Customer Churn Prediction on Credit Card Services using Random Forest Method

Xinyu Miao[1, *, †], Haoran Wang[2, †]

[1]*Beijing University of Technology, Beijing, 100000, China.*
[2]*Hefei University of Technology, Hefei, 230000, China.*
[*]*Email: xinyu.miao@ucdconnect.ie*
[†]*These authors contributed equally.*

ABSTRACT
With the continuous development of the Internet, more and more people are spending money using credit cards online, therefore, retaining customers in order to maintain profit margin becomes very important for many banks. This paper aims to make predictions on credit card customer churn through machine learning methods and to provide feasible solutions to deal with customer churn issue based on the results. Three models including Random Forest, Linear Regression and K-Nearest Neighbor (KNN) are applied to a dataset which contains more than 10000 pieces and 21 features. By tuning hyperparameters and evaluating models based on ROC & AUC and confusion matrix, it is concluded that Random Forest has the best performance with its accuracy reaching 96.25%. Total transaction amount in the last 12 months, total transaction count in the last 12 months and total revolving balance are the top three important features which have the significant impacts on the customer churn prediction. It shows that the more frequent customers use their credit cards, the less likely they are to leave, and by using this model, bank managers can proactively take actions to fight against customer churn.

Keywords: *credit card, customer churn, random forest, machine learning.*

1. INTRODUCTION

With the rise of Internet in the past decade, great changes have taken place in the market environment and user habits, resulting in the sharply increased use of credit cards [1]. The cumulative card issuing volume has reached 588 million in 2017 compared to 465 million in 2016, achieving an increase of 26.45% [1]. Since credit card plays an important part in banks' profit, many banks work very hard to offer better services and products. Competition between banks is really fierce because products are kind of homogeneous, so customers actually have many options and can compare various banks based on their past experiences of being served [2]. Therefore, many banks begin to realize the importance of customers and pay attention to customer relationship management (CRM) [3].

It is the case that getting new users is much more expensive and difficult than avoiding customer loss, given that the cost of selling to new customers is five times the cost of additional sales to existing customers [3]. Therefore, customer churn, becomes one of the main focuses of many banks. It is true if we look at AARRR or HEART framework used by many banks [4-5], both have some metrics to help them make decisions, and there is one metric called retention, i.e., to look at the n-day retention, monthly active users or session frequency. Studies have shown that a bank can increase its profits by 85% when the retention rate increases 5% [3].

The aim of this paper is to predict customer churn, once successfully predicted, banks can have enough time to proactively take actions to retain customers, by offering better services or giving more attractive discounts. As a consequence, it is of great significance, especially in today's world, when there are a lot of data related to customers, and with the spread use of big data, massive users' data have become valuable treasures for enterprises.

There are some paper discussed customer churns in the past, but most of them failed to use a specific collection of datasets or apply machine learning models, which is a branch of computer science, and is widely used in commercial applications with its goal to let computers "learn" without directly programmed [6]. With machine

learning approaches, it is possible to process and analyze large amounts of data. Some other paper, although used models to predict the outcome, mainly focus on unsupervised learning, which, in general, are not reliable and have relatively poor interpretability [7].

In this paper, we obtain credit card holders' information data from Kaggle, which has more than 10000 pieces of data and 21 various features. We do the exploratory data analysis to see its distribution and visualize relationships between features. Then we split the dataset into training and testing, followed by standardization. Three models are used including Random Forest, Logistic Regression and KNN to see their performance. After that, a method called grid search is used to tune hyperparameters. In order to select the optimum model, we use confusion metrics and ROC & AUC to evaluate models. It is concluded that Random Forest performs the best among the three models, and it is approximately 5% higher in accuracy and much higher in other metrics such as prediction and recall. We rank the importance of features based on Random Forest model and select the top three important features which exert meaningful impacts on banks' decision making. Specifically, they are the total transaction amount in the last 12 months, the total transaction count in the last 12 months and total revolving balance on the credit card. Therefore, we learn that the more frequent one customer uses his/her credit card, the less likely he/she will get churned, which is intuitive and makes sense in daily life, and the bank managers can adjust the card service accordingly.

