A Comparison of Machine Learning Algorithms for Customer Churn Prediction

Parth Pulkundwar
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, India
p.pulkundwar@somaiya.edu

Krishna Rudani
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, India
krishna.rudani@somaiya.edu

Omkar Rane
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, India
omkar.vr@somaiya.edu

Chintan Shah
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, India
cbs@somaiya.edu

Dr. Shyamal Virnodkar
Department of Computer Engineering
K. J. Somaiya Institute of Technology
Mumbai, India
shyamal@somaiya.edu

2023 6th International Conference on Advances in Science and Technology (ICAST) | 979-8-3503-5981-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAST59062.2023.10455051

***Abstract*— Today's fiercely competitive business environment has given significant importance to customer churn, a term used for the loss of customers, which possesses a significant challenge to organizations across various industries. To mitigate revenue loss and sustain growth, companies are increasingly turning to machine learning (ML) algorithms for customer churn prediction. This review paper provides a concise examination of ML algorithms' role in predicting customer churn, a pivotal concern for businesses seeking to sustain growth and profitability. The review begins by underlining the significance of customer churn in today's competitive landscape, highlighting the impact of data-driven approaches in this context. The paper then explores various ML algorithms suitable for churn prediction and comparing the results to find out the most optimal algorithm for a few real-world scenarios, namely telecommunication, banking and e-commerce. The review found that Decision Tree Classification, Random Forest Classification, AdaBoost and XGBoost Classification algorithms were optimal for churn prediction. Additionally, the review covers the implementation of the findings in a churn prediction application.**

Index Terms*— Machine Learning, Churn Prediction, Data-driven Approaches, Gradient Boosting Algorithms, Customer, churn

I. INTRODUCTION

Contemporary world businesses have loads of data to work and grow from. Every move of a human in this age generates data, straight from their smartwatches to their choice of turning on the ceiling fan at their homes. However what matters is how these companies handle this data. Data being the new oil, has numerous uses, which only need to be uncovered with innovative analytics, insightful interpretation, and strategic application to unlock its full potential for driving business growth and societal advancements.

It is widely recognized that maintaining a current customer is more cost-effective than acquiring a new one. [2, 5]. One of the key metrics in this regard is Customer Churn. In simple words, customer churn refers to the portion of your customer base that stops to engage with your products or services within a specified timeframe. Thus, when predicting a customer's future with respect to their company, the fact that they cease their involvement with any of the company's products or services is called Churn.

Companies may create their own datasets to keep track of customers who "churned" or stopped using their products or services. There are traditional statistical methods which were used for quite some time in the analysis of churn. However, today's world is blessed with advancements in computing technology, as well as the rapid increase of ML algorithms for churn prediction. These advanced algorithms enable businesses to not only identify churn patterns but also to harness the power of predictive analytics, allowing for more proactive and targeted retention efforts. Additionally, the scalability and adaptability of ML models make them invaluable in handling vast and complex datasets, providing businesses with a competitive edge in customer retention strategies. There are also Deep Learning Models to take advantage of. However, owing to their high computational power requirements, as well as higher model training time, it was decided not to include them in this comparative study, as all other models were comparatively less demanding.

This paper intends to discover the impact of various ML algorithms on real-world scenarios. This paper compares accuracy and time required for each of the nine algorithms to classify a new data item. The analysis made by this study will be utilized in a churn predictor application.

II. LITERATURE REVIEW

While looking for ML algorithms to process the real-world scenarios on, this study took care of two factors: first is accuracy of the model, for obvious reasons, and second is the time taken by the model to train. The latter was of equal importance as the ultimate intent was of the creation a client-centered portal for predicting customer churn.

The recent years have witnessed the use of Decision Trees (DT) based algorithms, as well as Ensemble Learning methods for Churn prediction [1]. Decision Tree algorithms

have been in use for a long time, because they are simple and easy to comprehend [3].

Logistic regression is another ML algorithm which is easy to implement and train a dataset on, while also ensuring no assumptions are made about the distributions [4] of classes in the feature space which is good for customer churn prediction.

Random Forest is one of the Ensemble Learning methods, which are deployed for use in regression as well as classification used for customer churn prediction [7, 10]. It avoids overfitting to a high extent, and also scales quite well on data.

Support Vector Machines (SVM) are mainly used for data that has unknown distribution. It also doesn't suffer from overfitting. SVM models are parametric hence it maximizes the effectiveness of churn prediction [16].

