# Evaluation of Customer Behaviour with Machine Learning for Churn Prediction: The Case of Bank Customer Churn in Europe

Pei Chen[1*†], Nian Liu[2 †], Binrui Wang[3 †]
*pxc5328@psu.edu, eee1909240@xmu.edu.my, binrui.wang@student.reading.ac.uk

[1]School of Liberal Arts and informational science Pennsylvania state university, University Park, United States

[2]School of Electrical Engineering and Artificial Intelligence Xiamen University Malaysia, Kuala Lumpur, Malaysia

[3]School of Management and Economics Beijing Institute of Technology, Beijing, China

[†]These authors contributed equally

**Abstract**—The way to extract valuable information from users' information data as well as provide correct and effective financial services is a topic worthy of attention. Machine learning scenarios are widely adopted to predict user churn, implanting corresponding marketing methods, and formulating appropriate retention measures is being applied by more and more companies to reduce customer churn. In this paper, the logistic regression classifier is chosen as the baseline model to predict customer churn based on 10,000 customer information of a European bank. Based on the SMOTE oversampling method to balance the data set and other processes, five models of logistic regression, SVC, GBDT, random forest and AdaBoost are finally trained. Thereinto, this paper only performs grid search on the four models except the baseline model to optimize the model parameters. After the evaluation through recall, precision and AUC scores, AdaBoost classifier is finally selected as the prediction model for this case, with recall and AUC score being 0.718 and 0.776, respectively. These results shed light on guiding further exploration of the prediction of bank customer churn.

**Keywords**-machine learning; customer churn prediction; logistic regression; AdaBoost.

## 1 INTRODUCTION

Contemporarily, banks have additional channels and ways for users' information due to the widespread usage of mobile devices and the Internet's rapid growth. It supports big data technology, data storage technology, the quick development of cloud computing technology and application, and the rapid rise of data and computing capacity, hence accelerating the development of artificial intelligence technology [1]. In this setting, the traditional financial sector confronts competition from its peers and the challenges posed by the Internet financial sector.

Achieving practical value from data, doing business more effectively for diverse user groups, investigating the potential worth of data, and obtaining more advantages are essential business

FFIT 2022, October 28-30, Shenzhen, People's Republic of China

DOI 10.4108/eai.28-10-2022.2328450

requirements for data-driven industries (e.g., banks). Banks seek channels of combining data analysis, data mining, and machine learning models with actual company operations to segment user groups, anticipate user churn, and evaluate user behavior data [2]. However, the majority of them are still in the research and early application phase, and the use of data in actual production is insufficient and illogical. Moreover, the implementation of big data and artificial intelligence technologies in the banking business is still in its infancy, and there is ample space for growth.

The path to extract useful and valuable information from any user information data and deliver correct and effective financial services is a topic worthy of attention. For instance, it is crucial to the bank's healthy and sound development to maintain existing customers and to take preventative steps to keep potential customers [3]. In real production, however, the danger of user loss is uncertain and inconsistent. Learning from a large amount of user data and anticipating the risk of user attrition is a crucial strategy and tool for addressing this issue. The fast growth of big data technology and artificial intelligence technology gives more trustworthy options for processing and analyzing enormous amounts of data. User churn prediction with machine learning techniques for banks is helpful in identifying customers at risk of churning using certain means and procedures, adopting relevant marketing means and designating appropriate retention measures, enhancing resource utilization, and decreasing customer turnover.

Previous studies have shown that churn has increased in recent years in all industries, including finance, and the cost of acquiring a new user is more than five times that of retaining an existing one [4]. Therefore, it is the key to enhancing the core competitiveness of banks to establish an effective forecasting model for customer in order to offer a path that churn customer managers can take effective measures in advance to prevent customer churn.

Contemporarily, the prediction of user churn has attracted wide attention in many industries such as banking [5-7], telecommunications [8], insurance companies [9], and academia. Many machine learning models (e.g., logistic regression and decision tree models) have achieved a good simulation effect and have been successfully applied in user churn prediction. In the study of logistic regression prediction of user churn, Gürsoy uses data mining technology, logistic regression and decision tree to determine the customer churn for the company [10]. Caigny et al. proposed the Logit Leaf Model with a mixture of logistic regression and decision tree, which better solved the problem that the decision tree could not deal with linear relations and logistic regression could not solve the interaction between variables [11]. Prasad et al. used data mining technology and the CART classification tree respectively, to model the purchasing behavior of bank users [12]. Meanwhile, Keramati et al. also simulated the user loss of A commercial bank by using a decision tree through data mining technology [6]. In addition, Sawangarreerak solves the data imbalance problem with SMOTE and improves the forecasting model of subscriber churn based on sample data set in Indonesia using weighted random forest which has better performance than previous prediction models [13].

