Comparing Different Local 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.

Data / Model

What kind of data will you use?

- Tabular data: Pima diabetes dataset

What kind of models and ML techniques will you use?

- Model: Neural Network
- Explanation Methods: SHAP, LIME, Integrated Gradient (IG), DeepLift

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.

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.

Tasks / Analytical Questions

Feature Contribution Summary:

- **T1:** What are top k most important features, and how do they contribute positively or negatively for each method?
- **T2:** What is the the distribution of the feature importances vs. the feature value varies for each method?

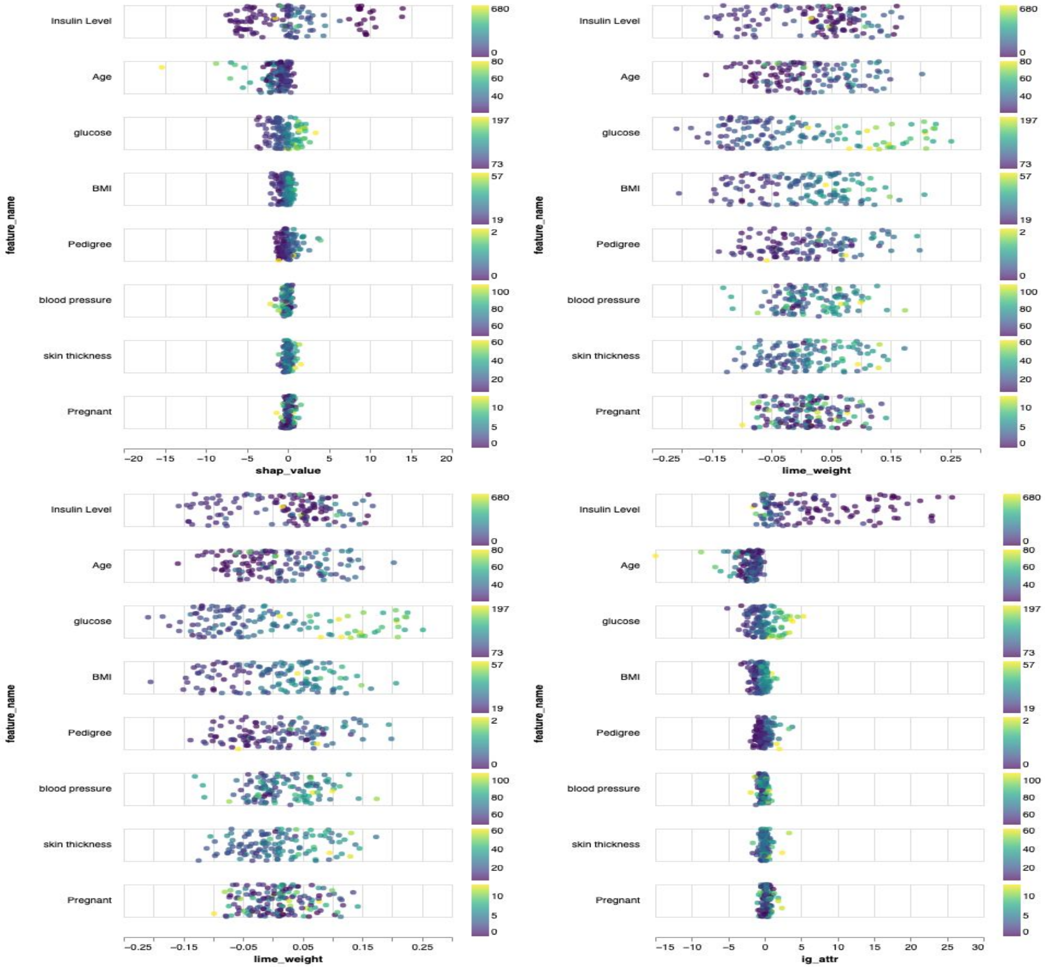
Feature Contribution at Instance and subset level:

- **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?
- **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

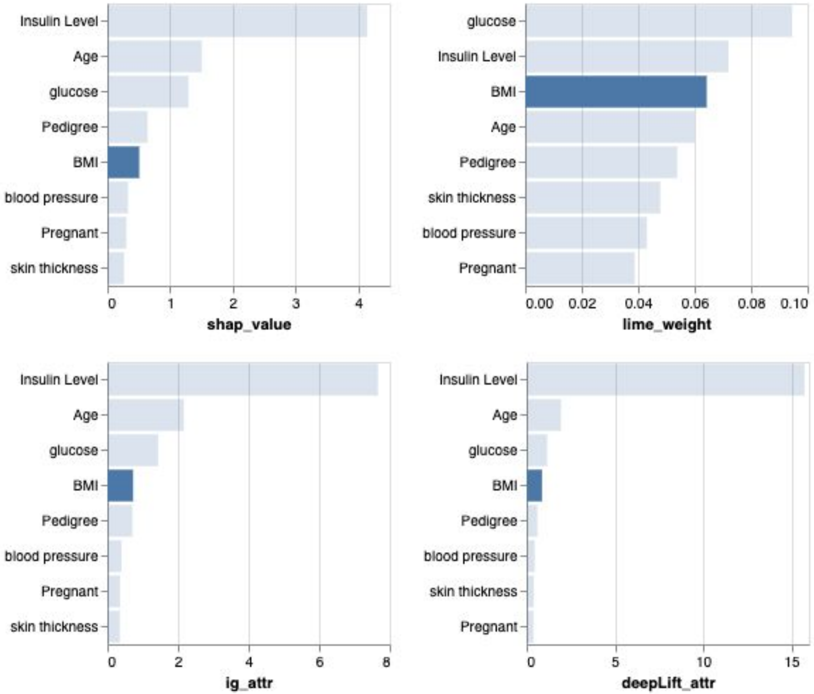
- **T5:** For each pair of methods, what do the instances that the methods' explanation agree and disagree the most look like?

