Mini Project Report on

Customer Churn Prediction Model

Submitted by:

Ajitkumar Gupta	54
Jayesh Patil	63
Tejas Patil	66
Niranjan Taware	74

Under the guidance of:

Prof. Praveen Barapatre

Department of Computer Engineering

Vishwaniketan's Institute of Management Entrepreneurship & Engineering Technology

2021-22

Vishwaniketan's Institute of Management Entrepreneurship & Engineering Technology

2021-22

CERTIFICATE

This is to certify that the project report titled, **"CUSTOMER CHURN PREDICTION MODEL",** duly submitted by the following student

Name	**Roll No.**
Ajitkumar Gupta	54
Jayesh Patil	63
Tejas Patil	66
Niranjan Taware	74

has been completed under my supervision in a satisfactory manner in a partial fulfillment of the requirements for the award of semester VI Bachelor's Degree in **Computer Engineering** to be conferred by the **University of Mumbai.** In my opinion, the work embodied in this report is comprehensive and fit for evaluation.

PROF. PRAVEEN BARAPATRE
HOD (COMPUTER DEPARTMENT)

EXTERNAL EXAMINER

PROJECT REPORT APPROVAL SHEET

The Project Report Titled **"CUSTOMER CHURN PREDICTION MODEL"** submitted by the students

Name	**Roll No.**
Ajitkumar Gupta	54
Jayesh Patil	63
Tejas Patil	66
Niranjan Taware	74

Is examined by the board of examiner and approved for further perusal.

Sign: Sign:

Name: Name:

(Examiner -I) (Examiner - II)

Date:

Place:

DECLARATION

We declare that this written submission represents our ideas in our own words and where others ideas or words have been included, we have adequately cited and referenced the original sources.

We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission.

We understand that any violation of the above will be cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Ajitkumar Gupta

Jayesh Patil

Tejas Patil

Niranjan Tawade

Date:

ACKNOWLEDGEMENT

We are profoundly grateful to Prof. Praveen Barapatre for his expert guidance and continuous encouragement throughout to see that this project meets its target since its commencement to its completion.

I would like to express a deepest appreciation towards Dr. B.R. Patil(Principal, ViMEET) and Prof. Praveen Barapatre, (Head of the Department of Computer Engineering, ViMEET) for their valuable guidance and support in completing this project.

We express our sincere gratitude to all the staff members of the computer engineering department who helped us directly or indirectly during this course of work.

Thank You

Content

Topics	Page. No
1. Abstract	7
2. Introduction	8
3. Objectives	9
4. Literature Review	9
5. Methodology	10
6. Results	13
7. Conclusion	14
8. References	15

Abstract

The number of service providers are increasing very rapidly in every business. These days, there is no shortage of options for customers in the banking sector when choosing where to put their money.

As a result, customer churn and engagement has become one of the top issues for most of the banks.

With increased competition in the banking business, banks must implement client retention tactics while attempting to improve their market share by gaining new customers.

This study evaluates the effectiveness of six supervised classification approaches to offer an efficient model to forecast customer turnover in banking business, using 10 demographic and personal data from 10000 customers of European banks.

The effects of feature selection, class imbalance, and outliers will be addressed for the two competing models, ANN and random forest.

As demonstrated, unlike random forest, ANN does not exhibit any severe concerns about overfitting and is also noise resistant.

As a result, the highest performing classifier is an ANN structure with five nodes in a single hidden layer.

Introduction

With increased competition in the banking business, banks must implement client retention strategies while attempting to improve their market share by gaining new customers.

It has been demonstrated that increasing the retention rate by up to 5% may improve a bank's profit by up to 85%.

Furthermore, recruiting new consumers is more expensive for any organization than maintaining old ones, who are more likely to make profit.

As a result, banks should preserve their competitive edge by using machine learning algorithms to forecast client turnover.

Utilizing multiple supervised classification approaches, this study attempts to provide an effective prediction model for customer turnover in the banking industry using a randomly chosen population of 10000 customers from three European-based banks.

The next sections will go into model performance, goodness of fit, feature selection, class imbalance, and dealing with outliers.

- The cost of attracting new customers can be five to six times more than holding on to an existing customer.
- Long term customers become less costly to serve, they generate higher profits, and they may also provide new referrals
- Losing a customer usually leads to loss in profit for the bank.

Objectives

Customer churn has become a major concern at many banks since it is far more expensive to gain a new customer than it is to keep an existing one.

A customer churn prediction model may be used to identify potential churners in a bank, allowing the bank to take measures to keep them from leaving.

• Availability of latest technology
• Customer-friendly bank staff
•Low interest rates
•Location
•Services offered
•Churn rate usually lies in the range from 10% up to 30%.

Literature Review

While most studies have been done for the telecommunications and communication sector, customer churn analysis has also found applications in a wide range of fields, including e-commerce, banking, insurance, retail trade, energy, games and entertainment, and the medical. Because this study is a customer churn analysis application in banking, it will be focused on this sector.

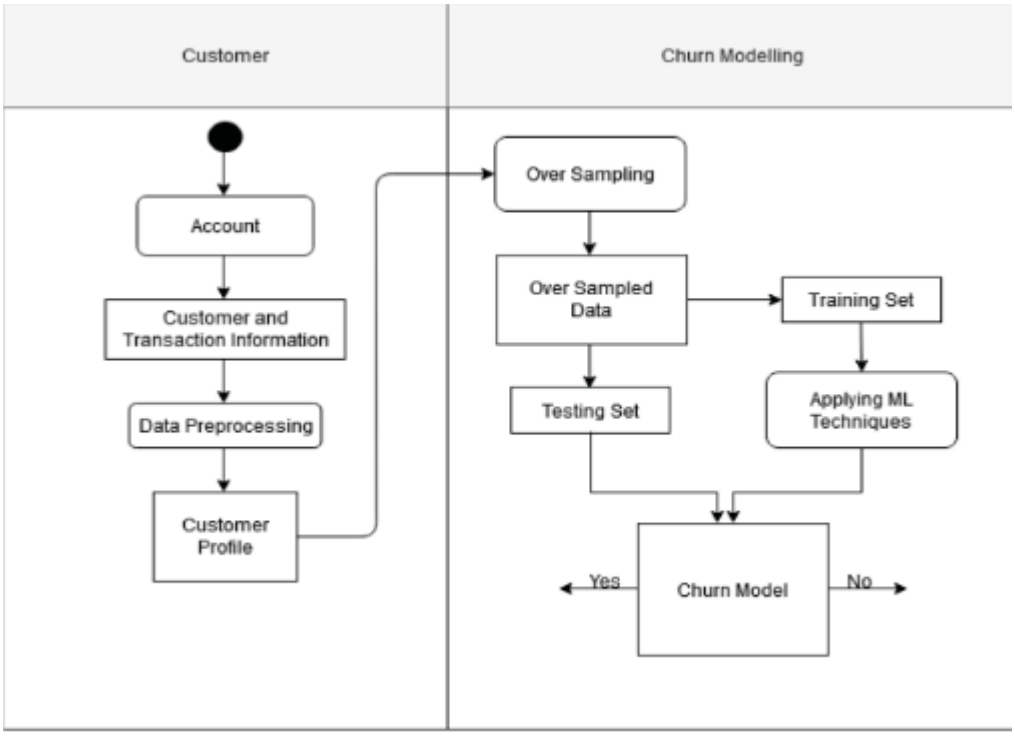
Customers are encouraged through strong customer relationships to be always more satisfied and loyal to the bank, according to many studies.

According to Ozatac, the most significant determinants of customer satisfaction in banking are the accuracy of the information, the responsiveness of employees, access to all services, ability of employees, reliability, security of financial transactions, personalization, and consistency.

In 2013, lead to customer satisfaction are punctuality, effective communication, direct and acceptable information, efficient employee services in banking. Customer satisfaction, brand loyalty, pricing policy, and service quality were found to be the most important determinants of customer loyalty in their study.

Methodology

This work aims to predict customer churn in a commercial bank as early as possible using efficient data mining methods. A diagrammatic representation of the proposed model is given in figure below:



A. Dataset description:

The dataset used in this analysis was obtained from Kaggle to model churns. The dataset includes information of 10000 bank clients, and the target parameter is a binary variable that represents whether the customer has left the bank or is still a customer.

Of this, 7963 were positive class(maintained) samples and 2037 were negative class(excited) samples. The target variable reflects the binary flag 1 when the client has a bank account closed, and 0 when the client is retained.

The dataset contains 13 feature vectors(predictors) that were reported from customer data and transactions processed by the customer. The details of these features are given in Table II.

Feature Name	Feature Description
Row Number	Row numbers from 1 to 10000.
Customer Id	Unique id 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	Numbers 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.
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 an account with the bank and 0 if the customer is retained.

B. Data Preprocessing

Preprocessing the data is a significant phase in the process of data mining. Since they have a direct effect on task success rate. It must deal with irrelevance, noisiness, and unreliability of data. And if necessary, the data conversion too. Predictors descriptions after preprocessing

are listed in Table III. These are the attributes taken for deciding on churn prediction in this Study.

1) Irrelevancy
2) Transformation

C. Feature Selection

In machine learning, the process of identifying a subset of appropriate predictors to be used in model building is called feature selection. The selection phase of the feature is very critical as it will help to shorten the training time, escape the high-dimensionality curse, and above all simplify the model.

1) mRMR (Minimum Redundancy Maximum Relevance)
2) Relief

D. Over Sampling:

In data processing, oversampling and undersampling are strategies used to configure the class distribution of given data. Since the data is highly imbalanced(7963 positive class sam- ples and 2037 negative class) and the size of the available data sample is small, this study will make use of the oversampling technique. Because if undersampling is preferred, the size of data will decrease in a way that enough data will not be there to build the model.

Hence, this study is using random oversampling by resampling the minority class(negative class).

E. Classification

The classification methods were applied over the prepro- cessed data. KNN, SVM, Decision Tree(DT) and RF classifiers 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 Neighbor (KNN)
2) Support Vector Machine (SVM)
3) Decision Tree (DT)
4) Random Forest (RF)

Results:

When the preprocessing of the data has been completed, the data will be in the operational form. And the 10 features which are obtained after preprocessing is taken for the remaining study. Among that, 70% of data will be used for training and the remaining 30%will be used for testing as random.

The classifiers will be used alone and along with the specified feature selection methods. And each model is evaluated by the accuracy which is obtained after a 10 fold cross-validation. And the random confusion matrix was also produced for each model.

The performance of classifiers varies when using different feature selection methods. The features selected in each feature selection method and the classifiers parameter details will describe in the following paragraphs.

For KNN, the k-value is set to 5. That is, the nearest five neighbors are considered for classifying the new data. By reducing the neighbors to 5, sometimes the accuracy is increasing and vice versa. But, since the data is taken randomly for the classification it is not a good practice to select fewer neighbors. But when the number of neighbors is greater than 5, the result is highly decreasing.

Hence, the value of k is selected as 5(in which the accuracy and the change is optimized). And the distance measure used is Euclidean distance. For SVM, the linear kernel function is used(LSVM).

In the case of RF, the number of trees in the forest is set as 100. All these parameters are selected based on the optimization of classification accuracy.

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 work with fixed and potential data sources that are independent of any service provider.
- Boosting has given the increased accuracy of 86.85 with low error, high sensitivity and specificity. Organizations periodically calculate customer churn in multiple aspects.
- Churning can be the number of customers lost, ratio or percentage of customers lost compared with total customers in the bank.
- Churn can be calculated on a quarter or annual basis.
- An accurate forecast can give insights on the future using which a strategy can be formulated.

References

1) D.-R. Liu and Y.-Y. Shih, “Integrating ahp and data mining for product recommendation based on customer lifetime value,” Information & Management, vol. 42, no. 3, pp. 387–400, 2005.

2) G. Canning Jr, “Do a value analysis of your customer base,” Industrial Marketing Management, vol. 11, no. 2, pp. 89–93, 1982.

3)R. W. Stone and D. J. Good, “The assimilation of computer-aided marketing activities,” Information & management, vol. 38, no. 7, pp. 437–447, 2001.

4) M.-S. Chen, J. Han, and P. S. Yu, “Data mining: an overview from a database perspective,” IEEE Transactions on Knowledge and data Engineering, vol. 8, no. 6, pp. 866–883, 1996.

5)M.-K. Kim, M.-C. Park, and D.-H. Jeong, “The effects of customer satisfaction and switching barriers on customer loyalty in Korean mobile telecommunication services,” Telecommunications policy, vol. 28, no. 2, pp. 145–159, 2004.