Customer churn prediction from Internet banking transactions data using an ensemble meta-classifier algorithm

Fatemeh Ehsani (ehsani@email.kntu.ac.ir)
K.N.Toosi University of Technology

Research Article

Keywords: customer churn, Internet banking, transactions data, meta-classifier

Posted Date: May 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1630808/v1

Customer churn prediction from Internet banking transactions data using an ensemble meta-classifier algorithm

Abstract

With the advancement of electronic markets platforms, customers show different purchase behaviors. Since they have a wide range of choices and low exit barriers, customer movement from one digital store to another has become a natural problem for every business. Customers are the most valuable property of every enterprise. Thus, Suppliers have to tie a strong relationship with their customers to prevent their leaving by providing the most appropriate produces and services based on their desires and trying to satisfy them. Customer churn prediction (CCP) is one of the customer relationship management (CRM) strategies to estimate the probability of their abandonment. Marketers use CCP to attract visitors, engage them in websites activities, convert them into customers, and retain them for a long time. Internet banking transactions dataset is a dependable resource to analyze customer interactions and their churn behaviors. Transactions data scrutinizes the customer's specifications and their payment details. In this paper, we introduced a meta-classifier algorithm to predict customer churn behavior according to the transactions data. We applied the four most used supervised classification algorithms. Then, we improved their performance by hyperparameters and tuning. We also performed RFECV feature selection to extract and rank the most critical variables. The experimental results represent that combining various machine learning developed algorithms in a funnel of meta-classifier can extract the highest prediction accuracy.

Keywords: customer churn, Internet banking, transactions data, meta-classifier

1. Introduction

The emergence of electronic commerce has developed massive alterations in customer behavior. Marketers have created various electronic retailing websites to present their products and services and find their target customers (Grubor & Jakša, 2018). Since users access too many options in electronic retailing websites presented, their choice and substitution power have increased. Also, they are exposed to large volumes of information, preparing to improve their knowledge of products and services (Isa & Nayan, 2020). Almost all electronic commerce organizations adopt customer relationship management (CRM) programs to communicate with visitors and turn them into customers. They show great perseverance to provide incentives and make strong connections to bind long-term relationships with their customer. CRM provides different facilities to serve the customers perfectly and influence their lifetime value (Anshari, Almunawar, Lim, & Al-Mudimigh, 2019). Customers' demands may alter with time variation. Thus, CRM programs encompass different understandable sets of business applications to improve the efficiency and effectiveness in managing their demands. The relationships and connections in CRM activities are perceivable by customer's lifetime or their life cycle. CRM has the purpose of ensuring customers about the enterprises' products and services. CRM brings enjoyment and satisfaction for firms and customers through giving more information about customer's demands and their behaviors (Mahajan & Gangwar, 2017). Another goal of CRM is to build, manage, and strengthen close and perpetual connections with customers. Many researchers have extensively used it in various sectors of e-commerce, such as e-retailing, e-banking, insurance, and telecommunications (Asthana, 2018). CRM always focuses on acknowledged customers who are the most profitable resource of the firms' dataset. This dataset contains the customers' behavior characteristics to evaluate their potential values and predict their following desires. Companies' income usually depends on the profits gaining from the customer purchase

decisions. Thus, customers are the most valuable asset of every business (Tsai & Lu, 2009). The CRM process tries to create perceptions by recognizing customers and prepare some motivations to retain them. Customer retention happens when the existing organization fulfils customers' demands (Mahajan & Gangwar, 2017). One of the main tasks of CRM for enterprises is customer churn analysis, which is necessary to maintain valuable customers in their electronic markets (Tsai & Lu, 2009). In this paper, we describe churn prediction to distinguish less attracted customers using the ensemble of classifiers approach. The concept is to identify the rate of customers transferring from a company to the competitors and persuade them to sustain. To address this issue, we decided to use the electronic banking transactions data to forecast customer abandonment. We presented a meta-classifier consisting of four more practical algorithms to show the highest prediction accuracy.

2. Customer Churn Prediction

The customer churns analysis is described as the research performed on the propensity of a customer's movement from an enterprise to another or perhaps their product (or service) abandonment (H. Jain, Khunteta, & Srivastava, 2020). It happened when consumers left the firm due to dissatisfaction or rivalry situations, like new products substitution. Even reducing the number of purchases by customers might lead to this problem for the businesses. (Çelik & Osmanoglu, 2019). Customers who decide to stop purchasing in the e-marketing stores separate into intentional and unintentional. Intentional churners refer to customers who decide to cease their relationship with the marketers. Unintentional churning happens when the firm or server withdraws the customers for reasons, such as delay in payment, non-payment of products, or misusing the firms' data and information (Berry & Linoff, 2004; Tsai & Chen, 2010). A resolutions to tackle this problem is predicting the customers who are prone to leave. Customer churn prediction (CCP) helps marketers adopt practical strategies to catch new customers and retain the existing ones. CCP might perform in reactive and proactive manners. In a reactive, marketers wait until the customers cancel requests. Then, they suggest exciting plans to attract and retain customers. While, in a proactive, marketers predict the probability of customer abandonment. Then, they offer their plans based on this prediction. Experts have separated leaving customers (churners) from loyal ones (non-churners) in a binary classification problem to resolve this issue (Lalwani, Mishra, Chadha, & Sethi, 2021). Customer churn prediction is incredibly considerable in the estimations of revenue models of businesses, such as electronic banking, insurance, or retailing (Çelik & Osmanoglu, 2019). Every business tries to acquire new customers, up-sell and cross-sell to them, and increase their retention time. Maintaining current customers is more economical (Lalwani et al., 2021) because, in today's saturated markets and intensive competition environment, new customers' acquisition is sometimes more than twenty times costlier for enterprises than keeping the previous ones. Also, the value proposition of the organizations is straightly proportionate to the number of active customers who commit transactions and perform buying several times in a given time. The profitability of satisfying and loyal customers is superior and depends on some parameters like the expenses, investment capacity, profitability, cash flow, and size of the firms. (Çelik & Osmanoglu, 2019). Concentrating on customer churn prediction implicates some significant financial and managerial implications. Firstly, enterprises can better render services to their existing customer through making and keeping communications, rather than spending time, effort, money, and energy to attain fresh ones, who sometimes distinguish with higher attrition rates (Richards & Jones, 2008; Zhao, Shi, Lee, Kim, & Lee, 2014). Secondly, satisfied customers expend more time and money on enterprises, purchase more products from them, advertise the products and services to others via affirmative word of mouth, and are more economical to render services (Nyilasy, 2007; Royo-Vela &

Casamassima, 2011). Thirdly, loyal customers are less price-sensitive and less exposed to the advertisement of other competitor markets (Santouridis & Trivellas, 2010). Finally, the number of customers' reduction may lead to irreparable financial expenses and decrease the up-selling and cross-selling possibility of the goods. On the other hand, this reduction might increase the essential costs of attracting and engaging new visitors to convert them into actual customers (Shaheen & Naseem, 2015).

