Customer churn prediction in telecommunication industry using data certainty

# Customer churn prediction in telecommunication industry under uncertain situation

Adnan Amin[1], Feras Al-Obeidat[2], Babar Shah[2], Awais Adnan[1], Jonathan Loo[3], Sajid Anwar[1]

[1]Center for Excellence in Information Technology, Institute of Management Sciences, Peshawar, 25000 Pakistan.
[2]College of Technological Innovation, Zayed University,144534 Abu Dhabi, United Arab Emirates.
[3]Computing and Communication Engineering, University of West London.

[1]{adnan.amin, awais.adnan, sajid.anwar}@imsciences.edu.pk, [2]{babar.shah, Feras.Al-Obeidat}@zu.ac.ae,
[3]*jonathan.loo*@uwl.ac.uk

## Abstract

With the terrific growth of digital data and associated technologies, there is an emerging trend, where industries become rapidly digitized. These technologies are providing great opportunities to identify and resolve different problems. In particular, the telecommunication industry is facing a serious problem of customer churn relating to, the customers who are going to abandon their established relation with the business/network in the near future. This problem cannot only affect the rapid growth of the business but can also affect the revenues. Therefore, many customer churn prediction (CCP) models have been introduced but not yielding the desired performance in CCP. This is because there can be many factors, that contribute to customer churn which are still unexplored. In this paper, we focus on determining the effectiveness of the factors, i.e. lower and upper distance between the samples, are considered by the proposed model for the CCP. Further, we demonstrate a novel solution pertaining to the telecommunication sector showing the hidden factors considered for predicting the customer churn. Finally, we investigate the effects of both types of samples: those samples that are low distance and the upper distance (in terms of relevance) to the majority samples in given publicly available dataset. As a result of the study, we found that lower distance test set (LDT) samples have obtained best performance as compare to upper distance test set (UDT) samples in term of increased in the accuracy, f-measures, precision and recall when the uncertain sample size increases. Because the classification performance on upper distance samples remain almost the same when the size of samples increased in the test set.

**Keywords: -** churn prediction, uncertain samples, classification, telecommunication, customer churn

## 1. Introduction

The customers are considered one of the most important asset for a business in numerous dynamic and competitive companies within a marketplace [1]. In competitive market companies in which customers have numerous choice of service providers they can easily switch a service or even the provider. Such customers are referred to as churned customer [2]. The causes of customer churn can be due to dissatisfaction, higher cost, low quality, lack of features, and privacy concerns [3].

Many organizations e.g., financial service [4], [5], airline ticketing services [6], social network analysis [2], [7], online gaming [8], banking [9] and telecommunication sector [10], are ever more focusing on establishing and maintaining the long-term relationships with their existing customers [10]. Loyal customers can be considered long-term customers that are not only profitable for the company but also are great ambassadors in the market [11]. One of the industry wherein this phenomenon is observed is the Telecommunication Communication Industry (TCI). CCP in TCI is an increasingly well-known domain and popular research problem in the literature in recent years [1], [10], [12]. It is reported that TCI is suffering from the substantial problem of customer churn due to fierce competition, saturated markets, dynamic condition, and launching new attractive offers [2].

It is observed that acquiring new customer can be more expensive for companies as compared to retention of the existing customer [13]. Also, the researchers have confirmed that CCP approaches can improve a company's revenue and good reputation in market [7]. Currently, companies in the TCI have abundant information about their customers−including local/international call records, short massages, voice mail, demographics, financial detail, and other usages behavior of the customers. This has created an opportunity for the machine learning (ML) researchers to perform analytical and predictive modeling techniques to handle the CCP in TCI. Therefore, a wide range of approaches based on ensemble techniques [14], probabilistic methods [15], Support Vector Machine (SVM) [4], K-nearest neighbor (KNN) , Rough Set Theory (RST) [10], Fuzzy Logic Systems [16], Neural Networks [17] etc., has been developed to identify customers with the highest tendency to churn. Accordingly, decision maker offers incentives to persuade such high risk customer to stay. These CCP approaches help in deciding measures from the TCI datasets to prevent their customers from churning by offering them promotions and better deals. [18]. However, these machine learning (ML) techniques lack the required effectiveness for the CCP [19]. Hence, there is greater challenge of which ML or data mining technique for prediction model should be chosen for CCP. This is partly because there are many uncertain factors that needs to be addressed which contribute in CCP for TCI. In this paper, we introduce a novel approach focusing on a factor of uncertain samples for addressing CCP in TCI yielding significant improvements in the performance of CCP model. We have obtained the uncertain samples based on lower and upper distance factors between the samples as these factors has not been considered for CCP in TCI yet. The proposed approach pertaining to the target industry, showing the discussed unexplored factor, i.e. lower and upper distance samples, can play a vital role in CCP. Furthermore, we empirically investigated the performance of CCP model under uncertain situation in TCI.

The rest of the paper is organized as follows: the next section presents literature review of CCP approaches and critical discussion on existing techniques; the propose methodology of this study is explored in Section 3. Section 4 presents the results, comparisons and discussion on our findings; the paper is concluded in Section 5.

## 2. Related work

The review in this section is primarily related to exploring the state-of-the-art techniques for CCP that have been adopted for CCP.

C.F. Tsai et al. [20] presented a hybrid neural networks approach for CCP in a CRM dataset of the American telecommunication company. They used an approach in which they have combined artificial neural network (ANN) and self-organized map (SOM) for CCP model. The ANN is used for data reduction in which unrepresentative data was filtered out from the training set. Then, the output of the first step is put into the SOM to build prediction model. The results indicate that combination of ANN+SOM outperform the single neural network with respect to accuracy. However, it can be observed that data reduction and filtering in first method (i.e., ANN) leads to loss of samples from the training set.

Wouter Verbeke et al. [21] explored the application of Ant-Miner+ which is based on Ant Colony Optimization. Anti-Miner+ can allow to add domain knowledge through monotonicity constraints on the final decision rules. As a result, it produces highly accurate prediction model. Further, they have also incorporated the results to SVM, RIPPER, logistic regression and decision tree; wherein, combined with RIPPER reported the highest accuracy, while higher sensitivity obtained through decision tree. However, it is also investigated that Ant-Miner+ produces low sensitive decision rules and also require domain knowledge to be incorporated in the final rules-set for obtaining higher accuracy in CCP. On the other hand, RIPPER also produces good results in small rule-sets as well as it is leads to unintuitive approach that violate the domain knowledge.

Benlan He et al. [22] proposed CCP methodology based on the SVM classifier and random sampling technique. The random sampling method is used to change the data distribution in order to minimize the imbalance class problem which is caused due to the lack of availability of data. However, the class imbalance issue does not improve the predictive performance of their CCP model [23]. In connection to this, Burez et al [23] suggested to use weighted random forests method. However, random forests are often criticized for being very difficult to interpret and understand, particularly to identify reasons of customers churns which are important to explore for preventing the risky customers from churning [24]. Thus, it is not the appropriate methods to address CCP [25].

In order to improve the predictive performance of the CCP model, researchers have also proposed ensemble techniques. An ensemble method is the combination of several member of classifiers into one aggregated model. Koen W. De Bock et al. [26] proposed ensemble technique based on rotation forest and Rotboost as two modeling methods for CCP. The rotation forests are used for features extraction while Rotboost method is applied in combination with rotation forest and AdaBoost to improve the performance of the CCP model using ensemble technique. The results shown that Rotboost outperform than rotation forests in term of accuracy while rotation forests obtained higher value of area under the curve (AUC) and lift measures. However, again there is problem of interpretability and understandability of the factors of customers churns. Therefore,

they suggested another study [27] based on generalized additive models (GAM) concepts and incorporated this concept into ensemble classifier (e.g., Bagging and random subspace, and semi parametric GAM). As a result, it obtained comparatively good predictive performance as compared to training classifier individually with logistic regression and GAM method.

Similarly, Pendharkar et al. [28] suggested to use genetic algorithm (GA) and neural network (NN) for CCP. Where the genetic algorithm was used to search features space while NN applied to predict the customer churn. The GA based NN CCP model increase the prediction accuracy of the customer churn. Furthermore, they have compared the performance of the GA-based NN model with z-score model and evaluated using receiver operating characteristics (ROC) curve and lift measures. It was found that the GA-based NN model performed better than z-score statistical mode.

