Methods for understanding the variable importance of local explanations of black-box models

Abstract

Artificial Intelligence (AI) has seen a revitalization in recent years from the use of increasingly hard-to-interpret black-box models. In such models, increased predictive power comes at the cost of opaque factor analysis, which has led to the field of explainable AI (XAI). XAI attempts to shed light on these models, one such approach is the use of local explanations. A local explanation of a model give a point-estimate of linear variable importance in the vicinity of one observation. We extract explanations for each observation, and approximate data and this attribution space side-by-side with linked brushing. After identifying an observation of interest its local explanation is used as a 1D projection basis. We then manipulate the magnitude of the variable contributions with a technique called the tour. This tour animates many projections over small changes in the projection basis. Doing so allows a user to visually explore the data space through the lens of this local explanation and interrogate its variable importance. The implementation of our framework is available as an R package called cheem available at github.com/nspyrison/cheem.

1 Introduction

Mathematically rigorous approaches to predictive modeling are attributed to the method of least squares, over two centuries ago by Legendre and Gauss in 1805 and 1809 respectively. In 1886 Francis Galton coined the term *regression* to refer to continuous, quantitative predictions. While *classification* refers discretion predictions as introduced by Fisher in 1936.

Breiman and Shmueli (**shmueli__explain__2010?**) introduce the idea of distinguishing modeling based on its purpose; *explanatory* modeling is done for some inferential purpose such as hypothesis testing, while *predictive* modeling is performed for to predict new or future out-of-sample observations. This distinction draws attention to the divide between interpretable models and black-box models. In explanatory modeling the interpretable is a key feature for drawing inferential conclusions. While predictive modeling may opt for potentially more accurate black-box models. The intended use of a model has important implications for which methods are used and the development of those models.

Predictive model and black box modeling is becoming increasingly common, but not without controversy and issues (**kodiyan__overview__2019?**). Applications have been known to reflect common biases against sex (**duffy__apple__2019?**), race (**larson__how__2016?**), and age (**diaz__addressing__2018?**). This is a common issue stemming from biases the in sample data are violate ethical principals. Another issue is that of data-drift, when new data is outside the support of latent or exogenous explanatory variables. Data-drift can lead to worse predictions (**salzberg__why__2014?**). Such issues highlight the need to make models fair, accountable, ethical, and transparent which has led to the movement of XAI Arrieta et al. (2020).

One branch of XAI is local explanations, which take a variable attribution approach to bring transparency to a model. Local explanations attempt to approximate a linear variable importance at the location of one observation. There are many such local explanations, any of which is works with our approach (assuming model-explanation compatibility).

However, to illustrate our work we apply the model-agnostic explanation SHAP (**strumbelj__explaining__2014?**). The exact details of SHAP are tangent to the ideas of this work, but suffice it to say that SHAP approximates variable importance by taking the median importance over permutations of the explanatory variables. To be exact we apply a variant that enjoys a lower computational complexity, known as tree SHAP (**lundberg__consistent__2018?**).

In multivariate data visualization a *tour* Lee et al. (2021) is a sequence of linear projections of data onto a lower dimensional space, typically 1-3D. Tours are viewed as an animation over small changes to a projection basis. Structure in a projection can then be explored visually to see which variables contribute to the formation of the structure. The intuition is similar to watching the shadow of hidden 3D object change as the object is rotated; watching the structural shape of the shadow change gleans insight into the shape and features of the object. There are various types of tours, which are distinguished be the generation of sequence of projection bases. In a *manual* tour Spyrison and Dianne Cook (2020) this path is defined by changing the contribution of a selected variable. Applying tours in conjunction with models has been previously done, *ie* for exploring various statistical model fits (**wickham_removing_2015?**), and using tree- and forest-based approaches as a projection pursuit index to generate a tour basis path (**da_silva_projection_2021?**).

The approach purposed below is to use the manual tour as means to interrogate a local explanation; it see if its variable importance are good explanation for the model predictions. We make R package `cheem` with an interactive application to facilitate analysis. By viewing approximations of data- and attribution- space side-by-side, with linked brushing an analyst can identify observations of interest whose explanations are then rendered at the initial projection basis and explored with a manual tour to further interpret the variable importance of the local explanation. We give case studies of toy and modern datasets for both classification and regression tasks.

The rest of paper is organized as follows. The next section SHAP covers the background of the local explanation SHAP and the traditional visuals produced from it. The section Application Design discusses the layout of the application, how it facilitates analysis. Following that, Software Instructure discusses the backend details of the package and preprocessing. The section Case Studies illustrates several applications of this method. We conclude with Discussion of the insights we draw from classification and regression tasks.

2 SHAP local explanation

SHaply Additive exPlanations, or SHAP (Lundberg and Lee 2017) approximates the variable importance in the vicinity of one observation by taking the median importance of a subset of permutations in the explanatory variables. This idea stems from the field of game theory where Shapley devised a method to evaluate individual's contribution to cooperative games by permuting the players contributing to the score (**shapley_value_1953?**).

An observation's SHAP values were originally used to additvely explain the difference from the intercept to the prediction. This sort of explanation is predicated on the contribution of the previous variables, making it asymmetric across variable ordering. However, viewing several of these can highlight the non-linear weightings within a single model. We will use soccer data from FIFA 2020 season (Leone 2020) to illustrate this. We have 5000 observations of 9 aggregated skill measures and use a use a random forest model to regress the wages, in 2020 Euros, from the skill measures. We then extract the SHAP values of a star offensive player (Messi) and defensive player (van Dijk). We expect to see a difference in attribution of the variable importance across the two positions of the players. Figure 1 highlight that while the players have quite different wages, when we normalize the attribution we see that skill aggregate importance is different and in a way that we would expect; offensive and movement are more important for the offensive player, while defensive and power skills are more important to the model for explaining the the prediction of the defensive player.

