جامعة نيويورك أبوظبي
NYU ABU DHABI

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Visualizing Machine Learning:
Creating explanations for black-box machine learning models

Oscar Gomez, Steffen Holter, Enrico Bertini

Abstract

Our project focuses on increasing the interpretability of machine learning models by creating black-box model explanations. By combining local instance level explanations and a global model interpretation we are creating an interactive web application to visualize the logic behind each decision. The solution identifies the most important features contributing to a decision and suggests the minimal set of changes needed to alter the model's output. It aggregates the individual explanations into an interactive hierarchical interface which allows for easy comparison and exploration of the single instances.

While white-box analysis techniques that allow for straightforward interpretation are available, they are usually limited to simple models that cannot achieve the accuracy of more complex ones such as Support Vector Machines (SVMs) or Deep Neural Networks (DNNs) [1]. To handle black-box models, we use modifications of the algorithms in [2-3] to create explanations for each sample (a credit application) and aggregate them to create a global explanation that gives an overview of the model's logic.

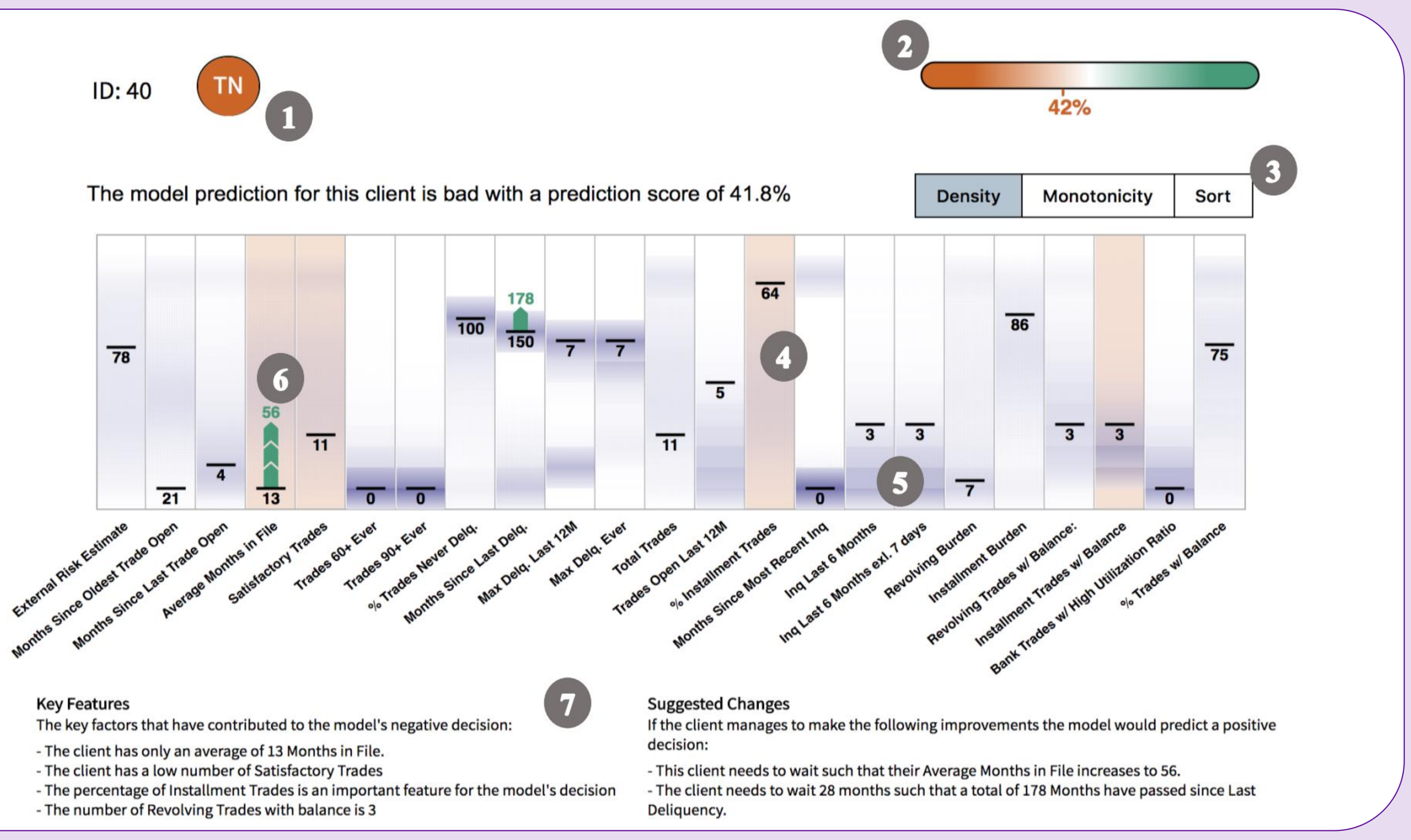
Background

Recent advancements in machine learning have allowed for the creation of models with great predictive accuracy in a variety of applications. However, the complexity of such models make understanding and interpreting them difficult, and it is often the case that neither the trained model nor its individual predictions are readily explainable. This poses a considerable problem for work concerning high-risk datasets and sensitive decisions where reliance on only the model's output is not feasible. Fields such as medicine require understanding the underlying logic behind each prediction as every decision can have serious and longstanding implications.

