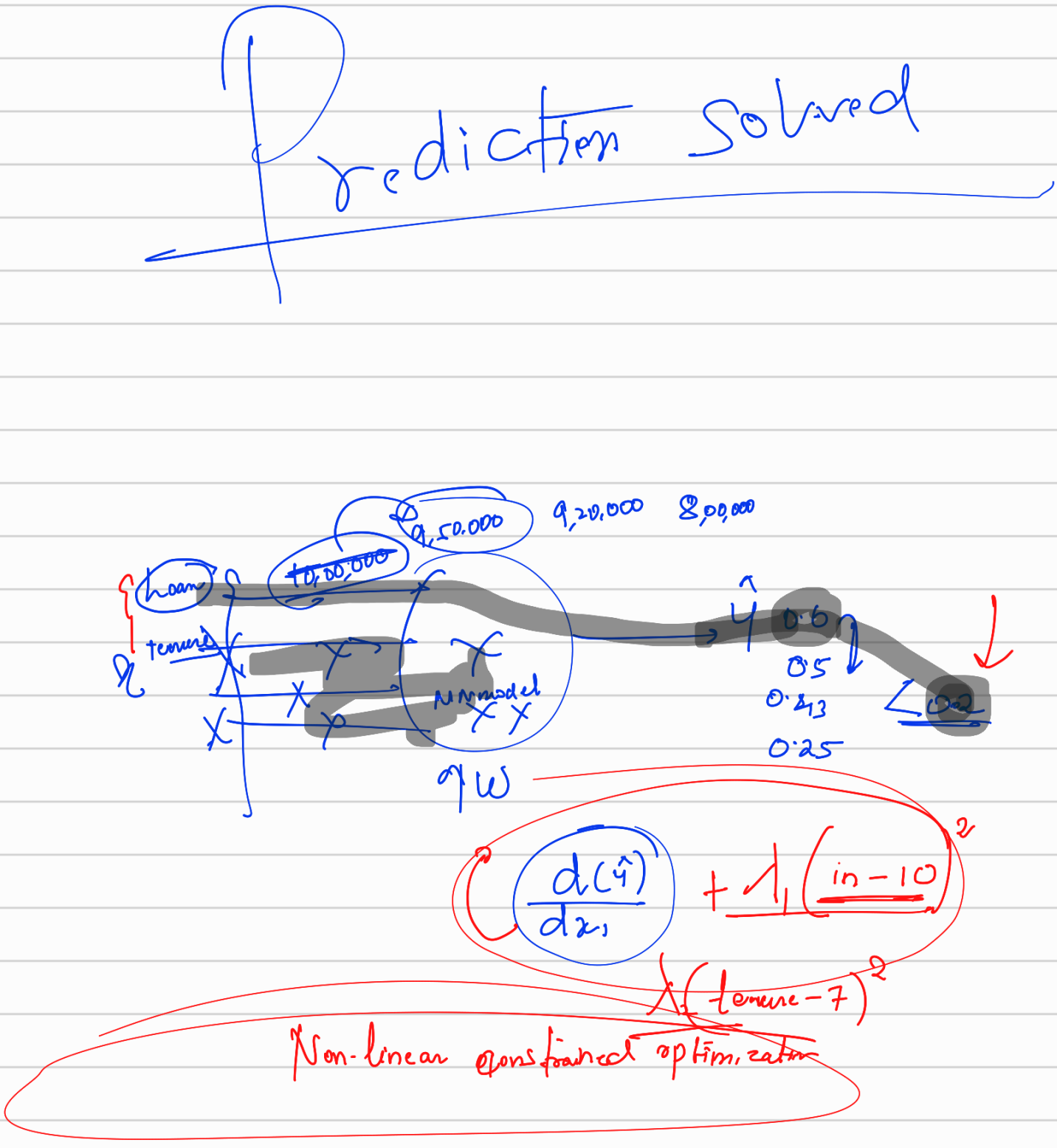
