Customer churn prediction in the Banking industry using the hybrid models

Behdad Ehsani, Negar Aminpour

Abstract

Customer retention plays a critical role in all industries because it minimizes the time, money, and workforce required to acquire new clients. Hence, banks should take strategic steps to preserve their current clients from churning. A company's ability to retain existing customers is critical to its total revenue and prestige in a highly competitive environment. As a result, every firm employs essential measures to maintain existing clients through customer management strategies. Hence, Instead of using single classification models for customer churn prediction, this study deploys the unsupervised techniques and incorporates the obtained results into supervised learning to predict churners. Based on the evaluation of proposed hybrid methods, the model enhanced its accuracy, and it can detect more churners than the baseline method.

Keyword: Customer churn prediction – Hybrid model – Clustering – Banking industry

1. Introduction

Customers are a company's most valuable asset because they are the primary source of revenue in any business. As a result of this, it has been seen that organizations have started to devote more resources to attracting new consumers and keeping the ones they already have. The industries' primary focus has shifted to reducing customer attrition. Therefore, customer churn is a rampant problem in the financial services business.

Churn is the action of a consumer transferring banks, typically due to dissatisfaction with financial services. The advent of new banking services and products, e.g., mobile banking, internet banking, various types of loans with a low-interest rate, and credit points, intensifies the fierce rivalry in this industry. To combat churn, financial institutions must forecast which clients are inclined to leave based on their behaviors. Customer churn prediction is considered a binary classification task, differentiating the churners from non-churners. Notably, it is not plausible to retain all probable churners. If we identify privileged customers (high-level clients) and prevent them from churning, we can better the problem of leaving.

The research reviewed several strategies for churn prediction that have been published in the literature. In particular, for the financial industry, comprehending the dataset and its characteristics is critical prior to building a churn prediction model. So this article describes methods that combine supervised and unsupervised machine learning techniques (hybrid method) to define the customer groups and predict the churners in these groups. The clustering technique is utilized in the unsupervised section. Then clustering results are incorporated in supervised learning models, encompassing Linear Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and XGBoost (XGB). Furthermore, this study illustrates the efficiency of these combined models and selects the best-fitted model with the bank customer dataset.

The rest of this study is as follows. In section two, a comprehensive literature review is conducted to find the research gaps. Next, in section three, the preprocessing techniques are utilized for the bank customer dataset. Additionally, the terminology of supervised and unsupervised methods is described in section three. Section four presents the proposed methodology and evaluation of these models, conducted by performance measures.

2. Literature survey

Many data mining approaches are regularly employed in credit card churn prediction. Here is a brief summary of the methods that have been developed so far. Kumar and Ravi (2008) deployed machine learning techniques to predict churners. In this study, Multilayer Perceptron, Logistic Regression, decision trees, Random Forest, Radial Basis Function Network, and Support Vector Machine are utilized. The final decision was conducted based on majority voting techniques. Furthermore, the under-sampling, over-sampling, and SMOTE method was hired in this work to cope with imbalance. Bose and Chen (2009) used the hybrid model for the telecommunication industry. This study proposed two methods for hybridization using the various clustering techniques, including hierarchical clustering, K-means, fuzzy c-means, and self-organizing map, which are evaluated by top decile lift. Additionally, only the decision tree algorithm is incorporated with them.

Lu et al. (2014) evaluated the model based on boosting method, Adaboost. This method outperformed other models based on ROC curve results. Rajamohamed and Manokaran (2018) proposed the hybrid model for credit card churn prediction. In this study, SVM, NB, RF, KNN, and RF are combined with the k-means clustering method. Jain et al. (2021) surveyed Customer Churn Prediction (CCP), CCP types, the reason for CCP in the telecommunication industry. Performance measures and all supervised learning methods used in CCP are reviewed with pros and cons. Also, this study concentrated on the Deep Learning method for the prediction of churners.

Regarding the meticulous analysis shown in Table 1, this study proposes two hybrid models for bank customer churns prediction based on the mixture of K-means and four classification models, including LR, SVM, RF (bagging method), and XGB (boosting strategy). The preprocessing phase includes balancing the dataset. Also, the validation set is utilized for model selection and hyper-parameter tuning.

