Ahad Zifain Miyanji

1) Section A

a) A Classification problem involves predicting a catogorical table for a given output based on the learned patterns from the Training data. The goal is to assign inputs to one of the predifined catogories or classes

Key Diffrences from Regression problem

1) Output Type : Classification : The output is ~~catogori~~ catogorical ~~[illegible]~~ variable such as spam or not spam.

Regression: The output is continious variable such as predicting house price or Temprature

2) Evaluation Metrics :

Classification: Performance is often valued using metrics like accuracy, precision recall, F_1 score or the area under ROC curve

Regression: Performance is typically measured using Metrics like mean squared error MSE, Mean Absolute error MAE or R squared.

Three Algorithms

1) Decision Tree
2) Support Vector Machines (SVM)
3) K nearest neighbours (K-nn)

b) In Logistic Regression, the odd ratio represent the ratio of the odds of an event occuring to the odds of it not occuring. It is used to measure the association Between the predictor variable and the outcome.

c) Factor Analysis

It is a statistical method used to describl variability among observed variables in terms of fewer unobserved variables called factors. The goal is to identify the underlying relationships Between the observed Variables

Applications

1) Data Compression: Simplifying Complex data set uncovering hidden patterns and guiding Strategic Decisions making.

i) Feature Extraction: ~~They can~~ It can identify the most significant features in data Set, Helps in identifying distict customer segments by uncovering factors in ~~customers~~ customer behaviour & preferences [customer Segmentation].

Section B

Part A

a)

Diferences

1) Nature of the Data: Time Series Problem: involves sequential data points ordered in Time The objective is often to predict future values based on past observations.

Regression: Problem: Involves predicting a continous target variable based on a set of independent variables. The order of data points is not inherently important.

2) Dependencies: Time Series Problem: Assumes that observation are dependent on the ~~per~~ previous time points, indicating temporal Defiencies.

Regression Problem: Assumes that observation are independent of each other.

Test-Train Split process

Time Series Problem: The data is typically split ~~chronologi~~ chronologically to preserve the time order. The Training set consists of ~~the~~ earlier time periods & the test set consists of later time periods to simulate future predictions

Regression Problem: The data can be split randomly since the order of observation does not affect the outcome. This ensures that both training & Test set represent the same distribution

b) Stationary: A time Series is stationary if its statistical properties such as Mean, variance & autocorrelation are Constant over time.

Importance: It is crucial because many time series modelling techniques like ARIMA assume the time series stationary. Non stationary data can lead to misleading results & poor model performance checking

- Visual Inspection : Plotting the time series to check for constant mean & variance

- Statistical Test : Using test like ADF test to formally check for stationary

Common Test : Augmented Dickey Fuller (ADF)

C) Formating Data object

In Time Series Modelling the date object is usually formatted as 'Datetime' To facilitate time based operations & Analysis.

Eg

```
import pandas as pd

# Example Data string

data_string : "25-07-2024"

# Convert to dateTime object

data_object : pd.to_datatime(data_string,
                  format:"%d-%m-%y")
```

Common Evaluation Metrics

1) Mean Absolute Error

2) Mean Squared Error