

Telco Customer Churn Cluster Analysis

Focused Customer Retention

Rudraksh Mishra




Context

- The *Telco customer churn* data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California.
- It indicates which customers have left, stayed, or signed up for their service.



Problem Statement

The aim of this project is to find interesting patterns and information from the past data collected using various data mining methods . This is done by forming clusters by identifying and grouping similar data points.



The Data

The data consists of 7,043 rows that represent the 7,043 customers with 21 attributes discussed below.

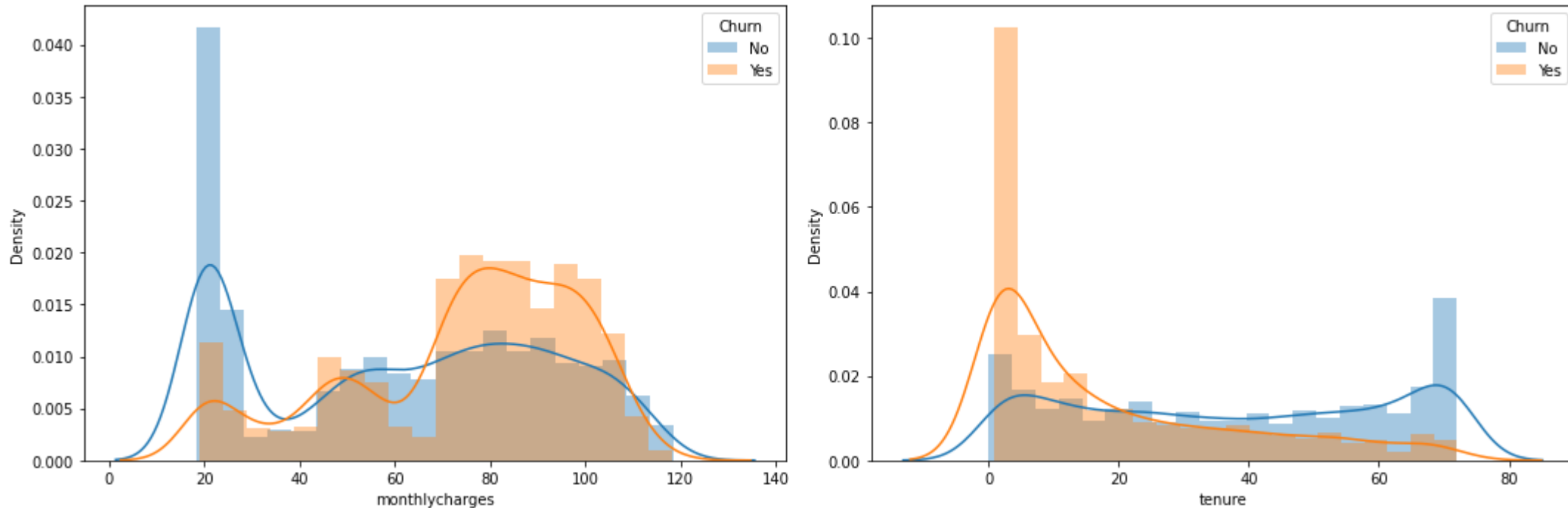
The data comes from the public repository on the Kaggle database, namely Telco Customer Churn data by [**BlastChar**](#).

