(SCMA 632) - (Statistical Modelling & Analysis) - FINAL EXAM

SATHWIK NAG CHANNAKESHI VENKATESH
2328626
V01107764

Section-A

Part-A :-

a) **Classification Problem** :- It is a type of supervised learning where the output variable is a category / class.

→ The goal is to accurately predict the category for each case in the data by learning from a training dataset of cases that already have category labels.

Key difference from Regression :-

	Classification	Regression
Output type →	Output variable is categorical Eg:- Yes/No, Red/Blue/Yellow, Male/Female etc	In regression output variable is continuous Eg: Height/Weight, Price, Temperature etc
Objective →	Predicts the belonging to a class / category (involving a decision boundary between classes.	Involves predicting a continuous value & aims to find the best line of fit through the data.

Algorithms to solve a Classification problem.

(i) Decision Tree (ii) Random Forest (iii) Support Vector Machines (SVM)

b) Odds ratio: In logistic regression, the odds ratio for a particular (e^{β}) coefficient represents how the odds of the dependent variable change with a one unit change in the corresponding independent variable (given all the other variables are held constant).

→ Determines / Quantifies the strength of association between a feature & the outcome.

If β → coefficient in logistic regression
e^{β} → Odds ratio

Eg: If coefficient $\beta = 2$ then Odds ratio $= e^2 \approx 7.39$, which indicates a strong positive effect on the odds.

c) Principal Component Analysis (PCA):-

PCA is a dimensionality reduction technique that transforms a large set of variables into a smaller one that still contains most of the information in the large set.

→ This is done by identifying the principal components that capture the maximum variance in the data.

→ Application in DA :- → To reduce dimensionality of large datasets
→ Improve visualizations.
→ Discover patterns in data.
→ Improve ML model performance by eliminating multicollinearity.

Factor Analysis: It is used to identifying latent variables that are not directly observed but are inferred from the correlations among observed variables.

(3)

→ It reduces the observed variables into a few latent factors.

Application in BA:- → Used in customer segmentation,
→ Risk management,
→ Market research
→ Identify underlying relationships between variables that can explain patterns in customer behaviour or preferences.

Section B - Part A

a)

	Time-Series	Regression
Nature of Data:-	Sequential & indexed by time ∴ they are inherently ordered	Here independent observations where the order of data does not matter.
Objective:-	Involves forecasting future values based on past data Eg: Seasonality of sales, trends & cycles.	Predict data/Dependent variable value without explicitly accounting for the order of data.
Test-train Split :-	→ Split is sequential → Training data consists of initial segments of time series → Testing data is subsequent segments. → Preserves temporal order & the model is validated on unseen future data.	→ Split is random → Datasets are assigned randomly to either training or testing set. → It is assumed that observations are independent of each other.

(4)

b) Stationarity → It means that statistical properties such as mean, variance, & autocorrelation are constant over time.

→ It implies that the time series does not exhibit trends or seasonality.

Importance in T.S modelling :-

→ Predictive modeling → Stationary data are easier to model & predict because future values have same properties as past.

→ Stability of time-series characteristics are important because most of the statistical modelling methods assume stationarity.

Checking Stationarity :- * Usually inspect by plotting time-series & look for consistent patterns in mean & variance.

* Apply statistical tests that can quantitatively determine whether a series is stationary.

i.e mean, SD, covariance are not a function of time.

ADF test (Augmented Dicky-Fuller test)

↳ checks for a unit root in T-S.

) H_0: TS can be represented by a unit root process (is non-stationary).

) H_A: Stationary

c) Date objects should be formatted to determine datatype that explicitly represents both Date & Time components.

(5)

* Convert DD-MM-YYYY to DateTime Object:-

→ Using pandas (Python).. import pandas as pd.

date_series = pd.to_datetime(date_series, format = '%d-%m-%Y')

* Common evaluation metrics:-

→ Root Mean Squared Error (RMSE)

↳ It is the measure of magnitude of error.

→ Mean Square error (MSE):-

↳ Average of squares of errors.

→ Mean Absolute Error (MAE):-

↳ Average magnitude of error (without direction considered)