**Problem 1: Clustering**

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.  
**1.1** Read the data and do exploratory data analysis. Describe the data briefly.

**1.2**  Do you think scaling is necessary for clustering in this case? Justify

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

**1.4** Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

**1.5** Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

1. **what is expected in terms of promotional strategies, is it about describing the business case as a whole**?

If you have few clusters where you can see some distinctive behaviour of customers, promotional strategies are basically how would you treat them. What are the offers you can design for these different group of customers. That is all required in this section

Dataset for Problem 1: [bank\_marketing\_part1\_Data.csv](https://olympus.greatlearning.in/courses/13598/files/1749852/download?verifier=RjCAv8QCqkhHUyA6GCcTP9jbmiUGZLAPbIROvguh&wrap=1)

**Data** **Dictionary** **for** **Market** **Segmentation:**

1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current\_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)

**Problem 2: CART-RF-ANN**

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

**2.1** Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

1. What all do we need to explain/show for “**null value condition check**”.

Check if NULL value exists or not. Mention your finding and provide technique to impute NULLs, if exist.

1. for the above question is univariate,  multivariate analysis, correlation and heat map needed?

Yes, everything is required in this section. You need to perform a complete EDA where Exploratory Data Analysis includes all the univariate / bivariate analysis, descriptive stats check, extracting important business insights. It needs to be a thorough examination of the data - keeping the business context in min

**2.2** Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network  
**2.3** Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model  
**2.4** Final Model: Compare all the model and write an inference which model is best/optimized.

1. W**hat is expected in this question as we are already describing the best model in 2.4?**

Provide business insights from EDA, Model outcome. Explain what are the factors influencing claims. Provide recommendations on how can these influencing factors be mitigated. 2.4 is all about comparing models while 2.5 is about the conclusion of the analysis that requires thorough explanation of business insights and recommendations to the business

**2.5** Inference: Based on the whole Analysis, what are the business insights and recommendations

Dataset for Problem 2: [insurance\_part2\_data-1.csv](https://olympus.greatlearning.in/courses/13598/files/1749851/download?verifier=sk8OxMboR7yasKxHHnYD32k3vv5DmTqqQ95ZIGsw&wrap=1)

**Attribute** **Information:**

1. Target: Claim Status (Claimed)  
2. Code of tour firm (Agency\_Code)  
3. Type of tour insurance firms (Type)  
4. Distribution channel of tour insurance agencies (Channel)  
5. Name of the tour insurance products (Product)  
6. Duration of the tour (Duration)  
7. Destination of the tour (Destination)  
8. Amount of sales of tour insurance policies (Sales)  
9. The commission received for tour insurance firm (Commission)  
10. Age of insured (Age)

Ask the Expert - Data Mining

The due date for this discussion topic has passed.

**Instructions**

Dear Participant,

You are free to ask the doubts related to Data Mining Project through Ask the Expert.  
This discussion forum is open from 30th Nov 2020**to 4th Dec 2020 11:59 PM** to answer all your queries related to the project.

Enjoy Learning!  
Program Office

**Comments**

340



**Sribindu Mavuri**

Dec 04, 10:47 PM

problem 2 :  should duplicates be eliminated before or after outlier treatment?  There is a noticable change in duplicate number after outliertreatment.  pls advise.



**Aviral Jain**

Dec 04, 11:43 AM



**Rashmi Dhar**

Dec 04, 2:36 AM

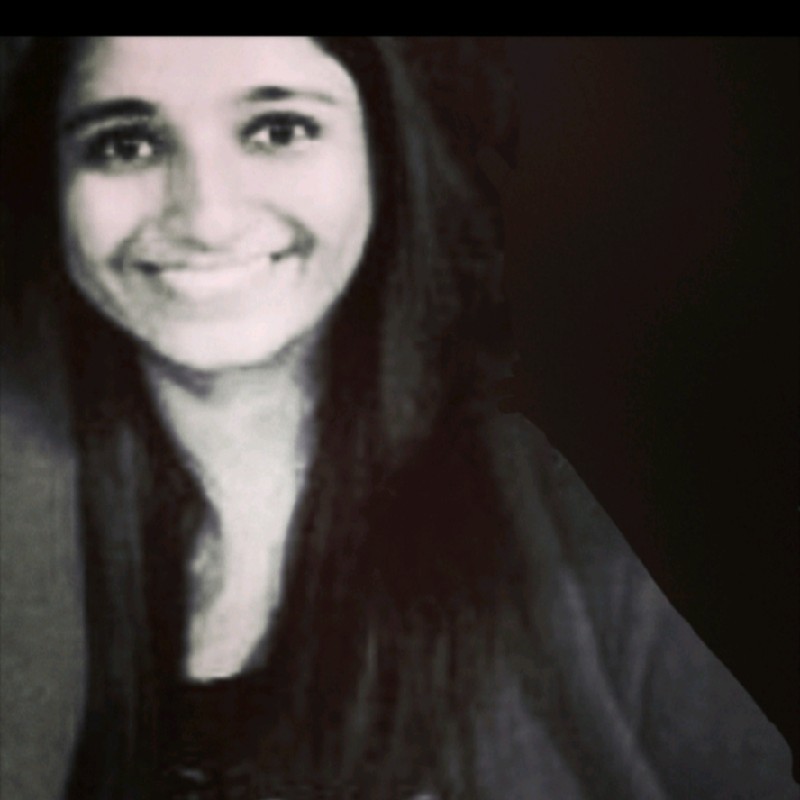
Do we need to keep Agency\_code column or can I drop it



