Chelonian Conservation And Biology

Vol. 18 No. 2 (2023) | https://www.acgpublishing.com/ | ISSN - 1071-8443

DOI: doi.org/10.18011/2023.12(2).2269.2289

ANALYTICAL CRM USING SOFT COMPUTING TECHNIQUES: A STUDY OF IMPLICATIONS OF FEATURE SELECTION

Kiran Babu Kommu
Assistant Professor, Dept. of Computer Science and Engineering (Data Science)
ACE Engineering College, Hyderabad, Telangana, INDIA -501301

Jaishri P. Wankhede
Assoc. professor, Dept. of Computational Intelligence, Malla Reddy College of Engineering and Technology , Maisammaguda, Medchal, Hyderabad, Telangana, INDIA -500100

Anitha Yajjala
Assistant Professor, Dept. of Computer Science and Engineering,Nalla Malla Reddy Engineering College,autonomous,Medchal, Hyderabad, Telangana, India- 500088.

Jaya Prakash Koyyalamudi
Assistant Professor, Dept. of CSE, Narasaraopeta Engineering College, Narasaraopet, Palnadu District, Andhra Pradesh, INDIA-522601

ABSTRACT

Analytical CRM has become the need of the hour as the data generated about customers is huge and imbalanced in nature. However, it become nearly impossible to analyse available data manually. Hence, soft computing community has reported various approaches to analyse and understand the data. In this research study we propose the use of Feature Selection approach prior to classification modelling in order to reduce the complexity of the classifier without compromising its performance. We have used Churn prediction in bank credit card customer data which is medium in size and highly imbalance in nature. Bank needs to understand about their customers (valuable) and device policies to retain them who are about to switch their loyalties to the competitor. As the problem at hand is about predicting possible churners, sensitivity is accorded priority in analysing the performance. Based on sensitivity yielded it is observed that the proposed approach i.e., using reduced features yielded better results compared to the results obtained using full feature data. Further, it is also observed that Support Vector Machine (SVM) performed better compared to Decision Tree (DT).

Keywords: *Feature Selection, Analytical CRM, Soft Computing, SVM, DT, Chi-Squared, Correlation, Gini Index, Information Gain.*

1. INRODUCTION

Customer Relationship Management (CRM) is has become one of the centre points for many industries such as Banking, Retail, Telecommunication and insurance. Recently CRM has become widely recognized as an important business approach, that enables systems supporting a business strategy to build long term, profitable relationship with specific customers (Ngai,Lixiu and chau, 2009).CRM is a process or methodology used to learn more about customers‟ needs and behaviours in order to develop stronger relationship with them. Business nowadays revolves around customers; it is observed that retaining an existing customer is easy and requires less efforts than acquiring new customers.

CRM involves the continuous use of refined information about current and potential customers in order to anticipate and respond to their needs. CRM draws on a combination of business process and Information Technology i.e. Data mining to discover the knowledge about the customers. And answer questions like, “who are the customers?”, “what do they do?” and “what they like? Therefore, the effective management of information and knowledge is central and critical to the concept of CRM. Based on (Parvatiyar and Sheth 2001, p., 5) (Kracklauer, Mills and Seifert 2004, p.4) CRM consists of four steps:

Customer Identification:

It includes target customer analysis and customer segmentation (Roung-Shiunn Wu and Po-Hsuan Chou., 2011). This phase involves targeting the population who are most likely to become customers or most portable to the company. Moreover, it involves analyzing customers who are being lost to the competition and how they can be won back (Kracklauer et al., 2004).

Customer Attraction:

It depends on direct marketing (Shu-hsien liao, 2011 et al). After identifying the customers, the organization has to make efforts in attracting the customers. For Example, providing gift coupons will be one of the ways of attracting the customer.

Customer Retention:

It is based on the satisfaction of customers (Roung-Shiunn Wu and Po-Hsuan Chou., 2011). This is the central concern for CRM. Customer satisfaction, which refers to the comparison of customers’ expectations with his or her perception of being satisfied, is the essential condition for retaining customers (Kracklauer et al.,2004). As such, elements of customer retention include one-to-one marketing, loyalty programs and complaints management. One-to-one marketing refers to personalized marketing campaigns which are supported by analyzing detecting and predicting changes in customer behaviors. Loyalty programs involve campaigns or supporting activities which aim at maintaining a long-term relationship with customers. Specifically, churn analysis, credit scoring, service quality or satisfaction form part of loyalty programs.

Customer Development:

Elements of customer development include customer lifetime value analysis, up/cross selling

and market basket analysis. (Kim, Y.H and Moon B.R 2006). This involves consistent expansion of transaction intensity, transaction value and individual customer portability.The first requirement for successful implementation of CRM is clarity about the concept of the CRM. Among the various approaches which exist in this area, the following three areas are separable (BehrouzianNejad et al., 2010).

Strategic CRM:

Strategic CRM in which the business puts the customers first. It collects, segregates and applies information about customers and market trends to come up with better value proposition for the customer This approach supports face to face operations and activities with customers. Product, production and selling are the three major business orientations identified by (kotler).

Analytical CRM:

This approach is based upon operational CRM.It provides analytical information about customer segmentation, customer's behavior and customer value.

Collaborative CRM:

This approach focusses on relationship integration with customers using the appropriate communication channel (J Edwards 2007).

Churn Prediction:

Churn is defined as customers leaving the services from one service provider and gets the services from other, due to reasons such as availability of latest technology, customer friendly bank staff, low interest rates, proximity of geographical lacation and dissatisfaction in services or getting better services. Hence need to develop the models that can predict which existing '*loyal*' customer is going to churn out in the near future. Banks would like to know their about-to-churn customers.The customer churn prediction problem is essentially a pattern classification problem. Data mining techniques have emerged to tackle the challenging problem of customer churn. A lot of research has been done in the field of CRM in various industries for retention of customers and develop strategies to build an efficient model so that specific group of customers can be targeted for retention. Various data mining and statistical techniques have been used for churn prediction of which some famous techniques include Decision Trees, Regression Models, Neural Networks, Bayesian Models, SVM, etc.

