Damascus University Journal for engineering sciences
Vol 40 No.1 (2024): 38-53

مجلة جامعة دمشق للعلوم الهندسية
المجلد 40– العدد 1 (2024) 38–53

Enhancing the methods of customer behavior analysis to increase customer satisfaction. Case study: Syriatel Telecom Company

Hala Alnemeh[1], Prof. Dr. Rakan Razouk[2]

[1]Master's student - Faculty of Informatics Engineering - Damascus University.
[2]Doctor - Faculty of Informatics Engineering - Damascus University.

Received: 9/6/2022
Accepted: 11/8/2022

Abstract

The telecommunication industry is in the strongest competition ever, as this sector gets disrupted by new arising competitors with high technical infrastructure as 5G networks. However, the current customer satisfaction measures are based on subjective questionnaires without utilizing the vast amount of objective network KPIs and telecom systems data into account. This work presents a model that tackles this lack of research and provides a high impact solution to survive in the tough competition of the telecom industry. The paper addresses two fundamental questions: 1) To what extent satisfied/dissatisfied customers can be classified based on telecom systems data that was produced during users' interactions? 2) Can satisfaction indicators be derived from telecom systems data? this study discusses a machine learning problem, and compare 7 classifiers and analyze data for 10,000 real users from the Syrian telecom company Syriatel. 120 extracted features were drawn from the most significant available sources: billing, network, customer service system, and customer demography data. The best result for customer satisfaction classification was 87%, achieved with XGBOOST classifier. Furthermore, the paper identifies the most 9 potential indicators for satisfaction. Our goal was to classify customer satisfaction/dissatisfaction based on the objective data that is generated from every service interaction on the network or customer care centers.

Key Words: Customer satisfaction prediction, Classification models, Machine learning, Telecom industry, Quality of service, Quality of experience.

تحسين طرق تحليل السلوك بغرض زيادة رضا الزبائن دراسة حالة شركة سيرتل

حلا النعمة، أ.د. راكان رزوق

طالبة ماجستير – كلية الهندسة المعلوماتية– جامعة دمشق.

دكتور – كلية الهندسة المعلوماتية– جامعة دمشق.

الملخص

يتميز قطاع الاتصالات بالمنافسة الشديدة بين الشركات التي تسعى لتقديم أفضل التقانات والخدمات في سعي حثيث للتفوق على المنافسين. يُعتمد في تحديد رضا العملاء في الغالب على الاستبيانات الشخصية دون دراسة العدد الكبير من المؤشرات والبيانات المتراكمة في أنظمة الاتصالات.

نسعى في هذا العمل إلى إيجاد حل عملي لدراسة رضا العملاء، حيث يتناول العمل الأسئلة التالية:

1) إلى أيّ مدى يمكن تصنيف العملاء إلى راضين / غير راضين بناءً على بيانات أنظمة الاتصالات التي يتم إنتاجها أثناء تفاعلات المستخدمين؟

2) هل يمكن اشتقاق مؤشرات الرضا من بيانات أنظمة الاتصالات؟

تناقش هذه الدراسة مسألة تعلم بالاعتماد على الآلة، حيث تم مقارنة نتائج سبع مصنّفات وتحليل البيانات لعشرة آلاف مستخدم من شركة الاتصالات السورية سيرتل.

قادت الدراسة إلى استخلاص 120 ميزة من أهم المصادر المتاحة: نظام الفوترة، مؤشرات أداء الشبكة، أنظمة خدمة العملاء، وبيانات العملاء. كانت أفضل نتيجة لتصنيف رضا العملاء هي 87٪ التي تم تحقيقها باستخدام مصنف XGBOOST. كما تم تحديد أكثر تسعة مؤشرات تأثيراً على الرضا، حيث كان الهدف الأساسي هو تصنيف رضا العملاء / عدم رضاهم بناءً على البيانات الموضوعية لأنظمة الاتصالات دون العودة إلى العميل وسؤاله بشكل مباشر.

الكلمات الرئيسية: توقع رضا العملاء، نماذج التصنيف، التعلم الآلي، صناعة الاتصالات، جودة الخدمة، جودة التجربة

تاريخ الإيداع: 2022/6/9

تاريخ القبول: 2022/8/11

Introduction

Telecommunication is a communication over distance by cable, telephone, telegraph or broadcast among others, but traditionally it refers to telephone services. Telecommunication also can be known as a transmission of information in order to provide communication between multiple parties.

Nowadays, Telecommunication advances are concurred with the growth of communication technology and mobile devices. The users can stay connected on a global scale by just having good telecommunication services.

Telecommunication, especially via mobile devices, has become a part of everyone's life, not only for making calls and messaging via text messages and multimedia, but also for connecting to the Internet, which increased the demand for mobile phone services continuously. The use of these technologies may cause the user to be unsatisfied with the provided service if it does not meet his/her needs, and therefore the dissatisfied customer will be inclined to switch the service provider to a better telecom operator service. [1]

[2] Showed that organizations are becoming more customer-centric and focusing on retaining existing customers rather than acquiring new ones, since the costs of attracting a new customer is higher than retaining an existing one. [1] also concluded that it is more profitable retaining old customers who are more likely to repurchase or reuse company's products or services and recommend them to others.

Promotion schemes advertised to attract customers or make them switch to a new service provider by targeting their expectations made customer satisfaction a main pillar in these schemes, especially when there are many competitors [2]. The competition and advancement in information communication technology exert a lot of pressure on Telecommunication companies to be customer-centric and provide continuous service improvement as a way to ensure customer satisfaction. [3].

In summary, the need to understand the factors leading to customer satisfaction has become an important research task not only to researchers but also to company executives. [3]

Therefore, the main objective this research focuses on is to study and understand the needs of the Syrian telecom operator Syriatel's customers, to identify the factors that may affect their satisfaction with the various services Syriatel provides, and to investigate the factors that can become the customer satisfaction index in the Syrian telecom industry.

1. Research objectives

Understanding customers in depth and distinguishing their behavior patterns and preferences is a great wealth that telecommunications companies aim to reach, and is an essential pillar in obtaining customer satisfaction and loyalty. This knowledge enables companies to provide what customers need or prefer without explicitly asking for it. It also enables to identify weakness points in customers' experience in order to fix it and provide them with services that fulfill all their needs.

