A Framework to Enhance Accuracy of Customer Churn Prediction in Telecom Industry

Kholoud T. Mahmoud
Department of Business Information System, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

Shimaa Ouf
Department of Business Information System, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

Manal A. Abdel-Fattah
Department of Information Systems, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

ABSTRACT
Customer Churn Prediction problem is a long-standing challenge for Different communities, there are many groups in the scientific and commercial communities like telecom sector trying to improve Predictions. The primary motivation is the dire need of businesses to retain existing customers, coupled with the high cost associated with acquiring new one.

The machine learning techniques have a significant impact onimproving and predicting customer data mining techniques to improve customer retention, but thesetechniques face a lot of challenges in terms of accuracy.

This study aimed to enhance prediction and detection using a comparative study on the most popular supervised machinelearning methods , Support Vector Machine (SVM) andextreme Gradient Boosting (XGBoost) model to detectcustomer churn in IBM Watson dataset of telecom company.

This paper provides XG boost classifier which less focused in the previousworks. XG boost classifier is applied on publicly available telecom dataset and experiential results are compared with SVM Classifier. XG boost classifier performs superior out of SVM.The evaluated metrics such as Precision, Recall, F1-score. It yielded an accuracy of the framework reached 84%.

General Terms
Classification problem , Accuracy , Machine learning Algorithms

Keywords
Customer churn, Telecommunication, XG boost classifier, Classification, Churn Prediction

1. INTRODUCTION
The customers are considered one of the most important assets for a business in numerous dynamic and competitive companies within a marketplace [1]

In competitive market, companies in which the customers have numerous choices of service providers they can easily switch a service or even the provider. Such customers are referred to as churned customer[2]

Acquiring a new customer not only costs more (5–6 times than retaining a customer) but also requires a time period for developing customer loyalty with the service provider subject to the satisfaction of demanded services. Whereas, retaining a customer does not involve any additional marketing or other expense, rather an attention to the resolution of customer's concerns is enough in most of the cases. Further, long term customers generate more profit because they are not easily attracted by other competitors and in addition, can refer new customers and eventually become less costly to attend. Thus, a fractional improvement in customer retention can considerably impact the growth and sustainability of telecom business [3].

The issue of customer churn or customer attrition is not unique to the telecom industry. Almost every company in every industry at some point in time faces the problem of losing their customer to a competitor. This is usually a source of huge financial loss for companies as it is considered easier and much cheaper to retain existing customer than to attract new ones. Several studies and surveys back up this fact. Van den Poel and Larivière (2004) validate this in their study on the importance of the economic value of customer retention in the context of a European Financial Services company. [4].

(Burez, Van den Poel,2007) [5]. showed that there are two types of approaches for managing customer churn: reactive and proactive. When the company follows a reactive approach, it waits until the customer requests the company to cancel their service relationship. In this scenario, the company will offer the customer an incentive to stay. When the company adopts a proactive approach, it tries to find customers who are likely to churn before they do so. Then, the company provides special incentives for these customers to keep them from churning.

Most of research used various applied Machine Learning techniques to extract hidden relationships between different entities and attributes in a flood of data banks. These facts have attracted many companies to invest in CRM to maintain customer information. Customer centric approach is very common, particularly, in them telecommunication sector for predicting customers' behavior based on historical data stored in CRM.

According to **[6]** There are two main categories of churners - voluntary or involuntary. As (Figure 1.)

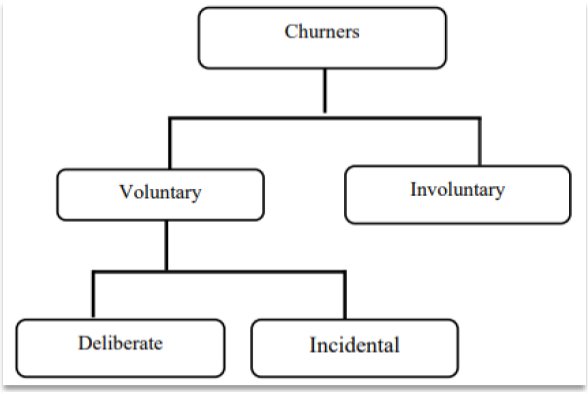

Figure 1.Types of churners

Involuntary Churners

Some customers are deliberately withheld service due to reasons which may include fraud, failure to pay bills and sometimes even non-utilization or insufficient utilization of the service.

