Exploring Local Explanations of Nonlinear Models Using Animated Linear Projections

Abstract

The increased predictive power comes at the cost of interpretability, which has led to the emergence of eXplainable AI (XAI). XAI attempts to shed light on how models use predictors to arrive at a prediction with a point estimate of the linear feature importance in the vicinity of each instance. These can be considered linear projections and can be further explored interactively to understand better the interaction between features used to make predictions across the predictive model surface. Here we describe interactive linear interpolation used for exploration at any instance and illustrate with examples with categorical (penguin species, chocolate types) and quantitative (football salaries, house prices) response features. The methods are implemented in the **R** package **cheem**, available on CRAN.

1 Introduction

There are different reasons and emphases to fit a model. Breiman (2001), reiterated by Shmueli (2010), taxonomize modeling based on its purpose; *explanatory* modeling is done for some inferential purpose, while *predictive* modeling focuses more on the predictions of out-of-sample instances. The intended use has important implications for model selection and development. In explanatory modeling, interpretability is vital for drawing inferential conclusions. While predictive modeling may opt for more accurate nonlinear models. The use of black-box models is becoming increasingly common, but not without their share of controversy (O'Neil 2016; Kodiyan 2019). However, the loss of interpretation presents a challenge.

Interpretability is vital for exploring and protecting against potential biases (e.g., sex (Dastin 2018; Duffy 2019), race (Larson et al. 2016), and age (Díaz et al. 2018)) in any model. For instance, models regularly pick up on biases in the training data that have observed influence on the response (output) feature, which is then built into the model. Feature-level (variable-level) interpretability of models is essential in evaluating such biases. It is also generally important for many problems, where it is not enough to predict accurately. Still, one must be able to explain which predictors are most responsible for generating a response value.

Another concern is data drift, a shift in support or domain of the explanatory features (variable or predictors). Nonlinear models are typically more sensitive and do not extrapolate well outside the training data domain. Better interpretability of the model means more transparency where models' predictions may be plausible or completely unreliable.

Explainable Artificial Intelligence (XAI) is an emerging field of research that tries to increase the interpretability of black-box models. A common approach is to use *local explanations*, which attempt to approximate linear feature importance at the location of each instance (observation), or the predictions at a specific point in the data domain. Because these are point-specific, it is challenging to visualize them to comprehensively understand a model.

In multivariate data visualization, a *tour* (Asimov 1985; Buja and Asimov 1986; S. Lee et al. 2021) is a sequence of linear projections of data onto a lower-dimensional space. Tours are viewed as an animation over minor changes to the projection basis. Structure in a projection can then be explored visually to see which features contribute to the formation of that structure. The intuition is similar to watching the shadow of a hidden 3D object change as the object is rotated; watching the shape of the shadow change conveys information about the structure and features of the object.

There are various types of tours distinguished by the generation of projection bases. In a *manual* tour (Cook and Buja 1997; Spyrison and Cook 2020), the path is defined by changing the contribution of a selected

feature. Applying tours to models has been done in a couple of contexts. Specifically for exploring various statistical model fits and classification boundaries (Wickham, Cook, and Hofmann 2015), and using tree- and forest-based approaches as a projection pursuit index to generate a tour basis paths (Y. D. Lee et al. 2013; da Silva, Cook, and Lee 2021).

We use the radial manual tour to scrutinize a local explanation in our proposed approach. Additional interactivity allows the user to identify an instance of interest, then explore its local explanation by changing feature contribution with the radial tour. The methods are implemented in R package **cheem**. Example datasets are provided to illustrate usage for classification and regression tasks.

Using a radial tour can be compared with counterfactual, what-if analysis, such as *ceteris paribus* profiles (Biecek 2020). Ceteris paribus, is Latin for "other things held constant" or "all else unchanged". These profiles show how an instance's prediction would change from a marginal change in one explanatory feature given that other features are held constant. It ignores correlations of the features and imagines a case that was not observed. In contrast, our approach is a geometric explanation of the factual; it varies contributions of the features by rotating the basis, a reorientation of the data object. A constraint in our approach is that the basis must remain orthonormal. When the contribution of one feature decreases, the contributions of others necessarily increase such that there is a complete component in that direction. This also ensures that what is seen is strictly a low-dimensional projection from high-dimensions and is thus an interpretable visualization.

The remainder of this paper is organized as follows. The following Section, 2, covers the background of the local explanation and the traditional visuals produced. Section 3 explains the animations of continuous linear projections. Section 4 discusses the visual layout in the interactive interface, how they facilitate analysis, data preprocessing, and package infrastructure. Then Section 5 illustrates the application to supervised learning with categorical and quantitative response features. We conclude with Section 6 of the insights gained and directions that might be explored in the future.

2 Local explanations

Consider a highly nonlinear model. It can be hard to determine whether small changes in a feature's value will make a class prediction change group or identify which features contribute to an extreme residual. Local explanations shed light on these situations by approximating linear feature importance in the vicinity of a single instance.

A comprehensive summary of the taxonomy and literature of explanation techniques is provided in Figure 6 of 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 explanations, of which LIME, (Ribeiro, Singh, and Guestrin 2016) SHAP, (Lundberg and Lee 2017), and their variants are popular.

These instance-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 indicating important pixels (Simonyan, Vedaldi, and Zisserman 2014). For instance, pixels corresponding to snow may be highlighted when distinguishing if a picture contains a wolf or husky. In text analysis, word-level contextual sentiment analysis can be used to highlight the sentiment and magnitude of influential words (Vanni et al. 2018). In the case of numeric regression, they are used to explain feature additive contributions from the model intercept to the instance's prediction (Ribeiro, Singh, and Guestrin 2016).

SHaply Additive exPlanations (SHAP) quantifies the feature contributions of one instance by examining the effect of other features on the predictions. The explanations of SHAP almost all refer to Shapley (1953)'s method to evaluate an individual's contribution to cooperative games by assessing the performance of this player in the presence or absence of other players. Strumbelj and Kononenko (2010) introduced the use of SHAP for local explanations in ML models. The attribution of feature importance depends on the sequence of the features already included. The SHAP values are the mean contributions over different feature sequences. The approach is related to partial dependence plots (Molnar 2020), used to explain the effect of a feature by predicting the response for a range of values on this feature after fixing the value of all other features to their

mean. Partial dependence plots are a global approximation of the feature importance, while SHAP is specific to one instance. It could also be considered similar to examining the coefficients from all subsets regression, as described in Wickham, Cook, and Hofmann (2015), which helps to understand the relative importance of each feature in the context of all other candidate features.

For our 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 features, tree SHAP calculates instance-level feature importance by exploring the structure of the decision trees. Tree SHAP is only compatible with tree-based models; we illustrate random forests. The following section will use normalized SHAP values as a projection basis (call this the *attribution projection*) will have coefficients varied to further scrutinize the feature contributions.

Following the use case *Explanatory Model Analysis* (Biecek and Burzykowski 2021), we use FIFA data to illustrate the use of SHAP. Consider soccer data from the FIFA 2020 season (Leone 2020). There are 5000 instances of 9 skill measures (after aggregating highly correlated features). A random forest model is fit regressing wages [2020 Euros], from the skill measures. We then extract the SHAP values of a star offensive player (L. Messi) and defensive player (V. van Dijk). The results are displayed in Figure 1. We expect to see a difference in the attribution of the feature importance across the two positions of the players, which would be interpreted as how the player's salary depends on this combination of skill sets. Plot (b) is a modified breakdown plot (Gosiewska and Biecek 2019) where the order of features is fixed, so the two instances can be more easily compared.

