VIJAYATHITHYAN B.B
V01107268.

29/07/24.

SCMA632 : Final Exam

SECTION : A :
PART : A :

Q.a. A classification problem involving predicting a categorical label based on input features, while a regression problem predicts a continous value. Key differences are in classification problem is for discrete categories where as regression is for continous values. classification is evaluated based on accuracy, precision, recall, F1 score, AUC. Regression is evaluated based on Mean square error (MSE), mean absolute ~~a~~ error (MAE), R-square.

Three Classification Algorithms are:

1. Logistic regression
2. Support Vector Machines (SVM)
3. Random Forest.

Q.b. In Logistic regression, the Odd ratio represents the ratio of the odds of an event occuring to the odds of not occuring It quantifies the change in odds resulting from a one-unit change in a predictor variable, holding all other variables constant.

Q.c Pr
re
Cor
unc
The
the
its
Var
wh
use
key

Q.d. In
Seq
pre
poin
and
seri
regr
and

Q.C Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the original correlated variables into a smaller set of uncorrelated variables called principal components. These components capture the maximum variance in the data, simplifying the dataset while retaining its essential patterns. PCA reduce the number of variables, simplifying models and visualizations while maintaining critical information. This is useful in areas like marketing analytics to identify key factors driving customer behaviour.

SECTION : B :

PART: A :

Q.A. In time-series problem the data points are sequentially ordered and often depend on previous values. In regression problem, data points are generally independent of each other and do not have an inherent order. In time-series, data is split based on time. In regression data is split randomly into training and test data sets.

Q.b. Stationarity implies that the statistical properties of time series (mean, variance, autocorrelation) are constant over time. Many time-series models (ARIMA) assume stationarity. Non-stationary data can lead to unreliable and spurious ~~data~~, results. Stationarity can be checked by plotting the data to check constant mean and variance over time; and by using formal tests to assess stationarity. Augmented Dickey-Fuller (ADF) test, a hypothesis test is used to determine if a time series is stationary.

Q.C. If the date is in DD-MM-YYYY formate, converting it to a datetime object is necessary for time-series analysis. For which the ~~pto~~ python code,

```
import pandas as pd
df['date'] = pd.to_datetime(df['date'], format=
                '%d-%m-%y')
```

is used. The time-series models are commonly evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute ~~Error~~ Percentage ERROR (MAPE).