K - Means Clustering

Instructions

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

Name: Nukala Ayyappa Bharthwaj

Batch Id: DSWDMCOS 21012022

Topic: K Means Clustering

1. Business Problem

1.1. Objective

1.2. Constraints (if any)

2. Work on each feature of the dataset to create a data dictionary as displayed in the below image:

Name of Feature	Description	Type	Relevance
ID	Customer ID	Quantitative, Nominal	Irrelevant, ID does not provide useful information

Using R and Python codes perform:

3. Data Pre-processing

2.1 Data Cleaning, Feature Engineering, etc.

4. Exploratory Data Analysis (EDA):

4.1. Summary

4.2. Univariate analysis

4.3. Bivariate analysis

5. Model Building

5.1 Build the model on the scaled data (try multiple options)

5.2 Perform the K- means clustering, visualize the clusters using scree plot

5.3 Validate the clusters (try with different no. of clusters) – label the clusters and derive insights (compare the results from multiple approaches)

© 2013 - 2020 360DigiTMG. All Rights Reserved.

6. Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

Note:

The assignment should be submitted in the following format:

- R code
- Python code
- Code Modularization should be maintained
- Documentation of the modules (elaborating on steps mentioned above)

1.) Perform clustering (K means clustering) for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.

Sol:

Business Objective: to perform K means clustering on the east west airlines data set

Data Types: the given data and its types are as follows:

Name of feature	Description	Data type	Relevance
ID	Id of the flight	Ordinal	Irrelevant since it's a id of the flight
Balance	Balance amount	Ratio	Relevant
Qual_miles	Number of miles travelled	Ratio	Relevant
cc1_miles	CC1 miles of flight	Ratio	Relevant
Cc2_miles	CC2 miles of flight	Ratio	Relevant
Cc3_miles	CC3 miles of flight	Ratio	Relevant
Bonus_miles	Bonus miles of the flight	Ratio	Relevant
Bonus_trans	Bonus transc. Of flight	Ratio	Relevant
Flight_miles_12mo	12 months flight miles	Ratio	Relevant
Flight_trans_12	12months flight transc.	Ratio	Relevant
Days_since_enroll	Number of days since flight enrolled	Ratio	Relevant
Award?	Whether it got award or not	Nominal	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Data Pre Processing: all the variables in the given data is used for applying the clustering except the ID column, since it is not useful.

Exploratory Data Analysis: mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

k-Means clustering: after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 3 and based on that the complete data is clustered.

	ID.	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enroll	Award.
1	1	28143	0	1	1	1	174	1	0	0	7000	0
2	2	19244	0	1	1	1	215	2	0	0	6968	0
3	3	41354	0	1	1	1	4123	4	0	0	7034	0
4	4	14776	0	1	1	1	500	1	0	0	6952	0
5	5	97752	0	4	1	1	43300	26	2077	4	6935	1
6	6	16420	0	1	1	1	0	0	0	0	6942	0
7	7	84914	0	3	1	1	27482	25	0	0	6994	0
8	8	20856	0	1	1	1	5250	4	250	1	6938	1
9	9	443003	0	3	2	1	1753	43	3850	12	6948	1
10	10	104860	0	3	1	1	28426	28	1150	3	6931	1
11	11	40091	0	2	1	1	7278	10	0	0	6959	0
12	12	96522	0	5	1	1	61105	19	0	0	6924	1
13	13	43382	0	2	1	1	11150	20	0	0	6924	0
14	14	43097	0	1	1	1	3258	6	0	0	6918	0
15	15	17648	0	1	1	1	0	0	0	0	6912	0

2.) Perform clustering for the crime data and identify the number of clusters formed and draw inferences. Refer to crime_data.csv dataset.

Sol:

Business Objective: to perform K means clustering on the crime data set

Data Types: the given data and its types are as follows:

Name of feature	Description	Data type	Relevance
X	Name of the murderer	Nominal	Irrelevant since name of the person
Murder	Murder rate	Ratio	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Assult	Assault rate	Ratio	Relevant
UrbanPop	Urban pop rate	Ratio	Relevant
Rape	Rape rate	Ratio	Relevant

Data Pre Processing: all the variables in the given data is used for applying the clustering except the X column, since it is not useful.

Exploratory Data Analysis: mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

k-Means clustering: after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 4 and based on that the complete data is clustered.

