# Customer Churn Prediction for Telecommunication: Employing various various features selection techniques and tree based ensemble classifiers

Adnan Idris
Pattern Recognition Lab., Department of Computer and Information Sciences
Pakistan Institute of Engineering and Applied Sciences
Nilore, Islamabad 45650, Pakistan
adnanidris@pieas.edu.pk

Asifullah Khan
Pattern Recognition Lab., Department of Computer and Information Sciences
Pakistan Institute of Engineering and Applied Sciences
Nilore, Islamabad 45650, Pakistan
asif@pieas.edu.pk

***Abstract*—Ensemble classifiers have received increasing attention for attaining the higher classification performance in recent times. In this paper, we present comparative performances of various tree based ensemble classifiers in collaboration with maximum relevancy and minimum redundancy (mRMR), Fisher's ratio and F-score based features selection schemes for a challenging problem of churn prediction in telecommunication. The large sized telecommunication dataset has been the main hurdle in achieving the desired classification performance in the contemporary proposed churn prediction models. Though, tree based ensemble classifiers are considered suitable for larger datasets, but we have found rotation forest and rotboost as effective techniques compared to random forest, which employ boosting through features selection and increased diversity by incorporating linear feature extraction method such as Principal Component Analysis. In addition to the features selection performed by used ensembles, we have also incorporated mRMR, Fisher's ratio and F-score techniques for features selection. mRMR returns a coherent and well discriminants feature set, compared to Fisher's ratio and F-score, which significantly reduces the computations and helps classifier in attaining improved performance. The performance evaluation is conducted using area under curve, sensitivity and specificity where Rotboost, an ensemble of rotation forest and Adaboost in collaboration with mRMR has shown competitive results for churn prediction in telecommunication as compared to other ensemble methods.**

***Keywords-churn prediction;teleommunication; Rotation Forest; RotBoost ; mRMR***

## I. Introduction

Customer churn prediction in telecommunication is attaining serious attention of the stakeholders in order to retain the customer's loyalty and improve the standard of customer relation management. The telecom operators also realize the importance of retaining the customers instead of striving for adding new customers. The cost incurred to add a new customer is far more then retaining a customer whose appetite is not being properly served [1]. The telecom operators not only stabilizes the customer base but save the customer churning by appropriately targeting the sect of customers, predicted unsatisfied by a churn prediction model. Customer churn prediction is a binary classification problem but the large dimensionality of the telecommunication dataset makes difficult for conventional binary classifiers like Support Vector Machines to show desired performance [2].

Similarly, the simplest of the classifiers, KNN shows good performance on various classification problems [3][4][5] and its hybridized form with Logistic Regression [6] also claims competitive performance for churn prediction. However, this performance is constrained to the application domains where datasets do not possess high dimensionality and imbalance distribution. Few ensemble classification algorithms have also been applied to model churn prediction in telecommunication .One of such algorithms uses AdaBoost and ANN Boosting to predict churners in telecommunication [7]. Similarly few other ensemble methods comprising of Logit and ANN ensembles [8] and, Bagging and Stochastic Gradient Boosting [9] are also used for modeling churn prediction in telecommunication. Another C5.0 Boosting ensemble [10] also predicts churners in telecommunication. Though ensemble classifiers are considered as better performers compared to single classification algorithm [11][12] but these ensemble approaches do not achieve desired accuracy for predicting churners in telecommunication which still creates margin for improvement.

In this study we have used tree and rotation based ensembles for modeling churn prediction for telecommunication. Rotations based ensembles employ rotations on the input data through linear feature extraction algorithms such as Principal Component Analysis (PCA), Independent Component Analysis (ICA) etc. and have been found to exhibit better performance compared to single classification algorithms [13]. We have used Random Forest, Rotation Forest and RotBoost ensembles to

978-1-4673-2252-2/12/$31.00 ©2012 IEEE

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on December 08,2025 at 05:27:36 UTC from IEEE Xplore. Restrictions apply.

