D212 K-Means Clustering

The purpose of this Data Mining Report is to be able to identify the principal variables of customers. The model used is a K-Means Clustering model. The goal of the report is to find valuable and actionable insights to satisfy business needs. The model is ideal since the data provided has many variables that require a model to determine any possible relationship. Once we have performed the Clustering then one question, we would ask is what marketing techniques can be applied or implemented.

The clustering technique used is called an Unsupervised Learning Model. This means that the data provided will be removed from all its labels and be fit into the Model to identify patterns or groups of data with similar characteristics and behaviors. The expected outcomes will be clusters that share similar characteristics. One of the assumption of the K-Means model is that the appropriate number of clusters need to be chosen. This assumption is determined by using what is called the “Elbow Method” to determine which number of clusters has the appropriate amount of the Sum of Squared Errors.

Eight python packages or libraries were used for the model. These packages included pandas and NumPy to import the data and be able to manipulate the data as a Data frame and then convert it to a 2d array. Seaborn and Matplotlib were used for visualization purposes. Sklearn was the library where the majority of packages were imported. These consisted of Kmeans, k\_means, plot\_tree, MinMaxScaler. These packages were used to prepare the data, instantiating the model, fitting the model, and evaluating the model.

One data preprocessing goal was to scale the feature variables. This step was crucial in preparing the data set as two of the variables consisted of extremely high dollar amounts. Scaling the variables consisted of determining the highest value in those columns and then dividing each value by the highest amount. This allows the ranges of the particular column to have a minimum value of 0 and the highest value to be 1. All other values are a decimal number below 1. This is extremely important in ensuring all features could be properly computed by the model in balanced way. Ensuring that the features that were scaled had values below 1 and the categorical features had values of 1s and 0s would ensure our Kmeans model would be balanced.

Of the 50 variables from the original spreadsheet, 58 were used for the model. The increase in variables from the model was due to so many columns from the original spreadsheet were categorical variables. This meant that many of these columns would get turned into dummy variables in order to be run into the model. These 58 variables were Population which was continuous variable, Children which was continuous variable, Age which was continuous variable, Income which was continuous variable, Outage\_sec\_perweek which was continuous variable, Email which was continuous variable, Contacts which was continuous variable, Yearly\_equip\_failure which was continuous variable, Tenure which was continuous variable,

MonthlyCharge which was continuous variable, Bandwidth\_GB\_Year which was continuous variable, State\_CA which was categorical variable, State\_NY which was categorical variable, State\_Other which was categorical variable, State\_PA which was categorical variable, State\_TX which was categorical variable, Area\_Rural which was categorical variable, Area\_Suburban which was categorical variable, Area\_Urban which was categorical variable, Marital\_Divorced which was categorical variable, Marital\_Married which was categorical variable, Marital\_Never Married which was categorical variable, Marital\_Separated which was categorical variable, Marital\_Widowed which was categorical variable, Gender\_Male which was categorical variable, Churn\_Yes which was categorical variable, Techie\_Yes which was categorical variable, Contract\_Annual which was categorical variable, Contract\_Month-to-month which was categorical variable, Port\_modem\_Yes which was categorical variable, Tablet\_Yes which was categorical variable, InternetService\_Fiber Optic/DSL which was categorical variable, InternetService\_None which was categorical variable, Phone\_Yes which was categorical variable, Multiple\_Yes which was categorical variable, OnlineSecurity\_Yes which was categorical variable, OnlineBackup\_Yes which was categorical variable, DeviceProtection\_Yes which was categorical variable, TechSupport\_Yes, StreamingTV\_Yes, StreamingMovies\_Yes which was categorical variable, PaperlessBilling\_Yes which was categorical variable, PaymentMethod\_Electronic Payment which was categorical variable, PaymentMethod\_Mailed Check which was categorical variable.

The steps used to prepare the data for analysis was to first remove all variables that were not relevant to the model. This consisted in dropping the ID number and many other variables such as latitude and longitude. The code segment used to drop these variables was: A picture containing diagram

Description automatically generated

The second step was to convert all the categorical variables into binary variables. This was done by using pd.get\_dummies on the model. However, this presented a problem as pd.get\_dummies creates two columns for Yes or NO. In order to prevent collinearity, the No columns were dropped. This was the code to convert the categorical variables into continuous variables and to drop all the No columns.

Text, letter

Description automatically generated

The analysis technique used for the model was to first fit the data to the model and run it using clusters from 1 to 40. This was used to find out the Inertia score for the model at different cluster amounts. Inertia tells how far away the points within a cluster are. Therefore, a small of inertia is aimed for. The range of inertia’s value starts from zero and goes up (Amelia, 2018). Attached is a screenshot of the code segment that ran the Inertia score by the number of clusters and the Elbow Plot to determine the best number of clusters.

Text

Description automatically generated with low confidence Graphical user interface

Description automatically generated with low confidence

The best amount of clusters to use was 7 as that was where the Inertia score began show little change. Another method used to analyze the data was to run Principal Component Analysis on the model. This was necessary since the amount of variables being used in the model was 58 and a method needed to be applied in order to reduce the Inertia score. Attached is the code for the PCA portion of the model. 26 components were chosen as 26 components were able to explain 95% of the variance.

A picture containing table

Description automatically generated

Chart, line chart

Description automatically generated

Once 26 principal components were selected, the model was then reran using 7 Clusters. Using PCA we were able to reduce the Inertia score from 1,396,322,222,298 to 50,451. This drastic change shows how valuable PCA was in ensuring the model would cluster the data in an appropriate manner.

Text

Description automatically generated

The accuracy of the model is determined by the reduced Inertia and using the elbow method to determine which number of Clusters would be accurate for the model. The elbow method shows us that 7 clusters was ideal for the model. The results show that of the 7 clusters, cluster 5 appears to have the biggest distribution from the dataset. Table below shows how the clusters are distributed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
| 985 | 1521 | 1520 | 1276 | 2036 | 1159 | 1503 |

Another observation is that clusters 1,3,4, and 7 have their contract set up on a month to month bases instead of a contract. Attached is a graph that displays this discrepancy.

Chart, bar chart

Description automatically generated

The implication that the clusters show would be to use Marketing to create traits for each cluster. Once these traits are highlighted for each cluster, then Marketing can decide on which cluster to focus on and make sure all the traits are specifically highlighted when targeting each cluster. For example, using the chart above, if we want to target cluster 5 then we should ensure that our marketing campaign specifically mentions how great our Annual Contracts are compared to our competitor. This would ensure that cluster 5 which had the highest distribution of our data is specifically targeted. Using the other insights from the data we would make sure the marketing campaign addresses any other insights from this cluster.

One limitation from the data was the amount of continuous and categorical variables. Usually, if a dataset has all continuous variables, then we would use specifically the K-Means algorithm. If we had all categorical variables, we would use Hierarchical clustering. Since we had a combination of both, then K-Means Clustering was chosen.

The recommended course of action is to use the Clusters for Marketing purposes. If the cable company were to specifically target certain clusters, then we would be able to see an increase in customers.