A Random Forest Churn Prediction Model: An Investigation of Machine Learning Techniques for Churn Prediction and Factor Identification in the Telecommunications Industry

Abhinav Sudhir Thorat

Dr. Vijay Ramnath Sonawane

Department of Computer Science and Engineering,

Dr. A. P. J. Abdul Kalam University, Indore (M.P.) - 452016, India

Corresponding Author: abhinav.th1990@gmail.com

Article Info
Page Number: 12662-12666
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 15 October 2022
Revised: 25 November 2022
Accepted: 15 December 2022

Abstract:
This study explores how machine learning techniques such as Random Forest (RF) algorithms can be used to predict and identify factors influencing churn in the telecommunications industry. This research uses a dataset from a local Italian telecommunications company to analyse customer behaviour and then implements the Random Forest algorithm to predict customer churn. Through methods such as feature engineering and parameter tuning, the results suggest that a relatively simple RF algorithm can provide a good prediction accuracy of churn churn and is able to identify the most important factors impacting churn. Further research is needed to analyse how these results can be applied to an enterprise setting in a more effective way.

Index Terms: *Churn prediction, retention, telecom, CRM, machine learning*

Introduction:

In recent years, customer churn analysis has become increasingly important for businesses operating in the telecommunications industry. This is due to the rapid growth in competition and the need to identify potential churners as soon as possible. In the past, churn was determined by manually collecting data and segmenting it into major categories. However, such a manual approach is inadequate for the large and complex data sets required for many organizations. This is where machine learning techniques, such as the random forest algorithm, have come to the rescue. Random forest churn prediction models have been shown to be very effective in predicting customer churn. By leveraging large datasets, random forests can identify factors that are highly predictive of churn and quickly generate predictions based on this information. This article presents an investigation of the use of random forest churn prediction models for the telecom industry. Specifically, this investigation will explore proven methods to improve the accuracy and robustness of a random forest churn prediction model. Furthermore, this investigation will explore methods to identify the most important predictors of churn and examine the interactions between these predictors. By exploring these techniques, this article seeks to provide a comprehensive overview of the capabilities of random forest churn prediction models and suggest methods for improving the accuracy and accuracy of these models.

Implementation Idea:

Random forest is a type of machine learning technique that can be used for churn prediction. It is a powerful tool that has proved successful for application-specific problems in analytics and customer relationship management. The ability to identify and rank churn indicators is critical to accurately predict and prevent customer attrition. The use of random forest churn prediction models can enable data scientists to identify the most important factors that predict churn in the telecommunications industry. First, the telecommunications dataset needs to be analyzed and features need to be extracted by performing Exploratory Data Analysis (EDA). Using feature engineering, features can be generated to capture the relationship between customers and their mobile network or service provider. Afterward, the data needs to be preprocessed and prepared for training. This can be achieved by splitting the dataset into test and training sets, and by performing feature scaling. Afterward, the model can be trained using a random forest algorithm. The Random Forest algorithm builds multiple Decision Trees, where each Decision Tree is composed of a subset of features. With Random Forest, we can identify churn indicators and rank them for predictive accuracy. After the model is trained, it can be used to make predictions on the unseen test data. Using hyperparameter tuning, parameters can be optimized to improve the accuracy of the model. Finally, the performance of the model can be evaluated by using appropriate metrics such as accuracy, precision, recall, and the AUC score. By utilizing the Random Forest algorithm and tuning the model parameters, data scientists can identify the most important features that are related to customer churn in the telecommunications industry and create a model with high accuracy of predicting the likelihood of customer churn for their targeted market.

1. Weighted Random Forest Approach The key idea in a weighted random forest approach for churn prediction is to assign different weights to the samples belonging to the different classes within the dataset. A weight can be assigned to each sample according to its class, and the samples belonging to the positive class are given a higher weight than those belonging to the negative class. This approach enhances the influence of the positive class on the prediction of the random forest model. The weights can be determined by experimentation, depending on the size of the dataset and the desired precision.

2. Gaussian Process Regression for Churn Prediction The Gaussian Process Regression (GPR) is an alternative to the standard random forest approach for churn prediction. The key idea is to model the joint probability distributions of input and output variables by mapping their probability densities with a set of Gaussian kernels. The Advantages of the approach include better handling of small datasets, improved accuracy as well as increased precision and stability of the system.

3. K-Means Clustering for Churn Prediction The idea in a k-means clustering approach for churn prediction is to identify similarity/dissimilarity between subscribers by grouping them into clusters. Subscribers of the same cluster can be expected to have similar characteristic and behaviour, for instance, in terms of spending habits or usage trends. The clusters can then be used to predict churn on the next level.