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMusic	Contract	PaperlessBilling	PaymentMethod	MonthlyChurn	TotalCharges	Churn
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic	29.85	29.85	No
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer	42.3	1840.75	No
9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic	70.7	151.65	Yes
9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-month	Yes	Electronic	99.65	820.5	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-month	Yes	Credit card	89.1	1949.4	No
6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	No	No	No	No	No	Month-to-month	No	Mailed check	29.75	301.9	No
7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic	104.8	3046.05	Yes
6388-TABGU	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer	56.15	3487.95	No
9763-GRSKD	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-month	Yes	Mailed check	49.95	587.45	No
7469-LKBCI	Male	0	No	No	16	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Two year	No	Credit card	18.95	326.8	No
8091-TTVAX	Male	0	Yes	No	58	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	One year	No	Credit card	100.35	5681.1	No
0280-XJGEX	Male	0	No	No	49	Yes	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Month-to-month	Yes	Bank transfer	103.7	5036.3	Yes
5129-JLPIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-month	Yes	Electronic	105.5	2686.05	No
3655-SNQYZ	Female	0	Yes	Yes	69	Yes	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card	113.25	7895.15	No
8191-XWSZG	Female	0	No	No	52	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	One year	No	Mailed check	20.65	1022.95	No
9959-WOFKT	Male	0	No	Yes	71	Yes	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Two year	No	Bank transfer	106.7	7382.25	No
4190-MFLUW	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-month	No	Credit card	55.2	528.35	Yes
4183-MYFRB	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-month	Yes	Electronic	90.05	1862.9	No
8779-QRDMV	Male	1	No	No	1	No	No phone service	DSL	No	No	Yes	No	No	Yes	Month-to-month	Yes	Electronic	39.65	39.65	Yes
1680-VDCWW	Male	0	Yes	No	12	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	One year	No	Bank transfer	19.8	202.25	No
1066-JKSGK	Male	0	No	No	1	Yes	No	No	No internet service	No internet service	No internet service	No internet service	No internet service	No internet service	Month-to-month	No	Mailed check	20.15	20.15	Yes
3638-WFARW	Female	0	Yes	No	58	Yes	Yes	DSL	No	Yes	No	Yes	No	No	Two year	Yes	Credit card	59.9	3505.1	No

Pre-Processing & Cleaning

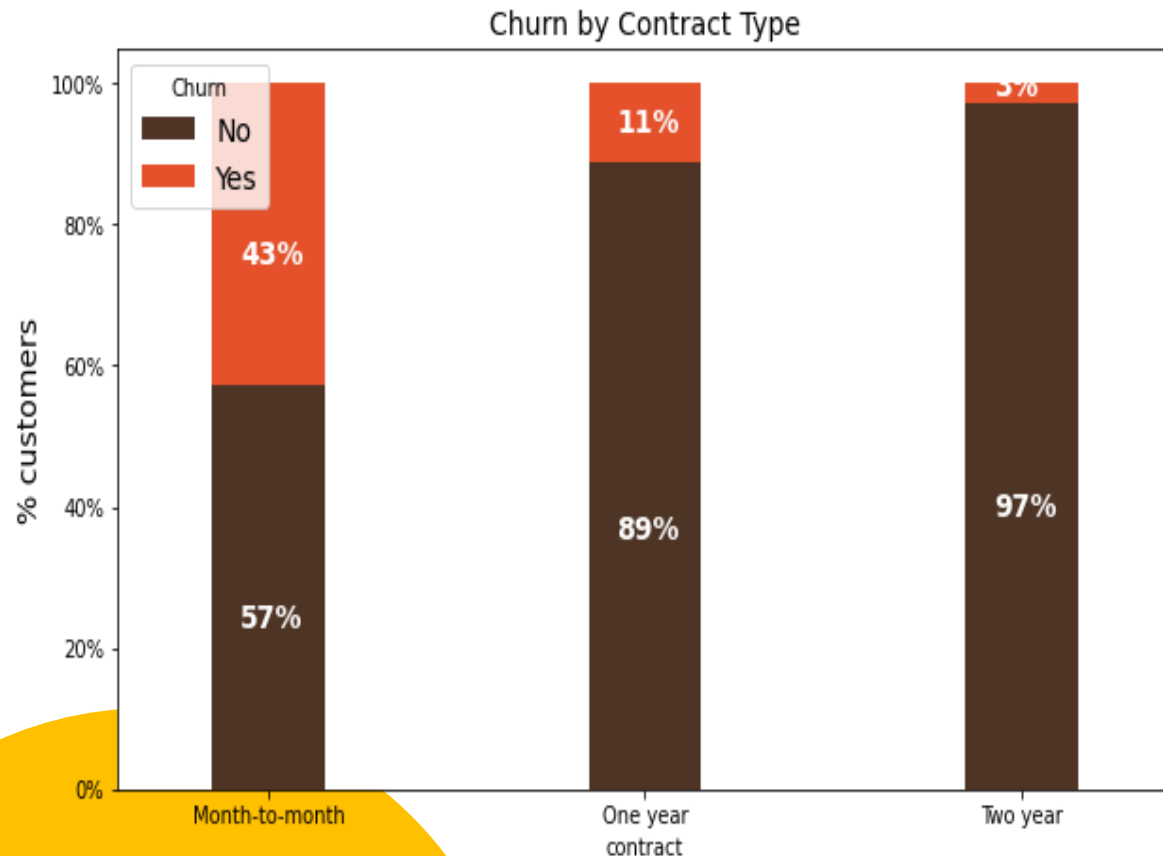
- The customer ID column has no relevant information that might be useful and so it is dropped.
- All the columns have been converted into categorical data
- Rows which has tenure value as 0 and rows which has null value in monthly charges has been dropped.
- One Hot encoding is done, where all the categorical data has been represented in a binary form

