Advances and Applications in Mathematical Sciences
Volume 20, Issue 12, October 2021, Pages 3049-3060

CHURN PREDICTION - A COMPARATIVE ANALYSIS WITH SUPERVISED MACHINE LEARNING ALGORITHMS

CHIKA K. GANGADHARAN[1], ROSHNI ALEX[2] and M. K. SABU[3]

[1,2]Department of Electronics
MES College Marampally, Kerala
E-mail: chika4895@gmail.com
roshnivalex@gmail.com

[3]Department of Computer Applications
Cochin University of Science and Technology, Kerala
E-mail: sabumk@cusat.ac.in

Abstract

Customer churn predictive model plays an indispensable role in all the industries since "churn is the rate at which the customers stop doing business with an organization". Machine Learning algorithms are used to build faultless models for prediction and classification. In this paper, a comparative analyzation of the performance of five different supervised machine learning algorithms namely Gaussian Naive Bayes, Support Vector Machine, *K* Nearest Neighbours, Decision Tree and Random Forest Classifiers in predicting churn is studied. Churn_Modelling dataset from Kaggle is used to test these classifiers. Experimental outcomes show that Random Forest Classifier outperforms all other algorithms in predicting the churn of a customer regarding accuracy, precision and recall.

1. Introduction

In the current circumstances, customer churn in the banking sector [1] is a big concern. This problem takes hold of dreadfully in the banking field. Preventing churn in companies lends a hand to develop the business by keeping a great extent of customers as possible. Hence customer churn prediction [2] is a requisite in every business. Thus we will get a clear idea about the customer exiting rate from the services of our business. The

2010 Mathematics Subject Classification: 68T07.
Keywords: Machine learning, supervised learning, churn modelling, support vector machine, random forest classifier.
Received October 7, 2020; Accepted October 27, 2020

customer churn prediction can be done with the aid of Machine Learning (ML). The types of machine learning are supervised learning [13], unsupervised learning [14] and reinforcement learning. We have chosen supervised learning since it is a prominent type of Machine Learning technique. In supervised machine learning our program learns some patterns from the data supplied, then this extracted knowledge is used for testing new cases. This is a recurring process. This will minimize the error in prediction and our model acquires enough knowledge from the training so that when an unseen data is given it will be able to classify the new data efficiently.

This literary work aims at evaluating the performance of five different machine learning algorithms. The algorithms we adopted for building the model are Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), K Nearest Neighbours (KNN), Decision Tree (DT) [5, 6] and Random Forest(RF) [7] Classifiers. Since churn prediction is a classification problem the performance metrics used for model estimation are accuracy, precision and recall. On the comparative analysis based on the three parameters, we observed that the Random Forest is predominant in terms of accuracy, precision and recall.

The paper is portrayed as, Section 2 focuses on the review of literature, Section 3 introduces Machine Learning Algorithms used in this paper, Section 4 gives the experimental approach and results and the paper is ended with Section 5.

II. Literature Review

This section offers an overview of different machine learning classifiers and their performance on various datasets.

Susmita Ray in her paper, highlights the virtues and flaws of machine learning algorithms from their application prospects, so that new learners can select the algorithm according to the specifications in their application. She has discussed nine algorithms and, to name some are SVM, KNN, DT, NB etc., which have contributed to our work also. She specifies that the Naïve Bayes gives good performance and it is easy to implement. For SVM, it handles both structured and semi structural data. Also with the use of appropriate kernel functions it can handle complex functions. Probability of

over fitting is less in SVM since generalization is adopted. KNN is a highly flexible classification algorithm and is finely suited for multi-model classes. Susmita identifies the advantages of Decision Tree as it is well suited for classification problems. It can easily handle both qualitative and categorical values, also capable of filling the attribute values which are missing with the most probable values. A Decision Tree is sometimes affected by the overfitting problem, in that case Random Forest which is based on the ensemble modelling approach is a better choice [5].