Researchers have also used tree based ensemble classification methods, such as Random Forest [13], Balanced Random Forest [14], Rotation Forest[15], RotBoost[16] and its variants[17] for dealing with the problem of churn prediction, but these models unfortunately lacks desired performance for predicting churn in telecommunication. The main hurdle with these tree based ensemble methods is handling high dimensionality of the telecommunication dataset. These tree based algorithms also suffer with the issue of higher memory resources, needed to meet the requirements of tree like data structures. Therefor a considerable margin of improvement exists to efficiently handle the preprocessing phase and reduce the dimensionality of the dataset which would extend better learning capabilities to the classifiers.

The initial preprocessing involved in these classification methods essentially improves the performance through increased diversity and discriminative features space. We have further supplemented the preprocessing phase with various feature extraction schemes and it is found that maximum relevancy and minimum redundancy (mRMR) [19] scheme provides a reduced coherent set of features that consequently lessen the computational cost and extends better learning capabilities. The mRMR based reduced features set overall improves the performances, where RotBoost shows competitive performance over Random Forest and Rotation Forest. Two standard telecommunication datasets have been used to evaluate the experimental results. AUC, sensitivity and specificity based measures are used to evaluate the performances of employed classification schemes.

### A. Novel Contribution

In this work, the application of mRMR as a features selection technique in collaboration with tree and rotation based ensembles to model churn prediction in telecommunication is a novel idea. We have demonstrated the capabilities of mRMR technique as a features selection tool for telecommunication dataset. mRMR reduces the interaclass and increases the interclass distance between the instances which introduces separability and impacts the classifier's performance. Rotation Forest, RotBoost and Random Forest show improvement in their respective performances on collaborating with mRMR. This collaboration makes a unique contribution in the domain of customer churn prediction in telecommunication.

## II. Methodology

In the current study, detailed experimentation is performed to evolve a churn prediction model based on mRMR as a features selection strategy and RotBoost as a classification approach. We have demonstrated how the proposed model attains competitive performance for the telecommunication dataset. The original dataset is initially processed to remove useless and missing values with the help of filters available in WEKA data mining tool. Then mRMR, Fisher's Ratio and F-Score feature reduction methods are employed in collaboration with three ensembles, Random Forest, Rotation Forest and RotBoost as shown in Figure 1. The performance of each ensemble is evaluated in the context of applied feature selection method. The features are selected by linearly searching the features space and finding a feature set which extends better learning capabilities to a classifier that ultimately achieves good prediction performance.

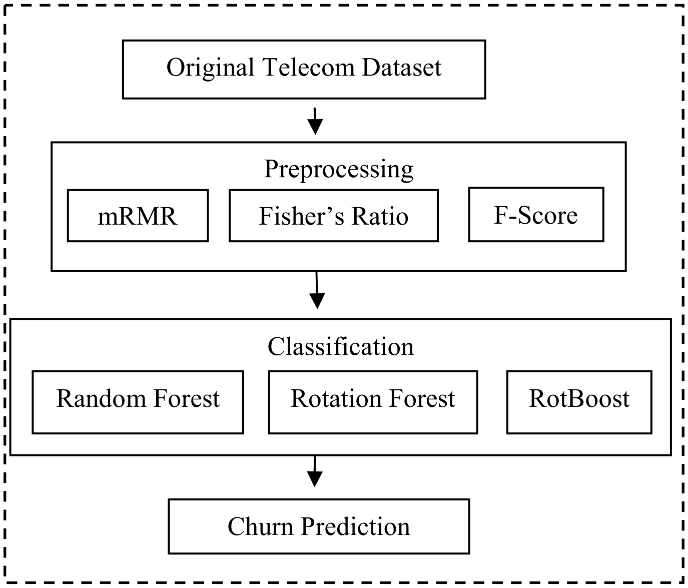

Figure.1 Proposed Churn prediction model

### A. *Maximum Relevancy and Minimum Redundancy (mRMR) based Features selection*

mRMR selects the features, which contain maximum discriminating information. This is accomplished by maximizing the interclass and minimizing the interaclass proximities. The mRMR works by selecting the features which show strong correlation with class labels while being not dependant on each other [19]. mRMR adopts broad criteria for features selection based on minimum redundancy and maximum relevance. The maximum relevance is implemented with the help of the expressions given in 1 and 2;