In this paper, we assess the efficacy of different machine learning models for churn prediction based on bank customer data. The data acquisition, feature engineering, model training, and evaluation phases of the application process are concluded. In this study, five classifiers, including Logistic Regression, Support Vector Machine, Random Forest Classifier, GDBT, and Adaboost, are employed to create predictions. To evaluate the performance of models, certain metrics (e.g., precision and recall) are calculated.

The rests part of the paper is organized as follows. The second part mainly introduces the model used in this paper. This paper discusses the history and principles of these models and shows their formulas. The third section is the results presentation. It outlines the data sources and preparation techniques utilized in the analysis. The EDA is then performed, which demonstrates a first comprehension of the customer's composition. The method of feature engineering is then demonstrated. Finally, the performance of the model is reviewed and its limits are introduced. The conclusion summarizes the work and provides an outlook for the future.

# 2 METHODOLOGY

## 2.1 Model

In the paper, the analysis is performed using the Python programming language. Python is an open-source programming language that makes use of Scikit-learn, a free software toolkit designed to operate with Python. There are various functionalities in Python, containing regression and classification algorithms, clustering, data cleaning and feature engineering [14]. In the article, there are many models used to analyze the data, including Logistic Regression, Random Forests Classifier, GDBT, Support Vector Machine and Adaboost.

Logistic Regression is a classification model rather than a regression approach [15]. Using the logistic sigmoid function, logistic regression transforms the linear function 0 + β1X as well as translates it into two or more discrete groups. As illustrated in the following equation, P is the probability, X is the input set, and b and b' are the coefficients determined using maximum-likelihood estimation during training. Logistic regression is one of the most common machine learning classification techniques due to its ease of implementation, extensibility, and effectiveness in categorizing new information [16].

$$P(X) = \left(\frac{(e)^{b+b'X}}{1+(e)^{b+b'X}}\right) \quad (1)$$

Random Forests Classifier is an ensemble of decision tree classifiers that constructs vast collections of decorrelated trees and aggregates the prediction of each tree to assign a class by majority vote [17]. The RFC is a supervised learning technique that individually trains each tree using a random subset of the whole training data. The nodes are then segregated according to feature selection in the random subset. The equation represents classification, hi(x) represents single decision tree models, Y represents the desired output, and i represents the indicator function.

$$H(x) = \arg\max \sum_{i=1} n\, I(hi(x) = Y) \quad (2)$$

Unlike random forest, GBDT is a mix of the decision tree and the Boosting algorithm. The GBDT model is given in the same manner as the 1951 Decision Trees in Gradient, which is an iterative Decision tree algorithm consisting of often hundreds of small-sized Decision Trees. Model forecasts, for a sample instance input, first will provide a starting value, then every decision tree traversal, each tree will modify forecast correction, and the results are obtained by accumulating the outcomes of each decision tree. The GBDT training process is linear. In contrast to the random forest algorithm, which trains multiple trees in parallel, the residual difference between the first $T_1$ training result and the actual value T is used as the sample for

the second $T_2$ decision tree, and the sample for the nth Tn decision tree is the training result of the n-1st $T_{N-1}$ decision tree. Consequently, the final classification result of this model is the sum of each decision tree's node values.

SVM is a novel approach for supervised machine learning that was created in the 1990s. SVM maps the input space of a nonlinear issue to a high-dimensional feature space using kernel functions. SVM can classify input data linearly after generating the best classification hyperplane in this high-dimensional space. SVM relies on linear separability in the ideal classification plane to determine the best classification line that is both closest to and parallel to the two sample sets. In addition to accurately separating the two sample types, the best classification line optimizes the classification interval [18].