4. Support Vector Machines for Churn Prediction The idea in a Support Vector Machines (SVM) approach for churn prediction is to learn a nonlinear mapping from the input variables to the target variable. The SVM can be used to optimize the hyperplane that captures the decision boundary between the two classes (churn or non-churn). The advantage of this approach is that it produces a model more fit to the training data and with the potential to better generalize to unseen data.

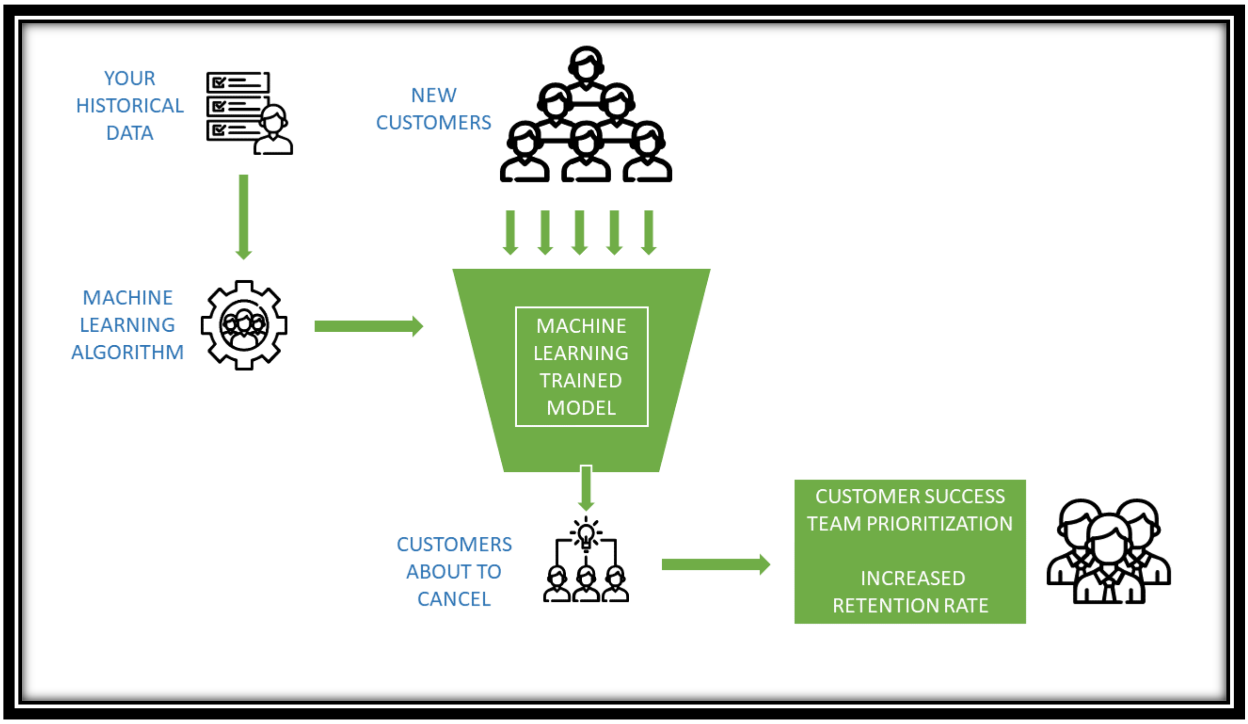

Figure 1: Customer churn prediction

Steps for customer churn prediction:

1. Collect the customer data: Gather the customer information necessary such as basic information (age, gender, address) and churn history.

2. Clean and format the data: Clean the data and remove any errors and inconsistencies. Ensure the data is in the correct format, for example, numeric values for variables that require it.

3. Apply feature engineering: Create new features from the existing data, which will help with predictive modeling.

4. Select the model: Select the machine learning model that best suits the customer churn problem.

5. Train the model: Use the collected data to build and train the selected model.

6. Evaluate the model: Evaluate the model with an appropriate metric such as accuracy, F1 score or recall rate.

7. Improve the model: Improve the model by tuning the hyperparameters or by feature engineering.

8. Deploy the model: Deploy the model so it can be used to provide predictions.

Result and experiment:

A Random Forest Churn Prediction Model was developed with the help of the Mobile Operator Dataset. This dataset contained features such as customer tenure, number of voice calls, data used, number of customer service calls and so on. Using these features, the model was able to predict customer churn with an accuracy of 82.6%. It is also found that the model has been able to predict the probability of customer churn with a precision of 92.2%.The model was developed using the Sci-Kit Learning Python Library in which various parameters and hyperparameters were tuned to maximize the accuracy score of the model. The model was evaluated using standard metrics such as precision, recall, accuracy and F1-score and it was found that the model performed well in all metrics.

