Int. J. Advance Soft Compu. Appl, Vol. 11, No. 2, July 2019
ISSN 2074-8523

A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry

Nilam Nur Amir Sjarif [1,*], Muhammad Rusydi Mohd Yusof [1], Doris Hooi-Ten Wong[1], Suraya Ya'akob[1], Roslina Ibrahim[1] and Mohd Zamri Osman[2]

[1]Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia Kuala Lumpur
Jalan Sultan Yahya Petra
54100 Kuala Lumpur, Malaysia
Corresponding e-mail: nilamnur@utm.my

[2]Soft Computing Research & Intelligent System Group (SPINT), Faculty of Systems Computer & Software Engineering, University Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia

Abstract

Customer churn in telecommunication industry is actually a serious issue. The Telco company needs to have a churn prediction model to prevent their customer from moving to another telco. Therefore, the objective of this paper is to propose the customer churn prediction using Pearson Correlation and K Nearest Neighbor algorithm. The algorithm is validated via training and testing dataset with the ratio 70:30. Based on experiment, the result shows that the K Nearest Neighbor algorithm performs well compared to the others with the accuracy for training is 80.45% and testing 97.78%.

Keywords: ***Customer Churn Prediction, Pearson Correlation, Machine Learning, K Nearest Neighbor.***

1 Introduction

Customers are one of the most important asset in dynamic and competitive business [1]. Slogan "The customer is always right" which exhorts the company to give a high priority and best service to customer satisfaction. Thus, a comprehensive strategy for developing, managing and strengthening long-lasting customer relationship is under Customer Relationship Management team [2]. One of the challenges for telco company is to maintain the loyal of the customer. This is because losing customers could cause critical loss of income. As per contemplate by [3-5], retaining the old customer is five to six times less expensive than finding a new customer. Therefore, it is necessary for telco company to have capacity to predict customer that will be loyal to them without intervention action that could cause loss of income, extra expenses of customer retention and reacquisition, extra advertisement costs, organizational as well as planning and budgeting chaos [5]. Apart from that, customers that leave the telco company could impact others to do the same [6]. Subsequently, in order to sustain the customers' proclivity or tendency towards the company, they should consider the customers' behavior and provide the best services in respect of the customers' preferences. This is called the 'customer churn prediction'. It is critical to implement the churn prediction in their approach to forecast high risk customers.

Machine learning could be the sort of tools that could help telco company in churn prediction field. Machine learning is a part of artificial intelligence that provide the ability to allow computer learns the algorithm automatically without human involvement. Machine learning task such as classification allows computers to use existing data to forecast future behaviors, outcomes and trends. Classification is described as an activity that denote either a meta-scientific area of organizing the knowledge of a phenomenon into a set of separate classes, to structure the phenomenon and relate different aspects of it to each other, or in a nutshell is a discipline of supervised classification, which developing rules for assigning class labels to a set of entities under consideration [13]. The label of classification in churn prediction model is churn and non-churn customer. There are some study on churn prediction using machine learning algorithms [5]–[11]. Most of the study use supervised learning such as Decision Trees, Support Vector Machine (SVM), MultiLayer Perceptron Neural Network (MLP-ANN) and Naïve Bayes.

The organization of this paper includes 5 section. Section 1 include introduction, for section 2 include some previous study on churn prediction in terms of different algorithm. In section 3 and 4, we briefly explain our proposed algorithm. As for the results and evaluations, we will discuss elaborately in section 5. Section 6 of this paper focusing on the conclusion.

2 Related Work

In telecommunication industry, churn is one of the key solutions to remain competitive. Churn is used to create a model that accurately encompasses the customer-survival and customer-hazard functions to acquire insights on churn rate

[14]. According to Iyakutti et al, [15] the general definition of churn is the action of a customer terminating the customer's service due to dissatisfaction of the service offered or other companies providing better offers within the customer's budget. However, Ahmed et al [14] said that the customer's dissatisfaction is the main cause rather than the latter. In other words, churn prediction is a process of identifying the existing customer who are likely to discard the services soon. It will be a significant impact to the company's revenue if it loses customers.