Similarly, even with their great potential, complex machine learning solutions are struggling to find widespread acceptance in the financial industry where the lack of explainability makes it hard to fulfill regulatory requirements. To incentivize research in this area, FICO has launched a challenge with a real home equity line of credit data set where the objective is to create models that are both accurate and interpretable, in which we will be participating with this project.

Instance Level Explanation

1) Classification correctness.
2) Model's percentage prediction.
3) Buttons that allow modifying the display.
4) Highlights a key feature for this decision.
5) Shows the density distribution.
6) Minimum changes needed to reverse the decision.
7) Text version of the explanation.



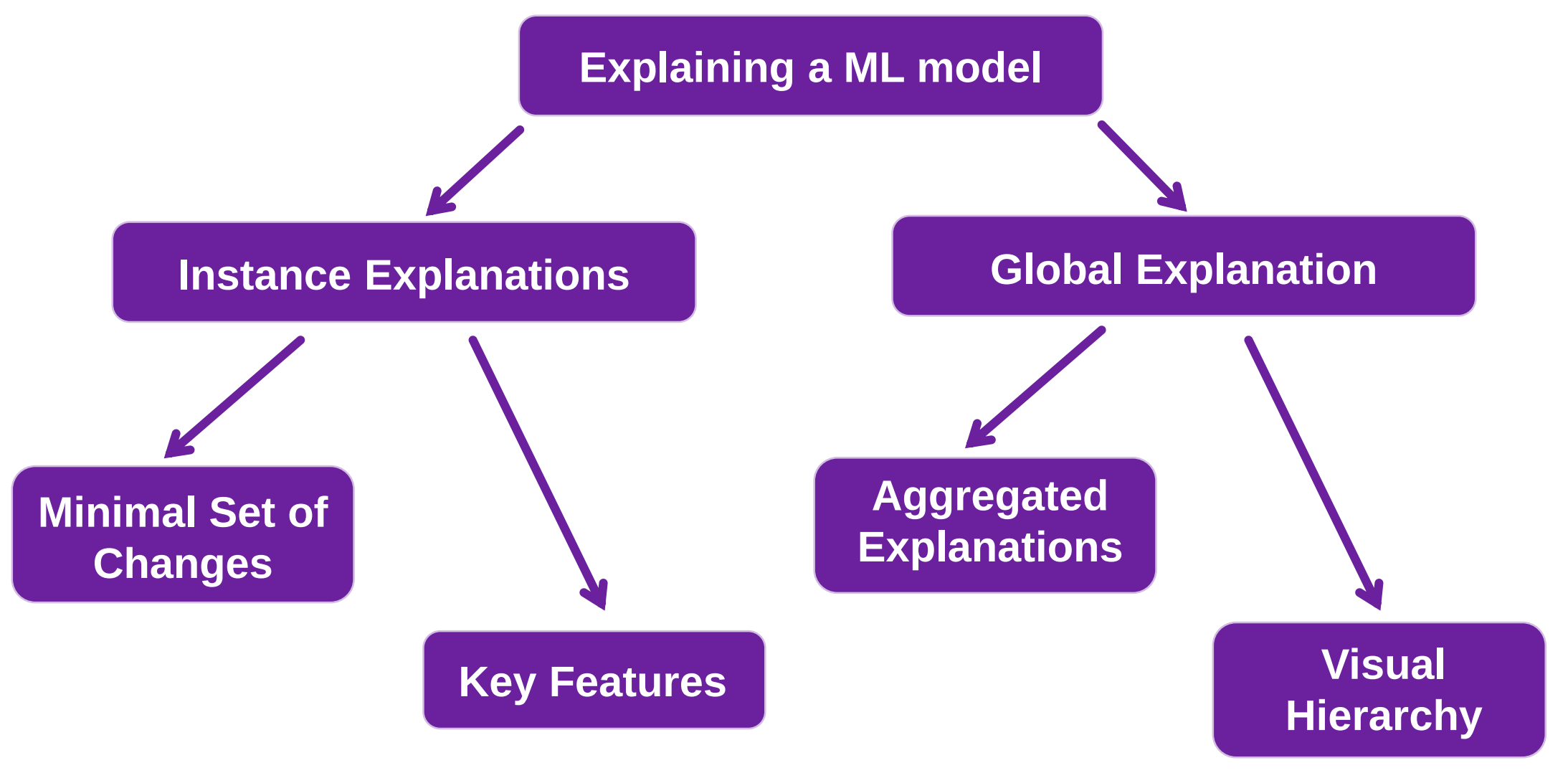
ID: 40

The model prediction for this client is bad with a prediction score of 41.8%

Density | Monotonicity | Sort

Key Features
The key factors that have contributed to the model's negative decision:
- The client has only an average of 13 Months in File.
- The client has a low number of Satisfactory Trades
- The percentage of Installment Trades is an important feature for the model's decision
- The number of Revolving Trades with balance is 3

