Wassouf *et al. J Big Data* (2020) 7:29
https://doi.org/10.1186/s40537-020-00290-0

Journal of Big Data

RESEARCH Open Access

Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study

Wissam Nazeer Wassouf[1*], Ramez Alkhatib[2], Kamal Salloum[1] and Shadi Balloul[3]

*Correspondence: w.wassouf@albaath-univ.edu.sy
[1] Faculty of Information Technology-Department of software engineering and information systems, Al-Baath University, Homs, Syria
Full list of author information is available at the end of the article

Abstract

Given the growing importance of customer behavior in the business market nowadays, telecom operators focus not only on customer profitability to increase market share but also on highly loyal customers as well as customers who are churn. The emergence of big data concepts introduced a new wave of Customer Relationship Management (CRM) strategies. Big data analysis helps to describe customer's behavior, understand their habits, develop appropriate marketing plans for organizations to identify sales transactions and build a long-term loyalty relationship. This paper provides a methodology for telecom companies to target different-value customers by appropriate offers and services. This methodology was implemented and tested using a dataset that contains about 127 million records for training and testing supplied by Syriatel corporation. Firstly, customers were segmented based on the new approach (Time-frequency-monetary) TFM (TFM where: Time (T): total of calls duration and Internet sessions in a certain period of time. Frequency (F): use services frequently within a certain period. Monetary (M): The money spent during a certain period.) and the level of loyalty was defined for each segment or group. Secondly, The loyalty level descriptors were taken as categories, choosing the best behavioral features for customers, their demographic information such as age, gender, and the services they share. Thirdly, Several classification algorithms were applied based on the descriptors and the chosen features to build different predictive models that were used to classify new users by loyalty. Finally, those models were evaluated based on several criteria and derive the rules of loyalty prediction. After that by analyzing these rules, the loyalty reasons at each level were discovered to target them the most appropriate offers and services.

Keywords: TFM, RFM, Customer loyalty, Classification algorithms, Customer behavior, Machine learning, Big data, CDR, CRM, Features selection

Introduction

The telecom sector is witnessing a massive increase in data, and by analyzing this massive data, telecom operators can manage and retain customers. It is also important for companies to be able to predict the amount of income they may receive from their active customers. For this purpose, they need models able to determine customer loyalty. The cost associated with customer gain is usually higher than the cost associated

Springer Open

© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

with maintaining it [1]. Prediction can be directed at customer loyalty to identify both customers who have great loyalty to their preservation as well as customers with intentions to change to the competitors. This capability is necessary, especially for modern telecommunications operators. Nowadays companies face more complexity and competition in their business and need to develop innovative activities to capture and improve customer satisfaction and retention [2]. Growing profitability is the goal of most companies, to reach this goal, companies must provide an analysis of customer relationship management (CRM) and provide appropriate marketing strategies [3]. Some studies provided a new model of transactions based on both the services and customer satisfaction and showed that the price is not the only measure affecting customer buying decisions, but it is also important that both the customer and the company agree on product value and good customer services. Therefore, organizations should not seek to develop a product to satisfy their customers, but they must follow the customer purchasing behavior and offer distinct products for each segment. In other words, segmenting customers based on purchasing behavior is necessary to develop successful marketing strategies, which in turn cause the creation and maintenance of competitive advantage. Current methods of customer value analysis which are based on past customer behavior patterns or demographic variables are limited to predict future customer behavior. So, better patterns were exch

Research objectives

Our goals of this research

- Customers value was Analyzed by segmenting them according to the new approach TFM and then determine, the level of loyalty for each segment in a big data environment in telecom.
- A set of features was derived from the telecom data.
- The best behavioral features for customers with their demographic information were Chosen, based on these features and the level of loyalty for each segment, the following classification algorithms were applied and the classification models were built: random forest classifier, Decision tree classifier, Gradient-boosted tree classifier, and Multiplexer perceptron (MLPC).
- These models were evaluated based on several criteria that evaluated and selected the most accurate model.
- The loyalty rules were derived from this model, these rules showed the characteristics of each level of loyalty and thus the loyalty reasons were identified in each segment to target them in a representative manner. The other advantage of classification algorithms application was building a model to classify new users by loyalty.

Related works

Various efforts have been made to build an effective prediction model for retaining customers using different techniques. To better understand how Many studies have built their own predictive models suggested by Oladapo et al. [4]. Logistic regression model design, a good model of customer data to predict customer retention in

a telecommunications company with 95.5% accuracy. This model predicts customer retention based on billing, value-added services, and SMS service issues.

Aluri et al. [5] have focused on using machine learning to determine the value of customers in the hospitality sectors of customers, such as restaurants and hotels, by engaging dynamic customers with the loyalty program brand. Their results also show that automated learning processes excel in identifying customers with greater value in specific promotions. They have deepened the practical and theoretical understanding of automated learning in the value chain of customer loyalty, in a structure that uses a dynamic model for customer engagement.

Wiaya and Gersang [6] predict customer loyalty at the National Multimedia Company of Indonesia, using three data mining algorithms, to form a customer loyalty classification model, namely: C4.5, Naive Bayes and Nearest Neighbor. These algorithms were applied to the set of data contained 2269 records and 9 attributes to be used. By comparing the analysis models, the C4.5 algorithm with its own data set segment (80% for training data and 20% of test data) has the highest accuracy results of 81.02% compared to algorithms and other data segments. In the attribute analysis, the disconnection attributes (the attribute that is interpreted as the reason why customers have stopped) get the most influential attribute on the accuracy of the results in the data extraction process to predict customer loyalty. This article does not discuss the algorithms of features selection, methods of obtaining important features, and its impact on model accuracy.

Wong and Wei [7] presented a research to develop a tool to analyze customer behavior and predict their upcoming purchases from Air Travel Company. They provided an integration tool between data mining Pricing for competitors, customer segmentation and predictive analysis. Results In customer segmentation analysis, 110,840 clients are identified and segmented based on their purchasing behavior. Customers' profiles are split using a weighted RFM model, and customer purchasing behavior is analyzed in response to competitor price changes. The following destinations are expected for high-value customers identified using pre-link rules and custom packages promoted to targeted customer segments.

Moedjionom et al. [8] have predicted customer loyalty in a multimedia services company, offering many services to win the market. This research contribution is to use data related to the segmentation and splitting of potential customers based on the RFM model, then applying the classification, Proportion of accuracy in customer loyalty rating research. Although the C4.5 algorithm with the k-mean segmentation give a better result, there are some important action that can be added to the search: using optimization algorithm to select the features or to adjust the value of the label to obtain a more accurate model.

Kaya et al. [9] have built a predictive model based on spatial, temporal and optional behavioral features using individual transaction logs. Our results show that proposed dynamic behavioral models can predict change decisions much better than demographics-based features and that this effect remains constant across multiple data sets and different definitions of customer leakage. They have examined the relative importance of different behavioral features in predicting leakage, and how predictive power differed across different population groups.

Cheng and Sun [10] have viewed other application of the RFM model (named TFM) to identify high-value customers in the communications industry. Use three main features to describe users who have accumulated a greater amount of service time (T), often purchase 3G services (F) and create large amounts of invoices per month (M).

This study proposes a comprehensive CRM strategy framework that includes customer segmentation and behavior analysis, using a dataset that contains about 500 million (full dataset in syriatel company). Al Janabi and Razaq [11] used intelligent big data analysis to design smart predictors for customer churn in the telecommunication industry. The goal of this research maintain customers and improve the level of revenue. The proposed system consists of three basic pashas: First Phase: an understanding of the company's data. This phase focuses on the initial processing of data that is fragmented and unbalanced. They addressed the problem of imbalance by building the DSMOTE algorithm. Second Phase: construct a GBM-based predictor after it was developed, replace its decision-making part, which is (DT) with a (GA) algorithm. The impact of this is able to overcome DT problems and reduce time implementation. Third Stage: The accuracy of the predictor results was verified by using the matrix of the conflict matrix. A comparison was made between the traditional method of initial treatment, which is SMOTE, DSMOTE in terms of error rate and accuracy. GBM-GA method has higher Accuracy than GBM.