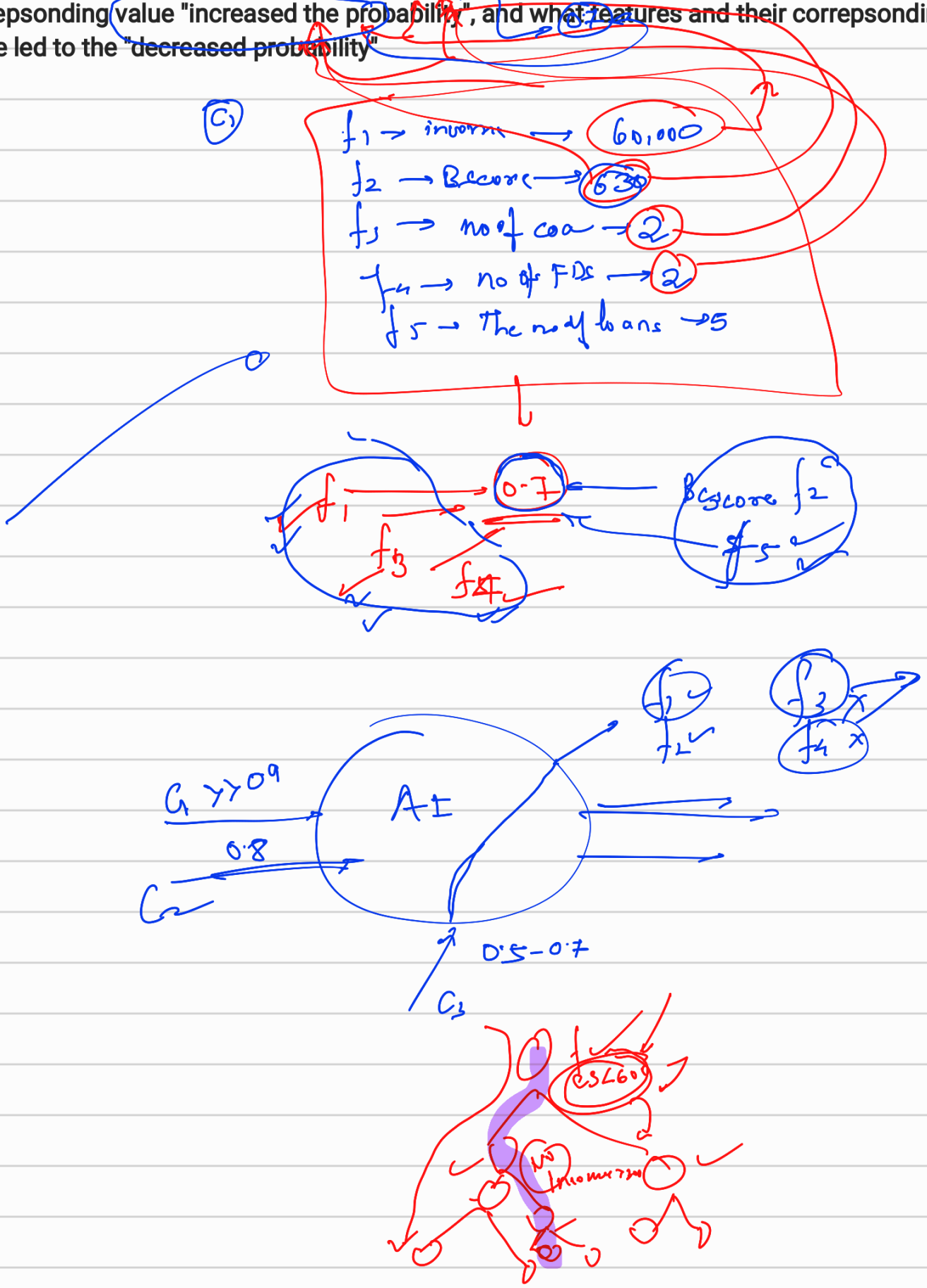