What not to do
Summary Strip Plot

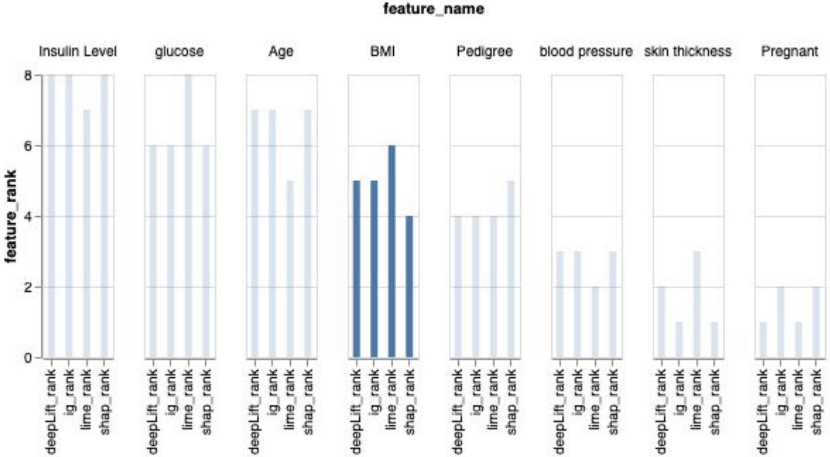


T1: What are top k most important features for each method?

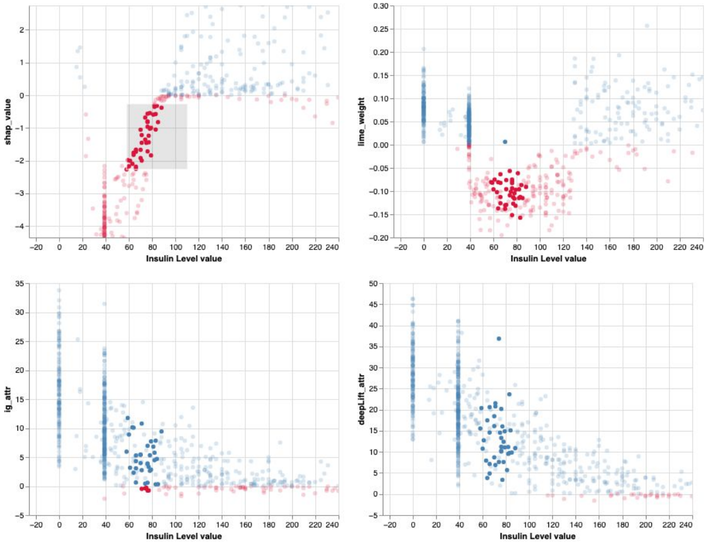
Mean Absolute Value of Feature Attribution