3. Related Works

Burez & Van den Poel, 2007 indicated reactive and proactive processes for customer churn management. In reactive adoption, the enterprise recommends some interesting choices to attract and engage the customers after they cancel their requests. While, in proactive adoption, the enterprise tries to detect customers with a high likelihood of churning. Then, it persuades them to stay by offering some particular intensives. They used the Markov chain and random forests to make a churn prediction plan for the European Pay-TV market. Also, they acknowledged that customers who do not repeat purchasing separate into commercial churners and financial churners. First, those who avoid renewing their temporary contract after its deadline. Second, those who do not pay within their legally committed subscriptions (Burez & Van den Poel, 2008). Creating an efficient practicable customer churn prediction algorithm has become a significant subject in academics and business in the past few years. Coussement & Van den Poel, 2008 applied support vector machines (SVM), random forests, and logistic regression to build a predictive churn model. Tsai & Lu, 2009 considered to improve a hybrid machine learning algorithm by combining classification and clustering approaches. They also filtered out unrepresentative training data during preprocessing and data reduction for normalization and feature engineering. Most of the customer churn datasets are imbalanced, resulting in low accuracies. Xie, Li, Ngai, & Ying, 2009 designed a balanced random forest algorithm to solve this problem and improve the accuracy rate. Wei-Yun, 2012 also used this model to predict customer churn in a commercial bank. Qiu & Li, 2010 presented random forests along with Proximity Matrix to extract useful features and prevent the curse of dimensionality problem. Nie, Rowe, Zhang, Tian, & Shi, 2011 introduced a churn prediction method combining decision trees and logistic regression. In their study, the influence of customers, credit cards, risks, and transactions information have computed. Guelman, Guillén, & Pérez-Marín, 2012 advocated random forests as an applicable method to uplift the customer retention model in the insurance company. They compared their model with SVM and achieved a higher accuracy rate. Researchers have widely used different types of decision trees in customer churn prediction. Kirui, Hong, Cheruiyot, & Kirui, 2013 mentioned C4.5 decision tree in the mobile telecommunication industry. They also compared this model with Naive Bayes and Bayesian Network to illustrate its highest precision rate. Coussement & De Bock, 2013 cited CART decision tree and random forests in some electronic gambling stores. In this plan, they used K-fold as a CV metric to improve the performance. Almana, Aksoy, & Alzahrani, 2014 acclaimed C5.0 and CART decision trees have outperformed neural networks and genetic algorithms in the telecommunication industry. Verbeke, Martens, & Baesens, 2014 described the Alternating Decision Tree to incorporate in social network effects. Saini, Monika, & Garg, 2017 explained CHAID and CART decision trees in the telecommunication to keep satisfied and loyal customers. Ding, Liu, & Li, 2015 constructed an expanded random forest to improve telecommunication churn rate. They divided its nodes to generate each tree. A. Jain, Menon, & Chandra, 2015 utilized an extreme gradient boosting (XGBoost) algorithm to strengthen the prediction rate of sales and forecast customer churn in the retailing chains. They uncovered new hidden patterns in their retailing dataset to improve the robustness of extracted features. Dalvi, Khandge, Deomore, Bankar, & Kanade, 2016 approved that

logistic regression and decision trees are the two most prominent classifiers in churn forecasting in the telecommunication industry. X. Wu & Meng, 2016 tried to enhance the accuracy of retention prediction by oversampling their B2C e-commerce dataset with SMOTE approach and then used the AdaBoost algorithm. Mahajan & Gangwar, 2017; Maheswari & Priya, 2017 asserted SVM as one of the most powerful classifiers applying sales and inventory datasets to forecast customers behavior in the e-retailing markets. Yanfang & Chen, 2017 proposed EBURM model, which is based on logistic regression to predict churner's behavior with a high confidence level. They measured the duration that visitors spent in the e-market, the number of their registration and logins, their interests, and other elements that impact churning. Amin et al., 2017 suggested rough set theory (RST) composed of Covering, Genetic, Exhaustive, and the LEM2 algorithms as the four rule-generation methods in the telecommunication sector to predict customers' retention rate. De Caigny, Coussement, & De Bock, 2018 composed a strong and comprehensible supervised model based on decision trees classifier and logistic regression to predict churners and non-churners. Bharadwaj et al., 2018 experimented with logistic regression and multilayer perceptron (MLP) in mobile networks to make a sustainable grid for e-marketing customization. L. Wu & Li, 2018 exploited the Conjugate Gradient (CG) model to categorize churners and non-churners and then integrated with Logistic regression to simulate a conductive analysis. O'Brien & Ishwaran, 2019 boosted random forests with quantile classifiers for multiclass imbalanced data to optimize TPR and NPR rates in churn prediction Shirazi & Mohammadi, 2019 integrated the structured archival big data of a Canadian bank with unstructured data of web pages, such as page views to create a predictive churn plan. Amin et al., 2019 provided better insights for customer retention in telecommunication with estimating the correlation between classifier's certainty and the distance element. Stripling, vanden Broucke, Antonio, Baesens, & Snoeck, 2018 engendered EMPC profitable churn classification model to maximize simulates genetic algorithm along with lasso-regularized logistic regression. They also applied feature selection to exclude less effectual variables. XGBoost is a strong and comprehensive predictor in the customer churn prediction (AL-Shatnwai & Altibbi, 2020; Hanif, 2020; Senthan, Rathnayaka, Kuhaneswaran, & Kumara, 2021). P. Tang, 2020 mixed XGBoost with K-means clustering algorithm to improve churn prediction in the telecommunication field. Q. Tang, Xia, Zhang, & Long, 2020 executed XGBoost to extract the number of customers according to numerical attributes. Then, they combined XGBoost with MLP and incorporated these two classifiers with discrete attributes.(H. Jain et al., 2020) carried out Logistic regression and Logit Boost in the WEKA environment to predict churn risk. Höppner, Stripling, Baesens, vanden Broucke, & Verdonck, 2020 improved the decision trees structure to maximize the accuracy of customer retention and distinguish the most profitable would-be churners. RB, 2021 invented Fine-tuned XGBoost algorithm using the imbalanced dataset to address the importance of retaining customers in the telecommunication sector. Vo, Liu, Li, & Xu, 2021 perused the spoken contents of more than two hundred thousand customers from two million call conversation logs to predict churn risk utilizing the unstructured data in phone connection.(Li et al., 2021)confirmed the significance of churn prediction in broadcast service. They conducted an extensive experiment on customers' preferences, consumption patterns, payment habits, and watching intensity. Pustokhina et al., 2021 designed a dynamic model to predict customer churn. They inspired SFO to tune the given hyperparameters, CPIO-FS to optimize feature selection, and CCPBI-TAMO as a metaheuristic text classification algorithm. As mentioned here, most of the previous researches has been performed on the telecommunication dataset, but we determined to conduct our research on the transactions data of a bank committed through the Internet. We believe that transactions data is a reliable resource to perform our experiment.

4. Internet Banking Transactions Data

Technological development and marketing globalization has resulted in a great revolution in the banking industry. This revolution has provided the Internet as the most practical and economical channel for every marketing enterprise to offer products and services according to the customer's desires (Firdous & Farooqi, 2017; Harahap, Hurriyati, & Amanah, 2020). Internet Banking demonstrates the banks' websites as the virtual gateways enabling subscribers to enter in. These subscribers can access information about the products, services, regulations, and payments using any intelligent electronic device. They can create accounts and log in to their profile to conduct a range of financial transactions compatible with their demands without any time and geographical constraints (Sarker, Podder, & Alam, 2020). Marketers can understand their customer's behavior when they analyze their transactions data. Transactions data refers to the information of transactions committed by the customers when they purchase products or services. They record customers' activities whenever they access their accounts and save these activities in the banks' databases. The transactions dataset consists of customer's characteristics, their payment information, their loan and task information, and all sequences of events that happened by customers (Hand, 2018). Here, we used a transactions dataset of a bank, which includes the information of churners and non-churners.

