Hybrid Model for the Customer Churn Prediction

Mansimar Anand, Irtibat Shaukat, Harnoor Kaler, Jai Narula, Prashant Singh Rana

Thapar Institute of Engineering and Technology, Patiala, India

Corresponding author: Mansimar Anand, Email: manand_be18@thapar.edu

The fast development of the showcase in each segment is driving to a prevalent endorser base for benefit suppliers. In such a quick setup, benefit suppliers have realized the significance of holding the on-hand clients. It is in this manner fundamental for the benefits suppliers to prevent churn - a phenomenon that states that the client wishes to quit the benefit of the company. The key here is to be motivated and have interaction with these clients. While simple, in theory, the realities worried about achieving this "proactive retention" goal are incredibly challenging. The most commitment of our work is to create a Churn expectation show that helps companies foresee clients who are most likely subject to churn. The model developed in this work employs machine learning strategies on the dataset and builds a robust training and testing model. The proposed model results are authenticated using K-fold cross-validation, and an accuracy of 91.48% is achieved. The main contribution is using K-means clustering to make clusters and then applying the Random Forest classifier for model prediction. The model was organized and tested operating on a data set created and supplied by a telecom organization from the United States of America. The model experimented with seven algorithms: Random Forest, Logistic Regression, Naive Bayes, K-nearest neighbors, Gradient Boosting, Ada Boosting, and K-means. However, the proposed model is experimented with by combining the K-means and Random Forest algorithm.

Keywords: Machine Learning, Hybrid Model, Predictive Model, Feature Selection, Customer Churn.

2021. In Prashant Singh Rana, Deepak Bhatia & Himanshu Arora (eds.), *SCRS Proceedings of International Conference of Undergraduate Students*, 85–94. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-95502-01-1-9

1 Introduction

Investigating client churn for gigantic information in client maintenance is an open investigation in machine learning innovation [1]. Client churn is the misfortune of clients who changes from one division, for illustration, a bank, media transmission arrange, among others, to another contender inside a specific time [2]. Client churn misclassification utilizing clustering can incite enormous monetary misfortunes and harm the association's advancement. Client churn administration is fundamental, especially for businesses like managing an account industry where data have been utilized; examination of colossal information having a considerable number of dreary, excess, and boisterous data must be wiped out [3]. The client relationship [4] board strategizes, directs, and looks at client associations and data all through the client's lifecycle within the organization to oversee the connections with clients in terms of trade and drive exchanges improvement [5]. Analyzed information of customers is accumulated to form proper and robust choices [6]. Client churn will result in the misfortune of organizations; client churn predicts the people who are progressing to churn. The client churn prognosticates got to a developing consideration amid the foremost, this later decade as one of the essential measures to hold clients. Churn expectation has been trepidation within the money-related and investigative world. As of late, different information mining strategies [7] have been gotten for churn forecast, counting standard measurable procedures, for case, calculated relapse, non-parametric models like for case logistic regression [8], k-nearest neighbor [9], Bayesian classifiers [10], choice trees, and neural systems [11]. In this examination, mining a client behavior is inspected by utilizing the k-means clustering [12] method and random forest classification [13] based on customer's highlights to assist the monetary businesses to recognize unmistakable sorts of clients and the churn behaviors.

2 Methods

Data set and its features

The customer churn data set is collected from a telecom company in the United States of America. Each row in the dataset reflects a specific customer's information, and each column in the dataset comprises the customer attributes indicated in the column Metadata. The dataset has 3333 rows (customers) and 20 columns (features). The data set includes information about the services that each customer has signed up for, account information, and demographic information. A glimpse of the data set is presented in Table 1 and Table 2.