The researcher in the subject area has manifested several approaches as elaborated above. However, we cannot consider which prediction approach can be set as the standard approach to address CCP in more appropriate fashion. It remains an open challenge for research community of machine learning. Although the existing work [29] argues that SVM is the best classification technique due to its ability to efficiently handle the arbitrary nonlinearities but M.A.H. Farquad et al. [4] reported about SVM that it also generates black-box illusion which can be considered its main disadvantage. Also, in literature [22] researchers have argued that the random sampling method is good approach before applying the classification technique. Random sampling minimizes the imbalance data distribution which is caused due to unavailability of the target class data in the dataset (i.e., customer churn); however, it is also reported in another study [23] that class imbalance does not significantly improve the performance of prediction model. On the other hand, Burez et al. [23] also suggested to apply weighted random forests and report its obtained results, which showed significance, in CCP model. Their technique, however, has been also criticized for its complexity in understanding and interpretation [24].

Apart from the above mentioned discussion and to the best of our knowledge, there is no study which has focused and considered the role of the uncertain samples in building CCP model for TCI. Therefore, this paper presents a novel CCP model based on the role of uncertain samples in a given TCI dataset. The next section introduces the propose methodology and empirical setup of this study.

## 3. Methodology

In this section, we provide detailed descriptions of the proposed empirical study. Section 3.1 explains the problem statement. Section 3.2 and 3.3 provide details of the empirical setup, and evaluation setup, respectively.

### 3.1 The problem statement

The CCP is binary classification problem where all the customers are divided into two possible behaviors: (i) Churn, and (ii) Non-Churn. Further, the churn behavior can be classified into the

following sub-categories (a) voluntary customer churn, in which a customer decides to leave the service or even company, and (b) involuntary customer churn, in which the company or service provider decides to terminate a contract with customer [10], [25]. This study only addresses the voluntary customer churns due to difficulty in predicting the later type of customer churn, and also because it is easier to filter out the voluntary customer churn by simple queries. On the other hand, the literature revealed that existing studies have been published but still there is no agreement on choosing the best approach to handle CCP problem. To the best of our knowledge, there is no state-of-the-art study to focus on the uncertain samples for building CCP model.

### 3.2 Empirical setup

We designed an empirical study to evaluate the proposed CCP model where we have focused on uncertain samples in the given dataset. Fig. 1 visualize the overall process of the framework.

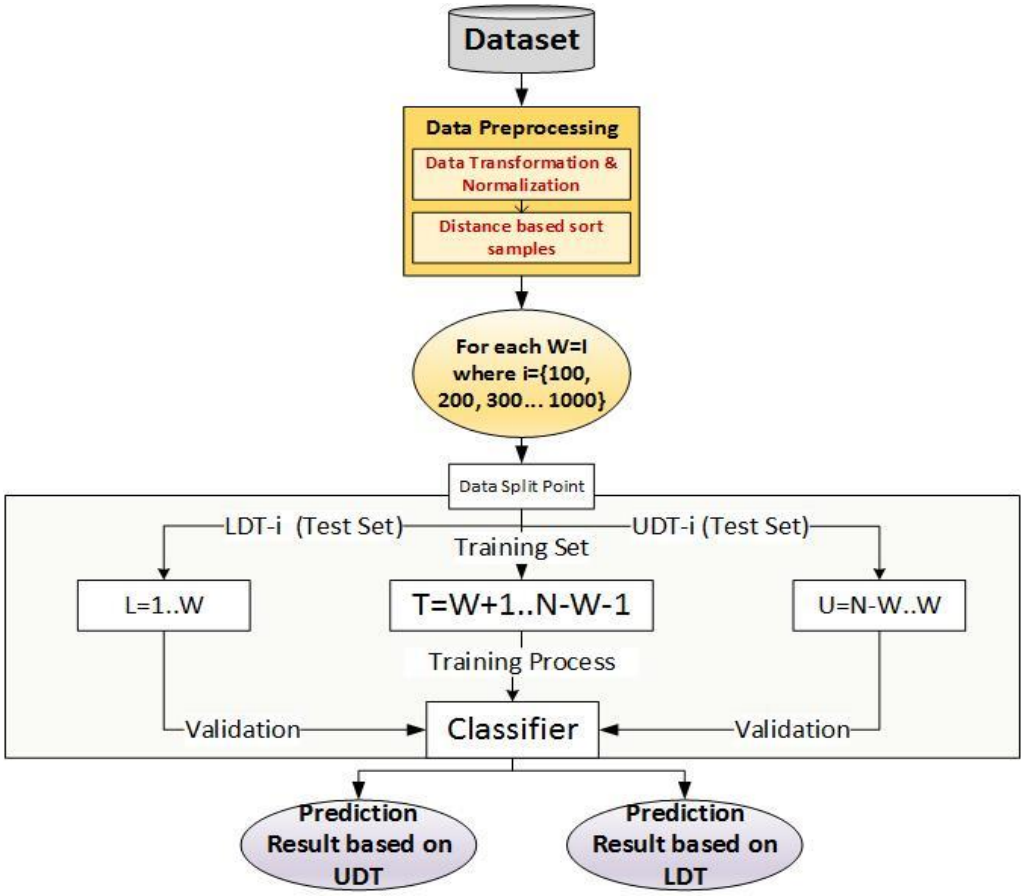



***Figure 1.*** *Visualization of the benchmark framework.* ***L, U*** *are referring to lower and upper side of the distance based sorted samples, respectively. Where* ***W*** *is representing the window size (sample of test set i.e., 100, 200, …1000) and* ***N*** *is the total number of samples in the training set. Where UDT and LDT stands for Upper and Lower distance samples, respectively.*

For this study, we used publicly available datasets. Table 1 describes detail about datasets. The dataset-1 consists of 3333 samples and each sample represent individual customer; whereas, the ratio of churn and non-churn customers is 85.5% and 14.49% respectively. Similarly, dataset-2 and 3 contains 7043 and 5783 samples, respectively. These datasets can be downloaded from data sources which are given in table 1.

**Table 1.** Summary of datasets

| | **Dataset-1** | **Dataset-2** | **Dataset-3** |
|---|---|---|---|
| No. of Samples | 3333 | 7043 | 5783 |
| No. of Attributes | 21 | 21 | 232 |
| No. of Class Labels | 2 | 2 | 2 |
| Percentage of Positive Samples | 14.49% | 73.46% | 87.84% |
| Percentage of Negative Samples | 85.51% | 26.54% | 12.14% |
| Data Sources | URL [1] | URL[2] | URL[3] |

### 3.2.1 Data preprocessing

In the preparation step, we have discretize, by size the values that exists in each attribute of the dataset, and then assigned certain labels e.g., Zero to Nine (0-9) possible values, to each discretized group. The discretizing by size leads to selecting the numerical attributes to nominal attributes and grouped them into specific size of bins. We then divide the total number of values in an attribute by size of bin. Ultimately, it produced specific list of values in different number of groups of an attribute. The step by step procedure for data preprocessing and discretization is elaborated as following:

1- Removed the attributes consisting unique values which represents identity of the sample.
2- Normalize the categorical values (such as ‘yes’ or ‘no’) into 0’s and 1’s where each value represents the corresponding category, then transformed the 0 and 1 into the same range and assign the same labels which applied for the rest of the attributes.
3- Find the distinct count of each value in every attribute, and also calculate the frequencies of these values in corresponding attributes.
4- Then divide the range of values into 10 possible groups and assigned 0-9 label to each group in all attributes.

[1] Data Source Link http://www.sgi.com/tech/mlc/db/

[2] http://www3.ntu.edu.sg/sce/pakdd2006/

[3] http://lamda.nju.edu.cn/yuy/dm07/assign_2.htm

### 3.3 Evaluation setup

In this section, a benchmarking framework is setup to present and evaluate the performance of the proposed study. These experiments were carried out using MATLAB toolkit[4] to fulfill the objectives of the proposed study by addressing the following research questions:

**RQ1:** Is the proposed approach capable to extract hidden most uncertain samples?

**RQ2:** Is it possible to design an effective CCP model under uncertain samples?

To address the aforementioned important research questions (RQ1, and RQ2), the following procedure is followed:

#### 3.3.1 Method for calculating the distances

To find the uncertain samples in the given dataset, first we calculated the distance of each instance from every other instance of the dataset using Manhattan distance formula. It can be generally expressed as;

$$P = [p_0, p_1, p_2 \dots p_n] \; and \; Q = [q_1, q_2, q_3 \dots . q_n]$$

$$D(p, q) = \sum_{i=1}^{i=n} |q_i - p_i| \tag{1}$$

Where *i=0, 1, 2, 3,…..n* and n is the total number of samples in given dataset. The Manhattan distance is the sum of absolute differences between points. Similarly, we have applied the Eq. 1 on the given dataset where ***p*** is considered one instance and ***q*** was considered as another instance from the same dataset. We calculated the distance between one instance with the rest of the instances and kept track of the distances. Fig. 2 illustrates the track of the obtained pair-wise distances of all samples. Wherein sample size of (i) and (j) are representing the instances of the same dataset while vertical dimension reflects the pair-wise ($p_i$ and $q_j$) distances between all the samples of the dataset. The representation in Fig.2 does not help in providing significant evidence to estimate the uncertain samples. It is, therefore, difficult to identify the lower and upper distances samples in a given dataset using the said figure. Therefore, we have calculated sum of the distance between individual instance and remaining samples using equation (2) in order to clearly investigate and visualize the uncertain samples in the dataset depicted using figure 3.

