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 gives 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 **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 to discrete predictions as introduced by Fisher in 1936.

Breiman and Shmueli Shmueli (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 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 (2019). Applications have been known to reflect common biases against sex Duffy (2019), race (Larson et al. 2016), and age (Díaz et al. 2018). This is a common issue stemming from biases in the in-sample, training data are violate ethical principles. 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 (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 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 Štrumbelj and Kononenko (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 (S. M. Lundberg, Erion, and Lee 2018).

In multivariate data visualization a *tour* S. 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 a 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 by the generation of the sequence of projection bases. In a *manual* tour Spyrison and 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, Cook, and Hofmann 2015), and using tree- and forest-based approaches as a projection pursuit index to generate a tour basis path Silva, Cook, and Lee (2021).

The approach purposed below is to use the manual tour as means to interrogate a local explanation; a means of evaluating if its variable importance is 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 this 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 a Discussion of the insights we draw from classification and regression tasks.

2 SHAP local explanation

SHaply Additive exPlanations, or SHAP (S. 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 that contribute to the score (Shapley 1953).

TO illustrate SHAP and its original use we use soccer data from FIFA 2020 season (Leone 2020). We have 5000 observations of 9 aggregated skill measures and 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 the attribution of the variable importance across the two positions of the players.

Figure 1 illustrates the SHAP values of these players. Panel b) shows the underlying distribution of the SHAP attributions while permuting the explanatory variables, with the medians being the SHAP values. In the light of the player position, the difference in the variable importance makes sense; offensive and movement are more important for the offensive player, while defensive and power skills are more important to the model for explaining the prediction of the defensive player. We would likewise expect the profile of variable importance to be unique for star players of other positions as well, such as goalkeepers or middle fielders. Panel c) shows a simplified breakdown plot (Gosiewska and Biecek 2019), where a local explanation is used to additively explain the difference from the intercept to the observations prediction. Such additive approaches will show an asymmetry with respect to the variable ordering, so we opt to fix the order to that of panel b), namely, by decreasing the sum of the SHAP values.

In summary, this highlights how local explanations bring transparency to a model at least in the vicinity of their observations. In this instance, we showed how two very different soccer players receive different profiles of variable importance to explain the prediction of their wages. In the following section, we will be using normalized explanations as the starting projection basis to interrogate the explanation further.

