MOCK TEST II

MCQs

You are given a multiple linear regression model: $Y=\beta0+\beta1x1+\beta2x2+\beta3x3$
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., $\beta1 = 0$.
In other words, the null hypothesis tests if the predictor's coefficient, i.e., $\beta i = 0$.
If the null hypothesis is rejected, then $\beta i \neq 0$.
Answer to Question 1 and 2 are related to above content.

Question 1
If $\beta1=\beta2=0$ holds and $\beta3 = 0$ fails to hold, then what can you conclude?

A. There is a high correlation between x1 and x2.
B. There is a linear relationship between the outcome variable(Y) and x3.
C. There is a linear relationship between the outcome variable and x1, x2.

Question 2
If $\beta1 = \beta2 = \beta3 = 0$ holds true, what can you conclude?

A. There is no linear relationship between y and any of the three independent variables.
B. There is a linear relationship between y and all of the three independent variables.
C. 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?

A. 4
B. 5
C. 6
D. 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?

A. 1.9
B. -1.9
C. 2.5
D. -2.5

Question 5

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.9382	1.546	-2.547	0.011	-6.969	-0.908
tenure	-1.5172	0.189	-8.015	0.000	-1.888	-1.146
PhoneService	0.9507	0.789	1.205	0.228	-0.595	2.497
PaperlessBilling	0.3254	0.090	3.614	0.000	0.149	0.502
MonthlyCharges	-2.1806	1.160	-1.880	0.060	-4.454	0.092
TotalCharges	0.7332	0.198	3.705	0.000	0.345	1.121
SeniorCitizen	0.3984	0.102	3.924	0.000	0.199	0.597
Partner	0.0374	0.094	0.399	0.690	-0.146	0.221
Dependents	-0.1430	0.107	-1.332	0.183	-0.353	0.067
Contract_One year	-0.6578	0.129	-5.106	0.000	-0.910	-0.405
Contract_Two year	-1.2455	0.212	-5.874	0.000	-1.661	-0.830
PaymentMethod_Credit card (automatic)	-0.2577	0.137	-1.883	0.060	-0.526	0.011
PaymentMethod_Electronic check	0.1615	0.113	1.434	0.152	-0.059	0.382
PaymentMethod_Mailed check	-0.2536	0.137	-1.845	0.065	-0.523	0.016
gender_Male	-0.0346	0.078	-0.442	0.658	-0.188	0.119
InternetService_Fiber optic	2.5124	0.967	2.599	0.009	0.618	4.407
InternetService_No	-2.7792	0.982	-2.831	0.005	-4.703	-0.855
MultipleLines_Yes	0.5623	0.214	2.628	0.009	0.143	0.982
OnlineSecurity_Yes	-0.0245	0.216	-0.113	0.910	-0.448	0.399
OnlineBackup_Yes	0.1740	0.212	0.822	0.411	-0.241	0.589
DeviceProtection_Yes	0.3229	0.215	1.501	0.133	-0.099	0.744
TechSupport_Yes	-0.0305	0.216	-0.141	0.888	-0.455	0.394
StreamingTV_Yes	0.9598	0.396	2.423	0.015	0.183	1.736
StreamingMovies_Yes	0.8484	0.396	2.143	0.032	0.072	1.624

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.)

A. Tenure
B. TotalCharges
C. MonthlyCharges
D. 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?

   Answer:

- In any text corpus, three main types of words often require cleaning: Stop words, Noise words, and Rare words.
- Stop words are common words like "the," "is," and "and" that do not carry significant meaning.
- Noise words may include slang, misspellings, or domain-specific jargon that does not add useful information.
- Rare words appear very infrequently and often do not contribute to general patterns in the data.
- Reasons to remove them:
- Removing these words helps reduce dimensionality and improve processing efficiency in text analytics tasks.
- It improves model accuracy by focusing on meaningful words, reducing noise that could distort insights or predictions.

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?

   Answer:

   - NLTK offers various tokenizers for splitting text into tokens, each serving specific purposes:
   - Word Tokenizer: Splits text into individual words, useful for applications that require word-level analysis.
   - Sentence Tokenizer: Divides text into sentences, helpful for analyzing sentence-level sentiment or summarization.
   - Reg exp Tokenizer: Allows custom tokenization based on regular expressions, useful for complex tokenization tasks where specific patterns need to be extracted.
   - Tweet Tokenizer: Specially designed to handle social media text, such as Twitter, accounting for hashtags, mentions, and emojis.
   - Why use them?
   - They are optimized for specific tasks, providing flexibility to handle various text structures.
   - They help standardize and preprocess text data efficiently according to the requirements of different NLP applications.

3. Why can't linear regression be used in place of logistic regression for binary classification?

   Answer:

   - Linear regression is unsuitable for binary classification because it predicts continuous values, which may fall outside the 0-1 range needed for binary outcomes. In contrast:
   - Logistic regression applies a logistic function to restrict the output between 0 and 1, making it suitable for binary probabilities.
   - Using linear regression for binary classification could result in unbounded predictions, leading to inaccurate classifications and ineffective probability interpretation

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.)
   Answer:
   - Factors that could significantly contribute to customer churn in a subscription-based streaming service include:
   - Content relevance: Lack of new or diverse content can drive customers to other platforms.
   - Pricing: High subscription costs or unexpected fee increases can make customers more likely to cancel.
   - User experience: Poor interface design, slow streaming speeds, or frequent technical issues could negatively impact the user experience.
   - Customer engagement: Low engagement due to irrelevant recommendations or lack of personalized content could indicate potential churn.
   - Competitor offerings: If competitors offer more compelling content or better deals, customers might switch to those alternatives.

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?"

   Answer:
   - The ROC (Receiver Operating Characteristic) curve is crucial in logistic regression and other classification models as it visually assesses the model's ability to distinguish between classes (True Positives and False Positives).
   - The Area Under the Curve (AUC) measures the model's performance. A higher AUC indicates a better model, with an AUC of 1 representing perfect classification and 0.5 indicating random guessing.
   - The ROC curve helps select an optimal threshold for classification by showing the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate), allowing the model to be tuned based on the desired balance between accuracy and recall.

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) :

bank_train.csv

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

2. Clustering K-Means

Description:

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

IndianStatesEdu.xlsx

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?