T. Vafeiadis et al., in their paper [8] gave an analyzation based on the comparison of various classification algorithms used for predicting the churn. For modelling, they use classifiers namely Naive Bayes, Artificial Neural Network (ANN), Decision Tree, Support Vector Machine and logistic regression. To enhance the performance of algorithms they use the popular boosting algorithm AdaBoost. Then they compare the results with the boosted version and non-boosted version. Their result shows that the models with boosted versions are superior to the non-boosted ones. SVM with Ada Boost showed best performance with 84% of F-measure and 97% accuracy [8].

Arno De Caigny et al. in the paper proposes an algorithm named Logit Leaf Model (LLM) for classifying the data more effectively. The LLM algorithm, rather than taking the entire dataset for modelling, builds several models using segmented dataset. To measure the performance of the predictive model they have considered the TOP Decile Lift (TDL) and AUC-Area under the receiver operating characteristics (ROC) curve, which shows a better result for LLM on Decision Trees and Logistic Regression [9].

Manjula C. Belavagi et al. in their paper [10] attempted to find the best model in predicting the intrusion detection in network data traffic. For classification, they use algorithms such as Gaussian Naive Bayes, Logistic Regression, Support Vector Machine and Random Forest. Plotting the reliability curve, the best result was shown by Random Forest classifier. Quality of the classifiers is analysed by the ROC curve. By observing the obtained graphs, they wrap up that in identifying attacks the Random Forest classifier has the low FPR and high TPR [10].

Dana Bazazeh and Raed Shubair have compared the RF, SVM and Bayesian Networks. The dataset used was the Wisconsin original breast

cancer dataset. They found out that based on the methods used, the classification performance of each algorithm was varied. From the results obtained Support Vector Machine classification showed the highest performance in terms of accuracy, specificity and precision [11].

Xia Guo-en, Jin Wei-dong in their paper uses structural risk minimization method on SVM for predicting Customer Churn. This method is applied with decision tree, logistic regression, ANN and Naive Bayes classifier. By comparing the results, they found SVM owns the highest accuracy rate, lift coefficient, hit rate and covering rate. Hence, they concluded that SVM is an efficient algorithm for churn prediction [12].

Muhammad Zain Amin and Amir Ali in their paper [13] studied the performance of different supervised machine learning classifiers used for predicting Healthcare Operational Decisions. These algorithms are evaluated by the performance evaluation factors such as accuracy, F1 rate, precision, MCC rate, ROC Area. On the comparative study, they concluded that by predicting 95 cases correctly both RF and KNN achieved the greatest accuracy rates [13].

Priyanka S. Patil et al. in their paper [14], has evaluated the performance of Artificial Neural Networks when applied to two different banking datasets. The datasets they used for modelling are the German credit dataset (dataset1), for fraud detection problems and another dataset (dataset2) for customer retention problems. They included all components required for an ANN. In training, they include two phases, feed-forward phase, and backpropagation phase. After that, their model is ready for prediction. Their algorithm gives 72% and 98% accuracy for dataset1 and dataset 2 respectively [14].

Hemlata Dalmia et al. have discussed the application of the KNN algorithm together with XGBooster algorithm for better accuracy. To check the classifier performance accuracy, error rate, sensitivity and specificity are considered. XGBooster showed an accuracy of 87% while KNN gave 84%. XGBooster has given high sensitivity and specificity and low error rate compared to KNN algorithm [15].

Ionut Brandusoiu, Gavril Toderean have applied the SVM algorithm with four kernel functions to implement the predictive model. They have used a

dataset of 3333 call details records containing 21 attributes for each record. The model with Polynomial kernel produced an overall accuracy of 88.56% [16].

To identify churn customers, Irfan Ullah et al. propose a predictive model using classification and clustering techniques. They have applied the feature selection methods like information gain (IG) and correlation attribute evaluation technique. For building the model, Decision tree algorithm, Random forest, Decision Stump and Random tree with 10-fold cross-validation are used. Compared to other algorithms Random Forest produced a better result of 88.6% accuracy. To pick out the basic causes of churn, customer profiling by employing k- means clustering is performed, so that the company can improve the business strategies [17].

In the paper written by A. Mishra et al., the authors have compared the classifiers that are conventional with Ensemble-based ones. Both types of algorithms performed well, having an accuracy of around 90% and Random Forest classifier outperformed with a low error rate and greater accuracy of 92% [17].

