Predictive Analysis on Churn Customers in Business Industries using Supervised Machine Learning Algorithms and Smote

Pallabi Baruah[1], Bhairab Sarma[2]

[1]Research scholar, Department of Computer Science, University of Science and Technology Meghalaya.
[2]Associate Professor, Department of Computer Science, University of Science and Technology Meghalaya.

Received: 15.04.2024 Revised : 13.05.2024 Accepted: 28.05.2024

ABSTRACT

The churn customers are the big loss for the enterprise and business industry. A customer which are loyal and using a particular products and services for along time, suddenly taking a decision to change their mind set to use another products and services. If the industries or enterprises know that these customers are planning to migrate and want to use another product and services in that situations they can use business strategies to protect them. This research work is related to a predictive model on churn customer based on their credit score, history of buying patterns , geography location, gender specification and estimated salary with proper balance. The proposed predictive model is based on the 10K customers data which were collected from the different geographical location. The predictive Models Used: Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosting, XGBoost as base learners. The results of this churn predictive model is comparaed with that of the logistic regression, SVM, Gradient Boosting , Random Forest algorithms. The predictive model is based on hybrid XGboost with SMOTE which is used for management of imbalanced data and also hypertuned and is having accuracy 96%. Here, it is assured that this predictive hybrid model could be utilized in the industries which are service oriented in predicting of whether a loyal customer will churn or not.

Keywords: Churn Customers, XGBoost, SVC, LR, Random Forest, Smote, hypertuned

INTRODUCTION

Customer churn occurs when customers or subscribers avoid taking the services of an organization or in. Customer changing service from one indusrty to another is also termed as customer attrition. It is a very sensitive parameter as it takes little effort to handle existing customer rather than to handle the new customers – earning business from new customer's means to start from the scratch which is a struggle factor to acquire trust in the market scenerio. Customer retention, on the other hand, is generally more cost-effective as trust and loyality is already earned by the customers(Liao S., 2012). .Just like most things that can be calculated and measured, it is also possible to find customer churn in a number of ways.

The researcher is focused on new technologies, and new competitors are opening up the business industry, portable speculation and management has become a major concern for mobile service providers. Ahn J.(2006) in his research states that the service provider of mobile who has the intention to retain its customers needs to be able to predict which of them may be at risk of switching services and will turn those subscribers into customer retention efforts(Ahn J,2006).

Su-YeonKima et al(2006) emphasized that in response to the limitations of existing churn forecasting systems and the unavailability of customer numbers at the investigated business provider, we propose, construct, and scrutinize the churn speculation process that predicts subscription contract information and telephone pattern changes from telephone details. This proposed approach is able to identify potential participants at the contract level over a period of prediction. In addition, the proposed process includes a multidisciplinary approach to address the challenge of highly distorted section distribution between churner and non-churner(Su-YeonKima et al, 2006) .

REVIEW OF LITERATURE

Research work of Brito J.B.G.(2024) discussed that the framework outperformed alternative methods reported in the literature in terms of precision-recall area under curve, accuracy, recall, and specificity. From a practical perspective, the framework provides managers with valuable information to predict customer churn and develop strategies for customer retention in the banking industry.(Brito J.B.G.,2024)

According to Ahmad A.K(2019) four tree based algorithms were used as it has some specific qualities of implementation and they were Decision Tree, Random Forest, GBM tree algorithm, and XGBOOST algorithm. Qureshii S.A.(2013) reported that if customer churn prediction could be done at the initial stage than that would save revenue loss. Rajendran S. et al.(2023) proposed the implementation of Random Forest combined with SMOTE-ENN and F1 score has the best result than other. Saleh S. et al(2023) explored the causes of churn in Danish telecommunication industry and also found the strategies for retention of customers.

Dias J.R et al.(2023) in his experimental results show that boosting techniques such as XGBoost present the best predictive performance. Suh Y(2023) in his research identified and calculated the influence of key variables on individual customer churn to enable a business person (rental care customer management staff) to carry out customer-tailored marketing to address the cause of the churn. Lalwani P et al.(2022) in their research work made a comparison of Customer Churn prediction in Telecommunication Industry using the different techniques such as Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest, XGBoost Classifier, CatBoost Classifier,AdaBoost Classifier and Extra tree Classifier.

Shu-Hsien et. al (2012) prescribes DMT, in terms of the following three areas: types of information, types of analysis, and types of structures, and their types of applications in various studies and functional domains. Bong -Horng et. al (2007) identified in the churn model to divide all 'churners' into distinct groups. Van den and Larivière (2004) focused on predictors becoming one complete model for retention including several 'new' types of various covariates related to real customer behavior by analysing churn performance.

Coussementab et. al (2010) used Generalized Additive Models (GAM) such as Logistic Regression, GAM lowers the linearity limit that allows for non-linear equilibrium of data. Contributions that better identify risky customers; Amal M. Almana et al (2014) stated in his research work that customer Fradulency is one of the prime issue. Su-YeonKima et.al (2006) the researcher has proposed a framework for analysing customer numbers and classifying clients. After the classification of customers, customer-building strategies will be demonstrated through case studies in wireless business company

Kandogan (2001) who specializes in multi-dimensional data is indicated by a point, where each attribute data contributes to its location by entering the same code. Star Coordinates interactive features give users the ability to apply a variety of dynamic conversions, aggregate and split sizes, multi-dimensional integration, view collections, styles, and locations for sale in data distribution, and query points based on data ranges.Pınar Kisioglu and Y. IlkerTopcu (2011) emphasized that management, companies try to keep their existing customers, instead of getting new ones. Previous researchers have focused on predicting customer trends in the business industry. A researcher mapped as a basis for the Bayesian Belief Network is presented with the results of integration analysis, multicollinearity testing and expert opinions.

Chen-FuChiena and Li-FeiChenab (2008) developed a data mining framework based on decision-making rules and organizational rules for making useful workers' selection rules. The results can provide decision-making rules relating to employee knowledge of work and performance. Chih-Fong and TsaiaYu-HsinLub (2009) focused on hybrid models by combining two different neural churn predictive network techniques, namely artificial distribution networks (ANN) and self-organizing maps (SOM).