In the rest of the paper, the methods used are discussed first, followed by introduction of dataset, preprocessing steps and results of applying models. After that, three conclusions are introduced in the end with some innovations and a touch of shortcomings and how we can further improve.

2. METHOD

2.1. Random Forest

Random forest is a supervised learning model. It was proposed by Breiman and Cutler in 2001, and is based on decision tree and ensemble learning [8]. Decision tree can describe complicated relationships between x and y rather than simple linear relationship, thus has a stronger modeling strength. However, a single tree model is very sensitive to the training set data, therefore is very likely to cause overfitting problem [9]. Ensemble learning, however, can solve this problem by a method called bagging, which is to train multiple learners, with each one's training data comes from a collection of bootstrapped samples selected randomly from original dataset with replacement. It decreases variance by introducing randomness into model framework, making the model more robust and the result more accurate and convincing. Specifically, each tree learns independently from random sub-dataset and sub-features, and the final outcome is drawn with the help of deterministic averaging process, or in other words, the average of predictions of individual trees [10]. A simple example of random forest tree is shown below as Figure 1.

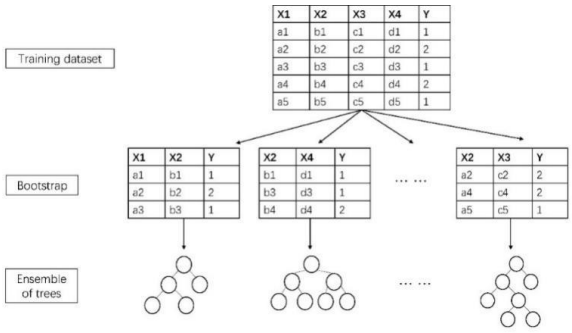

Figure 1. Description of Random Forest. *Notes*: X1 to X4 are all features, with Y is the outcome that needs to be predicted. The original dataset undergoes splits to several sub-sets with each one contains less features and less data. Then, each sub-set will be used to train a tree to make prediction, and deterministic averaging process is applied to determine the final result.

2.1.1. Random Forest Algorithm

Input: Dataset (*D*) with *N* features and numbers of trees *n*.

Output: A random forest.

For *i*=1 to *n*:

first step: Draw a bootstrap sample from original dataset *D*.

second step: Grow a random forest tree to the bootstrap data, and repeat the following steps until the minimum node size is reached.

(1) Select a subset of $\sqrt{N}$ features (variables).

(2) For *j*=1 to $\sqrt{N}$, pick the best variable from $\sqrt{N}$ and split the node into left and right child nodes.

To make a binary classification prediction of a new point x, we can use this formula with denotes the state prediction of the *m*-th random forest tree.

$$\hat{C}_{rf}^{n}(x) = \text{majority vote}\{\hat{C}_{m}(x)\}_{1}^{n} \quad (1)$$

2.2. Logistic Regression

Logistic Regression is a linear model, which connects X_1 ,…, X_p to the conditional probability $P(Y = 1|X_1, \ldots, X_p)$ through this formula:

$$P(Y = 1|X_1, \ldots, X_p) = \frac{exp(\beta_0+\beta_0X_1+\cdots+\beta_pX_p)}{1+exp(\beta_0+\beta_0X_1+\cdots+\beta_pX_p)} \quad (2)$$

where Y stands for the binary outcome that we are interested, and $X_1, \_? X_p$ are the features. $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients, which are derived from a method called maximum-likelihood using the dataset [11]. For a new instance, all the βs are replaced by their calculated counterparts and *X*s by their realizations, and if the probability *P* is greater than a threshold value, the new instance will then be assigned to *Y*=1 class accordingly. Otherwise, it will be assigned to *Y*=0. Usually, the threshold is set to be 0.5 and therefore is so-called Bayes Classifier.

