DATA MINING

PROJECT

Telco Customer Churn Using K-Modes Clustering

- *Rudraksh Mishra*

TABLE OF CONTENTS

1. Project Title

Telco Customer Churn Clustering

2. Team Members

Ambika Chundru (ajc7832@psu.edu)

Ashwath Ramesh (akr6021@psu.edu)

Rudraksh Mishra (rjm7016@psu.edu)

3. Project Contact Person

Rudraksh Mishra (rjm7016@psu.edu)

4. Problem Description

This project aims 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.

5. Mining Topics

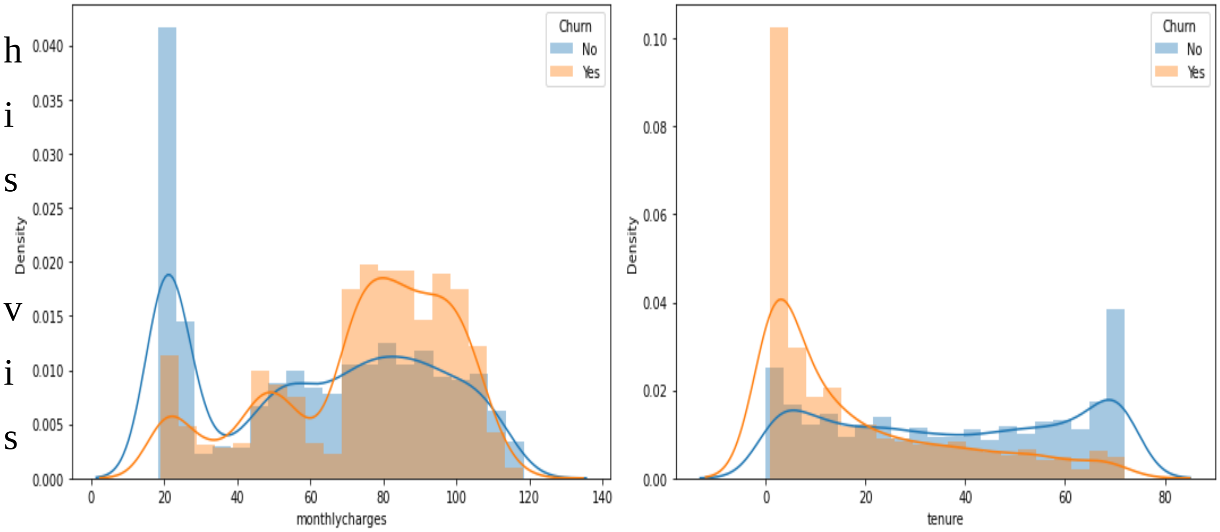
5.1 Data preprocessing steps and storing choices

The data preprocessing steps include: -

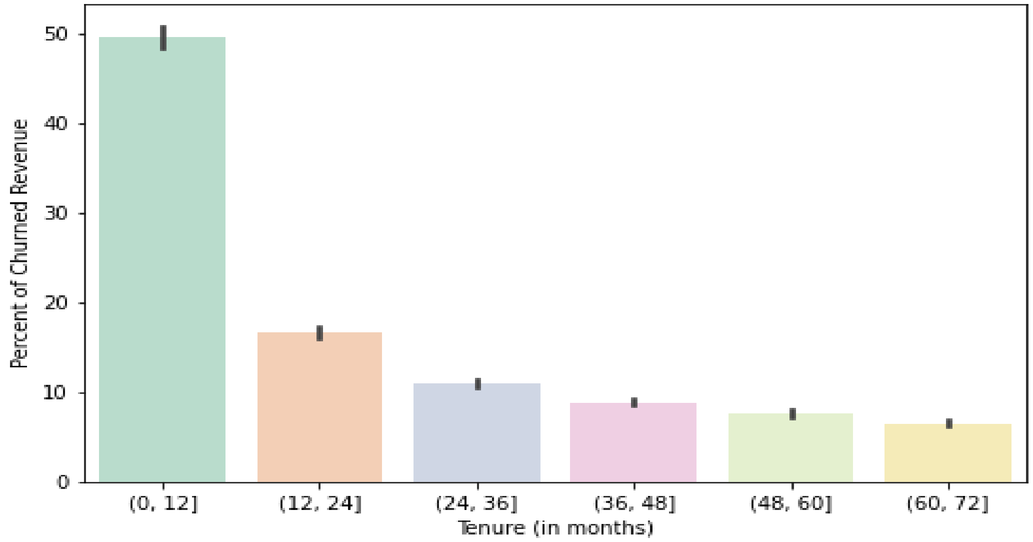
1. After importing the data and necessary libraries all the columns in the table are converted into categorical data with most of them being binary.
2. All the rows with ‘tenure’ equal to 0 were dropped and those also with a null value in monthly charge were dropped.
3. Tenure and monthly charges which were originally float and were converted into categorical data.
4. One Hot encoding is done, where all the categorical data has been represented in a binary form.