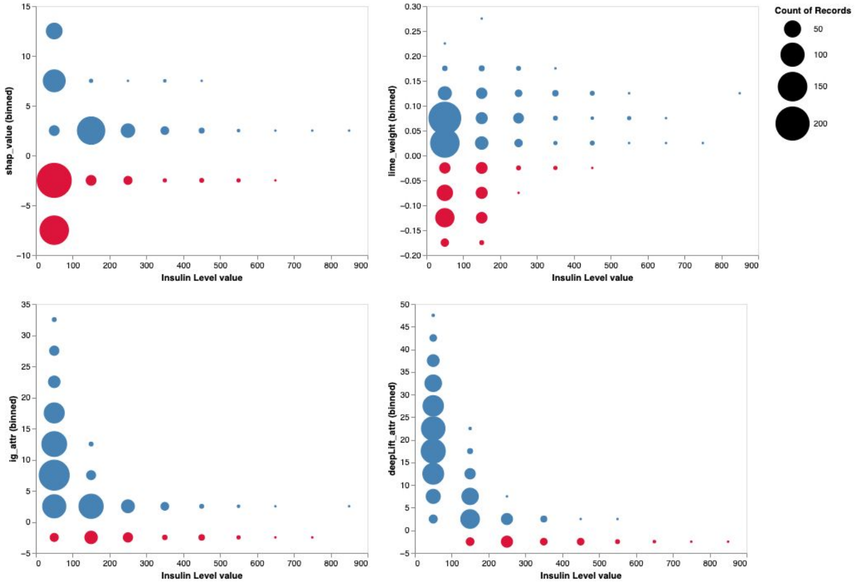


Mean Absolute Value of Feature Rank

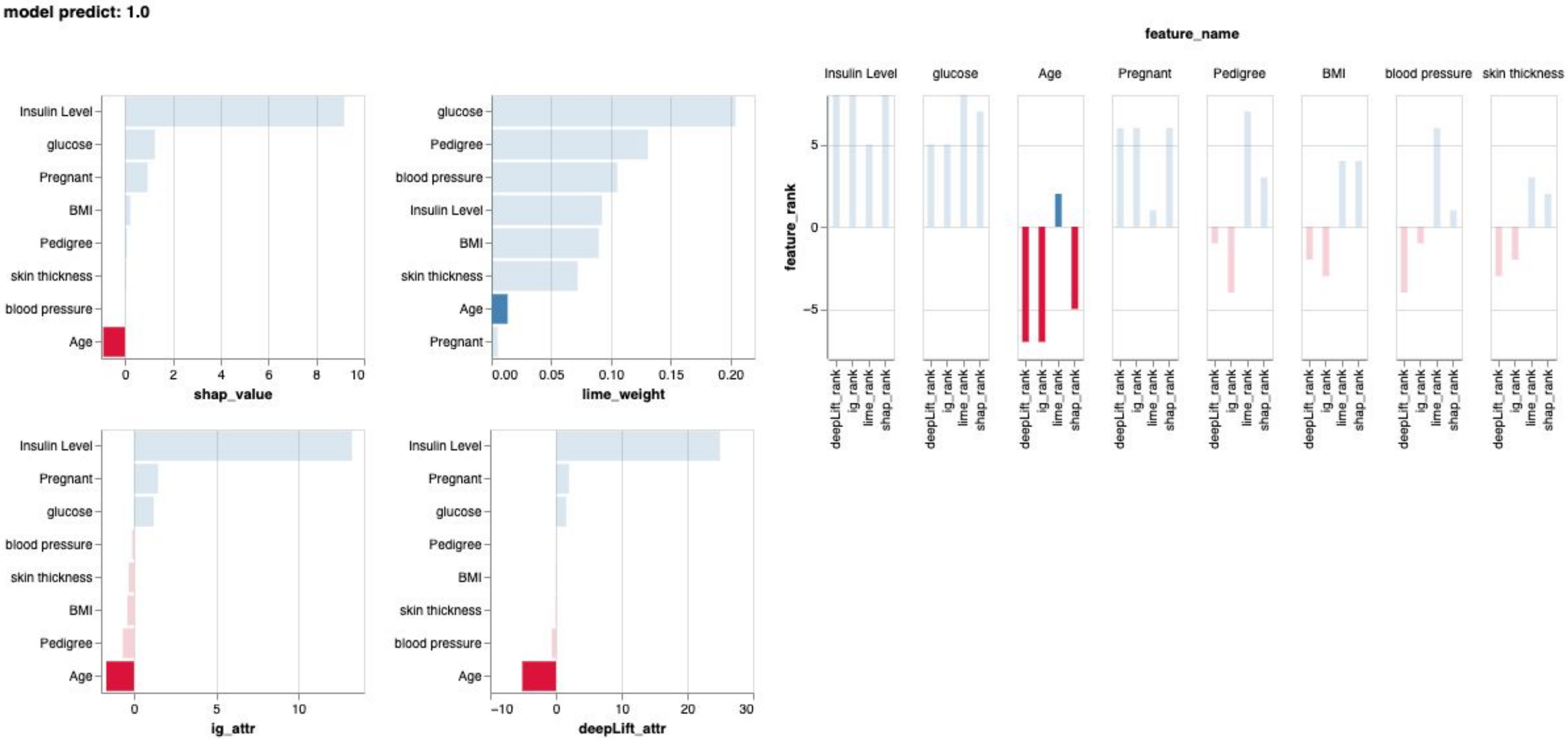


T2: How does features contribute positively or negatively for each method?

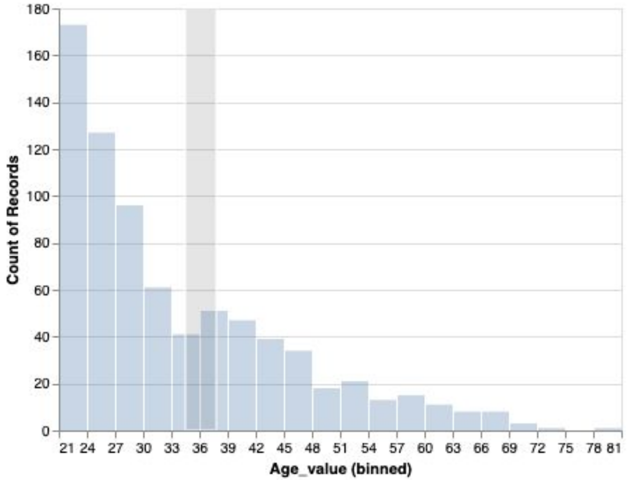


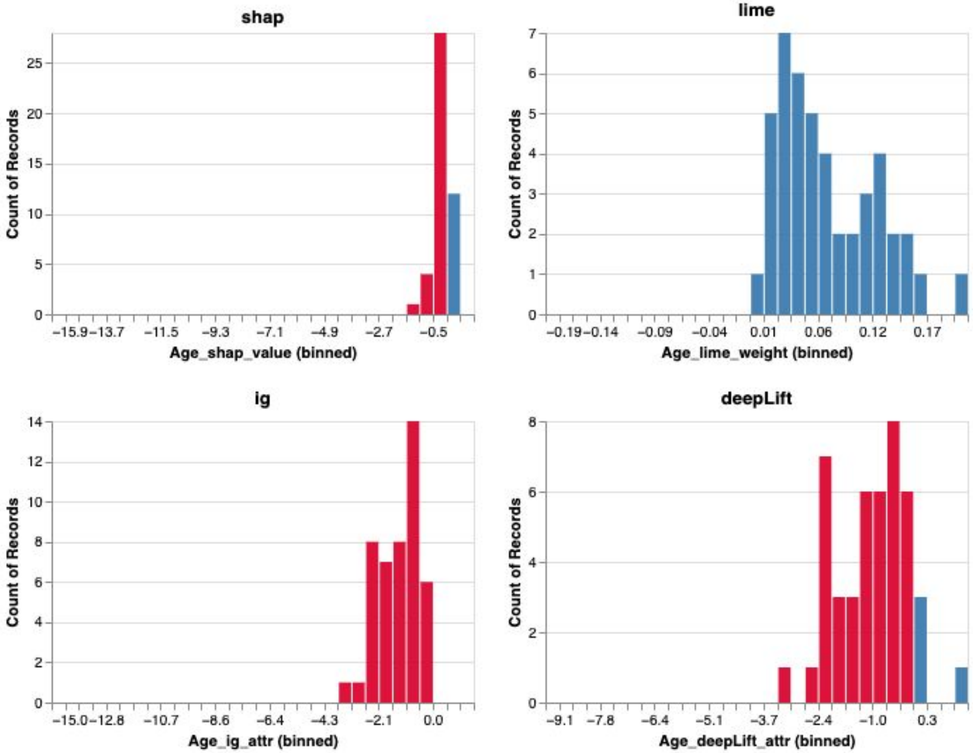


T3: Individual Data Point: for each method, what are the relative importances of each feature for the prediction? How much do the order and the sign of the contribution of the top k most important features for this instance agree across the various methods?

