Course Name: Programming for Data Analytics
Course Code: 04638
Assignment Title: Customer Segmentation and Classification using Machine Learning
Report Title: Data Analytics: Predictive Analytics: Customer Behavioral Segmentation using Machine Learning
Instructor: George Okeyo, FHEA, Ph.D.

Name: OLATUNJI Damilare Emmanuel
AndrewID: dolatunj
Submission Date: 15th December 2023.

Abstract

In today's highly competitive business landscape, understanding customers and effectively targeting them with personalized marketing strategies is essential for success. As such, customer segmentation play an important role for achieving this goal. With the advancements in artificial intelligence techniques, this research project explores the application of machine learning algorithms in customer segmentation, querying the given dataset to find the optimal number of clusters, explore the possibility of deploying a machine learning model and performing evaluation metrics to justify the efficiency of the chosen model performance. The utilized dataset (approx. 9000) records with 18 distinct variables suggested the optimal number of clusters to be three (30) and the evaluation metric (f1-score, recall, and precision) achieving 96.3% scoring score respectively. In conclusion, the study demonstrates the efficacy of machine learning in customer segmentation, offering a foundation for its application in marketing strategies to meet business objectives.

Background and Problem Description

In the face of growing competition, businesses have turned to analysing historical data using machine learning techniques [1][2]. By leveraging this technology, businesses now unlock hidden patterns and gain deeper insights into customer behaviour and needs as the develop knowledge can help develop more effective marketing strategies and ultimately increase customer satisfaction [3].

[4] demonstrated the effectiveness of clustering technique as an important step in data mining process to develop a real time and online system for a supermarket. The developed model receives inputs directly from sales data records and automatically updated segmentation statistics at the end of day's business. [5], on the other hand analysed Pakistan's largest e-commerce dataset by introducing k-means, Gaussian, and other algorithms for segmentation. The adopted cluster analysis method includes elbow, dendrogram and many others.

Consequently, in this report, an exploratory analysis of how to apply machine learning algorithm for customer segmentation and classification was performed. In addition, this report explores the importance of feature importance while using machine learning algorithms to classifier customers into clusters. Lastly, this project explores the possibility of deploying models using Flask.

Methods

The dataset contains information about 9,000 active credit card users' behaviour over six months. It includes 18 attributes describing their spending habits and purchase behaviour for analysis. Fig 1. explains the methods.

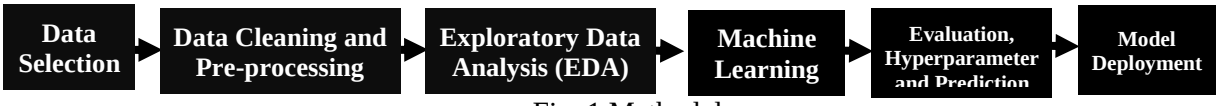


Fig. 1 Methodology

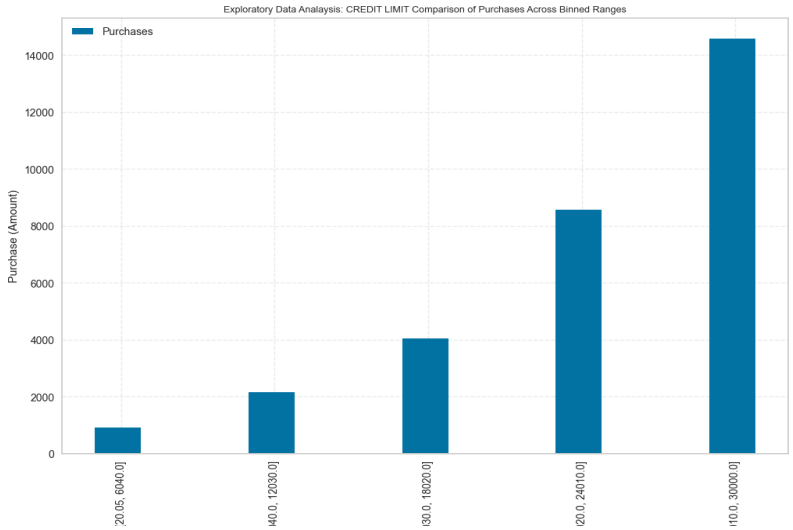
Explanation:

1. The foundational approach involved data download from Kaggle, followed by data cleaning and pre-processing for quality control. Some of the pre-processing activities include removing null values, checking for duplicated records, and checking for outliers.
2. In the exploratory data analysis phase, the data was analysed to uncover underlying patterns. This in-depth analysis provided information on about 15 features being positively skewed and two (2) being negatively skewed. Coupled with this about 448 records were identified as outliers reflecting extreme deviation from normal customer patterns.
3. For the segmentation, clustering algorithm K-means was employed to identify unique customer groups based on their similarities and the fact that we lack prior knowledge about the dataset. The effectiveness of the segmentation was evaluated using silhouette score and other evaluation metrics to ascertain the accuracy or equally precision of the customer groups formed.
4. In addition to this is the introduction of a supervised learning model – RandomForest Classifier to perform the classification operation. In addition to evaluate the accuracy of the clustering provided using k-Means. There also, important features for predicting customers behaviour were analysed.

Features known to be important are features with significance above the average and this includes six (5) features.

5. The culmination of this process was the deployment of the classification model into an operational environment using Flask.

Results and Discussion



Exploratory Data Analysis

Fig 2. show a trend where the purchase amounts increase as the credit limit ranges increase. Notably, the highest range, "(24000, 30000]", has the largest purchase amount, suggesting that customers with higher credit limits tend to make more purchases, or more expensive purchases, compared to those with lower limits.

Fig 2. EDA: Credit Card Limit Comparison with Purchases Across Binned Ranges

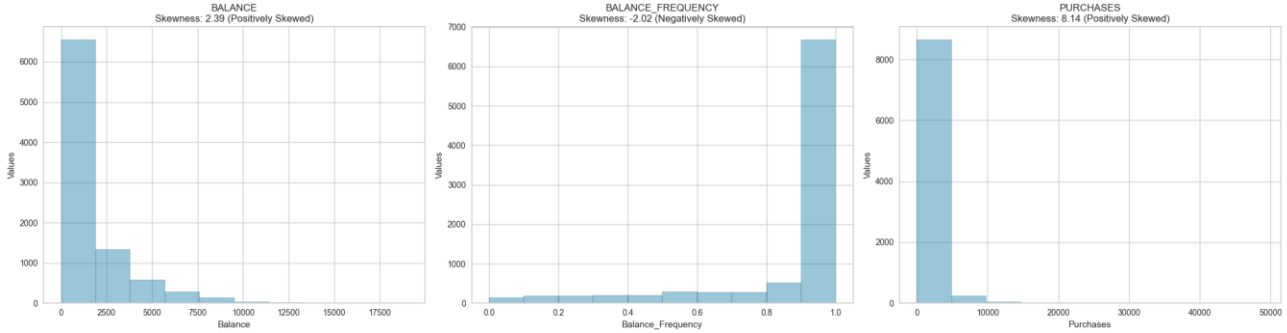


Fig 3 Data distribution that characterizes each feature.

Based on analysis, it became evident that 15 of the given features in the given dataset are positively skewed. Equally, about two (2) are negatively skewed. The positively skewed distribution show a more extended or bulkier tail compared to the left, with the bulk of the distribution's density leaning towards the left side while for the negatively skewed, the distribution density is predominantly situated on the right side.