One of the biggest challenges of the current big data landscape is our inability to process vast amounts of information in a reasonable time. Reyes-Ortiz et al. [12] explored and compared two distributed computing frameworks implemented on commodity cluster architectures: MPI/OpenMP on Beowulf that is high-performance oriented and exploits multi-machine/multi- core infrastructures, and Apache Spark on Hadoop which targets iterative algorithms through in-memory computing. The Google Cloud Platform service was used to create virtual machine clusters, run the frameworks, and evaluate two supervised machine learning algorithms: KNN and Pegasos SVM. Results obtained from experiments with a particle physics data set show MPI/OpenMP outperforms Spark by more than one order of magnitude in terms of processing speed and provides more consistent performance. However, Spark shows better data management infrastructure and the possibility of dealing with other aspects such as node failure and data replication.

There are several studies in the field of communication that deal with predicting the age and gender of the customer in big data platform by analyzing their personal data, including the study presented by Zaubi [13]. Where he designed a model using a reliable data set of 18,000 users provided by SyriaTel Telecom Company, for training and testing. The model applied by using big data technology and achieved 85.6% accuracy in terms of user gender prediction and 65.5% of user age prediction. The main contribution of this work is the improvement in the accuracy in terms of user gender prediction and user age prediction based on mobile phone data and end-to-end solution that approaches customer data from multiple aspects in the telecom domain.

Other studies have also dealt with the prediction of customer churn in telecom using machine learning in big data platform, including the study presented by Ahmad [14]. The main contribution of his work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. The model

developed in this work uses machine learning techniques on big data platforms and builds a new way of features' engineering and selection. In order to measure the performance of the model, the Area Under Curve (AUC) standard measure is adopted, and the AUC value obtained is 93.3%. Another main contribution is to use customer social networks in the prediction model by extracting Social Network Analysis (SNA) features. The use of SNA enhanced the performance of the model from 84 to 93.3% against the AUC standard.

With regard to how some studies approached customer value analysis, retention, and loyalty. A study in [4] did not apply to big data as it studied all customers according to some features and using a method of machine learning (a logistic regression model) to show the role of machine education in retaining and increasing customer loyalty. In the study [5]. Machine learning was implemented in a major hospitality location and compared to traditional methods to determine customer value in the loyalty program. In the study [6] predict customer loyalty at the National Multimedia Company of Indonesia, using three data mining algorithms, These algorithms were applied to the set of data obtained are 2269 records and contain 9 attributes to be used. By comparing the analysis models, the C4.5 algorithm with its own data set segment has the highest accuracy results of 81.02% compared to algorithms and other data segments. In my study, a model is built to increase customer loyalty predictions based on the new TFM methodology and machine learning. My experiences were demonstrated that TFM most appropriate for the telecom sector than RFM. The concept of the TFM is adjusted, where T is the sum of the duration of calls and the periods of internet sessions during a certain period. The set of data obtained is 127 million records and contains 220 features to be used. Binary and multi-classification are applied. After comparing the classifiers, the Gradient-boosted-tree classifier was found to be the best in binary and Random Forest Classifier is the best in multi-classification.

Research tools

Hortonworks data platform (HDP)

An open-source framework for distributed storage and processing of large and multi-source datasets [15]. HDP enables flexible application deployment, machine learning, deep learning workloads, real-time data storage, security, and governance. It is a key element in the modern data structure of data (Fig. 1).

The HDP framework was custom- installed to obtain only the tools and systems required to track all stages of this work. these tools and systems were: a distributed file system [16], Hadoop HDFS[1] for data storage, Spark[2] implementation engine for data processing [17], Yarn for resource management, Zeppelin[3] as a development user interface,Ambari for system monitoring, Ranger for system security and (Flume[4] System and Scoop[5]) for data acquisition from Syriatel company data sources to HDFS in our dedicated framework.

[1] https://hadoop.apache.org/.

[2] https://spark.apache.org/.

[3] https://zeppelin.apache.org/.

[4] https://flume.apache.org.

[5] https://sqoop.apache.org/.

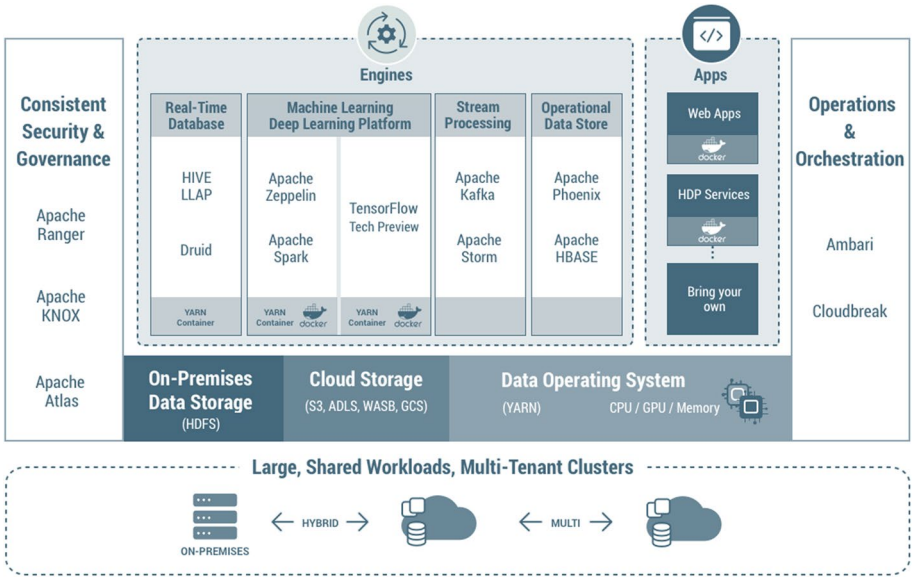

Fig. 1 Hortonworks data platform: HDP 3.1 [15]

Hive[6] is an ETL and data warehouse tool on top of the Hadoop ecosystem and used for processing structured and semi-structured data. Hive is a database present in the Hadoop ecosystem that performs DDL and DML operations, and it provides flexible query language such as HQL for better querying. Hive in Map reduce mode is used because data was distributed across multiple data, nodes to execute queries with better performance in a parallel way. The hardware resources were used included 12 nodes with 32 GB of RAM, 10 TB storage capacity, and 16-core processor per node. The Spark engine [17] was used in most phases of model building such as data processing, feature engineering, training and model testing because it is able to save its data in compute engine's memory (RAM) and also perform data processing over this data stored in memory, thus eliminating the need for a continuous Input/Output (I/O) of writing/reading data from disk. In addition, there are many other advantages. One of these advantages is that this engine contains a variety of libraries to implement all stages of the machine learning life cycle.

Syriatel data sources

Call details record of CDRs

Each time a call is made, a message is sent, the Internet is used, or an operation is performed on the network, the descriptive information is stored as a call details record (CDR). Table 1 illustrated some types of call logs, messages, and Internet details available in Syriatel that were used in this research to predict customer loyalty:

Rec: Call log, SMS message log, MMS Multimedia Message log MMS multimedia messaging log, DATA internet data usage log, Mon fee log,Vou recharge log, Mon monthly log information, web metadata information, EGGSK tab In roaming.

[6] https://hive.apache.org/.

Table 1 CDR sample fields in Syriatel company

Call type	GSM (A)	GSM (B)	Direction	Cell identifier	Duration	Date	...
Call	+963********8	+963********5	Out	C83	56 s	10/10/2018 23:30:26	...
Call	+963********5	+963********8	In	C203	56 s	10/10/2018 23:30:26	...
SMS	+963********9	+963********3	Out	C322	Null	10/10/2018 23:59:11	...
SMS	+963********3	+963********9	In	C164	Null	10/10/2018 23:59:11	...
...	...	...	...	...	...	...	...

