Churn Forecast Portal using Random Forest Classifier

Arpan Chakraborty[1], Manjula Sanjay Koti[2*]

[1,2]Dayananda Sagar Academy of Technology and Management, Karnataka, India

***Email:** hodmca@dsatm.edu.in

Abstract

The competitive scene inside the telecom and keeping cash businesses demands compelling client upkeep strategies. This request almost centres on making a energetic Client Churn Figure system utilizing machine learning strategies, especially the Subjective Forest Classifier, to recognize at-risk clients proactively. By analysing client data, tallying socioeconomics, advantage utilization plans, and charging information, the system predicts the likelihood of churn. The encounters picked up coordinate companies in actualizing centred on trade to make strides client steadfastness. The made system is affirmed utilizing datasets from the telecom and overseeing an account division, outlining tall precision and unflinching quality in churn figure.

Keywords

Customer Retention Strategies, Churn Prediction System, Machine Learning Techniques, Random Forest Classifier, Telecom and Banking Datasets

Introduction

Inside the present-day telecom and overseeing an account commerce, client upkeep has finished up crucial due to the tall taken a toll of getting unused clients compared to keeping up existing ones. Client churn, where clients cease their relationship with a advantage provider, stances a essential threat to commerce efficiency and viability. Understanding the factors driving to churn and tending to them proactively is imperative for companies indicating to diminish whittling down and make strides client constancy. This request approximately centres on making a Client Churn Desire system utilizing machine learning techniques to expect the likelihood of a client suspending their advantage and coordinate companies in executing centred on upkeep.

Many studies have investigated customer churn prediction using various machine learning algorithms. Traditional methods based on historical trend analysis or customer surveys have significant drawbacks, such as limited accuracy and reactive approaches. Recent advances in machine learning have introduced more advanced techniques such as neural networks and ensemble methods, which have improved accuracy and insight into the importance of features. In

Submission: 14 October 2024; **Acceptance:** 3 November 2024

this section, we review the relevant literature on customer churn prediction and highlight the advantages and limitations of different approaches.

The use of neural networks has shown promising results in predicting customer churn in the telecommunications sector. (Adwan et al., 2014) demonstrated the effectiveness of multilayer perceptron neural networks for this purpose, illustrating its capability to model complex patterns and trends in customer behavior. In a different study, (Nath & Behara, 2003) applied data mining techniques to analyze customer churn in the wireless industry. This approach identified significant factors influencing customer churn and provided insights into customer retention strategies (Nath & Behara, 2003.).

(Lalwani and Lalwani, 2021) presented a comprehensive machine learning framework for customer churn prediction across various sectors, including telecommunications and banking. Their work emphasized the adaptability of machine learning algorithms in addressing churn issues in different domains (Lalwani & Lalwani, 2021). Comparative studies have also been conducted to evaluate the performance of different machine learning techniques. (Vafeiadis et al., 2015) compared several algorithms for predicting customer churn, highlighting the advantages of ensemble methods such as Random Forest in terms of accuracy and performance (Vafeiadis et al., 2015).

In their paper, (Kumar and Chandrakala, 2016) present a detailed review of various machine learning techniques for predicting customer churn across multiple sectors, emphasizing the importance of retaining customers for business sustainability. They evaluate methods like neural networks, ensemble classifiers, and boosting algorithms, discussing their effectiveness in identifying customers likely to leave a service provider. Customer churn prediction has been explored using various data mining techniques, but existing algorithms face limitations due to imbalanced data, noise, and the need for customer ranking (Zhao et al., 2005), (Au et al., 2003). This study proposes an improved balanced random forests (IBRF) method that combines sampling techniques with cost-sensitive learning to address these challenges and achieve better performance in churn prediction (Chen et al., 2004).

