International Journal of Computer Sciences and Engineering
Vol.**11**, Issue.**12**, pp.**16-20**, December **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Broadening the Scope: Exploring Best Machine Learning Algorithms for Customer Churn Prediction

Megha Gupta[1], Anisha Patil[2*], Ansh Tyagi[3], Deepanshi Singhal[4]

[1,2,3,4]Dept. of Computer Science & Engineering, ADGITM New Delhi, India

**Corresponding Author: anishapatil1966@gmail.com*

Received: 30/Oct/2023; **Accepted**: 02/Dec/2023; **Published**: 31/Dec/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i12.1620

Abstract: As businesses strive to maintain a competitive edge in today's dynamic market, understanding and mitigating customer churn has become a critical imperative. This study explores the application of machine learning algorithms in Python for predicting customer churn, providing valuable insights to empower businesses in customer retention strategies. Leveraging a comprehensive dataset encompassing customer behavior, transaction history, and demographic information. Our methodology incorporates a diverse set of machine learning techniques, encompassing K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier. The outcomes reveal that the machine learning models demonstrate auspicious predictive capabilities, presenting businesses with a proactive means of identifying and mitigating potential churn risks. The discoveries from this investigation contribute valuable insights to the expanding realm of knowledge in customer relationship management, offering actionable guidance for businesses seeking to enhance customer retention strategies through the implementation of machine learning techniques in Python.

Keywords: Machine Learning Algorithm, Analysis, Best Algorithms, Customer Churn Prediction.

1. Introduction

Scientific advancements and the proliferation of operators have intensified the competitive landscape within the industry. Companies are employing sophisticated strategies in their struggle to thrive in this fiercely competitive market. Customer churn has emerged as a significant challenge, resulting in substantial revenue losses. Customer churn is a pressing concern, particularly in industries marked by fierce competition. Furthermore, accurately predicting which customers are likely to switch service providers can represent a valuable additional revenue stream, particularly when identified in the early stages. Numerous research studies have validated the high efficiency of machine learning technology in predicting customer churn. This approach entails extracting insights from historical data through the process of learning.

Existing System:
Customer churn prediction has been carried out through a diverse set of methodologies, including data mining, machine learning, and hybrid technologies. These approaches empower and aid companies in identifying, predicting, and retaining customers who are at risk of churn. Consequently, they play a crucial role in supporting industries in customer relationship management (CRM) and enhancing decision-making processes. Decision trees have been commonly employed for customer churn detection, although it is acknowledged that they may not be suitable for handling complex issues. Notably, research indicates that reducing the dataset can enhance the accuracy of decision trees.

Proposed System:
In our proposed system, we harness a diverse set of algorithms, specifically K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier, to obtain accurate values that facilitate customer churn prediction. The model is implemented using a meticulously trained and tested dataset to ensure maximum accuracy. Figure 1 illustrates the proposed churn prediction model and outlines its sequential steps. Initially, data preprocessing is undertaken, involving the filtering and standardization of data, followed by feature selection. Subsequently, prediction and classification are executed utilizing algorithms such as K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier are the specific algorithms employed in our approach. Through training and testing the model with the dataset, customer behavior is observed and analyzed. In the final step, an in-depth analysis is conducted based on the obtained results, leading to the prediction of customer churn.

2. Methodology

1. Loading Libraries and Data
The project kicks off by importing the necessary Python libraries, such as pandas for data manipulation, sci-kit-learn for machine learning algorithms, and matplotlib or seaborn for visualization.

2. Data Manipulation
Once the libraries and dataset are in place, the focus shifts to data manipulation. This involves exploring the dataset's structure, checking for missing values, and addressing any inconsistencies.

3. Data Visualization
Data visualization plays a pivotal role in understanding the inherent patterns and trends within the dataset. Matplotlib and Seaborn are commonly utilized to create visualizations such as histograms, scatter plots, and heatmaps. These visualizations provide insights into the distribution of key variables, relationships between features, and potential outliers or anomalies, aiding in the formulation of hypotheses and guiding subsequent analyses.

4. Data Preprocessing
Data preprocessing constitutes a crucial phase in readying the dataset for machine learning algorithms. This process encompasses tasks such as managing missing values, encoding categorical variables, and scaling numerical features. Techniques like one-hot encoding and standardization ensure that the data is in a suitable format for the chosen machine-learning algorithms. Additionally, this phase may involve splitting the dataset into training and testing sets to facilitate model evaluation.