5.2 Data Visualization

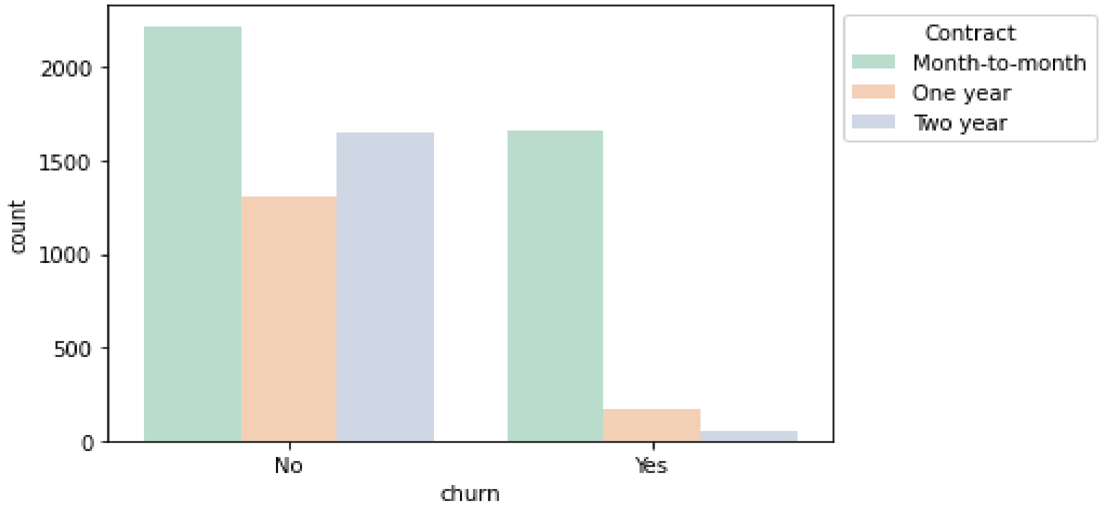
- This is a visualization of the frequency distribution of monthly charges and tenure. From this, we can observe that people with low tenure have a high chance of churning, at the same time people with high monthly charges have a high probability of churning too.
- This vis



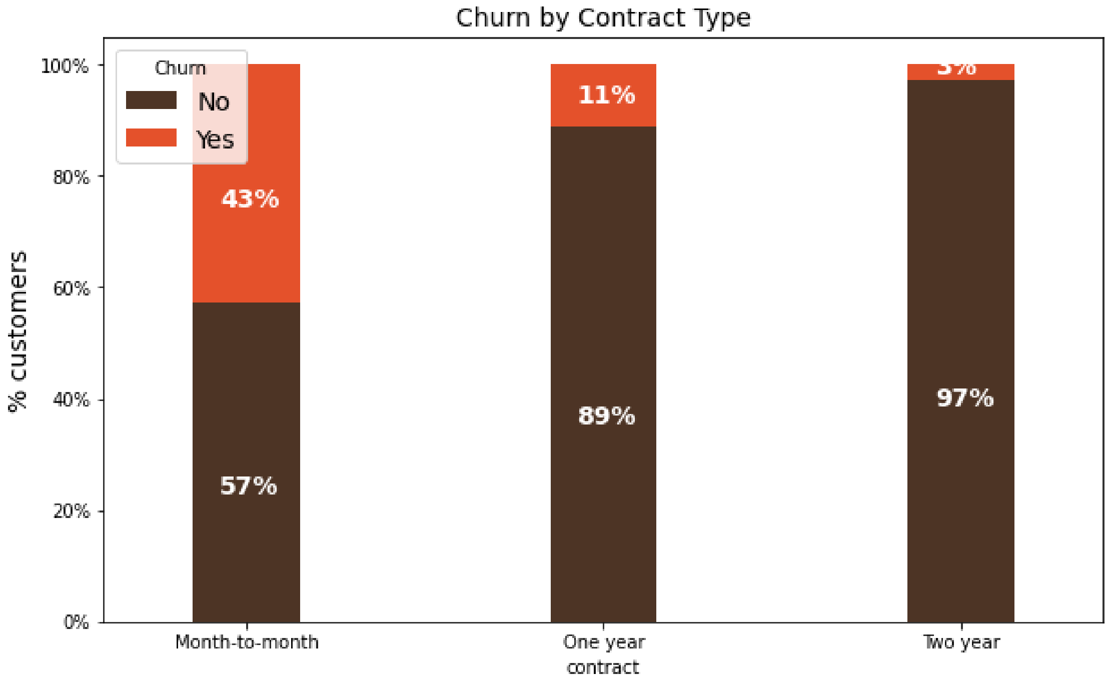
ualization shows the percentage of the people who churned sorted by their tenure in months.



- Through this visualization, we can observe that the count of the people who churned is high when the contract is month-to-month.



- Through this visualization, we can see that 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.