	X	Murder	Assault	UrbanPop	Rape
1	Alabama	13.2	236	58	21.2
2	Alaska	10.0	263	48	44.5
3	Arizona	8.1	294	80	31.0
4	Arkansas	8.8	190	50	19.5
5	California	9.0	276	91	40.6
6	Colorado	7.9	204	78	38.7
7	Connecticut	3.3	110	77	11.1
8	Delaware	5.9	238	72	15.8
9	Florida	15.4	335	80	31.9
10	Georgia	17.4	211	60	25.8
11	Hawaii	5.3	46	83	20.2
12	Idaho	2.6	120	54	14.2
13	Illinois	10.4	249	83	24.0

3.) Analyze the information given in the following 'Insurance Policy dataset' to create clusters of persons falling in the same type. Refer to Insurance Dataset.csv

Sol:

Business Objective: to perform K means clustering on the insurance data set

Data Types: the given data and its types are as follows:

Name of feature	Description	Data type	Relevance
Premiums Paid	Premiums paid by customer	Ratio	Relevant
Age	Age of customer	Ratio	Relevant
Days to Renew	Number of days to renew	Ratio	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Claims made	Claim amount made	Ratio	Relevant
Income	Income of customer	Ratio	Relevant

Data Pre Processing: all the variables in the given data is used for applying the clustering.

Exploratory Data Analysis: mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

k-Means clustering: after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 3 and based on that the complete data is clustered.

	Premiums.Paid	Age	Days.to.Renew	Claims.made	Income
1	2800	26	233	3890.076	28000
2	2950	27	130	2294.444	29500
3	3100	28	144	2564.545	31000
4	3250	30	65	1978.261	32500
5	3400	32	56	2009.091	34000
6	3550	35	89	2349.455	35500
7	3700	44	95	2503.346	37000
8	3850	45	48	2217.405	38500
9	4000	46	76	2527.778	40000
10	6225	56	200	6908.232	41500
11	6450	67	211	7672.549	43000
12	6675	69	245	10208.824	44500
13	6900	70	261	12192.233	46000

4.) Perform clustering analysis on the telecom dataset. The data is a mixture of both categorical and numerical data. It consists the number of customers who churn. Derive insights and get possible information on factors that may affect the churn decision. Refer to Telco_customer_churn.xlsx dataset.

Hint:

- Perform EDA and remove unwanted columns.
- Use Gower dissimilarity matrix and In R use daisy() function.

Customer ID	Count	Quarter	Referred a Friend	Number of Referrals	Tenure in Months	Offer	Phone Service	Avg Monthly Long Distance Charges	Multiple Lines	Internet Service	Internet Type	Avg Monthly GB Download	Online Security	Online Backup	Device Protection	Premium Tech	Streaming TV	Streaming Movies	Streaming Music
8779-QRDMV	1	Q3	No	0	1	None	No	0	No	Yes	DSL	8	No	No	Yes	No	No	Yes	No
7495-OOKFY	1	Q3	Yes	1	8	Offer E	Yes	48.85	Yes	Yes	Fiber Optic	17	No	Yes	No	No	No	No	No
1658-BYGOY	1	Q3	No	0	18	Offer D	Yes	11.33	Yes	Yes	Fiber Optic	52	No	No	No	No	Yes	Yes	Yes
4598-XLKNJ	1	Q3	Yes	1	25	Offer C	Yes	19.76	No	Yes	Fiber Optic	12	No	Yes	Yes	No	Yes	Yes	No
4846-WHAFZ	1	Q3	Yes	1	37	Offer C	Yes	6.33	Yes	Yes	Fiber Optic	14	No	No	No	No	No	No	No
4412-YLTKF	1	Q3	No	0	27	Offer C	Yes	3.33	Yes	Yes	Fiber Optic	18	No	No	Yes	No	No	No	No
0390-DCFDQ	1	Q3	Yes	1	1	Offer E	Yes	15.28	No	Yes	Fiber Optic	30	No	No	No	No	No	No	No
3445-HXXGF	1	Q3	Yes	6	58	Offer B	No	0	No	Yes	DSL	24	No	Yes	Yes	No	No	Yes	No
2656-FMOKZ	1	Q3	No	0	15	Offer	Yes	44.07	Yes	Yes	Fiber Optic	28	No	No	No	No	No	No	No
2070-FNEXE	1	Q3	No	0	7	Offer E	Yes	26.95	No	Yes	Fiber Optic	18	Yes	No	No	No	No	No	No