In summary, these plots highlight how local explanations bring interpretability to a model, at least in the vicinity of their instances. In this instance, two players with different positions receive different profiles of feature importance to explain the prediction of their wages.

3 Tours and the radial tour

A *tour* enables 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 of the features of the tour is the object permanence of the data points; one can track the relative change of instances in time and gain information about the relationships between points across multiple features. There are various types of tours that are distinguished by how the paths are generated (S. Lee et al. 2021; Cook et al. 2008).

The manual tour (Cook and Buja 1997) defines its path by changing a selected feature's contribution to a basis to allow the feature to contribute more or less to the projection. The requirement constrains the contribution of all other features 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 feature 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 the new work. In a radial tour, the selected feature 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 feature.

4 The cheem viewer

To explore the local explanations, an ensemble of plots (Unwin and Valero-Mora 2018) is provided, called the *cheem viewer*. 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. In addition, there are numerous user inputs, including feature selection for the radial tour and instance selection for making comparisons. There are different plots used for categorical and quantitative response. Figures 2 and 3 are

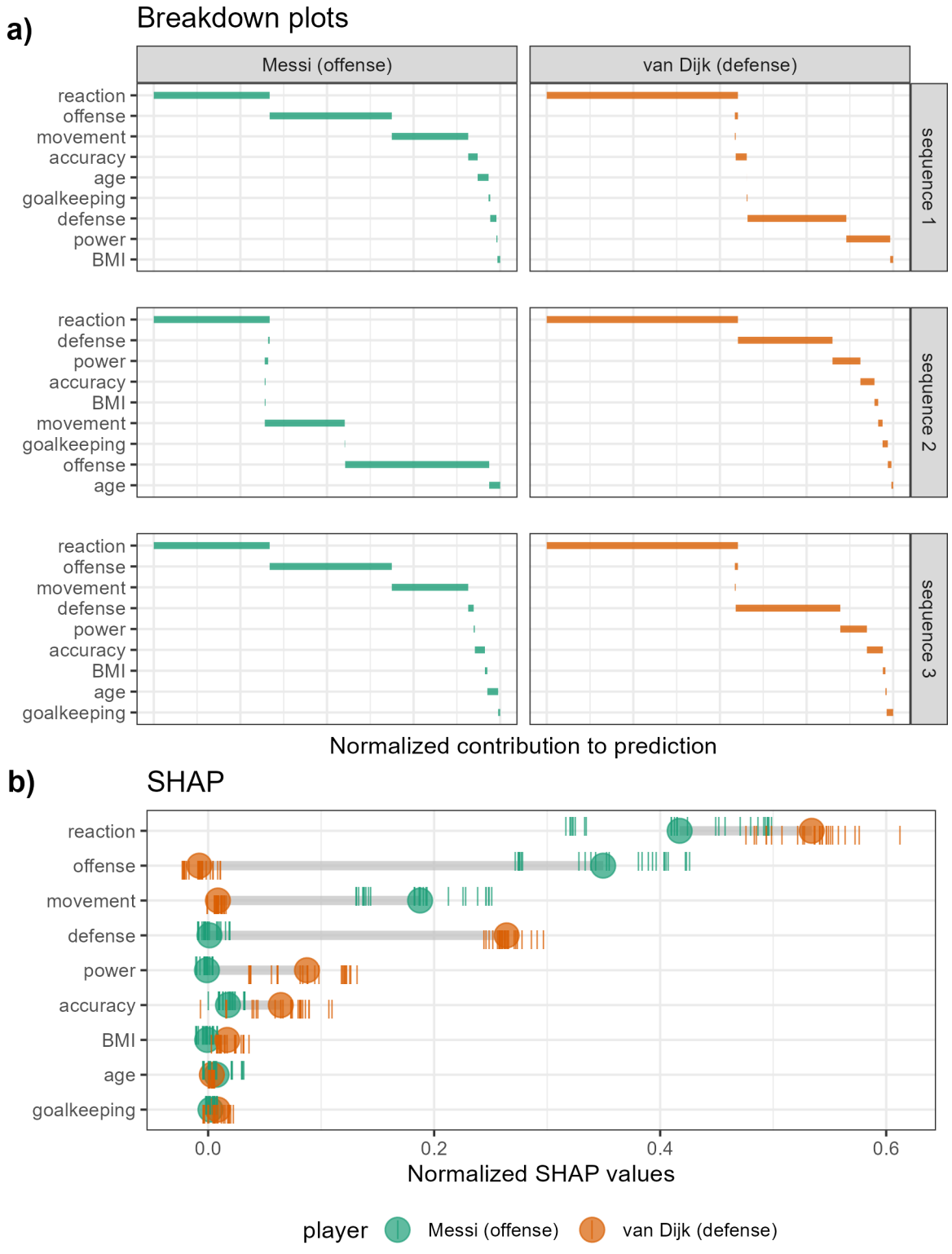

Figure 1: Illustration SHAP values for a random forest model for salaries of FIFA 2020 players based on nine predictors corresponding to different skills. A star offensive and defensive player are compared, L. Messi and V. van Dijk, respectively. Panel (a) shows breakdown plots of three sequences of the features, order and magnitude change. Panel (b) shows the distribution of attribution for each feature across 25 sequences of predictors, with the mean displayed as a dot, for each 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.

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 the context of all instances and facilitates the exploration of the separability of the data- and attribution-spaces. Both of these spaces are of dimension $n \times p$, where n is the number of instances and p is the number of predictors. The attribution space corresponds to the local explanations for each instance, which will have p values for each instance.

A visualization is provided by 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 useful visual summaries. In addition, a plot of the observed against predicted response values is also provided (Figures 2b, 2 XXX), to help identify instances poorly predicted by the model. For classification tasks, misclassified instances are circled in red if applicable. 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 2a,c).

While the comparison of these spaces is interesting, a main purpose of the global view is to enable the selection of instances to explore the local explanations. The projection attribution of the primary instance (PI) is examined and typically viewed with an optional comparison instance (CI). These instances are highlighted as asterisk and $\times$ respectively.

4.2 Radial tour

The local explanations for all observations are normalized (squared sum of values adds to 1) and thus the relative importance of features can be compared across all instances. These are depicted as a vertical parallel coordinate plot, where each line connects one instance's feature attribution (Figures 2e and 3e). The attribution projections of the PI and CI are shown as dashed and dotted lines, respectively. From this plot, we would read the range of importances across all instances can be interpreted. For classification, one would look at differences between groups on any feature, for example, Figure 2e suggests that `b_l` is important for distinguishing the green class from the other two. For regression, one might generally observe which features have low values for all instances (not important), for example, `BMI` and `pwr` in Figure 3e, and which have a range of high and low values (e.g. `off`, `def`) suggesting important for some instances and not important for other instances.

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 is an approximation of the feature importance, for the prediction of this instance. This combination of features best explains the difference between the mean response and an instance'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 2f), and as scatterplots for a quantitative response (Figure 3f). The PI and CI are indicated by vertical dashed and dotted lines, respectively. The user uses the radial tour to vary contribution of the selected feature between 0-1. Doing so tests the sensitivity of structure (class separation or strength of relationship) to the feature's contribution. For classification, if the separation between classes diminishes when the feature contribution is reduced then this suggests that the feature is important for the class separation. For regression, if the relationship scatterplot weakens when the feature contribution is reduced then this suggests that the feature is important for accurately predicting the response.

