Understanding Saving Habit of Individuals for the Varied Financial Measurement Categories using Cluster Analysis

MSc Research Project
FinTech

Rachita Patel
Student ID: x18121331

School of Computing
National College of Ireland

Supervisor: Prof. Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rachita Patel
Student ID:	x18121331
Programme:	FinTech
Year:	2019
Module:	MSc Research Project
Supervisor:	Prof. Noel Cosgrave
Submission Due Date:	12/08/2019
Project Title:	Understanding Saving Habit of Individuals for the Varied Financial Measurement Categories using Cluster Analysis
Word Count:	7649
Page Count:	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Understanding Saving Habit of Individuals for the Varied Financial Measurement Categories using Cluster Analysis

Rachita Patel
x18121331

Abstract

Saving being a rational behaviour, it was observed that many people cannot deal with their finances which affect their long-term financial needs. Also, the economic growth of a nation depends on the Gross Domestic Saving (GDS) which is the saving from the private, public, and household sector. Thereby, this research aims at understanding the saving behaviour or saving habit of the individuals. For this purpose, the financial well-being survey data from the Consumer Financial Protection Bureau (CFPB) is being used. The saving behaviour is observed for the six categories of financial measurement. Initially, the attributes falling into these categories are identified using Multiple Linear Regression (MLR) and Random Forest (RF). Having known, the significant factors, clustering which is a type of unsupervised machine learning is being used to discover the factors influencing saving habit of individuals. Five types of clustering algorithm- K-means, Partitioning around Medoids (PAM), Clustering Large Applications (CLARA), Hierarchical, and Fuzzy (Fanny) are evaluated using internal validation metrics to find the optimal number of clusters, thereby, interpreting those results to analyse the saving behaviour. The internal validation metrics like connectivity, the Dunn's index, and the Silhouette index has shown significant results for all the categories. Finally, the interpreted results reveal various factors leading to a good saving habit and they are education, income; also, the unmarried individuals, who owns a house, had no bad financial experience, and the ones who are goal-oriented have a good saving habit.

1 Introduction

What is financial saving? It is not just about spending less than the income earned but keeping the money for the future. Saving is different than investing because here in saving, the individual has access to the money at any time. This provides safety in case of any emergency. Although, investing is associated with the risk of losing money but there are chances to earn money.

Saving is basically done for any emergency or bad times, for retirement or anything else. Also, individuals or households are dependent on savings for achieving their financial goals and to maintain their financial well-being (Donnelly et al.; 2012). Thus, it is considered as a rational behaviour. Also, during the various financial crises, it has been noted that many people cannot deal with their finances, so it is important to create awareness amongst individuals' about their long-term financial needs and resources (Wärneryd; 1999).

Also, saving is important from the nation's standpoint because it relates economic growth of a country. The factors affecting the savings' behaviour in any country includes income, growth rate, foreign savings rate, dependency rate, and financial sector development (Agrawal et al.; 2009). Thus, after analysing the saving behaviour, several ways can be suggested to improvise the savings' rate and create sound policies for the country.

Asebedo et al. (2018) observed problems in rational decision making by the consumers due to the integration of psychological concepts within saving behaviour and the rise of behavioural finance. Having understood the importance of saving and the need of rational behaviour behind it, it is time to interpret the saving habit of the people and the attributes which influences such behaviour. This will further help in designing financial remedy or investment solution for the varied type of saving habit behaviour.

With this motivation, the research objectives are being addressed using the financial well-being survey data from the Consumer Financial Protection Bureau (CFPB)[1]. [2] According to the CFPB financial well-being survey report, there are six categories of financial measurement: (a) Individual characteristics, (b) Household and family characteristics, (c) Income and employment characteristics, (d) Savings and safety nets, (e) Financial experiences, and (f) Financial behaviours, skills, and attitudes. So, these six categories will be used for analysing the saving habit of individuals'. The CFPB financial well-being survey data has been used to derive a financial well-being score for the people of United States (U.S.) but in this research, it is used for analysing the saving habit of the individuals' to derive insights out of it.

The research question and the objectives are as follows:
How clustering can help in discovering group of factors that influence the saving behaviour of the people in various financial measurement categories?

- Initially, using Multiple Linear Regression (MLR) and Random Forest (RF) to select attributes relevant to the six financial measurement categories.

- Then performing cluster analysis for those categories to find out which type of factors are responsible for a specific saving habit.

Unsupervised machine learning is preferred here, as this research focuses on understanding saving habit or saving behaviour from various factors. Unsupervised learning helps in deriving hidden patterns from the data or assist in grouping the related elements of the data. With this objective, clustering which is a type of unsupervised learning is chosen for the analysis. Five types of clustering algorithms are evaluated for determining suitable clusters using internal validation metrics like *Connectivity, Dunn index, and Silhouette index.* This research is limited to understanding saving behaviour of the people related to the abovementioned six financial measurement categories.

Five types of clustering algorithms used are:

(a) Partitioning methods: k-means, Partitioning around Medoids (PAM), and Clustering Large Applications (CLARA)

(b) Hierarchical clustering

(c) Fuzzy clustering: Fanny

[1] `https://www.consumerfinance.gov/data-research/financial-well-being-survey-data`

[2] `https://files.consumerfinance.gov/f/documents/201705_cfpb_financial-well-being-scale-technical-r`
`pdf`

To address the research objective, the paper is subdivided into various sections. Various researches conducted in this area are critically evaluated in section 2. That is been followed by the methodology adopted to conduct this research in section 3. On understanding the previous work and with the research methodology, the current implementation is described in section 4 accompanied by evaluation and discussion of the results in section 5. Finally, the conclusion for this research is presented in section 6, leading to further understanding of saving behaviour using different characteristics and providing financial remedy thereafter in the future.