In this research, we focus only on the problem of feature selection in supervised learning by selected five techniques in feature selection: Chi Squared, Correlation, Gini Index, Support vector machine and Information gain ratio to find a set of best features of Churn prediction dataset for each technique. Then, we compared the sensitivity by two supervised classification algorithms: Support vector machine and Decision tree and for each feature selection techniques. Remaining paper is organized as follows. Section 2 provides the details about related work. In section 3. Datasets used in this study and experimental setup followed during this study are presented. In section 4 Provides the detailed empirical analysis and observations. Section5. Conclude this paper.

2. Review of Literature:

Feature selection has been widely studied in multiple fields and for various models (Dash and Liu 1997; Guyon and Elisseeff 2003; Liu and Yu 2005). For example, in the field of bio informatics where high-dimensional data is prevalent, feature selection has become a prerequisite for constructing mathematical models (saeys et al.2003). There is a clear and growing need for feature selection methods as the complexity of collected data grows. In 2003, at the dawn of the big data age, (Guyon and Elisseeff 2003) wrote are view of feature selection techniques and stated that most of the papers they cited with hundreds to tens of thousands of features. The number of features available in most domains has surely grown since then as a result of the many successful applications of artificial intelligence and machine learning and in correspondence with the capabilities of systems and sensors for collecting contextual and transactional data. (Caruana and Freitag 1994) introduces the idea of relevance with respect to a specific learning algorithm. (Guyon and Elisseeff 2003) lay out several steps for general feature selection. These are in the form of questions, and the answers lead to specific actions to be taken or a particular feature selection method to be used. For example, their fifth question is, “Do you need to assess features individually?” If the answer is yes, a procedure that ranks features should be used.

(Molina et al.2002) test and compare10 feature selection algorithms that require the use of supervised data. They characterize each algorithm based on search optimization (exploring the feature space), generation of successors (selecting the next feature to add or subtract from the proposed feature set), and the evaluation measure. The authors evaluate each algorithm based on the number of relevant or irrelevant features included in the final feature subset, as well as the number of redundant features included in the final subset. Their tests demonstrate that the algorithm that performs the best is highly dependent upon the data, leading to the conclusion that there is no optimal feature selection algorithm for all datasets. Feature selection approaches can be categorized into three types of models. Usually, the filter model evaluates each feature independent from the classifier, rank the features after evaluation and take superior ones (GuyonI, Elisseeff 2003).This evaluation may be don using entropy for example (Y.ozkan). Wrappers takes a subset of the feature set, evaluates the classifier’s performance on this subset, and then another subset is evaluated on the classifier has the maximum performance is selected. (J.Novakovic2010).The embedded techniques perform feature selection during learning process.

Filters

Filters treat feature selection as a pre processing step and select features with no regard to the model, i.e. filters only consider the properties of the collected features and how they distinguish themselves from one another or how they relate to a target class label. These methods are generally fast in terms of computation, but can result in feature subsets that do not yields at is factory predictive accuracy. The simplest filtering technique is to select features based on their correlation with the class or continuous response variable. More complex filters include the FOCUS (Almual-lim and Dietterich 1991) and Relief (Kira and Rendell 1992) algorithms. Extensions of Relief (Kononenko 1994; Robnik-Sikonja and Kononenko 2003) can be applied to multi class problems and unsupervised data.

Some filters rank or assign weights to features. Correlation-based methods (Yu and Liu 2003) rank features based on their association with a class label. (Forman.2003) compares several metrics for ranking features. The evaluation is specific to text classification, so results may not generalize to all classification problems.The author concludes that the binomial separation metric outperforms other filters on the chosen datasets.(Bins and Draper.2001) propose using both filters and a traditional wrapper in a three-stage feature selection method. The first stage removes the irrelevant features using a filter. The second stage removes the redundant features also using a filter. The first two stages reduce the number of features to appoint where a traditional wrapper can be applied without significant computational cost. The authors suggest using either forward or backward search depending on how many features remain after the first two stages.

Wrappers

In a wrapper approach, data are typically divided into three groups: training, evaluation, and testing. A model using a subset of the candidate features is trained on the training set, and then evaluated using the evaluation set. The feature subset is then augmented in some fashion and the process repeated. The feature subset that optimizes the evaluation function is chosen as the final feature subset and tested on the withheld testing set. When data are scarce, the evaluation set can be eliminated by evaluating the feature subset on the training data. However, this can cause a poor generalization error and increase the likelihood of over fitting to the training data. Crossvalidation can be used in the case of small datasets.

Sequential forward search and sequential backward search are two types of exploration algorithms (John et al.1994; Kohavi and John 1997). These are considered greedy algorithms, as opposed to an exhaustive search of the feature subsets. Sequential search methods are often referred to as hill-climbing strategies, because they look for improvement in the evaluation function. However, a non-exhaustive search cannot guarantee the optimal feature subset (Cover and Campenhout 1977).(Aha and Bankert 1995)compare forward and backward search algorithms to filters. They found that these wrappers, and some of the tested variants, generally outperform filters on the tested data sets and that backward search does not significantly outperform forward search as previously claimed by (Doak.1992). Another exploration method, the stepwise search, combines forward and backward sequential search so that a teach step in the algorithm a feature can either be added or removed. When compared to unidirectional methods and filters, stepwise search algorithms outperform unidirectional search and some filtering techniques (Caruana and Freitag 1994). (Kohavi and John.1998) propose a best-first-search feature-exploration technique with compound operators. Compound operators add or remove groups of features, as opposedto adding or removing a single feature in sequential forward search and sequential backward search.

The branch-and-bound algorithm (Narendra and Fukunaga 1977; Somol et.al 2004) can yield an optimal feature subset, but requires a monotonic evaluation function that is generally not practical. Floating-search methods (Pudil et al. 1994a, b) add and remove different numbers of features to avoid the nested-feature problem. (Ng 1998) presents theoretical bounds on the performance of wrappers and a new search procedure called ORDERED-FS that only searches over all subsets of the same size to reduce computational cost. When using wrappers, the choice of an evaluation

function greatly affects the outcome of the method. (Dash et al.2000) compare a consistency measure to distance measures, information measures, and dependence measures. The authors argue for a consistency-measure evaluation function, because it is monotonic and lacks search bias. Obviously, the type of evaluation function is based on the available data. Functions that require the class label, such as consistency or accuracy, cannot be implemented in unsupervised learning. While wrappers can be used to select features for both GMMs and HMMs, there do two primary drawbacks to these methods both have to do with the unsupervised data. First, the wrapper will need to use an unsupervised evaluation metric.