5. Dataset Description and Preparation

The dataset of this research encompasses 10000 unique customers and 14 attributes related to the customers of a bank. The samples are gathered from the bank's information and took from Kaggle. "Row Number" is the first column and a nondeterministic analytic function allocates a single number to each customer in the dataset. "Customer Id" assigns a unique identification value to each customer to prevent duplications. When the customers decide to register on the banks' website and participate, they select a "Surname" for themselves. This feature may use when they want to create an account and log in. These three mentioned features is only for customer identification, and they do not affect customer's attitudes analyses and their decision to leave the bank. According to the Correlation Matrix, they also have high correlations with each other. Thus, we determined to eliminate them. After dropping ineffective features, there are eight categorical and two continuous indicators that remained in Table 1.

Table 1: Features used in churn prediction dataset

Variable	Data Type	Min	Max	Mean	Standard Deviation
CreditScore	int64	350.000000	850.000000	650.528800	96.653299
Geography	object				
Gender	object				
Age	int64	18.000000	92.000000	38.921800	10.487806
Tenure	int64	0.000000	10.000000	5.012800	2.892174
Balance	float64	0.000000	250898.090000	76485.889288	62397.405202
NumOfProducts	int64	1.000000	4.000000	1.530200	0.581654
HasCrCard	int64	0.00000	1.00000	0.70550	0.45584
IsActiveMember	int64	0.000000	1.000000	0.515100	0.499797
EstimatedSalary	float64	11.580000	199992.480000	100090.239881	57510.492818
Exited	int64	0.000000	1.000000	0.203700	0.402769

"Credit Score" is a numerical attribute that indicates a consumer's creditworthiness based on their credit history. Credit history is related to the number of customer's open accounts, types of loans, overall levels

of debt, and their payment profiles. Lenders apply this feature to estimate the possibility of borrowers' loans repayments just in the predetermined time. Lenders earmark an integer number between 300 and 850 to each customer. They believe that a score above 700 refers to good promise borrowers, who may receive a lower interest rate (Julia Kagan, 2021). Those customers with higher credit scores are more likely to continue receiving the banks' services. "Geography" refers to the country of the customers participating in the bank. They come from France, Spain, and Germany. "Gender" shows whether the user is male or female. "Age" denotes how old they are. On most occasions, younger users are more likely to stay in the bank than older ones. "Tenure" infers the number of years that a client remains subscribed to the bank. Usually, younger clients are less loyal and more likely to churn. "Balance" comprises all deposits and withdrawals in a bank account to compute the whole amount of money. It is a perfect variable in churn prediction, as customers with a lower balance rate in their accounts are more likely to leave. "NumOf Products" detects the number of goods that customers have bought via the bank. "Has Cr Card" declares if or not a client possesses a credit card. This indicator is important because a customer with a credit card is more engaged and attached to the bank. "Is Active Member" denotes active customers who do shopping through the bank several times in a given time. The bank should keep them because they are loyal. "Estimated Salary" notifies the approximate amount of salary each customer earns. Like "balance," clients with higher salaries are prone to commit more transactions than those with lower ones. The customer churn prediction proposal is set as a binary classification subject matter, determining the "Exited" as a target variable to conclude the accuracy of the mentioned algorithms. "Exited" is the class label. As illustrated in Figure 1, 7963 customers belong to the positive class. It means that they did not abandon the bank and its services. The rest of them (2037) have been placed in a negative class because they decided to leave the bank. Because of this situation, our binary classification dataset seems to be imbalanced. To overcome this problem, we perform oversampling. We divided the dataset into two different data frames and converted categorical indicators into dummies. Then, we combined categorical and continuous variables. After that, we used the "Yeo-Johnson" and "RobustScaler" approaches to perform data normalization and data transformation. Fortunately, our dataset has been preprocessed before. So we did not need the phase of data cleaning.

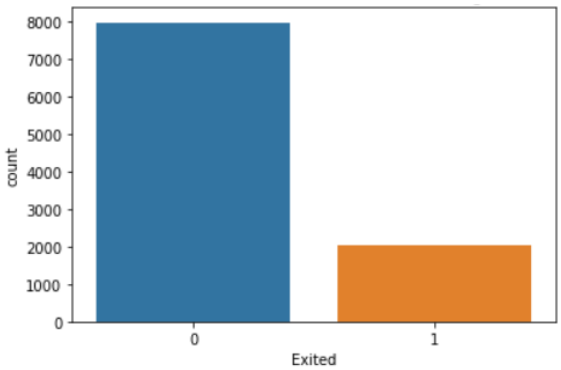

Figure 1: Count of customers based on churning

6. Methodology

Customer churn prediction (CCP) appears as a challenging subject in the e-marketing field. The probabilities to forecast customer churn has increased dramatically with the progression of artificial intelligence methods and machine learning algorithms. From the data mining perspective, CCP is the procedure by which customers anticipate belonging to the churner or non-churner target classes, according to their previous behavior information. Feature selection has a significant influence on

improving the classification performance after data preprocessing. Researchers have developed feature selection methods to decrease the dimension, calculation complexity, and over-fitting. These methods extract the most relevant features of users' input vectors for prediction (Coussement, 2014). Here, we did data preprocessing and feature engineering to remove the incomplete fields and duplicated rows. We performed four phases which have a significant influence on improving predictive accuracy. First, we selected Logistic regression, Decision Tree, Random Forest, and XGboost as four robust and practical supervised classification algorithms. Second, we try to find the most suitable parameters to optimize the performance of these algorithms, which are called hyperparameters. In this step, we preferred to apply the Grid search approach to discover and tune the hyperparameters. Many studies have proved that the Grid search is capable of detecting the most appropriate and fixed hyperparameters. Third, because the impact of the dataset attributes is not equal, we utilize the RFECV feature selection technique to extract the most influential features in the dataset and assign the higher weights to them in order of their importance. Finally, we proposed the Oracle algorithm as an ensemble of classifiers to achieve the highest precision rate. Table 2 explains the details of tuning the hyperparemeters for each classifier. We executed Logistic regression, C5.0 Decision Tree, and Random forest in the Sklearn library. We also implemented XGboost in the XGBoost library. XGboost is much more robust than the three mentioned ones. We extracted the RFECV method from the "sklearn.feature_selection" setting, which StratifiedKFold assigns as its cross-validation (CV) parameter. GridSearchCV allocates to the "sklearn.model_selection" setting of Jupiter notebook in the Anaconda environment. GridSearchCV always selects the most practical execution hyperparameters from the predetermined set of parameters and is the best optimizer tool for classifiers.

Table 2: Results of grid search for four classifiers

Model	Implementation	Optimizer Tool	Grid Parameters	Best Parameters
Logistic Regression (LR)	sklearn.linear_model (feature_selection = RFECV, step = 10)	GridSearchCV (cv = StratifiedKFold, n_jobs = -1, verbose = 0, scoring = 'accuracy)'	Solver = ['lbfgs','newton_cg','liblinear'], max_iter = [70,80,90,100], C = [10, 1.0, 0.1, 0.01], Penalty = ['l1','l2']	Solver = 'liblinear', max_iter = 70, C = 10, penalty= 'l1', accuracy = 0.835714
Decision Tree (DT)	sklearn.tree (feature_selection = RFECV, step = 10)	GridSearchCV (cv = StratifiedKFold, n_jobs = -1, verbose = 0, scoring='accuracy')	criterion = ['entropy'], class_weight = [dict,'balanced',None], max_depth = [5,6,7], max_leaf_nodes = [24,36,48], min_samples_leaf = [10,12,20], min_samples_split = [1,2,5], splitter = ['best']	criterion = 'entropy', class_weight = None, max_depth = 6, max_leaf_nodes = 48, min_samples_leaf = 20, min_samples_split = 2, splitter = 'best', accuracy = 0.822429
Random forest (RF)	sklearn.ensemble (feature_selection = RFECV, step = 10)	GridSearchCV (cv = StratifiedKFold, n_jobs = -1, verbose=0, scoring='accuracy')	n_estimators = [24,50], criterion = ['entropy','gini'], class_weight = [None,'balanced'], max_depth = [10,20], max_leaf_nodes = [2,5], max_features = ['auto',0.4], min_samples_leaf = [1,2,3], min_samples_split : [2,4,6], bootstrap = [True,False]	n_estimators = 50, criterion = 'gini', class_weight = None, max_depth = 10, max_leaf_nodes = 5, max_features = 0.4, min_samples_leaf = 1, min_samples_split = 2, bootstrap = True, accuracy = 0.817714