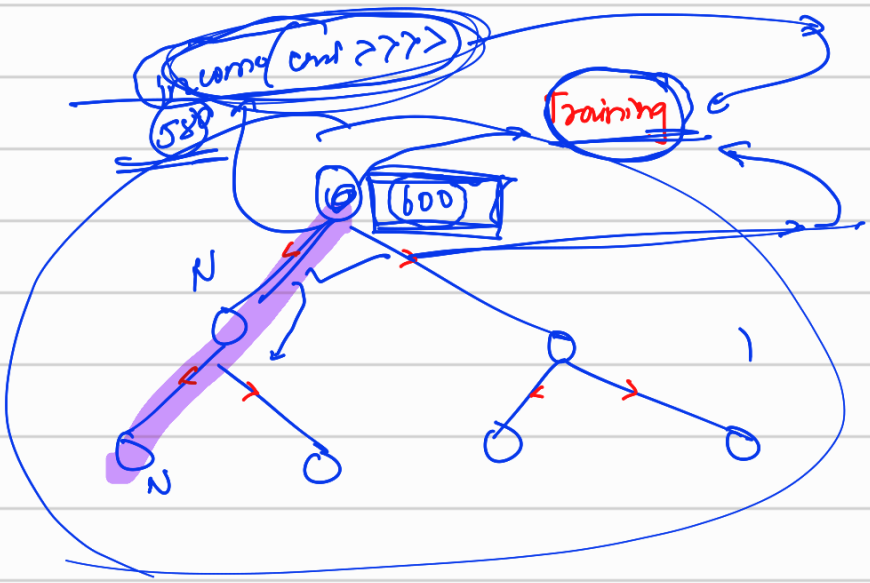
Model Explainability:
is a concept where in we are able to explain the prediction that we do, in the sense that we can atleast let the user know that what was the reason that the current sample got this prediction.

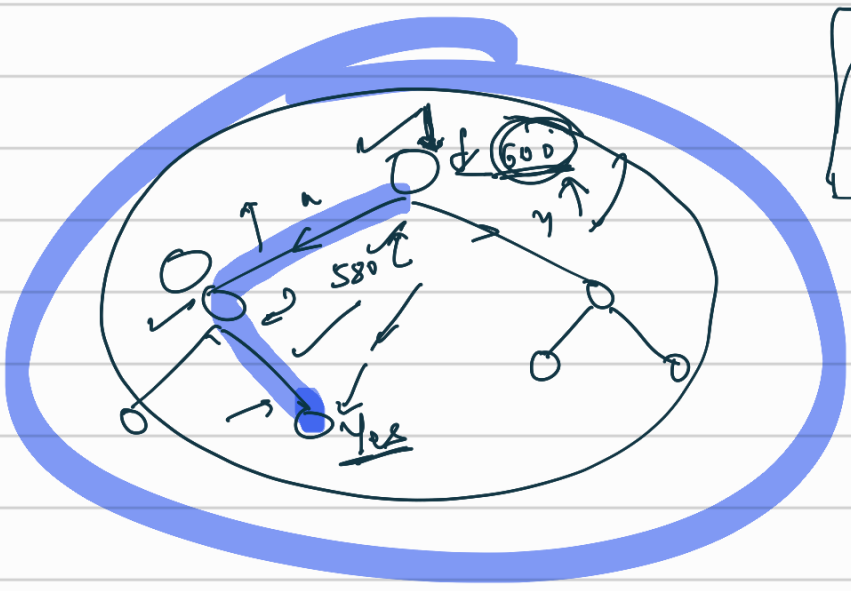
Feature importance- Model Explainability?
Feature importance from DT/XGB models can be considered as explainability however they

work on global level, they can help you to do feature selection

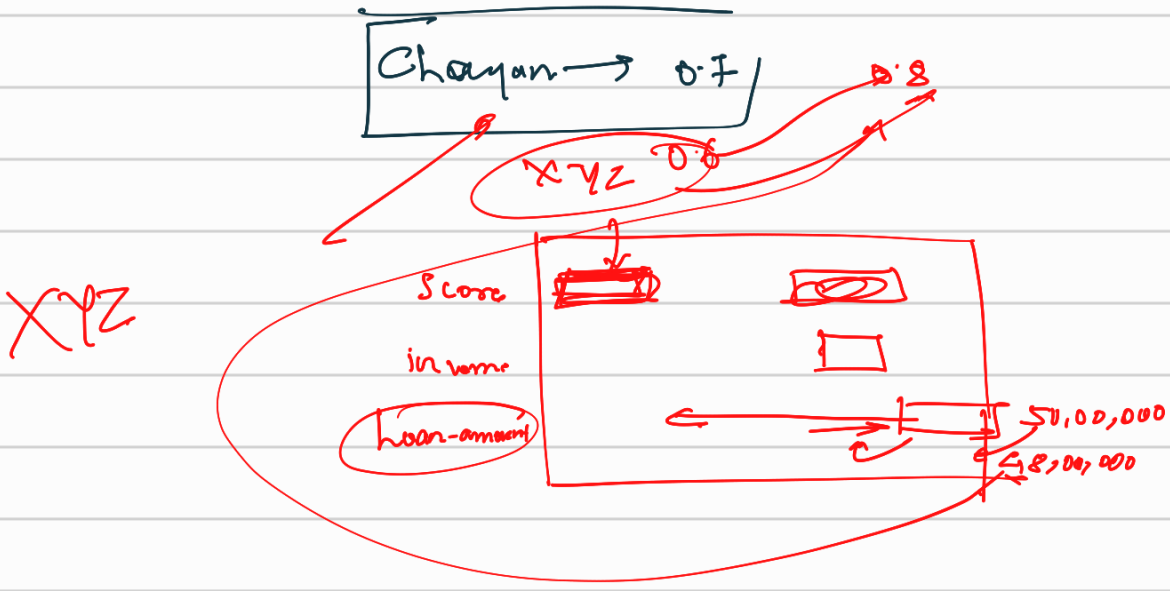
How do we define explainability?
Explainability is during the inference time/prediction has to work at a sample level, what features led to this value.
Classification Models- we should be able to tell at a sample level that what features and its correpsonding value "increased the probability", and what features and their correpsonding value led to the "decreased probability"





