Unraveling Waze User Churn: ML for Retention Enhancement

Dr. T. Swapna[1]*, T. Sathwik[2] , A.Yashwanth Sai[2] ,C. Sai Sukruth[2]
[1]Assistant Professor, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India.
[2]UG Students, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
Email swapnat@sreenidhi.edu.in , tumwarsathwik08@gmail.com , arukutiyashwanthsai@gmail.com , sukruthchatla@gmail.com

Abstract - This project delves into analyzing monthly user churn in Waze, a prominent navigation app, through advanced machine learning techniques. The primary objective is to develop a predictive model that can forecast user churn and reveal significant insights into the factors that drive this occurrence. Waze's vibrant community, including map editors, beta testers, translators, partners, and users, plays a pivotal role in refining the app continually. Understanding and mitigating user churn are vital for sustained growth, aligning with Waze's growth objectives. The project aims to build a robust machine learning model not only to predict and prevent churn but also to enhance overall user retention, thereby facilitating business expansion. Insights gleaned from the model will help identify high-risk user segments, enabling targeted retention strategies such as proactive engagement and special incentives. The anticipated outcomes will provide invaluable guidance to Waze's leadership in optimizing retention strategies, improving user experience, and making data-driven decisions regarding product development, bolstering Waze's competitive edge in the navigation app market.

Keywords: Waze, user churn, machine learning, predictive model, retention strategies, user experience, navigation app.

I.INTRODUCTION

Waze, a globally utilized navigation app, has revolutionized travel for drivers by providing a user-friendly platform enriched with collective intelligence. The collaborative efforts of its diverse community, comprising map editors, beta testers, translators, partners, and users, contribute to continuously refining and enhancing the driving experience. This project is deeply rooted in Waze's ethos of leveraging collective intelligence to make driving safer and more efficient worldwide while maintaining a commitment to providing a free and user-friendly navigation platform.

The main aim of this project is to create a machine learning model that predicts user churn, with a specific focus on monthly churn metrics. Churn, in this context, refers to users who uninstall or cease using the Waze app. By constructing an accurate predictive model, we aim not only to prevent churn but also to improve user retention, thereby contributing to the growth and success of Waze's business model.

The approach involves several key steps. Firstly, comprehensive data analysis is conducted, including data cleaning and exploration of key features to identify outliers. Subsequently, exploratory data analysis (EDA) is performed using descriptive statistics and visualizations to understand data distribution and relationships. Statistical testing is employed to compare mean rides for Android and iPhone users, determining statistical significance. Regression analysis is then utilized to interpret variables impacting user retention. Finally, machine learning models such as Random Forest and XGBoost are implemented and fine-tuned for optimal performance, with the aim of predicting and preventing user churn effectively.

By leveraging advanced machine learning techniques, this project aims to provide actionable insights for Waze's leadership, enabling them to optimize retention strategies, enhance user experience, and make informed decisions regarding product development. Ultimately, this endeavor seeks to solidify Waze's position in the navigation app market, ensuring continued success in an ever-evolving landscape of mobile applications.

I I. LITERATURE SURVEY

M.A.H. Farquad proposes an innovative solution to tackle a common issue associated with traditional SVMs, namely their tendency to generate complex, opaque models. The author introduces a three-phase methodology. In the first step, SVM Recursive Feature Elimination (SVM-RFE) is employed to streamline the feature set. Following this, the dataset with reduced features is subjected to classification

using SVM techniques. Finally, manual rule extraction is conducted, followed by amalgamating Naive Bayes with Decision Trees to produce outcomes [1].

The research focuses on a credit card dataset marked by significant volatility, with 93.24% loyal customers and 6.76% churned customers [2]. It's observed that the model's scalability to large datasets is limited. To predict credit card behavior within a telecom CRM dataset, the authors introduce two separate models: "Dual-ANN" and "SOM+ANN." These models combine back-propagation with neural networks and self-organizing maps (SOM) with neural networks.

The Dual-ANN method is utilized to remove irregular data through a data depletion technique, and its output is utilized as input for SOM+ANN. The evaluation of these models employs three testing strategies, including a "one general testing set" and fuzzy testing strategies. Findings emphasize the superiority of the hybrid model over the single baseline neural network model, with Dual-ANN demonstrating superior performance compared to SOM+ANN.

