SECTION - A
PART - A

a) CLASSIFICATION PROBLEM :

It involves predicting the category or class of a given data point based on its features. This type of problem is common in various fields such as healthcare, finance, marketing and more.

Eg:
- Email spam Detection.
- Disease diagnosis : predicting whether the patient has as specific disease based on symptoms and test results.
- Customer Churn prediction : predicting whether a customer will leave or continue using a service.

Differences between classification & regression problems

1. OUTPUT TYPE :
- Classification : The output is categorical. Eg: 'spam' or 'not spam', 'disease' or 'no disease', 'churn' or 'no churn"
- Regression : The output is continuous. Eg: predicting house prices, tempnature or stock prices.

2. PREDICTION GOAL :
- Classification : The goal is to determine the probability that a given input belongs to a specific class and classify the input accordingly.
- Regression : The goal is to predict a numerical value based on input features, finding a relationship between the ~~values~~ input and output variables to make accurate predictions.

papergrid

Date: / /

Three Algorithms used in Classification :

1. Logistic Regression .

- A statistical model that uses a logistic function to model a binary dependent ~~value~~ variable .
- Eg: predicting the probablity of a customer purchasing a product based on their browsing history .

2. Decision trees :

- A tree like model of decisions and their possible consequences .
- Eg: Classifying loan applicants as 'approved' or 'rejected' based on their credit history, income & other factors .

3. Support Vector Machines (SVM) :

- A powerful classification algorithm that finds the hyperplane that best ~~seats~~ seperates the data into different classes .
- Eg: Classifying images as 'cat' or 'dog' based on pixel values .

papergrid

Date: / /

b) Odds Ratio in Logistic Regression

In logistic regression, the odds ratio is a measure of association between an independent variable and the outcome. It represent how the odds of the outcome change with a one-unit increase in the predictor ~~value~~ variable.

Relation to Model co-efficients:

- The logistic regression model predicts the log-odds of the dependent variable as a linear combination of the independent variables.

- If β_i is the coefficient of the predictor ~~value~~ variable x_i then the ratio is given by e^{β_i}.

- Eg: I $\beta_i = 0.5$ then the odds ratio $e^{0.5} \approx 1.65$. This means that for one unit increase in x_i the odds of the outcome occuring increase by a factor of 1.65 for instance if x_i represents years of education a one-year increase in education might increase the odds of getting a job offere by 1.65 times.

c) Principal component Analysis (PCA) / Factor Analysis Applications.

PCA is a technique used to reduce the dimensionality of a dataset by transforming the original varibles into a new set of uncorrelated variables called principal components. These components are ordered such that the first few retain most of the variation present in the original variables.

papergrid

Date: / /

Steps in PCA:

1. Standardization: Standardize the data to have a mean of zero and a standard deviation of one.
2. Covariance Matrix computation: Calculate the covariance matrix to understand the relationships between the variables.
3. Eigenvalues and Eigenvectors: Compute the eigenvalues and eigenvectors of the covariance matrix to identify the principal components.
4. Principal components: select the top K principal components that explain the most variance and project the data onto the components.

Eg: In a marketing study involving 10 different customer traits, PCA can reduce these traits to a few principal components that capture most of the variance, ~~sg~~ simplifying the analysis.

Factor Analysis.

It is used to ~~analyse~~ identify underlying relationship between variables. It assumes that observed variables are influenced by a smaller number of unobserved variables (factors).

Steps in Factor analysis.

1. Extraction of factors: Identify the number of factors to extract, often using methods like Eigenvalues or scree plots.
2. Rotation: Apply rotation techniques like (Varimax) to make the factors more interpretable.
3. Factor Loadings: Analyse the factors loadings to understand which variables are most strongly associated with each ~~other~~ factor.

papergrid

Date: / /

Eg: In psychological testing, Factor analysis can revel underlying factors like intelligence, anxiety or extraversion that influence responses to different test items.

Application in Buisness Aanalytics:

1. Customer segmentation:
   - PCA: reduces the complexity of customer data by identifying the key component that explaing most of the variance in purchasing behaviour, demographics, etc.
   - Factor Analysis: Identifies underlying factors that influence customer preferences and segments customers based on these factors.