2 Related Work

In this section, various researches related to saving behaviour are being presented and critically analysed.

2.1 Need for understanding saving behaviour

Burton (2001) raised a fact that saving and investment behaviour is being under-researched and needs attention because it involves complex financial decision-making which makes it difficult for consumers to understand. Long after that, a study by Hanna et al. (2016) show that many people do not save for their retirement. Thus, Lee and Hanna (2015) highlight the importance of identifying attributes which influences the saving behaviour.

2.2 Importance of understanding saving behaviour

Zeller and Sharma (2000) say that savings act as a critical tool during times of crisis or shocks and helps in improving the financial well-being of an individual, whereas Attanasio and Szekely (2000) regard savings as an important factor for the economic growth of any nation and a crucial source for investments.

2.3 Savings behaviour in the world

There has been variation in the saving behaviour in different parts of the world. For the development of any country, financial stability is the most crucial factor. Whereas the savings from the public, private corporate and the household sector collectively termed as Gross Domestic Saving (GDS) is responsible for strengthening the country's financial stability. Palakvangsa-Na-Ayudhya et al. (2017) pointed out that the financial stability of any country is influenced by the long-term saving behaviour at any ages. Guiso et al. (2006) showed that country-specific characteristics and culture does affect the financial behaviour of an individual.

Data from different countries has been used by many researchers to test the relationship between people's origin and saving rates. For instance, Gatina (2014) has reported that the saving decision of the people who migrate in Australia is influenced by the country's characteristics from where they come from.

On the other hand, Thanoon and Baharumshah (2012) reveal that factors like economic growth, interest rates, dependency ratio, foreign capital inflows, and the export sector are responsible for the saving behaviour or saving ratio within any country.

Agrawal et al. (2009) agree partially with Thanoon and Baharumshah (2012) with regards to foreign saving rate and dependency rate for the savings in South Asia. In addition to these, Agrawal et al. (2009) highlight that savings in South Asia are also impacted by other factors like income and access to banking institutions. Also, income is proved to be an important factor for savings by Munozmoreno et al. (2014) in Mauritius. The results shown by Suppakitjarak and Krishnamra (2015) say that family and bank staff has much influence on the saving decision in Thailand.

An altogether different scenario was observed by Asare et al. (2018) in Ethiopia, Africa where illiterate members of households having learned from their experiences are more likely to save. But the study by Gaisina and Kaidarova (2017) shows financial literacy as an important factor for the increase in saving in Kazakhstan.

2.4 Characteristics that affect saving behaviour

There are various determinants like psychological characteristics, socio-demographic characteristics, household factors which affect an individual's saving behaviour. In the above section, the country-specific saving rate was determined whereas here factors influencing such behaviour is analysed.

How age or generation affects savings and the evidence of a decrease in savings with age is demonstrated by Brounen et al. (2016) whereas Heng-fu (1995) contradicts these findings. Relationship between saving behaviour and gender was investigated by Fisher et al. (2015) and their research demonstrates that there is a significant difference in financial behaviour based on gender.

Additionally, Gerhard et al. (2018) found relationship between saving habit and psychological characteristics for the two classes i.e. striving versus established households. Also, Heng-fu (1995) sheds lights on the family characteristics that there is no observable difference in saving behaviour of households having children or not.

Alike Agrawal et al. (2009), Binswanger (2010) verifies the fact that saving rates differ across various income groups and increases with income. While the impact of saving goals on saving behaviour using Maslow's theory is shown by Hanna et al. (2016). Cronqvist and Siegel (2011) and Sabri and MacDonald (2010) found that individual experiences play a significant role in providing learning towards savings.

Also, behavioural control or attitude towards savings affects saving behaviour (Minibas-Poussard et al.; 2018) whereas saving habit and self-control is indirectly related to saving behaviour and protects against impulsive spending as indicated by Allom et al. (2018), Kim (2017) and Strömbäck et al. (2017).

Apart from these characteristics, there are big five personality traits (agreeableness, conscientiousness, intellect, extraversion, and emotional stability) which indirectly support in understanding the financial behaviour of the individuals'.

2.5 Data mining approaches to predict saving behaviour

Gerhard et al. (2018) analysed the relationship between psychological characteristics and saving behaviour using a finite mixture model to estimate the class-specific regression coefficients whereas the household savings were predicted from the savings goals and other socio-demographics features using a logistic regression model by Lee and Hanna (2015). Also, to analyse the saving behaviour for men and women, Fisher et al. (2015) derives a logistic regression model and investigates the gender difference. Ordinal least squares (OLS) regression is used to find out important variables required for predicting saving behaviour (Thanoon and Baharumshah (2012); Brounen et al. (2016); Munozmoreno et al. (2014)). Again OLS regression is used by Strömbäck et al. (2017) to find the association between financial behaviour and self-control. Balasubramanian (2017) uses a decision tree approach to predict whether an individual saves regularly or not.

As it is seen from the researches that there are various factors which influence the saving habit of individuals'. Also, some research did contribute to analyse behaviour using a data mining approach. But to discover a group of determinants which influence this behaviour, this research aims at using cluster analysis to derive insights out of it.