Table 1: A comparison of relevant topic

Author	***Dataset***	***SL Method***	***Hybrid***	***Pre-processing***	***Validation***
Kumar and Ravi (2008)	Credit card churn prediction	MP, LR, DT, RF, RBF, and SVM	-	+	-
Bose and Chen (2009)	Telecommunication Industry	DT	+	+	-
Lu et al. (2014)	Telecommunication Industry	AdaBoost, LR	-	+	-
Rajamohamed and Manokaran (2018)	Credit card churn prediction	KNN, DT, RF, SVM, NB	+	+	-
Jain et al. (2021)	Telecommunication Industry	CNN, XGB, SVM, NB, LR, NN, DT	-	-	-
This study	Bank customer churn prediction	SVM, LR, XGB, RF	+	+	+

3. Terminology

In this section, the data structure, as well as data preprocessing, are explained, and the performance measure and summary of supervised and unsupervised methods utilized in this study are elaborated.

3.1. Data

The dataset contains customer churn prediction in the banking industry with 10000 observations. It is characterized as 14 input variables and one dependent variable, and it consists of 7963 churners and 2037 non-churners members. The mentioned dataset includes customer information for predicting which customers are likely to leave the bank for the next period. The structure of mentioned dataset are presented in Table 2.

Table 2: Structure of bank client's dataset

No	***Variable***	***Level***	***Nominal values***
1	RowNumber	ID	1 2 3 4 5 6 7 8 9 10 ...
2	CustomerId	integer	1 2 3 4 5 6 7 8 9 10 ...
3	Surname	character	"Hargrave" "Hill" "Onio" "Boni" ...
4	CreditScore	integer	between 350-850
5	Geography	character	"France" "Spain" "Germany"
6	Gender	character	"Female" "male"
7	Age	integer	between 18-92
8	Tenure	integer	between 0-10
9	Balance	numeric	between 0-250898
10	NumOfProducts	integer	between 1-4
11	HasCrCard	integer	0, 1
12	IsActiveMember	integer	0, 1
13	EstimatedSalary	numeric	between 11.58-199992.48
14	Exited	integer	0, 1

3.2. Preprocessing:

In this research, the dataset does not have any missing values; however, it has some categorical variables, so we encoded all categorical variables. Then we standardized the variables by using min-max normalization.

Since the data set used in this research was imbalanced, we used Synthetic Minority Oversampling Technique (SMOTE) algorithm to balance our data. SMOTE is one of the most commonly used algorithms for imbalanced data that oversamples the minority class by creating artificial data. (Amin et al., 2016)

3.3. K-means clustering algorithm

Clustering is an unsupervised algorithm. The primary goal of clustering is to group the data that are similar to each other in each cluster. K-means is one of the famous clustering algorithms, which minimize the sum of squared distance in each cluster. In this case, the algorithm aims to minimize a squared error function.

$$\sum_{j=1}^{k}\sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

However, determining the number of clusters is the challenging part, and in this research, we use the hold-out procedure to find K. The algorithm steps are as follows (Kodinariya & Makwana, 2013).

I. At first, it randomly selects k points, which will be the initial centroids for k groups.
II. It assigns each data to the group with the closest distance to the cluster centroids.
III. It computes the centroids of the new clusters, which will be the mean value of all the data in each cluster.
IV. It repeats this process until there is no change for each cluster data.

3.4. Classification algorithm

A classification is a tool for prediction and decision-making. It can be used on any dataset to extract some knowledge; it can work with audio, image, text, and numeric data sets and build a predictive model. A class label identifies the nature of the mining process. It can be binary or multilevel. In this research, we use four classifiers.

3.4.1. SVM

Support Vector Machine algorithm is used for classification or regression problems. It tries to find a hyperplane to separate the data into classes by considering maximum margins, which means maximizing the distance between the hyperplane and the nearest data on each class. (Rajamohamed & Manokaran, 2018)

3.4.2. Logistic Regression

Logistic regression works based on a mathematical approach to analyzing the relationship between existing independent variables to predict the occurrence probability of an event. In this research, the aim is to predict the occurrence probability of customer churn. The equations below can explain it (Hassouna et al., 2016):

$$(y = 1 \,|x_1 \,, \ldots, x_n) = f(y)$$

$$\text{Y} = \ \beta_0 + \beta_1 x_1 + \cdots + +\beta_n x_n$$

Where y is the target variable, which is binary (0, 1), x is the predictors (variables) for each customer, and β is the weight for the variable x_i for each customer.

