ARTICLE TEMPLATE

Exploring Local Explanations of Nonlinear Models Using Animated Linear Projections

Nicholas Spyrison, Dianne Cook, Przemyslaw Biecek

[a]Monash University, Melbourne, Australia; [b]Warsaw University of Technology and University, Warsaw, Poland

ARTICLE HISTORY
Compiled May 5, 2023

ABSTRACT
The increased predictive power of nonlinear models comes at the cost of interpretability of its terms. This trade-off has led to the emergence of eXplainable AI (XAI). XAI attempts to shed light on how models use predictors to arrive at a prediction. "Local explanations", which provide a point estimate of the linear variable importance in the vicinity of one observation, are one XAI method. These can be considered to be linear projections and can be further explored to understand the interactions between variables used to make predictions across the predictive model surface. Here we describe interactive linear interpolation used for exploration at any observation and illustrate with examples from categorical (penguin species, chocolate types) and quantitative (soccer/football salaries, house prices) response models. The methods are implemented in the **R** package **cheem**, available on CRAN.

KEYWORDS
explainable artificial intelligence; nonlinear model interpretability; visual analytics; local explanations; grand tour; radial tour

1. Introduction

There are different reasons and purposes for fitting a model. According to the taxonomies of Breiman (2001b) and Shmueli (2010), it can be useful to group models into two types: explanatory and predictive. Explanatory modeling is used for inferential purposes, while predictive modeling focuses solely on the performance of an objective function. The intended use of the model has important implications for its selection and development. Interpretability is critical in explanatory modeling to draw meaningful inferential conclusions, such as which variables most contribute to a prediction or whether some observations are less well fit. Interpretability becomes more difficult when the model is nonlinear. Nonlinear models occur in statistical models with polynomial or interaction terms between quantitative predictors, and almost all computational models such as random forests, support-vector machines, or neural networks (e.g. Breiman 2001a; Boser, Guyon, and Vapnik 1992; Anderson 1995).

In linear models interpretation of the importance of variables is relatively straight forward, one adjusts for the covariance of multiple variables when examining the re-

CONTACT Nicholas Spyrison. Email: spyrison@gmail.com, Dianne Cook. Email: dicook@monash.edu, Przemyslaw Biecek. Email: przemyslaw.biecek@gmail.com

lationship with the response. The interpretation is valid for the full domain of the predictors. In nonlinear models one needs to consider the model in small neighborhoods of the domain to make any assessment of variable importance. Even though this is difficult, it is especially important to interpret model fits as we become more dependent on nonlinear models for routine aspects of life to avoid issues described in Stahl (2021). Understanding how nonlinear models behave when usage extrapolates outside the domain of predictors, either in sub-spaces where few samples were provided in the training set, or extending outside the domain. It is especially important because nonlinear models can vary wildly and predictions can be dramatically wrong in these areas.

Explainable Artificial Intelligence (XAI) is an emerging field of research focused on methods for the interpreting of models (Adadi and Berrada 2018; Barredo Arrieta et al. 2020). A class of techniques called *local explanations*, provide methods to approximate linear variable importance at the location of each observation or the predictions at a specific point in the data domain. Because these are point-specific, it is challenging to comprehensively visualize them to understand a model. There are common approaches for visualising high-dimensional data as a whole, but what is needed are new approaches for viewing these individual local explanations, in relation to the whole.

For multivariate data visualization, a *tour* (Asimov 1985; Buja and Asimov 1986; Lee et al. 2021) of linear data projections onto a lower-dimensional space, could be an element of XAI, complementing local explanations. Applying tours to model interpretation is recommended in Wickham, Cook, and Hofmann (2015) primarily to examine the fitted model in the space of the data. Cook, Swayne, and Buja (2007) describe the use of tours for exploring classification boundaries and model diagnostics (Caragea et al. 2008; Lee et al. 2013; da Silva, Cook, and Lee 2021). There are various types of tours. In a *manual* or radial tour (Cook and Buja 1997; Spyrison and Cook 2020), the path of linear projections is defined by changing the contribution of a selected variable. We propose to use this to scrutinize a local explanation. This approach could be considered to be a counterfactual, what-if analysis, such as *ceteris paribus* ("other things held constant") profiles (Biecek 2020).

The remainder of this paper is organized as follows. Section 2 covers the background of the local explanation and the traditional visuals produced. Section 3 explains the tours and particularly the radial manual tour. Section 4 discusses the visual layout in the graphical user interface and how it facilitates analysis, data pre-processing, and package infrastructure. Illustrations are provided in Section 5 for a range of supervised learning tasks with categorical and quantitative response variables. These show how the local explanations can be used to get an overview of the model's use of predictors, and to investigate errors in the model predictions. Section 6 concludes with a summary of the insights gained. The methods are implemented in the **R** package **cheem**.

2. Local Explanations

Local explanations shed light on nonlinear model fits by estimating linear variable importance in the vicinity of a single observation. There are many approaches for calculating local explanations. A comprehensive summary of the taxonomy of currently available methods is provided in Figure 6 by Barredo Arrieta et al. (2020). It includes a large number of model-specific explanations such as deepLIFT (Shrikumar et al. 2016; Shrikumar, Greenside, and Kundaje 2017), a popular recursive method for estimating

importance in neural networks. There are fewer model-agnostic methods, of which LIME, (Ribeiro, Singh, and Guestrin 2016) SHaply Additive exPlanations (SHAP), (Lundberg and Lee 2017), are popular.

These observation-level explanations are used in various ways depending on the data. In image classification, where pixels correspond to predictors, saliency maps overlay or offset a heatmap to indicate important pixels (Simonyan, Vedaldi, and Zisserman 2014). For example, pixels corresponding to snow may be highlighted as important contributors when distinguishing if a picture contains a coyote or husky. In text analysis, word-level contextual sentiment analysis highlights the sentiment and magnitude of influential words (Vanni et al. 2018). In the case of numeric regression, they are used to explain additive contributions of variables from the model intercept to the observation's prediction (Ribeiro, Singh, and Guestrin 2016).

We will be focusing on SHAP values in this this paper, but the approach is applicable for any method used to calculate the local explanations. SHAP calculates the variable contributions of one observation by examining the effect of other variables on the predictions. The term "SHAP" refers to Shapley (1953)'s method to evaluate an individual's contribution in cooperative games by assessing this player's performance in the presence or absence of other players. Strumbelj and Kononenko (2010) introduced SHAP for local explanations in machine learning models. Variable importance can depend on the sequence in which variables are entered into the model fitting process, thus for any sequence we get a set of variable contribution values for a single observation. These values will add up to the difference between the fitted value for the observation, and the average fitted value for all observations. Using all possible sequence, or permutation, gives multiple values for each variable, which are averaged to get the SHAP value for an observation. It can be helpful to standardize variables prior to computing SHAP values, if they have been measured on different scales.

The approach is related to partial dependence plots (see for example Molnar (2020)), used to explain the effect of a variable by predicting the response for a range of values on this variable after fixing the value of all other variables to their mean. Though partial dependence plots are a global approximation of the variable importance, while SHAP is specific to one observation.

We use 2020 season FIFA data (Leone 2020) to illustrate SHAP following the procedures described in Biecek and Burzykowski (2021). There are 5000 observations of nine predictor variables measuring players' skills and one response variable, wages (in euros). A random forest model is fit regressing players' wages on the skill variables. In this illustration in Figure 1 the SHAP values are compared for a star offensive player (L. Messi) and a prominent defensive player (V. van Dijk). We are interested in knowing how the skill variables locally contribute to the wage prediction of each player. A difference in the attribution of the variable importance across the two positions of the players can be expected. This would be interpreted as how a player's salary depends on which combination of skills. Panel (a) is a version of a breakdown plot (Gosiewska and Biecek 2019) where just three sequences of variables are shown, for two observations. A breakdown plot shows the absolute values of the variable attribution for an observation, usually sorted from highest value to the lowest. There is no scale on the horizontal axis here because values are considered relative to each other. Here we can see how the variable contribution can change depending on sequence, relative to both players. (Note that the order of the variables is different in each plot, because they have been sorted by biggest average contribution across both players.) For all sequences, and for both players `reaction` has the strongest contribution, with perhaps more importance for the defensive player. Then it differs by player: for Messi `offense`

