**18th May session: Logistic regression**

**Exploration of EDA using plotly**

<https://www.plotly.express>

How to see overlapping in the data?

**Ans:** plotly shows but Lavi said that there are techniques to see te overlap & how to go about it.

If we add more data points, then we would be able to know overlap.

If we remove overlapped columns, then we get good model

For ex. If we remove sepal length & sepal width which is almost 50% of my data, then also we can do the prediction because other 2 columns also have 3 classes which we require to make predictions.

Good things:

* The prediction with 2 columns is always better than 4 columns because it gives more accurate results & faster than 4 columns. So its always better to work on less amount of data than higher amount.

Bad things:

* 2 columns are important in business point of view and business doesn’t want them to remove.

But there are 2 problems:

* If we remove 2 columns from 4, you are getting rid of the important information in regards to prediction.
* Domain perspective as domain expertise doesn’t want to remove 2 columns in business point of you, may be it will lose out some business information.

**Questions:**

1. **Why do we do we use log (p/1-p)?**
2. **What is (p/1-p)?**
3. **Why we use odds and not likelihood?**

We are looking for likelihood of success v/s failure or we are interested in likelihood of something happening to something not happening

1. **Why do we use log?**

To normalise bcz normalisation is simple and once we use odds then we get values ranges from 0 to infinity.

(p/1-p) – odds : probability between 0 to infinity

Log(odds) – Bets 0 + beta1x – probability between -infinity to +infinity

We want p value between 0 & 1 that’s why we use log(odds)

1. **What is logit function?**

log(p/1-p) = beta0+beta1 x is the logit function

It does not give s shaped curve

1. **What is sigmoid function?**

(P Values) (It gives s shaped curve)

Inverse of logit function is called the sigmoid function

P = e to the power (beta0+ beta1x)/(1 +e to the power (beta0+ beta1x))

P value ranges from 0 to 1

1. **Why do we need P value between 0 & 1**

Bcz it’s a classification problem/ category in 0 & 1

**Some Information:**

* The sigmoid function is used in deep leaning(Activation functions)
* We are using binary classification problem which has 2 class but we can use multinomial classification (where we can use more than 2 class)
* Ex. If we have 5 classes, x vs not x, 1 vs not 1,, 2 vs not 2…etc means we are building 5 models todo the classification of
* Dummy as the feature & dummy (x vs not x) as class perspective

1. **Why do logistic regression is benchmark model?**

It’s a standard, Simple & efficient, low variance (Probability & the normalize ranges, what happens is even f we have huge variation in the data, then we are still in the range of values between 0 & 1)

1. **Most common benchmark that we can use before the selection of model?**

Use dummy classifier

1. **How to address missing value**

less than 2-3%/<1 million missing value then its good enough to remove the column?

If we have more than 5% />1 million then best way is add mean/median, even we can use model to predict missing values

- Mean/Median

Global & local:

Global is get the mean/median of the particular column

Local is get the all column mean /median

- Removing rows

Model based: Global n local