**Week 2 Assignment**

Saloni Awale

Department of IT, Westcliff University

TECH 405a: Artificial Neural Network and Deep Learning

Professor Acharya

November 3, 2024

**EDA on Mall Customer Segmentation Dataset**

**Introduction**

Mall Customer Segmentation Dataset is one of the most famous datasets where all necessary details about the customers of a shopping mall are described along with their demographic details and their spending patterns. Gender, Age, Annual Income, Spending Score. These features allow for understanding customer, their preferences and their purchasing actions. This knowledge is especially important for organizations planning to construct effective advertising campaigns, customize the beneficial offers, and enhance buyer interactions. In this use case, it is important to predict the Spending Score based on different features like Annual Income, Age, etc. with the purpose of understanding their interdependencies for making business decisions. The Spending Score is important because it quantifies the amount a customer spends in the mall, measures customer activity and loyalty. The case study knowledge may be useful to retail and malls businesses in order to fill consumer group profiles in relation to income, age, and spending. Finally, such a segmentation allows companies to use more effective marketing techniques, as well as to consistently create better customer experiences that translate into better sales figures.

**Findings from Exploratory Data Analysis (EDA)**

1. **Data Overview:**

The dataset contains 200 records with the following features:

* **CustomerID:** A numerical or an alphanumeric code that uniquely identifies each customers.
* **Gender:** Binary nominal feature that specifies the gender of the customer; male or female.
* **Age:** The age of the customer.
* **Annual Income:** The income per year, from the month the customer becomes associated with the company, divided by 1000.
* **Spending Score:** A score given to the customer by the mall concerning his or her spending habit (ranging from 1 to 100, for a higher spending capacity).

1. **Missing Data:**

As with missing values, there is no data that needs to be imputed, or cleaned in any way, in the current dataset.

1. **Feature Distributions:**

* **Age:** The age of the customers is approximately equal, and they vary from 18 to 70 years old. It also reveals that the median age of the population is about 32 years.
* **Annual Income:** The annual customer income is also skewed right and ranges from just below $30,000 to slightly above $120,000, with the largest group earning between $30,000 and $80,000 annually.
* **Spending Score:** The spending scores distribution is uniform though a few appear in low (10-20), middle (40-60) and higher (80-90) categories.

1. **Gender Analysis:**

* In terms of gender share, the customers are evenly divided with 50% males and 50% females.
* It is clear from the statistical examination of the spending scores using the boxplot that females spend slightly more than males.
* The dispersion of adjusted expenditures is higher among females generalizing that there is higher variability in the spending pattern of female customers.

1. **Correlation Analysis:**

* **Annual Income vs. Spending Score:** Annual Income and Spending Score have a very low negative relationship of - 0.09. This implies that the higher income clientele, does not actually translate to increased numbers of expenditures within the mall.
* **Age vs. Spending Score:** Age is not influential in the spending behavior; the correlation between Age and Spending Score is just 0.03, which means there is a very poor positive correlation.
* **Annual Income vs. Age:** We have a positive linear relationship of 0.50 between Annual Income and Age, which is as expected since they are ages imply that customers belonging to the older generation earn more than their junior ones.

1. **Customer Segments (Clustering):**

The Clustering Model chose was K-means, on the basis of Annual Income and Spending Score. As per the elbow method used in this study, the number of clusters that yielded the best result of K=4 was considered most suitable.

* **Cluster 1:** High income, low spending – Customers with higher incomes but lower spending scores. These might be affluent individuals who are less engaged with the mall.
* **Cluster 2:** Low income, high spending – Customers with lower incomes but high spending behavior. These customers could be value-driven shoppers.
* **Cluster 3:** Medium income, medium spending – Customers with average income and average spending behavior. This is the largest segment of customers.
* **Cluster 4:** High income, high spending – Affluent customers who spend significantly more in the mall.

**7. Pairwise Relationships:**

* The positioning of the Annual Income vs. Spending Score also show no direct correlation of the score with spending though a close relationship is shown by the scores clustering.
* The Customer Segments plot helped to understand that customers can indeed be divided into particular segments, and this division can be made based on income and spending.

**Approach**

1. **Data Preprocessing**

The first way of the proposed approach, therefore, was to carry out data cleaning and create a clean dataset for analysis. For the second step, we conducted a feature encoding to deal with missing values and given that the data set was rather clean, and there were no many missing values, the process of imputation was not necessary. But we also did transform categorical data for example, Gender into numerical for analysis where 0 represented males and a 1 represented females. We also made sure that each of them was of a correct data type for further analysis of the data.

1. **Exploratory Data Analysis**

The analysis of the dataset continued with an EDA for the sake of getting an estimate of the distribution of the values of the envisioned dependent variables and the independent variables, as well as the nature of the correlation that could conceivably exist among these values. Descriptive statistics and exploration of variables Age, Annual Income, and Spending Score were done by means of histograms, while Gender was represented by boxplots which showed how it related to the above mentioned variables. Likewise, the correlation matrix was employed to check overall correlation between the numerical variables such as Annual Income, Age and Spending Score.

1. **Customer Segmentation for Kinds trailer Using K-Means Clustering**

After we were clear about the structure, we carry out K-means clustering in order to classify the customers according to their Annual Income and Spending Score. We decided on 4 clusters using the Elbow Method and then categorized customers into segmented sets depending on their total spending and total income. With the help of this segmentation, it became possible to distinguish high\_invite value customers, value\_invite driven consumers and Invite other target audiences with whom certain types of marketing strategies can be used.

1. **Visualization and Insights**

Lastly, we overlaid the actual clusters and then explored the customer segments by income, spending and demography. This was useful in giving the firm insights that could be utilised in the marketing of products, in the development of a specific marketing strategy and in the management of the customer relations.

**Conclusion**

In conclusion, it presents an exploratory data analysis of the Mall Customer Segmentation Dataset to provide insights for customer targeting and marketing strategies. Gender was concluded to have an effect on the spending behaviour, with the female gender getting a slightly higher spending score than the male gender. Moreover, as showed in the analysis, annual income and spending score are only slightly related to each other; therefore, income categorization does not necessarily indicate how a customer will spent money. The K-means clustering analysis segmented customers into four distinct groups based on their income and spending scores: category consumers, conscience-orientated consumers, rational buyers, routine-orientated customers, saturated consumers, total engaged consumers. These customer segments offer a strategic guide to the specific marketing approaches that should be implemented. For instance, one can easily promote the product or service to those with a tendency of choosing products based on their perceived value while on the other end, a luxury product may be promoted more easily among the high class population. In conclusion, the present paper emphasizes the need to consider factors apart from income when analyzing the customer’s behaviour and shows how segmentation may aid in making better marketing and business-related decisions.

**Reference**

*Mall Customer Segmentation data*. (2018b, August 11). Kaggle. <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>