Data Mining Project FAQs

Project Problem Part -1 - Clustering

If Elbow curve and silhouette_score both give different results for a number of clusters then which method we should use?

Actually, that can't be the case. Then one should choose the method based on the business problem, identify what can be an optimal cluster that any company would want in accordance with the associated problem, then try to match with Elbow curve or silhouette score. This will give you an answer. But the very important thing is the elbow curve and silhouette score have to match at any point in time. Also, check if you are using the same configuration/parameters in the KMeans algorithm for the calculation of Silhouette Scores and plotting Elbow Plot. It should give the same results for the same config.

Where to cut a dendrogram?

There is no one answer to this. You have to take a call looking at different metrics. Try to look at silhouette plots or maybe looking at the dendrogram and use the knowledge of the data and the clusters to make sure that the clusters are heterogeneous. That way you can optimize the same.

"I have a broad question regarding the 'Optimal Number' of clusters, referenced in questions 1.3 and 1.4. Is the assessment of optimal number a subjective decision, or is there 1 clear number in both these cases?

- **I ask because, for both hierarchical and k-means, I'm unable to form a clear assessment of what's optimal. That is, two options seem viable:**
- **In hierarchical: the dendrogram suggests an optimal number based on the colors that it assigns to clusters. At the same time, based on the distance criteria, I can form a different impression on what's optimal.**
- **In k-means: the elbow approach suggests one option, but the corresponding silhouette score isn't necessarily the highest. Can you please advise?**

To understand the optimal number of clusters based on hierarchical clustering, the color combination that comes from the dendrogram is just for representation purposes. Also based on the distance criteria, there is no 1 number that we can decide for the optimal cluster. The best way to understand is what is difference between the distances that are calculated for each cluster and is the difference in distance huge enough to consider the noise in the data and also the concept of intracluster distance is maximized. If this question is answered, then you shall get the optimum level of clusters. Now coming to the k-means approach, the silhouette score also defines the same concept where it tells us what is the difference between each cluster and is a difference in distance between the clusters maximized. Now I do agree that at one point it will very difficult to understand from the scree plot which is the optimal level of the cluster and the silhouette score will also not provide some information. Then my viewpoint would be, based on the business requirement and dataset offered if you can decide what can be an optimal level cluster that you can create, it may give you an answer. If not, then you can make some assumptions and analyze both the clusters that you feel can be the answer.

"I have a doubt in question 1.3. It says - " Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them" what is the question asking to describe?? is it asking to describe the cluster profile or is it asking to describe the dendrogram and the number of clusters we are choosing??"

Yes, It is asking about both. You have to explain why did you choose the optimal # of clusters - whatever you decide as to the optimal number. You also need to profile the clusters that you get from the hierarchical clustering technique. Give pointers in bullets, no need to write lengthy paragraphs.

In the first question you mentioned in one of your answers that we should avoid using 2 clusters but when I see my dendrogram I can clearly see two distinct clusters and thus I went ahead with it. So, should I go ahead with 2 clusters or should I create 3 clusters instead??

I agree that from Dendrogram you'll get 2 clusters, but generally speaking, 2 clusters really do not make much business impact as it is kind of implicit. For example, for the dataset, it is imperative that there will be some high spenders and some low spenders. Then why are we doing clustering at all? So in such cases, you need to use your business context to better judge the optimal number of clusters.

Do we need to replace 2 highly correlated variables with one of the 2 correlated variables, before performing hierarchical & k-means clustering?

I would recommend using both first and then remove one and assess the difference. Multicollinearity generally doesn't impact the clustering process but it also doesn't contribute to any good.

Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

. **Query 1: Can we not use f-cluster instead of AgglomerativeClustering**
. **Query 2: Is it mandatory to take an optimal number of clusters depending on the different number of colors the dendrogram shows as output or can we take the number of clusters based on our understanding.**

Yes, you should use an f-cluster, and additionally, you can do agglomerative. Usually, the colors suggest the cluster numbers but you may decide the appropriate cluster number by cutting the dendrogram at a particular height

"Q1.1: Read the data and do EDA Clarification1: what is expected to be done in EDA Clarification2: Are we supposed to treat outliers Q1.3: Apply hierarchical clustering to scaled data. Identify the no. of optimum clusters using dendrogram and briefly describe them.

. **Clarification1: Are we also supposed to do agglomerative clustering?**
. **Clarification2: What linkage method is expected to be used for both dendrogram and agglomerative"**

EDA comprises of all the pre-processing steps like checking the shape, info, converting the variables incorrect data type, check missing values, check for duplicates, check for outliers, impute missing values, outliers, univariate, bivariate, and multivariate analysis, feature creation, scaling if required, etc.

Yes, if there are outliers present they should be treated as clustering results are affected by the presence of outliers.

Yes, agglomerative clustering should be tried.

Linkage Method - This is subjective and depends on what you would like to choose.

Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them ----- While executing the algorithm with linkage method =" average" and "ward", the results are different. I mean the results in terms of the number of clusters are different for these two linkage methods. Which method should we take?

Please explain what is the magnitude of the difference you can see. If it is drastically different, you might need to check if you have done it correctly like the underlying dataset or other aspects of clustering. If it is a matter once 1 or 2, I think that should be fine, these two techniques are methodologically different, which can produce some variation at an overall level, but not significantly different.

What do you mean by cluster profiles?

Cluster profile means the definition of the clusters. When you do clustering you basically try to group look-alike customers into smaller cohorts. Now, these cohorts are typically identified or described by their centroids or cluster centers. Therefore if you try to write in simple English what could be the properties of these cluster centroids - that process is called Cluster Profiling.

If my elbow curve shows no significant drop in WCSS after around 2.5 clusters and the Silhouette score is highest for 2 clusters, then ideally what should be my value of K in K - means clustering?

Clusters cannot be in fractions. It is either 2 or 3. Moreover, 2 clusters do not make much business sense. Therefore, select the clusters meaningfully - keeping the business perspective as well as technicalities in mind.

"Could you please explain the variable " advance_payments - Amount paid by the customer in advance by cash", does it imply, nonusage of credit card?"

No. Here is what 'advance_payments' means - Let's say you have made the purchase of 1000/- in the current cycle of your credit card. However, before the credit card statement has been generated, let's say you have paid the entire credit amount of your outstanding balance or maybe a part of it. That way, when the credit card statement will be generated, only the balance amount will reflect in your statement. It has multiple +ve repercussions on your credit card usage, your credit score, etc. Hence, typically, customers who make Advance Payments are low-risk customers.

In question 1, there is multicollinearity among variables. Would you recommend treating it before clustering?

Not Required

"Question only asks about describing the clusters. Moreover, these would be different from K Means clusters. Should recommendations in Q1.5 be for clusters from KMeans and Dendrogram?"

Please provide recommendations using hierarchical clustering (dendrogram)

"Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. Should the clusters from Dendrogram be used for recommendations or only the KMeans clusters?"

The clusters from the dendrogram should be used for recommendations

But if we apply PCA here to treat multicollinearity, how can we understand the business importance of PCs.I mean they will be in form of PC1, PC2, etc so after clustering, we might not be able to write business implications on PCs. Please suggest.

You can do PCA. However, generally, PCA is performed when you have high dimensionality. For the second point, you can inverse transform the components to get the original form. That shouldn't be a problem.

How can we identify the number of optimum clusters using Dendrogram?

Try to look at the silhouette score. Try to look at the height of the tree and what makes more sense in terms of #of optimal clusters. Try to look at the clusters after choosing an optimal number. It's a mix of all different things to choose the optimal #of clusters.

Do we have to drop the column "probability_of_full_payment" for performing clustering?

You need to analyze if that variable is adding any value to the clusters or not. Based on that you can keep or remove the variable - probability_of_full_payment.

Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters." - This has to be done for the clusters obtained through just KMeans or both, KMeans and Hierarchical?

Yes, you need to profile the clusters found via hierarchical clustering as well

"Q 1.5: Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. In the above question, what does it mean to " Describe cluster profiles"? Please clarify."

Cluster profiling is describing the clusters in plain English. What are the traits of those customers in different clusters? If you look at the cluster means, you should see some differences. You need to explain to them.

"Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. Can you explain more on what is expected on this question?"

Please try to create a table with the cluster centroids and then try to profile them. You can plot scatterplots between clusters and continuous columns to visualize the customer segmentation. A table can be created having the cluster labels and applying 'group by' on the data and talk about the difference in means or other statistical methods to differentiate and explain customer segmentation. Basically applying any function on the whole of data and displaying customer segmentation on the basis of cluster labels with logical explanation backing up your analysis will suffice here. Profiling means you need to understand and explain how these clusters are different from each other. For example, one cluster with high average card spends and the high average balance is typically those customers who are economically stable and have high spending capacity, on contrary, in another cluster customers with a high average spend but low average balance are those customers who are economical may not be stable but have a high potential of purchasing/spending because of various reasons. That way there would be different strategies to deal with these customers. You can offer a loan to the second group while an expensive brand new credit card product to the first group. Offering loans to the first group won't make much difference as they have a high balance already. You need to identify such patterns in your clusters and then explain them with whatever strategies you prefer!.

In the clustering problem, most of the variables have a high correlation among them (which is expected, given that the variables point to the financial/spending profile of each customer) - does that impact resulting clusters, and do we need to take any steps to account for it?