Table 2 Example of customer's services in Syriatel company

GSM	Economy	Education	Health	Horoscopes	Duration	Sport	...
+963********9	0	1	0	0	1	0	...
+963********5	0	0	1	1	0	0	...
+963********8	1	0	0	0	0	0	...
+963********3	0	0	0	0	0	0	...
...	...	...	...	...	...	...	...

Detailed data stored in relational databases

The Call details record was linked to the customer detailed data stored in the relational databases using this GSM, which is detailed as follows: Customer Management Database, Customer Complaints Database, Towers Information Database, Towers Information Database, Mobile Phone Information Database.

Customer services

All services recorded by the client were collected and classified manually based on the type of service such as political related services News, sports news, horoscopes, etc. . ., these categories are treated as a customer Advantages. As a result, a customer service table is produced. Table 2 is a sample.

Customer contract information

Customer contract information was fetched from the CRM system, and contains basic customer information (gender, age, location . . .) and customer subscription information, as a single customer. You may have more than one subscription (two or more GSM networks) with different types of subscriptions: pre-subscription, prepay, 3G, 4G . . . subscription.

Database of cells and sites

Telecommunications companies related to location data and their components were stored in the relational database. This data was used to extract spatial features. Table 3 is a sample.

Demographics data for customers

Building such a predictive system requires a data containing the real demographics such as gender and age for each GSM, whatever the demographics of the GSM owner sometimes the real user and the GSM owner, not much. Table 4 is a sample.

Table 3 Sample of cells and sites database

Cell identifier	Site identifier	Longitude	Latitude	...
C147	S73	**.******2	**.******7	...
C23	S119	**.******0	**.******6	...
C64	S14	**.******1	**.******0	...
...	...	...	...	...

Table 4 Demographics data for customers

Age group	Year range	Percentage (%)
A	18–27	32
B	28–39	41
C	40–60	27

Extraction of features

The features were engineered and extracted based on our research and our experiment in the telecom domain. 223 features were extracted for each GSM. These features belong to 6 feature categories; each category provided with examples.

- *Segmentation Features T, F, M (3 features)*

  total of calls and Internet duration in a certain period of time (Fig 2).

  Frequency (F): use services frequently within a certain period (Fig. 3).

  Monetary (M): The money spent during a certain period (Fig. 4).

  Classification Features (220 features)
- *Individual Behavioral Features*
  Individual behavior can be defined as how an individual behaves with services.
  For example:
  Calls duration per day: calls duration per day for each GSM.
  Duration per day: calls and sessions duration per day for each GSM (Figs. 5, 6, 7).
  Entropy of duration
  High entropy means the data has high variance and thus contains a lot of information and/or noise.
  Daily outgoing calls: for each GSM the daily outgoing calls.
  Calls incoming daily night: for each GSM the daily outgoing calls at night (Fig. 8).
  SMS received daily at work time, · · · About (200 features).
- *Social behavior features*
  Is behavior among two or more organisms within the same species, and encompasses any behavior in which one member affects the other. This is due to an interaction among those members.
  Some examples about this features:

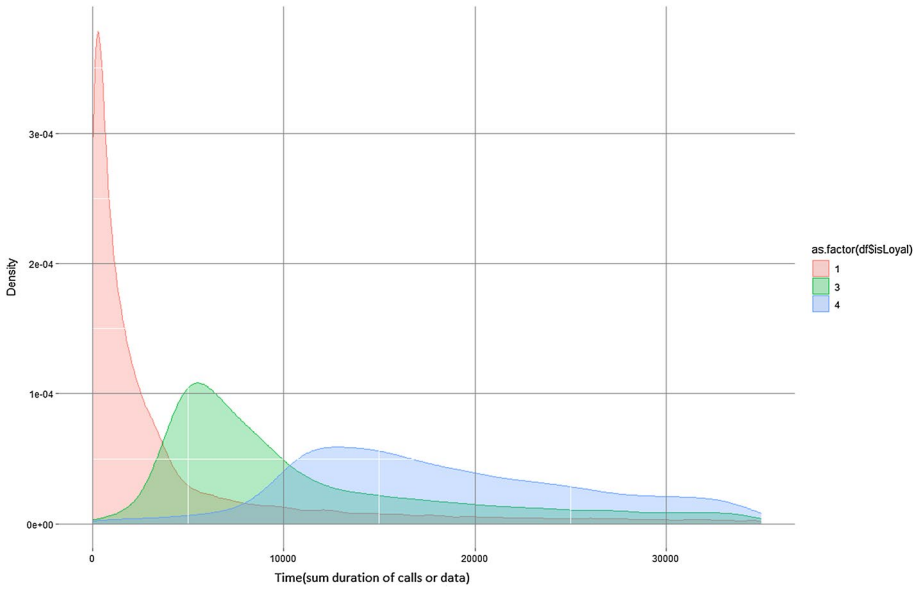

Fig. 2 Mass distribution for T in our Study

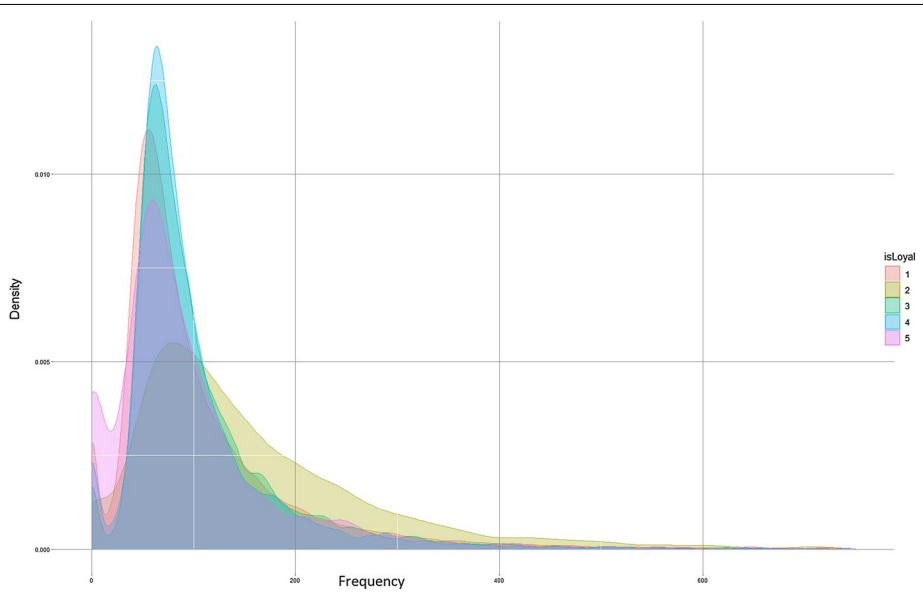

Fig. 3 Mass distribution for F in our Study

Number of contacts: for each customer number of contacts with other customers.
Transactions received per customer: number of calls, sms,Internet sessions received by each customer.
Transactions sent per contact, etc. (20 features).

