INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, ALLAHABAD
VII Semester B.Tech in Information Technology

Mini - Project

Customer Churn Analysis and Prediction

Under the supervision of :-
Dr. Pavan Chakraboty

By:
Trupti Pendharkar (IIT2018097)
Puja Kumari (IIT2018191)
Prabha Kumari (IIT2018195)

Tables Of Content

1.Introduction:

Customer churning which can also be referred as customer attrition refers to the situation when customers in the bank tend to leave that bank, so the relationship between customer and the bank comes to halt, and churn rate during a particular period is the degree to which customers stop using the brand of bank that is end the relation with the bank . As customer churning is rising with years and as it has a direct negative impact on the revenue of the organisation it is bad for the bank. Customer churn may be because of several reasons like some other bank providing financial services at low charges or due to low interest rates or due to location of bank branch etc. So in order to stop this, bank need to prepare a prediction model in advance which would be able to predict the behaviour of the customer in advance about the customer that if he/she is going to leave the organisation or not and hence the company will get the good opportunity to provide some advantages to such specific customers going to leave the company. And we can achieve all these aforesaid objectives with the help of data mining and Machine learning techniques.

In this paper we have used different models like ANN, XGBoost, Stacking Classifiers and presented the analysis of the different results obtained.

2.Motivation:

Nowadays, there are almost 1.5 million customers that are churning in a year that is rising every year.The Banking industry faces challenges to hold clients. The clients may shift over to different banks due to reasons like better financial services at lower charges, bank branch location, low-interest rates, etc. Thus, prediction models are utilized to predict clients who are probably going to churn in the future. Because serving long-term customers is less costly as compared to losing a client that leads to a loss in profit for the bank. Also, old customers create higher benefits and provide new referrals.

3.Problem statement:

With the rise in competitive market customer churn becomes an important issue for the banks. So it's very necessary for the bank to find an efficient method which warns the system of possible customer churn by mining information which can lead to churn from large existing customer data. The cost of finding a new customer is much higher than that of maintaining old customers. To resolve the problem the computing age will help in the prediction of attrition. Data Mining and Machine Learning Techniques can be used on the data available by analyzing data from different viewpoints such as cause of attrition, its frequency , age of a customer, gender , rate of interest, geography etc. to create a model to predict if an customer is going to leave the brand or not for given user details.

4. Literature survey:

S. No.	Paper Title	Name of the Conference/journal (Year)	Objective	Methodology	Results
1	Analysis and prediction of bank user churn based on ensemble learning algorithm[1].	2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA) Date of Conference: 22-24 Jan. 2021 Conference Location: Shenyang, China	The purpose of this paper is to analyze the quarterly user data of banks, and establish user churn prediction model by using ensemble learning so as to improve the accuracy of prediction, so as to achieve the purpose of helping banks save costs.	They used algorithm such as Catboost, Lightgbm, Random Forest to build their model.	at the end of each quarter, the customer churn rate with a large amount of deposit or financial products is quite low, so they should focus on those users who have little deposit.

2	Application of Machine Learning and Statistics in Banking Customer Churn Prediction [2].	2021 8th International Conference on Smart Computing and Communications (ICSCC) Date of Conference: 06 September 2021 Conference Location: Kochi, Kerala, India	They are willing to make a website which is useful for the bank managers and decision makers of the bank to get an idea of those customers who are likely to leave the services of the bank in future and can retain them by formulating some new policies.	They used Machine Learning Techniques like SVM and Statistical Analysis like Analysis of Numerical Data and Analysis of Categorical Data. Also they used 'Flask' framework to create the web application along with 'HTML' and 'CSS' .	The model predicts the probability of the customer's leaving the bank and continuing the services of bank . The accuracy is around 84.15 %.
3	Customer Churn Analysis and Prediction in Banking Industry using Machine Learning [3].	2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) Date of Conference: 6-8 Nov. 2020 Conference Waknaghat, India	They are aimed to use different models of machine learning to the bank dataset to predict the probability of customer who is going to churn. The comparison in terms of performance like accuracy, recall, etc. is presented.	They used algorithm such as Logistic regression (LR), decision tree (DT), K-nearest neighbor (KNN), random forest (RF).	They observed that stratified and cross validation performs better in each case among all classifiers. But DT classifier has a .4429 recall value and 85.20% accuracy that is better as