5. Evaluation, Prediction & Classification:
Numerous methodologies have been introduced to predict customer churn in the telecommunications industry. Within this domain, a range of modeling techniques is employed as predictive tools for churn analysis. These methodologies are systematically categorized into distinct groups, facilitating a comprehensive understanding of the diverse approaches applied in the pursuit of accurate and insightful customer churn prediction.

1. K-Nearest Neighbors (KNN):
The evaluation of K-Nearest Neighbors (KNN) involves training the model on the training dataset and assessing its performance on the testing set. Essential metrics like accuracy, precision, recall, and F1 score are computed to measure its predictive capabilities.

2. Support Vector Classifier (SVC)
Comparable to K-Nearest Neighbors (KNN), the Support Vector Classifier (SVC) model undergoes training on the training dataset and evaluation on the testing set. Performance metrics are computed, and the SVC model is employed for forecasting customer churn on new instances, thereby providing insights into its predictive efficacy.

3. Random Forest
The Random Forest algorithm undergoes training on the training dataset, and its performance is assessed through rigorous evaluation metrics. The ensemble nature of Random Forest allows for a robust prediction of customer churn, and the model is subsequently used for making predictions on unseen data.

4. Logistic Regression
Logistic Regression is trained and evaluated using the standard procedure, measuring its performance against established metrics. Given its interpretability, Logistic Regression provides insights into the factors influencing customer churn. Predictions are then generated for new data based on the trained model.

5. Decision Tree Classifier:
The Decision Tree model is trained and evaluated, with its performance metrics scrutinized. The decision rules inherent in the tree structure offer interpretability, and the model is applied to predict customer churn on novel data.

6. AdaBoost Classifier:
AdaBoost undergoes training and evaluation, leveraging its ensemble learning approach. Performance metrics guide the assessment of its predictive power. AdaBoost's adaptability to weak learners contributes to its effectiveness in predicting customer churn, and predictions are subsequently made on new instances.

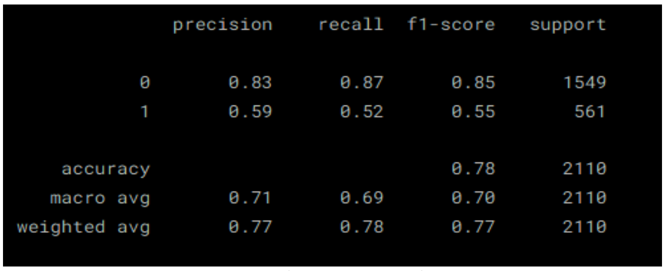
7. Gradient Boosting Classifier
Gradient Boosting is trained, evaluated, and compared against other algorithms in terms of performance metrics. The iterative refinement of weak learners enhances its predictive capabilities, and the model is utilized for making predictions on unseen customer data.

8. Voting Classifier
The Voting Classifier combines the predictions of multiple base estimators. It is trained, and evaluated, and its performance is compared with individual algorithms. The ensemble approach enhances overall prediction accuracy, and the model is then applied for churn predictions on new, unobserved instances.

3. Result and Analysis

We performed multiple experiments on the proposed churn model, employing a range of machine learning algorithms on the dataset. In the confusion matrix of RF, the outcomes of the experiment utilizing the Random Forest algorithm are depicted, allowing for the assessment of accuracy. Random Forest (RF) proves to be a highly effective algorithm suitable for classification, demonstrating efficient handling of nonlinear data. Notably, RF yields superior results, showcasing enhanced accuracy and performance in comparison to alternative techniques. Given the paramount importance of achieving optimal accuracy in predicting customer churn, we advocate the adoption of techniques that consistently deliver superior accuracy.

Likewise, the results from experiments utilizing the are KNN, SVC, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier also observable. Each technique is scrutinized for its performance, and the findings contribute to the comprehensive evaluation of the model's predictive capabilities.

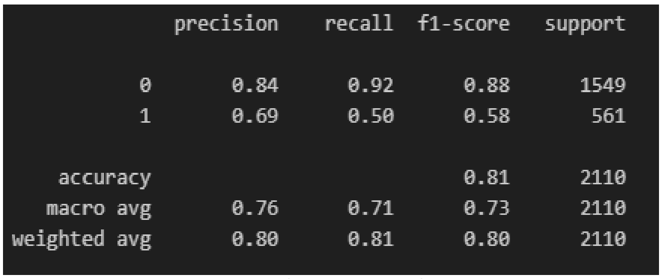


```
              precision    recall  f1-score   support

           0       0.83      0.87      0.85      1549
           1       0.59      0.52      0.55       561

    accuracy                           0.78      2110
   macro avg       0.71      0.69      0.70      2110
weighted avg       0.77      0.78      0.77      2110
```