- *spatial and navigation features*
  Features about the spatial navigation of customers

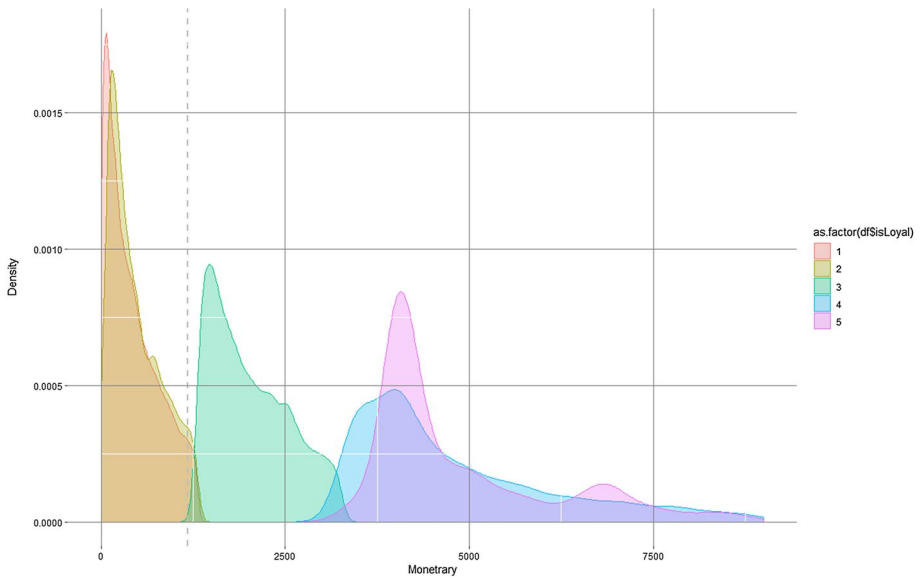

Fig. 4 Mass distribution for M in our Study

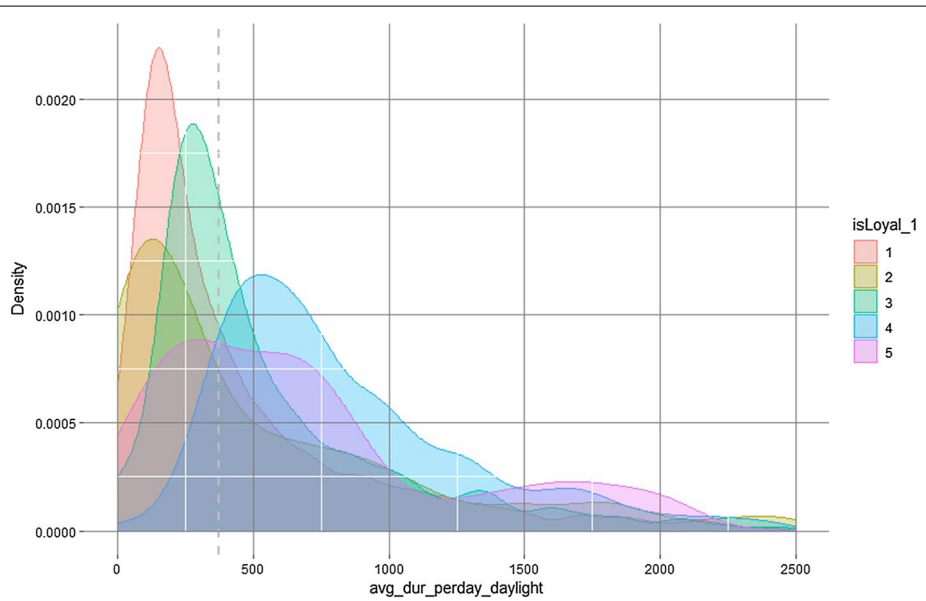

Fig. 5 Mass distribution for "Avg dur per day daylight"

holiday navigation: Customer Movements on holiday.
Home zone: location of customer home.
Antenna location: location of antenna.
Daytime antenna: antenna which are used by customer transactions in daytime.
Workday antenna: antenna which are used by customer transactions in workday.
Vacation antenna: antenna which used by customer transactions in vacation, . . ., etc.
Their number is about (21 features).

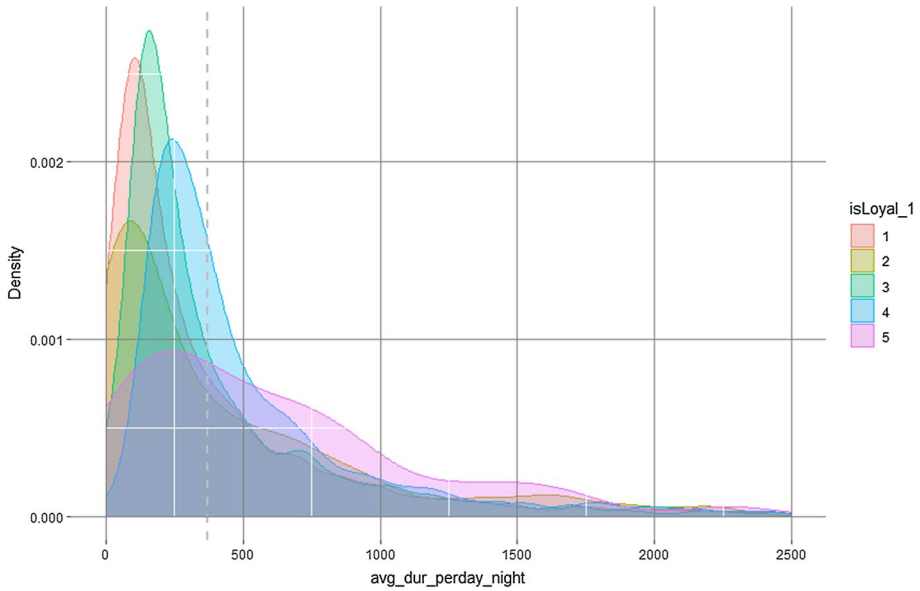

Fig. 6 Mass distribution for "Avg dur perday night"

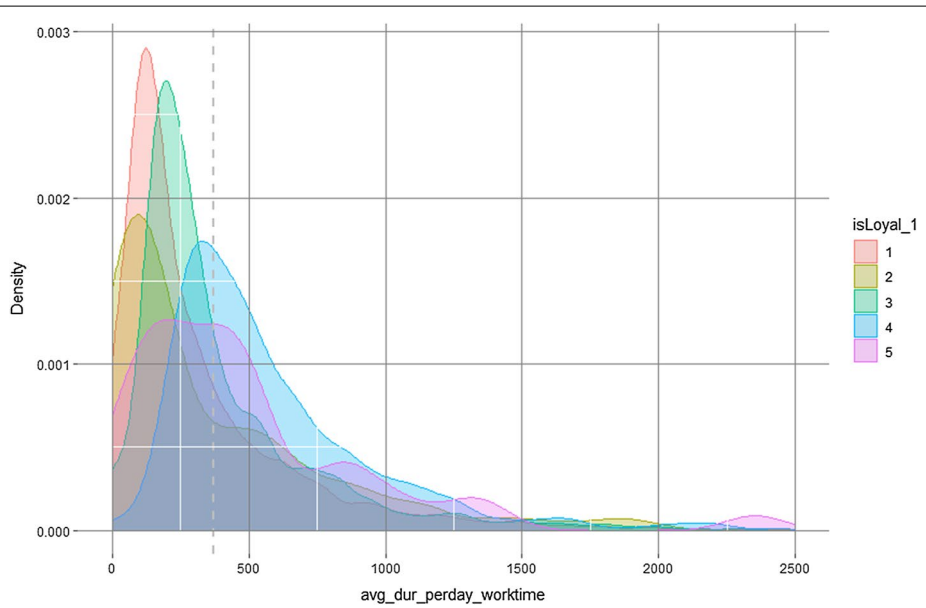

Fig. 7 Mass distribution for "Avg dur per day worktime"

- Timestamps for each working day (Sunday to Thursday), on holidays, or during the day (9 to 16) and at night: average number of SMS received per day (17 to 8) on holidays, etc. (165 features).
- *Types of services registered*
  technical news services, educational services, sports news services, political news services, entertainment services, etc., (13 features).
- *Contract information tariff type*
  GSM type, (2 features).