3.4.3. Random Forest

The Random Forest algorithm builds decision trees and takes their majority vote for classification. One of the central features of the Random Forest Algorithm is that it can classify data that contains categorical variables. (Hassouna et al., 2016)

3.4.4. XGBoost

XGBoost implements gradient boosted decision trees for speed and performance. One of the main advantages of using this algorithm is that XGBoost is fast.

Boosting is a technique where new models are added sequentially to fix the errors made by existing models until no further improvements can be made. Gradient boosting is a technique where new models create that predict the errors of prior models, and then those new models are combined to make the final prediction. They call it gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

3.5. Performance measures:

As a result of this work, the following performance measures were used for model selection, hyper-parameter tuning, and model evaluation (Verbeke et al., 2012):

	Churn	NonChurn
Churn	TP	FN
NonChurn	FP	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. The proposed algorithm

This section aims to point out the Baseline model, as well as two proposed hybrid models. Moreover, the results obtained by the validation and test set are proposed in the following.

4.1. Experiment 1

Using a hold-out procedure, the provided data set is partitioned into the train, validation, and test data sets. The system's efficiency is evaluated using precision, sensitivity, accuracy, and misclassification error. The first experiment is called a baseline model utilized for evaluating the efficacy of the proposed hybrid models, which is illustrated in Figure 1.

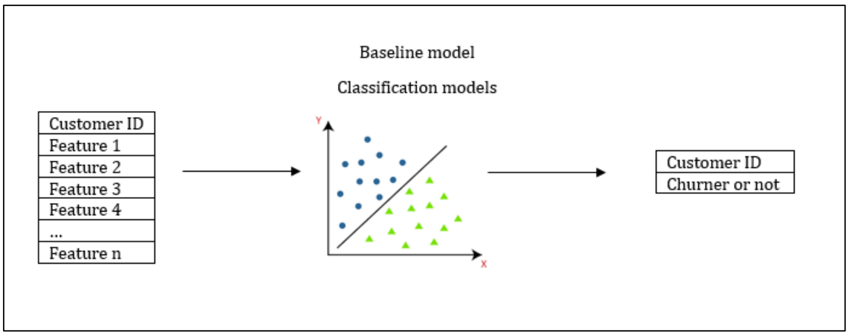

Figure 1: The scheme baseline model

The performance of four critical classifiers investigated in this work is depicted in Figure 2 and Table 3. After hyper-parameter tuning, RF performed better than the other three single classifier models, with a maximum accuracy of 83.5% based on test set results. The performance result of the test set is presented in Table 4.

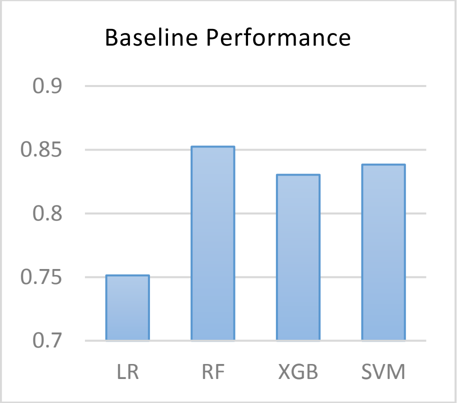

Figure 2: The performance of baseline model (based on validation set)

Table 3: The performance measure based on validation set

Models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7513	0.8524	0.8303	0.8383
Precision	0.3894	0.6366	0.5783	0.5867
Recall	0.4214	0.616	0.5711	0.6584

Table 4: the performance measure of selected model (based on test set)

Models	*RF*
Accuracy	0.8350
Precision	0.5880
Recall	0.6650

4.2. Experiment 2

The hybrid model is developed in this study by merging the k-means clustering approach with four primary classifiers. Initially, the whole dataset is split into train, validation, and test dataset. Figure 3 illustrates the procedure of data splitting. First, the k-means method is implemented on the train set. Then, the data points in the validation and test set are mapped to train set groups. The cluster labels are incorporated into the dataset as a predictor in the initial hybrid method. The comprehensive illustration of the hybrid model is depicted in Figure 4. The new feature is considered to classify the bank customers well. The principal classifiers hybridized with k-means include SVM, LR, RF, and XGB.

