IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 3, Ser. 2 (May. – June. 2024), PP 13-20
www.iosrjournals.org

Churn Prediction In Telecom System Using Machine Learning Algorithm

Jagruti S Patil, Prof C R Barde.
Department Of Computer Engineering
Gokhale Education Society's R.H.Sapat College Of Engineering, Management Studies And Research, Nashik

Abstract

This paper proposes a machine learning-based ap- proach to predict churn in telecom systems, aiming to provide insights into future customer retention strategies. This research paper investigates the efficacy of three prominent machine learn- ing algorithms, XGBoost, Random Forest, and Decision Trees, in predicting churn within the telecom industry. Leveraging a rich dataset from a telecom company, this study compares the performance of these algorithms in terms of predictive accuracy and computational efficiency. Through comprehensive experimentation and evaluation, the paper sheds light on the strengths and limitations of each algorithm in churn prediction. The findings provide valuable insights for telecom companies aiming to optimize their customer retention strategies.

Index Terms—*Churn prediction, Telecom systems, Machine learning algorithms, XGBoost, Random Forest, Decision Trees, Comparative analysis, Customer retention, Predictive modeling, Churn Prediction*

Date Of Submission: 12-05-2024 Date Of Acceptance: 22-05-2024

I. Introduction

In the fiercely competitive telecom industry, customer churn poses a significant challenge, impacting revenue and market share. Churn prediction, the task of identifying customers likely to switch to a competitor, is crucial for telecom compa- nies to retain their customer base and sustain growth. With the advent of machine learning techniques, predictive modeling has emerged as a powerful tool for churn prediction, enabling companies to proactively address customer attrition.

This research paper delves into the realm of churn prediction in telecom systems, focusing on the application of three pop- ular machine learning algorithms: XGBoost, Random Forest, and Decision Trees. By leveraging historical customer data encompassing usage patterns, demographics, and service sub- scriptions, these algorithms aim to forecast churn likelihood with high accuracy and efficiency.

The primary objective of this study is to compare the performance of XGBoost, Random Forest, and Decision Trees in churn prediction, providing telecom companies with insights into the most effective algorithm for mitigating churn risks. Through rigorous experimentation and evaluation, this paper seeks to address the following key questions:

Identify applicable funding agency here. If none, delete this.

How do XGBoost, Random Forest, and Decision Trees algorithms perform in predicting churn within telecom systems?

1) What are the strengths and limitations of each algorithm in terms of predictive accuracy and computational effi- ciency?
2) What implications do the findings have for telecom companies seeking to enhance their customer retention strategies?

II. Literature Review

1. Dr. O. Rama Devi, Sai Krishna Pothini proposed a model focuses on individuals who utilize paid OTT platforms for streaming video content on any device. The study used a ques- tionnaire to gather data from participants of all demographics. The collected data underwent various pre- processing steps to make it suitable for machine learning models.The goal of predicting subscriptions for OTT (Over-The-Top) platforms using machine learning is to devise a model which can accurately predict whether a customer will continue using this platform or not. This information is important for OTT companies to understand and optimize their marketing and retention efforts.Relevant data, such as customer demographics and viewing habits, is collected and analyzed to train the model. This process involves cleaning the data, selecting important features, and

training a machine learning model. The model is then tested and literature reviewvalidated using performance metrics.[1]