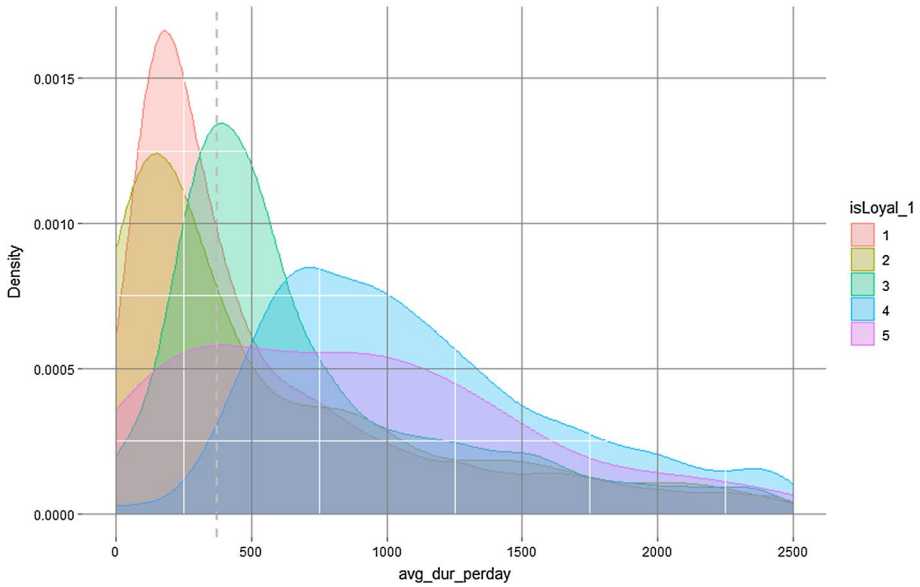

Fig. 8 Avg dur per call

The total number of features listed is 421, but there are about 201 features belonging to more than one category, so the total number of features is 220. Mass distribution for Some features with loyalty.

Features engineering-ways to choose features

Feature engineering is the process of using domain knowledge of data to create features that make machine learning algorithms work well. The most important reasons to use the selection of the features are:

- It enables machine learning algorithm to train faster.
- It reduces the complexity of the model and makes it easy to interpret.
- It improves the accuracy of the model if the correct subset is selected.
- It reduces overfitting.

Next, we'll discuss several methodologies and techniques that you can use to set your feature space and help your models perform better and more efficiently.

Attribute Selection Algorithms (Features)

Feature selection algorithms discovered and reported in the literature.

The feature selection algorithms are categorized into three categories such as filter model, embedding model (or aggregator) and embedded model according to mathematical models.

Filter model

It depends on the general characteristics of the data and the evaluation of features without involving any learning algorithm [13]. Filter model Algorithms are Relive F, Information Gain. Entropy (H) was used to calculate the homogeneity of a sample.

Information gain is decrease in entropy after dividing the data set by an attribute. Gain index, Chi-Squared, Gain Ratio [18].

Wrapper model

A predefined learning algorithm is required and its performance is used as a benchmark for evaluation and feature identification.

Embedded model

It chooses a set of features for training and building a model, then test its feature importance depending on the goal of learning model. You can get the feature importance of each feature of your dataset by using the feature importance property of the model. Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable. Feature importance is an inbuilt class that comes with Tree Based Classifiers.

The big data analytics is used technological advances are based on memory usage, data processing and scrutiny of big data include handling high volume data with decreasing cost of storage and CPU power, the effective usage of storage management for flexible computation and storage and Distributed computing systems through flexible parallel processing [19] with the development of new frameworks such as Hadoop. And also big data frameworks such as the Hadoop ecosystem, No SQL databases efficiently handle complex queries, analytics and finally extract, transform and load (ETL) which is complex in conventional data warehouses. These technological changes have shaped a number of improvements from conventional analytics and big data analytics.

The feature selection aims to select a feature subset that has the lowest dimension and also retains classification accuracy so as to optimize the specific evaluation criteria. feature selection methods can be filter and wrapper. Filter methods make a selection on the basis of the characteristics of the dataset itself. The selection may be fast but it leads to poor performance. Wrapper methods takes a subsequent classification algorithm as the evaluation index, it achieves high classification accuracy but results in very low efficiency for a large amount of computation. Backward feature elimination and Forward feature construction have been used. Feature selection methods such as backward elimination, Forward selection [19].

Backward elimination: In contrast to the forward selection strategy, the backward elimination strategy starts with the complete attribute set as initial subset and iteratively (and also heuristically) removes attributes from that subset, until no performance gain can be achieved by removing another attribute.

Forward selection initially uses only attribute subsets which exactly one attributes. Then additional attributes are added heuristically until there is no more performance gain by adding an attribute. Both of the methods are handled carefully Hadoop ecosystem.

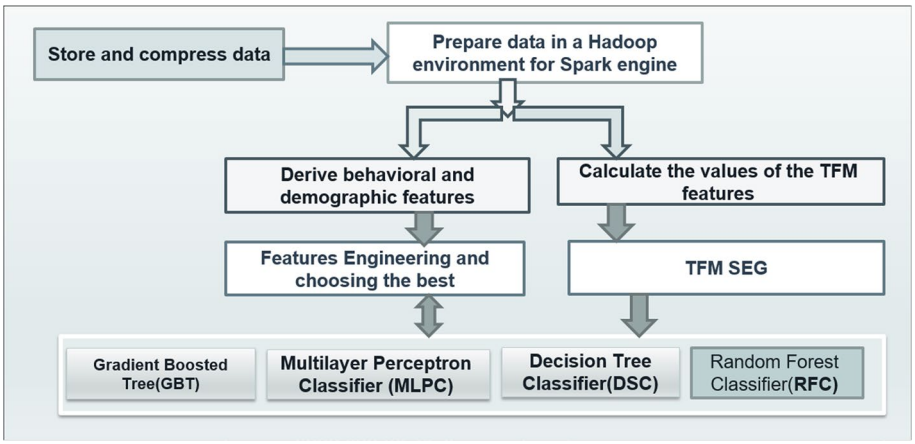

Fig. 9 Steps required for the proposed model

Implement methodology

The intended model was described to determine customer loyalty. This model relied on data mining and customer value analysis methods based on TFM to improve customer relationship management. Customers were divided using the Calculating TFM Score.

After the division of customers and determining the degree of loyalty in each department. The classification was performed based on descriptors expressing the level of loyalty in each segment and a set of behavioral and demographic features using several algorithms and evaluated their models to obtain the best model in terms of the highest accuracy. Steps required for the proposed model illustrated in Fig 9.

The loyalty features of these models were extracted to find out the causes of loyalty and the precise targeting of these segments.

Data preparation

- The data was collected from Syriatel data sources to the Hadoop environment.
- The Scala language was chosen to perform data preparation, attribute extraction, model training, and testing because it is the language in which the distributed implementation engine (spark) has developed. spark achieves a high-speed execution in addition to stability. In Spark, the library of automated learning provided by the Spark implementation engine, ML extension is used.
- The original data was divided into two sub-parts: the training group and the test group by 70/30, respectively.

Address missing and text values

Filling in the missing values with values of either zero or average of several nearby values was an advantage so that it enabled us to use the information in most attributes for the exercise. In this research, the following were applied:

- The attributes whose 70% at least include a missing value were deleted.
- The missing numerical values were replaced with the mean attribute itself.
- StringIndexer has been applied to String style attributes to convert them into numbers. This represents a common way to produce labels using StringIndexer, train a model with these pointers, and retrieve the original labels from the predicted pointers column using IndexToString. However, you are free to provide your own labels. Emphasis was placed on the development of the attribute preparation and attribute selection process.

Emphasis was placed on the development of the attribute preparation and attribute selection process.

Data processing and application of extraction and selection of attributes

- The T, F, M features were calculated for each customer, and the behavioral features were chosen.
- The most important attributes were chosen based on Chi-Squared function. this function was applied to groups of categorical characteristics and selected features to assess the probability of correlation or correlation between them using frequency distribution.

Compilation using calculating TFM

To calculate the TFM results, the three input parameters were divided into five subcategories. The Time, frequency and monetary scores were calculated and then combined, to obtain a comprehensive TFM analysis score.

Cumulative total duration (cumulative total duration) T

Time (T): total of calls and Internet sessions duration in a certain period of time.

The customers have been divided into different categories (Table 5). The cumulative total duration of service, for example, the total of calls and Internet sessions duration in 3 months (T).

Some research defines it as the average time you spend communicating or using application services in one month and in our study for three months. When the user's connection time is greater than the previously calculated average, the value of T is 2; otherwise, it is 1. In our study, There were 5 levels which were calculated after the maximum cumulative value of the total duration of calls and usage per GSM and the smallest value was calculated. Calculations were made and 5 levels of values T resulted. Where

Table 5 Assessment (cumulative total duration) T

T1	T2	T3	T4	T5
...	...	...	...	Greater ...
...	...	...	...	...

