Candidate Name: Tan Yin

Please note the following:

1. Both questions use the same dataset
2. All source code must be provided with comments along with instructions on how to execute the code.
3. R is preferred, but you are free to consider other tools. Regardless of the tool you use, source code / workings must be provided and clearly documented
4. Two datasets are provided **spenddata.csv** and **testdata.csv.** You are free to decide how best to use them.
5. You do not need to augment this dataset with any external data. However, if you choose to do so, you must document the reasons.

**Question 1**

You are given a set of survey data which captures spend amount among other data points. The Marketing team is curious to understand if there are similar groups of people visiting Singapore and they have approached the data science team to help them find out.

Using the set of data provided:

* 1. Determine how many groups of tourists have visited Singapore

|  |
| --- |
| Based on the clustering algorithm, it can be seen that there are **3** distinct groups of tourists that visited Singapore. |

* 1. Please explain the choice of metric used

|  |
| --- |
| All variables were considered for this question except for “month”, first column which represents the index column, “year”, and “respondent.id”. These are the metrics that do not add value to the computation and might cause variation in the results if they were used.  For the remaining data, as majority of them are masked, it is hard to determine what the values represent. As such, all these values are used for the computation of the clustering algorithm. |

* 1. What are the assumptions you made when building this solution?

|  |
| --- |
| 1. A dummy variable was created for the var9 column. In that column, I assumed that there were only 3 different values by eyeballing the dataset. These values are: Mono, Multi, and empty values. To perform the necessary computation, I needed to change these into dummy variables. As such, there are only 3 different dummy variables for this column.  2. For all other NULL values, I assumed that they are non-binary fields. This was because for binary fields, if the response was a FALSE or NEGATIVE response, the field will be labelled as 0. As such all NULL values were replaced with the value of 0 to show NULL fields. |

* 1. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.

|  |
| --- |
| The KMeans Clustering algorithm was used in this dataset. This algorithm helps to find the value of K, which will determine the number of similar groups of people visiting Singapore, as requested from the marketing team.  Using the KMeans Clustering algorithm and the elbow curve, it helped us easily identify the number of possible clusters. However, there are also cons as in some cases, the elbow curve does not clearly define the value of K. In such a case, it is hard to determine the number of clusters. |

* 1. Please explain under what conditions will the approach you choose **be not appropriate**

|  |
| --- |
| The KMeans Clustering algorithm will not be appropriate when the value of K cannot be easily determined. This is when the elbow curve shows a straight line or when there is no definitive spot where an arc can be seen. In such a case, there will be a need to look into other clustering algorithms to solve the question. |

* 1. Why is your approach performing well / not well?

|  |
| --- |
| In this question, the model was able to perform well as it was able to determine the value of K. Despite having some discrepancy in the choice of K, as k=3 and k=5 seemed like viable options, k=3 was chosen as it had a larger arc as compared to k=5. Ultimately, when plotting the results on a heatmap, the 3 groups were able to display different characteristics, which will aid in the profiling of tourists. |

* 1. Was any feature engineering required? If yes, what were they. If no, why?

|  |
| --- |
| 1. There was a need to create dummy variables for var9 column as the algorithm does not allow textual information. Thus, Mono, Multi, and empty values were converted into the values of 1, 2, and 0 respectively.  2. Nan values had to be replaced with 0. This was also because of the limitations where the algorithm does not allow NULL values. Thus, all NULL values were replaced with 0.  3. Data transformation was done for highly skewed variables. Variables with skewness of greater than 1 and lesser than -1 were transformed using log transformation. The transformation will convert the skewed distribution in a one that is normal, such that outliers can be prevented. However, there were issues with the transformation as log(0) will return as error. To resolve this issue, smoothing had to be done. The value of 1 was added to every cell value such that the minimum value of log(1) will not return an error.  4. Standardisation was done using the scaler. This will help to ensure that the variable values are on the same scale, such that variables with bigger values will not dominate the choice of clusters. |

**Question 2**

Some of the respondents have been tagged as belonging to group 1 – 6. However, due to a data calculation issue, some of the respondents have had their groups (**pov6**) missing.

Build a model that will classify these respondents back into one of the 6 groups.