Data Visualization



- Visualization of the frequency distribution of monthly charges and tenure.
- From this, we can observe that people with low tenure have a high chance of churning, at the same time people with high monthly charges have a high probability of churning too.

Data Visualization

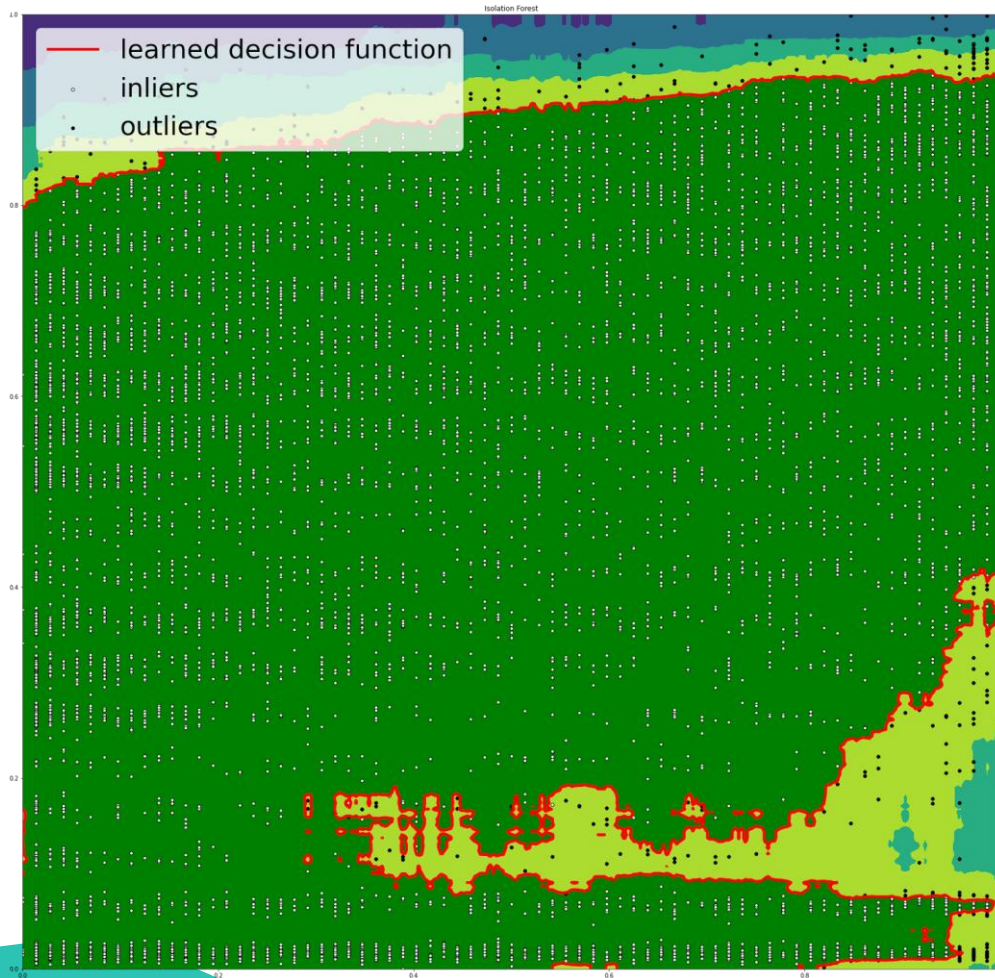


- Customers who have type month-to-month contracts are not yet loyal.
- Hence, the customers who use to have a month-to-month contract have a higher possibility of churn than the others.