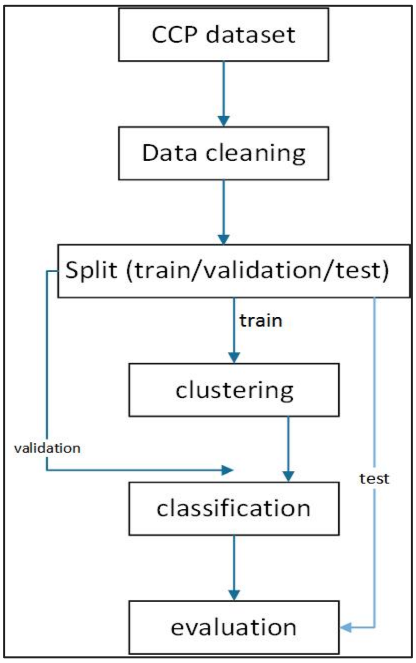

Figure 3: The hold-out procedure of data splitting

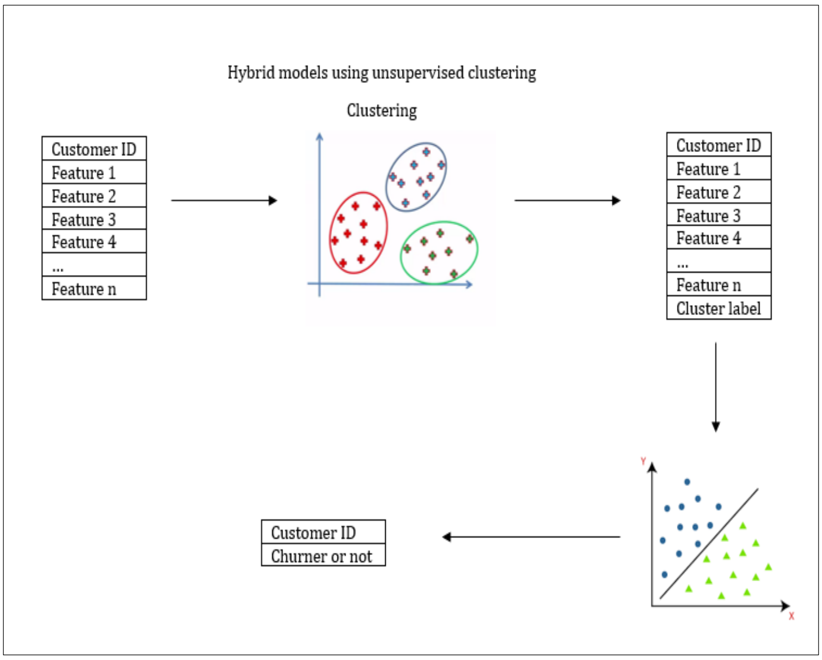

Figure 4: Illustration of first proposed hybrid method

Regarding the hybridization, for the hyper-parameter tuning of k-means, all K should be considered with the combination of models. Afterward, the performance measures of different Ks are considered together, shown in Table A1-A4, and the best K with the accompaniment of the best model is selected at the end. The best *K* is *three*, depicted in Table 5. Notably, the RF combined with k-means outperforms others, and this comparison is shown in Figure 5. Finally, the selected model is evaluated by the test set. The test set results are presented in Table 6.

Table 5: Performance measures of the first hybrid model (K=3)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7513	0.8529	0.8293	0.8383
Precision	0.3894	0.6354	0.5789	0.5874
Recall	0.4214	0.6259	0.5486	0.6534

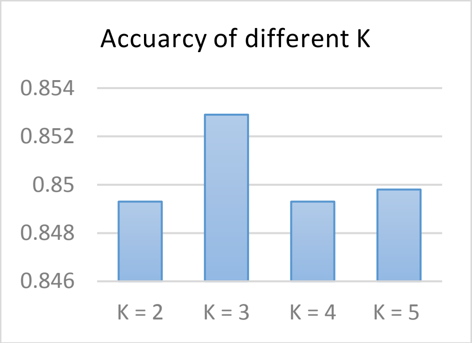

Figure 5: Comparison between different K for the best-selected model (RF)

Table 6: The performance measure of the first hybrid model (based on the test set)

Models	*RF & K=3*
Accuracy	0.8300
Precision	0.5738
Recall	0.6796

The discrepancy between test set and validations set results shows that the overfitting happened in RF. Note that ensemble techniques, including bagging and boosting, are probable for overfitting the model. Additionally, adding one attribute (feature) into the dataset can make the model complex and increase the model capacity. Based on the Bias-Variance trade-off, the generalization error rises once the model capacity increases. Therefore, this first hybrid method may worsen the performance measures in many cases. However, it depends on the dataset because some models require more complexity to reach the optimal points in the Bias-Variance figure.