4.3 Classification task

Selecting a misclassified instance as PI and a correctly classified point nearby in data space as CI, makes it easier to examine the features most responsible for the error. The global view (Figure 2c) displays the model confusion matrix. The radial tour is 1D, plotted as a density where color indicates class. A scroll bar here

enables the user to vary the contribution of each feature to explore the sensitivity the separation to that feature.

4.4 Regression task

Selecting an inaccurately predicted instance as PI and an accurately predicted instance, with similar feature values, as CI is a useful way to understand how the model is failing, or not. The global view (Figure 3a) 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 Mahalanobis distance) or correlation, to help with understand where the model fits better or worse.

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 features, and a good fit would be indicated when there is a strong relationship seen with the observed values. This can be viewed as a local linear approximation if the fitted model is nonlinear. As the contribution of a feature is varied, if the value of the PI doesn't change much it would indicate that the prediction for this instance is NOT sensitive to that feature. Conversely, if the predicted value varies substantially, the prediction is very sensitive to that feature, suggesting that the feature is very important for the PI's prediction.

4.5 Interactive features

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. The application also has more exploratory interactions to help link points across displays and reveal structure found in different spaces.

A tooltip displays instance 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. The information corresponding to the selected points is populated on a dynamic table. These interactions aid exploration of the spaces and, finally, identification of a primary and comparison instance.

4.6 Preprocessing

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

- **Data:** predictors and response are unscaled complete numerical matrix. Most models and local explanations are scale-invariant.
- **Model and explanation:** any model can be used with this method. Currently, we apply random forest models via the package **randomForest** (Liaw and Wiener 2002), compatibility tree SHAP. We use modest hyperparameters, namely: 125 trees, number features randomly sampled 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. Tree SHAP is calculated for *each* instance using the package **treeshap** (Kominsarczyk et al. 2021). This implementation aggregates over exhaustively over all trees' attribution, and we opt not to fit interactions of features.
- **Cheem view:** after the model and full explanation space are calculated, we scale each feature by standard deviations away from the mean to achieve common support for visuals. Statistics for mapping to color are calculated on the scaled spaces. Interactive tabular display reports the original values.

The time to preprocess the data will vary significantly with the model and local explanation. For reference, the FIFA data, 5000 instances of nine explanatory features, took 2.5 seconds to fit a random forest model of modest hyperparameters. Extracting the tree SHAP values of each instance took 270 seconds combined.

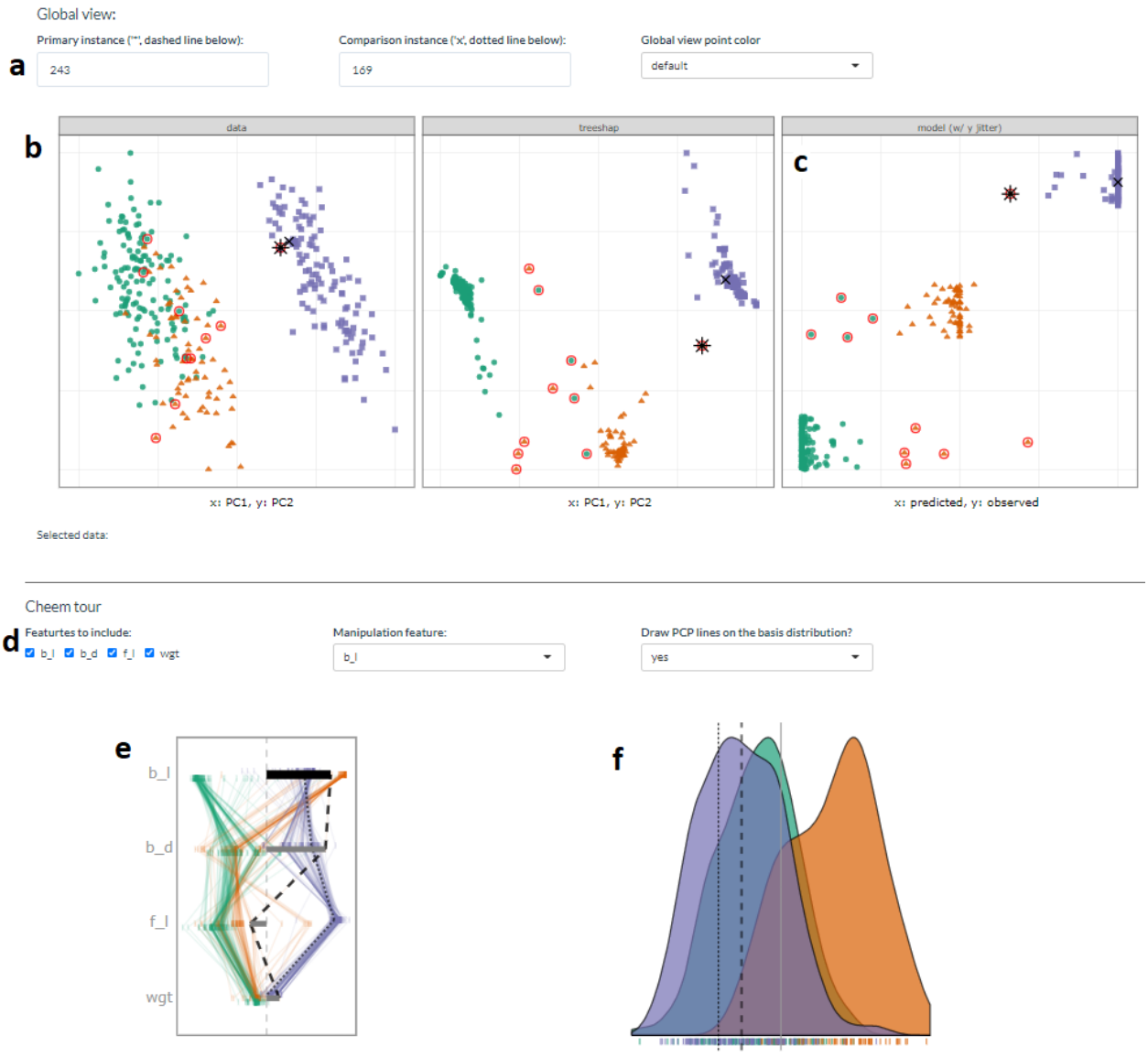

Figure 2: Overview of the cheem viewer for classification tasks. Global view inputs, (a), set the PI, CI, and color statistic. Global view, (b) PC1 by PC2 approximations of the data space and attribution space. (c) prediction by observed y (visual of the confusion matrix for classification tasks). Points are colored by predicted class, and red circles indicate misclassified instances. Radial tour inputs (d) select features to include and which feature is changed in the tour. (e) shows parallel coordinate display of the distribution of the feature attributions while bars dipict contribution for the current basis. The black bar is the variable being changed in the radial tour. Panel (f) is the resulting projection of the data indicated as density in the classification case.

