CUSTOMER CHURN PREDICTION

INTRODUCTION:

There are many competitors in the banking sector nowadays and those competitors are also ready to provide higher quality and lower prices for the same products and services.

So customer are shifting loyalties from bank to bank. When products or services do occur, customer churn or attrition occurs. There are two kinds of churns like voluntary or involuntary churning. The customer leaves the bank or stops using involuntary churners which the bank removes users. Voluntary churn arose when customer immediately stopped using products or services. This study is focused on the voluntary churn. To prevent this, the prediction of customer churn is necessary to determine whether customer can discontinue products. In recent years, data mining has become popular in the research industry and in society as a whole, due to the enormous availability of large amounts of data and the need to turn such data into useful information and knowledge. Using SVM, bank customer churn prediction is developed among data mining techniques, because it is widely used binary classification technique and performs good classification accuracy in small dataset. But it increases time complexity in large data set so that clustering time complexity in large data set so that clustering of **K-**means is used to effectively cluster high volume data before SVM to reduce operating time and improve prediction accuracy.

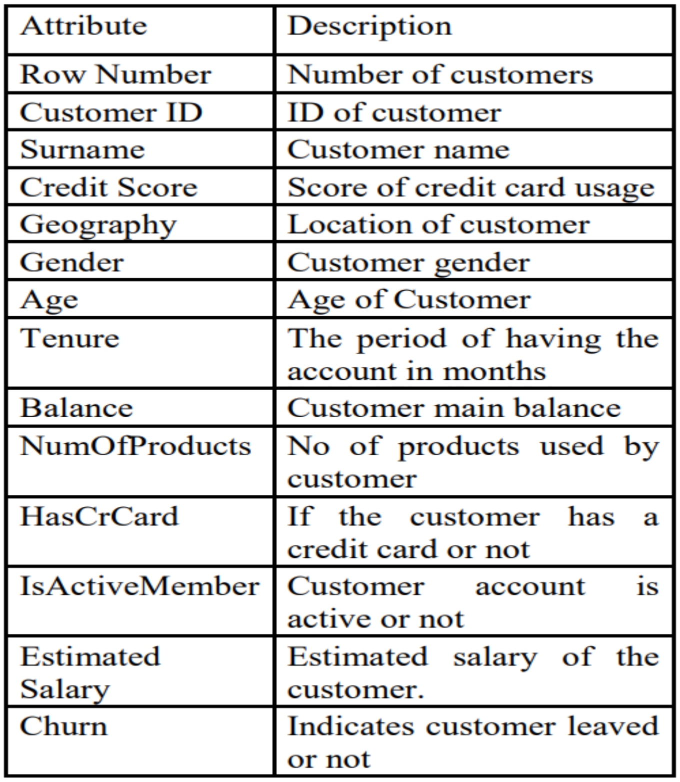
The purposes of this research are to identify the important factors that drive the churn, to support information on which marketing actions will have the greatest retention impact on each individual customer, to reduce advertising costs for new customers, and to avoid losing revenue that results from a customer abandoning the bank.

2. Related Works :

Many authors tackled customer churn analysis in a number of domains (both contractual and non-contractual) such as telecommunications, banking and insurance, subscription services, game businesses and also retail. This churn prediction is based on socio-demographic attributes, balance, account level and behavioural attributes of the customers. Models were developed using data mining techniques to forecast possible churns with satisfactory performance. Several data mining techniques were used, such as decision trees, neural network, K Nearest Neighbor, logistic regression, random forests, SVM, linear and quadratic discriminate analysis, GA, Markov model, cluster analysis and optimal analysis. The authors, B.He et al., proposed SVM on the customer data of Chinese commercial banks on both attribute indicators and business attribute indicators with random sampling method. They used random sampling method to solve the problem of class imbalance and get good results of prediction. Also, SVM (both linear and radia based function) was compared with or without sampling method and logistic regression. And they found the method they proposed provided good results[1]. The authors, A.P.Patil et al., made comparisons in the online retail dataset between SVM, Random Forest and Extreme Gradient Boosting and found that the accuracy of these three algorithms was extremely higher, but computation time increased in the same order[2].. The authors Fa-Gui Liu et al. proposed combined method on the telecom dataset of Fuzzy K-prototypes and SVM to reduce operating time and improve prediction accuracy[3]. The authors, A.O.Oyeniyi et al., proposed a combined K-means and JRip algorithm method on the Bank dataset attributes of demographic, balance, dormancy status. They found the model was capable of producing good results [4]. The authors, F.Abdi et al., proposed combined clustering technique (K-means) and classification techniques (both neural networks and decision trees) on telecom dataset attributes of socio-demographic and behavioural character. The combined model proposed may be more.

3. Architecture Design: The proposed model is detailed in the following sections. Data collection Data is from the online.

Table 1. Dataset Description



Attribute	Description
Row Number	Number of customers
Customer ID	ID of customer
Surname	Customer name
Credit Score	Score of credit card usage
Geography	Location of customer
Gender	Customer gender
Age	Age of Customer
Tenure	The period of having the account in months
Balance	Customer main balance
NumOfProducts	No of products used by customer
HasCrCard	If the customer has a credit card or not
IsActiveMember	Customer account is active or not
Estimated Salary	Estimated salary of the customer.
Churn	Indicates customer leaved or not

This dataset includes 10k bank customer data records with 14 attributes including socio demographic attributes, account level and behavioural attributes.

3. Data pre-processing:

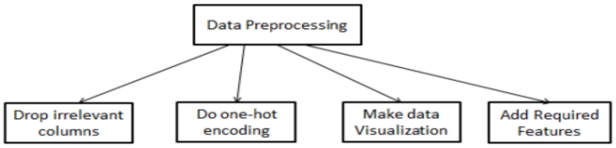


Fig 1. Data pre-processing

Pre-processing data consists of four parts, as shown in Figure 1. Firstly, sample data is pre processed by removing the irrelevant attributes (Row Number, Customer ID and Surname) that do not affect the prediction of churn. Then the strings or categorical attributes were converted using one-hot encoding to numerical values because K-means can only understand numerical attributes. And also make visualization of the data to know which features are mostly associated with churn. Based on the results of the visualization, add a few features to make the churn classification more effective. One thousand samples with 17 attributes are used amongst 10000 data records. And the pre processed data is then divided into 80% training and 20% testing data.

MODEL BULDING:

K-means clustering is used to cluster data with similar structure on trained data. After that, SVM is used to develop prediction model for classifying which customers are likely to be churner or non-churner on the output of clustering results.

Step 1: K-means clustering

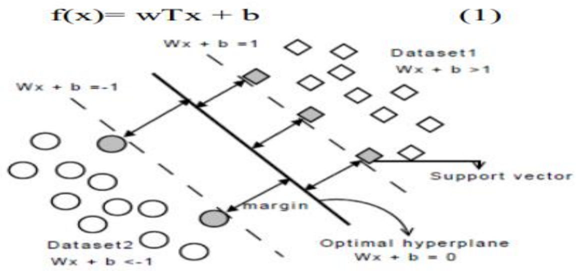
It is widely used and is a very simple technique to analyze huge amounts of data in clusters. And very quick, robust and easily comprehensible too. K-means Steps developed by Macqueenwhich

(1) Initialize K centre locations ($C_1, . . ., C_K$).
(2) Assign each data point (x_i) to its nearest cluster centre C_K.
(3) Update each cluster centre C_K as the mean of all xi that have been assigned as closest to it.
(4) Calculate D=? Using Euclidean distance.
(5) If the value of D has converged, then return ($C_1, . . ., C_k$)

Using the K-means clustering, 5 customer group similarities are performed on prepared data. And then analyze the results of the clustering to know the character of each group.

Step 2: Support vector machine (SVM)

SVM is used to develop a predictive model for classifying which customers on the output of clustering results are likely to be churner or non churner. Binary classification technique is widely used in model prediction. It can solve both linear and non-linear issues and work well for many practical issues. Hyper plane is used to separate data into classes, f(x) is function, w is weight, T is carriage, x is input and b is bias, as in (1) above.



As shown in Fig. (2), the margin between the classification face is defined as wTx+b=1 and wTx+b=-1 is 2/||w|| and minimizing the distance 2/||w|| is equivalent to maximizing ½ . Then the problem of seeking the optimal classification face is transformed into the following optimization problem:

$$Min_{w,b}\ \tfrac{1}{2}\ ||w||^2$$
$$\text{S.t }\ y_i\big((w.x)+b\big) \geq 1\ for\ any\ i = 1, \ldots, n$$

In this study, radius basis function (RBF) kernel trick is used to transform non-linear to linear classification. RBF kernel on two samples x and x´, represented as feature vectors in some input space, K (x, x´) = exp $\left(-\frac{||x-x'||2}{2a^2}\right)$

K (x, x´) = exp $(-\gamma||x - x'||^2)$

where,||x-x'||^2 may be recognized as the squared Euclidean distance between the two feature vectors, is a free parameter. And then, in SVM, GridSearchCV is used to find optimal hyper-parameters like (to use C or gamma values) that improve accuracy and predictive results. It searches the parameter grid by every combination of parameters. The first step is to create a dictionary of all parameters and their respective set of values which we want to test for best performance. The next step is to create a GridSearchCV-class instance. Once the GridSearchCV class is initialized, the final step is to call the class fit method and pass the training and test set to it. The next step after the method completes execution is to check parameters that return the highest precision.

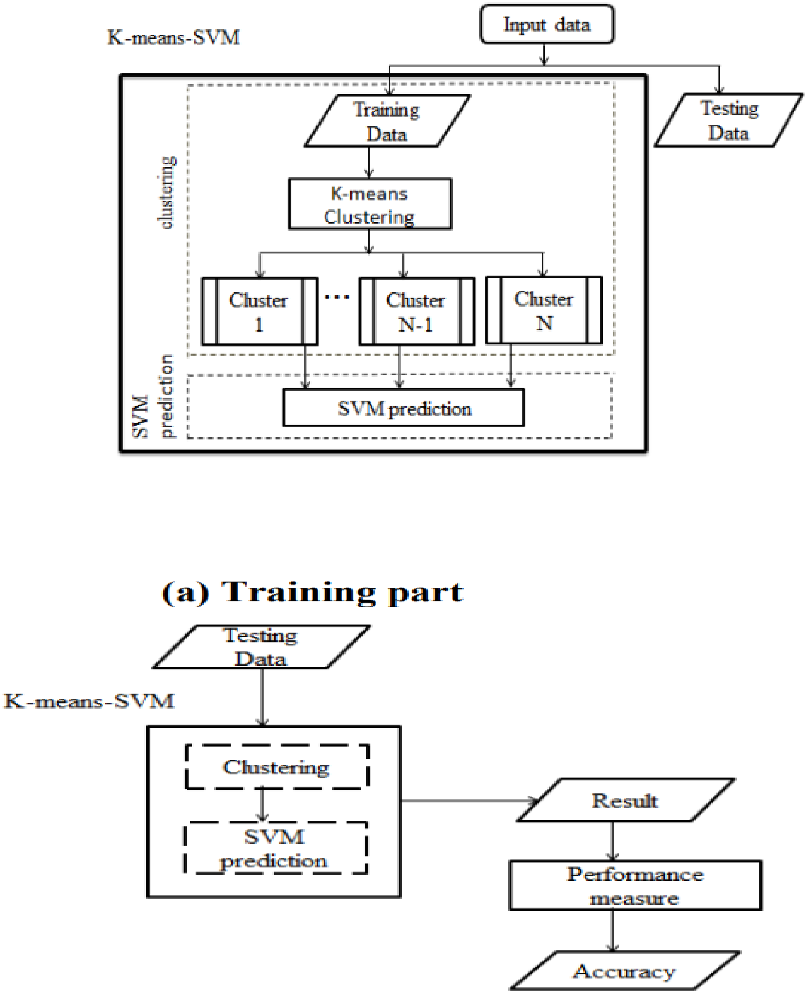
Testing:

The proposed combined k-means-SVM model is then used to test data on whether the model has properly classified churner or non-churner. Test results include clients who may be churner as 1 or clients who may be non-churner as 0.

Accuracy Measure:

Finally, F-measure also called (f1-score) is used to evaluate how much effectiveness proposed model is. It is a single matrix used to have good precision and recall. It provides the harmonic mean of precision and recall. F-measure = $\frac{2rp}{r+p}$

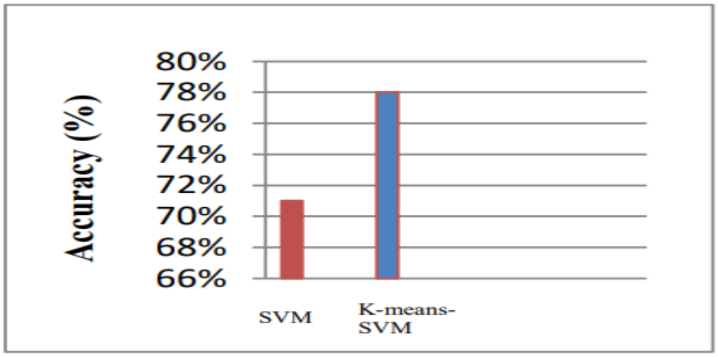
where, r= recall (true positive rate), p = precision in (4)



(a) Training part

(b) Testing part Fig. 3. Data Flow Diagram of K-means-SVM combined model (a) Training part and (b) Testing part

4. Customer Churn Prediction Expected Result



Comparison of SVM and K-means-SVM

As shown in figure 4, the accuracy of the proposed and only SVM algorithm is compared and it is analyzed that due to the clustering of uncluttered points from the dataset, the proposed model offers high accuracy. And the proposed model can not only decrease training time but can also react to marketing by targeting similar group of customers using clustering with K means**.**

<u>Conclusion</u>:

Customer churn analysis has become a major concern in almost every industry that offers products and services. The model developed will help banks identify clients who are likely to be churners and develop appropriate marketing actions to retain their valuable clients. And this model also supports information about similar customer group to consider which marketing reactions are to be provided. Thus, due to existing customers are retained, it will provide banks with increased profits and revenues. And also proposed combined model K-means-SVM reduces SVM training time by reducing support vectors