2.3. K Nearest Neighbors

K Nearest Neighbors takes less time than the other and therefore is a relatively simple model. It conducts predictions straightforward from training set data, by calculating the closest *k* objects on distance of dataset to the input data, where *k* is the hyperparameter and can be adjusted to affect the classifier performance, and it then assigns classification based on maximum voted classes out these adjacent classes [12].

There are many ways to calculate the distance, for example, Euclidean distance and distance Manhattan, with the former one the most popular. The distance *d* between two points *a* and *b* can be calculated through the formula below:

$$d(a,b) = \sqrt{\left(\sum_{i=1}^{N}(a_i - b_i)^2\right)} \quad (3)$$

The picture below shows the principle of K Nearest Neighbors.

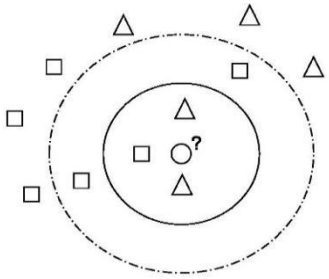

Figure 2. Sketch map of K Nearest Neighbors. *Notes*: The small circle in the middle stands for the newly input data which needs to be classified, whereas other small triangle and square represent original binary types' data, and we can assign the circle to either type according to the model. If we set k to 3, we will look at the inner solid line circle and assign my input circle to triangle, if k is 5 on the other hand, it will be assigned to square.

3. DATA AND EXPLORATORY DATA ANALYSIS

3.1. Basic Information about the dataset

We find relevant dataset about bank customer information from Kaggle, which consists of more than 10000 pieces and includes 21 features such as age, income, marital status, credit card limit and so on. 16 of them are numerical while 5 of them are categorical.

3.2. Exploratory Data Analysis

We conduct exploratory data analysis (EDA) to have a better understanding of the data by checking for missing and duplicated values, handling outliers, visualizing distributions and plotting graphs to see the relationships between features and our target, which is whether the customer get churned. Here are some important features that needed to illustrate.

3.2.1. Type of Card

It can be seen from the below table that the type of card held by majority of people is blue card with 93.2%. In the Figure 3, we split the data into two parts, thus clearly, we can visualize the relationships between card type and both currently existing customer and left customers. These two follows the same pattern, with the amount of blue card holders extremely surpasses the others.

Table 1. Proportion of different card categories

type	percentage
Blue	93.2%
Silver	5.48%
Gold	1.15%
platinum	0.197%

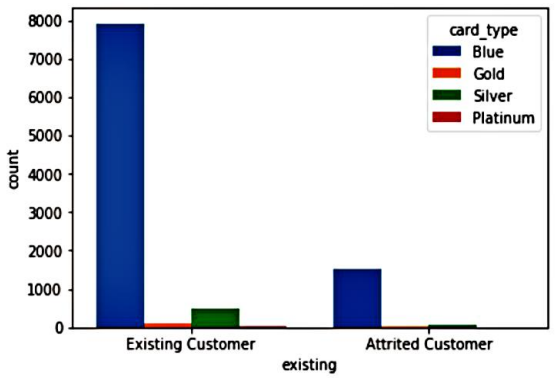

Figure 3. Relationship between customers and card type.

This kind of data distribution is not informative because it is unbalanced and thus, we cannot explain the

relationship between card type and our target. Therefore, this feature is deleted in the further analysis.

3.2.2. Credit Card Limit

The credit card limit is analyzed to see whether there are some extremely large values, or outliers. For example, if some people get a card limit of 1-2 million, which are significantly larger than the others, then we need to delete these data. Fortunately, there are no outliers, and all the limits are within a reasonable range.

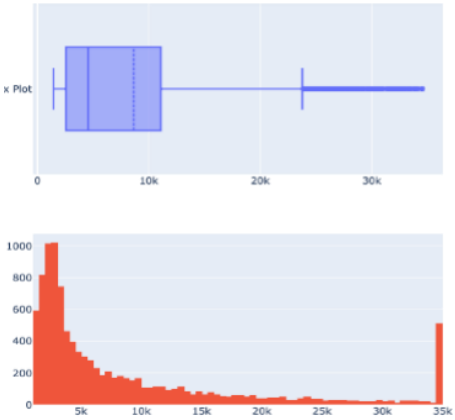

Figure 4. Distribution of the card limit.

