#ALI JAWED DELAWARI GH1024093

GITHUB: <https://github.com/alijaweddelawari/B198.C5>

Customer Spending Prediction and Segmentation for Direct Marketing

**Introduction**

In today's very highl competitive business environment, direct marketing has become an essential method and strategy for companies aiming to reach targeted customers efficiently and is also verycost effective in general. By using data driven insights, businesses can tailor campaigns to the right audience to improve conversion rates and reduce wasted marketing effort.

A central aspect of successful direct marketing is the ability to understand your customers and thier behavior, particularly when it comes to spending patterns and customer segmentation. Accurately predicting how much a customer is likely to spend and identifying groups of people with similar characteristics enables marketers to deliver personalized offers, allocate resources strategically and improve overall return on investment (ROI).

**This project explores the question:**

Can we use past customer attributes and behaviors to predict how much they might spend or segment them for marketing campaigns?

To investigate this, below we use a real world dataset consisting of 1,000 customer records from a direct marketing context. Each record includes all the datas including demographic and behavioral attributes such as age group, gender, income, marital status, home ownership, number of catalogs received, and historical customer value. The main target variable for us is the Amount Spent by the customers in the dataset.

**Objectives of this project are:**

To predict customers spending using regression based techniques.

To segment customers into meaningful groups using unsupervised learnings (clustering).

To generate business insights and ideas that can support more effective marketing strategies for us.

By applying appropriate data science methods to this dataset in hand , we aim to provide a very structured, data driven framework that helps our business make informed marketing decisions and improve campaign outcomes.

**Data Exploration & Cleaning**

To begin the analysis, the dataset below was loaded and inspected using the pandas library in Python coding system. The dataset contains 1,000 records and 10 columns, which consists mix of categorical and numerical features such as age, gender, marital status, salary, and amount spent. An initial inspection done by us revealed that the data was mostly clean, with the exception of the History column, which had approximately 30% missing values.

To address this issue we chose to fill the missing values with a new category labeled as Unknown. This approach is a commonly accepted method in categorical feature handling, especially when the missing values themslves may carry informative signals or when maintaining dataset size is important. This strategy helps to avoid unnecessary data loss and keeps the model's integrity intact. As recommended by Towards Data science (2020), using a distinct label for unknown categories in classification or segmentation tasks can improve model interpretability and preserve information quality.

Following the cleaning step done below, descriptive statistics and visualizations were generated using seaborn and matplotlib to explore data distributions and potential patterns for us. Key findings included a right skewed distribution in customer spending with a median around €1,000 and visual trends suggesting that middle aged, married homeowners tend to spend more. These early insights provide valuable context for the modeling and segmentation steps that follow.

A white background with blue text

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A close up of a word

AI-generated content may be incorrect.A computer screen shot of a couple of words

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A graph of a distribution of amount spent

AI-generated content may be incorrect.A chart of different colored squares

AI-generated content may be incorrect.A blue and orange rectangular bars

AI-generated content may be incorrect.

**Feature Engineering**

To ensure compatibility with machine learning models, the dataset then underwent a series of feature transformations below. First all categorical variables such as age, gender, ownhome, married, location, and history were transformed into numerical format using one hot encoding. This approach is commonly used to represent non numeric data in a shape that avoids implying any ordinal relationship between categories like: Own is not greater than Rent. One hot encoding is widely recommended for it’s simplicity and effectiveness overall in preserving the meaning of categorical data while enabling algorithm compatibility (Brownlee, 2021).

Next, min max scaling was applied to numerical variables below such as salary, children, catalogs,amountspent etc. This normalization ensures that features are on the same scale which is particularly beneficial for all models that rely on distance metrics, such as clustering algorthms. Without scaling variables with larger numeric ranges could somehow influence model output.

Finally a new binary feature named highincome was engineered in the below coding to identify customers whose salaries are above the dataset’s median. This feature may help in classifying or segmenting high value customers for targeted marketing strategies.

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A white background with red and black text

AI-generated content may be incorrect.

**Modeling & Analysis**

In this stage of the project in the code below we applied both unsupervised and supervised learning techniques to extract insights from the customer datasets. First for the customer segmentation we used the Kmeans clustering algorithm to group customers based on their demographic and behavioral features. The optimal number of clusters was determined using a method called the Elbow Method and we selected k = 4 as the best fit. Each customer was assigned to a cluster and the groups were visualized using principal component analysis (PCA). The resulting clusters revealed distinct customer profiles such as high-income low-spenders, frequent catalog users and younger customers with lower budgets but offering valuable insights for targeted marketing strategies.

A screenshot of a computer program

AI-generated content may be incorrect.A graph of a number of clusters

AI-generated content may be incorrect.A screenshot of a computer code

AI-generated content may be incorrect.A screen shot of a computer code

AI-generated content may be incorrect.A chart with many dots

AI-generated content may be incorrect.

**Regression (Spending Prediction)**

Alongside segmentation, we had to implement a spending prediction model using both Linear Regression and Random Forest Regression. The data was split by us into training and testing sets (80/20), and model performance was evaluated using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The Random forest model achieved better accuracy and provided insight into which features contributed most to spending behavior. Key drivers included salary, catalogs, and historical customer value (History). This dual approach segmenting customers while also predcting their potential value offers businesses a comprehensive way to both understand and also act on customer data. This modeling strategy that we did aligns with best practices in applied machine learning, where ensemble models and hybrid approaches are often recommended for richer insights (Koehrsen, 2018).

A computer screen shot of a program code

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.A graph with numbers and a bar

AI-generated content may be incorrect.

**Evaluation**

To assess model performance both the numerical accuracy metrics and business relevance were considered here. For the regression models we compared Linear Regression and Random Forest Regression. The Random Forest outperformed the Linear Regression, achieving a lower RMSE of 0.0612 and MAE of 0.0420, indicating higher predictive accuracy on scaled data. This supports the idea that ensemble models like Random Forest are better suited for capturing complex patterns in marketing datasets especially when relationships between features and outcomes are nonlinear (Brownlee, 2021).

For customer segmentation we did apply the Kmeans algorithm and evaluated cluster quality using the inertia score which in was 1176.79 for four clusters. The Elbow Method that we used confirmed that k = 4 produced a reasonable trade of between simplicity and accuracy. The segments then uncovered meaningful customer profiles which can be used by our marketing team to personalize outreach for example targeting high income customers who currently spend little amount or rewarding loyal frequent buyers.

Despite these promising results there are some limitations like always. The dataset is small with like 1,000 records and lacks time based behavior which prevents trend analysis for us. The column History also introduces ambiguity as it is not clearly defined. Nonetheless, the analysis was conducted in here with no use of personal data and fair treatment was maintained across all groups. These considerations reflect responsible and ethical data science practices which are very important for us and in commercial applications.

A screenshot of a computer

AI-generated content may be incorrect.A white rectangular object with blue text

AI-generated content may be incorrect.A screenshot of a computer

AI-generated content may be incorrect.

**Conclusion**

This project that we did demonstrated how data science techniques can be effectively applied to support decision making in direct marketing. By reading, exploring and analyzing a real world dataset of 1,000 customer records we were able to uncover meaningful infos and insights into customer spending behavior and segment the customer base into actionable groups.

Using Kmeans clustering we identified four distinct customer segments each characterized by unique demographic and behavioral patterns. These clusters can inform targeted marketing strategies allowing businesses to better allocate resources, personalize offers and improve campaign effectiveness in general. In parallel we built a very predictive models using Linear Regression and Random Forest Regression to estimate how much a customer might spend. The Random Forest model delivered the best results with a low RMSE and MAE, in the end highlighting the value of ensemble learning methods in business applications.

Throughout the project that we did, we ensured data quality through proper cleaning and preprocessing steps, handled all categorical variables using encoding, and normalized continuous values to enable robust modeling. Ethical considerations were also taken into account no personal identifiers were used by us and also fairness was maintained in all analytical processes.

Overall, this end to end analysis provided a practical framework for combining segmentation and prediction to enhance direct marketing performance. These techniques. Of us if applied in a real business environment could help companies identify high-value customers, reduce marketing waste, and improve customer retention through data driven personalization.

**References**

Patel, D. (2020) ‘Handling categorical data in machine learning using Pandas and Scikit-Learn’, *Towards Data Science*, 5 February. Available at: <https://towardsdatascience.com/handling-categorical-data-in-machine-learning-using-pandas-and-scikit-learn-5d8d8f77530f> (Accessed: 11 June 2025).

Brownlee, J. (2021) ‘How to one hot encode categorical data for machine learning’, *Machine Learning Mastery*, 27 January. Available at: <https://machinelearningmastery.com/how-to-one-hot-encode-categorical-data-for-machine-learning/>(Accessed: 11 June 2025).

Koehrsen, W. (2018) ‘A conceptual explanation of random forests’, *Towards Data Science*, 19 October. Available at: <https://towardsdatascience.com/random-forest-explained-9d566bde7839> (Accessed: 11 June 2025).

Brownlee, J. (2021) ‘How to choose a machine learning model’, *Machine Learning Mastery*, 2 April. Available at: <https://machinelearningmastery.com/how-to-choose-a-machine-learning-model/> (Accessed: 11 June 2025).

Towards Data Science. (2020) ‘Understanding KMeans clustering’, *Medium*, 7 April. Available at: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a> (Accessed: 11 June 2025).