T1 was the first category with the lowest value, T2 was the second category, T3 was the third category, T4 was the category Fourth was high value, T5 Class V was very high.

Frequency

Frequency (F): use services frequently within a certain period

Customers were divided into different frequency categories (Table 6). Totalize the number of times he/she performed with the company (communication, message, internet access) during the past 3 months.

F1 represents clients who have performed less than or equal to 2 transactions in the last 3 months and F5 represents customers who have performed more than or equal to 11 transactions in the last 3 months.

In our study, there were 5 levels calculated after calculating the maximum cumulative value of the total number of calls and the number of uses per GSM and the smallest value was calculated. Calculations were made and 5 levels of values F resulted.

Monetary

Monetary (M): The money spent during a certain period. Customers were divided into different cash categories (Table 7) according to the total amount he/she paid for transactions with the organization over the past 3 months. M1 as clients paid less than or equal to 100 transactions in the last 3 months M5 customers who have paid greater than or equal to 10,000 transactions in the last 3 months.

In our study, There were 5 levels calculated after calculating the maximum cumulative value of the total cash mass of calls and the number of times of use per GSM and the smallest value was calculated. Calculations were made and 5 levels of values M resulted.

Based on the TF results calculated above (Table 8), we calculate the TFM results (Table 9).

Segment and target customers

Customer categories

Customer segmentation involves splitting the customer base into different subsets. A specific subsets with the same interest and spending habits [20]. Based on the TFM results as calculated above, customers can be divided into five parts:

Table 6 Frequency

F1	F2	F3	F4	F5
Greater 2	3–4	4–6	6–10	Greater 11
...	...	...	...	...

Table 7 Monetary

M1	M2	M3	M4	M5
Greater 100	100–1000	1000–50,000	50,000–10,000	Greater 10,000
...	...	...	...	...

Table 8 Assessment criteria for TF score (time–frequency)

Frequency						
...	**Rank**	**F1**	**F2**	**F3**	**F4**	**F5**
	T1	T1F1	T1F2	T1F3	T1F4	T1F5
	T2	T2F1	T2F2	T2F3	T2F4	T2F5
Time	T3	T3F1	T3F2	T3F3	T3F4	T3F5
	T4	T4F1	T4F2	T4F3	T4F4	T4F5
	T5	T5F1	T5F2	T5F3	T5F4	T5F5
...	...	...	...	...	...	

- Very high value customers (greater loyalty) These are the customers who make the highest profit for the operator. Without them the operator will lose its market share and competitive advantage. These customers are given appropriate care and attention from the operator.
- High value customers (great loyalty) These are the customers who make the highest profit for the operator. Without them, the operator will lose its market share and competitive advantage. These customers are given appropriate care and attention from the operator.
- Medium value customers (average loyalty) These are the customers who make medium profitability.
- Low value customers (little loyalty) These are the clients who make very little profit.
- Customer churn from the company (very little loyalty).

Table 9 Assessment criteria for TFM score

	Frequency					
.....	Rank	M1	M2	M3	M4	M5
		Low Loyalty		Medium Loyalty		Very high loyalty
	T5F5	T5F5M1	T5F5M2	T5F5M3	T5F5M4	T5F5M5
	T5F4	T5F4M1	T5F4M2	T5F4M3	T5F4M4	T5F4M5
	T5F3	T5F3M1	T5F3M2	T5F3M3	T5F3M4	T5F3M5
	T5F2	T5F2M1	T5F2M2	T5F2M3	T5F2M4	T5F2M5
	T5F1	T5F1M1	T5F1M2	T5F1M3	T5F1M4	T5F1M5
	T5F1	T5F1M1	T5F1M2	T5F1M3	T5F1M4	T5F1M5
	T4F5	T4F5M1	T4F5M2	T4F5M3	T4F5M4	T4F5M5
	T4F4	T4F4M1	T4F4M2	T4F4M3	T4F4M4	T4F4M5
	T4F3	T4F3M1	T4F3M2	T4F3M3	T4F3M4	T4F3M5
	T4F2	T4F2M1	T4F2M2	T4F2M3	T4F2M4	T4F2M5
Time	T4F1	T4F1M1	T4F1M2	T4F1M3	T4F1M4	T4F1M5
						high loyalty
	T3F5	T3F5M1	T3F5M2	T3F5M3	T3F5M4	T3F5M5
	T3F4	T3F4M1	T3F4M2	T3F4M3	T3F4M4	T3F4M5
	T3F3	T3F3M1	T3F3M2	T3F3M3	T3F3M4	T3F3M5
	T3F2	T3F2M1	T3F2M2	T3F2M3	T3F2M4	T3F2M5
	T3F1	T3F1M1	T3F1M2	T3F1M3	T3F1M4	T3F1M5
		Very Low Loyalty				
	T2F5	T2F5M1	T2F5M2	T2F5M3	T2F5M4	T2F5M5
	T2F4	T2F4M1	T2F4M2	T2F4M3	T2F4M4	T2F4M5
	T2F3	T2F3M1	T2F3M2	T2F3M3	T2F3M4	T2F3M5
	T2F2	T2F2M1	T2F2M2	T2F2M3	T2F2M4	T2F2M5
	T2F1	T2F1M1	T2F1M2	T2F1M3	T2F1M4	T2F1M5
	T1F5	T1F5M1	T1F5M2	T1F5M3	T1F5M4	T1F5M5
	T1F4	T1F4M1	T1F4M2	T1F4M3	T1F4M4	T1F4M5
	T1F3	T1F3M1	T1F3M2	T1F3M3	T1F3M4	T1F3M5
	T1F2	T1F2M1	T1F2M2	T1F2M3	T1F2M4	T1F2M5
	T1F1	T1F1M1	T1F1M2	T1F1M3	T1F1M4	T1F1M5
...	....	....	....	...		
	Very Low Loyalty	Low Loyalty		Medium Loyalty	high loyalty	Very high loyalty.

Customers who have the least loyalty are those who have left the company or are about to leave. Endeavors were taken to prevent them from leaving, and if they leave the company, the cost of customer service will be calculated. The total cost (1) associated with these potential customers if they stop their relationship Total cost (customer leakage) = lost revenue Marketing cost (1) Lost revenue is the revenue that these customers can make if they do not cease their relationship with the operator. The cost of marketing is the cost associated with replacing these customers with new customers.

Target customers

By calculating the TFM score, individually the status of the total time spent on calls, sms and the total Internet data, high-value customers were recognized as well as potential customers to leave the company. Today, however, most people have access to a range of telecommunications operator services that include both the price of the connection and the amount of Internet data. The sudden exceptions to this were people who used only the operator's services for the Internet or calls and not both. Therefore, to target customers based on these things, considerations about both the TFM score for the total call time, SMS and the amount of Internet data were taken (Tables 10, 11).

In both TFM score for total call time and number of messages.

If TFM is High, then Customers who use large Internet data and spend a lot of time on calls and send a large number of messages. Average customers who use Internet data, spend time on calls and average messages. If TFM is low, then Customers use less Internet data and spend less on calls and fewer messages. Large users can be targeted using loyalty points and personalized offers tailored to them, as they are the key to the competitive operator advantage in the marketplace, for medium users who can make offers based on a combination of both call rates and data bundle and for low user operators can deliver Offers tend to make these users use more of the services provided by operators. for example, free local calls to the same telephone numbers of the operator late at night,

Table 10 TFM score for the total calls duration and the total amount of Internet data/SMS with (high, low) loyalty

	Rank	**Internet data/sms**	
		+	−
Total Time of call	+	High	Mid
	-	Mid	Low

Table 11 TFM score for the total call durations and the total amount of Internet data/sms with multi-level loyalty

	Internet data/sms				
	Rank	++	+	−	−−
Total time of call					
−	Very-high	High	Mid	Mid	Mid
	Mid		Mid	Low	Very low

etc. For Potential Customers churn, TFM points were integrated with unstructured data such as social media data and call center feedback data to accurately predict them. Customers who were likely to churn must be taken seriously into consideration by telecom operators because of the cost of revenue and marketing associated with each. Therefore, the telecom operator must be made offers such as free talk time, free packet data for a specified period and an additional number of messages, for example, 200 MB of data for 3 days, to retain it.