3.2.3. Number of Products held by customers

The number of bank products held by customers also has an impact on our research.

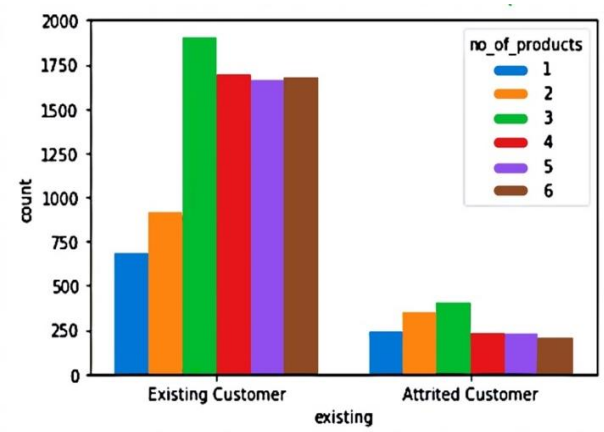

Figure 5. Relationship between customers and No. of products held.

As can be seen from Figure 5, customers with 4, 5, 6 numbers of products accounts for most of the existing personnel, and their losing rate is also lower than that of customers with only one or two products. It shows that in the future marketing process, perhaps it is a good idea for banks to implement bundle marketing.

After conducting exploratory data analysis of features, we gain deeper insights of our data and some results are shown in the table below.

Table 2. Comment of EDA.

Features	Notes
Education level	70.65% of customers gained a high school or higher education.
Marriage statues	Almost half of the customers are married, and single customers' number is up to 40%.
Income level	People with income of 40k-60k may be potential customers for banks.
Number of products held	Banks could give priority to bundling sales.
Card type	The distribution of the card type is too unbalanced to help in predicting whether a customer will get churned.

4. PREPROCESSING, APPLYING MODELS AND RESULTS

4.1. Data preprocessing

4.1.1. Label Encoding

Since the input data of the model need to be numerical numbers, thus we use label encoding to transfer all the categorical values into numerical, for example, for the gender feature, we use 0 to label female and 1 for male. One hot encoding is also a popular method, but it will create too many features in this case, making it more difficult and more time-consuming to find the results.

4.1.2. Correlation Matrix

Pearson correlation matrix can give us the information about relationships between features. It can help us do feature selection by removing some highly correlated features in the model training step. Here we plot correlation graphs of categorical features in our dataset.

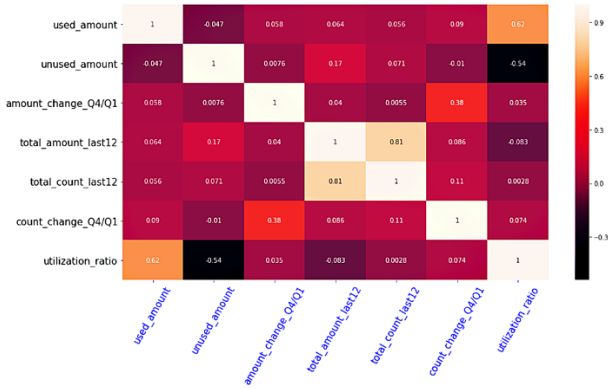

Figure 6. Pearson correlation graph of the first half categorical features.

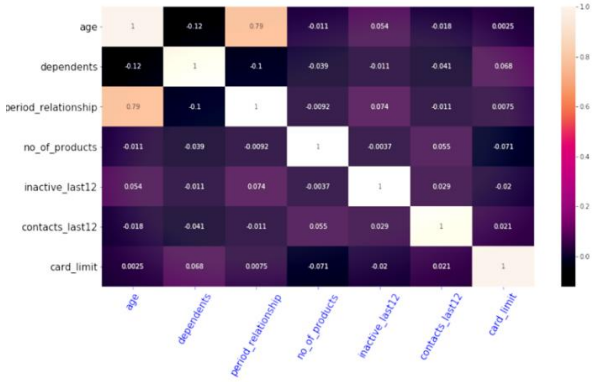

Figure 7. Pearson correlation graph of the second half categorical features.

