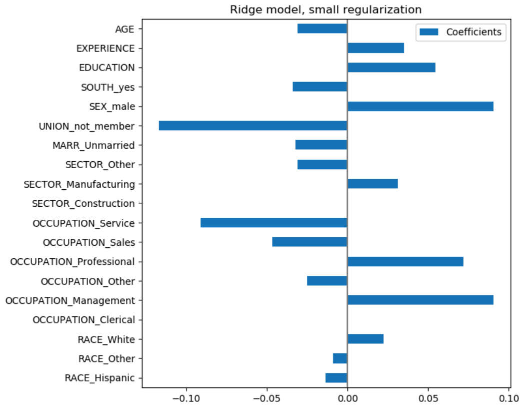
John Titus Jungao

Right to Explanation

- **Definition:** The right to be given an explanation for the output of an algorithm
- **Importance:** High-risk ML applications that **may significantly affect an individual** (e.g. Financial, Legal, Medical)
- **Example:** Being disapproved for a loan and given the reason "due to poor credit history" is **not actionable**

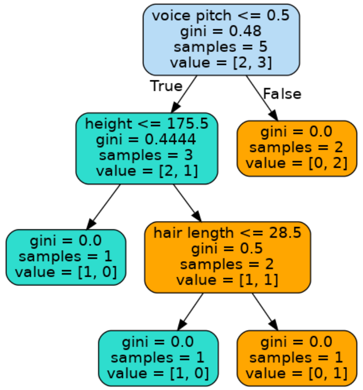
Model-Specific Methods

Linear Models



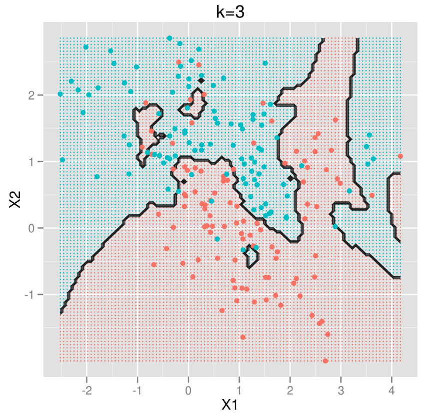
Use the coefficients

Decision Trees



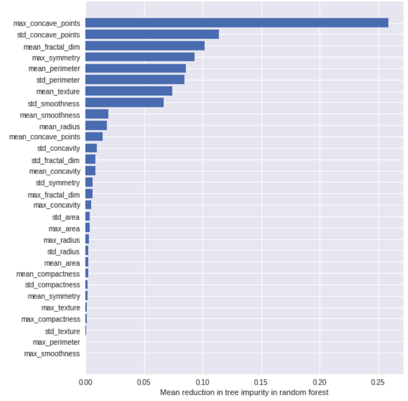
Graph the Tree

K-Nearest Neighbors



Show "most similar" samples

Tree-based Ensembles



Use feature importance

Model-Agnostic Methods

- **Description:** Methods that are **independent from any model**
- **Importance:** Allow **unified interpretation** across models
- **Examples:** Permutation Importance, LIME, SHAP, DiCE

Permutation Importance

- **Intuition:** **Measures increase in prediction error** after a feature's values are permuted
- **Theory:** If a feature increases the prediction error when shuffled, it is important since the **model relied on its association with the target** variable.
- **Pseudocode:**
    1. Compute original model error.
    2. For each feature j, **shuffle values** and keep the values of other features fixed.
    3. Estimate new error given predictions of the permuted data
    4. Compute Permutation importance using the new and original errors

Demo

Permutation Importance: Pros and Cons

Pros

- Very Intuitive
- Takes into account feature interactions
- Does not require retraining the model

Cons

- Use train or test data?
- Takes into account feature interactions
- Might require several repetitions to achieve stable error results

LIME

(**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations)

