# Understanding Black-Box Models

Yifei Sun

## Contents

**Variable importance** **2**

**Partial dependence plots (PDPs)** **3**

**Individual conditional expectation (ICE) curves** **4**

**lime** **5**

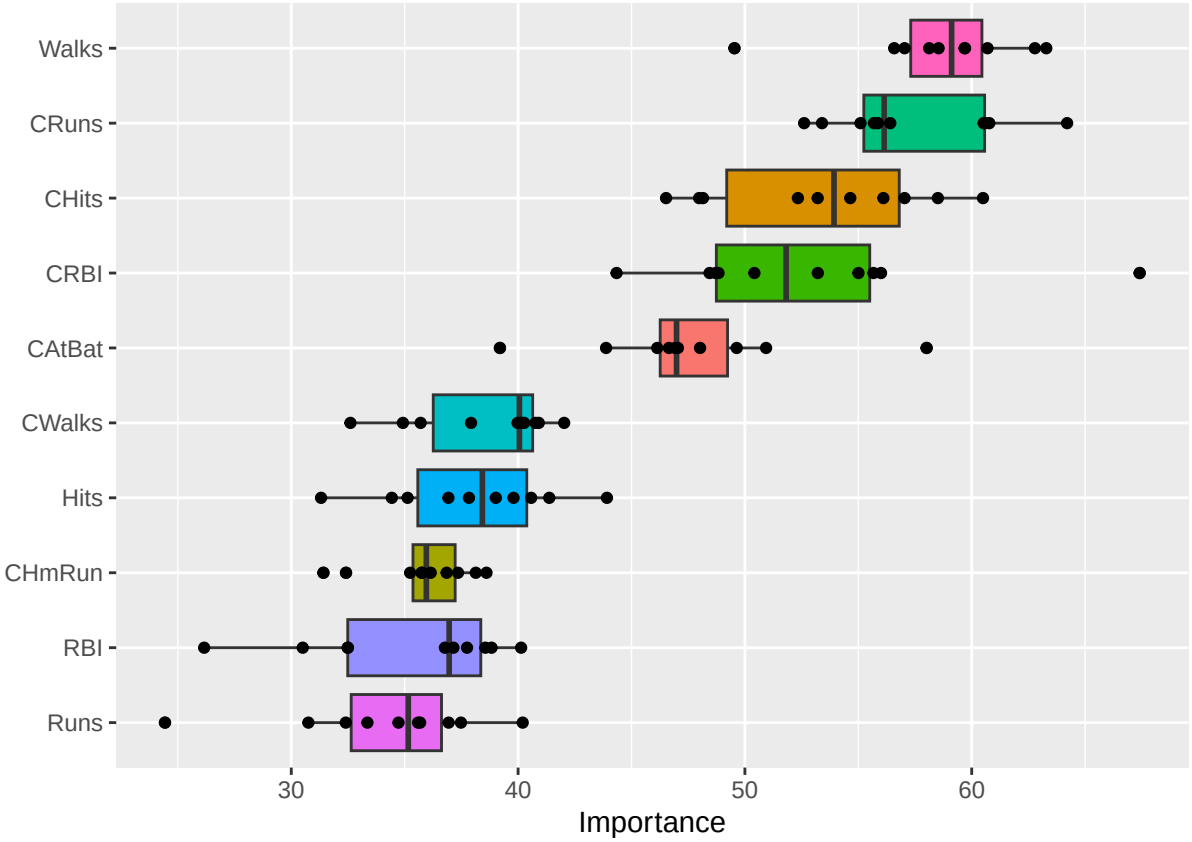
```
library(ISLR)
library(caret)
library(vip)
library(pdp)
library(lime)
library(gridExtra)
library(tidymodels)
```

Predict a baseball player's salary on the basis of various statistics associated with performance in the previous year.

```
data(Hitters)
Hitters <- na.omit(Hitters)

set.seed(2)
data_split <- initial_split(Hitters, prop = 0.8)

training_data <- training(data_split)
testing_data <- testing(data_split)


ctrl <- trainControl(method = "cv")

rf.grid <- expand.grid(mtry = 1:6,
                       splitrule = "variance",
                       min.node.size = 1:6)
set.seed(1)
rf.fit <- train(Salary~., training_data,
                method = "ranger",
                tuneGrid = rf.grid,
                trControl = ctrl)
```

