**MOCK TEST II**

**MCQs**

You are given a multiple linear regression model: Y=β0+β1x1+β2x2+β3x3

The null hypothesis states that the variable is insignificant. Thus, if we fail to reject the null hypothesis, you can say that the predictor is insignificant.

For example, if you fail to reject the null hypothesis for x1, you can say that x1 is insignificant. This would also imply that the coefficient for x1, i.e., β1 = 0.

In other words, the null hypothesis tests if the predictor's coefficient, i.e., βi = 0. If the null hypothesis is rejected, then βi≠0.

Answer to Question 1 and 2 are related to above content.

Question 1

If β1=β2=0 holds and β3 = 0 fails to hold, then what can you conclude?

1. There is a high correlation between x1 and x2.
2. There is a linear relationship between the outcome variable(Y) and x3.
3. There is a linear relationship between the outcome variable and x1, x2.

Question 2

If β1  = β2 = β3 = 0 holds true, what can you conclude?

1. There is no linear relationship between y and any of the three independent variables.
2. There is a linear relationship between y and all of the three independent variables.
3. There is linear relationship between x1, x2 and x3.

Question 3

Suppose you need to build a model on a dataset that contains 2 categorical variables with 2 and 4 levels, respectively. How many dummy variables should you create for model building?

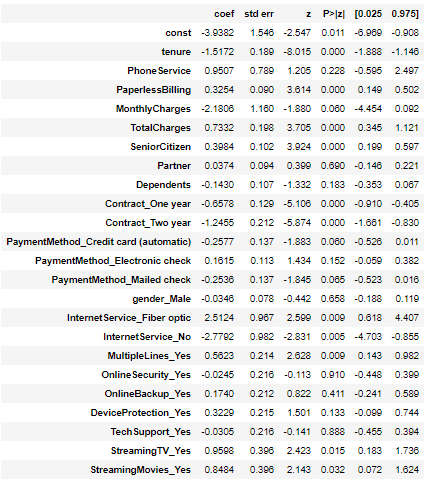
1. 4
2. 5
3. 6
4. 8

Question 4

In a dataset with mean 50 and standard deviation 12, what will be the value of a variable with an initial value of 20 after you standardise it?

1. 1.9
2. -1.9
3. 2.5
4. -2.5

Question 5



Which of the following variables are negatively correlated with the target variable based on the summary statistics report given above? (More than one option may be correct.)

1. Tenure
2. TotalCharges
3. MonthlyCharges
4. TechSupport\_Yes

**Subjective Questions**

1. To do text analytics, we need to clean it . There are three kinds of words present in any text corpus. What are they and give two reasons why they must be removed?
2. In NLTK, you have different types of tokenisers present that you can use in different applications. Explain briefly what are they and why one should use it?
3. Why can’t linear regression be used in place of logistic regression for binary classification?
4. Developing hypotheses will be a key part of your job role as a data scientist when you're working on real-world problems. You need to bring all your domain knowledge to the forefront and try to identify the potential root causes of the given problem. Your question is “What factors contribute most significantly to customer churn in a subscription-based streaming service?" (For ex:Netflix, Amazon Prime etc)
5. ROC stands for Receiver Operating Characteristic curve. This name has emerged from the domain of electrical engineering around the 2nd World War when electrical and radar engineers used such curve to detect enemy planes. Since then, this concept has found its application in many fields, machine learning being the latest one.

"What is the significance of the ROC curve in Logistic Regression, and how does it help in evaluating the model's performance?"

**Coding Problems**

1. **Decision Tree - Bank Marketing Dataset**

**Description**

You are given the 'Portuguese Bank' marketing dataset which contains data about a telemarketing campaign run by the bank to sell a product (term deposit - a type of investment product).  
  
Each row represents a 'prospect' to whom phone calls were made to sell the product. There are various attributes describing the prospects, such as age, profession, education level, previous loans taken by the person etc. Finally, the **target variable is**'purchased' (1/0), 1 indicating that the person had purchased the product. A sample of the training data is attached below (note that 'id' shouldn't be used to train the model) :



As an analyst, you want to predict whether a person will purchase the product or not. This will help the bank reduce their marketing costs since one can then target only the prospects who are likely to buy. **Build a decision tree with default hyperparameters**to predict whether a person will buy the product or not. You have to write the predictions in the file *bank\_predictions.csv* in the following format (note the column names carefully)

bank\_predicted id

0 2041

1 399

0 1400

0 3709

1 2111

1. **Clustering KMeans**

**Description:**

Given below is a data set on the education status of Indian states.



Which parameters do you think are the most important for segmenting the states? How did you decide this? How will you check if the segmenting is good or whether you need to use different factors for segmenting? How are the clusters different when we have not scaled compared to clusters formed after scaling?