3 Methodology

The Cross-industry standard process for data mining, also known as *CRISP-DM* provides a step-by-step guide (structured approach) for a data mining project (Chapman et al.; 2000). So, the *CRISP-DM* reference model is used for this research to discover useful knowledge from the data.

'R' is a statistical and data analytics platform, and therefore, preferred over other platforms. This research analysis is carried out based on the following steps using *'R platform'*.

A. Business understanding

The primary business objective for any institution is to keep their current customers. Thus, this research will help financial institutions in deriving financial remedy i.e. investment solutions for the varied type of saving behaviour and will connect them with their customers. Understanding the saving habit of the people based on varied financial measurement characteristics using data mining algorithms will contribute towards achieving the business objective.

B. Data understanding

Data collection and its description: The main problems encountered during collecting data from a public source for this research was most of the data were related to either transaction statements or financial saving data of banks or government organisations. That is why a financial well-being survey data from CFPB has been used for this research [3] which is a *.csv* file. The ultimate goal of CFPB was to find the financial well-being score for the U.S. adults using this data. In addition to that, it collected details related to an individual's cognitive psychology and financial planning attitude i.e. six categories of financial measurements. Thus, the survey data related to these characteristics were useful for this research. The CPFB financial well-being survey data was collected from the younger consumers (aged 18-61 years) and older consumers (aged 62 and older) from the 50 U.S. States and Washington DC. The data was published in the year 2017 having 217 variables and 6394 records.

Exploration of data: CFPB has already encoded the categorical description of the attributes to numeric labels or values. So, all the records are numeric for the collected data and there are no missing values so there is no need for any type of transformation.

Verifying Data Quality: The data quality report is thus examined to check for errors, missing values, and completeness of data. This step has verified the data exploration process.

[3] https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

C. Data preparation

This dataset comprises of survey related to varied category accounting to 217 attributes. So, depending on different financial measurement categories, the data is selected and processed. Attributes are selected taking into consideration the following criteria: (a) correlation, (b) parametric test, (c) normalisation test using histograms, (d) absence of outliers, and (e) supervised machine learning algorithms like multiple linear regressions as well as random forest. The final aim of this step is to cluster individuals based on their savings' habit. So, the response variable is 'saving habit'. The MLR is used here as it models the relationship between dependent and independent variables thereby helps in choosing attributes whereas RF helps in verifying the results given by MLR because it reduces over-fitting thereby leading to accurate results. Here, the dataset had attributes falling into six categories of financial measurement. Table 1 represents various categories and selected attributes in those categories.

So, there will be six use cases for the six respective categories and the saving habit will be predicted accordingly.

D. Modelling

Since the data is not labelled, so this research focuses on unsupervised machine learning algorithm i.e. clustering. And the business use here is to identify unknown groups in the dataset. Clustering will automatically divide the data into clusters or groups of similar items. Five types of clustering algorithm are used for this research. The numbers of cluster i.e. *'k'* for each of these methods has to be chosen appropriately. Therefore, the algorithm needs to be evaluated for a range of values of *'k'* and then results will be compared based on the validation index.

Five different types of clustering algorithms being evaluated are:

1. Partitioning methods: k-means, PAM, CLARA

   a) *K-means*
   K-means is one of the simplest and well-known clustering method which is based on the idea of 'centroids' (Zhu et al.; 2018) and thus chosen for this research. K-means algorithm follows an iterative process where a given set of data are partitioned into *'k'* clusters using the distance from each data point to *'k'* different centroids (Ganganath et al.; 2014). But, it is very sensitive to outliers and noisy data. K-means works computationally fast and produces tight clusters because it assumes that the variance of each attribute is same and spherical. That is why, each cluster has an almost equal number of observations.

   b) *PAM or k-medoids*
   PAM is computationally harder than k-means because it computes medoids and not centroid (Arbin et al.; 2015). Here, it selects data point in each cluster as a center or medoids. Alike k-means, PAM also partitions data into *'k'* groups intending to minimise the point to nearest center distance. PAM is computationally expensive but provides robustness and accuracy (Olukanmi et al.; 2019). Thus, it is not suitable for large data (Olukanmi et al.; 2019). This does not get much affected by the outliers.

Table 1: Six categories of financial measurement

Sr. No.	***Category***	***Selected attributes***
1	Individual characteristics	Education Generation Gender
2	Household and family characteristics	Housing status House satisfaction Marital Status Financially supporting children
3	Income and employment characteristics	Employment status Household income Household income volatility Military status
4	Savings and safety nets	Money in savings account Non-retirement investments Health insurance Friends or family will lend money with expectation of repayment or no expectation of repayment
5	Financial experiences	Have savings account Housing cost burden Experienced any negative financial shocks Have a student loan
6	Financial behaviours, skills, and attitudes	Recent financial goal Steps to achieve financial goal Confidence in own ability to achieve financial goals

c) *CLARA*
As the name suggests, it can deal effectively with large datasets with less computing time. It is an extension of PAM evaluating each medoid set to achieve the optimal set of centroids for the sample. Thereby, it relies on the sampling approach. So, in this algorithm, multiple samples of the dataset are taken, then medoids are found by applying PAM to each of these samples, and finally, it returns the best clustering.