- **Intuition**

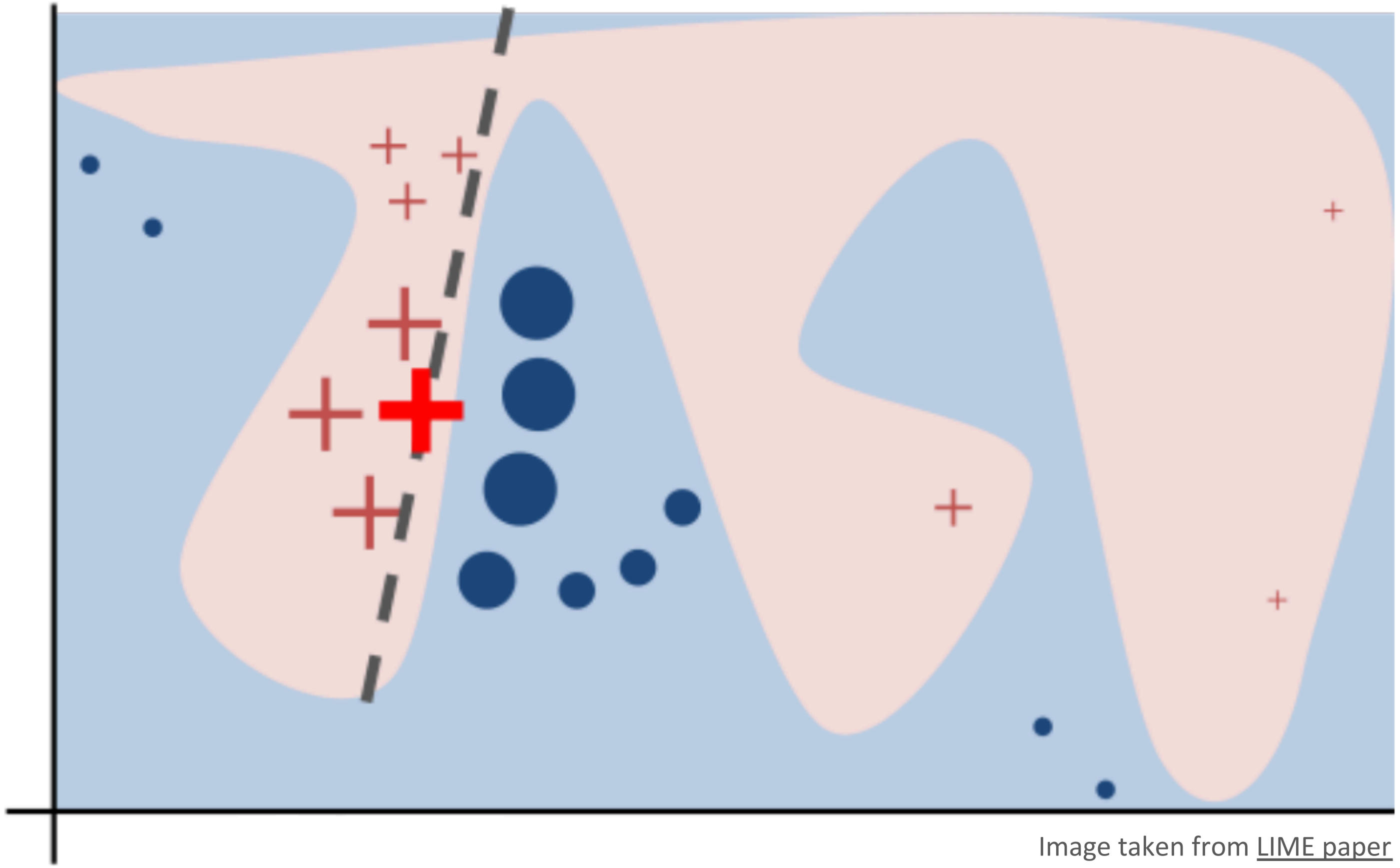


Image taken from LIME paper

- Blue/pink area represents the model's decision function *f*
- We're trying to explain prediction for **bold red cross**
- LIME generates perturbed samples, gets prediction using *f*, *and weighs these by proximity to target sample*
- Fit a model using new samples (dashed line)

LIME

(Local Interpretable Model-agnostic Explanations)