In unsupervised feature selection these metrics often trade off the likelihood the model fits the data with the number of parameters or features included in the model. These metrics can be inaccurate and uninformative for feature selection. Second, the computation required for the search procedure is amplified due to the fact that unsupervised learning techniques for GMMs and HMMs are often iterative and computationally expensive.

Embedded Techniques

Embedded techniques simultaneously select features and construct models. Therefore, these techniques have the wrapper's advantage of selecting feature subsets with respect to a specific learning algorithm, and the filter's advantage of being more computationally efficient than wrappers. Classification and regression trees (CART)(Murphy 2012)are one example of an embedded feature selection method. CART recursively divides the feature space to create a classification model. Irrelevant features will not be selected for inclusion in the tree. (Daelemans et al.2003) show that jointly selecting features and optimizing model parameters in a natural-language-processing context outperforms optimizing the two processes separately.

3. Techniques Employed: Feature Selection Techniques

Correlation Feature Selection

Correlation is a statistical method to calculate the relevance of the feature with respect to the class feature. The correlation coefficient takes a value between -1 and+1 that represent the strength of association between two features. The positive correlation means as one variable gets larger the other gets larger. Whereas, the negative correlation means as one gets larger the other gets smaller. If the correlation is zero, it means there is no association between the features. Suppose we have two features observations $\{x_1 \ldots x_n\}$ and $\{y_1, \ldots, y_n\}$ then that the correlation coefficient is calculated as in Equation-1.

$$r = r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_{xs_y}} \qquad (1)$$

Chi-Squared Feature selection

Usually, the chi-square feature selection is used to select significant features by evaluating the dependence between an input variable and classes. Let a number of classes be $\{C_1, \ldots, C_j, \ldots C_n\}$,and

the categorical values of feature F_p be $\{V_{p1}, \ldots, V_{pi}, \ldots, V_{pm}\}$.The chi-squared statistic can be defined by (2):

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \qquad (2)$$

Where $n_{ij}$is number of instances for which the categorical value of feature F_p is V_{pi} and belongs to j^{th}class and $E_{ij}$is the expected frequency of n_{ij} and is calculated by (3)

$$E_{ij} = \frac{N_{+j} X N_{i+}}{N} \qquad (3)$$

Where N is the total number of instances. N_{+j} and N_{i+} are marginal total of class $C = C_j$ and value $V = V_{pi}$ respectively. They are defined by (4).

$$N_{+j} = \sum_{i=1}^{m} n_{ij}, N_{i+} = \sum_{j=1}^{n} n_{ij} \qquad (4)$$

The cumulative probability (cp), which indicates the probability that values X^2falls within a specified range, can be calculated by (5).

$$Q_{x^2,d} = [2^{\frac{d}{2}} \tau\left(\frac{d}{2}\right)]^{-1} \int_{x^2}^{\infty} (t) \frac{d}{2} \text{-1 } e^{\frac{-1}{2}} dt \qquad (5)$$

Where $d=(m-1)(n-1)$ is the degree of freedom, τ is the generalization of factorial function (Gamma function) to real and complex arguments, it can be calculated as follows:

$$\tau_x = \int_0^{\propto} t^{x-1} e^{-1} dt \qquad (6)$$

If the absolute value of *cp* is less than a critical significant level coefficient $\propto$ (usually take 0.05 or 0.01), then there is 95% or more chance to say that the feature is important. Chi- Square only judges whether an input feature is helpful to classify the classes. However, it does not give us information about whether a value of a feature $V = V_{pi}$ indicates the potential correct class

Gini Index

Gini index is a feature selection method which measures the purity of the features with respect to the class (Shang, w., et al., 2007). The purity refers to the discrimination level of a feature to distinguish between the possible classes (Haralampieva V. And G. Brown 2016). This feature selection method measures the purity when using a chosen feature. For a feature, the Gini index is calculated by the Equation:

$$GI(t_i) = \sum_{j=1}^{m} p\ (t_i \mid c_j)^2\ P(c_j \mid t_i)^2 \qquad (7)$$

Where m is the number of classes, p ($t_i \mid c_j$) is the term t_i, probability given class c_j, p ($c_j|t_i$) is the class c_j probability given term t_i. The higher the weight of an attribute, the more relevant it is considered.

Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are a supervised learning methods used for classification and regression. This approach focusses on classifying the features depends on its weights. The idea of SVM is to mapping of the data to a high dimensional space. The operation of the SVM algorithm is based on finding the hyperplane which the distance between the training data and hyperplane will be maximized.

In nonlinear separable, using a kernel function (e.g. polynomial, fisher, and RBF kernel) to transform the data representation into higher dimensional feature space, in which it is probable that there is a linear separator (Bouazza et al., 2015). The use of selection of SVM parameters, kernel and kernel parameter optimization using the BAT algorithm in underwater target classification indicated that the result of classification accuracy ismorethan4%increased (Sherin et al., 2015).

Information Gain Ratio

Information Gain is one of filter method of feature selection which used statistical and entropy-based measurement. IG calculates the value of the features information by determining the entropy with respect to the class to which they belong. Entropy value ranges from 0 to 1, value 0 means that all instances of the variables have the same value and value 1 equals the number of instances of each value. The entropy of N is calculatedas:

$$Entropy(A) = \sum_{i=0}^{k} P_i \, log_2 \; P_i \qquad (8)$$

In equation (1), P is probability of class for which a particular value and this equation calculates the information of all classes.

$$Entropy\ (D_{i)} = \sum_{i=1}^{D_j} \frac{D_{ji}}{N} X\ Entropy(D_{ji}) \quad (9)$$

$$IG\ (D)_j = Entropy\ (N) - Entropy\ (D) \qquad (10)$$

In equation (2) means that the i^{th} feature contains kinds of different values. And equation (3), the IG of each feature is calculated by finding the difference of equation (1) and equation (2).

