

Analyzing Customer Churn with k-means Clustering

Introduction:

The goal of this project is to determine if it is possible to predict if a customer is likely to switch telecommunications providers. For this project I will use k-means clustering. According to EDUCBA (2020) "This algorithm is an iterative algorithm that partitions the dataset according to their features into K number of predefined non-overlapping distinct clusters or subgroups. It makes the data points of inter clusters as similar as possible and also tries to keep the clusters as far as possible."

Part One: Research Question

Can k-means clustering, with a data set containing both continuous and categorical variables, be used to determine the likelihood of an event occurring?

Step One: Install the necessary packages

In [3]:

```
library(readxl)
library(class)
library(tidyverse)
library(caret)
```

Warning message:

"package 'readxl' was built under R version 3.6.3"

Warning message:

"package 'class' was built under R version 3.6.3"

Warning message:

"package 'tidyverse' was built under R version 3.6.3"

-- Attaching packages

----- tidyverse 1.3.0 -----

```
v ggplot2 3.3.2    v purrr   0.3.4
v tibble  3.0.4    v dplyr   1.0.2
v tidyr   1.1.2    v stringr 1.4.0
v readr   1.4.0    v forcats 0.5.0
```

Warning message:

"package 'ggplot2' was built under R version 3.6.3"

Warning message:

"package 'tibble' was built under R version 3.6.3"

Warning message:

"package 'tidyr' was built under R version 3.6.3"

Warning message:

"package 'readr' was built under R version 3.6.3"

Warning message:

```
"package 'purrr' was built under R version 3.6.3"
Warning message:
"package 'dplyr' was built under R version 3.6.3"
Warning message:
"package 'stringr' was built under R version 3.6.3"
Warning message:
"package 'forcats' was built under R version 3.6.3"
```

```
-- Conflicts -----
----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
```

```
Warning message:
"package 'caret' was built under R version 3.6.3"
Loading required package: lattice
```

```
Warning message:
"package 'lattice' was built under R version 3.6.3"
```

```
Attaching package: 'caret'
```

```
The following object is masked from 'package:purrr':
```

```
lift
```

```
In [4]: install.packages("factoextra")
```

```
package 'factoextra' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
C:\Users\ContactTracer\AppData\Local\Temp\RtmpMtmgwI\downloaded_packages
```

```
In [5]: install.packages("cluster")
```

```
package 'cluster' successfully unpacked and MD5 sums checked
```

```
The downloaded binary packages are in
C:\Users\ContactTracer\AppData\Local\Temp\RtmpMtmgwI\downloaded_packages
```

```
In [6]: install.packages("corrplot")
```

```
Warning message:
"package 'corrplot' is in use and will not be installed"
```

```
In [7]: library(corrplot)
```

```
In [8]: library(factoextra)
library(cluster)
```

```
Warning message:
"package 'factoextra' was built under R version 3.6.3"
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
Warning message:
"package 'cluster' was built under R version 3.6.3"
```

```
In [9]: require(caTools)
```

Loading required package: caTools

Warning message:

"package 'caTools' was built under R version 3.6.3"

Part II: Technique Justification

K-means clustering uses a technique called customer segmentation, which divides customers into groups based on specific information. "Broadly speaking, the goal of customer segmentation is to divide customers into groups that share certain characteristics. It is crucial for a company to understand its customer behavior and categorize customers based on their user behavior. With good segmentation, companies could develop different and appropriate strategies for each group of customers accordingly and achieve the best effect such as targeting customers effectively and improving the customer experience etc." (Gong, 2020)

Step Two: Import/clean the data

```
In [10]: churn_clean <- read.csv("C:/Users/ContactTracer/Desktop/D212 Data Mining II Task 1/chur
```

```
In [11]: churn_clean_new <- churn_clean %>% select(15, 16, 17, 20, 21, 24, 40, 41, 42)
```

```
In [12]: str(churn_clean_new)
```

```
'data.frame':  10000 obs. of  9 variables:
 $ Children      : int  0 1 4 1 0 3 0 2 2 1 ...
 $ Age           : int  68 27 50 48 83 83 79 30 49 86 ...
 $ Income        : num  28562 21705 9610 18925 40074 ...
 $ Churn         : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 2 2 1 1 ...
 $ Outage_sec_perweek : num  7.98 11.7 10.75 14.91 8.15 ...
 $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
 $ Tenure        : num  6.8 1.16 15.75 17.09 1.67 ...
 $ MonthlyCharge  : num  172 243 160 120 150 ...
 $ Bandwidth_GB_Year : num  905 801 2055 2165 271 ...
```

```
In [13]: churn_clean_new$Churn <- as.numeric(as.factor(churn_clean_new$Churn))
```

```
In [14]: churn_data <- churn_clean_new[,4]
```

```
In [15]: churn_clean_new <- churn_clean_new[-4]
```

