IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

Churn Prediction using Various Machine Learning Algorithms

Leena Mandurkar[1], Sawali Khanke[2], Rani Khandaskar[3], Anmol Ukey[4]
Professor, Department of Computer Science & Engineering[1]
Students, Department of Computer Science & Engineering[2,3,4]
G.H Raisoni Institute of Engineering and Technology, Nagpur, Maharashtra, India
An Autonomous Institute Affiliated to Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur, Maharashtra, India
Accredited by NAAC With A+ Grade

Abstract: *In the era of big data, customer churn is a big problem faced by banks in the increasingly competitive market. The number of service providers are being increased very rapidly in every business. In these days, there is no shortage of options for customers in the banking sector when choosing where to put their money. In this paper, a method to predicts the customer churn in a Bank, using machine learning techniques, which is a branch of artificial intelligence is proposed. The research promotes the exploration of the likelihood of churn by analyzing customer behaviour. The KNN, SVM, Decision Tree, and Random Forest classifiers are used in this study. Also, some feature selection methods have been done to find the more relevant features and to verify system performance. The experimentation was conducted on the churn modelling dataset from Kaggle. The result gives us that in which algorithm the customer will stay or exits according to the data.*

Keywords: Customer Churn In Bank, Decision Tree, K-Nearest Neighbours, Logistic Regression , Random Forest, SVM , X-G Boost algorithm, Flask

I. INTRODUCTION

The market is very dynamic and highly competitive nowadays. It is because of the availability of a large number of service providers. The challenges of service providers are finding the changing customer behaviour and their rising expectations. The raising aspirations of current generation consumers and their diverse demands for connectivity and innovative, personalized approaches are very distinct from previous generations of consumers. They are well educated and better informed of emerging approaches.

The bank can analysed relationship between user churn and user's personal attributes, consumption attributes and financial attributes through the model, so as to formulate effective control strategies to minimize customer churn rate and save costs.The loss of commercial banks' customers is serious, and the loss rate can reach 20%. The cost of acquiring new customers is 5 times that of maintaining old customers. The bank can analyse the relationship between user churn and user's personal attributes, consumption attributes and financial attributes through the model, so as to formulate effective control strategies to minimize customer churn rate and save costs.

II. LITERATURE SURVEY

The analysis of the client churn in banking is a really broad area. In another study, a scientific study of the use of Machine learning in the extraction of information from repositories in the banking sector is presented. The findings show that customers who use more banking services (products) seem to be more loyal, so the bank can concentrate on those customers who use fewer than three products and sell them goods as per their needs.[3]

There are several churn predictions models in banking industry and other financial institutions. They mostly applied data mining and machine learning approaches to solve the problems identified churn customers using some criteria before they unsubscribe from a service or leave the business. [5]Customer churn prediction model was developed by

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

analysing historical behaviour data of defected customers, for early detection and retention purposes. Since the target class is unbalanced, the sample for learning is rebalanced by taking a sample of data to balance the two classes. The study began with oversampling by multiplying the churn class to fit with the other class. A random under-sampling approach was also used, which decreases the sample size of the broad class to be compared with the second class.

We studied many research papers in which various machine learning model like SVM, Linear Regression, Decision Tree, K-Nearest Neighbour, X-G Boost [2][3] etc. SVM is unable to handle large amount of Dataset. There are two type of Linear and Non Linear SVM. Various Machine learning and Like ANN(Artificial Neural network) [2]is also used for bank churning model. The highest Accuracy among all the machine learning model got from X-G Boost that is 82%. Likewise KNN ,SVM and LR also got accuracy around 81.37%,70.36%, 85% [1]

Authors	Title of Work	Year	Methodology	Remarks
Manas Rahman ,V Kumar	Machine Learning Based	ICECA-2020	KNN,SVM,	The dataset from kaggle is
	customer churn prediction		DT	oversampled and divided in
	in Banking			ratio 80:20,and in which
				decision tree got the
				highest accuracy of 91.98%
R Yahaya, O.A.Abisoye,	An Enhnaced Bank Customer	Cyber Nigeria	Genetic algorithm	From oversampled dataset we
S.A.Bashir	Churn Prediction Model	-2020	K-Means Filtering,	used various algorithms like
	Using a Hybrid G-Algorithm,		Artificial Neural	G-Means Algorithm,K-Means
	K-means Filter and ANN		Network.	Algorithm,Artificial Neural
				Network in which ANN
				got highest accuracy.
				which is 87.10%

III. METHODOLOGY

The data collection and experiments that were conducted in the study are described in this section.