Table 1. Illustrations of the features

Index	Feature	Feature Representation
1	State	F_a
2	Account length	F_b
3	Area code	F_c
4	Phone number	F_d

5	International plan	F_e
6	Voicemail plan	F_f
7	Number vmail messages	F_g
8	Total day minutes	F_h
9	Total day calls	F_i
10	Total day charge	F_j
11	Total eve minutes	F_k
12	Total eve calls	F_l
13	Total eve charge	F_m
14	Total night minutes	F_n
15	Total night calls	F_o
16	Total night charge	F_p
17	Total intl minutes	F_q
18	Total intl calls	F_r
19	Total intl charge	F_s
20	Customer service calls	F_t
21	Churn	F_u

Data pre-processing

Data cleaning: The data obtained was clean in the sense that it does not consist of missing values. At this phase dataset contains 3333 records and 20 features. The 2850 records have a 'False' churn value in this data set, and 483 have 'True'.

Feature Selection: It may contain data that is not necessary to improve our results. The best way to identify is by logical thinking or by creating a correlation matrix [14]. In this dataset, we have a phone number feature that will not influence our predicted outcome. Therefore, after dropping this column, we have 19 columns (features).

Data type conversion: The data obtained have different data types. Some columns have string values from which we cannot calculate anything. To give them meaning, convert string values into numeric ones. To achieve this, we used Label Encoding [15] which refers to changing over the labels into a numeric form to change over them into the machine-readable frame.

Table 2. Sample dataset of Customer Churn

State	Account length	Area code	—	Total intl charge	Customer service calls	Churn
KS	128	415	—	2.70	1	False
OH	107	415	—	3.70	2	False
NJ	137	415	—	3.29	0	False
OH	84	408	—	1.78	2	False
OK	75	415	—	2.73	3	False

Machine-Learning Methods

Machine learning approaches enable the creation of adaptable algorithms that can alter spontaneously in response to their inputs. Models' adaptability allows them to continue their demonstrations without being programmed. Within the show think about, seven models are investigated as mentioned in Table 3 with required packages and tuning parameters. Modes are adjusted to urge superior outcomes, but

default parameters have been utilized in the display. The brief detail of models includes Random Forest and K-means clustering, which are used in the hybrid model [16] given below:

Random forest is an ensemble approach, which means it comprises of many smaller decision trees called estimators, each of which makes its own predictions. As an ensemble method, the random forest provides the flexibility to accommodate problems like over-fitted/under-fitted/biased datasets, giving a robust and generalized method to handle unknown datasets and missing values in the dataset. Random forests also handle large datasets with high dimensionality and heterogeneous feature types.

Bagging – It generates a distinct training subset with replacement from sample training data, and the final output is determined by majority vote.

Boosting – It combines weaker learners into strong learners by building successive models [17] with the best accuracy as the final model.

K-means clustering [12] is a powerful unsupervised machine learning algorithm that tries to group similar items in clusters. It finds the similarity between the items based on feature similarity and groups them into clusters. The calculation works iteratively to allot each information point to one of the K groups based on the highlights that are given. Each centroid of the individual cluster represents a collection of attributes that defines the particular cluster. The asymptotic time complexity for finding the global minimum is NP-complete. The k-means algorithm updates cluster centroids till a local minimum is found.

Logistic Regression [8] is a classification algorithm that uses the logistic function, also called the sigmoid function, at its core. It's an S-shaped curve that can project any real-valued integer to a value between 0 and 1 but never exactly between those two points. Input values are combined linearly using weights or coefficient values to predict an output value, modeled to binary values, i.e., 0 or 1.

K-Nearest Neighbors [18] is a classification algorithm that makes predictions for a new data point by searching through the entire dataset for the K most similar instances and summarizing the output variable for those K instances. Distances used to find K nearest neighbors are Euclidean, Hamming, Manhattan, Tanimoto, Jaccard, Minkowski, Mahalanobis, Cosine, etc.

The Bayes theorem is used to create a classification method, Naive Bayes [19]. Furthermore, their time complexity scales linearly with the number of features, making naive bayes a suitable tool for high-dimensional data. The assumption in naive Bayes is that all characteristics are independent of one another, resulting in serious errors in predictions.

AdaBoost (Adaptive Boosting) [20] is an algorithm with access to a weak learner and finds a hypothesis with low empirical risk. Weak models are added one by one, and they are trained using weighted training data. The process is repeated until a predetermined number of weak learners has been produced (a user parameter), or the training dataset can no longer be improved. After you've finished, you'll have a pool of weak learners, each with a stage value.

