“Why Should I Trust You?”Explaining the Predictions of Any Classifier

M. T. Ribeiro, S. Singh, C. Guestrin

02/27/2020

Presenter: Yu Yang

https://qdata.github.io/deep2Read/

Introduction

Two problems:

How does one…

P1) Trust a prediction

Prediction of importance: medical diagnosis, terrorism attack

P2) Trust a model

Real-world data are often different from datasets

Metrics may not be indictive of the end goal

Solution:

Inspecting the prediction and explanation of sampled individual samples

- Explanation for each sample (P1)
- Multiple samples (P2)

Introduction

Solution:

Inspecting the prediction and explanation of sampled individual samples

- Explanation for each sample (P1)
- Multiple samples (P2)

LIME:

- provides explanation for individual samples (Solution to P1)

SP-LIME:

- selects a set of representative instances with explanation
- addresses ‘explanation of the model’ using explanations of the most representative samples (Solution to P2)

Comprehensive evaluation with…

- Simulated User Subjects
- Human Subjects

The Case for Explanations

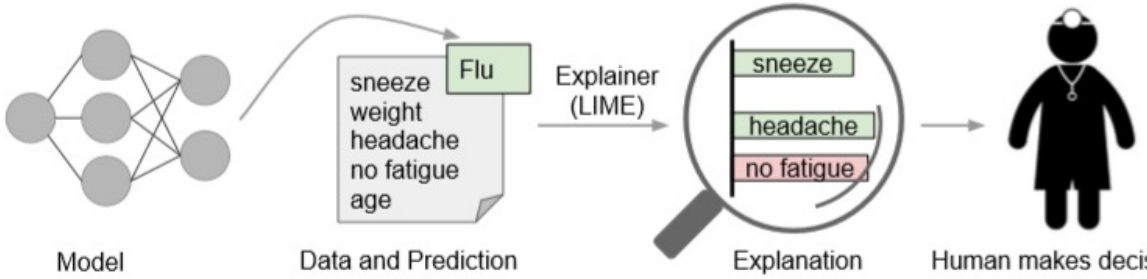


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient's history that led to the prediction. Sneeze and headache are portrayed as contributing to the "flu" prediction, while "no fatigue" is evidence against it. With these, a doctor can make an informed decision about whether to trust the model's prediction.

- When human make decisions with the help of predictions, trust is a fundamental concern.
- A model can go wrong in several ways:
  - Data leakage
  - Dataset shift
- Explanations can help us identify what went wrong and fix it fast.