Outlier Detection

- Here we used the PyOD library with an outlier factor of 6% to find the outliers in the data
- PyOD is a comprehensive and scalable Python toolkit for detecting outlying objects in multivariate data.
- We used Isolation Forest and ABOD methods for outlier detection.

Isolation Forest



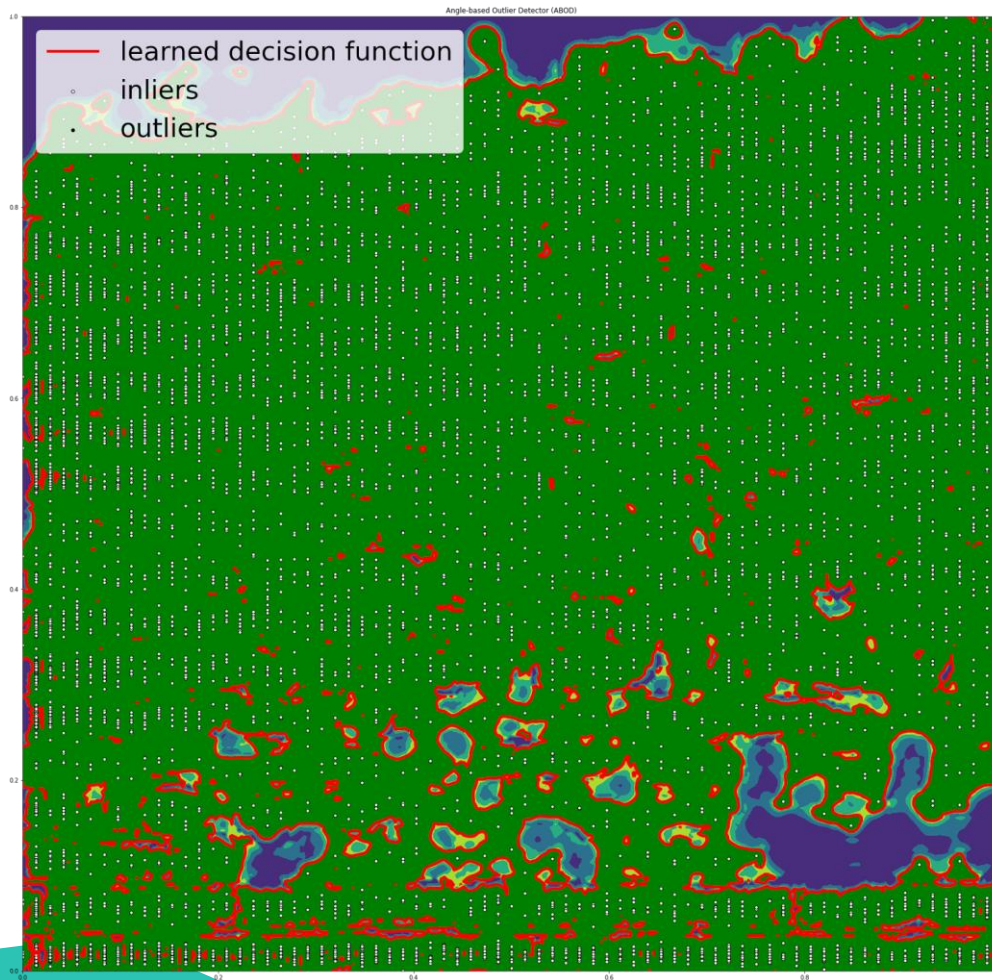
- Isolation Forest uses decision trees, in these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature.
- From Isolation Forest:
Outliers detected: 423
Inliers detected: 6620

