SRN | | | | | | | | | | | | |

PES UNIVERSITY	**PES University, Bengaluru** (Established under Karnataka Act No. 16 of 2013)	**UE20CS905**

: END SEMESTER ASSESSMENT (ESA)
M TECH DATA SCIENCE AND MACHINE LEARNING_ SEMESTER I

UE20CS905 - MACHINE LEARNING - I

Time: 3 Hrs	Answer All Questions	Max Marks: 100

Instructions

1. Answer all the questions.
2. Section A should be handwritten in the answer script provided and signed at the end of the same.
3. Section B and C are coding questions which have to be answered in the system and uploaded in Olympus Login.
4. Avoid using GridSearchCV as it might impact the system performance drastically.

Section A (20 marks)			
1	a)	What is Machine learning? State any two types of machine learning.	4
	b)	How can you handle overfitting and underfitting?	4
	c)	State the assumptions of linear regression algorithm.	4
	d)	If y = 2x1 + 12x2 + 3x3 + 5 is the linear regression equation, then explain how the coefficients of x1 and x2 affect the value of y.	4
	e)	Explain any two of the data preprocessing steps.	4
Section B (40 Marks)			
2	a)	DATA DESCRIPTION: Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. This dataset describes the listing activity and metrics in NYC, NY for 2019. ● id = listing ID ● name = name of the listing ● host_id = host ID ● host_name = name of the host ● neighbourhood_group = location ● neighbourhood = area ● latitude = latitude coordinates ● longitude = longitude coordinates ● room_type = listing space type ● price = price in dollars ● minimum_nights = amount of nights minimum ● number_of_reviews = number of reviews ● last_review = latest review	25

SRN | | | | | | | | | | | | | |

		● reviews_per_month = number of reviews per month ● calculated_host_listings_count = amount of listing per host ● availability_365 = number of days when listing is available for booking Perform Exploratory data analysis and summarize important observations from the data set Some pointers which would help you, but don't be limited by these i. Explore the types of variables (continuous, categorical etc.) (3 marks) ii. Calculate five point summary for numerical variables (2 marks) iii. Summarize observations for categorical variables – no. of categories, % observations in each category (5 marks) iv. Check for defects in the data. Perform necessary actions to 'fix' these defects. (10 marks) check for numerical/ categorical variable(wrong representation) missing values, outlier treatment, skewness, encoding v. Summarize relationships among variables. (5 marks)	
	b)	Fit a base model. Please write your key observations i. What is the overall R2? Please comment on whether it is good or not. (2 marks) ii. What is the adjusted R2? Is it different from R2? Why? (3 marks) iii. Which variables are significant? (4 marks) iv. Is there multicollinearity? (4 marks) v. Which other key model output parameters do you want to look at? (2 marks)	15
		Section C (40 marks)	
3	a)	How do you improve the accuracy of the model? Write clearly the changes that you will make before re-fitting the model. Fit the final model. Feature Engineering / Feature Selection Regularization Cross Validation Please feel free to have any number of iterations to get to the final answer. Marks are awarded based on the quality of the final model you are able to achieve.	20
	b)	Summarize as follows i. Summarize the overall fit of the model and list down the measures to prove that it is a good model ii.Write down a business interpretation/explanation of the model – which variables are affecting the target the most and explain the	20

		relationship. Feel free to use charts or graphs to explain. iii. What changes from the base model had the most effect on model performance iv.What are the key risks to your results and interpretation? Justification for selecting a model	