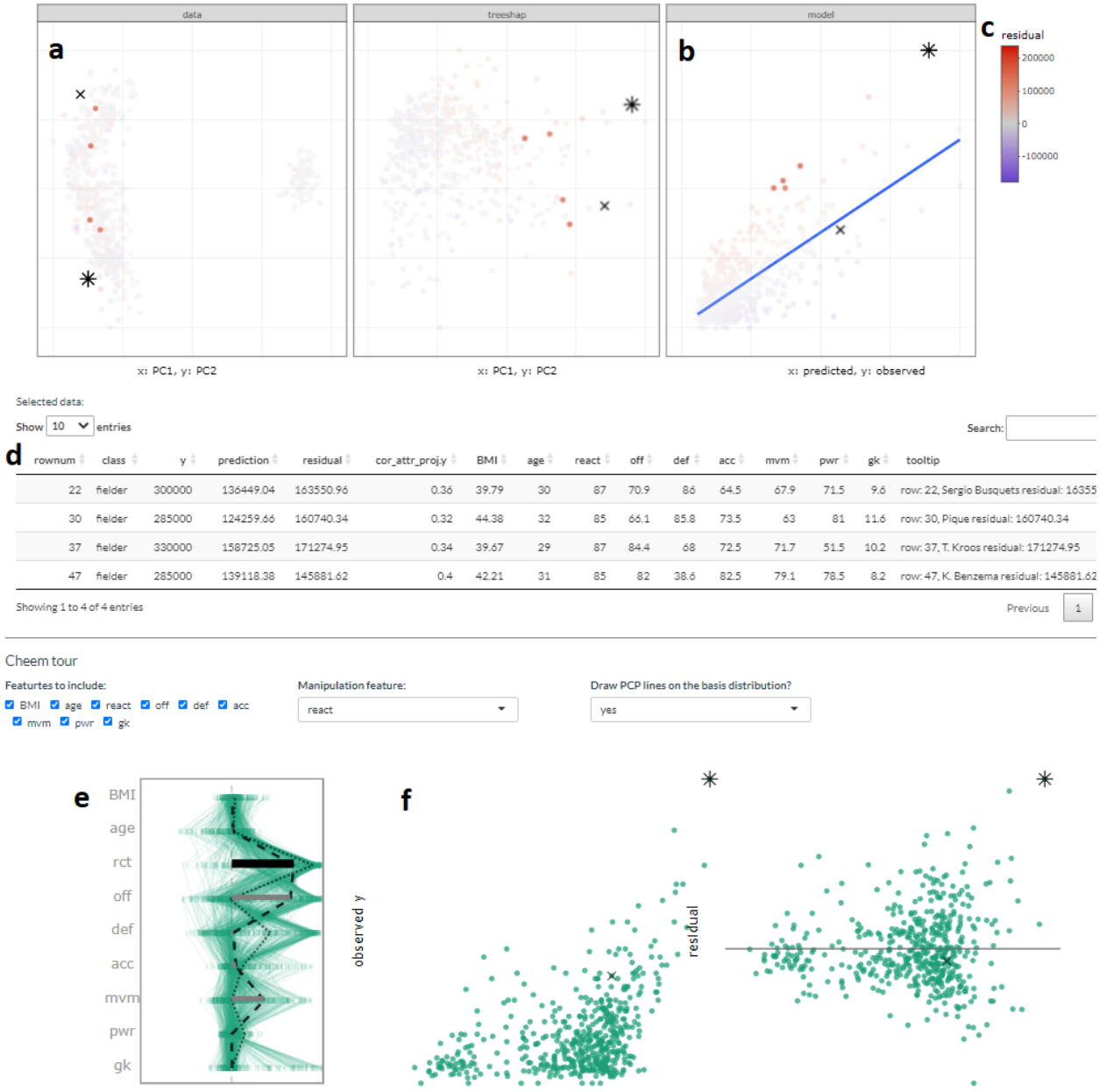

Figure 3: Overview of the cheem viewer for regression task highlighting the differences from the classification task and interactive features. Panel (a) PCA of the data and attributions spaces, (b), residual plot, predictions by observed values. Four points areselected points and highlighted in the PC spaces and tabularly displayed. Coloring on a statistic (c) highlights structure organized in the attribution space. Interactive tabular display (d) populates when instances are selected. Contribution of the 1D basis effecting the horizontal position (e) parallel coorindate display of the feature 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, (left and right).

PCA and statistics of the features and attributions took 2.8 seconds. These runtimes were from a non-parallelized R 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 a good candidate to start with. (Alternatively, the package **fastshap** (Greenwell 2020) claims extremely low runtimes, attributed to fewer calls to the prediction function, partial implementation in C++, and efficient use of logical subsetting.)

4.7 Package infrastructure

The above-described method and application are implemented as an open-source **R** package, **cheem** available on CRAN. 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 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 huge boon to understanding local explanations and how to apply them.

4.8 Installation and getting started

The package can be installed from GitHub using the following **R** code:

```
# Install remotes if absent
if(require("remotes") == FALSE) install.packages("remotes")
remotes::install.packages("cheem", dependencies = TRUE)
library("cheem")
run_app()
```

To process your own data, you will need to use the `treeshap` package, which can be installed from GitHub:

```
remotes::install_github('ModelOriented/treeshap')
```

Follow the examples provided with the package to compute the local explainers, (see `?cheem_ls`). The application expects the return of call from `cheem_ls()` which is then saved to an .rds file with `saveRDS()`. Alternatively, the cheem viewer shiny app can be directly accessed at https://ebsmonash.shinyapps.io/cheem_initial/.

5 Case studies

To illustrate the use of the cheem method, we apply it to modern datasets, 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 was publicly available to substitute for the overly-used iris data and is quite similar in form. After removing incomplete instances, there are 333 instances and we will use the four physical measurements, `bill_length_mm` (`b_l`), `bill_depth_mm` (`b_d`), `flipper_length_mm` (`f_l`), `body_mass_g` (`wgt`), for this illustration. A random forest model was fit with species as the response feature.

Figure 4 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).

There is more separation visible in the attribution space than the data space, as would be expected. The predicted vs observed plot reveals a handful of misclassified instances. We will explore why a Gentoo has

