Riddhi Rungta
2328623
V01107488.

Section A

9) Classification Problem involves predicting which category or class a given set of input data belongs to. The output is a discrete value representing different categories or labels, such as "spam" vs. "not spam", "yes" vs. "no", or different animal types like "cat", "dog", or "bird".

Difference between Classification Problem & Regression Problem

	Classification Problem	Regression Problem
Output Nature	The output is a categorical value. (eg. class labels)	The output is a continuous numerical value. (eg. predicting the price of the house)
Objective	The goal is to categorize the input data into one of the predefined classes based on the features.	The goal is to predict a continuous outcome based on the input features.

Algorithms used in classification:

1) Decision Trees
2) Support Vector Machines (SVM)
3) K- Nearest Neighbors (KNN)

b) Odds Ratio in logistic regression represents the change in the odds of the dependent event occuring for a one-unit change in the predictor variable, while keeping other variables constant. It is obtained by exponentiating the coefficient of the predictor variable. (e^{β_i}).

In logistic Regression, the coefficient (β_i) of a predictor variable indicates the change in the log-odds of the outcome for a one-unit increase in that predictor. The odds ratio is calculated by exponentiating this coefficient (e^{β_i}). An odd ratio greater than 1 indicates an increase in odds, less that 1 indicates a decrease, and equal to 1 indicates no change in odds concerning the predictor variable.

c) Principal Component Analysis is a technique for reducing the number of variables in a dataset by transforming the original variables into a smaller set of uncorrelated variables called principal components. These principal components capture the maximum variance in the data with the fewest components.

factor Analysis is used to identify underlying relationships between observed variables, assuming these observed variables are influenced by fewer unobserved variables called factors. The goal is to reveal the ~~talent~~ latent structure in the data.

Application in Business Analytics:

1) PCA and factor Analysis can reduce the dimensionality of large datasets (data compression), making it easier to visualise and analyse data without significant information loss.

2) Feature Extraction - These techniques help identify the most important features or components that explain most of the variance in the data, improving model performance & interpretability.

3) In market research, factor analysis can identify latent factors influencing customer preferences & behaviours, ~~addi~~ aiding in segmentation & targeting strategies.

Section B

(a)

Category	Time Series Problem	Regression Problem
Temporal Dependency	Data Points are sequential and depend on time, meaning each observation is influenced by previous ones. The objective is to predict future values based on historical data.	The goal is to predict a continuous outcome from a set of independent variables, with no inherent order or temporal dependency among the data points.
Test-Train Split Process	The test-train split must maintain the chronological order of the data to prevent data leakage. The training set comprises earlier time periods, & the test includes later time periods	The train-test split can be random since there is no temporal order, allowing the data to be shuffled.

b A time series is stationary if its statistical properties like mean, variance, and autocorrelation remain constant over time.

Stationarity is vital for time series modeling because many forecasting techniques assume the data is stationery, leading to more accurate & interpretable models.

→ Methods to check Stationary.

1) Visual Inspection: Plotting the time series to identify obvious trends or seasonal patterns.
2) Statistical Tests: Using statistical tests to verify Stationery.

Common test is Augmented Dickey Fuller (ADf) Test. This test is commonly used to check if a time series is stationary. It tests the null hypothesis that a unit root exists, indicating non-stationarity. A p-value below a certain threshold (eg: 0.05) means the null hypothesis is rejected, & the series is considered Stationary.

c) Date object is formatted in time series modelling by converting to a datetime format that is recognizable and manipulable by the software.

eg: DD-MM-YYYY to Datetime Object:

```
import pandas as pd
date_series = pd.series(['31-07-2024', '01-08-2024'])
date_series = pd.to_datetime(date_series, format="%d-%m-%Y")
```

→ Common metrics to evaluate time series models are:

1) Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions, ignoring their direction.
2) Mean Squared Error (MSE): Measures the average of the squared errors, penalizing larger errors more than MAE.
3) Root Mean Squared Error (RMSE)
4) Mean Absolute Percentage Error (MAPE)*