5.3 Data cleaning- cleaning false data

The dataset we inherited was filled with unwanted and false data which had to be removed.

The column ‘Customer_ID’ provided us with no useful information in predicting if the customer will churn or not and thus it was dropped.

All the rows with ‘tenure’ equal to 0 were dropped and those also with a null value in monthly charge were also dropped.

5.4 Outlier Detection/ Anomaly 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.

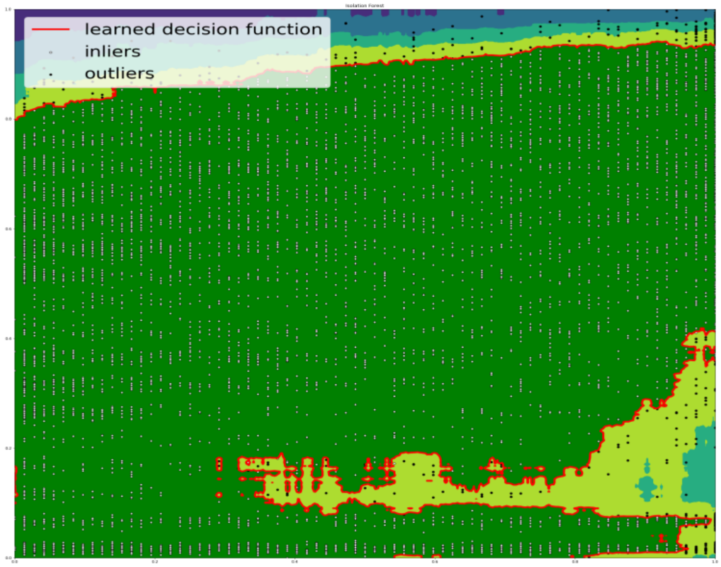
Isolation Forest uses decision trees, in these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature. Whereas, ABOD assesses the broadness of the angular spectrum of a point as an outlier factor.

Observations:

- <u>From Isolation Forest:</u>

  Outliers detected: 423

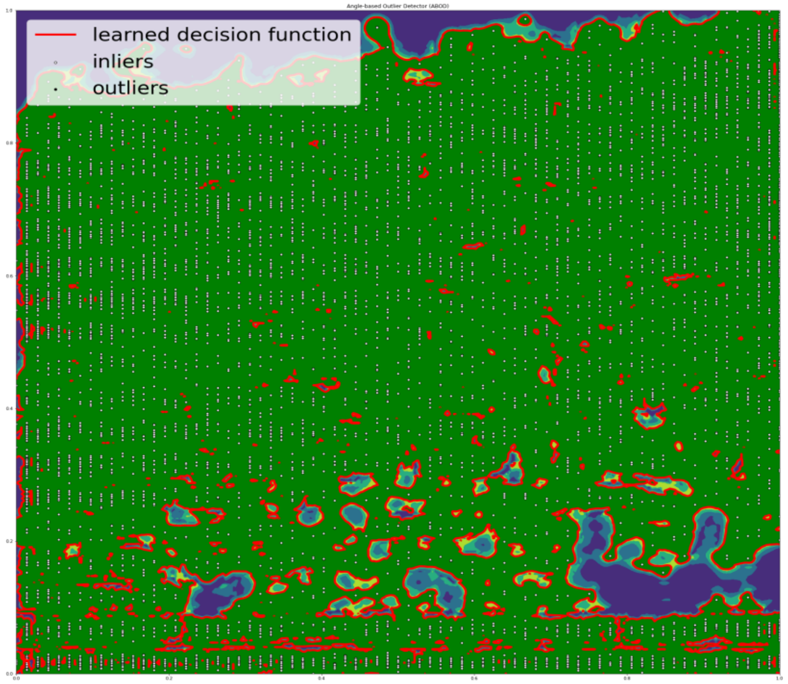
  Inliers detected: 6620



- <u>From Angle-based Outlier Detector (ABOD):</u>

  Outliers detected: 0

  Inliers detected: 7043



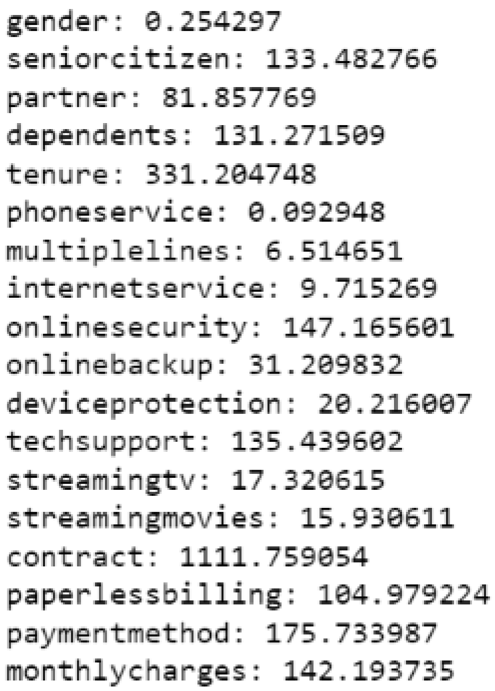
5.5 Feature selection- chi2:

Feature selection is a process where those features in a data frame that contribute most to the output are automatically selected. Having irrelevant features in the data could reduce the accuracy of the model and thus leading to poor output.