Cluster Segmentation

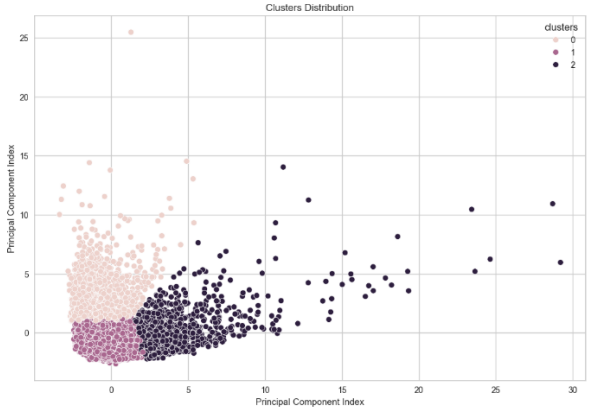


Fig. 3 Customers' Cluster Segmentation

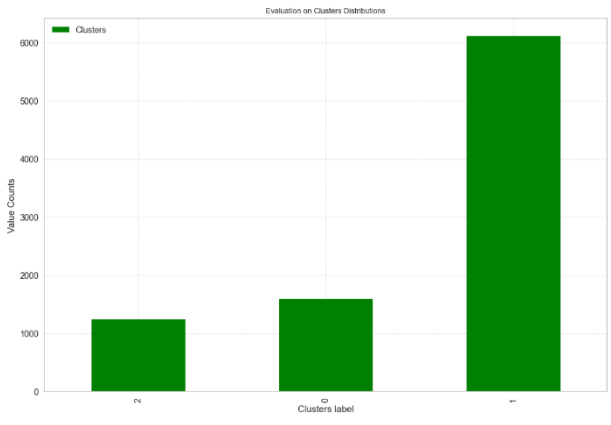


Fig. 4 Bar Chart of Cluster Sizes

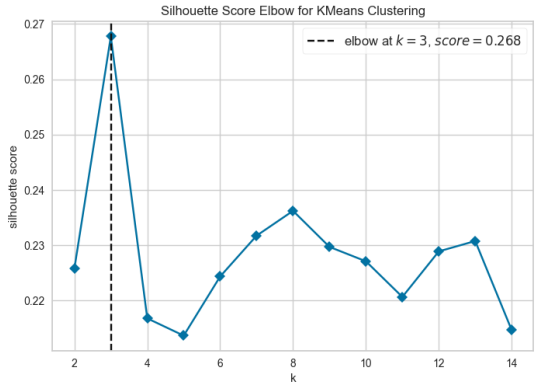
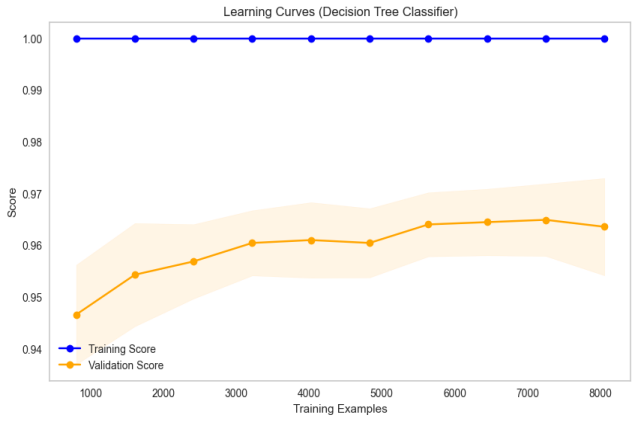


Fig. 3 – (Customers' Cluster Segmentation) shows how the customers behaviour fits into three distinct clusters (0, 1, and 2). Clusters 0 and 1 are more densely packed and located towards the lower end of the Principal Component Intensity axis, indicating tighter grouping and possibly more similarity within those clusters. Cluster 2 is spread out across a higher range of the Principal Component Intensity axis, suggesting more variation within that cluster.

Fig. 4 – Bar Chart of Cluster Sizes shows the count of data points within each of the three clusters. Cluster 1 contains majority of the data points, next to it is cluster 0 with moderate count and cluster 2 with the fewest. This distribution suggests that the dataset is heavily skewed towards the characteristics defined by Cluster 1.

Fig. 5 which plots the silhouette score against the number of clusters suggest k=3, where the silhouette score is approximately 0.268 as the optimal number of clusters for the given dataset. This is confirmed by the dashed line indicating the elbow point, which is used to choose the best value for k in KMeans clustering. A score of 0.268 indicates moderate separation between the clusters.

Fig. 6 Model Debugging



Based on analysis, both the training and validation scores increases and stabilize at a high value as more data is used for training. This indicates that the model is learning well from the data and generalizing effectively.

Did I notice overfitting or underfitting ? No

Overfitting is typically identified when the training score greatly surpasses the validation score, resulting in a substantial gap between them. Similarly, underfitting is characterized by both low training and validation scores that are closely aligned.