- **Use Case**

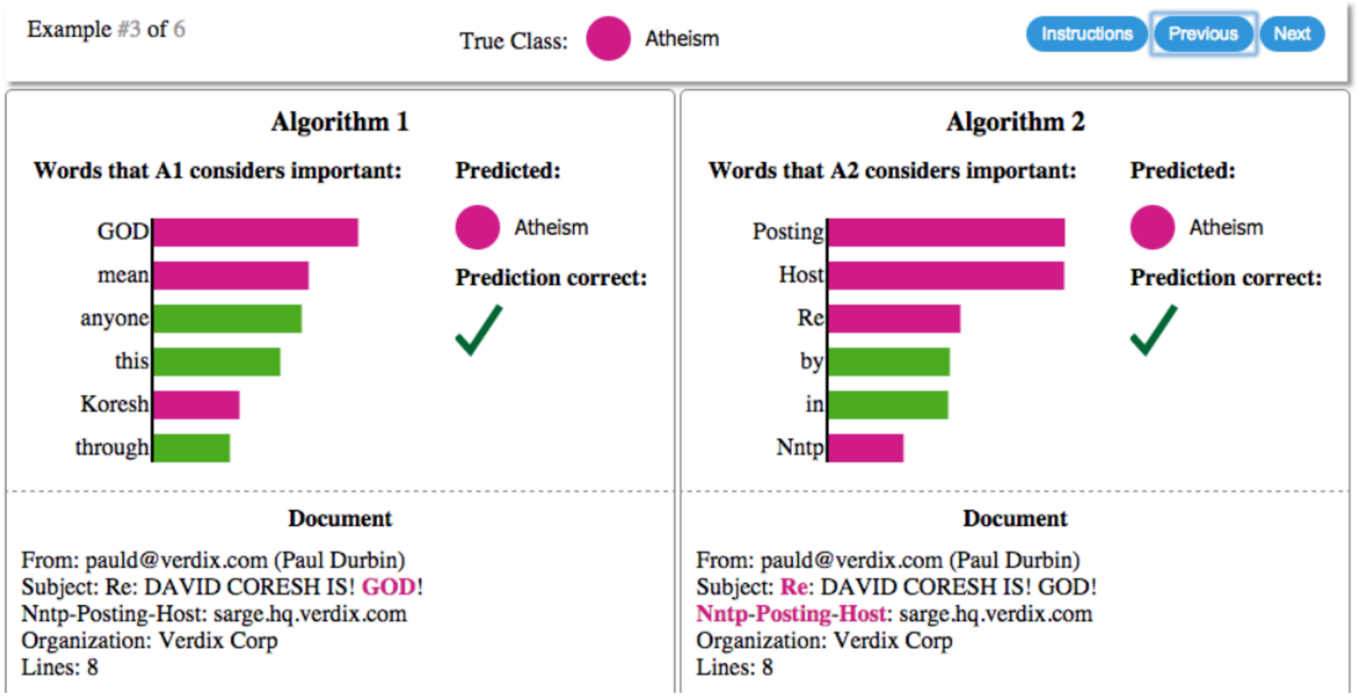


Image taken from LIME paper

- Which model should you trust?
- Models here considers different features as important
- Based on importance, we can see which model is better

DEMO

LIME: Pros and Cons

Pros

- Leverages simple ideas
- Relatively fast

Cons

- May not handle nonlinearity in local region
- Simple perturbed samples may not be enough

SHAP

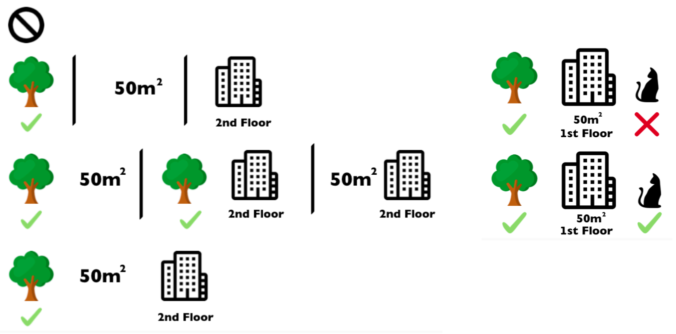
(SHapley Additive Explanations)

Shapley Value (Game Theory). Given a model with prediction on apartment price,
how much does each feature contribute?

Given: 

Solution:

1. Get similar samples.



2. For missing features, fill using expected value.
3. Compute predicted price using the model.
4. Fit a linear model and weigh data
   using sum of available features

SHAP

(SHapley Additive Explanations)

More similar* to human intuition than LIME

*based on a sample sized comparison

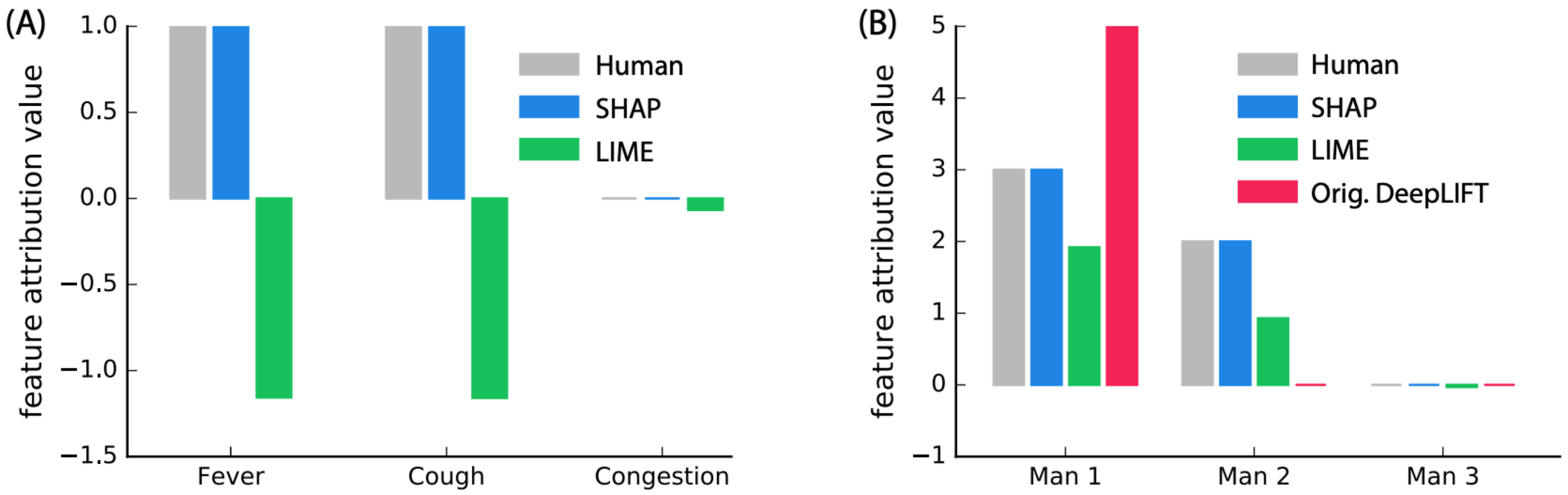


Figure 4: Human feature impact estimates are shown as the most common explanation given among 30 (A) and 52 (B) random individuals, respectively. (A) Feature attributions for a model output value (sickness score) of 2. The model output is 2 when fever and cough are both present, 5 when only one of fever or cough is present, and 0 otherwise. (B) Attributions of profit among three men, given according to the maximum number of questions any man got right. The first man got 5 questions right, the second 4 questions, and the third got none right, so the profit is $5.


DEMO

SHAP: Pros and Cons

Pros

- Solid theoretical foundation
- Prediction is fairly distributed across features
- SHAP connects LIME and Shapley values (unifies field of ML Explainability)
- Global model interpretation is feasible using Shapley values

Cons

- SHAP is slow and computing multiple Shapley values is impractical
- SHAP ignores feature dependence

DiCE

(**Di**verse **C**ounterfactual **E**xplanations)

- **Intuition:** Find feasible “what if” scenarios that could result to more favorable outcomes

  (e.g. qualifying for a loan if income is increased by X)

- **Theory:** Setup finding these scenarios as an optimization problem similar to finding adversarial examples

Demo

Summary

- **Explainability in ML models is important** given that it can have significant effects on individuals
- There are **Model Specific and Model-Agnostic Methods** that we can use
- Permutation Importance can be used for **global interpretation of a model**
- LIME and SHAP can be used for **local interpretation of sample predictions**
- DiCE can be used to provide **alternative scenarios that are potentially actionable**

References

- Interpretable ML free online book
- ML Explainability Kaggle Course
- Interpretml/interpret Repo
- My notebooks