Voluntary Churners

Voluntary churner occurs when a customer decides to terminate his/her service with the provider and switch to another company or provider. Telecom churn is usually of the voluntary kind. It can also be further divided into two sub-categories – deliberate and incidental [6]

Incidental churn can happen when something significant changes in a customer's personal lives which forces a customer to churn whereas deliberate churn can happen for reasons of technology, with customers always wanting newer or better technology, better service quality factors, social or psychological factors, and convenience reasons.

Among many approaches developed in the literature for predicting customer churn, supervised Machine Learning (ML) techniques are the most widely investigated. Supervised ML concerns the developing of models which can learn from labeled data. ML includes wide range of algorithms such as Decision trees, k-nearest neighbors, Linear regression, Naive Bayes, Neural Networks, Support Vector Machines (SVM), Genetic Programming and many others.[7]. Therefore, wide range of studies contributing to solve this Problem workaround churn aspects.

In [8]A comparative analysis of linear regression and two machine learning techniques: neural networks and decision trees for predicting customer churn behaviors to get why they leave.

Recent study argues proposed integrative model explain customer resistance to churn [9]. Some argue for Churn prevention by qualitive role involvement [10].

Widely considered building prediction models to anticipate the value that a random variable will assume in the future or to estimate the likelihood of future events Reviewed in literature. [11].

This paper contribution can be summarized as follows: Developing classification framework based on anaconda platform to predict churn

- Applying different ML classification algorithms

Applying ensemble learning such as XGBoost Classifier

- Applying feature selection algorithms to select the important features from the dataset
- Enhance performance by using cross-validation techniques.

The rest of this paper is structured as follows: Section 2 presents the previous studies to predict churn. Section 3 presents the main stages of a developing model to predict Churn based on Proposed Framework. Section 4 presents the experimental results and discussion. Finally, conclusions are presented in Section 5.

2. LITERATURE OF REVIEW

Most prior research has applied to Predict churners and what motivates customers Stop doing business with telecom companies and switch to another company. It is important for businesses to understand what contributes to churn to address those issues—and ultimately drive customer retention. So, Majority Studies pursuits on Churn Prediction models with varieties of tools and methodologies implementing techniques to achieve multiple level for accuracy investing those efforts in Predictions and retention plans.

The study addresses several further questions on, what are those factors that done through learning Process, how could be the Preprocessing phase contributing to diverse accuracy levels. The review in this section is primarily related to exploring the state of- the-art techniques for CCP that have been adopted for CCP towards improving Accuracy and Performance for Churn predictions.

Machine learning (ML) has shown outstanding results in a variety of applications, including speech recognition [12], computer vision [13], medical diagnostics [14], and engineering [15].

Among the many approaches developed in the literature for predicting customer churn, supervised Machine Learning (ML) techniques are the most widely investigated. Supervised ML concerns the developing of models which can learn from labelled data. ML includes a wide range of algorithms such as Decision trees, k-nearest neighbours, Linear regression, Naive Bayes, Neural Networks, Support Vector Machines (SVM), Genetic Programming and many others **[7].**

An Earlier study by (Awang et al., 2013.) [**16]** has emphasized that most customer churn prediction approaches only focuses on the classifier selection in improving the accuracy and performance of churn prediction, but rarely contemplate the feature reduction algorithms. . Experiment results show that the performance of classifiers improves with the application of features reduction of the customer churn data set. The best prediction model with the prediction accuracy of 92.08% produced by the Correlation based Feature Selection (CFS) method with the Decision Table classifiers.

Prior Study of Data Mining techniques by (Shaaban et al., 2012) [**17**]A Simple model based on DM techniques was introduced. Using 3 different techniques which are DT, SVM, and NN for classification and simple K Means techniques for clustering results indicate that the best output for the data set in hand is SVM techniqueAccuracy 83.7% and error rate 16.3%].