## Variable importance

```
# You can use built-in VIMP in random forests
# The following code is to illustrate vip() in calculating VIMP in a general case
set.seed(1)
vip(rf.fit,
    method = "permute",
    train = training_data,
    target = "Salary",
    metric = "RMSE",
    nsim = 10,
    pred_wrapper = predict,
    geom = "boxplot",
    all_permutations = TRUE,
    mapping = aes_string(fill = "Variable"))
```

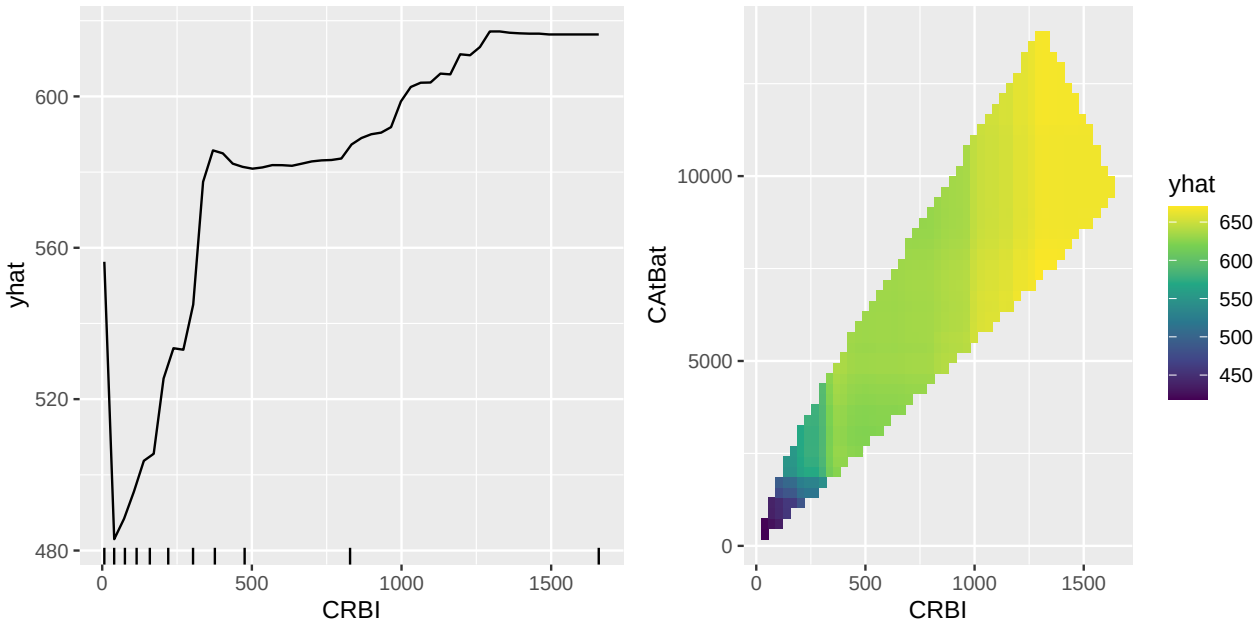



## Partial dependence plots (PDPs)

After the most relevant variables have been identified, the next step is to attempt to understand how the response variable changes based on these variables.

PDPs plot the change in the average predicted value as specified feature(s) vary over their marginal distribution.

