Clustering Analysis

Project Title: Customer Churn Analysis for Telecommunication Company

Project Sponsor: Doris Lee, CEO of Advanced Consultant Services

Project Manager: Maniram Luital

Project Start Date: July 1, 2024

Expected project End Date: September 15, 2024

Presented by : Kushal Basyal

Date: August 21, 2024

Role: Data Analyst - Clustering Analysis

Abstract:

This report aims to provide a comprehensive overview of the clusterimng analysis along with the resulting visualization techniques applied in customer churn analysis project. So, segmented the customers into distinct groups based on their behaviours and characteristics using k-means algorithms. In our first step, we assumed the number of clusters to be 3. We are assuming the number of clusters to get some ideas on how to use the K-Means algorithms. Then we will visualize the clusters based on the 3 clusters based on the customer segmentation features tenure and monthly charges. This step is just to figure out how to apply the K-Means and see the outcomes. In the next step, we will determine the optimal number of clusters using the elbow and silhouette analysis. After that, we utilised the clustering algorithms to segment customers. We selected the appropriate clustering algorithms K-Means on the both min-max and standard scaled datasets. We trained the clustering model on the dataset with the determined number of clusters. Interpreted the clustering results and labelled the clusters based on their characteristics. Finally, we developed the visualizations to aid in the interpretation of clustering results by using tools such as matplotlib or seaborn to create the visual representation of the clusters.

Finally, we developed the visualizations to aid in the interpretation of clustering results by using tools such as Matplotlib or Seaborn to create the visual representation of the clusters.

Introduction

Background of the customer analysis project

The main aim of the customer churn analysis is to predict the statistical report and presentation on the status of the customers. Basically, its very important to identify the customers likely to discontinue their subscription. Identifying at risk customers allows the company to take proactive measures to retrain them, which is way better than acquiring newly customers. So, whole project involved analysing historical customers data to define the patterns and behaviour that indicate a likelihood of churn. This whole analysis informed the development of predictive modelling that will help us to forecast the future churn. Finally, help us in creating the targeted retention strategies.

Importance of Visualization and Clustering

Clustering Analysis play a vital role in understanding the customer segments which are arranged within the dataset. One of the important advantage of clustering analysis is that it allows us to group customers with their similar characteristics based on their features. It will help us to identify the distinct segments with unique behaviours and character. All these information can be used to tailor the marketing and retention strategies more accurately. We start with determining the optimal number of clusters and once we identify the number of clusters, we utilize clustering algorithms to segment customers.

Visualizations play an important role in interpreting the results of clustering. The main thing of the visualization is to create a clear and insightful visual representation of the clusters. With the help of this process, we can better understand the nature of each segment, identify outliers, and communicate the findings to stakeholders. We will use different methods, such as scatter plots, box plots, cluster distribution, heat maps and many more from the customer segmentation features tenure and monthly charges. together clustering and visualizations form the foundation for data-driven decisions-making in customer.

Roles and responsibilities:

Specific role in the project:

I am taking the role of data analyst focusing on clustering analysis and visualizations. My primary responsibilities included utilise clustering algorithms to segment customers, determine the optimal number of cluster by using elbow method or silhouette analysis and validate the optimal number of clusters through internal and external validation methods, train the clustering model on the dataset with the determined number of clusters and interpret the clustering results and label the clusters based on their characteristics, develop the visualizations to aid in the interpretation of the clustering results by using the tools like Matplotlib or seaborn to create visual representation of the clusters.

Detailed description of responsibilities

1. Clustering Analysis

a. Data Pre-processing:

Feature Selection: The feature selection is done from the pre-processed datasets which includes both min-max and standard scaled dataset. The feature selection is done in such a way that it influence customer behaviour. Regarding this project, we selected the features of tenure and monthly charges from both scaled datasets. My role in this project is to apply clustering analysis and visualizations of the customer segmentations based on the tenure and monthly charges.

Data Scaling: My role is to utilized the scaled dataset for clustering and visualizations. In our project the scaled datasets include min-max scaled dataset and standard scaled dataset. These datasets are obtained by data pre-processing which includes feature engineering followed by the scaling.

b. K-Means Clustering

K-Means Algorithm Selection: Regarding this project, K-Means was chosen as the primary clustering algorithms. The main reason behind choosing K-Means algorithm was its efficiency with large datasets, its ability to create well-separated clusters, and its simplicity and interpretability. K-Means optimizes the centroids by minimizing the variance within clusters, and make sure about distinct and meaningful groupings. It assumes spherical clusters, which aligns with the characteristics and behaviours of the customer data.

Clustering Initial: Initial clustering represents the applying of K-Means clustering with assume number of cluster equal to 3. Applied K-Means for both scaled dataset with the assumed number of cluster and visualize the clusters.

c. Determining the optimal number of cluster

Elbow method: This project utilized the elbow method to plot sum of squared distance between each data point and its assigned cluster centre, identifying the point where the curve starts to level off, indicating the optimal number of clusters. The elbow point where additional clusters provide minimal improvement in reducing variance, ensuring the model remains both interpretable and effective without overfitting. The elbow visual nature makes it an intuitive tool for communicating findings to stakeholders.

Silhouette Analysis: We implemented this method to evaluate the cohesion and separation of clusters, offering insights how well each data points fits within its assigned clusters compared to others. A higher silhouette score indicates that clusters are well-defined and distinct, making it a valuable metric in determining the optimal number of clusters. By obtaining these scores, we ensured that the clusters were not only compact but also clearly separated, enhancing the model accuracy and interpretability. This analysis was crucial in validating the chosen clusters, reinforcing the overall robustness of the clustering solution.

d. Final Clustering

After using the elbow method and silhouette analysis, I determined the optimal number of clusters as 4. Now by using determined number of clusters, we trained the clustering model on the dataset by using K-Means algorithms. Interpret the clustering results and label the clusters based on their characteristics.

e. Interpretation of clusters:

We carefully analysed characteristics of each cluster by analysing the mean and median values of the key features tenure and monthly charges, leading to a clear understanding of the distinct customer segments. Based on this analysis we confirmed that certain clusters represent customers with high tenure and high charges, while the others are characterized by low tenure and low charges.

Based on distinct characteristics, each cluster was labelled to reflect the specific customers profile. For instance, each cluster identified as high tenure and high charges indicating the group of long term, premium customers. While another cluster labelled low tenure and low charges representing customers with basic plans. These labels provide the summary of the customers, making easier to tailor business strategies to each group.

2. Visualizations

a. Visualization of clusters:

Scatter plots: The main of creating the scatter plots is to visualize the distribution of the customer segmentation within the clusters based on their characteristic's tenure and monthly charges

Boxplots: The main aim of creating boxplots is to develop the boxplots to compare distribution of selected features and provide the insights into variability within each segment.

Heatmaps: The main aim of generating heatmaps is to develop the relationship between the selected features across the clusters which helps us to identify the patterns and correlations

b. Interpretation and communication

Visualisations interpretation helped us to draw the meaningful conclusions about the customers segments and their behavior. After visualisation to draw the conclusion, prepared the detailed reports and documentation in order to communicate the findings to the stakeholders. and most important thing is that while doing so ensure that visualisations conveyed the key insights effectively.

c. Reporting and documentation:

All the process and outcomes are documented that includes the clustering and visualization process. Clustering process includes applying the k-means algorithms, determining the optimal number of cluster and train the determined cluster based on their features. All the outcomes, plot along with the documentations were recorded in the respective folder for the future analysis. Similarly the detailed visualizations were done by using different techniques such as scatter plots, boxplots, heatmaps and many more. The insights gained from each visualizations were recorded in the respective folder along with the effective scripts, Jupyter notebooks and documentation separately for the easy access. The main aim of these visualizations is to show the insights gained from each visualizations, and how these visualizations supported the understanding the customer segments.

Results and outcomes:

Clustering Analysis:

We stated the K-Means algorithms by assuming the number of clusters equal to 3. Applied the K-Means clustering in both min-max scaled and standard scaled datasets on customer segmentation based on the features tenure and monthly charges. Visualized the clusters for both the scaled dataset, providing different insights into different customer segments.

Applied both elbow method and silhouette analysis on scaled datasets. Based on these both method, the optimal number of clusters for the dataset is 4. The consistency across both the scaling methods suggests that the natural grouping within the data is robust to the type of scaling used. After determining the optimal number of cluster, labelled the clusters, analysed their characteristics. These outcomes are recorded properly for future insights in areas such as marketing strategies and future analysis.

Trained the Clustering Model and interpret results from the determined number of cluster. Applied K-means clustering where number of k=4 for both scaled datasets. By clustering the customers into 4 groups, we identified the segments based on their tenure and monthly charges. We obtained the new scaled dataset including the clusters for both scaled dataset and saved in the clustering folder. Analysed and saved the cluster characteristics based on their features mean and median.

Clusters	Min-max scaled	Standard scaled	Character
High Tenure, High Charges (premium Customers)	Cluster 2	Cluster 1	Long-term high-value customers subscribed premium plans. Crucial for retention and upselling opportunities
Low tenure, Low Charges (New or Basic	Cluster 3	Cluster 0	New or lower value customers. prime targets for engagements strategies aimed at increasing value

Customers)			through promotion and upsell opportunities
High tenure, Low Charges (loyal but economical customers)	Cluster 0	cluster 3	Loyal customers with economical plans over time. Potential to increase lifetime value through premium services and rewards
Moderate Tenure and Charges	Cluster 1	Cluster 2	Mid-tier customers with stable tenure and charges. Potential for growth through targeted marketing and value-added services.

Visualizations

At the end generated visualizations to aid the interpretation. It helped us to understand the distribution of the clusters formed in the customer segmentation analysis. Visualized the distribution of customers within clusters, along with cluster centroids. Boxplots showed the distribution of tenure and monthly charges within each cluster. Cluster Distribution displayed the number of customers in each cluster. Heatmap of cluster Characteristics summarized the mean tenure and monthly charges for each cluster using heatmap.

How both steps enhanced the dataset for modelling

The Clustering and Visualizations method played an important role in enhancing the dataset for modelling in the context of customer segmentation into distinct, well-defined groups based on behaviour and characteristics. Both the clustering and visualization techniques provided valuable results, forming the natural groupings within the data that could be leveraged for targeted marketing and customer retention strategies. Visualizations method made these insights more clear and accessible, enabling the stakeholders to quickly grab the unique traits of each segment and make data driven decisions. Besides that, these method identified the key patterns, outliers, and influential factors helping refine the dataset for more accurate predictive modelling. These method contributed to improved customer satisfaction, loyalty, and overall performance by laying a solid foundation for tailored strategies and precise prediction.

Recommendations for future analysis:

Alternative Clustering Algorithms: We applied K-Means algorithms for our project which is one of the easiest and effective method. But it may miss some non-spherical clusters or clusters with various densities. So we can use alternative clustering algorithms such as DBSCAN or Agglomerative Clustering that would uncover additional insights and validate robustness in our clusters.

Introducing New Features: We can introduce more new features based on domain knowledges, such as customer segmentation levels or loyalty scores. While doing this we are adding more informative features that can lead to more meaningful clusters, and also provides deeper insights into customers behaviour.

Predictive Models for integration clusters: The main aim of this method is to validate the practical utility of clusters in predicting outcomes. Using this method not only tests the relevance of our clusters but also create more accurate and actionable predictions.

Sub-Clustering Analysis: The main purpose of this method is to discover more granular customer segments within existing clusters. Normally, we applied the clustering techniques within each of the identified clusters to see if there are any sub-clusters. Identified these small sub-clusters revealed the specific customer needs or preferences that can be focussed more precisely in marketing or any decision making.

Conclusion:

Applied K-Means clustering analysis based on tenure and monthly charges by assuming the number of clusters equal to 3. The clusters are visualized and saved, providing the insights into different customer segments. After that applied elbow and silhouette analysis, based on these methods, the optimal number of cluster is 4 for both min-max and standard scaled datasets. Then proceed to label the clusters, analysed their characteristics, and apply these insights to areas such customer segmentation, marketing strategies, and streamlined future analysis. After that fitted the K-Means clustering algorithms to our customer datasets and apply it for both scaled datasets. By clustering the customers into 4 groups we identified the segments based on tenure and monthly charges.

For Visualizations, we created detailed visualizations to interpret results and communicate the clustering findings, offering insights into customer segments and supporting data-driven decision making. We used different method for visualizations such as scatter plots, box plots, cluster distribution and heat maps.

Reflection Of Clustering and Visualization Importance in Overall project:

Clustering and visualization are very important components in the Customer Churn Analysis project, providing as the bridge between raw data and strategic business insights. The primary objective of this project is to understand the datasets and predict customer churn, a critical factor for maintaining profitability and customer satisfaction in the telecommunications industry. Clustering plays a fundamental role in this process by grouping customers into distinct segments based on their shared behaviors and characteristics, such as spending patterns, tenure, and service usage. These clusters help in identifying patterns and trends that may not be apparent through individual data points alone. The main aim of clustering analysis is to understand the data that can lead to make more informed and effective decision making.

The identification of these customer segments allows for a deeper understanding of the different types of customers the company serves. For example, some clusters might represent high-value, long-term customers who are less likely to churn, while others might highlight new customers who are more at risk. By understanding these segments, the business can tailor its marketing and retention strategies more effectively, targeting resources towards customers who are most likely to churn and offering personalized incentives to retain them.

References

Online References:

2.3. clustering scikit. Available at: 2.3. Clustering (Accessed: 25 August 2024).

Sharma, P. (2024) *Introduction to K-means clustering*, *Analytics Vidhya*. Available at: Introduction to K-Means Clustering (Accessed: 25 August 2024).

Habib, A.B. (2021) *Elbow method vs silhouette co-efficient in determining the number of clusters*, *Medium*. Available at: Elbow Method vs Silhouette Co-efficient in Determining the Number of Clusters (Accessed: 25 August 2024).

Atlassian *A complete guide to heatmaps*, *Atlassian*. Available at: A Complete Guide to Heatmaps | Atlassian (Accessed: 25 August 2024).

How to implement clustering algorithms in Python: A step-by-step approach (2024) *Data Headhunters*. Available at: How to implement clustering algorithms in Python: A Step-by-Step Approach (Accessed: 25 August 2024).

Project Specific references:

- Basyal, K.(2024). Customer Churn Analysis: Clustering Analysis. [Project Documentation]. Advanced Consultant Service.
- Basyal, K. (2024). clustering_analysis.py. [Script]. Advanced Consultant Service
- Basyal, K. (2024). visualizations.py. [Script]. Advanced Consultant Service
- Basyal, K. (2024). Clustering_analysis.ipynb. [Jupyter Notebook]. Advanced Consultant Service
- Basyal, K. (2024). Visualization.ipynb. [Jupyter Notebook]. Advanced Consultant Service