Creating Visual Explanations to Black-box Machine Learning Models

Oscar Gomez, Steffen Holter, Enrico Bertini

جامعة نيويورك أبوظبي
NYU | ABU DHABI

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Introduction
Approach

Client focused solution

- Useful feedback to clients
- Reasons for decision
- Suggestions for improvement / warnings

Visual Interface

- Aggregation / exploration of individual explanations
- Customizable screen

Machine Learning Model
Training & Pre-processing

Pre-processing Data

- Omit redundant data:
  - Samples with all the fields with -9 value (not investigated or not found)
- Linear Regression:
  - Samples with -9 values for External Risk Estimate
- k-NN Imputation:
  - Samples with -8 values (no usable / valid accounts)
- Approximation:
  - Samples with -7 values (condition not met)
- Standardization of categorical values

Model

- SVM (Linear Kernel)
- Test accuracy:
  - ~68% before pre-processing
  - ~74.8% after processing

Algorithms

Data discretization & Explanations

Minimal Set of Changes

- Suggest the fewest changes to flip a decision.
- Greedy procedure that optimizes the change in the model's prediction at each step.

Key Features

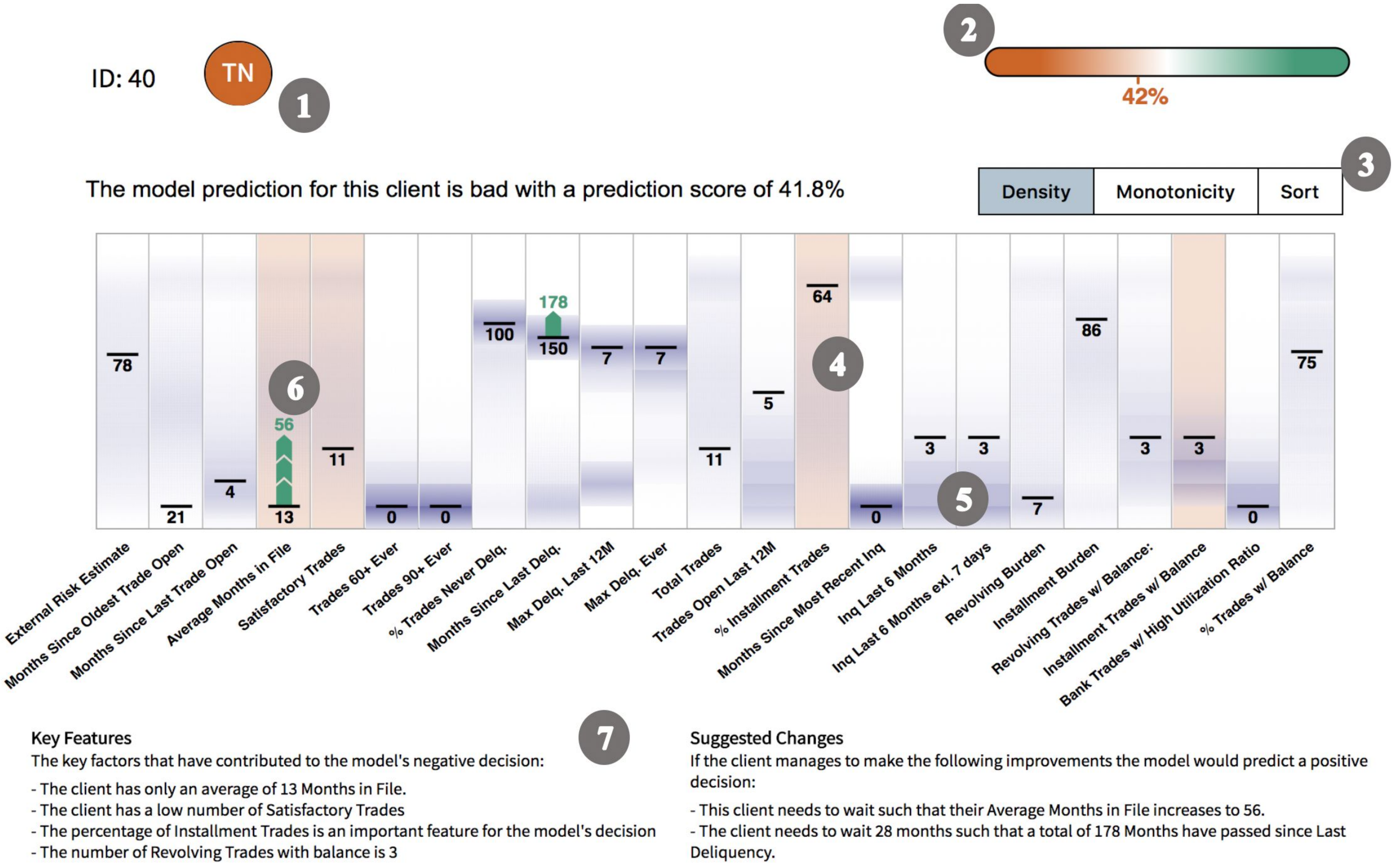
- Systematically perturbing a sample instance and measuring the resistance to change against a predetermined threshold.
- Highlighting the features that are of paramount importance for the model.
- Fixing one feature at a time and perturbing all the other columns by their respective Gaussians
- To add a dimension to the visualization a density estimation was performed to highlight the data distribution.