Values on both the vertical and perpendicular axis are our features, with the data in the middle shows the relationships between the corresponding axes. The color, obviously, reflects the strength of the correlation, and the lighter the color, the higher the positive correlation between two features, while the darker the color, the higher the negative correlation.

4.2. Modelling

4.2.1. Testing Models

We use 5-fold cross validation to test the performance of Random Forest, Linear Regression and KNN, and get the evaluation results of these three models as below.

Table 3. Accuracy score.

model	result of 5-fold cross validation
Random Forest	0.9610
Logistic Regression	0.9036
K-Nearest Neighbor	0.9032

Notes: Accuracy score is the accuracy of model prediction.

Among them, the model accuracy of Random Forest is the highest with 96.1%, and the rest two are very similar with 90.36% and 90.32% respectively. Therefore, we can choose to use Random Forest to make predictions, however, for comparison purposes, we still train the other models. We split the dataset into training and testing, with the formal contains 80%, and then use the training set data to find parameters of three models.

4.2.2. Optimal Hyperparameters

Hyperparameters are man-made parameters, and can exert significant impacts on the performances of models. If not fit, the model will either show weak prediction strength or cause overfitting problem. In this paper, we use grid search method to adjust parameters.

The results are shown in the table 4.

Table 4. Accuracy score.

Model	Result of 5-fold cross validation	After tuning hyperparameters
Random Forest	0.9610	0.9568
Logistic Regression	0.9036	0.9052
K-Nearest Neighbor	0.9032	0.9042

It can be found that the accuracy score on testing set data of Random Forest reached 95.68% after parameter adjustment, which is still the highest compared to the other two.

4.2.3. Model Comparison

To test the performance of models, we use ROC curve, AUC value and confusion matrix.

4.2.3.1. Confusion Matrix

The confusion matrix, also called error matrix, is a standard format for the accuracy evaluation and is represented by a matrix of 2 rows and 2 columns as it shown in figure 7 [13].

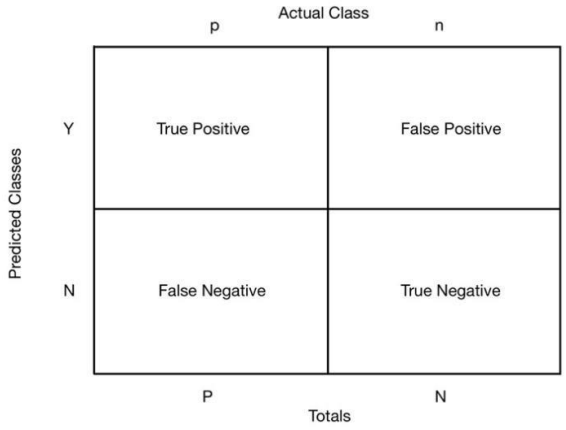

Figure 8. Confusion matrix.

Positive and negative refer to the result of model on whether the customer gets churned, while true and false indicate whether our model predicts correct. For example, true positive in the upper left means our model predicts this customer will leave our services and it is true. Similarly, we can define the other three matrices using the same rule.

We choose recall ratio to analysis model performances. It demonstrates how many churned customers are successfully predicted.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (4)$$

Since our goal is to predict consumer churn as many as possible, the recall ratio is the best matrix among others. Large proportion reflects high prediction accuracy, and the results are shown in table 5.

4.2.3.2. Receiver Operating Characteristics Curve (ROC curve)

Threshold is an important hyperparameter in ROC curve. It is a standard which helps to determine whether a customer will get churned. The result of our model is a numerical number lies in the range of 0 to 1, therefore, those under the threshold will be predicted staying while those surpass will be predicted churned.

