VIKRAM RADHAKRISHNAN

May 10th, 2020

DATA MINING FINAL PROJECT

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.

The data appears to be clean with all columns having continuous numerical values. There are no null values, no NaN values, missing values or special characters.

However, it appears that there are duplicate values. These values have been removed before processing the data.

Outliers are checked, and there are only a few. Therefore, no further processing is done with regards to outliers.

Heatmaps and pairplots are drawn to see the relationship between variables, and its observed that there are is a lot of correlation between spending and advance_payments. Current_Balance and Credit_limit also show high correlation to spending.

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

Scaling is definitely needed if the variables are in non-comparable units (like metres, kg, deg.F).

In our case however, the units are of the same measure. However the variances of the different varibles are not the same.

Leaving variances unequal is equivalent to putting more weight on variables with smaller variance, so clusters will tend to be separated along variables with greater variance.

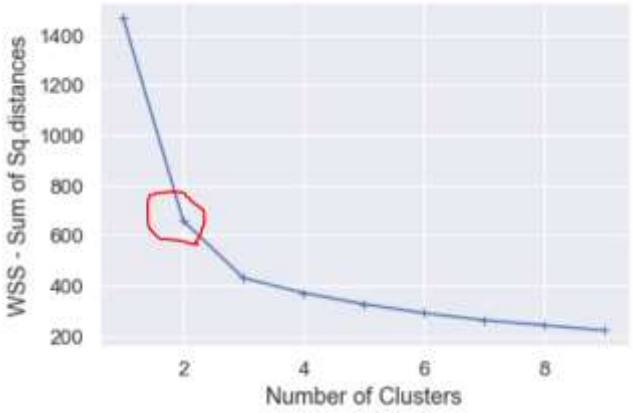
Therefore, scaling is performed in our case regardless of the variables having same units.

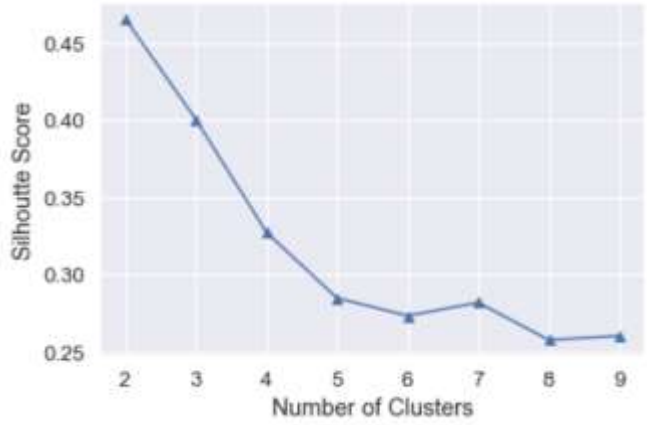
1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

The data is scaled and the 'ward' method is applied to create dendograms. The optimal number of clusters appears to be 2.

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

K –Means clustering is applied an a possible elbow is found at K=2 and K=3. However, since the silhouette score is more for K=2, we choose to go with 2 clusters.





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

Amongst the two clusters, the first cluster appears to have comparatively lower spending than the second cluster on an average (13 units vs 18 units). It is also observed that the average expenditure on a single shopping visit is 5 units for the first group against 6 units for the second.

From the credit card minimum payment amount, it is seen that the first cluster is more dependent on Credit to do the expenses. The second cluster does a larger value of advance payment. The average probability of full payment is also almost the same for both clusters. Therefore, increasing credit limits for the first cluster would be a promotional strategy to increase sales in the first cluster.

Differentiated marketing for each of the clusters as moderate and premium segments would lead to better promotional paths.

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.

The data-set is read and cleaning is attempted. The 'Agency' column is dropped as its not required for analysis. Null values and NaN values are investigated and are not found. Duplicated lines are removed. There are no odd characters in the data.

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

This process is done. Please refer to the code.

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

This process is done. Please refer to the code.

2.4 Final Model: Compare all the model and write an inference which model is best/optimized.

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.78	0.78	0.82	0.77	0.75	0.77
AUC	0.84	0.80	0.77	0.72	0.66	0.66
Recall	0.54	0.54	0.65	0.61	0.39	0.40
Precision	0.71	0.66	0.75	0.59	0.72	0.69
F1 Score	0.61	0.59	0.70	0.63	0.51	0.50

Based on the accuracy scores the CART, the Random Forest and ANN are almost equally good.

However, when it comes to the other measures such as AUC, Recall and F1 Score, ANN lags in this case. Its possible that the accuracy of the ANN is affected by a significant number of outliers to which the NN is sensitive to. Further tuning of the

NN could improve its performance, and using more hidden layers could reduce the effects of outliers. If one were to choose the best or optimized model in this scenario, it would be Random Forest.

2.5 Inference: Basis on these predictions, what are the business insights and recommendations

	Imp
Product Name	0.334224
Sales	0.213442
Commision	0.143992
Duration	0.132228
Age	0.089662
Type	0.066779
Destination	0.017030
Channel	0.002643

Based on feature importance, its seen that Product name plays the most important role in affecting the dependent variable or target, followed by Sales. Based on product name, the insurance company could look into which products are seeing higher claims and re-work on the offerings. In terms of sales, it needs to be seen which category of product sells more and whether that product has a high average claim rate. If so, the price of the insurance policy could be increased. Similarly, other features can be looked into.