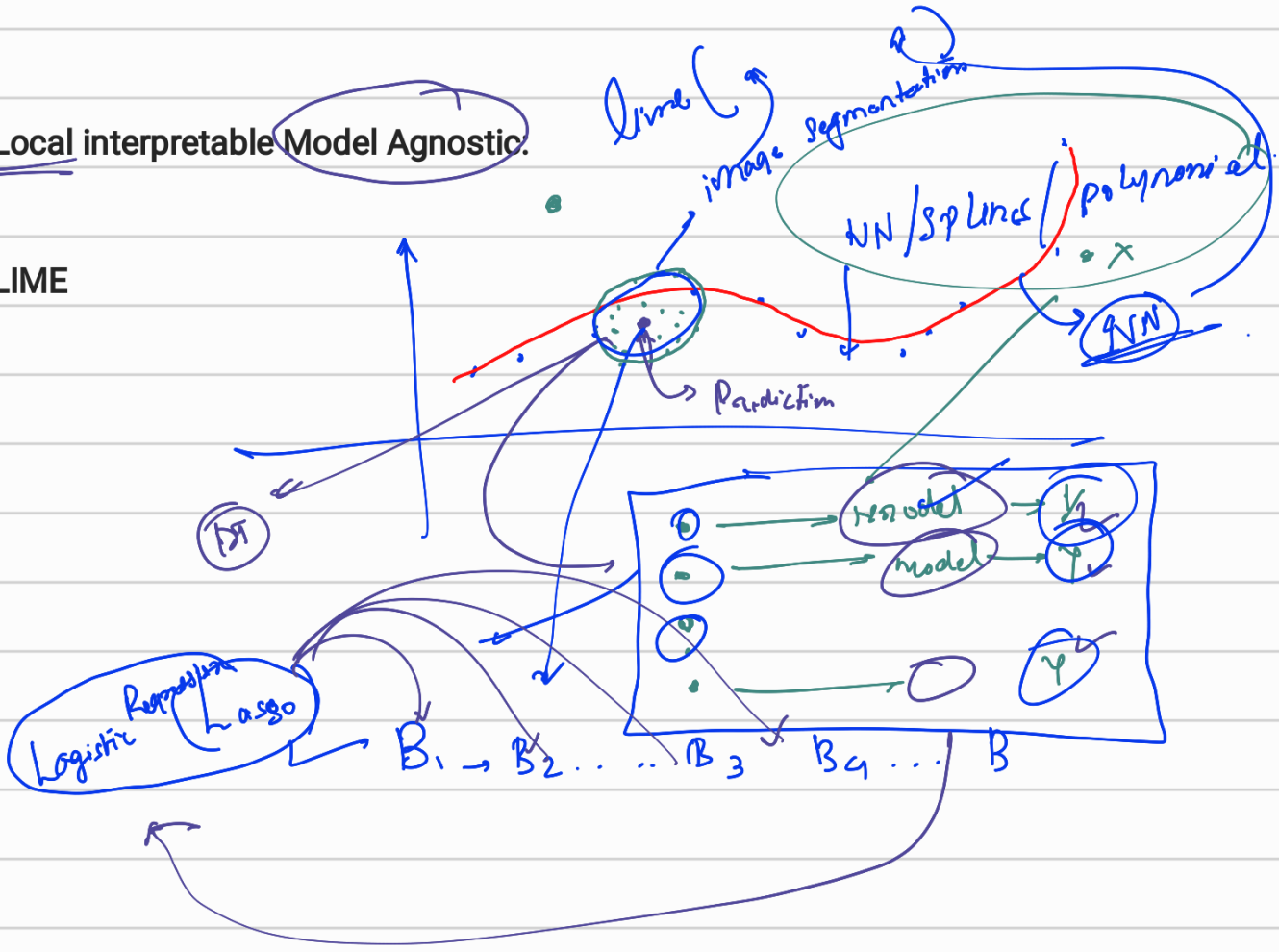


Local level / Sample level

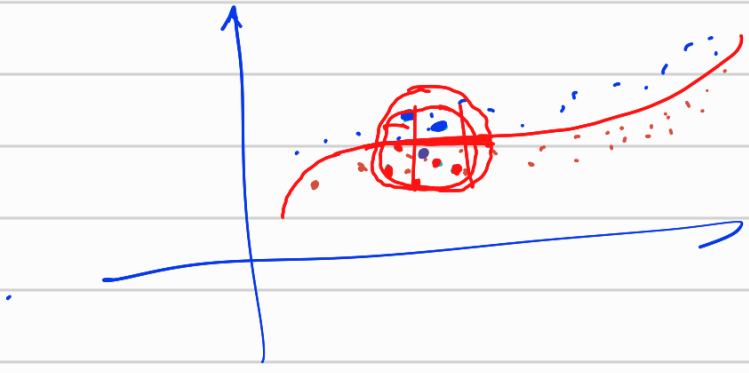


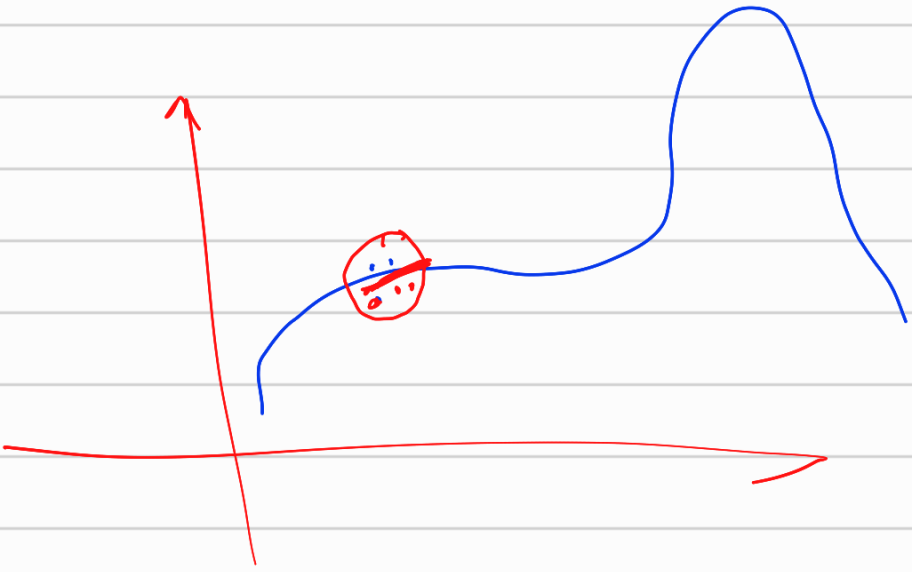
Local interpretable Model Agnostic.