Recently, a few of machine learning algorithm have been used for an accurate prediction in various fields. For example, in the banking sector, from the results of using Random Forest algorithm, an ensemble of Decision Trees has been reported to manage highly accurate prediction. However, the results in telecommunication sector are not as good as in the banking industry [15]. This is because the Random Forest algorithm works by generating trees on bootstrap samples which may not work well because of the imbalance of class distribution of telecommunication datasets.

In a separate study, customer churn prediction in telecommunication industry suffers from the eruption of enormous telecom dataset such as Call Detail Records (CDR) [15]. Telecom churn prediction has been recognized to be of different application domain to churn prediction in comparison to other subscription-based sector due to the variety, volume and biases of the dataset. The larger dimensionality comes when telecom operators gain information about customers, ranging from individual demographics to details of usage of service which results in many challenging problems regarding large number of features and imbalance class distribution of dataset.

In a previous study, Effendy et al, [16] proposed a method for handling imbalanced data problem to improve customer churn prediction. The incorporation of sampling and Weighted Random Forest (WRF) is being implemented in the proposed technique, which eventually result in balanced dataset and more accurate churn prediction. Incorporation of under-sampling and SMOTE (Synthetic Minority Oversampling Technique) is known as a sampling process whereby the process involves two vital processes which are sampling for imbalance data problem and classification of data through WRF for accurate prediction. Consequently, the F-measure and accuracy value increases shows that the elimination of data makes better prediction after employing the combined sampling process. However, it is not significant with the use of under-sampling technique.

Ning Lu et al, [17] have suggested a model that use a boosting technique for churn prediction. This methodology uses logistic regression as the foundation followed by boosting technique to improve model prediction. However, there a few sets of limitations for both techniques used which is non-consideration of class rarity and incompetence to conclude the reasons for churn respectively.

To enhance the customer churn and insurance fraud detection, Ravi et al, [18] suggested One Class SVM (OCSVM) based under-sampling technique. Sampling the data through OCSVM and applying machine-learning algorithm for classification resulted in a more accurate prediction of the Decision Trees than other classification algorithms with reduced system complexity.

An approach of neural network basis was suggested by Anuj Sharma et al, [19] for customer churn prediction. The data are modelled into nodes and employed using Clementine 12.0 and the over-training data problem is resolved in this approach by randomly selecting training data for network training. The evaluations prove high rate prediction of 92%. However, this model is performed for data reduction only without implementing reduction of dimensionality and eventually increase the model's complexity.

Another implementation of machine learning algorithm namely decision trees and logistic regression was carried out by Preeti K. Dalvi et al, [20]. The suggested approach is based on performing a combination and comparative analysis of machine learning algorithm. In order to determine the degree of which feature affects the decision, a machine learning algorithm called logistic regression is being used. Consequently, the decision tree is employed to deliver graphical overview of the data. The evaluation results prove that the prediction accuracy is improved and the time taken for churn prediction is reduced. However, the classification is limited to a few classes only.

Aleksandar Petkovski et al, [21] proposed four different type of machine learning algorithms for Macedonian telecom. The authors compared the four classifiers which are C4.5 Decision-Tree, k-nearest neighbor, Naïve Bayes and logistic regression in terms of area under the curve, execution time and rate of prediction. It is proven that all of these classifiers exceed 90% accuracy with logistic regression as the highest one. However, the long span of time to execute and the need of vast memory resources is the limitation for this classifier.

From the presented literature work, it can be concluded that the algorithm that could be applied on customer churn is machine learning. Machine learning can be used for increasing the performance of the prediction. Therefore, in this paper, we will propose a churn prediction using the Pearson Correlation Function with k-Nearest Neighbor. The propose methods will compare with other algorithm to define the best performance accuracy.

3 The Methodology

In this section, we briefly explain the steps of the churn prediction methodologies. Generally, in applying machine learning on Telco-Customer-Churn dataset, there are four steps need to be followed: 1) Data Preparation 2) Data Preprocessing, 3) Model Construction, 4) Prediction Accuracy. Fig 1 shows the flow chart of churn prediction model for this study.