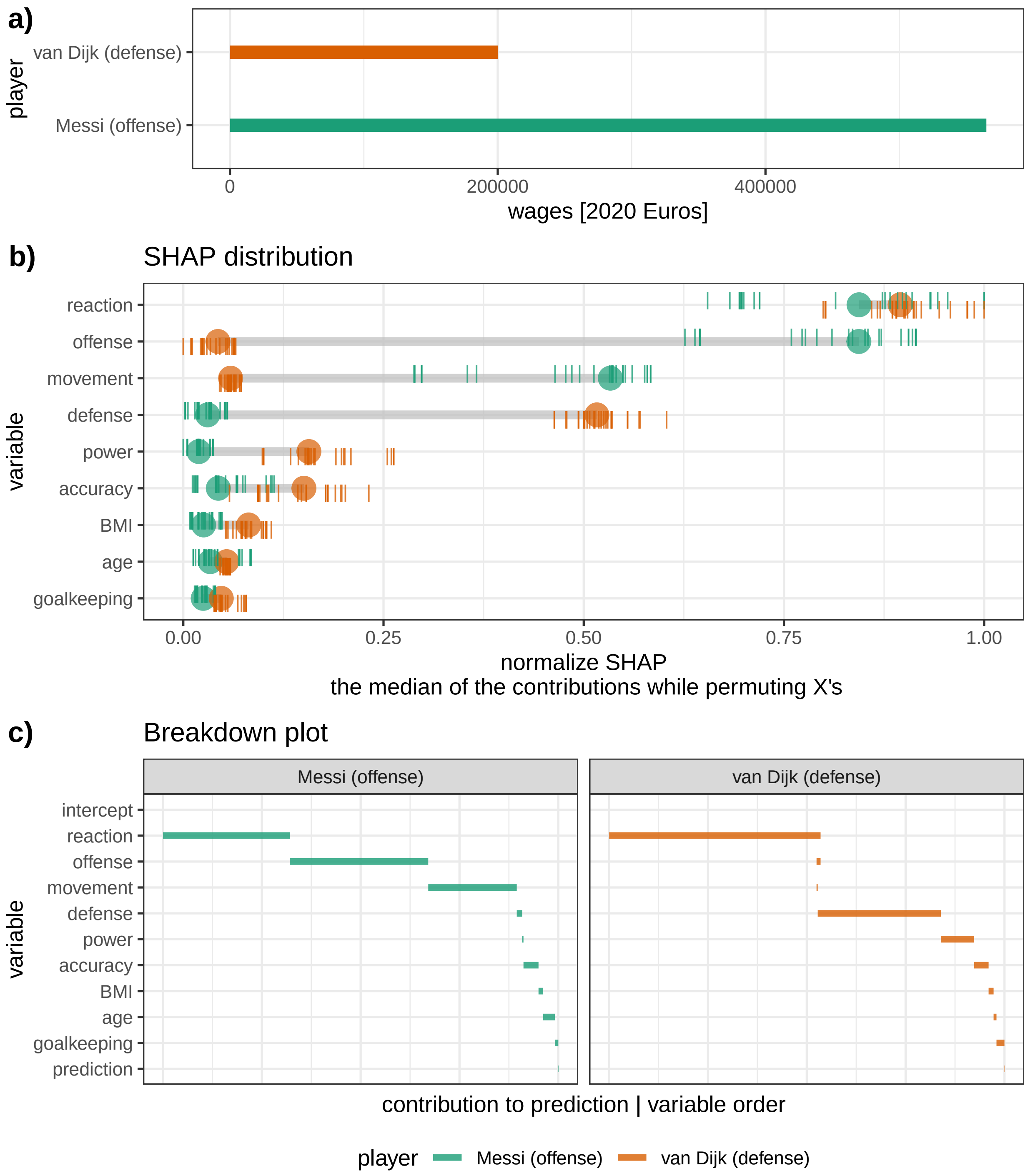


Figure 1: Illustration of the distribution of SHAP attributions, the SHAP values, and a breakdown plot, the typical visual of SHAP local explanations. For FIFA 2020 data, of a random forest model regressing wages from 9 skill attributes for a star offensive and defensive player. a) The players have very different wages. b) Shows the distributions of the attributions permuting over 25 permutations in the explanatory variables. The median of these distributions are the final SHAP values, notice that that the variable importance differs across the exogenous information of player position. These explanations make sense; the variable importances make sense in light of the position of the player. c) Breakdown plots of the observations the explanation used to additively explain the difference between the intercept and prediction

3 Application Design

Below we illustrate the two primary displays of the application: the global view and the tour view. Then we'll cover what we take away from the classification and regression tasks. Lastly, we discuss the preprocessing that needs to be done before display.

3.1 Global view

The global view is an important context for exploring the separability of the data- and the local explanation's attribution-spaces, and is crucial in the selection of explanation to further interrogate and explore the structural sensitivity of.

We show an approximation of these spaces with a projection through their first two principal components. The orientation of the variables are shown inscribed on a unit circle. While a single 2D projection will rarely encompass all of the structure of a higher-dimensional space, it provides a reasonable starting point for the real task at hand, the selection of observation and nearby comparison.