Angle-based Outlier Detector (ABOD):

- ABOD assesses the broadness of the angular spectrum of a point as an outlier factor.
- From Angle-based Outlier Detector (ABOD):

Outliers detected: 0

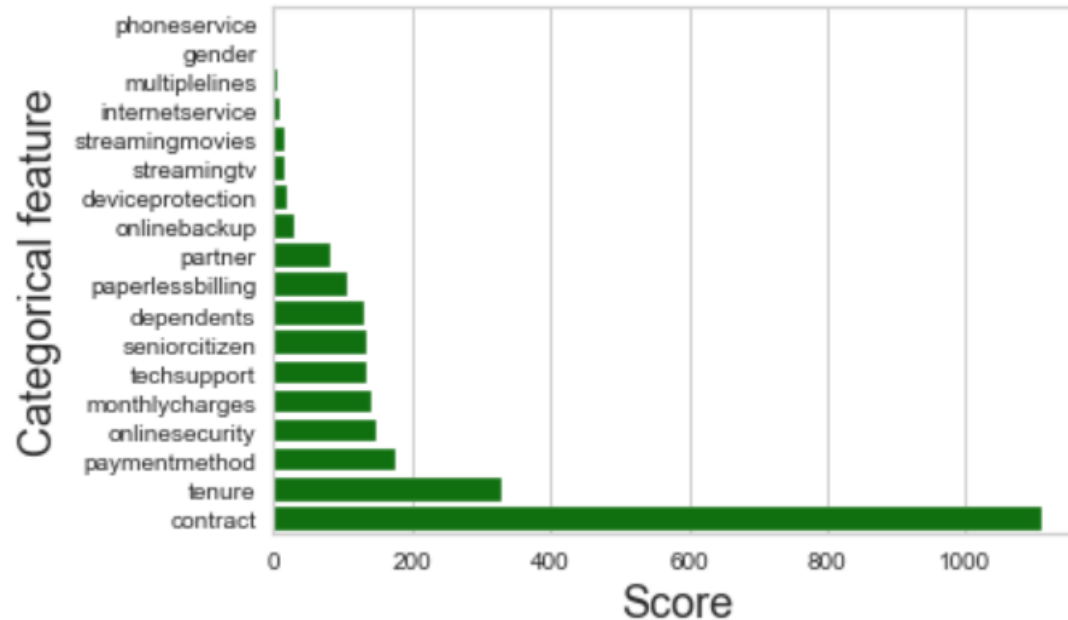
Inliers detected: 7043



Chi2 Feature Selection

- The test has been performed to test the dependency of the churn based on the other factors.
- We use the chi2 value to determine this.
- Higher the chi2 value the more the churn is dependent on that factor.
- Here churn is the target variable, and the rest are the input features.

Chi2 Feature Extraction



- Chi2 Feature selection has been performed and the resulting chi2 scores has been printed out as the output.
- Out of which top 10 has been selected, based on the chi2 values.
- The data has been reshaped by only choosing the 10 contributing features.

Association Rule Mining

- Association rule mining being done to find interesting frequent item sets.
- We use the apriori algorithm to find the frequent itemsets with a minimum support of 0.9 as the threshold.
- Association rules has been done with lift as 1 as the minimum threshold.

	support	itemsets
0	1.000000	(tenure)
2	1.000000	(monthlycharges)
3	1.000000	(churn)
5	1.000000	(tenure, monthlycharges)
6	1.000000	(churn, tenure)
9	1.000000	(churn, monthlycharges)
12	1.000000	(churn, tenure, monthlycharges)
1	0.909042	(phoneservice)
4	0.909042	(tenure, phoneservice)
7	0.909042	(monthlycharges, phoneservice)
8	0.909042	(churn, phoneservice)
10	0.909042	(tenure, monthlycharges, phoneservice)
11	0.909042	(churn, tenure, phoneservice)
13	0.909042	(churn, monthlycharges, phoneservice)
14	0.909042	(churn, tenure, monthlycharges, phoneservice)