$$max\ D(S,c), D=I\left(\left\{x_i, i=1,....,m\right\};c\right) \tag{1}$$

$$max\ D(S,c), D=\frac{1}{\left|S\right|}\sum_{x_i \hat{I} S} I(x_i;c) \tag{2}$$

Where *D* is dependency that is intended to be maximized in order to establish maximum relevance of the instances S with class labels *c*. The $I(x_i;c)$measures the mutual information between the instance xi and the corresponding class label c The maximum relevance is sort out by searching the features set which satisfy the criteria in equation 1 and approximates the *D*(*S,c*) in equation 2 with the mean value of all mutual information values between individual feature xi and class c. A features set *S* is chosen where the features have higher dependency on the respective class labels. Once the maximum relevant features are selected, then there can be redundancy between them, therefore one of the two redundant features are removed which would not change the discriminating power of the features set. The expression given in equation 3 minimizes the redundancy:

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on December 08,2025 at 05:27:36 UTC from IEEE Xplore. Restrictions apply.

$$min\ R(S),\ R = \frac{1}{\left[S\right]^2} \sum_{x_i,x_j\hat{I}S} I(x_i,x_j\ ) \quad (3)$$

The above both criterion of minimizing redundancy and maximizing relevance is then combined in one simple form, where Φ operator is defined as given in equation 4. This simplest form is used to optimize both D and R.

$$max\ \Phi(D,R),\ \Phi = D\text{-}R \quad (4)$$

The features set obtained using mRMR is expected to be optimal, for showing strong relevance with class targets and at the same time having features with maximum unique values.

### *B. Fisher's Ratio based features selection*

Fisher's Ratio is considered to be sensitive to the non-normality of the data and measures the discriminating power of the features in the dataset. Fisher's Ratio is computed as given in 5.

$$Fisher's\ Ratio = \frac{\left(\mu_1 - \mu_2\right)}{\sigma_1^2 - \sigma_2^2} \quad (5)$$

Where $\mu_1$, and $\mu_2$ being the means of binary classes involved and $\sigma_1^2$ and $\sigma_2^2$ the respective variances.

### *C. F- Score based features Selection*

F-score is a simple technique which measures the discrimination of two sets of real numbers. Given training vectors $x\ k$, $k = 1, \ldots ,m$, if the number of instances of churner and non-churners classes are $n+$ and $n-$, respectively, then the F-score of the $i$th feature is defined as:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (6)$$

Where $\bar{x}_i$ is mean value of $i$th feature, $\bar{x}_i^{(-)}$ for negative instances, and $\bar{x}_i^{(+)}$ is mean value of the $i$th feature for positive instances. The F-score minimizes the interaclass distance, whereas maximizes the interclass distance of the instances as shown in equation 6. The larger the F-score is, the more likely the feature is more discriminative.

### *D. Random Forests*

Random Forest [13] is an ensemble of decision trees, which are grown over the bootstrap samples of the training dataset, involving random features selection in the process of trees construction. The final predictions are made by aggregating the predictions of all individual trees. The Random Forest for being the ensembles of the decision trees, certainly exhibit substantial performance improvement over single tree based classifiers. Although, Random Forest is considered to be good choice in handling the large sized data but it also suffers in the case of imbalanced training dataset. Random Forest minimizes the overall error rate and therefore in the case of imbalanced dataset the higher total accuracy sometimes undermines the true prediction of the minority class. The telecommunication datasets normally suffer with higher degree of skewness therefore Random Forest sometimes suffers to show appreciable performance.

### *E. Rotation Forest*

The Rotation Forest [15, 17] is also an ensemble classifier that operates by simultaneously improving diversity and accuracy. Rotation Forest achieves high diversity by employing rotation through linear feature extraction methods such as PCA, ICA etc on the input data. The original datasets is divided in *K* subsets (K is a Rotation Forest parameter) derived from *L* original features space. The desired accuracy is attained by utilizing all the components of each subset during learning of the base classifier. This also preserves the variability information in the data. The original feature space L is split into K subsets. Then PCA is applied on each subset which results in K axis rotation and forms the new attributes for a base classifier. Rotation Forest encourages diversity to use PCA as a features extraction method for each base classifier. Whereas utilizing all principal components and using the complete dataset for training each base classifier, it seeks accuracy. Let the class of an instance X is predicted with the Rotation Forest ensemble $C^*$

$$C^*(X) = \underset{y\in\Phi}{\text{argmax}} \sum_{t=1}^{T} I(C_t(XR_t^a) = y) \quad (7)$$

$C_t$, $(1\ldots T)$ shows the base classifier and y corresponds to either 0 or 1 for the binary nature of the churn prediction problem. Whereas, $R_t^a$ shows the rotation matrix, derived for each of the feature subsets $(1\ldots T)$. *I* is an indicator function that assigns the instance X, 0 or 1.