Rajamohamed R. and Manokaran, J., have presented a predictive model for retaining customers in credit card churn prediction. In that, after preprocessing the dataset is split into clusters using unsupervised methods. To build the predictive models, supervised methods are used. Upon evaluation of performance metrics, the combination of SVM with the rough k-means clustering algorithm worked better with an accuracy of 96.85% [18].

In the paper, Dr B. Valarmathi et al. have selected a marketing data set of bank which is imbalanced. They have applied dimensionality reduction using CfsSubsetEval method and then the model is built using the Naive Bayes, J48, KNN and Bayes net. The results of performance evaluation unveiled that J48 alias Decision Tree scored the highest accuracy of 89% and 91.2% respectively with and without dimensionality reduction [19].

Review of the above papers shows that better performance is given by Random Forest Classifier and SVM on various datasets. The kNN, Decision Tree and Naive Bayes also performed well so that we chose these classification algorithms for our experiment. The various performance evaluation parameters used in the preceding papers are accuracy, specificity,

f1 score, precision, error rate. To evaluate the performance characteristic, a confusion matrix model was chosen [17]. A confusion matrix is a table that is used to illustrate the performance of a classification model on a test dataset for which the true values are known. From the confusion matrix the value of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) can be evaluated [20]. Then applying these values in the equation given below the performance measure of the model can be calculated.

$$\text{Accuracy} = (TP+TN)/\text{Total} \quad (1)$$

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

$$\text{Specificity} = TN/(TN+FP) \quad (4)$$

$$\text{Error rate} = 1 - \text{Accuracy} \quad (5)$$

Based on this study we inferred to use the accuracy, precision and recall metrics for our comparative analysis. The upcoming section concisely describes the machine learning algorithms used in this paper.

III. Machine Learning Algorithms

The in view of this paper is to evaluate the computational performance of different classifiers using a single dataset. By the literature review completed, we aim to investigate the performance of the classifiers specifically Gaussian Naïve Bayes, Support Vector Machine, *K* Nearest Neighbour, Decision Tree [5] and Random Forest Classifier. For modelling, we have used the Churn_Modelling dataset, which has been fetched from the Kaggle. This data set gives information about the customer data from a bank which shows the clients that are likely churn and not likely to churn. The following algorithms are executed in this data set and results are generated.

A. Gaussian Naïve Bayes (GNB) Classifier

Gaussian Naïve Bayes (GNB) algorithm is based on the Bayes Theorem [18, 21] in Probability and Statistics which focuses on determining the probability of an event occurring based on prior knowledge of conditions that might be related to the event. The general equation for Bayes theorem is

given in equation (6):

$$P(H/E)=(P(E/H)*P(H))/P(E) \quad (6)$$

In equation (6), P(H/E) is the conditional probability of an event H, given another event *E* is true and vice versa. P(H) and P(E) are the probabilities of events *H* and *E* respectively [22]. Gaussian Naïve Bayes (GNB) is the simplest classifier having the assumption that the data from each label is drawn from a simple Gaussian distribution

A. Support Vector Machine (SVM) Classifier

An SVM model represents different classes in a hyperplane in multidimensional space [5]. The SVM generates this hyperplane iteratively to minimize the errors. The goal of this is to divide the datasets into classes to find a maximum marginal hyperplane (MMH). Large margin is considered a good margin and a small margin as a bad margin [11]. Support vectors are the data points, which are closest to the hyperplane. These points will define the separating line better by calculating margins. These points are more relevant to build the classifier [20]. The linear SVM classifier model predicts the class of a new instance x by simply computing the decision function $w^T x+b=w_1\times 1+\ldots+w_n x_n+b$: if the result is positive, the predicted class ŷ is the positive class (1), or else it is the negative class (0) [20]. Many possible hyperplanes could be chosen to separate the two classes of data points. Our objective is to find a plane that has the maximum margin. To define the hyperplane, we can use equation (7). Maximum margin can be obtained from minimizing the value of weight vector w.

$$w, b \frac{1}{2}\|w\|^2$$

$$\text{s.t} y_i(w\cdot x_i+b)\geq 1, \forall x_i \quad (7)$$

B. K Nearest Neighbour (KNN) Classifier

K Nearest Neighbour (KNN) is another versatile supervised learning algorithm can be used for both regression and classification [5]. This classification uses Euclidean distance [23] to find feature similarity for prediction. When new data is given to the model it goes through all the original samples and finds the one that is most similar to the new data and

uses its class. One of the disadvantages of KNN is that prediction is slow if the dataset is large [5, 24].