For instance, the study in [**18]** was introduced the Practicability for Models which are based on combining multiple machine learning techniques.

In [19]Similarly, Lalwani (2022) argued that data preprocessing is important phase and highlighted the role of

feature selection to improve classification accuracy and applied gravitational search algorithm to preform Feature selection. The paper proven that ensemble The XGBoost and CatBoost Classifier have significant results in terms of accuracy rates.

Prior research in **[20]** suggests that by combining SVM with boosting algorithms for higher accuracy and performance.

According to study in**[21,22]** XG Boost classifier stands for extreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It is a supervised and ensemble learning method that combines trees to give a more generalizable Machine Learning model boost classifier.

SVM stands for Support Vector Machine is a machine learning technique that includes supervised learning. SVM aims to minimize structural risk and account for aspects of generalization by finding the best hyper plane to separate data from defined classes **[23,24].**

Another study in **[25]** shows application for GP Application , Author searched capabilities of genetic programming (GP) and classification capabilities of AdaBoost are integrated in order to evolve a high-performance churn prediction system having better churn identification abilities. this paper Applied particle swarm optimization (PSO) based under sampling method, which provides unbiased distribution of training set to GP-AdaBoost based prediction system to handle imbalanced data. The results show that the proposed ChPGPAB system yields 0.91 AUC and 0.86 AUC.

A recent study in **[4]** addressed the recent technology for Churn Prediction Platforms Rapid Miner and R using the Common techniques investing the research for accuracy level ,The results show that Decision Tree and Random Forest are the two most accurate algorithms in predicting customer churn. And Found that Most significant Attributes in predicting customer churn are Contract, Tenure, Monthly Charges and Total Charges.

Another study in [**23]** introduced a hybrid model improved the results of predictions based on SVM. They used preprocessing method that combines oversampling and under-sampling methods to achieve results that are more accurate.

A critical open question is whether is there no standard model which addresses the churning issues of global telecom service providers accurately or not. Absolutely Data preprocessing, data normalization and feature selection have shown to be prominently influential. This motivated us to develop a prediction model which not only predicts the potential churners accurately devise decision makers intelligently. In next section, overview of proposed model is presented for accurate prediction.

This paper proposed a Framework for Churn Prediction of learning model using XG boost ensemble classifier which is less focused in the previous studies. XG boost classifier is applied on the publicly available telecom dataset and the experiential results are compared with SVM Classifier.

In the next section, an overview of the proposed model is presented for accurate prediction.

3. EXPERMENTAL FRAMEWORK

3.1 Research methodology

The purpose of this study is to introduce new layer focused on accuracy factors affect Customer churn Predictions. The most important methodological choice researchers make is based on the distinction between qualitative and quantitative data. As mentioned previously, qualitative data takes the form of descriptions based on language or images, while quantitative data takes the form of numbers.

The paper Research methodology is Quantitative data, on the other hand, might be easier to collect and analyze and it is based on a large sample of participants. Quantitative methods are based on data that can be ‘objectively’ measured with numbers. The data is analyzed through numerical comparisons and statistical analysis. Online data sources are used in searching and Exploring data sets. The data set for this study acquired from Kaggle source, originally from database of telecom company, The criteria followed is data range.

3.2 Proposed Framework

In this section, this paper offered an overview of proposed model. See **(Figure 2.)** shows customer churn prediction proposed model.

3.2.1 Acquiring dataset

Telecom dataset is not publicly available more due to itscustomer’s personal privacy. Data set for this paper obtained
from IBM Watson dataset released in 2015. The data setcontains 7043 instances and 21 attributes. The last attributes
denote churn or not in which 5174 are not churners and 1869 are churners. The percentage of churners is 26.53% and non-churnersis 73.46%. This dataset helps to figure out customerProphecy and build retention possibilities.

3.2.2 Preprocessing data

is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modeling purposes. The records with unique values do not have any significance as they do not contribute much to predictive modeling. Fields with too many null values also need to be discarded.
Firstly, the missing value analysis had been performed and it is found that there some missing value exist in the dataset. Missing values handled Next, the conversion of data types i.e.. Object data type to numerical data type is implemented. Total charges attribute is converted to numerical data type from object data type for the classifiers had been used for analysis. also checked for duplication of customer id and it showed that all the rows are unique.