4.3. Experiment 3 and results

The second hybrid method is proposed based on implementing the classification in each cluster of similar clients. The procedure of data splitting is akin to the previous method. Regarding this method, first, the k-means aims to group clients. Then, the mentioned classification methods are utilized in each cluster to enhance the model's accuracy. The scheme of the second hybrid method is presented in Figure 6.

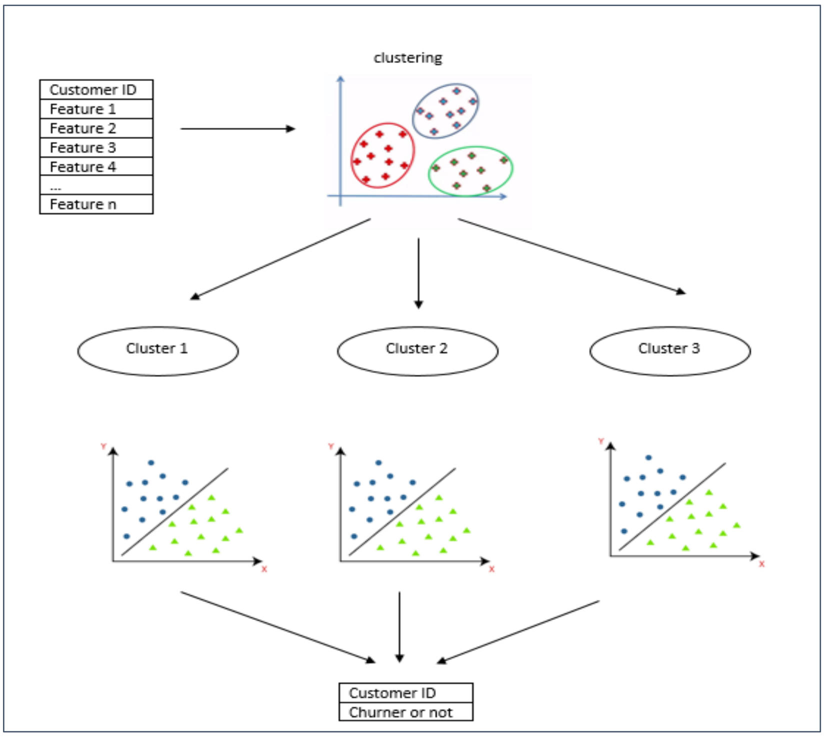

Figure 6: the second proposed hybrid algorithm

Similar to the previous method, the cluster hyper-parameter tuning is based on evaluating the various K. the performance measure of different *K*s is shown in Table A5-A8. The best K with the best-selected model is presented in Table 7. Overall, the RF outperforms other methods, and the best *K* equals *five* in the second version, shown in Figure 6. Next, the selected model is evaluated by the unobservable dataset. The obtained result from the test set is shown in Table 8.

Table 7: Performance measures of the first hybrid model (K=5)

Models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7618	0.8418	0.8278	0.8303
Precision	0.2968	0.5860	0.5586	0.6110
Recall	0.3802	0.6104	0.5729	0.5724

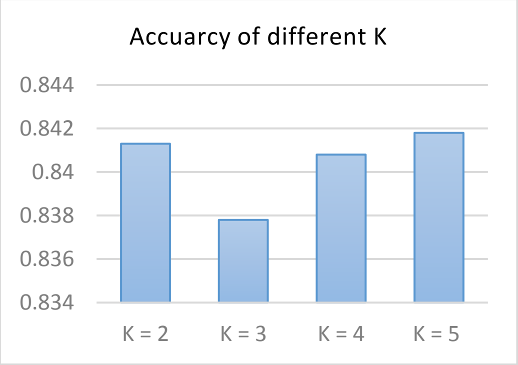

Figure 6: Comparison between different K for the best-selected model (RF)