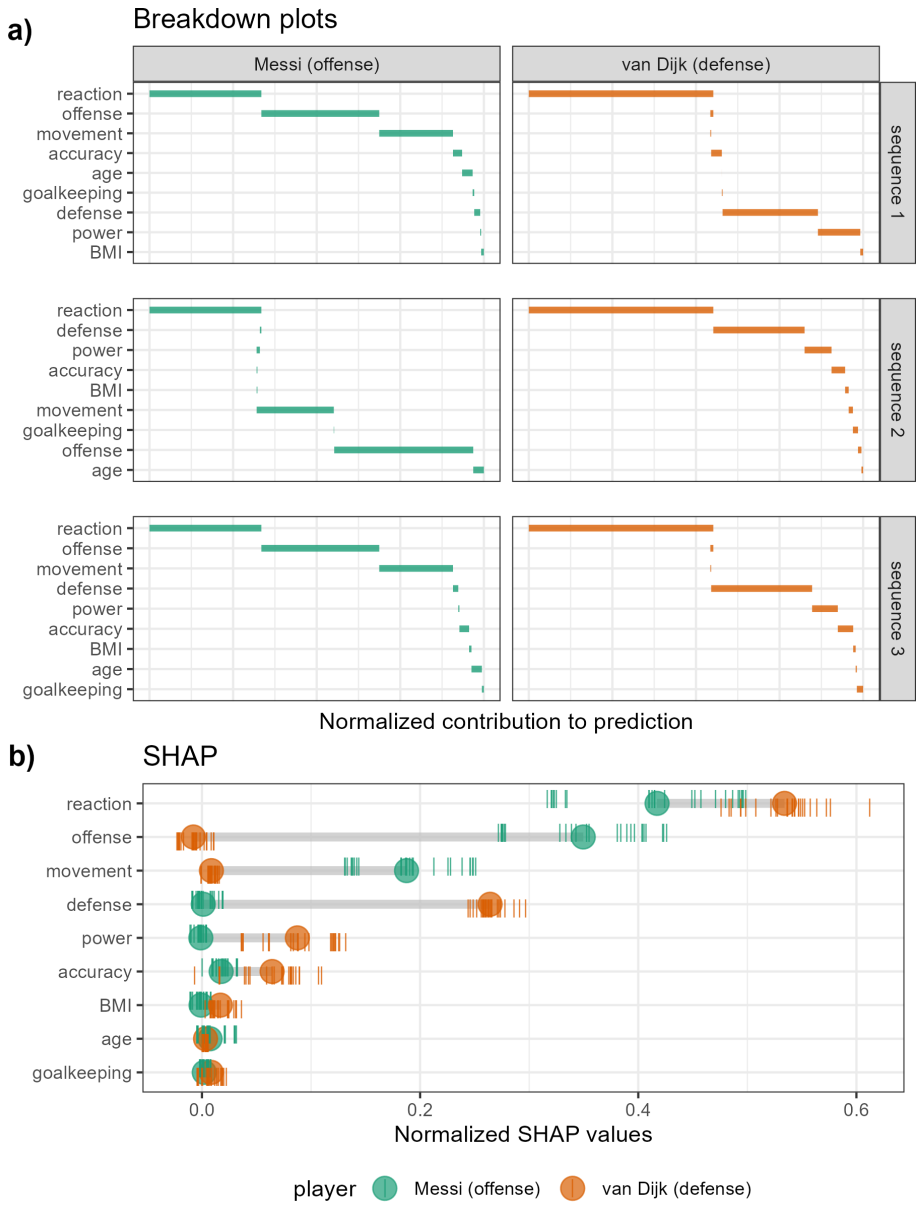

Figure 1. Illustration of SHAP values for a random forest model FIFA 2020 player wages from nine skill predictors. A star offensive and defensive player are compared, L. Messi and V. van Dijk, respectively. Panel (a) shows breakdown plots of three sequences of the variables. The sequence of the variables impacts the magnitude of their attribution. Panel (b) shows the distribution of attribution for each variable across 25 sequences of predictors, with the mean displayed as a dot for each player. Reaction skills are important for both players. Offense and movement are important for Messi but not van Dijk, and conversely, defense and power are important for van Dijk but not Messi.

and `movement` have the strongest contributions, and for van Dijk it is `defense` and `power`, regardless of the variable sequence.

Panel (b) shows the differences in the player's median values (large dots) for 25 such sequences (tick marks). We can see that the wage predictions for the two players come from different combinations of skills sets, as might be expected for players who's value on the team depends on their offensive or defensive prowess. It is also interesting to see from the distribution of values across the different sequence of variables, that there is some multimodality. For example, look at the SHAP values for `reaction` for Messi, and in some sequences reaction has a much lower contribution than others. This suggests that other variables (`offense`, `movement` probably) can substitute for `reaction` in the wage prediction.

This can also be considered similar to examining the coefficients from all subsets regression, as described in Wickham, Cook, and Hofmann (2015). Various models that are similarly good might use different combinations of the variables. Examining the coefficients from multiple models helps to understand the relative importance of each variable in the context of all other variables. This is similar to the approach here with SHAP values, that by examining the variation in values across different permutations of variables, we can gain more understanding of the relationship between the response and predictors.

For the application, we use *tree SHAP*, a variant of SHAP that enjoys a lower computational complexity (Lundberg, Erion, and Lee 2018). Instead of aggregating over sequences of the variables, tree SHAP calculates observation-level variable importance by exploring the structure of the decision trees. Tree SHAP is only compatible with tree-based models. so random forests are used for illustration.

3. Tours and the Radial Tour

A *tour* enables the viewing of high-dimensional data by animating many linear projections with small incremental changes. It is achieved by following a path of linear projections (bases) of high-dimensional space. One key variable of the tour is the object permanence of the data points; one can track the relative change of observations in time and gain information about the relationships between points across multiple variables. There are various types of tours that are distinguished by how the paths are generated (Lee et al. 2021; Cook et al. 2008).

The manual tour (Cook and Buja 1997) defines its path by changing a selected variable's contribution to a basis to allow the variable to contribute more or less to the projection. The requirement constrains the contribution of all other variables that a basis needs to be orthonormal (column correspond to vectors, with unit length, and orthogonal to each other). The manual tour is primarily used to assess the importance of a variable to structure visible in a projection. It also lends itself to pre-computation queued in advance or computed on-the-fly for human-in-the-loop analysis (Karwowski 2006).

A version of the manual tour called a *radial tour* is implemented in Spyrison and Cook (2020) and forms the basis of this new work. In a radial tour, the selected variable can change its magnitude of contribution but not its angle; it must move along the direction of its original contribution. The implementation allows for pre-computation and interactive re-calculation to focus on a different variable.

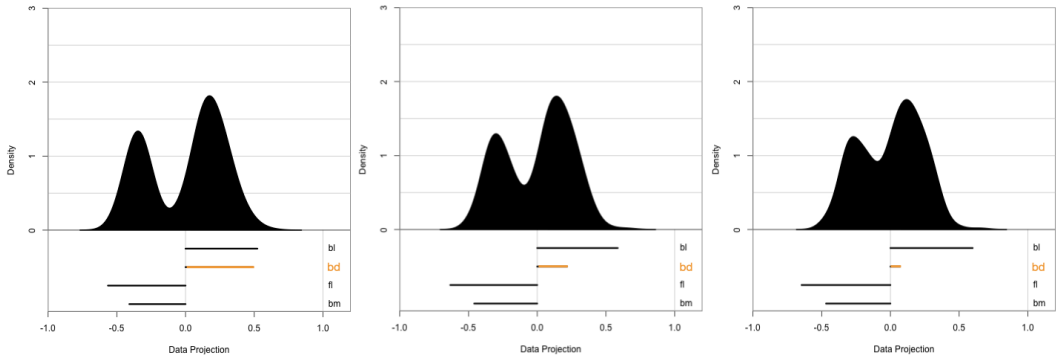

Figure 2. The radial tour allows the user to remove a variable from a projection, to examine the importance of this variable to structure in the plot. Here we have a 1D projection of the penguins data displayed as a density plot. The line segments on the bottom correspond to the coefficients of the variables making up the projection. The structure in the plot is bimodality (left), and the importance of the variable `bd` is being explored. As this variable contribution is reduced in the plot (middle, right) we can see that the bimodality decreases. Thus `bd` is an important variable contributing to the bimodal structure.

4. The Cheem Viewer

To explore the local explanations, coordinated views (Roberts 2007) (also known as ensemble graphics, Unwin and Valero-Mora 2018) are provided in the *cheem viewer* application. There are two primary plots: the **global view** to give the context of all of the SHAP values and the **radial tour view** to explore the local explanations with user-controlled rotation. There are numerous user inputs, including variable selection for the radial tour and observation selection for making comparisons. There are different plots used for the categorical and quantitative responses. Figures 3 and 4 are screenshots showing the cheem viewer for the two primary tasks: classification (categorical response) and regression (quantitative response).

4.1. *Global View*

The global view provides context for all observations and facilitates the exploration of the separability of the data- and attribution-spaces. These spaces both have dimensionality $n \times p$, where n is the number of observations and p is the number of variables. The attribution space corresponds to the local explanations for each observation; variable importance in the vicinity of the observation.

The visualization is composed of the first two principal components of the data (left) and the attribution (middle) spaces. These single 2D projections will not reveal all of the structure of higher-dimensional space, but they are helpful visual summaries. In addition, a plot of the observed against predicted response values is also provided (Figures 3b, 4a) to help identify observations poorly predicted by the model. For classification tasks, misclassified observations are circled in red. Linked brushing between the plots is provided, and a tabular display of selected points helps to facilitate exploration of the spaces and the model (shown in Figures 4d).

While the comparison of these spaces is interesting, the primary purpose of the global view is to enable the selection of particular observations to explore in detail. The projection attribution of the primary observation (PI) is examined and typically viewed with an optional comparison observation (CI). These observations are highlighted as asterisk and $\times$, respectively.

4.2. *Radial Tour*

The local explanations for all observations are normalized (sum of squares equals 1), and thus, the relative importance of variables can be compared across all observations. These are depicted as vertical parallel coordinate plots (Ocagne 1885) on the basis biplot (Gabriel 1971). 1D biplot displays the values of the current basis as bars. The parallel coordinate overlays lines connecting one observation's variable attribution (Figures 3e and 4e). The attribution projections of the PI and CI are shown as dashed and dotted lines. From this plot, the range and density of the importance across all observations can be interpreted. For classification, one would look at differences between groups on any variable. For example, Figure 3e suggests that `bl` is important for distinguishing the green class from the other two. For regression, one might generally observe which variables have low values for all observations (not important). For example, `BMI` and `pwr` in Figure 4e, have a range of high and low values (e.g., `off`, `def`), suggesting they are important for some observations and not important for others.