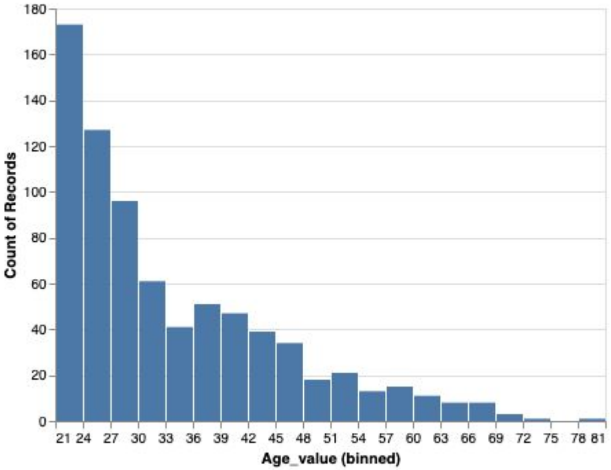


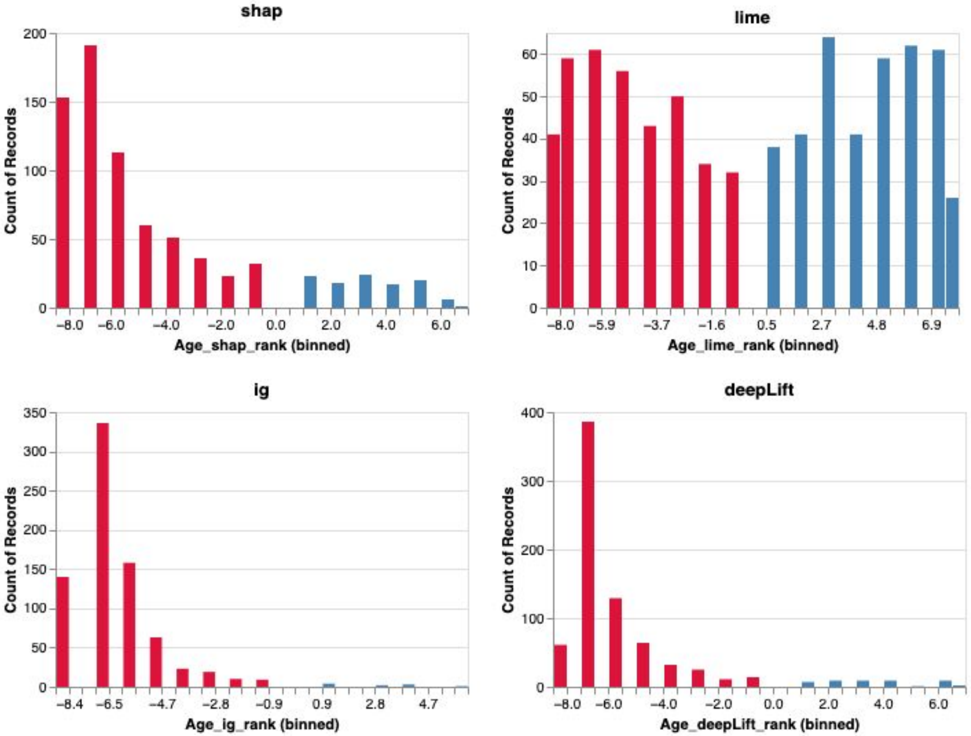
Subset Analysis: T2: What is the the distribution of the feature importances vs. the feature value varies for each method?



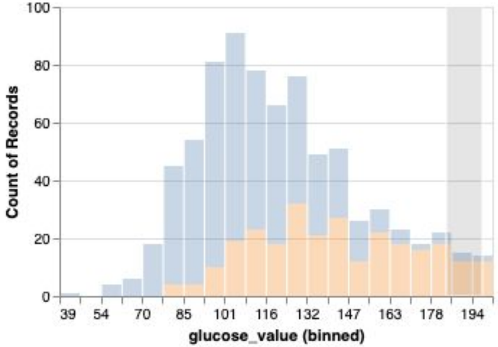


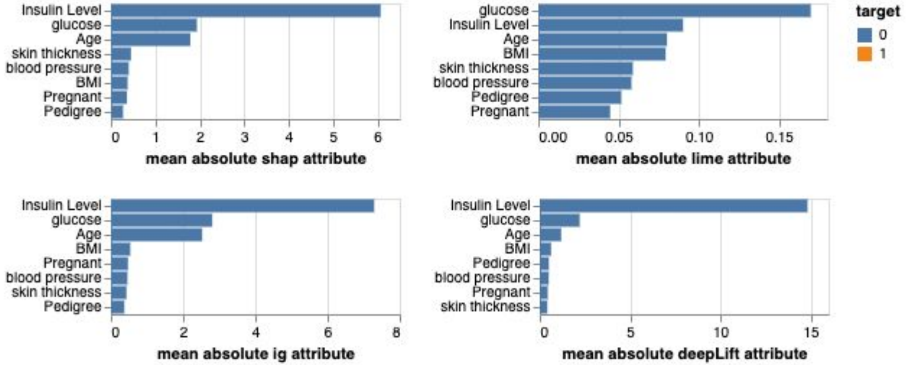
Subset Analysis: T2: What is the the distribution of the feature importances vs. the feature value varies for each method? Signed Feature Rank

