**Machine Learning**

# **Understanding Data**

7 questions you must ask yourself.

1. How big data is?
2. How does data look like?
3. What are data types of columns?
4. Are there any missing values?
5. How does data look mathematically?
6. Are there any duplicate values?
7. How is the correlation between columns?

# **EDA**

Exploratory Data Analysis

Asking question again and again and finding their answers.

## **Univariant Analysis**

### **Categorical Variable**

1. Count Plot, how much value were there in each
2. For percentage use Pie chart

### **Numerical Variable**

1. **Histogram**, to understand the distribution of the data
2. **Displot**, tell the probability of the number, it shows curve + histogram. On y-axis probability and on x-axis values. Also called PDF (Probability Density Functions). Also shows data is skewed or not.
3. **Boxplot**, shows the outliers and quartile of the data

## **Bivariant Analysis**

1. **Scatterplot** (Numerical - Numerical), relationship between columns to understand a relationship or trend, which would be impossible to see in almost any other form.
2. **Barplot** (Numerical – Categorical)
3. **Boxplot** (Numerical – Categorical)
4. **Displot** (Numerical – Categorical) you can check the probability of for example, survived not survived with age column.
5. **HeadMap** (Categorical – Categorical)
6. **ClutsterMap** (Categorical – Categorical), shows which value is closer to each other, it creates a dendrogram for relationship.
7. **Pairplot,** shows the scatter plot of each numerical column with others.
8. **Lineplot**, if on x-axis time-based data.
9. **PivotTable** draw table of two columns, based on index.

## **Profiling**

Is the python library which automatically do some of above works. And return HTML interactive page.

# **Feature Engineering**

Preparing dataset for machine learning model

## **Feature Transformer**

If we think a column will not give us well performance, we transform it in some other form by changing it.

### **Missing Values**

Either you fill missing values or remove.

### **Handling Categorical Values:**

By converting the categorical values into numerical

### **Outlier Detection**

### **Feature Scaling**

Converting the feature (columns) numerical values into given range.

Is last step before giving dataset to the model.

#### Standardization

Also called z-score normalization.

Create new column with given column example, age has some value, subtract each value of age from its mean and divide it on the standard deviation. This will give us new column, new columns mean will be = to zero and standard deviation will be = to 1.

In terms of mean your centering, and in terms of scaling you are scaling down the large values, or small values to scale up but to the 1.

Datapoints remains same to actual values, but data become centered and scaled.

Draw displot of actual data and scaled data you will understand.

Outliers behave same with or without standardization.

**We must do** standardization for these algorithm

* K-means (Euclidian distance)
* KNN (Euclidian distance)
* PCA (we need to control variance and do mean centering)
* ANN
* Gradient Decent

**We do not need scaling** for decision tree, random forest, XG boost, GB boost.

#### Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.

Normalization is the process of translating data into the range [0, 1] (or any other range) or simply transforming data onto the unit sphere.

* **Minmax scaling**, actual value – minimum value of column / maximum value of column – minimum value of column. [0 to 1]
* **Mean normalization**, actual value – mean value of column / maximum value of column – minimum value of column. [-1 to 1]. We use where centered data is needed.
* **Max absolute scaling**, where too many zeros we use max absolute scaling, actual value / |max value of column|.
* **Robust scaling**, actual value – median value of column / IQR (75% - 25%). Robust to outliers if data has a lot of outliers.

#### Normalization vs Standardization

Is feature scaling is required?

What to do normalization / standardization?

If we know min and max value of feature such as for CNN, we know for picture resolution 0 to 255, so we can use their max min scaler.

Else use standardization

## **Feature Construction**

Creating new column with given columns.

## **Feature Selection**

Select the main feature from the dataset and gives these features to model so model can perform better.

## **Feature Extraction**

Converting given input features into completely new features. We do feature reduction as well as feature creation.

# **Algorithms**

## **Linear Regression**

Line of equation Y = m \* x + b, here m is slope and b is point on y-axis.

We must draw a line which pass closely to all points for predicting output points, here m is weightage how much x depends on Y.

We can calculate m and b value with two methods.

* Closed form expression also called ordinary least square (OLS) and can find through direct formula. For lower dimension we use direct formula to find m and b.
* Un-closed form expression also called Gradient descent.

### **Simple Linear Regression Using Direct Formula**

If 1 input column and 1 output.

Y=mx+b, here m is slope and b is point on y-axis.

We must draw a line which pass closely to all points.

M is weightage how much x depends on Y.

Need to minimize the distance between points and the line we will draw.

If we assign each point a variable and then square each and sum their distance to find error to minimize the distance.

E = d1 \* 2 + d2 \*2 + dn \* 2.

Taking square means, we can find differentiation on any points.

Error function also called **loss function**.

The distance (difference) is what our model predicts and where it actually point is located. So, (Y actual point – Y predict) and their square, sum of all the distance for each point.

We can write this as below.

**Y predict = m \* x + b**



To minimize the error how would we know that the error is now minimum?

We can find through maxima and minima.

When we are at minimum slope is zero. To find slope we take derivative and make it zero. But we have two values m and b, we will take partial derivative of each from this we will get two equations and with these two we will get m and b values.

**Partial Derivative w.r.t b**

E = = 0

∂E / ∂b= ∂/∂b = 0

= 0

= 0

= 0

= 0

= b

As we have found equation of b, now putting equation of b in below equation

E =

E =

**Partial Derivative w.r.t m**

∂E / ∂m=

=

=

**If dividing -2 on both sides**

=

=

=

**Final Equation**  m =

### **Multiple Linear Regression Using Direct Formula**

= Xb

E = ==

E =

= Xb

E =

E= - Xb -

**Xb == if you calculate it their results will be same.**

E= – 2 (Xb)

Now we calculate minimum by taking partial derivative and making it zero

∂E / ∂b = ∂ / ∂b [ – 2 (Xb) ] = 0

= 0 - 2 X + ∂ / ∂b ) = 0

= - 2 X +

(

(

b= (

**Final equation.**  b= (

### **Polynomial Linear Regression**

If X and Y, then for 2 degrees, will be , , and y. Previous you have just 1 feature now

For higher degree of polynomial, it overfit and for lower its underfit.

### **Regularization**

## **Gradient Decent**

Is an optimization algorithm. We give differentiable function to gradient descent it gives us minimum of that function.

**How would I know I am at minimum?**   
You must find slope at that point to find out. If slope is negative than increment in b value and similarly if slope is positive than decrement in value of b.

To control long or short jump you multiply slop with learning rate, I represent it with alpha, its value range between 0 to 1, 0.0001, 0.1 etc.

α \*

**Where to stop?**

* The first step is, if you know that or close to zero and you are not doing any improvement means you reached to minimum. Or if the is minimum mainly 0.0001.
* Second step is you limit the iteration name as epoch. 1000, or 100 etc.

**Start with random value of b.**

for i in epoch:

α \*

If learning rate is = 0.01

To find slope you know

By taking derivative

=

α \*

**How to calculate m?**

Step 1 initialize random values for m and b as well as learning rate and epochs.

α \*

**What is slope in this m?**

**What if we have local minima and global and we reached local minima?**

**Data also effects on step size, if data is in range or standardized than it will reach faster to the minima.**

### **Batch Gradient Descent**

Will start tomorrow

### **Stochastic Gradient Descent**

Will start tomorrow

### **Mini-Batch Gradient Descent**

Will start tomorrow

# **Regression Matrices**

## **MAE**

Same unit for MAE and output column

Robust for outliers, but it is not differentiable for at zero.

Loss function

## **MSE**

It finds the square distance. We can use it as loss function because it is differentiable. It penalizes the outliers simply not robust to outliers.

Loss function

## **RMSE**

It is differentiable, not robust to outliers.

Loss function

## **R2 Score**

Also called coefficient of determination or goodness of fit.

= Sum of squared error regression line

= Sum of squared error mean line

To calculate you are going to 1 rather 0.

Can be in negative mean is making mistake more than .

R2 score means, this number of score of variance in output column is explained by the input columns

## **Adjusted R2 Score**

In some cases because of adding irrelevant column r2 score not decrease, and stay same, to handle this we use adjusted r2.

n is number of rows, k is number of independent columns

Good for adjusted r2 score.