2. *Hierarchical clustering*
Hierarchical clustering aims at building a hierarchy of clusters. Here, the agglomerative or divisive algorithm is used to find clusters. In agglomerative

approach, each element is present in a single cluster which is further combined into larger clusters. Contrary to that, the divisive approach starts with a large cluster having all the elements and then divides it into small subclusters (Łuczak; 2016). This can be visualised using a dendrogram which is a tree-like diagram that has a sequence of splits. This needs more space and is time-consuming so it cannot be used with huge datasets.

3. *Fuzzy clustering: Fanny*

   The goal of fuzzy clustering is to determine the dependence degree among the selected attributes (Hasanpour et al.; 2018). It is an extended version of k-means but not similar to k-means, here samples are divided into *'C'* clusters. Fuzzy does not focus on providing a boundary between the clusters; instead, here the clusters are allowed to interfere or overlap. Thus, it is also considered as a soft clustering. Another property of fuzzy clustering is that it provides a degree to which an element belongs to a cluster which is a value between 0 and 1. The points close to the cluster center are assigned a higher degree than those which are at the edge of a cluster. Thus, a single feature can belong to one or more clusters. And the centroid of a cluster is calculated by taking the mean of all these points and degree.

E. Evaluation

These clustering methods are tested for different numbers of clusters i.e. *'k'* using internal validation results, the optimal value of *'k'* is selected and then the saving habit is analysed according to that. The internal validation metrics for clustering used for this research includes connectivity, Dunn index, and Silhouette index as extracted from Brock et al. (2011). This validation metrics uses intrinsic information of the data to determine the quality of clustering by taking the actual data and the partitioned cluster as an input (Brock et al.; 2011).

(i) *Connectivity*

Connectivity checks whether the nearest and farthest neighbors are associated with the same cluster or not. It has a value between zero and infinity and it should always be minimised.

$$Conn(C) = \sum_{a=1}^{R} \sum_{b=1}^{S} x_{i,nn_{i(j)}} \tag{1}$$

where, R = total number of rows, S = number of nearest neighbors to use, and C = clustering partition

(ii) *Dunn's Validity Index*

The Dunn Validity Index finds out the degree of compactness and separateness of the cluster sets. It has a range of value from zero to infinity. This value should be maximised to get a dense and distinct clusters. It can also be termed as the ratio of the smallest distance between observations of different cluster to the largest intra-cluster distance (Brock et al.; 2011).

(iii) *Silhouette Validity Index*

"This index measures the silhouette width for each data point, average silhouette width for each cluster and the overall average silhouette width for the total

dataset." as given Ansari et al. (2011). The average of the Silhouette value of each observation is termed as Silhouette width. The values of this index go from +1 to -1. A well-clustered observation has value of +1 whereas poor ones are at -1. Thus, the number of clusters with the values close to +1 are considered as the optimal numbers of the clusters. For the ith observation, the Silhouette validity index is given as follows:

$$S(i) = \frac{b_i - a_i}{max(b_i, a_i)} \tag{2}$$

where, a = average distance between all the other observations in the same cluster and *i*, and b = average distance between all the other observations in the nearest neighboring cluster and *i*.

Apart from these, the quality of a clustering can be measured by taking into account the following:

- Maximising inter-cluster distance
- Minimising intra-cluster distance

Using these evaluation metrics, optimal numbers of clusters will be taken. Then the attribute values for the respective clusters leading to a particular savings' habit will be analysed. Depending on the results for the six case studies, the saving habit of the people will be discussed.

Apart from these, the metrics used to measure the performance of MLR and RF are as follows (Shcherbakov et al.; 2013). All these are errors so they should tend towards lower values. Let n be the total number of observations, x be the actual value and y be the predicted value.

- *Mean Error (ME):* For all the errors in a set, the calculated average is the ME. It is one of the most basic approaches for finding accuracy.
- *Mean Absolute Error (MAE):* Unlike ME, MAE calculates the average magnitude of errors. It does not take into account the direction of errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{3}$$

- *Mean Square Error (MSE):* The MSE is just like MAE, but it squares the differences before summing them.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{4}$$

- *Mean Absolute Percentage Error (MAPE):* The MAPE measures accuracy in terms of percentage. It works well if there are no outliers in the data.

$$MAPE = \frac{1}{n}\frac{\sum_{i=1}^{n}|x_i - y_i|}{x_i} \tag{5}$$

- *Root Mean Square Error (RMSE):* It is the square root of MSE. It measures the spreadness of the residuals.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n}} \tag{6}$$

4 Implementation

The implementation of the proposed solution for this research is as shown in Fig. 1. The following steps are implemented using 'R platform'.

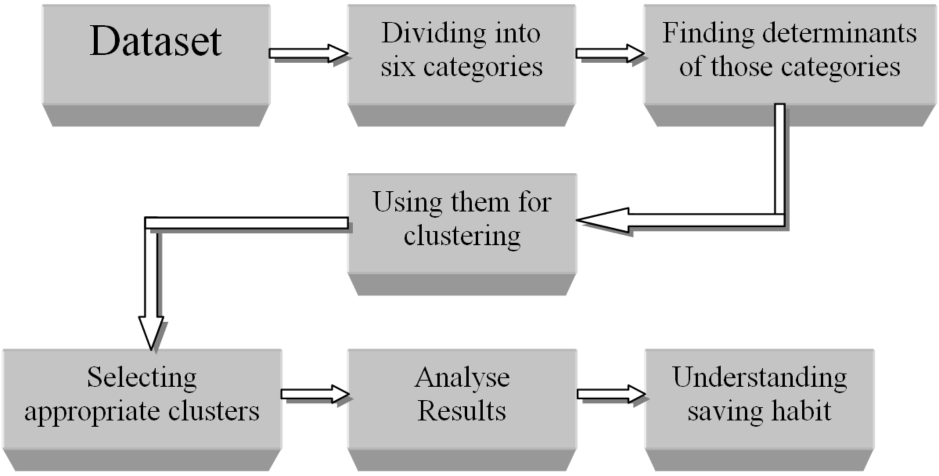
Dataset
Dividing into six categories
Finding determinants of those categories
Using them for clustering
Selecting appropriate clusters
Analyse Results
Understanding saving habit

Figure 1: Implementation flow diagram

Step 1: Dataset

As discussed in the previous section, the CFPB financial well-being survey data will be used for this research.