XGboost (XGB)	XGboost (feature_selection = RFECV, step = 10)	GridSearchCV (cv = StratifiedKFold, n_jobs = -1, verbose=0, scoring='accuracy')	learning_rate = [0.1], max_depth = [4,5,6], min_child_weight = [1,2], gamma = [0,0.1,0.2], subsample = [0.5,0.6], colsample_bytree = [0.7,0.8]	learning_rate = 0.1, max_depth =5, min_child_weight = 1, gamma = 0.2, subsample = 0.6, colsample_bytree = 0.8, accuracy = 0.854286

6.1. Logistic regression

Logistic regression is an improved supervised classification and generalized linear regression model indicating the relationship and discrimination between several independent features and dependent ones (Çelik & Osmanoglu, 2019).In binary classification problems, a logistic function is constructed by linearly merging input vectors with coefficients. This function has a sigmoid curve that assigns from 0 to 1 values to predict target attributes (Seippel, 2018). Logistic regression has frequently been applied in electronic marketing and Internet banking because of its easy concept and interpretation, particularly in analyzing consumer behavior. This approach results in precise and robust conclusions in customer purchase prediction intention and also their churn attitude. Sometimes logistic regression might outperform other complicated algorithms (Burez & Van den Poel, 2009).

6.2. Decision Trees

A decision tree is a tree-shaped decision structure producing classification rules from the training dataset by inductive learning. It divides up a massive amount of records into consecutive smaller sets. (Çelik & Osmanoglu, 2019). In a decision tree structure, leaves demonstrate class labels, for example, churner or non-churner. Branches describe connections of variables leading to these class labels. Although a decision tree cannot capture sophisticated and nonlinear links among the variables, it reaches a high accuracy in churn prediction problems (Asthana, 2018). Building and pruning are the two main stages to improve its structure. In the building phase, the dataset partitions recursively until all of the samples in every partition comprise identical values. In the pruning phase, some branches with the highest calculated error rate or noisy data are removed (Johny & Mathai, 2017).

6.3. Random forests

The random forest consists of some tree structures and combines subspaces and bagging elements of decision trees randomly. Random forest selects a subset of the variables at each node of the tree. Next, it extracts the best split among those variables for that node. (Çelik & Osmanoglu, 2019). Random forest uses the bagging technique to make the training dataset for every single tree by de-correlation. Then it fits them to the bootstrap resampled dataset. This structure selects a random sample of variables after splitting every tree for the next partition. The conclusions are according to the majority vote of the individual trees. Random forests apply the bagging technique for a large number of trees to overcome some week points of non-ensemble decision trees, like over-fitting and robustness (Seippel, 2018). As we mentioned before, many experts have applied this structure in binary classification problems, such as customer purchase behavior or customer retention prediction.

6.4. Extreme Gradient Boosting (XGBoost)

XGBoost is a supervised classification machine learning algorithm applying for tabular or structured datasets. XGBoost executes a decision tree classifier with a gradient boosting approach to accelerate the velocity of computation time, use less memory, and enhance the accuracy (Çelik & Osmanoglu, 2019). Gradient boosting is a method in which residuals and novel algorithms models implement to calculate the error concisely, and after that, both mix to use the existing resources optimally to train the prediction classifier. XGBoost also applies gradient descent to place the minimum or decrease the amount of loss function (Lalwani et al., 2021).

6.5. Oracle Ensemble Selection.

Oracle is a classifier composition method that chooses the preliminary algorithm to forecast the accurate class label for a current instance (Malmasi, Tetreault, & Dras, 2015). This ideal abstract meta-classifier often estimates the upper bound of classification precision (Cruz, Hafemann, Sabourin, & Cavalcanti, 2020). Oracle usually selects the classifiers that anticipate the correct class label. Therefore, it represents the perfect ensemble selection structure. Figure 2 illustrates the procedure of Oracle comprising three feasible stages: (1) Generation, (2) Selection, and (3) Aggregation. In Generation, a pool containing M classifiers $C = \{c_1, \ldots, c_M\}$ trains to produce a collection of precise, various, and informative classifiers. In Selection, an ensemble of the most suitable classifiers $(C' \subseteq C)$ is chosen based on the validation situation. In Aggregation, the chosen base classifiers generate the outputs through a compound rule to forecast the class labels (Cruz, Sabourin, & Cavalcanti, 2018).

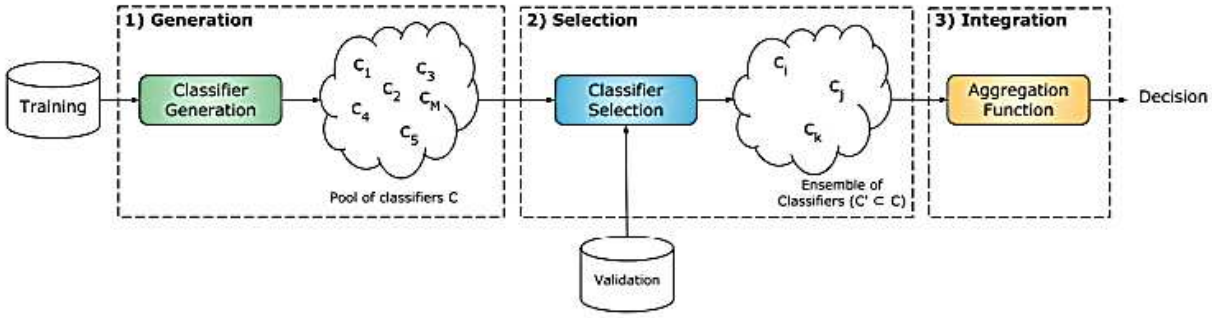

Figure 2: Static ensemble selection procedure

7. RFECV Feature Selection

Feature selection approaches are performed to designate an appropriate subset of the most relevant variables in a dataset. Some supervised classifiers may be misrouted by irrelevant input variables. Recursive Feature Elimination Cross-Validation (RFECV) is a wrapper-type feature selection method. It performs by eliminating the weakest variables recursively and creates a model on the remained ones. It uses the accuracy to diagnose which variables or a mixture of variables take part the most to forecast the target variable. RFECV searches for a subset of the feature by starting with all of them in the training dataset. Next, it discards them. Then, it fits the given classifier in the model's core. After that, it ranks the features by their importance and excludes the least significant ones. Finally, it re-fits the model to get the desirable count of variables. It uses a cross-validated selection of the best number of variables, which is more practicable in binary classification issues (Brownlee, 2016). RFECV assigns a weight for each variable using the classification function to evaluate subsets' accuracy on test data. RFECV avoids overfitting by measuring cross-validation. Figure 3 displays the RFECV method for four classifiers in this study. The optimal number of features selected in Logistic Regression, Decision tree, Random Forest, and XGboost are 46, 26, 46, and 346, respectively. The two best features selected in all of them are “Balance” and

“Estimated Salary” respectively. These two features do not change the cross-validation score in logistic regression. When RFECV selects them, the cross-validation score remains zero. After selecting the “Geography” related features, the cross-validation achieves more than a 0.835 score. It is the highest score in logistic regression and does not change after that. When RFECV chooses “Balance,” the cross-validation score increases dramatically in a decision tree, random forest, and XGboost because “Balance” is the best feature and these three algorithms are sensitive to the optimum feature. After taking “Balance,” the cross-validation in the decision tree and XGboost attains more than a 0.82 score. Despite some fluctuations, this score remains over 0.82 in the decision tree. This is the same trend in the random forests after achieving a 0.815 score. XGboost shows different behavior. When RFECV takes “Estimated Salary,” its cross-validation reaches more than 0.83. This score increases steadily and stays in the range of 0.85 when “CreditScore” related features are selected. XGboost chooses more features and considers their effects more than the other classifiers. It is the superiority of XGboost performance compared to the other classifiers, especially the logistic regression.