[4] http://www.mathworks.com/

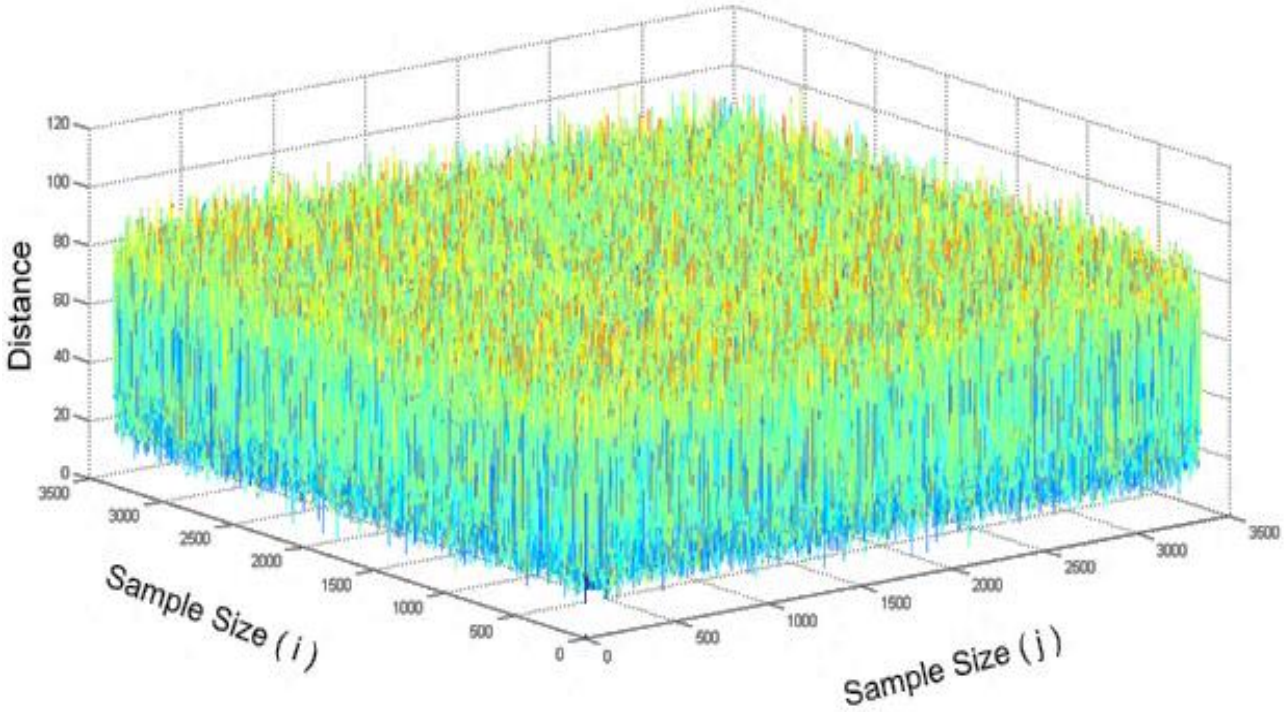



***Figure 2.*** *Shows the distances between one sample against the rest of instances.*

$$d(i,j) = \sum_{k=1}^{k=18} \left| r_{ik} - r_{jk} \right| \quad (2)$$

Where ***d(i , j)*** is the sum of distance between each instance with all instances while ***r*** means the row (sample) and ***k*** is referring to attribute of sample. Fig. 3 reflects the sum of the distances of each instance from all the instances.

As we can see in Fig. 3 that the distances of each instance vary greatly from each other and can be viewed as three visible groups of unordered samples which are; (i) samples at lower distance side, (ii) samples at upper distance side, and (iii) samples that are in the middle and can be seen as the major part of the samples.

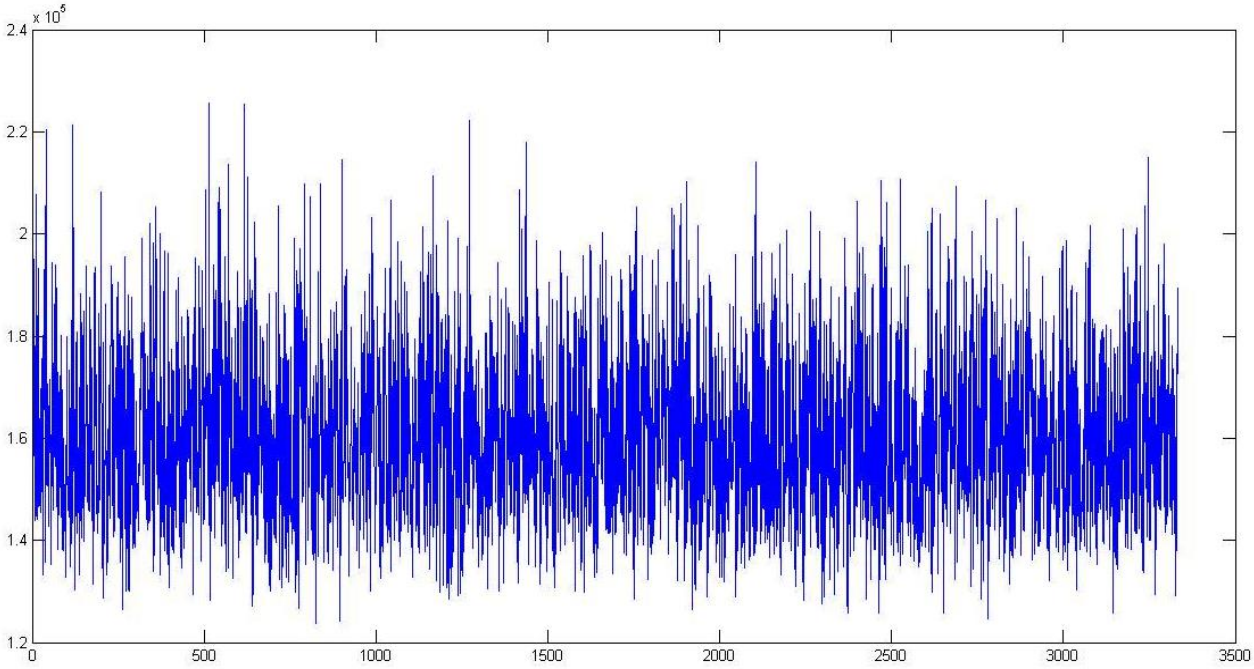



***Figure 3.*** *Distances of each instance from all the instances, where x-axis represents sample size while the y-axis reflects the distance between samples.*

Therefore, we have sorted all the instances based on the obtained distances ***d*** (the summation of distance of each instance against all the instances) in ascending order. Fig. 4 shows the sorted sum of distance of all instances from each other.