Step 2: Dividing into six categories

The saving habit of individuals depends on various characteristics and the dataset was chosen here comprises of attributes for all the categories i.e. 217 variables. So, the problems occurred when dealing with 217 attributes are: (a) clusters are not compact, (b) the number of clusters increases leading to misleading analysis, and (c) large processing time.

Due to these reasons, the six financial measurement categories are considered as reported in the CFPB financial survey report. Thus, the entire data is divided into these categories and then evaluated. The outcomes of such decision are: (a) better clusters, (b) saving habit for respective categories can be thoroughly analysed and understood, and (c) comparatively less processing time.

Step 3: Finding attributes influencing those categories

Now, each category has attributes relevant to that. But not all of them influences the saving habit. It is to be noted that this is a survey data, so it contains some variables which will not be relevant for this study. Thus, the influencing attributes are selected based on their: (a) correlation values, (b) p-values, (c) normalisation status, (d) outliers detection, and (e) using MLR and RF algorithms.

The MLR and RF algorithms are evaluated using performance metrics like ME, MAE, MSE, MAPE, and RMSE.

Step 4: Using the selected attributes for clustering

Those attributes which showed some relevance depending on the selection criteria and based on prior research were selected for cluster analysis. So, Table 1 enlists all the selected attributes.

Step 5: Selection of appropriate clusters

Clustering algorithms based on their simplicity and relevance was selected for this research. Algorithms like k-means, PAM, CLARA, hierarchical clustering, and fanny clustering were chosen for the analysis. The important step in clustering is to determine the optimal number of clusters and this cannot be selected randomly. So, the implementation for a varied number of clusters was conducted and then they were evaluated using internal validation metrics to check for their performance. The internal validation metrics include connectivity, the Dunn's index, and the Silhouette index. Based on that, the numbers of clusters relevant for a particular category were selected.

Step 6: Analysing the results from the above step

Having known the number of clusters for a case, their results were thoroughly analysed to understand the type of factors responsible for a particular saving habit.

Step 7: Understanding the saving habit

Lastly, the saving habit and their influencers were interpreted for the respective financial measurement categories thereby understanding the saving behaviour.

5 Evaluation

As discussed in the methodology section, here the saving behaviour will be analysed for each of the data category, so considering only the six case studies.

The Table 2 shows the performance evaluation measures for multiple linear regression and random forest. This is been done to select appropriate attributes for the cluster analysis.

Table 2: Validation metrics for selecting appropriate attributes using MLR and RF

Case Studies	ME		MAE		MSE		MAPE		RMSE	
	MLR	RF	MLR	RF	MLR	RF	MLR	RF	MLR	RF
1	0.15	0.05	1.16	1.18	2.11	2.04	0.4	0.42	1.45	1.43
2	0.05	0.04	1.12	1.16	2.04	1.97	0.4	0.42	1.43	1.4
3	0.008	0.08	1.08	1.14	1.98	1.89	0.39	0.39	1.41	1.37
4	0.04	0.004	1.09	1.07	1.97	1.76	0.41	0.4	1.4	1.33
5	-0.3	0.02	1.12	1.2	2.19	2.08	0.45	0.45	1.48	1.44
6	0.03	0.02	0.93	0.99	1.65	1.56	0.33	0.35	1.28	1.25

From Table 2, it has been observed that both MLR and RF show significant results so the attributes selected for the respective case studies are appropriate. On the other hand, Table 3 demonstrates the internal validation metrics for the five clustering algorithms having $k = 2, 4, 9$. Here Con. = Connectivity; Dunn = Dunn's Index; Sil. = Silhouette Index