B. Decision Tree Classifier

Decision tree structure can be thought of as a flow chart like tree construction with features represented as internal nodes, the decision rules as branches and outcomes as each leaf nodes. The top node also known as the root node is an attribute and based on the value of that attribute, we can partition the dataset. The partition is done in a recursive manner to be continued until we get a partition without confusion. This process is continued until we get a leaf node. This is the termination condition in the case of constructing a decision tree [5, 24].

C. Random Forest Classifier

The decision trees always show a tendency to overfit the training data, for which random forest is a remedy. A cluster of decision trees can be named as a random forest, in which each tree may differ from the other and will do the prediction. The tree with highest probability is taken for the final prediction [20].

These classifiers are used in the subsequent section which discusses the experimental approach.

IV. Experimental Analysis and Results

The paper mainly emphasizes the comparative study of machine learning algorithms. The assessment of the algorithms is accomplished using the performance metrics on the mentioned dataset. Figure 1 shows the methodology of the decision-making system.

Figure 1 can be illustrated in four steps that are required to be performed for predictions: data preprocessing, building the model, testing the dataset, and evaluating the performance of machine learning algorithms.

In our experiment, first, the data set is imported into the program. We are using the Churn_Modelling_dataset. Based on the features in the dataset, it gives a target variable which tells whether a customer has a chance to stay in the bank or not. It has twelve independent variables (features) and one

target variable.

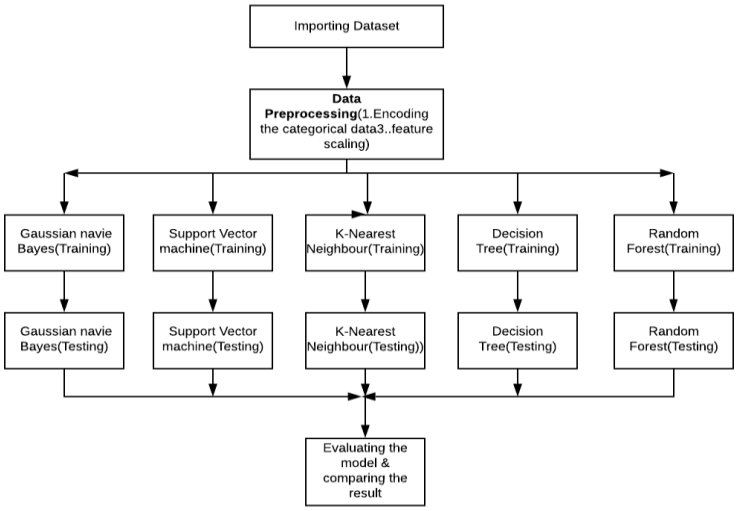

Figure 1. The predictive model.

Data preprocessing is performed after importing the dataset for obtaining a quality data for building the model. Categorical data encoding and feature scaling are performed in data preprocessing to build our model. The categorical features are encoded to a numerical array and the target labels are encoded to a natural number between 0 and n-1. We use Label Encoding() [25] and One Hot Encoder() [26] to encode the categorical data. The preprocessed dataset is divided into training set and testing set. Data preprocessing can be standardized using Feature Scaling [27] method. Standardization involves rescaling the features such that they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one [28]. To develop the models the ML algorithms are applied after the preprocessing of data. While completing the training with machine learning algorithms the model is tested against the test dataset. The performance of each algorithm is measured in terms of accuracy, precision and recall using equations (1), (2) and (3) respectively. The confusion matrix is also taken in each case as the accuracy alone can sometimes be misleading. Now the model is ready for prediction.