Results and discussion

Apply classification algorithms

Having segmented using grades and recognizing loyalty for each segment, at this stage, the causes of loyalty were needed, i.e. The behavioral features of customers in each segment. The behavioral 220 features were taken and the descriptions resulting from the segmentation process as an input to the classification algorithms to identify the causes of loyalty and to identify the influential features at each level of loyalty. The other benefit of applying classification algorithms was to build an accurate predictive model for classifying new users by loyalty. Multiple and binary classifiers were built and the results were compared using different criteria. It is the highest accuracy classifier that gave us the best correlation between behavioral features and loyalty categories and gave the best behavioral features that were described categories (classes) and thus assist in decision-making in building marketing presentations for each category thus increasing the company's profit.

Performance measurement

The correlation matrix shown in Table 12 contains information on the actual and predicted classifications made by the binary classification system where (Loyal 1, Not Loyal 0). Each term corresponds to a specific situation as follows:

- True Positive (TP) is expressed as an example when the prediction is yes (the customer has loyalty to the company), and the truth has loyalty to the company.
- True negative (TN): When the prediction is no (no customer loyalty to the company), in fact the customer has no loyalty to the company.
- False Positive (FP): The prediction is yes (the customer has loyalty to the company), but the customer left the company is also known as "Type 1 error".
- D False negative (FN): When the prediction is not (the customer has no loyalty to the company), but the customer has loyalty and did not actually leave the company. Also known as "Type 2 error".

Some performance measures can be calculated directly from the confusion matrix [21].

Table 12 TFM score for total call time and number of messages

	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

$$Rcall\ (TruePositiveRate\ (TPR)) = \frac{T_P}{T_P + F_N} \tag{1}$$

$$Precision\ (PositivePredictiveValue) = \frac{T_P}{T_P + F_P} \tag{2}$$

$$FalsePositiveRate\ (FPR) = \frac{F_P}{F_P + T_N} \tag{3}$$

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{4}$$

TPR is also known as recall or allergy.

The accuracy standard does not rate the rate of correctly classified cases from both categories. It is expressed by the following equation:

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{5}$$

Area under the curve (AUC): measures the effectiveness of the work can be calculated by [21]:

$$AUC = \int_0^1 TPR(x)dx \tag{6}$$

$$TPR = \frac{T_p}{T_p + F_n} \tag{7}$$

F1-measure: harmonic mean of the precision and recall. Can be calculated by:

$$F1\text{-}measure = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{8}$$

Compare binary classifiers

Confusion matrix

An example of the confusion matrix for the Multilayer Perceptron Classifier algorithm (Tables 13, 14).

Loyalty categories (loyal 1, not loyal 0)

After comparing the classifiers, Gradient-boosted-tree classifier was found to be the best.

Table 13 Confusion matrix for multilayer perceptron classifier

	Positive (1)	Negative (0)
Positive (1)	TP/1777.0	FP/108.0
Negative (0)	FN /261.0	TN/145.0

Table 14 Comparison of binary classes

Algorithm	Accuracy	Precision	Recall	areaUnderROC	F1-score
Multilayer Perceptron Classifier (MLPC) Confusion matrix 1777.0 108.0 261.0 145.0	0.83	Precision (0.0) = 0.92 Precision (1.0) = 0.64	Recall (0.0) = 0.93 Recall (1.0) = 0.61	0.64	F1-score (0.0) = 0.93 F1-score (1.0) = 0.62
Decision Tree Classifier (DTC) Confusion matrix 1757.0 134.0 151.0 234.0	0.87	Precision (0.0) = 0.92 Precision (1.0) = 0.63	Recall (0.0) = 0.92 Recall (1.0) = 0.60	0.76	F1-score (0.0) = 0.92 F1-score (1.0) = 0.62
Random Forest Classifier (RFC) Confusion matrix 1817.0 68.0 164.0 226.0	0.87	Precision (0.0) = 0.91 Precision (1.0) = 0.76	Recall (0.0) = 0.96 Recall (1.0) = 0.57	0.77	F1-score (0.0) = 0.93 F1-score (1.0) = 0.66
Gradient-Boosted-Tree (GBT) Confusion matrix 1796.0 83.0 137.0 265.0	0.87	Precision (0.0) = 0.92 Precision (1.0) = 0.76	Recall (0.0) = 0.95 Recall (1.0) = 0.65	0.80	F1-score (0.0) = 0.94 F1-score (1.0) = 0.70

Table 15 Example of confusion matrix for a multiple classes

	A	B	C	D
A	100	80	10	10
B	0	9	0	1
C	0	1	8	1
D	0	1	0	9

Comparison of multiple classes

Example of confusion matrix for binary Classes

Avrage recall = [R(A) + R(B) + R(C) + R(D)]/4 = 0.775: 4, number

Recall calculation from the correlation matrix of model TP: 100, FN:100 R(A) = 100/200 TP:9, FN: 1 R(B) = 9/10 TP:8, FN:2 R(C) = 8/10 TP:9, FN:1 R(D) = 9/10

Recall = TP/(TP + FN) Table 15.

Multi-classification (1, 2, 3, 4, 5)

It reflects the loyalty levels wherethe 5 is very high loyalty, 4 is high loyalty, 3 is medium loyalty, 2 is low loyalty, 1 is very low loyalty.

Note 1: Multilayer Perceptron Classifier Input: 220 feature with 4 layers, 5 node in each layer Output: 5 classes

Note 2: Gradient-boosted tree classifier currently only supports binary classification (Table 16).

After comparing the Classification algorithms, it turns out that Random Forest Classifier is the best. An example of the distinctive features of each level of loyalty derived from the binary classification model (Table 17):

Table 16 Comparison of multiple classes to predict loyalty

Algorithm	Accuracy	Precision	Recall	F1-score	Weighted
MLPC Confusion matrix 1246.0 0.0 0.0 0.0 0.0 97.0 0.0 0.0 0.0 0.0 616.0 0.0 0.0 0.0 0.0 334.0 0.0 0.0 0.0 0.0 17.0 0.0 0.0 0.0 0.0	0.55	Precision (1.0) = 0.53 Precision (2.0) = 0.0 Precision (3.0) = 0.0 Precision (4.0) = 0.0 Precision (5.0) = 0.50	Recall (1.0) = 1.0 Recall (2.0) = 0.0 Recall (3.0) = 0.0 Recall (4.0) = 0.0 Recall (4.0) = 0.50	F1-score (1.0) = 0.70 F1-score (2.0) = 0.0 F1-score (3.0) = 0.0 F1-score (4.0) = 0.0 F1-score (5.0) = 0.0	Weighted precision: 0.26 Weighted recall: 0.51 WeightedF1 score: 0.34 Weighted false positive rate: 0.51
DTC Confusion matrix 995.0 8.0 130.0 66.0 0.0 74.0 13.0 4.0 2.0 0.0 211.0 7.0 297.0 97.0 0.0 54.0 0.0 70.0 211.0 0.0 8.0 0.0 2.0 2.0 0.0	0.67	Precision (1.0) = 0.74 Precision (2.0) = 0.46 Precision (3.0) = 0.59 Precision (4.0) = 0.55 Precision (5.0) = 0.59	Recall (1.0) = 0.82 Recall (2.0) = 0.13 Recall (3.0) = 0.48 Recall (4.0) = 0.62 Recall (5.0) = 0.60	F1-score (1.0) = 0.78 F1-score (2.0) = 0.21 F1-score (3.0) = 0.53 F1-score (4.0) = 0.59 F1-score (5.0) = 0.60	Precision: 0.65 Recall: 0.53 F1 score: 0.65 False positive rate:0.22
RFC Confusion matrix 1073.0 2.0 80.0 44.0 0.0 73.0 12.0 7.0 1.0 0.0 185.0 3.0 341.0 83.0 0.0 40.0 0.0 46.0 249.0 0.0 1.0 0.0 6.0 5.0 0.0	0.74	Precision (1.0) = 0.74 Precision (2.0) = 0.70 Precision (3.0) = 0.71 Precision (4.0) = 0.65 Precision (5.0) = 0.74	Recall (1.0) = 0.89 Recall (2.0) = 0.12 Recall (3.0) = 0.55 Recall (4.0) = 0.74 Recall (5.0) = 0.74	F1-score (1.0) = 0.83 F1-score (2.0) = 0.21 F1-score (3.0) = 0.62 F1-score (4.0) = 0.69 F1-score (5.0) = 0.66	Precision: 0.73 Recall: 0.74 F1 score: 0.72 False positive rate:0.18

