Master Thesis Project Specifications

—

Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model

Marc Beillevaire
marcbei@kth.se

February 16, 2017

Abstract

This document describes the specifications of my master thesis. This project will be carried on at Dataiku, a french datascience start-up in Paris. Into their datascience team, I will be researching on how to better understand predictions made by a machine learning algorithm, which is quite difficult in most cases today. This document provides both the project description and practical information on how it will be conducted.

1 Supervisor at Dataiku and ressources

Pierre Gutierrez (pierre.gutierrez@dataiku.com), datascientist at Dataiku will supervise this project in the company. I will have weekly (at least) meetups with him to check how the project is going on.

Dataiku provides me with a computer and a *Dataiku Data Science Studio* license to test my results in the company's software.

2 Background and objectives

Dataiku is a software editor that develops Dataiku DSS (*Data Science Studio*), a collaborative data science platform that enables companies to efficiently process their data and build machine learning models. Dataiku also have an important datascience team that provide support in datascience, along with the software, for the clients.

Yet, most of these algorithms today still look like black boxes on the user side: it is not easy - if not impossible - to understand how a given prediction is made by, let's say, a Random Forest algorithm with several hundreds of trees. Several tools already exist to understand models but there are mostly global coefficients like the variable importance in a decision tree.

Yet in many cases having an insight locally would be really helpful to users. An insurance company would like to know why a particular client has been detected as a fraudster, or a manufacturer doing predictive maintenance would like to know not only which device is likely to break soon but also how this device's lifetime could be increased practically. Moreover a European regulation has been adopted by the EU parliament and will force companies to provide *"explanations"* on every decision made by an algorithm that impact their client. Even if the application conditions has not been stated clearly up to now, this will require good explanations system to understand a particular decision. For instance, someone who has been refused a load could ask the bank for a specific and logical reason, even if this decision has been obtained by a computer.

My Master Thesis is therefore focused on finding *explanations* on predictive models outputs, at the prediction level.

The main goals are: exploring bibliography on the subject, testing several methods on various datasets and linking theses methods to Dataiku DSS to make it easy to explain predictions.

3 Research question and methodology

As stated before, the research will focus on how to produce understandable explanations related to a prediction made by a machine learning algorithm.

The goodness of the output of such an automatic explainer is not easy to quantify, as it relies on the user's knowledge of the data, and its ability to judge the pertinence of a prediction and an explanation. So the qualitative question would be: *How to extract explanations on individual predictions from a predictive model ?*

Yet providing explanations can have a measurable impact when an algorithm is relying on the wrong features.

When an algorithm relies on the wrong variables to make a prediction, the problem is not the prediction in itself, that can be true, but the lack of generalizability of the algorithm. Thoses bad or wrong variables are called so because they are correlated to the target on the training sample dataset, but not on the whole individuals in real life. Thus such predictive model relying on the wrong variables are expect to generalize badly on a different dataset.

From this point comes the quantitative question: *Is it possible to improve the score of a bad model that relies on the wrong variables using a explanation algorithm ?*

4 Evaluation method

Several explainers will be tested, and for each of them the goal would be to improve the score of a bad predictive model. This will need one train dataset where "bad" features are correlated to the output as well as a validation set where these features change. Therefore the bad model score should decrease on

this validation set.

The various methods will be evaluated using on several dataset from websites such as UCI, or Kaggle.

Some datasets would probably better suits some methods, but to be genral enough, testing datasets should be diverse: regression and classification datasets, text data and tabular data, ...

5 Schedule

30 Jan / week 5:	Project specifications and pilot study
6 Feb / week 6:	Going on with experiments and testing of implementations already performed at Dataiku. The focus will be on the quantitative evaluation of the performances of the tested algorithms.
13 Feb / week 7:	Getting a plan set for the report, finishing testing algorithms and implementations.
20 Feb / week 8:	Writing the report.
27 Feb / week 9:	Writing the report, first Draft.
6 March / week 10:	Report read again and mistakes fixed.

6 Pilot study

The bibliography study features in the Reference part below. It shows that this topic was only sparsely studied more than 5 years ago, but that more and more professors and PhD student has started to think about it in the past few years.

References

[1] Christopher H. Achen. *Interpreting and using regression*. SAGE, Oct. 1982.

[2] D. Baehrens, T. Schroeter, and S. Harmeling. "How to Explain Individual Classification Decisions". In: *Journal of Machine Learning Research* 11 (2010), pp. 1803–1831.

[3] Z. Bosnić and I. Kononenko. "Estimation of individual prediction reliability using the local sensitivity analysis". In: *Journal of Applied Intelligence* 29.3 (Aug. 2007), pp. 187–203. DOI: 10.1007/s10489-007-0084-9.

[4] R. Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission". In: KDD '15. 2015, pp. 1721–1730.

[5] I. De Falco et al. "An evolutionary approach for automatically extracting intelligible classification rules". In: *Journal of Knowledge and Information Systems* 7.2 (Feb. 2015), pp. 179–201. DOI: 10.1007/s10115-003-0143-4.

[6] Mary T. Dzindoleta et al. “The role of trust in automation reliance”. In: *International Journal of Human-Computer Studies* 58 (2003), pp. 697–718.

[7] Alex Goldstein et al. “Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation”. In: *Journal of Computational and Graphical Statistics* 24.1 (2015), pp. 44–65. DOI: 10.1080/10618600.2014.907095.

[8] B. Goodman and S. Flaxman. “European Union regulations on algorithmic decision-making and a “right to explanation””. 2016. URL: https://arxiv.org/pdf/1606.08813v3.pdf.

[9] I. Konenko and Robnik-Šikonja. “Explaining Classifications For Individual Instances”. In: *IEEE Transactions on Knowledge & Data Engineering* 20 (May 2008), pp. 589–600.

[10] Zachary Chase Lipton. “The Mythos of Model Interpretability”. In: *CoRR* abs/1606.03490 (2016). URL: http://arxiv.org/abs/1606.03490.

[11] D. Martens and F. Provost. “Explaining data-driven document classifications”. In: *MIS Quarterly* 38 (Mar. 2014), pp. 73–100.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why Should I Trust You? Explaining the Predictions of Any Classifier”. In: KDD ’16. 2016. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

[13] A. Saba. “Interpreting random forests”. In: *Diving into data* (2015). URL: http://blog.datadive.net/interpreting-random-forests/.

[14] E. Strumbelj and I Kononenko. “An Efficient Explanation of Individual Classifications using Game Theory”. In: *Journal of Machine Learning Research* 11 (2010), pp. 1–18.

[15] E. Štrumbelj et al. “Explanation and reliability of prediction models: the case of breast cancer recurrence”. In: *Journal of Knowledge and Information Systems* 7.2 (Aug. 2010), pp. 305–324.