APRIL 4, 2024

D212

PERFORMANCE ASSESSMENT

ROBERT PATTON
WESTERN GOVERNORS' UNIVERSITY
Rpatt33@wgu.edu

Table of Contents

Part I: Research Question

A. Purpose of Data Mining Report

1) Based on the given medical data, are individuals living at poverty more likely to have longer hospital admissions because of financial ailments that prevent them from seeking medical treatment early on in an illness? I will be using k-means clustering to answer this question.
2) The main goal of this analysis is to determine if patients with higher income spend less time in the hospital than those living at or below the poverty line.

Part II: Technique Justification

B. Reason for Chosen Clustering Method

1) The k-means clustering model is a very efficient way to group unlabeled data into distinct groups, or clusters. The k-means model looks at given data points and attempts to group them based on similarity. In a k-means cluster model, a predetermined number of groups is established. Then a centroid for each group is assigned, allowing the algorithm to perform an iterative process that groups individual points based on distance to each centroid.
2) One assumption of k-means clustering is that there is no specific predefined or optimal number of clusters. The model is therefore only as good as determined by the analyst when selecting a number of clusters. This begins with an analyst's knowledge of the dataset and the goals for analysis (Roth & Amor 2017).
3)

Pandas as pd	Manipulate / create data frames.
Numpy as np	Create arrays and mathematical techniques.
Matplotlib.pyplot as plt	Visualizations.
Seaborn as sns	Visualizations.
Sklearn.cluster import KMeans	Clustering technique used.
Sklearn.preprocessing import StandardScaler	Normalize the data the that the clustering algorithm performs more smoothly.

Sklearn.metrics import silhouette_score	Clustering score the helps inform most efficient number of clusters that should be used.

Part III: Data Preparation

C. Preparing the Data

1) An important preprocessing goal, for performing the k-means clustering technique, is to ensure the data has been normalized/scaled. What this means is that the variables being used for analysis are standardized into similar scales to produce efficient results from the cluster model. The StandardScaler method from Sklearn is used for this preprocessing step. The goal in this step is to prevent skewing or misleading clustering results by improving distance calculations based on a scaled data frame.

2)

Income	Continuous Variable
Initial_days	Continuous Variable

3) Steps to prepare the data for analysis:
    i. Import necessary packages and import CSV file.
    ii. Ensure there are no missing or duplicate values, treat if so.
    iii. Select variables for analysis and examine the scatterplot to pre-determine the number of clusters.
    iv. Create a data frame for analysis, based on chosen variables, obtain z-scores and scale the data using StandardScaler.
    v. Review scaled data using .describe() to confirm data has been scaled.

