International Journal for Research Publication and Seminar
ISSN: 2278-6848 | Vol. 16 | Issue 1 | Jan-Mar 2025 | Peer Reviewed & Refereed Refereed
Special Edition : SPARK 2025 : XXI National Conference on Emerging Technology Trends in Engineering & Project Competition

Churn Prediction: A Predictive Modeling Approach

Prof. K. K. Ingole, Nikita Chindhalore, Vedanti Bele, Kartik Girde, Vinayak Chaudhari
Department of Artificial Intelligence and Data Science
K. D. K. College of Engineering Nagpur, India
kartik.ingole@kdkce.edu.in, nikitachindhalore23@gmail.com, vedantibele20@gmail.com,
vinaychaudhari060@gmail.com, girdekartik2002@gmail.com

ABSTRACT: Customer churn, the phenomenon of customers discontinuing their service or subscription, presents a significant challenge for businesses across various industries. Accurately predicting churn is crucial for proactive retention strategies and maximizing customer lifetime value. This research paper explores various predictive modeling techniques for churn prediction, including Logistic Regression, Support Vector Machines, Random Forest, and Neural Networks. We evaluate their performance on a publicly available telecom churn dataset, comparing metrics such as accuracy, precision, recall, F1- score, and AUC-ROC. The study aims to identify the most effective model for churn prediction and discuss the key factors contributing to customer churn, offering actionable insights for businesses.

KEYWORD: Churn Production, Churn Modeling, Productive Modeling, Predictive Modeling, Customer Churn, Machine Learning, Statistical Modeling.

I. INTRODUCTION

In today's competitive market, customer retention is paramount for sustainable business growth. Acquiring new customers is often significantly more expensive than retaining existing ones, making churn a costly problem. Churn prediction empowers businesses to identify at-risk customers and implement targeted interventions to prevent them from leaving. This paper investigates the application of machine learning algorithms to predict customer churn, providing a valuable tool for businesses to proactively address customer dissatisfaction, improve retention rates, and optimize resource allocation for customer relationship management.

This section presents the results of the experiments, comparing the performance of the different models based on the evaluation metrics. We will provide a detailed analysis of the results, including: Comparison of model performance: Which model performed best based on the chosen metrics? Are there trade-offs between different metrics (e.g., precision vs. recall)? Feature importance analysis: Which features are most influential in predicting churn? This can provide valuable insights for businesses to focus their retention efforts. We will use feature importance scores from models like Random Forest and permutation importance for other models. Error analysis: Examining the confusion matrices to understand the types of errors made by each model. Are certain types of customers more likely to be misclassified? Discussion of limitations: Acknowledging any limitations of the study, such as data limitations, model assumptions, or computational constraints.

Churn means the number of customers a company loses in a set time period. CC has a significant impact on the service sector with tough competition. Spotting customers who might leave can bring in a lot of money. The CC dataset helps to study customer marketing trends from big databases. We can think of customer attrition as a churn rate, which shows the percentage of users who stop using a service within a specific timeframe.

In mainland China, the government gave three mobile companies the go-ahead to run wireless communication services. The yearly growth rate of new mobile users slowed down a lot. It dropped from over 10% between 2009 and 2013 to 4.7% in 2014. Like in many other countries, China's mobile communication industry is getting close to its limit and becoming more Companies with subscription-based models often use this metric to check their financial health. These businesses have customers under contract. In developed countries, telecommunications is a big industry. New tech and more managers make the field more competitive. We'll test how a machine learning (ML) model affects churn using a new dataset. Our CC prediction model offers these benefits:

1. It uses different ways to clean up data, including SMOTE-ENN, to make it more normal.
2. It tries various ways to sort data to find the best model for predictions.

II. LITERATURE REVIEW

Farquad, H. &Vadlamani, Ravi & Surampudi, Bapi (2024) [1] proposed a mongrel approach to overcome the downsides of general svm model which generates a box model (i.e., it does n't reveal the knowledge gained during training in mortal accessible form). The mongrel approach contains three phases. In the first phase, SVM- RFE (SVM- recursive point elimination) is employed to reduce the point set. In the alternate phase, reduced features dataset is also used to get SVM model and support vectors are uprooted. In the final phase, rules are also generated using Naive Bayes Tree (NBTree which is combination of Decision tree with naive Bayesian Classifier). The dataset used then's bank credit card client dataset (Business Intelligence Cup
2004) which is largely unstable with 93.24 pious and 6.76 churned guests. The trial showed that the model was not scalable for large datasets. For vaticination, in this paper two different mongrel models have been established, which integrate back- propagation artificial neural networks and tone organizing charts for relating churning possibilities. The mongrel models include ANN ANN and ISOM ANN, which are appertained to as SOM ANN. In that case, one of the specific mongrel models has a data reduction process by performing filtering on data that's regarded as unrepresentative. thus, in the training of the coming stage, the labors which are attained, are further employed Ito achieve the vaticination model grounded on the alternate fashion. For performance evaluation of these models, three different kinds of testing sets, like one general testing set and two fuzzy testing sets grounded on the filtered- out data by the first fashion of the two mongrel models, i.e. ANN and Som independently are used. The result shows mongrel models outperform the single neural network birth model in terms of vaticination delicacy and the stylish set up was ANN ANN compared with SOM ANN.