Table 3: Internal Validation metrics for the five clustering algorithms

Case Study	Cluster	K-means			PAM			CLARA			Hierarchical			Fanny		
		Con.	Dunn	Sil.	Con.	Dunn	Sil.	Con.	Dunn	Sil.	Con.	Dunn	Sil.	Con.	Dunn	Sil.
1	2	16.25	0.18	0.32	20.23	0.15	0.27	25.67	0.15	0.29	24.43	0.17	0.32	15.4	0.15	0.28
	4	29.54	0.24	0.27	25.31	0.18	0.26	33.49	0.18	0.22	52.36	0.21	0.23	NA	NA	NA
	9	42.15	0.32	0.26	51.87	0.21	0.22	59.33	0.24	0.19	73.99	0.24	0.19	NA	NA	NA
2	2	99.56	0.09	0.31	89.43	0.10	0.24	41.55	0.10	0.29	43.32	0.10	0.24	40.5	NA	NA
	4	171.43	0.10	0.31	91.69	0.10	0.28	126.2	0.09	0.26	87.40	0.11	0.26	NA	NA	NA
	9	195.78	0.12	0.33	169.0	0.11	0.31	187.2	0.11	0.27	109.5	0.11	0.27	NA	NA	NA
3	2	148.5	0.09	0.38	233.7	0.08	0.37	196.8	0.08	0.38	144.4	0.09	0.38	221	0.08	0.35
	4	253.79	0.13	0.39	250.9	0.12	0.39	429.2	0.10	0.28	342.6	0.10	0.38	305	0.09	0.01
	9	495.58	0.12	0.31	513.1	0.14	0.31	651.0	0.13	0.22	456.2	0.13	0.26	NA	NA	NA
4	2	0.00	11.3	0.97	0.00	11.3	0.97	0.00	11.3	0.97	0.00	11.3	0.97	0.00	11.3	0.97
	4	99.44	0.17	0.40	131.2	0.15	0.39	113.6	0.15	0.37	136.5	0.14	0.40	NA	NA	NA
	9	184.70	0.24	0.27	181.4	0.22	0.22	292.5	0.18	0.17	355.0	0.20	0.23	NA	NA	NA
5	2	0.00	11.4	0.97	0.00	11.4	0.97	0.00	11.4	0.97	0.00	11.4	0.97	0.00	11.4	0.97
	4	17.68	0.16	0.47	35.84	0.16	0.38	21.76	0.16	0.41	60.33	0.15	0.46	NA	NA	NA
	9	56.89	0.30	0.37	38.87	0.25	0.35	105.9	0.19	0.40	149.4	0.22	0.33	NA	NA	NA
6	2	10.95	0.17	0.42	15.31	0.15	0.40	20.95	0.14	0.31	20.93	0.20	0.42	NA	NA	NA
	4	39.32	0.19	0.32	24.22	0.17	0.26	45.41	0.18	0.24	49.12	0.18	0.33	NA	NA	NA
	9	83.07	0.26	0.30	55.32	0.20	0.33	58.55	0.19	0.31	118.7	0.22	0.17	NA	NA	NA

For the five clustering algorithm, it is observed that as *k* increases, the connectivity between the cluster is maximised. Also, the Dunn's index increases with the increase in the value of *k* whereas the Silhouette's index decreases with the value of *k*.

The attributes for the six scenarios are used to carry forward the analysis, records with *'refused'* values in any of their attributes were removed from the dataset. The analysis of the results obtained is presented in the case studies below.

5.1 Case Study 1: Individual Characteristics

Based on connectivity, and the Dunn index, the number of clusters selected for this case were *(k=2)*. Also, nine clusters gave competitive results in some cases (like k-means). K-means, hierarchical, and fanny gave almost similar groups of factors with slight variations in their internal validation metrics. Thus, Table 4 enlists the clustering results by analysing the respective savings behaviour.

Table 4: Clustering of individual characteristic

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Males belonging to boomer generation and having some college degree or an associate's degree	Agree
2	Females of Gen X and having high school degree	Slightly disagree
3	Males or females having just high school degree and belonging to either boomer, gen X or millennial generation	Disagree

This shows that as the level of education is decreased, the saving habit also reduces which implies that education plays an important role contributing towards financial literacy whereas gender does not play any major role.

5.2 Case Study 2: Household and Family Characteristics

The results for this case were poor as compared to other case studies. So, depending on that, the value of *k=2* was selected. Alike, case study 1, k-means with nine clusters did give good results. Here it was observed that, k-means and hierarchical gave almost similar clusters. So, the analytical results are described in Table 5.

It is observed that unmarried individuals and those who own their house has more habit of saving compared to those who are married, widowed, or separated and live on rent. This is because people owning their houses are satisfied with their place and save on paying rents. Also, married individuals supporting children financially does not normally have a good saving habit.

Table 5: Clustering of Household and Family characteristic

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Individuals' who rent and have never married, have no children to support financially thereby they are somewhat satisfied with the place they currently live in	Agree slightly
2	Married individuals having their own house and no children to support financially are very satisfied with their current livelihood	Strongly agree
3	Individuals who are divorced or married, may have children to support financially or may not; whereas owning their own home are somewhat satisfied	Disagree
4	Widowed individuals with no children to support while living on rent are not very satisfied	Disagree slightly

5.3 Case Study 3: Income and Employment Characteristics

For this category, the internal validity metrics showed that the clusters are poor. So, out of them, clusters with *k=2 and k=4* had been considered. In Table 6, the results are analysed.

Table 6: Clustering of income and employment characteristic

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Homemakers earning \$50k to \$60k every month	Agree slightly
2	Spouse/dependent veteran who are permanently sick, disable, or unable to unable to work and earning between \$60k to \$75k every month	Agree
3	Retired independent individuals earning \$75k to \$100k every month	Agree
4	Independent unemployed or temporary laid off individuals earning \$20k to \$30k every money	Disagree

The analysis here reveals that independent unemployed with fewer earnings has less habit of saving. On the other hand, retired individuals or dependent veterans with more earnings have increased chances of saving.

5.4 Case Study 4: Savings and Safety nets

The attributes selected for savings and safety nets returned same group of factors for all the five algorithms when *k=2*. Apart from that, the understanding of these results is described in Table 7.

Table 7: Clustering of savings and safety characteristic

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Individuals having health insurance with or without non-retirement investments and having \$20k-\$75k in savings today.	Agree

The results in Table 7 describes the scenario where individuals have money in their savings account in addition to insurance. Although, non-retirement investments are necessary for the future safety so the individuals who does not have that should plan for investing their saving somewhere. Also, these individuals have friends or family who can lend them money during needs but they need to repay them.