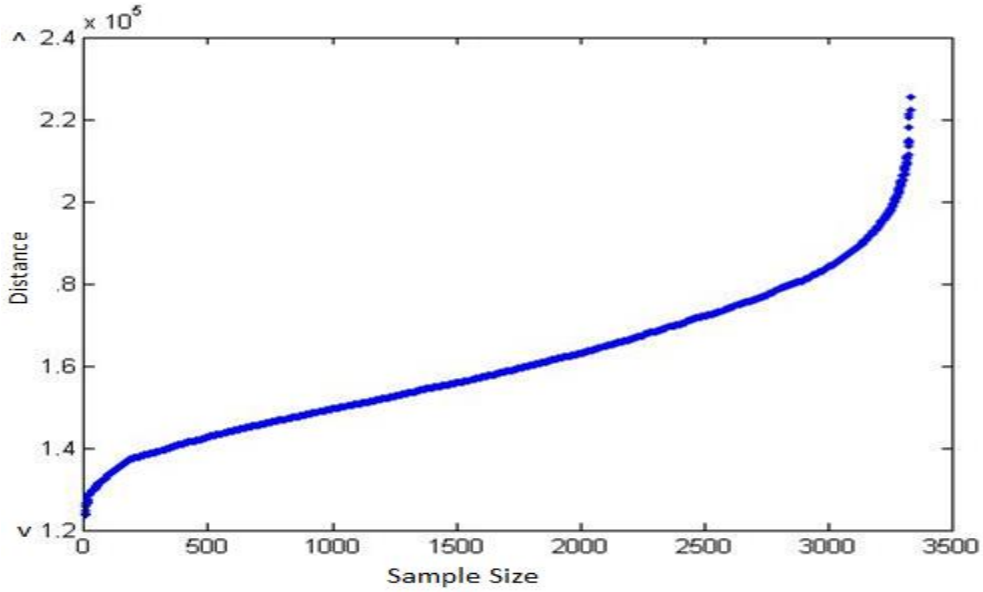
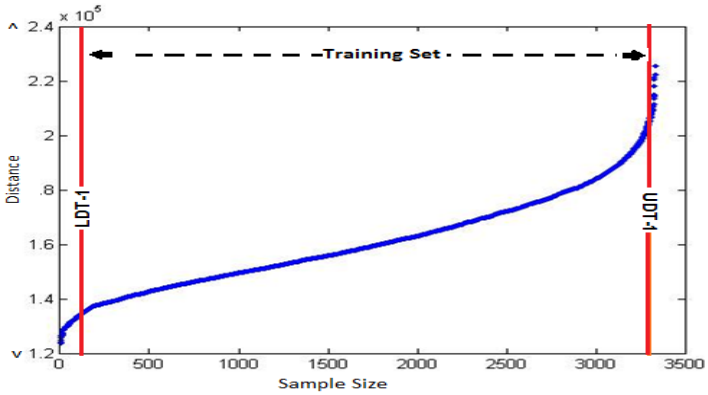



***Figure 4.*** *Sorted sum of distances of each instance against all instances, where x-axis represents the samples size and y-axis reflects the sum of distance. The values of both axes are in increasing order. Where sign ^ and ∨ representing upper and lower bound of distances, respectively.*

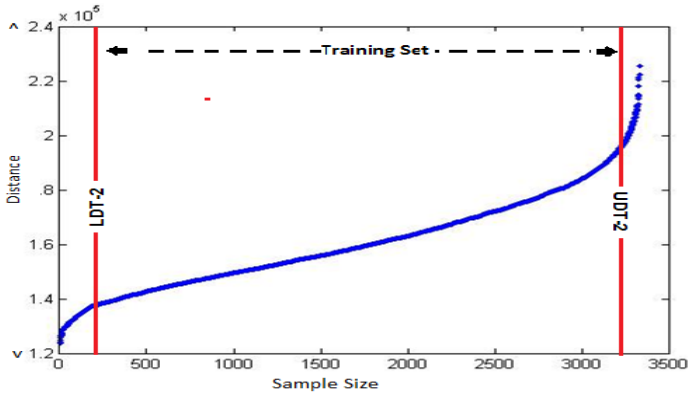
### 3.3.2 Selection of samples for building CCP model

The splitting of dataset into training and test sets is a common process of ML for building predictive model [30]. The training set is usually used to train the model while test set is used in order to estimate how well the proposed model has been trained (performance evaluation of the model).

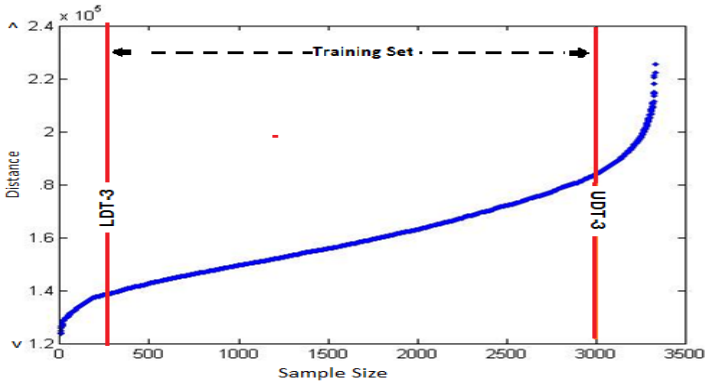
In this study, we have introduced a novel procedure for training and validation process for finding the uncertain samples as well as its impact on the classification performance. We pretend the test set is new data where the value of the class label is obtained d from the proposed predictive model. We then collected the predictions result from the trained classifier on the inputs from test sets samples and compared them to the obtained results of test set. This process allows us to evaluate the performance of the proposed model on the given test sets. For this purpose, initially, the first 100 samples from the lower distance are selected and last 100 samples are selected from the upper distance as the test set while the rest of the samples are used as training set. Then we have used Naïve Bayes (NB) as base classifier which is based on the simplest statistical Bayesian theorem [31], studied since 1950. It assumes that the entire input variables are mutually correlated and contribute in the binary classification problems. The base classifier is used to classify both non-churn and churn customer in the TCI dataset. For this purpose, then the base classifier "Naïve Bayes" is trained on the training sets to validate the performance of the classifier on the both test sets (LDT-i and UDT-i, where i=1) separately. After first iteration, samples are incremented by next 100 samples from the lower distance and 100 samples from the upper distance to the test sets (LDT-i and UDT-i, where i=2) and so on. At this stage, we are unable to address the RQ1 (Is the proposed approach capable to extract hidden most uncertain samples?) because there are two possibilities for declaring uncertain samples, they are: (i) can we consider uncertain samples the most wider distance samples from the majority samples, or (ii) can we consider uncertain samples the most closer distance samples to the majority samples. For this purpose, we recorded the performance of base-classifier on each iteration. Ultimately, it provided CCP model which predicted the class label of test sets samples with varying performances resulting in reflecting the predictive power of CCP model. The whole process of the training set and test set are visualized in Fig. 5.


x 10^5
Training Set
LDT-1
UDT-1
Distance
Sample Size

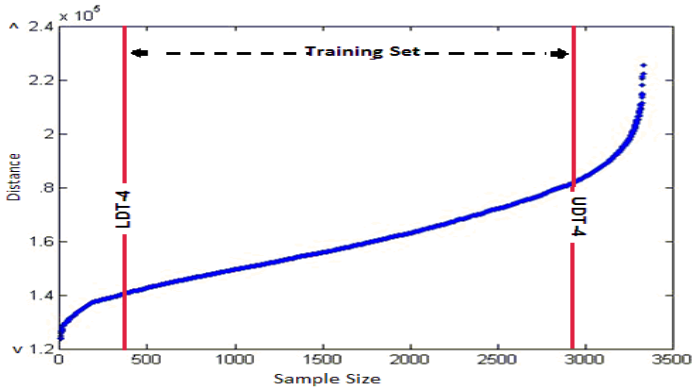
(a)


x 10^5
Training Set
LDT-2
UDT-2
Distance
Sample Size

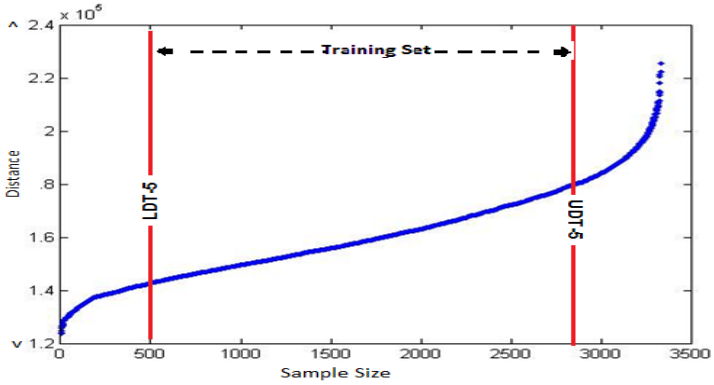
(b)


x 10^5
Training Set
LDT-3
UDT-3
Distance
Sample Size

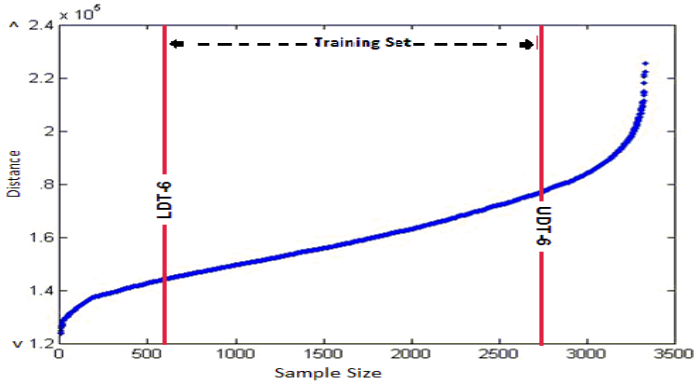
(c)


x 10^5
Training Set
LDT-4
UDT-4
Distance
Sample Size

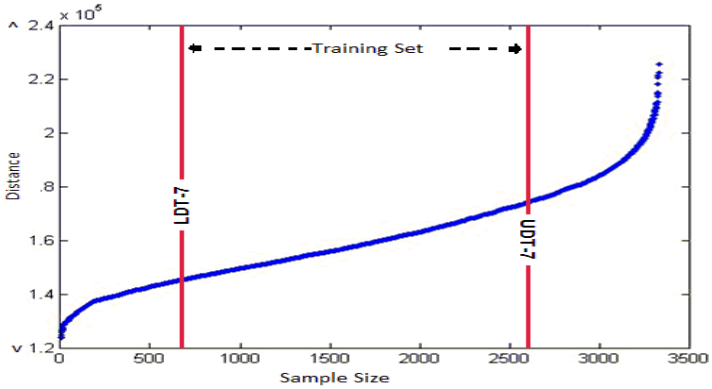
(d)