In 2024, Kumar Dudyala & Ravi. Vadlamani [2] suggested the use of algorithms with Ant-Miner and ALBA a publicly available churn prediction dotaset to develop accurate and explainable models for churn prediction, rule-sets. Ant-Miner is an efficient data mining algorithm for Ant- principle such that the resulting optimization problem can leverage the knowledge by adding monotonicity constraints on the end rule-set. The benefits of Ant-Minert are accuracy, readability of the resulting models, and the option of requesting visually interpretable predictive models. Active Learning Based Approach (ALBA) is an algorithm for the extraction of SVM rules, combining the very high predictive performance of a non-linear SVM model with the easy learnability of the rule-set representation.

The results which are benchmarked to RIPPER, SVM and logistic regression showed that Chih Fong Tsai (2023). ALBA in combination with RIPPER achieves the best accuracy, whilst sensitivity is highest and RIPPER on a oversampled dataset. Ant-Miner+ generates less sensitive rule-sets, but permits

the integration of domain knowledge together with generating rule-sers which are significantly more manageable (much shorter) than those generated using. RIPPER also results in short and understandable sets of rules, however, it produces models that are counter-intuitive and violate domain knowledge.

Ning Lu [3] suggested the application of boosting algorithms to improve a customer churn prediction model in which the customers are partitioned into two classes according to the weight given by the algorithm. B) As a result, a high risky customer cluster has been identified. Logistic regression is employed as a candidate learner, and a churn prediction model is developed within each cluster, respectively. Experimental results indicated that the boosting algorithm offers a sort of separation of the churn data, compared with the single logistic regression model.

Benlan He [4] proposed a customer churn prediction method using SVM model and applied random sampling method to enhance SVM model by taking the imbalance characteristic of customer data sets into account. A support vector machine build an hyper-plane in a high- infinity-dimensional space, which/or can be used for classification, Random sampling method can be employed to adjust the data distribution in practical sense to alleviate the dataset imbalance issue. Training data imbalance stems from the small amount of churners.

Using a finite mixture model for designing the reference value and decision interval of the chart and, at the same time, with a hierarchical Bayesian model for accounting for customer heterogeneity, recency - Yet another time interval variable which complements IAT - is also included in the model for keeping track of recent login status. Apart from that, benefits from the nature of its control charts, the graphical interface of each of the users is one of the advantages of the proposed approach. The findings have indicated that the accuracy rate (ACC) of the gamma CUSUM chart is 5.2% favoring the exponential CUSUM chart. The Average Time to Signal (ATS) of gamma CUSOM is approximately 2 days longer than exponential CUSUM.

Principal Component Analysis(PCA), Independent Component Analysis(ICA) and Sparse Random Projections(SRP) on the classification performance of RotBoost and Rotation Forest.

(i) performance metrics employed to quantify classification performance and (ii) the applied feature extraction algorithm.

Lee et al. [5] focused on constructing predictive an accurate yet concise predic model with the the g goal I of churn prediction by using a Partial Least Squares (PLS) based method on high correlated data sets among variables. They not only present a predictive model to forecast customers churning behavior but also a simple and implementable marketing program for for churning was adapted. With this suggested methodology the marketing managers will be able to do it in an effective and efficient manner! y keep u up to the optimal, at least nearly optimal, churners level using marketing programs. In this model, PLS is employed as a prediction model.

Recently, Koen W. De Bock [6] proposed GAMensplus, an ensemble classifier based on generalized additive models (GAMS) to reconcile not only performance but also to be applied for evaluation purposes and can be used for churn prediction modeling. The proposed GAMens, which is based on Bagging. the Random Subspace i Method, semi-parametric GAMs as constituent classifiers, is extended to include two instruments for model interpretability namely, the generalized feature importance scores, and bootstrap confidence bands for smoothing splines. The experimental comparison of the classification performance over the data sets of six real-life churn prediction projects showed that GAMensplus exhibits a strong classification performance at least as good as that shown by the two individual

classifiers, namely, logistic regression and GAM.