Association Rule Mining

The frequent item sets have been found using apriori algorithm for churn =1

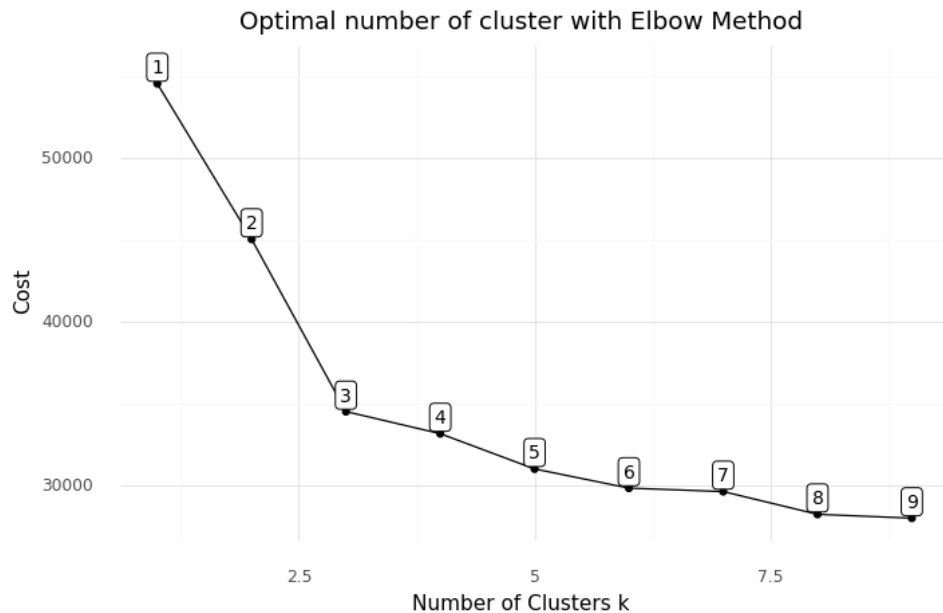


Association Rule Mining

	support	itemsets
0	1.00000	(tenure)
2	1.00000	(monthlycharges)
4	1.00000	(monthlycharges, tenure)
1	0.90122	(phoneservice)
3	0.90122	(phoneservice, tenure)
5	0.90122	(phoneservice, monthlycharges)
6	0.90122	(phoneservice, monthlycharges, tenure)

The frequent item sets have been found using apriori algorithm for churn =0

K Elbow



- Kmode clustering is done for k in range (1,10)
- The result is plotted to find the optimum number of clusters(K) given the cost of each cluster numbers.