5.5 Case Study 5: Financial Experiences

The selected attributed for this category resulted proper clusters (especially for *k=2*). Here, the internal validity metrics has shown significant results which is not observed for any other category. So, Table 8 describes the factors contributing towards saving habit.

Table 8: Clustering of financial experiences

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Individuals with or without savings account, with no experience to negative financial shocks, and no student loan	Agree slightly

The evaluation for financial experiences shows that savings account is necessary for any individual to get more benefits. But there are individuals with no experience to negative financial shocks and no debt on them but still have a habit of saving.

Table 9: Clustering of financial behaviour, skills, and attitude

Sr. No.	***Clustering Analysis***	***Putting money into savings is a habit for them***
1	Individuals with no recent financial goal	Disagree slightly
2	Individuals having current financial goal with a clear plan of action and are confident to achieve that	Agree

5.6 Case Study 6: Financial Behaviours, Skills, and Attitude

In this case, all the clustering algorithms gave almost similar clusters with slight variations in their internal validation metrics. Thus, Table 9 enlists the clustering results by analysing the respective savings behaviour.

The inference drawn for this category was that financial goal oriented individuals have a clear plan of action and are therefore focused on their saving habit.

6 Conclusion and Future Work

The designed methodology helps in understanding the saving behaviour of individuals for the six categories of financial measurement i.e., (1) individual characteristics, (2) household and family characteristics, (3) income and employment characteristics, (4) savings and safety nets, (5) financial experiences, and (6) financial behaviours, skills, and attitudes. For successful cluster analysis, the MLR and RF algorithms have contributed in selecting appropriate attributes. The performance metrics used for evaluating MLR and RF justifies that the attributes selected were significant thereby leading to good cluster analysis. Ultimately, unsupervised machine learning (i.e. clustering in this case) did play a significant role in discovering a hidden group of factors which are the determinants of such behaviour. The internal validity metrics were used for cluster evaluation and the observations from this were: (a) k-means outperforms in all the six categories, (b) case study 4 and case study 5 gave best results in comparison to the rest, (c) Hierarchical and Fanny clustering algorithm consumes more time, and (d) the NA's were observed in Fanny's performance. The NA's introduced by Fanny were due to the following reasons: (i) it cannot handle outliers efficiently, (ii) problem due to the high dimensionality of the data, and (iii) at times the cluster membership depends on other cluster centers membership value which leads to unaccepted results.

Depending on these results, inferences were drawn for the varied factors leading to a particular saving habit. It was observed that income and education do have a positive impact on saving behaviour as noted by many researchers. Apart from that, married individuals have less tendency to save than the unmarried ones. Similarly, individuals financially supporting children do not show good saving habit whereas retired individuals have the best savings. Also, house proprietors exhibit good saving behaviour. Furthermore, a good saving habit is found in the people who have not gone through any negative financial shock. One of the major observation is goal-oriented individuals have an increased tendency to save.

So, these varied types of behaviours or factors will further help in designing financial investment solutions for the people thereby contributing towards their financial wellness. Also, the research is limited to the six financial measurement categories, so this research can be further extended to understand the saving behaviour of the individuals depending on the other characteristics as well. A hybrid clustering can be used in the future to improve upon the results.

References

Agrawal, P., b, P. S. and Dasha, R. K. (2009). Savings behaviour in south asia, *Journal of Policy Modeling* **31**(2): 208–224.

Allom, V., Mullan, B., Monds, L., Orbell, S., Hamilton, K., Rebar, A. and Hagger, M. (2018). Reflective and impulsive processes underlying saving behavior and the additional roles of self-control and habit, *Journal of Neuroscience, Psychology, and Economics* **11**(3): 135–146.

Ansari, Z., Azeem, M., Ahmed, W. and Babu, A. (2011). Quantitative evaluation of performance and validity indices for clustering the web navigational sessions, **1**(5): 217–226.

Arbin, N., Suhaimi, N. S., Mokhtar, N. Z. and Othman, Z. (2015). Comparative analysis between k-means and k-medoids for statistical clustering, *2015 3rd International COnference on Artificial Intelligence, Modelling and Simulation* pp. 117–121.

Asare, E., Segarra, E., Gertrude, N. and Asiseh, F. (2018). Explaining the saving behavior of households' in ethiopia, africa, *Applied Economics and Finance* **5**(2): 143.

Asebedo, S. D., Wilmarth, M. J., Seay, M. C., Archuleta, K., Brase, G. L. and MacDonald, M. (2018). Personality and saving behavior among older adults, *Journal of Consumer Affairs* **53**(2): 488–519.

Attanasio, O. and Szekely, M. (2000). Household saving in developing countries - inequality, demographics and all that: how different are latin america and south east asia?, *Inter-American Development Bank* p. 2.

Balasubramanian, C. (2017). Predicting regular saving behaviour of the poor using decision trees- an importnt input to financial inclusion in india, *Scholedge International Journal of Management and Development* **2**(8).

Binswanger, J. (2010). Understanding the heterogeneity of savings and asset allocation: A behavioral-economics perspective, *Journal of Economic Behavior and Organization* **76**(2): 296–317.

Brock, G., Pihur, V., Datta, S. and Datta, S. (2011). clvalid , an r package for cluster validation.