Adaboost is a decentralized method for data segmentation, clustering, prediction, and regression [19]. Adaboost modifies the original training set's distribution in a new method [20]. If t is the misclassification rate at trial t, then $\beta t=(1-t)/t$ will be applied to the weights of misclassified samples in the training set. Normalized to 1 is the entire sum of the corrected weights. The classifiers C1~Ct are integrated using weighted voting, with Ct being weighted by log(t). If $t > 0.5$ (indicating that the classifier's accuracy is worse than a random guess), the trials will be ended and trial T will be modified to t 1. If $t = 0$ (corresponding to 100 percent accuracy), then trial T becomes t. Adaboost forces the newly created ensemble classifiers to prioritize "harder" situations in this manner. Voting for the final labels is weighted according to each classifier's precision. Adaboost is implemented in this work using C5.0, and the trees are generated using the information gain ratio. In the feature selection stage, one-hot encoding, SMOTE and normalization are used to remove irrelevant features and redundant features. The scope of the parameters for tuning in grid search is shown in Table 1.

Table 1 Scope of the parameters for tuning in grid search

| **Model** | **Parameters** |
|---|---|
| SVC | kernel = ["linear", 'rbf'] |
| | gamma = [1e-3, 1e-4] |
| | C = [100, 1000, 2000, 5000, 10000] |
| GBDT | learning_rate = [0.01, 0.1] |
| | max_depth = [4,6,8,10] |
| | Subsample = [0.3,0.5,0.8,1.0] |
| RandomForest | max_depth = range(10, 18, 1) |
| | min_samples_leaf = [1, 2, 4, 6] |
| | min_samples_split = [5, 6, 7] |
| | n_estimators = [20, 40, 50, 60] |
| AdaBoost | n_estimators = [20, 40, 50, 60] |

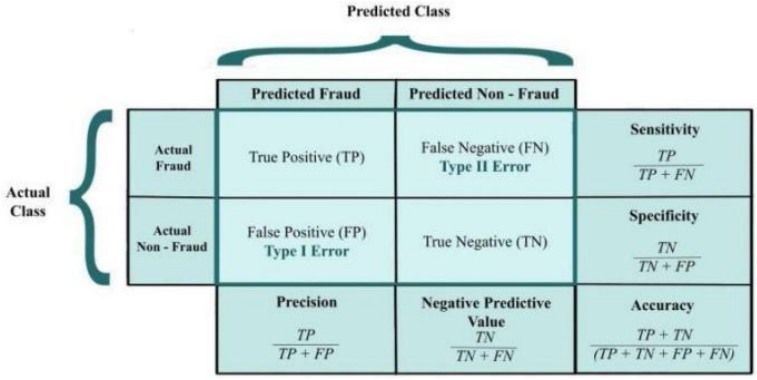


**Figure 1.** Correlation matrix

## 2.2 Metrics

For illustration purposes, a correlation matrix is made. As shown in Fig. 1, each object is associated with a binary label l which accounts for the correctness of the object with respect to the task. As depicted in Fig. 1, recall is defined as the true positive rate (TPR), whereas precision measures the fraction of items classified as positive that are truly positives. Therefore, it describes how good a model is as predicting the positive class.

$$\mathrm{RC} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

$$\mathrm{PC} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4}$$

To balance precision and recall, the analysis also refers to F1 score. F1 score is a harmonic mean of precision and recall, which can be expressed as a function of counts of true positives, false positives and false negatives. The value of F1 score is at most 1 and at least 0.

$$\mathrm{F1} = \frac{2\mathrm{tp}}{2\mathrm{tp+fp+fn}} \tag{5}$$

The AUC measures the model's discriminating ability mapped between values of 0.5 and 1. The AUROC is a measure of the usefulness of the test, where the greater the area under the curve, the more useful the test [21].

## 2.3 Data

The data used in this paper is from 10,000 pieces of customer information from a European bank provided by Kaggle website. The original data contains 14 columns, and the specific name and description of each field are shown in Table 2. "Excited" is the dependent variable that needs to be predicted in business, and the value 1 in the column indicates that the customer is an exited one, accounting for 20.4% of all records; while the value 0 represents a retained user, taking up 79.6% of the total. The data set is imbalanced and should be processed properly before the training.