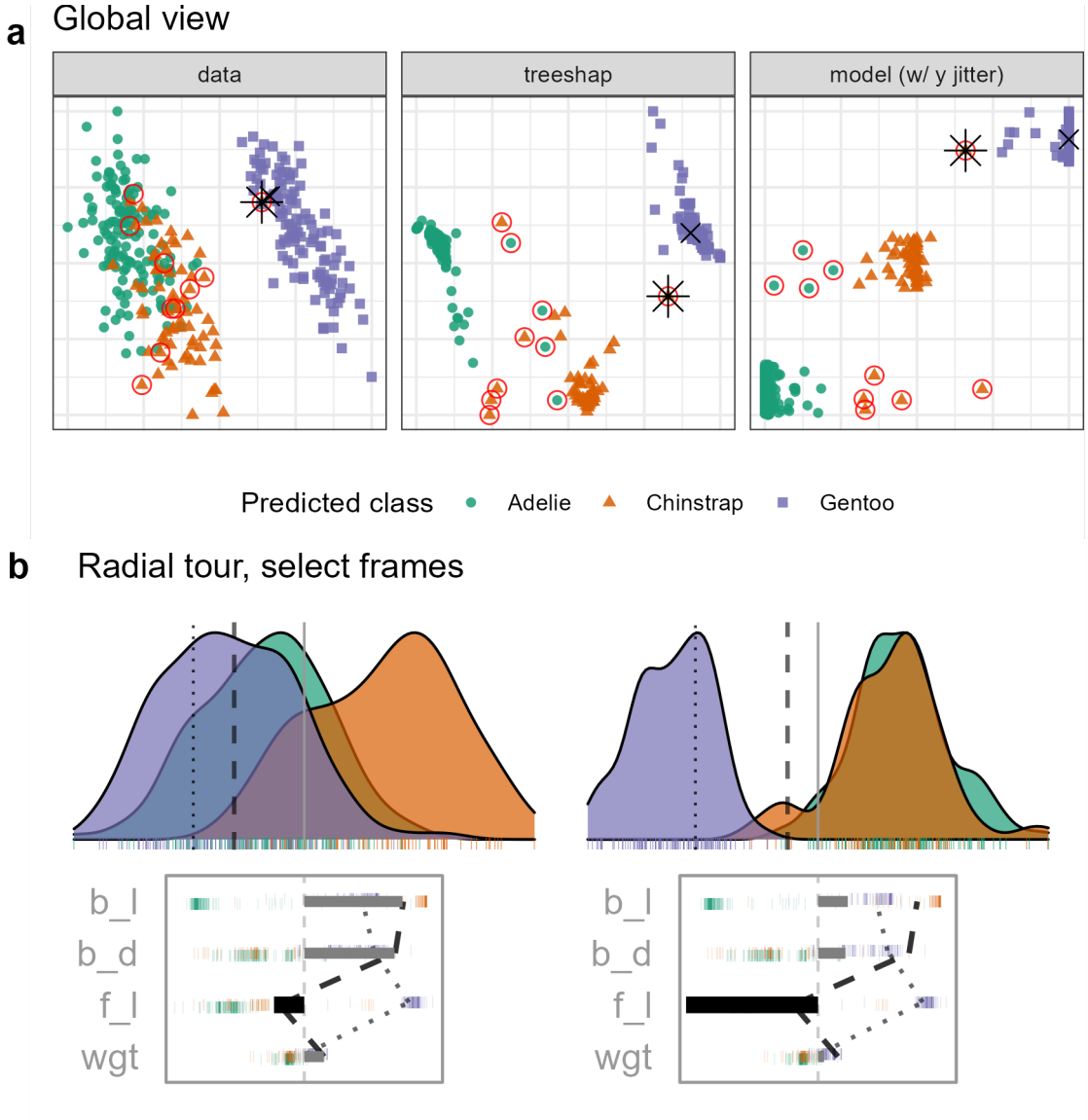

Figure 4: Examining the SHAP values for a random forest model classifying Palmer penguin species. The PI is an Chinstrap (orange) penguin that is misclassified as a Gentoo (purple), marked as an asterisk in (a), and the dashed vertical line in (b). The radial view shows varying the contribution of `f_l` from the initial attribution projection (b, left), which produces a linear combination where the PI is more probably a Chinstrap than a Gentoo (b, right). (The animation of the radial tour is at vimeo.com/666431172.)

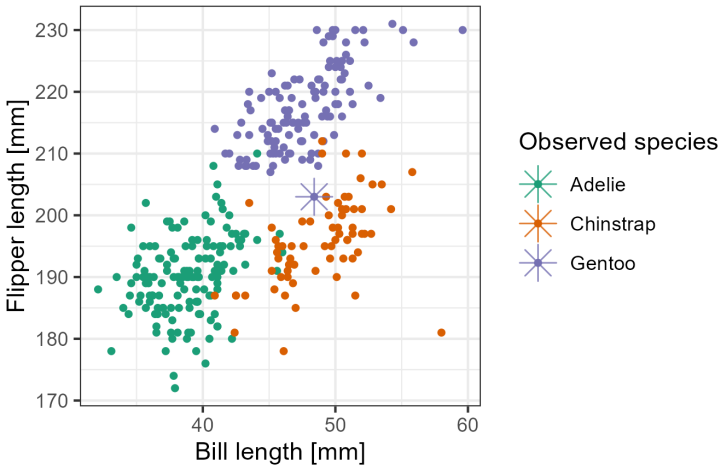

Figure 5: Checking what is learned from the cheem viewer. This is a plot of flipper length (`f_l`) and bill length (`b_l`), where an asterisk highlights the PI. A Gentoo (purple) misclassified as a Chinstrap (orange). The PI has an unusually small `f_l` length which is why it is confused with a Chinstrap.

been wrongly labeled as a Chinstrap for this illustration. The PI is a misclassified point (represented by the asterisk in the global view and as 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 instance (b, left). The important features identified by SHAP in the (wrong) prediction for this instance are mostly `b_l` and `b_d` with small contributions of `f_l` and `wgt`. This projection is a view where the Gentoo (purple) looks much more likely for this instance than Chinstrap. That is, this combination of features is not particularly useful because the PI looks very much like other Gentoo penguins. To explore this, we use the radial tour to vary the contribution of flipper length (`f_l`). (In our exploration, this was the third feature explored. It is typically useful to explore the features with larger contributions, here `b_l` and `b_d`, but when doing this, nothing was revealed about how the PI differed from other Gentoos). On varying `f_l` as it contributes more to the projection (b, right), we see that more, and more, this penguin looks like a Chinstrap. This suggests that `f_l` should be considered an important feature for explaining the (wrong) prediction.

Figure 5 confirms that flipper length (`f_l`) is important for the confusion of the PI as a Chinstrap. Here, flipper length and body length are plotted, and we can see that the PI is closer to the Chinstrap group in these two features, mostly because it has an unusually low value of flipper length relative to other Gentoos. From this view it makes sense that its a hard instance to account for as decision trees can only partition only vertical and horizontal lines.

5.2 Chocolates, milk/dark chocolate classification

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

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 that is nutritionally more like milk should not be considered a healthy alternative. It is interesting to explore which of the nutritional features contribute most to misclassification.

This type of exploration is shown in Figure 6, where a chocolate labeled dark but predicted to be milk is chosen as the PI (instance 22). It is compared with a CI that is a correctly classified dark chocolate (instance 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 instances. The misclassifications are clearer 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, instances 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 the interpretation of a point with high values in this plot. The dark chocolates (orange) are mostly on the left, and this is a reason why they are considered to be healthier: high fiber and low sugar. The density for milk chocolates is further to the right, indicating that they generally have low fiber and high sugar.

The instance of interest (dashed line) can be viewed against the comparison instance (dotted line). Now one needs to pay different attention to the parallel plot of the SHAP values, which are local to a particular instance, and the density plot, which is the same projection of all instances as specified by the SHAP values of the instance of interest.

We can quickly compare the feature contributions to the two different predictions from the parallel coordinate plot. The instance of interest differs with the comparison primarily on the Fiber feature, which suggests that this is the reason for the incorrect prediction.

From the density plot, which is the attribution projection corresponding to the instance of interest, both instances are more like dark chocolates. If we vary the contribution of Sugars, and completely remove Sugars from the projection, this is where the difference becomes apparent. When primarily Fiber is examined, instance 22 looks more like a milk chocolate.

It would also be interesting to explore the inverse case. This would describes which features lead to a milk chocolate being misclassified as dark and how those attributions differ from the previous misclassification. Chocolate 84 is just this case, and we compare it with a correctly predicted milk chocolate (instance 71). This exploration is shown in Figure 7.

