Assignment 2: Classification and Cluster Analysis of Internet Users; due date 2021-04-16

Student: Warren Byron (20091892)

Marking Scheme

- 10%: 1 CLUSTER choose candidate feature subsets
- 15%: 2 CLUSTER use clustering to identify possible groups
- 15%: 3 CLUSTER discuss pros and cons of clustering technique
- 10%: 4 CLUSTER which features perform best and why
- 25%: 5 CLASS fit the data using 3 techniques
- 10%: 6 CLASS rank the classification algorithms
- 15%: 7 CLASS repeat the analysis with PCA reduced data

Marks and Comments

1 CLUSTER choose candidate feature subsets

- The student's use of EDA was exemplary - by following a very structured approach, he was able to identify aspects that were new to us.
- The student used (count)plots and other techniques to select features, grouping values where necessary, excluding features that were correlated with others…
- The student also uses information from other sources, e.g., on internet adoption, and this guided his choices.
- There is a very nice diagram showing PCA explained variance as a function of number of components.
- The student carefully analyses each feature, deciding whether to include or exclude; all decisions are well-justified.
- It was a nice idea to see if the multi-value columns could coded as categorical.
- Good approach to deal with missing values for Age.
- Excellent EDA - organised, logical and to the point.
- Good attempt of feature engineering to extract language from Primary_Language.
- You are too aggressive in dropping features that are not balanced, for example Gender was only 61/38 split ratio.
- Interesting experiment with Household_Income - it is probably better to use the original categorical feature, but interesting attempt.
- Rather than going from categorical to numerical we often go in the reverse direction. This is called Binning. By converting a numerical feature to categorical we do lose information but the simpler feature can help model training. See Kaggle Bike Hire competition for a successful application of binning.

- Excellent comments - comments were both based on observations of the data and whether those observations should be expected or not.
- Marks: 10/10

2 CLUSTER use clustering to identify possible groups

- The student used clustering techniques (AGNES, k-means, GMM and DBSCAN) very well.
- Perhaps because of the good choice of features, and limiting the dendrogram depth, it was relatively easy to compare the different linkage types and to make more reliable estimates of k as a result.
- The "Comments Regarding The Clustering Findings" and associated table, were perceptive and clear.
- The only thing that might be added would be to apply EDA to each cluster subset, but everything else was exceptionally well done.
- The organisation of features into feature subsets is excellent, and the resulting clusters limited to the various feature subsets are novel and interesting.
- Marks: 14/15

3 CLUSTER discuss pros and cons of clustering technique

- Each clustering techniques was appraised, with its advantages and disadvantages.
- Metrics (inertia, silhouette score, and distance (for DBSCAN) were used to interpret the success of each technique for given parameter settings, and the discussion was very well-informed.
- Excellent
- Marks: 15/15

4 CLUSTER which features perform best and why

- The discussion on which features performed best looked at the various feature groups and identified what clusters were meaningful and what represented noise.
- As elsewhere in this assignment, all observations were supported by analysis.
- This was not answered directly since features were put into subgroups, and in the ALL case the ranking of features was not given.
- Marks: 10/10

5 CLASS fit the data using 3 techniques

- Even the choice of classification procedure was made in a careful and well-supported way.
- Pipelines are used to organise the experiments, and cross-validation is used to estimate prediction accuracy.
- Timings are collected and various settings are changed so this is an investigation of the classification techniques too.
- Hyperparameter tuning is applied and improvements, when they arise, are noted.

- Student mentioned that “categorical features were converted to numerical features. I’m going to keep these as numerical attributes because it’ll make the classification algorithms better”. This is not always true.
- Standardising, while still a good idea, only had a minor impact since the individual feature ranges did not vary greatly.
- The intention was to develop two separate classification models. Combining them made the classification harder, and its implementation harder.
- Liked the mangling of the two targets into a single target.
- Marks: 25/25

6 CLASS rank the classification algorithms

- The student compared and contrasted the algorithms, before and after tuning, with and without PCA.
- The student noted where they differed and gave plausible explanations.
- All observations were supported with extensive investigations and experiments.
- Excellent work.
- Marks: 10/10

7 CLASS repeat the analysis with PCA reduced data

- The use of pipelines made it easier to incorporate PCA.
- As was the case with clustering, the student noted how accuracy varied with number of components, by analysing some well-chosen plots.
- The student was able to choose the number of components to keep to maximise the accuracy performance of *each* classifier - very impressive!
- This task was completed to a very high standard.
- PCA was included as part of the pipeline correctly and its impact on the various classifiers was correctly discussed.
- Marks: 15/15

Overall: 99/100 This was an exceptional attempt, with excellent use of module and online resources, benefiting from the student’s own intuition and investigative zeal. Well done!. Excellent project, the approach to referencing was something special. Scaled: 49.5% of overall mark