Gradient Boosting algorithms work on somewhat similar lines as Random Forest. They involve an ensemble of weak prediction models, wherein the newer model learns on the shortcomings of the previous model making more accurate churn predictions [14].

There are two very commonly used gradient boosting algorithms, which was in this review, namely AdaBoost and XGBoost. AdaBoost (Adaptive Boosting) is one of the earliest and most well-known gradient boosting algorithms. XGBoost (Extreme Gradient Boosting) is a more recent and highly optimized gradient boosting algorithm [8].

K-Nearest Neighbors (KNN) classification offers several advantages. It doesn't require a training period, making it highly time-efficient and suitable for quick modeling on existing data for fast churn prediction [6].

Naive Bayes is an algorithm for probabilistic classification that relies on the principles of Bayes' theorem. It assumes that features are independent, simplifying calculations which makes the churn prediction as accurate as possible [17].

III. ML Classification Models

Based on the review of ML algorithms for classification, the following models were chosen to classify the datasets with:

1. Logistic Regression (LR)
2. Random Forest Classification (RF)
3. Support Vector Machines (SVM)
4. AdaBoost (ADAB)
5. XGBoost (XGB)
6. Decision Tree Classification (DT)
7. Naïve Bayes Classification (NB)
8. K-Nearest Neighbors (KNN) Classification
9. A basic artificial neural network (ANN)

A basic artificial neural network was also created to classify the datasets to compare the accuracy score as well as the time it takes to classify data on a deep learning model, both of which were key factors in the choice of models for the end-user application.

IV. Data Preprocessing

The customer churn datasets have columns like Price, Geography, Tenure and so on. These columns are expected to have values in them differing from each other in terms of datatypes and extremes, it is required to alter these data items in order to make them workable.

This means that, preprocessing of the datasets is required. Preprocessing involves Cleaning, Transformation and Reduction of the data.

1. Data cleaning involves finding and rectifying errors or inconsistencies in the data [15].
2. Transformation: This process entails preparing data for analysis by employing various techniques. Common methods for data transformation include normalization, standardization, and discretization.
3. Data Reduction involves reduction of volume of data, but still keeping essential information in a dataset.

The datasets had a few missing values, columns with categorical variables, as well as imbalanced classes. Hence, before training the models on the datasets, it was important to deal with these issues first.

4.1 Missing Values: In the datasets, the records with missing values were removed.

4.2 Categorical Variables: As numeric inputs are a must for a majority of ML algorithms, categorical variables were encoded into numeric values using One Hot Encoding.

4.3 Normalization: The features have different units of measurements, therefore MinMax Scaling method was used for uniform data normalization as shown in equation 1.

$$x_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where min(x) is the minimum value in x, and max(x) is the maximum value in x.

4.4 Class Imbalance: In some cases, some variables are imbalanced. These variables will be balanced using the OverSampling method, so the size of the minority values is increased to a size similar to the majority before balancing.

4.5 Feature Selection: In the final stage of data preprocessing, the task is to choose the most suitable features that serve as indicators for churn.

V. Datasets

This review analysed the aforementioned ML algorithms on the following datasets

A. Telecom Company Dataset

Telecom companies today have to keep up with huge competition, as a lot of companies have sprung up, providing services and programs at prices which aim to capture the price-sensitive consumer. They need to be aware of the patterns of modern-day consumers and adapt their strategies in order to stay afloat in this dynamic industry.

The Telecom company dataset [16] has 7043 rows and 21 columns of customer data about their usage of the company's

phone and internet services. This dataset includes features (columns) like "PhoneService", "InternetService", "StreamingTV", "StreamingMovies" etc., which gives the description of the services a customer has subscribed for, from the company. The only irrelevant data column in this dataset was the "customerID" column, hence it was dropped.

B. Bank Customer Dataset

Banks benefit from understanding the factors that influence a client's decision to depart from the company. Churn prevention allows banks to develop loyalty programs and retention campaigns to keep as many customers as possible.

This dataset [17] has 10000 values and 18 columns, RowNumber, CustomerId, Surname, CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary, Exited, Complain, Satisfaction, Score, Card Type, Point Earn. The variables, RowNumber, CustomerId, Surname will be dropped as these will not useful for model training.

C. E Commerce Dataset

Customer churn in the E-Commerce industry is a big problem for the organizations. It is beneficial for these organizations to know what makes the customers exit or not use their platform or services. This helps them to design attractive discounts accordingly to retain as many customers as possible.