In addition, IBRF also produces good results than other random algorithms such as moderate forests and random forests. Shin-Yuan et., Al (2006) stated that this study shows that both tree and network decisions can produce accurate predictor models using customer statistics, payment details, contract / service status, call details, and service log changes. Hyunseok Hwang, and Taesoo Jung EuihoSuh (2004), a study conducted to calculate the number of customers based on the value of the Customer's life (LTV). However, there are limitations. It's hard to imagine a customer revolt. The types of forecasts focus on the expected future cash flows based on past customer benefits.

John Hiddenites et. al (2007) focuses on a variety of advanced churn management strategies in response to the above requirements. The focus of this paper is to review some of the most popular technologies found in the literature for the construction of customer management platform. Scott A. Neslin et.al (the accuracy of the 2006 speculation on all submissions could alter the profitability of a fraudulent management campaign by hundreds of thousands of dollars. Second, beauticians have the ability to stay

Jadhav R., Pawar U.(2011) argued that data mining could be incorporated into the communications sector to identify Churn's predictions of customers leaving the company's platform called Churn Prediction. The company must take the necessary steps to maintain it. Gary M. Weiss (2005) who researches arrogant clients and finds a solution to research problems using data mining, is the first step in understanding data. Yang L., Chiu C. (2006) focused on preliminary research on the use of data mining techniques to solve the real-world customer problem in the communications sector. IonutBrandusoiu I., Toderean G.(2013) emphasized that a high-performance, four-kernel Support Vector Machines algorithm has been used by churn and non-churn buyers. Kaur J.(2019) referred that customer churn is one of the significant research issues in business industries. The customers affected by business services such call rate, voice message, voice mail, international plan, and customers services in case of any issues.

Problem Statement And Research Objectives

Customers churn is one of the most important study in service industries. The customers affected by business services such call rate, voice message, voice mail, international plan, and customers services in case of any issues. Churning customers are big loss for the service related industries and it is this issue which is critical globally. The predictive model is also capable to handle biased and imbalance dataset. The researcher discusssed some of the research objectives which are stated as:

1. To study the churn/non-churn customers in business industries and its related issues
2. To develop a predictive model with high accuracy to classify churn/ non-churn customers.
3. To evaluate the predictive model results and optimized it.

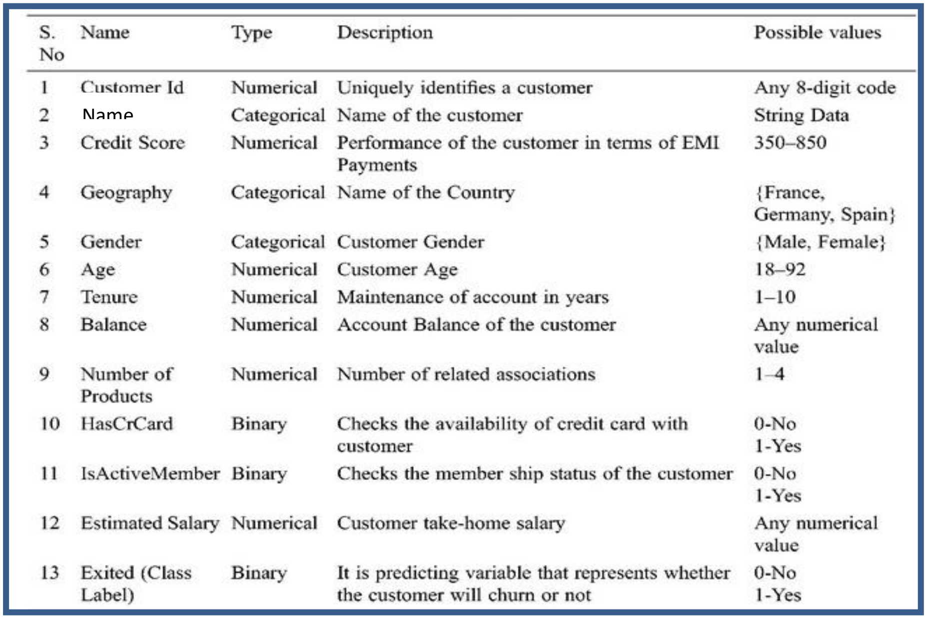
Dataset Analysis

In this research the data set of 10000 entries was used where 8000 for training dataset and 2000 entries for testing dataset for classifier to classify the data in terms of 80:20 ratio.

The dataset and the related features were used for study and anaysus of the study. The dataset is splitted on random basis and it is divided automatically into two different dataset with a ratio 80:20. The standard mapping parameter is used with respect to the training and testing dataset in the traditional algorithm. Various techniques are used so that the features are taken out which would contribute towards the development of a model.

The dataset where the model is fiited is the data source from Kaggle.com, industry: Business . This Dataset is accessible in the web and is widely used during the estimation of the proposed model and the structure is given below(Table 1)

Table 1. Metadata description of the dataset



S. No	Name	Type	Description	Possible values
1	Customer Id	Numerical	Uniquely identifies a customer	Any 8-digit code
2	Name	Categorical	Name of the customer	String Data
3	Credit Score	Numerical	Performance of the customer in terms of EMI Payments	350–850
4	Geography	Categorical	Name of the Country	{France, Germany, Spain}
5	Gender	Categorical	Customer Gender	{Male, Female}
6	Age	Numerical	Customer Age	18–92
7	Tenure	Numerical	Maintenance of account in years	1–10
8	Balance	Numerical	Account Balance of the customer	Any numerical value
9	Number of Products	Numerical	Number of related associations	1–4
10	HasCrCard	Binary	Checks the availability of credit card with customer	0-No 1-Yes
11	IsActiveMember	Binary	Checks the member ship status of the customer	0-No 1-Yes
12	Estimated Salary	Numerical	Customer take-home salary	Any numerical value
13	Exited (Class Label)	Binary	It is predicting variable that represents whether the customer will churn or not	0-No 1-Yes

Research Design And Methodology

Here, the researcher used the hybrid XGboost technique with hyperparameters to extract the customers who are likely to churn in business industry. SMOTE is implemented to handle the imbalanced dataset and gives the higher level accuracy for predictive model. With respect to the predictive model of machine learning the researcher used the comparative study between logistic regression and random forest, support vector algorithms, Gradient Boosting, XGBoost to classify the category of churn customers and Non-churn customers in Business Industries. There are several factors which are directly affected to the business services such as proper connectivity, international calls rate, day and night call rate, customer care/support services etc. Internally, XGBoost models represent all problems as a regression predictive modeling problem that only takes numerical values as input. If the data is in a different form, it must be prepared into the expected format.