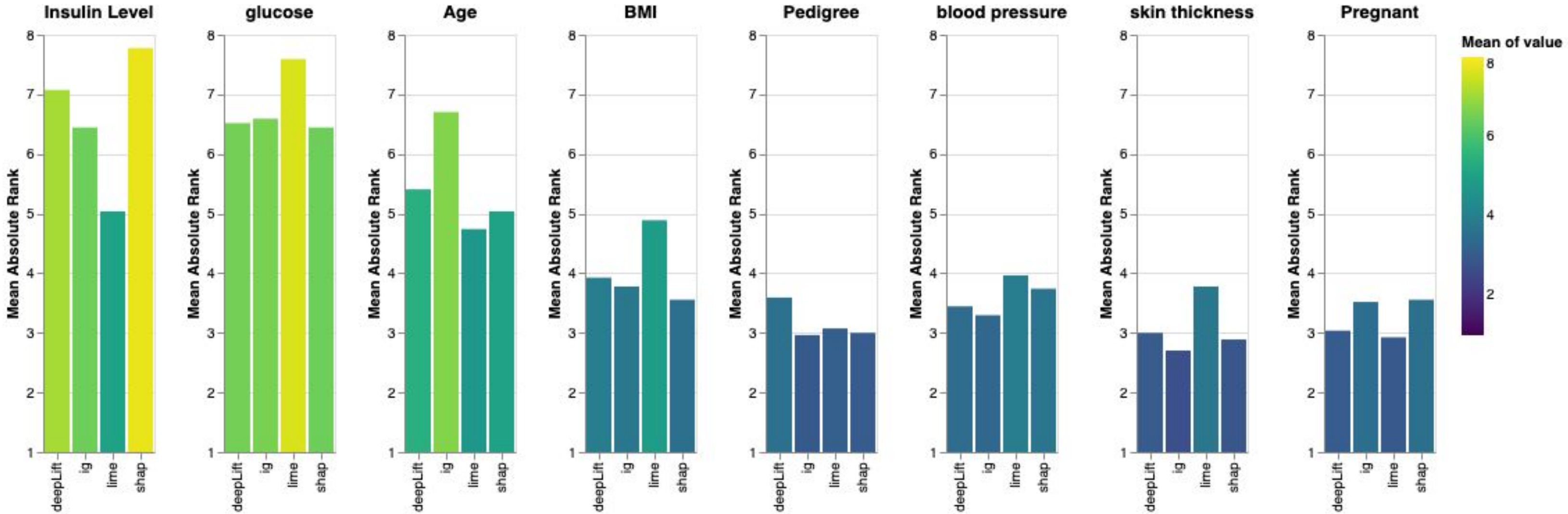




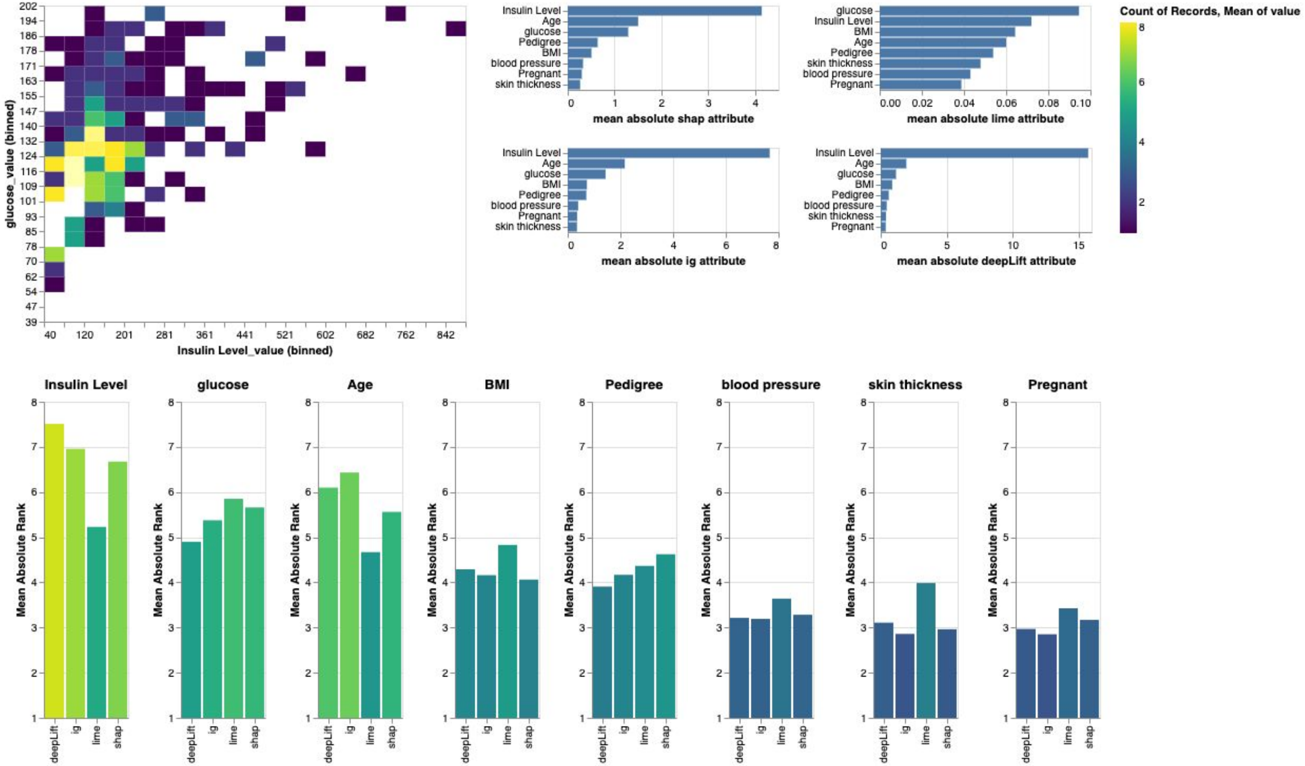
Subset Analysis: T2: feature value vs. All feature importances