3.1 Data Collection and Description of Dataset

It is common knowledge that banks do not reveal their customers' transaction or profile information because of its sensitive nature. Consequently, the study's dataset was downloaded from Kaggle which was uploaded on sep 24,2021.Using a source of 10,000 bank records, we created an app to demonstrate the ability to apply machine learning models to predict the likelihood of customer churn. Kaggle is a data science and machine learning community where students, professionals, researchers, and enthusiasts compete and share machine learning techniques as well as datasets. The dataset represented a collection of information from a fictitious bank.

DATASET DESCRIPTION

Feature Name	Feature Description
Row number	Row numbers from 1 to 10000.
Customer Id	Unique Ids for bank customer identification.
Surname	Customer's last name.
Credit Score	Credit score of the customer.
Geography	The country from which the customer belongs.
Gender	Male or Female.
Age	Age of the customer.
Tenure	Number of years for which the customer has been with the bank.
Balance	Bank balance of the customer.
Num of Products	Number of bank products the customer is utilizing(savings account, mobile banking, internet banking etc.).
Has Cr Card	Binary flag for whether the customer holds a credit card with the bank or not.
Is Active Member	Binary flag for whether the customer is an active member with the bank or not.
Estimated Salary	Estimated salary of the customer in Dollars.
Exited	Binary flag 1 if the customer closed account with bank and 0 if the customer is retained.

Table 2: Preprocessed Dataset description

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0

Table 3: Dealing with Imbalanced *Data*

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

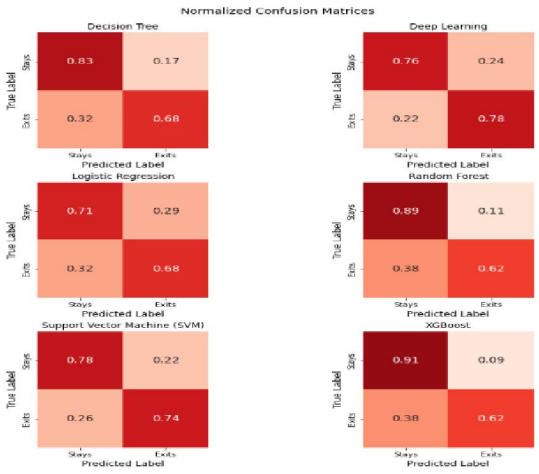
Today any machine learning practitioner working with binary classification problems must have come across this typical situation of an imbalanced dataset. This is a typical scenario seen across many valid business problems like fraud detection, spam filtering, rare disease discovery, hardware fault detection, etc. Class imbalance is a scenario that arises when we have unequal distribution of class in a dataset i.e. the no. of data points in the negative class (majority class) very large compared to that of the positive class (minority class).

Generally, the minority/positive class is the class of interest and we aim to achieve the best results in this class rather. If the imbalanced data is not treated beforehand, then this will degrade the performance of the classifier model. Most of the predictions will correspond to the majority class and treat the minority class features as noise in the data and ignore them. This will result in a high bias in the model.

Resampling data is one of the most commonly preferred approaches to deal with an imbalanced dataset. There are broadly two types of methods for this i) Undersampling ii) Oversampling. In most cases, oversampling is preferred over undersampling techniques. The reason being, in undersampling we tend to remove instances from data that may be carrying some important information. In this article, I am specifically covering some special data augmentation oversampling techniques: SMOTE and its related counterparts.

3.3 SMOTE: Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.



3.3.1. Clean the Data

By reading the dataset into a dataframe using pandas, we removed unnecessary data fields including individual customer IDs and names. This left us with a list of columns for Credit Score, Geography, Gender, Age, Length of time as a Bank customer, Balance, Number Of Bank Products Used, Has a Credit Card, Is an Active Member, Estimated Salary and Exited.

3.3.2.Analyse Initial Dataframe

Utilizing Matplotlib, Seaborn and Pandas, we next analyzed the data. We can see that our dataset was imbalanced. The majority class, "Stays" (0), has around 80% data points and the minority class, "Exits" (1), has around 20% data points. To address this, we utilized SMOTE in our machine learning algorithms (Synthetic Minority Over-sampling Technique). More on that later on. In percentage, female customers are more likely to leave the bank at 25%, compared

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

to 16% of males. The smallest number of customers are from Germany, and they are also the most likely to leave the bank. Almost one in three German customers in our sample left the bank.

3.3.3. Machine Learning using 7 Different Models

We tested seven different machine learning models (and used six in the final application) to predict customer churn, including Logistic Regression, Decision Tree, Random Forest, Deep Learning (TensorFlow), K-Nearest Neighbour, Support Vector Machine and X-G Boost. As mentioned earlier, we also used SMOTE to handle issues with the imbalanced data on the Support Vector Machine model. SMOTE (Synthetic Minority Over-sampling Technique) is an over-sampling method that creates new (synthetic) samples based on the samples in our minority classes. It finds the k-nearest-neighbours of each member of the minority classes. The new samples should be generated only in the training set to ensure our model generalizes well to unseen data. We used imblearn python package. Using SMOTE gave us better recall results which is a general goal for customer churning tasks.

