An efficient enhanced k-means clustering algorithm for best offer prediction in telecom

Malak Fraihat[1], Salam Fraihat[2], Mohammed Awad[3], Mouhammd AlKasassbeh[1]
[1]Computer Science Department, King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan
[2]Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University, Ajman, United Arab Emirates
[3]Department of Computer Science and Engineering, American University of Ras Al Khaimah, Ras Al Khaimah, United Arab Emirates

Article Info

Article history:

Received Jun 24, 2021
Revised Jan 3, 2022
Accepted Jan 23, 2022

Keywords:

Clustering
Customer best offer
Data mining
K-means
Telecom

ABSTRACT

Telecom companies usually offer several rate plans or bundles to satisfy the customers' different needs. Finding and recommending the best offer that perfectly matches the customer's needs is crucial in maintaining customer loyalty and the company's revenue in the long run. This paper presents an effective method of detecting a group of customers who have the potential to upgrade their telecom package. The used data is an actual dataset extracted from call detail records (CDRs) of a telecom operator. The method utilizes an enhanced k-means clustering model based on customer profiling. The results show that the proposed k-means-based clustering algorithm more effectively identifies potential customers willing to upgrade to a higher tier package compared to the traditional k-means algorithm. Our results showed that our proposed clustering model accuracy was over 90%, while the traditional k-means accuracy was under 70%.

Corresponding Author:

Salam Fraihat
Artificial Intelligence Research Center (AIRC), College of Engineering and Information Technology, Ajman University
Ajman, United Arab Emirates
Email: s.fraihat@ajman.ac.ae

1. INTRODUCTION

Customers are increasingly demanding their service providers to be in tune with their needs and interests. Hence, customers expect these providers, including telecom companies, to offer them new products and offers that align with their needs and expectations. Upselling is one of the approaches used for building a customer-centric culture. It is defined as "a useful sales technique, referring to the attempt to persuade customers to purchase a product or service at a higher level" [1]. Upselling can be achieved by applying the best customer offer concept [2]. The idea of the best customer offer is simply used to express that the customer is subscribed to the offer package that matches his needs and desires. The best offer can be determined by analyzing and studying customer behavioral data. Data mining techniques can facilitate producing the best offers and ultimately magnify the customer upselling approach. Customer profiling and segmentation techniques can be used to determine which type of customers will most properly respond to the company's offers. Such techniques will flag a specific group of customers that the company's marketing campaigns should target.

This paper used clustering and segmentation techniques to divide the customers into homogeneous clusters based on their characteristics. Customer segmentation can help marketing decide what campaign or activities to run [3]. The customer profiling phase requires the knowledge of telecom business experts.

Before customer segmentation, profiling is applied where customers are divided over predefined business rules. After that, customer segmentation is applied to the resultant profiles using the k-means clustering algorithm. K-means is one of the most popular clustering data mining techniques. It is based on partitioning the data into k clusters where each cluster group must contain at least one object, and each object belongs to exactly one group. k-means clustering increases the similarity between the same cluster samples lowering the distances between homo-cluster samples [4]–[6]. k-means is considered a computationally fast clustering algorithm [7]. However, some of k-means drawbacks include its: i) unsuitability for data with noise and ii) dependence on the random initial position of the centroids. It is crucial not to be far away from data points; since a wrong position can result in an infinite number of iterations, which could lead to incorrect clustering [8].

The proposed approach in this paper handles the limitation of randomly initializing the centroid of each cluster in k-means. In our course, k-means is developed so that the positions of the centroids are based on well-defined customers' profiles. The used data in this paper is real telecom data. The data can be described as very noisy data that defines the behavior of a very large customer base. The proposed approach will also handle this limitation.

The proposed approach solves business and technical problems faced by telecom companies in achieving the "best customer offer" marketing approach. On the business side, many customers may not be aware of the existing packages/plans offered. Therefore, many customers may incur additional costs buying additional individual services such as minutes and/or internet bundles (known as top-ups). Usually, the cost of the original package and the additional incurred top-up amount is higher than the cost of the next-tier package, which in this case will be the "best customer offer". Customers who are likely to pay fees outside their package may feel dissatisfaction, which will increase customer churn.

From the technical aspect, the original dataset is big and quite noisy. Consequently, applying a clustering technique such as k-means will not yield the desired results to solve the business problem [9]. Therefore, this paper will develop an enhanced k-means model, which will overcome the limitations mentioned above. The rest of this paper is organized as follows: section 2 presents the related work, section 3 describes the research method and implementation, section 4 presents the results and analysis, and section 5 concludes the paper.

2. LITERATURE REVIEW

Customer profiling and segmentation are common and frequently used techniques to understand customer behavior. Numerous papers addressed this problem. Tripathi *et al*. [10] studied the importance of customer segmentation of the customer relationship management (CRM) data using clustering techniques. They used k-means and hierarchical clustering for CRM data of a mall. The data consisted of customer name, gender, age, annual income, and spending score. Due to the small dataset size, k-means delivered better performance in terms of time and accuracy. Tripathi *et al.* [10] found that k-means can also deal better with larger datasets than hierarchical clustering. However, the limitation of k-means is the selection of the number of clusters (k).

Badase *et al.* [11] presented a comparative analysis of available clustering models including k-means, single linkage, average linkage, complete linkage, balanced iterative reducing and clustering using hierarchies (BIRCH), and density-based spatial clustering of applications with noise (DBSCAN) using their basic hyper parameters. The authors used multiple data known datasets were in their study, including iris, car, WDBS, yeast, wine, glass, diabetics, optical digits, and musk. It was found that k-means can work with many variables and is computationally faster than hierarchical clustering, if the number of clusters (k) is small. On the other hand, the study concluded that k-means is very sensitive to the initial random position of each clusters' centroids [11]. Abdi and Abolmakarem [12] proposed a customer behavior mining framework in a telecom company using data mining techniques. The dataset of this work contains 1,000 records and includes information about the customers of a telecom company. It consists of 25 attributes and one target variable, which is the churn status. The framework works to build portfolio analysis for customers based on socio-demographic features using k-means. The customers are grouped into five clusters. Then, each cluster is analyzed based on two features: the number of hours the telecom services used and the number of services selected by customers. Consequently, and based on clustering portfolio analysis, six clusters of customers are divided and identified into three levels of attractiveness. Ultimately, this will help in establishing a lasting relationship with customers.

Gopi and Sumalatha [13] analyzed customer behavior, preferences, and attributes in a telecom company to determine the homogeneity of individual clusters and the dissimilarity of customers within the same cluster. They used two-layer clustering to grasp both macro and micro customer patterns and behavior. The first layer examines customer value, which is the amount of consumption for each customer, and the second layer uses consumer-behavior features for further grouping. The first layer is grouped into multiple clusters, and then each cluster in the first layer is sub clustered in the second layer based on business expert rules.

