MN50572 Coursework 2024

Data Analytics Report

Word Limit 2500 (±10%)

A social media company, Z, wants to better understand their user-base and have obtained behavioural data from their platform. You have been hired as a data science consultant to help them understand what insights these data reveal about their customers.

These data contain the following variables/features:

ID	user identifier
InDegree	total number of unique network neighbours replying to (or quoting) a user
OutDegree	total number of unique network neighbours receiving posts from (or being quoted by) a user
TotalPosts	total number of posts of a user
MeanWordCount	mean average word count for all user's posts
LikeRate	mean average number of likes per post. Calculated as: total number of likes received / total number of posts made
PercentQuestions	percentage of a user's posts that contain question marks (excluding within URLs)
PercentURLs	percentage of a user's posts that contain URLs
MeanPostsPerThread	total number of posts / number of threads participated in
InitiationRatio	number of threads initiated / number of threads participated in
MeanPostsPerSubForum	total number of posts / number of sub forums participated in
PerBuNeighbours	number of neighbours a user has both received posts from and posted replies to / total number of unique network neighbours
AccountAge	age of account in months

Your Tasks

1. Demonstrate you have explored, processed, and understood these data (30%)
2. Conduct a cluster analysis to understand the user base (30%)
    a. This can be any model of your choice, but have clear rationale behind choices made including but not limited to: data cleaning, data transformation, the generation of new variables or exclusion of others.
    b. Interpret these clusters and explain what they mean and tell you about the users.
    c. Include model evaluation and how you came to the numbers of clusters you did and interpretations. Consider what this means for the client.
3. Train ***a*** supervised learning model to predict the clusters you've identified (30%)
    a. This can be any model of your choice, but have clear rationale behind choices made including but not limited to: data cleaning, data transformation, the generation of new variables or exclusion of others.
    b. Provide insights into the evaluation of the model.
    c. Consider how predictive model could be helpful for the client.

Please present your findings as a report, which constitutes as the final 10% of your marks in terms of style, writing, and clarity in reporting. This needs to be a well written report in the style you'd present to a client assuming they have little technical knowledge and need clarification on the methods and approaches you're taking. Think carefully about what information is important to include and the sections you will present. Ensure this is well presented (e.g., ensure you only include necessary information in text). Remember, you can use an appendix but there are not marks allocated there, the report should be understandable and clear, with clear recommendations for the client (e.g., new approaches to expand user-bases, how to retain users, etc.)

Submit your report as a PDF with associated scripts (e.g., R, Python).

Due Wednesday 17th April 2024 at 2pm UK time.