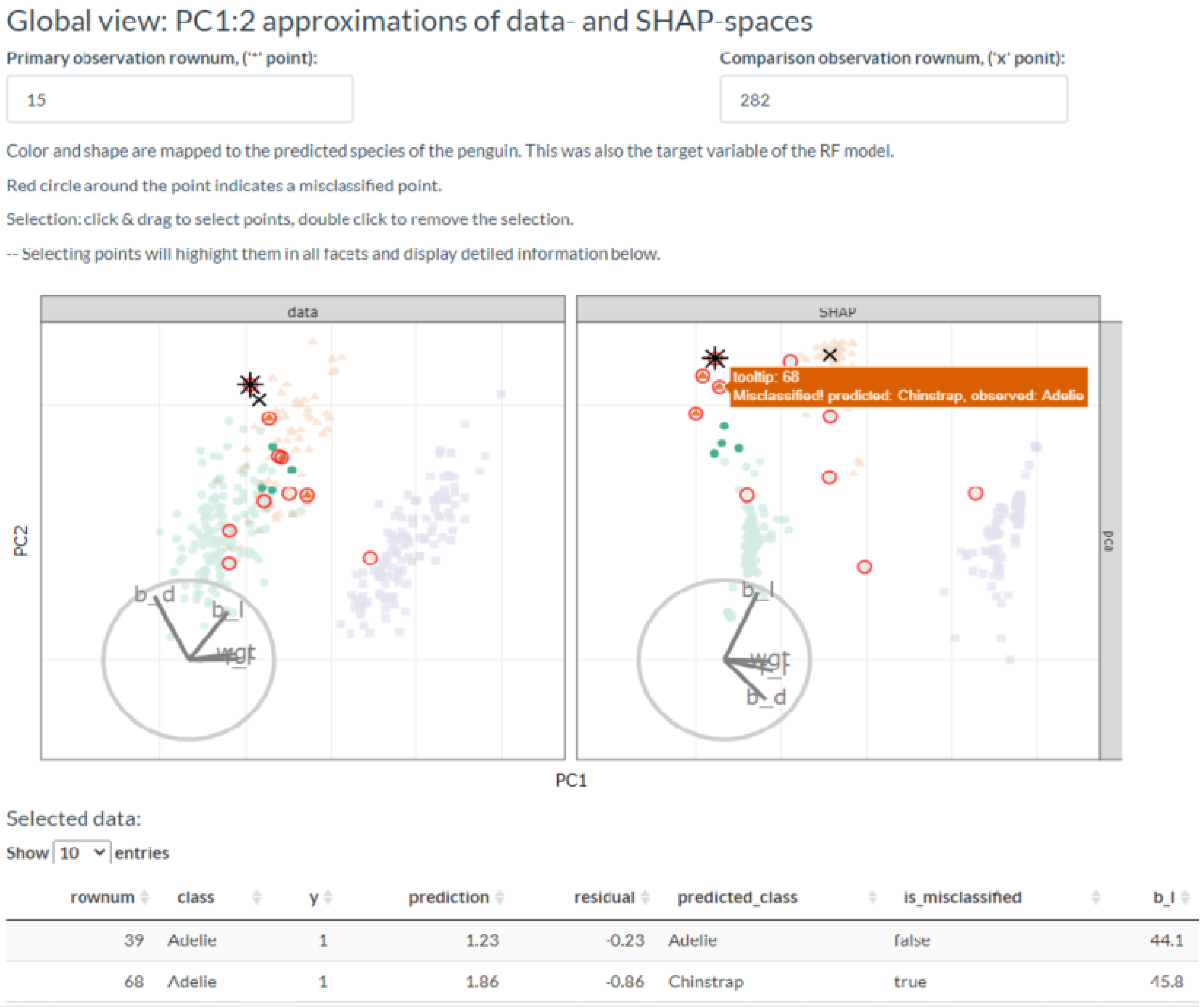
This view offers dynamic interaction in several ways. A tooltip on hovering over a point that displays the row number/name and classification information if appropriate. Linked brushing allows for the selection of points (by click and drag) where those points will be highlighted in both plots. The information corresponding to the selected points is populated on a sortable table and the data powering the proceeding tour will also subset the data to the current selection.

3.2 Cheem tour

The primary observation identified via the global view is foundational to the production of the cheem tour. Namely, the linear attribution of that variable is used as a 1D projection basis. This is the approximate contributions of the variables that this model uses to justify its prediction for the observation.

That normalized attribution of the primary observation is depicted as stacked bars where the horizontal width is the contribution. The bottom of this display is divided by the use case. In the classification case, 1D density curves with underlying rug marks are drawn and colored according to their predicted classes. In the regression case, the horizontal position of the points comes from projection through the 1D attribution basis while the vertical position of the observations is fixed to its prediction or residual.

\begin{figure}



Global view: PC1:2 approximations of data- and SHAP-spaces

Primary observation rownum, ('*' point): 15

Comparison observation rownum, ('x' ponit): 282

Color and shape are mapped to the predicted species of the penguin. This was also the target variable of the RF model.

Red circle around the point indicates a misclassified point.

Selection: click & drag to select points, double click to remove the selection.

-- Selecting points will highight them in all facets and display detiled information below.