Gradient Boosting [21] is a boosting technique driven by the principle of increasingly refined approximations. Gradient Boosting starts with a base model and each predictor corrects its predecessor's error. Each predictor is trained with the predecessor's residual errors as labels, leading in a versatile technique for regression and multi-class classification.

Table 3. Machine learning models

Model	Package	Tuning Parameters	Ref.
Logistic Regression	sklearn.linear_model	random_state, solver, max_iter	[22]
K-nearest Neighbors	sklearn.neighbors	n_neighbors	[23]
Random Forest	sklearn.ensemble	max_depth, n_estimators	[24]

Naive Bayes	sklearn.naive_bayes	None	[25]
K-Means	sklearn.cluster	n_clusters, random_state	[26]
Ada Boost	sklearn.ensemble	n_clusters, random_state	[27]
Gradient Boosting	sklearn.ensemble	max_depth, n_estimators, learning_rate	[28]

Motivation

In today's mercurial and operative setup, retailers have understood the importance of holding customers primed. The biggest challenge in our job is creating an abandonment expectations program to help businesses predict which customers are most likely to be abandoned. The authors adopt different anomaly detection and machine learning, but minimal works have taken substantial data sets and features to check the validity of their proposal. The massive use of classification, regression, and clustering analysis is a significant challenge during the implementation. The existing classification algorithm's performance has not been satisfactory. The reviews exhibit that little, or no work has been done to originate robust techniques for customer churn prediction. The problem needs in-depth analysis and improved methods to predict accurate results. Further, customer churn predictions are one of the techniques to understand customer behavior. The accurate predictions could lead to the deterioration of the gap between customers and suppliers. This research aimed to secure an understanding and improve customer behavior prediction. This hybrid model will acknowledge companies to understand customer behavior better and enhance their services.

Contribution

It is beneficial to handle the customer churn for any organization to grow or survive. This hybrid model technique helps to predict results accurately. The data set was collected from the UC Irvine Machine Learning Repository. The following is the contribution of the proposed completion:

1. Different models and strategies are compared and tested, like Logistic Regression, Random Forest, N-Nearest Neighbors, Naive Bayes, and K-Means are analyzed with different parameters and performance metrics.
2. Different boosting strategies like Ada boosting and Gradient boosting are discussed, and the results are analyzed.
3. Novelty of the work lies in developing a hybrid model using K-Means and Random Forest models.

Our major contributions are described as follows:

1. The accuracy of the prediction of customer churn is improved.
2. The Precision of the results is improved.
3. The space and time complexity reduction for the model.

Methodology

In the beginning, customer churn data is collected. The data set contains customers' data, including their churn value, either True or False. In machine learning, algorithms have multiple data repeatedly and allow the computer to search for a problem without human interference. A practical methodology is required to perform this task effectively and proposed, as described in Fig. 1. The next stage utilizes data to illustrate and comprehend the issue statement's complexity. Then, like a phone number in the prediction, delete the attributes that contribute little or nothing. The acquired data is then used to train classifiers using the best tuning settings available. Table 3 lists the machine learning models that were

employed. Figure 1 depicts the proposed model. Finally, the model's performance was validated using the process of repeated k-fold cross-validation and diverse performance parameters like AUC, recall, precision, Gini, and accuracy.

Flow of proposed scheme

Fig 1 describes the customer churn prediction using the features mentioned in Table 1 and trains the machine learning models. Two models are integrated to create the proposed model, explained in the next section. The proposed model produces the final prediction. The proposed model is trained for classifying customer churn for both negative and positive cases.