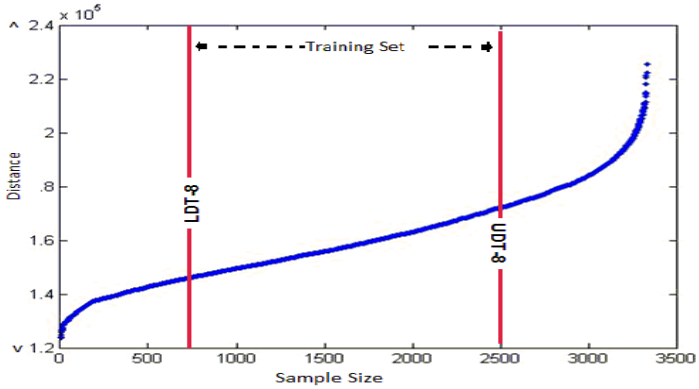

x 10^5
Training Set
LDT-5
UDT-5
Distance
Sample Size

(e)

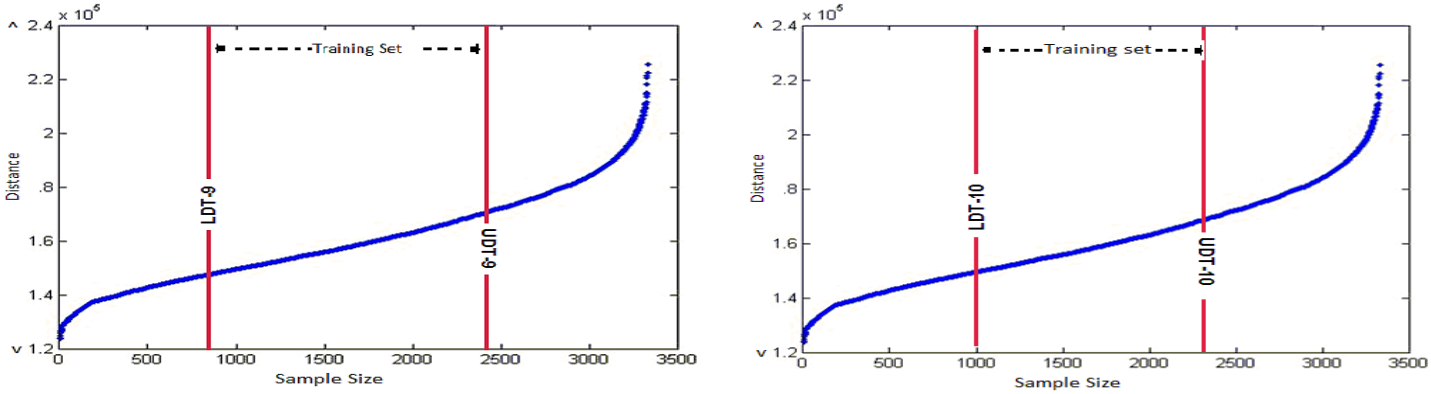

x 10^5
Training Set
LDT-6
UDT-6
Distance
Sample Size

(f)


x 10^5
Training Set
LDT-7
UDT-7
Distance
Sample Size


x 10^5
Training Set
LDT-8
UDT-8
Distance
Sample Size

(g) (h)




(i) (j)

***Figure 5.*** *Visualization of the training and test samples. Where (a) to (j) represents the training and test sets on each iteration while LDT-i and UDT-i (i=1, 2, 3,….10) represents the lower distance test set samples and upper distance test set samples respectively.*

## 4. Results and discussion

In this section, we have explored the results of the proposed empirical study and evaluated these results through the state-of-the-art evaluations measures (i.e., through precision, recall, f-measure, and accuracy). The detail about these evaluation measure given in Eq. 3, 4, 5, and 6 can be obtained from the studies referred as [10], [32]. The Eq. 3 mathematically expresses the precision measure, used to evaluates the correct degree of prediction power of the proposed model. On the other hand, Eq. 4 represents the recall measure, which is very important because prediction models intend to predict true churn customers as much as possible. However, there exists trade-off between precision and recall. Therefore, a comprehensive measure is required for precision and recall. Here, we use Eq. 6 (F-measure) that computes the harmonic mean of these two measures (e.g., precision and recall) resulting in achieving the balance between the said trade-off.

Table 2, 3 and 4 shows the CCP model performance on UDT and LDT samples on dataset 1, 2 and 3 respectively.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

$$F-measure = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{6}$$

**Table 2.** The performance CCP model based on LDT/UDT samples using Dataset 1

| Iteration Sequence | Sample Sizes | Precision % | | Recall % | | F-Measure % | | Accuracy % | |
|---|---|---|---|---|---|---|---|---|---|
| | | UDT | LDT | UDT | LDT | UDT | LDT | UDT | LDT |
| LDT-1 | 100 | 52.38 | 33.33 | 64.71 | 35.29 | 57.89 | 34.29 | 84.00 | 77.00 |
| LDT-2 | 200 | 48.72 | 30.56 | 61.29 | 42.31 | 54.29 | 35.48 | 84.00 | 80.00 |
| LDT-3 | 300 | 50.00 | 30.19 | 59.09 | 43.24 | 54.17 | 35.56 | 85.33 | 80.67 |
| LDT-4 | 400 | 52.78 | 36.00 | 58.46 | 49.09 | 55.47 | 41.54 | 84.75 | 81.00 |
| LDT-5 | 500 | 52.81 | 34.00 | 60.26 | 50.00 | 56.29 | 40.48 | 85.40 | 80.00 |
| LDT-6 | 600 | 51.38 | 37.93 | 58.95 | 52.38 | 54.90 | 44.00 | 84.67 | 81.33 |
| LDT-7 | 700 | 52.94 | 37.04 | 63.16 | 53.19 | 57.60 | 43.67 | 84.86 | 81.57 |
| LDT-8 | 800 | 53.46 | 37.58 | 63.43 | 56.19 | 58.02 | 45.04 | 84.63 | 82.00 |
| LDT-9 | 900 | 51.76 | 36.57 | 61.54 | 56.14 | 56.23 | 44.29 | 84.78 | 82.11 |
| LDT-10 | 1000 | 53.19 | 38.95 | 59.17 | 57.36 | 56.02 | 46.39 | 84.30 | 82.90 |

**Table 3.** The performance CCP model based on LDT/UDT samples using Dataset 2

| Iteration Sequence | Sample Sizes | Precision % | | Recall % | | F-Measure % | | Accuracy % | |
|---|---|---|---|---|---|---|---|---|---|
| | | UDT | LDT | UDT | LDT | UDT | LDT | UDT | LDT |
| LDT-1 | 100 | 47.73 | 45.45 | 87.50 | 83.33 | 61.76 | 58.82 | 74.00 | 65.00 |
| LDT-2 | 200 | 53.09 | 47.66 | 87.76 | 86.44 | 66.15 | 61.45 | 78.00 | 68.00 |
| LDT-3 | 300 | 53.03 | 48.00 | 90.91 | 80.90 | 66.99 | 60.25 | 77.00 | 68.33 |
| LDT-4 | 400 | 50.56 | 45.55 | 88.24 | 82.08 | 64.29 | 58.59 | 75.00 | 69.25 |
| LDT-5 | 500 | 49.77 | 47.44 | 85.94 | 81.02 | 63.04 | 59.84 | 74.20 | 70.20 |
| LDT-6 | 600 | 49.41 | 48.24 | 85.14 | 81.07 | 62.53 | 60.49 | 74.83 | 70.17 |
| LDT-7 | 700 | 49.49 | 47.65 | 82.95 | 78.76 | 62.00 | 59.38 | 74.43 | 70.29 |
| LDT-8 | 800 | 48.52 | 46.69 | 83.25 | 78.24 | 61.31 | 58.48 | 74.13 | 70.00 |
| LDT-9 | 900 | 50.13 | 47.82 | 81.66 | 79.44 | 62.13 | 59.70 | 74.67 | 70.44 |
| LDT-10 | 1000 | 50.48 | 47.38 | 82.42 | 80.37 | 62.61 | 59.62 | 74.80 | 70.60 |