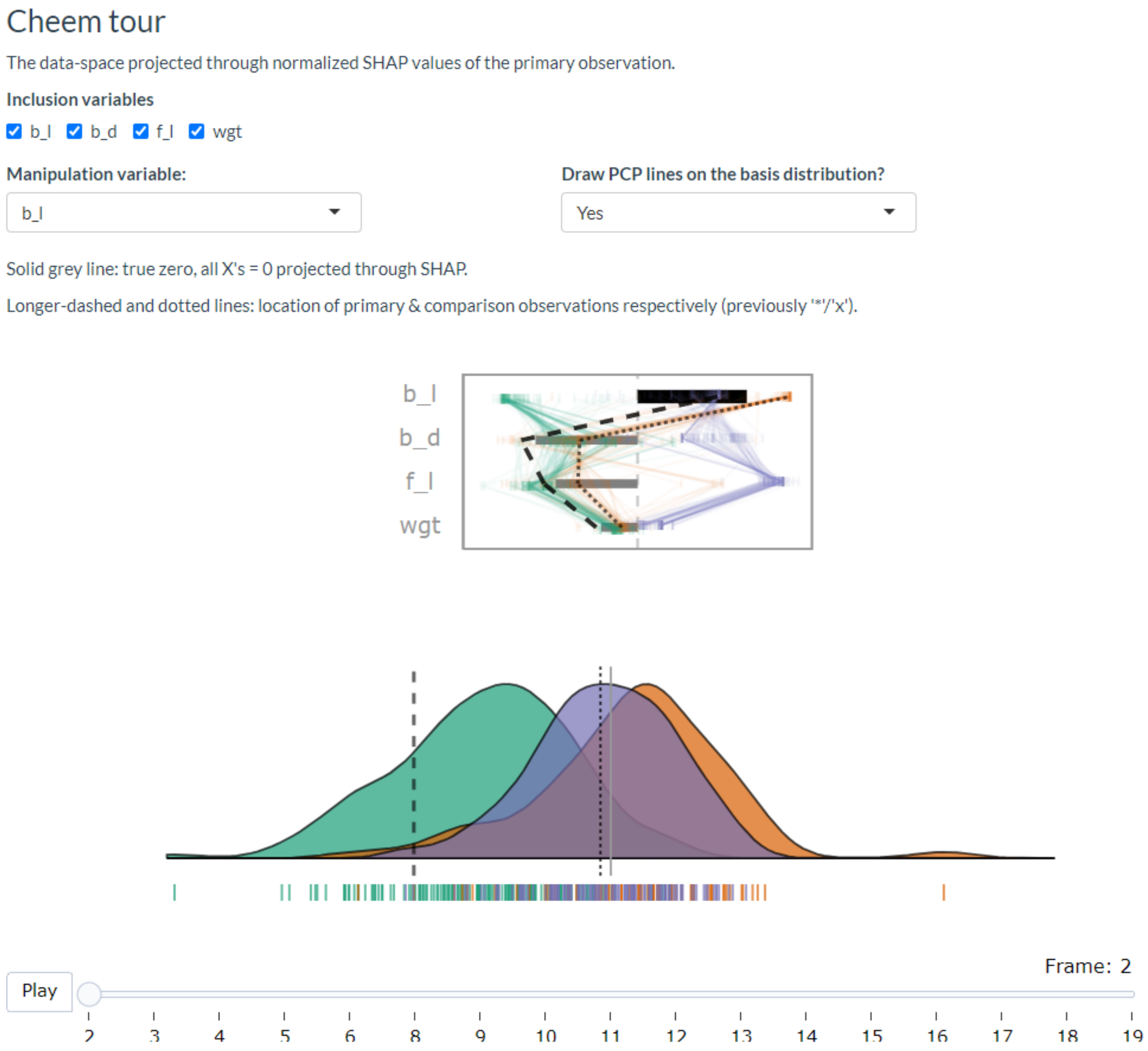
tooltip: 68
Misclassified! predicted: Chinstrap, observed: Adelie

Selected data:

Show 10 entries

rownum	class	y	prediction	residual	predicted_class	is_misclassified	b_l
39	Adelie	1	1.23	-0.23	Adelie	false	44.1
68	Adelie	1	1.86	-0.86	Chinstrap	true	45.8

Figure 2: Global view screen capture; the approximations of the data and SHAP-spaces of the penguins data. Orientation of the basis contributions is illustrated on a unit circle. Linked brushing allows observations brushed in one plot to be selected in others. This selection is also used in the proceeding view and their corresponding information is displayed in an interactive table. Hovering the cursor over an observation displays a tooltip with row number/name information. In the classification case, misclassified points are circled in red. This view prpvides an orientation to select a primary and comparison observation, key targets in the following tour.