3.4.4 Load Models to Display Predictions on App

Finally, using Flask and HTML/CSS, we created the user-facing app to add information to our data set matching our initial data frame to predict the likelihood of a customer departing the bank.

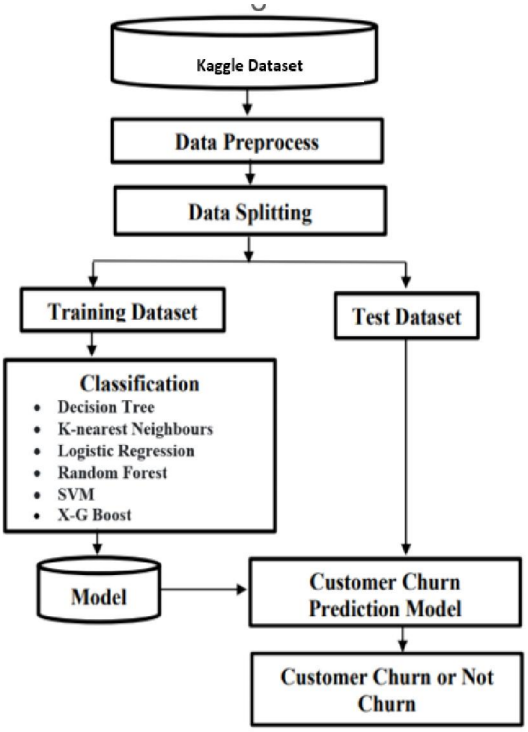


Figure: Activity Diagram of Proposed Diagram

3.4.Classifications

The classification methods were applied over the pre-processed data. KNN, SVM, Decision Tree (DT) and RF classifier, X-G Boosts, Logistic Regression are used for comparison of results. And also the comparison of results of different classifiers has been carried out over the selected features by different feature selection methods.

1. **K-nearest Neighbour (KNN) :** The KNN method is one of the easiest and most efficient non-parametric ways of classification, based on supervised learning . KNN works by identifying the k nearest samples from an existing dataset and when a new unknown sample appears, classify the new sample in the most similar

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

class. That is, the classification algorithm determines the test sample group by the k training samples that are the nearest neighbours. to the test sample and assign it to the class with the highest likelihood.

2. **Support Vector Machine (SVM):** Support Vector Machine is an efficient, supervised machine learning algorithm derived from wapnik theory of statistical learning. This has proved its success in the fields of classification , regression ,time series prediction, and estimation in geotechnical practice and mining science. SVM's main objective is to find an efficient distinguishing hyperplane that precisely categorizes data points and as far as possible, and distinguishes the points of two classes by reducing the possibility of misclassifying the training samples and unknown test samples. This implies that there is the maximum distance between two classes and the separating hyperplane. The Linear support vector machine (LSVM) model is used in this work. LSVM was originally developed to deal with binary class Problems.
3. **Decision Tree (DT):** A decision tree is a procedure that slices a collection of data into various branch-like segments. A tree of decisions is easy to read. This advantage makes explanations for the model simple. While another algorithm (like a neural network) can generate a much more accurate model in a given scenario, a decision tree could be trained to predict the neural network's predictions, thus opening up the neural network's "black box". Another benefit is that, in the correlation between the target variables and the predictor variables it can model a high degree of nonlinearity. A decision tree is composed of two major strategies Tree creation and Classification.
4. **Random Forest (RF):** Brei man presented RF as an ensemble classifier for tree learners. The method employs several decision trees so that each tree relies on the values of an individually selected random vector with the same distribution for all trees. Right choice for the tendency of decision trees to overfit their training collection. In short, Random forests are actually a way to combine many deep decision trees which are learned on various sections of the same dataset with the target of decreasing the variance. The real advantage of using RF is it comes with quite high dimensional data, with no need to perform dimensionality reduction and feature selection. The training rate is also higher and ease to use in parallel models.
5. **X -G Boost:** XG Boost is an implementation of Gradient Boosted decision trees. XG Boost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XG Boost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.
6. **Logistic Regression (LR):** Logistic regression models discover the relationship among qualitative and other variables. In most models established with logistic regression, dependent variable has only two results. Usually, the emphasized event that is being realized is indicated by 1 and the one which is not realized by 0. The scientific society in the domains of economics, financial sector, and other social and environmental sciences gets now incorporated these models . The LR model is used to estimate the likelihood of an occurring event based on a set of predictors**.** LR have several Statistical flaws. Muti-collinearity and decreased performance accuracy are two of them.