The difference of position in the tree SHAP PCA with the previous case is quite large; this gives an approximate feel that the attribution should be quite different. Looking to the initial contribution we find this to be the case. Previously fiber was very important while it is absent from the attribution in this case. Conversely, calories from fat and total fat are highly attributed here, while unimportant in the preceding case.

Comparing the attribution with the CI (dotted line) discrepancies in Sodium and Fiber are identified. We opt to vary Sodium. Even in the initial projection, the instance looks slightly more like its observered milk than predicted dark chocolate. The misclassification seem 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. From a plot of the correlation matrix, nine higher level skill groupings were identified and aggregated. A random forest model if fit from these aggregations and regress player wages [2020 euros]. The model was fit from 5000 instances before being thinned to 500 players to mitigate occlusion and render time. Continuing from the exploration in section 2, 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

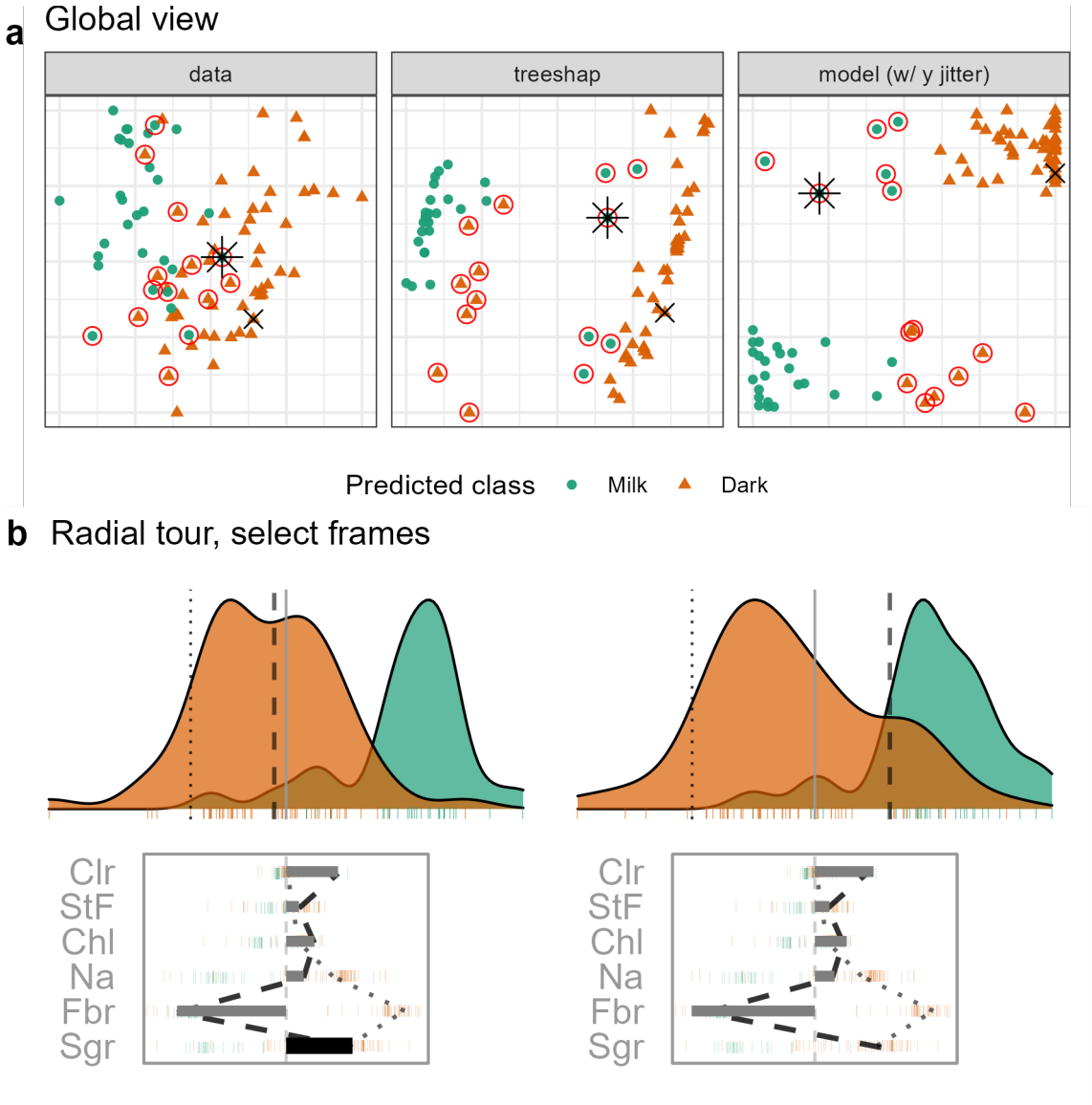

Figure 6: Examining the local interpretation for a PI which is dark (orange) chocolate incorrectly predicted to be milk (green). From the attribution projection this chocolate correctly dark more than milk, which suggests that the local explanation does not help to understand the prediction for this instance. So, we vary the contribution of Sugar – reducing it corresponds primarily with increasing Fiber. When Sugar is zero, Fiber is contributing strongly towards the left. In this particular 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 vimeo.com/666431143.)