### *F. RotBoost*

RotBoost is an ensemble classifier generation technique that is developed by combining AdaBoost and Rotation Forest[16][17]. Adaboost operates in a sequential manner where each new classifier is constructed keeping in mind the performance of previous classifier. In this method a set of weights are maintained over the original training set where initially they all are kept equal. In subsequent iterations the weights are adjusted so that instances which are misclassified in previous iteration are given more weights and the ones which are correctly classified are given less weights. In this way hard instances are better handled by subsequently trained classifiers. For an ensemble classifier to attain better generalization capabilities, it is essential that ensemble classifier consist of highly accurate classifiers which at the same time differ in their decisions as much as possible. In RotBoost the weight updation over the training data distribution is taken from AdaBoost while rotation matrix is

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on December 08,2025 at 05:27:36 UTC from IEEE Xplore. Restrictions apply.

computed in similar fashion as in Rotation Forest. Decision trees are used as base classifiers in RotBoost.

## III. Experimental Evaluation

In current work, we have performed the detailed experimentation involving aforementioned features selection and classification approaches to model churn prediction for telecommunication. Eventually mRMR and RotBoost appear as best performing tools for modeling churn prediction in telecommunication. AUC, sensitivity and specificity based measures are used to evaluate the performance with 10 folds cross validation.

### *A. Dataset*

Duke University has made available a telecommunication dataset [18], provided by cell2cell telecom company. The dataset has as many as 40000 instances, for which labels are readily available. The balanced form of the dataset is provided after preprocessing and establishing the equal number of instances for both classes. Table 1 describes the characteristics of the used dataset. This dataset is used in this study to analyze the experimental results.

Table 1. The characteristics of the used Telecom dataset.

| | **Cell2Cell** |
|---|---|
| ***Total Instances*** | 40k |
| ***Total Features*** | 76 |
| ***Numerical Features*** | 68 |
| ***Nominal Features*** | 8 |
| ***Data Distribution*** | Balanced |
| ***Missing values*** | No |

### *B. Feature Reduction*

Feature reduction is performed in the preprocessing phase in order to provide the most meaningfull and discriminating features to the classifiers. Mostly classifier suffer when provided with irrelevant features space.We applied mRMR,Fisher's ratio and F-Score feature extraction methods to analyze the impact on prediction performance of used ensembles. For each feature reduction method, a linear search is performed to select the features which provide maximum discriminating information to the classifiers and hence produce better performance. A linear search is separately conducted for mRMR, F-Score and Fisher's Ratio with Random Forest. Only thirty four features linearly searched using mRMR produced the highest AUC of 0.742 for Random Forest as shown in Table 2. Random Forest demonstrates the improved performance with mRMR reduced features set compared to best searched features returned by F-Score and Fisher's Ratio. Rotation Forest encourages diversity by employing PCA and achieves higher accuracy by utilizing all principal components. A mRMR reduced set of only thirty five features attain 0.762 AUC for predicting churners in the telecommunication dataset as shown in Table 3.

Table 2. Effect of features extraction techniques on Random Forest.

| | **D*** | **Random Forest** | | |
|---|---|---|---|---|
| | | ***AUC*** | ***Sensitivity*** | ***Specificity*** |
| **mRMR** | **34** | **0.742** | 0.690 | 0.601 |
| **F-Score** | 37 | 0.716 | 0.666 | 0.646 |
| **Fisher's Ratio** | 37 | 0.713 | 0.656 | 0.647 |

D* represents the number of features giving best result

Table 3. Effect of features extraction techniques on Rotation Forest.

| | **D*** | **Rotation Forest** | | |
|---|---|---|---|---|
| | | ***AUC*** | ***Sensitivity*** | ***Specificity*** |
| **mRMR** | **35** | **0.762** | 0.721 | 0.583 |
| **F-Score** | 37 | 0.691 | 0.670 | 0.629 |
| **Fisher's Ratio** | 37 | 0.652 | 0.603 | 0.610 |

