Explaining and Trusting Predictions

SOEN 691: Engineering AI-based Software Systems

Emad Shihab, Diego Elias Costa
Concordia University

Introduction

- Users need to trust a model or prediction they use
  - Trusting a **prediction**: trusting a prediction (or outcome) to act on it
  - Trusting a **model**: trusting that the model will behave “reasonably” if deployed, on real-world data

Introduction
- contributions

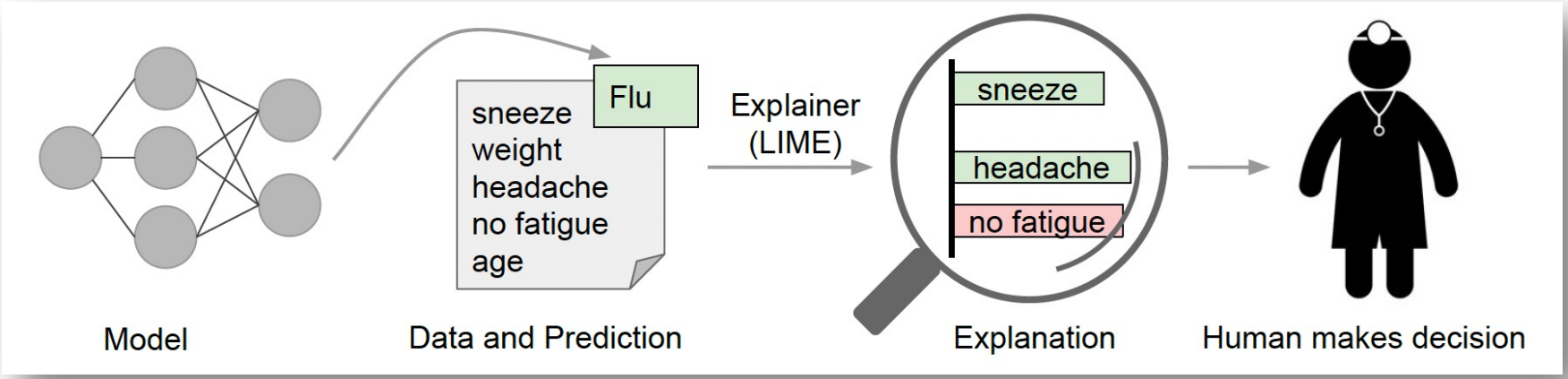
- Explain an individual **prediction** -> trusting the prediction (LIME)

- Select and explain **multiple predictions** -> trusting the **model** (SP-LIME)

Case for Explanations

Explanations can help in many scenarios:

- Data leakage: e.x. IDs are used in the prediction
- Dataset shift: training and testing are different



Q: What if the human does not have the knowledge?

Desired Characteristics for Explainers

- **Interpretable**: provide qualitative understanding between the input variables and the response
  - Should not have 1000s of features for example
  - Can depend on the target audience

- **Local fidelity**: an explanation must correspond to **how the model behaves in the vicinit**y of the instance being predicted

Desired Characteristics for Explainers

- **Model-agnostic explainer**: explain **any** model

- **Providing a global perspective**: important to ascertain trust in the model

LIME

- **Interpretable explanations**: understandable by humans (e.g., using binary vectors)

- **Local fidelity**: using L, which measures how unfaithful a model is in approximating a probability (f) in a locality defined (pi)

- **Goal: minimize L while keeping the model interpretable (i.e., complexity low)**

LIME

- **Sampling**: needs to be model-agnostic. Sample uniformly, at random, weighted by defined locality (pi)

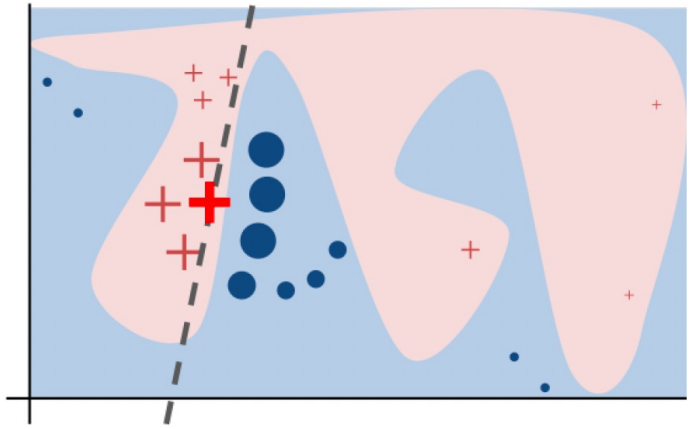


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Examples

What is the main idea here?