4. CLASSIFICATION TECHNIQUES

Decision tree:

Decision trees (DT) is referred to as a non-parametric supervised learning method used for Classification and prediction (Nefeslioglu et.al, 2010). In general, the DT is classified into two Types: classification trees and regression trees. For predicting a discrete variable, the classification trees are used; while for a continuous Variable, regression trees are used. The basic idea here is to develop a model that forecasts the value of a dependent factor by learning several decision rules inferred from the whole data (Yeon et.al, 2010). This set of rules is called a Decision Tree.

For this total population or sample is split into two or more homogeneous sets based on the most significant splitter in input variable. The main advantage of the DT is that the variable transactions not required because the tree structure will remain the same with or without the transactions thus saves the overall processing time. Further it is capable of complex modeling relationship among variables and also quick to build and easy to interpret for the decision makers.

Support Vector Machines (SVM):

It is a group of supervised learning methods that can be employed for classification or regression (Ivanciuc, 2007; Zhenzhou, 2012). In a two-class learning task, the SVM goal is to discover the best classification function to differentiate between members of the two classes in the training data. For that purpose, SVM construct a hyper plane or a set of hyper planes in a high or infinite dimensional space for separating dataset and SVM find the best function by maximizing the margin between the two classes.

LibSVM:

It is a library for Support Vector Machines (SVMs). Chih-Chung Chang and Chih Jen Lin have been actively developing this package since the year 2000. The goal is to help users to easily apply SVM to their applications. LIBSVM has gained wide popularity in machine learning and many other areas. A typical use of LIBSVM involves two steps: training dataset to obtain a model and second, using the model to predict information of a testing dataset (Chih-ChungandChih-Jen,2011).

5.Dataset Description: Empirical Analysis

Churn Prediction Dataset

The churn prediction dataset is obtained from Business Intelligence Cup, 2004, which is obtained from a Latin American bank that suffered from an increasing number of churns with respect to their credit card customer. Table 1 presents the dataset information including attribute description and the values of the attributes. This dataset consists of 14814 instances which include 13812 instances representing loyal customers i.e., 93% and 1002 instances representing churners i.e., 7%. Hence, the dataset is highly unbalanced in terms of the proportion of churners versus non-churners.

Table 1: Feature Set description of Churn Prediction dataset

#	Attribute	Description
1	Target	Target Variable (*0 – Non-Churner; 1- Churner*)
2	CRED_T	Credit in Month T
3	CRED_T-1	Credit in Month T-1
4	CRED_T-2	Credit in Month T-2
5	NCC_T	Number of credit cards in month T
6	NCC_T-1	Number of credit cards in month T-1

7	NCC_T-2	Number of credit cards in month T-2
8	INCOME	Customers Income
9	N_EDUC	Customers Educational Level (1-University Student; 2-Medium degree; 3-Technical degree; 4-University degree)
10	AGE	Customers age
11	SX	Customer Sex (1 - Male; 0 - Female)
12	E_CIV	Civilian Status (1-Single; 2-Married; 3-Widow; 4-Divorced)
13	T_WEB_T	Number of web transaction in months T
14	T_WEB_T-1	Number of web transaction in months T-1
15	T_WEB_T-2	Number of web transaction in months T-2
16	MAR_T	Customers margin for the company in months T
17	MAR_T-1	Customers margin for the company in months T-1
18	MAR_T-2	Customers margin for the company in months T-2
19	MAR_T-3	Customers margin for the company in months T-3
20	MAR_T-4	Customers margin for the company in months T-4
21	MAR_T-5	Customers margin for the company in months T-5
22	MAR_T-6	Customers margin for the company in months T-6

Experimental Setup

In this paper we have carried out the experiments using 80:20 and 70:30 evaluation methods, where 80 and 70 percent of the data is used as training and tested against 20 and 30 percent data, respectively. Prior to modelling using the training data, 20% of the data is taken out and stored separately for later validation purposes and remaining 80% of the whole data is then used for modelling purposes.

The validation dataset is later used to analyse the efficiency of the classifier when original data is provided as the validation data is never used earlier neither for training nor test. Whereas, the parameters of the classifiers are tweaked based on the results obtained i.e. test set play a role in tweaking the parameter of the classifiers. Once the parameters are set them that classification model is tested against validation set. Table 2 presents the details of the datasets prepared for empirical analysis in this research study. Dataset prepared includes (i) original dataset with full features, (ii) 20% Validation set, (iii) 80% Training set, (iv) 20% Test set, (v) 70% Training set and (vi) 30% Test set. Dataset (iii) to (vi) are prepared from remaining 80% data from original data after taking out validation set. The information presented her includes number of instances,

number of churner samples, number of non-churner samples and distribution ratio. Figure 2 depicts the dataset preparation for experimental analysis and modelling in this research study.

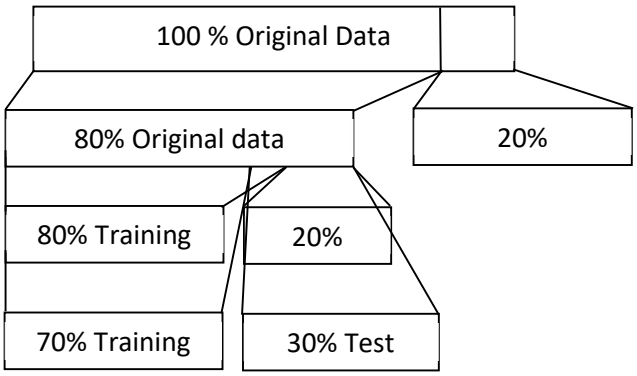

Fig. 2: Experimental Setup used in this research study

Table 2: Churn Prediction dataset Preparation

Data Set	Number of Instances	Churn	Non-Churn	Distribution Ratio
Full dataset	14814	1002	13812	93:7
Validation Set	2962	200	2762	93:7
80 Training	9482	642	8840	93:7
20 Test Set	2370	160	2210	93:7
70 Training	8297	562	7736	93:7
30 Test Set	3555	240	3315	93:7