This dataset [18] has 5630 values and 20 columns, describing a customer's profile on this e-commerce platform. Some of these columns are "PreferredLoginDevice", "CityTier", "PreferredOrderCat", "DaySinceLastOrder" and "HourSpendOnApp" to name a few. The variable "CustomerID" is dropped from the dataset as it will not be useful for model training.

VI. Analysis and Discussion

Below mentioned are the accuracy and runtime comparison of the 3 datasets, and their analysis.

A. Telecom Company dataset Analysis

TABLE I. Results of different algorithms' performance on the Telecom Company dataset

Algorithm	Accuracy	Run Time
Logistic Regression	80.75%	0.961s
Random Forest	80.88%	3.3201s
SVM	82.02%	1.4811s
AdaBoost	81.59%	0.4151s
XGBoost	82.94%	0.8828s
DecisionTree	73.77%	0.0498s
Naive Bayes	69.79%	0.0055s
KNN	74.84%	0.0312s
ANN (DL)	90.33%	29.0029s

The aforementioned selection of ML algorithms performed as expected on this dataset.

1. Naïve Bayes and KNN Classification took the least time to train but provided low accuracies on testing.
2. A similar result was observed for Logistic Regression, in terms of training time and accuracy.
3. Gradient Boosting algorithms (AdaBoost and XGBoost) did take slightly more time to train (Fig. 2), however, they provided good accuracy scores (Fig. 1).
4. From TABLE I, it is evident that the best performer out of the ML algorithms is Random Forest.
5. Even though the ANN provided the highest accuracy, it took a comparatively large amount of time to train, which reinforced our pre-assumptions about its performance.

Figure 1 shows the accuracy score of various algorithms applied on telecom churn.

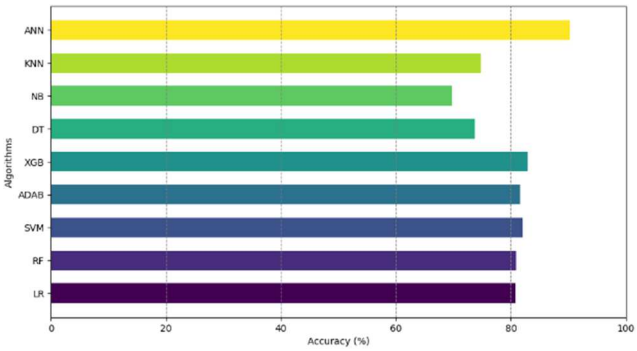

Fig. 1. Accuracy scores of various algorithms on Telecom Churn dataset

Figure 2 shows the model training time of different algorithms on telecom company dataset.

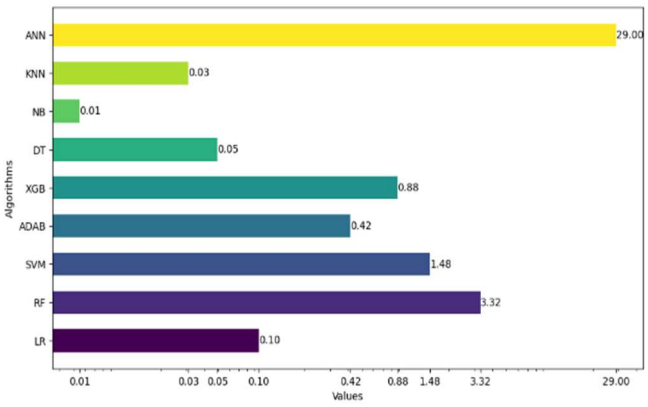

Fig. 2. Model Training time of various algorithms on Telecom Company dataset

B. *Bank Churn dataset Analysis*

TABLE II. Results of different algorithms' performance on Bank Churn dataset

Algorithm	Accuracy	Run Time
Logistic Regression	65.76%	0.033s
Random Forest	83.63%	1.82s
SVM	70.40%	9.35s
AdaBoost	81.59%	0.65s
XGBoost	78.00%	1.64s
DecisionTree	76.63%	1.41s
Naive Bayes	70.46%	0.009s
KNN	67.10%	0.004s
ANN (DL)	95.33%	42.19s

The results observed in TABLE II, were similar to that observed in the Telecom dataset.

