10 Questions & Answers

Q: Since nothing significant was found from the exploratory data analysis, should the business use the K-Means model now to start their targeted advertisement campaign?

A: They can but the model is close to random guessing, so it's highly suggested that they do not. With it being close to random guessing the company may find that they're spending more money without increased profit.

Q: You seem to bring up the data in reference to the bad silhouette score, could this just be that the wrong model was chosen for this data?

A: Yes, it could be that the model wasn't the best choice for data. However, I bring the data up more because there are even splits with in columns of the data which could mean it's more manufactured than actual data.

Q: The K-Means visual seems to show 8 distinct groupings. Wouldn't the 8 clusters work better than 2?

A: If the only option at hand was to evaluate the model based upon the visual then we may have gone with the 9 clusters. However, we can see in the silhouette scores that the accuracy of the decreased significantly if we have more than two clusters.

Q: Can you show that what you're saying about the categories is correct and that these aren't a generalized assumption?

A: This would be more detailed than just showing. You would need to compare the K-Means results to the converted data frame and understand what the integer values represent in each category. For gender its easy because it's either 0 or 1 so if more males are in the category the value would be closer to 0 and 1 for more females than males. The rest can become tricky but knowing what the values represent really help to interpret the model.

Q: I see you mention the ID column, but this doesn't show up in the categories you discussed. What happened to this column and the Var_1 column?

A: I ran the model with these in initially, but the ID column cause the model only to fit based on the ID. The var_1 column was an anonymized category for the customer and didn't make much sense in the model. Once this was removed the model improved significantly.

Q: Since the work experience visual doesn't make sense the way it's displayed currently, could this be displayed with the age values reversed?

A: It makes sense in some ways, like maybe the older generation only needed one member of the family to provide an income. Also, the values were set up correctly with 0 being the youngest and 10 being the oldest. This being the case the visual is set up correctly.

Q: Why are there no values in either category greater than 10? If these were converted, how were they converted?

A: All the values were scaled to be between 1 and 10. This was because there were multiple columns with less than 10 options and a few greater than 10. All the data points went through the same scale. Each data point had the minimum value subtracted from it then was divided by the difference of the maximum and minimum data point. This value was them multiplied by 9 and had 1 added to the result.

Q: Is the silhouette score the only test for accuracy for k-means? The model could have had better results if a different test for accuracy was used.

A: Silhouette score is not the only way to check for this model's accuracy; there is also within-cluster sum of squares as well that can be used to test for accuracy. However, I felt the result would be similar since the silhouette score was so low. There are additional tests as well to check for the stability of the model which I did not move onto as the silhouette score for this model was unsatisfactory.

Q: With the low silhouette score, can the model be underfitting?

A: while the model didn't perform well, it is not underfitting since the model was far from 0 and it wasn't negative. Underfitting would occur is we chode any other data point that wasn't 2 since point was the place where the distortions decreased most rapidly.

Q: If the dataset gets larger and new features are added, can the model handle this scalability?

A: Yes, the K-Means method can handle many values. However, if more columns are added we may need to make sure that there is no multicollinearity occurring. Also, increasing the features increases dimensionality which can spread points out further. It would be best to complete a feature selection and reduction to avoid this issue.