Comparing of model interpretability tools on dataset with sensitive clients data using LIME, SHAP and ANCHOR

Team members:
Maksym Opirskyi **Anton Bilchuk**
Bohdan Matviiv **Oleksandr Smyrnov**

PROJECT MOTIVATION

Information about why the model produces a certain result is extremely important for the decision-making process since it describes the main reasons why the model provides a particular score. This enables the end user to understand whether it is worth relying on the model, and generally causes more confidence in the algorithm.

Quite a lot of frameworks have appeared that allow us to interpret the results of machine learning models, and all of them are based on different mathematical approaches.

This was the motivation of our team in this work to develop a machine learning model based on data that clearly discriminates and compare the interpretation of modern frameworks.

MODELLING

As a predictive model, we considered ensembles algorithms. The reason is that these models are not easy to interpret as logistic regression or a decision tree, and are often used in applied problems.

In the process of working on a project, since many variables in the data are categorical, we considered CatBoost, a boosting algorithm from Yandex, which processes categorical variables. But it is impossible to proceed to interpretation, for example, using LIME, since numerical variables are required.

So the final algorithm that we trained - Random Forest classifier. The algorithm is not prone to overfitting, has a random effect (bootstrap of samples and a random subset of features) and the final prediction is a mix of a few of decision trees.

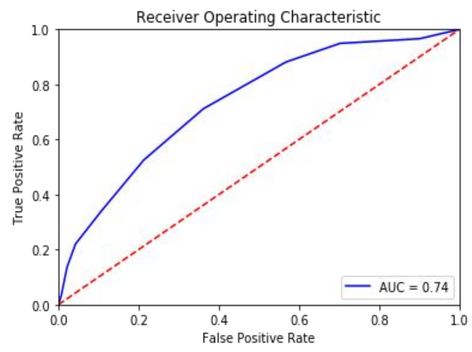
CONSIDERED DATA

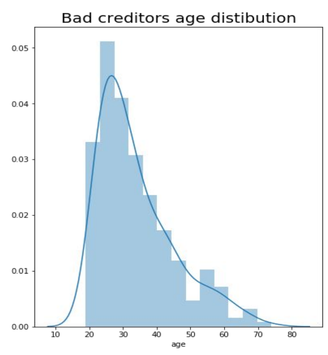
Data was taken from the German credit bureau. We are interested in this set because it contains private information about a person - gender/age/work - that can commonly occur in official documents and may be potential predictors for models.

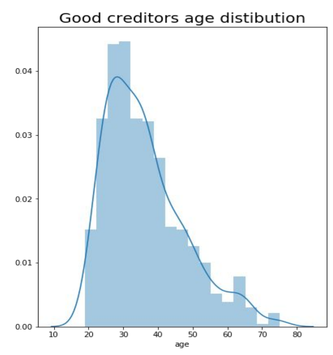
There are 1000 clients of the credit bureau, 700 of which took out a loan and returned (good), 300 got a loan and did not return (bad). For each sample, there is such data:

- Status on client accounts (balance, salary)
- Information about the loan application (duration, amount, appointment)
- Personal data of the client (gender, age, work, etc.)

Source - "UCI Machine Learning Repository - extracted from the German Credit Rating Agency SCHUFA"

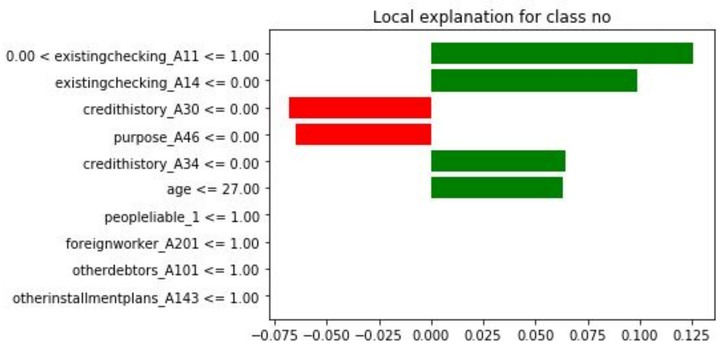






Final accuracy score: 75%

LIME



Lime is a framework which tries to solve for model interpretability by producing locally faithful explanations for machine learning models. It explains individual predictions of text, image and table data.

For explanation, Lime treats the model as a black-box and so the only way to understand it is perturbing the input and see how the prediction changes.

In LIME explanation model is defined by g, real classification model by f(x) and use x(z) to define locality around x. As LIME is model agnostic we do not know anything about actual f(x).

After we would get a measure of how our explanation model is unfaithful of approximating f(x).

