URBAN EDGE

SCMA 632, Final Exam

Tony Balaseran

Section A

Part A

a) A classification problem involves predicting a categorical outcome, such as determining whether an email is spam or not, based on the given input data. A regression problem, on the other hand, involves predicting a continous outcome, such as predicting house prices based on features like size, location, etc.

Key differences:

1. Output Nature: Classification outputs discrete labels (classes), while regression outputs continuous values.

2. Model Evaluation: Classification outputs discrete labels (classes), while

URBAN EDGE

Classification often uses metrics like accuracy, precision, recall, and F1-score, whereas regression uses metrics like mean squared error (MSE), mean absolute error (MAE), and R-squared.

- Algorithms for classification.

1. Logistic regression
2. Decision Trees
3. Support Vector Machines (SVM)

6. Odds Ratio in Logistic Regression:

- The odds ratio in logistic regression represents the change in the odds of the outcome occurring for a one-unit increase in the predictor variable, holding all other variable constant. It is calculated by exponentiating the coefficient of the predictor (Exp(B)).

c. Principal Component Analysis

PCA is a statistical technique used to simplify a dataset by reducing its number of dimensions without losing much information. It does this by transforming the original variables, the principal components, which are orthogonal (uncorrelated) and ordered so that the first few retain most of the variation present in the original dataset.

Applications in Business Analytics

- Customer Segmentation
- Market Basket Analysis
- Financial Modeling
- Operational Efficiency
- Fraud detection:

URBAN EDGE

Section. B

Part. A

(a) A Time Series problem involves predicting future values based on past observations, where the data points are time-ordered. The key aspect is the temporal structure and potential autocorrelation within the data.
Whereas A regression problem predicts a continous outcome variable based on one or more predictor variables, without necessarily considering the order of the observations.

Test-train Split in Time Series Vs Regression

- In a typical regression problem, data is randomly split into training and testing sets, as the order of data does not affect the outcome.

- In a time series problem, the data is split based on time. The earlier part of the data is used for training, and the latter part is used for testing to maintain the temporal sequence and prevent data leakage.

URBAN EDGE

b. Stationarity in Time Series Data:

- Stationarity: A time series is stationary if its statistical properties (mean, variance, autocorrelation) do not change over time. Stationarity is crucial because many time series forecasting methods assume the series is stationary.

- Stationarity simplifies the analysis and forecasting process. It ensures that the models parameters are consistent over time, making predictions more reliable.

- Stationarity can be checked using visualizations (like line plots and ACF plots) and statistical tests.

- Common Test: The Augmented Dickey-Fuller (ADF) test is commonly used to test for stationarity. It checks for the presence of a unit root in the series

URBAN EDGE

6. Formatting and converting Date objects in Time Series Modelling.

- Formatting Date objects: Dates in time series data are often formatted in a consistent format, such as "DD-MM-YYYY" or "YYYY-MM-DD". They are converted to ~~determine~~ datetime objects for time-based operations.

- Conversion to datetime objects: in Python, you can convert a date in "DD-MM-YYYY" format to a datetime object using the 'pd.to_datetime' function from the pandas ~~to~~ library, specifying the formats.

Common evaluation metrics for time series data.

- ~~MF~~ Mean Absolute Error
- Mean Squared Error
- Root Mean Squared Error
- Mean absolute percentage Error
- R-Squared (Coefficient of determination)