On another work, (Ebrah and Elnasir, 2019) explored the use of Naïve Bayes, SVM, and Decision Tree algorithms for churn prediction in the telecom sector, achieving high accuracy on IBM Watson and cell2cell datasets. Their study demonstrated that SVM outperformed other models, with AUC values reaching up to 0.99 on the cell2cell dataset, highlighting the model's predictive strength in identifying customer churn. (Xie et al., 2009) proposed an improved balanced random forests (IBRF) algorithm that addresses data imbalance in customer churn prediction. By integrating sampling techniques with cost-sensitive learning, the model demonstrated superior accuracy compared to traditional algorithms like decision trees and artificial neural networks (Xie et al., 2009).

Besides that, (Rahman and Kumar, 2020) investigated various machine learning classifiers, such as KNN, SVM, Decision Tree, and Random Forest, for predicting customer churn in the banking sector. Their results demonstrated that Random Forest, particularly after oversampling, achieved the highest accuracy, making it the most effective model for churn prediction (Rahman and Kumar, 2020).

The paper discusses customer churn prediction in the retail banking sector and telecom sector, emphasizing the significance of identifying customers at risk of switching to competitors. Through logistic regression and customer value analysis, the study demonstrates how predictive models can help in creating targeted marketing strategies for customer retention, showcasing the effectiveness of conventional statistical methods in addressing churn.

Methodology

There are a few steps within the preparation of making a solid Client Churn Expectation framework: information collection, preprocessing, include designing, preparing the show, and arrangement. To ensure the expectation model's precision and constancy, each step is basic. A broad summation of the methodology's steps is given in this segment.

For the information securing and preprocessing, any prescient demonstrate depends on high-quality information as its establishment. Datasets from the telecom and keeping money segments were utilized for this ponder, and they included socioeconomics, benefit utilization designs, charging data, and client benefit interactions. The introductory arrange within the information arrangement stage was to handle lost values by either ascribing them or killing the influenced records to ensure a clean dataset. Categorical factors were then encoded utilizing approaches like one-hot encoding to create them worthy for machine learning models. This approach created a full dataset appropriate for consider.

To include building, highlight designing may be a crucial arrange that incorporates choosing and altering factors to move forward the model's expectation potential. Account length, add up to day minutes, add up to nighttime calls, and recurrence of client bolster calls were found to be noteworthy indicators of client turnover in this consider. To capture complicated client behaviors, unused features were created by consolidating existing factors. For case, interaction terms between unmistakable utilize measurements and statistic characteristics were made. These fabricated highlights were at that point evaluated for pertinence utilizing approaches such as the Irregular Woodland Classifier's highlight significance scores, guaranteeing that as it were the foremost important characteristics were included within the wrapped-up show.

Next to demonstrate preparing, the arbitrary timberland classifier are chosen for its strength and precision, was at the heart of the methodology's preparation. The dataset was partitioned into preparing and testing sets using 80-20 parts, guaranteeing that the demonstration was prepared on a huge rate of the information whereas keeping a isolated set for appraisal. The preparing strategy included tweaking hyperparameters such as the number of trees within the woodland and each tree's greatest profundity. Cross-validation strategies were utilized to maintain a strategic distance from overfitting and ensure that the model is generalizable. The prepared model's execution was surveyed utilizing measures such as exactness, exactness, review, and F1-score, with the objective of striking an adjust between these measurements to supply precise forecasts.

To demonstrate assessment, assessing the forecast model's execution is basic to deciding its convenience in real-world circumstances. The Arbitrary Timberland Classifier was assessed

utilizing the testing dataset, and a few execution pointers were computed. Precision gave a wide degree of precision, while exactness and review provided data approximately the model's capacity in recognizing genuine positive churn circumstances whereas dodging untrue positives. The F1-score, which could be a consonant cruel of precision and review, was exceptionally successful in adjusting these two measures. The model's execution was too compared to pattern models and conventional approaches to illustrate its predominance in anticipating client attrition.