6. RESULTS AND DISCUSSION

Keeping the churn prediction problem in mind, *sensitivity* is accorded high priority in this research study. It is observed that the loss to the organization is huge when an *expected churner* is predicted *as loyal* when compared to the prediction of *loyal customer as churner*. *Sensitivity* (also called Recall or True Positive Rate): Sensitivity is the proportion of actual positives which are correctly identified as positives by the classifier.

$$Sensitivity = \frac{TP}{(TP+FN)}$$

Specificity (also called True Negative Rate): Specificity relates to the classifier's ability to identify negative results. Consider the example of medical test used to identify a certain disease. The

specificity of the test is the proportion of patients that do not to have the disease and will successfully test negative for it. In other words:

$$Specificity = \frac{TN}{(TN+\quad)}$$

Accuracy: This is the simplest scoring measure. It calculates the proportion of correctly classified instances.

$$Accuracy = \frac{TP+T}{(TP+TN+FP+F\quad)}$$

Table 3 through Table 6 present the results obtained using 80:20 data (refer Table 2). Table 3 and Table 4 present the results obtained using SVM against Test data and Validation Set, respectively. The results obtained implies that feature selection prior to classifier modelling helps the classifier learn better about the churner class and predict better when compared to the results yielded using full feature data. Hence, it is reported that despite getting simpler modelling with reduced features the classifier's learning is improved. Further, it implies that features selected using Correlation has yielded best sensitivity of 77.5% against Test set and 80.5% against Validation set. Comparative analysis of the sensitivity yielded using SVM with full features and reduced features against Test set and Validation set is depicted in Figure 3.

Table 3: Results obtained using SVM against Test Set (80:20)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	50	90.15	87.44	7007.5
FS Correlation	**77.5**	79.37	79.24	7843.5
FS Chi-Squared	61.25	33.17	30.06	4721
FS SVM	70.62	76.47	76.08	7354.5
FS Gini Index	**77.5**	72.4	72.74	7495
FS Information Gain	76.25	79.05	78.86	7765

Table 4: Results obtained using SVM against Validation Set (80:20)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	52	82.37	80.32	6718.5
FS Correlation	**80.5**	78.17	78.33	7933.5
FS Chi-Squared	61.5	32.84	34.77	4717
FS SVM	74.5	74.87	74.85	7468.5
FS Gini Index	78	73.14	73.46	7557
FS Information Gain	77	79.94	79.74	7847

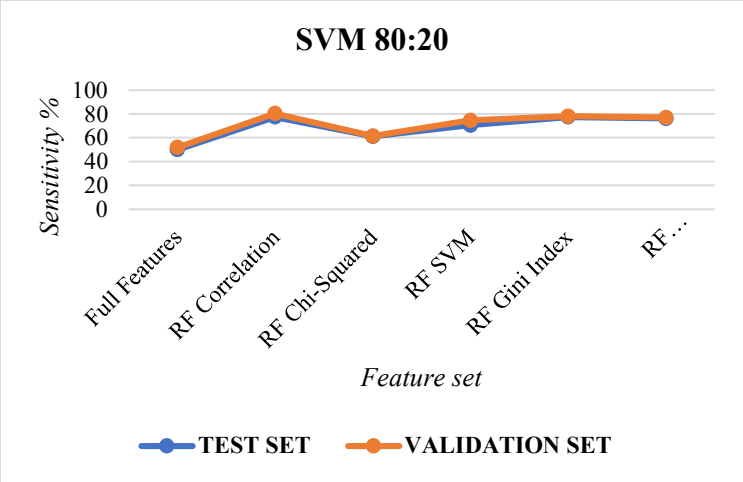

Figure 3: Results of SVM against Test set and Validation Set

Table 5 and Table 6 present the results yielded using Decision Tree against Test set and Validation set, respectively. It is observed that better sensitivity is yielded when reduced feature data is employed compared to sensitivity yielded using full feature data. Further, classifier DT yielded best sensitivity of 63.12% against Test set and 64% against Validation set using FS Information Gain and outperformed every other variant employed in this category. Further, once again it is implied that selection of most important features' subset is helping the classifier learn better about Churner class. Comparative analysis of the sensitivity yielded using DT with full features and reduced features against Test set and Validation set is depicted in Figure 4.

Table 5: Results obtained using DT against Test Set (80:20)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	43.75	99.41	95.65	7158
FS Correlation	46.25	98.37	94.85	7231
FS Chi-Squared	43.12	95.66	92.11	6939
FS SVM	59.38	97.51	94.94	7844.5
FS Gini Index	59.38	97.65	95.06	7851.5
FS Information Gain	**63.12**	96.7	94.43	7991

Table 6: Results obtained using DT against Validation Set (80:20)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	45.5	99.31	95.68	7240.5
FS Correlation	47	97.76	94.33	7238
FS Chi-Squared	44.5	95.37	91.93	6993.5
FS SVM	62.5	97.43	95.07	7996.5
FS Gini Index	62.5	97.43	97.07	7996.5

FS Information Gain	**64**	96.52	94.33	8026

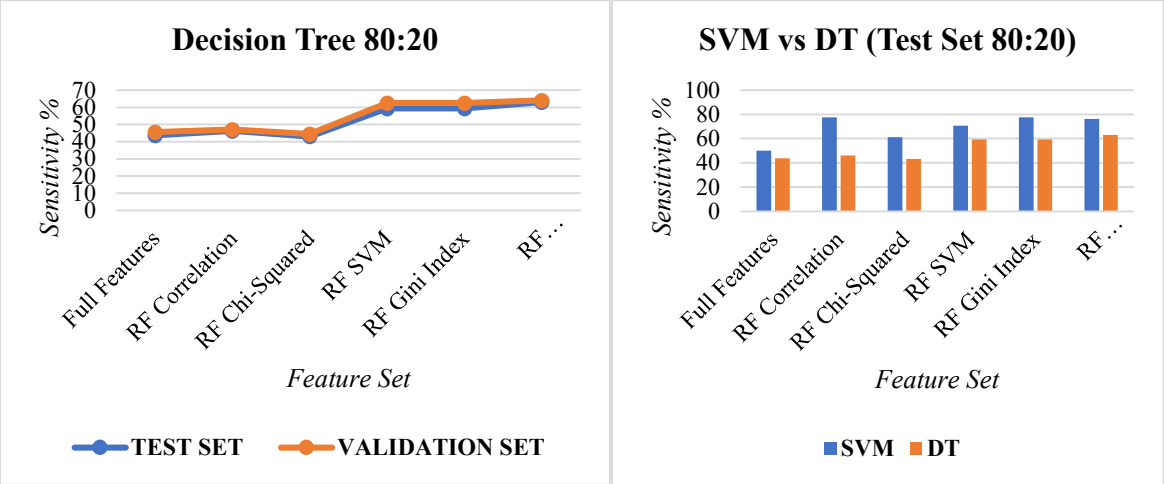

Fig.4: Results of Decision Tree against Test set and Validation SetFigure 5: SVM and DT performance comparison against Test Set