Silhouette Method

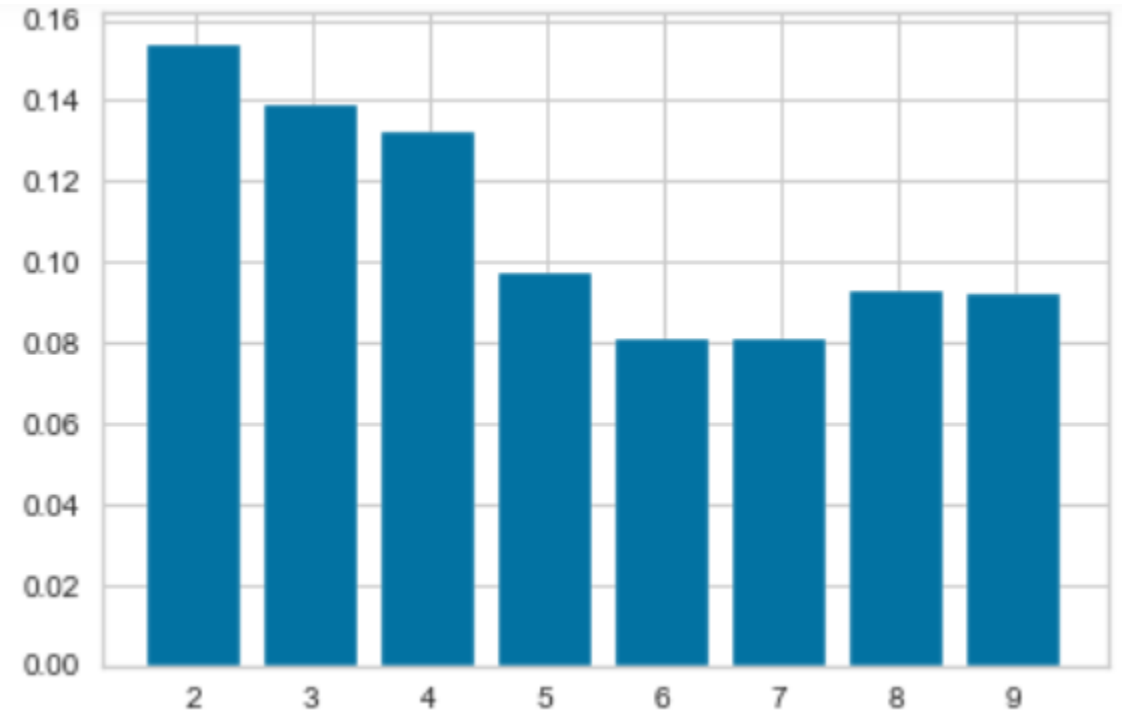
- It is another method used to find the optimum value for K.
- It is used to analyze the mean distance between the clusters and is considered a better method than Elbow since it is less ambiguous.
- It is done by using the silhouette score which is calculated using the mean intra cluster distance and the mean nearest cluster distance

Silhouette Method Observations

Silhouette scores of different cluster numbers

```
For n_clusters = 2, silhouette score is 0.15393652474912434)
For n_clusters = 3, silhouette score is 0.13919418718591922)
For n_clusters = 4, silhouette score is 0.13219882522469917)
For n_clusters = 5, silhouette score is 0.09753912402917762)
For n_clusters = 6, silhouette score is 0.08081128935552943)
For n_clusters = 7, silhouette score is 0.08121526655911628)
For n_clusters = 8, silhouette score is 0.09248224313438029)
For n_clusters = 9, silhouette score is 0.091896225895231)
```

Visual Representation of the Silhouette scores of different Cluster numbers



KModes Clustering

- Kmeans is generally the most commonly used clustering technique. But it only works for the numerical data. It's actually not suitable for the data that contains the categorical data type.
- So, Huang proposed an algorithm called k-Modes which is created in order to handle clustering algorithms with the categorical data type.

KModes Clustering

After the K modes clustering has been done. Each cluster has been given a label and they have been added to the dataframe.

	seniorcitizen	partner	dependents	tenure	onlinesecurity	techsupport	contract	paperlessbilling	paymentmethod	monthlycharges	churn	Cluster Labels
0	0	1	0	0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	0	1	0	0	0	0	2
2	0	0	0	0	1	0	0	1	0	0	1	1
3	0	0	0	1	1	1	1	0	1	0	0	0
4	0	0	0	0	0	0	0	1	0	1	1	1
...
7038	0	1	1	0	1	1	1	1	0	1	0	2
7039	0	1	1	1	0	0	1	1	1	1	0	2
7040	0	1	1	0	1	0	0	1	0	0	0	2
7041	1	1	0	0	0	0	0	1	0	1	1	1
7042	0	0	0	1	1	1	2	1	1	1	0	0

