Comparing Different Feature Importance Methods

Searidang Pa

Motivation / Problem Statement

What is the problem you want to solve?

- Methods that calculate feature importances such as SHAP and LIME could give an intuitive explanation of a model's prediction at an instance level. However, as this paper has shown, they could disagree with each other. I hope to build visualization tools that could help the data scientists gain a more granular understanding of how the explanation agree and disagree with each other.

Who has this problem

- Data scientists and ML developer who are trying to explain their models to stakeholders. A survey done by this paper shows that most practitioners use more than one explanation methods to explain their model's prediction.

Why is it relevant / interesting?

- I think it is important to build tools that could help assist data scientists and ML developer gain a deeper understanding of when the explanation methods agree and disagree. It could give the data scientists more confidence when more than one explanation methods agree in their account of how the features contribute to the prediction, and it could serve as a warning to use the explanation with caution in the case where the explanation disagree strongly.

Background / Related Work

What have others done?

- This paper by Krishna et al shows that feature importance explanation methods often disagree. They design different evaluation metrics to gauge how similar two explanation methods are on average.

What can you learn from existing efforts?

- I think I could draw inspiration from the SHAP feature contribution summary plot and SHAP scatter dependence plot.

How is your work related and different from existing works?

- Rather than just focusing obtaining a single scalar value that roughly compute how similar two explanations are on average, I wish to take a more granular look at when and how the explanations diverges.

Data / Model(s)

What kind of data will you use?

- Tabular data: Pima diabetes dataset

What kind of models and ML techniques will you use?

- Model: Random Forest
- Explanation Methods: SHAP, LIME, Smooth Gradient, Integrated Gradient

Are these data and models available or you still have to find or produce them (better to have them already)?

- There are some missing values, so I would have to use some heuristic to do data imputing such as filling in the missing value with the median by the output class.

Tasks / Analytical Questions

Feature Contribution Summary:

- **T1:** What are top k most important features, and do they contribute positively or negatively on average for each method?
- **T2:** For each method, what is the relationship and the distribution of the feature importances as the value of the feature varies for each feature?

Feature Contribution at Instance level:

- **T3:** For each method, for a selected instance, what are the relative importances of each feature for the prediction? How much do the order and the sign of the contribution of the top k most important features for this instance agree across the various methods?

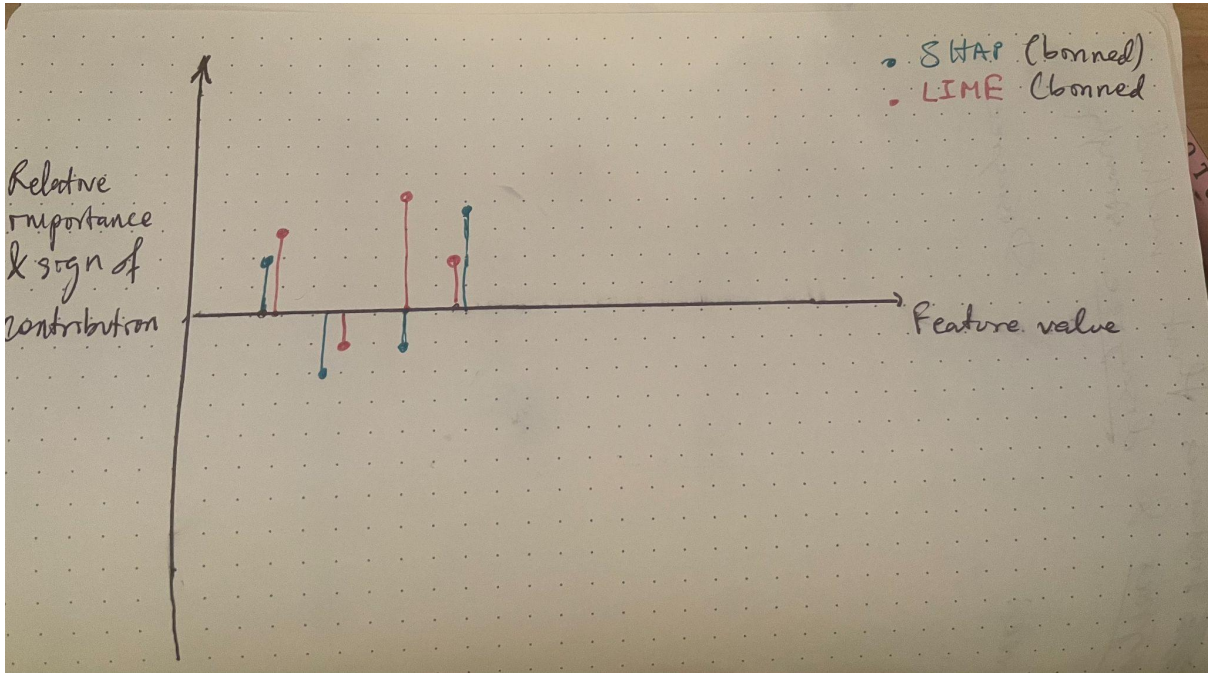
Stability of Explanation:

- **T4:** For each method, does the feature importances change dramatically for instances with similar input values?

Clustering Instances With Similar Explanation and Instances with Diverging Explanation

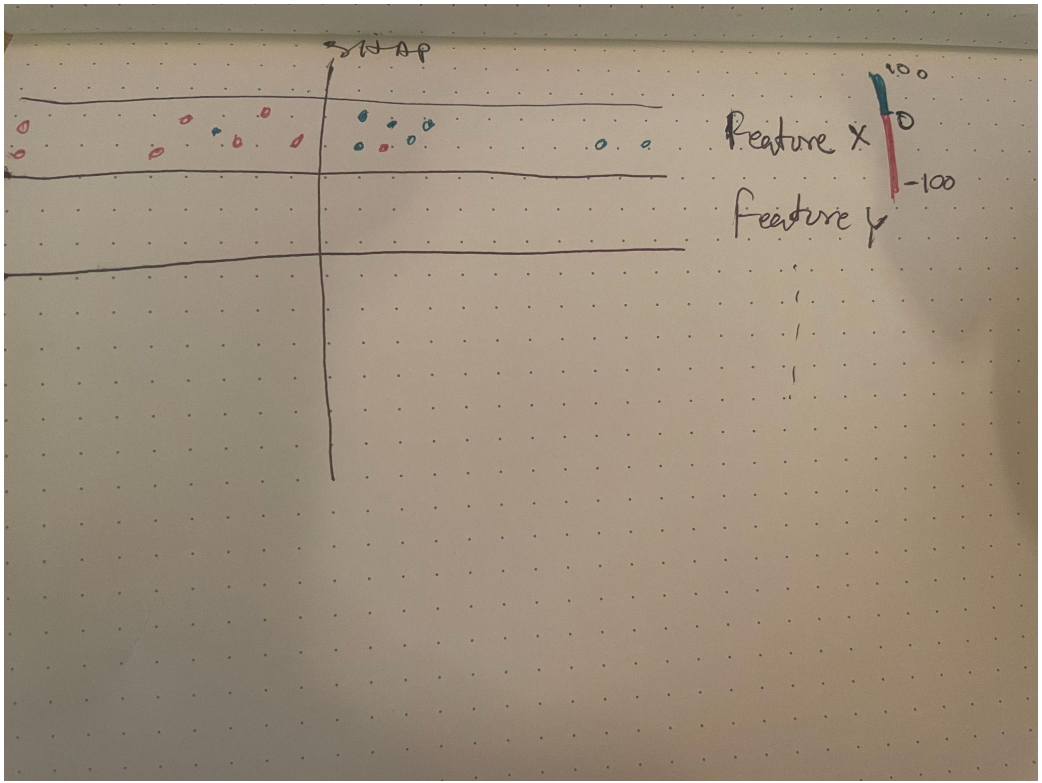
- **T5:** For each pair of methods, what do the instances that the methods' explanation agree and disagree the most look like?

Mock-Ups: Aim to solve T1



T1: For each method, what is the relationship and the distribution of the feature importances as the value of the feature varies for each feature?

