Interpretability of Machine Learning

Konstanty Kraszewski

May 26, 2024

1 Introduction

As machine learning models become more and more popular, their interpretability becomes crucial. Explainable artificial intelligence (XAI), focuses on making the decision-making processes of ML models understandable to humans. This is particularly vital in sensitive areas like healthcare, finance, and biology, where understanding the reasoning behind predictions can foster trust and facilitate more informed decisions.

2 Why XAI?

The need for interpretability arises from several key concerns:

- **Trust**: Stakeholders need to trust the models to accept their decisions.
- **Compliance**: In regulated industries, it is essential to comply with legal requirements by providing explanations for automated decisions.
- **Debugging**: Interpretability helps in diagnosing and fixing issues within models.
- **Insight**: Gaining insights into the models' behavior can reveal novel scientific findings or business insights.

3 Overview of XAI methods

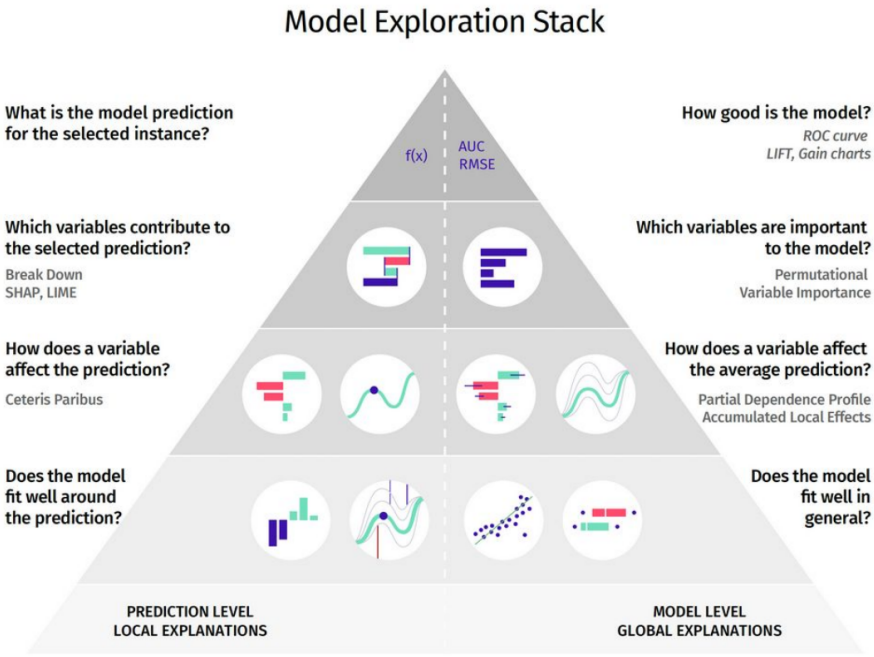

Figure 1: Visualization of different XAI methods.

As you can see in Figure 1, there are many different methods used to explain the output of the model. They are so called 'model agnostic', i.e. they can be used with different kinds of ML models to delve deeper into their results.

4 Local Explanations

Local explanations focus on individual predictions, explaining the output in a specific case. Key methods include:

- **Break Down**: This method deconstructs the prediction into contributions of each feature, as in Figure 2.
- **SHAP (Shapley Additive Explanations)**: SHAP values provide a unified measure of feature importance, as in Figure 3.
- **Ceteris Paribus Profiles**: These profiles show how changing one feature at a time affects the prediction, holding other features constant, as in Figure 4 and 5.

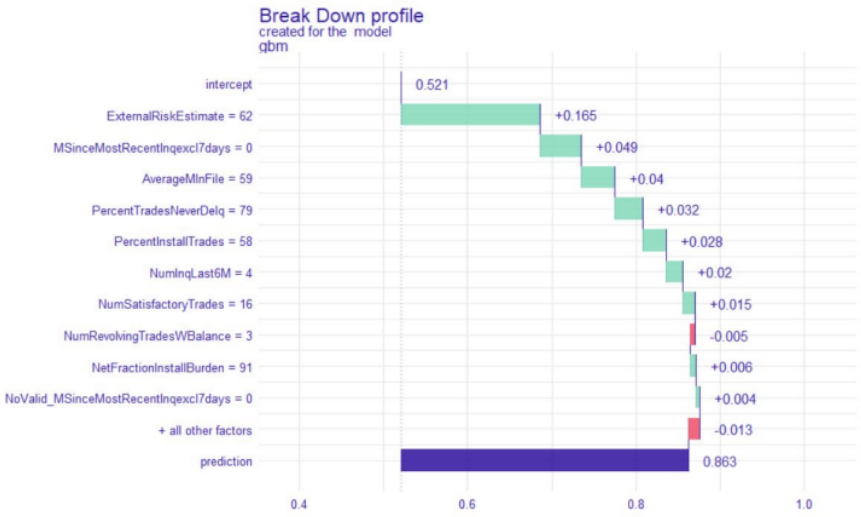

Figure 2: Example of a Break Down profile.

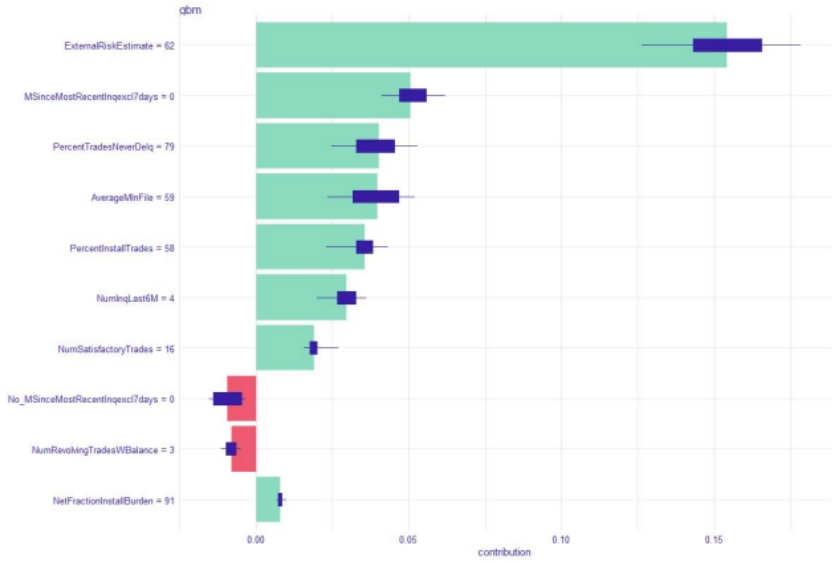

Figure 3: Example of SHAP values for a single prediction.

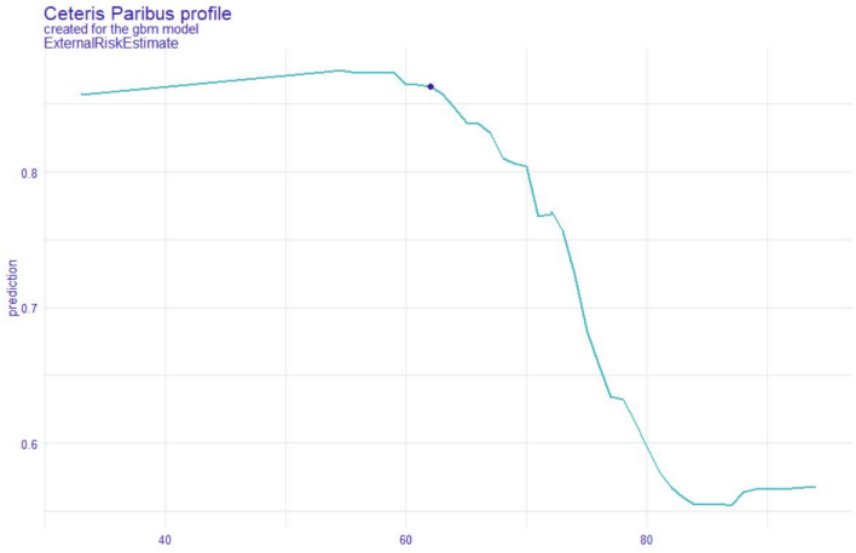

Figure 4: Example of a Ceteris Paribus profile for one feature.

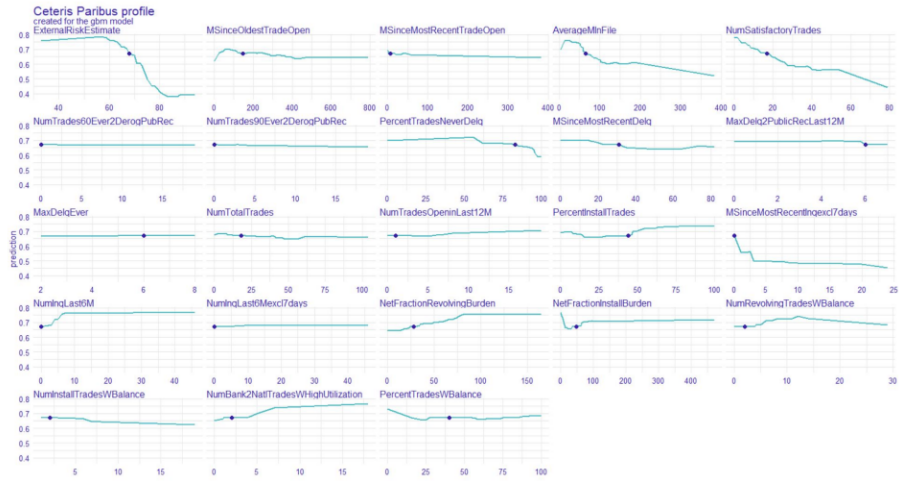

Figure 5: Ceteris Paribus profiles for all features.

5 Global Explanations

Global explanations provide insights into model behavior across the entire dataset:

- **ROC and LIFT Curves**: Visualize model performance and effectiveness, an example of a LIFT curve is presented in Figure 6.
- **Permutation Variable Importance**: Measures changes in model performance by permuting feature values and checking the changes in the model's outputs, a plot of importances is in Figure 7.
- **Partial Dependence Plots (PDPs)**: Illustrate the relationship between a feature and the predicted outcome, averaged over the dataset, visible in Figure 8.

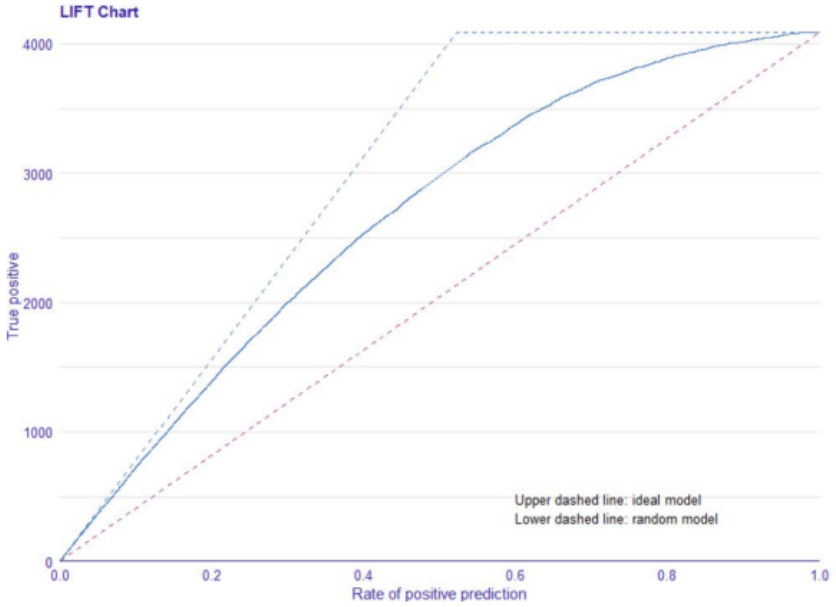

Figure 6: Example of a LIFT curve.

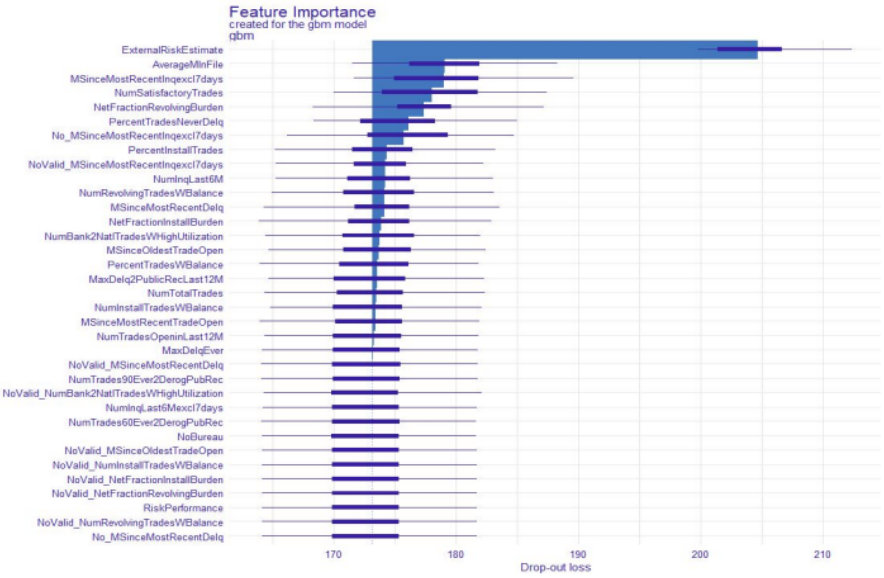

Figure 7: Example of a feature importance plot.

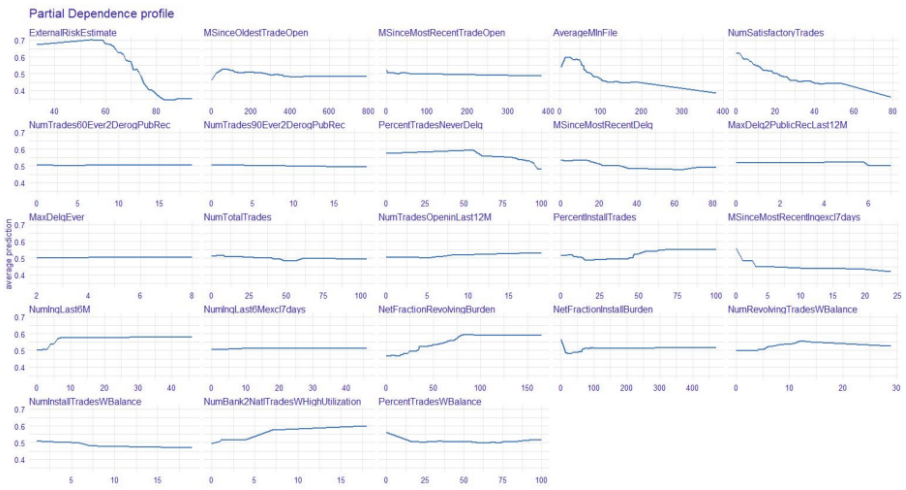

Figure 8: Partial Dependence Profiles for all features.

6 Conclusion

Interpretability of ML models becomes more and more in modern times, when we encounter such models on a daily basis. Apart from the policies made by different countries (also EU), we would all like to know, why a certain decision that may concern us, is being made. For these reasons, there is a lot of different techniques designed especially for XAI and there will probably be even more in the future to make sure that the models are not biased and no one is being discriminated.

References

[1] Biecek, Przemysław, and Tomasz Burzykowski. 2019. Explanatory Model Analysis. http://ema.drwhy.ai/

[2] Przemysław Biecek. XAI Stories. https://pbiecek.github.io/xai_stories/

[3] Christoph Molnar. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/