Cheem tour

The data-space projected through normalized SHAP values of the primary observation.

Inclusion variables

☑ b_l ☑ b_d ☑ f_l ☑ wgt

Manipulation variable: b_l

Draw PCP lines on the basis distribution? Yes

Solid grey line: true zero, all X's = 0 projected through SHAP.

Longer-dashed and dotted lines: location of primary & comparison observations respectively (previously '*'/'x').

{

}

\caption{Cheem tour screen capture; the primary observation's normalized SHAP values are the initial basis. This is the explanation of the linear variable importances for the model at this observation. The point in question was misclassified, it was predicted to be from the green cluster, while it was observed to be from the orange cluster. The top shows the contributions of the variables, with the dashed line being the primary observation's SHAP, its shape is closer to the orange cluster (observed group) while the position is pulled more toward the green (predicted group). The bottom shows the 1D projection with density and rug marks below. The dashed line is plausibly in the middle of the green density, the story that the explanation trying to sell us. Yet, when we play the tour animating on the contribution of bill length (b_l), the bottom dashed line is more regularly in the center of the observed orange cluster.} \end{figure}

Data visualization tours animate many linear projections over small changes to the basis. The manual tour creates a basis path by varying the contribution of a selected variable, fully into and out of a projection frame. Doing allows an analyst to test an individual variable's sensitivity to the structure identified in the frame. The default variable selected is the one with the largest discrepancy between the primary and comparison observation's attribution. In the following sections we elaborate on the takeaways we draw from applying this approach in classification and regression tasks respectively.

3.3 Classifcation task

What information do we glean from using this method on a classifcation task? Typically we select a misclassified observation in comparison with correctly classified point that is nearby in data space. We start by seeing the data projected through the linear attribution, the combination that best justifies that prediction. By default the manual tour varies the contribution of the variable with the largest difference between the primary and comparison observation. That is, we can test the sensitivity of each variable to structure identified by the local explanation, we are exploring the support of the explanation, evaluating the support or robustness of the prediction.

Another way of thinking about this would be: what would the predictions be if the explanation were different. This is similar to the idea of *ceteris paribus* profiles (**biecek_ceterisparibus_2020?**). *Ceteris paribus* is Latin for "other things held constant" or "all else unchanged." The profiles visualize what 'what-if' analysis, showing how an observation's prediction would change from a change in one explanatory variable given that other other variables are all held constant. In contrast to *ceteris paribus* profiles, touring methods visualize all observations rather than one, and variables are not treated as independent, that is the projection has an orthonormal basis; when one variable is rotated out-of-frame, other variables are effectively rotated into-the-frame.

3.4 Regression task

The regression case is not as discrete a feature. Instead the prediction or the residual becomes the comparison of accuracy.

3.5 Preprocessing

The benefit of having dynamic interaction with data is predicated on a reasonably small render time. It is important to preprocess as much work as possible so that application resources can be used efficiently. Below we discuss the steps and details of the reprocessing.

- **Data:** a complete numerical matrix; explanatory and response variable, an optional aesthetic (color/shape) variable can be mapped typically a categorical variable exogenous to the model.
- **Model:** any model can be used with this method. Currently, we apply random forest models via the package **randomForest** (Liaw and Wiener 2002) to mitigate the runtime of our local explanation which requires tree-based models.
- **Local explanation:** any model-compatible linear explanation could be used. We apply tree SHAP, a more computationally efficient variant of SHAP applicable to tree-based models. This is done with the package **treeshap** (Kominsarczyk et al. 2021), hosted on GitHub only]. The global view shows all observations in attribution space requiring that we must extract the variable weightings from *all* observations rather than just one.
- **Global view:** The data- and attribution-spaces are approximated as their the first two principal components.