Wouter Verbeke et al. propose the use of Ant-Miner and ALBA to develop precise and interpretable credit card prediction models. Ant-Miner, which relies on Ant Colony Optimization (ACO), incorporates domain expertise and moderate monotonicity constraints, resulting in highly accurate and understandable models. Additionally, the authors juxtapose the performance of different models, including RIPPER and C4.5 [4].

Ning Lu suggests utilizing boosting algorithms to improve customer churn prediction models, with a focus on clustering customers according to the weights generated by the boosting algorithm. Logistic regression serves as the learner, with each cluster having its churn prediction model. The efficacy of boosting algorithms in segregating churn data is highlighted compared to a single logistic regression model [5].

H. Karamollaogluetal perform a comparative examination of different machine learning techniques aimed at attaining favorable f1-scores. The assessment encompasses Multilayer Perceptrons, Logistic Regression, AdaBoost, Decision Trees, among others, where the random classifier emerges as the top performer without the need for data augmentation methods[6].

The author envisions a credit card prediction method based on SVM models, utilizing random sampling to tackle data imbalance concerns. SVM constructs a hyperplane for classification, while random sampling adjusts data distribution to alleviate dataset imbalance resulting from fewer churners [7].

Ssu-Han Chen presents a fresh approach employing the gamma Cumulative SUM (CUSUM) chart to monitor Inter-Arrival Times (IAT) of customers. This method utilizes a limited mixture model for reference value and decision interval, alongside a ranked Bayesian model to account for diverse customers. Moreover, it incorporates a parallel time interval variable to IAT to track recent login behavior, providing a user-friendly graphical interface [8].

Koen W. De Bock advocates for the use of Rotation Forest and Rot Boost in churn prediction, employing an ensemble classifier to amalgamate outputs from multiple member classifiers. Rotation Forests require feature extraction for training base classifiers, while Rot Boost demonstrates enhanced AUC and top decile lift precision compared to Rotation Forests [9].

The author introduces a novel model, "impact learning," derived from CNN, to predict client attrition, yielding superior performance compared to ANN and Logistic regression. Furthermore, A hybrid approach known as "Generalized Additive Models (GAMs)" is developed, combining bagging concepts with the Random Subspace method, exhibiting superior performance over logistic regression and GAM [11].

Ning et al. explore credit card prediction and advocate boosting methods to enhance performance, categorizing the model partitions customers into clusters based on the weights generated by the boosting algorithm. It suggests an Implementation Zone for targeting high-churn customers, facilitating retention efforts [14].

The paper aims to delve into machine learning techniques employed by experts in recent years, summarizing the performance of prediction methods across varied datasets. Analysis reveals that hybrid and ensemble methods have significantly enhanced model performance, underlining the importance of clear evaluation guidelines [16].

V. Geetha et al. scrutinize feature extraction in credit card prediction methods, employing distance zone methods to explore characteristics. Despite existing techniques demonstrating satisfactory accuracy, they are time-intensive. The authors propose an ML approach to enhance accuracy and expedite evaluation, ensuring optimal services for non-churning customers in the telecom sector. Additionally, the SMOTE method effectively addresses dataset imbalance in the credit card dataset [17].

III. PROPOSED METHODOLY

Dataset :

The dataset, sourced from Kaggle, comprises information on 14,999 users, with each entry containing 14 distinct features. These features include 'ID','label','sessions','drives','total_sessions','n_days_after_onboarding','total_navigations_fav1','total_navigations_fav2','driven_km_drives','duration_minutes_drives','activity_days', 'driving_days', and 'device'. Prior to applying supervised classification techniques, it's imperative to preprocess the dataset appropriately. This involves generating new features derived from recurrent user behaviors, which are crucial for predicting and understanding user usage patterns. These derived features serve as essential insights for the model, aiding in forecasting user behavior effectively.