The overlaid bars on the parallel coordinate plot represent the attribution projection of the PI. (Remember that the PI is interactively selected from the global view). The attribution projection approximates the variable importance for predicting this observation. The combination of variables best explains the difference between the mean response and an observation's predicted value. It is not an indication of the local shape of the model surface. That is, it is not some indication of the tangent to the curve at this point.

The attribution projection of the PI is the initial 1D basis in a radial tour, displayed as a density plot for a categorical response (Figure 3f) and as scatterplots for a quantitative response (Figure 4f). The PI and CI are indicated by vertical dashed and dotted lines. The radial tour varies the contribution of the selected variable between 0 and 1. This is viewed as an animation of the projections from many intermediate bases. Doing so tests the sensitivity of structure (class separation or strength of relationship) to the variable's contribution. For classification, if the separation between classes diminishes when the variable contribution is reduced, this suggests that the variable is important for class separation. For regression, if the relationship scatterplot weakens when the variable contribution is reduced, indicating that the variable is important for accurately predicting the response.

4.3. *Classification Task*

Selecting a misclassified observation as PI and a correctly classified point nearby in data space as CI makes it easier to examine the variables most responsible for the error. The global view (Figure 3c) displays the model confusion matrix. The radial tour is 1D and displays as density where color indicates class. An animation slider enables users to vary the contribution of variables to explore the sensitivity of the separation to that variable.

4.4. *Regression Task*

Selecting an inaccurately predicted observation as PI and an accurately predicted observation with similar variable values as CI is a helpful way to understand how the model is failing or not. The global view (Figure 4a) shows a scatterplot of the observed vs predicted values, which should exhibit a strong relationship if the model is a good fit. The points can be colored by a statistic, residual, a measure of outlyingness (log

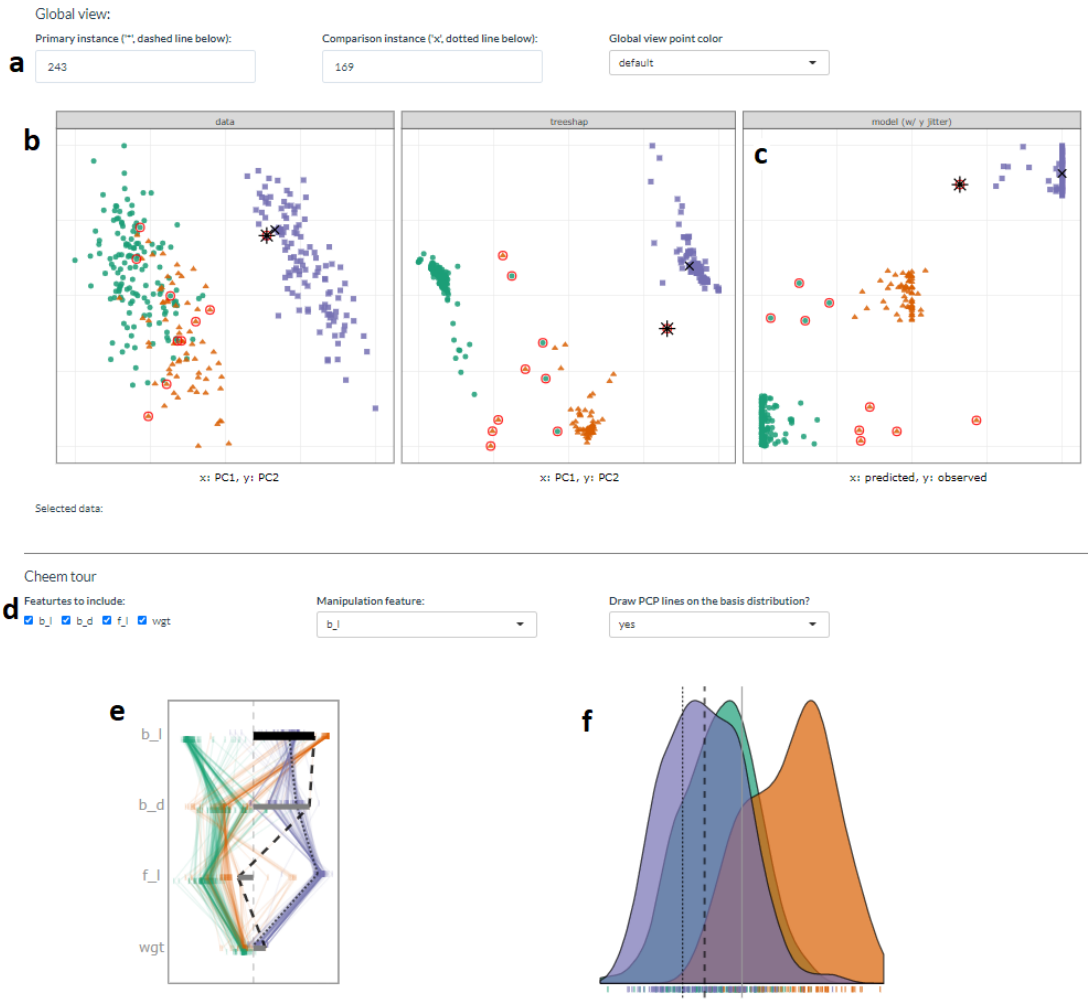

Figure 3. Overview of the cheem viewer for classification tasks (categorical response). Global view inputs, (a), set the PI, CI, and color statistic. Global view, (b) PC1 by PC2 approximations of the data- and attribution-space. (c) prediction by observed y (visual of the confusion matrix for classification tasks). Points are colored by predicted class, and red circles indicate misclassified observations. Radial tour inputs (d) select variables to include and which variable is changed in the tour. (e) shows a parallel coordinate display of the distribution of the variable attributions while bars depict contribution for the current basis. The black bar is the variable being changed in the radial tour. Panel (f) is the resulting data projection indicated as density in the classification case.

Mahalanobis distance), or correlation to aid in understanding the structure identified in these spaces.

In the radial tour view, the observed response and the residuals (vertical) are plotted against the attribution projection of the PI (horizontal). The attribution projection can be interpreted similarly to the predicted value from the global view plot. It represents a linear combination of the variables, and a good fit would be indicated when there is a strong relationship with the observed values. This can be viewed as a local linear approximation if the fitted model is nonlinear. As the contribution of a variable is varied, if the value of the PI does not change much, it would indicate that the prediction for this observation is NOT sensitive to that variable. Conversely, if the predicted value varies substantially, the prediction is very sensitive to that variable, suggesting that the variable is very important for the PI's prediction.

4.5. *Interactive variables*

The application has several reactive inputs that affect the data used, aesthetic display, and tour manipulation. These reactive inputs make the software flexible and extensible (Figure 3a & d). The application also has more exploratory interactions to help link points across displays, reveal structures found in different spaces, and access the original data.

A tooltip displays observation number/name and classification information while the cursor hovers over a point. Linked brushing allows the selection of points (left click and drag) where those points will be highlighted across plots (Figure 3a & b). The information corresponding to the selected points is populated on a dynamic table (Figure 3d). These interactions aid exploration of the spaces and, finally, the identification of primary and comparison observations.

4.6. *Preprocessing*

It is vital to mitigate the render time of visuals, especially when users may want to iterate many explorations. All computational operations should be prepared before runtime. The work remaining when an application is run solely reacts to inputs and rendering visuals and tables. Below discusses the steps and details of the reprocessing.

- **Data:** predictors and response are unscaled complete numerical matrix. Most models and local explanations are scale-invariant. Keep normality assumptions of the model in mind.
- **Model:** any model and compatible explanation could be explored with this method. Currently, random forest models are applied via the package **randomForest** (Liaw and Wiener 2002), compatibility tree SHAP. Modest hyperparameters are used, namely: 125 trees, the number of variables at each split, mtry = $\sqrt{p}$ or $p/3$ for classification and regression, and minimum size of terminal nodes $max(1, n/500)$ or $max(5, n/500)$ for classification and regression.
- **Local explanation:** Tree SHAP is calculated for *each* observation using the package **treeshap** (Kominsarczyk et al. 2021). We opt to find the attribution of each observation in the training data and not fit to fit variable interactions.
- **Cheem viewer:** after the model and full explanation space are calculated, each variable is scaled by standard deviations away from the mean to achieve common support for visuals. Statistics for mapping to color are computed on the scaled spaces.