Fig 1: Confusion Matrix of KNN

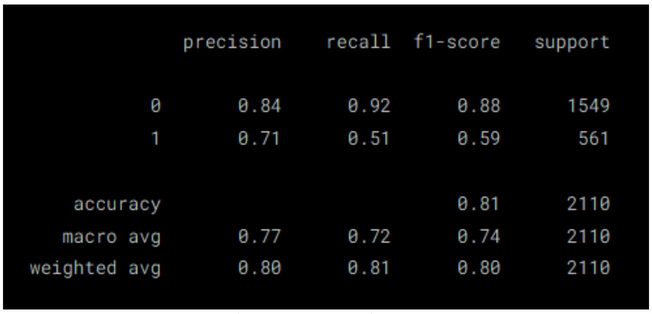


```
              precision    recall  f1-score   support

           0       0.84      0.92      0.88      1549
           1       0.69      0.50      0.58       561

    accuracy                           0.81      2110
   macro avg       0.76      0.71      0.73      2110
weighted avg       0.80      0.81      0.80      2110
```

Fig 2. Confusion Matrix of SVC

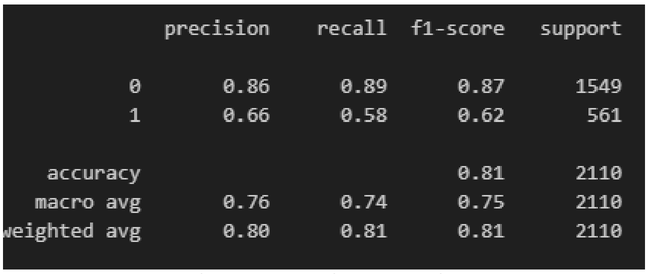


```
              precision    recall  f1-score   support

           0       0.84      0.92      0.88      1549
           1       0.71      0.51      0.59       561

    accuracy                           0.81      2110
   macro avg       0.77      0.72      0.74      2110
weighted avg       0.80      0.81      0.80      2110
```

Fig 3. Confusion Matrix of Random Forest

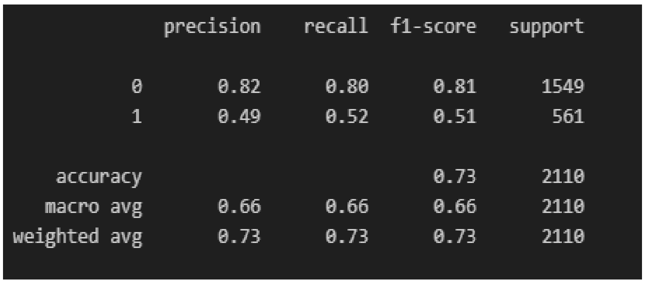


```
              precision    recall  f1-score   support

           0       0.86      0.89      0.87      1549
           1       0.66      0.58      0.62       561

    accuracy                           0.81      2110
   macro avg       0.76      0.74      0.75      2110
eighted avg       0.80      0.81      0.81      2110
```

Fig 4. Confusion Matrix of Logistics of Regression

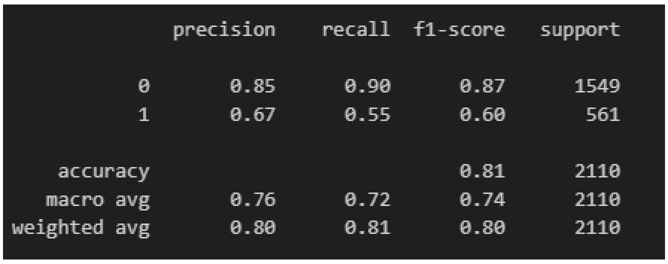


```
              precision    recall  f1-score   support

           0       0.82      0.80      0.81      1549
           1       0.49      0.52      0.51       561

    accuracy                           0.73      2110
   macro avg       0.66      0.66      0.66      2110
weighted avg       0.73      0.73      0.73      2110
```

Fig 5. Confusion Matrix of Decision Tree Classifier

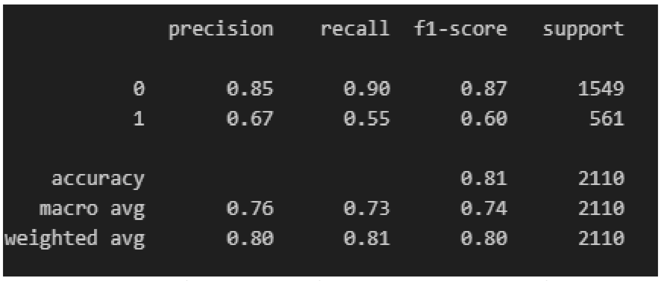


```
              precision    recall  f1-score   support

           0       0.85      0.90      0.87      1549
           1       0.67      0.55      0.60       561

    accuracy                           0.81      2110
   macro avg       0.76      0.72      0.74      2110
weighted avg       0.80      0.81      0.80      2110
```