Suggested Changes
If the client manages to make the following improvements the model would predict a positive decision:
- This client needs to wait such that their Average Months in File increases to 56.
- The client needs to wait 28 months such that a total of 178 Months have passed since Last Deliquency.



https://github.com/5teffen/Fico-Challenge
www.ml-explainer.com

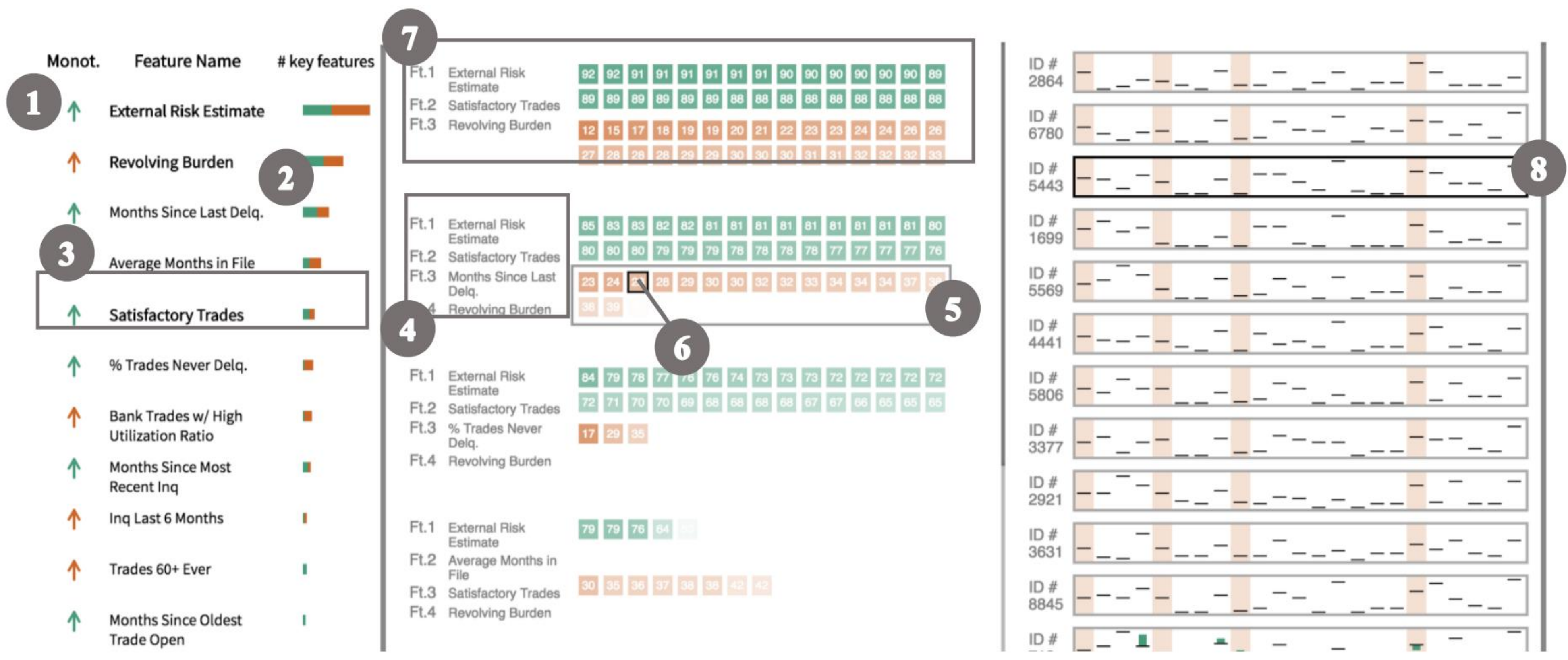
Algorithms

The application makes use of a Minimal Set of Changes (MSC) algorithm and a Key Features algorithm. By analyzing each feature column the algorithms suggest the fewest changes that would prompt a different decision by the model. Key Features are found by systematically perturbing a sample instance and measuring the resistance to change against a predetermined threshold. The MSC is found by a greedy procedure that optimizes the change in the model's prediction at each step. To add a dimension to the visualization a density estimation was performed to highlight the data distribution.

In the global explanation the individual results are aggregated into an interactive hierarchical interface. Features are ranked by relevance and can be filtered to explore similar individual explanations.