The objective of this research is to:

- study and analyze the methods used to measure customer satisfaction within telecommunications sector.
- study and analyze customers' behavior and their experiences in the Syrian telecom company, Syriatel.
- develop a framework for modeling and processing different customer data to obtain value-added knowledge that contributes to determining which factors have the biggest impact on customer satisfaction.

Most studies determining customer satisfaction by relying on satisfaction characteristics are based on the opinions of customers directly (via questionnaires or phone calls), or by analyzing submitted complaints, or provided quality of service.

This paper aims to predict customer satisfaction by relying on the quality of services provided to each customer in a comprehensive way, by covering and measuring all the indicators that may affect his/her satisfactory level during each interaction, whether activating an offer or service, recharging, payment, service center call or visit, and network indicators when customer made a call or establish an internet session. Customer demography was also extracted (age group, geographic location, gender,

etc.) to take into account the individuality of the experience for each customer, for more accurate customer experience measurement. This research results can contribute in taking both business and technical decisions in easier and more accurate way.

2. Related works

Previous researches in various fields such as marketing, communications and data science gave great importance to the analysis and measurement of customer satisfaction. Satisfaction can be defined as the features or characteristics that can satisfy the need or desire of the consumer in a better way than the competitors. The aim of these researches was to measure participants' satisfaction and to identify the factors influencing customer satisfaction the most.

2.1 Studies based on subjective methods

Traditional studies measure satisfaction degree by asking the customer directly through questionnaires or via direct contact (subjective way). Study [4] identified six main hypotheses responsible for customer satisfaction in the telecommunications sector, which are shown in Figure 1.

Figure (1) Customer Satisfaction Indicators

A structured questionnaire was distributed randomly to a sample of students. To analyze the results, the statistical study relied on descriptive statistics, correlation analysis and regression. The result of this study was that all the mentioned factors have a significant impact on customer satisfaction and that price fairness and coverage are the main factors that contribute to customer satisfaction among university students. In a similar study conducted in India [5], the SERVQUAL model used to measure service quality, and customer satisfaction was adopted according to the following dimensions (Reliability, Assurance, Tangibility, Empathy, Responsiveness, and Network), the overall perception of service quality was evaluated using a questionnaire asking about the general opinion of the user about the quality of service. The results of the study indicated that service quality is affected by all six dimensions, but network quality is the most important in subscriber satisfaction, followed by Responsiveness.

The concept of customer satisfaction is closely related to customer loyalty and keeping them loyal in the long run. Therefore, in some studies, customer satisfaction with services has been studied as an indicator to determine customer loyalty. [6] showed that sales promotion is not related to customer loyalty, While the quality of communication, coverage, and customer service centers have a direct impact. In addition to the low price and value-added services.

The subjective method (survey, interview, etc.) is the most common and reliable method for analyzing customer satisfaction. However, it is expensive, time-consuming, lacks real-time redundancy and may not capture the technical aspect of telecom network service performance in the telecom industry. Indeed, the subjective method can provide a clear picture of customer satisfaction levels and the factors that led to satisfaction in the past few weeks or months, but studying the results takes a long time, and customer preferences change rapidly, which makes this as one of the biggest disadvantages of subjective methods in assessing customer satisfaction. [7]

3.2 Studies based on social media

Social media studies were adopted after the development of social media applications to support customer satisfaction. Many researches relied on social data to analyze and support customer satisfaction instead of traditional means. [8] measured customer satisfaction for Saudi telecom companies using sentiment analysis based on a set of 20,000 Arabic tweets. machine learning approach using support vector machine (SVM) was tested, with two deep learning approaches: long short-term memory (LSTM) and gated recurrent unit (GRU). As a result of this research, the extent of customer satisfaction with telecom companies in the Kingdom of Saudi Arabia was determined. The

study also showed that the best methodology for measuring satisfaction based on Arabic tweets is (binary-GRU) with the attention technique, where the accuracy of the model reached 95.16%. In a similar study [9] based on customer sentiment analysis, the comments of Jordanian telecommunications companies' customers on the social networking site Facebook were analyzed, where 14,332 comments were processed and categorized as positive or negative. In this study, KNN (K Nearest Neighbor), SVM (Support Vector Machine), NB (Naïve Base), and DT (Decision Tree) classifiers were relied upon, and the result of this study showed that the SVM classifier outperformed the other three classifiers with an accuracy of 95%.

Customers' comments may include their preferences and opinions about the service (or their complaints about a particular aspect of it), which helps to assess the extent of customer satisfaction and show potential for service improvement [10]. However, there are some drawbacks in adopting this method due to the disorganization of comments compared to the directed questionnaire. Furthermore, analyzing customers' satisfaction using their comments may be biased as they are usually emotional or in a negative mood, and most of these comments are colloquial and linguistically uncertain, which leads to semantic ambiguity [11]. The characteristics of strong emotionality and professional weakness in customer comments are the biggest obstacle to the accuracy of satisfaction models shown based on this method.

3.3 Studies based on quality-of-service

Since service quality is one of the most important factors influencing customer satisfaction, many researchers used service quality as a proxy to customer satisfaction. The term Quality of Service (QoS) has been defined in largely different applications in different ways, but most definitions refer to end-user satisfaction, expectations, or fulfillment of requirements. Monitoring the quality of service for any telecom network requires continuous operations that measure the values of Key Performance Indicator (KPI) parameters in real time and analyze the measured empirical data of KPIs that determine the quality of service provided to subscribers. In the study [12], Call Setup Success Ratio (CSSR), Call Drop Ratio (CSR), and Traffic Channel Congestion Ratio (TCH) were measured, and performance was evaluated based on the threshold values for these parameters that were specified by the vendor. In a similar study [13], the Answer Seizure Ratio (ASR) was measured, which represents the probability that a call experience will lead to resonance, the Answer Bid Ratio (ABR) is the ratio of answered calls to contact attempts, the ratio of unsuccessful calls (UCR), Network Efficiency Ratio (NER), Call Setup Time (CST), Average Length of Call (ALOC). The results were analyzed and compared with the target values in order to diagnose and fix network problems and thus increase the quality of service provided. Some studies have evaluated service quality based on analysis of data stored in Call Detail Record (CDR). Every time a call is made, a detailed log is generated. As the detailed records are the tickets whose data provides information related to the relevant system elements, such as the time and duration of the call, types and phone numbers, as in [14] an algorithm was shown to monitor the quality of service, and to detect the occurrence of malfunctions based on these records, once a call is made, the behavior is analyzed for every element in the communication network, which helps reduce complaints resulting from poor service, and increase the quality of service for customers and thus increase their satisfaction.