Figure 5 and Figure 6 depict the performances of SVM and DT classifiers against Test and Validation set. It is observed that when the classifiers are compared SVM outperform DT in all the experiments carried out and analysed in this research study.

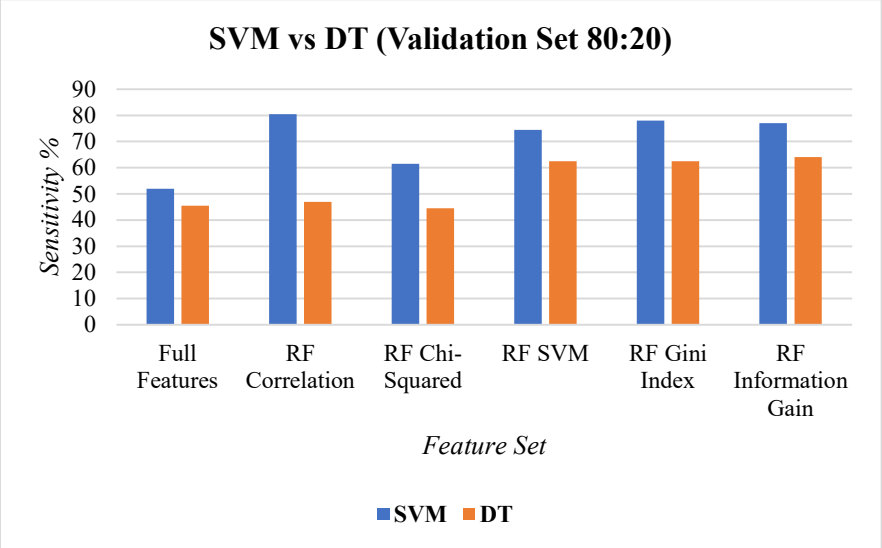

Figure 6: SVM and DT performance comparison against Validation Set

Table 7 through Table 10 present the results obtained using 70:30 data (refer Table 2). To further evaluate the classifiers' results extensively, experiments are carried out using 70:30 data. Table 7

and Table 8 present the results yielded using SVM against Test and Validation set, respectively. Similar to SVM's earlier results, it has outperformed (with 80.83% sensitivity against Test set and 82.5% sensitivity against Validation set) every other SVM model when compared to the SVM model created using Correlation based selected features. Further, it can be implied from the results that reducing features helps the classifier learn better about the Churn customers as the result with reduced feature is better compared to the results with full features. Comparative analysis of the sensitivity yielded using SVM with full features and reduced features against Test set and Validation set is depicted in Figure 7.

Table 7: Results obtained using SVM against Test Set (70:30)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	42.08	88.45	85.32	6526.5
FS Correlation	**80.83**	76.68	76.96	7875.5
FS Chi-Squared	53.75	37.13	38.26	4544
FS SVM	62.92	34.21	36.12	4856.5
FS Gini Index	73.33	87.06	86.13	8019.5
FS Information Gain	61.67	31.89	33.9	4678

Table 8: Results obtained using SVM against Validation Set (70:30)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	51.5	89.25	86.7	7037.5
FS Correlation	**82.5**	76.1	76.54	7930
FS Chi-Squared	59.5	38.6	40.01	4905
FS SVM	68	35.16	37.37	5158
FS Gini Index	75.5	78.86	78.63	7718
FS Information Gain	60	33.17	34.98	4658.5

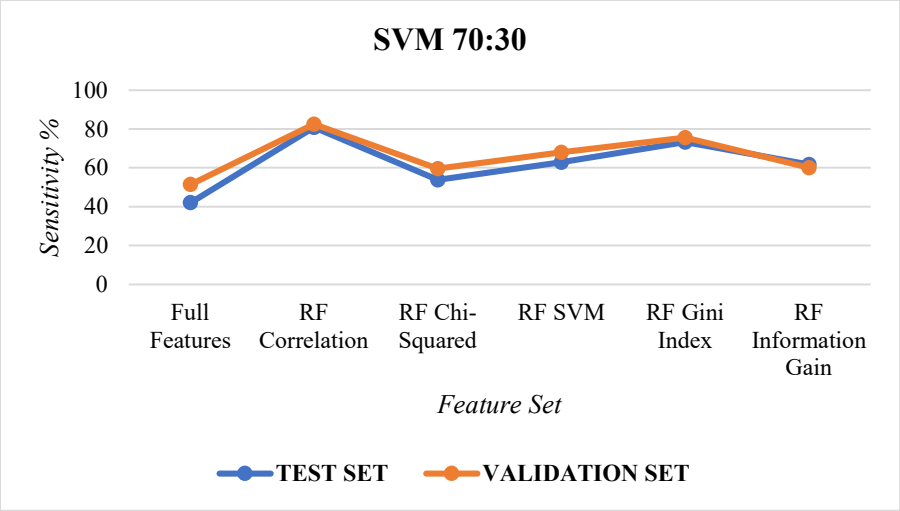

Figure 7: Results of SVM against Test set and Validation Set

Table 9 and Table 10 present the results obtained using DT against Test and Validation set, respectively. It is observed that DT achieved better sensitivity using reduced features than using full feature data. Further, it is observed that DT using FS Information Gain yielded best sensitivity of 62.92% against Test set and 63% sensitivity against Validation set. Comparative analysis of the sensitivity yielded using DT with full features and reduced features against Test set and Validation set is depicted in Figure 8.