3.2.3 Analysis data and ML application

In a comparatively manner SVM and XG Boost had been applied in machine learning cycle on test and train split dataset. The validation done by cross validation. The hyperparameters of the algorithms were optimized using K-fold cross-validation.While the evaluation of model performance according to classification measurements . The comparison between models were made using accuracy, precision , f-score, and recall.

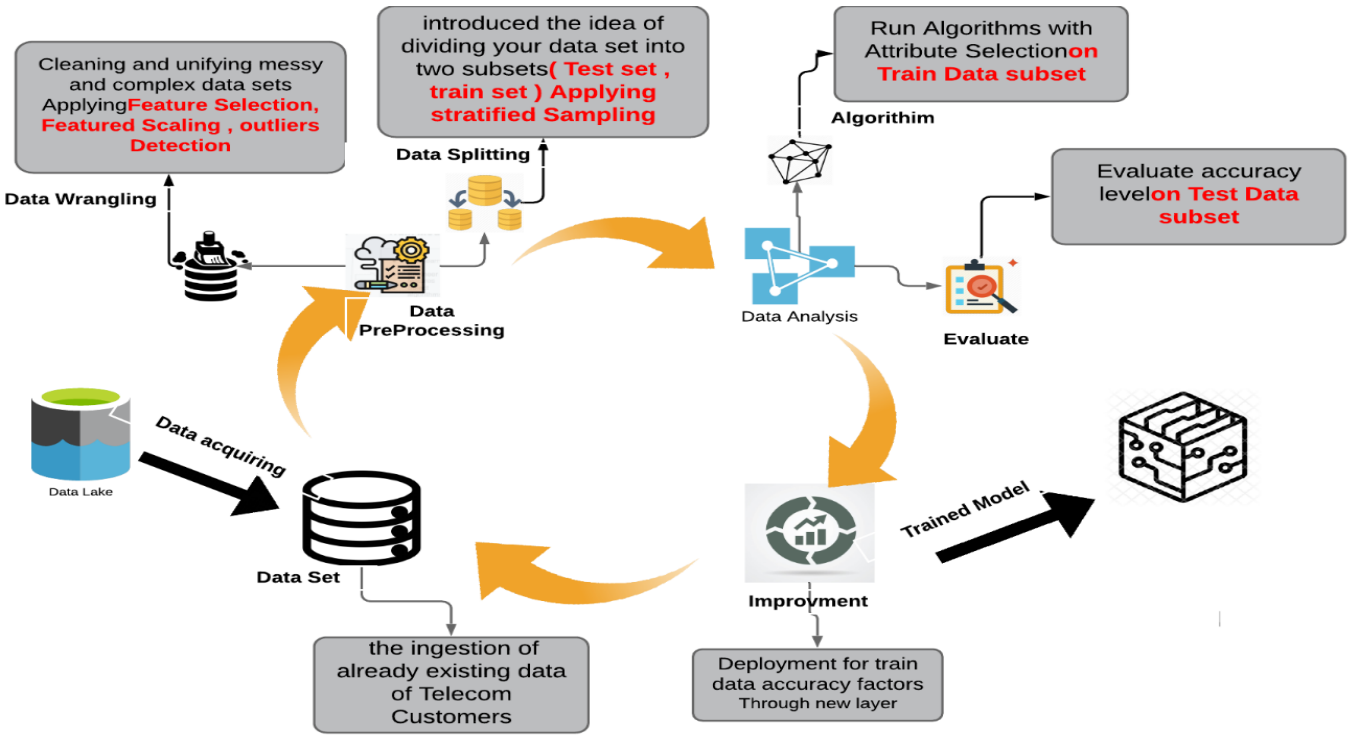

Figure 2. A Proposed framework for customer churn prediction

3.2.4 Performance measures

Accuracy: It can be described as the ratio of correctly predicted observations to the total number of observations **[26].** The formula is as follows: $(TP + TN)/(TP + FP + TN+FN$ (1)

Precision: It can be described as the ratio of correctly predicted positive observations to the total predicted positive observations. The formula is as follows:

$TP/(TP + FP)$ (2)

where TP is true positive, and FP is false negative

Recall: It can be described as the ratio of correctly predicted positive observations to all observations in the actual class. The formula is as follows:

$TP/(TP + FN)$ (3)

F1-Score: It can be described as the weighted average of precision and recall. Therefore, this score takes both false positives and false negatives into account.