Table 8: The performance measure of the second hybrid model (based on the test set)

models	*RF & K=5*
accuracy	0.8440
precision	0.6699
recall	0.6106

Based on the test set result of the second hybrid model, the accuracy of the model increases 1% approximately in comparison with the baseline model. The second hybrid model aims to group customers and then uses the classification models. Therefore, the size of the train set decreases noticeably because it is split into smaller datasets. The division into small datasets leads to change in the dataset distribution and may increase variance and generalization error because it learns the model very well. It should be noted that the dataset's bias does not decrease when the test set size decreases. In conclusion, both proposed hybrid models may increase the model's variance and capacity. There is no definite solution for selecting one experiment for all datasets in this case. For instance, in this study, the Baseline model has an overfitting problem, and the first hybrid model worsened this problem due to adding one feature. Still, the second method remedied the result because models in small datasets can learn their distributions perfectly. However, model complexity plays an important role in hybrid models' accuracy, which ultimately depends on the dataset and its distribution. Regarding Table 5 and 7, in this case, the second hybrid model enhances its complexity appropriately, whereas the first method increases it more than enough. The test sets' results prove the previous claim, shown in Tables 8 and 6.

5. Conclusion

In this research, we proposed two different hybrid models to predict churners, which combined supervised and unsupervised techniques. The techniques include k-means, Linear Regression, Support Vector Machine, Random Forest, and XGBoost. At first, in the preprocessing, the variables were standardized using the min-max normalization. And Synthetic Minority Oversampling Technique (SMOTE) was used to balance our data. The related features for ranking bank customers are selected in the clustering stage, and groups are categorized based on them. The results show that the first method worsened the performance, but the second method efficiently predicted churners, and Random Forest was the best among the supervised techniques. Overall, the accuracy of the two proposed methods depends on the dataset and its distribution. For future works, soft clustering techniques, e.g., Gaussian mixture clustering and fuzzy clustering, can be used due to their robustness to outliers.

References

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J& Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access*, *4*, 7940-7957.

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

Hassouna, M., Tarhini, A., Elyas, T., & AbouTrab, M. S. (2016). Customer churn in mobile markets a comparison of techniques. arXiv preprint arXiv:1607.07792.

Jain, Hemlata & Khunteta, Ajay & Srivastava, Sumit. (2021). Telecom churn prediction and used techniques, datasets and performance measures: a review. Telecommunication Systems. 76. 10.1007/s11235-020-00727-0.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Kumar, Dudyala & Ravi, Vadlamani. (2008). 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.

Lu, N., Lin, H., Lu, J., & Zhang, G. (2014). A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, *10*(2), 1659-1665.

Rajamohamed, R., & Manokaran, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. *Cluster Computing*, *21*(1), 65-77.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach. European journal of operational research, 218(1), 211-229.

Appendix

Table A1: Performance measure of the first hybrid model (for k=2)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7497	0.8493	0.8343	0.8408
Precision	0.3857	0.6289	0.5956	0.5928
Recall	0.4165	0.6085	0.5436	0.6608

Table A2: Performance measure of the first hybrid model (for k=3)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7513	0.8529	0.8293	0.8383
Precision	0.3894	0.6354	0.5789	0.5874
Recall	0.4214	0.6259	0.5486	0.6534

Table A1: Performance measure of the first hybrid model (for k=4)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7492	0.8493	0.8328	0.8378
Precision	0.3858	0.6289	0.5884	0.5873
Recall	0.4214	0.6085	0.5561	0.6459

Table A4: Performance measure of the first hybrid model (for k=5)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7508	0.8498	0.8388	0.8328
Precision	0.3875	0.6272	0.6076	0.5724
Recall	0.4165	0.6209	0.5561	0.6608

Table A5: Performance measure of the second hybrid model (for k=2)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7528	0.8413	0.8348	0.8303

Precision	0.4264	0.5611	0.5586	0.6384
Recall	0.3931	0.6148	0.5942	0.5689

Table A6: Performance measure of the second hybrid model (for k=3)

models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7497	0.8378	0.8368	0.8328
Precision	0.4364	0.5885	0.5835	0.6409
Recall	0.3898	0.5975	0.5954	0.5749

Table A7: Performance measure of the second hybrid model (for k=4)

Models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7623	0.8408	0.8268	0.8308
Precision	0.4389	0.5711	0.5511	0.6409
Recall	0.4131	0.6107	0.5711	0.5698

Table A8: Performance measure of the second hybrid model (for k=5)

Models	*LR*	*RF*	*XGB*	*SVM*
Accuracy	0.7618	0.8418	0.8278	0.8303
Precision	0.2968	0.5860	0.5586	0.611
Recall	0.3802	0.6104	0.5729	0.5724