Ning et al. [7] experimentally studied customer churn prediction in the telecommunication sector, and suggested the introduction of boosting to better model churn prediction in the domain of telecommunication. In contrast to other boosting techniques which refine the prediction of an existing basis learner, the authors proposed to divide customers into two subsets using the weight assigned by the boosting algorithm. The "Implementation Zone" that is proposed by the current model enables retention actions for customers, who are at their highest churn potential.

Ver-braken et al. [8] proposed a new performance measure known as the expected maximum profit criterion, which aligns with the primary goals of end users. This framework not only helps companies choose the classifier that maximizes profit but also offers insights into the portion of the customer base to include in retention campaigns.

P.C. Pendharkar [9] introduced two genetic algorithm (GA) based neural network (NN) models for predicting customer churn. The first model utilized cross-entropy criteria for churn prediction, while the second aimed to enhance prediction accuracy directly. They compared these GA-based models against a statistical z-score model using a real customer dataset, evaluating performance through precision, top 10% lift, and the area under the Receiver Operating Characteristic curve. The experimental results indicated that both NN models outperformed the z-score model across all performance metrics.

Y. Xie et al. [10] employed an improved balanced random forest (IBRF) model, which combines balanced and weighted random forests to address data distribution issues. The IBRF model learns the best features iteratively by adjusting class distribution and imposing a higher penalty for misclassifying the minority class. Experiments conducted on a Chinese bank dataset demonstrated that IBRF achieved higher accuracy than artificial neural networks, decision trees, and support vector machines.

III. METHEDOLOGY AND DATASET PREPROCESSING

This study utilizes the publicly available Telecom Churn dataset from Kaggle. This dataset is a popular choice for churn prediction research due to its comprehensive information about customers, including demographics (age, gender), service usage (data plan, voice mail), billing details (monthly charges, total charges), contract type, and churn status (yes/no). The data preprocessing steps are crucial for ensuring data quality and model performance:

Data Cleaning: Handling missing values is critical. We will explore various imputation techniques (mean, median, K-NN) and assess their impact on model performance. We will also identify and handle outliers, which can skew model predictions. Feature Engineering: Creating new features from existing ones can significantly improve model accuracy.

Examples include:

Tenure in months: Converting contract start date to tenure. Charges to
monthly charges ratio-Portrays spending behavior. Average monthly
spend-Reflects customer value.
Service calls-Forums where customer service interactions are logged. Months of
contract duration-A proxy for the commitment of the customer.

One-Hot Encoding: All categorical variables like type of contract/internet service are transformed into numerical representations through one-hot encoding.

Data Transformation: Scaling numeric features using standardization (z-score normalization) or min-max scaling to ensure features with large values do not overwhelm the model and to increase the speed of convergence of algorithms like gradient descent applied in logistic regression and neural networks.

Data Splitting: The dataset is divided into training, validation, and testing sets using stratified sampling to maintain class balance (proportion of churned and non-churned customers) across all sets.

Fig. Customer Churn Prediction

Predictive Modeling:

This research explores the following machine learning models for churn prediction:

Logistic Regression: A linear model that predicts the probability of churn using a sigmoid function. It's relatively simple and interpretable, providing insights into feature importance.

Support Vector Machines (SVM): A powerful algorithm that finds the optimal hyperplane to separate churned and non-churned customers in a high-dimensional feature space. We will explore different kernel functions (linear, polynomial, RBF) and their impact on performance.

Random Forest: It is an ensemble method that uses many decision trees combined to make better predictions with increased robustness. It works well with non-linear relationships and also gives an estimate of feature importance.

Neural Networks: It is a deep learning model, which uses many layers of interconnected nodes that can learn complex patterns in the data. We will try out different network architectures (number of layers, neurons per layer), activation functions, and optimization algorithms.

For each model, we use hyperparameter tuning techniques such as GridSearchCV or RandomizedSearchCV and use the validation set to discover the best combination of

hyperparameters. We will use cross-validation within the training set to make sure that performance estimates are robust.

Evaluation Metrics:

Accuracy: Overall correctness of the model's predictions.

Precision: The proportion of correctly predicted churned customers out of all customers predicted as churned. Important when minimizing false positives (predicting churn when the customer doesn't).

F1-score: The harmonic mean of precision and recall, providing a balanced measure of performance, especially when dealing with imbalanced datasets.

AUC-ROC: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between churned and non-churned customers across different classification thresholds. A higher AUC-ROC indicates better performance.

Confusion Matrix: A table summarizing the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. This helps in understanding the types of errors made by each model.