The model comprises of three parts- The preprocessing phase, Training phase and the testing Phase. In the preprocessing phase, data with the null values are removed before converting the categorical data to numerical data. Also an empty value in age attribute is filled with a mean value of the age. Again converting the categorical data into numerical data with an encoding technique known as "Label Encoding," which assigns a numerical value from 0 to n-1 takes place here. Attributes with wide range of values are solved by the a technique known as "Standard Scalar" that redistributes the mean and standard deviation always to be 0 and 1 respectively.

In the training phase, logistics regression, gradient boosting and XGB as base learners and hperparameters with search algorithms are used to check the accuracy of the model. SMOTE is used handle imbalance data.

The test phase involves the cross validation of the proposed model with the test dataset and finally the different metrics are used to check the accuracy of the churn rate.

Logistic Regression

Logistic regression (also known as logit regression or logit model) is a widely used multidimensional method for modeling dichotomous results(Bennouna,2019)(Strzelecka A., 2020) . Whren decision makinng of a research work is concerned, it can be corelated with this algorithm. It is associated with statistical analysis in case of financial issues. Bagley(2001) states that the regression model serves two purposes: (1) it can predict the result variable for new values of predictive variables, and (2) it can help answer questions about the studied phenomenon, because the coefficient of each predictive variable clearly describes the relative contribution of this variable to the result variable, automatically controlling the influence of other predictive variables(Strzelecka A.,2020).

$$\text{Sigmoid Function} = \frac{1}{(1+e^{-\text{Value})}} \qquad (1)$$

Where e is the base of the natural algorithms (Euler's number or EXP) and the value is the actual numeric value that it is going to transform between into the range 0 and 1 using logistic function.

Random Forest Classifiers

Random forest is a supervised learning algorithm which is used for both classification as well as regression. It is mainly used for classification problems. Kumar A.(2021)emphasized that forest is made up of trees and more trees means more robust forest.

Support Vector Machine (Svm)

The SVM algorithm is widely used in machine learning as it can handle both linear and nonlinear classification tasks. However, when the data is not linearly separable, kernel functions are used to transform the data higher-dimensional space to enable linear separation. This application of kernel functions can be known as the "kernel trick", and the choice of kernel function, such as linear kernels, polynomial kernels, radial basis function (RBF) kernels, or sigmoid kernels, depends on data characteristics and the specific use case.

$$G(x)=w^{T}x+b$$

K is to be maximized such that

$$- w^{T}x + b>=k \text{ for } di==1 \qquad (2)$$

$$- w^{T}x + b <=k \text{ for } di==-1 \qquad (3)$$

Value of g(x) dependent of ||w||

1. Keep || w|| =1 and maximize g(x) or ,
2. g(x) >=1 and minimize ||w||

SMOTE first select a minority class instance a at random and finds its k nearest minority class neighbours. The synthetic instance is then created by choosing one of the k nearest neighbours b at random and connecting a and b to form a line segment in the feature space. The synthetic instance are generated as a convex combination of the two chosen instances a and b.

Gradient Boosting Algorithm

Gradient boosting is a powerful techniques for predictive models. It builds a better model by merging earlier models until the best model reduces the total prediction error. It is also known as a forecasting model, and it is used to attain a model that removes the problems of the previous models. Gradient Boosting is named so that the set target outcomes depend on the gradient of the inaccuracy vs the forecast. Every new model created using this method moves closer to the path that lowers prediction error in the range of potential outcomes for every ML training case. Gradient Boosting is mainly of two types depending on the target columns:

1. Gradient Boosting Regressor: It is implemented in case the columns are continuous
2. Gradient Boosting Classifier: It is implemented in case the target columns are with classification problems.

XG-BOOST

The XGBoost method is founded on gradient-boosting trees, which can be very useful for gradient enhancement Kazemi M.(2023).]. A regression and classification problem can be very effectively solved using XGBoost based on the concept of regression and classification trees(Nabavi,2020) . Also, XGBoost combines the novel algorithm with the GBDT method to represent a soft computing library. It is an ensemble additive model that is composed of several base learners. XG-Boost uses the Taylor series to approximate the value of the loss function for a base learner ft(xi), thus , reducing the load on Emily to calculate the exact loss for different possible base learners.

This research involves hybrid XGBoost with hyper parameters to predict customer churn. Hence, this study aims to fill this gap by proposing a novel approach to optimize XGBoost using base models like that of Logistic regression, Decision trees and Gradient Boosting Model and various algorithms, including random search and grid search. The study was conducted using data collected from a local bank as well as a state of art dataset where the predictive models were developed by considering various properties of the data. Then a comparison of the developed model was made with the traditional XGB model to evaluate their effectiveness in predicting variations in the churn rate. Further the regularisation technique L1 , L2 are used to maintain the under fitting and overfitting issues as processing takes place from the root to the child scuccesors of the tree. The proposed technique enhances the predictive accuracy of hybrid XGB contribute to the field of business by providing a prediction of customer churn.

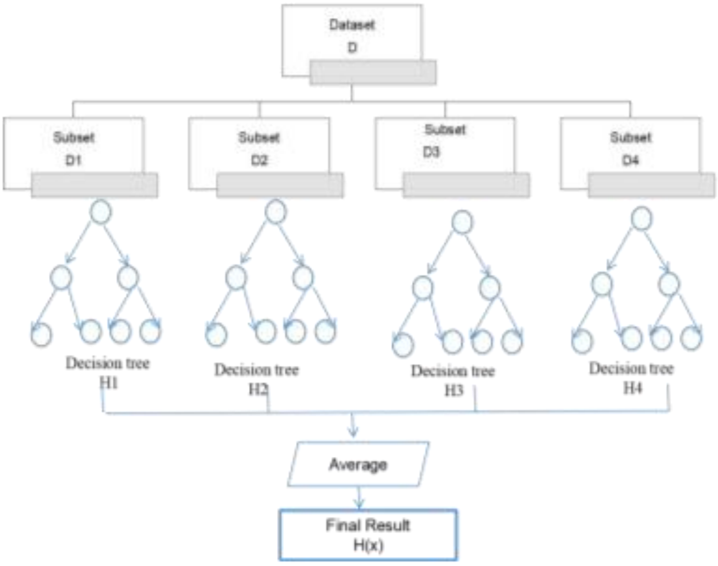


Fig 1. Extreme Gradient Boosting structure

XG-Boost starts with an initial prediction and use the loss function to evaluate if the prediction works well or not. In this equation the first part represent the loss function which calculates the pseudo residuals of predicted value yiwith hat and true of yi in each leaf whereas xi represents the number of features which works like as independent variable.

