Capstone Project Report:-

**Customer Segmentation**

Table of Contents

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons For Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

11. Conclusion

12. References

Problem Statement:

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence based insights to provide the same.

Project Objective

The primary aim of this project is to leverage data-driven strategies to understand customer purchase pattern. This objective can be summarized as follows:

* **Data-Driven Insights**: Our goal is to extract valuable insights from the data that will inform and enhance customer segmentation. These insights will be based on a comprehensive analysis of the dataset, shedding light on patterns and trends that can be used to optimize inventory.
* **Segment customers**: We seek to create robust forecasting models capable of clustering customers as per their purchase pattern.

Data Description

The dataset used in this project is "OnlineRetail (3).csv", which contains the following columns:

|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Invoice | Invoice number |
| StockCode | Product ID |
| Description | Product Description |
| Quantity | Quantity of the product |
| InvoiceDate | Date of the invoice |
| Price | Price of the product per unit |
| CustomerID | Customer ID |
| Country | Region of Purchase |

The dataset comprises 541909 rows and 8 columns.

Data Pre-processing Steps and Inspiration:

The pre-processing of the data included the following steps:

* **Data Loading**: The dataset, sourced from the "OnlineRetail (3).csv" file, was loaded into the analytical environment using the pandas library.
* **Exploratory Data Analysis (EDA):** A critical component of data pre-processing, EDA was undertaken for a deeper understanding of the data characteristics unique to each. This process enabled the identification of pertinent insights and trends. We also changed **'InvoiceDate'** to date time format for sales trend. We added an extra feature '**Amount**' by multiplying quantity and price for easy analysis
* **Data Cleansing**: Missing values were carefully checked in whole data and addressed to ensure the dataset's integrity. Additionally, the presence of duplicate entries was also examined. We found null and duplicate values in our data and removed them since we have a huge data removing them will not cause any Impact on our model. We also observed negative values in '**Quantity**' and removed them since they are outliers and will confuse our model.

Choosing the Algorithm for the Project

**K-Means Clustering Algorithm**:

K-Means clustering is a widely used unsupervised machine learning algorithm that serves as a powerful tool for grouping similar data points into clusters. Its primary objective is to partition a dataset into K clusters, where each data point is assigned to the cluster with the nearest mean.

The algorithm operates in the following steps:

1. **Initialization**: K-Means begins by selecting K initial cluster centroids. These centroids represent the center of each cluster. Often, the initial centroids are chosen randomly from the data points, or other strategies can be employed for seeding the centroids.
2. **Assignment**: In this step, each data point is assigned to the cluster with the nearest centroid. The distance between data points and centroids is typically calculated using the Euclidean distance, although other distance metrics can be used based on the nature of the data.
3. **Update**: After assigning all data points to clusters, the next step is to recalculate the centroids of each cluster. The new centroid is determined by computing the mean of all data points within the cluster. This new centroid represents the updated center of the cluster.
4. **Reassignment**: Steps 2 and 3 are iteratively repeated until convergence is reached. Convergence is defined by a predetermined criterion, such as the centroids no longer changing significantly between iterations or reaching a specified number of iterations.

Key Points and Considerations:

**Unsupervised Learning**:

K-Means is an unsupervised learning algorithm, meaning it doesn't require labelled data to perform clustering. It identifies patterns and groups within the data autonomously.

K selection:

One critical aspect of K-Means is selecting the number of clusters, K. This number must be specified in advance. Various methods, such as the elbow method, silhouette score, or domain knowledge, can be used to determine an appropriate value for K.

**Scalability**:

K-Means is computationally efficient and can handle relatively large datasets, making it applicable to datasets with a substantial number of data points.

**Interpretability**:

The clusters formed by K-Means are easy to interpret, as they group data points based on similarity. This interpretability is valuable for understanding the characteristics of each cluster.

**Applications**:

K-Means clustering is applied in various domains, including customer segmentation, image compression, anomaly detection, and recommendation systems.

Model Evaluation and Technique

In this crucial phase of our analysis, we meticulously evaluated the performance of the K-Means clustering model and leveraged various techniques to dissect the characteristics of each customer cluster.

The following steps and methodologies were applied to gain a comprehensive understanding of our customer segments:

1. Data Preparation and Feature Scaling:

Prior to initiating the K-Means clustering process, we undertook a meticulous data preparation procedure. Specifically, we selected pertinent features that are essential for customer segmentation. These included recency, frequency, and monetary value, all of which are highly relevant for discerning customer behaviour.

To ensure that the clustering process remains equitable across all features, we applied the StandardScaler. This standardization technique rescales the features to have a mean of 0 and a standard deviation of 1, preventing any single attribute from dominating the clustering results due to its scale.

1. Choosing the Number of Clusters (K):

Selecting the optimal number of clusters (K) is a pivotal decision in K-Means clustering. To make this determination, we harnessed the elbow curve method.

By subjecting the dataset to multiple iterations of K-Means with varying values of K, we scrutinized the sum of squared distances between data points and cluster centroids, commonly known as the inertia.

The resultant graph, plotting the inertia against K, enabled us to pinpoint the pivotal "elbow point." In our case, this demarcation indicated that K=4 was the ideal number of clusters for our customer segmentation.

1. Applying K-Means:

With the optimal K value identified, we executed the K-Means clustering algorithm, carefully applying it to the standardized data with K=4. This phase involved the iterative assignment of data points to the nearest cluster centroid, followed by the recalibration of these centroids, as outlined in the core K-Means algorithm.

1. Cluster Analysis:

Post-clustering, a comprehensive cluster analysis was conducted to fathom the characteristics of each customer cluster. This facet was pivotal in gaining a profound understanding of the behavior and attributes characterizing each segment.

The analysis encompassed the calculation and interpretation of mean values for key attributes within each cluster. This permitted insights into recency, frequency, and monetary value traits of customers in each segment.

1. Visualizations:

To enhance the interpretability of our cluster results, we judiciously employed data visualizations. These visual aids included:

Bar Plots: We harnessed bar plots to visually depict the average recency, frequency, and monetary value by cluster. This provided a succinct and accessible overview of the distinguishing characteristics of each segment.

Scatter Plots: Scatter plots were instrumental in illustrating the distribution of data points (customers) across clusters. This visual representation allowed us to assess the distinctiveness of customer segments from each other.

3D Scatter Plot: For a more intricate examination, a 3D scatter plot was employed, affording a multidimensional perspective of customer clusters. This advanced visualization was instrumental in offering a profound insight into how clusters diverged concerning recency, frequency, and monetary value.

1. Recommendations:

Building upon the in-depth analysis of each cluster, we were able to craft highly customized recommendations for our online retailer. These recommendations encompassed strategies for marketing and product offerings that were meticulously tailored to align with the unique behavioural patterns characterizing each customer segment.

By skilfully applying these techniques and methodologies, we not only meticulously evaluated our K-Means clustering model but also proffered actionable insights that can propel our online retailer toward informed decision-making, strategical enhancement, and the nurturing of customer satisfaction.

Inferences from the Project

1. Customer Segmentation:

- By using elbow-curve we determined K=4 for KMeans Clustering and segment our customers in 4 clusters

2. Customer Behaviour Understanding:

- Further analysis on our clusters revealed:

1. Avg. Recency was highest in cluster 1.
2. Avg. frequency was highest for customers in cluster 3.
3. Avg. Amount spent was highest for cluster 4 customers.
4. Highest number of people belonged to 2nd cluster
5. Maximum number of transaction were from 3rd cluster
6. Most countries have customers from cluster 2 except Netherland and EIRE
7. Top 5 countries by total sales are : United Kingdom(82%), Netherlands(3.2%),EIRE(3.0%),Germany(2.6%) ,France(2.4%)

Future Possibilities

1. Advanced Clustering Algorithms:

Experiment with advanced clustering algorithms beyond K-Means, such as Hierarchical Clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), or Gaussian Mixture Models. These algorithms may capture more complex patterns in customer behaviour.

1. Ensemble Clustering:

Consider the adoption of ensemble clustering to bolster the model's reliability. Ensemble clustering involves combining results from multiple clustering techniques, mitigating the limitations of individual methods and enhancing the overall understanding of customer behaviour.

1. Deep Learning:

Investigate the integration of deep learning techniques like auto encoders and neural networks. These approaches have the capability to reveal hidden data patterns that conventional methods might overlook, offering the potential for a more detailed insight into customer behaviour.

Conclusion

We were successfully able to implement KMeans clustering to segment of customers based on their purchasing pattern by clustering them In 4 categories using traditional clustering approach

References

* Kaggle (https://www.kaggle.com/): Kaggle, a prominent platform, provided access to valuable datasets and kernels related to your online retail analysis, including clustering techniques. It served as a valuable resource for obtaining real-world data and insights specific to your project.
* Analytics Vidhya (https://www.analyticsvidhya.com/): Analytics Vidhya, a renowned platform for data science and AI enthusiasts, furnished articles, tutorials, and datasets relevant to clustering techniques applied in this project. These resources supported in understanding and applying clustering concepts in the retail domain
* Medium (https://medium.com/): Medium, as a publishing platform, hosted insightful articles, case studies, and tutorials that specifically discussed clustering methods in data science. These resources were valuable in gaining deep mathematical insights into clustering concepts as applied here.
* Towards Data Science (https://towardsdatascience.com): "Towards Data Science," a respected publication on Medium, provided a wealth of articles and insights related to data science and clustering. This publication enriched my understanding of clustering techniques, including the use of K-Means clustering for customer segmentation