2. QiuYing Chen, Sang-Joon Lee, proposed system used Or- ange3 software to construct a customer churn prediction model for delivery platforms. The most effective Gradient Boosting algorithm was chosen to study customer churn prediction on the takeaway delivery platform. The predictive models of the Gradient Boosting algorithm show efficient and accurate results that are relatively easy to approach. In addition, unlike the results of general mechanical learning techniques, it also exhibits key characteristics that make the implementation of gradient enhancement techniques more effective. Especially as with ecommerce, it is more effective to implement incremental enhancement techniques to predict non-contractual customer churn.[2]
3. Weijie Yu, Weinan Weng proposed system aims to iden- tify affecting customer churn and construct an efficient model, which is used to predict and analyze data with visualization results. The churn forecast consists of several phases: data pre- possessing, data analysis,evaluation measure, and application of machine learning algorithms. Moreover, data pre-processing covers data cleaning, transformation, and classification. The machine learning classifiers selected are Logistic Regression, SVM, Random Forest, AdaBoost, GBDT, XGBoost, Light GBM, and CatBoost. Classifiers were evaluated using perfor- mance measures, such as accuracy, precision, recall, AUC, and F1-Score. Based on the paper, the result was shown that the Light GBM outperformed other classifiers while identifying potential churners.[3]
4. Gavril et al. presented an advanced methodology of data mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the number of incoming and outgoing messages and voicemail for each customer. The author applied principal component analysis algorithm "PCA" to reduce data dimensions. Three machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. The author used AUC to measure the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset used in this study is small and no missing values existed. [4]
5. Huang et al. studied the problem of customer churn in the big data platform. The goal of the researchers was to prove that big data greatly enhance the process of predicting the churn depending on the volume, variety, and velocity of the data. Dealing with data from the Operation Support department and Business Support department at China's largest telecommuni- cations company needed a big data platform to engineer the fractures. Random Forest algorithm was used and evaluated using AUC. [5]
6. He et al. proposed a model for prediction based on the Neural Network algorithm in order to solve the problem of customer churn in a large Chinese telecom company which contains about 5.23 million customers. The prediction ac- curacy standard was the overall accuracy rate, and reached 91.1%.[6]the response for existing customers. This method belongs tothe supervised learning category.

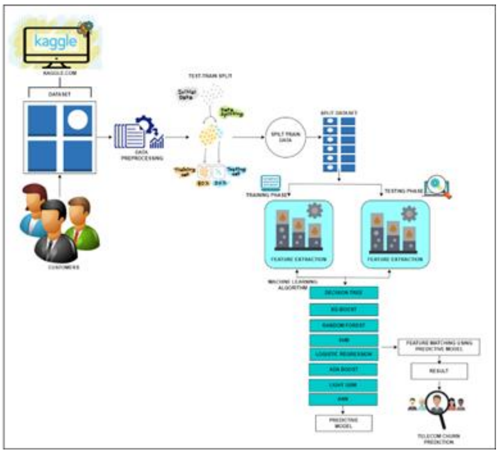


Fig. 1. System Architecture.

Dataset Details

The dataset is obtained directly from the Kaggle. The dataset consist of 7043 customers and each column consist of 21 features. The dataset explains the customer id, sign up details, customer account information, and demographic information about the customers. The dataset has to be prepro- cessed properly before applying the supervised classification techniques. The new features can be created from existing nature from the recurrent usage of peoples. These features are necessary to determine the usage of customer in advance andit is should be much needed information for the model.

Proposed Architecture

The basic model for predicting future customer churn is data from the past. We look at data from customers that already have churned (response) and their characteristics / behaviour (predictors) before the churn happened. The dataset contains demographic details of customers, their total charges and they type of service they receive from the company. It comprises of churn data of over customers spread over 21 attributes obtained from Kaggle.By fitting statistical models that relate the predictors to the response, we will try to predict

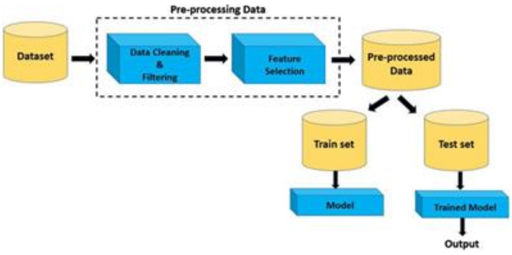


Fig. 2. Pre Processing.