1. Random Forest Classification stands out as the best ML algorithm, recording an accuracy score of 83.63%, while Decision Tree Classification provided an acceptable 76.63% accuracy.
2. Gradient Boosting Algorithms follow-up behind Random Forest, taking up 0.65s (AdaBoost) and 1.64s (XGBoost) for training (Fig. 4), while providing accuracy scores of 81.59% and 78% respectively (Fig. 3).
3. SVM Classification did not perform as expected, as it returned a comparatively low accuracy, while taking up 9.35s in training.
4. Logistic Regression had a relatively dismal accuracy score of 65.76%.

When compared with the accuracy scores on the Telecom Churn dataset (Fig.1), the selection of ML algorithms have shown a similar pattern of performance as observed in Fig. 3, with the exception of Logistic Regression. This pattern isn't observed in the next dataset.

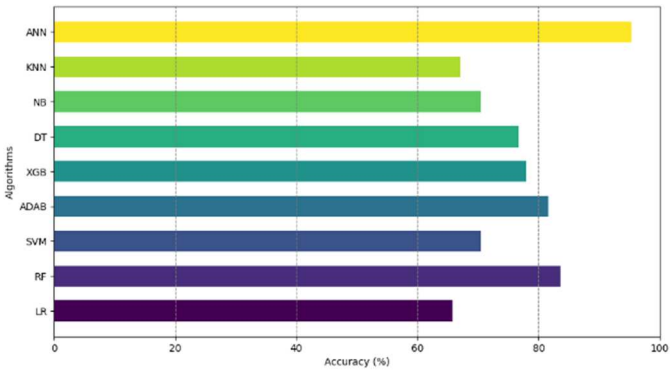

Fig. 3. Accuracy scores of various algorithms on Bank Churn dataset.

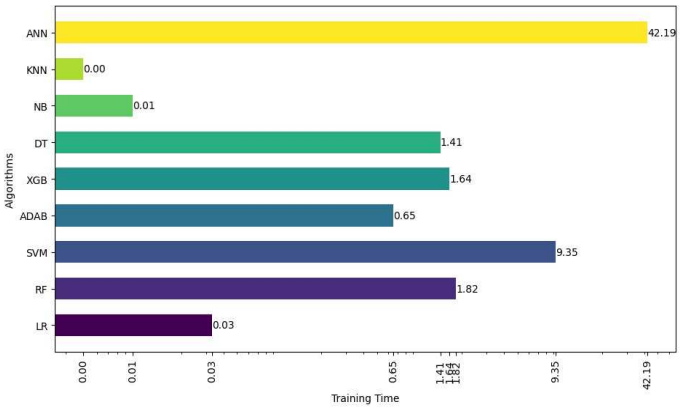

Fig. 4. Model Training time of various algorithms on Bank Churn dataset

C. *E-Commerce dataset Analysis*

TABLE III. Results of different algorithms' performance on E-Commerce dataset

Algorithm	Accuracy	Run Time
Logistic Regression	86.00%	0.8570s
Random Forest	81.18%	2.8990s
SVM	83.77%	1.5101s
AdaBoost	85.79%	0.2151s
XGBoost	83.12%	1.2828s
DecisionTree	93.00%	0.0384s
Naive Bayes	83.77%	0.0189s
KNN	79.00%	0.1275s
ANN (DL)	87.54%	27.0029s

The performance of ML algorithms (TABLE III) on E-commerce dataset was comparatively different than those on Telecom dataset and bank dataset.

1. While Decision Tree Classification couldn't provide a good accuracy score on the previous datasets, it gave an impressive 93% accuracy score on this dataset.
2. Random Forest Classification and Gradient Boosting Algorithms (AdaBoost and XGBoost) displayed consistent levels of accuracies, when compared to their previous performances.
3. Support Vector Machines (SVM) showed an improved accuracy when compared to its performance on the Bank dataset.
4. A tremendous improvement was recorded by Naïve Bayes and KNN Classification (Fig. 5).
5. Logistic Regression, which wasn't among the top models for the earlier cases, showcased an exceptional improvement in this scenario.

Overall, the ML algorithms performed at par with the ANN (DL) model (Fig. 6).