**Somak Sengupta**

Dec 04, 3:50 PM

Please check the freq distribution first for all the variables. It is not a good practice to drop any variable in the beginning unless they are id fields.



**Divya Aggarwal**

Dec 03, 11:20 PM

Hi, Please help in the below mentioned queries:

* what all do we need to explain/show for “**null value condition check**”.
* If Variable Importance comes as zero for many variables post regularizing the tree, what could be the possible reason?
* 2.5 - Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. – **What is expected in this question as we are already describing the best model in 2.4.**
* 1.5- Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters - **what is expected in terms of promotional strategies, is it about describing the business case as a whole**?

Thanks!



**Somak Sengupta**

Dec 04, 4:02 PM

* what all do we need to explain/show for “**null value condition check**”. - Check if NULL value exists or not. Mention your finding and provide technique to impute NULLs, if exist.
* If Variable Importance comes as zero for many variables post regularizing the tree, what could be the possible reason? - Those variables are useless in predicting the target class. Alternatively you can also check with RFE or logit function to see which all variables are significant.
* Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. – **What is expected in this question as we are already describing the best model in 2.4. -**Provide business insights from EDA, Model outcome. Explain what are the factors influencing claims. Provide recommendations on how can these influencing factors be mitigated. 2.4 is all about comparing models while 2.5 is about the conclusion of the analysis that requires thorough explanation of business insights and recommendations to the business.
* Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters - **what is expected in terms of promotional strategies, is it about describing the business case as a whole**? - If you have few clusters where you can see some distinctive behavior of customers, promotional strategies are basically how would you treat them. What are the offers you can design for these different group of customers. That is all required in this section.



**Sribindu Mavuri**

Edited on Dec 04, 4:04 PM

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

for the above question is univariate,  multivariate analysis, correlation and heat map needed?



**Somak Sengupta**

Dec 04, 4:04 PM

Yes, everything is required in this section. You need to perform a complete EDA where Exploratory Data Analysis includes all the univariate / bivariate analysis, descriptive stats check, extracting important business insights. It needs to be a thorough examination of the data - keeping the business context in mind.



**Aviral Jain**

Dec 03, 10:47 AM

What is the most important parameter while selecting no of clusters in K-means, is it Silhouette score, elbow curve or inertia values?



**Somak Sengupta**

Dec 03, 11:15 AM

It needs to be judgmental. Look at all the different charts / scores and apply the right business context before finalizing it.



**Parnoshree**

Dec 02, 11:25 PM

 Please explain what steps are they looking under descriptive statistics.



**Somak Sengupta**

Dec 03, 10:34 AM

Exploratory Data Analysis is what you need to perform in this section.That includes all the univariate / bivariate analysis, descriptive stats check, extracting important business insights. It needs to be a thorough examination of the data - keeping the business context in mind.



**akshay kumar**

Dec 02, 6:16 PM

**2.2** Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

the above mentioned process are done. as an answer what details we need to mention here



**Somak Sengupta**

Dec 03, 10:37 AM

1. Provide how you have split the data in terms the parameters you have used to split.

2. Show the train test size post split and show the target class proportion after splitting in both train and test.

3. Show how you applied the algorithm - how the hyper-parameters you have used to run the algorithm

4. Write couple of lines for each model, how you applied and what's your take on the same.



**Sripriya Elisetty**

Dec 02, 6:09 PM

For question 2 :- Is it wrong approach treating the outliers before building the three models.



**Somak Sengupta**

Dec 03, 10:37 AM

It's okay if you have treated the outliers for CART and RF, though it is not necessary. It's not wrong if you have already done it.



**Divya Sarika Amjuri**

Dec 02, 5:23 PM

for problem 1, perform eda means do we have to find mean,iqr and all or is finding missing,duplicate and outliers are enough. can anyone please clarify this ?