Those previous studies focused on evaluating the technical aspect of the quality of the network provided by telecom companies through the parameters of Quality of Service (QoS), where these parameters are measured on network nodes in order to evaluate and improve them to provide a service that meets the needs of customers. But despite this, what matters to customers is the subjective and non-technical experience of the service, as the quality of service does not necessarily indicate the quality of the experience and therefore does not indicate customer satisfaction.

3.4 Studies based on customer care

Proceeding from the fact that customer service is one of the methods of direct communication with

users, many studies [15], [16], [17], [18], and [19] have studied the predictability of customer satisfaction based on incoming phone calls in call centers or the ongoing conversations between subscribers and agents on the company's communication websites.

In the study [16], in order to build a model to assess satisfaction, chat conversations were extracted from Orange customer service call center records for a period of one month. Net Promoter Score (NPS) model of CRM (Customer Relationship Management) is used, where at the end of the conversation customers have the option to answer the question: "Given your connection to our company, how likely are you to recommend us to your friends or family?" the customer is asked to provide an answer on a scale from 0 to 10. These ratings are then grouped into 3 categories: detractors (0 to 6), negatives (7 or 8), and promoters (9 or 10). Several models were trained and tested with the aim of studying the possibility of predicting NPS value directly from chat logs without the need for direct customer contact, such as linear support vector machine (SVM), convolutional neural network (CNN), and Recurrent neural networks (RNN). where the best accuracy reached was 57.5%.

In the study [17] a machine learning-based system was designed and implemented to automatically predict customer satisfaction after incoming phone calls to a US insurance company's call center. After the call ended, a transcript of the call was automatically generated by the speech-to-text system, and to monitor customer satisfaction, they asked customers to take a survey with four topics measured by the questionnaire, the scores ranged from 1 to 10 and the average of the four scores was calculated as Representative Satisfaction Index (RSI). In addition to these data, metadata for calls duration, waiting time, customer information, and policy information were used. To evaluate the results of the proposed approach in predicting the RSI, it was compared with the results of three different linear and non-linear regression methods (Ridge regression, Lasso regression, and random forest regression), and the SVM Linear Support Vector Machine classification method.

In the study [18], customer service data from a British telecom company was used to assess customer satisfaction for each incoming call. This data contained traditional call information (customer profile, error/complaint details, and service type), in addition to NPS customer assessment value (0 Not Completely Satisfied to 10 Fully Satisfied), as well as two types of UGCs (User-Generated content) on the call, CC Customer Comment and AN Agent Note. The AN agent's note is given by the call center employee who handled the call, who briefly summarizes the call and records the main technical comments about errors/complaints, while the customer leaves a CC comment through which his opinion of the call. This study proposed a machine learning-based approach (CAMP) to perform CC and AN matching analysis in order to predict the value of NPS. In this approach first, the Convolutional Latent Semantic Model (CLSM) was used to extract the latent aspects in order to assign a pair of CC and AN to two indicative latent vectors of the acquired dimension space and calculate the distance between them with the aim of determining the matching ratio between client and agent comments. Then the characteristics were extracted and CC sentiment was calculated using Bayesian classifier. Hence, based on both the feelings of CC and ratio of matches between the CC and the AN, a prediction model was built to estimate the value of NPS with 55% prediction accuracy.

studies based on service evaluation after the call ends face many challenges, with only a small percentage of customers communicate with service centers and fill out surveys, which may cause bias in the studied sample, and make it difficult to cover the opinions of different segments. Also, the data extracted from the conversations will contain weak objective evidence about these subjective opinions, and may not cover all factors influencing customer satisfaction.

Nearly 42% of customers prefer chat conversations with customer service rather than calling, so some studies have focused on improving the conversation with chatbot if the chat support is automated, or with agent if the chat support in person. In the study [19], customer satisfaction with

a customer support conversation was measured using SVM machine learning technology, based on a set of data that represents the customer's feelings and personality extracted from his conversation.

studies of customer satisfaction measurement based on agent conversations are challenging because it requires the conversation between agent and customer to be long enough to extract the customer evaluation through it.

3.5 Studies based on QoE

The International Telecommunication Union (ITU) defines Quality of Experience (QoE) as: "the general acceptance of an application or service, as personally perceived by the end user". [20]. The importance of measuring the quality of experience was documented in a survey of clients of 362 companies, where only 8% of the clients of these companies describe their experience as superior, while 80% of the companies surveyed believe that the provided experience to clients is a superior experience [21] that what called the service delivery gap. Several recent studies in various fields have relied on QoE to measure customer satisfaction. In telecommunications, QoE is a measure of customer satisfaction with a service or services that they have tested with a service provider, which can be a measurement of the quality of a specific service or all services. A survey of telecom operators showed that proactive customer experience management is the biggest opportunity for data analytics applications [22].

QoE differs from the traditional QoS metric that measures the quality of network layer services as seen in previous studies. When measuring the quality of the experience, it is the personal customer experience that matters. QoS primarily focuses on what happens within network parameters (such as latency, throughput, packet loss, and delay), while QoE focuses on why customers behave in a particular way. A problem with QoS parameters can lead to a QoE problem such as excessive waiting time. Based on this relationship between service quality and experience quality, customer satisfaction can be inferred by finding a relationship between network parameters during the service session, and the customer's opinion of the quality of the provided session. It also enables the identification of how changes in QoS parameters can affect customer experience as well as the impact of both quality of service and customer experience on customer satisfaction. Therefore, in the study [23], a method was proposed to study the relationship between service quality and quality of experience for mobile Internet in order to measure the correlation between them while measuring service quality and estimating the quality of customer experience. Used QoS parameters consist of delay, information loss, availability, and data throughput. While QoE parameters consist of network response time, waiting time, and response time for customer complaints.