Assumption of k-means clustering:

According to (Dabbura, 2020) "Since clustering algorithms including kmeans use distance-based measurements to determine the similarity between data points, it's recommended to standardize the data to have a mean of zero and a standard deviation of one."

This means that due to the different variables having different scales, it is crucial to use the built-in scale function in R to scale all the data in the same way. A limitation of k-means clustering is that you must manually choose the number of clusters, i.e. the value of k in your model.

```
In [16]: churn_clean_scaled <- scale(churn_clean_new)
```

```
In [17]: churn_clean_final <- cbind(churn_clean_scaled, new_col = churn_data)
```

```
In [34]: write.csv(churn_clean_final, "C:\\Users\\ContactTracer\\Desktop\\churn_clean_final.csv",
```

Data Selection:

Step Three: Determine the Optimal Number of Clusters/Prepare the model

According to (Nallathambi, 2018), the best way to determine the optimal value of k is by using the `fviz_nbclust` function from the `factoextra` package in R.

```
In [ ]: fviz_nbclust(churn_clean_new, kmeans, method = "wss");
```

There is a problem with my version of Jupyter Notebook where it will not update, so I cannot run `fviz_nbclust` within Jupyter Notebook. I used R-Studio to run this part of the code and I was able to determine from the output graph that $k = 2$ is the ideal number of clusters.

```
In [18]: km <- kmeans(churn_clean_final, centers = 2, nstart = 25)
```

```
In [19]: km
```

K-means clustering with 2 clusters of sizes 5001, 4999

Cluster means:

	Children	Age	Income	Outage_sec_perweek	Yearly_equip_failure
1	-0.003779767	0.01942636	0.001751832	0.003240510	0.008364352
2	0.003781279	-0.01943413	-0.001752532	-0.003241806	-0.008367698

Tenure MonthlyCharge Bandwidth_GB_Year new_col

[illegible]

[4573]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4609]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4645]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4681]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4717]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4753]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4789]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4825]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4861]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4897]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4933]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4969]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	1
[5005]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5041]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5077]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5113]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5149]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5185]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5221]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5257]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5293]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[5329]	1	1	1	1	1																									

[illegible]

Within cluster sum of squares by cluster:
[1] 31531.17 31680.97
(between_SS / total_SS = 22.9 %)

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Analyzing the results of the k-means clustering model, Tenure and Bandwidth_GB_Year have the largest effect on if a customer will switch providers.

Next, a correlation matrix will be created as a check to make sure that the results of the k-means clustering make sense. (Facer, 2020)

```
In [21]: mydata.com
```