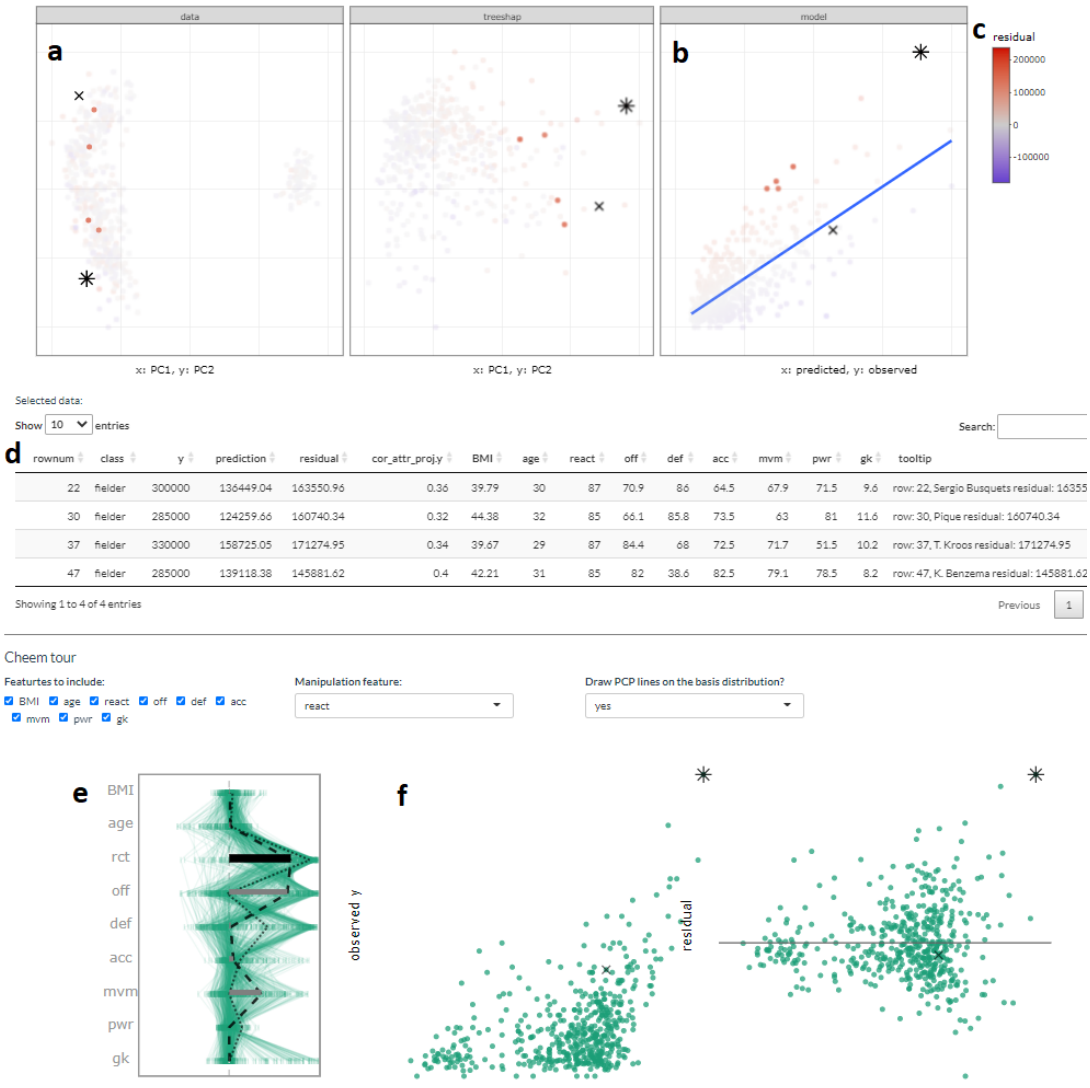

Figure 4. Overview of the cheem viewer for regression tasks (quantitative response) and illustration of interactive variables. Panel (a) PCA of the data- and attributions- spaces and the (b) residual plot, predictions by observed values. Four selected points are highlighted in the PC spaces and tabularly displayed. Coloring on a statistic (c) highlights structure organized in the attribution space. Interactive tabular display (d) populates when observations are selected. Contribution of the 1D basis affecting the horizontal position (e) parallel coordinate display of the variable attribution from all observations, and horizontal bars show the contribution to the current basis. Regression projection (f) uses the same horizontal projection and fixes the vertical positions to the observed y and residuals (middle and right).

The time to preprocess the data will vary significantly with the complexity of the model and local explanation. For reference, the FIFA data contained 5000 observations of nine explanatory variables took 2.5 seconds to fit a random forest model of modest hyperparameters. Extracting the tree SHAP values of each observation took 270 seconds total. PCA and statistics of the variables and attributions took 2.8 seconds. These runtimes were from a non-parallelized session on a modern laptop, but suffice to say that most of the time will be spent on the local attribution. An increase in model complexity or data dimensionality will quickly become an obstacle. Its reduced computational complexity makes tree SHAP an excellent candidate to start. (Alternatively, the package **fastshap** claims extremely low runtimes, attributed to fewer calls to the prediction function, partial implementation in C++, and efficient use of logical subsetting, Greenwell 2020)

4.7. *Package Infrastructure*

The above-described method and application are implemented as an open-source **R** package, **cheem** available on CRAN at `https://CRAN.R-project.org/package=cheem`. Preprocessing was facilitated with models created via **randomForest** (Liaw and Wiener 2002) and explanations calculated with **treeshap** (Kominsarczyk et al. 2021). The application was made with **shiny** (Chang et al. 2021). The tour visual is built with **spinifex** (Spyrison and Cook 2020). Both views are created first with **ggplot2** (Wickham 2016) and then rendered as interactive `html` widgets with **plotly** (Sievert 2020). **DALEX** (Biecek 2018) and the free ebook, *Explanatory Model Analysis* (Biecek and Burzykowski 2021) were a boon to understanding local explanations and how to apply them.

4.8. *Installation and Getting Started*

The package can be installed from CRAN, and the application can be run using the following **R** code:

```
install.packages("cheem", dependencies = TRUE)
library("cheem")
run_app()
```

Alternatively,

- A version of the cheem viewer shiny app can be directly accessed at `https://ebsmonash.shinyapps.io/cheem_initial/`.
- The development version of the package is available at `https://github.com/nspyrison/cheem`, and
- Documentation of the package can be found at `https://nspyrison.github.io/cheem/`.

Follow the examples provided with the package to compute the local explanation (using `?cheem_ls`). The application expects the output returned by `cheem_ls()`, saved to an `rds` file with `saveRDS()` to be uploaded.

5. Case Studies

To illustrate the cheem method it is applied to modern data sets, two classification examples and then two of regression.

5.1. *Palmer Penguin, Species Classification*

The Palmer penguins data (Gorman, Williams, and Fraser 2014; Horst, Hill, and Gorman 2020) was collected on three species of penguins foraging near Palmer Station, Antarctica. The data is publicly available to substitute for the overly-used iris data and is quite similar in form. After removing incomplete observations, there are 333 observations of four physical measurements, bill length (`bl`), bill depth (`bd`), flipper length (`fl`), and body mass (`bm`) for this illustration. A random forest model was fit with species as the response variable.

(ref:casepenguins-cap) Examining the SHAP values for a random forest model classifying Palmer penguin species. The PI is a Gentoo (purple) penguin that is misclassified as a Chinstrap (orange), marked as an asterisk in (a) and the dashed vertical line in (b). The radial view shows varying the contribution of `fl` from the initial attribution projection (b, left), which produces a linear combination where the PI is more probably (higher density value) a Chinstrap than a Gentoo (b, right). (The animation of the radial tour is at `https://vimeo.com/666431172`.)

Figure 5 shows plots from the cheem viewer for exploring the random forest model on the penguins data. Panel (a) shows the global view, and panel (b) shows several 1D projections generated with the radial tour. Penguin 243, a Gentoo (purple), is the PI because it has been misclassified as a Chinstrap (orange).

(ref:casepenguinsblfl-cap) Checking what is learned from the cheem viewer. This is a plot of flipper length (`fl`) and bill length (`bl`), where an asterisk highlights the PI. A Gentoo (purple) misclassified as a Chinstrap (orange). The PI has an unusually small `fl` length which is why it is confused with a Chinstrap.

There is more separation visible in the attribution space than in the data space, as would be expected. The predicted vs observed plot reveals a handful of misclassified observations. A Gentoo that has been wrongly labeled as a Chinstrap is selected for illustration. The PI is a misclassified point (represented by the asterisk in the global view and a dashed vertical line in the tour view). The CI is a correctly classified point (represented by an $\times$ and a vertical dotted line).

The radial tour starts from the attribution projection of the misclassified observation (b, left). The important variables identified by SHAP in the (wrong) prediction for this observation are mostly `bl` and `bd` with small contributions of `fl` and `bm`. This projection is a view where the Gentoo (purple) looks much more likely for this observation than Chinstrap. That is, this combination of variables is not particularly useful because the PI looks very much like other Gentoo penguins. The radial tour is used to vary the contribution of flipper length (`fl`) to explore this. (In our exploration, this was the third variable explored. It is typically helpful to explore the variables with more significant contributions, here `bl` and `bd`. Still, when doing this, nothing was revealed about how the PI differed from other Gentoos). On varying `fl` as it contributes increasingly to the projection (b, right), more and more, this penguin looks like a Chinstrap. This suggests that `fl` should be considered an important variable for explaining the (wrong) prediction.