Proposed Hybrid model

The hybrid is used to deal with the worst case of the model prediction. In the present work, the focus is on the model's false prediction and accurate prediction, and the hybrid model is used to deal with false and accurate predictions. Two models, i.e., Random Forest Classifier and K-Means clustering, are combined to improve accuracy, as mentioned in Fig 1. All the models are trained on 75% of the data set, and 25% is used for testing. The phases of the hybrid model are discussed below:

Phase 1:

1. Training Machine learning model for making clusters using K-Means clustering.
2. Adding the cluster number to the data set as a new feature.
3. Training Machine learning model for each cluster using Random Forest classifier.

Phase 2:

1. Find the cluster for the test data point using K-Means.
2. Testing the models and evaluating the predicted results.

3 Results

Performance Evaluation

Evaluation metrics like precision, recall, accuracy, AUC, and GINI are utilized to assess the robustness of the proposed hybrid model and each separate model. In the next section, these parameters will be briefly described.

Gini coefficient is a statistic for evaluating inequality in distribution. The Gini coefficient has a value between [0, 1]. The Gini coefficient of 1 denotes inequality, whereas the Gini coefficient of 0 suggests equality.

$$\square\square\square\square = 2 * \square\square\square - 1 \qquad (1)$$

AUC is a useful indicator for assessing the performance of classification models. The AUC statistic definitely aids in determining and informing us about a model's capacity to discriminate across classes. The AUC is the judging criterion, and the higher the AUC, the better the model.

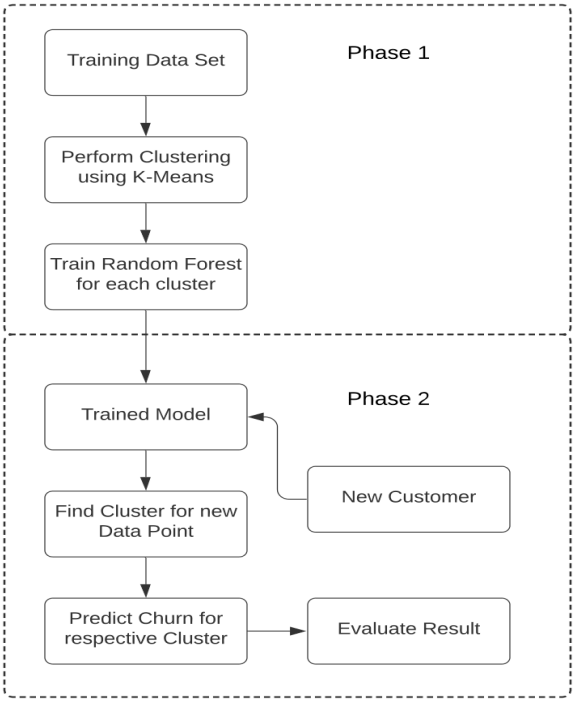

Fig. 1. Flow of the proposed hybrid model

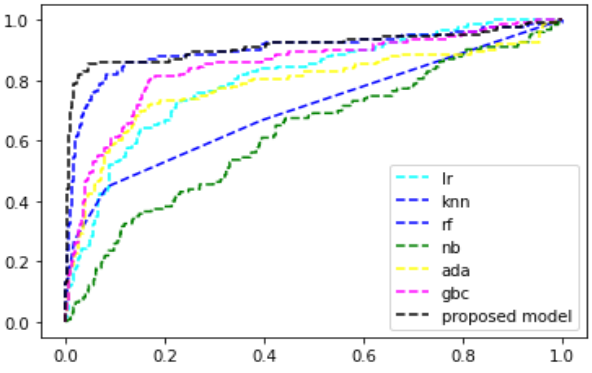

Fig. 2. ROC curve individual and proposed model

Accuracy is a performance metric, informally, the fraction of correct predictions out of the total number of predictions made by the classifier:

$$\square\square\square\square\square\square\square\square = \frac{\square\square + \square\square}{\square\square + \square\square + \square\square + \square\square} \qquad (2)$$

where TP is the number of true positive decisions, TN is the number of true negative decisions, FN is the number of false negative decisions and FP is the number of false positive decisions.