The time to preprocess the data will vary significantly with the choice of model and local explanation. However, for reference, the FIFA data, 5000 observations of 9 explanatory variables, took 0.6 seconds to create PCA for both the data and attribution spaces. On the same data, a modestly hyper-parametered random forest model fit in 2.9 seconds, while extracting the tree SHAP values of each observation took 254 seconds combined. These runtimes were from a non-parallelized R session on a modern laptop, but suffice it to say that the bulk of the run time will be spent on the local attribution. This makes tree SHAP a good candidate to start with. The package **fastshap** (Greenwell 2020) claims extremely low runtimes that are attributed to fewer calls to the prediction function, partial implementation in C++, and efficient use of logical subsetting.

The work remaining at runtime consists mostly in rendering the frames of the tour as specified by the selection of the parameters. The application is made with **shiny** (Chang et al. 2021). The global view is created first with **ggplot2** (Wickham 2016) which is then rendered to an interactive html widget with **plotly** (Sievert 2020). The tour view is created in **spinifex** (Spyrison and Cook 2020), which creates the manual tour basis array, and facilitates similar rendering to html widget.

4 Software Infrastructure

4.1 Extend spinifex, consume DALEX & treeshap

4.2 Preprocess

4.3 Runtime rendering

5 Case Studies

5.1 1) Penguins species classification

5.2 2) FIFA wage regression

5.3 3)?

5.4 4)?

6 Discussion

7 Acknowledgments

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60.

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. 1985. "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. https://doi.org/https://doi.org/10.1137/0906011.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Buja, Andreas, and Daniel Asimov. 1986. "Grand Tour Methods: An Outline." In *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics*, 63–67. New York, NY, USA: Elsevier North-Holland, Inc. http://dl.acm.org/citation.cfm?id=26036.26046.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

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.

Dastin, Jeffrey. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*, October. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Díaz, Mark, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. "Addressing Age-Related Bias in Sentiment Analysis." In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14.

Duffy, Claire. 2019. "Apple Co-Founder Steve Wozniak Says Apple Card Discriminated Against His Wife." *CNN*, November. https://www.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html.

Gosiewska, Alicja, and Przemyslaw Biecek. 2019. "IBreakDown: Uncertainty of Model Explanations for Non-Additive Predictive Models." *arXiv Preprint arXiv:1903.11420*.

Greenwell, Brandon. 2020. *Fastshap: Fast Approximate Shapley Values*. https://CRAN.R-project.org/package=fastshap.

Kodiyan, Akhil Alfons. 2019. "An Overview of Ethical Issues in Using AI Systems in Hiring with a Case Study of Amazon's AI Based Hiring Tool." *Researchgate Preprint*.

Kominsarczyk, Konrad, Pawel Kozminski, Szymon Maksymiuk, and Przemyslaw Biecek. 2021. "Treeshap." Model Oriented. https://github.com/ModelOriented/treeshap.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=RPR1E2qtzJltfJ0tS-gB_41kmfoWZAu4.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–5.

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 Preprint arXiv:2104.08016*.

Lee, Yoon Dong, Dianne Cook, Ji-won Park, and Eun-Kyung Lee. 2013. "PPtree: Projection Pursuit Classification Tree." *Electronic Journal of Statistics* 7: 1369–86.

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

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. 2018. "Consistent Individualized Feature Attribution for Tree Ensembles." *arXiv Preprint arXiv:1802.03888*.

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

O'neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Salzberg, Steven. 2014. "Why Google Flu Is A Failure." *Forbes*, March. https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/.

Shapley, Lloyd S. 1953. *A Value for n-Person Games*. Princeton University Press.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Silva, Natalia da, Dianne Cook, and Eun-Kyung Lee. 2021. "A Projection Pursuit Forest Algorithm for Supervised Classification." *Journal of Computational and Graphical Statistics*, 1–21.

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): 243. https://doi.org/10.32614/RJ-2020-027.

Strumbelj, Erik, and Igor Kononenko. 2010. "An Efficient Explanation of Individual Classifications Using Game Theory." *The Journal of Machine Learning Research* 11: 1–18.

Štrumbelj, Erik, and Igor Kononenko. 2014. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3): 647–65.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. "Visualizing Statistical Models: Removing the Blindfold." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8 (4): 203–25. https://doi.org/10.1002/sam.11271.