Manoj Kashyap

Northeastern university | EAI 6010

Final Project

Course Information

Course Title: Applications of Artificial Intelligence

Course Number EAI6010

Term and Year: Winter Part B:

Start and End Dates: March 2 – April 10, 2021

STUDENT NAME: Manoj Kashyap

ASSIGNMENT: 6 (Final)

**Look into Data:**

We shall work in python environment; the dataset is downloaded from Canvas website. The dataset has 5000 observation and 59 attributes. A glimpse of dataset and attributes are given below:

Table

Description automatically generated

Graphical user interface, application

Description automatically generated

**FIG: DATASET**

Text

Description automatically generatedGraphical user interface

Description automatically generated with low confidenceA close-up of a page

Description automatically generated with low confidence

Text

Description automatically generated

**FIG: ATTRIBUTES**

As a **marketing manager**, I would define market segmentation as “**dividing the customers into segment of groups** ” this data set based on,

1. Demographic Information – Gender, Age,
2. Geographical Information – Region, Town size
3. Behavioral Data – Education Years, Employment length

Lets choose the following attributes

* **Age** –. People between 15 – 30 are more focused on social media. To access social media internet is required. The quantity of internet used by teenagers is on a higher side. It makes sense for the company to target a high data volume to teenagers. People above 30 may look for broader and high-speed internet connectivity.
* **Gender** - Both the genders have different needs, for Example - Female customers tends to speak a greater number of minutes in phone than male customers. It makes sense for company to provide a plan with free talk time to Female customers
* **Credit Debt** and **Employment length**

This is a behavioral data; I feel the number of education years can decide the loyalty of customers. If the person can study for so many years, the person might be loyal to particular brands. The probability of a person being loyal to the telecommunication company is on a higher side.

* The mathematical model *K – Means* clustering is an unsupervised machine learning. Since it works on **Euclidian distance,** the attributes should be numerical.
* Town Size – Number of people play an important role because it decides the number of towers that needs to be installed

Now that the attributes are decided let’s have a look at the statistics of each features/attributes.

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram, box and whisker chart

Description automatically generated

**FIG: SUMMARY STATISTICS OF ATTRIBUTES USING BOX PLOT**

**Correlation graph**

Chart

Description automatically generated

**FIG: Pearson correlation**

Observation

A correlation between Region and town size is observed. A strong correlation between Age and Employment length is observed A correlation between age and cardtenure is observed A strong correlation between age and incometenure is also observed A strongest correlation between Credit debt and other debt is also observed.There are some negative correlations observed as well.

The data was standardized using sklearn module for the selected attributes

Table

Description automatically generated

Table

Description automatically generated

**FIG : THE DATA THAT WE ARE CONSIDERING FOR CLUSTTERING AND THE DATA THAT’S SCALED USING SK LEARN**

After scaling

Chart, line chart

Description automatically generated

**FIG: A graph between Inertia vs number of clusters. The elbow observed at 5**

Table

Description automatically generated

By setting the k value = 5, 5 clusters were generated 0,1,2,3,4, the cluster is the targeted variable.

***Hence we have created a target variable (i.e. cluster) , when plotted on a 2D graph the following plot was obtained.***

Text, table

Description automatically generated

The CODE lines were added and modified to generate a graph on a 2D plane.

Chart, scatter chart

Description automatically generated

**FIG: Clustered centered graph of 5-clusters. The code for this graph is mentioned above**

**Chart, line chart

Description automatically generated**

**FIG: Silhouette score for K means clustering, as you see the silhouette is max at k = 3.0 and min at k = 2.0**

*Chart

Description automatically generated*

**FIG: Feature importance and confusion matrix using Random forest**

Segmentation and Profiling: The word segmentation means “dividing people into groups based on an attribute”. Market segmentation aims for homogenous group from a heterogenous market based on attributes like geographic, psychographic, demographic etc.

Profiling answers the question “How do these segments differentiate each other based on variables”?