Table 2 Variable type and description

| Variable | Type | Description |
|---|---|---|
| RowNumber | int64 | Row number |
| CustomerID | int64 | Customer ID |
| Surname | object | Surname |
| CreditScore | int64 | Credit score, varying from 350 to 850 |
| Geography | object | The country that the customers live in |
| Gender | object | Customer gender |
| Age | int64 | Customer age |
| Tenure | int64 | The years that customer stayed with the bank |
| Balance | float64 | Total money in the account |
| NumOfProducts | int64 | Product number that the customer use |
| HasCrCard | int64 | Whether has a credit card in this bank |
| IsActiveMember | int64 | Whether the customer is active or not |
| EstimateSalary | float64 | Yearly salary |
| Exited | int64 | Churn or not |

# 3 RESULTS & DISCUSSION

## 3.1 Exploratory Data Analysis

The experiment begins with an exploratory analysis of the data to get a preliminary understanding of which groups are more likely to become churn customers. Some of the valuable findings are that female customers churn at a higher rate than male customers, most of the churn comes from customers with credit cards from the bank, and the churn rate is higher for inactive customers. In addition, it can be found that most consumers using product 3 and product 4 had changed to an excited one, and in fact, all customers using product number 4 were gone. Furthermore, customers aged from 40 to 65 and those with credit scores below 450 are more likely to quit the bank.

## 3.2 Feature evaluation

As mentioned in section 3.1, the data used in this paper have 13 independent variables X, and one dependent variable y (Exited). Among the features, RowNumber, surname, and customerID are irrelevant to the loss of the customer and can be deleted directly.

A simple correlation analysis is necessary before proceeding to a series of operations such as feature engineering and modeling. On the one hand, whether there is a high correlation coefficient between each feature is observed to avoid the occurrence of multicollinearity problems in the linear model. On the other hand, this step could calculate the correlation coefficient between independent and dependent variables to preliminarily determine which features are of better importance to y.

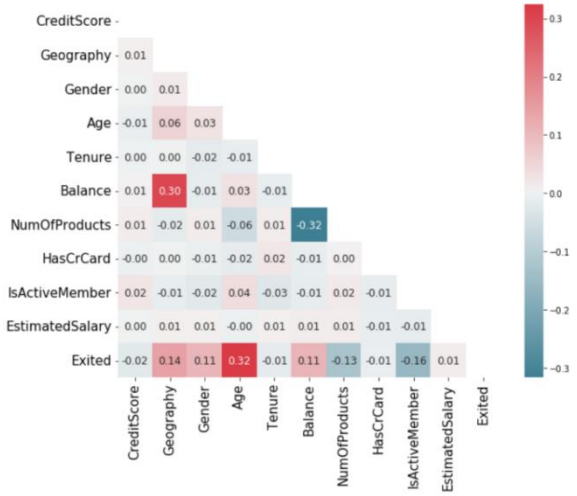


**Figure 2.** Correlation coefficients between original variables.

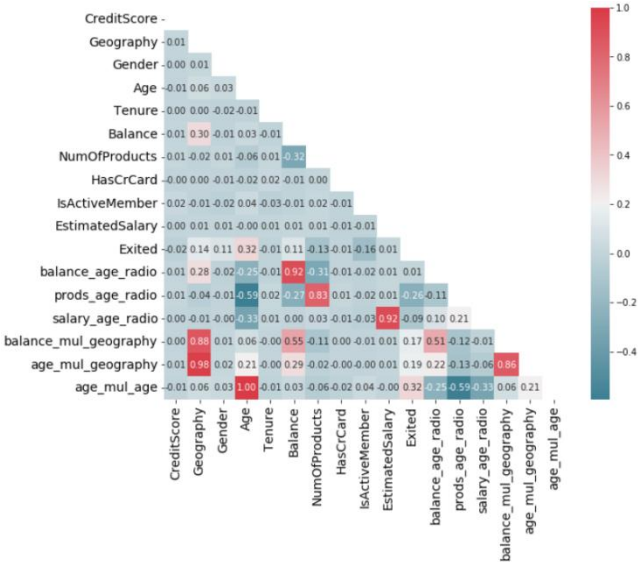


**Figure 3.** Correlation coefficients between present variables.

Fig. 2 is the Spearman correlation coefficient thermal diagram among 11 columns, in which the classification variable has been transformed into numerical variables through target encoding. According to the results, the highest correlation coefficient between independent variables is only 0.30, followed by 0.06, which shows little collinearity between independent variables. Moreover, Age, Geography, Balance, and Gender can be considered as four features that are more valuable to the dependent variable by observing the correlation coefficient between X and y. Thus, six new variables such as Age_mul_geography are established based on this conclusion, and the new inter-column correlation coefficient thermal map is shown in Fig. 3. As presented

in the diagram, the correlation between the derived and the original variables is very high, therefore, when applying linear models such as logistic regression and SVM linear kernel, the newly added variables should be removed. So far, there are 16 characteristic variables and 1 dependent variable ‘Exited’ in the data set for building the model.