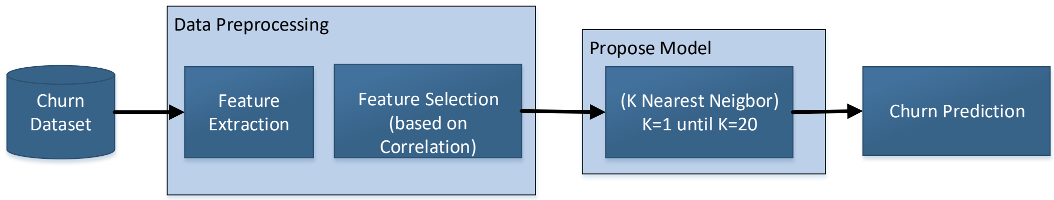

Fig 1: Churn Prediction Model Flow Chart

3.1 Data Preparation

The dataset used in this study is based on historical data. It is a public dataset from Kaggle dataset (https://www.kaggle.com/blastchar/telco-customer-churn) in CSV format and specifically stated as telco customer churn data. The dataset consists 7043 rows with 21 attributes. Details of attribute are listed in Table 1.

Table 1: Telco customer churn attribute list

Attributes	Type	Description
customerID	object	Unique number to represent customer
gender	object	Customer gender
SeniorCitizen	int64	Customer status based on age
Partner	object	Customer status based on partner
Dependents	object	Customer status based on dependency
tenure	int64	Customer period on using services from telco company
PhoneService	object	Status of having phone service
MultipleLines	object	Status of having multiple lines
InternetService	object	Status of having internet service
OnlineSecurity	object	Status of having online security service
OnlineBackup	object	Status of having online backup service
DeviceProtection	object	Status of having device protection service
TechSupport	object	Status of having technical support service
StreamingTV	object	Status of having streaming tv service

StreamingMovies	object	Status of having streaming movies service
Contract	object	Status of having contract
PaperlessBilling	object	Status of billing
PaymentMethod	object	Method of payment by customer
MonthlyCharges	float64	Customer monthly charges
TotalCharges	object	Customer total charges
Churn	object	Customer status

Based on Table 1, there is 18 attributes categorized as object. Object type attribute consist categorical data which put each customer into certain group. Other type of data in dataset is int64 and float64 which means the data of the attribute can be calculated.

3.2 Data Preprocessing

In this phase, the selection of attributes is done using the standard correlation function method called 'Pearson Correlation Coefficient'. As shown at Table 1 above, only three attributes are set as numeric value while the others are set as object type. Therefore, during the preprocessing, the features TotalCharges is converted into a numerical data type. During preprocessing also, 11 data values are missing from TotalCharges rows. Hence, it is necessary to clean up any unwanted data before next prediction model. Next, training and testing of dataset will be used in this study to model the data using propose algorithm. The ratio is 70:30. Based on the theory, the technique will evaluate the predictive models by splitting the original sample into a training set to train the model and a test set to evaluate it. The average of all the process is produced as the result of the model.

4 The Proposed Method

4.1 K Nearest Neighbor (KNN)

Churn prediction model is built using K Nearest Neighbor (KNN) algorithm. One of the advantages of this algorithm is it can do classification task without prior knowledge about distribution of data. KNN method help to predicts the property of a substance in relation to the experimental data for most similar compounds [22]. In a nutshell, KNN is an instance based learning or lazy learning. As a lazy algorithm, KNN is best suited when having the entire training data. Choosing the best value for k is depending on the given data. Generally, larger values of k reduce effect of the noise on the classification but make boundaries between classes less distinct [23].