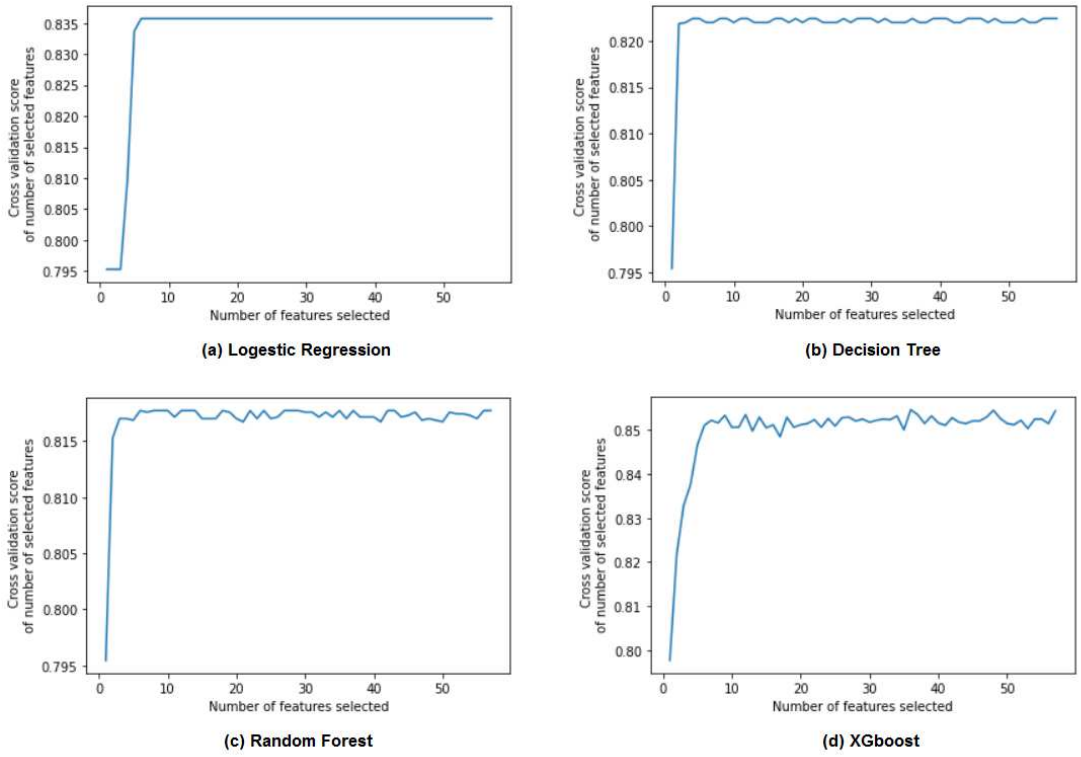

Figure 3: RFECV feature selection for classifiers

We also demonstrated the importance of graded variables by some bagged decision trees like the C5.0 decision tree, random forest, and XGboost, consisting of the feature importance for the scored features. Feature importance marks as the decrement in node impurity weighted by the feasibility of arriving at that node. We applied decision tree, random forest, and XGboost to rank the top twenty variables in Figure 4, Figure 5, and Figure 6 consecutively. As displayed in Figure 4, Figure 5, and Figure 6, “NumOfProducts” is the most critical variable in estimating the performance and has a conspicuous discrepancy with others. It signifies that the total number of products that customers have purchased through the bank impacts their churn decision. “NumOfProducts” indicates the relationship between customer churn intention and the target value (Exited). If the customers are not satisfied with the banks’ products or services, they stop buying and move to another bank.