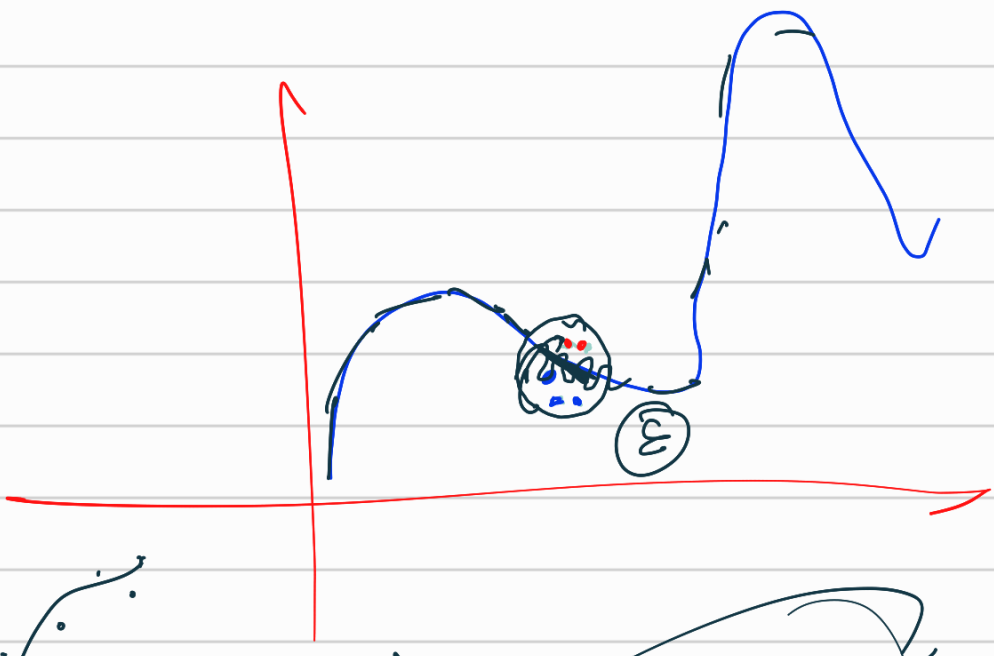
LIME

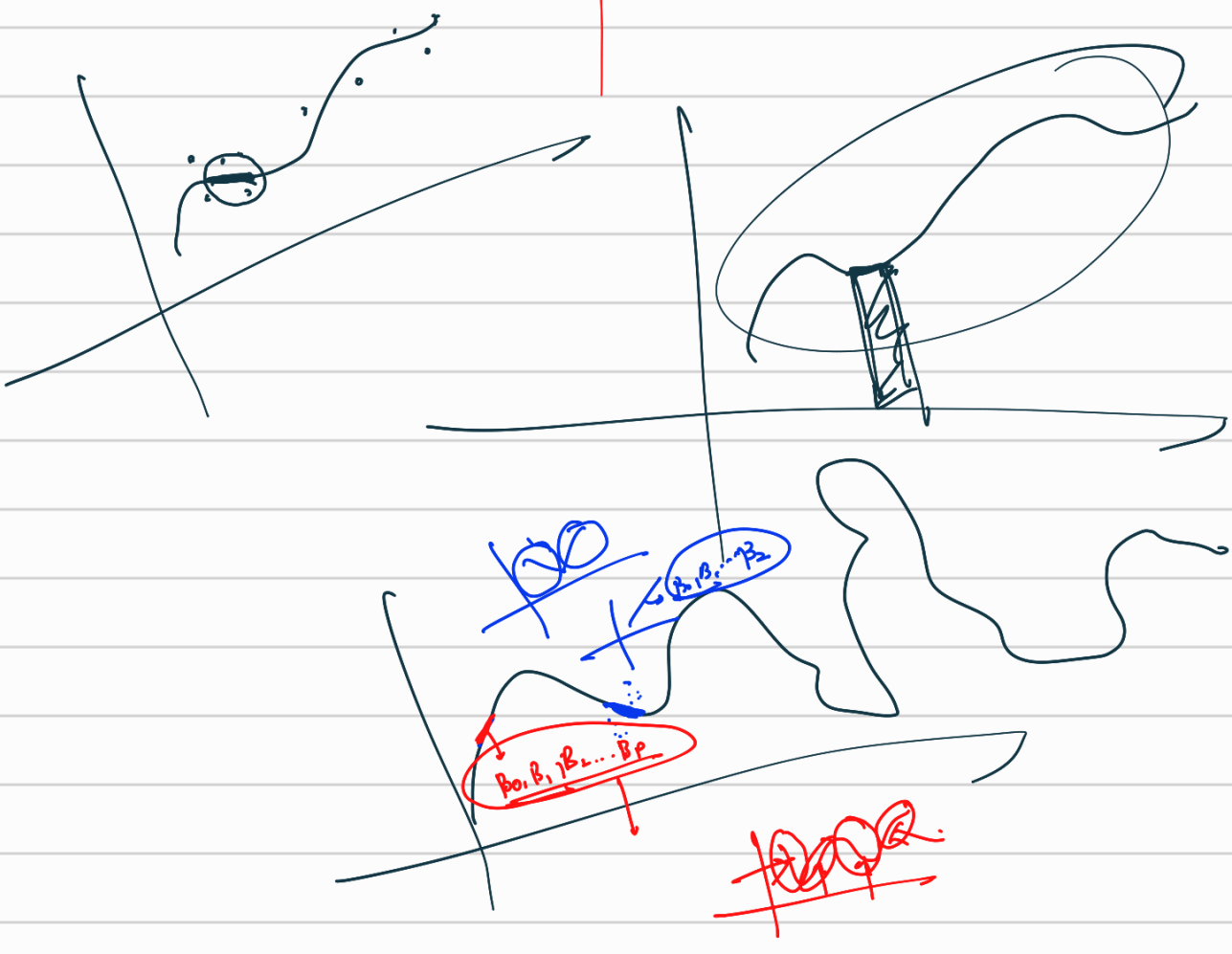


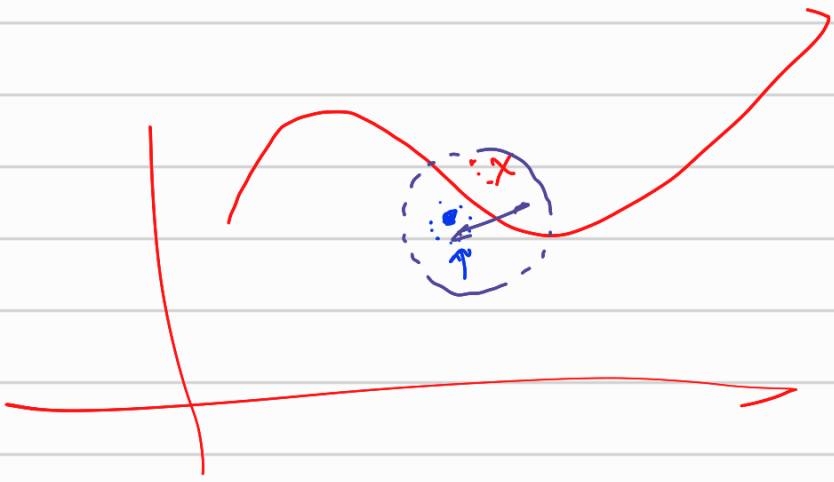
- We will be given a single point (purple Point) for prediction
- We will sample random points around the given point within a epsilon radius
- All these random points will be passed through the main model (red curve/NN mode)
- thereby we would have created a sample data of green points and there corresponding predictions
- This above data becomes a training data for me.
- I select a simple (explainable model like DT/Logistic Regression) to fit the above data
- The feature importance/coefficients become the explainability factors of the given purple point