Result analysis of each of the classifiers is discussed below. The results of

the experiment are plotted against the five machine learning algorithms on x-axis and performance evaluating factors on the y-axis in figure 2. Table 1 gives the comparison metrics of the classifiers.

Table 1. Comparison metrics.

	GNB	SVM	KNN	DT	RF
Accuracy	81	86	83	80	87
Precision	79	85	81	81	86
Recall	81	86	83	80	87

Figure 2. Performance evaluation of different classifiers.

Five different classification algorithms were tested for their performance. Observing figure 2, it can be wrapped up that the Random Forest Classifier has the highest accuracy. SVM Classifier also gives an accuracy which is not much less than the performance of Random Forest Classifier. Best results on Precision and Recall are also shown by the Random Forest Classifier.

V. Conclusion

ML techniques have been extensively used in banking and business sectors and have served as a useful predicting tool that helps the organization in analyzing the available data and taking necessary action to avoid the customer churn. Our work elucidates five supervised machine learning algorithms namely Gaussian Naive Bayes, Support Vector Machine, K-Nearest Neighbour, Decision Tree Classifier, Random Forest, used for customer churn prediction. We have measured accuracy, precision and recall for each algorithm. Simulation results obtained show that Random forest Classifier is the best performing supervised machine learning algorithm with an accuracy of 87%, precision 86% and recall 87% by using the default parameter values on Chrun_Modelling dataset.

References

[1] P. Hemalatha and G. M. Amalanathan, A Hybrid Classification Approach for Customer Churn Prediction using Supervised Learning Methods: Banking Sector, Proc. - Int. Conf. Vis. Towar. Emerg. Trends Commun. Networking, ViTECoN 2019, pp. 1-6, 2019, doi: 10.1109/ViTECoN.2019.8899692.

[2] K. G. M. Karvana, S. Yazid, A. Syalim and P. Mursanto, Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry, 2019 Int. Work, Big Data Inf. Secur. IWBIS 2019, pp. 33-38, 2019, doi: 10.1109/IWBIS.2019.8935884.

[3] G. Santafe, I. Inza and J. A. Lozano, Dealing with the evaluation of supervised classification algorithms, Artif. Intell. Rev., vol. 44, no. 4, pp. 467-508, 2015, doi: 10.1007/s10462-015-9433-y.

[4] L. Bandura, A. D. Halpert and Z. Zhang, Machine learning in the interpreter's toolbox: Unsupervised, supervised, and deep learning applications, 2018 SEG Int. Expo. Annu. Meet. SEG 2018, pp. 4633-4637, 2019, doi: 10.1190/segam2018-2997015.1.

[5] S. Ray, A Quick Review of Machine Learning Algorithms, Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35-39, 2019, doi: 10.1109/COMITCon.2019.8862451. [6] R. (Eds. Kalita, J. Balas, V.E., Borah, S., Pradhan, Recent Developments in Machine Learning and Data Analytics, Springer Singapore, 2019. https://www.springer.com/gp/book/9789811312793.

[7] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector, IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[8] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis and K. C. Chatzisavvas, A comparison of machine learning techniques for customer churn prediction, Simul. Model. Pract. Theory, vol. 55, pp. 1-9, 2015, doi: 10.1016/j.simpat.2015.03.003.

[9] A. De Caigny, K. Coussement and K. W. De Bock, A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, Eur. J. Oper. Res., vol. 269, no. 2, pp. 760-772, 2018, doi: 10.1016/j.ejor.2018.02.009.

[10] M. C. Belavagi and B. Muniyal, Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection, Procedia Comput. Sci., vol. 89, pp. 117-123, 2016, doi: 10.1016/j.procs.2016.06.016.

[11] D. Bazazeh and R. Shubair, Comparative study of machine learning algorithms for breast cancer detection and diagnosis, Int. Conf. Electron. Devices, Syst. Appl., pp. 2-5, 2017, doi: 10.1109/ICEDSA.2016.7818560.

