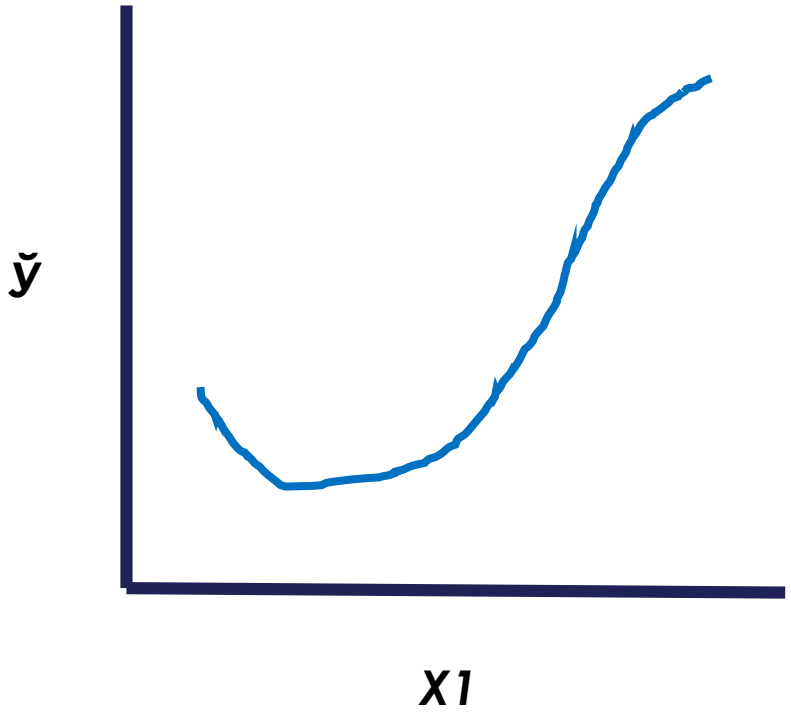
Accumulated local effects

Intuition

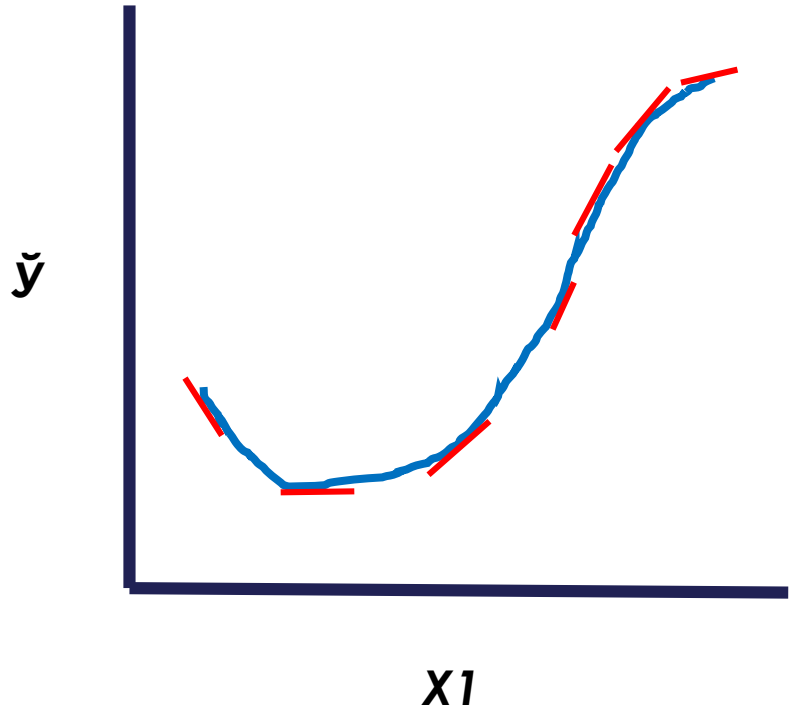
ALE - Intuition



Aim: Understand how a feature influences the prediction of a machine learning model on average.