Since it is a hyperparameter, we can set its value to adjust the strictness of this method. Specifically, the higher the threshold, the more difficult a customer to be assigned to lost customer, the higher the accuracy of this result is. Points on the ROC curve show the true positive rate and false positive rate achieved by specific decision thresholds, and the monotone curve is obtained by scanning all possible decision thresholds, and the area under the curve (AUC) corresponds to the proportion of correctly arranged positive and negative sample pairs [14]. Peter et al. had found that if we want to consider only the optimal threshold, then the higher AUC is, the stronger the predicted ability the model will have [15]. The below table clearly shows that Random Forest has the highest AUC which means this model have the best prediction ability.

Table 5. AUC of rocand recall ratio.

	Random Forest	Logistic Regression	K-Nearest Neighbor
AUC	0.9889	0.9175	0.9057
Recall ratio	0.9906	0.9700	0.9806

In conclusion, we have used several different ways to test which model performs best on credit card customer churns. We first use cross validation and then grid search to make some adjustments to the parameters. After that, both AUC value and recall ratio prove that Random Forest is the best.

4.3. Feature Selection

We can use Random Forest to rank the importance of features. It can be seen from the table below that the three most important features are total transaction amount in the past 12 months, total transaction count in the past 12 months, and total revolving balance on the credit card.

Table 6. Important features selected by random forest.

Features	Importance	Features	Importance
Total transaction amount in the past 12 months	0.1854	unused amount	0.0310
Total transaction count in the past 12 months	0.1797	Number of contacts in the past 12 months	0.0278
used amount	0.1112	period of relationship with the bank	0.0241
change in transaction count (Q4 over Q1)	0.1076	number of moths inactive in the past 12 months	0.0241
number of products held	0.0686	dependents number	0.0122
average card utilization ratio	0.0649	education	0.0105

Amount change	0.0633	Income level	0.0099
Card limit	0.0319	marital status	0.0086
age	0.0315	gender	0.0079

As for total transaction amount and count, they are very similar. Both can reflect usage situation of a customer, because the bill could either be several big expenses or frequent small-amount pay out. It is quite intuitive that the more one customer uses his credit card, the less likely he will leave bank's services. Through using process, customers may get more dependent on credit card or be more satisfied with the services and products, therefore, obviously they will keep using the card.

Identifying the factors affecting customer churn has always been popular research, because it can help banks better grasp existing customers and improve profits. By using logistic regression and decision tree, Abbas et al. found that customer relationship length, customer age, customer gender and the number of mobile banking transactions have an impact on customer churn [16]. Moreover, in the study of Mahdi et al., they found that the loss of bank customers had something to do with their careers through Neural Network model [17]. To determine the causes of customer churn in banking and e-banking services, Chiang et al. used association rules and analysis of customer transactions to find the most important customer churn patterns and their result shows that blind promotion is a major cause of customer loss [18].

Those findings seem differ from what we get, and the reason behind is because the databases, methods and models used in each study are different. In Abbas' study he used decision tree while Mahdi used Neural Network, whereas in our study, we mainly use Random Forest which is a combination of several decision trees. Therefore, each study draws different conclusions.

5. CONCLUSION

This paper aims at predicting the loss of bank credit card customers. We get 10,000 dataset containing age, salary, marital status, credit card limit, etc., and do analysis and research based on it. We firstly preprocess the dataset, then apply three rational classification models, specifically, Random Forest, Logistic regression, KNN by using 5-fold cross validation. We adjust the hyperparameters in each model to improve the accuracy and use ROC & AUC and confusion matrix to evaluate the model performance. Both two agree that Random Forest has the strongest predictive ability, and by using this, we find out three features which have the greatest impact on our prediction.

Totally three conclusions are obtained from the research. To begin with, Random Forest model is the best among the other two, although it has relatively low computational speed due to its complexity, its performance is approximately 5% higher in accuracy and 2% higher in recall. Secondly, using a better combination of parameters can improve model's performance. Finally, we check the feature importance of the dataset and find that the total transaction amount in the last 12 months, total transaction count in last 12 months and total revolving balance on the credit card have significant impacts on model forecasting. It shows that the more frequent customers use their credit cards, the less likely they are to leave, therefore, the bank managers can adjust credit card service based on it to fight against customer churn and increase retention rate. And the increase of retention rate brings about a greater profit growth. By using this model, they have plenty of time prior to taking actions to retain customers, i.e., by making promotions, offering coupons to encourage people to use their credit card and cultivate their using habits.