[12] G. E. Xia and W. D. Jin, Model of customer churn prediction on support vector machine," Xitong Gongcheng Lilun yu Shijian/System Eng. Theory Pract., vol. 28, no. 1, pp. 71-77, 2008, doi: 10.1016/s1874-8651(09)60003-x.

[13] M. Z. Amin and A. Ali, Performance Evaluation of Supervised Machine Learning Classifiers for Predicting Healthcare Operational Decisions, no. 1, pp. 1-8, 2017, doi: 10.13140/RG.2.2.26371.25127.

[14] P. S. Patil and N. V. Dharwadkar, Analysis of banking data using machine learning, Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017, pp. 876-881, 2017, doi: 10.1109/I-SMAC.2017.8058305.

[15] H. Dalmia, C. V. S. S. Nikil and S. Kumar, Churning of Bank Customers Using Supervised Learning, Lect. Notes Networks Syst., vol. 107, pp. 681-691, 2020, doi:

10.1007/978-981-15-3172-9_64.

[16] I. Brandusoiu and G. Toderean, Churn Prediction in the Telecommunications Sector Using Support Vector Machines, Ann. ORADEA Univ. Fascicle Manag. Technol. Eng., vol. XXII (XII), no. 1, 2013, doi: 10.15660/auofmte.2013-1.2772.

[17] A. Mishra, A Comparative Study of Customer Churn Prediction Classifiers, no. Icici, pp. 721-725, 2017, doi: 10.1006/brln.2000.2429.

[18] R. Rajamohamed and J. Manokaran, Improved credit card churn prediction based on rough clustering and supervised learning techniques, Cluster Comput., vol. 21, no. 1, pp. 65-77, 2018, doi: 10.1007/s10586-017-0933-1.

[19] B. Valarmathi, T. Chellatamilan, H. Mittal, J. Jagrit and S. Shubham, Classification of imbalanced banking dataset using dimensionality reduction, 2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019, no. Iciccs, pp. 1353-1357, 2019, doi: 10.1109/ICCS45141.2019.9065648.

[20] A. C. Mu ller and S. Guido, Introduction to Machine Learning with Python and Scikit-Learn. 2015.

[21] H. Altwaijry and S. Algarny, Bayesian based intrusion detection system, J. King Saud Univ. - Comput. Inf. Sci., vol. 24, no. 1, pp. 1–6, 2012, doi: 10.1016/j.jksuci.2011.10.001.

[22] M. Pirooznia, J. Y. Yang, M. Q. Qu and Y. Deng, A comparative study of different machine learning methods on microarray gene expression data, BMC Genomics, vol. 9, no. SUPPL. 1, pp. 1-13, 2008, doi: 10.1186/1471-2164-9-S1-S13.

[23] J. Pamina et al., An effective classifier for predicting churn in telecommunication, J. Adv. Res. Dyn. Control Syst. 11(1) (2019), 221-229.

[24] M. A. Hassonah, A. Rodan, A. K. Al-Tamimi and J. Alsakran, Churn Prediction: A Comparative Study Using KNN and Decision Trees, ITT 2019 - Inf. Technol. Trends Emerg. Technol. Blockchain IoT, no. 1, pp. 182-186, 2019, doi: 10.1109/ITT48889.2019.9075077.

[25] R. Guedrez, O. Dugeon, S. Lahoud and G. Texier, Label encoding algorithm for MPLS Segment Routing, Proc. - 2016 IEEE 15th Int. Symp. Netw. Comput. Appl. NCA 2016, pp. 113-117, 2016, doi: 10.1109/NCA.2016.7778603.

[26] K. Potdar, T. S., and C. D., A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers, Int. J. Comput. Appl., vol. 175, no. 4, pp. 7-9, 2017, doi: 10.5120/ijca2017915495.

[27] Y. Fu, P. Chen, S. Yang and J. Tang, An Indoor Localization Algorithm Based on Continuous Feature Scaling and Outlier Deleting, IEEE Internet Things J., vol. 5, no. 2, pp. 1108–1115, 2018, doi: 10.1109/JIOT.2018.2795615.

[28] Pedregosa et al, Scikit-learn: Machine Learning in Python, JMLR 12, 2011. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html.