Min and Lin [14] used signaling call detail record (CDR) data to train the clustering k-means model to discover hidden characteristics of fraud phones. Hence, detect fraud and identify fraud numbers. The data is from the CDR and contains 26,670 records grouped into seven clusters. The study concluded that as the data increased, the accuracy increased. The accuracy of the clusters was achieved by studying the characteristic differences between each cluster, and therefore the clustering results were found to have strong interpretability.

Insani and Soemitro [15] built a data mining technique to profile the customers using the recency, frequency, and monetary (RFM) model and k-means clustering. The data used is from a telecommunication company in Indonesia. The input for the k-means model is the RFM model. Using k-means, customer profiling is built based on customer segmentation, and according to customer's usage of services, customer invoice and customer payment, which allows the model to identify profitable customers. The results show that the implementation of the method is feasible and that customers can be classified as loyal, profitable, and likely to churn.

Kapil and Chawla [16] studied different types of distance functions used in building k-means clusters, mainly Euclidean distance and Manhattan distance. The data is a dummy dataset of 200 people's online behavior on a specific social network site. They concluded that choosing the distance metric function plays an important role, and that Euclidean distance outperforms Manhattan distance in several aspects.

Kamalraj and Malathi [17] applied a semi-supervised constraint-based churn clustering (SCCC) technique to identify the different types of churns. The semi-supervised learning method in SCCC works by using larger amounts of unlabeled data and smaller amounts of labeled data in the clustering process. The results show that the SCCC outperforms the association that used as a benchmark.

Ye *et al.* [18] applied k-means on Changzhou telecom customers to cluster them based on values to understand consumption patterns and behavior for each group of customers. The categorization will help in the decision-making process to achieve the strategic objective of profit improvements. Once customers are categorized, each customer group will be approached with a tailored marketing program. Other studies analyzed customer's opinions rather than their behavior to understand their demands. Another study analyzed customers' opinions rather than their behavior to understand their demands [19].

Additionally, several research papers utilized different machine learning techniques to predict customer churn in the telecom industry [20]–[23]. Other studies aimed to measure the degree of Telecom operator's sincerity [24]. On the other hand, our paper aims to indirectly mitigate customer churn by satisfying customers and strengthening their loyalty.

3. METHOD AND IMPLEMENTATION

This section elaborates the method and detailed implementation behind the proposed approach for predicting customer best offer in the telecom industry. The cross-industry standard process data mining (CRISP-DM) model is widely used in the business domain [25]–[27]. This model is followed throughout our study, as it has a structured approach in using data mining to solve and deal with business problems. In this study, CRISP-DM is followed to find and recommend the best customer offer.

Figure 1 shows an overview of the whole work process in the proposed approach. The figure starts with business understanding and data understanding, then data extraction and cleaning. Afterward, the algorithm is implemented in the following three steps: first, customer profiling, then k-means training, and lastly, k-means ground truth. This algorithm will result in a list of customers who are subscribed in Offer_A, but have the potential to switch and upgrade their bundle to Offer_B.

3.1. Business understanding

Telecom companies' bundles can be either postpaid or prepaid. As the name indicates, postpaid bundles are payable at the end of a specific period, for instance, at the end of every month. Prepaid bundles are paid in advance before using any of the telecom company services. In Jordan, most of the telecom customers (89%) are subscribed to prepaid services [28]. The proposed Best Customer Offer prediction approach aims to detect customers whose behavior does not match their subscribed offer. This detection can help telecom companies in several ways. For example, having a relevant, comprehensive subset of customers to focus on will utilize the company's resources. The marketing team will focus their efforts on clients who are likely to upgrade their package. Also, this approach will be less bothersome to clients who are not interested in the offer. In this study, Offer_A and Offer_B are used. Both offers are prepaid packages, which provide the customers with monthly services. Table 1 shows the services included in each bundle (Offer_A and Offer_B).

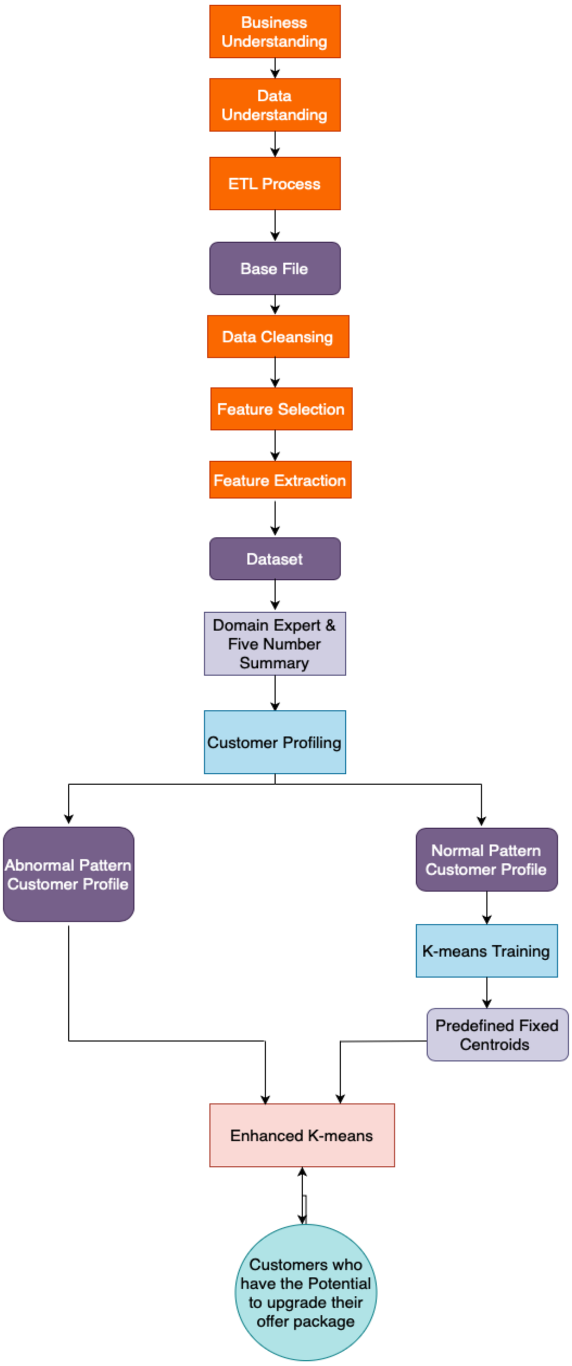
Business Understanding
Data Understanding
ETL Process
Base File
Data Cleansing
Feature Selection
Feature Extraction
Dataset
Domain Expert & Five Number Summary
Customer Profiling
Abnormal Pattern Customer Profile
Normal Pattern Customer Profile
K-means Training
Predefined Fixed Centroids
Enhanced K-means
Customers who have the Potential to upgrade their offer package

Figure 1. Process overview flowchart

Table 1. Bundle services description

Service	Internet Bundle (GB)	Outside the network Minutes	Within the Network Minutes	SMS	Monthly Subscription Fees
Offer_A	15 GB	1000 minutes	Unlimited	Unlimited	6 JOD (8.50 USD)
Offer_B	30 GB	1750 minutes	Unlimited	Unlimited	9 JOD (12.70 USD)

3.2. Data understanding

The dataset used to apply the proposed best customer offer is actual data extracted and aggregated from CDRs of a telecom company in Jordan. Each call made in the telecommunication network, whether received or initiated, has its descriptive information that is saved into the CDR. Thus, the number of stored

CDRs is extremely huge. The data is extracted from the CDRs and prepared in a specific format, then stored in the “base file,” where each record represents a unique customer. The base file has a total of 70 attributes. Table 2 shows some of the base file attributes along with their definitions. The CDRs’ data is considered as high quality data for its accuracy, completeness, timeline, and validity [29].

This study will target only the youth segment of the prepaid customers and determine their likelihood to upgrade their package. Increasing the retention rate of the youth segment can benefit the telecom company in the short and long runs as it will contribute towards customers’ satisfaction, which in return will improve customers’ loyalty. The total number of customers is approximately 100K, 78K of them are subscribed to Offer_A, and about 20K customers are subscribed to Offer_B. The proposed enhanced k-means is applied to three different consequent months.

Table 2. Base file list of the most important attributes

Field	Definition
Fees	The charged amount for subscription fees
Onnet Minutes	The duration of calls measured in minutes on the same network
Onnet Fees	The charged number of calls measured in minutes on the same network
Offnet Minutes	The duration of calls measured in minutes to another local network
Offnet Fees	The charged number of calls measured in minutes on the same network
International Minutes	The duration of calls measured in minutes to an international destination
International Fees	The number of calls measured in minutes to an international destination
Data Usage	The data consumed by the customer in MBs
Data Usage Fees	The fees of consumed data by the customer (in MBs)
Local SMS Fees	The charged amount of fees for an SMS sent to other local operators
International SMS Fees	The charged amount of fees for an SMS sent to other international operators
Top ups	The frequency of how many times the customer has made any top ups
ARPU	The Average Revenue per User.
PackageType	Type of the package post-paid/ prepaid
Segment	The segment of the packager type
Rate plan	Offer that the customer is subscribed to
DeviceModel	The device model of the customer
Age on network	Membership duration (customer since)
Status	If the line is active or inactive
Minutes to operator x	The duration of calls measured in minutes to operator x
Minutes to operator y	The duration of calls measured in minutes to operator y

3.3. Data preparation

Data preparation is an essential step before applying any data mining models. The data preparation phase starts with a crucial pre-step, known as the extract, load, transform (ELT) step, to migrate data from respective data sources into a database or data warehouse. Since the amount of CDR files is enormous, the ELT process is more suitable than extract/transform/load (ETL). Also, the servers’ specifications are capable of loading the data before transforming it. Figure 2 illustrates the ELT process pipeline.

a. Extraction: the billing system generates several CDRs daily. The file size is small (less than 1 MB). The data is loaded into the data warehouse (DWH) through the mediation layer, which facilitates communication between different services. In this case, between the billing system and the DWH.
b. Loading: all files that were previously generated from the billing system and loaded into the DWH are merged into one large file and loaded into the staging tables. The staging tables are used to hold the data temporarily before loading them to the target tables, which are the production tables.
c. Transformation: in this process, the data is reshaped, cleaned, and transformed into a suitable format. For instance, some fields are saved in float format while others as integers or strings. Lastly, all data is aggregated monthly to produce the base file.
d. Filtering: as the base file consists of all active, inactive, and suspended users, the customers are filtered based on their “Network Status.” Hence, only active users are filtered and kept as they are the targeted ones. Active users are defined as the users who pay their subscription fees regularly every month. In this study, as mentioned earlier, the focus is on the youth segment.
e. Selection: in this process, the base file is filtered based on the offered packages that represent the youth segment, which are Offer_A and Offer Offer_B.
f. Feature importance and selection: once the data is clean, a feature selection process is applied to the data, where only specific features are selected. The selected features are the most important features. Feature importance is defined based on the following:
    - The correlation with the offer packages (pearson correlation measure is calculated),
    - The domain expert who understands the general behavior of the customer base and the offered packages.

Pearson correlation is a statistical test that measures the association between continuous variables; it is based on the method of covariance. The correlation coefficient gives the magnitude of the correlation and the direction of the relationship [30]. It is used to measure the correlation between the offer packages and other fields in the base file. The correlation along with the domain expert will help in the feature selection process. Figure 3 is a heat bar chart that shows the features, which have the highest correlation with offer packages. Clearly, data usage (internet), number of calls outside the network, and the ARPU are the features with the highest correlation with offer package. The following features were selected from the base file:

- MSISDN: Customer’s phone number
- Total_Offnet_MIN: Calls duration to another local network (in minutes)
- TotalData_MB: Internet data usage per customer measured in MBs
- Competitor #1_MOU: Calls duration to operator #1 (in minutes)
- Competitor #2 _MOU: Calls duration to operator #2 (in minutes)
- ARPU_JOD: The Average Revenue per User (Jordanian Dinar)
- Package_Name: Bundle A or B.

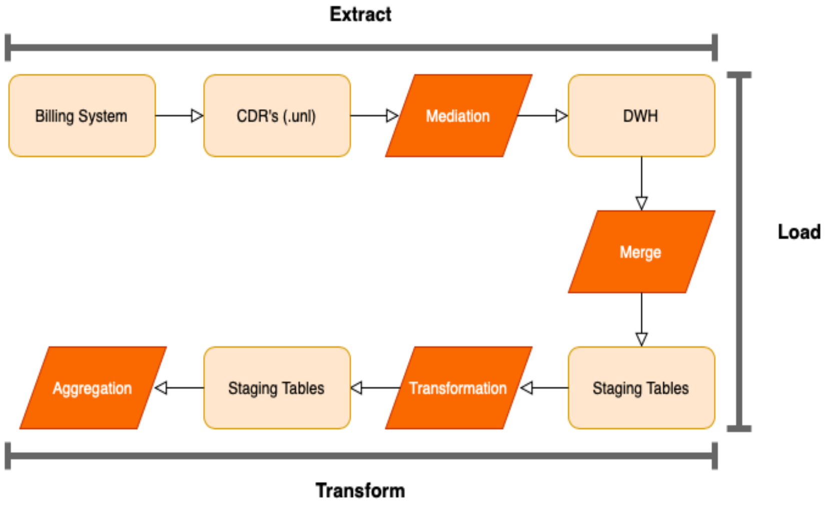

Figure 2. ELT process pipeline

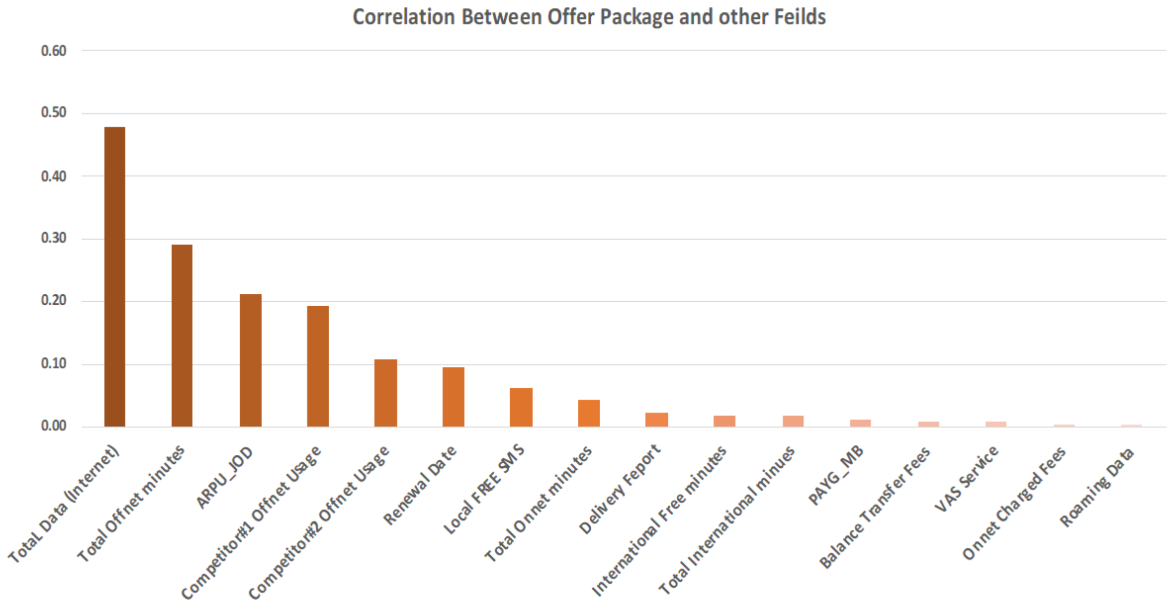

Figure 3. Pearson correlation between offer packages and other fields

3.4. Customer profiling

Knowing customers' behavior through customer profiling will allow companies to design and make better decisions concerning their provided services and products. Customer profiling will be used as the base to enhance the data mining algorithm, which is used to detect potential customers to upgrade their package into the higher one (from A to B). Customer profiling is built based on business and data understanding, as well as some basic statistics such as a five-number summary. Table 3 represents the summary of the basic statistics of Offer_A and Offer_B. It is clear when looking at the mean and the 50th percentile that most of Offer_A customers are using their full internet bundle. Furthermore, some customers are even overspending to buy additional internet bundles to meet their needs, this is apparent when looking at the 75th percentile as it exceeds the amount given to Offer_A subscribed customers.

The customers' profiling allows identifying the customers who behave according to their package and do not exceed it. Each offer profiled customer group represents the normal behavior expected based on the customers' subscription offer. For Offer_A customers, the normal internet data usage behavior should fall between 2 to 8 GB, whereas for Offer_B customers, the normal behavior of internet data consumption should fall between 15-20 GB per month. A dataset of the normal behavior is constructed to be used in the enhanced k-means algorithm.

Table 3. Summary of Offer_A and Offer_B customer base statistics

	Offer_A		Offer_B	
Summary Statistics	Data Usage (GB)	Off-net (minutes)	Data Usage (GB)	Off-net (minutes)
mean	15.82	227.30	27.82	396.11
std	5.12	259.38	7.67	441.00
50%	15.43	134.48	28.56	243.00
75%	33.38	302.75	40.32	509.00

3.5. Enhancement of k-means algorithm

The classical k-means algorithm starts by randomly initializing the clusters' centroids. K-means is based on two repeating steps, which are:

a. Cluster assignment step: k-means goes through each data point and depending on which cluster centroid is closer, it assigns the data point to one of the centroids.
b. Move centroid step: K-means calculates the average of all the points in a cluster and moves the centroid to that average location.

K-means is an iterative algorithm; it keeps repeating the above steps until there is no change in the centroids' location. The proposed customer best offer works by enhancing k-means so that the clusters' centroids are not initialized randomly but are initialized based on the customer profiling step. So that, the "moving centroid step" does not exist when applying the proposed k-means to the whole customer base.

The idea behind the change in the centroid definition process is because, with the traditional way, the algorithm tends to find a centroid of customers who have an abnormal consummation behavior in terms of bundle services. So the generated centroid represents a center of abnormal customer behavior. In contrast, the real need is to measure the closeness of the customer to normal behavior in terms of consumption of bundled services. In other words, if the customer is close to the centroid, this means he does not exceed his package, and there is no low chance of upgrading. On the other hand, the farther the customer moves away from the centroid, the higher the probability of changing the offer because his need does not match the current offer services.

Figure 4 shows an overview of the expected results and the differences between the traditional k-means and the proposed k-means. It is clear that the traditional k-means algorithm fails to distinguish the two groups of yellow and orange data points. This is due to the inaccurate centroids' locations (shown as stars in the figure). On the other hand, classifying the data points into two clusters using the proposed enhanced k-means yields better results due to the centroids' more accurate location.

3.5.1. K-means training

The used dataset in the customer profiling step explained in section 3.4 is used to train the k-means algorithm. The k-means clustering algorithm attempts to split the given unlabeled dataset of Offer_A and Offer_B into two clusters. Algorithm 1 illustrates the k-means training and ground truth steps. Training the k-means model on the profiled customer's dataset will output two centroids points. Each centroid point can be seen as a representative of each customer group. In other words, each centroid represents the actual expected behavior of each offer group. The resultant customer clusters from the k-means training step are only two clusters: i) group Offer_A that has only customers who are subscribed to Offer_A and ii) group Offer_B that has only customers who are subscribed to Offer_B. These two customer clusters behave according to their given bundles' features. Below are the steps of the k-means training phase working mechanism:

a. The algorithm starts by first initializing two centroids, which are centroid Offer_A and centroid Offer_B (k=2).
b. The centroids' values and their positions are placed randomly. The algorithm then starts iterating over the data points one by one and measuring the distance between the data points and each of the two centroids. The data points closer to centroid Offer_A will be grouped into cluster Offer_A, whereas the data points closer to centroid Offer_B will be grouped into cluster Offer_B.
c. When each data point is grouped and associated with the closest centroid to its position, the new value of the centroid is recalculated. This value is equal to the sum of data points associated with a centroid divided by the number of those data points. This step is repeated, and the centroids' values are recalculated until there is no change in any of the centroids' values.

"Closest" in this context means the smallest distance from a data point to the centroids. The used distance similarity measure is Euclidean Distance. In Euclidean Distance, if p=(p1, p2,..., pn) and q=(q1, q2,..., qn) are two points in a space on the Cartesian coordinates, then the distance (d) from p to q, or from q to p is given by the Euclidean distance, shown in (1).

$$d(p,q) = q(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

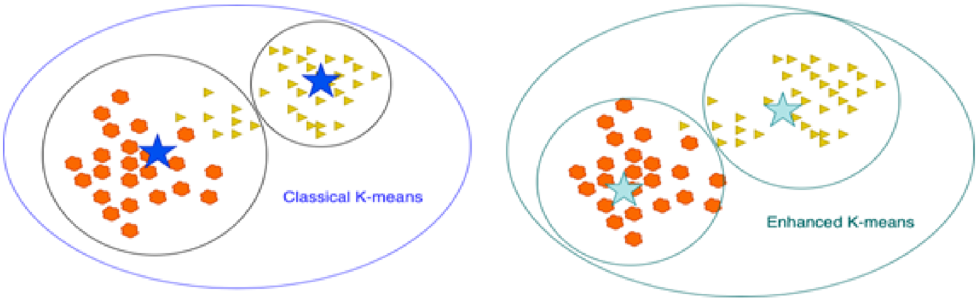

Figure 4. Traditional k-means vs. proposed k-means. The stars represent the centroids; Triangles represent Offer_A; Hexagon represent Offer_B

Algorithm 1: Enhanced k-means algorithm using customer profiling technique

```
Input: S' set of input data points of normal behavior customers, $x \in R^d$ and an integer k≥2,
S' is a subset of S which is the total data points.
Output: k clusters $C_1$, $C_2$,…, $C_k$
//Here the centroids of the normal behavior customers are extracted.
Select randomly K clusters centers (centroids) $c_1$, $c_2$,..., $c_k$, where k=2; $C_1$:Offer_A, $C_2$:
Offer_B.
Repeat
        For each data point x in S' do
        For all centroids $c_j$, 1≤j≤k do
              If $\|x-c_j\|<\|x-c_l\|$, j≠l, 1≤j, l≤k then,
                Assign x to $C_j$
        Consider the center of points of every cluster as a centroid: $c_j = \frac{1}{n_j}\sum_{x \in C_j} x, n_j = |C_j|$
Until the objective function is minimized
// Here the centroids of the normal behavior customers fixed above are used to classify all
the data point in S.
For each data point x in S do
For all centroids $c_j$, 1≤j≤k do
      If $\|x-c_j\|<\|x-c_l\|$, j≠l, 1≤j, l≤k then,
        Assign x to $C_j$
```

3.5.2. K-means ground truth

The above training step resulted in two centroids, one for Offer_A and another one that represents Offer_B. These two centroids represent the normal behavior of each offer. The two centroids will be used to apply enhanced k-means to the rest of the data. The rest of the data is all customers' data records, excluding those customers with normal behavior who were used in the customer profiling and k-means training step. In other words, the used data for the k-means ground truth is about customers who do not follow the normal behavior of their subscribed offer, so that they mostly buy top-ups to meet their needs. The following steps explain how the enhanced k-means works on ground truth data works:

a. The centroids of bothOffer_AandOffer_B are already fixed by applying the classical k-means algorithm on the customer profiling dataset.
b. The algorithm starts iterating over the data points one by one and measuring the distance between the data points and each of the two centroids. The data points closest to centroid Offer_A will be grouped into cluster Offer_A, whereas the data points closest to centroid Offer_B will be grouped into cluster Offer_B.

In the proposed enhanced k-means, the profiling step followed by the k-mean straining step has great importance in building a reliable model. When directly building the classical k-means model, the initial value of centroids is not placed accurately. A bad random initial position of a centroid might lead to clustering the data in a completely not precise and wrong way. Therefore, having fixed centroids built based on understanding the data and the business itself will help produce more accurate results.

4. RESULTS AND DISCUSSION

Figures 5 and 6 show that the data distribution of the two offer plans is similar. It is unusual to have two offer plans with different bundle features with the same data distribution. This very noisy data makes it challenging for the classical k-means model to differentiate between the two types of customers. Hence, the process of finding out the customers who buy top-up bundles to meet their needs and thus finding Offer_A customers who have the probability to upgrade to Offer_B is proving to be difficult.

Outliers increase the variability in a dataset, which decreases statistical power. Consequently, excluding outliers can cause results to become statistically significant [30]. Figure 7 shows the distribution of the data before and after applying the anomalies reduction. It is clear that for Offer_B, the lower and higher anomalies are removed. Still, for Offer_A, only lower anomalies are removed since the higher anomalies can be counted as customers who may upgrade to a higher offer package, which is Offer_B.

The used data in this study is for three different months. Thus, three different trials are built. A confusion matrix is used to present the k-means training step results based on the customer profiling step. The confusion matrix will allow us to predict customers who are actually subscribed to Offer_A accurately, and k-means training predicted them as Offer_A customers and the same scenario for Offer_B customers. The equations for measuring accuracy of predicting both Offer_A and Offer_B customers are:

- Accuracy of predicting Offer_A actual customers=Predicted as Offer_A Customers out of the total number of Offer_A actual customer base.
- Accuracy of predicting Offer_A actual customers=TP/(TP+FP)
- Accuracy of predicting Offer_B actual customers=Predicted as Offer_B Customers out of the total number of Offer_B actual customer base.
- Accuracy of predicting Offer_B actual customers=TN/(TN+FN)

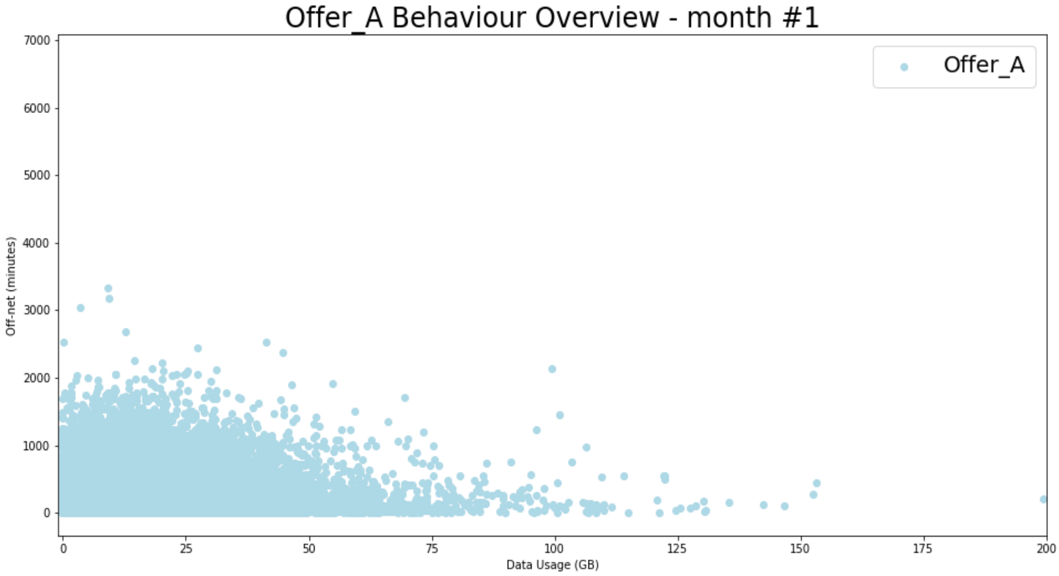

Figure 5. Offer_A data distribution- month#1

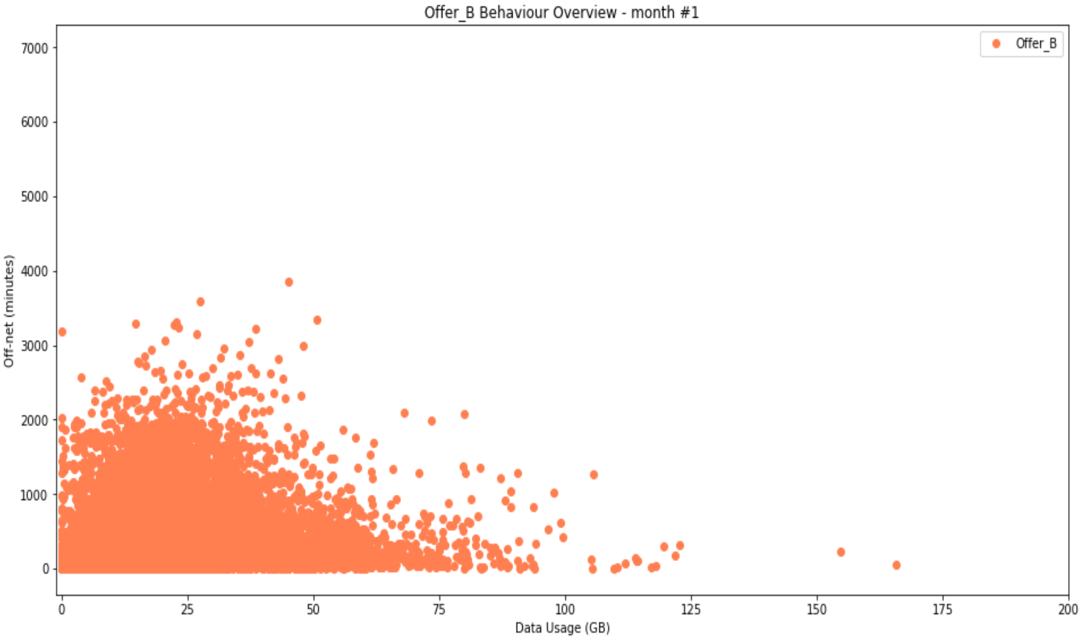

Figure 6. Offer_B data distribution- month#1

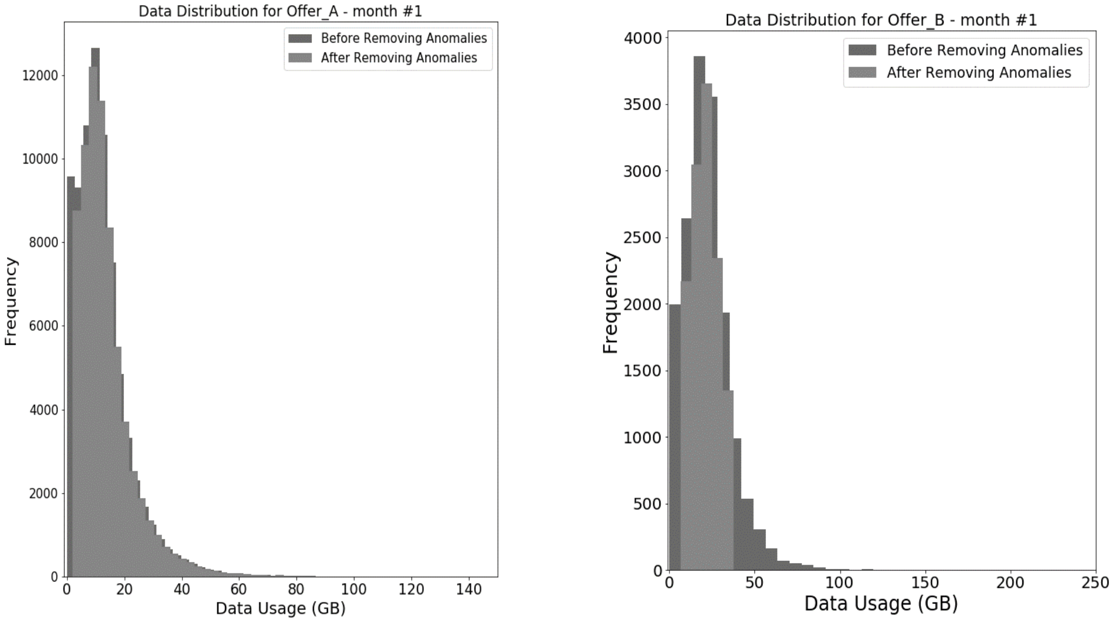

Figure 7. Data of offers A (left) and B (right) for month #1

Figure 8 represents the confusion matrix result from the k-means training step built upon profiled customer data for one month. The number 20 in Figure 8 represents the subscribed customers to Offer_A, but the k-means training step predicted them as Offer_B customers. The misclassification rate is insignificant (about 0.001%). This indicates that the customer profiling step is very precise since the model was able to classify the customers according to their subscribed offers with very high accuracy. Note that the data of the profiled customers, which was used in the k-means training, is unlabeled so that k-means is running as an unsupervised model, and then the results are checked against the actual label/class of each customer. Matching customers can be briefly defined as the number of customers predicted to have the potential to upgrade their offer package out of the actual total number of migrations. Figure 9 illustrates the matching customers.

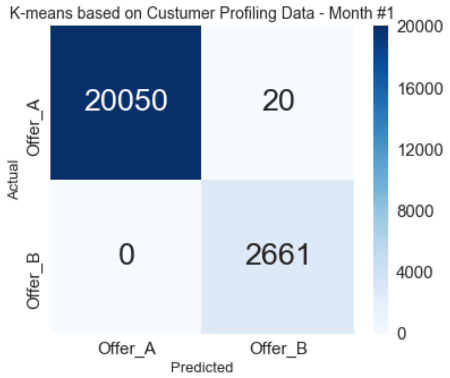

Figure 8. K-means based on customer profiling data-month#1

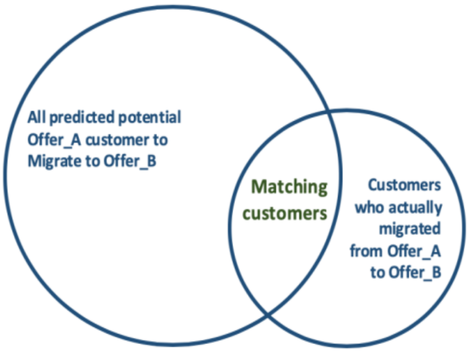
All predicted potential Offer_A customer to Migrate to Offer_B
Matching customers
Customers who actually migrated from Offer_A to Offer_B

Figure 9. Matching customers

Table 4 represents a summary of the evaluation of the findings against actual migrations; two approaches are presented:

a. Classical k-means: The matching accuracy is in the range of 60’s%. The classical k-means can relatively distinguish the customers of Offer_A who have the potential to upgrade their rate plan to Offer_B.
b. Proposed approach: Total Offer_A customers base for month #1 is 77,599 customers. Our proposed approach found that out of the 77,599 total base, 45,502 customers have the potential to upgrade their offer plan from Offer_A to Offer_B. In reality, 478 customers (per the change package file) upgraded their bundles from A to B. Out of these 478 customers, our approach accurately classified 435 of them as having the “potential to upgrade.” This yields a matching accuracy for month #1 equals 91%, which is relatively very high. Same for month #2, total Offer_A customer base is 79,949, and from the total base the proposed enhanced k-means predicted that 47,127 of them have the potential to upgrade their offer package, and the matching accuracy for month #2 is 93.1%, which is also consistently high. Also, for month #3, the matching percentage is satisfyingly very high. On the other hand, using the classical k-means, the matching accuracy is in the range of 60’s%, classical k-means can relatively distinguish the customers of Offer_A who have the potential to upgrade their rate plan to Offer_B. It is clear that the anomalies reduction as a data preparation step and the use of the customer profiling techniques to fix the k-means centroids increase the accuracy of the best next offer system considerably.

Table 4. Classical K-means as benchmark vs. proposed enhanced k-means approach

Month	Total Offer_A Customer Base	Total Actual Migrations (From Offer_A to Offer_B)	Classical K-means as benchmark			Proposed approach		
			Offer_A Actual Customers Predicted as Offer_B Customers	Total Migrations (From Offer_A to Offer_B)	Matching Percentage	Offer_A Actual Customers Predicted as Offer_B Customers	Total Migrations (From Offer_A to Offer_B)	Matching Percentage
#1	77,599	478	55,565	326	68.2%	45,502	435	**91.0%**
#2	79,949	377	54,004	231	61.3%	47,127	351	**93.1%**
#3	79,956	259	58,004	169	65.3%	48,293	244	**94.2%**

5. CONCLUSION

Customers of telecom companies are increasingly demanding their companies to be in tune with their needs and interests. Hence, they expect new products, promotions, and offers that fulfill their needs and expectations without explicitly sharing much. Companies can maintain their customers’ base and attract the attention of other potential new customers by becoming proactive in approaching business opportunities of satisfying their customers. This study proposes a practical way to detect the group of customers who are likely to upgrade their offer package using the enhanced k-means clustering algorithm along with customer profiling. The used dataset is actual data that describes around 100K real telecom customers. The data is extracted and prepared before use in the proposed model. The data showed very noisy unnatural customers' behavior, making it hard to use in any data mining model, including the classic k-means model. The proposed enhanced k-means overcame this very noisy data distribution issue. The results showed that the proposed enhanced k-means could effectively detect customers of a specific offer package that behave more like a higher tier offer package, hence offering this group of customer suitable campaigns that match their needs. The evaluation results confirmed that the proposed method is accurate and effective in a telecom company or any other business sector that relies on customers' subscriptions.

REFERENCES

[1] B. Denizci Guillet, “Online upselling: Moving beyond offline upselling in the hotel industry,” *International Journal of Hospitality Management*, vol. 84, p. 102322, Jan. 2020, doi: 10.1016/j.ijhm.2019.102322.

[2] M. Kjelin, “What makes the difference?: a study of the purchase process from a buyer’s perspective.” University of Borås, pp. 1–45, 2015.

[3] C.-I. Chang and J.-C. Ho, “A two-layer clustering model for mobile customer analysis,” *IT Professional*, vol. 19, no. 3, pp. 38–44, 2017, doi: 10.1109/MITP.2017.54.

[4] P. K. Sharma and G. Holness, “L^2-norm transformation for improving k-means clustering,” *International Journal of Data Science and Analytics*, vol. 3, no. 4, pp. 247–266, Jun. 2017, doi: 10.1007/s41060-017-0054-1.

[5] D. Cheng, X. Ding, J. Zeng, and N. Yang, “Hybrid K-means algorithm and genetic algorithm for cluster analysis,” *TELKOMNIKA Indonesian Journal of Electrical Engineering*, vol. 12, no. 4, Apr. 2014, doi: 10.11591/telkomnika.v12i4.4805.

[6] A. D. Indriyanti, D. R. Prehanto, and T. Z. Vitadiar, “K-means method for clustering learning classes,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 835–841, May 2021, doi: 10.11591/ijeecs.v22.i2.pp835-841.

[7] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Elsevier, 2011.

[8] Z. Zhang, J. Zhang, and H. Xue, “Improved K-Means clustering algorithm,” in *2008 Congress on Image and Signal Processing*, 2008, pp. 169–172, doi: 10.1109/CISP.2008.350.