4) *Clean data included in task 1 submission.*

Part IV: Analysis

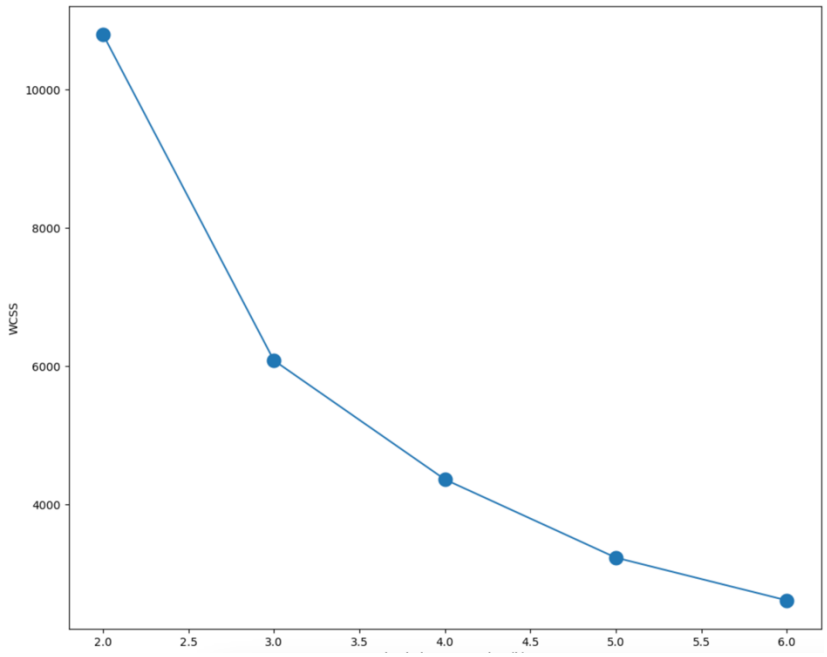
D. Data Analysis Report

1) After preparing the data and performing the preprocessing steps, I was able to determine the optimal number of k-clusters for this analysis. I first ran a k-means model based on my pre-determined number of k-clusters, 2. I then

created a centroid data frame to determine cluster centers, which allowed the clustering model to select and assign data points to each group. After creating a visualization to examine my cluster groups, I used two techniques to help me confirm what an ideal number of clusters would be for this analysis. First, I used an elbow plot. An elbow plot allows me to graph the wcss (within-cluster sum of square) and my k-values. The result is a line plot that at some point will begin to abruptly decrease in value (see below). The elbow of my graph began to drop off at 3 clusters, thus showing me that 3 clusters, rather than 2, should be used for analysis. Second, I used a silhouette plot to confirm the results from the elbow plot. The silhouette plot indeed confirms that 3 clusters are optimal for analysis. A screenshot is also provided below.

```
#Determine ideal number of clusters using elbow plot
wcss=[]
for k in range(2, 7):
    model=KMeans(n_clusters=k, n_init=50, random_state=300)
    model.fit(scaled_data)
    wcss.append(model.inertia_)
wcss_s=pd.Series(wcss, index=range(2, 7))

plt.figure(figsize=(12, 10))
ax=sns.lineplot(y=wcss_s, x=wcss_s.index)
ax=sns.scatterplot(y=wcss_s, x=wcss_s.index, s=200)
ax=ax.set(xlabel='Optimal Clusters Number (k)',
          ylabel= 'WCSS')
```



2) The code is included with the task 1 submission as a Jupyter notebook file.

Part V: Data Summary and Implications

E. Summarize the Data Analysis

1) When it came to determining the ideal number of clusters for analysis in the k-means model, I used the silhouette plot and calculated a silhouette score to understand the quality of my cluster model. A silhouette score will be in the range of -1 to 1. For my model, the silhouette score came to 0.50, indicating a good model of clusters as it is closer to positive one. The three clusters are therefore classified as good, appear dense, and are distinguishable until the cluster groups start getting close together.

```
#Get silhouette score to reconfirm ideal number of clusters for model
from sklearn.metrics import silhouette_score
sil_score= silhouette_score(scaled_data, k_model.labels_)
sil_score
```

```
0.5015402316355737
```

2) With the clusters divided into three groups now, hospital executives can now take a deeper look into income versus length of hospital admission. It seems that income and length of admission are correlated, with those making more money seeing less time spent in the hospital. The reason for such a correlation would require further research and most likely cannot be determined by one factor. As I mentioned before, those living at or below the poverty line may delay medical treatment because of financial constraints, making their medical conditions more severe and when they eventually present to the hospital, require more in depth treatment and longer hospital admission.
3) One limitation of this analysis could be inaccurately utilizing the most efficient number of clusters. If an analyst only considers their predetermined cluster numbers, the model could poorly guide decision making for stakeholders. Therefore, it would be wise to use an elbow plot and silhouette score to confirm how many clusters would be most effective for running a k-means cluster model.

4) With this information, hospital executives could team up with local community leaders and come up with a plan to help those of lower income get treatment they need and encourage community members to not wait out an illness, via community outreach and education. Lengthy hospital admissions can negatively impact a facility, either by fines from Medicare and Medical services, increasing hospital acquired infections, and overall operations due to inpatient rooms being occupied long-term.

Part VI: Demonstration

F. Panopto Video

1) Link included with task 1 submission.

G. Web Sources

Roth, P., & Amor, A. (2007). Demonstration of K-means assumptions. scikit-learn. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html