$\hat{y} = f(x1, x)$

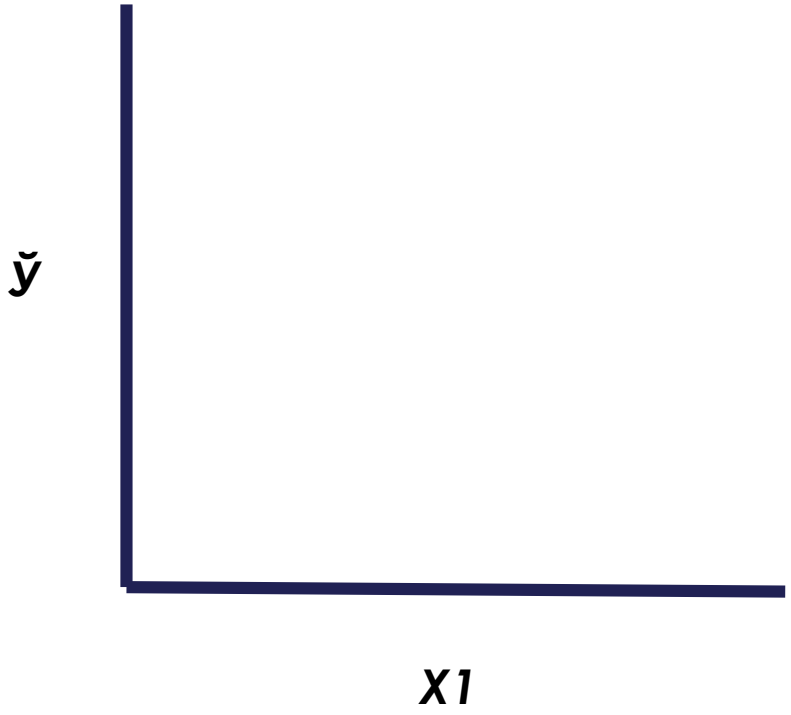
ALE - Intuition



Calculate the slope at each point.

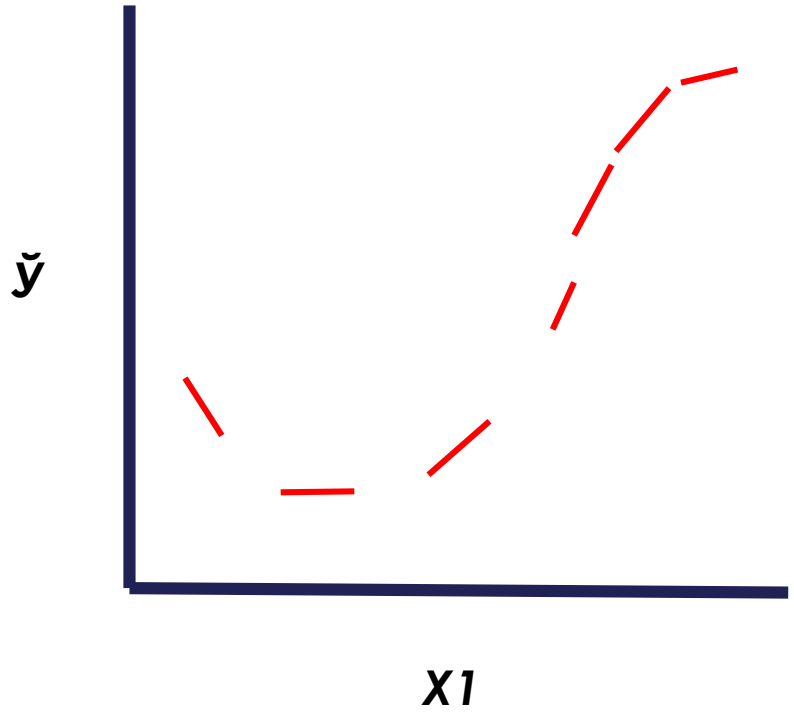
Aggregate the slopes.