There are some deficiencies as well. Firstly, machine learning has numerous algorithms in classification such as neural network, but we merely use a few of them. Next, only one dataset which is collected from a specific bank is used, thus it might bring limitations to our model because it just represents a part of the industry. Lastly, we use a single model to do predictions. In fact, ensemble learning can combine multiple models' advantages and give a better performance.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Xinyu Miao conducted the research, Liguo Tang analyzed the data, and Haoran Wang wrote the paper. All authors had approved the final version.

REFERENCES

[1] N. X. Hong, and L. Yi, "Standing at the crossroads-credit card," *Reporters' Notes*, vol. 5, pp. 41-43, 2020. (in Chinese)

[2] R. Rajamohamed, and J. Manokaran, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," *Cluster Computing*, vol. 21, pp. 65-77, June 2017.

[3] G. L. Nie, W. Rowe, L. L. Zhang, Y. J. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, pp. 15273-15285, 2011.

[4] J. Liao, and Y. F. Ruan, "Research on APP Intelligence Promotion Decision Aiding System

Based on Python Data Analysis and AARRR Model," *Journal of Physics: Conference Series*, vol. 1856, pp. 1-7, 2021.

[5] M. Kehoe, H. B. Taylor, and D. Broderick, "Developing student social skills using restorative practices: a new framework called H.E.A.R.T," *Social Psychology of Education*, vol. 21, pp. 189-207, 2017.

[6] B. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable Machine Learning for Scientific Insights and Discoveries," *IEEE Access*, vol. 8, pp. 42200-42216, 2020.

[7] Q. F. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *American Journal of Epidemiology*, vol. 188, pp. 2222-2239, October 2019.

[8] Y. A. Amrani, M. Lazaar, and K. E. E. Kadiri, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, vol. 127, pp. 511-520, 2018.

[9] S. Y. Xuan, G. J. Liu, Z. C. Li, L. T. Zheng, S. Wang, and C. J. Jiang, "Random Forest for Credit Card Fraud Detection," *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 1-6, 2018.

[10] T. Hengl, M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler, "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables," *PeerJ*, vol. 6, pp. e5518, 2018.

[11] R. Couronné, P. Probst, and A. L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC Bioinformatics*, vol. 19, pp. 270-283, July 2018.

[12] A. Singh, M. N. Halgamuge, and R. Lakshmiganthan, "Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 8, pp. 1-10, 2017.

[13] N. Yang, Y. Qian, H. S. EL-Mesery, R. Zhang, A. Wang, and J. Tang, "Rapid detection of rice disease using microscopy image identification based on the synergistic judgment of texture and shape features and decision tree–confusion matrix method," *Journal of the Science of Food and Agriculture*, vol. 99, no. 14, pp. 6589-6600, 2019.

[14] J. H. Orallo, P. Flach, and C. Ferri, "ROC curves in cost space," *Machine Learning*, vol. 93, no. 1, pp. 71-91, 2013.

[15] P. Flach, J. H. Orallo, and C. Ferri, "A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance," *ICML*, pp. 657-664, Jane 2011.

[16] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, "Investigating factors affecting customer churn in electronic banking and developing solutions for retention," *International Journal of Electronic Banking*, vol. 2, no. 3, pp. 185-204, November 2020.

[17] S. H. Iranmanesh, M. Hamid, M. Bastan, G. H. Shakouri, and M. M. Nasiri, "Customer churn prediction using artificial neural network: An analytical CRM application," *In Proceedings of the International Conference on Industrial Engineering and Operations Management, Pilsen, Czech Republic*, pp. 23-26, July 2019.

[18] D. Chiang, Y. Wang, S. Lee, and C. Lin, "Goal-oriented sequential pattern for network banking churn analysis," *Expert Systems with Applications*, vol. 25, no. 3, pp. 293-302, 2003.