X_i=x_1,x_2,x_3,x_4……………………..x_n ………………………(4)

Y_i=y_1,y_2,y_3,y_4,……………………..y_n ………………………(5)

Where Xi represent independents variables datasets and Yi represents dependents datasets.

Hyperparameter Tuning

Based on existing data, tuning can be used to learn an algorithm that finds the optimal hyperparameters.A hyperparameter can determine an algorithm's optimum performance in supervised learning This research used three tuning methods to find the optimal hyperparameter: grid search, random search, and metaheuristic algorithms. XGBoost is an algorithm with great potential and many hyperparameters. The following are the hyperparameters that adjusted in this study:

learning_rate: This hyperparameter sets the step size for each boosting iteration. Smaller values may improve performance, but may also make training longer.

max_depth: This hyperparameter controls the maximum depth of the decision tree, which affects the model's complexity. Higher values may cause overfitting, while lower values may cause underfitting.

n_estimators: This hyperparameter determines the number of boosting iterations or trees to build.

Grid Search

Belete D.M & Huchaiah M.D(2021) states that Grid search can be described as an exhaustive exploration or a method of brute force that tests all the combinations of hyperparameters given to the grid configuration.GS operates by assessing the Cartesian product of a finite set of values defined by the user operates by assessing the Cartesian product of a finite set of values defined by the user (] Hutter F. et al,2019). The most widely used method to learn hyperparameter configuration space is grid search (GS)(Zoller M-A, Huber MF(2019).

Additionally, in the research approach, the prediction models were prepared using the XGBoost library, and the programming environment used for model development and evaluation was Python with the scikit-learn and XGBoost libraries.

Smote

Classification related to binary has issues of dataset which may be imbalanced. It usually is found in practical business applications like detection, filtering, prediction issues. To rectify this problem, one popular technique is Synthetic Minority Oversampling Technique (SMOTE). SMOTE is specifically designed to tackle imbalanced datasets by generating synthetic samples for the minority class. It is devised to ameliorate the challenges associated with imbalanced datasets (Zhou et al., 2022).

Fine-tuning the parameters not only enhances the model's overall performance but also mitigates the risk of overfitting, thereby improving the model's generalization capability. The following depicts the hypertuned parameters used in python.

```
params = {
    'min_child_weight': [1, 5, 10],
    'gamma': [0.5, 1, 1.5, 2, 5],
    'subsample': [0.6, 0.8, 1.0],
    'colsample_bytree': [0.6, 0.8, 1.0],
    'max_depth': [3, 4, 5]
    }
```

RESULTS AND DISCUSSION

The main objective of this research study is to predict churn customers for a Business Industries which are in future planning to switch with other products or services. Specifically, we will initially perform Exploratory Data Analysis (EDA) to identify and visualise the factors contributing to customer churn. This analysis will later help us build Machine Learning models to predict whether a customer will churn or not.

1. Skills: Exploratory Data Analysis, Data Visualization, Data Preprocessing (Feature Selection, Encoding Categorical Features, Feature Scaling), Addressing Class Imbalance (SMOTE), Model Tuning.
2. Models Used: Logistic Regression, Support Vector Machines, Random Forests, Gradient Boosting, XGBoost, and Light Gradient Boosting Machine.
3. Our Data Frame has 11 features/attributes and 10K customers/instances. The last feature, 'Exited', is the target variable and indicates whether the customer has churned (0 = No, 1 = Yes). The meaning of the rest of the features can be easily inferred from their name.

The most important things to note are:

1. The age of customers ranges from 18 to 92, with a mean value approximately equal to 40,
2. The mean (and median) tenure is 5 years, so the majority of customers is loyal (tenure > 3), and
3. Approximately 50% of customers are active.

EDA will help us understand our dataset better. However, before we look at the data any further, we need to create a test set, put it aside, and use it only to evaluate our Machine Learning models.

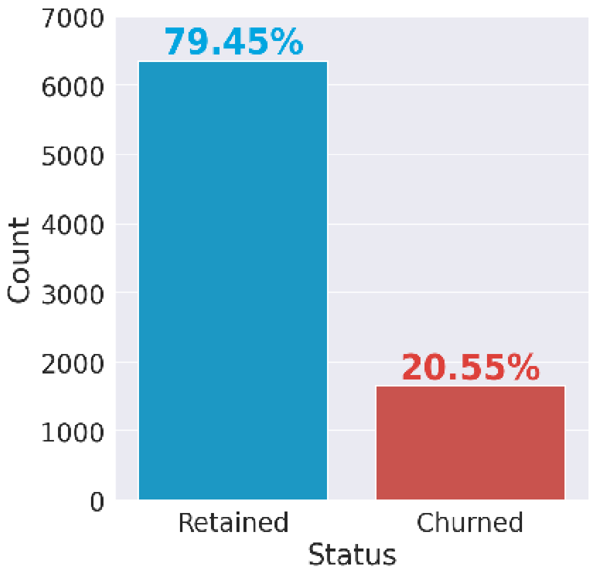


Fig 2. Segmentation of Customers and Churn Rate

Notice that our dataset is skewed/imbalanced since the number of customer in the 'Retained' class outnumbers the number in the 'Churned' class by a lot. Therefore, accuracy is probably not the best metric for model performance. Different visualization techniques apply to different types of variables, so it's helpful to differentiate between continuous and categorical variables and look at them separately(Fig.2).