The formula is as follows:

$2 \times (P \times R)/(P + R)$ (4)where P is precision and R is recall.

4. EXPERMENTAL RESULTS

Exploratory data analysis was used to discover dataset characteristics and prepare data for machine learning the dataset was explored using the dataset using library seaborn to visualize churn rate as shown in (**Figure 3**.) Churn Rate plot.

26.536987079369588 % of Customer left with the company

73.4630129206304 % of Customer stayed with the company.

4.1 Data correlation

Before building the churn, prediction model using classifiers of SVM and XGBoost the correlation among the features is identified and sorted in the descending order. Above table indicates the correlation values of the attributes and it is plotted as a graph (**Figure4.)** Based on this assumption, it shows clearly that there is a strong correlation between the attributes churn and Month-to-month contracts. Fibre Optic ISP and monthly charges are the second and third attributes that shows strong correlation.

4.2 Evaluation of results

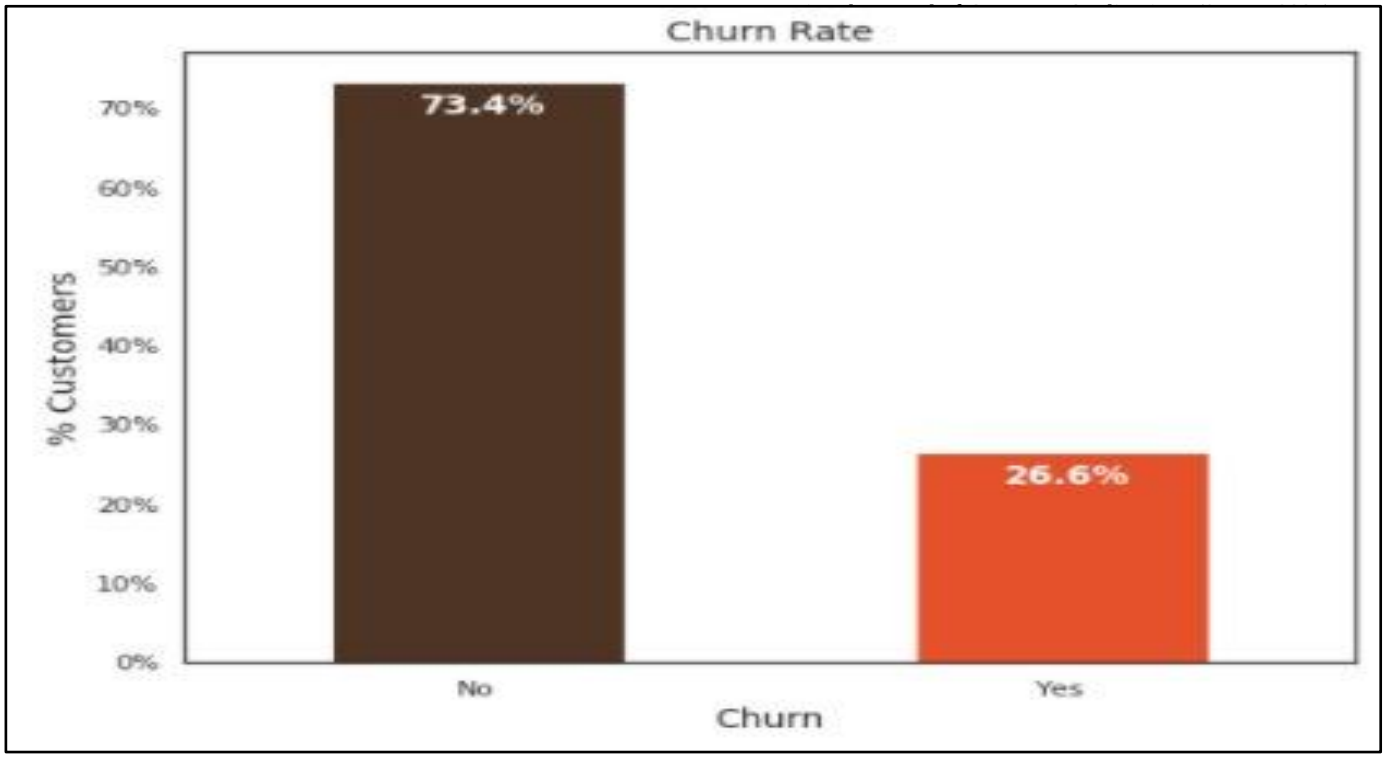

Figure 3. Churn rate plot

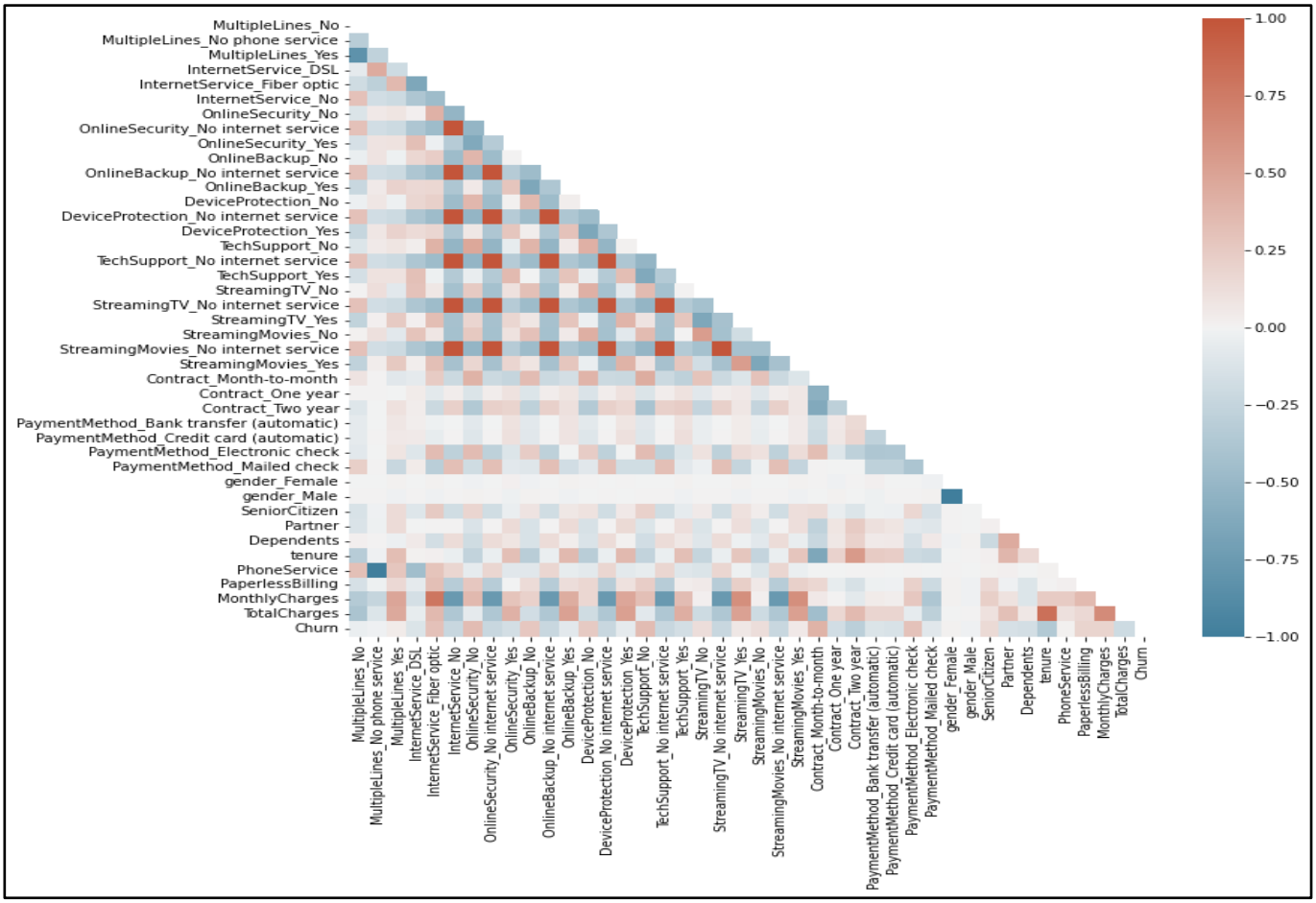

Figure 4. Feature correlation heatmap

A comparison was conducted on classification algorithms level to test the efficiency of the selected model within this dataset without any preprocessing tasks . The comparison was in terms of accuracy score, Precision ,Recall, and F score.

The**Table 1** below shows that XGBoost classifier gives higher accuracy score compared to SVM.

Table 1. Models classification report between classifiers ontrain results

Classifiers	Classification Report %			
	Accuracy	**Precision**	**Recall**	**F1-Score**
SVM	81%	64%	54%	59%
XGBOOST	84%	75%	62%	68%

5. CONCLUSION

5.1 Overall summary and contribution