Figure 6 confirms that flipper length (`fl`) is vital for the confusion of the PI as a

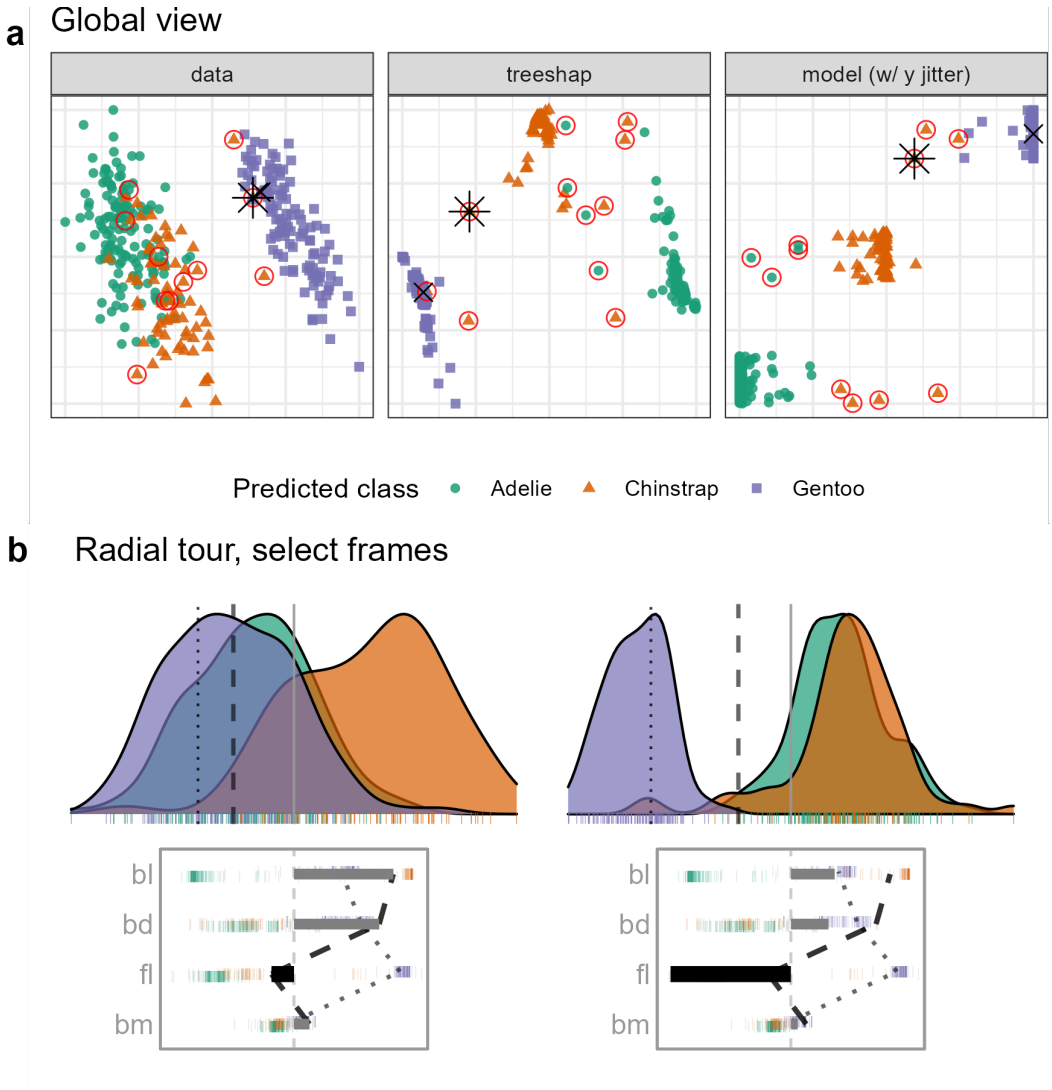

Figure 5. (ref:casepenguins-cap)

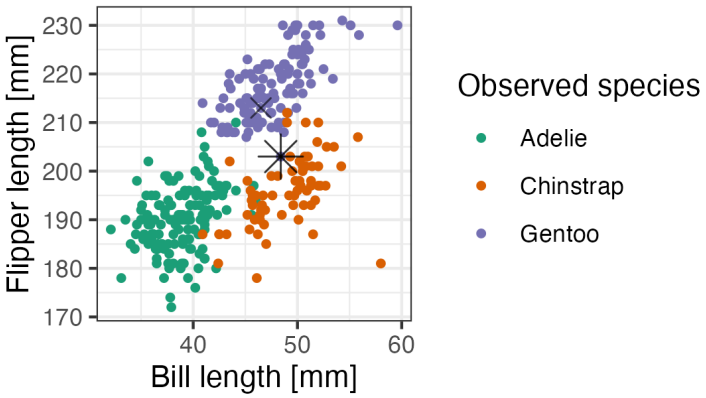

Figure 6. (ref:casepenguinsblfl-cap)

Chinstrap. Here, flipper length and body length are plotted, and the PI can be seen to be closer to the Chinstrap group in these two variables, mainly because it has an unusually low value of flipper length relative to other Gentoos. From this view, it makes sense that it is a hard observation to account for, as decision trees can only partition only vertical and horizontal lines.

5.2. *Chocolates, Milk/Dark Classification*

The chocolates data set consists of 88 observations of ten nutritional measurements determined from their labels and labeled as either milk or dark. Dark chocolate is considered healthier than milk. Students collected the data during the Iowa State University class STAT503 from nutritional information on the manufacturer's websites and were normalized to 100g equivalents. The data is available in the **cheem** package. A random forest model is used for the classification of chocolate types.

It could be interesting to examine the nutritional properties of any dark chocolates that have been misclassified as milk. A reason to do this is that a dark chocolate, nutritionally more like milk should not be considered a healthy alternative. It is interesting to explore which nutritional variables contribute most to misclassification.

(ref:casechocolates-cap) Examining the local explanation for a PI which is dark (orange) chocolate incorrectly predicted to be milk (green). From the attribution projection, this chocolate correctly looks more like dark than milk, which suggests that the local explanation does not help understand the prediction for this observation. So, the contribution of Sugar is varied—reducing it corresponds primarily with an increased magnitude from Fiber. When Sugar is zero, Fiber contributes strongly towards the left. In this view, the PI is closer to the bulk of the milk chocolates, suggesting that the prediction put a lot of importance on Fiber. This chocolate is a rare dark chocolate without any Fiber leading to it being mistaken for a milk chocolate. (A video of the tour animation can be found at `https://vimeo.com/666431143`.)

This type of exploration is shown in Figure 7, where a chocolate labeled dark but predicted to be milk is chosen as the PI (observation 22). It is compared with a CI that is a correctly classified dark chocolate (observation 7). The PCA plot and the tree SHAP PCA plots (a) show a big difference between the two chocolate types but with confusion for a handful of observations. The misclassifications are more apparent in the observed vs predicted plot and can be seen to be mistaken in both ways: milk to dark and dark to milk.

The attribution projection for chocolate 22 suggests that Fiber, Sugars, and Calories are most responsible for its incorrect prediction. The way to read this plot is to see that Fiber has a large negative value while Sugars and Calories have reasonably large positive values. In the density plot, observations on the very left of the display would have high values of Fiber (matching the negative projection coefficient) and low values of Sugars and Calories. The opposite would be interpreting a point with high values in this plot. The dark chocolates (orange) are primarily on the left, and this is a reason why they are considered to be healthier: high fiber and low sugar. The density of milk chocolates is further to the right, indicating that they generally have low fiber and high sugar.

The PI (dashed line) can be viewed against the CI (dotted line). Now, one needs to pay attention to the parallel plot of the SHAP values, which are local to a particular observation, and the density plot, which is the same projection of all observations as specified by the SHAP values of the PI. The variable contribution of the two different

