Интерпретация моделей машинного обучения

Пётр Болотин

Март 2020

План на семинар

- Зачем нужна интерпретация
- Shapley values в теории игр
- Shapley values для интерпретации ml
- SHAP теория
- SHAP примеры использования

Зачем нужна интерпретация

Цель интерпретации - понять, как в конкретном примере разные признаки влияют на ответ модели.

- Требует заказчик
- Самопроверка

Shapley values в теории игр

Задача

Пусть вокалист в переходе на трубной зарабатывает 10 у.е в час, гитарист 5, а вместе 10 у.е, как им стоит делить прибыль при совместном выступлении?

Shapley values в теории игр

Задача

Пусть вокалист в переходе на трубной зарабатывает 10 у.е в час, гитарист 5, а вместе 10 у.е, как им стоит делить прибыль при совместном выступлении?

Что изменится, если вместе они зарабатывают 20 у.е?

Shapley values в теории игр

Задача

Пусть вокалист в переходе на трубной зарабатывает 10 у.е в час, гитарист 5, а вместе 10 у.е, как им стоит делить прибыль при совместном выступлении?

Что изменится, если вместе они зарабатывают 20 у.е?
Какую формулу можно предложить для универсального расчета в таких ситуациях?

Shapley values в теории игр

Definition

Shapley values это такое распределение выигрышей, когда каждый игрок получает средний вклад в выигрыш коалиции

Shapley values в теории игр

Definition

Shapley values это такое распределение выигрышей, когда каждый игрок получает средний вклад в выигрыш коалиции

В случае с музыкантами shapley value для вокалиста равен (10+15)/2, а для гитариста (5+10)/2

Свойства shapley values

- Симметричность, то есть нет зависимости от номера игрока
- Если игрок не приносит прибыли, он не получает ничего
- Эффективность, сумма shap values равна выигрышу максимальной коалиции
- Линейность. Если коалиция участвует в двух играх, то shapley values общей игры можно получить как сумму shapley values соответствующих игр.

Свойства shapley values

- Симметричность, то есть нет зависимости от номера игрока
- Если игрок не приносит прибыли, он не получает ничего
- Эффективность, сумма shap values равна выигрышу максимальной коалиции
- Линейность. Если коалиция участвует в двух играх, то shapley values общей игры можно получить как сумму shapley values соответствующих игр.

Theorem

Shap values это единственный способ разделить прибыль, удовлетворяющий написанным выше аксиомам.

Shapley values в машинном обучении

Какие аналоги у музыкантов и прибыли, заработанной в переходе?

Shapley values в машинном обучении

Какие аналоги у музыкантов и прибыли, заработанной в переходе?

- Признаки и predict модели, коалиция это просто подмножество признаков.

Shapley values в машинном обучении

Какие аналоги у музыкантов и прибыли, заработанной в переходе?

- Признаки и predict модели, коалиция это просто подмножество признаков.

Признаки и predict модели, коалиция это просто подмножество признаков.

Алгоритм расчета Shapley values

- M - число итераций, x - объект из датасета, j - индекс фичи, f - модель, X - матрица признаков.
- For all $\mathrm{m} = 1, \ldots, \mathrm{M}$:
  - Возьми случайный объект z из X
  - Берём случайный набор признаков F
  - Конструируем два новых объекта:
    - Случайные признаки F вместе с признаком j заполняем значениями из x, остальные берем из z: $x_{+j} = \left(x_{(1)}, \ldots, x_{(k)}, x_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
    - Аналогично, но признак j тоже берем из z: $x_{-j} = \left(x_{(1)}, \ldots, x_{(k)}, z_{(j)}, z_{(j+1)}, \ldots, z_{(p)}\right)$
  - Вычисляем вклад признака j в ответ модели: $\phi_j^m = \hat{f}\left(x_{+j}\right) - \hat{f}\left(x_{-j}\right)$
- Compute Shapley value as the average: $\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

SHAP (SHapley Additive exPlanations)

Общая идея в том, чтобы приблизить в точке x исходную модель линейной и получить shapley values из коэффициентов линейной модели.
Более подробное объяснение теории:

- https://christophm.github.io/interpretable-ml-book/shap.html

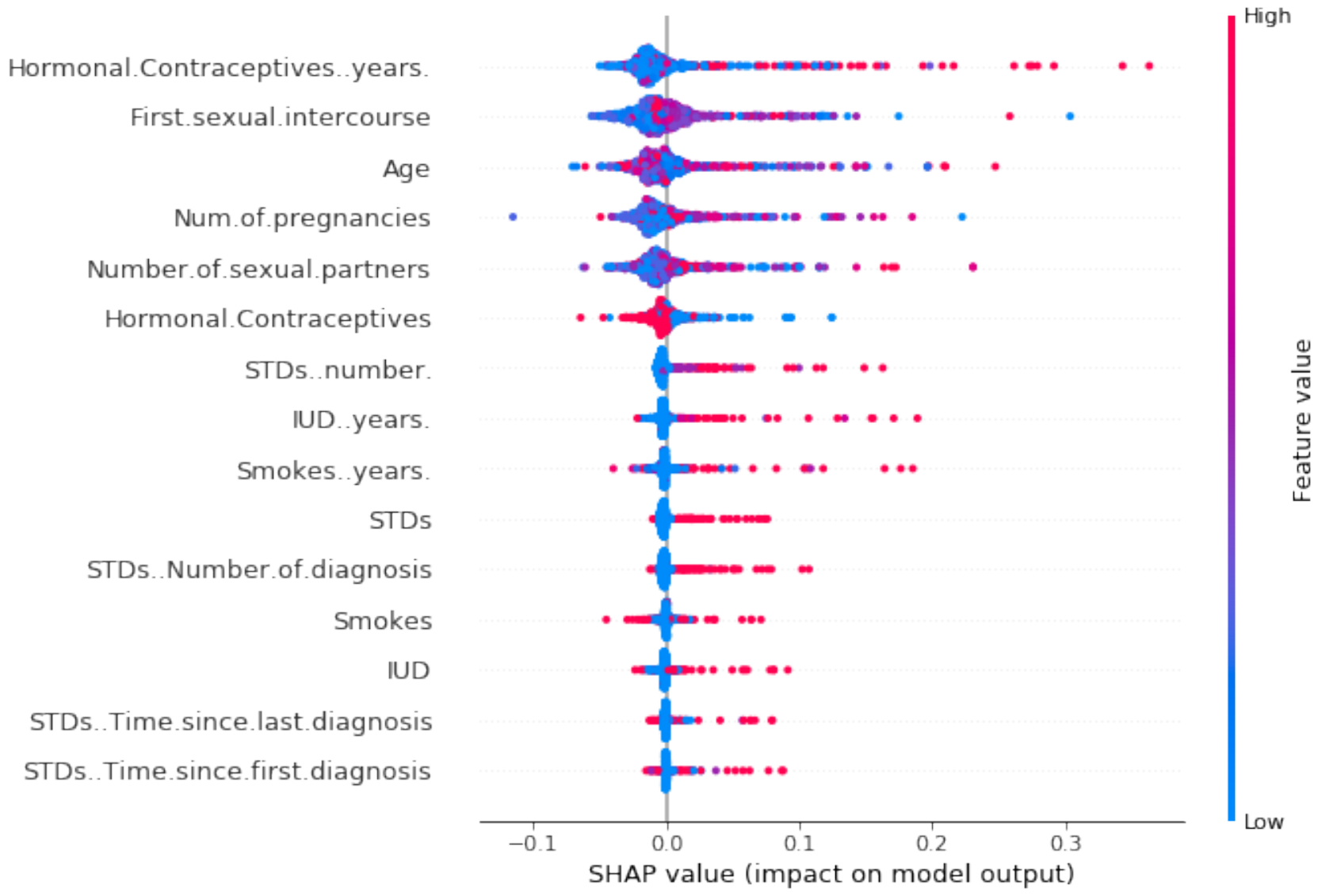
Библиотека для расчета:

- https://github.com/slundberg/shap

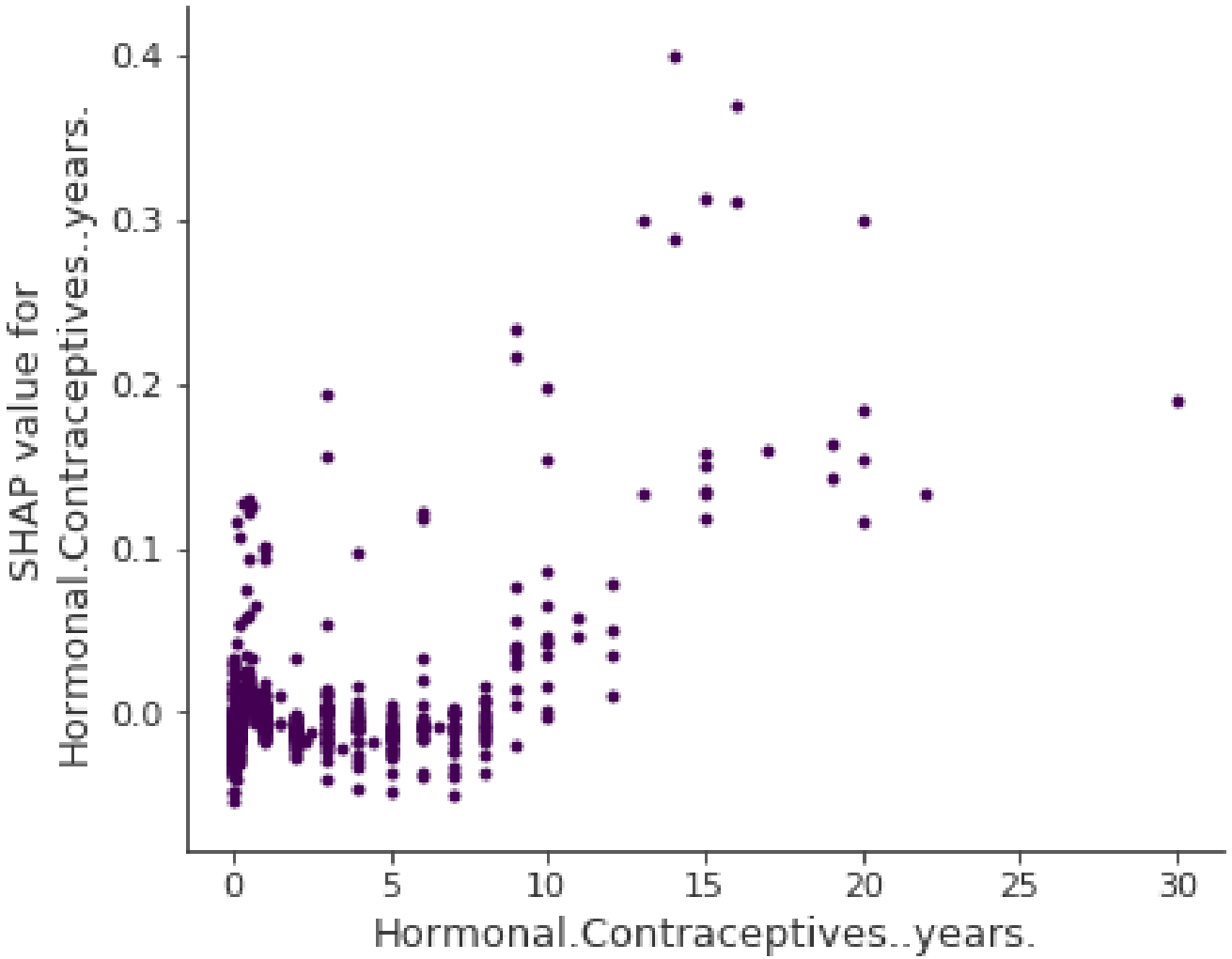
SHAP (SHapley Additive exPlanations)

```python
import xgboost
import shap

# train an XGBoost model
X, y = shap.datasets.boston()
model = xgboost.XGBRegressor().fit(X, y)

# explain the model's predictions using SHAP
# (same syntax works for LightGBM, CatBoost, scikit-learn, transformers, Spark, etc.)
explainer = shap.Explainer(model)
shap_values = explainer(X)

# visualize the first prediction's explanation
shap.plots.waterfall(shap_values[0])
```

SHAP Summary Plot



SHAP Dependence Plot



SHAP Dependence Plot