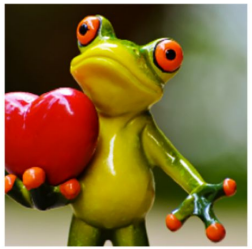
Local Interpretable Model-Agnostic Explanations (LIME)

Aidan Donnellan, Garrett Kemper
ELE392

What Is Lime

- LIME stands for Local Interpretable Model-Agnostic Explanations.
- Improves interpretability of your model.
- Explains the individual predictions that your model makes.
- This method allows it to be supported by all types of models.

What Does It Do

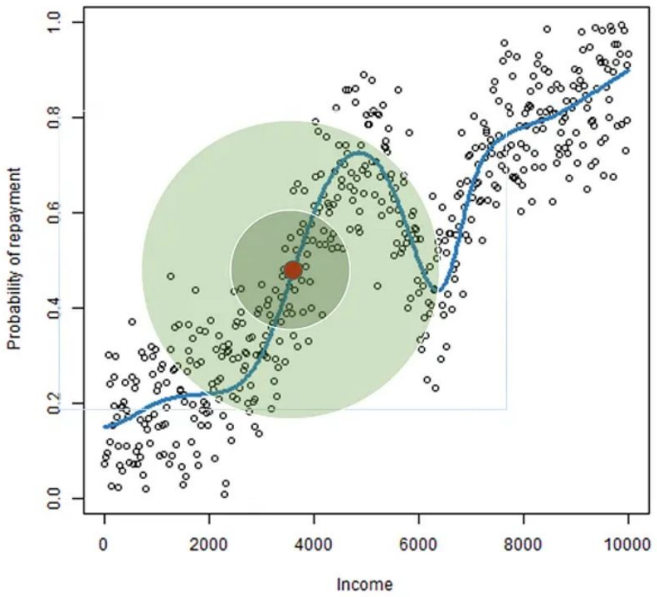


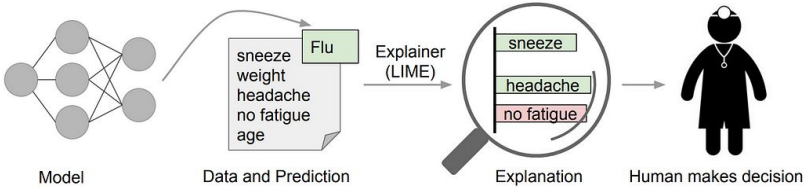
Original Image

Interpretable Components

- Explains any black-box model
- Supports tabular, text and image based datasets
- Visualisations created make the models more transparent to the maker and are easier to explain.
- Gives insights into which features are the most significant for predictions

How Does It Work





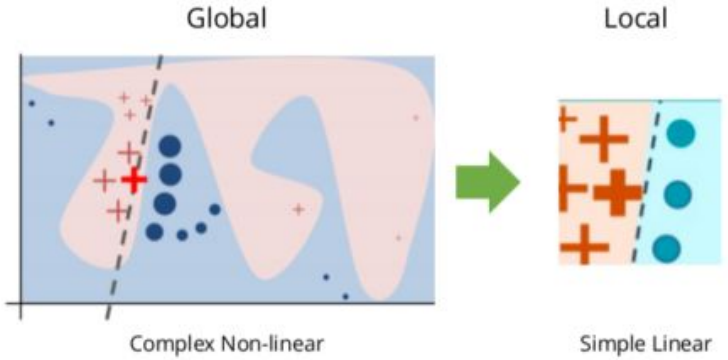
$$RBF(x^{(i)}) = \exp\left(-\frac{||x^{(i)} - x^{(ref)}||^2}{kw}\right)$$

Gaussian Kernel Formula

How Does It Work

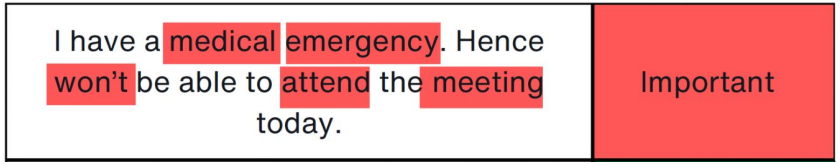
- Local
  - Approximates model in a local window
  - Focus on a single prediction
- Interpretable
  - Approximates complex model with simpler, readable model
- Model-Agnostic
  - Can be applied to any machine learning model
- Explanations
  - Explains your model’s explanations

How Does It Work



- Perturbation
  - Create small variations in the input
- Model Sampling
  - Model makes predictions on the new inputs
- Sample Weighting
  - Assigns higher weights to samples closer to original
- Model Training
  - Trains a more interpretable model (ex. Linear regression or decision tree)
- Generating Explanations
  - Determines which features contributed the most to a given prediction

Applications



Can be used on any classification or regression model

Model Debugging: spotting spurious correlations of biases

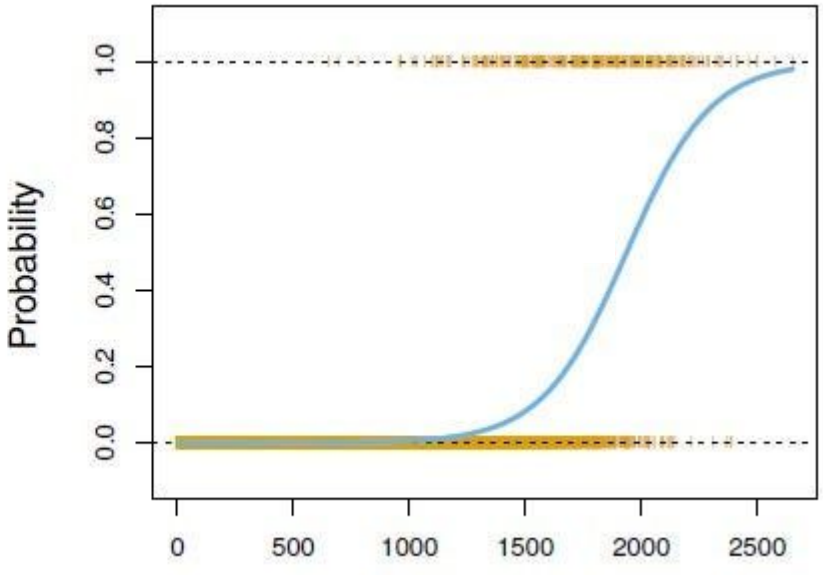
Image Recognition: highlight most important pixels

Text Explanation: highlight most important words

Regulatory Compliance: Making black-box models explainable to end users

End User Application

- Credit Card Company
- Classifying risk of defaulting
- Use LIME to inform the user why they were classified as high risk
- Inform user how they can improve their risk classification



Limitations

- Approximations
  - Models only make local approximations
  - May not be perfect for highly complex models
- Computational Costs
  - Generates multiple models
  - May be slow for large datasets

Other Explanation Techniques

- SHAP (SHapley Additive Explanations) - Computationally Expensive
- Integrated Gradients
- Grad-CAM - Better for CNNs in vision tasks

Colab Workshop

Workshop_LIME > ELE392_LIME_Workshop.ipynb

Resources

- https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe
- https://medium.com/intel-student-ambassadors/local-interpretable-model-agnostic-explanations-lime-the-eli5-way-b4fd61363a5e
- https://lime-ml.readthedocs.io/en/latest/lime.html