As a marketing manager I would count of each segment, then find overall average and segment wise average for each attribute and perform calculation for each K value, to find the best profiling

**Summary:**

K-means clustering is one of the simple ways of well-known unsupervised machine learning algorithms. In unsupervised algorithms, we make inferences from datasets using input vectors where the predicting class labels are unknown. Here we group similar data points together and discover underlying patterns, coming to K-means, at a certain stage we choose the k-value which we consider as the number of output labels.

* So in the initial step of K-means clustering, we decide the number of clusters we need to go with
* The algorithm randomly selects k-objects from the data which is taken as initial center centroids
* Then after each of the data points will be assigned to the class of the nearest centroid based on the distance calculated, where Euclidian distance is considered as a common to calculate the distance
* This process keeps happening iteratively over each of the data points over all of the centroid selected, and will be the computational multiples of n\*m where ‘n’ is the number of centroids we considered i.e. the selected k-value and ‘m’ is the total number of data points
* This iterative process keep continuing until clusters formed in the current iteration are almost same as of the previous iteration, indicating there is not much change in the clusters

Chart, scatter chart

Description automatically generated

The above example shows the way the clusters change over iteration and finalize it. Sine the considered example is a two-class clustering, only two classes are considered. For the multiclass classification, the same process iterates with the changing hyperplane in the multidimension, and the dimensionality depends on the cluster size we consider.

The dataset we considered is a telecommunication data where the task is to develop customer segmentation that can support effective, and economically sound, customer retention efforts.

* The dataset we considered has 5000 rows and 59 columns, which does not contain any duplicate values
* Null values of the numeric values are changed accordingly, for the continuous numeric attributes the null values are replaced with the mean value and for the finite value numeric attributes the null values are chosen randomly since we had no background information over those
* We observe that mean value of the customer’s age is around 48 that varies over a span of 30 to 63, it is also observed few are around 20 and near 80
* Town size seems to be label encoded since the values are ranged between 1 to 5
* In total we have 21 categorical variables, 16 numeric variables and 22 Boolean variables
* We also see there is a correlation between certain features such as ‘PhoneCoTenure’ and ‘CardTenure’
* We do not have Null values as Null, but they are in the form of ‘#Null!’ which are detected and handled
* The variables with currency values are cleaned, and the converted to float data type since they were object type earlier so that we can perform computation upon it
* Required categorical variables such as ‘gender’, ‘LoanDefault\_num’ are numeric coded for the computational flexibility, converting Nominal variables to Ordinal variables is idea behind bringing categorical attributes under computation
* Variables such as ‘DebtToIncomeRatio’, ‘Age’ and ‘PhoneCoTenure\_Coded’ are binned into require range of bins for the compatible visualization and to know the distribution underlying the attributes
* There after we select the 5 attributes for the clustering based on different segmentation for Demographics ‘Age’ is considered, for the Geographic ‘Region’ and ‘TownSize’ are considered, for the Behavioral ‘CreditDebt’ and ‘EmploymentLength’ are considered
* Now the data that is of shape 5000 \* 5 is Standardized to keep the scalability and make sure one attribute will not dominate over the other
* Upon plotting the inertia for multiple cluster solutions against the number of clusters, the elbow graph is generated where we choose 5 to be the number of clusters we consider since the elbow bend appears at approximately 5
* We fit the data for K-means clustering with a k-value of 5, where the determined inertia on the fitted data is 11668.08
* Upon fitting the data with 5-means clustering, the class-wise datapoints obtained are 2-1405, 0-1373, 3-1079, 1-1058, 4-85
* So now we have converted the Unsupervised problem statement to a Supervised problem statement upon obtaining the Target labels
* Now we predict the class labels using k-means, upon considering the clusters we created as true values
* Upon checking the Silhouette Score k-Means clustering, the elbow appears to be at 3, where it says with a k of 3 and score 0.37, will be the optimal solution to consider to predict the values, considering score of 1 best and -1 as worst
* Upon applying Random forest to predict the generated cluster values, considering n\_estimators=100, we get an accuracy of 97.6%, where ‘Age’ and ‘TownSize’ are seen to be top most important features among considered 5 features