K-Nearest Neighbor (KNN) is an algorithm that classify group of data point which is based on similarity measure. KNN algorithm tries to determine if a point is in group A or B depends on one point on a grid. The range is arbitrarily determined, but the point is to take a sample of the data. If most of the points are in group A, then it is likely that the data point will be group A rather than group B, and vice versa. The equation could be defined as;

Theorem 3.1 *KNN*

$$y(d, C_j) = \sum_{d_i is\ in\ kNN} Sim(d, d_i) \times\ y(d_i, C_j) - b_j\ ; \quad (1)$$

$$(d, c) = \begin{bmatrix} 0, if\ \ d\ is\ in\ class\ C \\ 1, if\ \ d\ is\ not\ in\ class\ C \end{bmatrix}$$

Where C_j is the class j (in this field we have two classes, spam and ham) and the amount of y (d, C_j) is achieved from the right hand of first formula and shows that d is in class C_j or not. Moreover, b_j is the predetermined threshold of C_j. The similarity of two entities in the KNN algorithm $Sim(d, d_i)$ is generally calculated according to the Euclidean distance. The evaluation accuracy is based on the different values for K (the number of neighbors to use) which would vary from 1 until 20 for the analysis. The error rate and misclassification rate for various k value is calculate with validation data to pick the k value that has the best classification performance. Normally, values of k fall between the range 1 to 20. The optimum value of k is lower as the complexity and irregularity of data structure increase [26]. KNN classifiers perform well with large dataset due to their simplicity and free from parametric assumptions. The drawback of KNN is lack of data generalization [27]. This paper will compare the model using two types of evaluation which are accuracy and confusion matrix. Accuracy is measured by the ratio of correct predictions to the total number of cases evaluated. Thus, the higher the accurate the result. Table 2 shows the calculation for the confusion matrix performance.

Table 2: Confusion matrix Churn prediction Model

Hypothesized Class	Actual Non-churners	Actual Churners
Non-churners (N)	True Negative (TN)	False Positive (FP)
Churners (P)	False Negative (FN)	True Positive (TP)

5 Results, Analysis and Discussions

The experiment of the propose method is compared based on accuracy and confusion matrix. Two class of customer churn prediction include such as churn ‘Yes’ and No churn ‘No’. As illustrate in Fig 2, total number of customer defined as not churn is 5174 and 1869 is churn. While, the result of correlation between attributes based on Pearson Correlation is presented in Fig 3.

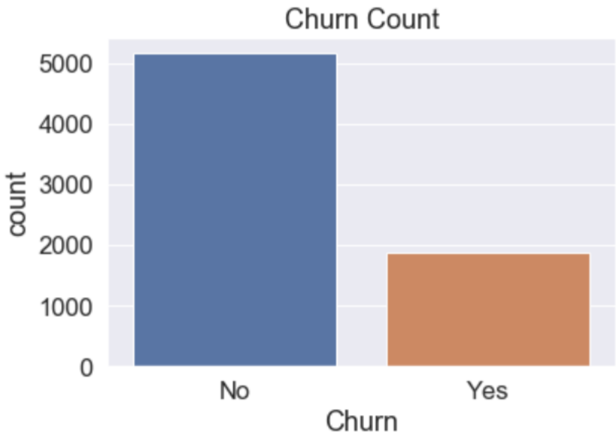

Fig 2: Total Customer Churn or Not Churn

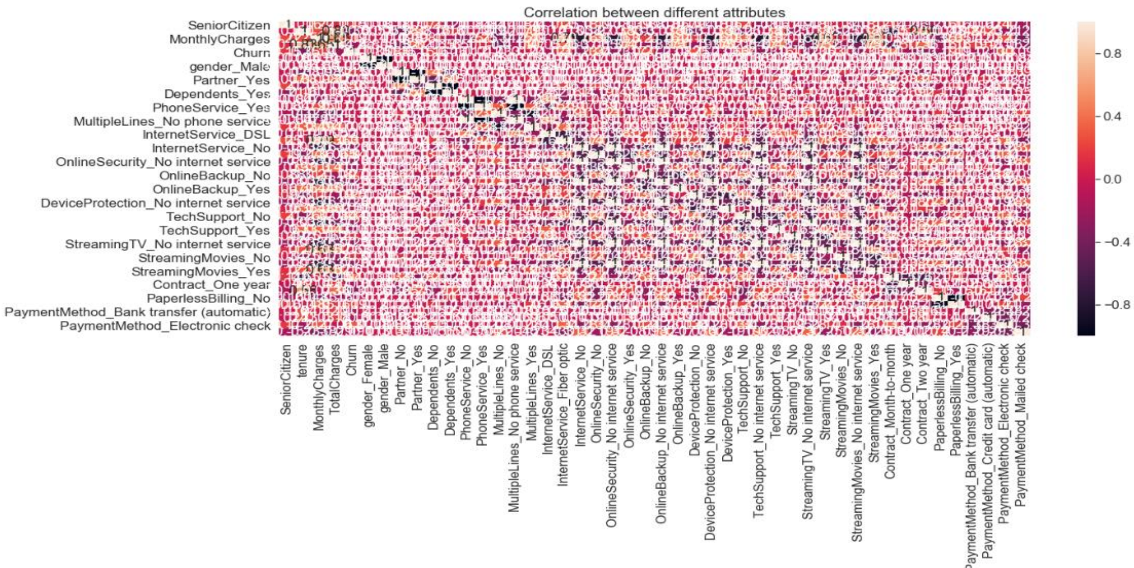

Fig 3: Correlation between attributes

From the result correlation between attribute, we select the relevant feature attribute for further discuss. Fig 4 below shows the four features correlation that possible for customer to churn. Based on the correlation analysis between features of SeniorCitizen (No=0, Yes=1) and features tenure, it shows that for senior citizen (1) will stay longer with the same telco rather than the young citizen (0). If we look at to the correlation between features SeniorCitizen with features MonthlyCharges, the results show that the young citizen will churn if the rate for monthly charges is high compared to the senior citizen.

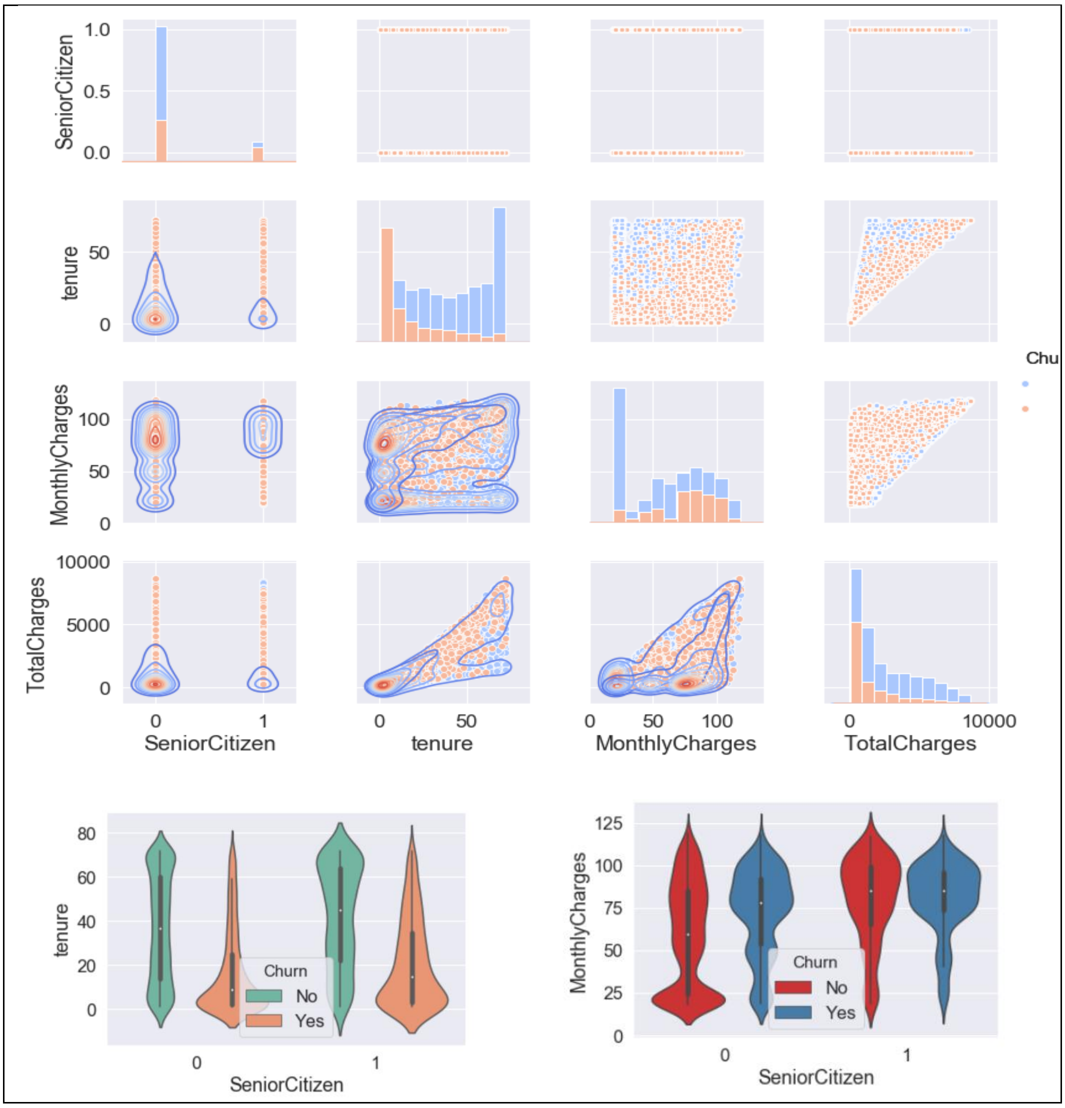

Fig 4: Correlation Features between SeniorCitizen, Tenure, MonthlyCharges and TotalChargers

We analyze also the correlation features between gender and MonthlyCharger. And next the correlation features between SeniorCitizen and Internet Service. Fig 5 present the result of the features. Based on the results, the young people either female and male could be churn to other telco. For next results, the correlation shows tha Fiber optics is the most favorable services especially for the young citizen.

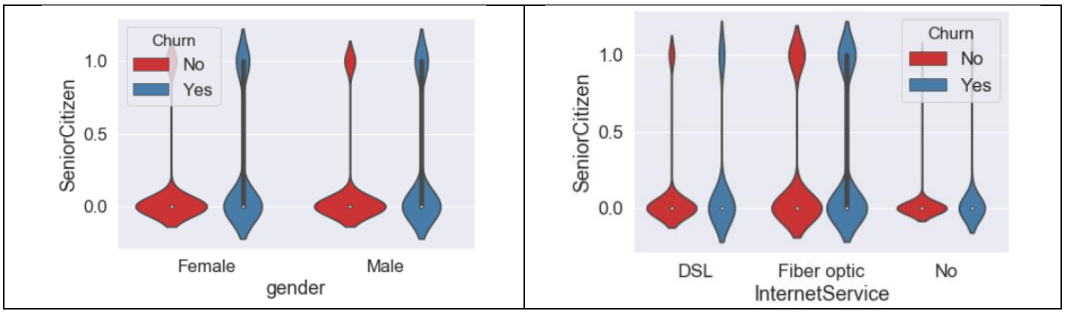

Fig 5: Correlation between gender, Senior Citizen and Internet Service

Next, Fig 6 shows the result accuracy based on different number of k. The number of k is set between the range of 1 until 20. Based on the graph, it shows that the best result of accuracy 80.45% is achieved for the training when k=18. Meanwhile, the best result accuracy 97.78% is achieved for the testing when k=1. Fig 7 shows the result of the confusion matrix for churn prediction using KNN algorithm.

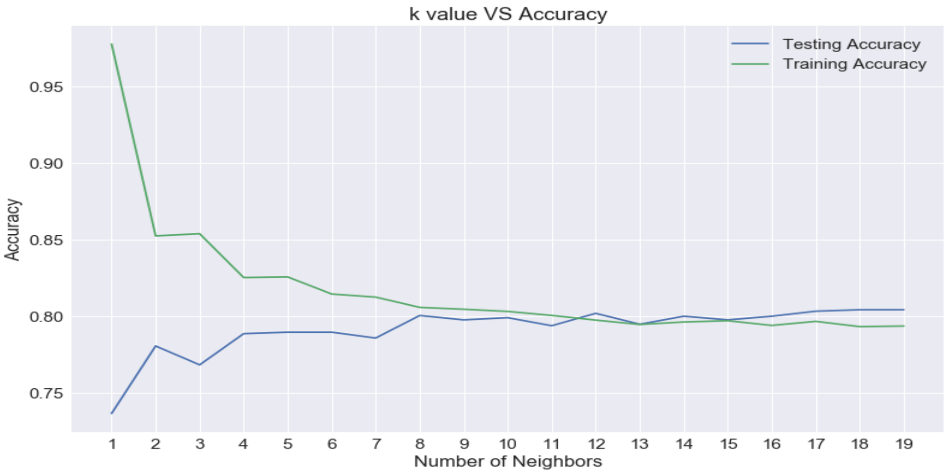

Fig 6: Result Accuracy for Training and Testing with k=1 until k-20

Confusion matrix plot of KNN

1287	298
258	270

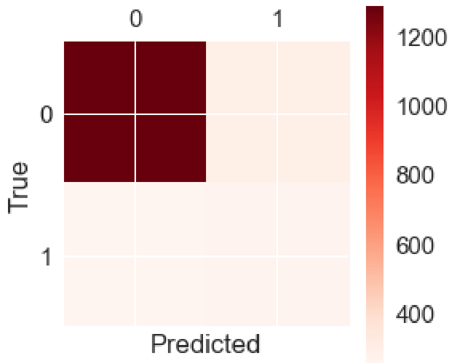

Fig 7: Confusion Matrix for KNN

Next, Table 3 present the evaluation comparison of the performance other algorithm include Random Forest and Support Vector Machine.

Table 3: Result Comparison Model

Algorithm	Training	Testing
Propose KNN	80.45%	**97.78%**
Random Forest	76.36%	76.85%
SVM	**83.67%**	79.41%

Based on Table 3, it shows the propose algorithms generate high prediction accuracy over than 80% at the training phase. The churn prediction using SVM generates 83.67%, KNN 80.45% and Random Forest 76.36%. During the testing phase, the best accuracy is achieved by KNN algorithm with the result 97.78%, while the two algorithms is achieved below 80%, where Random Forest 76.85% and SVM 79.41%.

6 Conclusion

In this paper, we have compared several algorithms that can predict whether the customer will terminate the service and choose another organization or not. The comparison of several classifiers will help us to accurately predict customer churn as well as addressing the main factor that leads to customer retention. Initially, the Pearson Correlation Function is performed in pre-processing phases. Then, we are analyzing all the classifiers in several aspects. Based on the results and evaluation, we observed that the KNN algorithm outperforms the others with the accuracy for training is 80.45% testing 97.78%. Also, it is proven that it is vital for the organization to run more promotion for any day plan services.

ACKNOWLEDGEMENTS

The authors would like to thank Ministry of Higher Education (MOHE) and Universiti Teknologi Malaysia (UTM) for their educational and financial support. This work is conducted at Razak Faculty of Technology and Informatics, under Cyberphysical Sytems Research Group (CPS RG) and funded by Universiti Teknologi Malaysia Grant (GUP Tier 1: Q.K130000.2538.18H42)

References

[1] K. Coussement, S. Lessmann, and G. Verstraeten (2017). 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.

[2] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas (2015). A comparison of machine learning techniques for customer churn prediction. Simul. Model. Pract. Theory, vol. 55, pp. 1–9,

[3] P. Kotler (1994). Marketing management: analysis, planning, implementation, and control, 8th ed. Englewood Cliffs, N.J: Prentice Hall.

[4] Almana, Mehmet Sabih Aksoy Amal, and Rasheed Alzahrani (2014). A Survey on Data Mining Techniques In Customer Churn Analysis For Telecom Industry.