Lift Curve: Visualizes the performance of a model by showing how many more true positives are predicted compared to a random model. This is useful for evaluating the effectiveness of targeting a certain percentage of customers based on their churn probability.

IV. CONCLUSION

This research demonstrates the effectiveness of predictive modeling for churn prediction. We identify the most suitable model for the given dataset based on the evaluation metrics and provide a justification for the chosen model. The study highlights the importance of data preprocessing, feature engineering, and hyperparameter tuning in achieving optimal model performance. The findings of this research can help businesses develop targeted retention strategies, reduce customer churn, and improve customer lifetime value. We will also discuss the practical implications of the findings for businesses.

V. REFRENCES

1. Farquad, H. &Vadlamani, Ravi & Surampudi, Bapi. (2024). Churn Prediction using Comprehensible Support Vector Machine: an Analytical CRM Application. Applied Soft Computing. 19. 10.1016/j.asoc.2014.01.031

2. Kumar, Dudyala& Ravi. Vadlamani. (2024). Predicting credit card customer churn in banks using data mining. International Journal of Data Analysis Techniques and Strategies. 1. 4-28. 10.1504/IJDATS.2008.020020.

3. Chih Fong Tsai (2023), "Customer churn prediction through the hybrid neural networks", Expert Systems with Applications 12764-12534.

4. Ning Lu, Bart Baesens "Constructing intelligible customer churn prediction models with advanced rule induction techniques", Expert Systems with Applications 2378-2394,

5. Benlan He, Hua Lin, Jie Lu, Guangquan Zhang "A Customer Churn Prediction Model in Telecom Industry Using Boosting", IEEE Transactions on Industrial Informatics, vol. 10, no. 2, may 2019.

6. Lee et al., İ. Yücedağ and I. A. Doğru, "Customer Churn Prediction Using Machine Learning

Methods: A Comparative Analysis." 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 139. 144, doi: 10.1109/UBMK52708.2021.9558876.

7. Koen W. De Bock, D. Chandrawat and D. Rajeswari, "Smart Farming Techniques for New Farmers Using Machine Learning", Proceedings of 6th International Conference on Recent Trends in Computing, vol. 177, 2021.

8. Ning et al., "The gamma CUSUM chart method for online customer churn prediction", Electronic Commerce Research and Applications, 17 (2020) 99-111.

9. Ver-braken et al., Dirk Van den Poel, "An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction", Expert Systems with Applications 38 (2021) 12293- 12301.

10. P.C. Pendharkar, R. D and P. M. "Prediction of Delamination Size in Composite Material Using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022. pp. 1228-1232, doi: 10.1109/ICEARS53579.2022.9752123.

11. Y. Xie et al. , M. D. S. Alam and M. D. I. Hosen, "To Predict Customer Churn By Using Different Algorithms," 2022 International Conference on Decision Aid Sciences and Applications (DASA). 2022. pp. 601-604. doi: 10.1109/DASA54658.2022.9765155.

12. Koen W. De Bock, Dirk Van den Poel, "Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models", Expert Systems with Applications 39 (2022) 6816-6826.

13. Sangamnerkar, S., Srinivasan, R., Christhuraj. M.R., Sukumaran, R.," An ensemble technique to detect fabricated news article using machine learning and natural language processing techniques", 2020 International Conference for Emerging Technology, INCET 2020, 2020, 9154053

14. L. Ning, L. Hua, L. Jie, Z. Guangquan, "A customer churn prediction model in telecom industry using boosting", IEEE Trans. Ind. Inform. 10 (2020) 1659-1665.

15. K. Goyal, K. Kanishka, K. Vasisth, S. Kansal and R. Srivastava, "Telecom Customer Churn Prediction: A Survey," 2021 3rd International Conference on Advances in Computing,-wnloaded on April 05,2023 at 06:56:34 UTC from IEEE Xplore. Restrictions apply.

16. Communication Control and Networking (ICAC3N), 2021. pp. 276-280, doi: 10.1109/ICAC3N53548.2021.9725621.

17. S. De, P. P and J. Paulose, "Effective ML Techniques to Predict Customer Churn," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 895-902, doi: 10.1109/ICIRC A51532.2021.9544785.

18. V. Geetha, A. Punitha, A. Nandhini, T. Nandhini, S. Shakila and R. Sushmitha, "Customer Churn Prediction In Telecommunication Industry Using Random Forest Classifier," 2020
International Conference on System, Computation. Automation and Networking (ICSCAN), 2020, pp. 1-5, doi: 10.1109/ICSCAN49426.2020,9262288

19. Srinivasan, R., Subalalitha, C.N. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distrib Parallel Databases https://doi.org/10.1007/s10619- 021-07331-4. (2021).