Data Discretization

- Distribute numerical features into ten bins.
- Range of two standard deviations below the mean to two above it.

Individual Explanation
Client Overview

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



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

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

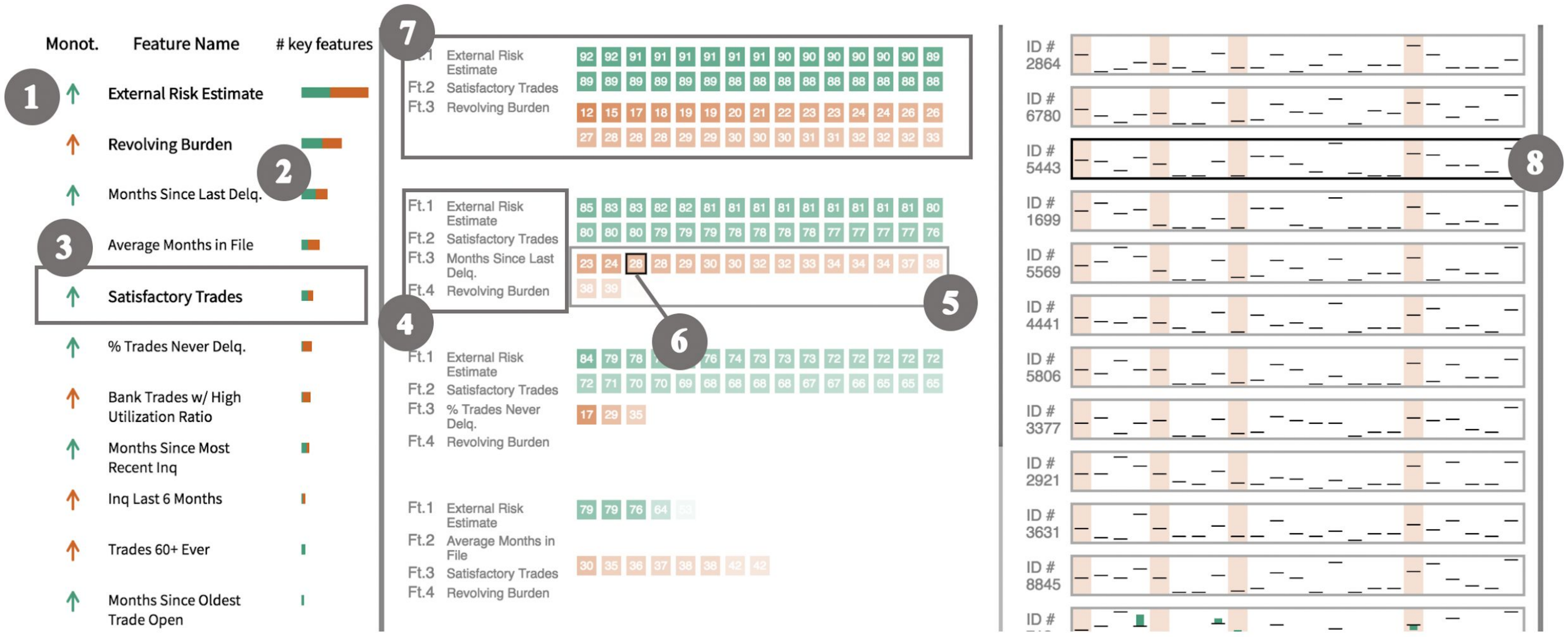
Suggested Changes

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.