Table 17 Results for gender prediction

Model tree (rules)	Features

If (feature 79 $\leq$ 3.3846153846153846)
If (feature 103 $\leq$ 2.2285714285714286)
If (feature 214 $\leq$ 2093.0)
If (feature 105 $\leq$ 0.17407765595569782)
If (feature 78 $\leq$ 3.0344827586206895)
Predict: 0.0 Else (feature 78 > 3.0344827586206895)
Predict: 1.0 Else (feature 105 > 0.17407765595569782)
If (feature 184 $\leq$ 2518.617404647386) Predict: 0.0
	Feature79 = std dur per day holiday out
	Feature103 = std dur per call work time out
	Feature214 = avg trans per cnt
	Feature105 = avg dur per callworkday
	Feature78 = std dur per day holiday

Conclusion

TFM segmentation and setting loyalty levels were been relied on. The classification algorithms were applied based on the loyalty levels as classification categories and the selected attributes, compared the results and selected the best classification model in terms of accuracy. Then the rules of loyalty prediction were derived from this model, which expressed the correlation of behavioral features with classification categories and thus known the causes of loyalty in each segment. target customers were optimized with appropriate offers and services. The other benefit of applying the classification algorithms was to build an accurate predictive model for classifying new users by loyalty.

Abbreviations
GSM: Global system for mobile communications; CDR: Call detail record; CRM: Customer relationship management; SMS: Short message service; QoE: Quality of experience; HDFS: Hadoop distributed file system; LDA: Linear discriminant analysis; XGBoost: Extreme gradient boosting; GBM: Gradient boosting machine; Bagged CART: Bagging classification and regression trees; AUC: Area under the curve; avg: Average; std: Standard deviation; in: Received by the customer; send by the customer; T_p: *True positive*; T_n: *True negative*; F_n: *False negative*; P: *True positive* + *false negative*; N: *True negative* + *false positive*; *TPR*: *True positive rate*.

Acknowledgements
This research was sponsored by SyriaTel telecom Co. We thank our colleagues Mem. Mjida (SyriaTel CEO), Mr. Adham Troudi (Big Data manager) who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. Also, the authors thank Housam Wassouf, Hazem S Nassar, Majdi Msallam, Mahmoud Eissa, Nasser abo saleh, Mustafa Mustafa for great ideas, help with the data processing and their useful discussions. the authors thank Moral help Rama Saleh, Marita wassouf, Amal Abbas, Nazier wassouf, weaam wassouf, walaa wassouf, Rawad wassouf.

Authors' contributions
WNW-W took on the main role so he performed the literature review, implemented the proposed model, conducted the experiments and wrote the manuscript. RA and KS took on a supervisory role and oversaw the completion of the work. All authors read and approved the final manuscript.

Funding
The authors declare that they have no funding.

Availability of data and materials
The data that support the findings of this study are available from SyriaTel Telecom Company but restrictions apply to the availability of this data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of SyriaTel Telecom Company.

Ethics approval and consent to participate
The authors Ethics approval and consent to participate.

Consent for publication
The authors consent for publication.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Faculty of Information Technology-Department of software engineering and information systems, Al-Baath University, Homs, Syria. [2] Faculty of Applied Sciences, Hama University, Hama, Syria. [3] Faculty of Information Technology, Higher Institute for Applied Sciences and Technology, Damascus, Syria.

Received: 19 October 2019 Accepted: 17 February 2020
Published online: 23 April 2020

References
1. Saini N, Monika Garg K. Churn prediction in telecommunication industry using decision tree. 2017
2. Qiasi R, Baqeri-Dehnavi M, Minaei-Bidgoli B, Amooee G. Developing a model for measuring customer's loyalty and value with rfm technique and clustering algorithms. J Math Comput Sci. 2012;4 (2):172–81.
3. Kim S-Y, Jung T-S, Suh E-H, Hwang H-S. Customer segmentation and strategy development based on customer lifetime value: a case study. Expert Syst Appl. 2006;31 (1):101–7.
4. Oladapo K, Omotosho O, Adeduro O. Predictive analytics for increased loyalty and customer retention in telecommunication industry. Int J Comput Appl. 2018;975:8887.
5. Aluri A, Price BS, McIntyre NH. Using machine learning to cocreate value through dynamic customer engagement in a brand loyalty program. J Hosp Tour Res. 2019;43 (1):78–100.
6. Wijaya A, Girsang AS. Use of data mining for prediction of customer loyalty. CommIT J. 2015;10 (1):41–7.
7. Wong E, Wei Y. Customer online shopping experience data analytics: integrated customer segmentation and customised services prediction model. Int J Retail Distrib Manag. 2018;46 (4):406–20.
8. Moedjiono S, Isak YR, Kusdaryono A. Customer loyalty prediction in multimedia service provider company with k-means segmentation and c4. 5 algorithm. In: 2016 international conference on informatics and computing (ICIC), IEEE. 2016:210–5.
9. Kaya E, Dong X, Suhara Y, Balcisoy S, Bozkaya B, et al. Behavioral attributes and financial churn prediction. EPJ Data Sci. 2018;7 (1):41.
10. Cheng L-C, Sun L-M. Exploring consumer adoption of new services by analyzing the behavior of 3G subscribers: an empirical case study. Elect Comm Res Appl. 2012;11 (2):89–100.
11. Janabi S, Razaq F. Intelligent big data analysis to design smart predictor for customer churn in telecommunication industry. In: Farhaoui Y, Moussaid L, editors. Big data and smart digital environment. Cham: Springer; 2019. p. 246–72.
12. Reyes-Ortiz JL, Oneto L, Anguita D. Big data analytics in the cloud: spark on hadoop vs mpi/openmp on beowulf. Proc Comput Sci. 2015;53:121–30.

13. Al-Zuabi IM, Jafar A, Aljoumaa K. Predicting customer's gender and age depending on mobile phone data. J Big Data. 2019;6 (1):18.
14. Ahmad AK, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6 (1):28. https://doi.org/10.1186/s40537-019-0191-6.
15. Hortonworks Data Platform (HDP) Kernel Description. https://www.cloudera.com/products/hdp.htm. Accessed 2019 Cloudera.
16. Shvachko K, Kuang H, Radia S, Chansler R, et al. The hadoop distributed file system. MSST. 2010;10:1–10.
17. Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I. Spark: cluster computing with working sets. HotCloud. 2010;10 (10–10):95.
18. El-Hasnony IM, El Bakry HM, Saleh AA. Comparative study among data reduction techniques over classification accuracy. Int J Comput Appl. 2015;122:2.
19. Seelammal C, Devi KV. Hadoop based feature selection and decision making models on big data. Middle-East J Sci Res. 2017;25 (3):660–5.
20. Singh I, Singh S. Framework for targeting high value customers and potential churn customers in telecom using big data analytics. Int J Educ Manag Eng. 2017;7 (1):36–45.
21. Bradley AP. The use of the area under the roc curve in the evaluation of machine learning algorithms. Patt Recogn. 1997;30 (7):1145–59.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com