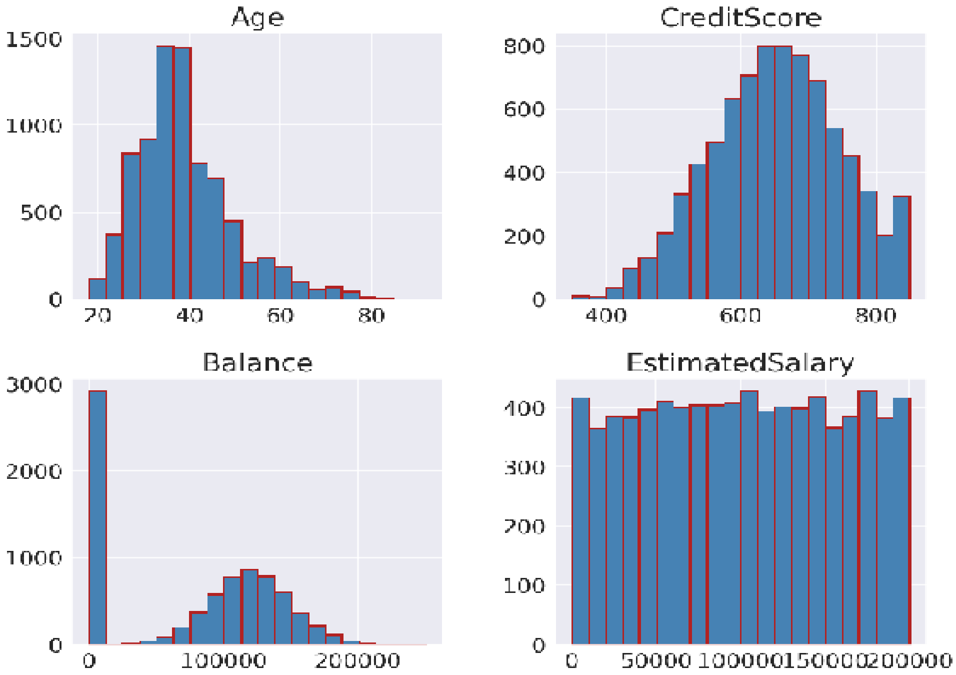


Fig 3. Categories of Customers

Descriptions

1. 'Age' is slightly tail-heavy, i.e. it extends more further to the right of the median than to the left,
2. Most values for 'Credit Score' are above 600,
3. If we ignore the first bin, 'Balance' follows a fairly normal distribution, and
4. The distribution of 'Estimated Salary' is more or less uniform and provides little information.

Exploratory Analysis

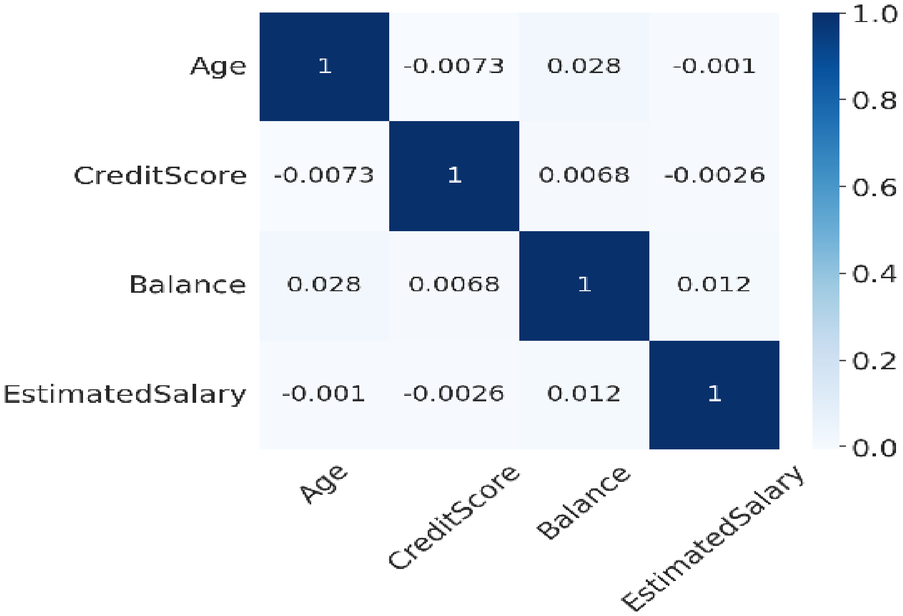


Fig 4. Correlations between features of customers

The above figure shows there is no intercorrelation amongst the features so there is no multicollinearity issue.

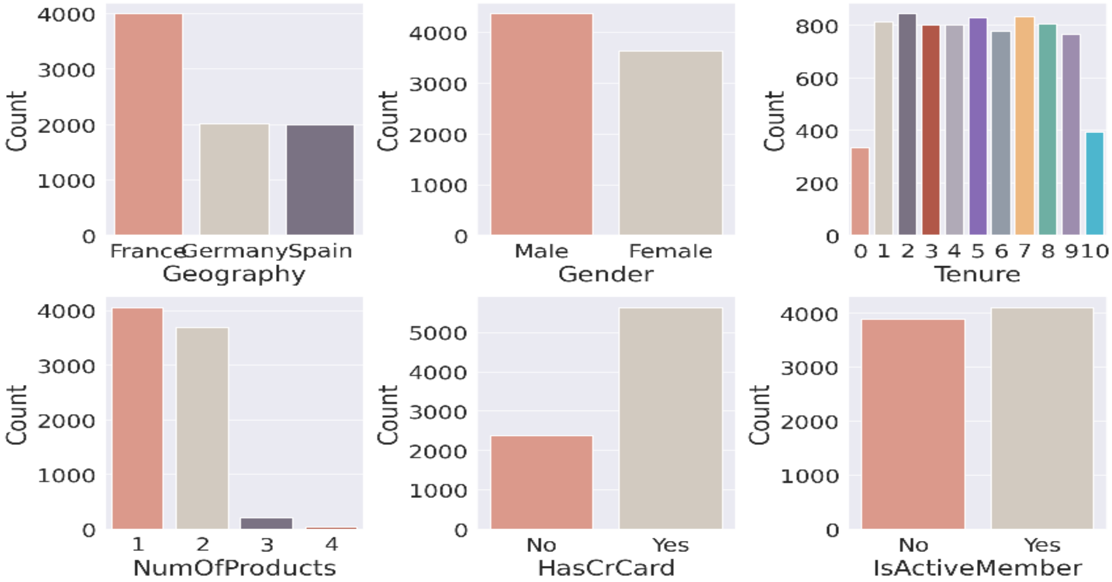


Fig 5. Correlations between features of customers

Descriptions

1. The Business Industries has customers in three countries (France, Spain, and Germany). Most customers are in France.
2. There are more male customers than females,
3. Only a small percentage leaves within the first year. The count of customers in tenure years between 1 and 9 is almost the same,
4. Most of the customers have purchased 1 or 2 products, while a small portion has purchased 3 and 4,
5. A significant majority of customers has a credit card, and
6. Almost 50% of customers are not active.