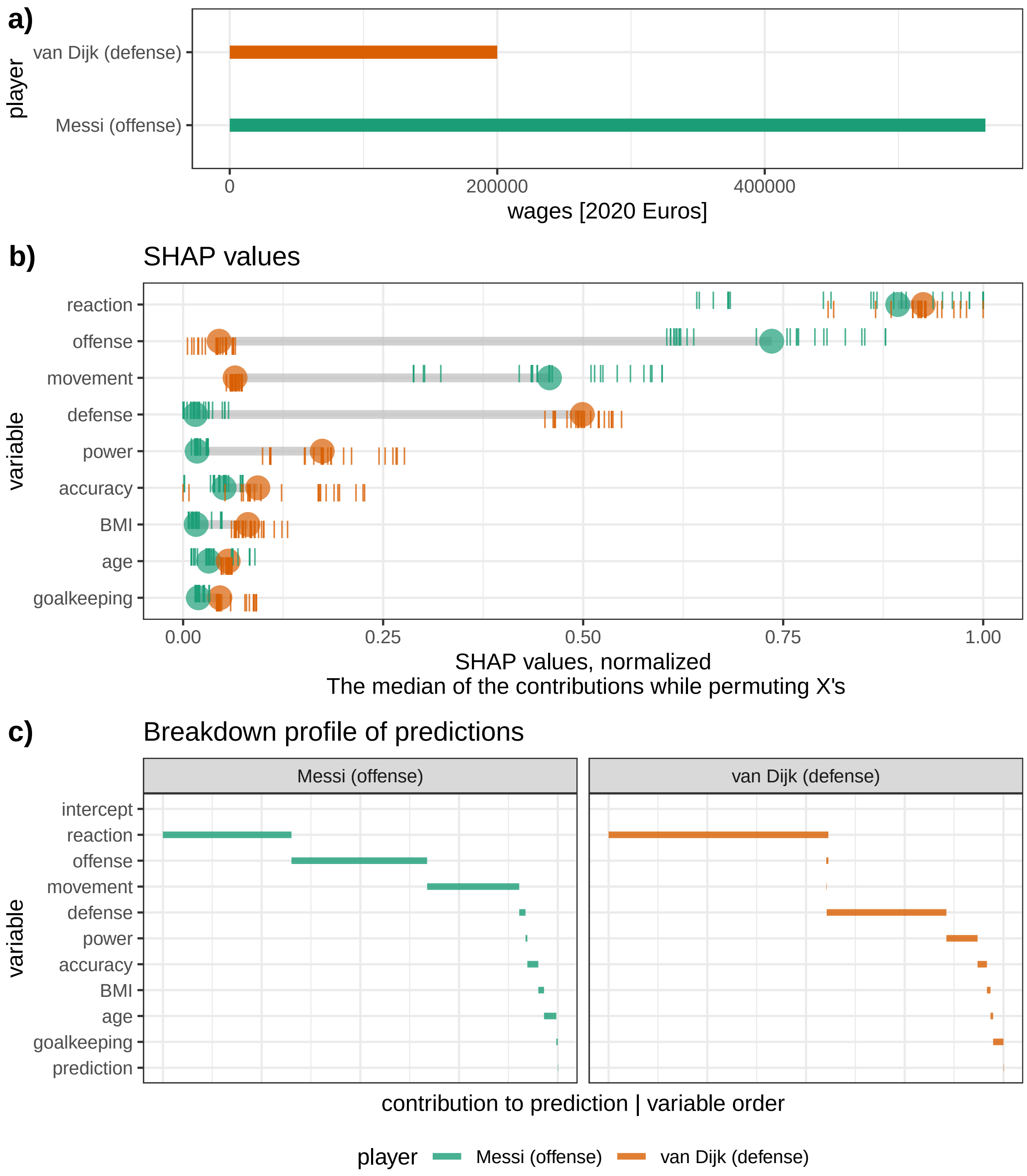

Figure 1: TODO:XXX Figure reference

2.1 Variable importance, permuting over the X's included

2.2 Visualizing and break down plots

3 Application Design

3.1 Preprocessing

3.2 Global view

3.3 Cheem tour

3.4 Classifcation task

3.5 Regression task

4 Software Infrastructure

4.1 Extend spinifex, consume DALEX & treeshap

4.2 Preprocess

4.3 Runtime rendering

5 Case Studies

5.1 1) Penguins speicies classification

5.2 2) FIFA wage regression

5.3 3)?

5.4 4)?

6 Discsussion

7 Acknoledgements

References

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Asimov, Daniel A., and Andreas Buja. 1994. "Grand Tour via Geodesic Interpolation of 2-Frames." In *Visual Data Exploration and Analysis*, 2178:145–53. International Society for Optics; Photonics. https://doi.org/10.1117/12.172065.

Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–80. https://doi.org/10.2307/1390747.

Lee, Stuart, Dianne Cook, Natalia da Silva, Ursula Laa, Earo Wang, Nick Spyrison, and H. Sherry Zhang. 2021. "A Review of the State-of-the-Art on Tours for Dynamic Visualization of High-Dimensional Data." *arXiv:2104.08016 [Cs, Stat]*, April. http://arxiv.org/abs/2104.08016.

Leone, Stefano. 2020. "FIFA 20 Complete Player Dataset." https://kaggle.com/stefanoleone992/fifa-20-complete-player-dataset.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv Preprint arXiv:1705.07874*.

Spyrison, Nicholas, and Dianne Cook. 2020. "Spinifex: An r Package for Creating a Manual Tour of Low-Dimensional Projections of Multivariate Data." *The R Journal* 12 (1): (accepted).