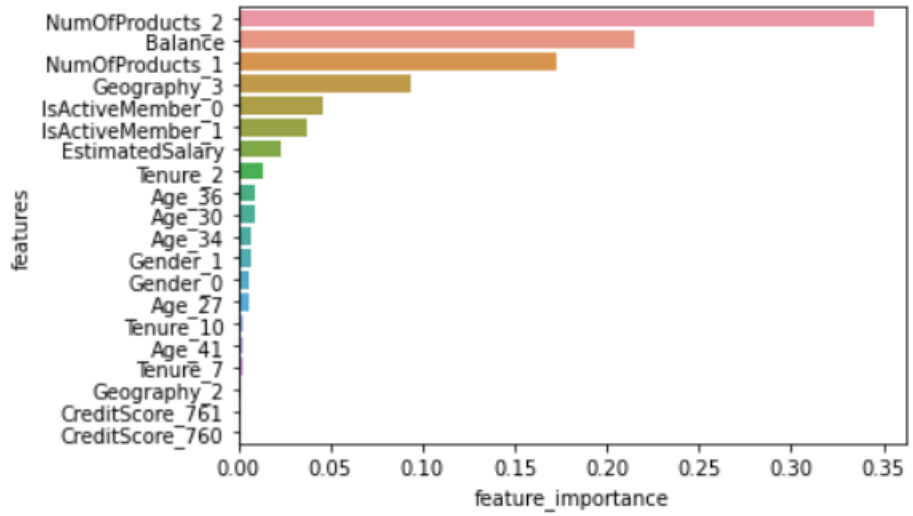

Figure 4: Feature importance of decision tree classifier in transactions dataset

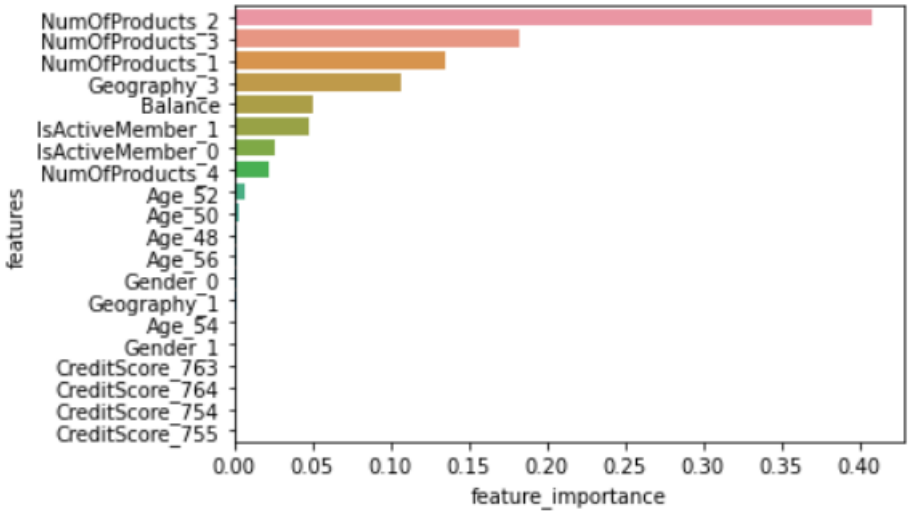

Figure 5: Feature importance of random forest classifier in transactions dataset

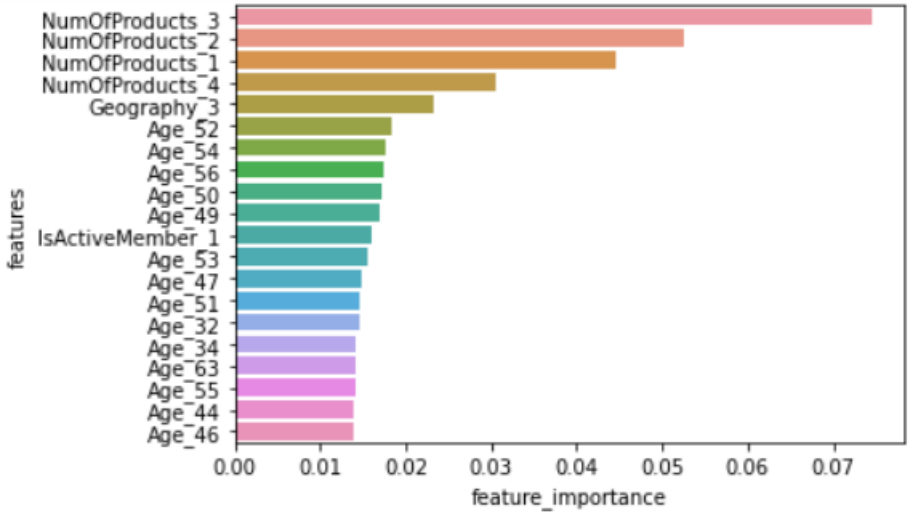

Figure 6: Feature importance of XGboost classifier in transactions dataset

8. Results

This study focuses on developing customer churn prediction by leveraging the data of transactions committed through a banks' gateway and supervised classification algorithms. We implemented Logistic Regression, C5.0 Decision tree, and Random Forest in the Sklearn library because it presents and supports different necessary applications for supervised classification algorithms. We fetched the XGboost algorithm from the XGboost library. To benchmark models, we selected four well-established classifiers. We applied hyperparameter optimization for each of them to discover the most suitable parameters. Next, we determined to execute RFECV feature selection in logistic regression, C5.0 decision tree, random forest, and XGboost because they support the RFECV method. Then, we picked out the Oracle meta-classifier as a static ensemble selection approach. We decided to add Logistic Regression, Decision tree, Random Forest, and XGboost along with the RFECV method in Oracle. It means that we incorporated the Oracle meta-classifier with RFECV feature selection to develop churn prediction accuracy. After compiling, the accuracy of Oracle achieves 90.73%, which is far more than each of logistic regression, C5.0 decision tree, random forest, and XGboost. We imported Oracle from the DESlib library. DESlib is an entirely documented library that facilitates the execution of classification models to achieve high accuracy (Cruz et al., 2020). Table 3 indicates the predicted scores of the most prominent models to forecast whether a customer will leave the bank or not.

Table 3: Summary of the results for supervised classification models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Cohen Kappa Score	Null Accuracy
LR	0.8533	0.7570	0.4017	0.5248	0.6845	0.4476	0.7983
DT	0.8327	0.6711	0.3339	0.4459	0.6463	0.3602	0.7983
RF	0.8243	0.9149	0.1421	0.2461	0.5694	0.2028	0.7983
XGB	0.8687	0.8168	0.4496	0.5800	0.7121	0.5098	0.7983
Oracle	0.9073	0.9910	0.5455	0.7036	0.7721	0.6541	0.7983

We also applied a confusion matrix and classification report to discover errors and interpret the results of the binary churn classification subject. TPR distinguishes the percentage of customers who did not leave the bank. On most occasions, this proportion is more than TNR, discerning the percent of the customers who stop receiving the bank's services and move to another one. PPV detects the percent of customers who did not churn, but they have been recognized to be disinterested and dissatisfied at a specific time. They may decide to leave the bank in future. Bank shareholders should adopt some new strategies to prevent their churning as much as possible because new customer acquisition is much more costly for them than keeping the existing one. NPV diagnoses the percentage of satisfied customers and prone to become loyal to the bank at a particular duration of time. They are engaged in purchasing products, receiving services, and committing the transactions. They might be the returning customers who repeat shopping several times. Table 4 indicate that the Oracle meta-classifier has a higher TPR, TNR, PPV, and NPV than other classifiers. FPR represents the ratio of customers who did not leave the bank, but analysts detect them as churners incorrectly. FNR infers the ratio of actual churners who are not distinguished correctly. FDR indicates all clients except the potential churners, while FOR signifies all clients except the potential loyal and repeated customers. FPR, FNR, FDR, and FOR demonstrate the error ratios. Table 4 acknowledges that Oracle has a lower error ratio than the others. ACC affirms that all churners and non-churners are classified in a proper way. Oracle has the highest rate of accuracy (ACC) among all mentioned algorithms. Finally, MCC calculates the competence of the algorithm in identifying the observed and predicted binary classifications, where Oracle attains an admissible rate of 69.51%.

Table 4: Summary of the predictions for model classifiers

Model	TPR	TNR	PPV	NPV	FPR	FNR	FDR	FOR	ACC	MCC
LR	0.8649	0.7570	0.9674	0.4017	0.2430	0.1351	0.0326	0.5983	0.8533	0.4791
DT	0.8507	0.6711	0.9587	0.3339	0.3289	0.1493	0.0413	0.6661	0.8327	0.3907
RF	0.8214	0.9149	0.9967	0.1421	0.0851	0.1786	0.0033	0.8579	0.8243	0.3197
XGB	0.8751	0.8168	0.9745	0.4496	0.1832	0.1249	0.0255	0.5504	0.8687	0.5417
Oracle	0.8969	0.9910	0.9987	0.5455	0.0090	0.1031	0.0013	0.4545	0.9073	0.6951

Since the churn dataset is imbalanced inherently, we preferred to apply the ROC curve (AUC) to compare the effectiveness of supervised classification models. The area under the ROC curve (AUC) for each algorithm exhibits in Figure 7. AUC is a standard criterion to appraise performance. According to Figure 7, the AUC of the random forest is less than 0.6. So, we considered it as week discrimination detected between two classes of churner and non-churner. Logistic regression and decision tree are between 0.6 and 0.7, distinguished as acceptable discrimination. The excellent discrimination is related to XGboost, which is more than 0.7. Considering a 77% AUC score, the Oracle meta-classifier has an outstanding performance compared to the other classifiers.