**Table 4.** The performance CCP model based on LDT/UDT samples using Dataset 3

| Iteration Sequence | Sample Sizes | Precision | | Recall | | F-Measure | | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|
| | | UDT | LDT | UDT | LDT | UDT | LDT | UDT | LDT |
| LDT-1 | 100 | 43.75 | 21.52 | 06.03 | 72.65 | 10.61 | 33.20 | 88.20 | 65.80 |
| LDT-2 | 200 | 46.67 | 19.72 | 06.67 | 72.16 | 11.67 | 30.97 | 88.22 | 65.33 |
| LDT-3 | 300 | 42.86 | 19.56 | 06.45 | 73.81 | 11.21 | 30.92 | 88.13 | 65.38 |
| LDT-4 | 400 | 33.32 | 19.06 | 05.00 | 71.62 | 08.70 | 30.11 | 88.00 | 64.86 |
| LDT-5 | 500 | 27.27 | 20.66 | 04.55 | 73.53 | 07.79 | 32.26 | 88.17 | 65.00 |

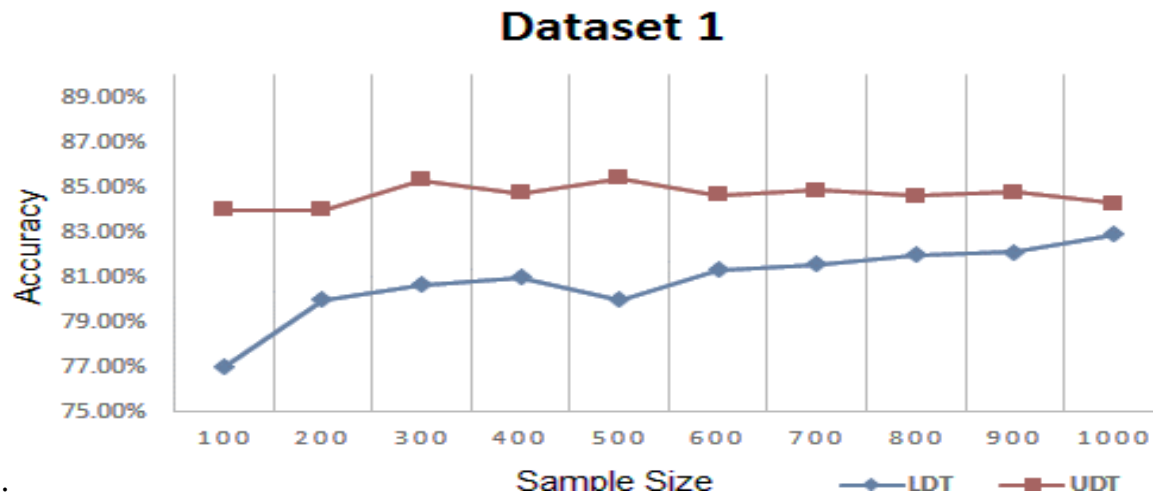
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| LDT-6 | 600 | 12.50 | 22.63 | 01.92 | 72.88 | 03.33 | 34.54 | 88.40 | 67.40 |
| LDT-7 | 700 | 16.67 | 21.15 | 00.22 | 70.21 | 03.92 | 32.51 | 87.75 | 65.75 |
| LDT-8 | 800 | 25.00 | 22.03 | 03.03 | 78.97 | 05.41 | 34.44 | 88.33 | 67.00 |
| LDT-9 | 900 | 25.00 | 22.78 | 05.00 | 81.82 | 08.33 | 35.64 | 89.00 | 67.50 |
| LDT-10 | 1000 | 04.76 | 19.44 | 01.10 | 87.50 | 01.79 | 31.82 | 89.01 | 70.00 |

It is investigated that CCP model provided different results on both UDT and LDT samples. However, when LDT samples were used as a test set in increasing order of the sample size. It obtained higher predictive power as compared to UDT samples of the same size of samples. Table 5 reflects overall performances of the base-classifier on multiple datasets and samples sizes (e.g., 100 to 1000 samples with 100 samples increased in test sets (i.e., UDT and LDT) samples on each iteration). The table 5 consists of five columns; whereas, first column represents the datasets used, second column describe the labels for both test sets samples, while third and fourth columns refers to the performance of the base-classifiers starting from initial samples size 100 for test sets up to 1000 samples. The fifth column "difference" in table 5 reflects the obtained accuracy when the size of the samples increases.

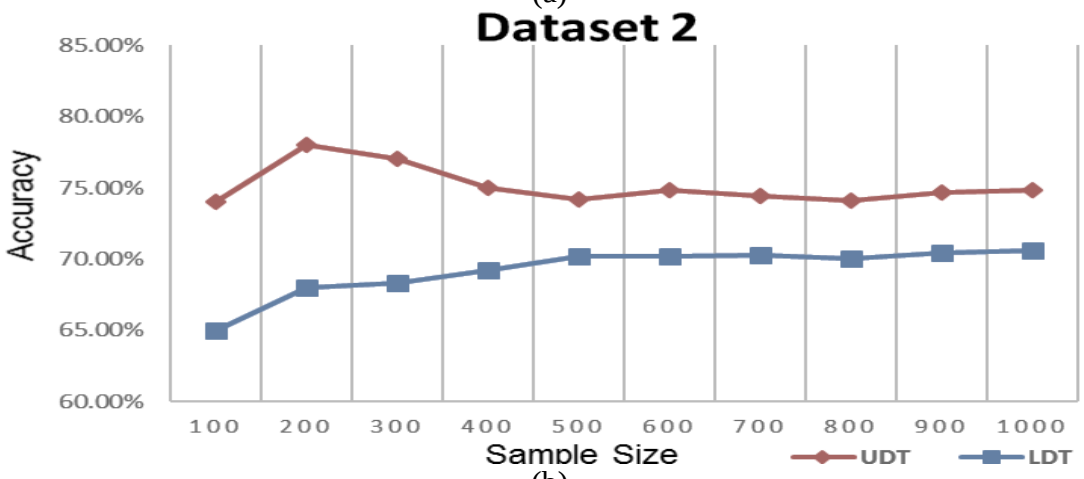
**Table 5.** Overall performance of the base-classifier on 100 to 1000 samples of UDT and LDT

| Datasets | Test Sets | Start | Last | Difference |
|---|---|---|---|---|
| Dataset 1 | LDT | 77.00% | 82.91% | 5.91% |
| | UDT | 84.00% | 84.30% | 0.30% |
| Dataset 2 | LDT | 65.00% | 70.60% | 5.60% |
| | UDT | 74.00% | 74.80% | 0.80% |
| Dataset 3 | LDT | 65.80% | 70.00% | 4.20% |
| | UDT | 88.20% | 89.01% | 0.81% |

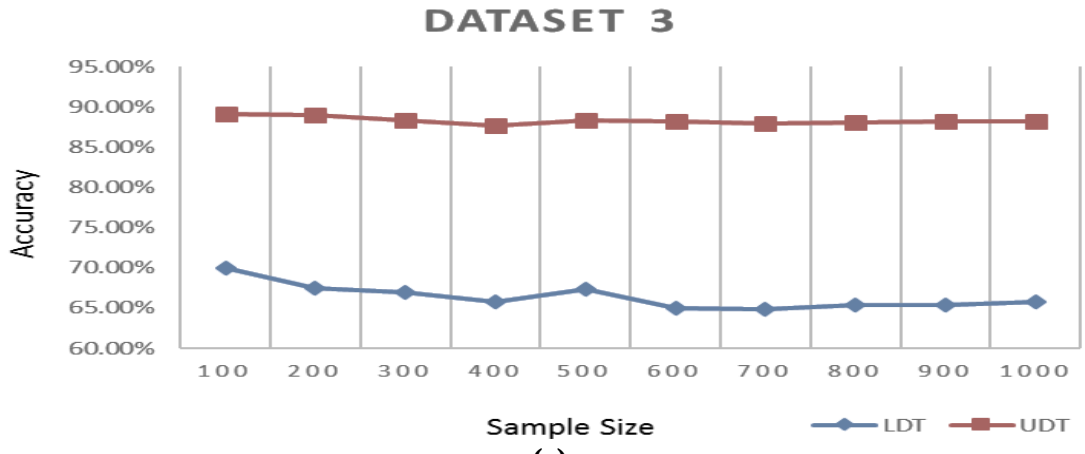
Finally, we have observed that UDT samples cannot be more affected on the performance of CCP model in TCI because if we see the differences obtained in the accuracies are 0.30%, 0.80% and 0.81% in dataset 1, 2 and 3 respectively. On the other hand, LDT achieved dramatic changes in the accuracies when the sample size increases such as 5.91%, 5.60% and 4.20% accuracy in dataset 1,2 and 3 respectively. Due to above analysis we also investigate some of the important reasons such as: (i) the accuracy of the CCP model using UDT samples cannot improve as compare to the LDT samples when the size of the samples increased, and (ii) the UDT samples are at upper or larger distance from the rest of the majority samples in term of distance. Therefore, the LDT are uncertain samples which is also reports to **RQ1**. On the other hand, the LDT samples are at lower or very close to the rest of the majority samples. It is investigated that the LDT samples have impact on the CCP model performance in TCI because the performance of the CCP model improved very smoothly in increasing order, when the size of the samples increased. Such finding also helps answer **RQ2**. Further, the performance of CCP model in uncertain situation in a dataset is inversely proportional to the accuracy. Additionally, the performance of CCP model under uncertain situation in a dataset (it is inversely proportional to accuracy) shown in Fig. 6.




(a)




(b)




**(c)**

***Figure 6.*** *The performance of CCP Models in term of accuracies on different size of UDT and LDT samples. Where (a), (b), and (c) represents the accuracies versus samples sizes of dataset 1, 2, and 3 respectively.*

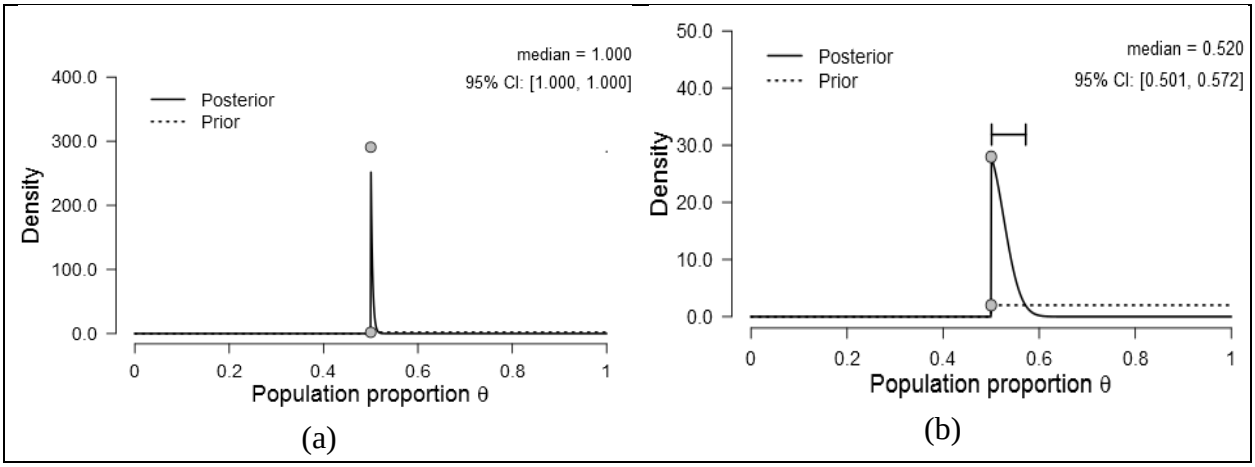
### 4.1 Bayesian Binomial Test

To statistically analyze all the empirical results of the proposed study of three datasets, which are reported in table 2, 3, and 4 while the detailed statistical information provided in table 1. These results are evaluating the performance of classifier, trained on different samples sizes in term of state-of-the-art evaluation measures (e.g., precision, recall, f-measure and accuracy) [10], [32]. However, to know whether the classification model resulting in purely guessing or whether the proportional correct is better than by chance. To do this, we also conduct the Bayesian binominal test [33].

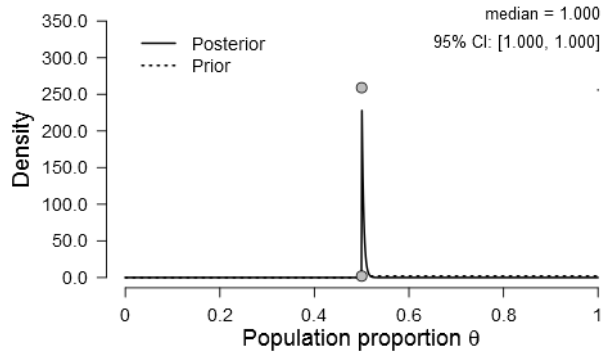
The Bayesian binomial test has only single parameter and an easy test to understand the distribution for the data. The Bayesian approach starts the estimation with prior distribution on the *p* parameter of interest, *p* could be a value between 0 and 1 with equal chance. Then, the posterior distribution of *p* is given through Bayes theorem given in eq.6 [33].

$$\pi_1(p|x_1 \dots \dots \ x_n) = \frac{f(x_1,\dots\dots,x_n|p)\pi_0(p)}{\int_0^1 f(x_1 \dots \dots x_n|p)\pi_0(p)dp} \tag{6}$$

The following configuration is set for applying the Bayesian binomial test; (i) first of all we set the test value to 0.5 which correspond to the traditional null hypothesis ($H_0$). In the proposed study, since there are two possible responses (i.e., churn and non-churn) and we want to know if the samples label purely guessing. Therefore, the test value is set to 0.5, then (ii) we have specified the alternative hypothesis ($H_1$) greater than test value because we interested to know the predicted label values through base-classifier is performing better than by chance, and (iii) finally, plots the prior and posterior analysis which is illustrated in figure 7.



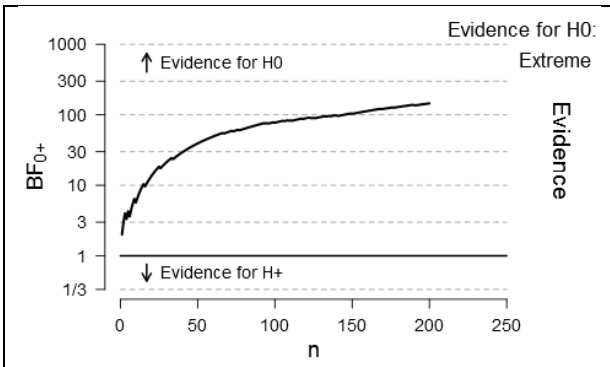

(a) (b)



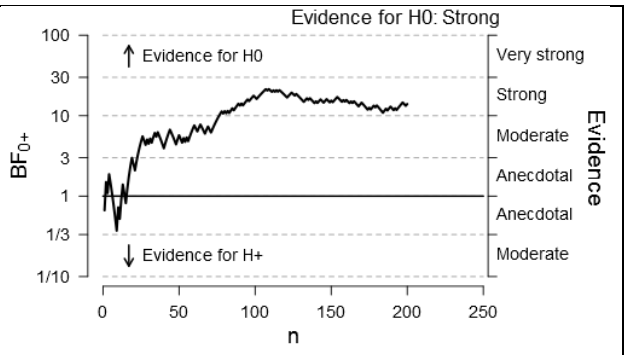

(c)

***Figure 7.*** *Plots of prior and posterior analysis where (a), (b) and (c) represent the pictorial representation of obtained analysis after applying Bayesian binomial test on dataset-1, 2 and 3 respectively. The prior distribution is the dotted lines and the posterior distribution is the solid lines.*

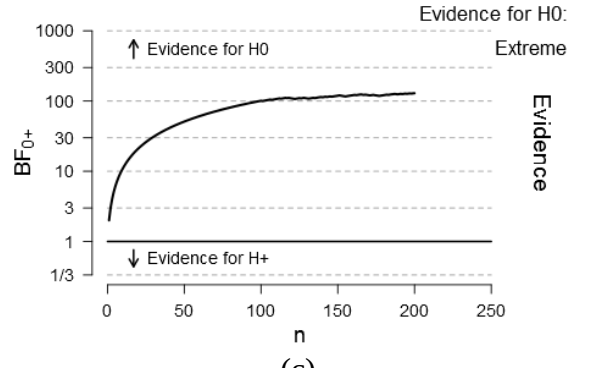
In figure 7, it is observed that the most of the posterior distribution is falls in between 0.4 and about 0.5. it is also noticed that there are two small circles on the plots. These small circles represent the height of the curve at the test value. Where the first small circle on the prior distribution is lower than on the small circle of the posterior distribution. This mean the Bayes factor supports the $H_0$. If the small circle on the posterior distribution had lower than the small circle of prior distribution, then the alternative hypothesis would be supported. It is also noticed that 95% confidence interval (CI) is considered for this test.




(a)




(b)




(c)

***Figure 8.*** *The sequential analysis plot represents the performance on both UDT and LDT samples on 1st iteration of base-classifier. Where (a), (b) and (c) reflects the dataset 1, 2 and 3 respectively.*

Finally, figure 8 shown the sequential analysis plotting where we can see graphical representation of Bayes factor along following by the CI=95%. The x-axis is the samples and y-axis is the Bayes factor and through figure 8 it can be tracked the Bayes factor as it changes after every data point. If the Bayes factor is the above 1 represents the evidence in the favor of the $H_0$ and it is below 1 then it shows the evidence in favor of the alternative hypothesis. So it can be observed from figure 8 (a) to 8(c) the obtained Bayes factor above 100 which shows the statistical evidence as extremely in the favor of $H_0$. On the other hand, figure 8 (b) obtained the Bayes factor as strong which also in the favor of the $H_0$ and rejects the $H_1$.

### 4.2. Threats to validity

- **Data preprocessing:** applying different discretizing technique instead of the proposed steps for data preprocessing may leads to variance in evaluation and classification process which may also impact on the predictive performance.
- **Base-classifier:** every classification algorithm has different mechanisms to classify the instances. We have arbitrarily chosen the Naïve Bayes classification algorithm as base-classifier and self-implemented it in MATLAB for the propose study. All the experiments are carried out using the same base-classifier. However, applying different classification algorithm may produce different results.

## 5. Conclusion

Uncertain samples have been shown to make a difference when applied to CCP in TCI. Through a novel method, we have extracted insightful uncertain samples, and build a CCP models. Further, we have empirically evaluated the impact of the uncertain samples on the performance of the predictive model in the context of TCI. In this study, we have also addressed the two important research questions (see Section 3.3) with empirical proof. Overall the propose study offers two main contributions to the existing literature such as: (i) introduced a novel method to identify efficiently uncertain samples in the TCI dataset, and (ii) an empirical comparison of the lower distance and upper distance samples in the context of their impact on the performance of the prediction model. Moreover, the performance of the resulting models was evaluated with four measures (see Section 4) which gave consistent and robust results. A benchmarking of the propose model for CCP in TCI using uncertain samples on this scale, does not have precedence in the current studies to the best of our knowledge. While an overview of the empirical results can be seen in Table 2, 3, 4 and 5 followed by fig 6. to show the effectiveness of our study in terms of four state-of-the-art evaluation measures e.g., precision, recall, accuracy, and f-measure.

Future studies might be able to provide empirical results on the balanced dataset with multiple base-classifier. It would be interesting to see what will be the effect on the CCP model if we apply the feature selection method by assigning weights to the features. Another future direction can be to test more comprehensive study with other types of models would offer the possibility to compare our results and eventually help to evaluate this effect statistically.

## References


[1] K. Coussement, S. Lessmann, and G. Verstraeten, "A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A case study in the telecommunication industry," *Decis. Support Syst.*, vol. 95, pp. 27–36, 2017.

[2] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture," *Expert Syst. Appl.*, vol. 85, pp. 204–220, 2017.

[3] R. R. Sharma, "Evaluating Prediction of Customer Churn Behavior Based On Artificial Bee Colony Algorithm," *Int. J. Eng. Comput. Sci.*, vol. 6, no. 1, pp. 20017–20021, 2017.

[4] M. A. H. Farquad, V. Ravi, and S. B. Raju, "Churn prediction using comprehensible support vector machine: An analytical CRM application," *Appl. Soft Comput.*, vol. 19, pp. 31–40, Jun. 2014.

[5] C.-S. Lin, G.-H. Tzeng, and Y.-C. Chin, "Combined rough set theory and flow network graph to predict customer churn in credit card accounts," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 8–15, Jan. 2011.

[6] C. J. C. Burges, D. Efimov, and R. Darrow, "Airline new customer tier level forecasting for real-time resource allocation of a miles program," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, Jun. 1998.

[7] O. Maria, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "A Comparative Study of Social Network Classifiers for Predicting Churn in the Telecommunication Industry," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1151–1158.

[8] M. Suznjevic, I. Stupar, and M. Matijasevic, "MMORPG Player Behavior Model based on Player Action Categories," in *Proceedings of the 10th Annual Workshop on Network and Systems Support for Games. IEEE Press*, 2011, pp. 1–6.

[9] A. O. Oyeniyi and A. B. Adeyemo, "Customer Churn Analysis In Banking Sector Using Data Mining Techniques," *African J. Comput. ICT*, vol. 8, no. 3, pp. 165–174, 2015.

[10] A. Amin *et al.*, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, 2017.

[11] G. Jaishankar, J. A. Mark, and E. R. Kristy, "Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers," *J. Mark.*, vol. 64, no. 3, pp. 65–87, 2000.

[12] L. Zhao, Q. Gao, X. Dong, A. Dong, and X. Dong, "K- local maximum margin feature extraction algorithm for churn prediction in telecom," *Cluster Comput.*, vol. 20, no. 2, pp. 1401–1409, Jun. 2017.

[13] A. D. Athanassopoulos, "Customer satisfaction cues to support market segmentation and explain switching behavior," *J. Bus. Res.*, vol. 47, no. 3, pp. 191–207, 2000.

[14] A. Idris and K. Asifullah, "Ensemble based Efficent Churn Prediction Model for Telecom," in *12th International Conference on Frontiers of Information Technology (FIT)*, 2014, pp. 1–7.

[15] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining," *IJCSI Int. J. Comput. Sci. Issues*, vol. 10, no. 2, pp. 165–172, 2013.

[16] H. Abbasimehr, "A Neuro-Fuzzy Classifier for Customer Churn Prediction," *Int. J. Comput. Appl.*, vol. 19, no. 8, pp. 35–41, 2011.

[17] Z. Kasiran, Z. Ibrahim, M. Syahir, and M. Ribuan, "Customer Churn Prediction using Recurrent Neural Network with Reinforcement Learning Algorithm in Mobile Phone Users," *Int. J. Intell. Inf. Process.*, vol. 5, no. March, pp. 1–11, 2014.

[18] S. Jamil and A. Khan, "Churn comprehension analysis for telecommunication industry using ALBA," in *ICET 2016 - 2016 International Conference on Emerging Technologies*, 2016, pp. 1–5.

[19] A. Idris, A. Khan, and Y. Soo, "Intelligent churn prediction in telecom : employing mRMR feature selection and RotBoost based ensemble classification," in *Springer Science Business Media New York.*, 2013, pp. 659–672.

[20] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 12547–12553, Dec. 2009.

[21] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2354–2364, Mar. 2011.

[22] B. He, Y. Shi, Q. Wan, and X. Zhao, “Prediction of customer attrition of commercial banks based on SVM model,” in *Procedia Computer Science*, 2014, vol. 31, pp. 423–430.
[23] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009.
[24] Y. Richter and N. Slonim, “Predicting customer churn in mobile networks through analysis of social groups,” in *Proceedings of the 2010 SIAM International Conference on Data Mining*, 2010, pp. 732–741.
[25] N. Lu, H. Lin, J. Lu, and G. Zhang, “A Customer Churn Prediction Model in Telecom Industry Using Boosting,” *J. Ind. Informatics*, vol. 10, no. 2, pp. 1–7, 2014.
[26] K. W. De Bock and D. Van den Poel, “An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction,” *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12293–12301, Sep. 2011.
[27] K. W. De Bock and D. Van den Poel, “Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models,” *Expert Syst. Appl.*, vol. 39, no. 8, pp. 6816–6826, Jun. 2012.
[28] P. C.Pendharkar, “Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6714–6720, Apr. 2009.
[29] I. Brandusoiu and G. Toderean, “Churn Prediction in the Telecommunications Sector using Support Vector Machines,” *Ann. ORADEA Univ. Fascicle Manag. Technol. Eng.*, no. 1, 2013.
[30] I. H. Witten, E. Frank, and M. a Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. 2011.
[31] J. Burez and D. V. d. Poel, “Data Mining Concepts and Techniques,” in *Pearson Education Asia Inc*, 2012.
[32] A. Amin *et al.*, “Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study,” *IEEE Access*, vol. 4, 2016.
[33] M. Zhu and A. Y. Lu, “The Counter-intuitive Non-informative Prior for the Bernoulli Family,” *J. Stat. Educ.*, vol. 12, no. 2, pp. 1–10, 2004.