**Somak Sengupta**

Dec 03, 10:38 AM

Exploratory Data Analysis includes all the univariate / bivariate analysis, descriptive stats check, extracting important business insights. It needs to be a thorough examination of the data - keeping the business context in mind.



**Rashmi Dhar**

Dec 02, 9:56 AM

Q1.1 Read the data and do EDA.  
Do we have to do Histogram boxplot for all individual columns ?



**Somak Sengupta**

Dec 03, 10:39 AM

Exploratory Data Analysis includes all the univariate / bivariate analysis, descriptive stats check, extracting important business insights. It needs to be a thorough examination of the data - keeping the business context in mind.



**Hemanshu Sardana**

Dec 01, 3:32 PM

Hello

In Problem 1- Clustering, Do we need to just perform the scaling on the whole data or we need to convert the different columns on same scale (like 100, 1000, 10000 on one scale) and then do scaling of data?

Please clarify as it is getting confusing.



**Somak Sengupta**

Dec 01, 6:10 PM

Yes, that would be the best approach. First convert the day in their original scale and perform scaling.



**Syed Rizwan**

Dec 03, 8:04 PM

Hi,

How to convert data in original scale?

we do scaling for converting all the variables into the same scale, so why we need to convert and then do scaling?



**Somak Sengupta**

Dec 04, 4:06 PM

The data provided are in different scales - 10's,100's etc. So recommendation would be first you should convert them to their respective scale and then perform scaling which could either be StandradScaler or MinMaxScaler etc.



**Syed Rizwan**

Dec 04, 11:26 PM

Hi Sir,

 How to convert them into the respective scales ?

(1)



**Sumodh Pallikunnath**

Dec 01, 5:07 AM

Do we need to scaling before  CART – RF - ANN



**Somak Sengupta**

Dec 01, 6:11 PM

For ANN, yes. For others no.



**Sumodh Pallikunnath**

Dec 01, 1:20 AM

Do we need to treat outliers before  CART – RF - ANN



**Somak Sengupta**

Dec 01, 6:11 PM

For ANN, yes, for others No.



**Parnoshree**

Edited on Dec 01, 6:13 PM

The truth value of an array with more than one element is ambiguous. Use a.any() or a.all()  
 I get this error when performing hierarchical clustering.  
Please help my code is.  
dend = dendrogram(link\_method,labels = labelList)  
  
dend = dendrogram(link\_method,  
                 truncate\_mode='lastp',  
                 p = 10,  
                 labels = labelList   
                 )



**Somak Sengupta**

Dec 01, 6:13 PM

I cannot really understand what is the issue here. Please try to check the dataset you are providing in the function. Also, you can read the function details by putting a? front of the function.



**Syed Rizwan**

Nov 30, 4:41 PM

Hello Sir,

Q1.) With respect to which variable we have to carry out the clustering?

Is it preferable to take spending as a variable for clustering?

Q2.) Is it necessary to drop the Agency\_code for 2nd question even though there are only 4 unique values?

Q3.) Can we predict the performance metrics directly after doing grid search CV first after defining the classifier for the models or first we have to do the coding giving inputs from our end then calculate the y variable and then performance metrics after that again check for grid serach CV and repeat the streps again for best parameter.



**Somak Sengupta**

Nov 30, 10:11 PM

1. Yes you can. In unsupervised learning i.e. Clustering, you don't know until you try to use the variable. So better start with it and see what importance does it carry.

2. Why do you think you need to drop the variable - Agency Code?

3. The second option is the best way to proceed.



**Syed Rizwan**

Dec 01, 6:11 PM

Q1.) Means twice we have to check performance metrics?

Do we only do the grid search CV for RF and ANN model?

Q2.) Importance check just like we did for CART and RF, right?

Do we have to check with respect to each and every variable?



**Aviral Jain**

Nov 30, 4:00 PM

1. Some values in the dataset are in 100s while some are in 1000s, do we first need to convert the dataset before starting our analysis?



**Somak Sengupta**

Nov 30, 10:07 PM

Yes. Please convert them to their respective scales and then do scaling and proceed further.



**Sumodh Pallikunnath**

Nov 30, 2:46 PM

1. Is removing duplicates necessary for project CART-RF-ANN
2. Is it mandatory to convert Object type  to Float or integer before CART-RF-ANN



**Somak Sengupta**

Nov 30, 10:06 PM

1. Yes, it is recommended.

2. Can you please elaborate a little what type of data it is? Not quite sure I understand your question properly.



**Simran Shah**

Nov 30, 1:54 PM

 if we have float data types in the data set, do we need to convert them as well for tree based analysis?

also why do silhouette score keep on changing whenever i run them



**Somak Sengupta**

Nov 30, 10:04 PM

I didn't get what do you mean by convert them as well.

For the second question, set a fixed value for random state, then the values won't change.