Global Explanation
Key Features

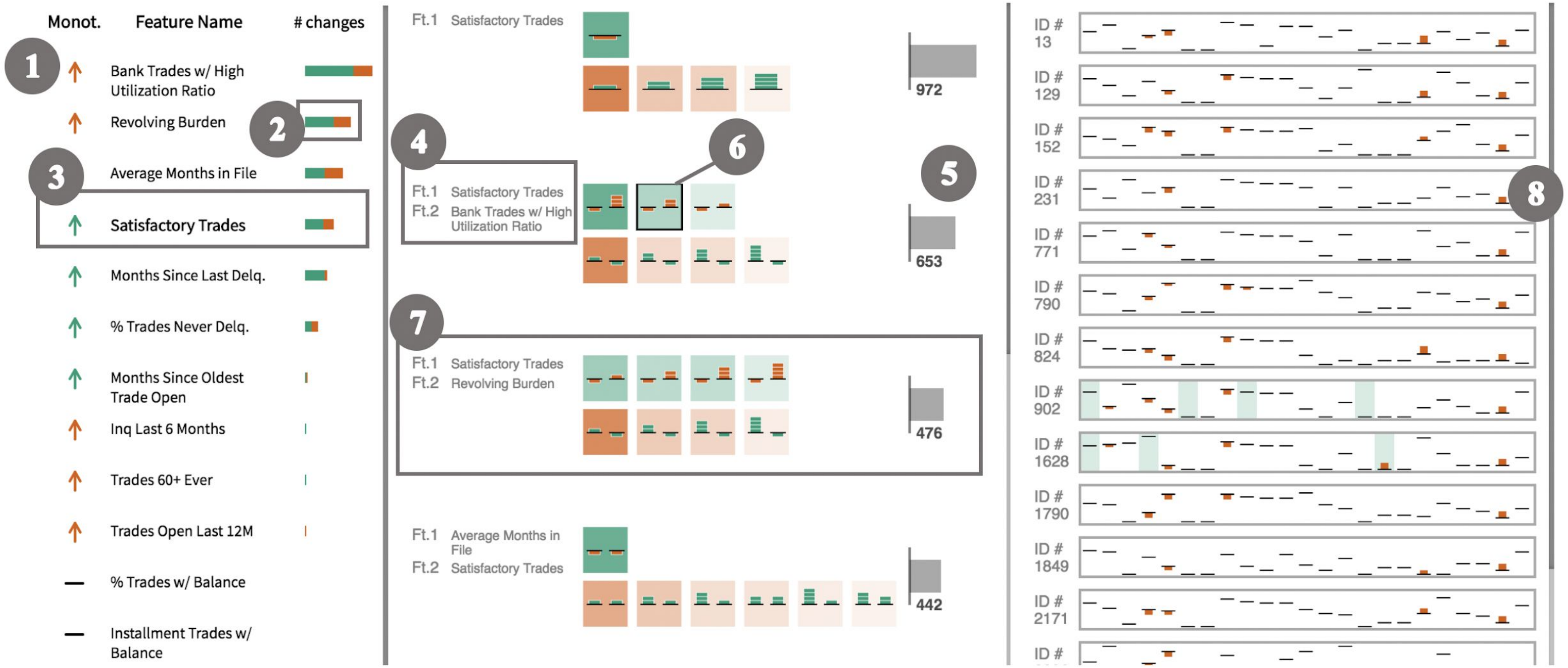
1. Monotonicity of the feature.
2. Number of samples where this feature is key
3. Selected feature(s)
4. Combination of features used for explanation
5. Total number of samples with these changes
6. All samples where such combination of changes is present
7. Set of samples explained by 4)
8. Miniature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



Global Explanation

Necessary Changes

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 Plans
Possible improvement

Similar samples

○ Improve current basic solution

Global visualization of data points

○ Create interactive visualization combining all individual results

Aggregation of explanations

Other datasets

Project site:

www.ml-explainer.com

Report:

http://www.ml-explainer.com/static/images/FICO_paper.pdf

References:

- [1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016

- [2] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Anchors: High Precision Model-Agnostic Explanations." AAAI Conference on Artificial Intelligence. 2018.

- [3] Tamagnini, Paolo, et al. "Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations." Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics. ACM, 2017.