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Sol:

Business Objective: to perform K means clustering on the tele customer data set

Data Types: the given data and its types are as follows:

Name of feature	Description	Data type	Relevance
Customer ID	Id of the customer	Nominal	Irrelevant
Count	Count number	Nominal	Relevant
Quarter	Type of quarter	Nominal	Relevant
Referred a Friend	Whether yes or no	Nominal	Relevant
Number of Referrals	Number of referals	Internal	Relevant
Tenure in Months	Tenure of the package	Ratio	Relevant
Offer	Type of offer given	Nominal	Relevant
Phone Service	Whether yes or no	Nominal	Relevant
Avg Monthly Long Distance Charges	Long distance charges for the udser	Ratio	Relevant
Multiple Lines	Whether yes or no	Nominal	Relevant
Internet Service	Whether yes or no	Nominal	Relevant
Internet Type	Type of internet	Nominal	Relevant
Avg Monthly GB Download	It is total GB downloaded in month	Ratio	Relevant
Online Security	Whether yes or no	Nominal	Relevant
Online Backup	Whether yes or no	Nominal	Relevant
Device Protection Plan	Whether yes or no	Nominal	Relevant
Premium Tech Support	Whether yes or no	Nominal	Relevant
Streaming TV	Whether yes or no	Nominal	Relevant
Streaming Movies	Whether yes or no	Nominal	Relevant
Streaming Music	Whether yes or no	Nominal	Relevant
Unlimited Data	Whether yes or no	Nominal	Relevant
Paperless Billing	Whether yes or no	Nominal	Relevant
Payment Method	Type of payment	Nominal	Relevant
Monthly Charge	Monthly charge to user	Ratio	Relevant
Total Charges	Total charge to user	Ratio	Relevant
Total Refunds	Total refunds to user	Ratio	Relevant
Total Extra Data Charges	Total extra data charge to user	Ratio	Relevant
Total Long Distance Charges	Total long distance chrge to user	Ratio	Relevant
Total Revenue	Total revenue to user	Ratio	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Data Pre Processing: all the variables in the given data is used for applying the clustering except the customer id and the columns which are non numeric are converted into numeric data using dummy variables.

Exploratory Data Analysis: mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

k-Means clustering: after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 4 and based on that the complete data is clustered.

5.) Perform clustering on mixed data convert the categorical variables to numeric by using dummies or Label Encoding and perform normalization techniques. The data set consists details of customers related to auto insurance. Refer to Autoinsurance.csv dataset.

Sol:

Business Objective: to perform K means clustering on the auto insurance data set

Data Types: the given data and its types are as follows:

Name of feature	Description	Data type	Relevance
Customer	Id of the customer	Nominal	Irrelevant
State	State of the customer	Nominal	Relevant
Customer Lifetime Value	Life time value of the customer	Ratio	Relevant
Response	Whether yes or no	Nominal	Relevant
Coverage	Type of coverage	Nominal	Relevant
Education	Education of cust.	Nominal	Relevant
Effective To Date	Insurance effective date	Nominal	Relevant
EmploymentStatus	Employment status of cust	Nominal	Relevant
Gender	Gender of cust.	Nominal	Relevant
Income	Income of the cust.	Ratio	Relevant
Location Code	Location code of the cust.	Nominal	Relevant
Marital Status	Marital status of cust.	Nominal	Relevant
Monthly Premium Auto	Monthly premium auto amount	Ratio	Relevant
Months Since Last Claim	No of months since last claim	Ratio	Relevant
Months Since Policy Inception	No of months since policy inception	Ratio	Relevant

© 2013 - 2020 360DigiTMG. All Rights Reserved.

Number of Open Complaints	No of open complaints on cust.	Ratio	Relevant
Number of Policies	No of policies of the cust.	Ratio	Relevant
Policy Type	Policy type of cust.	Nominal	Relevant
Policy	Policy amount of cust.	Ratio	Relevant
Renew Offer Type	Renew offer type of cust.	Nominal	Relevant
Sales Channel	Type of sales channel	Nominal	Relevant
Total Claim Amount	Total claim amount of the cust.	Ratio	Relevant
Vehicle Class	Vehicle class of cust.	Nominal	Relevant
Vehicle Size	Vehicle size of cust.	Nominal	Relevant

Data Pre Processing: all the variables in the given data is used for applying the clustering except the customer column, since it is not useful. The data which is in the form of categorical type is converted into numeric type so that it can be used for the clustering.