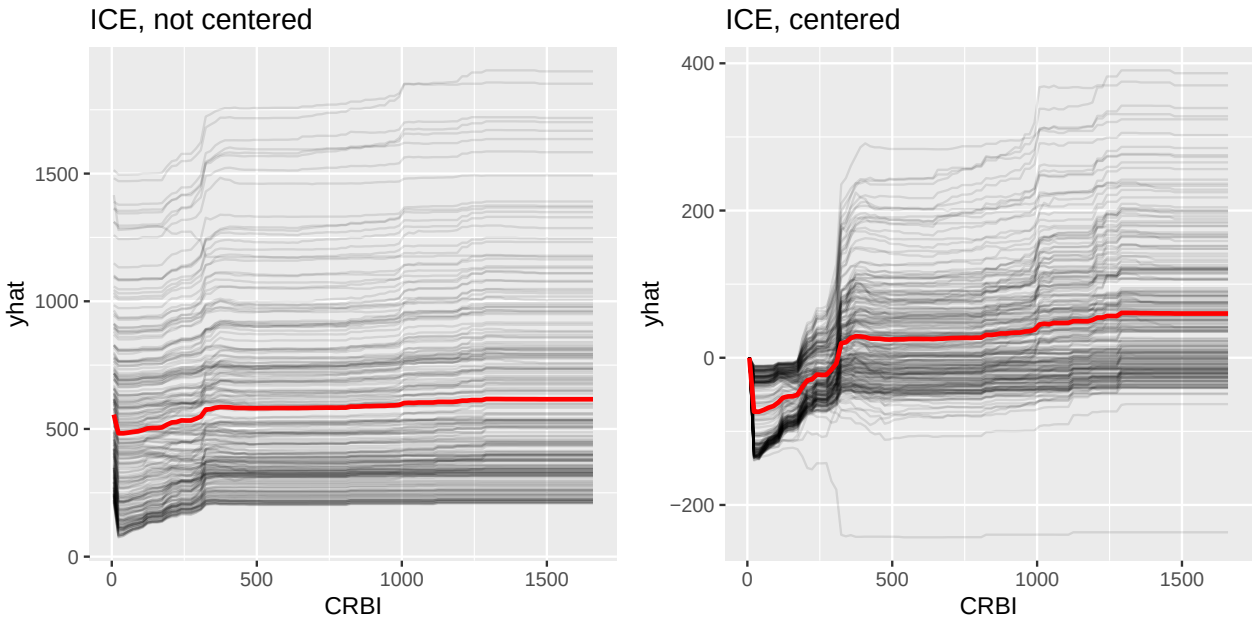
```
pdp1.rf <- rf.fit %>%
  pdp::partial(pred.var = c("CRBI")) %>%
  autoplot(train = training_data, rug = TRUE)

pdp2.rf <- rf.fit %>%
  pdp::partial(pred.var = c("CRBI","CAtBat"), chull = TRUE) %>%
  autoplot(train = training_data, rug = TRUE)

grid.arrange(pdp1.rf, pdp2.rf, nrow = 1)
```




## Individual conditional expectation (ICE) curves

ICE curves are an extension of PDP plots but, rather than plot the average marginal effect on the response variable, we plot the change in the predicted response variable for each observation as we vary each predictor variable.