For sending, Jar was utilized to form a web-based application that made the expectation show more open and client wonderful. The apparatus empowers telecom and budgetary companies to enter client data by means of a basic interface and get real-time churn projections. The backend server rationale forms the approaching information, applies the learned Irregular Woodland Classifier, and produces forecasts. The program is expecting to be energetic and adaptable, with the adaptability to overhaul the demonstrate as unused information gets to be accessible. The arrangement step too includes interfacing the application with a SQLite database to oversee client information successfully and guarantee smooth working.

To proceeds enhancement, the method incorporates continuous audit and refinement of the forecast show. The demonstrate will be overhauled on a customary premise with new information to adjust to changing client behaviors and advertise circumstances. Input from telecom and keeping money administrators who utilize the app will be collected to propose regions for change. Furthermore, exploring distinctive machine learning algorithms and outfit strategies is portion of the continual advancement handle to extend the model's exactness and steadfastness.

Using this exhaustive technique, the venture trusts to supply a vigorous and versatile client churn prediction system that gives significant data for telecom and keeping money organizations, permitting them to proactively hold at-risk clients and progress in general client dependability

Results and Discussion

The Irregular Timberland Classifier was prepared on pre-processed datasets from the telecom and managing an account business, yielding extraordinary execution comes about. The model's precision, exactness, review, and F1-score were calculated to decide its convenience in estimating client turnover. The discoveries are displayed within the table 1 underneath:

Table1: Discoveries found in both the Datasets

Metric	Telecom Dataset	Banking Dataset
Accuracy	92.5%	89.8%
Precision	90.2%	88.5%
Recall	87.4%	85.9%
F1-score	88.8%	87.2%

These measurements appear that the Arbitrary Timberland Classifier performs well in both divisions, with especially tall exactness and F1-score, demonstrating its Vigor and steadfastness in foreseeing client churn.

Significance of Highlights: Highlight noteworthiness appraisals revealed which characteristics were the foremost critical in predicting client whittling down. The foremost imperative characteristics within the telecom dataset were add up to day minutes, client benefit call recurrence, and account length. Within the keeping money dataset, credit score, adjust, and residency were among the most grounded indicators. The include centrality scores are outlined within the taking after figures:

- Telecom Dataset Include Significance:
  - Add up to Day Minutes
  - Client Benefit Call Recurrence
  - Account Length
  - Add up to Night Minutes
  - Add up to Evening Minutes

- Banking Dataset Include Significance:
  - Credit Score
  - Adjust
  - Residency
  - Age
  - Number of Items

These bits of knowledge are basic for telecom and managing an account organization in deciding which factors contribute most significantly to client turnover and fitting their maintenance procedure suitably. Comparison with Conventional Strategies: Conventional approaches for evaluating client steady loss, such as historical slant investigation and client studies, were too utilized for comparison. In comparison to the Arbitrary Timberland Classifier, these procedures created much lower exactness and expanded untrue positive rates. The table 2 emphasizes the aberrations in execution.

Table 2: Result comparison after implementation of the algorithm

Metric	**Random Forest (Telecom)**	**Traditional Methods (Telecom)**	**Random Forest (Banking)**	**Traditional Methods (Banking)**
Accuracy	92.5%	75.4%	89.8%	70.6%
False Positive Rate	3.2%	15.8%	4.1%	18.5%

The Arbitrary Woodland Classifier beat past approaches by a expansive edge, illustrating the benefits of utilizing machine learning strategies for anticipating client whittling down.

For the real-world application, the built prescient show was executed as a web-based application utilizing Carafe, giving telecom and monetary administrators an easy-to-use interface. Clients may enter client data and get real-time churn projections. Administrators have given favourable input on the model's exactness and down to earth bits of knowledge. The capacity to

figure whittling down proactively permits businesses to embrace focused on mediations, bringing down churn rates and upgrading client maintenance.