Table 9: Results obtained using DT against Test Set (70:30)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	41.67	99.46	95.56	7056.5
FS Correlation	43.33	95.47	97.47	6940
FS Chi-Squared	45	95.99	92.55	7049.5
FS SVM	54.17	96.59	93.73	7538
FS Gini Index	57.5	98.43	95.67	7796.5
FS Information Gain	**62.92**	95.72	93.5	7932

Table 10: Results obtained using DT against Validation Set (70:30)

Features Set	Sensitivity	Specificity	Accuracy	AUC
Full Features	43.5	99.35	95.58	7142.5
FS Correlation	45	97.94	94.36	7147
FS Chi-Squared	46.50	95.58	92.27	7104
FS SVM	55.5	96.78	93.99	7614
FS Gini Index	57.5	97.86	95.14	7768
FS Information Gain	**63**	97.25	94.94	8012.5

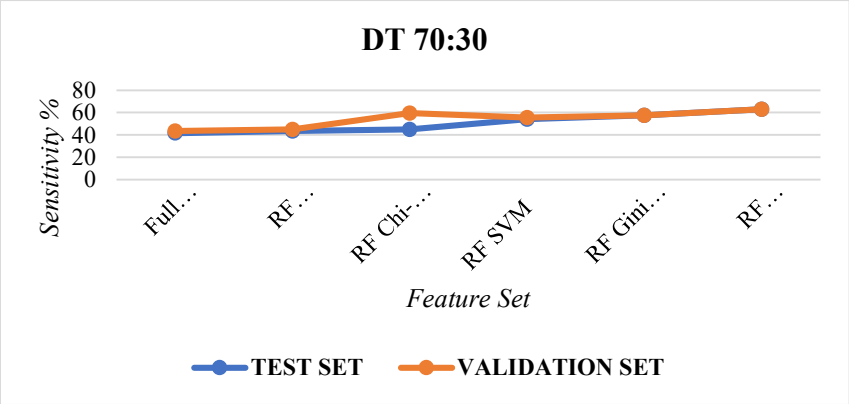

Figure 8: Results of Decision Tree against Test set and Validation Set

Figure 9 and Figure 10 depict the performances of SVM and DT classifiers against Test and Validation set using 70:30 data. It is observed that when the classifiers are compared SVM outperform DT in all the experiments carried out and analysed in this research study. Further, results show that there is negligible improvement in SVM's performance compared to DT's performance when full feature data is employed. Otherwise, when reduced feature data is used, SVM outperformed DT.

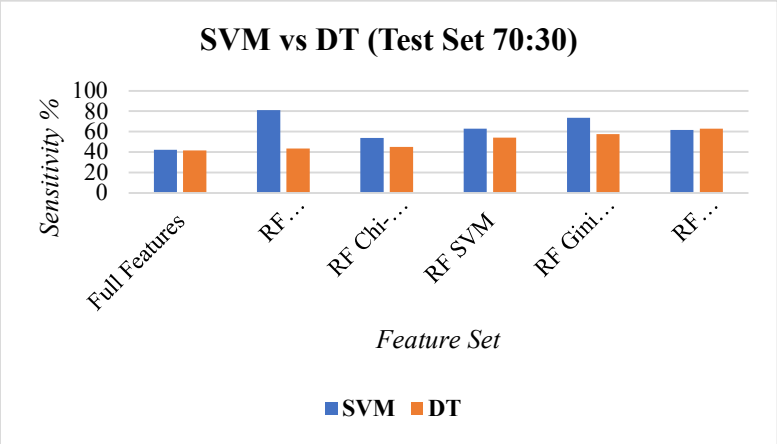

Figure 9: SVM and DT performance comparison against Test Set

7. CONCLUSION

With increased number of customers for banking industry it become almost impossible for the bank to evaluate, analyse and understand their customers. Analytical CRM provides an alternative to understand the behaviour of their customers and accordingly take necessary actions on time. Further, the bank cannot take chances of losing their most valuable customers to their competitors. Soft computing techniques have been in place and are reported to be employed for analytical CRM purposes. In this paper we propose the implementation of Feature Selection prior to classification.

The proposed approach can be divided into three parts. First part will make use of the full feature data and based on weights feature subset (viz., Correlation, Chi-squared, SVM, Information Gain and Gini Index) is considered for classification purposes. During second part modelling is done (viz., SVM and DT) and the best model is then evaluated against Validation set in final part of the proposed approach. The dataset considered for analysis in this research pertain to Churn prediction in bank credit card customers. The data at hand is highly imbalance in nature with only 7% samples representing churner class and it is slightly medium in size with around 15k samples in total. Keeping the problem of study in mind, sensitivity is accorded highest priority for classification modelling. It is observed from the results that classifiers SVM and DT perform better with reduced feature data compare to that of full feature data. Further, it is observed that SVM performed better than DT as classifier. Hence, it is concluded that feature selection prior to classification modelling for churn prediction problem is advisable. It also reduced the complexity of the classifier with reduced features with improved performance of the classifiers. In future other classifiers like Neural Network, Logistic Regression and Naïve Bayes could also be analysed for more extensive research analysis.

8. REFERENCES

Aha DW, Bankert RL (1995) "A comparative evaluation of sequential feature selection algorithms". In: Proceedings of the fifth international workshop on artificial intelligence and statistics.

Almuallim H, Dietterich TG (1991),"Learning with many irrelevant features." In: AAAI, vol 91. Citeseer, pp 547–552.

Bins J, Draper BA (2001),"Feature selection from huge feature sets". In: Eighth IEEE international conference on computer vision, 2001. ICCV 2001. Proceedings, vol 2. IEEE, pp159–165

B.M.Sherin and M.H.Supriya,"Selection and parameter optimization of SVM kernel function for under water target classification", under water Technology (UT), pp. 1-5,2015CaruanaR,FreitagD(1994)Greedyattributeselection.In:ICML.Citeseer,pp28–36.