					compared to others.
4	Machine Learning Based Customer Churn Prediction In Banking[4].	2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) Date of Conference: 5-7 Nov. 2020 Conference Location: Coimbatore, India	In this paper, a method to predicts the customer churn in a Bank, using machine learning techniques, is proposed.	The KNN, SVM, Decision Tree, and Random Forest classifiers are used in this paper. Also, some feature selection methods have been done to find the more relevant features and to verify system performance.	Result shows that the DT and RF classifiers accuracy increased after oversampling, but there is no change in KNN accuracy with regard to oversampling and SVM is not suitable for huge amounts of data. For KNN accuracy is 81.65%, SVM : 79.63% ,DT-78.99% , RF-85.18% .
5	An Enhanced Bank Customers Churn Prediction Model Using A Hybrid Genetic Algorithm And K-Means Filter And Artificial Neural Network[5].	2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA)Conference: 23-25 Feb. 2021 Conference Location: Abuja, Nigeria	They are proposed to used data mining techniques ANN and shown the performance of model .	They used Artificial Neural Networks (ANNs) and two filters were applied to the data, the Genetic Algorithm (GA) and K-means filter .	The results show that the training performance improved as the noise in the data reduces while the testing results were not improved with filtering.
6	Prediction of Customer Status in	2020 International Joint Conference	This paper presents a computer system	They used two different classifiers	This study shows that the data mining

	Corporate Banking Using Neural Networks[6].	on Neural Networks (IJCNN) Date of Conference: 28 September 2020 Conference Location: Glasgow, UK	that is based on the application of artificial neural networks and support vector machine and used to predict the future status of corporate banking clients.	and compared their result : a multilayer perceptron and a support vector machine.	techniques based on proper definition of input attributes and application of artificial neural networks provides a good tool for supporting the prediction of customer behaviour in corporate banking.

5. Dataset Description:

We will use the dataset available at kaggle named “Bank Customer churn Prediction”
Link : Dataset
Dataset consist of total 14 features

- Total record in dataset : 10000
- Total features of each record : 14

First ten features of the dataset are described as follows:

1. CustomerID : Describes Id of customer (Numerical Value)
2. Surname : Target variable
3. Credit Score : Integer Value.
4. Geography : Location Of Customer
5. Gender : String value
6. Age : Numerical value

7. Tenure : Numerical values.
8. Balance: Numerical Value
9. HasCrCard: Binary Value
10. IsActiveMember: Binary Value

And there are many more important features like estimated salary , exit status etc.

Language and tools:

Language: Python
Libraries: numpy,pandas
Web Interface using Flask

METHODOLOGY:

1. Exploratory Data Analysis and Feature engineering
2. Building Model
3. Performance evaluation
4. Integrating Model with Interface via Flask

6.1 Exploratory Data Analysis and feature engineering :

1. Customers with credit score less than 400 have higher chances to churn

Data exploration is the core part of a DM and ML project. With the help of Exploratory Data Analysis we have found that there are no missing values and we have dropped the irrelevant features from the dataframe like in our case roll no., customer id, and surname are irrelevant so we drop those features.

Also some data of some features like Geography and Gender were transformed from categorical data into numerical form.

The results from the Data Analysis were like:

2. Customers with 3-4 products have higher chances to churn.

3. Customers lying in the Age-Gap of 40-70 have higher chances to churn.

After analysis we have done feature scaling using the MinMax scaling algorithm. In this paper 'CreditScore', 'Age', 'Tenure', 'Balance', 'EstimatedSalary' are some of the features which were scaled down.

6.2 Data Preprocessing

While building the model it was found that data was highly imbalanced so we also performed under sampling , oversampling and used a smote classifier inorder to deal with an imbalanced dataset.

SMOTE(Synthetic Minority Oversampling Technique) : The basic problem of an imbalanced dataset is that the model has very few examples of the minority class in order to effectively design a decision boundary, therefore SMOTE is the most used technique in order to built(Synthesize) new examples. SMOTE works by choosing examples which are close to feature space thereafter marking a line among the examples and thereby drawing a new sample at a point along the line.

