The article *"Why Should I Trust You?" Explaining the Predictions of Any Classifier* by Ribeiro et al. (2016) aims to solve the problem of trusting a machine model, which are largely considered as a blackbox nowadays. In the article, the group of authors describe the problem and explains the needs for users to understand the machine learning models. Then, characteristics for an explainer are presented and two explanation algorithms, which are Locally Interpretable Mode-agnostic Explanation (LIME) and Submodular Pick – LIME (SP-LIME), are proposed for the problem. These algorithms are demonstrated with text model and image classification models for their usefulness. The article concludes with its limitation and potential future works.

In the beginning, the authors explain why the needs for understanding machine learning models exist. Several reasons are listed as (1) establishment of trust from users on both model and prediction, (2) possibility to improve untrustworthy models, (3) better model selection and (4) insight to prediction of model. Decisions makers can make decisions with understanding, rather than blind trust and have confidents when deploying the model to work with real world data.

When the need for model explanation is clear, authors propose the set of characteristics for an explainer. There are four characteristics that an explainer should have, which are (1) **interpretable**, (2) **local fidelity**, (3) **model-agnostic** and finally (4) **global perspective**. An explainer should be able to provide qualitative understanding between input and output of the model. Besides, the explanation should correspond to how the model behaves in the vicinity of the instance being predicted. The third characteristic presents the aim of authors to have an explainer that able to explain any classification model. Lastly, the explainer should provide global perspective for users, rather than just a single accuracy metric, which is not the goal for every application area.

Based on the characteristics, two algorithms are proposed: LIME and SP-LIME. LIME is an explanation technique to explain a prediction of any classifier. The goal of this technique can be presented by the equation

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

where:

- x is the original representation of an instance being explained
- g is the explanation model in quest
- G is a class of potentially interpretable models
- f(x) is the probability that x belongs to a certain class
- $\boldsymbol{\pi}_x(z)$ is a proximity measure between an instance z to x, so as to define locality around x
- $\mathcal{L}$ is locally weighted square loss, which measures how unfaithful g is in approximating f in the locality defined by $\boldsymbol{\pi}_x$
- Ω(g) is the complexity of g.

The variable "g" should satisfy the equation to minimize $\mathcal{L}$ to ensure **local fidelity** characteristic, while keeping Ω is low enough for **interpretability**. The **model-agnostic** characteristic is allowed by not making any assumption about f, but approximate its behavior by recovering the randomly selected *interpretable representation* to *original representation.*

The explanation model g can explain a single prediction x for better understanding, but not enough to assess trust as a whole i.e.: lacking **global perspective** characteristic. SP-LIME is proposed as a method to select a set of representative instances with explanations to address this trusting problem, via submodular optimization. The method can be represented by

$$Pick(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}} c(V, \mathcal{W}, I)$$

where:

- W is the explanation matrix that represents the local importance of the interpretable components for each instance. E.g.: when using linear model g as explanation that $g_i = \xi(x_i)$, then $W_i = w_i$
- B denotes the number of explanation that humans are willing to look at to understand the model.
- I_j denotes the global importance of the feature j in explanation space. I_j should have higher value if such feature j can explain more instances.
- V denotes the finding set of B instances for user to inspect.
- c is a set function define coverage that computes the total importance of the features that appear in at least one instance in a set V.

These two algorithms are demonstrated with simulated user experiment and with human subjects. The goal of these demonstrations is to answer three questions

1. Are explanations faithful to the model i.e.: higher recall %
2. Should I trust this prediction i.e.: higher F1 of trustworthiness
3. Can I trust the model i.e.: higher % of correct classifier selection?

In the case of user experiment, a classifying task for sentiment analysis challenges decision tree, L2 regularization, nearest neighbors, random forests, support vector machines models. Those models go through simulated user experiment process with explanations LIME and other methods including parzen, greedy and random. The results indicate that LIME and SP-LIME outperform other methods in all three questions.

The second demonstration is done with human subjects for the three questions:

1. Can users choose which of two classifiers generalized better?
2. Based on explanations, can users perform feature engineering to improve the model
3. Are users able to identify and describe classifier irregularities by looking at explanations?

The first two questions are done with classifying tasks for sentiment analysis while the third one is challenged with image classification task. The results for the first two scenario show the preference of SP-LIME in helping humans to choose better classifier and make improvement on a bad model. The third experiment shows that prediction explanation can provide insights into classifiers whether to trust them or even why.

In the conclusion, the authors listed out several limitations for the work and at the same time identify potential future works. They include

1. Demonstration is limited for sparse linear model only for LIME algorithm so more demonstration from real user for other explanation families e.g.: decision tree.
2. Applying SP-LIME algorithm for images due to the limitation that global importance I must work across super-pixels in different images, such as histograms.
3. Demonstration in more application areas, such as speech, video, medical and recommendation system
4. Theoretical properties study e.g.: Number of samples
5. Computational optimization for real time usage

Ribeiro M. T., Singh S. & Guestrin C. (2016). *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. https://doi.org/10.1145/2939672.2939778