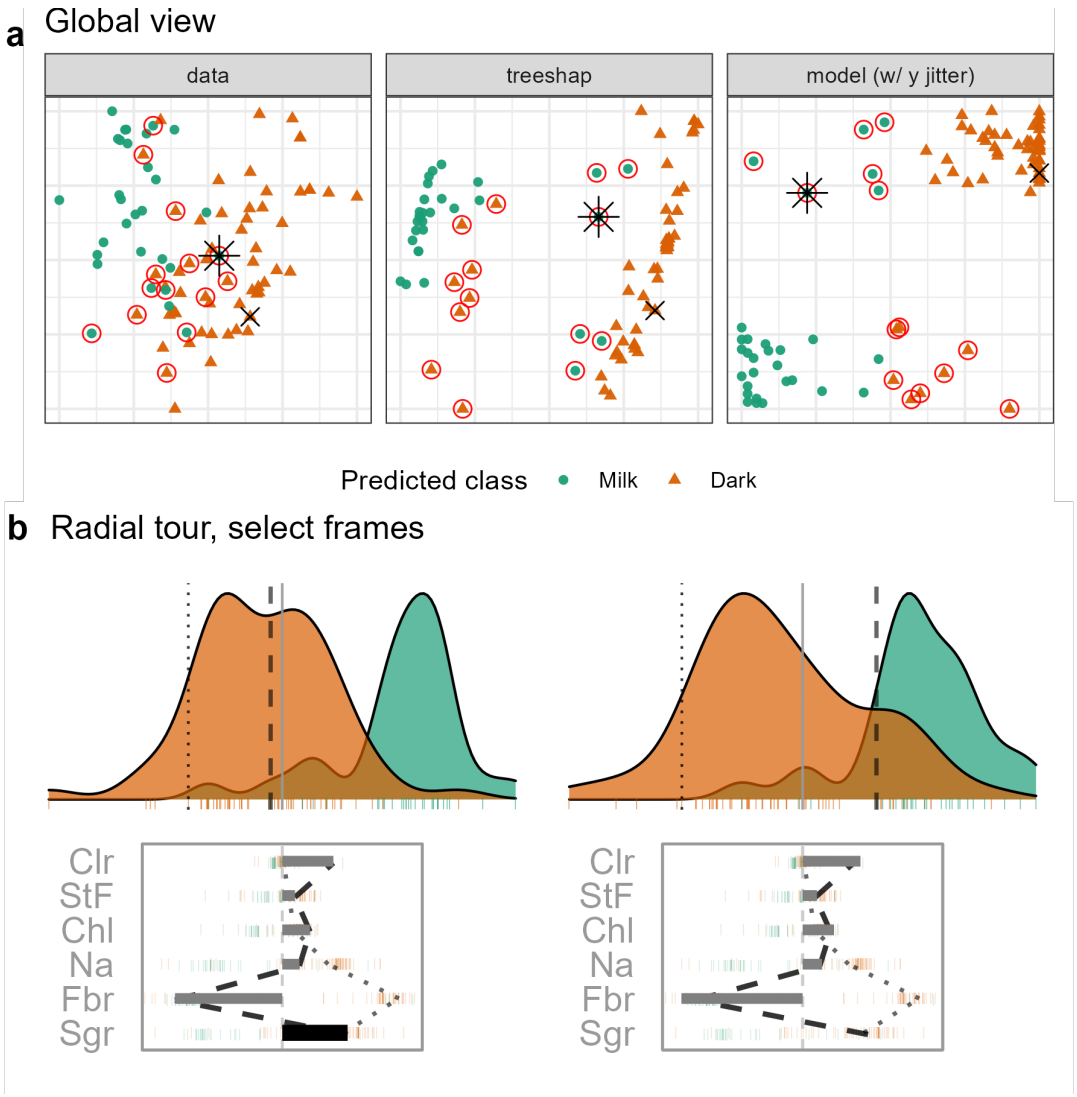

Figure 7. (ref:casechocolates-cap)

predictions can be quickly compared in the parallel coordinate plot. The PI differs from the comparison primarily on the Fiber variable, which suggests that this is the reason for the incorrect prediction.

From the density plot, which is the attribution projection corresponding to the PI, both observations are more like dark chocolates. Varying the contribution of Sugars and altogether removing it from the projection is where the difference becomes apparent. When a frame with contribution primarily from Fiber is examined observation 22 looks more like a milk chocolate.

It would also be interesting to explore an inverse misclassification. In this case, a milk chocolate is selected while it was misclassified as a dark chocolate. Chocolate 84 is just this case and is compared with a correctly predicted milk chocolate (observation 71). The corresponding global view and radial tour frames are shown in Figure 8.

(ref:casechocolatesinverse-cap) Examining the local explanation for a PI which is milk (green) chocolate incorrectly predicted to be dark (orange). In the attribution projection, the PI could be either milk or dark. Sodium and Fiber have the largest differences in attributed variable importance, with low values relative to other milk chocolates. The lack of importance attributed to these variables is suspected of contributing to the mistake, so the contribution of Sodium is varied. If Sodium had a larger contribution to the prediction (like in this view). the PI would look more like other milk chocolates. (A video of the tour animation can be found at `https://vimeo.com/666431148`.)

The difference of position in the tree SHAP PCA with the previous case is quite significant; this gives a higher-level sense that the attributions should be quite different. Looking at the attribution projection, this is found to be the case. Previously, Fiber was essential while it is absent from the attribution in this case. Conversely, Calories from Fat and Total Fat have high attributions here, while they were unimportant in the preceding case.

Comparing the attribution with the CI (dotted line), large discrepancies in Sodium and Fiber are identified. The contribution of Sodium is selected to be varied. Even in the initial projection, the observation looks slightly more like its observed milk than predicted dark chocolate. The misclassification appears least supported when the basis reaches sodium attribution of typical dark chocolate.

5.3. *FIFA, Wage Regression*

The 2020 season FIFA data (Leone 2020; Biecek 2018) contains many skill measurements of soccer/football players and wage information. Nine higher-level skill groupings were identified and aggregated from highly correlated variables. A random forest model is fit from these predictors, regressing player wages [2020 euros]. The model was fit from 5000 observations before being thinned to 500 players to mitigate occlusion and render time. Continuing from the exploration in Section \ref{sec:explanations), we are interested to see the difference in attribution based on the exogenous player position. That is, the model should be able to use multiple linear profiles to better predict the wages from different field positions of players despite not having this information. A leading offensive fielder (L. Messi) is compared with a top defensive fielder (V. van Dijk). The same observations were used in Figure 1.

(ref:casefifa-cap) Exploring the wages relative to skill measurements in the FIFA 2020 data. Star offensive player (L. Messi) is the PI, and he is compared with a top defensive player (V. van Dijk). The attribution projection is shown on the left,

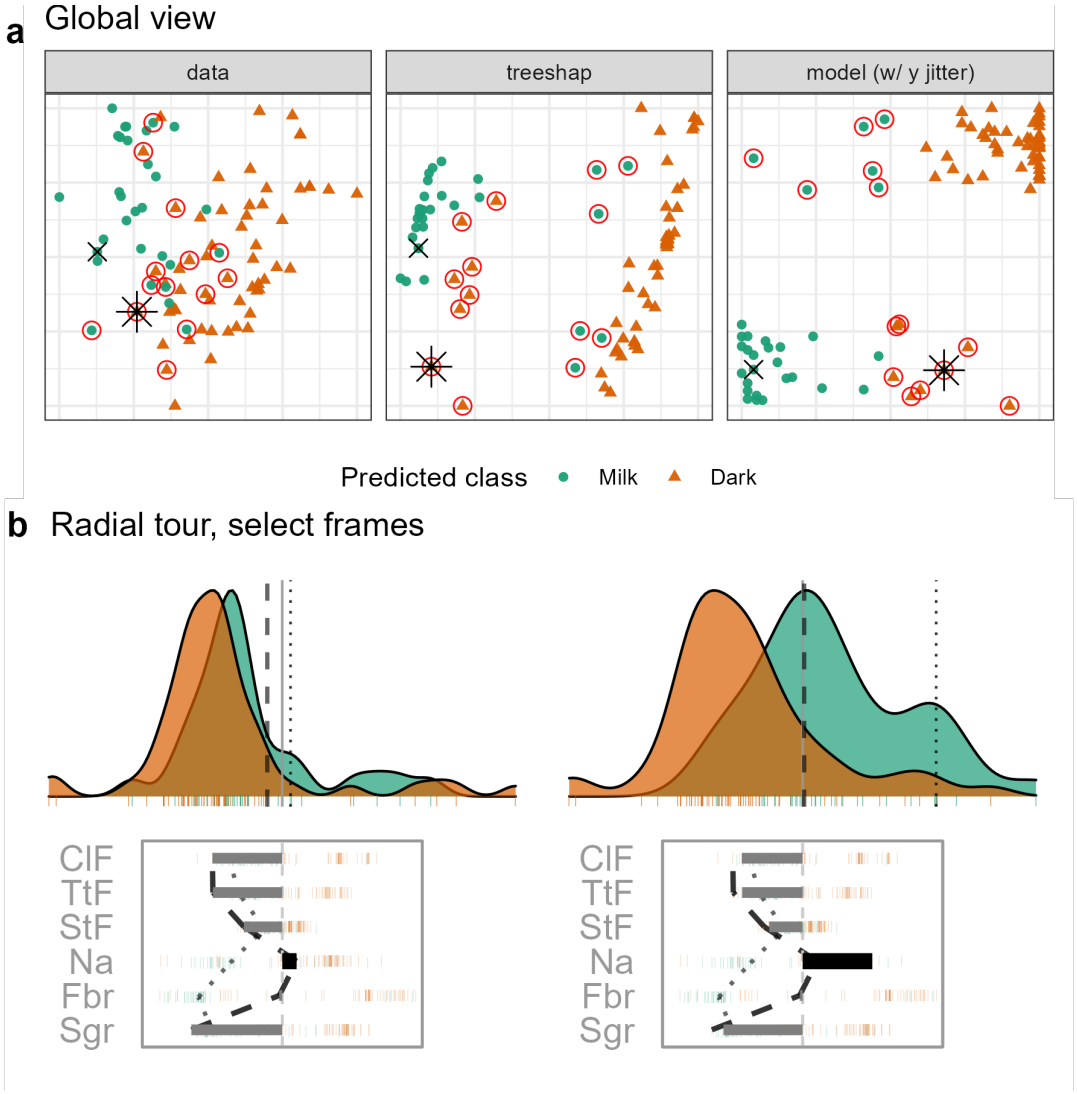

Figure 8. (ref:casechocolatesinverse-cap)

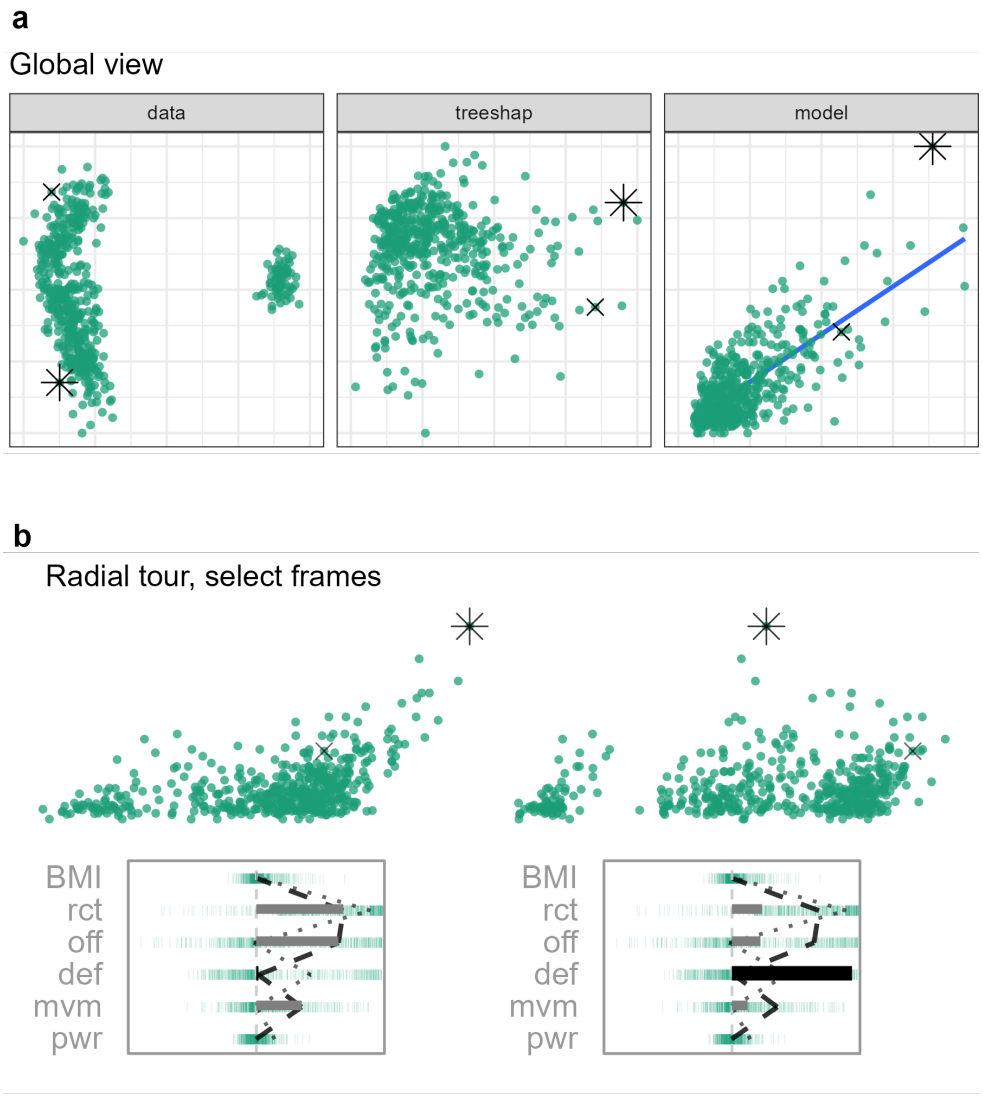

Figure 9. (ref:casefifa-cap)

and it can be seen that this combination of variables produces a view where Messi has very high predicted (and observed) wages. Defense (`def`) is the chosen variable to vary. It starts very low, and Messi's predicted wages decrease dramatically as its contribution increases (right plot). The increased contribution in defense comes at the expense of offensive and reaction skills. The interpretation is that Messi's high wages are most attributable to his offensive and reaction skills, as initially provided by the local explanation. (A video of the animated radial tour can be found at `https://vimeo.com/666431163`.)

Figure 9 tests the support of the local explanation. Offensive and reaction skills (`off` and `rct`) are both crucial to explaining a star offensive player. If either of them were rotated out, the other would be rotated into the frame, maintaining a far-right position. However, increasing the contribution of a variable with low importance would rotate both variables out of the frame.

The contribution from `def` will be varied to contrast with offensive skills. As the contribution of defensive skills increases, Messi's is no longer separated from the group. Players with high values in defensive skills are now the rightmost points. In terms of what-if analysis, the difference between the data mean and his predicted wages would be halved if Messi's tree SHAP attributions were at these levels.

(ref:caseames-cap) Exploring an observation with an extreme residual as the PI in relation to an observation with an accurate prediction for a similarly priced house in a random forest fit to the Ames housing data. The local explanation indicates a sizable attribution to Lot Area (`LtA`), while the CI has minimal attribution to this variable. The PI has a higher predicted value than the CI in the attribution projection. Reducing the contribution of Lot Area brings these two prices in line. This suggests that if the model did not value Lot Area so highly for this observation, then the observed sales price would be quite similar. That is, the large residual is due to a lack of factoring in the Lot Area for the prediction of PI's sales price. (A video showing the animation is at `https://vimeo.com/666431134`.)

5.4. *Ames Housing 2018, Sales Price Regression*

Ames housing data 2018 (De Cock 2011; prevek18 2018) was subset to North Ames (the neighborhood with the most house sales). The remaining are 338 house sales. A random forest model was fit, predicting the sale price [USD] from the property variables: Lot Area (`LtA`), Overall Quality (`Qlt`), Year the house was Built (`YrB`), Living Area (`LvA`), number of Bathrooms (`Bth`), number of Bedrooms (`Bdr`), the total number of Rooms (`Rms`), Year the Garage was Built (`GYB`), and Garage Area (`GrA`). Using interactions with the global view, a house with an extreme negative residual and an accurate observation with a similar prediction is selected.

Figure 10 selects the house sale 74, a sizable under prediction with an enormous Lot Area contribution. The CI has a similar predicted price though the prediction was accurate and gives almost no attribution to lot size. The attribution projection places observations with high Living Areas to the right. The contribution of Living Area contrasts the contribution of this variable. As the contribution of Lot Area decreases, the predictive power decreases for the PI, while the CI remains stationary. This large importance in the Living Area is relatively uncommon. Boosting tree models may be more resilient to such an under-prediction as they would up-weighting this residual and force its inclusion in the final model.

6. Discussion

There is a clear need to extend the interpretability of black box models. This paper provides a technique that builds on local explanations to explore the variable importance local to an observation. The local explanations of an attribution projection from which variable contributions are varied using a radial tour. Several diagnostic plots are provided to assist with understanding the sensitivity of the prediction to particular variables. A global view shows the data space, explanation space, and residual plot. The user can interactively select observations to compare, contrast, and study further. Then the radial tour is used to explore the variable sensitivity identified by the attribution projection.

This approach has been illustrated using four data examples of random forest models with the tree SHAP local explanation. Local explanations focus on the model fit, and

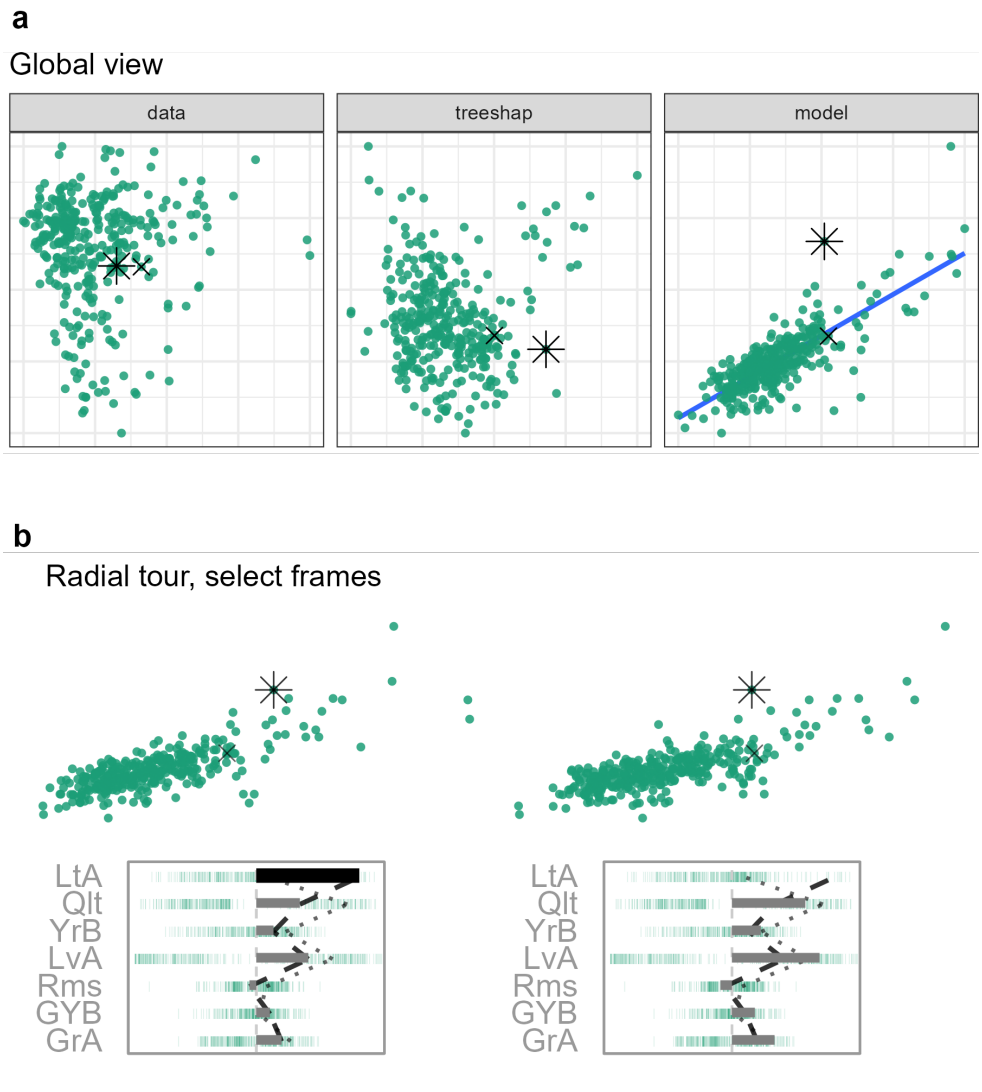

Figure 10. (ref:caseames-cap)

help to dissect which variables are most responsible for the fitted value. They can also form the basis of learning how the model has got it wrong, with when the observation is misclassified or has a large residual.