6.3 Building Model

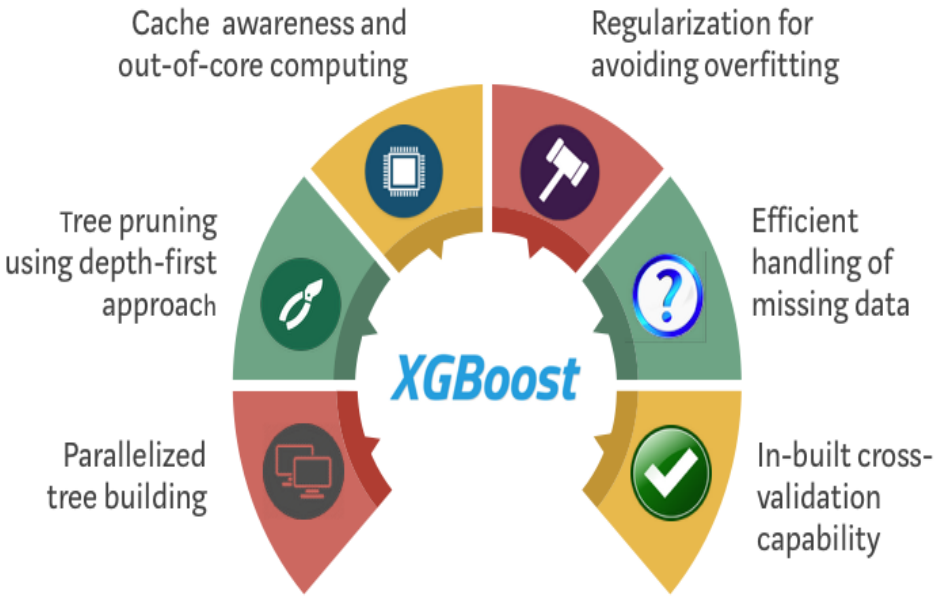
As per the literature review we have seen that in most of the research papers the training of data has been done using Random Forest, Decision Trees ,SVM or Logistic Regression. We know that when it comes to efficient classifiers ANN and xgboost have always proved themselves most of the times in terms of accuracy.So we have trained the dataset using ANN, XGBoost and Stacking Classifier as an ensemble model . In stacking Classifier we have used ANN, XGBoost as sub models and XGBoost as meta model.

1. XgBoost (Extreme Gradient Boosting): is one of the best known ensemble techniques and shows great performance and speed in tree based (sequential decision trees) machine learning algorithms. It proceeds towards the process of building sequential trees by using parallelized implementation. It is an

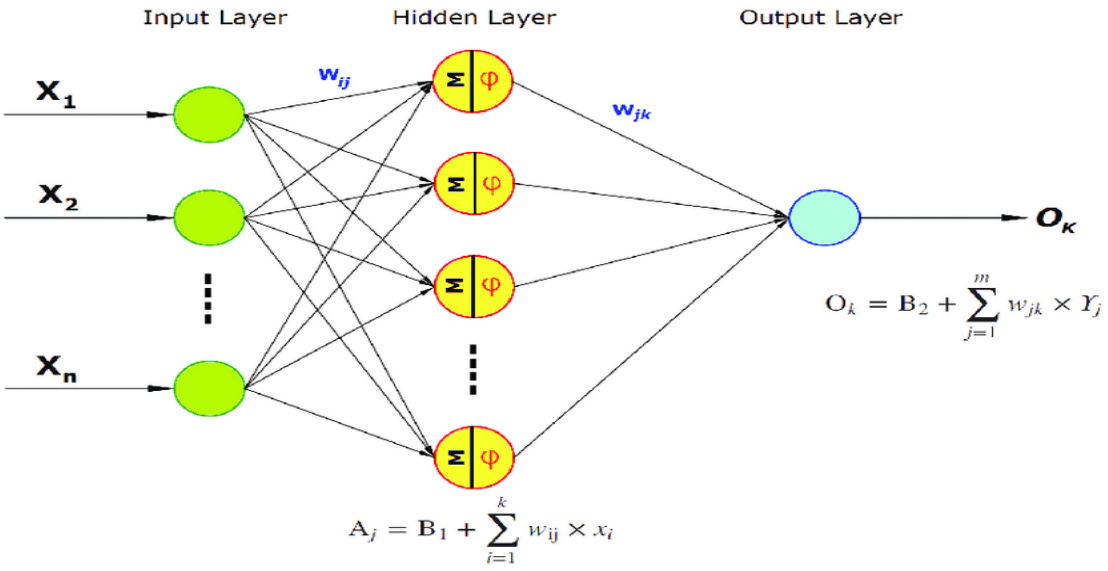
advanced implementation of gradient boosting algorithm along with some regularization factors.

Xgboost has an option to penalize complex models through both L1 and L2 regularization which helps in preventing overfitting