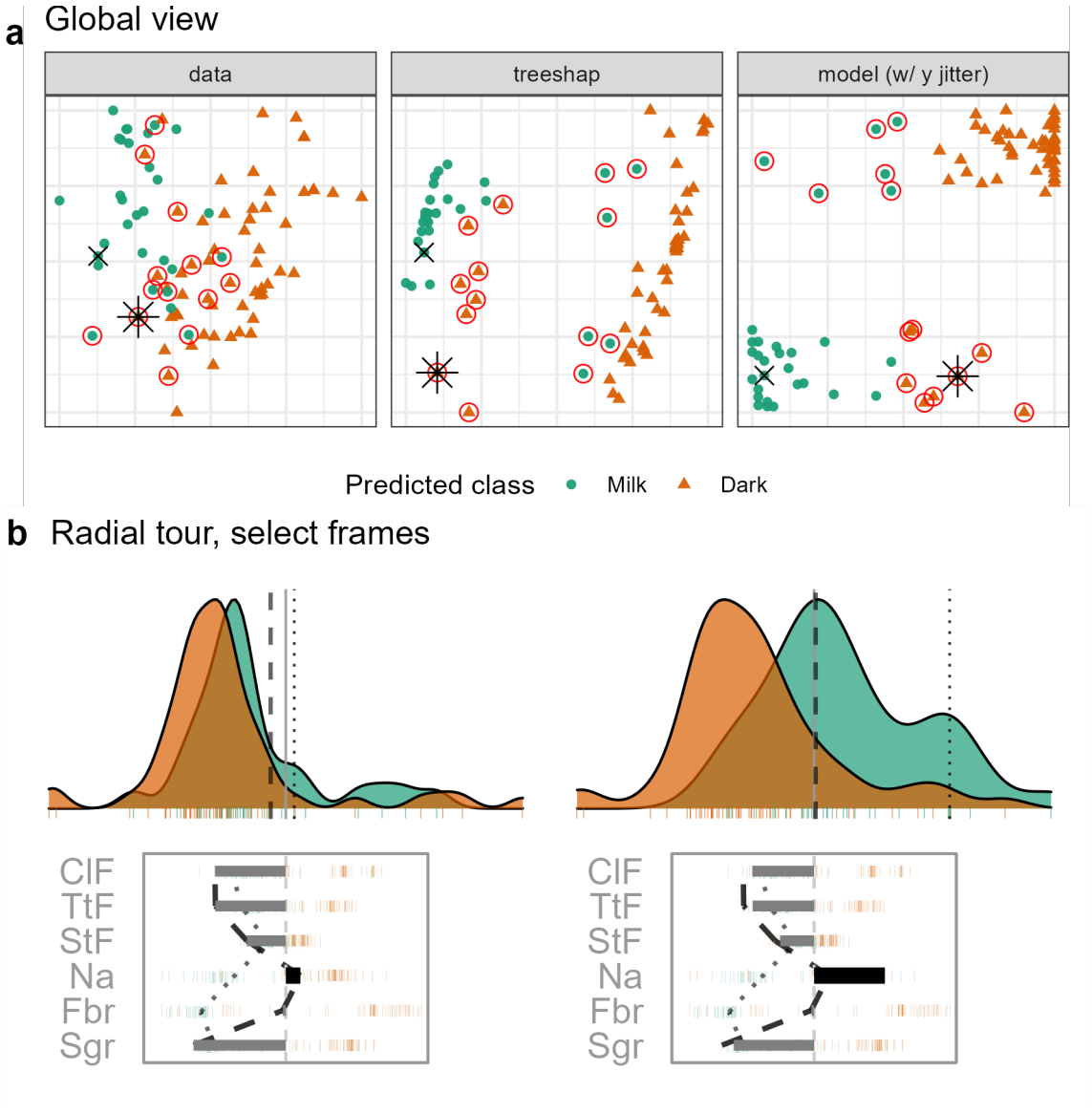

Figure 7: Examining the local interpretation for a PI which is milk (green) chocolate incorrectly predicted to be dark (orange). In the attribution projection the PI could be either milk or dark. Sodium and Fiber have the largest differences in attributed feature importance, both with low values relative to other milk chocolates. We suspect the lack of use of these variables is what has contributed to the mistake, so we vary the contribution of Sodium. 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 vimeo.com/666431143.)

profiles to better predict the wages from different field position of players. We compare a leading offensive fielder (L. Messi) with that of a top defensive fielder (V. van Dijk). The same instances were used in figure 1.

With figure 8, we will test the sopport of the local explanation. Offensive and reaction skills (`off` and `rct`) are both crucial to explaining a star offensive player. If either of them are rotated out the other will be rotated into frame maintaining a far-right position. However, when we vary the defensive skills, the other skills are rotated out of the frame.

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 right most 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 at these levels.

5.4 Ames housing 2018, sales price regression

Ames 2018, housing data was subset to North Ames (the neighborhood with the most house sales). The remaining are 338 house sales. A random forest model has regressed this price with the features Lot Area (`LtA`), Overall Quality (`Qlt`), Year the house was Build (`YrB`), Living Area (`LvA`), number of Bathrooms (`Bth`), number of Bedrooms (`Bdr`), total number of Rooms (`Rms`), Year the Garage was Build (`GYB`), and Garage Area (`GrA`). Using interaction from the global view, we select a house with an extreme negative residual and an accurate instance close to it in the data.

Figure 9 selects the house sale 74, a sizable under prediction that has a large contribution from lot area. The CI has a similar predicted price though the prediction was accurate and gives almost no attribution to lot size. The attribution projection is places instances with high living areas to the right. We control the contribution of this feature. As the contribution of lot area decreases, the predictive power decreases for the PI, while the CI remains stationary. This large of an importance is living area is relatively uncommon. Boosting tree models may be more resilient to such an under prediction as up-weighting this residual would force its inclusion in the final model.

6 Discussion

The need to maintain the interpretability of black-box models is evident. One aspect uses local explanations of the model in the vicinity of an instance. Local explanations approximate the linear feature importance to the model. Our contribution is to assess explanations by examining the support by varying the contributions with a radial tour. First, a global view visualizes approximations of the data space, explanation space, model predictions side-by-side, using dynamic interaction to compare and contrast and identify instances of interest. The normalized linear importance from the explanation of the PI becomes the feature of interest to further explore with the radial tour. The tours explore the feature sensitivity to the structure identified in the explanation.

We have illustrated this method on random forest models using the tree SHAP local explanation. While this analysis could be generally used with any compatible model-explanation pairing. We have created an open-source **R** package **cheem**, available on CRAN, to facilitate preprocessing and exploration with the described interactive application. Toy and real data are provided, or upload your data after preprocessing.

One limitation of the **cheem** packages is that it currently only calculates tree SHAP for tree-based models supported by **treeshap** (tree based models from **randomForest**, **ranger**, **gbm**, **xgboost**, **lightgbm**, or **catboost**). The SHAP and oscillation explainations could be added with use of the `DALEX::explain()` seems to be an excellent direction to extend (Biecek 2018; Biecek and Burzykowski 2021). Tree SHAP was selected because of its reduced computational complexity; processing runtime of the explanation will be a looming concern.

Attribution space does tend to be more separable than data space at least in the first couple principal components. Different statistics should be explored to convey a clearer understanding what structure the

