IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133142

CONNECTIVITY CRISIS: TACKLING TELECOM CHURN WITH MACHINE LEARNING

V. Chandana[1], S. Manohari[2], Y. Lakshmi Prasanna[3], SK. Feroze Moinuddee4,

DR. K.Pavan Kumar[5]

PVP SIDDHARTHA INSTITUTE OF TECHNOLOGY, KANURU VIJAYAWADA,
ANDHRA PRADESH, INDIA -520007.[1]

PVP SIDDHARTHA INSTITUTE OF TECHNOLOGY, KANURU VIJAYAWADA,
ANDHRA PRADESH, INDIA-520007.[2-5]

Abstract: Customer churn is when customers stop using a particular telecom service and switch to a competitor or cancel their contract altogether. This presents a major challenge in the telecommunication industry, as acquiring new customers is typically more expensive than retaining existing ones. To reduce the churn rate, businesses can analyze large volumes of customer data to gain insightful knowledge about customer behavior, preferences, and potential churn tendencies using machine learning algorithms. By utilizing machine learning models, telecom companies can gain an understanding of their customers' preferences and implement retention strategies that can increase customer satisfaction. In this study, we aim to illustrate the effectiveness of Random Forest, Cat Boost, and XG Boost models in accurately predicting customer attrition.

Keywords: Customer churn, Telecommunication Industries, Machine Learning algorithms, Retention Strategies, Insightful Knowledge.

I. INTRODUCTION

The telecommunications sector plays a central role in developed countries, facing challenges such as customer churn, where valuable customers switch to competitors. With scientific advancements and increased competition among operators, survival in this aggressive market relies on complex strategies. Customer churn poses a significant issue, leading to a loss of telecom services. Predicting customers likely to leave early can offer a valuable opportunity for additional revenue. Customer churn prediction plays a pivotal and really important role in the telecommunications industry as it helps in identifying customers who might be inclined to switch to a competitor's service or leave altogether. This analytical procedure is of utmost importance, as it involves the process of pinpointing those customers who are at a higher risk of churning in the near future, perhaps. The aim is to take proactive measures to retain these customers, leveraging data-driven methodologies. This predictive approach carries significant financial implications for the telecom sector. Acquiring new customers is typically a more expensive endeavor than retaining the existing ones, reducing churn becomes a key driver of enhanced profitability, definitely. Therefore, telecom companies stand to gain substantially from the ability to anticipate and mitigate customer churn.

II. LITERATURE STUDIES

According to paper [1],it offers a detailed overview of the methods, datasets, and performance metrics used in predicting telecom churn. The review provides valuable insights into the strategies and tools employed in this area, helping researchers and practitioners understand the current state of churn prediction in the telecom industry.

Likewise, in reference [2], the work in this paper is focusing on the complexities associated with predicting customer churn, viewed as a classification problem sensitive to costs. The authors propose a partition cost-sensitive CART model, designed to consider the diverse costs associated with misclassifications among distinct customer segments. By conducting empirical validation using actual data, Wang et al. illustrate the effectiveness of their method in improving classification accuracy while reducing overall misclassification expenses.

DOI: 10.17148/IJARCCE.2024.133142

By assessing these models using measures like true positive rate, false positive rate, and accuracy, businesses can accurately assess their predictive capabilities. Conducting an extensive examination of customer churn prediction within the telecommunications sector utilizing data mining methodologies presents an opportunity to explore prevalent approaches and optimal strategies in churn prediction tailored to this industry[7].

Highlighting the practical importance of customer churn prediction in telecommunications, particularly in enhancing customer retention and revenue, is emphasized. Utilizing machine learning techniques such as logistic regression, the research explores the development and assessment of customized classification models for telecom datasets. Essential evaluation criteria like true positive rate, false positive rate, and accuracy are employed as pivotal benchmarks for evaluating the predictive capabilities of these models [3].

The Journal of Big Data provides an extensive overview, as indicated by reference [4], presenting a wide-ranging perspective on common methodologies and optimal approaches in predicting customer churn within the telecommunications sector. By conducting a thorough examination of various machine learning techniques, such as logistic regression, decision trees, and ensemble methods, this review offers invaluable insights for professionals aiming to comprehend the intricacies of churn prediction.

The investigation delves into churn prediction using machine learning algorithms tailored specifically for the telecom sector, providing insights into the algorithms utilized, their parameter configurations, methods for feature selection, and the evaluation metrics utilized to gauge predictive performance [5].

As per reference [6], the research advocates for the adoption of hybrid data mining models that amalgamate diverse techniques to enhance the accuracy of churn prediction. By combining decision trees, neural networks, and clustering algorithms, this study showcases the potential of hybrid models in augmenting the effectiveness of churn prediction. Moreover, another investigation highlighted in reference [10] makes a noteworthy contribution to the realm of business analytics within telemarketing. This study furnishes practical insights and methodologies for executing cost-sensitive analyses of bank campaigns utilizing artificial neural networks (ANNs). Collectively, these findings underscore the significance of utilizing advanced analytical methods, such as hybrid data mining models and ANN-based analyses, to optimize marketing strategies and elevate predictive accuracy in telemarketing campaigns.

A study presented at a big data conference offers methods for predicting customer churn, particularly useful for telecom companies. Using advanced data analysis, the study suggests ways for businesses to anticipate when customers might leave their services, emphasizing the importance of good data, feature selection, and model choice for accurate predictions[8].

III. MOTIVATION

The motivation behind our research stems from the profound impact of customer churn on the telecommunications industry and the pressing need for effective churn prediction models. In the real world, telecom companies face significant challenges in retaining customers amidst intense competition and evolving consumer preferences. Customer churn not only leads to revenue loss but also undermines customer satisfaction and market competitiveness.

IV. PROBLEM DOMAIN

The problem domain of predicting customer churn in the telecom industry using machine learning involves a multidisciplinary approach that integrates data collection, analysis, modeling, and deployment to effectively address the challenge of retaining customers and reducing churn rates. This includes considering certain features like account length, international plan, customer service calls etc.

V. PROBLEM DEFINITION

Our perception of the problem highlights its multifaceted nature and its broader implications for society and business sustainability. By addressing the challenge of churn prediction, we aim to empower telecom companies to proactively manage customer attrition, optimize resource allocation, and enhance customer satisfaction. Ultimately, reducing churn can contribute to the long-term viability and profitability of telecom businesses, thereby benefiting both stakeholders and consumers. Being able to predict when customers may leave or churn helps businesses prioritize where they put their time and money. Predicting customer churn in the telecom industry is motivated by the desire to protect revenue, reduce costs, improve customer satisfaction, gain a competitive advantage, and make informed, data-driven decisions to ensure long-term business sustainability.

VI. STATEMENT

The problem addressed in this research is to develop effective machine learning models for predicting customer churn in the telecom industry, enabling proactive retention strategies to mitigate customer attrition and enhance profitability.

VII. INNOVATIVE CONTENT

Innovative Content Relative to Main References:

- Farhad Shaikh et al. (2021):

This study introduces a novel approach that combines natural language processing (NLP) techniques with machine learning for customer churn prediction in the telecom sector. By leveraging NLP on customer feedback data, the model gains deeper insights into customer sentiments and preferences, enhancing the accuracy of churn prediction.

- Chuanqi Wang et al. (2019):

The partition cost-sensitive CART model proposed in this research presents an innovative solution to address the varying misclassification costs across different customer segments. By integrating customer value considerations into the classification process, the model achieves enhanced performance in predicting churn while minimizing overall misclassification costs.

- Gaur and Dubey (2018):

This study explores various machine learning techniques tailored to the telecom sector for churn prediction. The innovative aspect lies in the comprehensive comparison and evaluation of these techniques, providing insights into their strengths and weaknesses in different predictive scenarios, thus guiding practitioners in selecting the most suitable approach.

- Journal of Big Data Review (2019):

The review offers an innovative synthesis of prevalent methodologies in telecom customer churn prediction, providing a comprehensive understanding of the landscape. By consolidating insights from various studies, it serves as a valuable resource for researchers and practitioners seeking to navigate the complexities of churn prediction.

- Amjad Hudaib et al. (2015):

This study advocates for the adoption of hybrid data mining models in churn prediction, combining different techniques to improve accuracy. The innovative aspect lies in the integration of decision trees, neural networks, and clustering algorithms, showcasing the potential of hybrid approaches to enhance predictive capabilities in telecom churn prediction.

VIII. DESIGN

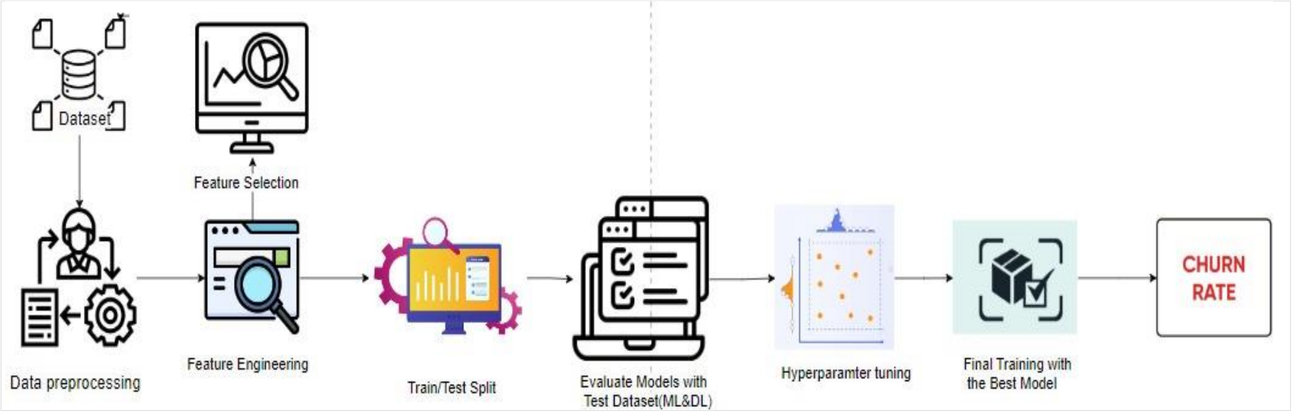


Fig 1: Workflow of predicting customer churn.

DOI: 10.17148/IJARCCE.2024.133142

IX. SOLUTION METHODOLOGIES

A. DATASET:

Datasets are fundamental components in various fields, aiding research, analysis, model development, and decision-making processes. They serve as the cornerstone for extracting data-driven insights and driving innovation forward. The dataset under consideration for our research contains 3,333 records and includes 21 attributes, sourced from Kaggle. This dataset offers valuable insights into customer characteristics, telephone usage patterns, and churn status. By leveraging predictive modeling techniques, researchers can gain deeper insights into the factors influencing customer behaviour. Combining both categorical and numerical attributes, this dataset is suitable for machine learning and statistical analyses.

B. DATA PREPROCESSING:

The original dataset contained 21 attributes. To prepare the data for analysis, we first cleaned it, transformed it, and organized it. This involved tasks like handling missing values, encoding categorical data, scaling numerical features, and removing outliers to improve data quality and model performance.

We paid special attention to dealing with missing values and made other adjustments like renaming columns and removing unnecessary attributes to simplify the dataset. Categorical variables were converted into binary numeric values to work better with analytical models. We used a tool called "missingno" to visualize and address any missing data. This helped us effectively handle gaps in the dataset.

VISUALIZATION OF MISSING VALUES:

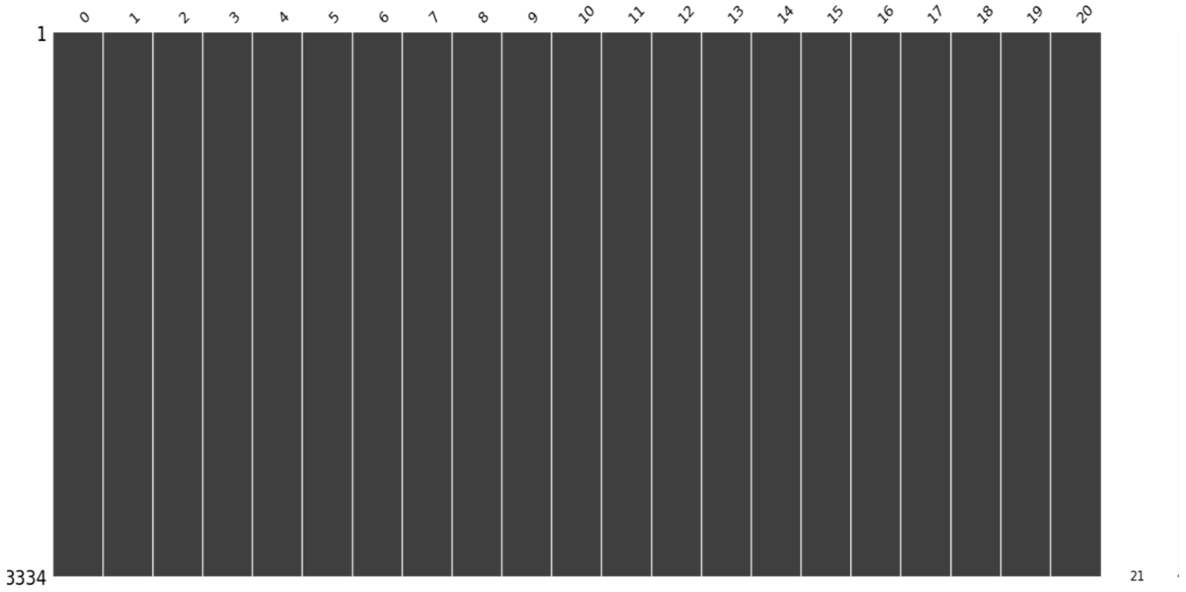


Fig 2: Matrix representation of missing data.

FEATURE ENGINEERING:

Feature engineering, a crucial aspect of machine learning, involves transforming raw data into meaningful features that significantly enhance predictive model performance. This process plays a pivotal role in improving the accuracy and interpretability of models.

In this specific section, we implemented several essential steps to extract and engineer relevant features. Initially, we adjusted the data types of specific columns and may have set error parameters to "coerce," effectively handling and replacing non-numeric values with NaN. Additionally, we performed transformations on certain columns to convert them into numerical representations. Furthermore, we created new features by deriving insights from existing attributes, with these engineered features proving to be essential for our final predictions.

DOI: 10.17148/IJARCCE.2024.133142

C. CLASSIFICATION:

• RANDOM FOREST:

The Random Forest algorithm functions as a collective decision-making process, aggregating the opinions of multiple individual decision trees rather than relying on a single source. This approach involves averaging the predictions of numerous decision trees, resulting in a robust and flexible method for predicting customer churn in the telecom industry. To enhance the performance of the Random Forest algorithm, we fine-tuned its parameters using a technique known as hyperparameter tuning. Additionally, we employed K-Fold cross-validation to assess the model's performance across different datasets, thereby reducing the risk of overfitting -and ensuring its effectiveness in diverse situations. These steps aim to enhance the accuracy of the model, thereby increasing its reliability for predictive tasks.

• XG BOOST:

XG Boost is a robust ensemble learning method that constructs potent predictive models by aggregating predictions from weaker models, typically decision trees. Its effectiveness lies in iteratively learning from errors during each training round, thereby enhancing accuracy. In our model development process, we extensively leveraged the capabilities of XG Boosting. We meticulously trained the model using our extensive dataset, enabling the extraction of valuable insights and patterns from the data. Additionally, to further improve the model's predictive performance, we conducted hyperparameter tuning.

This meticulous fine-tuning process was conducted through RandomizedSearchCV, a tool that systematically explores various hyperparameter configurations to identify settings that optimize model accuracy. Through these efforts, we significantly enhanced the model's predictive capabilities, making it adept at handling real-world data and producing highly accurate results.

• CAT BOOST:

Cat Boost is a gradient boosting algorithm that constructs predictive models by amalgamating numerous weak learners. It is specifically engineered to adeptly handle categorical features, eliminating the necessity for one-hot encoding or label encoding. In the telecom industry, customer data frequently contains categorical variables such as subscription plans, location, and types of services. Cat Boost's capability to directly handle these variables simplifies preprocessing and enhances model accuracy.

It employs a unique technique known as ordered boosting, which transforms categorical features into numerical ones. Additionally, Cat Boost incorporates advanced methodologies to mitigate overfitting and enhance generalization performance. With its efficient implementation of gradient boosting coupled with decision trees, Cat Boost is well-equipped to handle vast datasets comprising millions of examples and thousands of features.

E. HYPER PARAMETER TUNING:

In developing our machine learning model, hyperparameter tuning emerges as a crucial step, significantly impacting model performance. We rigorously explored hyperparameter optimization through two distinct approaches.

Firstly, K-fold cross-validation provided robust and unbiased assessments of hyperparameter choices.

Secondly, we employed RandomizedSearchCV, structuring a hyperparameter grid resembling a dictionary to randomly sample configurations. This method contrasts with traditional grid search by autonomously exploring the hyperparameter value space.

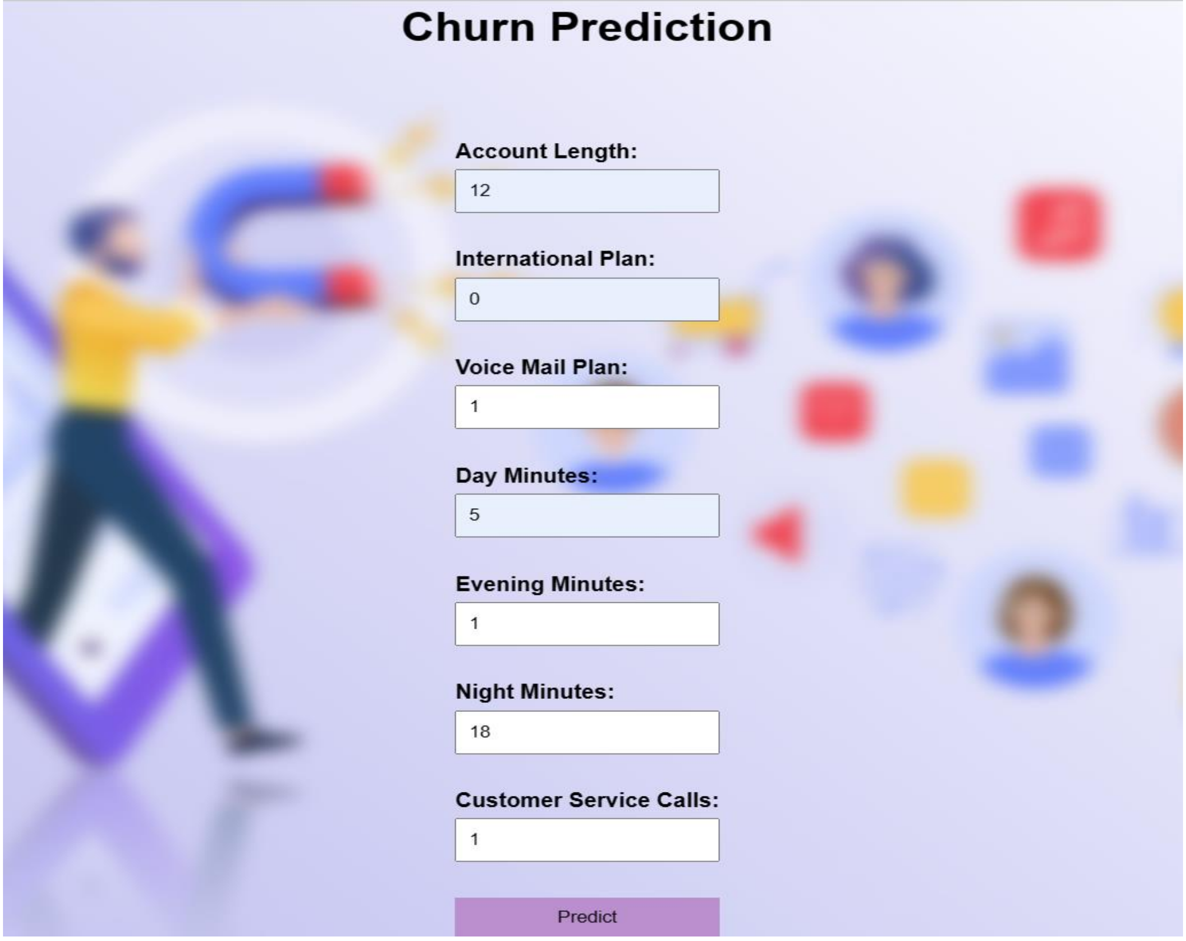
X. RESULTS AND ANALYSIS

A web application has been created using flask frame work featuring a form prompting users to input their customer information. Within this form, key features extracted from our dataset are included. Based on the provided data, machine learning models will generate predictions regarding whether a customer is likely to churn.

This empowers telecom companies to take preemptive measures to retain their customers, allowing them to focus resources on those who are at risk of churning, thus optimizing their time and investments.

DOI: 10.17148/IJARCCE.2024.133142

INPUT 1:



Churn Prediction

Account Length: 12

International Plan: 0

Voice Mail Plan: 1

Day Minutes: 5

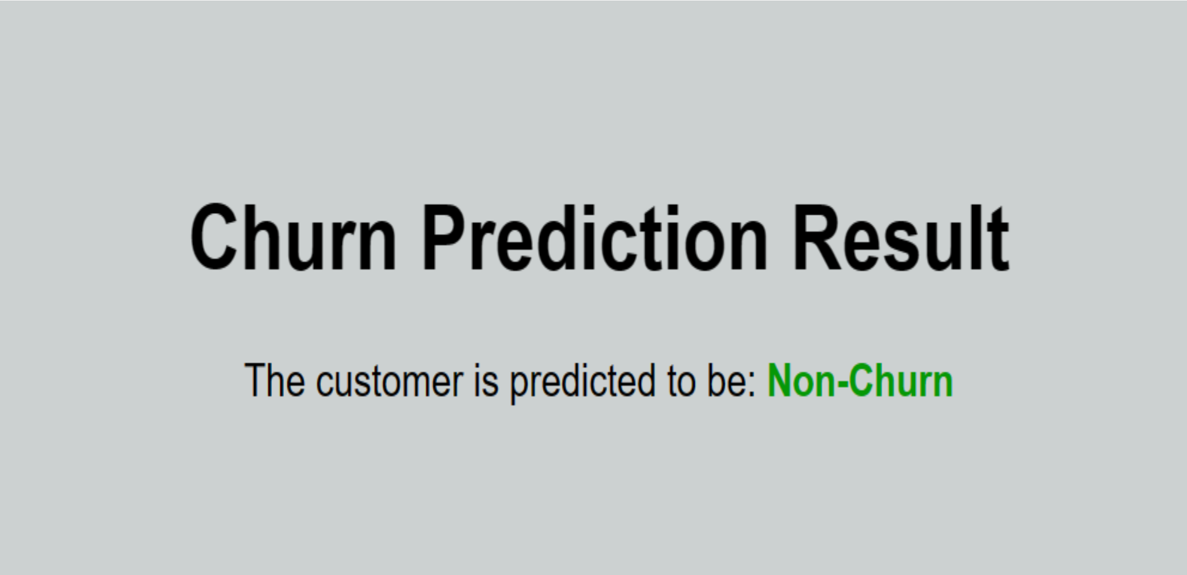
Evening Minutes: 1

Night Minutes: 18

Customer Service Calls: 1

Fig 3: Form to input customer data.

OUTPUT 1:



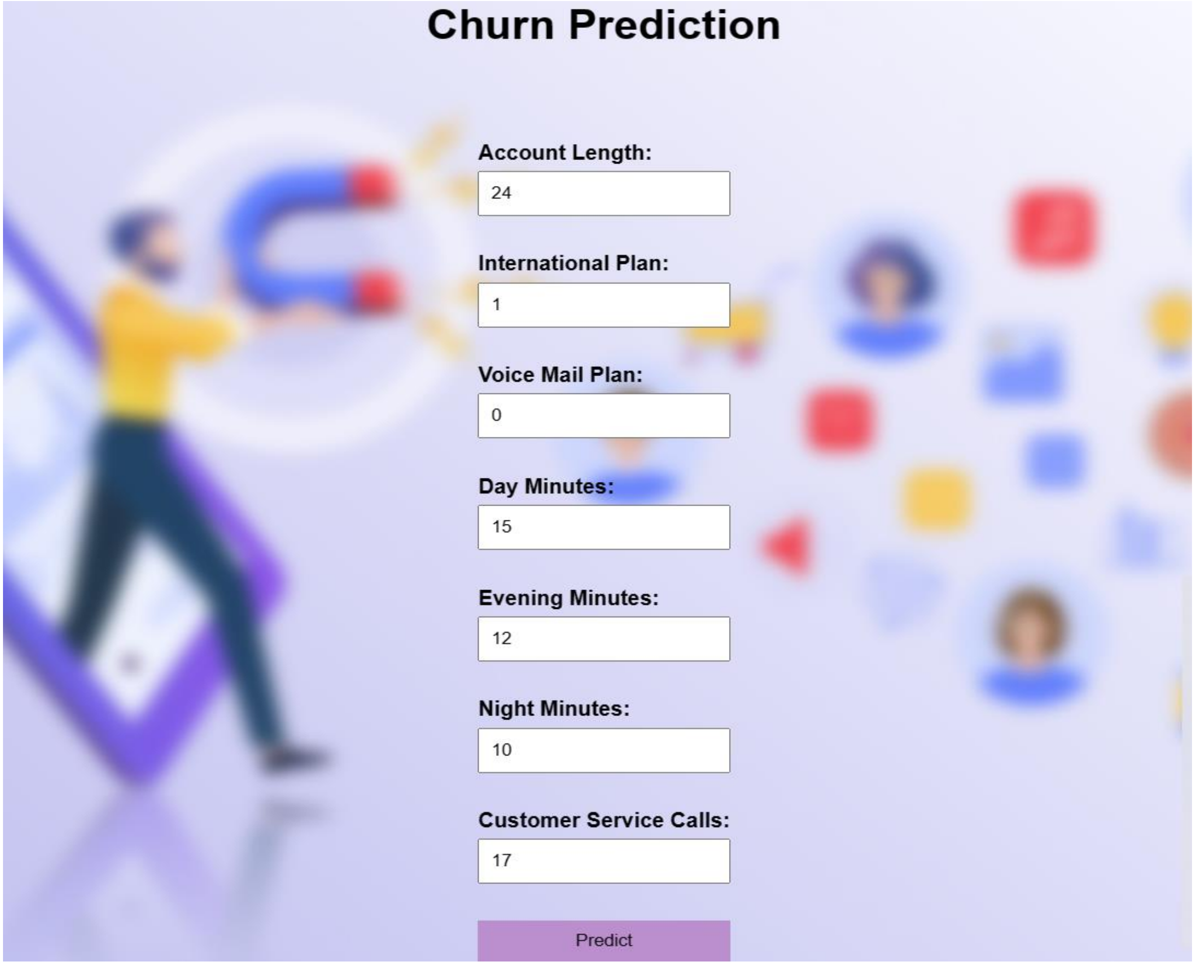
Churn Prediction Result

The customer is predicted to be: Non-Churn

Fig 4: Prediction result.

DOI: 10.17148/IJARCCE.2024.133142

INPUT 2:



Churn Prediction

Account Length: 24

International Plan: 1

Voice Mail Plan: 0

Day Minutes: 15

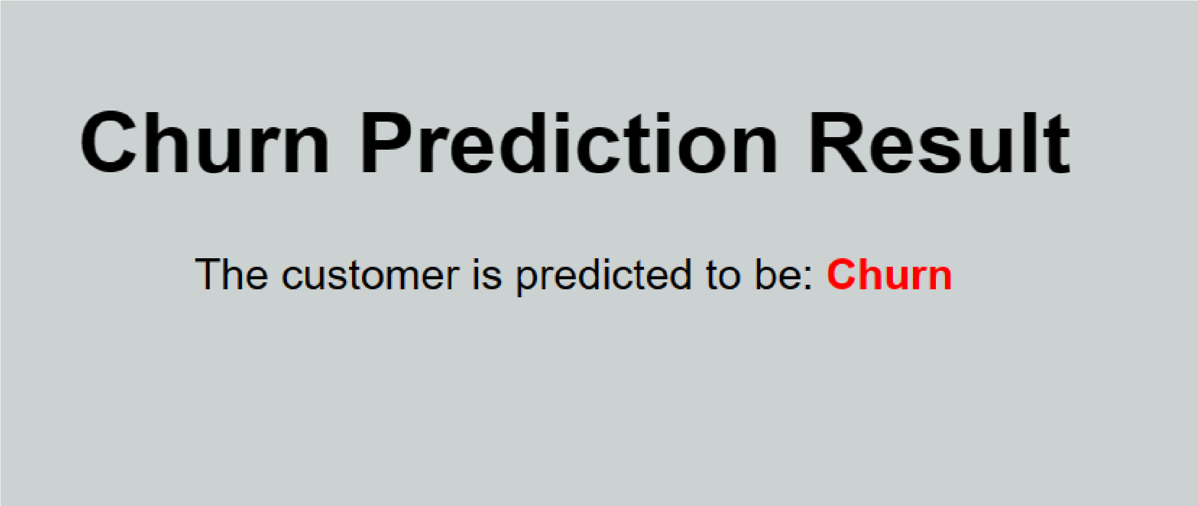
Evening Minutes: 12

Night Minutes: 10

Customer Service Calls: 17

Fig 5: Form to input customer data.

OUTPUT 2:



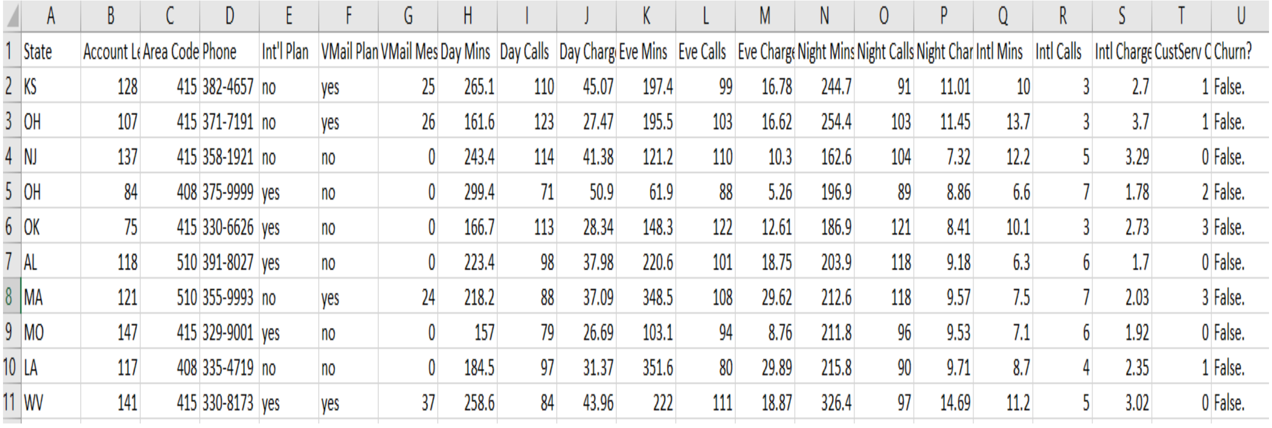
Churn Prediction Result

The customer is predicted to be: Churn

Fig 6: Prediction result.

DOI: 10.17148/IJARCCE.2024.133142

XI. DATA MODEL



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	State	Account Le	Area Code	Phone	Int'l Plan	VMail Plan	VMail Mes	Day Mins	Day Calls	Day Charg	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Char	Intl Mins	Intl Calls	Intl Charge	CustServ C	Churn?
2	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	11.01	10	3	2.7	1	False.
3	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	11.45	13.7	3	3.7	1	False.
4	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	121.2	110	10.3	162.6	104	7.32	12.2	5	3.29	0	False.
5	OH	84	408	375-9999	yes	no	0	299.4	71	50.9	61.9	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False.
6	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False.
7	AL	118	510	391-8027	yes	no	0	223.4	98	37.98	220.6	101	18.75	203.9	118	9.18	6.3	6	1.7	0	False.
8	MA	121	510	355-9993	no	yes	24	218.2	88	37.09	348.5	108	29.62	212.6	118	9.57	7.5	7	2.03	3	False.
9	MO	147	415	329-9001	yes	no	0	157	79	26.69	103.1	94	8.76	211.8	96	9.53	7.1	6	1.92	0	False.
10	LA	117	408	335-4719	no	no	0	184.5	97	31.37	351.6	80	29.89	215.8	90	9.71	8.7	4	2.35	1	False.
11	WV	141	415	330-8173	yes	yes	37	258.6	84	43.96	222	111	18.87	326.4	97	14.69	11.2	5	3.02	0	False.

Fig 7: Dataset containing customer details.

XII. COMPARISION OF RESULTS

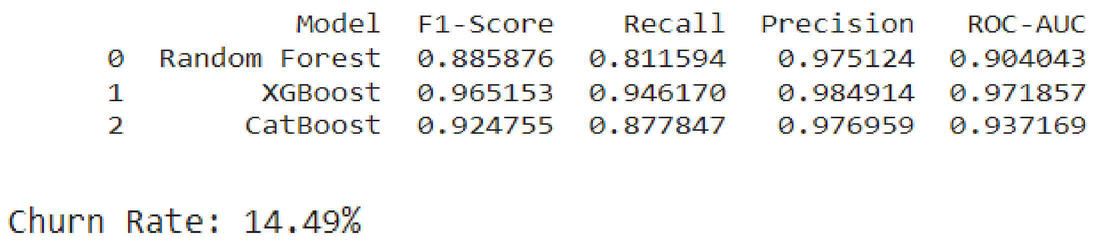
The previous research focus on utilization of traditional models like SVM, Random forest and logistic regression to analyse the behaviours of the customers and the hidden patterns in the dataset in order to predict the churn. The present study focuses on advanced machine learning models like Random Forest, XG Boost, and Cat Boost.

These models are specifically chosen for their effectiveness in handling complex telecom datasets and providing more accurate predictions. Additionally in this research we implemented methodologies like feature engineering and hyper parameter techniques to enhance the performance of the models and use metrics such as accuracy, precision, recall, and F1-score, to compare which model performs best in predicting customer churn within the telecom industry.

So the previous study uses some traditional models which are used to predict the churn. Our present study uses newer and more advanced machine learning techniques and performs some advance methodologies to yield outcomes with greater accuracy.

The summary table of the present study—

ACCURACY METRICS:



```
           Model  F1-Score    Recall  Precision   ROC-AUC
0  Random Forest  0.885876  0.811594   0.975124  0.904043
1        XGBoost  0.965153  0.946170   0.984914  0.971857
2       CatBoost  0.924755  0.877847   0.976959  0.937169


Churn Rate: 14.49%
```

Fig 8: Summary table on evaluation metrics.

© IJARCCE This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2024.133142

BARCHART:

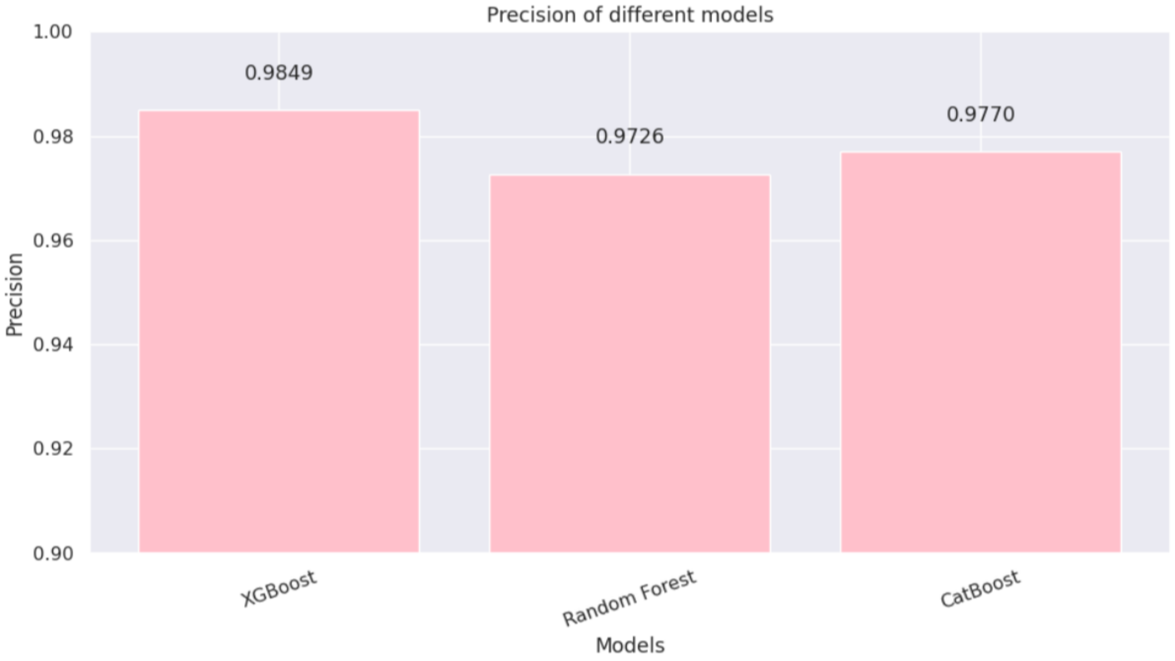


Fig 9: Bar chart on precision metrics.

XIII. JUSTIFICATION OF RESULTS

The justification provided highlights how each reference aligns with and reinforces the techniques and approaches used in our study on predicting customer churn. By citing relevant references, we demonstrated that our methodologies like performing feature engineering and hyper parameter tuning for improving the performance of our machine learning models is grounded in established research and best practices within the field. This strengthens the validity and credibility of our results, as they are supported by existing literature and methodologies proven effective.

XIV. CONCLUSION

In conclusion, our study focused on predicting customer churn within the telecom industry utilizing machine learning models, here we specifically used Random Forest, XG Boost, and Cat Boost. Through rigorous analysis, we found that XG Boost consistently outperformed the other models, exhibiting the highest accuracy rate around 97.18% among them. This indicates its superior capability in discerning patterns within the data and accurately predicting customer churn. The adoption of XG Boost holds significant promise for telecom companies seeking to proactively retain their customer base by identifying potential churners with greater precision. Such insights empower companies to allocate resources more effectively, thereby enhancing customer retention strategies and ensure long-term sustainability in a highly competitive market.

Future work can concentrate on automating the analysis of customer feedback from surveys, reviews, and support interactions using NLP techniques. Sentiment analysis can extract sentiment polarity like (positive, negative, neutral) from text data, providing valuable insights into customer sentiment and satisfaction levels related to telecom services. This can help identify common issues and trends that might lead to churn, prompting action from the retention team.

REFERENCES

[1]. Jain, Hemlata, Ajay Khunteta, and Sumit Srivastava. "Telecom churn prediction and used techniques, datasets and performance measures: a review." Telecommunication Systems 76 (2021): 613-630.

[2]. Chuanqi Wang, Ruiqi Li, Peng Wang, Zonghai Chen,“Partition cost-sensitive CART based on customer value for Telecom customer churn prediction” in Proceedings of the 36th Chinese Control Conference 2019 IEEE.]

This work is licensed under a Creative Commons Attribution 4.0 International License

DOI: 10.17148/IJARCCE.2024.133142

[3]. Gaur, A., & Dubey, R. (2018). Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques. 2018 International Conference on Advanced Computation and Telecommunication(ICACAT).

[4].Customer churn prediction in telecom using machinelearning https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6.

[5]. V. Kavitha, G. Hemanth Kumar, S.V Mohan Kumar, M. Harish, "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms", May 2020 (IJERT).

[6]. Amjad Hudaib, Reham Dannoun, Osama Harfoushi, Ruba Obiedat, Hossam Faris "Hybrid Data Mining Models for Predicting Customer Churn", J. Communications, Network and System Sciences, May 2015.

[7]. Kiran Dahiya, Surbhi Bhatia, "Customer Churn Analysis in Telecom Industry" in IEEE 2015, 978-1-4673-7231-2/15.

[8]. Vadakattu, B. Panda, S. Narayan, and H. Godhia, ''Enterprise subscription churn prediction,'' in Proc. IEEE Int. Conf. Big Data, Nov. 2015, pp. 1317–1321.

[9]. Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271-294.

[10].Nazeeh Ghatasheh, Hossam Faris, AlTaharwa Ismail, Yousra Harb, Ayman Harb, "Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks" in MDPI 2020.

[11]. Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1196-1201). IEEE.

[12]. Dingli, A., Marmara, V., & Fournier, N. S. (2017). Comparison of deep learning algorithms to predict customer churn within a local retail industry. *International journal of machine learning and computing*, *7*(5), 128-132.

[13]. Irfan Ullah, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, Sung Won Kim: "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Industry", in IEEE May 2019.

[14]. Nazeeh Ghatasheh, Hossam Faris, AlTaharwa Ismail, Yousra Harb, Ayman Harb, "Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks" in MDPI 2020.