In chi2 feature selection, we intend to find out the dependency of the churn, which is the target variable based on other factors. We use the chi2 value to determine this. The hypothesis taken at first is that the churn is independent of the feature and when the chi2 value is high, this hypothesis is proven to be incorrect. Using this we can select the features that have a higher effect on the churn rate and focus on that.

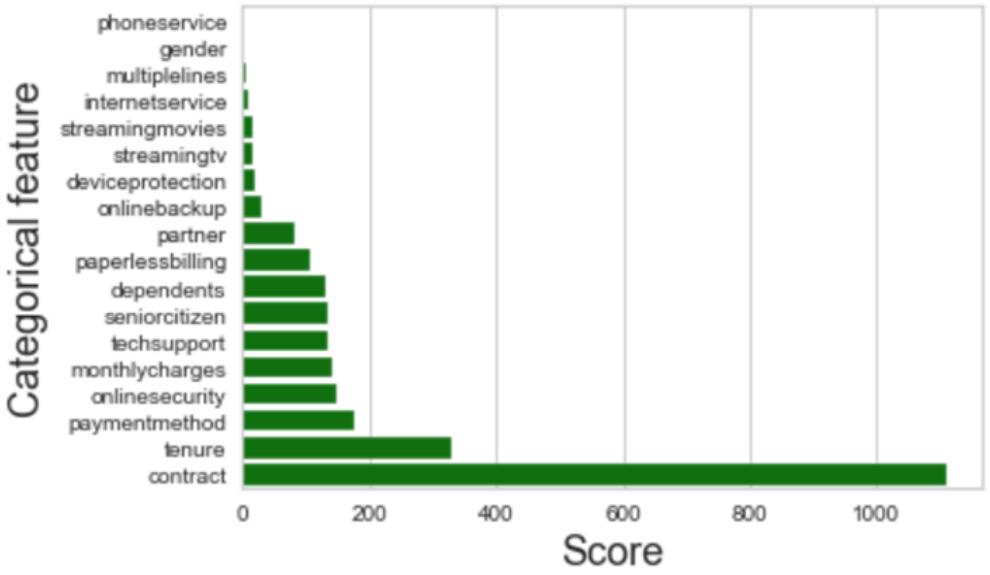
Observations:

The chi2 feature selection scores of the input features



```
gender: 0.254297
seniorcitizen: 133.482766
partner: 81.857769
dependents: 131.271509
tenure: 331.204748
phoneservice: 0.092948
multiplelines: 6.514651
internetservice: 9.715269
onlinesecurity: 147.165601
onlinebackup: 31.209832
deviceprotection: 20.216007
techsupport: 135.439602
streamingtv: 17.320615
streamingmovies: 15.930611
contract: 1111.759054
paperlessbilling: 104.979224
paymentmethod: 175.733987
monthlycharges: 142.193735
```

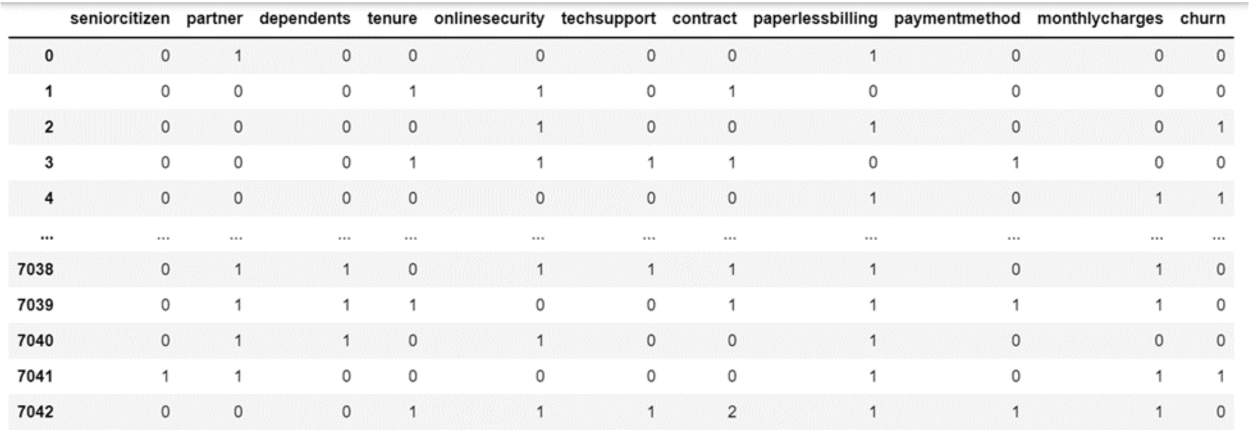
Plotting the result of the test:



From the results above we chose the top 10 features and reshaped the data frame accordingly.