In a similar study [24], an approach to building a QoE model was proposed by establishing a quantitative relationship between a set of KPIs and an objective measure (such as duration of voice calls or video playback). The same approach has been followed in other studies, duration of the call for service calls [25], and the percentage of playing videos for the Internet service [26], based on the assumption that the length of the call indicates the clarity of the voice, the absence of interference and interruptions, and thus indicates on the customer's enjoyment of good communication service.

Since the most accurate method for assessing perceived quality is self-assessment, as there is no better indicator of quality than that provided by the customer, some studies have combined the subjective and objective method, as in [27] the discrepancy between the quality of multimedia streams transmitted on the SDN network and the assessment of this quality for users. This variance was measured between the subjective MOS scale expressed by 20 users who viewed the video, and three objective measures, the Video Quality Metric (VQM), the Structural Similarity Index Measure (SSIM), and the blocking effect. To measure the discrepancy between these subjective and objective measures, supervised learning algorithms were used.

In Summary, machine learning techniques were not applied in order to classify overall satisfied/dissatisfied customers or to identify the most critical indicators based on continuously created objective data in telecom systems. Current

research within the area of artificial intelligence and machine learning is focusing on network data in order to improve quality of service. between the limitations of determining customer satisfaction using subjective questionnaires and the overall customer satisfaction as a fundamental management target in the telecom industry, there is an unmet need for a system to classify satisfied/dissatisfied customers from objective technical data. Large amount of data that is produced every day by the various telecom systems which can be used to achieve this goal. Such an approach has not been presented so far as the literature review expressed.

3. Proposed Model

After reviewing the methods and techniques for analyzing and measuring customer satisfaction in previous studies, we proposed a model for predicting customers' satisfaction for the Syrian telecom operator Syriatel based on the quality of the experience perceived by customers through their interactions with the various operator services.

Unlike previous studies that measured satisfaction based on social media comments, network quality parameters, or analyzing incoming calls to call centers, the proposed model relies on the objective factors extracted from telecom systems that are affecting customer satisfaction in a comprehensive way. This is done by measuring customer experience in an objectively through all customer interactions (request to activate a service, activate offers, call the call center, visit the service center), and projecting a set of network performance indicators on every customer call or internet session, with the aim of determining the factors that affect customers' experience. The methodology also determines each factor satisfactory threshold. the proposed methodology overcome the main drawbacks of all previous illustrated methods by capturing the technical aspects behind satisfaction.

We studied customer's experience based on three dimensions as following:

-Human dimension: due to the variation between the customer's perspectives and their requirements, we did not rely only on demographic features (age group, gender, line type, line age, location, etc.), but we also extracted a set of behavioral characteristics to illustrate the variation between different customers (average days in which the user makes a call within the network or to other networks, average days in which the user receives calls from within the network or from other networks, average days of internet consumption, etc.).

- Service dimension: to measure the quality of the subscriber's experience within the various services such as customer service, whether visiting the service point or calling the service center, activating offers, recharging, payment, internet session, making calls, and others.

- Context dimension: features were measured at the time and place in which the service was requested by the subscriber.

While measuring the quality of service based on objectively good network performance indicators does not reflect the users' subjective perception of the quality of service they experienced. NPS index was relied upon as a criterion for measuring the satisfaction of subscribers, where a text message (SMS) was sent to sample of subscribers to determine their satisfactory level with the services provided by the operator with a value ranging between 1 and 10. Different customer segments will show different behavior and different perception on the services they received. Therefore, a sample of 10,000 subscribers was selected, taking into account the different segments of subscribers. The random selection of the sample to take into account the normal distribution led to the loss of some segments due to the small number of users belonging to those segments (such as post-paid lines for females under 18 years old in the northern region).

Therefore, the following characteristics and limits were relied upon in order to divide the participants into approximately equal sub-slices (80 different segments) from the selected sample:

- Subscriber location: the southern region, the northern region, the coastal region, and the central region.
- Subscriber age: where the participants were divided into 5 age groups as follows: younger than 18, between 18 and 24, between 25 and 35, between 36 and 50, and older than 50.
- Subscriber gender: male or female.
- Subscriber line type: prepaid or postpaid.

We trained and tested the model using seven machine learning algorithms based on subscribers' satisfaction values (NPS values) and a set of network and telecom systems objective features, the most influential features were identified to predict NPS value, in order to be able to predict customer satisfaction based on interactions. Figure 2 shows the followed steps within the study.

4. Methodology

This section describes the materials and methods to classify customer satisfaction/dissatisfaction and to identify indicators for customer satisfaction. First, necessary data sources are described. Second, data preprocessing is presented. Third, the experimental study to prepare and model the data is explained.

5.1. Data Sources

5.1.1. Customer Relationship Management (CRM)

Through CRM system, contracts information for customers was obtained, such as gender, age, line age, line type, and others.

5.1.2. Unstructured Supplemental Service Data (USSD)

USSD protocol is used in GSM mobile phones to communicate with the service provider. It can be used to check balance, browse services available on the content menu, or subscribe to a specific service. In this study, this data was relied on to determine the times that the subscriber requested to activate a service through USSD, in order to integrate this data with the data of the services activated in the OCS system to determine the actual time it took to activate a service for a specific subscriber.

5.1.3. Online Charging System (OCS)

OCS is used to allow the service provider to charge the user for services in real time. As it deals with the subscribers' account balance. CDR (Call Detailed Record) is OCS's main component. For every event or communication made by the customer (making a call, sending a text message, or connecting to internet), events are captured and stored in files, which called CDR. In this study, we used CDR data to determine the used cells by the subscriber to make a call or internet session, and integrate this data with cell's KPIs at that time. Activation data for services was extracted for services, packages, and offers, to integrate it with activation requests data via USSD mentioned earlier.

5.1.4. Equipment Identity Register (EIR)

network entity used in GSM (Global System for Mobile Communication) that stores lists of IMEI (International Mobile Equipment Identity) numbers, which identifies the actual physical phone used by customer and its information. In this study, we used EIR database to obtain subscriber's device type, device specifications, operating system and 5.1.5. Call Center data

Is a system for managing calls and improving customer service by letting customer directly reach company's agent. In this study, this system's database was used to obtain data related to subscribers' communications to call centers, such as waiting time, length of service, and the rate of unanswered and serviced calls.