Customer churn predictions (CCP) requires understanding churn attributes and customers behaviors. This understanding enables decision makers developing retention plans accurately based on the results. While customer retention is concerned with retaining those correctly identified churners by the prediction models. It has been highlighted in the past research that acquiring a new customer is very expensive as compared to retaining the existing customer [**27].**

In this paper presents a comparative study of different Machine learning algorithms and focuses on the importance of preprocessing the dataset as well as comparative study between SVM and XGBoost on model performance.

The model yielded on classification results that the XG boost classifier performs superior out of SVM to be 84% without any preprocessing.

On Exploratory data analysis there was a key finding which dataset suffering from class imbalance issue that affecting the accuracy results which can be a future work scope for further research.

5.2 Future work scope

This provides a good starting point for discussion and further research of the following:

5.2.1 To investigate the Gridsearch of XGBOOST with cross-validation was used to optimize the parameters of ML

5.2.2 To optimize Classifier by tuning hyperparameters could be a means of better accuracy.

5.2.3 To investigate class imbalance technique could be a means of better accuracy.

5.3 Research Limitations

There may be some possible limitations in this study, that should be taken into consideration.

The first limitation concerns the research involved surveying certain National organizations, it had been faced the problem of having limited access to these respondents. Due to this limited access, research was redesigned data collection through online sources. As might affect data quality by for outdated data, Collected data for some other purpose.