[9] J. M. Peña, J. A. Lozano, and P. Larrañaga, “An empirical comparison of four initialization methods for the K-Means algorithm,” *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027–1040, Oct. 1999, doi: 10.1016/S0167-8655(99)00069-0.

[10] S. Tripathi, A. Bhardwaj, and E. Poovammal, “Approaches to clustering in customer segmentation,” *International Journal of Engineering and Technology*, vol. 7, no. 3.12, Jul. 2018, doi: 10.14419/ijet.v7i3.12.16505.

[11] P. S. Badase, G. P. Deshbhratar, and A. P. Bhagat, “Classification and analysis of clustering algorithms for large datasets,” in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Mar. 2015, pp. 1–5, doi: 10.1109/ICIIECS.2015.7193191.

[12] F. Abdi and S. Abolmakarem, “Customer behavior mining framework (CBMF) using clustering and classification techniques,” *Journal of Industrial Engineering International*, vol. 15, no. S1, pp. 1–18, Dec. 2019, doi: 10.1007/s40092-018-0285-3.

[13] Y. Gopi and V. Sumalatha, “Tele comm. customer data analysis using multi-layer clustering model,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 1, 2018.

[14] X. Min and R. Lin, “K-Means algorithm: fraud detection based on signaling data,” in *2018 IEEE World Congress on Services (SERVICES)*, Jul. 2018, pp. 21–22, doi: 10.1109/SERVICES.2018.00024.

[15] R. Insani and H. L. Soemitro, “Business intelligence for profiling of telecommunication customers,” *Asian Pacific Journal of Contemporary Education and Communication Technology*, vol. 2, no. 2, pp. 151–161, 2016.

[16] S. Kapil and M. Chawla, “Performance evaluation of K-means clustering algorithm with various distance metrics,” in *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, Jul. 2016, pp. 1–4, doi: 10.1109/ICPEICES.2016.7853264.

[17] N. Kamalraj and A. Malathi, “Semi-supervised churn clustering for fault and constraints prediction in telecom industry,” *Indian Journal of Science and Technology*, vol. 9, no. 33, Sep. 2016, doi: 10.17485/ijst/2016/v9i33/79842.

[18] L. Ye, C. Qiuru, X. Haixu, L. Yijun, and Z. Guangping, “Customer segmentation for telecom with the k-means clustering method,” *Information Technology Journal*, vol. 12, no. 3, pp. 409–413, Jan. 2013, doi: 10.3923/itj.2013.409.413.

[19] M. Gupta and S. Sebastian, “Framework to analyze customer’s feedback in smartphone industry using opinion mining,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3317–3324, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3317-3324.

[20] A. Khede, A. Pipliya, and V. Malviya, “A novel approach for predicting customer churn in telecom sector,” in *Social Networking and Computational Intelligence*, Springer Singapore, 2020, pp. 295–304.

[21] H. Jain, A. Khunteta, and S. Srivastava, “Churn prediction in telecommunication using logistic regression and logit boost,” *Procedia Computer Science*, vol. 167, pp. 101–112, 2020, doi: 10.1016/j.procs.2020.03.187.

[22] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” *Journal of Big Data*, vol. 6, no. 1, p. 28, Dec. 2019, doi: 10.1186/s40537-019-0191-6.

[23] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, “Clustering prediction techniques in defining and predicting customers defection: the case of e-commerce context,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2367–2383, Aug. 2018, doi: 10.11591/ijece.v8i4.pp2367-2383.

[24] D. A. Omar, M. Baslam, M. Nachaoui, Fakir, and Mohamed, “Customers’ perception towards services of telecommunications operators,” *International Journal of Informatics and Communication Technology (IJ-ICT)*, vol. 6, no. 3, pp. 146–154, Dec. 2017, doi: 10.11591/ijict.v6i3.pp146-154.

[25] B. AlArmouty and S. Fraihat, “Data analytics and business intelligence framework for stock market trading,” in *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICTCS.2019.8923059.

[26] L. AlWreikat, A. AlShawa, D. Al-Rimawi, and S. Fraihat, “Business intelligence and data analytics system for mobile money,” in *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems-DATA ’19*, 2019, pp. 1–6, doi: 10.1145/3368691.3368729.

[27] B. Abutahoun, M. Alasasfeh, and S. Fraihat, “A framework of business intelligence solution for real estates analysis,” in *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems-DATA ’19*, 2019, pp. 1–9, doi: 10.1145/3368691.3368730.

[28] B. Al-Sarhan, “Market statistics and indicators,” *Telecommunications Regulatory Commission TRC*. https://trc.gov.jo/Pages/viewpage?pageID=86 (accessed Oct. 12, 2020).

[29] H. Ramapriyan, G. Peng, D. Moroni, and C.-L. Shie, “Ensuring and improving information quality for earth science data and products,” *D-Lib Magazine*, vol. 23, no. 7/8, Jul. 2017, doi: 10.1045/july2017-ramapriyan.

[30] P. Schober, C. Boer, and L. A. Schwarte, “Correlation coefficients,” *Anesthesia and Analgesia*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

BIOGRAPHIES OF AUTHORS

Malak Fraihat earned her MSc in Data Science from Princess Sumaya University for Technology (PSUT). She received a BSc in Computer Information Systems from the University of Jordan. Currently, she is working at Microsoft in the Data and AI team. Her research interests include pattern recognition, data science, and machine learning. She can be contacted at email: fraihatmalak@gmail.com.

Salam Fraihat is an associate professor in the computer Science and Engineering School, Ajman University, UAE. He received his PhD from Aix-Marseille III University, France in 2010. His main research interests lie in the area of information system, information retrieval, data science, and artificial intelligence. He has published many papers in refereed journals and conference proceedings. He can be contacted at email: s.fraihat@ajman.ac.ae

Mohammed Awad joined the Department of Computer Science and Engineering at the School of Engineering in the American University of Ras Al Khaimah in Ras Al Khaimah, UAE in 2013. Dr. Awad earned his PhD in Computer Science from the University of Houston in the United States in 2011. He earned a MSc in Computer Science at the same university in 2006. Dr. Awad’s research interests include machine learning, e-learning, CubeSats, and security, more specifically E-voting and I-voting security. He can be contacted at email: mohammed.awad@aurak.ac.ae.

Mouhammd Alkasassbeh graduated from the school of computing, Portsmouth University, UK in 2008. He is currently a full professor in the Computer Science Dept. Princess Sumaya University for Technology. His research interests include network traffic analysis, network fault detection, classification network fault and abnormality and machine learning in the area of computer networking and network security. He can be contacted at email: m.alkasassbeh@psut.edu.jo.