ALE - Intuition



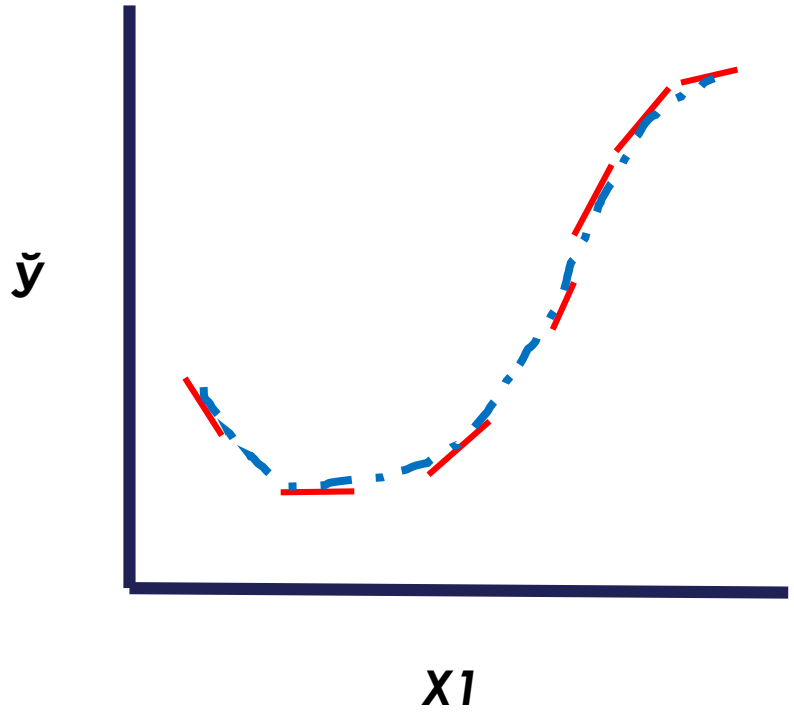
Understand the relationship between ŷ and X1.

ALE - Intuition



Calculate the slope at each point.

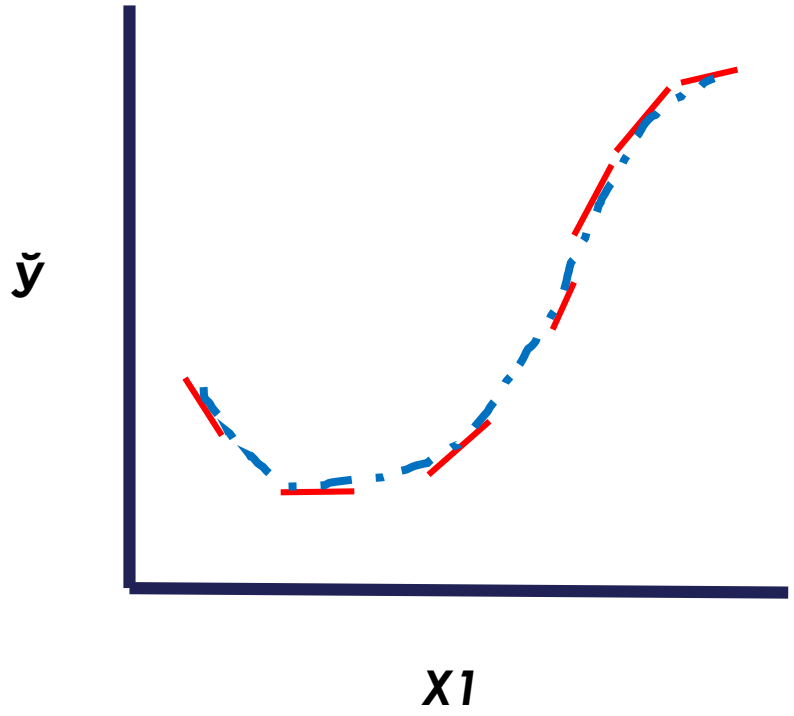
ALE - Intuition



Calculate the slope at each point.

Aggregate the slopes.