Building Machine Learning Models

We start this section by first creating two simple models to estimate the baseline performance on the training set. Specifically, we will use Gaussian Naïve Bayes and Logistic Regression. We will use their default parameters and evaluate their (mean) recall by performing k-fold cross-validation. The idea behind k-fold cross-validation, which is illustrated in this figure, is simple: it splits the (training) set into k subsets/folds, trains the models using k-1 folds, and evaluates the model on the remaining one fold. This process is repeated until every fold is tested once.

Apart from a confusion matrix, a plot of the learning curves will be provided for each classifier. Learning curves are plots of a model's performance on the training set and the validation set as a function of the training set size. They can help us visualise over fitting/under fitting and the effect of the training size on a model's error.

Table 2. Accuracy of different predictive models

	Accuracy	Precision	Recall	AUC
LR	0.683	0.667	0.730	0.734
SVC	0.797	0.808	0.780	0.876
RF	0.796	0.813	0.770	0.883
GBC	0.788	0.808	0.755	0.875
XGB	0.793	0.800	0.781	0.876
H-XGB	0.956	0.913	0.913	0.964

In trhe above Table 2, all other classifiers have a recall higher than 70% (baseline performance). XGB is the model with the highest recall (78.5 %). However, the LGBM classifier has the best overall performance with the highest accuracy, precision, and AUC.

Using single metrics is not the only way of comparing the predictive performance of classification models. The ROC curve (Receiver Operating Characteristic curve) is a graph showing the performance of a classifier at different classification thresholds. It plots the true positive rate (another name for recall) against the false positive rate.

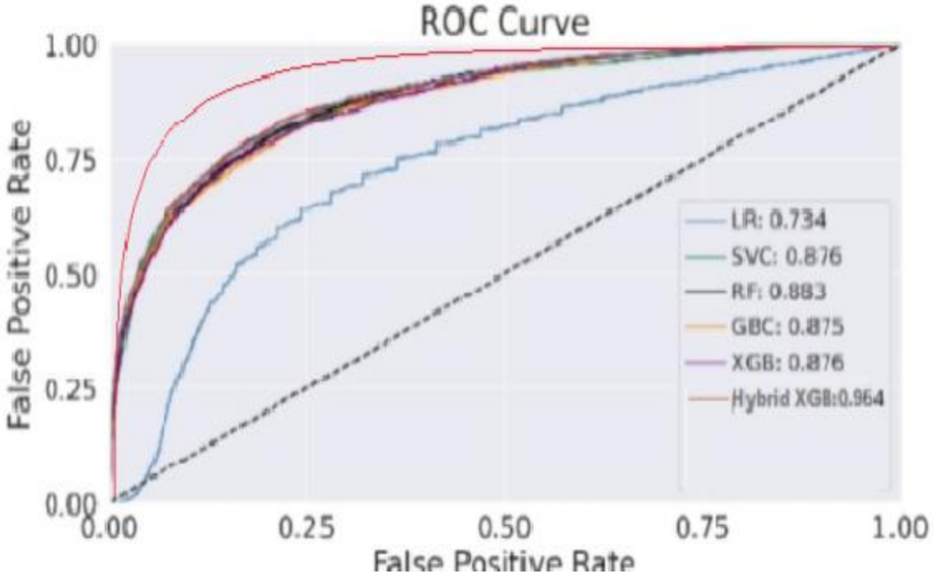


Fig 6. Accuracy of different predictive models with ROC Curve

The dashed diagonal line represents a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). In our case, all classifiers, apart from Logistic Regression, perform similarly. It seems that LGBM performs marginally better, as evidenced by the slightly higher AUC (0.888). Recently, we came across another tool for assessing the performance of a classifier model. Simply put, a Cumulative Gain shows the percentage of targets reached when considering a certain percentage of the population with the highest probability to be target according to the model

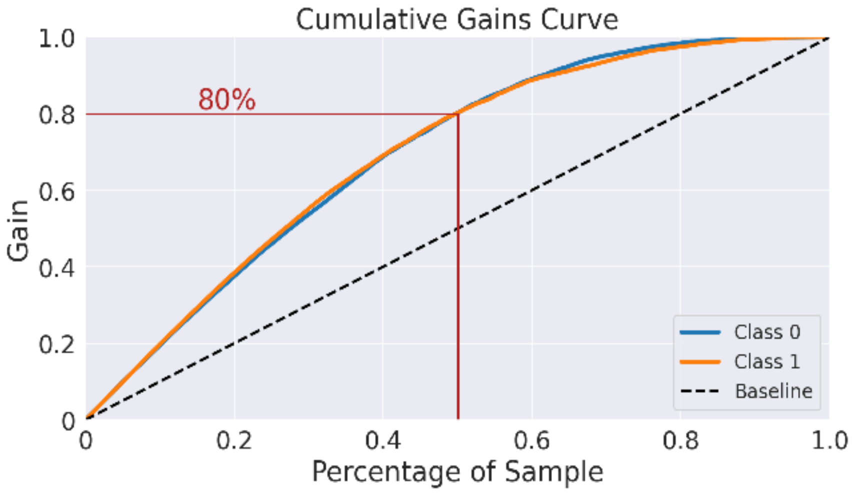


Fig 7. Cumulative Gains of Accuracy of different predictive models with ROC Curve

This chart shows that if we target 50% of the customers most likely to churn (according to the model), the model will pick 80% of customers who will actually churn, while the random pick would pick only 50% of the targets. The performance on the test set for all models is fairly similar to the training set, which proves that we do not over fit the training set. Therefore, we can predict customer churn with a recall approximately equal to 78%.

It is seen that Hybrid XGB performs marginally better with SMOTE, as evidenced by the higher AUC (0.964) in case of state of art dataset.