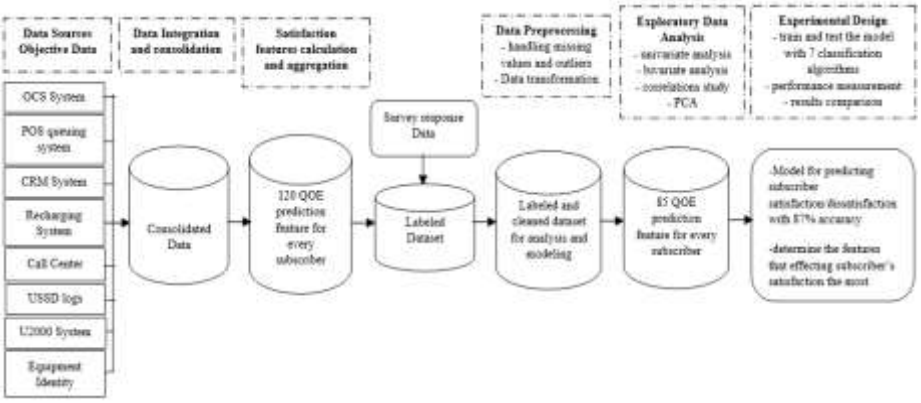

Figure (2) followed steps for Customer Satisfaction prediction

OCS each 2 minutes or a new service activation request is assumed, however different windows for different systems' data integration and consolidation were used based on the integrated systems working logic. The aggregation procedure was split up in four steps and is explained in the following.

Step 1: Data sources consolidation

(1) Data sets for all subscribers were extracted from the data sources introduced in Section 5.1.

(2) Check all data sources for matching subscribers (MSISDN).

(3) subscribers with survey responses were selected.

Step 2: Specified matching

(1) Data sets with identical timestamps and MSISDN were matched.

(2) subscribers with different timestamps but identical MSISDN were transferred to Step 3.

Step 3: Tolerance matching

(1) Biased timestamps with identical MSISDN were filtered and the difference between the timestamps was calculated.

(2) Records with a difference less than the window were consolidated into one request or interaction.

(3) Records that did not fulfill these criteria were considered as separate request or interaction.

Step 4: Integrate subscriber data

Since we study the quality of the customer's experience during three months, we have calculated the final expressive features as the arithmetic mean of all his experiences for a particular service during the study period (for example, the average time to activate a service via USSD), or calculate the feature as the worst experience of the service (for example, the longest time to activate a service via USSD). Where in the final dataset the customer experience was represented by a single view.

5.2.2. Data cleaning

Aims to clean up data by filling the missing values, dealing with outliers, smoothing out noisy data, and fixing inconsistencies. The common methods for handling missing/outlier values are to ignore the attributes or observations, or to fill the missing value with the mean or median value. In this study, these methods were not used because the

5.1.6. Point of Service (POS) queuing system

Queuing system can be described as a system with service facilities that customers access to request a specific service, that is, the input to the system consists of customers who request service, and the outputs are customers who have been served. In this study, the database of this system was relied on to obtain data related to subscribers' visits to service centers, such as waiting time and length of service, with the aim of studying the quality of this service and its impact on subscribers' satisfaction.

5.1.7. Abili Recharging System

electronic recharge system that allows subscribers to recharge their prepaid lines, pay their bills and recharge their electronic payment account. In this study, we used this system database to extract subscriber's recharge requests or payment requests, to further been integrated with actual recharging time on OCS system, to calculate request's duration time.

5.1.8. Central Management System for Mobile Network Elements (U2000)

This system centrally manages the elements of the mobile network. In this study, this system data was used to calculate number of networks KPIs (Key Performance Indicators) at the level of one cell per hour during the study period, in order to integrate them with subscriber calls data that was previously mentioned within the OCS system, for the aim of projecting these indicators to the single subscriber's experience to calculate the quality of the network provided at the level of one subscriber.

5.2. Data Preprocessing

5.2.1. data integration and consolidation

Telecom systems store the information of a customer interaction in combination with the subscriber identification number (MSISDN) and a timestamp, which is not unique in all data sources as there can be multiple timestamps identifying one customer request (for example timestamp of service activation request by USSD and the actual timestamp of service activation on OCS). This causes a biased timestamp, which needs to be integrated into the consolidation logic. A window of 2 minutes was tolerated to consider the data as one request, while the USSD send batch of requests to

followed by IOS by 8 %. In some other features, we grouped a number of low-frequency categories into one new category ‘other’, as in the mobile device brand where the most popular brands (Samsung, Apple, Nokia, Xiaomi, Huawei) were preserved and the rest of brands were placed under ‘other’. Some numerical features were also converted to categorical attributes since it was found that their values are few and have high frequencies within the records when analyzing these attributes. In iShow service activations feature for example, we found that 65% of observations had a value of 0 (not iShow users), while most of the service users activate it once or twice per month, and during analysis of the number of activation times we noticed that it did not show any effect on the variation of satisfaction/dissatisfaction, therefore this feature was converted to be categorical, indicating whether the subscriber is a user of this service (1) or not (0). We also used Bi-Variate analysis, that deals with causes and relationships to find out the relationship between two variables. Through this analysis, we were able to show the relationship between the studied categorical features and the satisfactory variables. as shown in Figure 3.

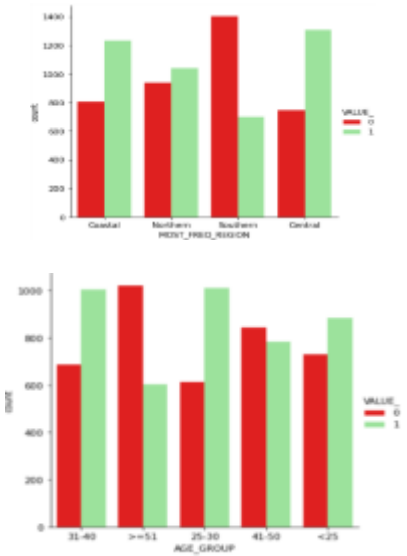

Figure (3) Bi-variate analysis samples between studied categorical features and satisfaction dependent feature

sample is small, so we were not able to delete observations that contain missing values or outliers in its features, and we couldn’t replace them with features mean values due to the presence of many features that are positively skewed from the mean. An alternative approach was used which is called Nearest Neighbor (KNN) algorithm, which calculates missing values or outliers by finding nearest neighbors using Euclidean distance scale.