Cover TM, Van Campenhout JM (1977),"On the possible orderings in the measurement selection problem". IEEE Trans Syst Man Cybern 7(9):657–661

Chih-Chung, C. and L. Chih-Jen, 2011. LIBSVM: "A library for support vector machines". *ACM Transactions on Intelligent System and Technology. 2: 1-27.*

Dash M, Liu H (1997) "Feature selection for classification". Intell Data Anal 1(3):131–156

E.W.T Ngai, LiXiu, D.C.K. chau (2009) "Application of data mining techniques in customer relationship management: A literature review and classification", *Expert Systems with Applications* 36 (2009) 2592-2602.

Forman G (2003),"An extensive empirical study of feature selection metrics for text classification". J Mach Learn Res 3:1289–1305.

GuyonI, Elisseeff (2003),"An introduction to variable and feature selection". JMach Learn Res3:1157–1182.

http://www.insidecrm.com/features/collabarative-crm-112907

Haralampie.VandG.Brown (2016),"Evaluation of Mutual Information versus Gini index for stable feature Selection", 2016, University of Manchester.

Ivanciuc, O., 2007. "Applications of support vector machines in chemistry". Reviews Computational Chemistry, 23: 291-400.

J.Edwards,"Get It Together with Collaborative CRM,inside CRM",2007.Tippit

J.Novakovic(2010),"The Impact of Feature Selection on the Accuracy of Naïve Bayes Classfier",18th telecommunications forum TELFOR,2010

John GH, Kohavi R, Pfleger K (1994) "Irrelevant features and the subset selection problem". In: Machine learning: proceedings of the eleventh international conference, pp 121–129

Kira K, Rendell LA (1992),"The feature selection problem: traditional methods and a new algorithm". AAAI 2:129–134.

Kononenko I (1994),"Estimating attributes: analysis and extensions of Relief". *In: Machine learning: ECML-94.Springer*, pp 171–182.

KracklauerA.H,Mills D.Q., and Seifert D(2004),"Collaborative Customer Relationship Management(CCRM)", *Springer*, Berlin, Heidelberg, pp 25-45.

Kohavi R, John GH (1997) Wrappers for feature subset selection. ArtifIntell 97(1):273–324

KP.Murphy (2012) "Machine Learning: A probabilistic perspective"2003 MIT Press, Cambridge, Massachusetts, London, England.

Kim, Y.H and Moon, B.R., (2006),"Multi campaign Assignment Problem" IEEE Transactions on Knowledge and Data Engineering",2006,vol 18, pp 405-414.

MolinaLC,BelancheL,NebotÀ(2002),"Feature selection algorithms survey and experimental evaluation".In:2002 IEEE international conference on data mining, 2002.ICDM2003.Proceedings.IEEE,pp306–313.

MolinaLC,BelancheL,NebotÀ(2002),"Feature selection algorithms survey and experimental evaluation".In:2002 IEEE international conference on data mining, 2002.ICDM2003.Proceedings.IEEE,pp306–313.

Nefeslioglu,H.A.,Sezer,E.,Gokceoglu,C.,Bozkir,A.S.,Duman,T.Y.,2010. "Assessment of land slide susceptibility by decision trees in the metropolitan area of Istanbul, Turkey.*Mathematical Problems in.Engineering 2010,1—15.https://doi.org/10.1155/2010/901095.*

Parvatiyar,A., and Sheth,J.N.(2001) "Customer Relationship Management: Emerging practice, Process, and Discipline", *Journal of Economic and social Research*,2001,3,1-34.

P.Ktler, "marketing management" (2000): The Millennium Edition.Englewood cliffs,NT: Prentice-Hall International,2000..

Roung-Shiunn Wu., and Po-Hsuan Chou(2011) "Customer segmentation of multiple category data in e-commerce using a soft clustering approach", *Electronic Commerce Research and Applications* 2011 vol-10.,issue 3,p331 – 341.

Shu-hsienLiao, Yin-juchen, Hsin-hua Hsieh, (2011)"Mining customer knowledge for direct selling and marketing", *Expert Systems with Applications*,2011,38,6059-6069.

SaeysY,InzaI,LarrañagaP(2007),"A review of feature selection techniques in bio informatics. Bioinformatics 23(19):2507–2517.

YuL,LiuH (2003),"Feature selection for high-dimensional data: a fast correlation-based filter solution". ICML 3:856–863.

S. H. Bouazza, N. Hamdi, A. Zeroual, and K. Auhmani, "Gene- expression-based cancer classification through feature selection with KNN and SVM classifiers", *Intelligent Systems and Computer Vision (ISCV)*, pp. 1-6,2015.

Ron kohavi, George H.John (1998),"The Wrapper Approach" *feature extraction, construction and selection* 1998, pp 33-50.

P.Somol, P.pudil and J.Kittler (2004),"Fast branch and bound algorithms for optimal feature selection"2004, vol 26, pp900-912, *IEEE Transactions on Pattern Analysis and machine Intelligence.*

P.Pudil, J.novovcova and J.kottler (1994),"Floating search methods in feature selection methods", 1994, vol 15 pp1119-1125, *Pattern Recognition Letters*.

Shang et.,al (2007),"A novel feature selection algorithm for text categorization",2003,vol 33,pp 1-5, *Expert systems with Applications*.

S. H. Bouazza, N. Hamdi, A. Zeroual, and K. Auhmani, "Gene- expression-based cancer classification through feature selection with KNN and SVM classifiers", *Intelligent Systems and Computer Vision (ISCV*), pp. 1-6,2015.

W Daelemans, V.Hoste, F.Demeulder (2003),"Combined Optimization of Feature Selection and Algorithm parameters in Machine learning of Language" European Conference on Machine Learnig, 2003, pp 84-95.

Yeon, Y.K., Han, J.G., Ryu, K.H., 2010. "Landslide susceptibility mapping in Injae, Korea, using a decision tree". *Journal of Engineering Geology.116, 274—283.* *https://doi.org/10.1016/j.enggeo.2010.09.009.*

Y.ozkan "very Madenciligi Yontemleri",Papatya publication, istanbul,2008

Zhenzhou, C., 2012. "Local support vector machines with clustering for multi modal data". Advances Inform. Sci. Service Sciences, 4: 266-275.DOI: 10.4156/AISS.vol4.issue17.30