[5] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui (2013). Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining," IJCSI Int J Comput Sci Issues, vol 10, pp. 165–172

[6] I. Nitzan and B. Libai. (2011). Social Effects on Customer Retention. Journal of Marketing, 75(6), pp. 24–38

[7] P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar, and V. A. Kanade (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. in 2016 Symposium on Colossal Data Analysis and Networking (CDAN), pp. 1–4.

[8] A. Ahmed and D. M. Linen (2017) "A review and analysis of churn prediction methods for customer retention in telecom industries," in 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1–7.

[9] U. Yabas, H. C. Cankaya, and T. Ince (2012). Customer Churn Prediction for Telecom Services. in 2012 IEEE 36th Annual Computer Software and Applications Conference, pp. 358–359.

[10] A. Amin et al. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, vol. 237, pp. 242–254

[11] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi (2014). Improved churn prediction in telecommunication industry using data mining techniques. Appl. Soft Comput., vol. 24, pp. 994–1012.

[12] S. J. Russell, P. Norvig, and E. Davis, Artificial intelligence: a modern approach, 3rd ed. Upper Saddle River: Prentice Hall, 2010.

[13] B. Mirkin, “Data analysis, mathematical statistics, machine learning, data mining: Similarities and differences,” in 2011 International Conference on Advanced Computer Science and Information Systems, 2011, pp. 1–8.

[14] Ahmed, Ammar & Maheswari Linen, D. (2017). A review and analysis of churn prediction methods for customer retention in telecom industries. 1-7

[15] K. Iyakutti and V. Umayaparvathi (2016). Attribute selection and Customer Churn Prediction in telecom industry. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, 2016, pp. 84-90.

[15] A. Backiel, Y. Verbinnen, B. Baesens, and G. Claeskens (2015). Combining Local and Social Network Classifiers to Improve Churn Prediction. 2015 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Min. Comb., pp. 651–658, 2015

[16] V. Effendy, Adiwijaya, and Z. K. A. Baizal. (2014)Handling Imbalanced Data in Customer Churn Prediction Using Combined Sampling and Weighted Random Forest,” 2014 2nd Int. Conf. Inf. Commun. Technol. Handl., pp. 325–330

[17] N. Lu, H. Lin, J. Lu and G. Zhang. (2014).A Customer Churn Prediction Model in Telecom Industry Using Boosting. in IEEE Transactions on Industrial Infor/matics, 10(2), pp. 1659-1665

[18] G. G. Sundarkumar, V. Ravi and V. Siddeshwar. (2015). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). pp. 1-7.

[19] A. Sharma (2013). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. arXiv:1309.3945

[20] P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar and V. A. Kanade (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. 2016 Symposium on Colossal Data Analysis and Networking (CDAN). pp. 1-4.

[21] A. J. Petkovski, B. L. R. Stojkoska, K. V. Trivodaliev, and S. A. Kalajdziski (2016). Analysis of churn prediction: A case study on telecommunication services in Macedonia. 2016 24th Telecommun. Forum, pp. 1–4.

[22] A. Manganaro, F. Pizzo, A. Lombardo, A. Pogliaghi, and E. Benfenati, (2016). Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm. Chemosphere, vol. 144, pp. 1624–1630.

[23] B. Everitt, Ed., Cluster analysis, 5th ed. Chichester, West Sussex, U.K: Wiley, 2011

[24] A. J. Petkovski, B. L. R. Stojkoska, K. V. Trivodaliev, and S. A. Kalajdziski (2016). Analysis of churn prediction: A case study on telecommunication services in Macedonia. 2016 24th Telecommun. Forum, pp. 1–4.

[26] G. Shmueli, N. R. Patel, and P. C. Bruce (2016). Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner. John Wiley & Sons, 2016.

[27] R. J. Roiger (2017). Data Mining: A Tutorial-Based Primer, Second Edition. CRC Press.

[28] Ahmed Qureshi, Saad & Saleem Rehman, Ammar & Qamar, Ali & Kamal, Aatif & Rehman, Ahsan. (2013). Telecommunication Subscribers' Churn Prediction Model Using Machine Learning. 8th International Conference on Digital Information Management, ICDIM 2013. 10.1109/ICDIM.2013.6693977.