Exploratory Data Analysis: mean, median, variance, standard deviation, skewness, kurtosis is calculated for all the variables of the given data then normalization for all the variables of the data to apply k Means clustering.

k-Means clustering: after cleaning the complete data K means clustering is applied on the data and number of clusters is in front decided as 5 and based on that the complete data is clustered.

Customer	State	Customer	Response	Coverage	Education	Effective To Date	Employme	Gender	Income	Location C	Marital Sta	Monthly Pi	Months Sir	Months Sir	Number of	Number of	Policy Type	Policy	Renew Off	Sales Char	Total Clair	Vehicle Cla	Vehicle Siz
BU79786	Washingto	2763.519	No	Basic	Bachelor	2/24/2011	Employed	F	56274	Suburban	Married	69	32	5	0	1	Corporate	Corporate	Offer1	Agent	384.8111	Two-Door	Medsize
QZ44356	Arizona	6979.536	No	Extended	Bachelor	1/31/2011	Unemploy	F	0	Suburban	Single	94	13	42	0	8	Personal A	Personal L	Offer3	Agent	1131.465	Four-Door	Medsize
AI49188	Nevada	12887.43	No	Premium	Bachelor	2/19/2011	Employed	F	48767	Suburban	Married	108	18	38	0	2	Personal A	Personal L	Offer1	Agent	566.4722	Two-Door	Medsize
WW63253	California	7645.862	No	Basic	Bachelor	1/20/2011	Unemploy	M	0	Suburban	Married	106	18	65	0	7	Corporate	Corporate	Offer1	Call Cente	529.8813	SUV	Medsize
HB64268	Washingto	2813.693	No	Basic	Bachelor	3/2/2011	Employed	M	43836	Rural	Single	73	12	44	0	1	Personal A	Personal L	Offer1	Agent	138.1309	Four-Door	Medsize
OC83172	Oregon	8256.298	Yes	Basic	Bachelor	1/25/2011	Employed	F	62902	Rural	Married	69	14	94	0	2	Personal A	Personal L	Offer2	Web	159.383	Two-Door	Medsize
XZ87318	Oregon	5380.899	Yes	Basic	College	2/24/2011	Employed	F	55350	Suburban	Married	67	0	13	0	9	Corporate	Corporate	Offer1	Agent	321.6	Four-Door	Medsize
CF85061	Arizona	7216.1	No	Premium	Master	1/18/2011	Unemploy	M	0	Urban	Single	101	0	68	0	4	Corporate	Corporate	Offer1	Agent	363.0297	Four-Door	Medsize
DY87989	Oregon	24127.5	Yes	Basic	Bachelor	1/26/2011	Medical Le	M	14072	Suburban	Divorced	71	13	3	0	2	Corporate	Corporate	Offer1	Agent	511.2	Four-Door	Medsize
BQ94931	Oregon	7388.178	No	Extended	College	2/17/2011	Employed	F	28812	Urban	Married	93	17	7	0	8	Special Au	Special L2	Offer2	Branch	425.5278	Four-Door	Medsize
SX51350	California	4738.992	No	Basic	College	2/21/2011	Unemploy	M	0	Suburban	Single	67	23	5	0	3	Personal A	Personal L	Offer1	Agent	482.4	Four-Door	Small
VQ65197	California	8197.197	No	Basic	College	6/1/2011	Unemploy	F	0	Suburban	Married	110	27	87	0	3	Personal A	Personal L	Offer2	Agent	528	SUV	Medsize
DP39365	California	8798.797	No	Premium	Master	6/2/2011	Employed	M	77026	Urban	Married	110	9	82	2	3	Corporate	Corporate	Offer2	Agent	472.0297	Four-Door	Medsize
SJ95423	Arizona	8819.019	Yes	Basic	High Schoo	10/1/2011	Employed	M	99845	Suburban	Married	110	23	25	1	8	Corporate	Corporate	Offer2	Branch	528	SUV	Medsize

© 2013 - 2020 360DigiTMG. All Rights Reserved.