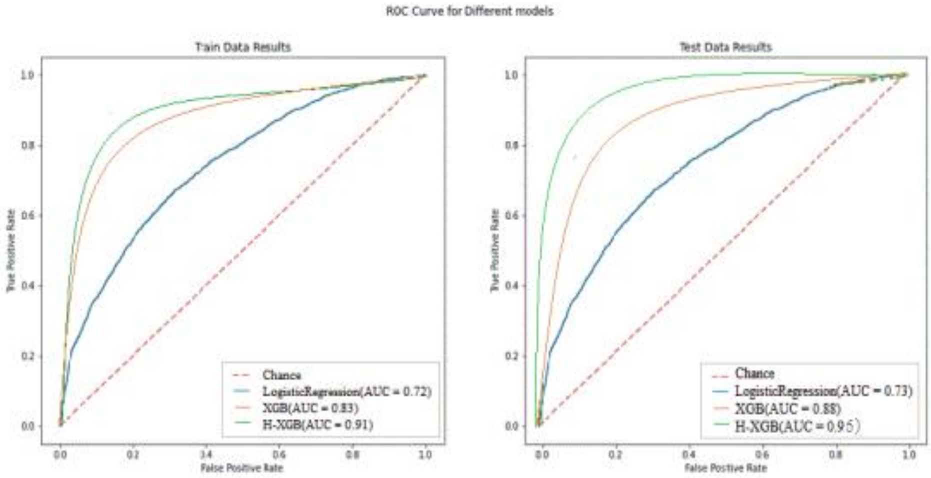


Fig 8. ROC for different models using the state of art dataset

Based on the dataset, the hybrid XGBoost algorithm is modeled and the results are compared with the AUC value and ROC curve of the existing techniques. SMOTE is necessary to obtain a balanced dataset. The ROC curve of the Train data and Test data in case of state of art dataset is plotted in the above graph(fig-13). It can be observed that the curve for training set is more smooth as compared to the curves for the testing sets. For the XG-Boost model trained without cross validation and hyperparameter tuning, the training AUC score is 0.83 but while testing it rises to 0.88, suggesting that the model must have overfit to the data. For the Hybrid XG-Boost model which used cross validation during training and was hyper tuned, the training score was 0.91 and the testing score was 0.96, suggesting that the model did not overfit to the training data and hence the testing score was the highest among all three models. The hyperparameters are tuned and fed to maintain accracy rate. Overall, the Hypertuned Hybrid XG-Boost model performs the best among other model.

CONCLUSION

Finally the researchers concluded that the predictive model is more accurate to predict the category of customers whether they belong to churn or non-churn. The researcher used machine learning algorithms to build a predictive model on churn / non –churn customers in business industries. The predictive model which is based on logistic regression having 81% accuracy where as at 19% does not predict accurately about the customers status. The accuracy score of the random forest classifiers shows that the predictive model is having 90% accuracy where as at 10% does not predict accurately about the customers status. At this level the researcher concluded that the random forest classifiers produced the more accurate result than logistic regression. As per the statistics given the score of the confusion matrix are such as accuracy 80%, Precision for " did not leave " is 0.96 where as recall is 0.93, and F1score is 0.94 with the support features of 10000. For the left customers precision is 0.66, recall is 0.76, F1score is 0.70 and support features is 160, the data analysis report shows that the predictive model is having 90% accuracy where as at 10% does not predict accurately about the customers status. For more accuracy and to handle the imbalance dataset the researcher used the XG-Boost algorithms which started at iteration number 0 with value 0.805 and passes through every 50 interval such as 0, 50, 100, 150,200,250, and 300 and until the predictive model get the stable result. Finally the confusion matrix and the classification report got great improvement. The accuracy of predictive model is now 80.5% and the Precision and Recall has improved drastically.

REFERENCES

[1] Liao S., Chu P., Hsiao P. (2012). Data mining techniques and their use - Tenth Review from 2000 to 2011, Expert Systems with Applications, Volume 39, Issue 12, 15 September 2012, Pages 11303-11311, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2012.02.063.

[2] Chu B., Tsai M., Ho C.(2007). Toward a hybrid data mining model for retention customer', Knowledge Based Knowledge, Volume 20, Issue 8, December 2007, Pages 703 -718, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.knosys.2006.10.003.

[3] Kim S., Jung T.,Suh E., Hwang H.(2006). Customer segregation and strategic development according to the value of customer life: Case studies, Application System, Volume 31, Issue 1, July 2006, Pages 101- 107, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2005.09.004.

[4] Kandogan E. (2001). Visualizing multi-dimensional cluster, trends, and outliers using star coordinates, KDD '01 Proceedings of ACM SIGKDD's seventh conference on data acquisition and data mining, Pages 107-116, San Francisco , California - August 26 - 29, 2001, ACM New York, NY, USA © 2001, ISBN: 1-58113-391-X, doi. 10.1145 / 502512.502530.

[5] Ahn J., Han S., Lee Y.(2006). Customer Churn Analysis: Churn Symptoms and Consequences of Few Discrimination in the Korean Business Industry, Communication Policy, Volume 30, Issues 10 -11, November - December 2006, Pages 552-568, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.telpol.2006.09.006.

[6] Kisioglu P., and Topcu Y.L.(2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey, Expert Systems with Applications Volume 38, Issue 6, June 2011, Pages 7151 -7157, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2010.12.045.

[7] Chen-Fu C., and Chenab L.(2008). Data mining to improve staff selection and improve human performance: Study studies in the high technology industry, Systems System with Applications, Volume 34, Issue Pages 280-290, 2008 Elsevier Ltd, https://doi.org/10.1016/j.eswa.2006.09.003.

[8] Tsai CF,Lu YH(2009). Customer churn prediction by hybrid neural network, Expert Systems with Applications, Volume 36, Issue 10, December 2009, Pages 12547-12553, Copyright © 2008 Elsevier Ltd, https: //doi.org/10.1016/j.eswa.2009.05.032

[9] Van den P, Lariviere,.Bart(2004). Customer analysis of financial services using risky forms, European Journal of Operational Research, Volume 157, Issue 1, 16 August 2004, Pages 196-217, Copyright © 2008 I -Elsevier Ltd, https://doi.org/10.1016/S0377-2217 (03) 00069-9.

[10] Coussementab K., Benoit D. Poel V. (2010). Improving marketing decisions in the context of customer speculation using standard add-on models, Expert Systems with Applications, Volume 37, Issue 3, 15 March 2010, Pages 2132 -2143, Copyright © 2008 Elsevier Ltd., https://doi.org/10.1016/j.eswa.2009.07.029.

[11] Amal M. Almana et al (2014). Research on Data Mining Methods in Churn For Customer Analysis For Industry Telecom, Int. Engineering Research and Applications Journal www.ijera.com,ISSN : 2248-9622, Vol. 4, Issue 5(Version 6), May 2014, pp.165-171.

[12] Madan M., Dave M. Kapoor V.N (2015). A Review on: Data Mining for Telecom Customer Churn Management, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 9, September 2015.