Precision is a performance indicator defined as the quality of positive results made by the model. The precision is calculated as follows:

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \qquad (3)$$

Recall is a performance indicator that indicates the fraction of correct positive predictions from all the given examples of a certain class. In other words, precision calculates the coverage of positive data class. The recall is calculated as follows:

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \qquad (4)$$

Repeated k- Fold Cross- Validation

Cross-validation is a technique for evaluating and comparing models that divide the data set into two segments, one for training and the other for validating the model. Cross-major validation's purpose is to ensure that each data segment of the dataset has an equal probability of appearing in both the training and testing dataset samples. In k-fold cross-validation, each division of the dataset is tested precisely once, and the rest of the time is considered part of the training dataset. Repeated k-fold cross-validation is used to increase the probability of each division in the testing dataset sample. To make the system more robust, repeated k-fold cross-validation is deployed as it provides more iteration. The dataset was partitioned into ten folds in this investigation, with each fold being run five times.

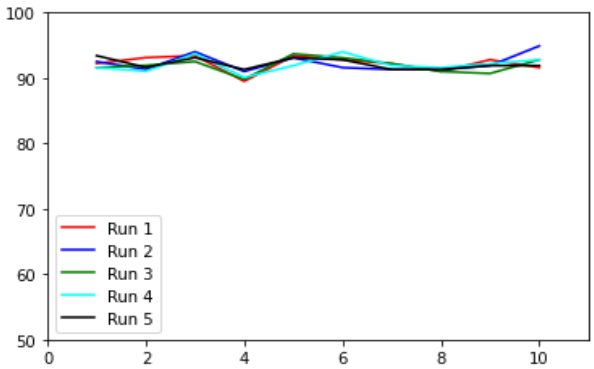

Fig. 3. Repeated ten-fold cross-validation of proposed hybrid model

Evaluation of Results

As mentioned in Sect. 2 (Methods), the machine learning models were efficiently trained on the dataset. Because data are very important in the training model, the data partition in this study was done effectively. Individual models are chosen based on their abilities. We used the classification algorithm that performs the best on the stand-alone dataset, and then we used the K-Means clustering algorithm to make clusters for the data set and used the best classification algorithm to predict the result. In comparison to individual models, the developed hybrid model eliminates incorrect predictions and improves predictability. Table 4 shows the results, which show that the performance of

the proposed hybrid model outperforms the individual models. It's possible that the model will be over-fitted/under-fitted/biased once it's been trained. To address these concerns and ensure the robustness of the proposed hybrid model, five-time tenfold cross-validation is performed. The dataset is partitioned into ten divisions for repeated k-fold cross-validation. Fig. 4 displays the accuracy for each run. In five-time ten-fold cross-validation, the developed hybrid model has an average accuracy of 91.48%. As a result, the prescribed hybrid model solves problems like over-fit/under-fit/bias. Figure 3 shows the ROC curve plots for the proposed and individual models. Because the quality of the model is determined by the bend of the curve towards the top left corner, it is clear from the plot in Fig. 3 that the described hybrid model outperforms the other six models.

Table 4. Performance evaluation on various parameters of the individual and proposed hybrid models

Index	Model	Gini	AUC	Accuracy	Recall	Precision
M1	Logistic Regression	0.609	0.805	0.861	0.874	0.977
M2	K-nearest Neighbors	0.412	0.705	0.878	0.886	0.983
M3	Random Forest	0.803	0.902	0.875	0.873	0.998
M4	Naive Bayes	0.251	0.626	0.603	0.899	0.602
M5	Ada Boost Classifier	0.569	0.785	0.873	0.898	0.961
M6	Gradient Boosting Classifier	0.692	0.846	0.880	0.905	0.961
M7	Proposed Hybrid Model	0.828	0.914	0.915	0.913	0.994

4 Conclusion

In comparison to the previous method, the suggested model improves the accuracy of customer churn prediction. Data division is used to refine incorrect predictions when training the suggested hybrid model. The advantage of utilizing this method is that it improves the suggested model's performance and efficiency. In the present study, seven models were trained and tested, i.e., random forest, logistic regression, k-nearest neighbors, naive bayes, ada boosting, gradient boosting, and k-means. The data are circulated to seven individual models, which train them adequately to provide reliable and accurate results. The random forest and k-means are being used to create a hybrid model. A novel hybrid model has been developed for the prediction, and it produces high accuracy, Gini, AUC, precision, and recall. The proposed model working has been divided into two phases. The data is traveled through clustering to make clusters of data set, then classification to produce reliable and accurate results. The exactness of the proposed demonstration is approved utilizing the repeated k-fold cross-validation procedure.