In the penguins example, we showed how the misclassification of a penguin arose due to it having an unusually small flipper size compared to others of its species. This was verified by making a follow-up plot of the data. The chocolates example shows how a dark chocolate was misclassified primarily due to its attribution to Fiber, and a milk chocolate was misclassified as dark due to its lowish Sodium value. In the FIFA example, we show how low Messi's salary would be if it depended on defensive skills. In the Ames housing data, an inaccurate prediction for a house was likely due to the Lot Area is not being effectively used by the random forest model.

This analysis is manually intensive and thus only feasible for investigating a few observations. The recommended approach is to investigate an observation where the model has not predicted accurately and compare it with an observation with similar predictor values where model fitted well. The radial tour launches from the attribution projection to enable exploration of the sensitivity of the prediction to any variable. It can be helpful to make additional plots of the variables and responses to cross-check interpretations made from the cheem viewer. This methodology provides an additional tool in the box for studying model fitting.

An implementation is provided in the open-source **R** package **cheem**, available on CRAN at `https://CRAN.R-project.org/package=cheem`. Example data sets are provided, and you can upload your data after model fitting and computing the local explanations. In theory, this approach would work with any black box model, but the implementation currently only calculates tree SHAP for tree-based models supported by **treeshap** (tree based models from **gbm**, **lightgbm**, **randomForest**, **ranger**, or **xgboost** Greenwell et al. 2020; Shi et al. 2022; Liaw and Wiener 2002; Wright and Ziegler 2017; Chen et al. 2021, respectively). Tree SHAP was selected because of its computational efficiency. The SHAP and oscillation explanations could be added with the use of the `DALEX::explain()` and would be an excellent direction to extend the work (Biecek 2018; Biecek and Burzykowski 2021).

Acknowledgments

Kim Marriott provided advice on many aspects of this work, especially on the explanations in the applications section. We would like to thank Professor Przemyslaw Biecek for his input early in the project and to the broader MI 2 lab group for the DALEX ecosystem of **R** and Python packages. This research was supported by the Australian Government Research Training Program (RTP) scholarships. Thanks to Jieyang Chong for helping proofread this article. The namesake, Cheem, refers to a fictional race of humanoid trees from Doctor Who lore. **DALEX** pulls on from that universe, and we initially apply tree SHAP explanations specific to tree-based models.

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)." *IEEE access* 6: 52138–52160.

Anderson, James A. 1995. *An introduction to neural networks*. MIT press.

Asimov, Daniel. 1985. "The Grand Tour: a Tool for Viewing Multidimensional Data." *SIAM journal on scientific and statistical computing* 6 (1): 128–143.

Barredo Arrieta, Alejandro, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58: 82–115. https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Biecek, Przemyslaw. 2018. "DALEX: explainers for complex predictive models in R." *The Journal of Machine Learning Research* 19 (1): 3245–3249.

Biecek, Przemyslaw. 2020. *ceterisParibus: Ceteris Paribus Profiles.* https://CRAN.R-project.org/package=ceterisParibus.

Biecek, Przemyslaw, and Tomasz Burzykowski. 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models.* CRC Press.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. "A training algorithm for optimal margin classifiers." In *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Breiman, Leo. 2001a. "Random forests." *Machine learning* 45 (1): 5–32.

Breiman, Leo. 2001b. "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*, New York, NY, USA, 63–67. Elsevier North-Holland, Inc. Accessed 2019-01-09. http://dl.acm.org/citation.cfm?id=26036.26046.

Caragea, Doina, Dianne Cook, Hadley Wickham, and Vasant Honavar. 2008. *Visual Methods for Examining SVM Classifiers*, 136–153. Berlin, Heidelberg: Springer Berlin Heidelberg.

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.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. 2021. "xgboost: Extreme Gradient Boosting." https://CRAN.R-project.org/package=xgboost.

Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–480. Accessed 2018-04-15. http://www.jstor.org/stable/1390747.

Cook, Dianne, Andreas Buja, Eun-Kyung Lee, and Hadley Wickham. 2008. "Grand Tours, Projection Pursuit Guided Tours, and Manual Controls." In *Handbook of Data Visualization*, 295–314. Berlin, Heidelberg: Springer Berlin Heidelberg. Accessed 2018-08-21. http://link.springer.com/10.1007/978-3-540-33037-0_13.

Cook, Dianne, Deborah F. Swayne, and A. Buja. 2007. *Interactive and Dynamic Graphics for Data Analysis: With R and GGobi.* Springer Science & Business Media.

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

De Cock, Dean. 2011. "Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project." *Journal of Statistics Education* 19 (3). Publisher: Taylor & Francis.

Gabriel, Karl Ruben. 1971. "The biplot graphic display of matrices with application to principal component analysis." *Biometrika* 58 (3): 453–467.

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. 2014. "Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis)." *PloS one* 9 (3): e90081.

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.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, and G. B. M. Developers. 2020.

"gbm: Generalized Boosted Regression Models." https://CRAN.R-project.org/package=gbm.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B. Gorman. 2020. "palmerpenguins: Palmer Archipelago (Antarctica) penguin data." https://allisonhorst.github.io/palmerpenguins/.

Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set*. CRC Press.

Kominsarczyk, Konrad, Pawel Kozminski, Szymon Maksymiuk, and Przemyslaw Biecek. 2021. "treeshap." Oct. Accessed 2021-10-29. https://github.com/ModelOriented/treeshap.

Lee, Stuart, Dianne Cook, Natalia da Silva, Ursula Laa, Nicholas Spyrison, Earo Wang, and H. Sherry Zhang. 2021. "The state-of-the-art on tours for dynamic visualization of high-dimensional data." *WIREs Computational Statistics* n/a (n/a): e1573. Accessed 2021-12-10. https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.1573.

Lee, Yoon Dong, Dianne Cook, Ji-won Park, and Eun-Kyung Lee. 2013. "PPtree: Projection pursuit classification tree." *Electronic Journal of Statistics* 7: 1369–1386.

Leone, Stefano. 2020. "FIFA 20 complete player dataset." Oct. Accessed 2021-08-23. 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 M., and Su-In Lee. 2017. "A unified approach to interpreting model predictions." In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777.

Molnar, Christoph. 2020. *Interpretable machine learning*. Lulu. com. https://christophm.github.io/interpretable-ml-book/.

Ocagne, Maurice d'. 1885. *Coordonnées parallèles et axiales. Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles, par Maurice d'Ocagne, ...* Paris: Gauthier-Villars.

prevek18. 2018. "Ames Housing Dataset." Sep. Accessed 2022-01-27. https://kaggle.com/prevek18/ames-housing-dataset.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, Aug., 1135–1144. Association for Computing Machinery. Accessed 2022-01-07. https://doi.org/10.1145/2939672.2939778.

Roberts, Jonathan C. 2007. "State of the art: Coordinated & multiple views in exploratory visualization." In *Fifth international conference on coordinated and multiple views in exploratory visualization (CMV 2007)*, 61–71. IEEE.

Shapley, Lloyd S. 1953. *A value for n-person games*. Princeton University Press.

Shi, Yu, Guolin Ke, Damien Soukhavong, James Lamb, Qi Meng, Thomas Finley, Taifeng Wang, et al. 2022. "lightgbm: Light Gradient Boosting Machine." https://CRAN.R-project.org/package=lightgbm.

Shmueli, Galit. 2010. "To explain or to predict?" *Statistical science* 25 (3): 289–310.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. "Learning important features through propagating activation differences." In *International Conference on Machine Learning*, 3145–3153. PMLR.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. "Not just a black box: Learning important features through propagating activation differences." *arXiv preprint arXiv:1605.01713* .

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

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep inside convolutional networks: Visualising image classification models and saliency maps." In *In Workshop at*

International Conference on Learning Representations, Citeseer.

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. Accessed 2020-10-16. https://journal.r-project.org/archive/2020/RJ-2020-027/index.html.

Stahl, Bernd Carsten. 2021. "Ethical Issues of AI." *Artificial Intelligence for a Better Future* 35–53.

Strumbelj, Erik, and Igor Kononenko. 2010. "An efficient explanation of individual classifications using game theory." *The Journal of Machine Learning Research* 11: 1–18.

Unwin, Antony, and Pedro Valero-Mora. 2018. "Ensemble Graphics." *Journal of Computational and Graphical Statistics* 27 (1): 157–165. Accessed 2021-01-22. https://doi.org/10.1080/10618600.2017.1383264.

Vanni, Laurent, Mélanie Ducoffe, Carlos Aguilar, Frédéric Precioso, and Damon Mayaffre. 2018. "Textual Deconvolution Saliency (TDS): a deep tool box for linguistic analysis." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 548–557.

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): 203–225. Accessed 2018-03-16. http://doi.wiley.com/10.1002/sam.11271.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1–17.