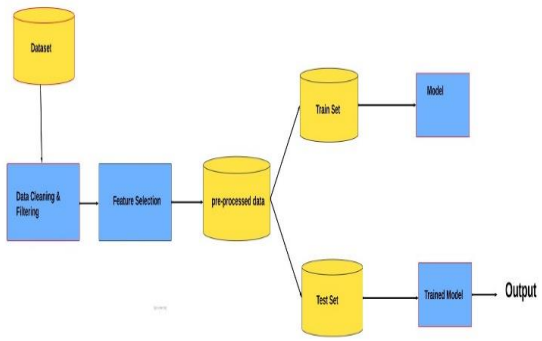

Fig 3.1: Proposed system Architecture

1. Data Preprocessing:

Data preprocessing involves preparing the raw user data from the Kaggle dataset for analysis and modeling. This includes handling missing values, removing duplicates, performing feature engineering, encoding categorical variables, Rescaling numerical features and dividing the dataset into training and testing subsets.

2. Exploratory Data Analysis (EDA):

EDA aims to understand the dataset's characteristics through descriptive statistics and visualizations. It helps identify data distributions, trends, outliers, and relationships between variables, guiding feature selection and model development decisions.

3. Hypothesis Testing:

Formulating hypotheses and applying statistical tests to compare relevant metrics or characteristics between different groups or categories within the dataset. This step helps identify potential predictors of churn and informs further analysis and modeling decisions.

4. Model Implementation:

Binomial Logistic Regression:

Binomial Logistic regression is a statistical model commonly utilized for binary classification tasks, including churn prediction. It models the probability of a binary outcome (e.g., churn or no churn)

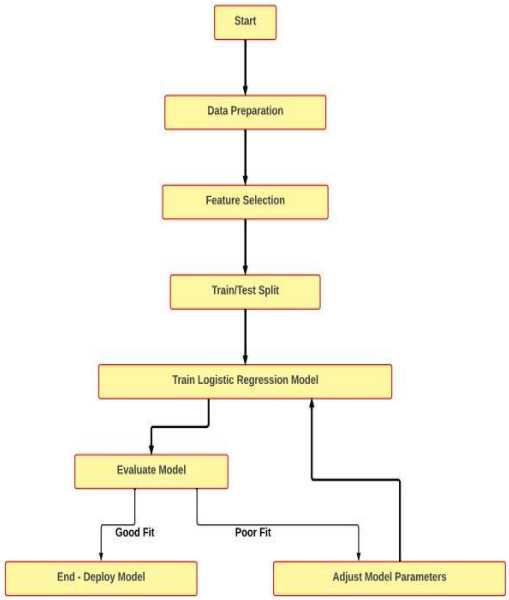

Fig 3.2: Binomial Logistic Regression Architecture

based on one or more predictor variables. In the realm of churn prediction, binomial logistic regression gauges the likelihood of a user churning by considering variables like session frequency, drive duration, and device type. It calculates the log odds of the probability of churn as a linear combination of the predictor variables, which is then transformed into a probability using the logistic function. Binomial logistic regression provides interpretable coefficients that indicate the influence of each predictor variable on the probability of churn.

Random Forest:

RF is a clever method that builds many decision trees together during training. It then combines the results of these trees to figure out the most common class for

classification tasks or the average prediction for regression tasks. Each tree in the RF is trained on a random part of the training data and some random features, which helps avoid overfitting and makes the model more flexible. In churn prediction, RF is great at understanding complex interactions between different factors and handling situations where the relationship between variables isn't straightforward. It also provides feature importances, indicating which features are most influential in predicting churn.

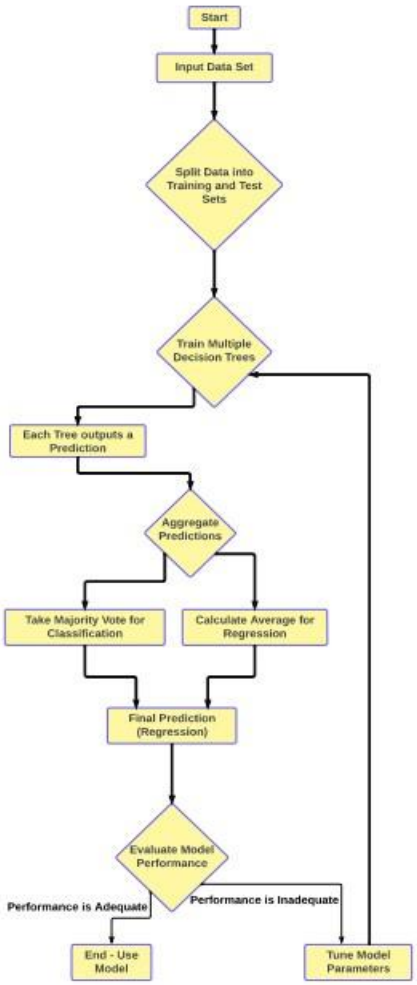

Fig 3.3: Random Forest Architecture

XGBoost:

XGBoost, short for Extreme Gradient Boosting, represents a highly optimized version of gradient boosting, a machine learning approach that constructs a predictive model through an ensemble of weak learners, often decision trees. XGBoost iteratively builds trees to minimize a predefined loss function, optimizing both the model's predictive performance and complexity. It is known for its speed, scalability, and high predictive accuracy, making it suitable for churn prediction tasks . Additionally, it offers feature importance scores, allowing users to identify the most influential features in predicting churn.

5. Model Evaluation:

Comparing the performance of implemented models using evaluation metrics on the testing set, assessing metrics such as accuracy, precision, recall, and F1-score helps determine the best-performing model for churn prediction.

6. Deployment and Monitoring:

Putting the chosen churn prediction model into a production environment for real-time predictions. Setting up monitoring systems to observe model performance as time progresses and periodically retraining the model with fresh data to sustain accuracy.

IV Experimental Results

The data reveals a notable relationship: increased distance driven per driving day is positively associated with user churn, signifying that as users cover greater distances during individual driving sessions, their probability of churning rises accordingly.

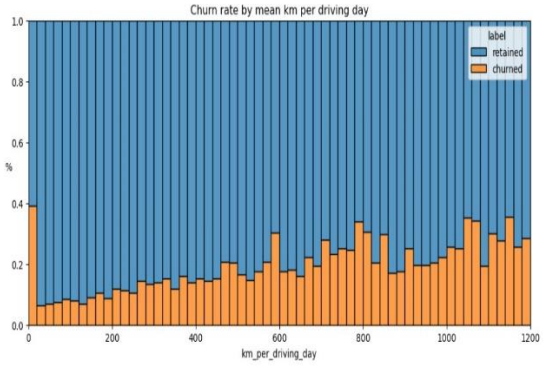

Fig 4.1: Killometer per driving day vs Churn

Conversely, a negative correlation is observed between the frequency of driving days and churn, indicating that users who participate in driving activities on more days within the preceding month exhibit a reduced likelihood of churning.

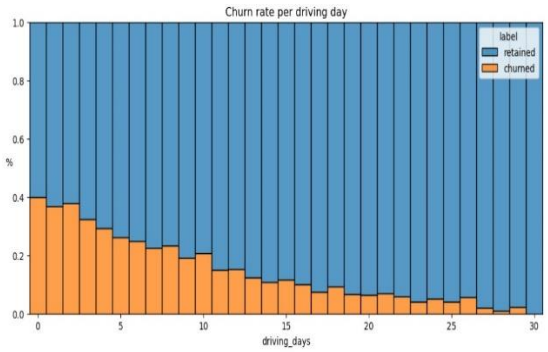

Fig 4.2: Driving days vs Churn

Binomial Logistic Regression :

The experiments are conducted using the Scikit-learn library. Given the dataset's imbalance, it is recommended to employ Accuracy and Recall scores as appropriate metrics for evaluating different classification algorithms.

When predictor variables exhibit a Pearson correlation coefficient exceeding the absolute value of 0.7, it indicates strong multicollinearity among these variables.

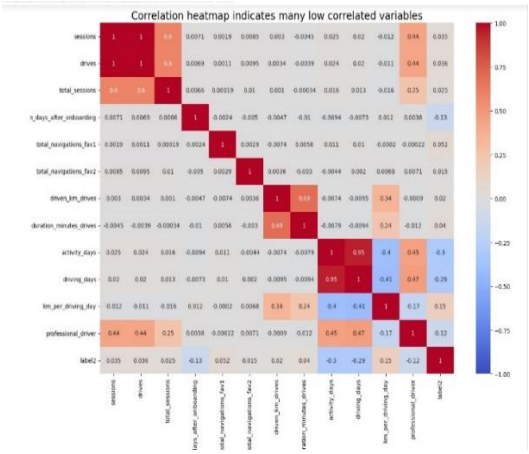

Fig 4.3: Correlation heatmap between Features

This statistical method is ideal for scenarios with binary outcome variables, like churn prediction, where the objective is to gauge the probability of an event occurring (such as churn or no churn) based on a given set of predictors. By examining the association between these predictors and the binary outcome, binomial logistic regression aids in identifying the factors influencing churn likelihood, thus enabling the project to develop effective prediction and intervention strategies.

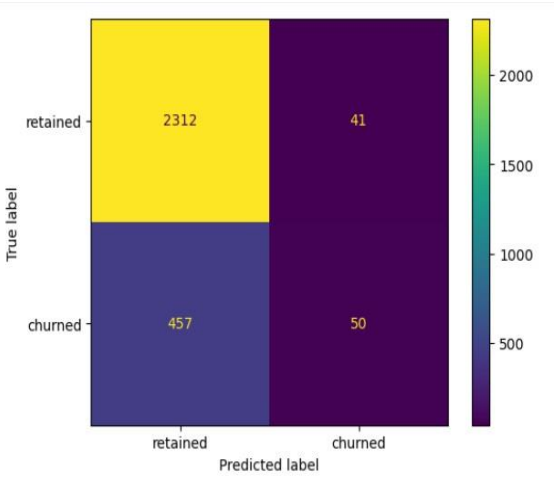

Fig 4.4: Logistic Regression Confusion Matrix

	precision	recall	f1-score	support
retained	0.83	0.98	0.90	2353
churned	0.55	0.10	0.17	507
accuracy			0.83	2860
macro avg	0.69	0.54	0.53	2860
weighted avg	0.78	0.83	0.77	2860

Fig 4.5: Logistic Regression Output

The model demonstrates mediocre precision and notably low recall, indicating a substantial number of false negative predictions and a failure to identify users who are likely to churn.

Random Forest and XGBoost Algoritthm:

	model	**precision**	**recall**	**F1**	**accuracy**
0	RF cv	0.45726	0.126782	0.198447	0.81851

Fig 4.6: Random Forest Output

Apart from accuracy, the overall performance of the scores remains subpar. However, it's noteworthy that during the construction of the logistic regression model in the previous iteration, the recall stood at approximately 0.09. This indicates an enhancement in recall by about 33% in the current model, alongside a comparable accuracy level, despite being trained on a smaller dataset.

	model	precision	recall	F1	accuracy
0	RF cv	0.457260	0.126782	0.198447	0.818510
0	XGB cv	0.442149	0.180033	0.255812	0.814198

Fig 4.7: XGBoost Output

The performance of this model surpasses even that of the random forest model. Its recall score nearly doubles that of the logistic regression model from the previous course, and it outperforms the random forest model by almost 50% in terms of recall, all while maintaining similar accuracy and precision scores.

	model	precision	recall	F1	accuracy
0	RF cv	0.457260	0.126782	0.198447	0.818510
0	XGB cv	0.442149	0.180033	0.255812	0.814198
0	RF val	0.438849	0.120316	0.188854	0.816783
0	XGB val	0.392670	0.147929	0.214900	0.808392
0	XGB test	0.421569	0.169625	0.241913	0.811538

Fig 4.8: Random Forest Output

While the recall remained consistent with the validation data, there was a notable decline in precision, resulting in slight drops across all other scores. Nevertheless, these deviations fall within an acceptable range for performance variation between validation and test scores.

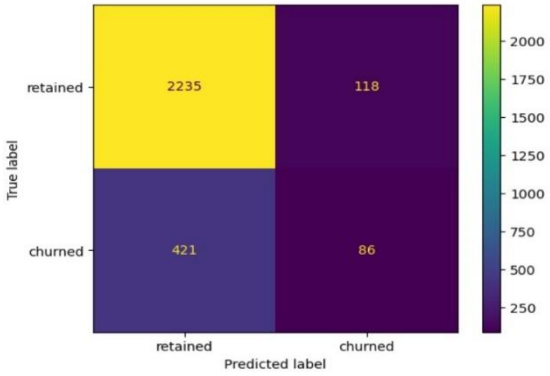

Fig 4.9 XGBoost Confusion Matrix

The model exhibited a significantly higher count of false negatives compared to false positives, approximately three times as many. Moreover, it precisely pinpointed only 16.6% of the users who genuinely churned.

	precision	recall	f1-score	support
Predicted would not leave	0.84	0.95	0.89	2353
Predicted would leave	0.42	0.17	0.24	507
accuracy			0.81	2860
macro avg	0.63	0.56	0.57	2860
weighted avg	0.77	0.81	0.78	2860

Fig 4.10 XGBoost Output

V. Conclusion

In summary, the Waze churn prediction project yielded notable insights. The developed model outperformed previous ones by achieving higher recall scores alongside acceptable precision and accuracy levels. However, its accuracy in identifying churned users was modest, standing at only 16.6%. Moreover, the substantial difference between false negatives and false positives underscores areas for improvement in future iterations. Although promising, the model requires further refinement to enhance its predictive prowess and effectively tackle churn prediction challenges within the Waze user community.

VI. REFERENCES

[1] Farquad H , Ravi Vadlamani, & Bapi Raju Surampudi. (2014). "Churn Prediction using Comprehensible Support Vector Machine: an Analytical CRM Application." Applied Soft Computing, 19, 10.1016/j.asoc.2014.01.031.

[2] Anil Kumar Dudyala,& Ravi Vadlamani.(2008). "Predicting credit card customer churn in banks using data mining." International Journal of Data Analysis Techniques and Strategies, 1, 4-28, 10.1504/IJDATS.2008.020020.

[3] Chih-Fong Tsai, Yu-Hsin Lu (2009). "Customer churn prediction through hybrid neural networks." Expert Systems with Applications, 36(2009) 12547-12553

[4] Wouter Verbeke, David Martens,chrstophe Mues& Bart Baesens. (2011). "Constructing intelligible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications, 38(3):2354–2364, 10.1016/j.eswa.2010.08.023

[5] Ning Lu, Hua Lin, Jie Lu,Guangquan Zhang. (2014). "A Customer Churn Prediction Model in Telecom Industry Using Boosting. 10.1109/TII.2012.2224355

[6] Karamollaoglu Hamdullah, Ibrahim Yucedag Yücedağ, & Ibrahim Alper Dogru(2021). "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis." 10.1109/UBMK52708.2021.9558876

[7] Ramena Venkata Satya Rohit, Chandrawat Dhrati & D. Rajeswari (2021). "Smart Farming Techniques for New Farmers Using Machine Learning." Proceedings of 6th International Conference on Recent Trends in Computing, Vol. 177.

[8] Ssu-Han Chen,(2016). "The gamma CUSUM chart method for online customer churn prediction." Electronic Commerce Research and Applications, 17, 99–111.

[9] Koen W. De Bock, & Dirk Van den Poel, (2011). "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction." 10.1016/j.eswa.2011.04.007

[10] Dhruv Sikka, Shivansh, Rajeshwari D., & Pushpaltha M (2022). "Prediction of Delamination Size in Composite Material Using Machine Learning." 2022 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1228-1232, doi: 10.1109/ICEARS53579.2022.9752123.

[11] Md Sayedur Rahman, Md Sahrial Alam, & Md Ikbal Hosen (2022). "To Predict Customer Churn By Using Different Algorithms." 2022 International Conference on Decision Aid Sciences and Applications (DASA), pp. 601-604, 10.1109/DASA54658.2022.9765155

[12] Koen W. De Bock,& Dirk Van den Poel, (2012). "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models." Expert Systems with Applications, 39, 6816–6826.

[13] Sarthak Sangamnerkar, Srinivasan Rajendran., Rajeev Sukumaran. (2020). "An ensemble technique to detect fabricated news articles using machine learning and natural language processing techniques.". 10.1109/INCET49848.2020.9154053

[14] Ning Lu, Hua Lin, Jie Lu, & Guangquan, Zhang. (2014). "A customer churn prediction model in the telecom industry using boosting." 10.1109/TII.2012.2224355

[15] Kartekay Goyal,Kumar Kanishka, Kanisk Vasisth, Sahil Kansal,& Ritesh Srivastava.(2021). "Telecom Customer Churn Prediction: A Survey." 10.1109/ICAC3N53548.2021.9725621

[16] Soumi De,Prabu P, & Joy Paulose (2021). "Effective ML Techniques to Predict Customer Churn." 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 895-902, doi: 10.1109/ICIRCA51532.2021.9544785.

[17] V Geetha, A Punitha, A Nandhini, T Nandhini, S Shakila, & R Sushmitha.(2020). "Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier."doi: 10.1109/ICSCAN49426.2020.9262288.