[13] Hashmi N., Anwer N.B and Iqbal M. (2013). Customer Churn Prediction in Business A Decade Review and Classification, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 5, No 2, September 2013.

[14] Jadhav R. J., Pawar U.T.(2011). Churn Prediction in Business Using Data Mining Technology, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 2, February 2011.

[15] Yang L., Chiu C (2006). Subscriber Churn Prediction in Business, 2006.

[16] Brandusoiu I., Toderean G.(2013). Churn Prediction In The Business Sector Using Support Vector Machines Issue # 1, May 2013.

[17] Hudaib A., Dannoun R., Harfoushi O., Obiedat R., Faris H. (2015). Hybrid Data Mining Models for Predicting Customer Churn, J. Communications, Network and System Sciences, May 2015, 8, 91-96.

[18] Xiea Y.,Lie X. Ngai E.W.T., Ying W.(2008). Customer Predicting using Advanced Informal Forests, Expert Systems with Applications, Volume 36, Issue 3, Part 1, April 2009, Pages 5445-5449, Copyright © 2008 Elsevier Ltd. https: //doi.org/10.1016/j.eswa.2008.06.121.

[19] Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying Data Mining To Telecom Churn Management, Expert Systems with Applications, Volume 31, Issue 3, October 2006, Pages 515-524, Copyright © 2008 Elsevier Ltd. https://doi.org/10.1016/j.eswa.2005.09.080.

[20] Hwang H., Jung T., Suh E.(2004). An LTV model and customer segmentation based on value value: a case study on the wireless business industry, Expert Systems with Applications, Volume 26, Issue 2, February 2004, Pages 181- 188, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0957-4174(03)00133-7

[21] Miguel A.P.M. Jejune, (2001). Measuring the impact of data mining on churn management, Internet Research, Vol. 11 Issue: 5, p. 375-387, https://doi.org/10.1108/10662240110410183.

[22] Chih-Ping, Weia-Tang Chiub (2002). Transforming business data to predict prediction: data mining method, Application Systems System, Volume 23, Issue 2, August 2002, Pages 103-112, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0957-4174 (02) 00030-1.

[23] Haddena J., Tiwari A., Kumar R., and Rutab R.D(2007). Computer assisted customer churn management: State-of-the-art and future trends, Computers & Operations Research, Volume 34, Issue 10, October 2007, Pages 2902-2917, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.cor.2005.11.007.

[24] Neslin S.A., Gupta S., Kamakura W., Lu J., and Mason C.H.(2006). Error Detection: Measuring and Understanding the Predictable Accuracy of Churn Customer Models. Marketing Research Journal: May 2006, Vol. 43, no. 2, pages 204-211, https: //doi.org/10.1509/jmkr.43.2.204.

[25] Rygielskia C., Wang J., Yen D.C.Y(2002). Methods of data mining customer management data, Technology in Society, Volume 24, Issue 4, November 2002, Pages 483-502, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/S0160-791X (02) 00038-6.

[26] Ahmad A.K., Jafar A., & Aljoumaa K.(2019). Customer churn prediction in telecom using machine learning in big data platform. Springer Open Journal of Big Data.

[27] Brito J.B.G. , Bucco. G.B., Heldt R.,Becker J.L., & Silveira C.S.(2024) A framework to improve churn prediction performance in retail banking. Spinger Open Access.

[28] Qureshii S., Rehman A.S., Qamar A.M., & Kamal A. (2013), Telecommunication subscribers' churn prediction model using machine learning. In: Eighth international conference on digital information management. 2013. p.131–6

[29] Rajendran S., Devarajan R., & Elangovan G(2023)Customer Churn Prediction Using Machine Learning Approaches, Conference: 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), January 2023

[30] Saleh S., & Saha S.(2023)Customer retention and churn prediction in the telecommunication industry: a case study on a Danish university, Springer Open Access, June,2023.

[31] Dias J.R., & Antonio N.(2023). Predicting customer churn using machine learning: A case study in the software industry. Journal of Marketing Analytics. Springer Open Journal. December, 2023

[32] Suh Y.(2023), Machine learning based customer churn prediction in home appliance rental business, Springer Open Journal of Big Data, https://doi.org/10.1186/s40537-023-00721-8.

[33] Lalwani P., Mishra M.K., Chadha J.S., & Sethi P.(2022), Customer churn prediction system: a machine learning Approach,Springer Collection https://doi.org/10.1007/s00607-021-00908

[34] Bennouna, Ghita, and Mohamed Tkiouat. (2019) "Scoring in microfinance: credit risk management tool –Case of Morocco-." Procedia Computer Science 148: 522-531

[35] Bagley, Steven C., White Halbert, Golomb Beatrice A. (2001) "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain." Journal of Clinical Epidemiology, 54: 979–985.

[36] Zhou, C., Li, M., & Yu, S. (2022). Intelligent Grouping Method of Science and Technology Projects Based on Data Augmentation and SMOTE. Applied Artificial Intelligence, 36(1), 2145637.

[37] Belete D.M & Huchaiah M.D(2021).Grid search in hyperparameter optimization of Machine learning models for prediction of HIV/ AIDS test results. International Journal of Computers and Applications.

[38]] Hutter F, Kotthoff L, Vanschoren J.(2019). Automated machine learning: methods, systems, challenges. Springer Nat. 2019;1:219

[39] Zoller M-A, Huber MF(2019). Benchmark and survey of automated machine learning frameworks, arXiv preprint arXiv: 1904.12054, 2019

[40] Kaur J.(2019). Factors affecting customer attrition a study of customer defection in banks. http://hdl.handle.net/10603/336505.

[41] Kumar A.(2021)Ensemble Technique on Predictive Analysis and Fraud Orders Detection using Supervised Machine Learning Algorithms in Supply Chain Management. Volume 12, Issue 7, July 2021: 12161-12175.

[42] Ahn J., Han S., & Lee Y.(2006). Customer Churn Analysis: Churn Symptoms and Consequences of Few Discrimination in the Korean Business Industry. Communication Policy, Volume 30, Issues 10 - 11, November - December 2006, Pages 552-568, Copyright © 2008 Elsevier Ltd, https://doi.org/10.1016/j.telpol.2006.09.006.

[43] Strzelecka A., Kurdyś-Kujawska A., Zawadzka D.(2020).Application of logistic regression models to assess household financial decisions regarding debt, Elzevier, https://doi.org/10.1016/j.procs.2020.09.055