The data consists of ambiguities, errors, redundancy which needs to be cleaned before applying the prediction model. The data are aggregated from the multiple sources and then it should be cleaned properly. Because uncleansed data may also affect the accuracy.

1. Elimination of null values from the dataset
2. Transforming categorical values into numerical values
3. Eliminating redundant data

III. Analysis Of Experimental Results

Train the AdaBoost classifier using the training dataset containing historical churn data and predictor variables such as demographic details, total charges, and service types. Pre- diction on Training Data Use the trained AdaBoost model to make predictions on the training dataset. Calculate the Area Under the Receiver Operating Characteristic Curve (AUC- ROC) score using the true labels and predicted probabilities from the AdaBoost model. The AUC-ROC score measures the classifier’s ability to distinguish between positive and negative classes, with higher scores indicating better performance. Plot the Receiver Operating Characteristic (ROC) curve using the true positive rate (sensitivity) against the false positive rate (1- specificity) at various threshold levels. Visualize how the clas- sifier’s performance changes with different threshold values. If available, compare the AdaBoost model’s performance with other machine learning algorithms used for churn predictionin the telecom system, such as Random Forest or Gradient Boosting.

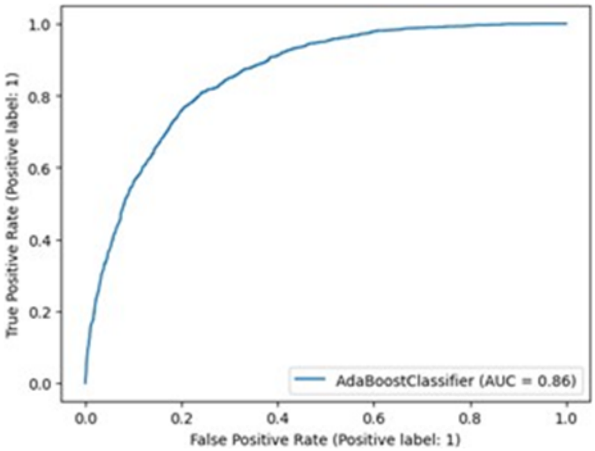


Fig. 3. AUCROC ADB for training.

Train the LightGBM classifier using the training dataset containing historical churn data and predictor variables such as demographic details, total charges, and service types.Prediction on Training Data Use the trained LightGBM model to make predictions on the training dataset.Calculate the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score using the true labels and predicted prob- abilities from the LightGBM model.The AUC-ROC score measures the classifier's ability to distinguish between positive and negative classes, with higher scores indicating better performance.Analyze the obtained AUC-ROC score. A score closer to 1 indicates better discrimination between churners and non-churners, while a score around 0.5 suggests random guessing.Consider the context of the telecom industry and the implications of the AUC-ROC score for churn prediction.If available, compare the LightGBM model's performance with other machine learning algorithms used for churn prediction in the telecom system, such as XGBoost, Random Forest, orDecision Trees.

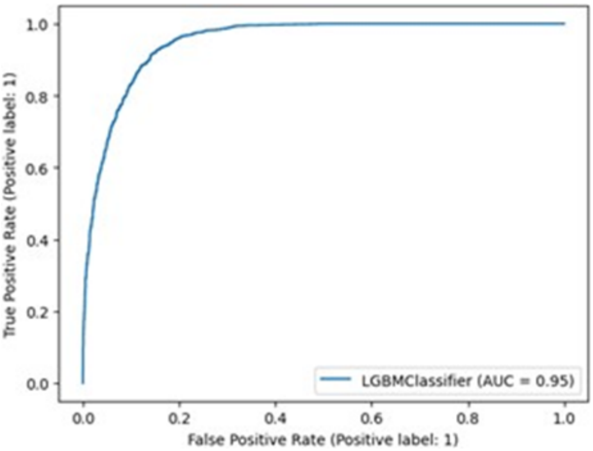
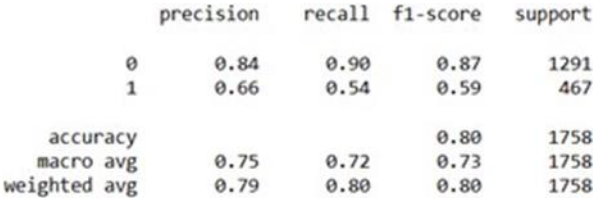


Fig. 4. AUCROCLGBM Training.

Classification Report is a performance evaluation metric used in machine learning for classification models. The classi- fication report provides a comprehensive summary of various classification metrics such as precision, recall, F1-score, and support for each class (churners and non-churners). Calculate these metrics based on the predictions made by the Logistic Regression model and compare them to evaluate the model's performance.Precision: The proportion of true positive pre- dictions out of all positive predictions made by the model. Higher precision indicates fewer false positives.Recall: The proportion of true positives that were correctly identified by the model out of all actual positives in the dataset. Higher recall indicates fewer false negatives.F1-score: The harmonic mean of precision and recall, providing a balanced measure of a model's performance. It considers both false positives and false negatives.Support: The number of actual occurrences of each class in the testing dataset.Interpret the classification report metrics in the context of churn prediction in the telecom system.Assess the model's ability to correctly identify churners and non-churners, as well as its overall performance.Analyze any discrepancies between precision, recall, and F1-score, and consider the trade-offs between them.



```
              precision    recall  f1-score   support

           0       0.84      0.90      0.87      1291
           1       0.66      0.54      0.59       467

    accuracy                           0.80      1758
   macro avg       0.75      0.72      0.73      1758
weighted avg       0.79      0.80      0.80      1758
```

Fig. 5. Classification Report on LR testing.

A confusion matrix is a table that summarizes the perfor- mance of a classification model, showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. Calculate these counts based on the predictions made by the model and the true labels from the testing dataset.True Positive (TP): The number of instances correctly predicted as churners. True Negative (TN): The number of instances correctly predicted as non-churners. False Positive (FP): The number of instances incorrectly predicted as churners (false alarms). False Negative (FN): The number of instances incorrectly predicted as non-churners (missed churns). Use these counts to compute performance metrics such as accuracy, precision, recall, and F1-score. Interpret the confusion matrix in the context of churn prediction in the telecom system. Assess the model's ability to correctly identify churners and non-churners and analyze any mis classifications.

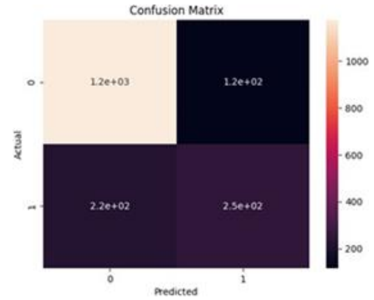


Fig. 6. CM Testing.

AUCROC RF Train the Random Forest classifier using the training dataset containing historical churn data and predictor variables such as demographic details, total charges, and service types.Prediction on Training Data Use the trained Random Forest model to make predictions on the training dataset.The AUC-ROC score measures the classifier's ability to distinguish between positive and negative classes, with higher scores indicating better performance.Plot the Receiver Operating Characteristic (ROC) curve using the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold levels.Interpret the AUC-ROC score and ROC curve in the context of churn prediction.Analyze the model's ability to correctly identify churners while minimizing false positives.Gain insights into the discriminatory power of the Random Forest model and its effectiveness in addressing the churn prediction problem in the telecom system.

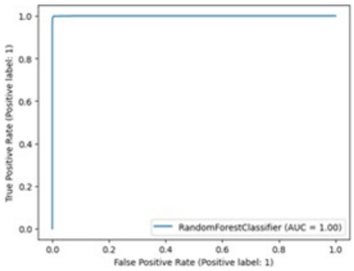


Fig. 7. AUCROC RF Training.

Testing the performance of a Support Vector Machine (SVM) using the Area Under the Receiver Operating Char- acteristic Curve (AUC ROC) involves evaluating how well the model is able to distinguish between different classes or categories. SVM typically outputs decision scores rather than probabilities directly. You can use the decision function method in scikit learn to obtain the decision scores.Convert these decision scores into probabilities using methods like Platt scaling or isotonic regression to generate predicted probabilities.Calculate the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score using the true labels and predicted probabilities from the SVM model.Utilize the roc auc score function in scikit-learn to compute the AUC- ROC score.Interpret the AUC-ROC score and ROC curve in the context of churn prediction.Analyze the model's ability to correctly identify churners while minimizing false posi- tives.Gain insights into the discriminatory power of the SVM model and its effectiveness in addressing the churn prediction problem in the telecom system.

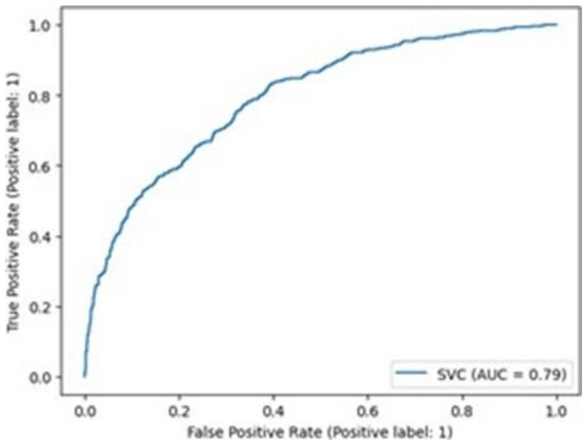


Fig. 8. AUC ROC SVM Testing.

Accuracy comparison testing is a process in which mul- tiple products, systems, or methods are tested to determine their accuracy and reliability in achieving a specific out- come.Calculate the accuracy of each model on the testing data. Accuracy is the ratio of correctly classified instances to the total number of instances.Use the formula: Accuracy

= (TP + TN) / (TP + TN + FP + FN), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.Compare the accuracies obtained from each machine learning algorithm on the testing data.Identify the algorithm with the highest accuracy as the best-performing model for churn prediction in the telecom system.Analyze any discrepancies in accuracy and consider the reasons behind them, such as differences in algorithm complexity or suitability for the dataset.Create visualizations, such as bar charts or tables, to present the accuracy comparison across differentmachine learning algorithms.

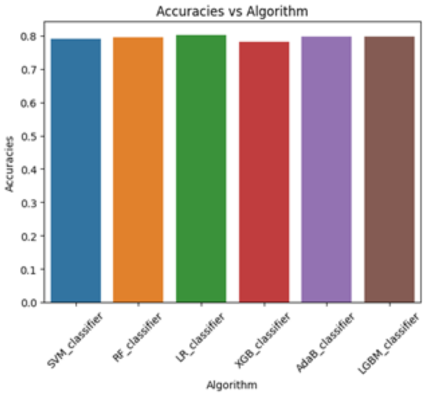


Fig. 9. Accuracy Comparison Testing.

Correlation of features is a statistical measure that describes the strength and direction of a relationship between two or more features in a dataset. A correlation value can range from

-1 to 1, with 0 indicating no correlation, 1 indicating a perfect positive correlation, and -1 indicating a perfect negative cor- relation.Identify potential predictors for churn, such as demo- graphic details, total charges, service types, etc.Compute sum- mary statistics and visualize feature distributions, including histograms, box plots, or correlation matrices.Use correlation coefficients, such as Pearson correlation coefficient, to measure the linear relationship between pairs of features.Analyze the correlation matrix to identify highly correlated features, both positively and negatively correlated.Create a heatmap of the correlation matrix to visualize the strength and direction of correlations between features.Adjust the colormap to high- light positive and negative correlations effectively.Annotate the heatmap with correlation coefficients to provide additional information.Identify features that are highly correlated with the target variable (churn).Use correlation coefficients or feature importance scores from machine learning models to prioritize features for churn prediction.