5.2.3. Data transformation

At this stage, the data has been reformulated according to a new model that is more suitable for the task of analysis and forecasting. Several types of data transformation can be used, normalization is one of the most used techniques for data transformation, and in the simplest cases the values measured at different levels can be tuned to a common scale, in many cases after averaging.

5.2.4. Data Labeling

survey data and customers’ features dataset were mapped. The overall satisfaction of a particular customer was considered to be the class label. The overall satisfaction rate was measured using NPS scale from 1 to 10. In order to guarantee a two-class classification problem, the results of this question have been transferred into a binary coding. satisfaction rates of 1 to 6 were identified as dissatisfactory and labeled with 0, whereas 9 and 10 represent the satisfied customers that were labeled with 1. We dropped all observations with 7 and 8 rates, and ended up with 8,180 observations. This is the common interpretation and presentation of NPS in a management manner.

5.2.5. Data analysis and feature extraction

During feature extraction stage, we analyzed categorical data univariately (univariate analysis), and applied feature engineering techniques to it, where the number of categorical features reached 20 (14 of which were preserved). missing categorical attributes were replaced by a new class that we called ‘other’. During analysis of the frequency of categories within each attribute, it was found that some of the features have categories with a frequency of more than 85%, so these attributes were deleted, such as the feature that determine the type of operating system in the subscribers’ mobile phones, where 85% of Android devices were,

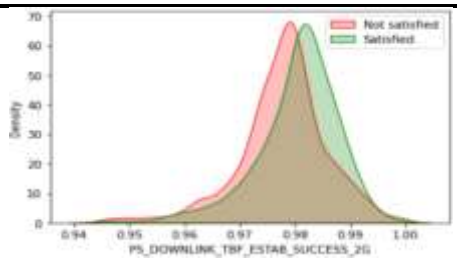

Figure (6) Bi-variate analysis for establishment success ratio over 2G and satisfaction dependent feature

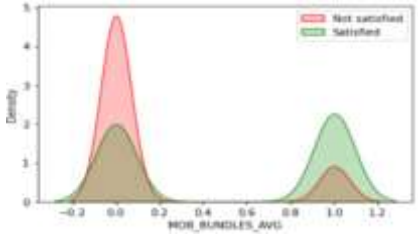

Figure (7) Bi-variate analysis for mobile bundle activation and satisfaction dependent feature

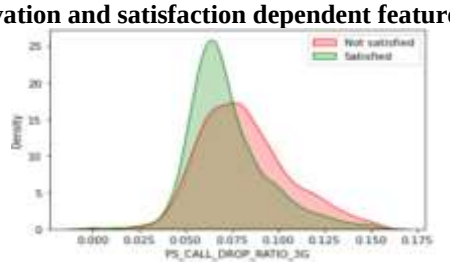

Figure (8) Bi-variate analysis for call drop ratio over 3G and satisfaction dependent feature

Spearman correlation coefficient was used to analyze the relationships between the different features, where the correlation matrix was calculated for all features, then features with a correlation value greater than 0.95 were ignored. As shown in Table 1, iShow package average consumption feature has a strong correlation with the attribute that indicate if the subscriber is a user of iShow service or not, therefore this attribute has been omitted.

Table (1) Correlation between the average iShow package consumption attribute with the attribute that indicates that the subscriber is a user of this service

	ISHOW_VOL_MB_AVG
ISHOW_USER	0.98
ISHOW_BUNDLES_AVG	0.60
VOLUME_MB_AVG	0.37

Principal Component Analysis is an unsupervised machine learning technique used to reduce the number of features. This method is part

Although numerical data can be used directly in machine learning models, we need to analyze and engineer these features in terms of scenario, problem, and domain before creating a model to reach more accurate results. Most of the numerical features have different distributions. We have studied and analyzed these distributions individually for each one (Univariate analysis) using (Histograms) curves, where the X-axis expresses the different values of the attribute, and the Y-axis expresses the frequency of these values. We observed that some features have a limited number of discrete values. For example, 89% of Abili service average waiting time feature observations have a value of 1 second, as in Figure 4. It is also observed that some features have large values grouped around zero, as the average activated ringtones per month feature shown in Figure 5. These results indicate that these features with fixed or close to constant values can be removed.

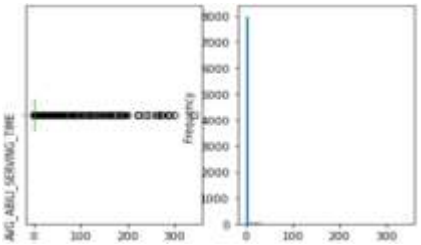

Figure (4) Histogram for Average waiting time for Abili recharging

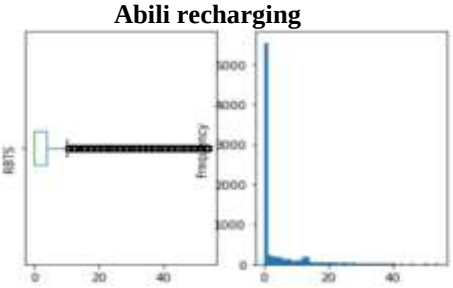

Figure (5) Histogram for average activated RBTs per month

We also used Bi-Variate analysis to analyze and show the relationship between the studied numerical features and the satisfactory variable. As shown in Figures 6,7 and 8.

classification algorithms. We have validated the different models by splitting the data into 70% for training and 30% for testing, all algorithms were trained using cross-validation using 10_fold_cross_validation.

5.3.1. Performance measurement

Confusion matrix, which contains actual and predicted classifications made by the binary classification system was calculated, where satisfied is 1, dissatisfied is 0.

below performance measures can be calculated directly from the confusion matrix:

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

Precision = True Positives / (True Positives + False Positives)

True Positive Rate = Recall = True Positives / (True Positives + False Negatives)

F1 = $\frac{2}{\frac{1}{precision}+\frac{1}{recall}} = \frac{2\,.\,precision\,.\,recall}{precision+recall}$

False Positive Rate = False Positives / (False Positives + True Negatives)

5.3.2. Classification Algorithms

We trained and tested the model with a set of linear, non-linear, and ensemble algorithms, different algorithm compared using the previously mentioned performance measurements as shown in table 2.

of the dimensionality reduction techniques. In this study, PCA algorithm was used to reduce the number of features and apply it on 10 and 50 features, but that did not increase the accuracy of the model. We also applied the algorithm on two features so we can draw and analyze the observations as in Figure 9. we noticed that observations' distribution indicating that these observations are not linearly separable. However, we have applied a set of linear, non-linear, and ensemble algorithms that will be explained in the next experimental study section.