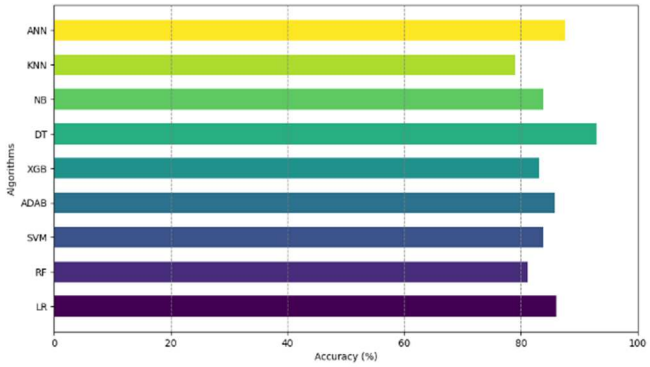

Fig. 5: Accuracy scores of various algorithms on E Commerce dataset.

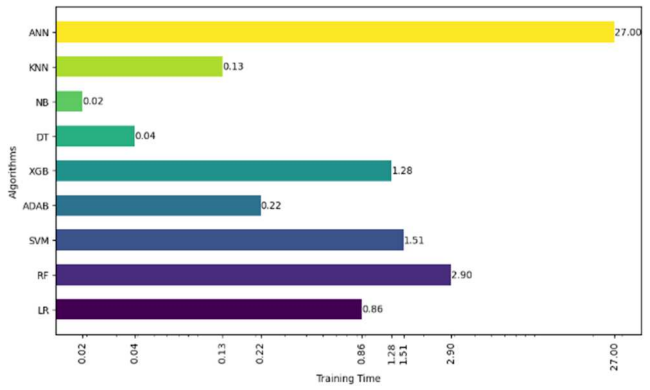

Fig. .6. Model Training time of various algorithms on E-Commerce dataset

VII. Results

Based on the results obtained on training the models, it was decided to incorporate Decision Tree, Random Forest, and XGBoost in the churn prediction application. The availability of models with the best performance in churn prediction would enable the users of this application to obtain the best possible insights about their data. Given below is an excerpt (Fig. 7) developed using Streamlit demonstrating how the models have been provided for the user's consideration.

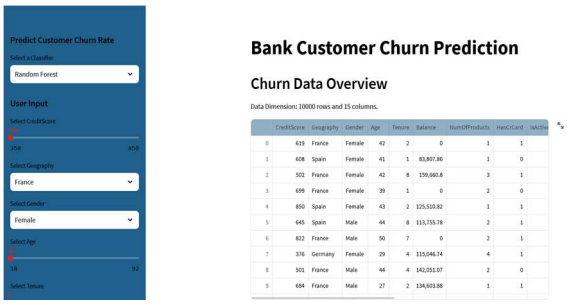

Fig. 7. Streamlit web application

The application asks the user to select an ML model of their choice. Next step involves selecting the features of their choice to provide data to the model for predicting churn. Once the data is entered, the application moves on to the final result, providing the churn status as shown in Fig. 8, as well as the probability that churn occurs.

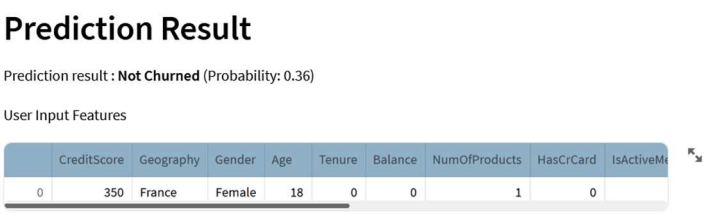

Fig. 8. Output/Prediction on the input features

VIII. Conclusion

This paper involved the training of ML models discussed earlier with respect to three real-world scenarios, for churn prediction. It also reviewed a churn prediction application which implemented the findings of this paper. As apparent, it was found out that on an average, gradient boosting algorithms i.e AdaBoost and XGBoost, Random Forest Classification as well as Decision Tree Classification performed the best, while other classification algorithms couldn't provide as much accuracy, even though they took much less time to train on the dataset.

In contrast, even though the basic ANN (Deep learning model) took a lot more time to train in comparison to ML models, it provided a much better accuracy score than the former.

Therefore, it only came down to the availability of resources mainly in terms of systems with better computing power, and the situation at hand. The main concerns were the accuracy of the result as well as the time taken for making the result available to the client, when they use the churn prediction application. Hence, deep learning techniques would not be suitable for this situation. This may not be the case with other applications of customer churn prediction, however when a client-centered application is considered, it is important to provide the client, the desired output as quickly as possible. Therefore, as ML models provided us with satisfactory accuracy scores while taking up only a fraction of time as compared to a DL model, it was decided to move forward with ML models to be implemented in the application.