But this paper's Proposed framework designated to enhance data accuracy that the findings is still reliable and validate despite this limitation.

Second many research efforts done in this field of study which reflects multiple points of view multiple purposes due to the limitations of time and manpower, it had been only surveyed research papers published between 2016 and 2020. Therefore, if the research had been extended to cover other journals such as those focused on Prediction modelling and, comparative studies, the results might have been different.

6. ACKNOWLEDGMENTS

This research did not receive any specific grants from funding agencies in the public, commercial, or non-profit sectors.

7. REFERENCES

[1] Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, 95, 27–36. https://doi.org/10.1016/j.dss.2016.11.007.

[2] Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation, and network architecture. *Expert Systems with Applications*, 85, 204–220. https://doi.org/10.1016/j.eswa.2017.05.028

[3] Huang, Y., Kechadi, T.: An effective hybrid learning system for telecommunication churn prediction. *Expert Syst. Appl*. 40, 5635– 5647 (2013)

[4] Applications of Data Mining Techniques for Churn Prediction and Cross-selling in the telecommunications Industry. (2019).

[5] Burez, Jonathan & Van den Poel, Dirk. (2007). CRM at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Syst. Appl.. 32. 277-288. 10.1016/j.eswa.2005.11.037.

[6] Saraswat, S. & Tiwari, A. (2018), 'A New Approach for Customer Churn Prediction in Telecom Industry', *International Journal of Computer Applications*, Vol. 181(11), pp. 40-46.