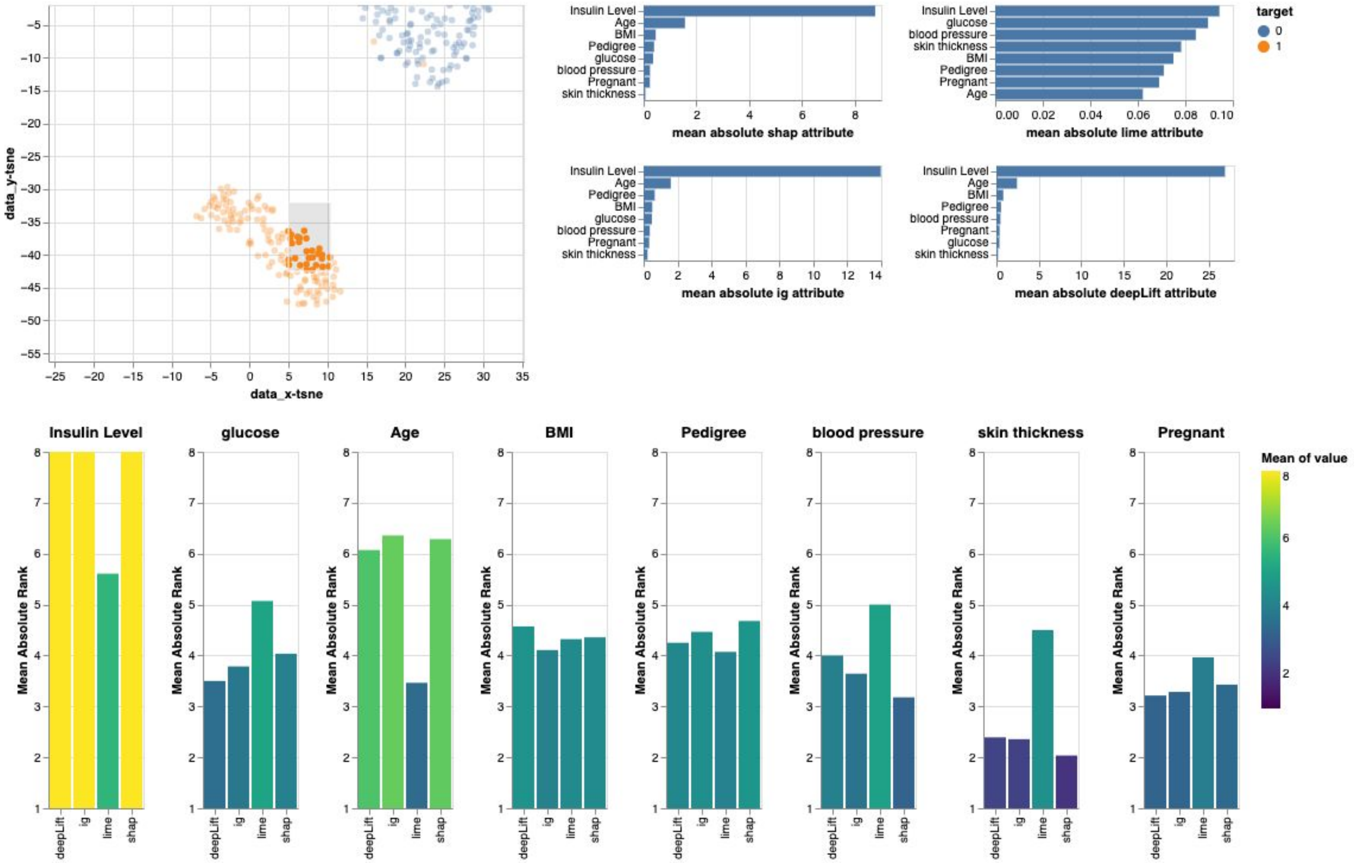




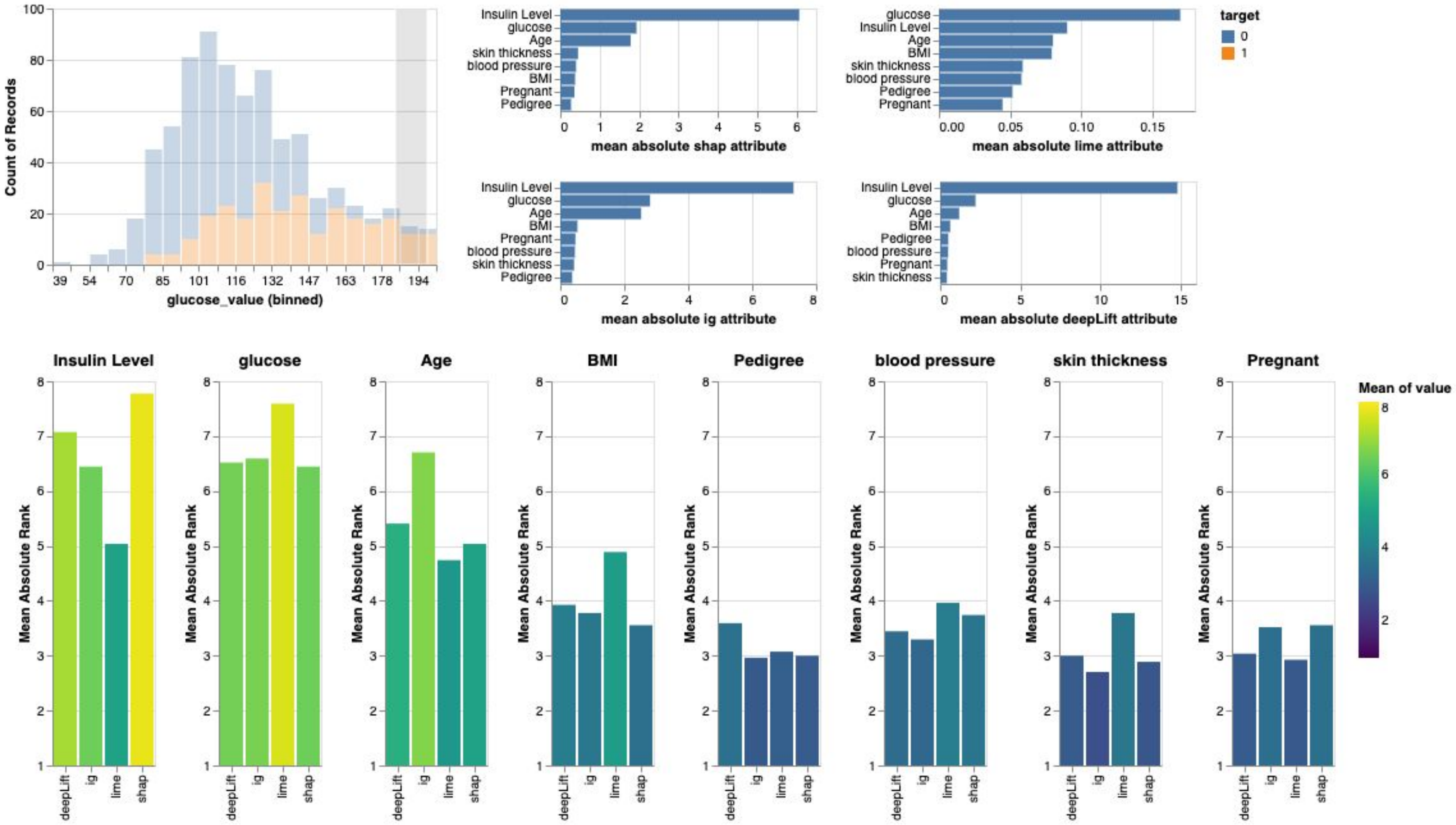
Subset Analysis: T2: What is the the distribution of the feature importances vs. the feature value varies for each method? Re: Heatmap