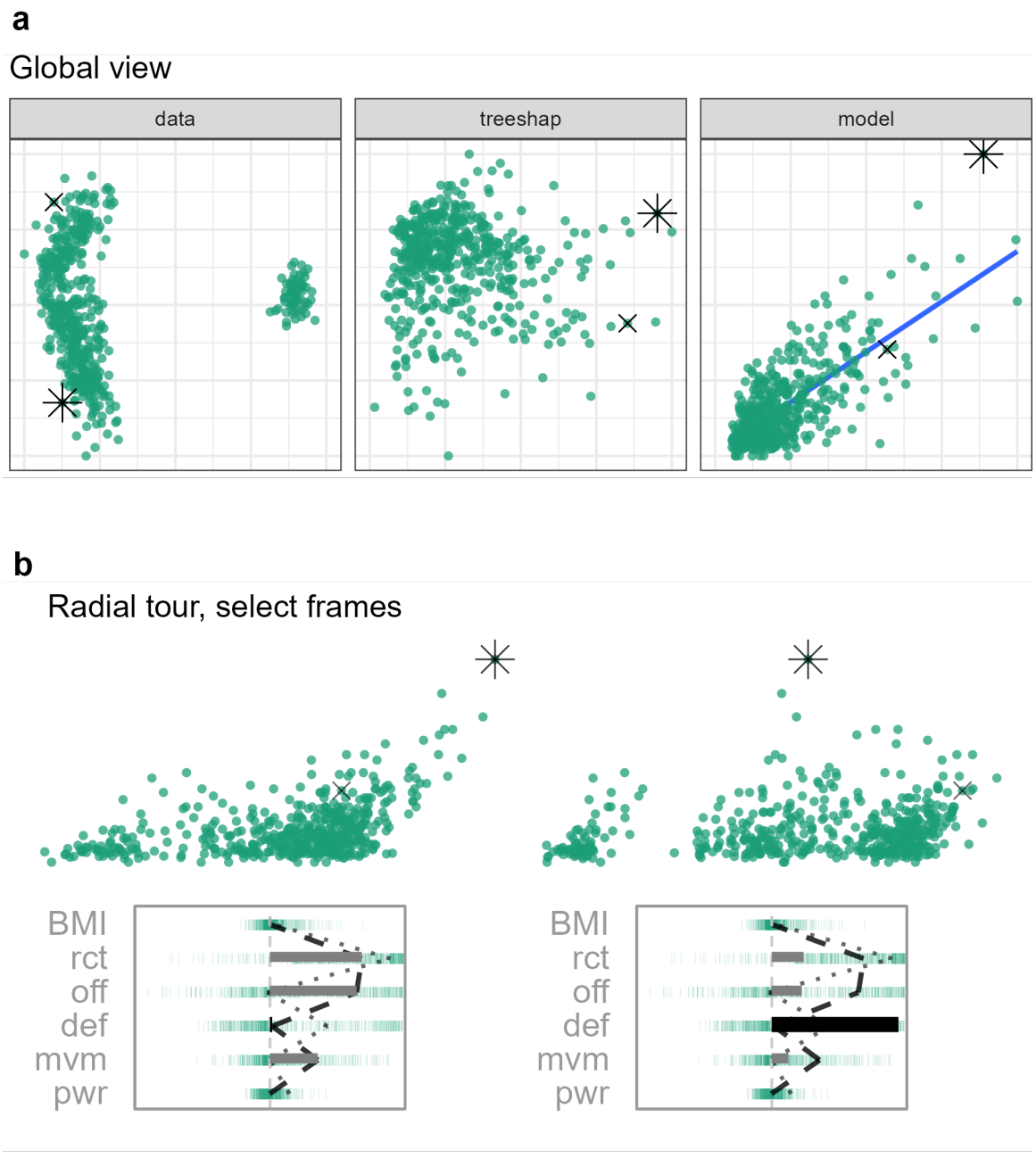

Figure 8: FIFA 2020 data, a random forest model, regresses wages [2020 Euros] from nine aggregated skill measurements. The PI is a star offensive player (L. Messi) compared with a top defensive player (V. van Dijk). We remove three features with low attribution from both players. The attribution projection starts with the selected instance on the right. We vary the contribution from defense (`def`), the star offensive player is not distinguished in the horizontal direction. At this point, defensive players have been rotated to the highest horizontal value. (A video of the animated radial tour can be found at vimeo.com/666431163.)

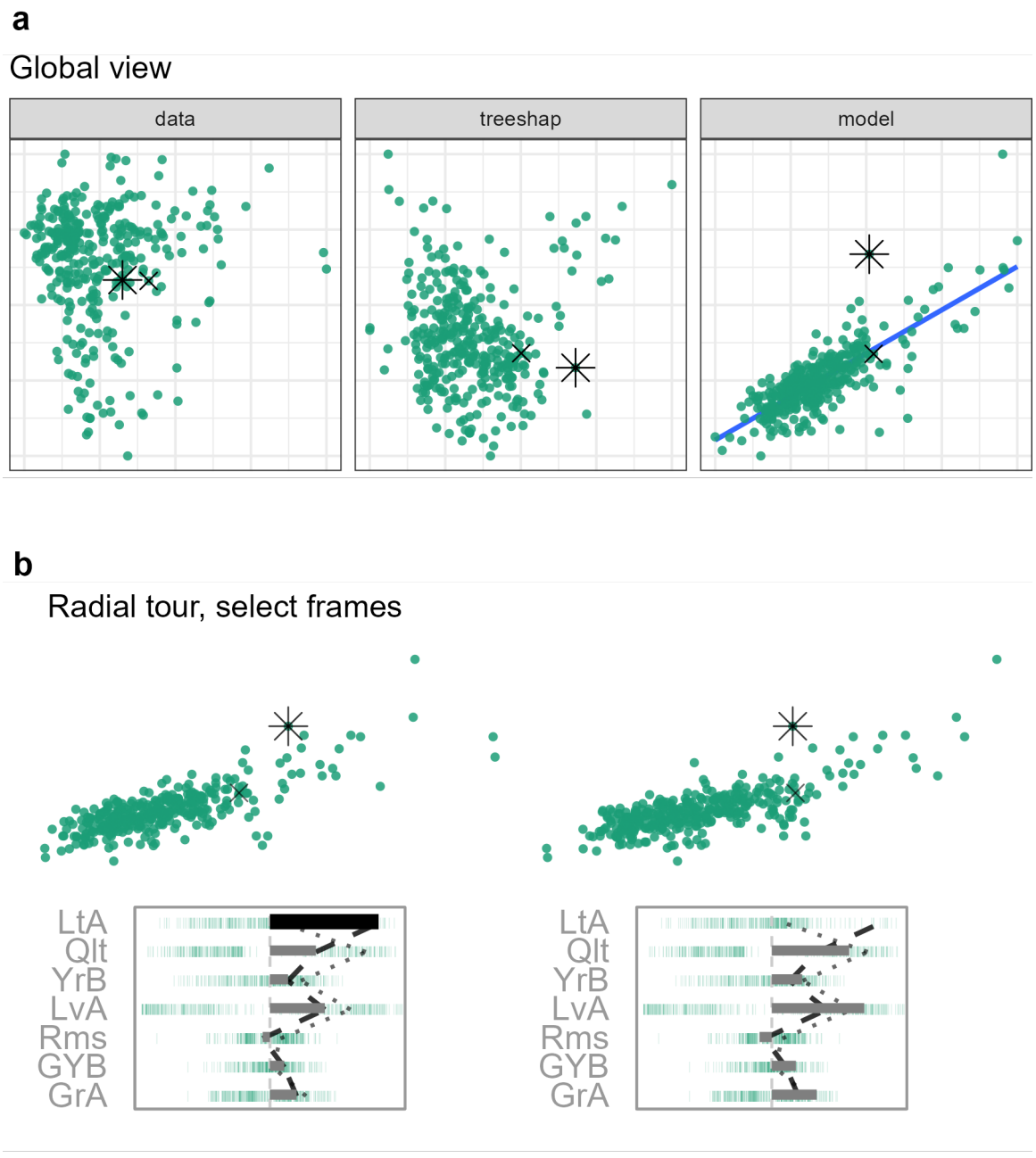

Figure 9: Ames housing 2018 regressing sales price [USD]. The PI sale price was under predicted and had sizable attribution to lot area (`LtA`). The CI was predicted sales price was similar and much more accurate with its observed sales price while it has very little attribution to lot area. Varying the contribution lot area the separation between these house sales crosses when there is a low contribution of `LtA`, which is important to explaining the PI and near invariant to the sales price of the CI. (A video showing the animation is at vimeo.com/666431134.)

explanation distinguishes. It is curious that the attribution space the same dimensionality of the data with a seemingly better separation. It sounds like a candidate to fit another model. However, such an attempt should want to be vigilant not to over-fit the training data.

7 Acknowledgments

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 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

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.

Biecek, Przemyslaw. 2018. "DALEX: Explainers for Complex Predictive Models in R." *The Journal of Machine Learning Research* 19 (1): 3245–49.

———. 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.

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.

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. https://doi.org/10.1007/978-3-540-33037-0_13.

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.

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.

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.

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.

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.

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. https://doi.org/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: 1369–86.

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 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*, 4768–77.

Molnar, Christoph. 2020. *Interpretable Machine Learning*. Lulu. com. christophm.github.io/interpretable-ml-book/.

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

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*, 1135–44. KDD '16. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

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.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. "Learning Important Features Through Propagating Activation Differences." In *International Conference on Machine Learning*, 3145–53. 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; 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. 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.

Unwin, Antony, and Pedro Valero-Mora. 2018. "Ensemble Graphics." *Journal of Computational and Graphical*

Statistics 27 (1): 157–65. 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)*, 548–57.

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.