[7] Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting Customer Churn in Telecom Industry using MLP Neural Networks: Modeling and Analysis. Life Science Journal, 11(3), 1097–8135.https://doi.org/10.7537/marslsj110314.11

[8] H.S. Kim and C.H. Yoon, "Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market," *Telecommunications Policy*, vol. 28, no. 9, pp. 751–765, 2004.

[9] Kim, S., Chang, Y., Wong, S. F., & Park, M. C. (2020). Customer resistance to churn in a mature mobile telecommunications market. International Journal of Mobile Communications, 18(1), 41. https://doi.org/10.1504/ijmc.2020.104421

[10] Joseph3, H., & Sumaya, K. M. (2020.). Role of qualitative research in preventing customer churn: a case study of mobile network operators in Kenya.

[11] Carlo Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, Ltd. 2009 ISBN: 978-0-470-51138-1

[12] L. Deng and X. Li, "Machine learning paradigms for speech recognition: an overview," IEEE Transactions on Audio Speech and Language Processing, vol. 21, no. 5, pp. 1060–1089, 2013.

[13] M. Q. Huang, J. Nini´c, and Q. B. Zhang, "Bim, machine learning and computer vision techniques in underground construction: current status and future perspectives," Tunnelling and Underground Space Technology, vol. 108, Article ID 103677, 2021.

[14] P. Oza, P. Sharma, and S. Patel, "Machine learning applications for computer-aided medical diagnostics," in Proceedings of the Second International Conference on Computing, Communications, and Cyber-Security

Springer, New York, NY,USA, 2021.

[15] T. Bikmukhametov and J. J¨aschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," Computers & Chemical Engineering, vol. 138, Article ID 106834, 2020.

[16] Awang, M. K., Makhtar, M., Nordin, M., & Rahman, A. (2013.). Improving Accuracy and Performance of Customer Churn Prediction Using Feature Reduction Algorithms.

[17] Shaaban, E., Helmy, Y., & Khedr, A. (2012). A Proposed Churn Prediction Model. Mona Nasr / International Journal of Engineering Research and Applications (IJERA), 2(4), 693–697. www.ijera.com

[18] Tsai, C.-F., & Lu, Y.-H. (2012). Data Mining Techniques in Customer Churn Prediction. *Recent Patents on Computer Sciencee*, *3*(1), 28–32. https://doi.org/10.2174/2212375911003010028

[19] Lalwani, P., Sethi, P., Kumar, M., Jasroop, M., & Chadha, S. (2022). Customer churn prediction system : a machine learning approach. Computing, 104(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y.

[20] Kumar, S., & D., C. (2016). A Survey on Customer Churn Prediction using Machine Learning Techniques. *International Journal of Computer Applications*, *154*(10), 13–16. https://doi.org/10.5120/ijca2016912237

[21] Beschi Raja J, Chenthur Pandian S. An optimal ensemble classification for predicting Churn in telecommunication. J Eng Sci Technol Rev. 2020;13(2):44-49. doi:10.25103/jestr.132.07.

[22] Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci. 2021;2(3):1-21. doi:10.1007/s42979-021-00592-x.

[23] Rustam, Z., Utami, D. A., Hidayat, R., Pandelaki, J., & Nugroho, W. A. (2019). Hybrid preprocessing method for support vector machine for classification of imbalanced cerebral infarction datasets. International Journal of Advanced Science, Engineering and Information Technology, 9(2), 685–691. https://doi.org/10.18517/ijaseit.9.2.8615

[24] J.Liu, E.Zio, "Integration of Feature Vector Selection and Support.

[25] Idris, A., Iftikhar, A., & Rehman, Z. ur. (2019). Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling. *Cluster Computing*, *22*. https://doi.org/10.1007/s10586-017-1154-3

[26] KALABALIK G, OKUR MC. a Comparison of the Performance of Ensemble Classification Methods in. Grad Sch Nat Appl Sci. Published online 2016.

[27] J. Hadden, A. Tiwari, R. Roy, and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," Computers & Operations Research. 34(10) (2007) 2902-2917.