The Case for Explanations

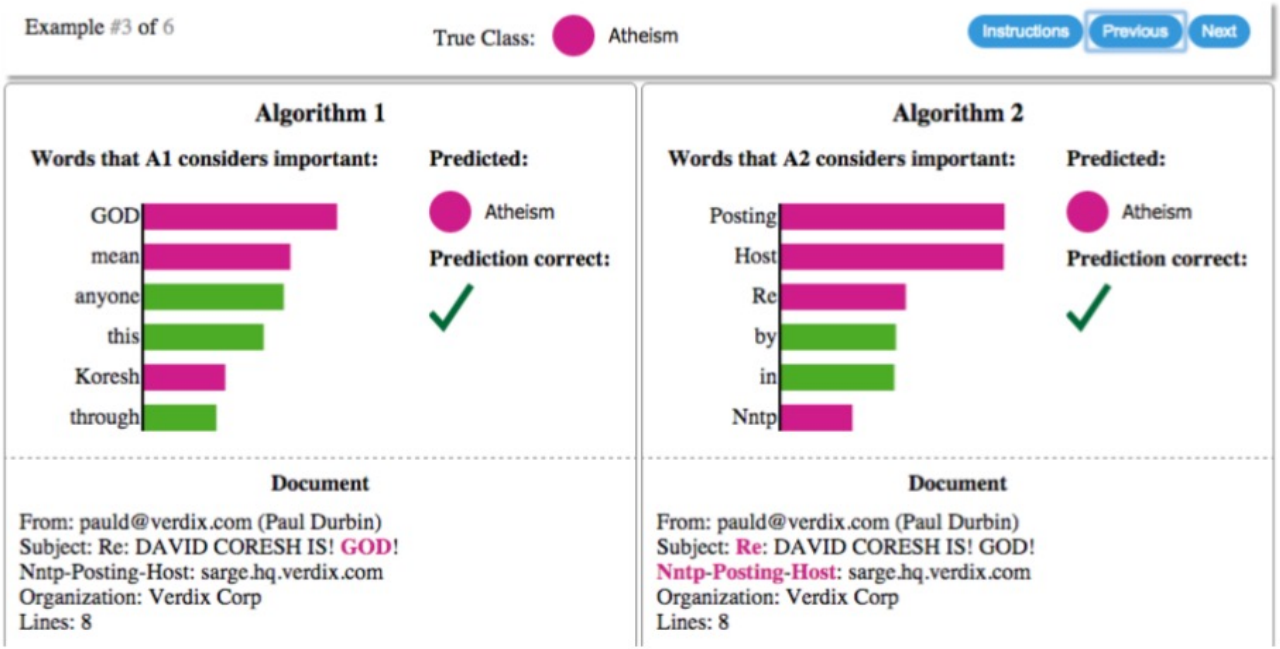


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about "Christianity" or "Atheism". The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green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

Proposed Solution

Desired Characteristics for Explainers:

- Interpretable
  - Should be easy to understand by human
- Local fidelity
  - Reflects how the model behaves in the vicinity of the instance being predicted
- Model-agnostic
- Providing global perspective
  - Accuracy may not be sufficient to explain a model
  - We want to explain the model, not just individual predictions.

3 LIME

LIME: Local Interpretable Model-Agnostic Explanations

Overall goal: identify an ***interpretable*** model over the *interpretable representation* that is ***locally faithful*** to the classifier.

$x \in \mathbb{R}^d$ the original representation of an instance

$x' \in \{0,1\}^{d'}$ A binary vector for its interpretable representation

Fidelity-Interpretability Trade-off

Definitions:

$g \in G$	g is the explanation, G is a class of potentially interpretable models
$\Omega(g)$	A measure of complexity of explanation g
$f : \mathbb{R}^d \rightarrow \mathbb{R}$	The model being explained
$\pi_x(z)$	Proximity measure around the instance x (locality)
$\mathcal{L}(f, g, \pi_x)$	Measures how unfaithful g is in approximating f in the locality defined by $\pi_x(z)$

We want to minimize *locally-aware* loss:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g) \qquad (1)$$

Sampling for Local Exploration

In order to learn the local behavior of f as the interpretable inputs vary, we approximate $\mathcal{L}(f, g, \pi_x)$ by drawing samples, weighted by πx.

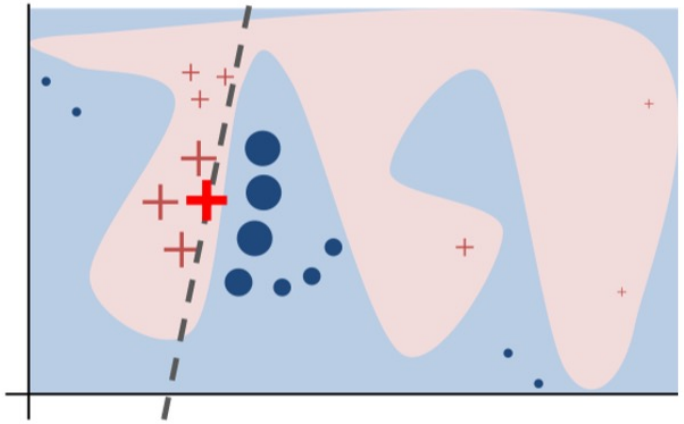


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions 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.

$$\Pi_x(z) = exp(-D(x,z)^2/\sigma^2)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2 \qquad (2)$$

LIME Algorithm

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z'_i \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
end for
$w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
return w