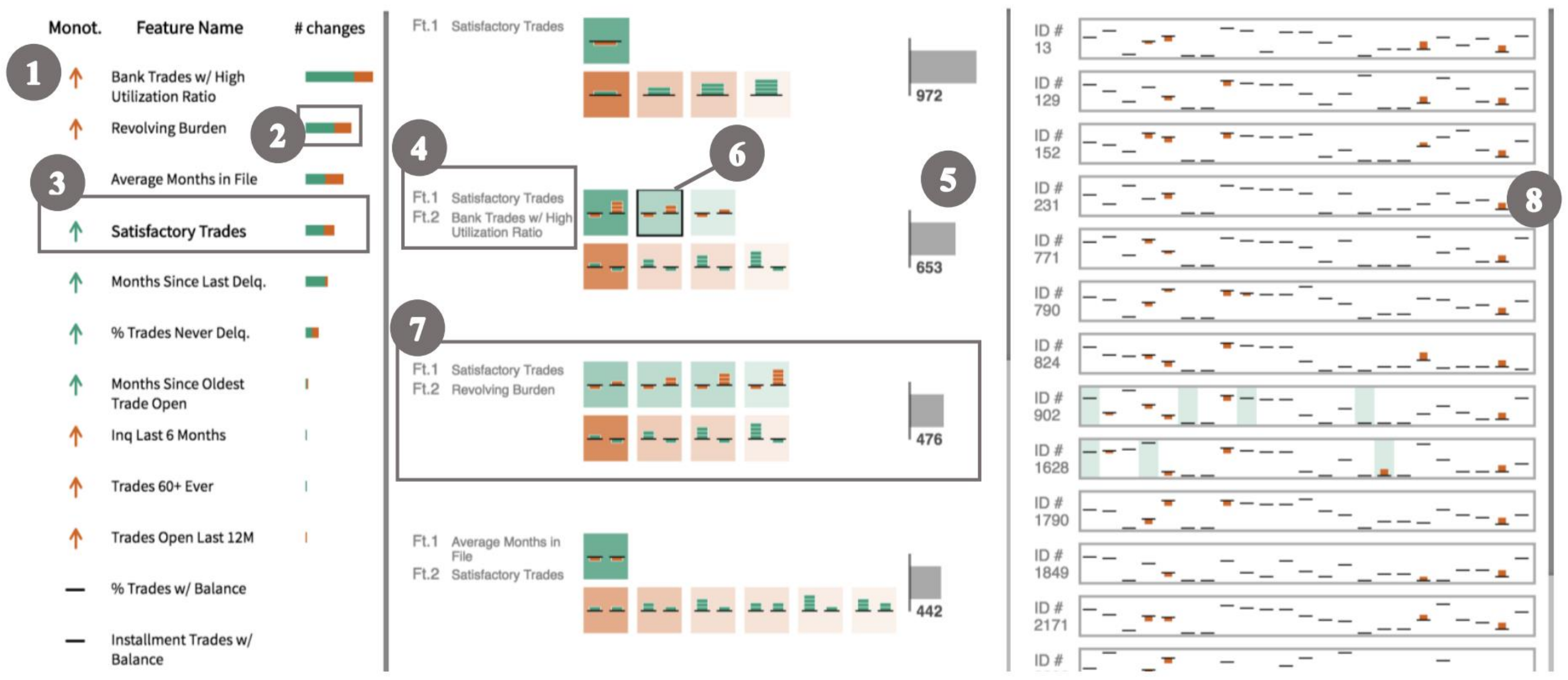
Global Explanation

1) Monotonicity of the feature.
2) Number of samples where this feature is key.
3) Selected feature(s).
4) Combination of features used for explanation.
5) Highlights the most positive/negative samples.
6) A single sample.
7) Set of samples explained by 4).
8) Miniature individual explanation.



Global Explanation

1) Monotonicity of the feature.
2) Number of samples where this feature is key.
3) Selected feature(s).
4) Combination of features used for explanation.
5) Total number of samples with these changes.
6) All samples where such combination of changes is present.
7) Set of samples explained by 4).
8) Miniature individual explanation.



Future work

Further work on the project includes expanding it to other datasets such as multiclass classification problems, as well as incorporating other explanation algorithms that can complement the ones already implemented and other visualization tools such as 2D projections of the space of explanations.

References

[1] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17). ACM, 1-6.

[2] David Martens and Foster Provost. 2014. Explaining Data-driven Document Classifications. MIS Q.38, 1 (March 2014), 73–100.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. AAAI Conference on Artificial Intelligence (AAAI-18). 1-9.

Acknowledgement

The authors thank NYU Tandon School of Engineering's Office of Undergraduate Academics and the NYUAD Undergraduate Research Program for generous funding of the project.