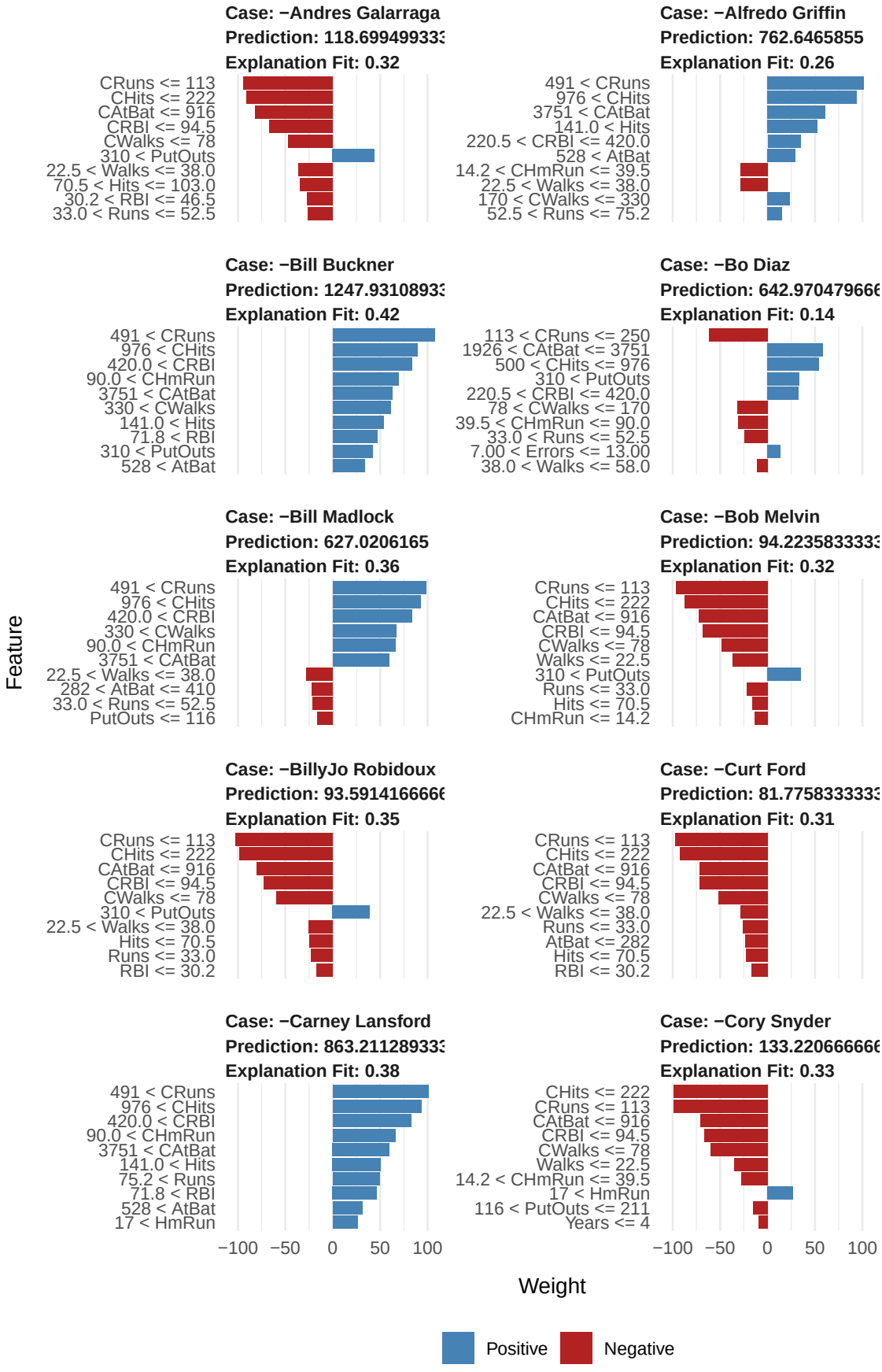
```
ice1.rf <- rf.fit %>%
  pdp::partial(pred.var = "CRBI",
               grid.resolution = 100,
               ice = TRUE) %>%
  autoplot(train = training_data, alpha = .1) +
  ggtitle("ICE, not centered")

ice2.rf <- rf.fit %>%
  pdp::partial(pred.var = "CRBI",
               grid.resolution = 100,
               ice = TRUE) %>%
  autoplot(train = training_data, alpha = .1,
           center = TRUE) +
  ggtitle("ICE, centered")


grid.arrange(ice1.rf, ice2.rf, nrow = 1)
```