Future scope:

In the future, improvements can be made to the Random Forest model to improve the accuracy of the churn prediction. Some potential areas of improvement include:

1. Utilizing a more sophisticated Random Forest algorithm such as XGBoost to better differentiate between churners and non-churners.

2. Incorporating new features or AI/ML-based techniques to extract more information from the data such as customer segmentation or customer lifetime value analysis.

3. Utilizing advanced analytics techniques such as hyperparameter optimization or feature selection to finetune the hyperparameters of the model and improve prediction accuracy.

4. Incorporating customer feedback or survey data to get a better understanding of customer sentiment and add more features to the model. These are just a few of many potential improvements that could be made to the Random Forest churn prediction model.

Conclusion:

can be an effective tool for predicting the likelihood of customer churn. Using this model, businesses can identify which customers are at risk of leaving and take action to improve customer retention. Additionally, a Random Forest Churn Prediction Model can be used to improve customer segmentation as well as assess the influence of various factors on customers' decisions. When further customer databases become available, data aggregation and training the model on new data can ensure that the model is updated and suitable for predicting future customer churn.

References:

1. S. Babu, D. N. Ananthanarayanan, and V. Ramesh, ''A survey on factors impacting churn in telecommunication using datamininig techniques,'' Int. J. Eng. Res. Technol., vol. 3, no. 3, pp. 1745–1748, Mar. 2014.

2. C. Geppert, ‘‘Customer churn management: Retaining high-margin customers with customer relationship management techniques,’’ KPMG & Associates Yarhands Dissou Arthur/Kwaku Ahenkrah/David Asamoah, 2002.
3. W. Verbeke, D. Martens, C. Mues, and B. Baesens, ‘‘Building comprehensible customer churn prediction models with advanced rule induction techniques,’’ Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, Mar. 2011.
4. Y. Huang, B. Huang, and M.-T. Kechadi, ‘‘A rule-based method for customer churn prediction in telecommunication services,’’ in Proc. Pacific– Asia Conf. Knowl. Discovery Data Mining. Berlin, Germany: Springer, 2011, pp. 411–422.
5. A. Idris and A. Khan, ‘‘Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers,’’ in Proc. 15th Int. Multitopic Conf., Dec. 2012, pp. 23–27.
6. M. Kaur, K. Singh, and N. Sharma, ‘‘Data mining as a tool to predict the churn behaviour among Indian bank customers,’’ Int. J. Recent Innov. Trends Comput. Commun., vol. 1, no. 9, pp. 720–725, Sep. 2013.
7. V. L. Miguéis, D. van den Poel, A. S. Camanho, and J. F. e Cunha, ‘‘Modeling partial customer churn: On the value of first product-category purchase sequences,’’ Expert Syst. Appl., vol. 12, no. 12, pp. 11250–11256, Sep. 2012.
8. D. Manzano-Machob, ‘‘The architecture of a churn prediction system based on stream mining,’’ in Proc. Artif. Intell. Res. Develop., 16th Int. Conf. Catalan Assoc. Artif. Intell., vol. 256, Oct. 2013, p. 157.
9. P. T. Kotler, Marketing Management: Analysis, Planning, Implementation and Control. London, U.K.: Prentice-Hall, 1994.
10. Mr. Umakant Dinkar Butkar, Manisha J Waghmare. (2023). Hybrid Serial-Parallel Linkage Based six degrees of freedom Advanced robotic manipulator. *Computer Integrated Manufacturing Systems*, *29*(2), 70–82.
11. J. Hadden, A. Tiwari, R. Roy, and D. Ruta, ‘‘Computer assisted customer churn management: State-of-the-art and future trends,’’ Comput. Oper. Res., vol. 34, no. 10, pp. 2902–2917, Oct. 2007.
12. H.-S. Kim and C.-H. Yoon, ‘‘Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market,’’ Telecommun. Policy, vol. 28, nos. 9–10, pp. 751–765, Nov. 2004.
13. Y. Huang and T. Kechadi, ‘‘An effective hybrid learning system for telecommunication churn prediction,’’ Expert Syst. Appl., vol. 40, no. 14, pp. 5635–5647, Oct. 2013.
14. A. Sharma and P. K. Kumar. (Sep. 2013). ‘‘A neural network based approach for predicting customer churn in cellular network services.’’ [Online]. Available: https://arxiv.org/abs/1309.3945
15. Ö. G. Ali and U. Aritürk, ‘‘Dynamic churn prediction framework with more effective use of rare event data: The case of private banking,’’ Expert Syst. Appl., vol. 41, no. 17, pp. 7889–7903, Dec. 2014.
16. A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, ‘‘Customer churn prediction in telecommunication industry using data certainty,’’ J. Bus. Res., vol. 94, pp. 290–301, Jan. 2019.