1. Please explain the choice of metric / evaluation criterion used

|  |
| --- |
| The intersect of variables in *spenddata* and *testdata* were considered for this question. All variables were included except for “month”, first column which represents the index column, “year”, and “respondent.id”. These are the metrics that do not add value to the computation and might cause variation in the results if they are used.  For the remaining data, as majority of them were masked, it is hard to determine what the values represent. As such, all these values are used for the computation of the classification algorithm.  For the evaluation criteria, the accuracy and F1 metrics were used to determine which was the better model. The accuracy was the first level of comparison that was being done between models to determine the general performance of the model. Following which, the F1 metric was used to determine the performance for each of the 6 groups. |

1. What are the assumptions you made when building this model?

|  |
| --- |
| 1. A dummy variable was created for the var9 column. In that column, I assumed that there were only 3 different values by eyeballing the dataset. These values are: Mono, Multi, and empty values. To perform the necessary computation, I needed to change these into dummy variables. As such, there are only 3 different dummy variables for this column.  2. For all other NULL values, I assumed that they are non-binary fields. This was because for binary fields, if the response was a FALSE or NEGATIVE response, the field will be labelled as 0. As such all NULL values were replaced with the value of 0 to show NULL fields. |

1. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.

|  |
| --- |
| The Naïve Bayes approach was being considered. This is because Naïve Bayes uses the actual values to make predictions. Naïve Bayes also requires less training data as compared to the other algorithms.  However, the Naïve Bayes model requires the correct classification in the training dataset. Also, there should be sufficient datasets to train the data with, despite it needing lesser training data as compared to some other methodologies.  In Naïve Bayes, there are 3 different models which I had tried. These models include the Gaussian model, Multinomial model, and Bernoulli model.  1. The Gaussian model will generally work for most cases, with an assumption of normal distribution. If there is no normal distribution, it might affect the accuracy of the predicted results.  2. The Multinomial model is mostly used for text classification purposes. However, I thought that it was worth giving the model a try to determine the performance.  3. The Bernoulli model is normally used when there are binary features. This might be useful in this case where there were a number of fields that contain binary fields.  In the end, the Bernoulli model was chosen as it was able to obtain the highest accuracy score and the best f1 scores for each of the 6 groups. |

1. Please explain under what conditions will the model you choose **be not appropriate**

|  |
| --- |
| The Bernoulli model will not be appropriate if there the dataset contains a large number of variables which are continuous data. This is because the Bernoulli model will generally work well when discrete variables (mostly binary) are used. In a case where majority of the data is categorical, the Gaussian model would be a better fit.  If the dataset contains textual data, it will also be more viable to use the multinomial model as compared to the Bernoulli or Gaussian model. |

1. How confident are you of the model’s robustness and how would you explain the model’s performance?

|  |
| --- |
| The Bernoulli model was able to obtain an accuracy of 0.786, with F1 scores of 0.87, 0.83, 0.82, 0.38, 0.15, and 0.15 for groups 1 to 6 respectively. Based on these scores, we can conclude that the model was able to correctly predict 78.6% of the dataset. As for the group level, we see that it was generally able to predict groups 1 to 3 well, and was not able to predict groups 4 to 6 well. This might be due to the lack of training datasets for groups 4 to 6. |

1. Why is your model performing well / not well?

|  |
| --- |
| The model was able to predict well for groups that has a high number of training datasets. This was the case for groups 1 to 3. However, the model was not able to perform as well for groups 4 to 6 due to the lack of training data. To better improve the model, a larger dataset could be utilised. Considerations such as the dataset size for each of the groups might be taken into account such that the model will be able to better predict values and obtain better performance. |

1. Was any feature engineering required? If yes, what were they. If no, why?

|  |
| --- |
| 1. There was a need to create dummy variables for var9 column as the algorithm does not allow textual information. Thus, Mono, Multi, and empty values were converted into the values of 1, 2, and 0 respectively.  2. Nan values had to be replaced with 0. This was also because of the limitations where the algorithm does not allow NULL values. Thus, all NULL values were replaced with 0.  3. Data transformation and standardisation was not needed as this question uses a supervised machine learning methodology. |