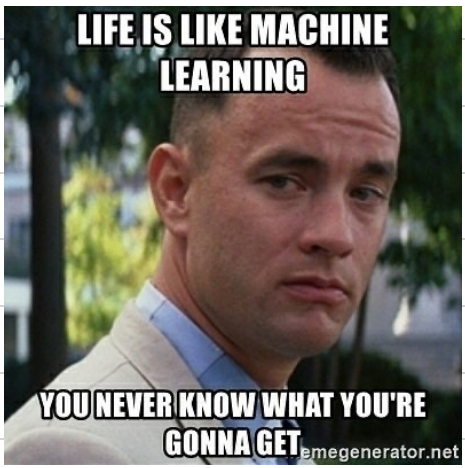
Session 12

Model Interpretability

March 04, 2024



Agenda

1. Lime
2. Shap

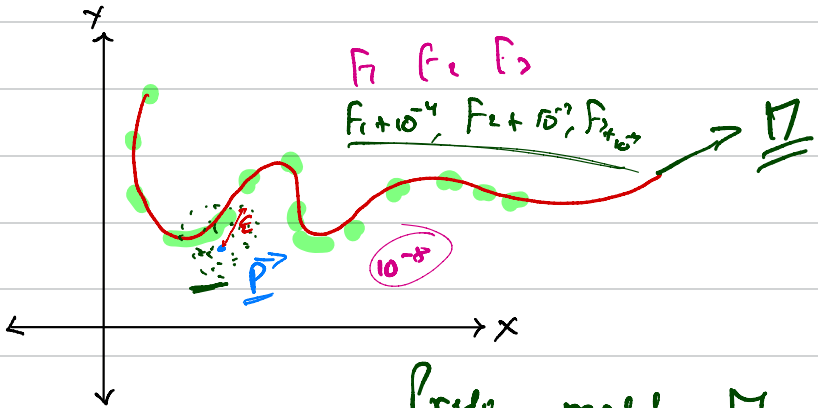
Why do we need it??

① Hedge fun: Mathematic model -> exploit

Feature Importance	Model Explainability
look at global level	look at local level
Entire dataset	sample level

Lime -> deals at local level
Shap -> " " global ".

LIME



Predict model M on new generated Points

1, 10 → Some noise → 50, 10

x_1 — — — y_1

— — y_2

x_{50} — — y_{50}

Fit a Linear model + lasso

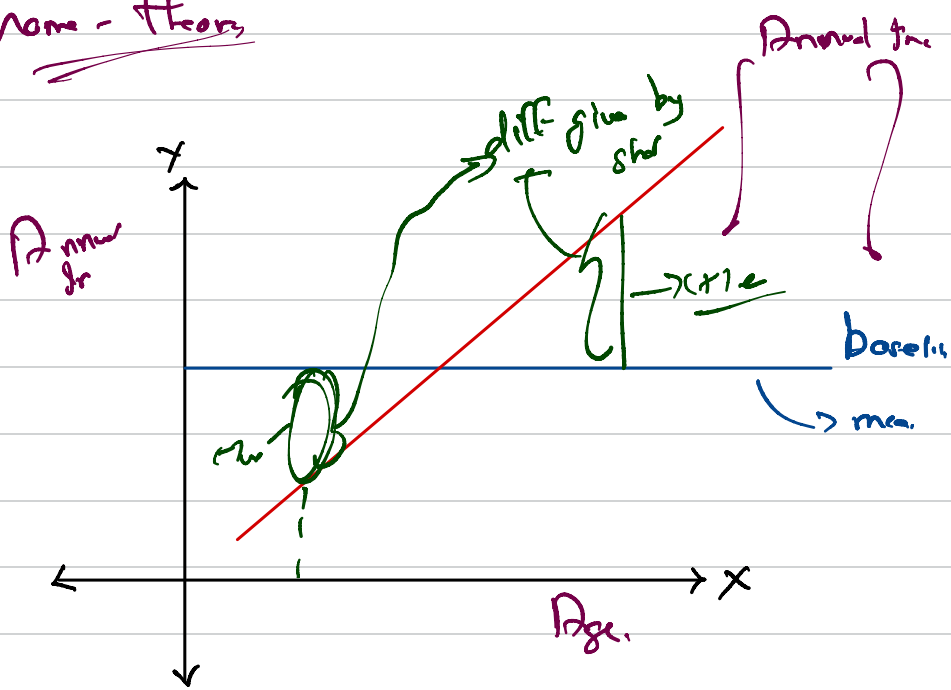
coeff of my linear model → help me understand why my model