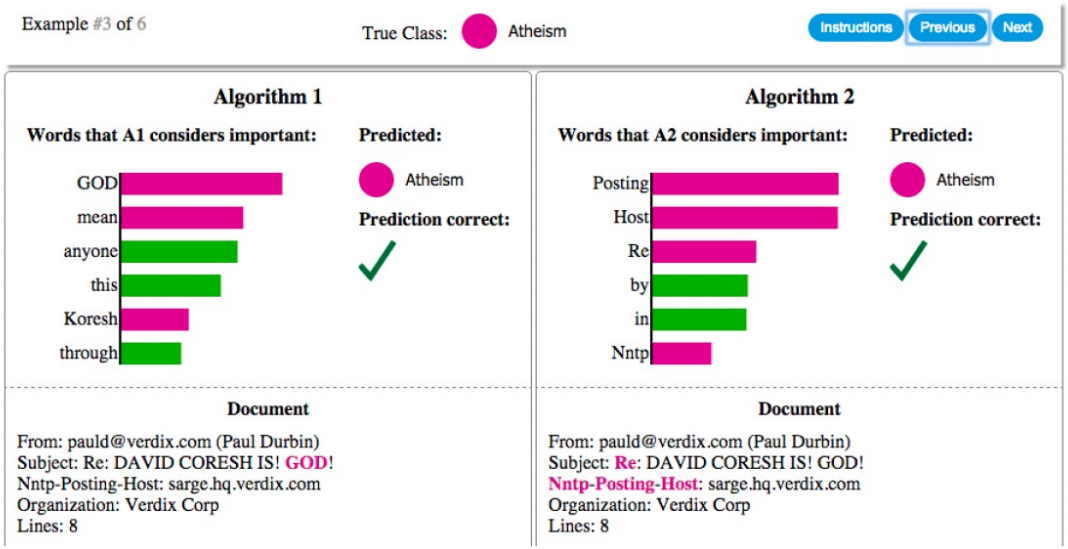
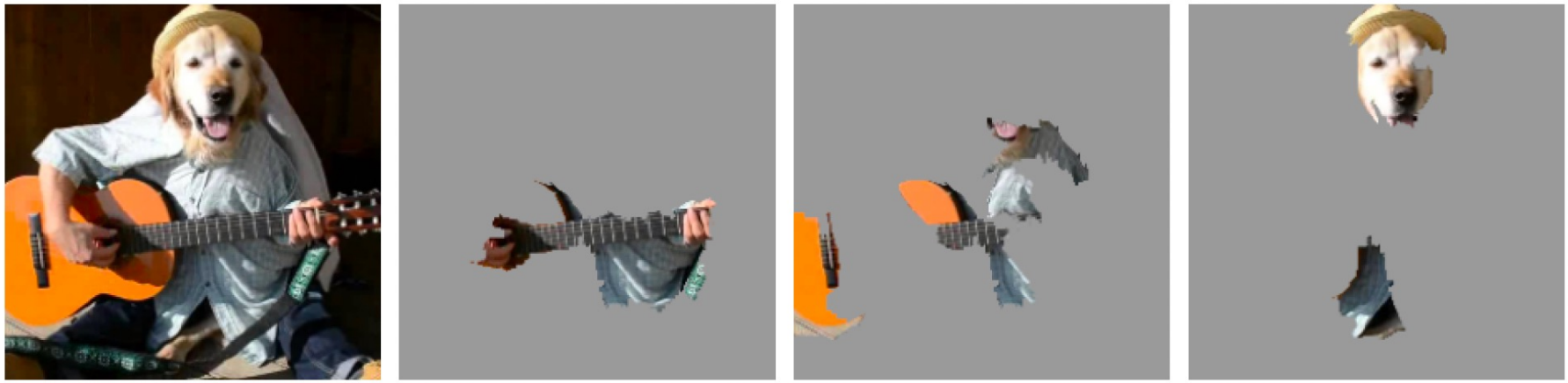


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for "Christianity", magenta for "Atheism").

Examples

What is the main idea here?



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Submodular Pick for explaining models

- Need to select explanation instances judiciously.
- **Pick step**: picking the right number of explanations within a time budget.
- **Idea**: Give a higher weight to instances that can explain many different instances.

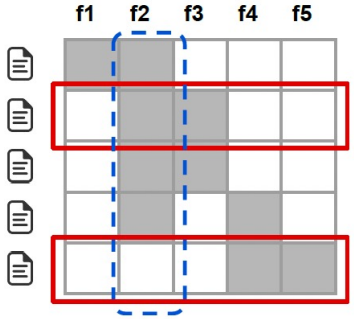
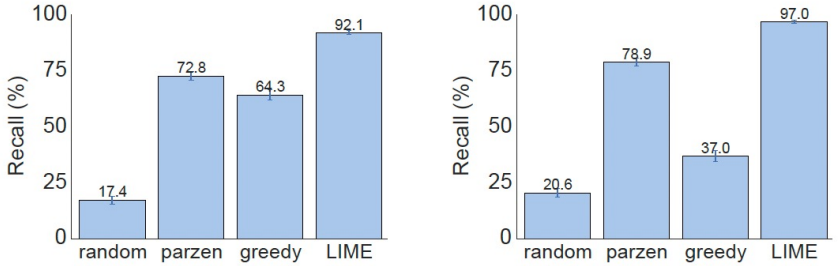


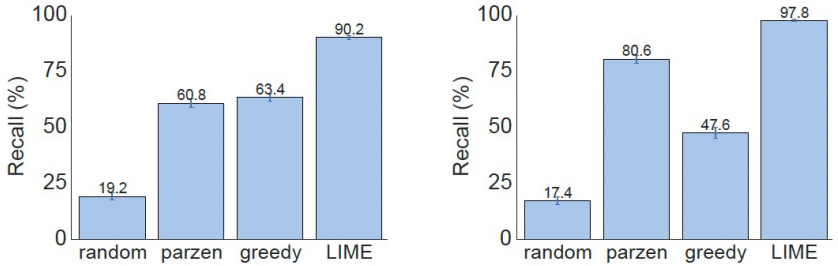
Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Validation
Are explanations faithful to the model?



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.



(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

1. Train models so that max # of features for any instance is 10 (gold set)
2. Generate explanations and measure overlap with gold set.

Validation
Should I trust this prediction?

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	**96.6**	**94.5**	**96.2**	**96.7**	**96.6**	**91.8**	**96.1**	**95.6**

1. 25% randomly selected features as untrusthworthy
2. **Oracle**: untrustworthy if the prediction changes when untrustworthy features are removed from the instance

LIME/parzen: untrustworthy if prediction changes when **ALL** untrustworthy features appear in the explanations are removed
Random/greedy: untrustworthy if **ANY** untrustworthy explanations appear in the explanations.

Validation
Can users select the best classifier?

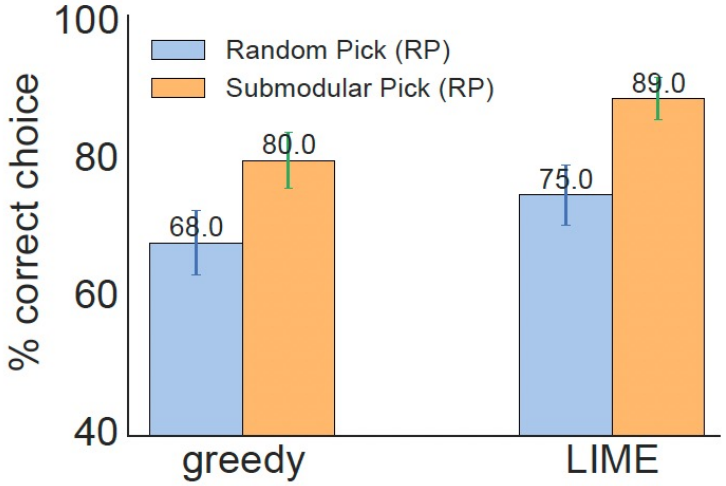


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

1. Compare two classifiers, one trained on cleaned dataset. Test accuracy is lower on cleaned dataset.
2. Users are given words and documents to inspect and determine which classifier will perform best in the real world.

Validation
Do explanations lead to insights?

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

1. Wolves (snow background) vs. Husky - considered a bad classifier.
2. Asked graduate students if they trusted? Why? How?
3. Three independent evaluators read the reasons to summarize the results.

Open Discussion