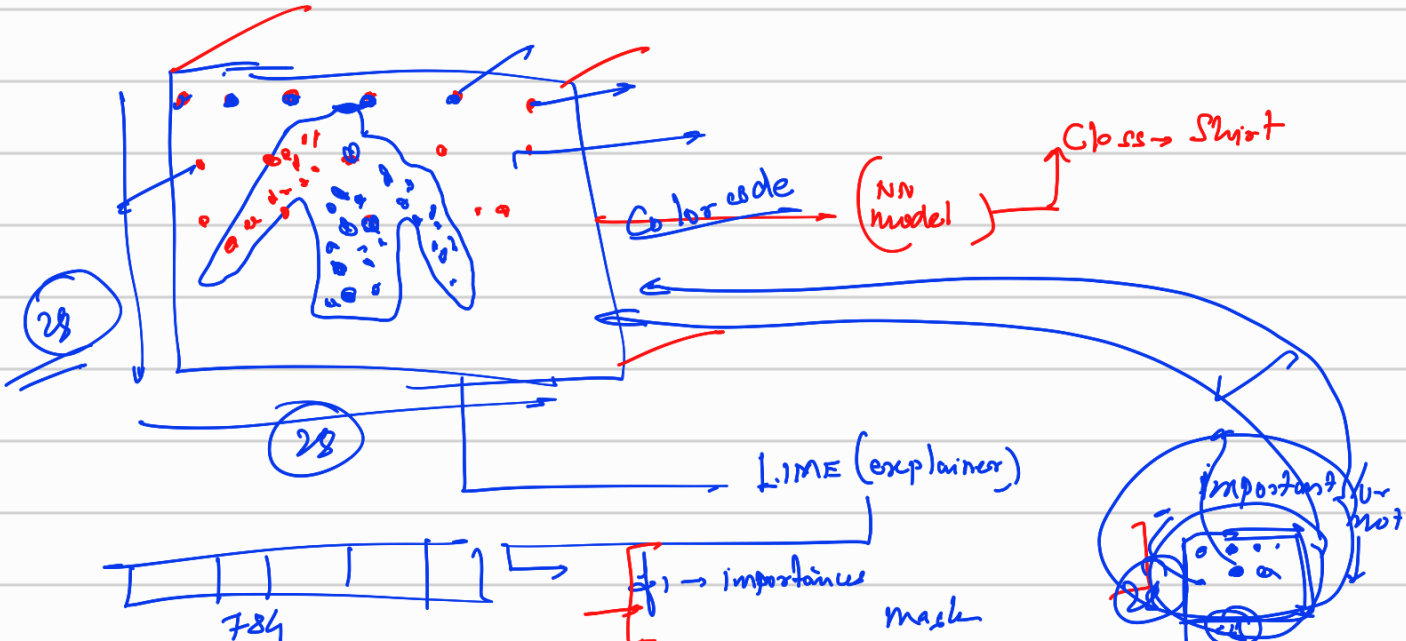


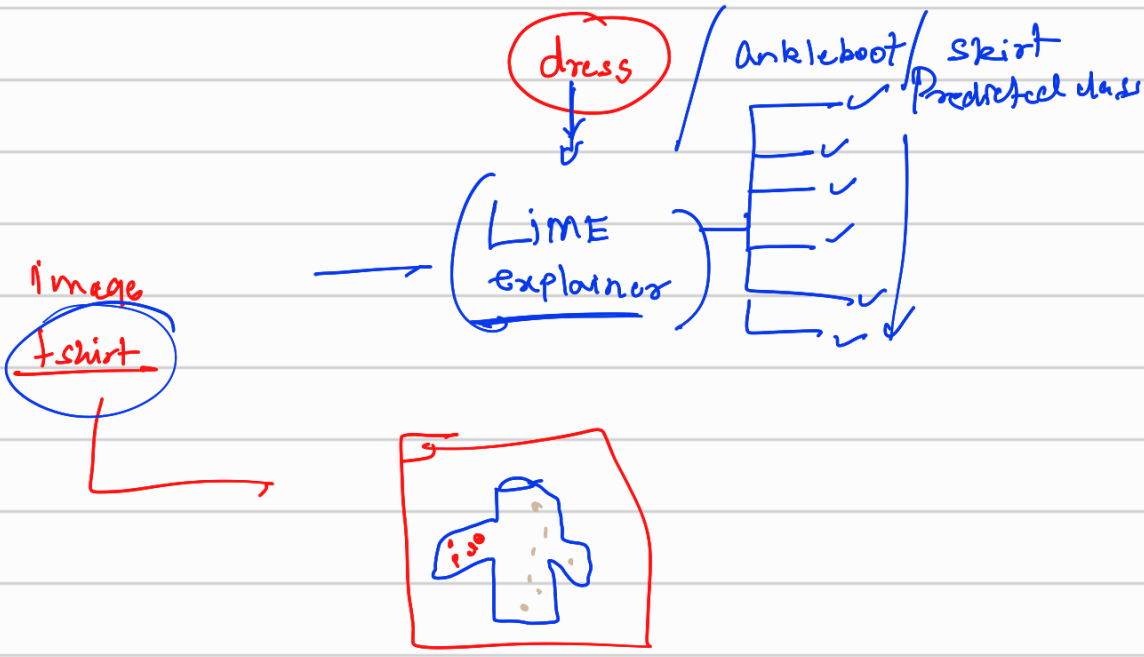

ε

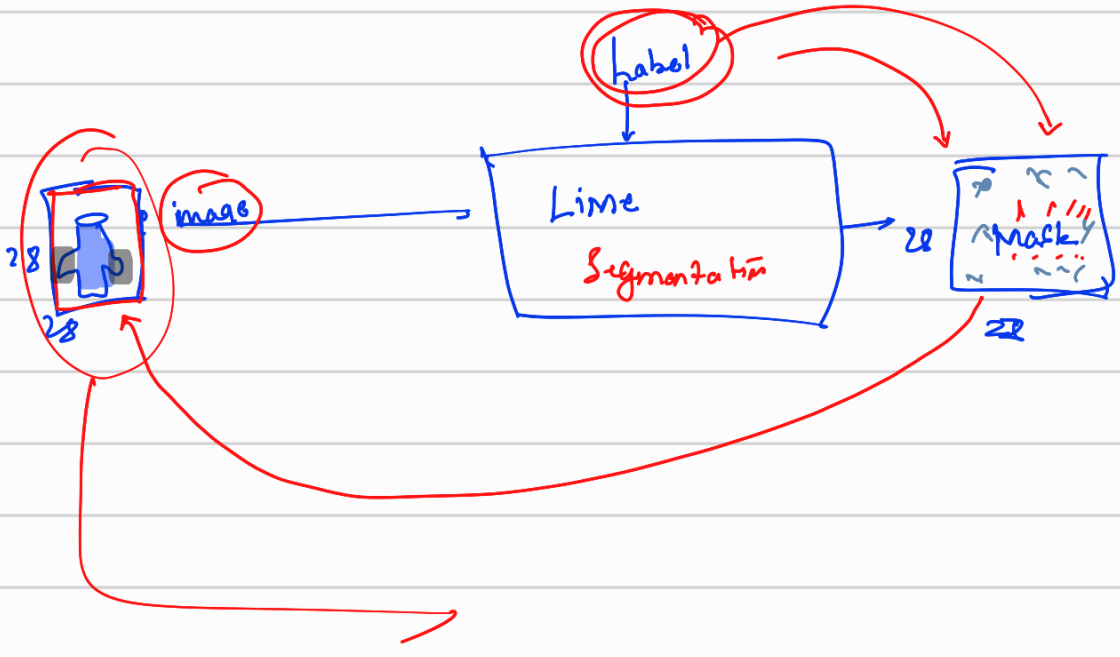

$\beta_0, \beta_1, \beta_2 \ldots \beta_p$


x

Shapley ' O /


Color code
NN model
Class → Shirt
28
28
LIME (explainer)
784
→ importance
mask
important
not


dress
ankleboot
skirt
Predicted class
LIME
explainer
image
tshirt


label
image
28
28
LIME
Segmentation
28
mask
28