DAV 6150 Project 2 (Module 9)

Clustering + KNN: Can We Predict Purchases from Web Sites?

***** You may work in small groups of no more than three (3) people for this Project *****

Most online retailers are prolific users of data science methods for a variety of purposes, including fraud detection, personalized marketing, price optimization, upselling, and inventory management. Of particular interest to most online retailers is whether or not a site visitor ends up executing a purchase while engaged with the web site. Web sites are capable of capturing a wide variety of metrics any time someone accesses one of their web pages, including the recording of the ID of the specific web page visited, the ID's of any items the user either hovered over or clicked on, the elapsed time spent by the user on the page, etc. Online retailers often use such data to try to determine whether or not a given site visitor will actually make a purchase.

For this Project, you will be working with a data set comprised of a variety of such web site metrics. The data is sourced from the UCI Machine Learning Repository :

- https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

Please refer to the web page cited above for further details on the variables contained within the data set.

Your objective for Project 2 is to use clustering algorithms to create groupings of similar data observations within the provided data set, apply labels to the data observations assigned to those groupings, and then, after completing the necessary EDA and data prep work, construct a KNN model to predict the most likely group/categorization of any previously unseen data items. It will be up to you as the data science practitioner to determine which features to include in your KNN models. To get started on the Project:

1) Load the provided **Project2_Data.csv** and **Project2_Data_Labels.csv** files to your DAV 6150 Github Repository.
2) Using a Jupyter Notebook, read the **Project2_Data.csv** data set from your Github repository and load it into a Pandas dataframe.
3) Prior to embarking on the implementation of a clustering algorithm, perform EDA on the data set. Of particular interest to you should be whether or not any data standardization techniques should be applied to the variables within the data set. Also, are any of the variables categorical rather than numeric? If so, you will need to think about how such variables can be effectively incorporated into a clustering algorithm.
4) Perform any required data preparation work, including any feature engineering adjustments you deem necessary for your work. For example, we know that clustering algorithms don't work well with non-numeric data: how should such attributes be prepped for use within a clustering algorithm? Also, do data standardization techniques need to be applied to any attributes?
5) Apply a hierarchical clustering algorithm to the data. What does its Dendrogram show you? How many clusters do you think should be imposed on the data based on what the Dendrogram shows?
6) Implement a K-means clustering algorithm. Start by using a range of values for K to create an elbow plot and a silhouette plot for the data set and use the plots to select an appropriate value for K. Is the

output of these plots indicative of a K value that is in line with the number of clusters you selected from the output of the hierarchical Dendrogram?

7) Based on the results of your Dendrogram, elbow plot, and silhouette plot, select what you believe should be an appropriate number of groupings to be applied to the data. Apply a K-means clustering algorithm to the data set using a value that matches the number of groupings that you have selected (e.g., if you believe that 5 groupings should be applied to the data, apply the K-means clustering algorithm using a K value of K = 5). At the conclusion of that process you will have K groupings of data.

8) Perform EDA on these different groupings: what do the summary statistics tell you? Are the groupings noticeably different from one another?

9) Add a new column to your Pandas dataframe with the name **Group**. Within this new column, insert the grouping assignments identified by your K-means clustering for each observation within the data set.

10) Read the **Project2_Data_Labels.csv** file from your Github repository and add its content to another new column within your Pandas dataframe. Assign a column name of **Revenue** to this new column. This new column contains the actual classification labels for the observations contained within the data set and indicates whether or not a web site visitor executed a purchase before leaving the web site. Based on the content of the **Revenue** column, calculate some basic comparative statistics between the groups resulting from your clustering work: What percentage of each of your groups actually made a purchase while visiting the web site?

11) Apply your knowledge of feature selection and/or dimensionality reduction techniques to identify explanatory variables for inclusion within your KNN models (NOTE: both the **Group** and **Revenue** columns represent response variables, not explanatory variables). You may select the features manually via the application of domain knowledge, use forward or backward selection, or use a different feature selection method (e.g., decision trees, etc.). It is up to you as the data science practitioner to decide upon the most appropriate feature selection and/or dimensionality reduction techniques to be used with the data set.

12) Separate the dataframe into model training and testing subsets.

13) Construct at least two different KNN models using different explanatory variables (or the same variables if they have been transformed via different transformation methods) for purposes of predicting the value of the **Group** classifier (i.e., **Group** is the response variable for your model).

14) After training your KNN models, decide how you will select the "best" model from those you have constructed. For example, are you willing to select a model with slightly lower performance if it is easier to interpret or less complicated to implement? What metrics will you use to compare/contrast your models? Evaluate the performance of your models via cross validation using the training data set. Then apply your preferred model to the testing data set and assess how well it performs on that previously unseen data for purposes of predicting the value of the **Group** classification.

15) Finally, compare the predicted values for the **Group** attribute from the testing data set to the actual values of the **Group** variable. How well does the KNN model derived from the results of your clustering work match up to the actual **Group** classifications you assigned to the dataset?

Your first deliverable for this Project is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem
2) **Pre-Clustering Exploratory Data Analysis (10 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis regarding the integrity + usability of the data in its raw state. This section should include any Python code used for the EDA
3) **Pre-Clustering Data Preparation (5 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering you have applied to the data set. This section should include any Python code used for Data Preparation.
4) **Cluster Modeling (20 Points)**: Explain + present your hierarchical and K-means clustering work, including your interpretation of the outputs of the models (e.g., comment on the Dendrogram + elbow + silhouette plots; how many clusters should be imposed on the data?, etc.).
5) **Post-Clustering Exploratory Data Analysis (5 Points)**: Explain + present your post-clustering EDA work including any conclusions you draw from your analysis regarding groupings generated by the K-means algorithm based on the number of groupings you specified. Comment on whether you can identify any noticeable differences between the two groupings. This section should include any Python code used for the EDA
6) **Clustering Output vs. Actual Labels (5 Points)**: Compare the content of **Revenue** to the content of the **Group** column generated by your clustering algorithm: does the output of the clustering algorithm appear to be indicative of whether or not a web site visitor made a purchase? Calculate some basic comparative statistics between the two columns to demonstrate their similarities and differences and comment on your findings.
7) **KNN Modeling (20 Points)**: Explain + present your KNN modeling work, including any feature selection methods used + the use of any kernel functions.
8) **Select Models (10 Points)**: Explain how you selected your model selection criteria. Identify your preferred model. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the testing data set and discuss your results. Did your preferred model perform as well as expected?
9) **Clustering vs. KNN Output (5 Points)**: Compare your initial **Group** column content to the **Group** predictions generated by your KNN algorithm: does the KNN algorithm appear to be effective at properly assigning customers to the groupings you created as a result of your clustering analysis? Calculate some basic comparative statistics between the actual and predicted **Group** values to demonstrate their similarities and differences and comment on your findings.
10) **Conclusions (5 Points)**

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook within the provided Project 2 Assignment Canvas submission portal. Be sure to save your Notebook using the following nomenclature: **first initial_last name_Project2**" (e.g., J_Smith_Project2). ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***

Your second deliverable for this Project **(10 Points)** is a short (approx. 10 minute) video presentation of your work. Your presentation should include a brief overview of your clustering work, a high-level explanation of your data preparation + feature selection process + KNN models, a summary of your model selection process, an explanation of why you chose your preferred model, and comments on the performance of your preferred model when applied to the testing data set. Finally, you should comment on how well the results of the combined clustering + KNN modeling process aligned with the actual **Group** data values.