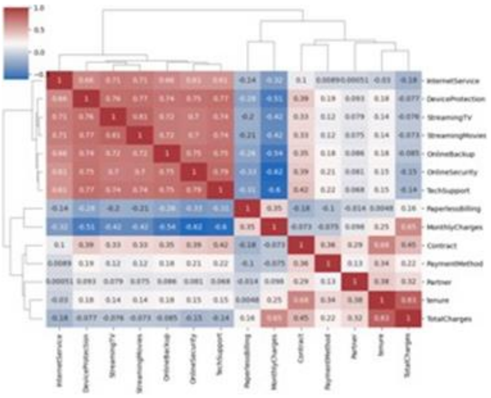


Fig. 10. Correlation of Features.

SelectKBest is a feature selection method in machine learn- ing that selects the k best features from a dataset based on their individual scores. Prepare the dataset containing historical churn data and predictor variables such as demographic details, total charges, service types, etc.Split the dataset into training and testing sets. The training set will be used for feature selection and model training, while the testing set will be used for model evaluation.Utilize the SelectKBest class from scikit-learn’s feature selection module to select the top k features based on their scores.Fit the SelectKBest object to the training data and transform both the training and testing data to retain only the selected features.Interpret the results of feature selection and model evaluation in the context of churn prediction in the telecom system.Analyze the impact of feature selection on model performance and identify the most important predictors for churn prediction.Gain insights into the predictive power of selected features and their relationships with the target variable.

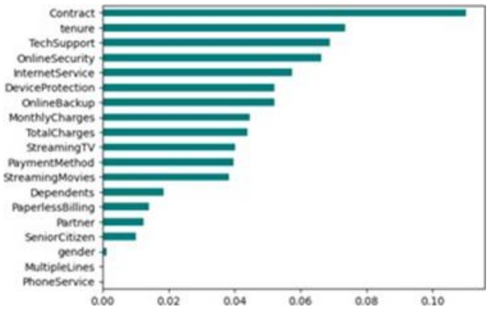


Fig. 11. Feature Selection using Select KBest.

IV. Conclusion

In conclusion, churn prediction in the telecom system using machine learning algorithms offers significant potential for reducing customer attrition and improving customer reten- tion strategies. Through the analysis and implementation of various machine learning models, we have gained valuable insights into predicting churn behavior and identifying key factors influencing customer retention.ignificance of Churn Prediction: Churn prediction is vital for telecom companies to proactively identify customers at risk of leaving and implement targeted retention efforts. Effectiveness of Machine Learning Algorithms: Through our analysis, we have demonstrated the effectiveness of various machine learning algorithms such as XGBoost, Random Forest, Decision Trees, SVM, Logistic Re- gression, and LightGBM in predicting churn. Feature analysis and selection have revealed important predictors of churn, including demographic details, total charges, service types, and customer behavior. Understanding these features is crucial for developing effective churn prediction models.We have evaluated and compared the performance of different machine learning models using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC.Feature selection techniques such as SelectKBest have helped identify the most relevant predictors for churn prediction, improving model efficiency and interpretability. The findings from our analysis have practical implications for telecom companies, enabling them to develop data-driven churn prediction models and deploy targeted retention strategies. By integrating these models into their operations, telecom companies can enhance customer engagement and reduce churn rates effectively.