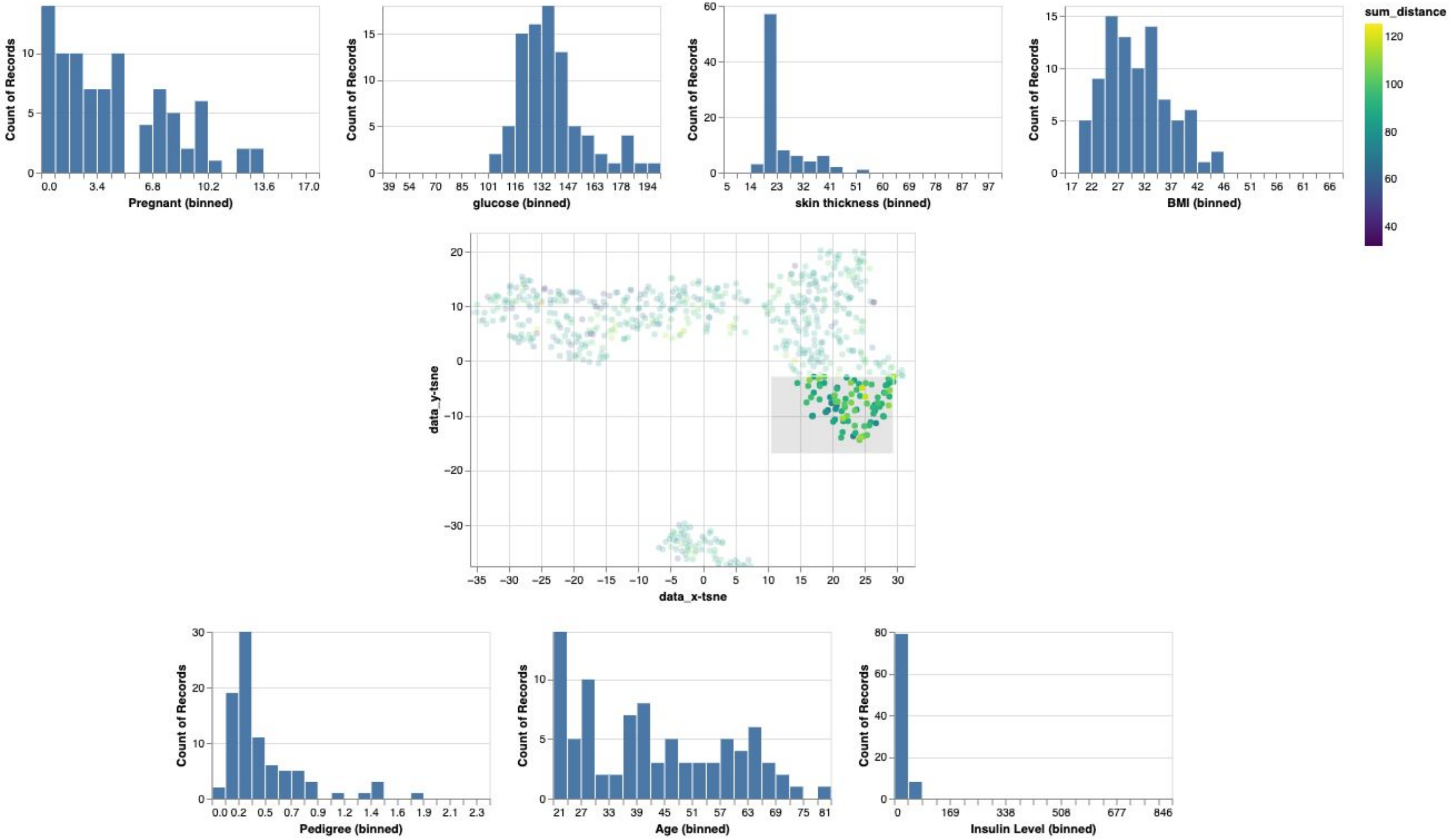
T4: For each method, does the feature importances change dramatically for instances with similar input values?



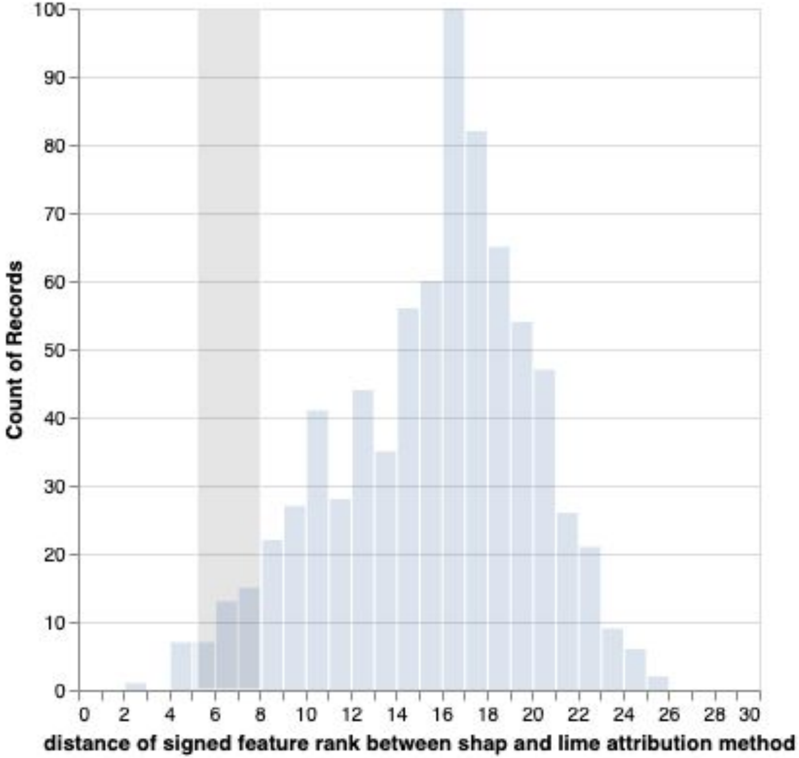
T4: For each method, does the feature importances change dramatically for instances with similar input values?