D* represents the number of features giving best result

Table 4. Effect of features extraction techniques on RotBoost.

| | **D*** | **RotBoost** | | |
|---|---|---|---|---|
| | | ***AUC*** | ***Sensitivity*** | ***Specificity*** |
| **mRMR** | **31** | **0.816** | 0.765 | 0.746 |
| **F-Score** | 36 | 0.726 | 0.679 | 0.627 |
| **Fisher's Ratio** | 37 | 0.724 | 0.675 | 0.629 |

D* represents the number of features giving best result

RotBoost is an efficient ensemble of Adaboost and Rotation Forest. The inclusion of boosting through Adaboost and higher diversity offered by Rotation forest make RotBoost a high performing classification approach. mRMR returns a coherent diversified set of features which have maximum discriminating information. mRMR adopts a broad criteria for features selection based on minimum redundancy and maximum relevance. Features having strong dependency with class labels are extracted. Ultimately a reduced feature set with only 31 features having maximum mutual information, calculated on the basis of entropy, is sorted out. This reduced features set extends an improved training level to RotBoost, which finally shows highest of the performance in predicting telecom churners compared to Random Forest and Rotation Forest as shown in Table 4.

### *C. mRMR vs Fisher's Ratio vs F-Score*

mRMR targets to select the features subset, by maximizing the mutual dissimilarity within the class and minimizing the marginal similarity with the class labels. mRMR based features selection produces highest AUC of 0.8161 with RotBoost using only 31 selected features as shown in Table 4. Such a significant improvement, in terms of AUC with fewer features, shows the effectiveness of selecting appropriate and most discriminating feature subset. Fisher's ratio and F-score based feature selections are based on ranking and weighting the individual feature. F-score and Fisher's ratio rank the features based on mutual information without considering relationships among features while mRMR is a different features selection method, which selects the features having strong correlation with class variable and mutually different from each other. This is the reason mRMR method substantially reduces the features set to 31 features and improves the AUC performance to the highest score of 0.816 as shown in Table 4.

### *D. mRMR – RotBoost based Churn Predictor*

The unique capabilities of mRMR which bifurcates the features space into most discriminating set of features, certainly

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on December 08,2025 at 05:27:36 UTC from IEEE Xplore. Restrictions apply.

provides a squeezed features space to RotBoost which in turn significantly improves the performance as shown in Figure 2. (RF and RotF represent Random Forest and Rotation Forest respectively in graph given in Fig.f) RotBoost carries its capabilities of attaining high diversity by constructing rotation matrix from Rotation Forest. Adaboost extends the weights updation capabilities to handle the hard instance iteratively, which ultimately helps RotBoost achieve higher accuracy in predicting churners in Telecommunication. Thus the unique collaboration of mRMR and RotBoost results into a high performing churn prediction model for Telecommunication.

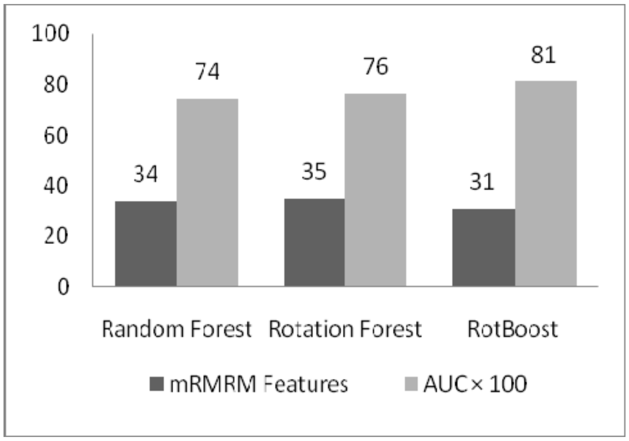

Figure2. Performance comparison of RF, RotF and RotBoost

## IV. CONCLUSION

mRMR and RotBoost appear as promising feature extraction and classification tools to model the challenging problem of churn prediction in Telecommunication. Specifically mRMR reduces the feature space to a mere set of 31 features which in turn provides better learning capabilities to RotBoost. Hence mRMR results as an efficient features reduction technique to mitigate the high dimensionality of the telecommunication data set compared to Fishers' Ratio and F-Score. mRMR not only reduces the features space but select the most discriminating features which eventually help RotBoost to attain highest prediction accuracy. RotBoost operates in an iterative manner thus small size feature space also lessens the computations involve in training and testing phases. Thus our proposed model uniquely employs mRMR and RotBoost that effectively handles the main hurdle of high dimensionality of telecommunication dataset and attains good prediction performance. Therefore, our proposed churn prediction model can be beneficial for predicting the churners in Telecommunication industry.