Mock-Ups: For Contribution Summary Task (T1 & T2)

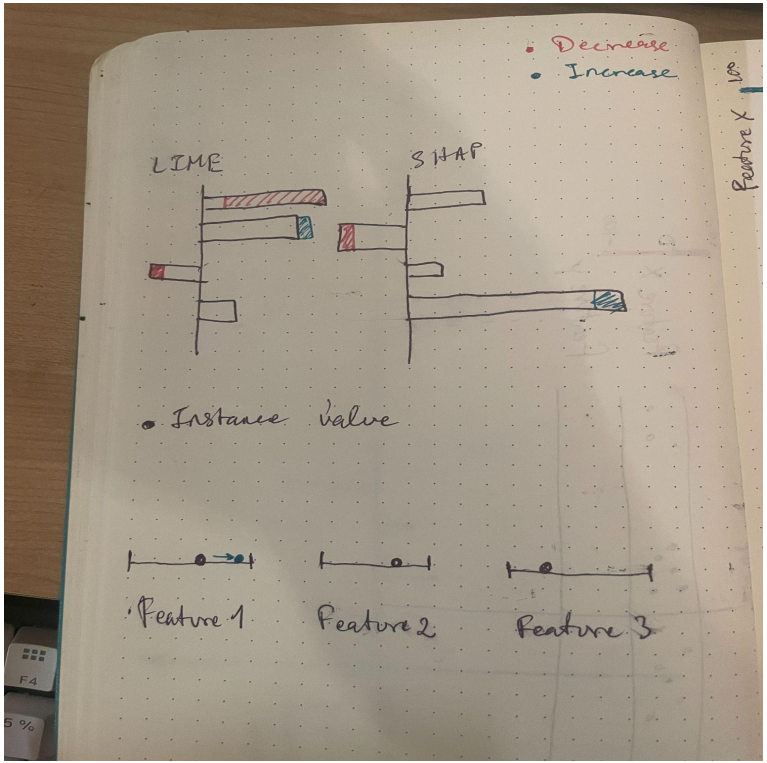


T1: For each method, what is the relationship and the distribution of the feature importances as the value of the feature varies for each feature?

T2: What are top k most important features, and do they contribute positively or negatively on average for each method?

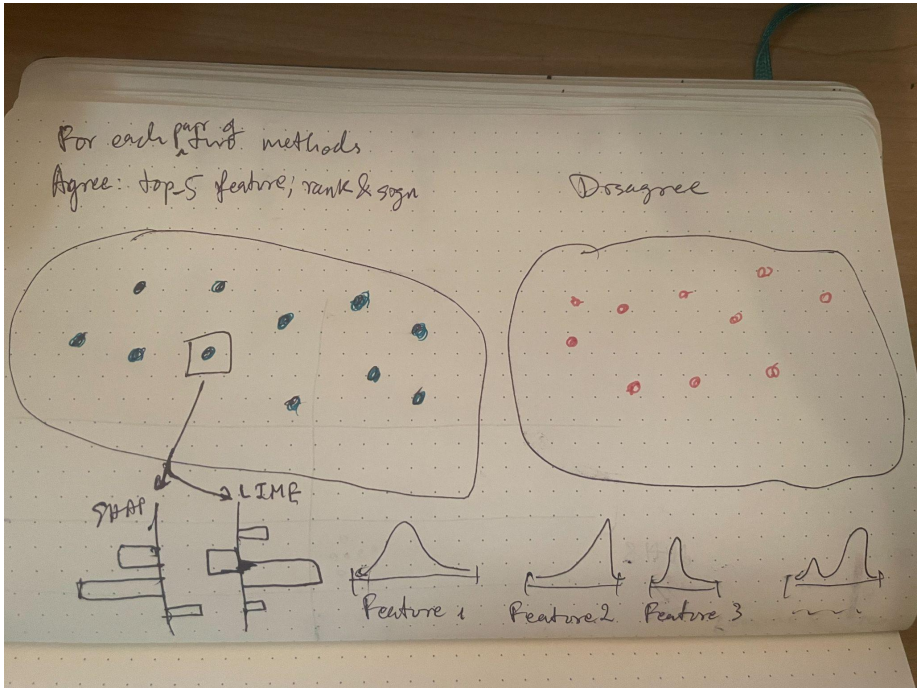
(Feature Contribution Summary Chart for each method)

Mock-Ups: Aim to solve T3& T4



- **T3:** For each method, for a selected instance, what are the relative importances of each feature for the prediction? How much do the order and the sign of the contribution of the top k most important features for this instance agree across the various methods?
- **T4**: For each method, does the feature importances change dramatically for instances with similar input values?

Mock-Ups: Aim for T5



T5: For each pair of methods, what do the instances that the methods' explanation agree and disagree the most look like?