

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech.

SEM: V - THEORY EXAMINATION (2022 - 2023)

Subject: Predictive Analytics

Time: 3 Hours

Max. Marks: 100

General Instructions:

IMP: Verify that you have received the question paper with the correct course, code, branch etc.

1. This Question paper comprises of three Sections -A, B, & C. It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.
2. Maximum marks for each question are indicated on right -hand side of each question.
3. Illustrate your answers with neat sketches wherever necessary.
4. Assume suitable data if necessary.
5. Preferably, write the answers in sequential order.
6. No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.

SECTION A

20

1. Attempt all parts:-

- | | | |
|------|--|---|
| 1 | A regression analysis is inappropriate when (CO1) | 1 |
| | (a) you have two variables that are measured on an interval or ratio scale. | |
| | (b) you want to make predictions for one variable based on information about another variable. | |
| | (c) the pattern of data points forms a reasonably straight line | |
| | (d) there is heteroscedasticity in the scatter plot. | |
| 1 | Least square method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the _____ deviations. (CO1) | 1 |
| | (a) Vertical | |
| | (b) Horizontal | |
| | (c) Both of these | |
| | (d) None of these | |
| 1-c. | _____ is rapidly being adopted for computing descriptive and query types of analytics on Big data. (CO2) | 1 |

- (a) EDR
- (b) Hadoop
- (c) Azure
- (d) InfoSight

- 1-d. When the error terms have a constant variance, a plot of the residuals versus the independent variable x has a pattern that (CO2) 1
- (a) fans out
 - (b) funnels in
 - (c) fans out, but then funnels in
 - (d) forms a horizontal band pattern
- 1-e. Logistic Regression is a Machine Learning algorithm that is used to predict the probability of a ____ (CO3) 1
- (a) categorical independent variable
 - (b) categorical dependent variable.
 - (c) numerical dependent variable
 - (d) numerical independent variable
- 1-f. Which of the following is used where the target variable is of categorical nature? (CO3) 1
- (a) Keras
 - (b) Knime
 - (c) Logistic Regression
 - (d) MXNet
- 1-g. What does autocovariance measure? (CO4) 1
- (a) Linear dependence between multiple points on the different series observed at different times
 - (b) Quadratic dependence between two points on the same series observed at different times
 - (c) Linear dependence between two points on different series observed at same time
 - (d) Linear dependence between two points on the same series observed at different times
- 1-h. Which of the following cross validation techniques is better suited for time series data? (CO4) 1

- (a) k-Fold Cross Validation
- (b) Leave-one-out Cross Validation
- (c) Stratified Shuffle Split Cross Validation
- (d) Forward Chaining Cross Validation

1	Which of the following is characteristic of exploratory graph? (CO5)	1
	(a) Made slowly	
	(b) Axes are not cleaned up	
	(c) Color is used for personal information	
	(d) All of the mentioned	
1	How to handle missing or corrupted data in a dataset? (CO5)	1
	(a) Drop missing rows or columns	
	(b) Replace missing values with mean/median/mode	
	(c) Assign a unique category to missing values	
	(d) All of the above	

2. Attempt all parts:-

2.a.	List out the critical assumptions of linear regression (CO1)	2
2.b.	What are the applications of predictive model? (CO2)	2
2.c.	List two examples of Multiple logistic regression. (CO3)	2
2.d.	Forecasts are frequently produced for different horizons. Give examples of their use (CO4)	2
2.e.	Explain how overfitting happens during model training? (CO5)	2

SECTION B

30

3. Answer any five of the following:-

3-a.	Explain the method of maximum likelihood in detail (CO1)	6
3-b.	Does correlation and dependency mean the same thing? If two events have correlation of zero, does this convey they are not dependent and vice-versa? Comment. (CO1)	6
3-c.	Explain the significance of Cost Function in Regression Graph (CO2)	6
3-d.	What do you understand by K-fold Cross-Validation? Explain using diagram about 5 fold cross validation (CO2)	6
3.e.	Is the decision boundary Linear or Non-linear in the case of a Logistic Regression model? Explain (CO3)	6
3.f.	How to use autoregressive integrated moving average (ARIMA) with seasonal series? (CO4)	6

3.g. Explain in detail about Feature Encoding Techniques (CO5) 6

SECTION C 50

4. Answer any one of the following:-

4-a. Consider the simple linear regression model $y = 10 + 25x + \varepsilon$ where the random error term is normally and independently distributed with mean zero and standard

deviation 2. Generate a sample of eight observations, one each at the levels $x = 10, 12, 14, 16, 18, 20, 22$, and 24.

(a) Fit the linear regression model by least squares and find the estimates of the slope and intercept.

(b) Find the estimate of σ^2 .

(c) Find the standard errors of the slope and intercept.

(d) Now generate a sample of 16 observations, two each at the same levels of x used previously. Fit the model using least squares.

(e) Find the estimate of σ^2 for the new model in part (d). Compare this to the estimate obtained in part (b).

(f) Find the standard errors of the slope and intercept using the new model from part (d). Compare these standard errors to the ones that you found in part (c). What impact has the increase in sample size had on the estimated standard errors? (CO1)

4-b. Explain the differences between Random Forest and Gradient Boosting machines. (CO1) 10

5. Answer any one of the following:-

5-a. What do you understand by Regularization Techniques? What is Shrinkage method and highlight the relationship between the target and available independent variables? (CO2) 10

5-b. What do you understand by overfitting, underfitting, good fit in regression? Support your answer along with example of any high variance algorithm? (CO2) 10

6. Answer any one of the following:-

6-a. How to interpret the results of a logistic regression model? What are the meanings of alpha and beta in a logistic regression model? (CO3) 10

6-b. Why can't linear regression be used in place of logistic regression for binary classification? (CO3) 10

7. Answer any one of the following:-

7-a. The sales of a company in million of rupees for the year 1994 - 2001 are given below 10

Year	1994	1995	1996	1997	1998	1999	2000	2001
Sales	550	560	555	585	540	525	545	585

- (i) Find the linear trend equation
- (ii) Estimate the sales for the year 1993
- (iii) Find the slope of the straight line trend
- (iv) Do the figures show a rising trend or a falling trend? (CO4)

7-b. Can Non-Sequential Deep Learning Models outperform Sequential Models in Time-Series Forecasting? 10
 Compare some Forecasting Techniques for Stationary and Non-stationary Time-Series (CO4)

8. Answer any one of the following:-

8 How can we determine whether the correlation is positive or negative and also comment on its degree or extent? (CO5) 10

8 Why is normalisation required before applying any machine learning model? What module can be used to perform normalisation? (CO5) 10