The 10 features were SeniorCitizen, Partner, Dependents, Tenure, Onlinesecurity, Techsupport, Contract, Paperlessbilling, Paymentmethod, Monthlycharges

The reshaped data frame:



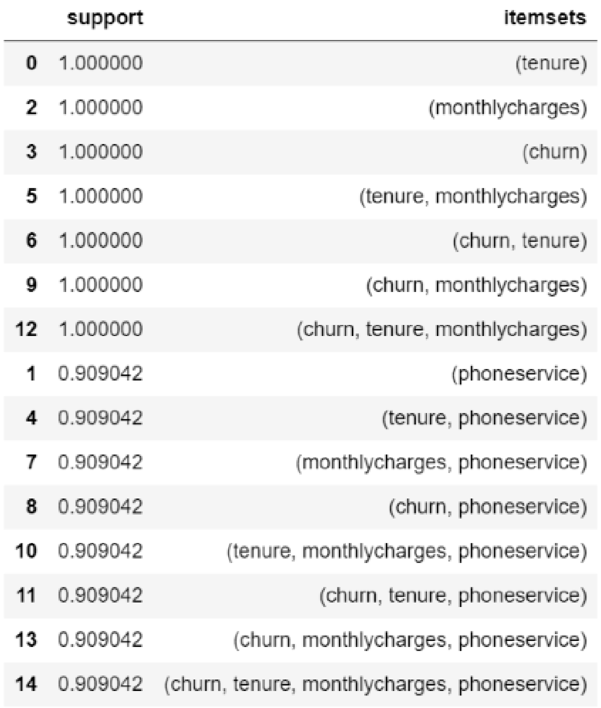
	seniorcitizen	partner	dependents	tenure	onlinesecurity	techsupport	contract	paperlessbilling	paymentmethod	monthlycharges	churn
0	0	1	0	0	0	0	0	1	0	0	0
1	0	0	0	1	1	0	1	0	0	0	0
2	0	0	0	0	1	0	0	1	0	0	1
3	0	0	0	1	1	1	1	0	1	0	0
4	0	0	0	0	0	0	0	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...
7038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

5.6 Association Rule Mining

Association rule mining is done to find interesting frequent itemsets. We can use the findings to better predict if the customer will churn or not. We use the apriori algorithm to find the frequent itemsets with minimum support of 0.9 as the threshold. Association rules have been done with lift as 1 as the minimum threshold.

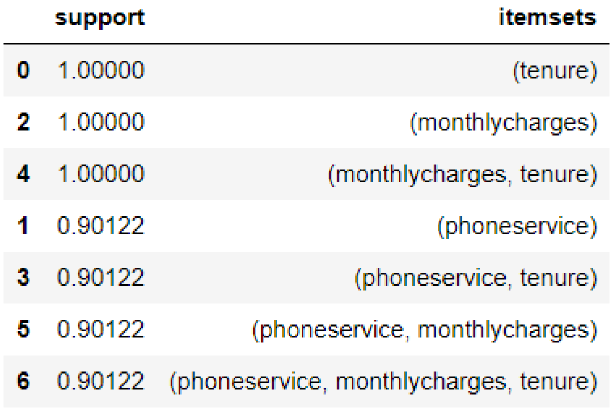
<u>Observations:</u>

Using the apriori algorithm, these were the frequent itemsets when churn=1:



	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)

Using the apriori algorithm, these were the frequent itemsets when churn=0:



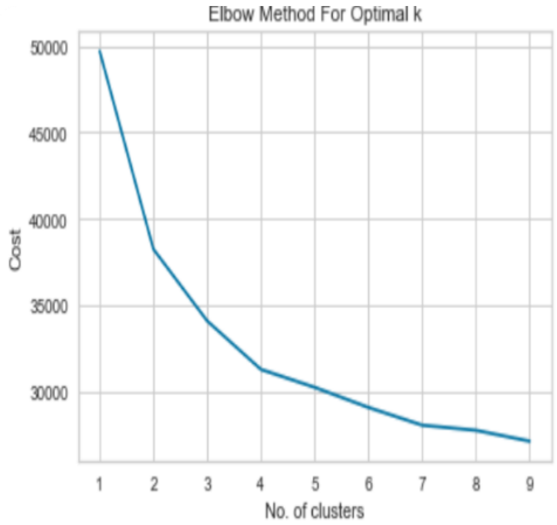
	support	itemsets
0	1.00000	(tenure)
2	1.00000	(monthlycharges)
4	1.00000	(monthlycharges, tenure)
1	0.90122	(phoneservice)
3	0.90122	(phoneservice, tenure)
5	0.90122	(phoneservice, monthlycharges)
6	0.90122	(phoneservice, monthlycharges, tenure)

5.7 K Elbow Method:

This method uses the optimum value of the number of clusters. This runs the kmode clustering on a loop each time increasing the value of the number of clusters and then plotting the result against the number of clusters to get a curve.