Great question. Ideally, you should handle it just to avoid any confusion. Clustering by default does not have any major impact on collinearity. But think of a scenario where you have three variables A, B, and C where A and B are highly correlated. During clustering, let's say you have 4 clusters. Now whenever you get a centroid value for A, B will automatically resonate with the same. There will be no surprise like a high average of A and a low average of B because that can't happen since A and B are highly correlated. So the best way to avoid confusion, create a transformed variable or use one instead of two. Take the variable which makes more sense in the business context.

how do I identify which is the best score as we see the score is high, however, the sample min value is negative, and in another cluster, the score is lower than the first one, however, the sample min value is positive. for e.g. Cluster 2 labels Silhouette score is 0.46 however, the silhouette sample minimum value is negative -0.016. And for the Cluster 3 label, the Silhouette score is 0.40 however the sample minimum value is positive 0.026. How to identify which one suitable cluster.

You have to take some judgmental call here. Although the Silhouette score is higher for cluster#2, it probably makes more sense to use #3 as that carries more business sense. Generally speaking, 2 clusters really do not make much business impact as it is kind of implicit.

For example, for the dataset, it is imperative that there will be some high spenders and some low spenders. Then why are we doing clustering at all? So in such cases, you need to use your business context to better judge the optimal number of clusters.

"On Data set no 1 - the last question is " Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters." It does not mention for which clustering, Hireachical or Kmeans"

It means you have to take the final version of the clusters you chose to use in the problem. The technique is immaterial here. You may use either of them and finally when you get the set of clusters, how do you profile them. That's what the question is all about.

Project Problem Part -2 - CART-RF-ANN

Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check (3 pts), Interpret the inferences from the descriptive statistics in a detailed manner (2 pts). -> 5 points.

. **Should Outlier treatment be done for this question?**
Ans: *Prefer not to treat outliers here. An observation is considered to be an outlier if that particular has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on the generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results. On the other hand, it is perfectly fine if you are building your models only once i.e. either after treating or not treating the outliers.*

. **Should any kind of scaling be done before we are fitting the data into any of the models?**
Ans: *It is not mandatory to scale your data before fitting it into all the models. You can choose to scale the data for the Neural Networks model.*

. **There are a few duplicate observations in the data. Should the duplicate observations be removed?**
Ans: *Manipulating the data in any way is necessary only when it is an absolute certainty that the data given contains some kind of erroneous observations. Here, if you think that there cannot be any duplicate observations as the travel company cannot sell the same kind of tour package to similar demography then please do go ahead and remove the duplicate observations.*

. **What should be the inference to be written in the business report after doing the null value condition check?**
Ans: *After performing the null value condition check if there are no null values present in the data set, please do mention the same in the business report. However, if there are null values present in the data, do mention the appropriate imputation techniques for the variable(s).*

. **Should we drop the variable 'Channel'?**
Ans: *A variable should only be dropped from the models if we are absolutely certain that that particular variable should not be having any effect on the target variable. Firstly, we should perform extensive Exploratory Data Analysis (EDA) to understand the data set, and then based on domain knowledge as well as EDA we can choose to drop a variable from the model.*

. **Should variables that have a *significant correlation* between themselves be dropped before building models?**
Ans: *There is no strict need to drop variables that have a significant correlation among themselves before building the models.*
Data Split: Split the data into test and train (1.5 pts), build classification model CART (2 pts), Random Forest (2 pts), Artificial Neural Network (2 pts). -> 5 points.

. **What should be written in the business report in the Data Split section?**
Ans: *In this section, you can mention the percentage of the classes in the target variable both in the training and the test set. Also, do mention the percentage split used in splitting the training and the test set.*
Please do make sure that the training-test split is the same for all three models.

. **How to perform encoding the data before fitting the data into the models?**
Ans: *The categorical variables can be treated in various ways. Here, different kinds of encoding for the categorical variables can be used. Do try out various iterations in the model building by treating the categorical variables differently.*
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy (1.5 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC_AUC score for each model (2 pts), Write inferences on each model (2 pts). -> 5 points

. **How should write the insights in the business report from the classification report and the confusion matrix?**
Ans: *For the question, please do elaborate on the confusion matrix. Also, mention the insights derived from the 'classification_report' about the various model performance parameters along with accuracy. You can put the confusion matrix in the business report as well.*

. **Should the ROC curve be mentioned in the business report?**
Ans: *It is always a good practice to include the various graphics which helps in understanding the models better.*
Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner (3 pts). Describe which model is best/optimized (2 pts). -> 5 points

. **What different things should be written in the business report for Question 2.3 and 2.4?**
Ans: *In question 2.3, all the reports and the values of the model evaluation metrics should be mentioned in detail whereas for question 2.4, the comparison of different models should be mentioned and the best or the most optimized model on the test set should be mentioned.*
Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. -> 5 points

. **What kind of business recommendations and business insights are should be put in the report?**

Ans: *For this question, it is expected that the EDA results and the model results are summarized and put in a proper format. Also, do put in any additional pointers that you feel are necessary for the business to understand from the data.*