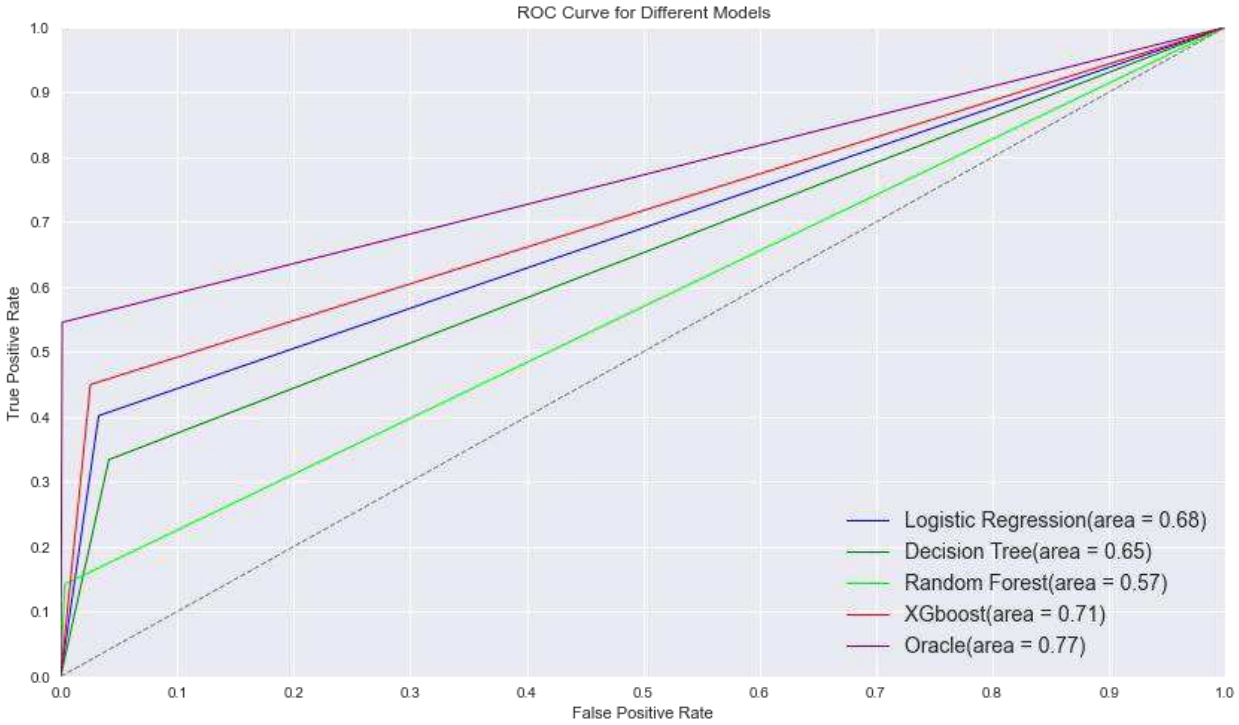

Figure 7: ROC Curve of churn prediction ensemble models

9. Theoretical Contribution and Future Extensions

In this research, we introduced the Oracle meta-classifier algorithm to predict customer churn intention from the transactions dataset. Oracle is a static and robust algorithm with a high degree of accuracy, particularly when incorporating with feature selection process. We implemented Oracle in the "DESlib" library, which has high test coverage. We chose the best parameters in the establishment of its classifiers. We used Logistic Regression, Decision Tree, Random Forest, and XGboost to forecast customer churn attitude intention in Internet banking. We performed the RFECV feature selection method to improve the proficiency of the selected algorithms. The main idea of RFECV is to generate several variable selectors,

control them, and combine their outputs, which operates better than the other feature selection techniques in many perspectives. Concerning the process of extracting hyperparameters, we operated Grid search as the most stable tool for tuning. Every time the tuning program runs, the values of hyperparameters extracted from the Grid search are constant and not changed. The outputs of Grid search are meticulous and explicit, which can improve the algorithms' veracity extraordinarily. Considering the effects of more significant variables, we preferred to apply RFECV. Analysts can perform SelectfromModel, which is an embedded approach instead of using RFECV as a wrapper one for feature selection. They also can apply Lasso CV or Ridge Classifier CV when they want to implement cross-validation. We assigned the StratifiedKFold as the cross-validation parameter to prevent overfitting. We depicted the top twenty features picked out by RFECV in the structure of the C5.0 decision tree, random forest, and XGboost in Figure 4, Figure 5, and Figure 6, respectively. Whereas there is not such this capability for logistic regression in illustrating feature importance. We understood that "NumOfProducts" has a significant impact on predicting the customers' behaviors. We selected "C5.0" as a decision tree algorithm because it acts much more flexible and accurate than other types of decision trees, like CART, ID3, and C4.5. C5.0 ranks variables with the Entropy criteria and reduces the tree size by excluding branches that engender inconsequential rules. C5.0 applies pre-pruning to prohibit the earlier tree growth and diminish the complexity. Another prior functionality of C5.0 are lesser memory space occupation and lesser execution time (Rajeswari & Suthendran, 2019).C5.0 can also control numerical and categorical features by dividing the training set correctly. So, it is much more proper for our transactions dataset. We calculated Gini criteria to organize the random forest structure. The random forest chooses the variable with superior Gini for the internal node. In a random forest, the mean decline impurity of all decision trees computes the feature importance (Płoński & Zaremba, 2014). Specialists can apply a more complicated method of gradient boosting on decision trees such as LightGBM, Catboost, or AdaBoost to improve churn prediction efficiency and effectiveness. Experts can use the Oracle meta-classifier not only with feature selection approaches but also in different classification issues, such as confidence estimation and discovering the missing features. They may also apply the Oracle algorithm in some other decision-making fields, such as conversion rate prediction or business intelligence.

10. Implication and Further Directions

Electronic marketing platforms provide a wide range of purchase choices. In these platforms, if the customers are not satisfied with the products or services of suppliers, they may leave the market and move towards the competitors. Thus the customers' bargain power has increased dramatically, especially in the retailing sector. It creates high pressures on suppliers to keep their customers satisfied and make a permanent relationship with them. Many retailers have decided to change their concentrations from product-centric into customer-centric because of the intense competition, which is available between organizations in new customer acquisition and retains the existing ones. Every organization uses CRM strategies to attract visitors, keep customers, and build some exciting hindrances to prevent them from leaving. Customer churn prediction is one of the enterprises' retention strategies, which has emerged as a critical and challenging domain in business studies. In the Internet banking industry, Bankers have to shift their product-centered strategies to customer-centered ones. This change has accrued because of the various situations c provided by the competitive online environment. Shareholders need to keep their existing customers because customer acquisition is much more expensive than retaining the previous ones. This expenditure considers the aspects of money, time, resources, and energy. Thus, they have to adopt some new strategies to prevent their churning as much as possible because customers bring the maximum profit and interest for their banks. Also, predicting customer churn behavior and buying patterns from transaction data may bring many competitive advantages for suppliers and marketers. For

example, distinguishing bestselling products, returning or new customers, customer purchase rate, and reasons are the benefits of forecasting customer churn patterns. They can take a practical approach to satisfy and retain the current customers and engage visitors to convert them to new customers. Many business analysts apply machine learning algorithms to recognize customer needs and consumption patterns. In this paper, we presented the Oracle meta-classifier, which is a static ensemble selection algorithm to predict customer churn intention. We utilized the transactions data of a bank. We applied Logistic Regression, Decision Tree, Random Forest, and XGboost in the Oracle structure. The results show that Oracle can predict customer churn much more accurate than each of the classifiers alone. Oracle is a precise meta-classifier, which may have practical usage in other prediction scopes, like e-marketing, e-retailing, Stock market, or even health areas.

References

AL-Shatnwai, A. M., & Altibbi, M. (2020). Predicting Customer Retention using XGBoost and Balancing Methods. International Journal of advanced computer Science and applications, 11(7), 704-712.

Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. International Journal of Engineering Research and Applications, 4(5), 165-171.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 94, 290-301.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Alawfi, K., Hussain, A., & Huang, K. (2017). Customer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242-254.

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. Applied Computing and Informatics, 15(2), 94-101.

Asthana, P. (2018). A comparison of machine learning techniques for customer churn prediction. International Journal of Pure and Applied Mathematics, 119(10), 1149-1169.

Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management: John Wiley & Sons.

Bharadwaj, S., Anil, B., Pahargarh, A., Pahargarh, A., Gowra, P., & Kumar, S. (2018). Customer Churn Prediction in Mobile Networks using Logistic Regression and Multilayer Perceptron (MLP). Paper presented at the 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT).

Brownlee, J. (2016). Feature Selection For Machine Learning in Python.

Burez, J., & Van den Poel, D. (2007). CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications, 32(2), 277-288.

Burez, J., & Van den Poel, D. (2008). Separating financial from commercial customer churn: A modeling step towards resolving the conflict between the sales and credit department. Expert Systems with Applications, 35(1-2), 497-514.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626-4636.

Çelik, O., & Osmanoglu, U. O. (2019). Comparing to techniques used in customer churn analysis. Journal of Multidisciplinary Developments, 4(1), 30-38.

Coussement, K. (2014). Improving customer retention management through cost-sensitive learning. European Journal of Marketing.

Coussement, K., & De Bock, K. W. (2013). Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research, 66(9), 1629-1636.
Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313-327.
Cruz, R. M., Hafemann, L. G., Sabourin, R., & Cavalcanti, G. D. (2020). DESlib: A Dynamic ensemble selection library in Python. Journal of Machine Learning Research, 21(8), 1-5.
Cruz, R. M., Sabourin, R., & Cavalcanti, G. D. (2018). Dynamic classifier selection: Recent advances and perspectives. Information Fusion, 41, 195-216.
Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. Paper presented at the 2016 Symposium on Colossal Data Analysis and Networking (CDAN).
De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2), 760-772.
Ding, J.-M., Liu, G.-Q., & Li, H. (2015). The application of improved random forest in the telecom customer churn prediction. Pattern Recognition and Artificial Intelligence, 28(11), 1041-1049.
Firdous, S., & Farooqi, R. (2017). Impact of internet banking service quality on customer satisfaction. Journal of Internet Banking and Commerce, 22(1).
Grubor, A., & Jakša, O. (2018). Internet marketing as a business necessity. Interdisciplinary Description of Complex Systems: INDECS, 16(2), 265-274.
Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random forests for uplift modeling: an insurance customer retention case. Paper presented at the International Conference on Modeling and Simulation in Engineering, Economics and Management.
Hand, D. J. (2018). Statistical challenges of administrative and transaction data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(3), 555-605.
Hanif, I. (2020). Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction.
Harahap, D. A., Hurriyati, R., & Amanah, D. (2020). Conceptual Model of E-Service Quality at Branchless Banking in Indonesia. Journal of Internet Banking and Commerce, 25(2), 1-11.
Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. European Journal of Operational Research, 284(3), 920-933.
Isa, S. I. H. S., & Nayan, S. M. (2020). WOW Your Customers: Tips to Retain Customers. Journal of Undergraduate Social Science and Technology, 2(2).
Jain, A., Menon, M. N., & Chandra, S. (2015). Sales forecasting for retail chains. San Diego, California: UC San Diego Jacobs School of Engineering.
Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101-112.
Johny, C. P., & Mathai, P. P. (2017). Customer churn prediction: A survey. International Journal of Advanced Research in Computer Science, 8(5).
Julia Kagan, T. B. (2021). Credit Score. Debt Management Guide. Retrieved from https://www.investopedia.com/
Kirui, C., Hong, L., Cheruiyot, W., & Kirui, H. (2013). Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI), 10(2 Part 1), 165.
Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2021). Customer churn prediction system: a machine learning approach. Computing, 1-24.

Li, Y., Hou, B., Wu, Y., Zhao, D., Xie, A., & Zou, P. (2021). Giant fight: Customer churn prediction in traditional broadcast industry. Journal of Business Research, 131, 630-639.
Mahajan, D., & Gangwar, R. (2017). Improved Customer Churn Behaviour By Using SVM. International Journal of Engineering and Technology, 2395-0072.
Maheswari, K., & Priya, P. P. A. (2017). Predicting customer behavior in online shopping using SVM classifier. Paper presented at the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS).
Malmasi, S., Tetreault, J., & Dras, M. (2015). Oracle and human baselines for native language identification. Paper presented at the Proceedings of the tenth workshop on innovative use of NLP for building educational applications.
Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), 15273-15285.
Nyilasy, G. (2007). Word of mouth: what we really know–and what we don't. In Connected marketing (pp. 197-220): Routledge.
O'Brien, R., & Ishwaran, H. (2019). A random forests quantile classifier for class imbalanced data. Pattern recognition, 90, 232-249.
Płoński, P., & Zaremba, K. (2014). Visualizing random forest with self-organising map. Paper presented at the International Conference on Artificial Intelligence and Soft Computing.
Pustokhina, I. V., Pustokhin, D. A., Aswathy, R., Jayasankar, T., Jeyalakshmi, C., Díaz, V. G., & Shankar, K. (2021). Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms. Information Processing & Management, 58(6), 102706.
Qiu, Y., & Li, H. (2010). Application of Feature Extraction Method in Customer Churn Prediction Based on Random Forest and Transduction. J. Convergence Inf. Technol., 5(3), 73-78.
Rajeswari, S., & Suthendran, K. (2019). C5. 0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud. Computers and Electronics in Agriculture, 156, 530-539.
RB, D. (2021). Customer churn prediction in telecommunication industry through machine learning based Fine-tuned XGBoost algorithm.
Richards, K. A., & Jones, E. (2008). Customer relationship management: Finding value drivers. Industrial marketing management, 37(2), 120-130.
Royo-Vela, M., & Casamassima, P. (2011). The influence of belonging to virtual brand communities on consumers' affective commitment, satisfaction and word-of-mouth advertising: The ZARA case. Online Information Review.
Saini, M. N., Monika, G. K., & Garg, K. (2017). Churn prediction in telecommunication industry using decision tree. Streamed Info-Ocean, 1.
Santouridis, I., & Trivellas, P. (2010). Investigating the impact of service quality and customer satisfaction on customer loyalty in mobile telephony in Greece. The TQM Journal.
Sarker, B., Podder, P., & Alam, R. (2020). Progression of Internet Banking System in Bangladesh and its Challenges. International Journal of Computer Applications, 975, 8887.
Seippel, H. S. (2018). Customer purchase prediction through machine learning. University of Twente,
Senthan, P., Rathnayaka, R., Kuhaneswaran, B., & Kumara, B. (2021). Development of Churn Prediction Model using XGBoost-Telecommunication Industry in Sri Lanka. Paper presented at the 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).
Shaheen, I., & Naseem, N. (2015). A review of customer satisfaction, employee satisfaction and their impact on firm performance. Studies, 4(1), 21-31.
Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. International Journal of Information Management, 48, 238-253.

Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., & Snoeck, M. (2018). Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation, 40, 116-130.

Tang, P. (2020). Telecom Customer Churn Prediction Model Combining K-means and XGBoost Algorithm. Paper presented at the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE).

Tang, Q., Xia, G., Zhang, X., & Long, F. (2020). A Customer Churn Prediction Model Based on XGBoost and MLP. Paper presented at the 2020 International Conference on Computer Engineering and Application (ICCEA).

Tsai, C.-F., & Chen, M.-Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. Expert Systems with Applications, 37(3), 2006-2015.

Tsai, C.-F., & Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10), 12547-12553.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. Applied Soft Computing, 14, 431-446.

Vo, N. N., Liu, S., Li, X., & Xu, G. (2021). Leveraging unstructured call log data for customer churn prediction. Knowledge-Based Systems, 212, 106586.

Wei-Yun, Y. (2012). The Research on Random Forests and the Application in Customer Churn Prediction. Management Review, 24(2), 140.

Wu, L., & Li, M. (2018). Applying the CG-logistic Regression Method to Predict the Customer Churn Problem. Paper presented at the 2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS).

Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. Paper presented at the 2016 13th International conference on service systems and service management (ICSSSM).

Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems with Applications, 36(3), 5445-5449.

Yanfang, Q., & Chen, L. (2017). Research on E-commerce user churn prediction based on logistic regression. Paper presented at the 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC).

Zhao, X., Shi, Y., Lee, J., Kim, H. K., & Lee, H. (2014). Customer churn prediction based on feature clustering and nonparallel support vector machine. International Journal of Information Technology & Decision Making, 13(05), 1013-1027.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ChurnModelling.csv