Ukrainian Catholic University

Faculty of Applied Sciences

Data Science Master Programme

???

Responsible Data Science project report

Authors:
Konstantyn Ovchynnikov
Dmitrii Glushko
Yuriy Lizak
Olena Shevchenko

May 14, 2019

1 Introduction

Today we are surrounded by applications of Machine learning. It helps us quickly write letters, corrects grammar, recommends movies and products, unlocks our phone. We are getting used to it; some of us do not really care how such algorithms make decisions, others do. Though some applications are harmless, meaning a mistake will not have serious consequences (e.g., a movie recommender system), others might heavily influence our life by making a mistake (e.g., autonomous driving). Generally, we might want to know 'what was predicted?' and 'why the prediction was made.' To answer the second question, the interpretability comes in handy.
According to *Miller*[3]: **Interpretability is the degree to which a human can understand the cause of a decision.** Another definition[1] is **Interpretability is the degree to which a human can consistently predict the models result.** So the interpretability is a measure of an answer to 'why' question, is it only? Various teams around the world work on the understanding of model decision making to make black box models more transparent. Different algorithms had been developed and open sourced. In this work, we want to explore the functionality of popular libraries and analyze their interpretability power.

2 Related work

The easiest way to achieve interpretability is to use only a subset of algorithms that create interpretable models. Linear regression, logistic regression, and the decision tree are commonly used interpretable models.[4] Such models are well described in terms of their explanatory power, so we will not pay additional attention to these models.
With the increasing popularity of deep learning, various teams work of mechanisms to achieve model agnostic interpretability[5].The team from the University of Washington *Ribeiro et al.*[6] presented the local interpretable model-agnostic explanations (LIME) that work with any classifier.

3 Problems

We want to investigate the interpretability systems based on two different tasks - Credit Scoring and Image Classification.

3.1 Credit Scoring

A credit score is a numerical expression based on a level analysis of a person's credit files, to represent the creditworthiness of an individual. Traditionally, a credit score was primarily based on credit report information typically sourced from credit bureaus.
About the Dataset: Lending Club Dataset **HERE SHOULD BE CITATION**
Lending Club is a US peer-to-peer lending company, headquartered in San Francisco,

California. Lending Club is the world's largest peer-to-peer lending platform. The company states that $15.98 billion in loans had been originated through its platform up to 31 December 2015.

4 Model-agnostic systems

In their work *Ribeiro et al.*[5] describes desirable properties of model-agnostic explanation systems. Authors point out three such features:

- Model flexibility - can work with both linear models and deep networks
- Explanation flexibility - not limited to a certain form of explanation
- Representation flexibility - should be able to use a different feature representation as the model being explained.

4.1 LIME

HERE SHOULD BE INFO ABOUT LIME: WHEN PRESENTED, BY WHOM, BASIC IDEA, HOW DOES IT WORK

4.2 SHAP

SHAP (SHapley Additive exPlanations) presented by *Lundberg and Lee* [2] at NIPS 2017 as a unified way to interpret model predictions. SHAP is a unified approach to explain the output of any machine learning model. SHAP connects game theory with local explanations, representing the only possible consistent and locally accurate additive feature attribution method based on expectations.
We will use Python Shap library on two tasks (credit scoring, image classification), to understand It's abilities, advantages and disadvantages.
Shap has model flexibility - this library can work with Tree ensembles, Linear models, and Deep Learning models also. By the time, Shap is one of the most famous libraries, it has good explanation flexibility. There is an ability to look at single observation (or multiple) to understand, which features influenced the result in which way (higher/lower).
HERE SHOULD BE MORE SPECIFIC INFO ON HOW DOES IT WORK

HERE SHOULD BE THE DISCRIPRIPTION OF THE GRAPH

4.3 SKATER

HERE SHOULD BE INFO ABOUT SKATER: WHEN PRESENTED, BY WHOM, BASIC IDEA, HOW DOES IT WORK

5 Results

WE CAN EITHER GROUP THE RESULTS BY TASKS OF BY MODELS

6 Conclusions

Here we should make small analysis on how it works

7 Peer Review

table with peer review

References

[1] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. "Examples are not enough, learn to criticize! Criticism for Interpretability". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 2280–2288. URL: http://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf.

[2] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 4765–4774. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[3] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *CoRR* abs/1706.07269 (2017). arXiv: 1706.07269. URL: http://arxiv.org/abs/1706.07269.

[4] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/. 2019.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[6] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *CoRR* abs/1602.04938 (2016). arXiv: 1602.04938. URL: http://arxiv.org/abs/1602.04938.