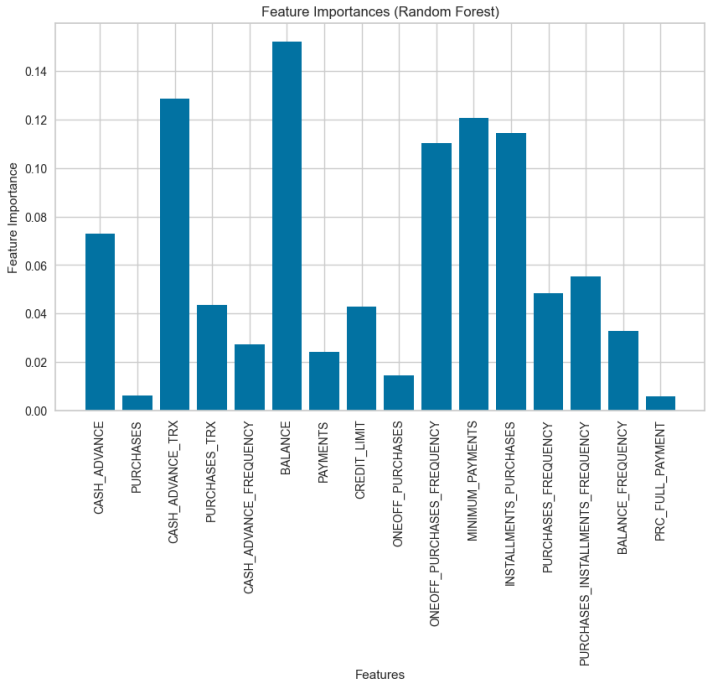


Fig 6b Feature Importance using Feature Importance on RandomForest Classifier

Fig 6b, Random Forest Model Features Importance indicates 'balance', 'purchases', 'cash_advance', 'cash_advance_frequency', 'cash_advance_trx', 'purchases_trx' are the top six (6) features above the mean. The was set as the threshold.

On the contrary, ‘purchases’, ‘payments’, ‘prc_full_payment’are among the least influential features.

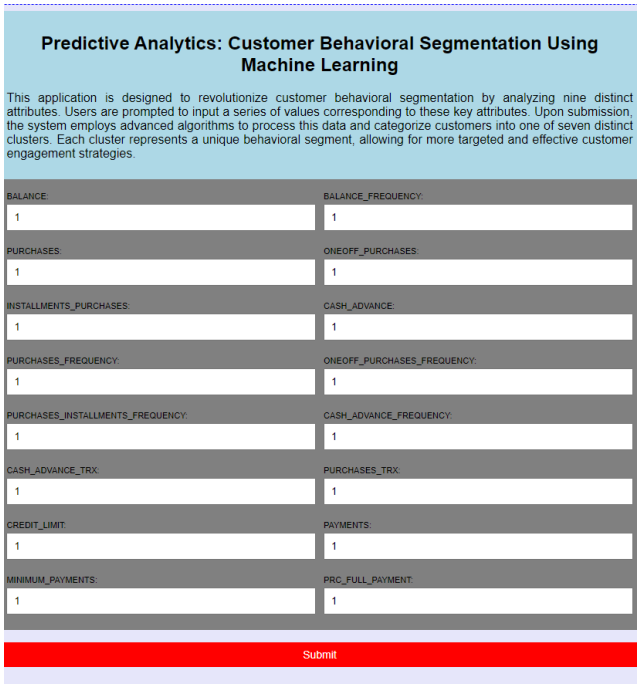


Fig. 7. Screenshots of the web application's input and result pages

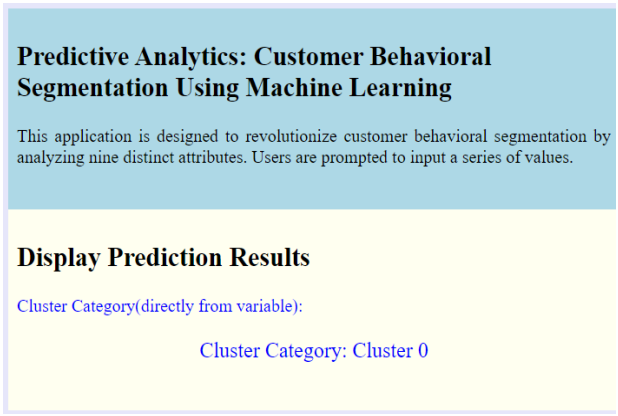
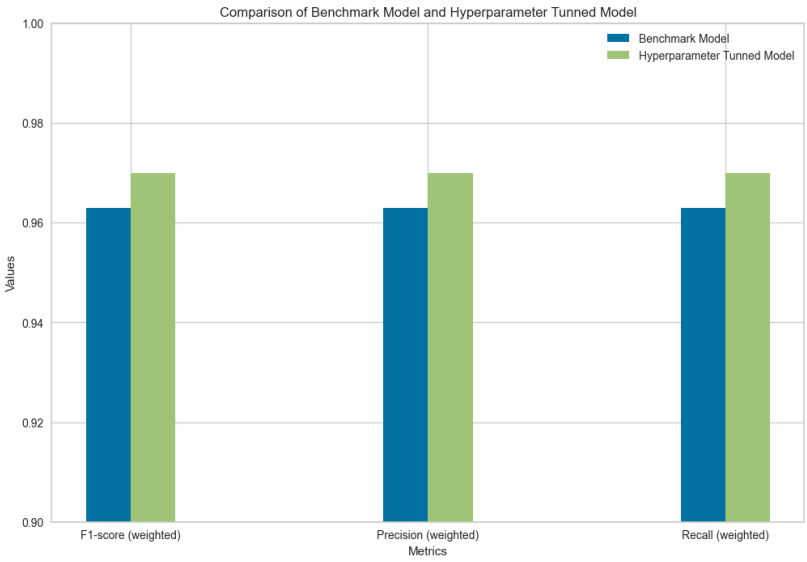


Fig. 8. Screenshots of the web application's result page

Fig 7 and 8 provides an overview of the user interface for cluster's prediction and the result. Users are expected to enter numeric values only. The possible clusters include – Cluster 0 and Cluster 1 and Cluster 3



Considering the F1-score, which balances Precision and Recall, the Hyperparameter Tuned Model exhibits a modest enhancement relative to the Benchmark Model. The advancement in the scoring with the Hyperparameter Tuned Model could be attributed to optimal parameters being passed in as estimators. Hence the hyperparameter tuned model can be considered as the final benchmark model.

Conclusion

"Clustering provides an effective way to understand and anticipate customer behaviors and expectations. Based on our analysis, leveraging cutting-edge machine learning techniques can help establish a solid foundation for marketing strategies targeting customers. Findings from this study reflect that to improve the evaluation metric of an algorithm process, hyperparameter tuning is important. The F1-score, Recall, and Precision value without hyperparameter tuning is 96.3%; however, after a grid search operation, the best estimators yielded an increase in the corresponding values of the model's predictive ability. The highest achieved result is 97.1%. Additionally, scaling data in a production environment provides easy access to predict new customer behaviors based on the historical dataset."

References

[1] A. Hagiu and J. Wright, “When Data Creates Competitive Advantage,” Harvard Business Review, Feb. 2020. https://hbr.org/2020/01/when-data-creates-competitive-advantage

[2] C. Brewis, S. Dibb, and M. Meadows, “Leveraging big data for strategic marketing: A dynamic capabilities model for incumbent firms,” Technological Forecasting and Social Change, vol. 190, p. 122402, May 2023, doi: https://doi.org/10.1016/j.techfore.2023.122402.

[3] “The Leverage Customer Data Business Model Explained | Untaylored,” www.untaylored.com. https://www.untaylored.com/post/the-explained-leverage-customer-data-business-model (accessed Dec. 17, 2023).

[4] K.R. Kashwan and C.M. Velu, “Customer Segmentation Using Clustering and Data Mining Techniques,” ResearchGate, 2013. https://www.researchgate.net/publication/271302240_Customer_Segmentation_Using_Clustering_and_Data_Mining_Techniques

[5] A. Ullah et al., “Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time,” Sensors, vol. 23, no. 6, p. 3180, Jan. 2023, doi: https://doi.org/10.3390/s23063180.