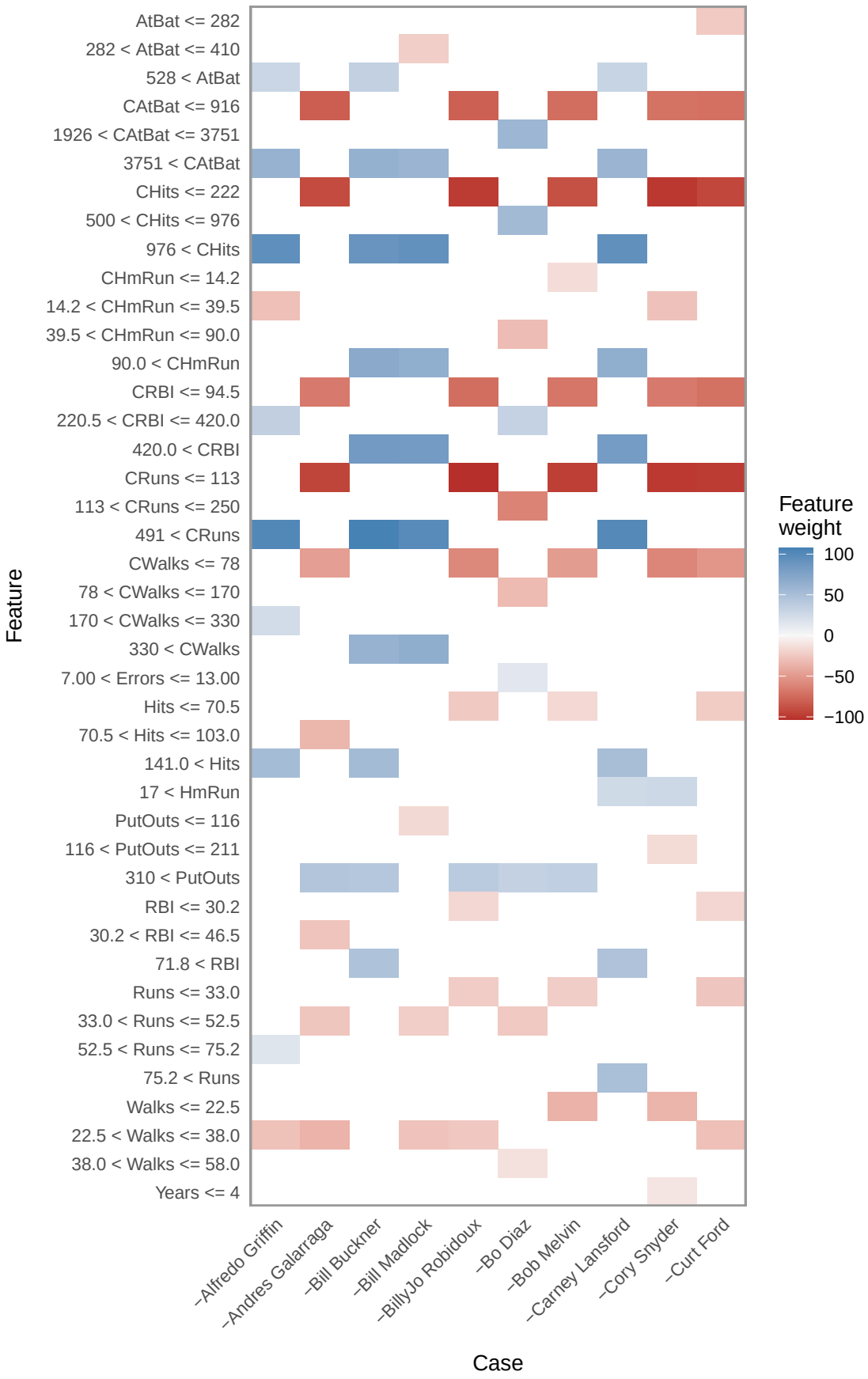

## lime

Once an explainer has been created using the `lime()` function, it can be used to explain the result of the model on new observations. The `explain()` function takes new observation along with the explainer and returns a data.frame with prediction explanations, one observation per row. The function `plot_features()` creates a compact visual representation of the explanations for each case and label combination in an explanation. Each extracted feature is shown with its weight, thus giving the importance of the feature in the label prediction.

```
explainer.rf <- lime(training_data[,-19], rf.fit)

new_obs <- testing_data[,-19][1:10,]
explanation.obs <- lime::explain(new_obs,
                                 explainer.rf,
                                 n_features = 10)

plot_features(explanation.obs)
```


Case: −Andres Galarraga
Prediction: 118.69949933
Explanation Fit: 0.32
CRuns <= 113
CHits <= 222
CAtBat <= 916
CRBI <= 94.5
CWalks <= 78
310 < PutOuts
22.5 < Walks <= 38.0
70.5 < Hits <= 103.0
30.2 < RBI <= 46.5
33.0 < Runs <= 52.5
Case: −Alfredo Griffin
Prediction: 762.6465855
Explanation Fit: 0.26
491 < CRuns
976 < CHits
3751 < CAtBat
141.0 < Hits
220.5 < CRBI <= 420.0
528 < AtBat
14.2 < CHmRun <= 39.5
22.5 < Walks <= 38.0
170 < CWalks <= 330
52.5 < Runs <= 75.2
Case: −Bill Buckner
Prediction: 1247.9310893
Explanation Fit: 0.42
491 < CRuns
976 < CHits
420.0 < CRBI
90.0 < CHmRun
3751 < CAtBat
330 < CWalks
141.0 < Hits
71.8 < RBI
310 < PutOuts
528 < AtBat
Case: −Bo Diaz
Prediction: 642.970479666
Explanation Fit: 0.14
113 < CRuns <= 250
1926 < CAtBat <= 3751
500 < CHits <= 976
310 < PutOuts
220.5 < CRBI <= 420.0
78 < CWalks <= 170
39.5 < CHmRun <= 90.0
33.0 < Runs <= 52.5
7.00 < Errors <= 13.00
38.0 < Walks <= 58.0
Case: −Bill Madlock
Prediction: 627.0206165
Explanation Fit: 0.36
491 < CRuns
976 < CHits
420.0 < CRBI
330 < CWalks
90.0 < CHmRun
3751 < CAtBat
22.5 < Walks <= 38.0
282 < AtBat <= 410
33.0 < Runs <= 52.5
PutOuts <= 116
Case: −Bob Melvin
Prediction: 94.2235833333
Explanation Fit: 0.32
CRuns <= 113
CHits <= 222
CAtBat <= 916
CRBI <= 94.5
CWalks <= 78
Walks <= 22.5
310 < PutOuts
Runs <= 33.0
Hits <= 70.5
CHmRun <= 14.2
Case: −BillyJo Robidoux
Prediction: 93.591416666
Explanation Fit: 0.35
CRuns <= 113
CHits <= 222
CAtBat <= 916
CRBI <= 94.5
CWalks <= 78
310 < PutOuts
22.5 < Walks <= 38.0
Hits <= 70.5
Runs <= 33.0
RBI <= 30.2
Case: −Curt Ford
Prediction: 81.7758333333
Explanation Fit: 0.31
CRuns <= 113
CHits <= 222
CAtBat <= 916
CRBI <= 94.5
CWalks <= 78
22.5 < Walks <= 38.0
Runs <= 33.0
AtBat <= 282
Hits <= 70.5
RBI <= 30.2
Case: −Carney Lansford
Prediction: 863.211289333
Explanation Fit: 0.38
491 < CRuns
976 < CHits
420.0 < CRBI
90.0 < CHmRun
3751 < CAtBat
141.0 < Hits
75.2 < Runs
71.8 < RBI
528 < AtBat
17 < HmRun
Case: −Cory Snyder
Prediction: 133.220666666
Explanation Fit: 0.33
CHits <= 222
CRuns <= 113
CAtBat <= 916
CRBI <= 94.5
CWalks <= 78
Walks <= 22.5
14.2 < CHmRun <= 39.5
17 < HmRun
116 < PutOuts <= 211
Years <= 4
Feature
−100
−50
0
50
100
Weight
Positive
Negative

```
plot_explanations(explanation.obs)
```


AtBat <= 282
282 < AtBat <= 410
528 < AtBat
CAtBat <= 916
1926 < CAtBat <= 3751
3751 < CAtBat
CHits <= 222
500 < CHits <= 976
976 < CHits
CHmRun <= 14.2
14.2 < CHmRun <= 39.5
39.5 < CHmRun <= 90.0
90.0 < CHmRun
CRBI <= 94.5
220.5 < CRBI <= 420.0
420.0 < CRBI
CRuns <= 113
113 < CRuns <= 250
491 < CRuns
CWalks <= 78
78 < CWalks <= 170
170 < CWalks <= 330
330 < CWalks
7.00 < Errors <= 13.00
Hits <= 70.5
70.5 < Hits <= 103.0
141.0 < Hits
17 < HmRun
PutOuts <= 116
116 < PutOuts <= 211
310 < PutOuts
RBI <= 30.2
30.2 < RBI <= 46.5
71.8 < RBI
Runs <= 33.0
33.0 < Runs <= 52.5
52.5 < Runs <= 75.2
75.2 < Runs
Walks <= 22.5
22.5 < Walks <= 38.0
38.0 < Walks <= 58.0
Years <= 4
Feature
−Alfredo Griffin
−Andres Galarraga
−Bill Buckner
−Bill Madlock
−BillyJo Robidoux
−Bo Diaz
−Bob Melvin
−Carney Lansford
−Cory Snyder
−Curt Ford
Case
Feature weight
100
50
0
−50
−100