Eg: A retail company use PCA to reduce data from hundreds of products preferences into a few principal components, then uses Factor analysis to identify key purchasing motivations (price sensitivity, brand loyalty).

2. Risk Management.
   - PCA: used in financial markets to reduce the number of risk factors and to identify the principal components that explain the most variance in assets returns.
   - Factor Analysis: Helps in understanding the underlying risk factors affecting investments and can be used to construct diversified portfolios.
   - Eg: A bank uses PCA to identify the main components of credit risk from numerous financial indicators, then uses factor analysis to understand the underlying factors driving these components.

3. Market Research :

• PCA : simplifies survey data by reducing the number of variables while retaining the most important information, making it easier to visualise and analyse.

• Factor Analysis : Identifies key factors that drive customer satisfaction and preferences helping businesses to focus on the most influential aspects

• Eg : A company conducts a customer satisfaction survey with 50 questions. PCA reduces these to a few principal components, and factor analysis reveals that product quality, customer service, and price are the main factors influencing satisfaction.

papergrid

Date: / /

SECTION-B.

PART-A

a) Time series problem

It involves predicting future values based on previously observed values. The data points are ordered in time, and the primary focus is on the temporal dependencies and trends.

Eg: Forecasting stock prices
Predicting weather conditions.
Estimating future sales of a product.

Differences between Time series problem & Regression Problem.

1 Nature of Data:

• Time Series: data points are sequentially ordered over time and order matters. Each observation is dependent on previous observations.

• Regressions: data points are not necessarily ordered. The primary focus is on the relationship between independent variables and dependendent variable without considering the order of observations.

2. Handling Temporal Dependencies:

• Time Series: Modal account for tempora dependencies, seasonality and trends. Techniques like ARIMA, Exponential smoothing and LSTM (for deep learning) are used.

• Regression: Models focus on capturing the relationship between the variables. Techniques like linear regression, Decision Trees, and Random Forests are used.

papergrid

Date: / /

Test- Train split Process :

• Time Series: the test -train split must respect the temporal order of data. typically, earlier data points are used for training and later data points are used for testing. A rolling forecast or walk-forward validation is often employed.

• Regression: The test-train split can be random since the order of data points doesnt matter. Cross validation techniques like k-fold can be used.

b. Stationarity in Time Series Data.

Stationarity:

A time series is considered stationary if its statistical properties, such as mean variance and autocorrelation, do not change over time. Stationarity is crucial for time series modeling because many forecasting methods assume the data is stationary.

Importance of time series modelling:

• Predictibility: they are easier to model and predict because ~~of~~ their properties are constant over time.

• Model assumption: many models like ARIMA assume that the series is statationary. Non stationary data can lead to unreliable and spurious results.

checking Staiontarity:

• Visual Inspection: plotting the time series and looking for trends or seasonaality.

• Statistical Tests: Calculating metrics like autocorrelation and partial autocorrelation.

papergrid

Date: / /

• Differencing : Transforming the data by subtracting previous observations from the current observation to achieve stationarity.

Common Test for stationarity :

· Augmented Dickey - Fuller (ADF) Test : A statistical test that checks for the presence of a unit root in time series sample. The null hypothesis is that the series is non - stationary.

c). Formating Date Objects & Evaluation Metrics in Time Series Modelling

date objects are typically formatted to facilitate time - based indexing and operations.

Converting DD-MM-YYYY to time series datetime object.

Python Example :

```
import pandas as pd

#sample date in DD-MM-YYYY format
date_str = '25-12-2020'

#convert to datetime object
date_obj = pd.to_datetime(date_str, format='%d-%m-%Y')

#output
print(date_obj)
```

Output : '2020-12-25 00:00:00'

Common Evaluation Metrics for time series models.

1. MAE

· measures the average magnitude of errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$$

2. MSE

· measures the average of the squares of the errors

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3 Root Mean Squared Error (RMSE)

· square root of mean squared error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4. Mean absolute percentage error (MAPE):

accuracy of a forecast as percentage.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

· R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$