## ACKNOWLEDGEMENT

This work is supported by the Higher Education Commission of Pakistan (HEC) as per award no. 17-5-6(Ps6-002)/HEC/Sch/2010

## REFERENCES

[1] W.J. Reinartz and V. Kumar,"The impact of customer relationship characteristics on profitable lifetime duration," Journal of Marketing, vol.67(1),pp.77-99,2003.

[2] T. Lee, C. Chiu and Y.Chou, " Mining the customer credit using classification and regression tree and multivariate adaptive regression splines," Journal of Computational Statistics and Data Analysis, vol.50(4),pp.1113-1130,2004.

[3] D. Ruta, D. Nauck and B. Azvine, "K nearest sequence method and its application to churn prediction," Lecture Notes in Computer Sciences, vol.4224/2006, pp.207-215,DOI: 10.1007/11875581_25.

[4] A.Khan, M. F. Khan, and T. Choi," Proximity based GPCRs prediction in transform domain," Biochemical and Biophysical Research Communications, vol.371 (3), pp.411-415, April 2008.

[5] S. Tan," An effective refinement strategy for KNN text classifiers," Expert Syst. Appl., vol.(30),pp.290-298, 2006.

[6] Y. Zhang, J. Qi, H. Shu and J. Cao, "A hybrid KNN-LR classifier and its application in customer churn prediction," in IEEE Int. Proc. SMC.,pp.3265-3269, 2007.

[7] "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," IEEE Trans. Neural Net., vol.11(3),PP.1045-9227,2000.

[8] Y. Kim, "Toward a successful CRM: variable selection, sampling, and ensemble," Decision Support Systems, vol.41(2), pp. 542-553, January 2006.

[9] A.Lemmens and C. Croux," Bagging and boosting classification trees to predict churn," Journal of Marketing Research, vol.43(2),pp.276-286,2006.

[10] I.Bose and X.Chen, "Hybrid models using unsupervised clustering for prediction of customer churn," Journal of Organizational Computing and Electronic Commerce, vol.19(2),pp.133-151, 2009.

[11] T.G. Dietterich,"Ensemble methods in machine learning," Lecture Notes in Computer Science, vol.(1857),pp.100-115,2000.

[12] E.Bauer and R. Kohavi,"An empirical comparison of voting classification algorithms: bagging, boosting and variants," Machine Learning, vol.36(2),pp.105-139,doi: 10.1023/A:1007515423169.

[13] L.B.Statistics and L.Breiman,"Random forests," Machine Learning, vol.(45), pp.5-32,2001.

[14] Y. Xie, X.Li,E.W.T Ngai and W.Ying, "Customer churn prediction using improved balanced random forests," Expert Syst. Appl., vol.36(3/1), pp.5445-5449,April 2009.

[15] J.J. Rodriguez, L.I. Kuncheva and C.J. Alonso, "Rotation forest: a new classifier ensemble method," IEEE Trans. Pattern Anal. Mach. Intell., vol.28(10),pp.1619-630,2006.

[16] C.Zhang and J. Zhang, "RotBoost: a technique for combining rotation forest and adaboost ," Patter Recognition Letters,vol.29(10),pp.1524-1536,2008.

[17] K.W.D.Bock and D.V.D.Poel,"An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction," Expert Syst. Appl.,vol.38(10),pp.12293-12301,September 2011.

[18] http://www.fuqua.duke.edu/centers/ccrm/

[19] H. Peng, F. Long and C.Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell.,vol.27(8),pp.1226-1238, August 2005.

Authorized licensed use limited to: ULAKBIM UASL - Pamukkale University. Downloaded on December 08,2025 at 05:27:36 UTC from IEEE Xplore. Restrictions apply.