The case think about, a few cases ponder were carried out to illustrate the model's adequacy in real-world circumstances. In one telecom commerce, the calculation accurately anticipated 85% of clients who were likely to churn within another three months. Based on these estimates, the organization propelled centred maintenance measures, coming about in a 20% drop in attrition over time. Within the managing an account industry, a case considers with a mid-sized bank appeared that the demonstrate may legitimately figure turnover for high-value clients. The bank utilized this data to supply focused on administrations and motivating forces, coming about in a 15% decrease in turnover inside this statistic.

Some restrictions and challenges, whereas the comes about are empowering, the inquire about did go up against a few limitations and issues. One limitation is the dependence on the quality and completeness of the provided information. Information that's lost or off-base might have an effect on the model's forecasts. Moreover, the model's execution may vary among areas and client socioeconomics, requiring visit retraining with upgraded datasets.

Another impediment was interfacing the approach with current client relationship administration apparatuses. Guaranteeing consistent information stream and interoperability with different CRM stages requires a significant sum of work. In any case, amid the arrangement stage, these issues were overcome utilizing thorough information pretreatment methods and cautious testing.

For future changes, a number of improvements are arranged to move forward the forecast show. Incorporating extra information sources, such as social media intuitive and buyer input, might grant a more total picture of client behaviour. Progressed machine learning strategies, such as profound learning and fortification learning, may also offer assistance to extend the model's exactness and adaptability.

Besides, real-time information preparing capabilities will allow the show to convey up-to-date estimates, moving forward its value to telecom and monetary companies. Persistent observing and client input will lead nonstop changes, guaranteeing that the show remains fruitful beneath energetic advertise circumstances.

Conclusion

The section shows that the Random Forest classifier can successfully predict customer attrition in the telecommunications and banking sectors. The model's high accuracy and insights into key predictive factors make it a powerful tool for companies looking to improve their customer retention efforts. The successful deployment and positive comments from early users highlight the model's practical applicability and potential real-world impact. Future improvements will focus on expanding data sources, exploring advanced algorithms, and improving real-time capabilities to maintain and improve the model's performance. Additionally, ongoing monitoring and feedback loops will be implemented to ensure the model remains effective and up-to-date in predicting

customer attrition. Regular updates and refinements will be made based on new data and evolving customer behaviors to enhance the model's accuracy and relevance over time.

Acknowledgement

The researcher did not receive any funding for this study, and the results have not been published in any other sources.

References

Adwan, O., Faris, H., Jaradat, K., Harfoushi, O., & Ghatasheh, N. (2014). Predicting customer churn in telecom industry using multilayer preceptron neural networks: Modeling and analysis. Life Science Journal, 11(3), 75-81. https://www.lifesciencesite.com/lsj/life1103/011_22840life110314_75_81.pdf

Ebrah, K., & Elnasir, S. (2019). Churn prediction using machine learning and recommendations plans for telecoms. Journal of Computer and Communications, 7(11), 33-53. https://doi.org/10.4236/jcc.2019.711003

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing, 104(2), 271-294. https://doi.org/10.1007/s00607-021-00908-y

Mutanen, T., Nousiainen, S., & Ahola, J. (2010). Customer churn prediction–a case study in retail banking. In Data mining for business applications pp. 77-83. IOS Press. https://doi.org/10.3233/978-1-60750-633-1-77

Nath, S. V., & Behara, R. S. (2003). Customer churn analysis in the wireless industry: A data mining approach. In Proceedings-annual meeting of the decision sciences institute v561, pp. 505-510. https://download.oracle.com/owsf_2003/40332.pdf

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. https://doi.org/10.1109/ICECA49313.2020.9297529

Saran Kumar, A., & Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. International Journal of Computer Applications, 975, 8887. http://dx.doi.org/10.5120/ijca2016912237

Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, 1-9. https://doi.org/10.1016/j.simpat.2015.03.003

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449. https://doi.org/10.1016/j.eswa.2008.06.121