References

[1] Hadden, J. et al. (2006). Churn Prediction: Does Technology Matter. International Journal of Intelligent Technology, 1: 104–110.

[2] Michel Ballings and Dirk Van den Poel. (2012). Customer event history for churn prediction: How long is long enough?. Expert Systems with Applications, 39: 13517–13522.

[3] Junxiang Lu and Overland Park. (2021). Modeling Customer Lifetime Value Using Survival Analysis, An Application in the Telecommunications Industry.

[4] Stewart Arnold and D. Nguyen. (2006).The value of consultant-client relationships: Perspectives from both sides of the fence. Academy of Management Proceedings.

[5] Mosa Alokla et al. (2019). Customer Relationship Management: A Review and Classification. Transnational Marketing Journal, 7: 187–210.

[6] Nabgha Hashmi, Naveed Anwer Butt, and Muddesar Iqbal (2013). Customer Churn Prediction in Telecommunication A Decade Review and Classification. IJCSI, 10: 271–282.

[7] A.K. Ahmad, A. Jafar, and K. Aljoumaa. (2019). Customer churn prediction in telecom using machine learning in big data platform. J Big Data.

[8] Joanne Peng, Kuk Lee, and Gary Ingersoll. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research - J EDUC RES, 96: 3–14.

[9] Gongde Guo et al. (2004). KNN Model-Based Approach in Classification.

[10] Pat Langley, Wayne Iba, and Kevin Thompson. (1998). An Analysis of Bayesian Classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence 90.

[11]Enzo Grossi and Massimo Buscema. (2008). Introduction to artificial neural networks. European journal of gastroenterology and hepatology, 19: 1046–54.

[12] Youguo Li and Haiyan Wu. (2012). A Clustering Method Based on K-Means Algorithm. Physics Procedia, 25: 1104–1109.

[13] L. Breiman. (2001). Random Forests. Machine Learning, 5–32.

[14] Thu Pham-Gia and Vartan Choulakian. (2014). Distribution of the Sample Correlation Matrix and Applications. Open Journal of Statistics, 04: 330–344.

[15] Kedar Potdar, Taher Pardawala, and Chinmay Pai. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. International Journal of Computer Applications, 175: 7–9.

[16] Giorgio Valentini and Francesco Masulli. (2002). Ensembles of Learning Machines. 2486: 3–22. isbn: 978-3-540-44265-3.

[17] Peter Bühlmann. (2012). Bagging, Boosting and Ensemble Methods. Handbook of Computational Statistics.

[18] Padraig Cunningham and Sarah Delany. (2007). k-Nearest neighbour classifiers. Mult Classif Syst, 54.

[19] Pouria Kaviani and Sunita Dhotre. (2017). Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management, 04.

[20] Tu Chengsheng, Liu Huacheng, and Xu Bing. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences, 139: 00222.

[21] Alexey Natekin and Alois Knoll. (2013). Gradient Boosting Machines, A Tutorial. Frontiers in neurorobotics, 7: 21.

[22] Logistic Regression Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.linear_mode l.Logistic Regression.html.

[23] KNeighborsClassifier Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.K NeighborsClassifier.html.

[24] RandomForestClassifier Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble .RandomForestClassifier.html.

[25] GaussianNB Homepage. url: https://scikit-learn.org/stable/modules/naive_bayes.html.

[26] KMeans Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

[27] AdaBoostClassifier Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.Ada BoostClassifier.html.

[28] GradientBoostingClassifier Homepage. url: https://scikit-learn.org/stable/modules/generated/sklearn.ensem ble.GradientBoostingClassifier.html.