Fig 6. Confusion Matrix of AdaBoost Classifier

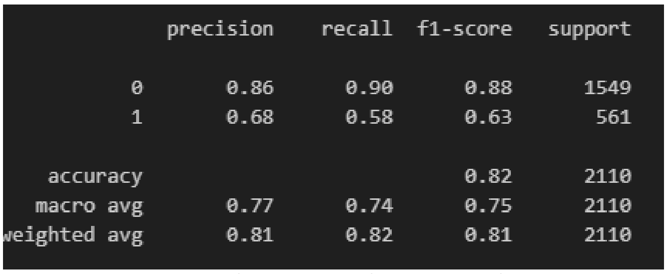


```
              precision    recall  f1-score   support

           0       0.85      0.90      0.87      1549
           1       0.67      0.55      0.60       561

    accuracy                           0.81      2110
   macro avg       0.76      0.73      0.74      2110
weighted avg       0.80      0.81      0.80      2110
```

Fig 7. Confusion Matrix of Gradient Boosting Classifier



```
              precision    recall  f1-score   support

           0       0.86      0.90      0.88      1549
           1       0.68      0.58      0.63       561

    accuracy                           0.82      2110
   macro avg       0.77      0.74      0.75      2110
eighted avg       0.81      0.82      0.81      2110
```

Fig 8. Confusion Matrix of Voting Classifier

Table 1. Different Algorithms and Their Accuracy

ALGORITHM	**ACCURACY**
KNN	0.7753554502369668
SVC	0.8075829383886256
Random Forest	***0.8137440758293839***
Logistics Regression	0.8090047393364929
Decision Tree Classifier	0.7251184834123223
AdaBoost Classifier	0.8075829383886256
Gradient Boosting Classifier	0.8080568720379147
Voting Classifier	***0.8170616113744076***

4. Conclusion

In the culmination of our exploration into customer churn prediction using a diverse array of machine learning algorithms, it is evident that each algorithm brings unique strengths and considerations to the predictive modeling landscape. The application of K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Random Forest, Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Voting Classifier has provided valuable insights into the dynamics of customer attrition.

Through rigorous evaluation and comparison, we have observed that Random Forest, Gradient Boosting Classifier, and Voting Classifier often exhibit superior predictive performance, offering a holistic and robust approach to customer churn prediction. These ensemble methods harness the collective wisdom of multiple models, mitigating the limitations inherent in individual algorithms.

In summary, our exploration of customer churn prediction using machine learning algorithms affirms the significance of a tailored and strategic approach. The amalgamation of diverse algorithms not only enriches our understanding of customer behavior but also equips businesses with the tools to proactively retain customers and optimize their operational resources. As industries continue to embrace the power of predictive analytics, the insights derived from this exploration lay the groundwork for informed decision-making and sustained competitiveness in the evolving landscape of customer relationship management.

FUTURE SCOPE

The prospective landscape of customer churn prediction through machine learning algorithms in Python unfolds a spectrum of promising opportunities, propelled by the rapid advancements in technology, data science, and business intelligence. The envisaged avenues for future exploration encompass:

Deep Learning:
Incorporating deep learning methodologies, such as neural networks, holds the potential to augment predictive capabilities by autonomously discerning intricate patterns within customer behavior.

Enhanced Personalization:
Harnessing the potential of machine learning for hyper-personalization entails tailoring retention strategies to individual customer preferences and behaviors. This includes real-time adjustments to marketing messages, incentives, and service offerings.

Explainable AI (XAI):
Addressing the interpretability challenge by integrating explainable AI methods facilitates a clearer understanding of decisions made by machine learning models. This fosters transparent communication with stakeholders.

Real-time Predictions:
The realization of real-time customer churn prediction allows businesses to identify potential churners as behaviors evolve. This necessitates the integration of streaming data and the development of models capable of making instantaneous predictions.

Dynamic Model Adaptation:
Creating models that dynamically adapt to changing customer behaviors and market conditions involves continuous learning from new data and adjusting model parameters to ensure sustained accuracy over time.