The point at which the curve bent and gives a shape of the ‘elbow’ is chosen as the optimum number of clusters. We choose that point since; it is the point at which the Sum of the Square of Errors first starts to diminish.

Observations:

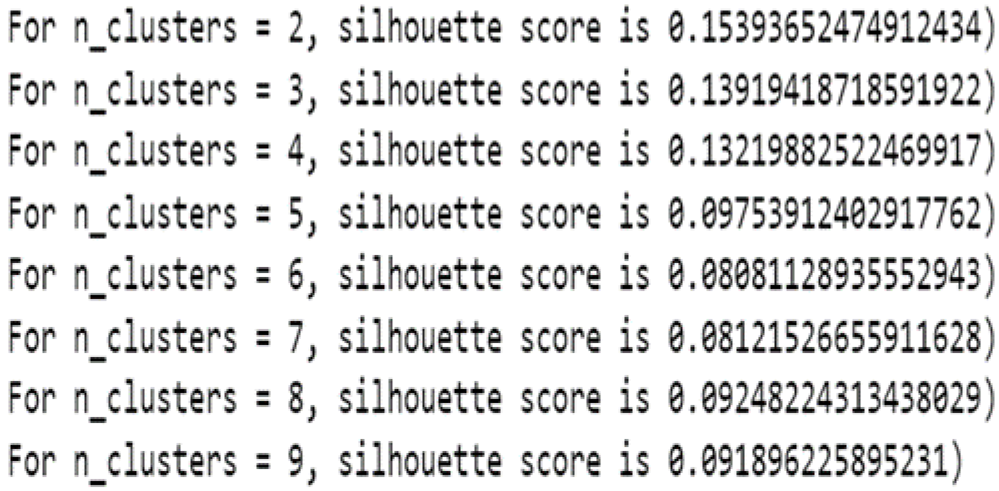


5.8 Silhouette Method:

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 intracluster distance and the mean nearest cluster distance

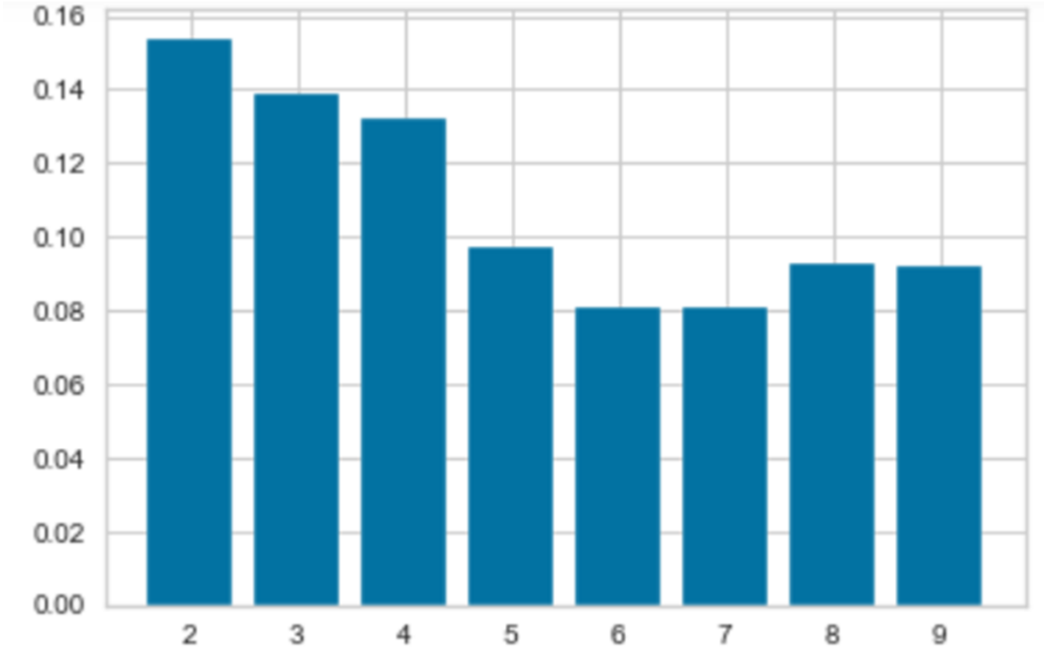
Observations:

The silhouette score is:



```
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)
```

The result of the silhouette score plotted against the number of clusters:

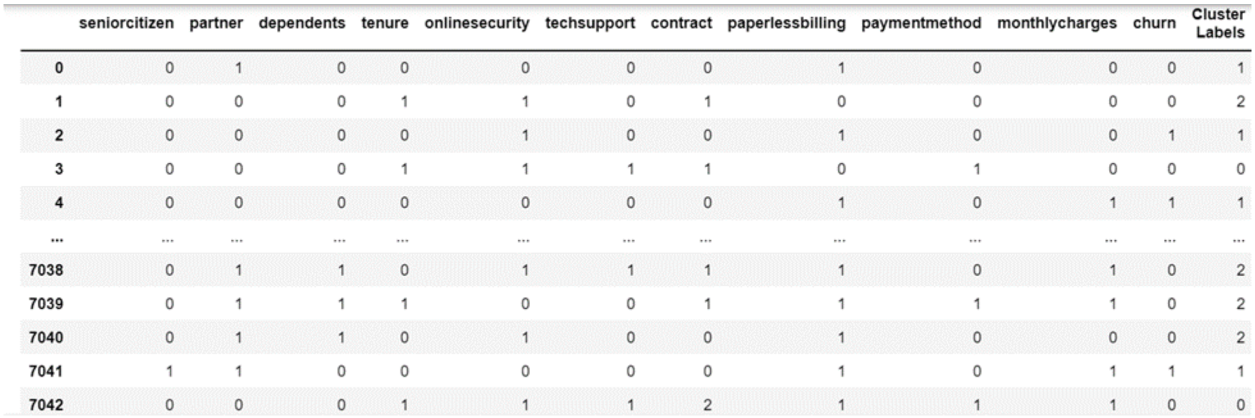


5.9 KModes Clustering & MCA:

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, in 1998, proposed an algorithm called k-Modes which is created in order to handle clustering algorithms with the categorical data type.

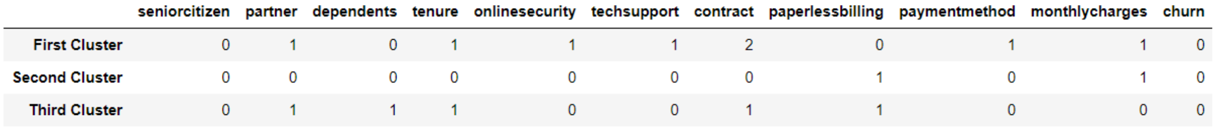
Observation:

After the K modes clustering has been done with the number of clusters chosen to be 3. Each cluster has been given a label and they have been added to the data frame:



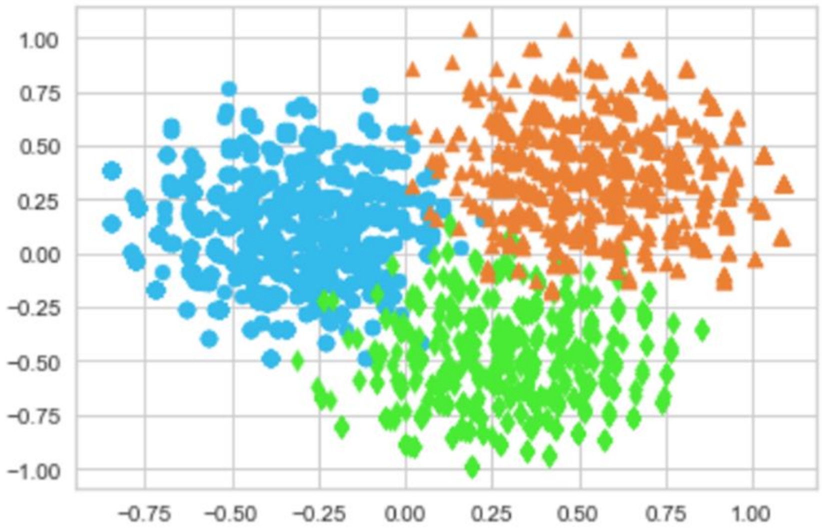
	seniorcitizen	partner	dependents	tenure	onlinesecurity	techsupport	contract	paperlessbilling	paymentmethod	monthlycharges	churn	Cluster Labels
0	0	1	0	0	0	0	0	1	0	0	0	1
1	0	0	0	1	1	0	1	0	0	0	0	2
2	0	0	0	0	1	0	0	1	0	0	1	1
3	0	0	0	1	1	1	1	0	1	0	0	0
4	0	0	0	0	0	0	0	1	0	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...
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

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

The resulting clusters were also plotted using MCA. 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.



6. 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.

7. Dataset source and attributes

The data is taken from the public repository in the Kaggle database. It was posted by *BlastChar* and is called Telco Customer Churn Data. The link is as follows: - https://www.kaggle.com/blastchar/telco-customer-churn

The data consists of 7043 records, each representing a customer, and 21 fields namely: -

a) Customer ID – Its object Data type. All the values in it are unique hence it is used as a primary key

b) Gender – It's a binary categorical (object) Data type. It represents the gender of the customer either male or female.

c) Seniorcitizen – It's an Integer data type. It shows if the customer is a senior citizen or not. Yes, represents they are and No represents they aren't.

d) Partner - It's a binary categorical (object) Data type. It shows if the customer has a partner or not. Yes, represents they have and No represents they don't.

e) Dependents - It's a binary categorical (object) Data type. It shows if the customer they have any dependents or not. Yes, represents they have and No represents they don't.

f) Tenure - It's an Integer data type. It shows the number of years the customer has held the service.

g) PhoneService: - It's a binary categorical (object) Data type. It shows if the customer they have phone service or not. Yes, represents they have and No represents they don't.

h) MultipleLines: - It's a nominal categorical (object) Data type. It shows if the customer they have multiple phone services or not. Yes, represents they have, No represents they don't, and No phone service represents they don't even have 1 phone service.

i) PaperlessBilling: - It's a binary categorical (object) Data type. It shows if the customer they have enrolled for paperless billing. Yes, represents they have and No represents they don't.

j) MonthlyCharges: - It’s a float data type. It shows the amount of monthly charge of the service the customer pays.

k) Churn: - It’s a binary categorical (object) Data type. It shows if the customer they have churned or not. Yes, represents they have and No represents they don’t.

l) InternetService: - It’s a nominal categorical (object) Data type. It shows if the customer they have phone service or not. DSL represents they use DSL service, Fiber Optics represents they use fiber optics and No represents they don’t use any service.

m) OnlineSecurity: - It’s a nominal categorical (object) Data type. It shows if the customer they have an Online Security to their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

n) OnlineBackup: - It’s a nominal categorical (object) Data type. It shows if the customer they have an Online Backup to their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

o) DeviceProtection: - It’s a nominal categorical (object) Data type. It shows if the customer they have a Device Protection to their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

p) TechSupport: - It’s a nominal categorical (object) Data type. It shows if the customer they have a Tech Support to their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

q) StreamingTV: - It’s a nominal categorical (object) Data type. It shows if the customer can stream TV using their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

r) StreamingMovies: - It’s a nominal categorical (object) Data type. It shows if the customer can stream movies using their internet service or not. Yes, represents they have, No represents they don’t, and No internet service represents that they don’t have any internet service.

s) Contract: - It’s a nominal categorical (object) Data type. It represents the type of contract the customer has. Month to month, represents they have a month-to-month contract, One year represents they must pay yearly, and Two years represents that they have a two-year contract.

t) PaymentMethod: - It’s a nominal categorical (object) Data type. It represents the payment method that the customer use. A mailed check means the user mails the payment, an electronic check means the user sends an e-check, a Credit card means the user uses a credit card and a Bank transfer means the user sends the money via a bank transfer.

u) TotalCharges: - It’s a nominal categorical (object) Data type. Each value represents the total amount paid by the customer.

The dataset is in .csv format.

8. Deliverables

8.1 Roles and responsibilities

- Rudraksh Mishra: pilot presentation, data visualization, use elbow method, and Silhouette method, apply k modes clustering to see optimum k values, Feature extraction- MCA, final presentation, writing the report, creating the PowerPoint

- Ambika Chundru: Data preprocessing, data cleaning, applying chi-square feature selection poster preparation, writing the report, creating the PowerPoint, final presentation.
- Ashwath Ramesh: readme document, writing the report, PowerPoint, final presentation.

8.2 Deliverables and tentative schedule:

- Chose a data mining project and deliver a pilot presentation (Week 1)
- First, we import and clean the data. (Week 2)
- Then we pre-process it by standardizing (Week2)
- Relationships between different attributes (correlations in the data) and build visualization to support them. (Week 2)
- Feature selection using chi-square feature selection (Week 2)
- Apply k-mode clustering algorithm, Association rule mining, k++ elbow method, silhouette method to see optimum k values. (Week 3 & 4)
- Feature extraction- MCA (week 4)
- Final presentation and plotting observed results (Week 5)

9. References

a) https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

b) https://www.kaggle.com/blastchar/telco-customer-churn

c) https://www.kaggle.com/ashydv/telecom-churn-prediction-logistic-regression

d) https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

e) https://amva4newphysics.wordpress.com/2016/10/26/into-the-world-of-clustering-algorithms-k-means-k-modes-and-k-prototypes/comment-page-1/

f) https://www.kaggle.com/supratimhaldar/telco-customer-churn-exploratory-data-analysis

g) https://medium.com/@techtangent.in/visualizing-the-telco-churn-dataset-and-picking-up-the-important-features-94bc154e4153

h) https://www.kaggle.com/nasirislamsujan/an-eda-on-telco-customer-churn-prediction

i) https://docs.tibco.com/data-science/GUID-F0FC2493-A73D-4BE8-A3F8-6E26C6057371.html

j) https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

k) https://towardsdatascience.com/silhouette-method-better-than-elbow-method-to-find-optimal-clusters-378d62ff6891#:~:text=The%20silhouette%20method%20computes%20silhouette,each%20object%20has%20been%20classified.