3.5. Web Application Using Flask Framework

Flask is a lightweight Python web framework that provides useful tools and features for creating web applications in the Python Language. It gives developers flexibility and is an accessible framework for new developers because you can build a web application quickly using only a single Python file. Flask is also extensible and doesn't force a particular directory structure or require complicated boilerplate code before getting started.

Learning Flask will allow you to quickly create web applications in Python. You can take advantage of Python libraries to add advanced features to your web application, like storing your data in a database, or validating web forms.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

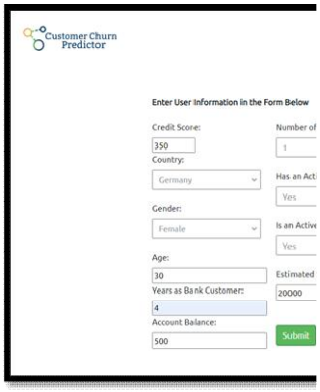
We are creating web application using Flask the python library for creating web page. In this Web page , various machine learning model is compared where the customer stays or exit .. You can see the web page on this url (https://bank-churn-predictions.herokuapp.com/) .

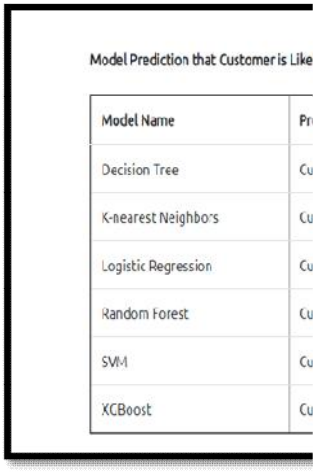
IV. RESULT AND DISCUSSION

When the pre- processing of the data has been completed, the data will be in the operational form. And the 10 features which are obtained after pre- processing is taken for the remaining study .Among that, 70% of data will be used for training and the remaining 30% will use for testing as random. The performance of classifiers varies when using different feature selection methods. There are 14 features given in our sampled datasets.

The best features selected to for getting accuracies is through analytical churn .csv files are Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Cr Card, Is Active Member, Estimated Salary, Exited. We have used various Machine learning model in our model like KNN (K – Nearest Neighbour), Decision Tree, RF (Random Forest Classifier), X - G Boost, SVM (Support Vector Machine), Logistic Regression. Various Machine learning Algorithms are used because in every algorithm we are testing that through which algorithm the Customer is likely to Exit and Stay.

We have used the python library Flask which is used for web applications. We are using Flask to create web page for our sample dataset to enter various features like Credit course, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Cr Card, Is Active Member, Estimated Salary. By entering various features in our web page we can identify that customer will stay or exit.





IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

V. CONCLUSION

While the banking sector is considered, like any other organization, customer engagement has become one of the primary concerns. To resolve this crisis, banks need to identify customer churn possibilities as quickly as possible. There are various studies ongoing in banking churn prediction. Different entities measure the churn rate of customers in various ways using different bits of data or information. The need for a system that can forecast the client churning in banking in a generalized way in the early stages is really important. The system needs to works with fixed and potential data sources that are independent of any service provider. And also the model must be in a form in which; can use minimal information and can give maximum throughput for the prediction. The model examined KNN, SVM, Decision Tree, Random Forest Classifier, X -G Boost, Logistic Regression.

REFERENCES

[1]. MACHINE LEARNING BASED CUSTOMER CHURN PREDICTION IN BANKING. (Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020))

[2]. An Enhanced Bank Customers Churn Prediction Model Using A Hybrid Genetic Algorithm And K-Means Filter And Artificial Neural Network (Proceedings of the 2020 IEEE 2nd International Conference on Cyberspace (Cyber Nigeria))

[3]. Analysis and prediction of bank user chum based on ensemble learning algorithm (2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA))

[4]. A study on Customer Churn of commercial banks based on Learning from Label Proportions (2018 IEEE International Conference on Data Mining Workshops (ICDMW))

[5]. Customer churn analysis in banking sector: Evidence from explainable machine learning models (JAME, Volume :1 - Issue :2 - Year: 2021)

[6]. PREDICTING CUSTOMER CHURN IN BANKING INDUSTRY USING NEURAL NETWORKS (Interdisciplinary Description of Complex Systems 14(2), 116-124, 2016)

[7]. Kaggle - Churn Modelling Classification Data Set

[8]. How to save a scikit-learn pipeline with Keras regressor inside to disk?

[9]. Churn Prediction in Banking System using K-Means, LOF, and CBLOF - Irfan Ullah; Hameed Hussain; Iftikhar Ali; Anum Liaquat (2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE))

[10]. Analysis and prediction of bank user churn based on ensemble learning algorithm- Yihui Deng; Dingzhao Li; Lvqing Yang; Jintao Tang; Jiangsheng Zhao (2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA))