**Data Science Project**

# **Executive Summary**

# **Literature Review**

## **Machine Learning & Data Mining**

Literature tells us that machine learning is categorised into 2 types, supervised and unsupervised learning (Kouhalvandi et al, 2020; Akmal, 2020). According to Pradana and Ha (2021) supervised learning is usually used to solve problems such as classification and regression, meaning that there is a target label to predict the future. On the other hand, unsupervised learning techniques do not always have a target label or targets to predict, for example, clustering is based on it's mathematical model (Hidayat et al, 2018).

“By means of the internet and database technologies, huge amounts of data about markets and customers has now become available to be exploited and this enables researchers and practitioners to make use of sophisticated data analysis techniques apart from the traditional multivariate statistical tools. These sophisticated techniques are a family of either data mining or machine learning research.” (Hiziroglu, 2013).

Kansal et al (2019) says that today, businesses run on the basis of such innovation that they have the ability to captivate their customers with products, but they can sometimes be at such large volumes that customers can become overwhelmed. They go on to say that customers can become confused on what to buy and what to not and so companies become nonplussed about what section of customers to target their products (Kansal et al, 2019). Machine learning helps with this as applying various algorithms helps unravel the hidden patterns in the data for better decision making in the future (Kansal et al, 2019). Mar (2017) goes onto say that there is no training data set when it comes to machine learning and therefore is it not known what the outcome would be; the AI can’t see the problem, it only has its faultless logical operations to guide it.

## **Market Segmentation**

“In this era, increasing the level of consumer consumption is very reasonable, this is based on the very fast development of production. This makes each person feel like they have an obligation to spend something to enjoy these developments” (Pradana and Ha, 2021). Calvo-Porral and Lévy -Mangin (2018) state that Increasing the number and variance of products is not a bad thing for the market, but when increasing customers, it’s important to tailer the strategy in order to not waste resources aiming a strategy at the wrong customer.

Smith (1956), an American marketing researcher, first brought up the concept of market segmentation. Chiu et al (2009) goes on to say that Smith’s concept was further promoted by following academics and researchers and applied by many companies. Nairn & Berthon (2003) define a market segment as “a set of customers who have similar characteristics of demography, behaviours, values and so on”.

“Research shows that detecting where a customer is going to meet their shopping needs is highly dependent on the service from the provider and the characteristics of the place they’re going to” (Pradana and Ha, 2021). From the perspective of McGoldrick & Thompson (1992) they indicate that the level of price, crowd, convenience and service are very vital factors in a consumers choice to buy.

## **K-Means Clustering**

The goal of K-Means is to group all the data available into non-overlapping sub-groups that are distinct from each other. That means each sub-group/cluster will consist of features that distinguish them from other clusters. (Selvaraj, 2021)

The selection of segmentation techniques has become more important due to the fact that the developments in information and communication technologies, especially database management systems and data mining have changed the way of marketing. The vast availability of data and the inefficient performance of traditional statistical techniques (or statistics-oriented segmentation tools) on such voluminous data have stimulated researchers to find effective segmentation tools in order to discover useful information about their markets and customers. Thus, knowledge discovery (KD) and data mining (DM) have been seen as a solution to this problem. (Hiziroglu, 2013)

K-means is the most frequently used customer segmentation method in our surveyed literature (41 of 105). One answer is that k-means is simple to implement and an established approach. In contrast, other approaches like EAs, hierarchical clustering, or SOMs are more complex according to how the run time or space requirements grow as the input size grows and it needs more effort to implement them (Firdaus and Uddin 2015). Considering the data dimensionality which is used in the publications we see that k-means approaches can handle a larger amount of data and is in pair with latent class approaches. As we mentioned, the hybrid approach that uses the largest amount of data is a combination of the latent class model. However, concerning the number of customers in the data which are the objective of the clustering, the numbers rarely exceed the 10,000. This indicates that clustering approaches need an appropriate feature selection method to deal with a larger amount of data. Based on our findings and analysis, we recommend using k-means or rule-based segmentation approaches which are easy to use and implement, to partition different customers for e-commerce use cases (Gomes and Meisen 2023).

One of the major problems of the k-means algorithm is that it may produce empty clusters depending on initial center vectors. For static execution of the k-means, this problem is considered insignificant and can be solved by executing the algorithm for a number of times (Pakhira. 2009).

# **Data**

The data used for this project was sourced from Kaggle. Pryzbyla (2020) describes some benefits to using Kaggle as it’s vast amount of easily searchable data sets, being able to take inspiration from the community/existing projects and learn from best practice from existing code notebooks.

Link to dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

## **Dataset features:**

*The data is divided into 4 categories*

### **1. Customer Information**

* ID : Customer’s unique identifier
* Year\_Birth : Customer’s year of birth
* Education : Customer’s level of education
* Marital\_Status : Customer’s marital status
* Income : Customer’s annual household income
* Kidhome : Number of children within the customer’s household
* Teenhome : Number of teens within the customer’s household
* Recency : Number of days since the customer’s last purchase
* Complain : If the customer complained within the last 2 years; 1 = yes, 0 = no

### **2. Shopping Habits**

* MntWines : Amount spent on wine in the last 2 years
* MntFruits : Amount spent on fruit in the last 2 years
* MntMeatProducts : Amount spent on meat in the last 2 years
* MntFishProducts : Amount spent on fish in the last 2 years
* MntSweetProducts : Amount spent on sweets in the last 2 years
* MntGoldProds : Amount spent on gold in the last 2 years

### **3. Promotional Interaction**

* NumDealsPurchases : Number of purchases made with discount applied
* AcceptedCmp1 : If the customer accepted the offer in the 1st campaign; 1 = yes, 0 = no
* AcceptedCmp2 : If the customer accepted the offer in the 2nd campaign; 1 = yes, 0 = no
* AcceptedCmp3 : If the customer accepted the offer in the 3rd campaign; 1 = yes, 0 = no
* AcceptedCmp4 : If the customer accepted the offer in the 4th campaign; 1 = yes, 0 = no
* AcceptedCmp5 : If the customer accepted the offer in the 5th campaign; 1 = yes, 0 = no
* Response : If the customer accepted the offer in the last campaign; 1 = yes, 0 = no

### **4. Buying Method**

* NumWebPurchases : Number of purchases the customer has made via the website
* NumCatalogPurchases : Number of purchases the customer has made via the catalogue
* NumStorePurchases : Number of purchases the customer has made via the store
* NumWebVisitsMonth : Number of time the customer has visited the website in the last month

## **ETL Pipeline**

(Diagram)

Once the dataset was downloaded from Kaggle in the form of a CSV, I began the transformation of the data to ensure my model wasn’t impacted by any bad data. Excel was my chosen tool for the the majority of the data transformation as I’m most familiar and comfortable with this tool. “Excel itself has a vast number of functions that allow you to transform data, change the spreadsheet’s appearance, perform various mathematical operations, and many others” (Yung, 2022), these functions meet the requirements to effectively perform the transformation. Once loaded, it was clear that Excel didn’t recognise the delimiters and so included all of the fields within the same column. I used Excel’s ‘Text to Columns’ wizard to separate the data fields into columns to give me effective visibility of the data to be able to carry out the transformation.

The transformation stage began with a comprehensive data quality audit. “Performing data quality audits is crucial for maintaining high-quality datasets, which in turn drive better decision-making and success” (Warsame, 2023). Battson (2019) explains that poor quality data may mean that any marketing campaign derived from the findings of this project may be ineffective and could even lead to being unlawful in terms of data protection. The data quality audit was conducted in accordance with the Government Data Quality Hub (2021) 6 data quality dimensions:

|  |  |
| --- | --- |
| Dimension | Action |
| Accuracy | Given the chosen analysis method, K-Means clustering, is an unsupervised machine learning technique its important to identify and remove any outliers during the audit. Franklin (2019) tells us that “since K-Means algorithm is about finding the mean of clusters, the algorithm is influenced by outliers” potentially impacting the model’s predictive modelling performance. The data was scanned, using the filters in Excel to ensure all values were within the expected range. |
| Completeness | According to Patil et al (2010), complete datasets are a requirement for most statistical and data mining techniques, and therefore these technique’s accuracy is impacted by missing values. There were 24 blank fields within the income column, this column is considered essential for this project as income is an important way to segment customers. Therefore, these rows were removed from the dataset for completeness. |
| Uniqueness | Conditional formatting was used to identify any duplicate values within the dataset where unique values were required e.g. the ‘ID’ field. No duplicate values were identified. |
| Consistency | The data was manually scanned and spot checked for consistency errors, however, none were found. |
| Timeliness | The majority of the data included in the set, is sales data from the last 2 years of trading and is therefore relevant for this project. No timeliness issues were found. |
| Validity | All values included within the dataset were valid. No issues found. |

Following the data quality audit, I converted the customers year of birth to their current age so that there were no date formats within the dataset to aid with the next steps of the transformation performed.

The next step within the transformation process was to standardise the data. Dabbura (2018) tells us that “since clustering algorithms, including k-means, use distance-based measurements to determine the similarity between data points, it’s recommended to standardize the data to have a mean of zero and a standard deviation of one”. The data set came with varying units of measurement across the variables, for example, income vs whether the customer has complained within the last 2 years. One of the most popular standardisation methods, according to Jaadi (2023), is using the Z-score. By subtracting the mean and dividing by the standard deviation for each value, the recommended mean of zero and standard deviation of one is achieved. This was conducted for all numerical variables within the data set.

The rest of the data transformation was conducted using Python and so the data was loaded into Jupyter Notebook. Jupyter Notebook was the chosen IDE because I’m new to using Python and the majority of online learning I had done in preparation for this task used Jupyter Notebook and so to set myself up for success, using a new form of code, I used a tool I was familiar with. A benefit of this tool is that just by typing ‘Shift’ and ‘Enter’ you can view the results of the code, making it easier for me to see if my code works or not. There are however some down sides to using this IDE that will need to be considered as I progress throughout the project. Tran (2020) talks about some of the disadvantages including it being easy to lose track of code as the code gets bigger, needing to find and rerun cells to experiment and it being difficulty to debug.

I then moved conducted the Categorical Variables Encoding as “Most of the Machine learning algorithms cannot handle categorical variables unless we convert them to numerical values” (Roy, 2019). Some of the variables, such as ‘Marital Status’ have categorical values such as ‘Single’ or ‘Married’ and so need to be coded into a numerical value in order for the algorithm to work. One hot encoding was the chosen method for this, Fawcett (2021) tells us that this method is useful when there is no relationship between the values.

**Analysis** - this could be in the form of an explanatory text with screenshots evidencing your analysis, slides, or mark-up explaining your decisions, techniques, and key outcomes.

* I will firstly provide a hypothesis (There is a relationship between customer shopping behaviour and their personality traits) and a null hypothesis (There is not a relationship between customer shopping behaviour and their personality traits).
* I will talk about experimental design elements such as the aim of the study, discussed by Dabbura (2018), as to “try to find homogeneous subgroups within the data such that in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance”. I will talk about this being considered an unsupervised technique due to there not being a version of the truth to compare the output with.
* How the value of K was establied/elbow method.
* The advantages of segmentation i.e. helping the marketing team tailor their efforts/develop products for customers based on their needs and improve their experience. As well as any disadvantages e.g. sensitivity to outliers and potential difficulty of finding K.
* Screenshots of the code and discussion of outputs. I will talk about whether the model met the requirements and whether the hypothesis was confirmed.
* Finally, I will discuss any risk of bias and any mitigation that was taken to mitigate this. Also any privacy factors (although there aren’t any).
* Elbow plot (to help us choose the optimum amount of clusters)
* Silhouette coefficient (used to evaluate the quality of the clusters created by the algorithm.

**Visualisations** - visualisations, dashboards, and any other relevant tools of effective data communication.

* I’m planning to build a dashboard to provide the ability to filter between the different cluster groups to give information on their differing personality traits. I will include screenshots of the dashboard.
* I will discuss the different elements of the dashboard as well a discussion on decisions that were made to aid the story telling, any trade-offs etc.

# **Recommended Improvements**

The data quality audit was conducted manually in excel which is the most robust process and is open to potential human error. In order to ensure all data quality issues are identified, the transformation stage of the project could have been done with Python like the rest of the project.

With more time, the dimensionality reduction algorithm could have been evaluated on it’s performance to ensure the data is representative of the dataset and the relationships within it.

# **References**

Akmal, A,. (2020). Predicting Dropout on E-Learning using Machine Learning. Journal of Applied Data Sciences. 1(1), pp.29-34. [Online]. Available at: https://doi.org/10.47738/jads.v1i1.9 [Accessed 31 July 2023].

Battson, D. (2019). Why data auditing is a necessity. [Online]. Data HQ. Last Updated: 19 February 2019. Available at: https://datahq.co.uk/knowledge-hub/blog/why-data-auditing-is-a-necessity [Accessed 1 August 2023].

Brownlee, J. (2020). 6 Dimensionality Reduction Algorithms With Python. [Online]. Machine Learning Mastery. Last Updated: 17 August 2020. Available at: https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/ [Accessed 10 August 2023].

Calvo-Porral, C. Levy-Mangin, J. (2019). Profiling shopping mall customers during hard times. Journal of Retailing and Consumer Services. 48(x), pp.238-246. [Online]. Available at: https://doi.org/10.1016/j.jretconser.2019.02.023 [Accessed 31 July 2023].

Chiu, C. Chen, Y. Kuo, I. Ku, H. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. Expert Systems with Applications. 36(3), pp.4558-4565. [Online]. Available at: https://doi.org/10.1016/j.eswa.2008.05.029 [Accessed 24 July 2023].

Dabbura, I. (2018). K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks. [Online]. Towards Data Science. Last Updated: 17 September 2018. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-draw [Accessed 2 July 2023].

Fawcett, A. (2021). Data Science in 5 Minutes: What is One Hot Encoding?. [Online]. Educative. Last Updated: 11 February 2021. Available at: https://www.educative.io/blog/one-hot-encoding [Accessed 14 August 2023].

Franklin, S. (2019). Effect of outliers on K-Means algorithm using Python. [Online]. Medium. Last Updated: 16 November 2019. Available at: https://medium.com/analytics-vidhya/effect-of-outliers-on-k-means-algorithm-using-python-7ba85821ea2 [Accessed 3 August 2023].

Gallup. (2023). What Is Employee Engagement and How Do You Improve It?. [Online]. Gallup Workplace. Available at: https://www.gallup.com/workplace/285674/improve-employee-engagement-workplace.aspx [Accessed 27 June 2023].

Gomes, M. Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. Information Systems and e-Business Management. x(x), p.x. [Online]. Available at: 10.1007/s10257-023-00640-4 [Accessed 31 July 2023].

Government Data Quality Hub. (2021). Meet the data quality dimensions. [Online]. Gov.uk. Last Updated: 24 June 2021. Available at: https://www.gov.uk/government/news/meet-the-data-quality-dimensions [Accessed 2 July 2023].

Hidayat, S. Matsouka, M. Baja, S. Rampisela, A. (2018). Object-Based Image Analysis for Sago Palm Classification: The Most Important Features from High-Resolution Satellite Ima. Remote Sensing. 10(8), p.1319. [Online]. Available at: 10.3390/rs10081319 [Accessed 31 July 2023].

Hiziroglu, A. (2013). Soft computing applications in customer segmentation: State-of-art review and critique. Expert Systems with Applications. 40(16), pp.6491 - 6507. [Online]. Available at: https://doi.org/10.1016/j.eswa.2013.05.052 [Accessed 24 July 2023].

Jaadi, Z. (2023). When and Why to Standardize Your Data. [Online]. Built In. Last Updated: 4 August 2023. Available at: https://builtin.com/data-science/when-and-why-standardize-your-data [Accessed 10 August 2023].

Kansal, T. Bahuguna, S. Singh, V. Choudhury, T. (2019). Customer Segmentation using K-means Clustering. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). x(x), pp.135-139. [Online]. Available at: 10.1109/CTEMS.2018.8769171 [Accessed 31 July 2023].

Kouhalvandi, L. Ceylan, O. Ozoguz, S. (2020). Automated Deep Neural Learning-Based Optimization for High Performance High Power Amplifier Designs. IEEE Transactions on Circuits and Systems I: Regular Papers. 67(12), pp.4420-4433. [Online]. Available at: 10.1109/TCSI.2020.3008947 [Accessed 31 July 2023].

Mar, B. (2017). Supervised V Unsupervised Machine Learning -- What's The Difference?. [Online]. Forbes. Last Updated: 16 March 2017. Available at: https://www.forbes.com/sites/bernardmarr/2017/03/16/supervised-v-unsupervised-machine-learning-whats [Accessed 8 August 2023].

McGoldrick, P. Thompson, M. (1992). The role of image in the attraction of the out-of-towncentre. International Review of Retail, Distribution and Consumer Research. 2(1), pp.81-98.

Nairn, A. Berthon,. (2003). Creating the Customer: The Influence of Advertising on Consumer Market Segments – Evidence and Ethics. Journal of Business Ethics. 42, pp.83-100. [Online]. Available at: https://doi.org/10.1023/A:1021620825950 [Accessed 24 July 2023].

Pakhira, M. (2009). A Modified k-means Algorithm to Avoid Empty Clusters. [Online]. ResearchGate. Last Updated: January 2009. Available at: https://www.researchgate.net/publication/228414762\_A\_Modified\_k-means\_Algorithm\_to\_Avoid\_Empty\_Clust [Accessed 3 August 2023].

Patil, B.M., Joshi, R.C., Toshniwal, D. (2010). Missing Value Imputation Based on K-Mean Clustering with Weighted Distance. In: Ranka, S., et al. Contemporary Computing. IC3 2010. Communications in Computer and Information Science, vol 94. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14834-7\_56 [Accessed 3 August 2023].

Pradana, M. Ha, H. (2021). Maximizing Strategy Improvement in Mall Customer Segmentation using K-means Clustering. Journal of Applied Data Sciences. 2(1), pp.19-25.

Przybyla, M. (2020). Why Does Everyone Use Kaggle?. [Online]. Towards Data Science. Last Updated: 9 July 2020. Available at: https://towardsdatascience.com/why-does-everyone-use-kaggle-db1bdf1f1b1a [Accessed 2 July 2023].

Roy, B. (2019). All about Categorical Variable Encoding. [Online]. Towards Data Science. Last Updated: 16 July 2019. Available at: https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02 [Accessed 10 August 2023].

Selvaraj, N. (2021). Customer segmentation with Python. [Online]. Medium. Last Updated: 7 June 2021. Available at: https://towardsdatascience.com/customer-segmentation-with-python-31dca5d5bdad [Accessed 24 July 2023].

Smith, W. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing. 21(1), pp.3-8. [Online]. Available at: https://doi.org/10.2307/1247695 [Accessed 24 July 2023].

Tran, K. (2020). 5 Reasons why you should Switch from Jupyter Notebook to Scripts. [Online]. Towards Data Science. Last Updated: 24 August 2020. Available at: https://towardsdatascience.com/5-reasons-why-you-should-switch-from-jupyter-notebook-to-scripts-cb35 [Accessed 21 August 2023].

Warsame, M. (2023). Practical Data Quality Auditing: A Comprehensive Guide. [Online]. Towards Data Science. Last Updated: 2023. Available at: https://towardsdatascience.com/data-quality-auditing-a-comprehensive-guide-66b7bfe2aa1a#:~:text=Perf [Accessed 1 August 2023].

Yung, Z. (2022). How to Transform Data in Excel? Ultimate Guide. [Online]. Coupler.io. Last Updated: 11 March 2022. Available at: https://blog.coupler.io/transform-data-in-excel/#:~:text=third%2Dparty%20tools.-,Excel%20functions,d [Accessed 1 August 2023].

Screenshots

A screenshot of a table

Description automatically generated

Missing values

A screenshot of a computer

Description automatically generated

Standardisation