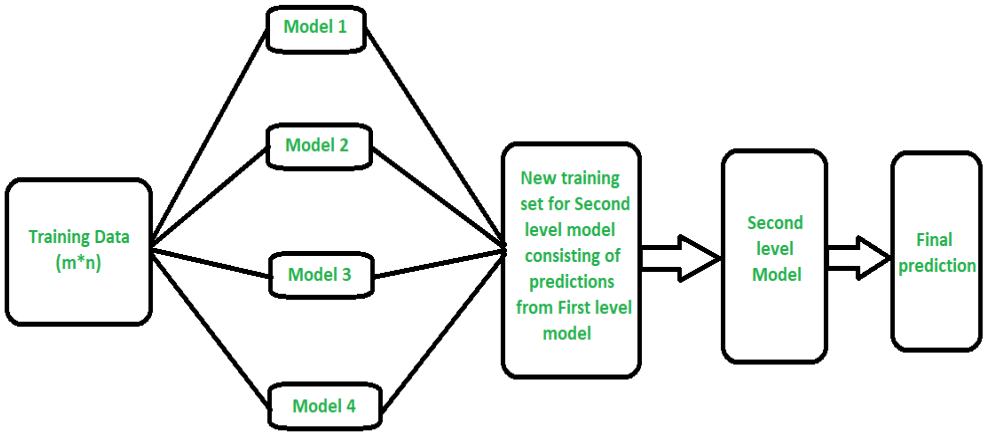
Xgboost comes with a built-in Cross Validation function.



2. ANN (Artificial Neural Network): We will also build a model using Feed Forward ANN also known as MultiLayer Perceptron. It consists of Input, hidden and output layers, and these multiple layers of nodes are like directed graphs where each layer is fully connected to the next one and works on the concept of weight optimization using backpropagation algorithm.



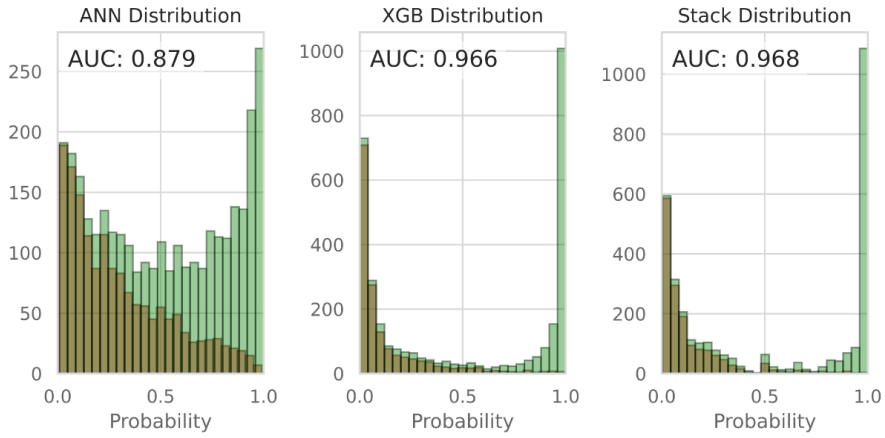
3. Stacking Classifier : Stacking is an ensemble technique where 2 or more heterogeneous(different) learners are combined and trained by meta-model. Input to this meta-model is actually outputs predictions based on the multiple predictions returned by these heterogeneous sub models .
   So basically we need to define 2 things in order to build our stacking model n(>=2) sub models we want to fit and a meta model that combines them.



PERFORMANCE EVALUATION

We have analysed the result on the basis of AUC score:

Data	ANN	XGBoost	Stacking Classifier
Using SMOTE data preprocessing	0.879	0.966	0.968



Output(AUC score for data preprocessed with smote)

7. Design of Web Application

Now we will make a web application using flask as backend framework and our model will run on a flask server and via web interface inputs will be provided to the model through APIs which will predict the output.

Activity Schedule:

Steps	Time Required	Predicted Date and time
Requirement Verify	Done	5th September 2021
Project Planning	Done	8th September 2021
System Design	Ongoing	20th September 2021
Details Design	10 days	30th September 2021
Coding	15 days	15th October 2021
Debugging and coding	15 days	30th October 2021
Testing	5 days	5th November 2021
Documentation and Final	8 days	13th November 2021

References:

[1]. Saadat M Alhashmi 2019 International Conference on Digitization (ICD) https://ieeexplore.ieee.org/document/9105767

[2].Richard Joseph; Shreyas Udupa; Sanket Jangale; Kunal Kotkar; Parthesh Pawar 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) https://ieeexplore.ieee.org/document/9432259

[3].Rachna Jain; Anand Nayyar, 2018 International Conference on System Modeling & Advancement in Research Trends (SMART)
https://ieeexplore.ieee.org/document/8746940

[4]. Sepideh Hassankhani Dolatabadi; Farshid Keynia 2017 2nd International Conference on Computer and Communication Systems (ICCCS)
https://ieeexplore.ieee.org/document/8075270

[5]. A Rohit Hebbar; Sanath H Patil; S. B Rajeshwari; S S M Saquaf
2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)
https://ieeexplore.ieee.org/document/9012243

[6]. Sandeep Yadav; Aman Jain; Deepti Singh
2018 IEEE 8th International Advance Computing Conference (IACC)
https://ieeexplore.ieee.org/document/8692137