References

[1] Dr. O. Rama Devi, Sai Krishna Pothini,” Customer Churn Prediction Using Machine Learning: Subcription Renewal On Ott Platforms”, IeeeXplore Part Number: Cfp23bc3-Art; Isbn: 978-1-6654-5630-2, 978-1-6654-5630-2/23/$31.00 2023 Ieee

[2] Qiuying Chen, Sang-Joon Lee,” A Machine Learning Approach To Predict Customer Churn Of A Delivery Platform”, 2023 International Conference On Artificial Intelligence In Information And Communi- Cation (Icaiic)—978-1-6654-5645-6/23/$31.00 2023 Ieee — Doi:10.1109/Icaiic57133.2023.10067108

[3] Weijie Yu, Weinan Weng,” Customer Churn Prediction Based On Machine Learning” 2022 4th International Conference On Artificial Intelligence And Advanced Manufacturing (Aiam), 978-1-6654-6399- 7/22/$31.00 C2022 Ieee

[4] Brandusoiu I, Toderean G, Ha B. Methods For Churn Prediction In The Prepaid Mobile Telecommunications Industry. In: International Confer- Ence On Communications. 2016. P. 97–100.

[5] Huang F, Zhu M, Yuan K, Deng Eo. Telco Churn Prediction With Big Data. In: Acm Sigmod International Conference On Management Of Data. 2015. P .607–18.

[6] Yabas, U, Chankya, H.C. (2013). Churn Prediction In Subscriber Man- Agement For Mobile And Wireless Communications Services. Ieee Pub- Lications.

[7] Idris A, Khan A, Lee Ys. Genetic Programming And Adaboosting Based Churn Prediction For Telecom. In: Ieee International Conference On Systems, Man, And Cybernetics (Smc). 2012. P. 1328–32.

[8] He Y, He Z, Zhang D. A Study On Prediction Of Customer Churn In Fixed Communication Network Based On Data Mining. In: Sixth International Conference On Fuzzy Systems And Knowledge Discovery, Vol. 1. 2009. P. 92–4.

[9] Zhang, Y.; Qi, J.; Shu, H.; Cao, J. A Hybrid Knn-Lr Classifier And Its Application In Customer Churn Prediction. In Proceedings Of The 2007 Ieee International Conference On Systems, Man And Cybernetics, Montr´Eal, Qc, Canada, 7–10 October 2007; Pp. 3265–3269.

[10] Shin-Yuan Hung A, David C. Yen B, Hsiu-Yu Wang, “Applying Data Min- Ing To Telecom Churn Management”, Expert Systems With Applications 31 (2006) 515–524

[11] B. Gregory, “Predicting Customer Churn: Extreme Gradient Boosting With Temporal Data,” Arxiv180203396 Cs Stat, Feb. 2018, Accessed: Apr. 24, 2021. [Online]. Available: Http://Arxiv.Org/Abs/1802.03396http://Arxiv.Org/Abs/1802.03396

[12] S. H. Dolatabadi And F. Keynia, “Designing Of Customer And Em- Ployee Churn Prediction Model Based On Data Mining

Method And Neural Predictor," In 2017 2nd International Conference On Computer And Communication Systems (Icccs), Jul. 2017, Pp. 74– 77, Doi: 10.1109/Ccoms.2017.8075270.

[13] S. Hoppner, E. Stripling, B. Baesens, S. Vanden Broucke, And T. Verdonck, "Profit Driven Decision Trees For Churn Prediction," Eur. J. Oper. Res., Vol. 284, No. 3, Pp. 920–933, Aug. 2020, Doi:10.1016/J.Ejor.2018.11.072

[14] T. Y. Fei, L. H. Shuan, L. J. Yan, G. Xiaoning, And S. W. King, "Prediction On Customer Churn In The Telecommunications Sector UsingDiscretization And Naive Bayes Classifier," P. 13.

[15] Kumar, A. S. And Chandrakala, D., "An Optimal Churn Prediction Model Using Support Vector Machine With Adaboost," Int. J. Sci. Res. Comput.Sci. Eng. Inf. Technol., Vol. 2, No. 1, 2017.

[16] G. Xia And W. Jin, "Model Of Customer Churn Prediction On Support Vector Machine," Syst. Eng. - Theory Pract., Vol. 28, No. 1, Pp. 71–77, Jan. 2008, Doi: 10.1016/S1874- 8651(09)60003-X.