is giving that output

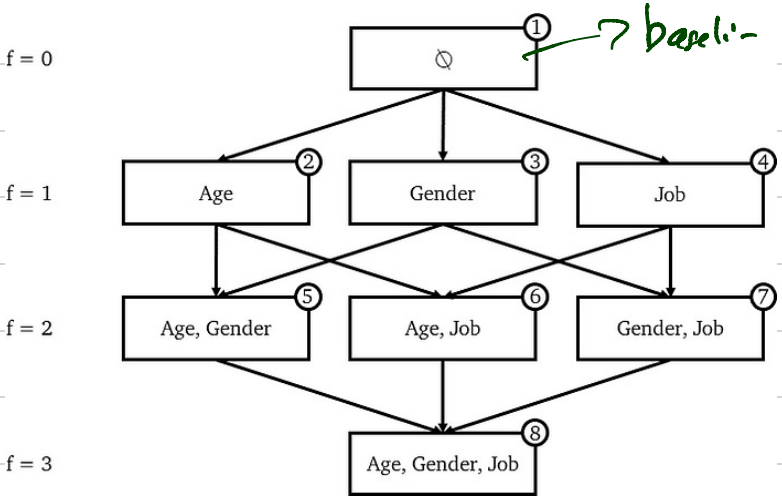
→ Points closer to blue point will be given higher weightage

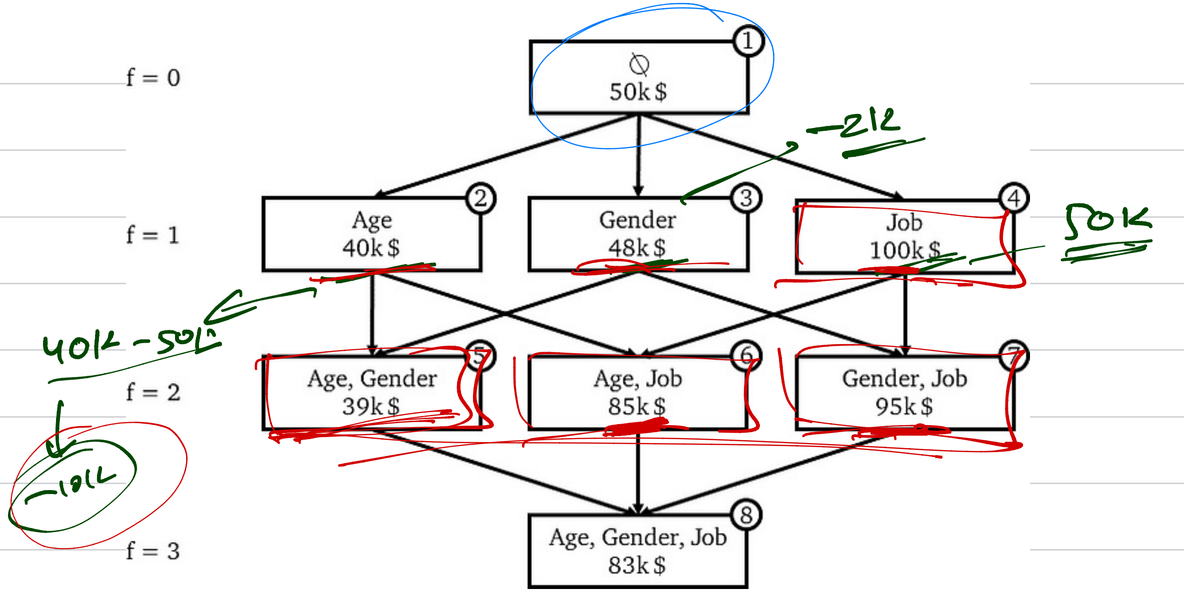
SHAP

Game theory



NN mod


f = 0
f = 1
f = 2
f = 3
1
∅
7 baseli-
2
Age
3
Gender
4
Job
5
Age, Gender
6
Age, Job
7
Gender, Job
8
Age, Gender, Job



~9K ~15K } Age

~2K

Age Gender Job } NN → Output

Prediction

Shapl ← Training data mean — Age, Gender, Job

10

1 → Age, Gen Avg., Job Avg. → Output

450K, 400?

Age — 50k

avg
of
across
rows

(2) Age Gender Job
avg avg

Job

Age → 20k

Baseline + 20k → 70k

Shapley additive Explanation