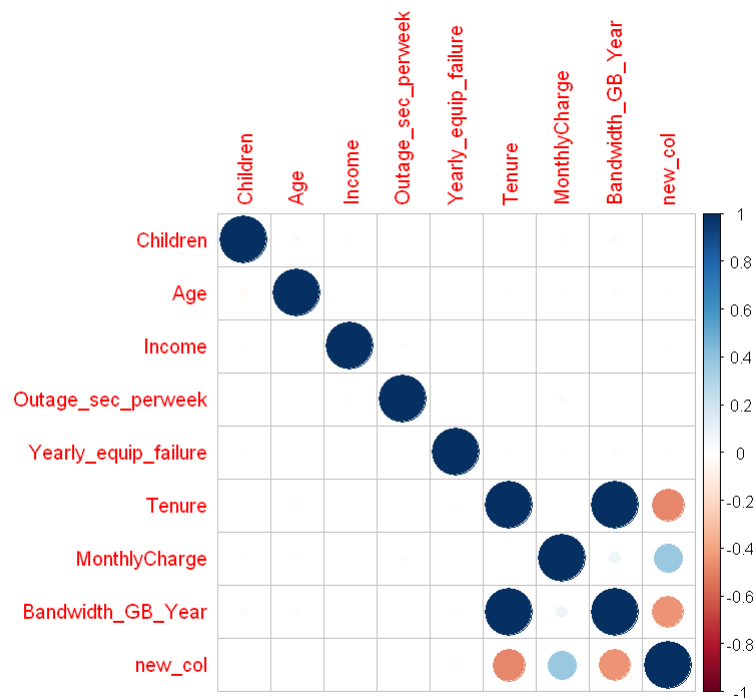
A matrix: 9 × 9 of type dbl

	Children	Age	Income	Outage_sec_perweek	Yearly equip_failure
Children	1.000000000	-0.029731540	0.009942354	0.0018892554	0.007320587
Age	-0.029731540	1.000000000	-0.004090602	-0.0080467191	0.008577348
Income	0.009942354	-0.004090602	1.000000000	-0.0100105457	0.005423276
Outage_sec_perweek	0.001889255	-0.008046719	-0.010010546	1.0000000000	0.002908726
Yearly equip_failure	0.007320587	0.008577348	0.005423276	0.0029087255	1.000000000
Tenure	-0.005091318	0.016979273	0.002114367	0.0029319584	0.012434911
MonthlyCharge	-0.009781399	0.010728512	-0.003013965	0.0204960735	-0.007172276

	Children	Age	Income	Outage_sec_perweek	Yearly_equip_failure
Bandwidth_GB_Year	0.025584816	-0.014723648	0.003673550	0.0041756614	0.012033693
new_col	-0.004264004	0.005629555	0.005937383	-0.0001564081	-0.015927129

The correlation matrix shows that Tenure and Bandwidth_GB_Year have the largest correlation with Churn. This result aligns perfectly with the result of the k-means clustering model.

```
In [22]: corrplot(mydata.cor)
```



Step Four: Split the data into A testing set, and a training set

Following (Kedia, 2018), the data is split into a training set and a test set.

```
In [23]: split1=sample.split(churn_clean_final[4:],SplitRatio=2/3)
```

```
In [24]: train=subset(churn_clean_final,split1==TRUE)
```

```
In [25]: test=subset(churn_clean_final,split1==FALSE)
```

```
In [26]: kmtrain <- kmeans(train, centers = 2, nstart = 25)
```

kmtrain

Cluster means:

Clustering vector:

[illegible]

[illegible]

[219]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4256]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4293]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4330]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4367]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4404]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4441]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4478]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4515]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4552]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4589]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4626]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4663]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4700]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4737]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4774]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4811]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4848]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4885]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4922]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4959]	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
[4996]	2	2	2	2	2																									


```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Step Five: Determine the accuracy of the model

```
In [30]: predicted_value <- c(1.467066, 1.060132)
```

```
In [31]: expected_value <- c(1.476877, 1.054540)
```

```
In [32]: acc <- abs(predicted_value - expected_value)*100
```

```
In [33]: mean(acc)
```

0.7701500000000003

The k-means clustering model is 77% accurate in predicting if a customer will switch providers.

Conclusion:

I have shown that it is possible to create a k-means clustering model that is highly accurate in predicting if a customer is at risk of switching service providers. From this analysis it is clear that tenure and monthly data usage are the two biggest factors driving customers to switch providers. My advice would be for the providers to offer a loyalty incentive to customers. This could be as simple as some small free bonus each month. There should also be an unlimited data plan offered to all customers. The reason high data usage is causing customers to leave is due to the extra fees that come with it. If customers were not getting those extra costs, they would not feel punished each month and they would be far less likely to switch providers.

References:

S. (2020, November 13). K- Means Clustering Algorithm. EDUCBA.
<https://www.educba.com/k-means-clustering-algorithm/>

Nallathambi, J. (2018, June 20). R Series — K means Clustering (Silhouette) - CodeSmart. Medium.
<https://medium.com/codesmart/r-series-k-means-clustering-silhouette-794774b46586>

Facer, C. (2020, December 3). How to Create a Correlation Matrix in R. Displayr. <https://www.displayr.com/how-to-create-a-correlation-matrix-in-r/>

Gong, S. (2020, October 27). K-means Clustering for Customer Segmentations: A Practical Real-world Example. Medium.
<https://medium.com/@sygong/k-means-clustering-for-customer-segmentations-a-practical-real-world-example-196a10323b9f>

Dabbura, I. (2020, August 10). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. Medium.
<https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>