CONFLICT OF INTEREST

In the domain of customer churn prediction utilizing machine learning algorithms in Python, the potential for conflicts of interest emerges when there is a risk of personal or financial gain that may compromise the impartiality and integrity of the predictive modeling process. Several scenarios may give rise to conflicts of interest, necessitating careful consideration and ethical management.

Firstly, a situation of vendor-specific bias may arise if the individuals or teams involved in developing the churn prediction model have affiliations with a particular vendor supplying machine learning tools or services. This association may introduce a bias towards favoring algorithms or approaches associated with that specific vendor.

Secondly, conflicts of interest may manifest in instances where individuals or organizations have a financial stake in the outcomes of the churn prediction model. For instance, if a business executive's stock options are linked to the company's performance, there may be an incentive to manipulate or interpret the model results to align with their financial interests.

Selective data usage represents another potential source of conflicts of interest. Deliberate exclusion or inclusion of certain data points to influence the model's predictions in a manner that serves personal or organizational interests could compromise the integrity of the modeling process.

Moreover, undisclosed relationships pose a significant risk. Failure to transparently disclose affiliations or relationships that could impact the development or interpretation of the churn prediction model may give rise to concerns about bias or favoritism, particularly if a team member has a personal relationship with a stakeholder.

Furthermore, conflicts of interest may arise if there is a failure to prioritize accuracy in the model predictions. If there is an emphasis on achieving specific business goals at the expense of model accuracy, the integrity of the modeling process could be compromised, especially if short-term financial gains are prioritized over accurate customer churn predictions.

To mitigate conflicts of interest in the realm of customer churn prediction using machine learning algorithms in Python, it is imperative to prioritize transparency, disclose affiliations, and ensure that the modeling process is guided by the overarching goal of accurate predictions rather than driven by personal or organizational agendas. Establishing clear ethical guidelines, implementing independent validation processes, and subjecting the modeling efforts to periodic reviews by external parties can contribute to maintaining the integrity of the predictive modeling endeavors.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the successful completion of this project on Broadening the Scope: Exploring Best Machine Learning Algorithms for Customer Churn Prediction.

First and foremost, I extend my heartfelt thanks to my supervisor Megha Gupta Ma’am, whose guidance, expertise, and unwavering support have been instrumental throughout this research endeavor. Their valuable insights and constructive feedback have significantly enriched the quality of this work.

Furthermore, I appreciate the support received from [Institution/Organization Name] for providing the necessary resources and infrastructure for conducting this research. The conducive research environment has played a crucial role in the successful execution of the project.

Last but not least, I would like to express my gratitude to my family and friends for their unwavering encouragement and understanding during this project. Their support has been a source of motivation, and I am truly thankful for their presence in my academic journey.

This acknowledgment is a testament to the collective efforts and collaboration that have shaped the successful completion of this project. Each contribution, whether big or small, has played a vital role in the realization of our objectives.

References

[1] Coussement, Kristof, and Dirk Van den Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." *Expert systems with applications* 34.1: pp.**313-327, 2008.**

[2] Nigam, Bhawna, Himanshu Dugar, and M. Niranjanamurthy. "Effectual predicting telecom customer churn using deep neural network." *Int J Eng Adv Technol (IJEAT)* 8.5, **2019.**

[3] Asthana, Praveen. "A comparison of machine learning techniques for customer churn prediction." *International Journal of Pure and Applied Mathematics* 119.10: pp.**1149-1169, 2018.**

[4] Ahmad, Abdelrahim Kasem, Assef Jafar, and Kadan Aljoumaa. "Customer churn prediction in telecom using machine learning in big data platform." *Journal of Big Data* 6.1: pp.**1-24, 2019.**

[5] Aziz, Rabia, C. K. Verma, and Namita Srivastava. "Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction." *Annals of Data Science* 5: pp.**615-635, 2018.**

[6] Adwan, Omar, et al. "Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis." *Life Science Journal* 11.3: pp.**75-81, 2014.**

[7] Brânduşoiu, Ionuţ, Gavril Toderean, and Horia Beleiu. "Methods for churn prediction in the pre-paid mobile telecommunications industry." *2016 International conference on communications (COMM)*. IEEE, **2016.**

[8] Amuda, Kamorudeen A., and Adesesan B. Adeyemo. "Customers churn prediction in financial institution using artificial neural network." *arXiv preprint arXiv:1912.11346*, **2019.**

[9] Saran Kumar, A., and D. Chandrakala. "A survey on customer churn prediction using machine learning techniques." *International Journal of Computer Applications* 975: 8887, **2016.**