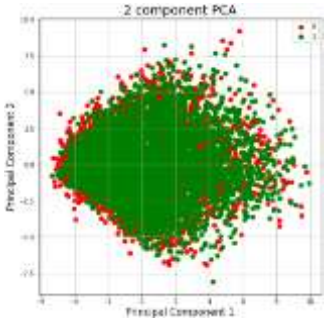

Figure (9) PCA study which indicates that our dataset is not linearly separable

5.3. Experimental Study

In this research, reliable data for 8,180 participants were analyzed through the extracted features (85 features after reduction were made in the previous section). We tested and applied seven

Table (2) Algorithm comparison using performance measurements

Algorithm	ROC-AUC	Accuracy	Sensitivity/Recall	Precision	F1 score
Logistic regression	87%	80%	81%	81%	81%
Kernel SVM	90%	82%	86%	81%	83%
KNN	86%	78%	83%	77%	80%
Naive Bayes	77%	71%	73%	71%	72%
Decision Tree	87%	82%	89%	80%	84%
Random Forest	91%	84%	90%	81%	85%
XGBOOST	93%	87%	91%	84%	87%

min_child_weight =1, alpha=0.9, subsample=1, n_estimators=200.

"Sensitivity/Recall" value was 91%, meaning that the most satisfied cases were correctly predicted, and the "Precision" value in the false prediction of the dissatisfied participants was 84%, which is a very good percentage. AUC-ROC percentage was 93%.

5. Results comparison and discussion

The prediction results shown by the satisfaction prediction models show that ensemble learning

XGBoost classifier achieved the best result, with an average accuracy for the training set of (84.7%) and a standard deviation of (1.5), meaning that model accuracy on training data ranged from 83.2% to 86.2%. On testing data set, the accuracy was 86%. Following values were set for the algorithm parameters, where several values were tested (by applying grid search) and these were the values with highest accuracy for the model:

learning_rate=0.05, eval_metric='logloss', max_depth=10, eta=0.01, gamma=0,

of satisfaction once satisfaction indicators are identified. The presented methodology for linking technical cellular data to customer perceptions is a new approach that can be used for the service industry. For example, we can identify dissatisfied customers and try to satisfy them to reduce churn through offers, or fix the bad experiences indicated by the features that caused their dissatisfaction.

The presented work answers two research questions:

- to what extent can satisfied/dissatisfied customers be categorized based on the data produced during the customer's interactions with the network? the results of the research showed that it is possible to classify customers based on customer interaction data, and also showed that it is not a single event that transforms a customer from satisfied to unsatisfied but rather the entire history of the customer's experiences. In the presented data analysis procedure, the challenge of integrating technical data from different cellular systems with personalized questionnaire responses to standardize the data was resolved. An experimental design consisting of seven modeling approaches was performed to answer the research question mentioned above. The algorithm that predicts customer satisfaction/dissatisfaction the best was XGBOOST classifier with an accuracy of 87% for the test data.

- The second research question: Can satisfaction indicators be derived from cellular data? This question was answered by the results of analytical and empirical studies, and indicators of satisfaction were identified. Knowing these indicators allows for proactive customer treatment as dissatisfied customers can be directly identified. Thus, actions to increase satisfaction can be implemented immediately and secure a competitive advantage. The features that most affect the satisfaction are the local minutes packages, the number of answered calls at customer service centers, the success establishment rate of internet session in the 3G network, and the fact that the customer's line is postpaid and located in the southern region.

algorithms such as XGBOOST and random forest outperform other classification algorithms in terms of accuracy, AUC and F1-measure criterion.

The information gain is a measure of the importance of a feature; therefore, it was used to detect features with the greatest influence on satisfaction. Gain criterion: includes the relative contribution of the attribute corresponding to the computed model by taking the contribution of each attribute in each tree in the model. A higher value of this metric when compared to other features means that it is more important for forecast generation.

Figure 10 shows the features that most influence the prediction of satisfaction in terms of the gain criterion for the XGBOOST model. The most influential features of satisfaction are the local minutes packages, the number of answered calls at customer service centers, the setup success rate of internet session in the third-generation network, user's line to be postpaid, and the user to be located in the southern region.

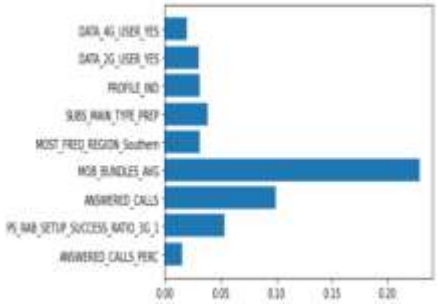

Figure (10) most influential features on satisfaction prediction based on gain criterion for XGBOOST

6. Conclusion and future work

This paper demonstrated a method for predicting satisfied/dissatisfied customers in the Syrian telecom operator Syriatel based on the data produced within telecom systems with a maximum accuracy of 87% with XGBOOST classifier.

It also contributed to predicting customer satisfaction based on objective data only. This allows proactive dealing with customers in the future in order to achieve the management goal of increasing customer satisfaction. Moreover, sustainability can be achieved to maintain the level

in-depth capture of the best picture of the community as a whole. Due to the reliance of this study on the questionnaire, some of the responses provided by the participants in the study may be biased. In order to generalize our findings, comparative studies should be conducted in different contexts to test the accuracy of the results.

Funding information: this research is funded by Damascus university – funder No. (501100020595).