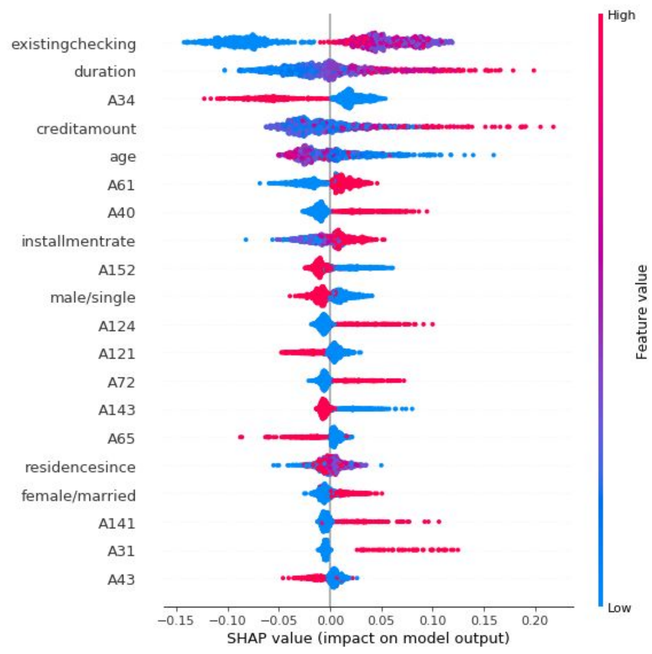
$$\mathcal{L}(f, g, \pi_x)$$

The approximation produced by LIME is obtained by:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

In general, LIME explanation is based on local surrogate models like a linear model or decision tree that are learned on the predictions of the original black box model. But instead of trying to fit a global surrogate model, LIME focuses on providing local surrogate models to explain why single predictions were made.

SHAP

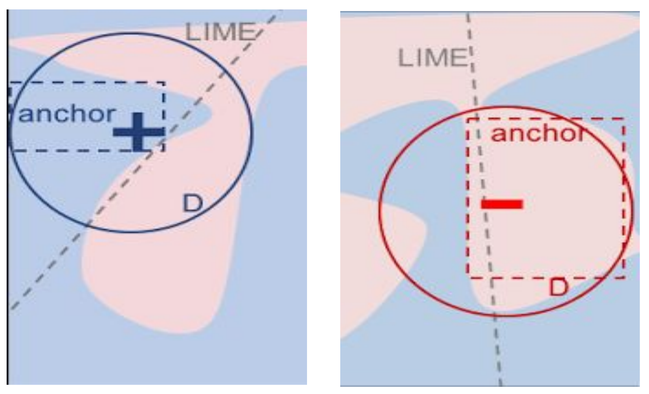


SHAP is a python package which can be integrated in an existing data pipeline to produce additional insightful information.

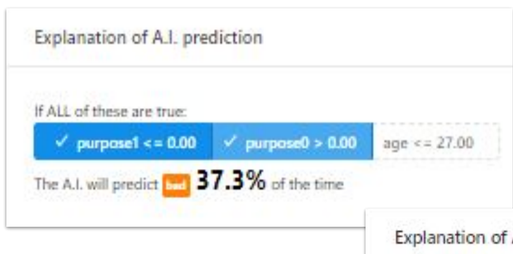
The idea is derived from cooperative game theory to calculate contribution of each factor to a common work or problem and interpret by given values target model.

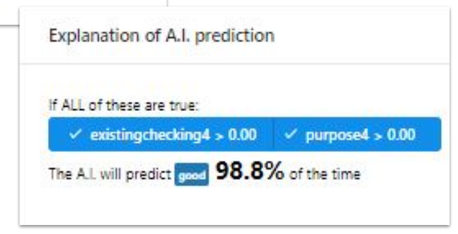
It utilizes shapley value formula as a core component where all features are “contributor” and trying to predict the task which is “game” and the “reward” is actual prediction minus the result from explanation model. In SHAP, feature importance is assigned to every feature which is equivalent to mentioned contribution. SHAP provides explainers for DL models, RF models , kernels, etc.

Anchor



Anchor is interpretability tool based on high precision local explanations. It produces a set of simple if-then rules about input features that "anchors" a prediction - meaning that if rules are satisfied, then no matter how other input features would change the prediction will still be the same with high probability. This means, that is is possible to determine the coverage of the explanations produced, because explanations apply only to instances where the rules hold. Unlike e. g. LIME, Anchor is constructed in such a manner that allows user to apply it to unseen instances, because not only explanation for input instance is known, but also its local region.





RESULTS

After constructing the model, we used each of the frameworks to interpret each of the samples from the test sample (the test sample consists of 200 elements) to identify age discrimination. Since the algorithms are based on different mathematical approaches, therefore, the computational time and the results obtained turned out to be very different. SHAP and Lime, in comparison with Anchor, were much more often interpreted based on the sign “age” (57 samples - SHAP (28.5%), 55 samples - Lime (27.5%), 23 - Anchor (11.5%)).

Based on the obtained results, we can say that the frameworks for interpreting models on the same data provide different explanations for the samples. This should be taken into account using these tools in decision-making.

In our case, it turned out that Lime and SHAP were much more often based on the sign of age than Anchor.

REFERENCES

1. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. https://arxiv.org/abs/1602.04938
2. Anchors: High-Precision Model-Agnostic Explanations.Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin.https://homes.cs.washington.edu/~marcotcr/aaai18.pdf
3. Interpreting your deep learning model by SHAP https://towardsdatascience.com/interpreting-your-deep-learning-model-by-shap-e69be2b47893
4. Lundberg S. M., Lee Su-In. A Unified Approach to Interpreting Model Predictions. 2017. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.