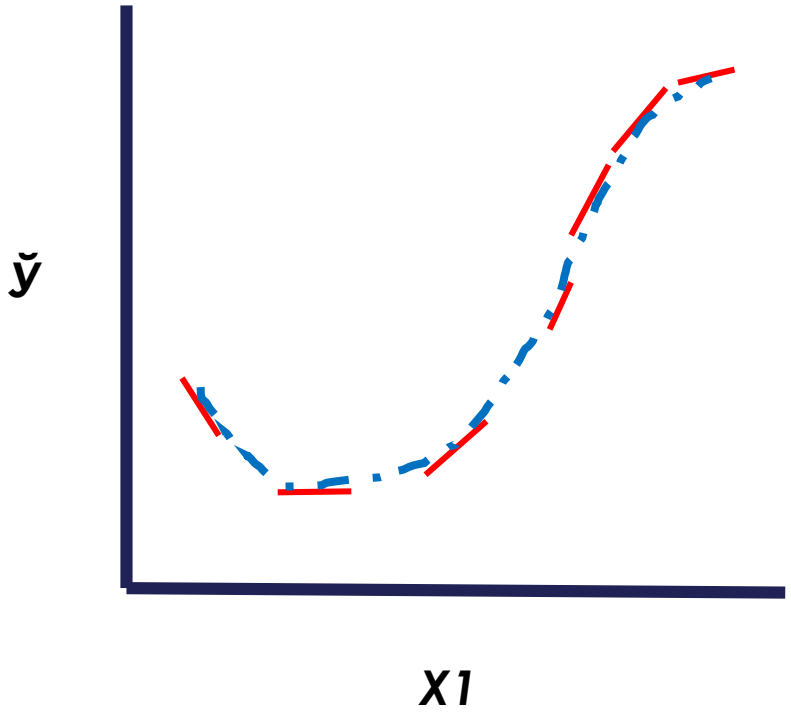
ALE - Formulation



Slopes: Partial derivative with respect of x1.

Aggregation: Integral of those partial derivatives.

ALE - Formulation



$$\widetilde{ALE}_{\hat{f},j}(x) = \int_{z_{0,j}}^{x} E_{X_c|X_j}[\hat{f}^j(X_j, X_c) \mid X_j = z_j]\, dz_j,$$

Xj is the variable we are examining.

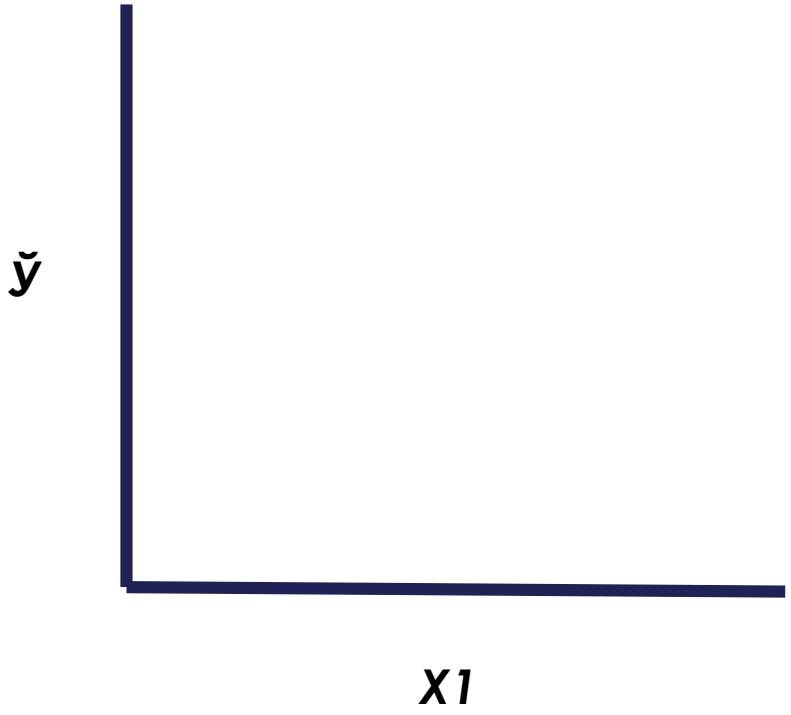
Xc represents all the other variables.

$\hat{f}^j$ is the partial derivative when Xj = z

z are X1 possible values.

We integrate (∫) over all z.

ALE – Calculation?



How can we calculate the ALE when we do not have / know ŷ?

Train In Data

THANK YOU

www.trainindata.com