KModes Clustering

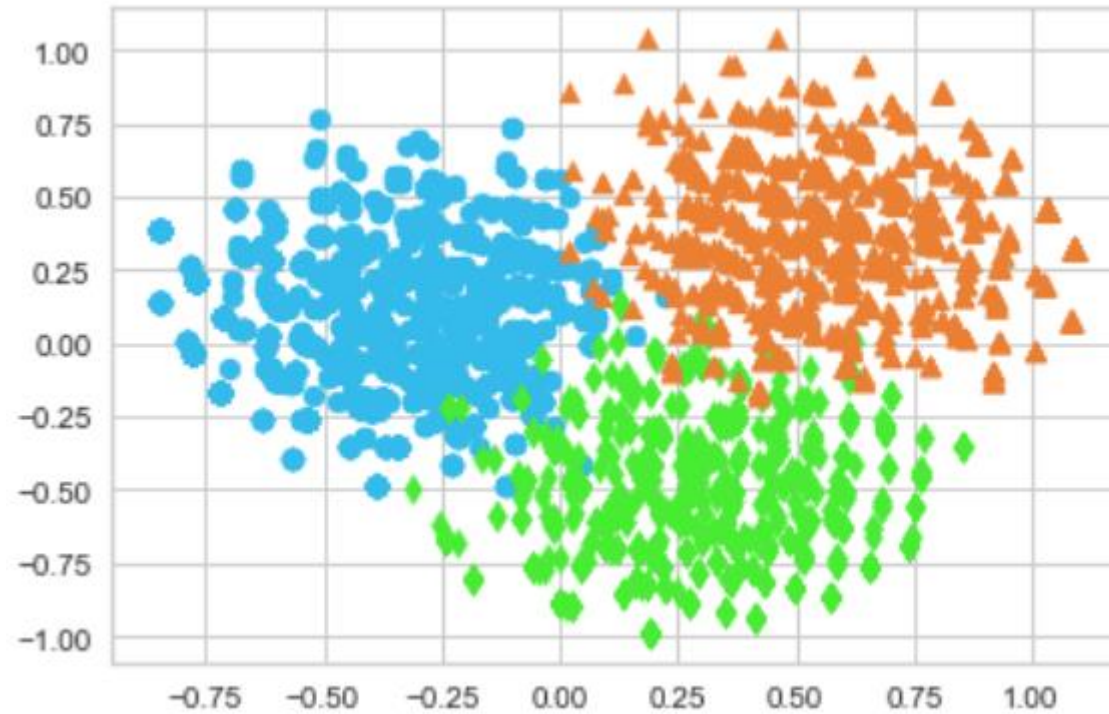
- The properties of the centroids of three clusters were noted and observed:

	seniorcitizen	partner	dependents	tenure	onlinesecurity	techsupport	contract	paperlessbilling	paymentmethod	monthlycharges	churn
First Cluster	0	1	0	1	1	1	2	0	1	1	0
Second Cluster	0	0	0	0	0	0	0	1	0	1	0
Third Cluster	0	1	1	1	0	0	1	1	0	0	0

Multiple Correspondence Analysis

- MCA is applicable specific to categorical features.
- The idea of MCA is to apply CA into the one-hot encoded version of the dataset. The result is like the PCA or CA result, two principal components with SVD result as the values.
- The results are then plotted in a 2-dimensional plot based on the cluster they belong to.

Plotting the result for Kmodes



Limitations of the clustering methods

- **KModes:**
 - Data types can only be categorical, cannot be a mix of categorical and numerical.
 - The feature on which the disagreement matters. Because K-Modes simply counts the number of dissimilarities, it doesn't matter to the algorithm on which "features" points are different. If a given category is particularly prevalent, this may become an issue as the algorithm will not take it into account when clustering.
- **Association Rule Mining:**
 - They have a large number of discovered rules with low comprehensibility.

References

- <https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113>
- <https://www.kaggle.com/blastchar/telco-customer-churn>
- <https://www.kaggle.com/ashydv/telecom-churn-prediction-logistic-regression>
- <https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>
- <https://amva4newphysics.wordpress.com/2016/10/26/into-the-world-of-clustering-algorithms-k-means-k-modes-and-k-prototypes/comment-page-1/>



THANK YOU