Brounen, D., Koedijk, K. and Pownall, R. (2016). Household financial planning and savings behavior, *Journal of International Money and Finance* **69**: 95–107.

Burton, D. (2001). Savings and investment behaviour in britain: More questions than answers, *The Service Industries Journal* **21**(3): 130–146.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide, *SPSS* .

Cronqvist, H. and Siegel, S. (2011). The origins of savings behavior, *SSRN Electronic Journal* .

Donnelly, G., Iyer, R. and Howell, R. T. (2012). The big five personality traits, material values, and financial well-being of self-described money managers, *Journal of Economic Psychology* **33**(6): 1129–1142.

Fisher, P. J., Hayhoe, C. R. and Lown, J. M. (2015). Gender differences in saving behaviors among low- to moderate-income households, *Financial Services Review* .

Gaisina, S. and Kaidarova, L. (2017). Financial literacy of rural population as a determinant of saving behavior in kazakhstan, *Rural Sustainability Research* **38**(333): 32–42.

Ganganath, N., Cheng, C.-T. and Tse, C. K. (2014). Data clustering with cluster size constraints using a modified k-means algorithm, *2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* .

Gatina, L. (2014). The saving behaviour of immigrants and home-country characteristics: Evidence from australia, *Australian Economic Review* **47**(2): 157–172.

Gerhard, P., Gladstone, J. and Hoffmann, A. (2018). Psychological characteristics and household savings behavior: The importance of accounting for latent heterogeneity, *Journal of Economic Behavior and Organization* **148**: 66–82.

Guiso, L., Sapienza, P. and Zingales, L. (2006). Does culture affect economic outcomes?, *Journal of Economic Perspectives* **20**(2): 23–48.

Hanna, S. D., Kim, K. and Chen, S. (2016). Retirement savings. in j. xiao, *Handbook of Consumer Finance Research, Springer Publishing, 2nd edition* pp. 33–43.

Hasanpour, Y., Nemati, S. and Tavoli, R. (2018). Clustering system group customers through fuzzy c-means clustering, *Proceedings - 2018 4th Iranian Conference of Signal Processing and Intelligent Systems* pp. 161–165.

Heng-fu, Z. (1995). The spirit of capitalism and savings behavior, *Journal of Economic Behavior and Organization* **28**(1): 131–143.

Kim, G. J. (2017). Do self-control measures aff ect saving behavior?, *Journal of Personal Finance* **16**(2).

Lee, J. M. and Hanna, S. D. (2015). Savings goals and saving behavior from a perspective of maslow's hierarchy of needs, *Journal of Financial Counseling and Planning* **26**(2): 129–147.

Minibas-Poussard, J., Bingol, H. and Roland-Levy, C. (2018). Behavioral control or income? an analysis of saving attitudes and behavior, credit card use and buying on installment, *Revue Européenne de PsychologieAppliquée* **68**(6): 205–214.

Munozmoreno, R., Tandrayen-Ragoobur, V., Seetanah, B. and Sannassee, R. V. (2014). Demographic transition and savings behavior in mauritius, *Emerging Markets and the Global Economy* .

Olukanmi, P. O., Nelwamondo, F. and Marwala, T. (2019). Pam-lite: Fast and accurate k-medoids clustering for massive datasets, *Proceedings - 2019 Southern African Universities Power Engineering Conference* pp. 200–204.

Palakvangsa-Na-Ayudhya, S., Pongchandaj, S., Kriangsakdachai, S. and Sunthornwutthikrai, K. (2017). Keptaom: Savings management system to increase long term savings behavior of children, *Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia,* .

Sabri, M. F. and MacDonald, M. (2010). Savings behavior and financial problems among college students: The role of financial literacy in malaysia, *Cross-cultural Communication* **6**(3): 103–110.

Shcherbakov, M. V., Brebels, A., Nataliya Lvovna Shcherbakova, A. P. T., Alexandrovich, T., Janovsky and Kamaev, V. A. (2013). A survey of forecast error measures, *World Applied Sciences Journal 24 (Information Technologies in Modern Industry, Education and Society)* pp. 171–176.

Strömbäck, C., Lind, T., Skagerlund, K., Västfjäll, D. and Tinghög, G. (2017). Does self-control predict financial behavior and financial well-being?, *Journal of Behavioral and Experimental Finance* **14**: 30–38.

Suppakitjarak, N. and Krishnamra, P. (2015). Household saving behavior and determinants of the forms of saving and investment in thailand, *Journal of Economics, Business and Management* pp. 326–330.

Thanoon, M. and Baharumshah, A. (2012). Comparing savings behavior in asia and latin america: The role of capital inflows and economic growth, *The Journal of Developing Areas* **46**(1): 113–131.

Łuczak, M. (2016). Hierarchical clustering of time series data with parametric derivative dynamic time warping, *Expert Systems with Applications* **62**: 116–130.

Wärneryd, K. E. (1999). The psychology of saving: A study on economic psychology, *Northampton, MA: Edward Elgar Publishing* .

Zeller, M. and Sharma, M. (2000). Many borrow, more save and all insure: Implications for food and micro-finance policy, *Food Policy* **25**(2): 160–161.

Zhu, E., Wen, P., Zhu, B., Liu, F., Wang, F. and Li, X. (2018). Effective clustering analysis based on new designed clustering validity index and revised k-means algorithm for big data, *2018 IEEE Intl Conf on Parallel and Distributed Processing with Applications* pp. 96–102.