IX. Future Scope

Customer churn prediction, as mentioned earlier, has its pre-defined importance in this era of dynamically changing markets. With the advent of ML models, paired with the increasing reliance on data-driven decision making, customer churn prediction is only expected to grow in importance. Today's world is a witness to the staggering growth of computing technology, with faster and more efficient components making way into the consumer markets sooner than ever. This indirectly means that computing power is going to grow significantly, for a given amount of power supply and time.

What is inferred from this future is DL models will be treated in the future just like how ML models are treated and

used in today's world. This would help in the realization of better and faster predictions, for Customer churn as well as for many other purposes. DL models would even enable the client to use many more parameters for calculation, making predictions as real as possible.

In the meantime, there could be studies for optimization of DL models to meet the time limitations as well as performance benchmarks.

REFERENCES

[1] Wang, Xing & Nguyen, Khang & Nguyen, Binh. (2020). Churn Prediction using Ensemble Learning. 56-60. 10.1145/3380688.3380710.

[2] A. De Caigny, K. Coussement, and K.W. De Bock. 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees.

[3] European Journal of Operational Research 269, 2 (2018), 760–772Hu, Xin & Yang, Yanfei & Chen, Lanhua & Zhu, Siru. (2020). Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network. 129-132. 10.1109/ICCCBDA49378.2020.9095611.

[4] Martínez-García, M. et al. (2023). Learning Logistic Regression with Unknown Features. In IEEE CAI 2023, pp. 298-299. doi: 10.1109/CAI54212.2023.00133.

[5] Rani, K. Sandhya and., Shaik Thaslima and., N.G.L. Prasanna and ., R.Vindhya and ., P. Srilakshmi, Analysis of Customer Churn Prediction in Telecom Industry Using Logistic Regression (JUNE 10, 2021). International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-9, Issue-4, July 2021.

[6] Hassonah, M. A. et al. (2019). Churn Prediction: KNN vs. Decision Trees. In Sixth HCT ITT 2019, pp. 182-186. doi: 10.1109/ITT48889.2019.9075077.

[7] Feng, L. (2022). Customer Churn Prediction: Borderline-SMOTE and Random Forest. In IEEE ICPICS 2022, pp. 803-807. doi: 10.1109/ICPICS55264.2022.9873702.

[8] Zhang, J., & Dong, Y. (2022). Customer Loss Identification and Factor Analysis in Mobile Operators with XGBoost. In 2022 NetCIT,.

[9] Wu, X., & Meng, S. (2016). E-commerce Customer Churn Prediction with Enhanced SMOTE and AdaBoost. In 2016 ICSSSM.

[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[11] A. Alamsyah and N. Salma, "A Comparative Study of Employee Churn Prediction Model," 2018 4th International Conference on Science and Technology (ICST), Yogyakarta, Indonesia, 2018, pp. 1-4, doi: 10.1109/ICSTC.2018.8528586.

[12] K. Gupta, A. Hardikar, D. Gupta and S. Loonkar, "Forecasting Customer Churn in the Telecommunications Industry," 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 2022, pp. 1-5, doi: 10.1109/IBSSC56953.2022.10037334.

[13] A. Raj and D. Vetrithangam, "Machine Learning and Deep Learning technique used in Customer Churn Prediction: - A Review," 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 2023, pp. 139-144, doi: 10.1109/CISES58720.2023.10183530.

[14] Y. Y. Win and C. G. Vung, "Churn Prediction Models Using Gradient Boosted Tree and Random Forest Classifiers," 2023 IEEE Conference on Computer Applications (ICCA), Yangon, Myanmar, 2023, pp. 271-275, doi: 10.1109/ICCA51723.2023.10181933.

[15] D. Dasari and P. S. Varma, "Employing Various Data Cleaning Techniques to Achieve Better Data Quality using Python," 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2022, pp. 1379-1383, doi: 10.1109/ICECA55336.2022.10009079.

[16] Rodan, Ali & Faris, Hossam & Al-sakran, Jamal & Al-Kadi, Omar. (2014). A Support Vector Machine Approach for Churn Prediction in Telecom Industry. International journal on information.

[17] D. T. Barus, R. Elfarizy, F. Masri and P. H. Gunawan, "Parallel Programming of Churn Prediction Using Gaussian Naïve Bayes," 2020 8th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia, 2020, pp. 1-4, doi: 10.1109/ICoICT49345.2020.9166319.

[18] https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[19] https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn

[20] https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction