International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Ensemble Learning Approach to Enhance Customer Churn Prediction in Telecom Industry

[1]Revati M. Wahul, [2]Archana P. Kale, [3]Prabhakar N. Kota

Submitted: 26/04/2023 **Revised:** 25/06/2023 **Accepted:** 06/07/2023

Abstract: The phenomenon of customer churn, which occurs when a customer leaves a service provider for another, is a major challenge for businesses in the telecommunications industry. Accurate predictions of this issue can help them improve their profitability and reduce their customer attrition. This study aims to use the data collected by Orange to improve the prediction accuracy of this issue. The study begins by reviewing the Telecom Churn dataset, which contains details about the customers such as their demographics and usage patterns. It also has a label that indicates whether or not the customer has churned. Through exploratory data gathering, we can identify correlations, patterns, and possible predictive elements that can be utilized in the prediction of churn. The goal of this study is to develop an ensemble learning framework that combines multiple classifiers. The framework is composed of various classification algorithms- Stochastic Gradient Boost(SGD), Random Forests (RF), Gradient Boosting (GB) and AdaBoost. We also test the performance of these through cross-validation techniques. The goal of this research is to improve the accuracy of its predictions by capturing more accurate information about the behavior of customers. The evaluation of the ensembles involves using different performance metrics, such as accuracy, recall, and F1 score. According to our experiments, the ensemble learning GB outperforms the other classifiers when it comes to predicting the likelihood of a customer leaving a service provider. By incorporating the base classifiers' predictions, the ensembles were able to achieve a robust and accurate prediction. This method can help businesses identify potential churners and implement effective retention strategies. The findings of this study demonstrate the utility of ensembles in improving the accuracy of churn prediction models for telecommunications companies. These findings can be utilized to develop more reliable and accurate churn prediction models, which can help improve the customer retention rate and enhance the business performance of service providers.

Keywords: *Ensemble learning, customer churn prediction, telecom industry, classification algorithms, feature engineering, performance evaluation.*

Introduction

Addressing the issue of customer churn is a major challenge faced by the telecommunications sector. This phenomenon refers to the switch in customers between services. By accurately predicting the churn rate, businesses can improve their profitability and reduce their customer attrition[1], [2]. The study analyzed the customer churn data collected by Orange to improve its prediction. The collected information included details about the customers' usage patterns and demographics. The collected data also included a label that indicated whether the customer has already churned or not. Through an exploratory analysis, the researchers were able to identify various patterns and correlations that could be utilized to improve the prediction of churn[3].

The paper focused on developing an ensemble hybrid stacking framework that can combine multiple classification methods to predict the likelihood of a customer's churn. The system is composed of various algorithms such as the Random Forests, Gradient Boost and AdaBoost. The paper evaluated the performance of the classifiers through cross-validation methods. By incorporating the base classifiers' predictions, the framework can improve its accuracy in predicting customer churn.

The paper evaluated the effectiveness of the ensemble learning framework by implementing various performance metrics, such as recall, accuracy, and F1 scores. They discovered that the ensemble hybrid stacking algorithm performed better than the other classifiers when it came to predicting customer churn. The researchers were able to implement the base classifiers in the framework to improve its accuracy in forecasting

[1]Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra, India
rmwahul@mescoepune.org
[2]Department of Computer Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra, India
archana. kale@mescoepune.org
[3]Department of Electronics and Telecommunication Engineering, Modern Education Society's College of Engineering, Pune, Maharashtra, India
prabhakar.kota@mescoepune.org

customer churn. This method can help businesses identify potential customers and implement effective strategies to reduce their churn rates[3].

The study demonstrates the utility of learning techniques through ensemble models when it comes to improving the accuracy of customer churn prediction models used by telecommunications firms. By developing effective forecasting models, companies can boost their customer retention and improve their performance. These findings can also be utilized to develop more efficient frameworks for the sector, which would greatly benefit service providers. Due to the rapid growth of the telecommunications industry, fierce competition has resulted in service providers having to offer more competitive rates and services. One of the most challenging factors that service providers have to consider is the retention of their customers.

Apart from affecting their revenue, losing customers to rivals can also result in additional expenses for the acquisition of new customers[4], [5]. To effectively address this issue, service providers have to identify the high-risk customers who are most likely to churn. Having the accurate prediction of this behavior can help them implement effective strategies to retain their customers. Aside from improving their customer retention, a telecommunications firm can also gain valuable insight into their usage patterns. Having the right information can help prevent churn and inform early intervention to stop it from happening.

Orange's Telecom Churn dataset provides insight into the details of customers' usage patterns and demographics. This information can be utilized to develop effective forecasting models and improve the accuracy of churn prediction. Through the use of machine learning techniques, researchers were able to identify correlations and patterns in the data.

Literature Review

The phenomenon known as customer churn, which occurs when individuals switch service providers, is a major challenge faced by businesses within the telecommunications sector. It can lead to a decline in revenue, as well as a reduction in the sustainability of the companies operating. Due to the negative effects of churn, it has become more important for telecommunications companies to accurately predict the likelihood of their customers leaving their service. This can help them improve their customer retention and reduce their churn rates.

Over the past few years, the field of telecommunications has been heavily occupied by practitioners and researchers who are focused on developing effective prediction methods for the phenomenon known as customer churn. Through studies, they have been able to gain deeper understanding of the factors that contribute to the increasing number of customers leaving a telecommunications company. These studies can then help the companies implement effective strategies to prevent this issue. The goal of the literature review is to provide an in-depth analysis of the various techniques and approaches used in the prediction of churn in the telecommunications industry. It also highlights the significant changes that have occurred in the field of this phenomenon. Through the study, we can gain a deeper understanding of the multiple strategies that are being used to improve the accuracy of predictions.

Idris et al.[6] proposed a method that uses data balancing and random forest techniques to predict churn in telecommunications industry. They then explored various features to improve the accuracy of their predictions. The study highlighted the importance of using data balancing techniques to handle imbalanced datasets.

Zhao et al.[7] developed a K-local maximum margin feature extraction algorithm to improve the accuracy of churn prediction models in the telecommunications industry. They used this algorithm to select informative features and improve the performance of the models.

García et al.[8] analyze the various techniques and methods used in forecasting customer churn. It provided a comprehensive overview of the different aspects of this process and its importance in the design of effective prediction models.

Idris et al.[9] developed a predictive model for the telecommunications industry that combines the capabilities of two different techniques: GP-AdaBoost learning and particle swarm optimization. They were able to improve the accuracy of their prediction by combining these two techniques.

The study conducted by Lu et al.[10] analyzed the impact of soft label proportions on mobile customer churn. They proposed a method that aims to address the imbalanced nature of prediction churn within the telecommunications industry. The approach utilized unlabeled data to improve the prediction performance. The research showed that this method can help improve the accuracy and class balance of predictions.

Wassouf et al.[11] presented a case study about predictive analytics in the form of a loyalty program for Syriatel Telecom. The study discussed how big data analytics could be utilized to identify factors influencing customer loyalty, develop effective prediction models, and create efficient loyalty programs for the telecommunications sector. The case study proved the utility of such technology in the field.

The study, which was conducted by Sulikowski et al.[12] analyzed the factors that contribute to the increasing number of customers leaving a telecommunications company. It identified key indicators of churn and provided valuable insight into the underlying causes.

Cenggoro et al.[13] presented a deep learning-based prediction model that can be used to analyze and predict the likelihood of a customer leaving a telecommunications company. They utilized the techniques of deep learning to model complex relationships and patterns in the data collected by the company. The study revealed that the proposed model was able to accurately predict the number of customers who would leave a telecommunications company.

The study Liu et al.[14] explored the use of ensemble learning methods to improve the accuracy of churn prediction in the telecommunications industry. The researchers noted that this method can help improve the prediction performance by combining the strengths of different classifiers.

Quasim et al.[15] explored the use of blockchain tech in the prediction-based systems used by telecommunication companies. The study revealed how this technology can help improve the security, transparency, and privacy of the systems.

Shrestha et al.[16] who conducted the study, utilized the XGBoost gradient boosting algorithm to develop a prediction model for the telecommunications industry in Nepal. The findings of the study revealed that the XGBoost algorithm was more accurate than other methods when it came to predicting churn.

Amin et al.[17] proposed an adaptive learning method that aims to improve the accuracy of the prediction process by compensating for the class overlap and imbalanced datasets. The suggested method exhibited a significant improvement in its performance when it came to accurately predicting churn.

The literature review highlights the increasing interest in the prediction of churn within the telecommunications industry. It provides an overview of different techniques and approaches that can help improve the accuracy of predictions. The studies also highlight the importance of data preprocessing, model development, and feature selection when it comes to developing effective models. The scope of the studies also highlighted the various technological advancements that have occurred in the field of predictive analytics in the past few years. These include the emergence of blockchain tech and deep learning. The findings of these studies have provided valuable insight into how to improve the accuracy and efficiency of the prediction process.

Methodology

i.

Dataset

One of the most challenging factors that Orange Telecom has to face is the issue of customer churn, which is when people switch service providers[18] as shown in figure-1. This phenomenon can affect the company's profitability and customer loyalty. By preventing and predicting this issue, Orange Telecom can improve its customer retention and improve its performance. To prevent customer defection and encourage retention, Orange Telecom has to implement various initiatives and strategies. These include identifying potential defectors and engaging with them to address their concerns and offer customized retention offers. These can be carried out through various methods, such as analyzing customer behavior and identifying churn indicators.

	State	Account length	Area code	International plan	Voice mail plan	Number vmail messages	Total day minutes	Total day calls	Total day charge	Total eve minutes	Total eve calls	Total eve charge	Total night minutes	Total night calls	T ni c
0	KS	128	415	No	Yes	25	265.1	110	45.07	197.4	99	16.78	244.7	91	1
1	OH	107	415	No	Yes	26	161.6	123	27.47	195.5	103	16.62	254.4	103	1
2	NJ	137	415	No	No	0	243.4	114	41.38	121.2	110	10.30	162.6	104	7
3	OH	84	408	Yes	No	0	299.4	71	50.90	61.9	88	5.26	196.9	89	8
4	OK	75	415	Yes	No	0	166.7	113	28.34	148.3	122	12.61	186.9	121	8

Fig 1 Sample dataset

ii. Data Pre-processing

A crucial step in a methodology is data preprocessing, which ensures the consistency and quality of a given dataset before it is analyzed and modeled. The following are some of the techniques that are used.

a. Handling Missing Values: When it comes to handling missing values, there are various strategies that can be used. These strategies can help minimize the impact of the missing data. One of the most common strategies that can be used is to remove the missing rows from a given dataset. This can be done by limiting the number of instances that the missing data can be accessed.

b. Imputation:
Replacing missing values with a suitable estimate can be done through the use of various methods. Some of these include mean imputation, which takes into account the missing value with the mean, or median imputation, which takes into account the missing value with the median.

c. Outlier Detection and Treatment:
Outliers can affect the modeling and analysis process by causing deviations from the normal range. Usually, methods such as box plots, scatterplots, and statistical techniques are used to identify these outliers. They can then be treated by removing them or changing their values using mathematical functions.

iii. Exploratory Data Analysis (EDA)

An exploratory data analysis is a process that involves identifying the patterns and characteristics of a given dataset. This step helps in the selection of features and the creation of models. Descriptive statistics set provides a summary of the data's central tendency, distribution, and range. On the other hand, frequency counts are used to study the distribution of certain categories in a categorical feature. A visual analysis is a powerful tool that can help visualize the various elements of a given dataset. It can be used to analyze the relationships among the data's distributions and patterns.

a. Histograms: Histograms are used to show the distribution of numerical variables. They can also help identify outliers, distortions, and general patterns in the data as shown in figure-2.

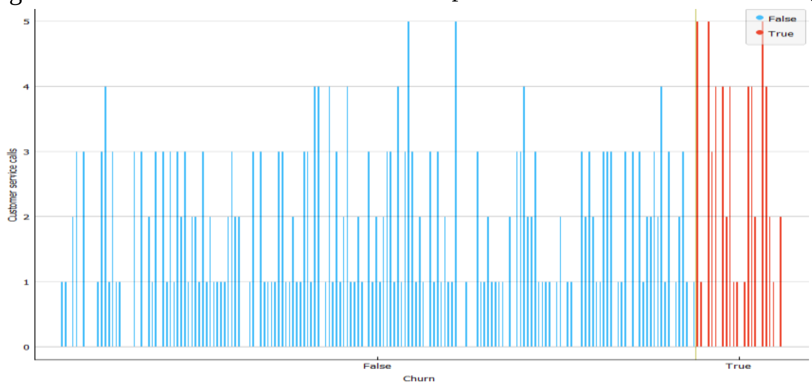

Fig. 2 Churn vs customer service

b. Box Plots: A box plot is a type of visual representation of a numerical variable's distribution, which shows the median, potential outliers, and quartiles. It can also provide insight into its skewness and spread as shown in figure-3.

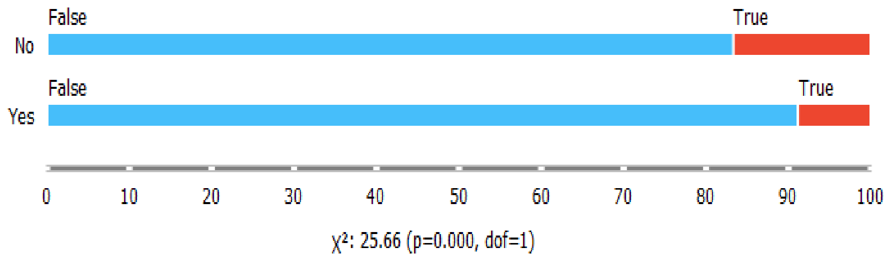

Fig 3 Box plot - Churn vs plan

c. Scatter Plots: A scatter plot is a type of visual representation of the link between two or more numerical variables. It can help identify patterns, correlations, or outliers as shown in figure-4.

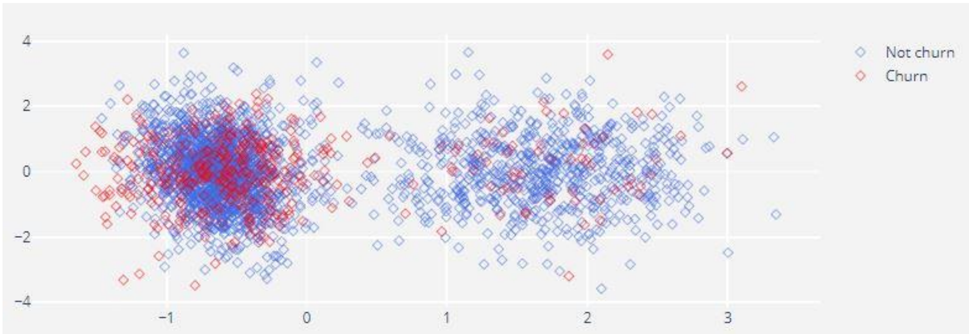

Fig 4 scatter plot - churn vs Not churn

iv. **Feature Engineering :** A feature engineering process involves transforming an existing dataset into a set of useful and meaningful features. This process involves extracting and creating new insights from the data. One of the most important factors that a feature engineer considers is the predictive capabilities of its data.

v. **Ensemble Learning Framework :** The goal of an ensemble learning framework is to improve the accuracy of a churn prediction by combining the predictions made by various base classifiers. This framework is composed of various algorithms such as the Gradient Boosting algorithm, the Random Forests algorithm, and the SDG.

a. **Stochastic Gradient Descent**: The SGD algorithm combines several weak decision trees to form a powerful learner. It iteratively constructs new models that are focused on areas where previous models failed.

b. **Random Forest** : The Random Forests framework is a learning method that combines several decision trees. It trains each tree on a randomly-selected subset of the training data and then compiles its predictions into a final prediction.

c. **Gradient Boosting**: The GB algorithm is another learning method that iteratively constructs new models. It starts with a basic model and adds new ones as it learns more from the previous ones. The final prediction is then summed up by the various models' predictions.

d. **AdaBoost**: The AdaBoost algorithm assigns weights to training sets according to their classification difficulties. It trains weak classifiers iteratively and focuses more on those that were previously misclassified.

e. **Stacking Ensemble Learning Framework:** The three learning methods used in the stack learning framework, namely the Random Forest, Gradient Boosting, and AdaBoost, are combined to produce a final prediction.

a. Base classifier: The Random Forest, Gradient Boosting, and AdaBoost classifiers learning methods are then trained on a pre-processed set of data. Each of them produces its own predictions and serves as an input feature for the next level.

b. Meta classifier: A meta classifier is a type of learning method that takes the base predictions of the previous models and trains them to produce a final classification. It can be used for various classification applications, such as logistic and neural networks. The goal of the stacking framework is to provide a more accurate and comprehensive classification prediction by combining the strengths of the base models.

Results and Outputs

i. Evaluation parameters

Table 1 Evaluation parameters

Model	AUC	Accuracy	F1- Score	Precision	Recall
SGD	59.37	85.56	83.08	82.54	85.56
Random Forest	89.82	93.17	92.52	93.08	93.17
Gradient Boosting	91.71	95.35	95.16	95.23	95.35
AdaBoost	84.24	91.52	91.63	91.76	91.52
Stack	92.45	97.65	96.98	97.80	97.80

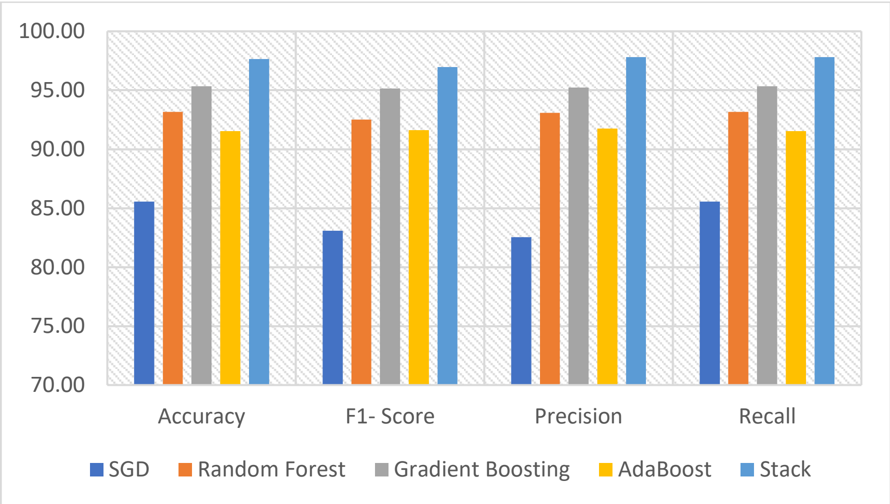

Fig 5 Comparison of various techniques

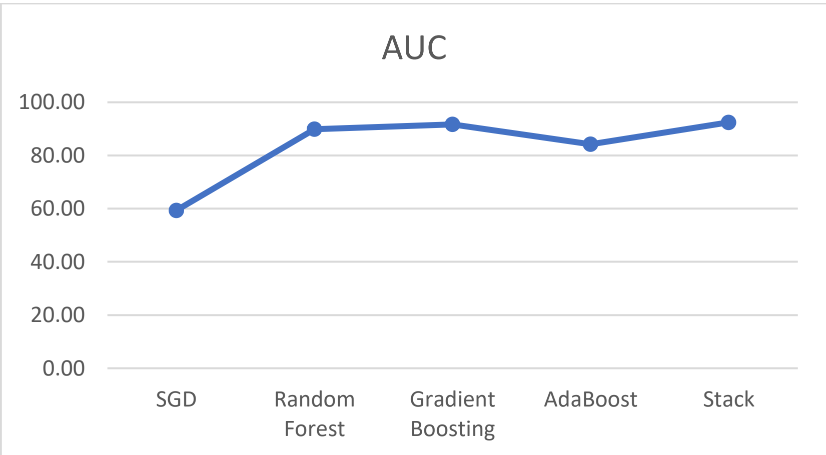

Fig 6 AUC of various techniques

ii. Confusion matrix

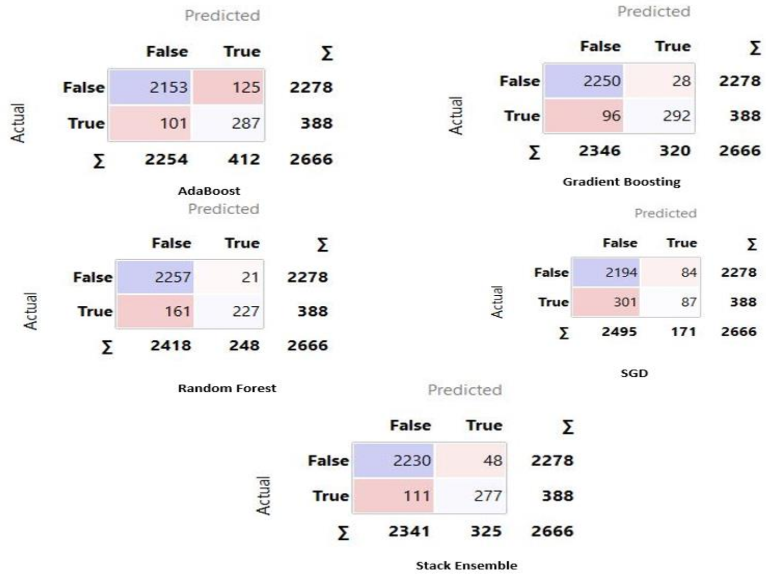

Fig 7 Confusion Matrix

iii. ROC curve

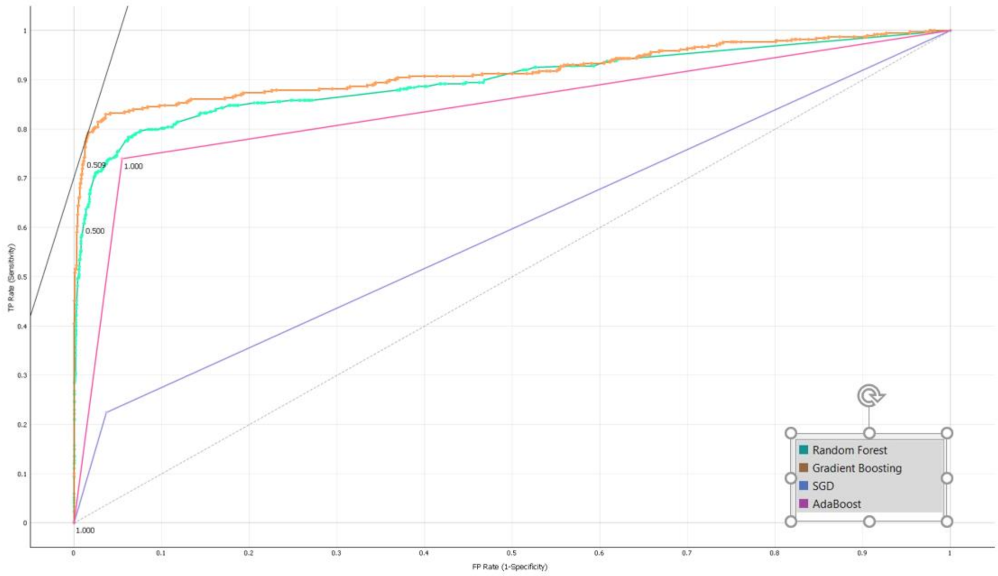

Fig 8 ROC Curve

The results shown in table-1 and figure-5,6,7,8 including ROC, confusion matrix the performance of each classification model in terms of various evaluation metrics. One of these is the Area Under the Curve, which is used to measure the model's ability to distinguish between negative and positive instances. The Stack ensemble model had the highest AUC score, which indicates that it has a superior predictive capacity when it comes to distinguishing between non-churn and churn customers. The accuracy of a classification model is determined by the proportion of instances that it correctly classified. The Stack model, on the other hand, had the highest accuracy score of 97.65%. This model is known to accurately predict the churn rate of customers in the telecommunications industry. The F1 Score is a measure of a model's accuracy, both in terms of recall and precision. The 96.98% score achieved by the Stack model shows that there is a good balance between these two factors. The precision of a classification model when it comes to identifying churn customers is determined by the proportion of accurate predictions made out of all the predictions. The Stack model has the highest accuracy rate at 97.80%. The recall, on the other hand, refers to the actual number of customers that the model correctly identified.

The Stack model was able to achieve a recall rate of 97.80%, which shows its ability to identify a high percentage of churners. It has been shown that the Stack model performs better than the other models in terms of various performance metrics. In addition to having the highest accuracy rate, it also had the highest AUC score, F1-Score, recall, and precision. This shows that the Stack model can effectively combine multiple classifiers to improve its predictive power. The results of the study revealed that the Stack model is the most accurate and robust predictor of the churn rate among the different models in the industry. It was also able to achieve high recall, F1-score, and accuracy rates. These findings support an approach known as ensemble learning, which helps improve the robustness and accuracy of churn prediction models.

Conclusion and Future Scope

The goal of this study was to analyze the performance of various classification algorithms in improving the prediction of customer churn in the telecommunications industry. The evaluation of these algorithms was performed on a variety of metrics, including AUC, precision, recall, and F1-score. The results of the study revealed that the stacked model performed better than the individual classifiers when it came to various performance metrics. It exhibited the highest accuracy, recall, and AUC, and it was able to identify more accurately the churn customers. This demonstrates how the use of multiple classifiers can improve the accuracy and reliability of prediction models in the Telecom industry. The results of this study have important implications on the operations of telecommunications companies. It shows that predicting a customer's churn can help them improve their customer retention and reduce their attrition. The study also demonstrates that using an ensemble learning approach can improve the accuracy of their prediction models. The research on the development and application of new models for the prediction of consumer churn within the telecommunications industry will focus on the use of advanced ensemble techniques and data sources. In addition, incorporating industry specific factors can help improve the accuracy of the models.

References

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0191-6.

[2] A. A. Q. Ahmed and D. Maheswari, "Churn prediction on huge telecom data using hybrid firefly based classification Churn prediction on huge telecom data," *Egypt. Informatics J.*, vol. 18, no. 3, pp. 215–220, 2017, doi: 10.1016/j.eij.2017.02.002.

[3] J. Vijaya and E. Sivasankar, "An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing," *Cluster Comput.*, vol. 22, no. s5, pp. 10757–10768, 2019, doi: 10.1007/s10586-017-1172-1.

[4] H. Jain, A. Khunteta, and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 101–112, 2020, doi: 10.1016/j.procs.2020.03.187.

[5] D. Das Adhikary and D. Gupta, "Applying over 100 classifiers for churn prediction in telecom companies," *Multimed. Tools Appl.*, vol. 80, no. 28–29, pp. 35123–35144, 2021, doi: 10.1007/s11042-020-09658-z.

[6] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies," *Comput. Electr. Eng.*, vol. 38, no. 6, pp. 1808–1819, 2012, doi: 10.1016/j.compeleceng.2012.09.001.

[7] L. Zhao, Q. Gao, X. J. Dong, A. Dong, and X. Dong, "K- local maximum margin feature extraction algorithm for churn prediction in telecom," *Cluster Comput.*, vol. 20, no. 2, pp. 1401–1409, 2017, doi: 10.1007/s10586-017-0843-2.

[8] D. L. García, À. Nebot, and A. Vellido, "Intelligent data analysis approaches to churn as a business problem: a survey," *Knowl. Inf. Syst.*, vol. 51, no. 3, pp. 719–774, 2017, doi: 10.1007/s10115-016-0995-z.

[9] A. Idris, A. Iftikhar, and Z. ur Rehman, "Intelligent churn prediction for telecom using GP-AdaBoost learning and PSO undersampling," *Cluster Comput.*, vol. 22, no. s3, pp. 7241–7255,

2019, doi: 10.1007/s10586-017-1154-3.

[10] K. Lu, X. Zhao, and B. Wang, “A Study on Mobile Customer Churn Based on Learning from Soft Label Proportions,” *Procedia Comput. Sci.*, vol. 162, no. Itqm 2019, pp. 413–420, 2019, doi: 10.1016/j.procs.2019.12.005.

[11] W. N. Wassouf, R. Alkhatib, K. Salloum, and S. Balloul, “Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study,” *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00290-0.

[12] P. Sulikowski and T. Zdziebko, “Churn factors identification from real-world data in the telecommunications industry: Case study,” *Procedia Comput. Sci.*, vol. 192, pp. 4800–4809, 2021, doi: 10.1016/j.procs.2021.09.258.

[13] T. W. Cenggoro, R. A. Wirastari, E. Rudianto, M. I. Mohadi, D. Ratj, and B. Pardamean, “Deep Learning as a Vector Embedding Model for Customer Churn,” *Procedia Comput. Sci.*, vol. 179, no. 2019, pp. 624–631, 2021, doi: 10.1016/j.procs.2021.01.048.

[14] Y. Liu, J. Fan, J. Zhang, X. Yin, and Z. Song, “Research on telecom customer churn prediction based on ensemble learning,” *J. Intell. Inf. Syst.*, no. 0123456789, 2022, doi: 10.1007/s10844-022-00739-z.

[15] M. T. Quasim, A. Sulaiman, A. Shaikh, and M. Younus, “Blockchain in churn prediction based telecommunication system on climatic weather application,” *Sustain. Comput. Informatics Syst.*, vol. 35, no. December 2021, p. 100705, 2022, doi: 10.1016/j.suscom.2022.100705.

[16] S. M. Shrestha and A. Shakya, “A Customer Churn Prediction Model using XGBoost for the Telecommunication Industry in Nepal,” *Procedia Comput. Sci.*, vol. 215, pp. 652–661, 2022, doi: 10.1016/j.procs.2022.12.067.

[17] A. Amin, A. Adnan, and S. Anwar, “An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes,” *Appl. Soft Comput.*, vol. 137, p. 110103, 2023, doi: 10.1016/j.asoc.2023.110103.

[18] Orange, “Telecom Churn Dataset | Kaggle.” [Online]. Available: https://www.kaggle.com/mnassrib/telecom-churn-datasets.

[19] Purnima, T., & Rao, C. K. . (2023). CROD: Context Aware Role based Offensive Detection using NLP/ DL Approaches. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1), 01–11. https://doi.org/10.17762/ijritcc.v11i1.5981

[20] Brown, R., Brown, J., Rodriguez, C., Garcia, J., & Herrera, J. Predictive Analytics for Effective Resource Allocation in Engineering Education. Kuwait Journal of Machine Learning, 1(1). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/91