SP-LIME

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X, Budget B

```
for all x_i ∈ X do
    W_i ← explain(x_i, x'_i)                ▷ Using Algorithm 1
end for
for j ∈ {1 ... d'} do
    I_j ← sqrt(Σ_{i=1}^{n} |W_ij|)          ▷ Compute feature importances
end for
V ← {}
while |V| < B do                            ▷ Greedy optimization of Eq (4)
    V ← V ∪ argmax_i c(V ∪ {i}, W, I)
end while
return V
```

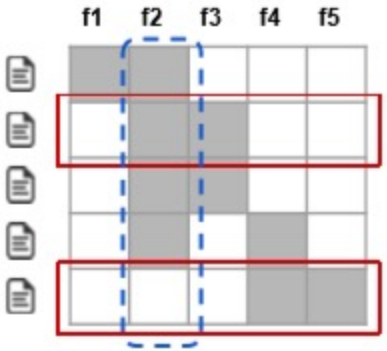


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

5 - Simulated User Experiment

Experiment Set-up:

Two datasets: *Books*, *DVDs* (2000 samples each)

Different models:

- Decision trees (DT)
- Logistic regression with L2 regularization (LR)
- Nearest neighbors (NN)
- SVM with RBF kernels (SVM)
- Random forest w/ 1000 trees (RF)

Feature: bag of words

5.1 – Experiment Set up

Methods to Compare:

- LIME
- parzen
- Greedy procedure
- Random procedure

Where there is a Pick procedure

- Random pick (RP)
- Submodular pick (SP)

Parameters:

K – # of features with highest absolute gradients

N – Cross validation

K = 10

N = 15000

Questions to ask:

5.2 Are explanation faithful to the model?

5.3 Should I trust this prediction?

5.4 Can I trust this model?

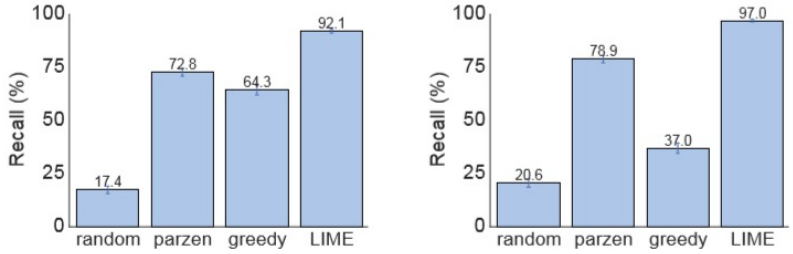
5.2 Are predictions faithful?

Measure faithfulness of explanation on classifiers that are interpretable:

- Sparse logistic regression
- Decision trees

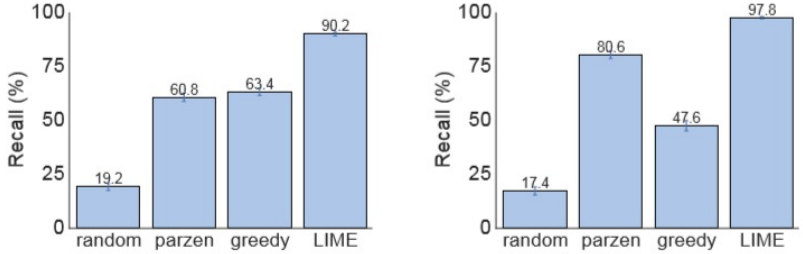
Procedure:

- Train both classifiers s.t. maximum number the model use is 10
- Compute the fraction of "gold" features that are recovered by the explanation.
- Report averaged recall over all the test instances (Fig.6. Books; Fig. 7. DVDs)



(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.

5.3 Should I trust this prediction?

Procedure:

- Randomly select 25% of the features to be "untrustworthy"
- Assumption:
  - Users can identify these "untrustworthy" features
  - Would not want to trust these features
- Develop the oracle "trustworthiness" of a prediction
  - Untrustworthy if the prediction changes after removing all untrustworthy features
  - Trustworthy, otherwise.
- Deem if a prediction is trustworthy or not
  - LIME or parzen:
    - *untrustworthy* if prediction from linear approximation changes after removing **untrustworthy features** from explanation
  - Greedy or random:
    - *untrustworthy* if any **untrustworthy features** show up in explanation
- Report F1 on trustworthy predictions (averaged over 100 runs)

5.3 Should I trust this prediction?

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	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**

Results

- LIME dominates by p =0.01, on both dataset, on all models
- Other methods achieved either
  - Low recall; or
  - Low precision

  while LIME maintain high recall and precision.

5.4 Can I trust this model?

Goal:

to evaluate whether a user can identify the better classifier based on the explanations of *B* instances from the validation set

Procedure:

- Add 10 artificially "noisy" features
  - On training/validation (80/20) sets, each artificial feature appears in 10% of the examples in one class, and 20% of the other.
  - On test set, each artificial feature appears in 10% of the examples in each class.
- Create pairs of competing classifiers by repeatedly training pairs of model
  - Random forest with 30 trees
  - Their validation accuracy is within 0.1% of each other
  - But their test accuracy differs by at least 5%
- Simulated user choose which predictions are untrustworthy
  - Untrustworthy if artificial features appear in explanation
- Count # of trusted predictions

5.4 Can I trust this model?

Procedure (Continued):

- Choose the model with fewer untrusted predictions
- Compare if the choice is consistent with test set performance
  - Inconsistent if the chosen model performs worse on test set
- Present Accuracy of picking correct classifier (avrg. over 800 runs) as *B* varies

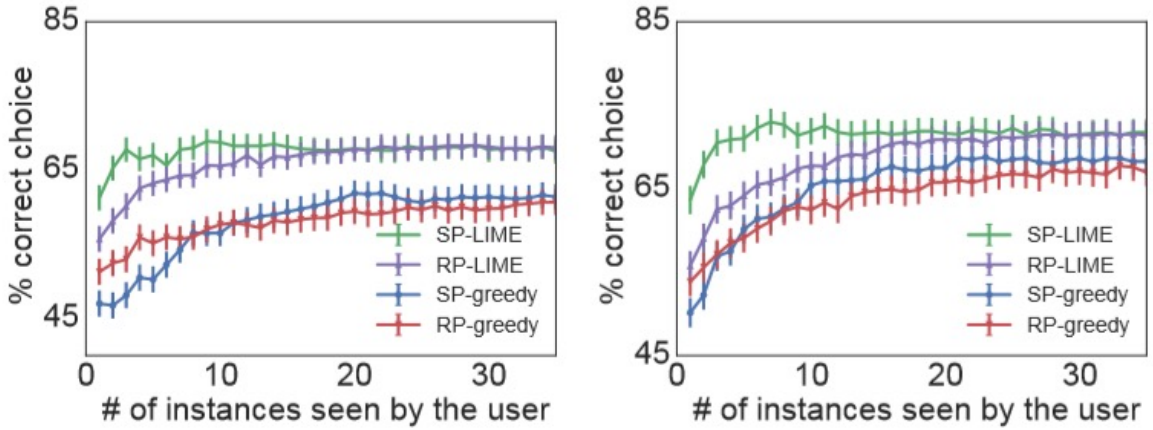


Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

6 Evaluation with Human Subjects

Goal: recreate three scenarios in machine learning that require trust and understanding of predictions and models

Three questions/situations:

- 6.2 Can users choose which of two classifiers generalizes better
- 6.3 Can users perform feature engineering to improve the model, based on the explanations
- 6.4 Are users able to identify and describe classifier irregularities by looking at explanations

6.1 Set up

For first two questions

- Data: 20 newsgroup
- Task: distinguish “Christianity” and “Atheism”
- Evaluation dataset:
  - A new *religion dataset*
  - Downloaded from DMOZ 819 websites in each class (Christianity, Atheism)
- Model:
  - SVM with RBF kernel
  - Hyperparameters turned via cross-validation

6.2 Can users select the best classifier?

Goal: to evaluate whether explanations can help users decide which classifier generalizes better.

Task: User decide between two classifiers:

- SVM w. RBF trained on 20 newsgroup dataset
  - Accuracy on test set during train/test split: 94.00%
  - Accuracy on religion: 57.3%
- Same classifier trained on a “cleaned” dataset (features that do not generalize well are removed)
  - Accuracy on test set during train/test split: 88.60%
  - Accuracy on religion: 69.0%

Human subjects:

Recruited on Amazon Mechanical Turk
(no ML experts, but with basic religion knowledge)

6.2 Can users select the best classifier?

Procedure:

- Show **B** predictions, each with **K** explanations
  - B = 6, K = 6
  - Order of showing samples are randomized
  - Explanations are produced by ***greedy*** or ***LIME***
  - Instances are selected by either ***RP*** or ***SP***
- Users examine samples
- Users asked to select which algorithm will perform better in real world, and explain why

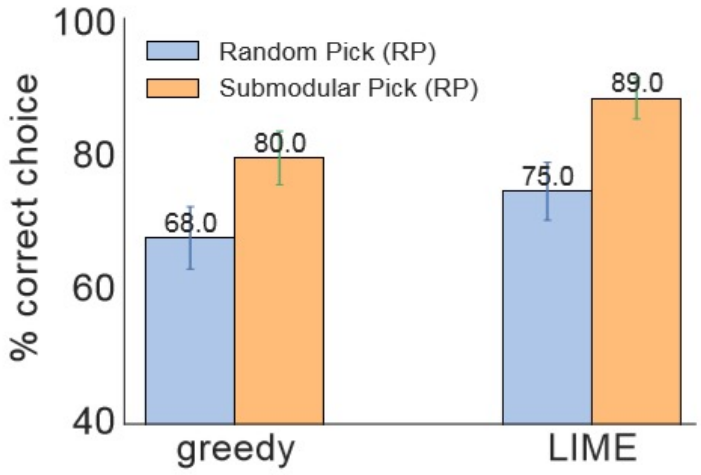


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

6.2 Can users select the best classifier?

Result:

- All models good at identifying the better classifier
  - Explanations are useful in determining which to trust
- SP performs better than RP
- User's reason for their selection:
  - If the model utilizes more semantically meaningful words

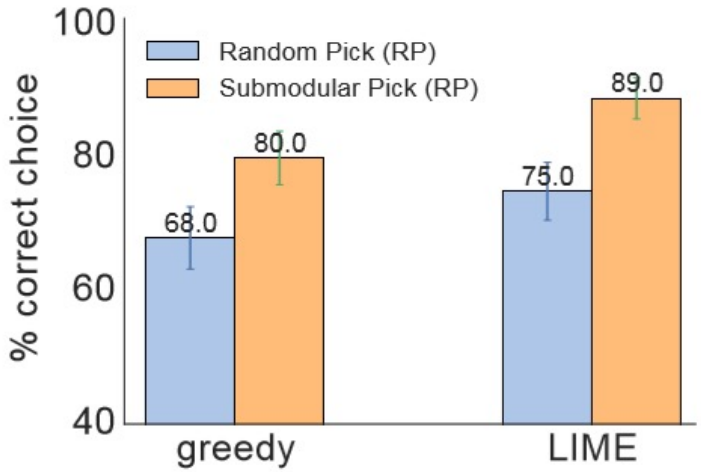


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

6.3 Can non-expert improve a classifier?

Basis: removing features that the users feel do not generalize to improve generality.

Procedure (3 rounds):

(Round 1)

- Show **B** predictions, each with **K** explanations
  - B = 10, K = 10
  - Explanation instances are produced by ***SP-LIME*** or ***RP-LIME***
- Users examine samples (10 users)
- User marks words for deletion
  - No access to religion dataset
- Train 10 classifiers for each subject (with their modification)

6.3 Can non-expert improve a classifier?

Procedure (Continued)

(Round 2)

- ***Same set up***
- 5 users examine each of the 10 new classifiers
- Resulting in 50 classifiers in total

(Round 3)

- Repeat Round 2
- Resulting in 250 classifier

- Report averaged accuracy on *religion* at each round for paths originating from the original 10 subjects (shaded line), and average across all path (solid line)

6.3 Can non-expert improve a classifier?

Procedure (Continued)

- Report averaged accuracy on *religion* at each round for paths originating from the original 10 subjects (shaded line), and average across all path (solid line)

Results:

- Non-experts can improve a classifier
- SP performs better than RP
- Explanations make it easy to improve untrustworthy model

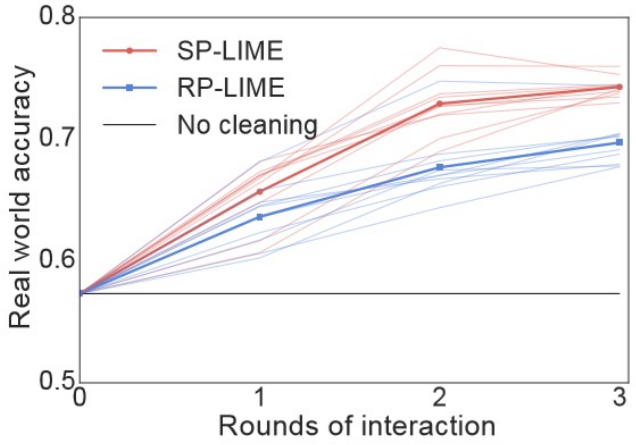


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

6.4 Do explanations provide insights?

Problem: undesirable correlations that the classifiers pick up during training are difficult to identify by looking at raw data and predictions.

Set up:

Dataset: photos of Wolves and Huskies (Eskimo Dogs)

- Training set: 20 images (manually-selected)
- All wolves pictures have snow in the back
- All huskies pictures do not

Features: max-pooling layer of Google's Inception NN

Human subjects: graduate students who have taken at least one graduate machine learning course

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

6.4 Do explanations provide insights?

Procedure:

- Present to user a balanced set of 10 test predictions w/o explanations
  - Where one wolf is not in snowy background and one husky is
  - Other 8 examples are classified correctly.
- Ask the subject the following questions
  - Do they trust this algorithm to work well in the real world, (2)
  - why, and (3)
  - how do they think the algorithm is able to distinguish between these photos of wolves and huskies.
- Show the subjects samples w/ explanations
- Ask the same questions
- Have 3 independent evaluator read the responses and determine if each subject mentioned snow, background, or equivalent as a feature the model may be using

6.4 Do explanations provide insights?

Procedure (Continued):

- Report the majority to decide whether the subject was correct about the insight

	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.

Results:

- Changes number of subjects who noticed in snow pattern
- Drop in trust in classifier
- Demonstrates the utility of explaining individual predictions for getting insights into classifiers knowing when not to trust them and why.

Conclusion and Future Work

Conclusion & Summary:

- Trust is crucial for effective human interaction with machine learning systems
- Explaining individual predictions is important in assessing
- Proposing LIME for a modular and extensible approach to faithfully explain the predictions of any model in an interpretable manner.
- Introducing SP-LIME for selecting representative and non-redundant predictions
- Explanation are useful for trust-related tasks

Conclusion and Future Work

Future Work:

- Did not mention ho to perform pick step for images
- More exploration in domain and model agnosticism, and its application in speech, video and medical domains
- Exploration in theoretical properties and computational optimizations

References

[1] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In Human Factors in Computing Systems (CHI), 2015.

[2] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. M¨uller. How to explain individual classification decisions. Journal of Machine Learning Research, 11, 2010.

[3] A. Bansal, A. Farhadi, and D. Parikh. Towards transparent systems: Semantic characterization of failure modes. In European Conference on Computer Vision (ECCV), 2014.

[4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Association for Computational Linguistics (ACL), 2007.

[5] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. Dataset Shift in Machine Learning. MIT, 2009.

[6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Knowledge Discovery and Data Mining (KDD), 2015.

[7] M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. Neural information processing systems (NIPS), pages 24–30, 1996.