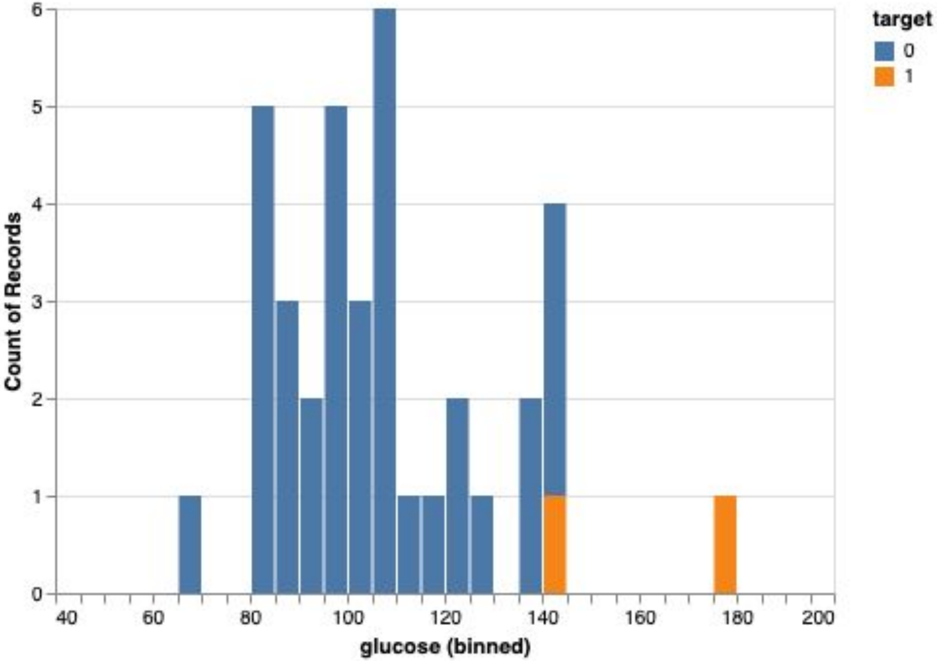


T5: For each pair of methods, what do the instances that the methods' explanation agree and disagree the most look like?

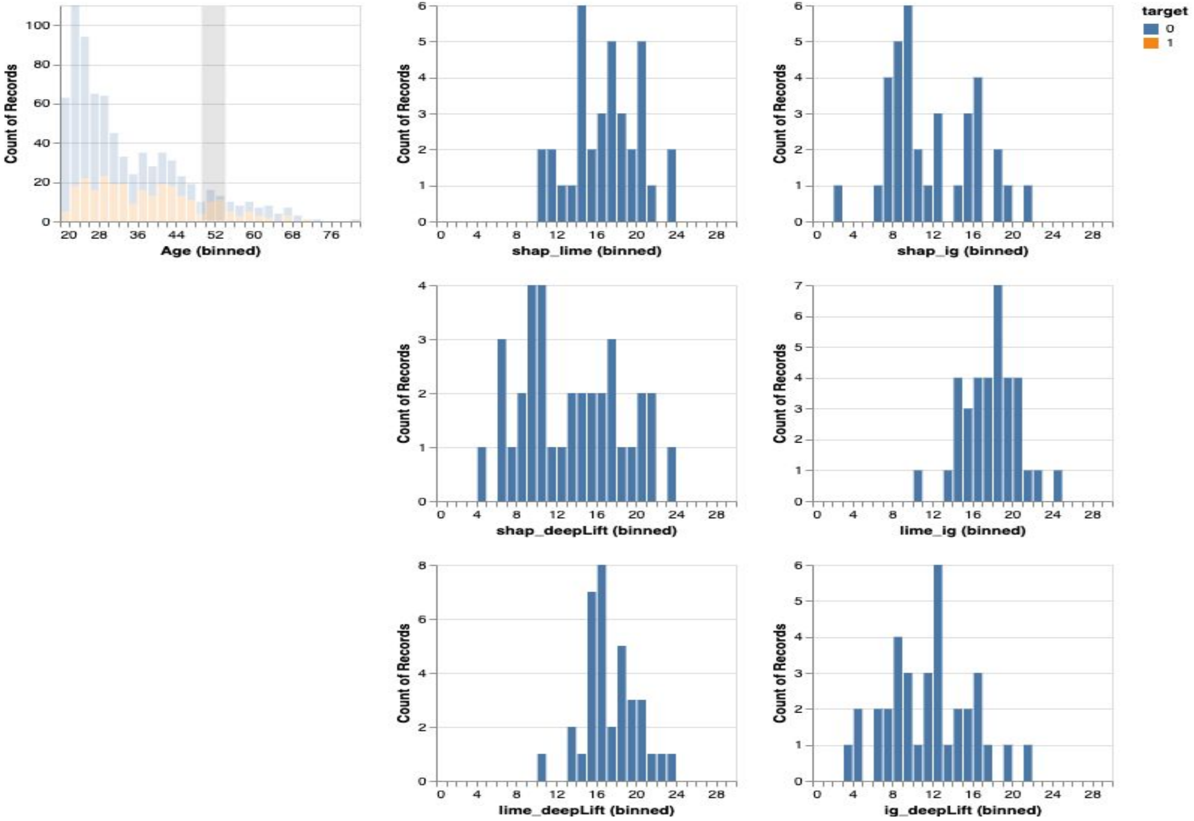


T5: For each pair of methods, what do the instances that the methods' explanation agree and disagree the most look like?





T5: Distribution of Feature value vs. Signed Rank distances



Result: What are the results of applying your solution to the problem stated above?

- Enable more granular understanding of how the various feature attribution method agree or disagree on a subset level
    - On the whole dataset, the aggregate metric might agree, but on a subset level, they might differ significantly

- Helps answer T1- T5 to some extent

Future Work

- More controls and customization for the users
    - Number of top features to be shown
    - How many and which feature attribution methods to be shown
    - Which feature attribution methods to be taken into account for sum of distance on t-sne plot
    - Let the user combine and cross reference various charts
- Support for visualizing the various metrics on nearest neighbors
    - **Stability of Explanation**: For each method, does the feature importances change dramatically for instances with similar input values?
- Scaling to more than 4 attribution methods
    - **Inspiration from Emblaze**: Have a tooltip recommend interesting regions of data between the various methods
- User-case study to understand the effectiveness of different designs. E.g. a simple box-plot be more intuitive and powerful to understand the various distribution?

Lesson Learned

- Altair is an empowering tools, but it does not provide customization at a lower granular level that is required for more complex interaction.
- The problem of choosing appropriate baseline for gradient methods is hard to address
- It can be difficult to draw any explainable conclusion from feature attribution alone. Similar Interpretability tools need to be combined with each other.