Besides, this paper considers the five features of 'Geography', 'Gender', 'Tenure', 'HasCrCard', and 'IsActiveMember' as categorical variables and conducted one-hot coding. The remaining 11 variables were divided into numerical variables, treated by Min-Max scaling. The data set is divided into a training set and test set by 8:2, in which, only the training set is over-sampled using SMOTE method, and the test set remains untouched.

### 3.3 Modeling and Hyperparameter Optimization

In this paper, logistic regression with default parameters is selected as the baseline model. The baseline recall and precision scores obtained are 0.0.7226 and 0.3859, respectively. Subsequently, models of SVC, Random Forest, GDBT, and AdaBoost are selected to train the samples, and the grid search is used to tune the models.

During parameter tuning, several super parameters with the greatest influence are selected to conduct grid search according to the principle of ‘first large and then small. In this case, we first use "large mesh" to search the general range, and then use "small mesh" to search for more detailed parameters based on the previous searching range. The reasonable model parameters finally obtained are shown in Table 3.

Table 3 Parameters used for each model

| Model | Parameters |
|---|---|
| SVC | kernel = “linear” |
| | gamma = 0.001 |
| | C = 2000 |
| GBDT | learning_rate = 0.1 |
| | max_depth = 6 |
| | Subsample = 0.5 |
| RandomForest | max_depth = 15 |
| | min_samples_leaf = 1 |
| | min_samples_split = 5 |
| | n_estimators = 60 |
| AdaBoost | n_estimators = 60 |

### 3.4 Model evaluation

Precision, recall, F1 score, accuracy, and AUC were selected as the evaluation metrics of model performance. Table 4 summarizes the evaluation results. In real business, the cost of retaining a customer is usually lower than the cost of finding new users, and the success rate of the former is often higher. Hence, it is more critical to discover a customer who is about to lose than to misjudge a user who is retained. As reflected in the model indicators, a high recall rate should be the primary concern. The recall scores of logic regression, SVC, and AdaBoost are all above 0.71, which is better than other models. While with the small difference in the recall, the precision, F1 score and AUC values of AdaBoost are 0.514, 0.599, and 0.776, respectively,

which are significantly improved compared with the other two. Considering all these factors, AdaBoost is a better choice for this data set, which can highlight approximately 72% of all those who will churn.

Table 4 Model performance

| | precisio | recall | f1 | accuracy | roc_au |
|---|---|---|---|---|---|
| Logistic | 0.386 | 0.723 | 0.503 | 0.720 | 0.721 |
| SVC | 0.403 | 0.712 | 0.515 | 0.736 | 0.727 |
| GBDT | 0.478 | 0.697 | 0.567 | 0.791 | 0.756 |
| RF | 0.573 | 0.639 | 0.604 | 0.836 | 0.761 |
| AdaBoost | 0.514 | 0.718 | 0.599 | 0.811 | 0.776 |

### 3.5 Limitation

It can be seen from Table 3 that the precision of all trained models is quite low. In other words, the model has a high error kill rate. For example, for AdaBoost, a retained customer will be predicted as an exited customer about half of the time. Additionally, this paper does not focus on Neural Networks which may present a decent and better performance. Furthermore, the understanding of this data is limited, and there is no way to get in touch with the business people to combine the models with the actual business. To be specific, a total of 1537 churn customers still have a large amount of balance in their accounts, with the 25% and 75% quantile being 103,700 and 137,986. Since the percentage of these confused records in all entries is not a small number, the paper does not delete these data during the training process.

## 4 CONCLUSION

In summary, this paper investigates the bank customer churn based on 10000 pieces of data from Kaggle. According to the correlation coefficient between independent variables and dependent variables, 6 variables are created to participate in modeling. After oversampling, normalization and one-hot encoding, this paper used logistic regression, SVC, GBDT, Random Forest and AdaBoost to predict customer churn. Due to the relatively high recall (at 0.7176), high AUC score, and reasonable precision, AdaBoost is finally selected as the machine learning model for this business problem to make the prediction.

In real business, the selections of evaluation metrics of models are based on need according to the application scenarios, such as customer churn prediction. In such cases, more attention needs to be paid to the recall score, since the purpose of the model is to find the actual exited customers as more as possible. On this basis, the ability of finding more churn customers is the key indicators of the evaluation of the model. Therefore, to optimize the model in the future, improving the recall score should still be put at the priority. In addition, the way to construct more precisely model is of vital importance. Low precision makes it take more time for users to determine whether a customer is really going to be lost or not, and hence increases the bank's labor and marketing costs. Therefore, it is necessary to balance precision and recall in the future. Overall, these results offer a guideline for the bank churn prediction in the data processing, modeling process, and model selection.

## REFERENCES


[1] Expert Tips – NGDATA. “How Banks Can Gain Customer Insights: 28” online information, available at: https://ngdata.com/how-banks-can-gain-customer-insights/.

[2] P. Sidelov, “Machine & Deep Learning Forecasting for Banking Industry.” SDK.finance - White-Label Digital Core Banking Software, SDK.finance, 19 Feb. 2022, online information, available at: https://sdk.finance/machine-learning-deep-learning-forecasting-for-banking-industry/.

[3] K. Coussement, et al. “A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry.” Decision Support Systems, vol. 95, 2017, pp. 27–36.

[4] S. F. Slater, and J. C. Narver. “Intelligence Generation and Superior Customer Value.” Journal of the Academy of Marketing Science, vol. 28, no. 1, 2000, pp. 120–27.

[5] B. He, et al. “Prediction of Customer Attrition of Commercial Banks Based on SVM Model.” Procedia Computer Science, vol. 31, 2014, pp. 423–30.

[6] A. Keramati, et al. “Developing a Prediction Model for Customer Churn from Electronic Banking Services Using Data Mining.” Financial Innovation, vol. 2, no. 1, 2016.

[7] T. Bagheri, et al. “Using Data Mining for Prediction of Retail Banking Customers Churn Behaviour.” International Journal of Electronic Banking, vol. 2, no. 4, 2020, p. 1.

[8] B. Huang, et al. “Customer Churn Prediction in Telecommunications.” Expert Systems with Applications, vol. 39, no. 1, 2012, pp. 1414–25.

[9] C. C. Günther, et al. “Modelling and Predicting Customer Churn from an Insurance Company.” Scandinavian Actuarial Journal, vol. 2014, no. 1, 2011, pp. 58–71.

[10] A. Amin, et al. “Customer Churn Prediction in the Telecommunication Sector Using a Rough Set Approach.” Neurocomputing, vol. 237, 2017, pp. 242–54.

[11] A. de Caigny, et al. “A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees.” European Journal of Operational Research, vol. 269, no. 2, 2018, pp. 760–72.

[12] S. Sawangarreerak, and P. Thanathamathee. “Random Forest with Sampling Techniques for Handling Imbalanced Prediction of University Student Depression.” Information, vol. 11, no. 11, 2020, p. 519.

[13] J. A. Cook, and V. Ramadas. “When to Consult Precision-Recall Curves.” SSRN Electronic Journal, 2019.

[14] M. Lokanan, and S. Liu. “Predicting Fraud Victimization Using Classical Machine Learning.” Entropy, vol. 23, no. 3, 2021, p. 300.

[15] D. Salillari, and L. Prifti. “A Multinomial Logistic Regression Model for Text in Albanian Language.” JOURNAL OF ADVANCES IN MATHEMATICS, vol. 12, no. 7, 2016, pp. 6407–11.

[16] M. Aria, et al. “A Comparison among Interpretative Proposals for Random Forests.” Machine Learning with Applications, vol. 6, 2021, p. 100094.

[17] M. Jändel, “A Neural Support Vector Machine.” Neural Networks, vol. 23, no. 5, 2010, pp. 607–13.

[18] S. Bhattacharya, et al. “Comparison of Supervised and Unsupervised Approaches for Mudstone Lithofacies Classification: Case Studies from the Bakken and Mahantango-Marcellus Shale, USA.” Journal of Natural Gas Science and Engineering, vol. 33, 2016, pp. 1119–33.

[19] X. Bu, et al. “Distributed Filtering for Time-Varying Systems over Sensor Networks with Randomly Switching Topologies under the Round-Robin Protocol.” Neurocomputing, vol. 346, 2019, pp. 58–64.

[20] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine.” The Annals of Statistics, vol. 29, no. 5, 2001.
[21] J. A. Hanley, and B. J. McNeil. “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve.” Radiology, vol. 143, no. 1, 1982, pp. 29–36