Despite the valuable results of this study, as it is the first study where the factors affecting the satisfaction of the customers of Syrian telecom companies were determined, the sample that was approved may not represent the entire population of subscribers to the Syrian telecom service despite the relatively large sample collected considered large enough to generalize compared to previous studies. It remains a question whether the sample was ideally or objectively representative, or allowed an

References

1. Munyanti, I., & Masrom, M. (October, 2018). Customer Satisfaction Factors towards Mobile Network Services. Journal of Advanced Research in Business and Management. Vol. 13. P-P: 9-18.
2. Viet, P. Q. (March, 2019). Assessment of the Satisfaction Customers Using Mobifone Mobile Service in Ho Chi Minh City. Business and Economic Research. Vol. 9. P-P: 228-235.
3. Ampomah, Y. K. (September, 2012). Factors Affecting Customer Satisfaction and Preference in the Telecommunications Industry: A Case Study of MTN Ghana. Common Wealth Executive Masters of Business Administration. Institute Of Distance Learning, Kwame Nkrumah University of Science and Technology. Ghana. P: 92.
4. Khan, S., & Afsheen, S. (2012). Determinants of customer satisfaction in telecom industry a study of telecom industry peshawar KPK Pakistan. Journal of Basic and Applied Scientific Research. Vol: 2. P-P: 12833-12840.
5. Kalita, B. (May, 2019). A study on the dynamics that influence customer satisfaction and loyalty towards various telecom service providers. Journal of Management in Practice. No: 1. Vol: 4.
6. Hossain, M. M., & Suchy, N. J. (2013). Influence of customer satisfaction on loyalty: A study on mobile telecommunication industry. Journal of Social Sciences. Vol: 9. P-P: 73-80.
7. Subramanian, P., & Palaniappan, S. (2015). Telecom Data Integration and Analytics-Proposed Model to Enhance Customer Experience. International Journal of Conceptions on Computing and Information Technology. Malaysia University of Science and Technology. Malaysia, P:6.
8. Almuqren, L. A., Qasem, M. M., & Cristea, A. I. (2019). Using deep learning networks to predict telecom company customer satisfaction based on Arabic tweets. In 2019 28th International Conference of Information Systems Development. Toulon, France.
9. Najadat, H., Al-Abdi, A., & Sayaheen, Y. (May, 2018). Model-based sentiment analysis of customer satisfaction for the Jordanian telecommunication companies. In 2018 9th International Conference on Information and Communication Systems (ICICS). Irbid, Jordan.
10. [10] Zhang, K., Narayanan, R., & Choudhary, A. (2010). Voice of the customers: mining online customer reviews for product feature-based ranking. In 3rd Workshop on Online Social Networks (WOSN 2010). P-P: 11-11.
11. Coleman, D., Georgiadou, Y., Labonte, Y. (2009). Volunteered Geographic Information: The nature and motivation of producers. International journal of spatial data infrastructures research. Vol:4. P-P: 332–358.
12. Idigo, V. E., Azubogu, A. C. O., Ohaneme, C. O., & Akpado, K. A. (2012, October). Real-Time Accessments of QoS of mobile cellular Networks in Nigeria. International journal of engineering inventions. Vol: 1. P-P: 64-68.

13. Basic, N., Lipovac, V., & Lipovac, A. (2012, April). Practical Analysis of xDR Based Signaling Network Performance and End-to-End QoS. In International Conference on Networked Digital Technologies. P-P: 34-45. Springer, Berlin, Heidelberg.
14. Breda, G. D., & de Souza Mendes, L. (2006). QoS monitoring and fault detection using call detail records. In Proceedings of the International Conference on Wireless Information Networks and Systems. P-P: 95-100.
15. Cabarrão, V., Julião, M., Solera-Ureña, R., Moniz, H., Batista, F., Trancoso, I., et al. (September, 2019). Affective analysis of customer service calls. In 2019 10th International Conference of Experimental Linguistics. Lisbon, Portugal.
16. Auguste, J., Charlet, D., Damnati, G., Béchet, F., & Favre, B. (2019, May). Can we pre dict self-reported customer satisfaction from interactions? In 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP). P-P: 7385-7389. IEEE.
17. Bockhorst, J., Yu, S., Polania, L., & Fung, G. (2017, September). Predicting self-reported customer satisfaction of interactions with a corporate call center. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. P-P: 179-190. Springer, Cham.
18. Wei, Q., Shi, X., Li, Q., & Chen, G. (2020). Enhancing Customer Satisfaction Analysis with a Machine Learning Approach: From a Perspective of Matching Customer Comment and Agent Note. In 53rd Hawaii International Conference on System Sciences. Grand Wailea, Hawaii.
19. Adl, A., & Yacoup, K. M. (March, 2019). Determination of Telecom Customers satisfaction from their Personality Traits using Natural-language understanding and SVM. In The Fourth Student Conference of the Education and Student Affairs Sector entitled Innovation and Creativity for Bachelor Students in Egyptian, Arab and African Universities.
20. Chiara Gentile, C., Spiller, N., & Noci, G. (October, 2007). How to Sustain the Customer Experience: An overview of experience components that co-create value with the customer. European Management Journal. Vol: 25. P-P: 395-410.
21. Allen, J., Reichheld, F. F., Hamilton, B., & Markey, B. (2005). Closing the delivery gap. Bain & Company Report.
22. [22] Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., ... & Gani, A. (2014). Big data: survey, technologies, opportunities, and challenges. The scientific world journal.
23. Yusuf-Asaju, A. W., Md Dahalin, Z., & Ta'a, A. (August, 2016). A proposed framework for mobile Internet QoS customer satisfaction using big data analytics techniques. In Knowledge Management International Conference (KMICe). Chiang Mai, Thailand.
24. Aggarwal, V., Halepovic, E., Pang, J., Venkataraman, S., & Yan, H. (2014, February). Toward quality-of-experience estimation for mobile apps from passive network measurements. In Proceedings of the 15th Workshop on Mobile Computing Systems and Applications. P-P: 1-6.
25. Chen, K. T., Huang, C. Y., Huang, P., & Lei, C. L. (2006). Quantifying skype user satisfaction. ACM SIGCOMM Computer Communication Review. Vol: 36. P-P: 399-410.
26. Dobrian, F., Awan, A., Joseph, D., Ganjam, A., Zhan, J., Sekar, V., ... & Zhang, H. (2013). Understanding the impact of video quality on user engagement. Communications of the ACM. Vol: 56. P-P: 91-99.
27. Abar, T., Letaifa, A. B., & Asmi, S. E. (2020). Quality of experience prediction model for video streaming in SDN networks. International Journal of Wireless and Mobile Computing. Vol: 18. No: 1. P-P: 59-70.