Projection pursuit Random Forest using discriminant feature analysis model for churners prediction in telecom industry

Asia Mahdi Naser Alzubaidi[1], Eman Salih Al-Shamery[2]
[1]Department of Computer Science, University of Karbala, Iraq
[2]Department of Software Engineering, University of Babylon, Iraq

Article Info

Article history:

Received Sep 6, 2019
Revised Oct 11, 2019
Accepted Oct 20, 2019

Keywords:

Churn prediction
Decision classification trees
discriminant random forest
Linear discriminant analysis
oblique tree
Project pursuit index
Support vector machines

ABSTRACT

A major and demand issue in the telecommunications industry is the prediction of churn customers. Churn describes the customer who attrites from the current provider to competitors searching for better service offers. Companies from the Telco sector frequently have customer relationship management offices it is the main objective in how to win back defecting clients because preserve long-term customers can be much more beneficial than gain newly recruited customers. Researchers and practitioners are paying great attention to developing a robust customer churn prediction model, especially in the telecommunication business by proposed numerous machine learning approaches. Many approaches of Classification are established, but the most effective in recent times is a tree-based method. The main contribution of this research is to predict churners/non-churners in the Telecom sector based on project pursuit Random Forest (PPForest) that uses discriminant feature analysis as a novelty extension of the conventional Random Forest for learning oblique Project Pursuit tree (PPtree). The proposed methodology leverages the advantage of two discriminant analysis methods to calculate the project index used in the construction of PPtree. The first method used Support Vector Machines (SVM) while, the second method used Linear Discriminant Analysis (LDA) to achieve linear splitting of variables during oblique PPtree construction to produce individual classifiers that are robust and more diverse than classical Random Forest. It is found that the proposed methods enjoy the best performance measurements e.g. Accuracy, hit rate, ROC curve, Lift, H-measure, AUC. Moreover, PPForest based on LDA delivers effective evaluators in the prediction model.

Corresponding Author:

Asia Mahdi Naser Alzubaidi,
Department of Computer Science,
University of Karbala, Karbala, Iraq.
Email: asia.m@uokerbala.edu.iq

1. INTRODUCTION

The Telecom industry is a highly technological sector that has developed tremendously over the past two decades as a result of the emergence and commercial success of both mobile telecommunication and the internet [1]. Customer churn or customer attrition is a great challenge for many telecom companies. It happens when a customer ends his subscription and switch to another competitor. There are many factors affect the customer's decision to change to another competitor. In general, such factors related to the high cost, bad customer service-related work, fraud and privacy concerns [2, 3]. Customer churn causes serious profit loss when exceeds certain limits. On the other hand, companies realize that attracting new customers is much more expensive than preserving existing ones. The initial and foremost step in curtailing outbound churn and establishing loyalty of the prevailing customers is to understand the reasons for churning. In this situation, the churn prophecy is a useful and helpful tool to forecast customer at churn risk. The only remedy to overcome churn business hazards and to retain in the company [4].

Customer Churn Prediction (CCP) has been raised as a key issue in many fields such as Telecom providers, credit cards, internet service providers, electronic commerce, retail marketing, newspaper publishing companies, banking and financial services [5]. CCP in Telecommunication companies has become an increasingly popular research issue in recent years and therefore, Telecom providers using widely strategies to identify the potential churn customers based on their historical information, prior behaviors and offering some services to persuade them to stay. On other hand, Long-terms customers are more profitable for the service providers, since they are more dependency to buy additional products and spread the customer's satisfaction in their circle, thus procedure will indirectly attract more and more customers [6].

Stockholders forced to search for alternative approaches for using machine learning techniques and statistical tools to recognize the cause of churn in advance and to yield instantaneous efforts in response. This is possible if the historical data of the potential customers analyzed systematically [7]. Fortunately, telecom sectors produce and preserve a large volume of data, they include non-relational data i.e. billing information, demographic, customer care, customer behavior, and relational data i.e. Call Detail Records data (CDR) and network data. Moreover, not all the features of the telecom database used by all the prediction methods only the relevant features that really contribute to the CCP used in data mining (DM) techniques [8].

The statistical learning model discovers methods of approximating functional dependency from a given assortment of data. It covers significant issues in classical statistics such as discriminant analysis, regression methods, and the density estimation problem [9]. Statistical learning is a kind of statistical inference, also called inductive statistics. Recently, statistical learning methods such as Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) have an important role in describing the differences between a reference collection of patterns and the population under exploration [10].

The main contribution of this research is to develop a new ensemble learning method for churn prediction method based on Random Forest constructed but with oblique trees principal using an optimal and linear association of randomly chosen predictors, which increases the predictive performance when the cutoff hyperplane between classes is in a linear collection of variables. The suggestion method called a Projection Pursuit Random Forest (PPForest). Moreover, using a visualization tool of Constructed PPForest and compare those with the Random Forest graph in order to understand how the PPForest model summarizes datasets.

The main difference with the known Random Forest approach is that the oblique partitions of variables not selected randomly. Nevertheless, the linear association in each tree construction is calculated by improving a projection pursuit index depend on a linear discriminant analysis (LDA) or Support Vector Machine (SVM) to discover the projection data of the variables that best splits the classes taken into account the correlation between the target variable and other dataset variables. PPForest outperforms a traditional Random Forest when separations between groups occur in Linear combinations of variables.

The PPForest uses the Project Pursuit tree (PPtree) as an oblique model for classification problem where the response variable is categorical and the method is define to use the quantitative feature, which built the tree from the available variables to enhance its performance in multi-class problems and in the presence of nonlinear separations [11]. Two project pursuit indexes, LDA and SVM used in this research, PPtree as based on optimized the projection pursuit index to find low-dimensional projections that separate classes of the group. At each node, the PPtree uses the best projection to separate two groups of classes using LDA or SVM projection pursuit indices with class information. One class assigned to only one final node with the condition that the depth of the oblique PPtree cannot be greater than the number of classes. Therefore, the PPtree constructs a simple but more understandable tree for classification. The projection coefficients of each node represent the importance of the variables to the class separation of each node. To enhance the performance accuracy of the ensemble PPForest method and to improve the generalization of this model, a novel weak tree remover used to ignore the trees with lower out of a bag and tune the PPtree in order to enhance the performance accuracy of the PPForest in general. Chi-square method used for feature selection to prove if running the PPForest algorithm with a relatively small size of the dataset could improve the performance of the PPForest [12]. After analysis the outcome of the proposed method based on classification performance metrics regarding different Telecom datasets in the number of observations and attributes, it has been shown that the proposed ensemble method using PPForest with LDA Indice has robust results of overall churners prediction system. Far from complexity computational in the terms of time and saving complexities, there are no differences in churn classification output of wither using feature selection method or not. The structure of the suggested paper prepared in sections as illustrate in follows: Section 1, present the introduction and previous studies about customer churn prediction in the Telco sector. Methodology, model building, data preprocessing, chi-square test, executed methods are described in Section 2. Section 3, illustrated the experimental implementation and outcomes of churn system are discussed. PPForest graph and huber plot of pptree visualize in sections 4 and 5.Conclusions are considered in Section 6.

Churners and non-churners classification regard as predominant trouble for telecom providers and is defined as the missing of customers because they leave for competitors. Being able to classify customer churn in advance, provides the Telco company an appreciated insight to retain its customer base. Wide ranges of churn classification methods have investigated in recent years. Most innovative models make use of state-of-the-art machine learning classifiers and identified that the origins of customer churn related to the quality of services, demographic factors, customer satisfaction/dissatisfaction, and economic value factors. Table 1. The Literature Review of recent research related to the suggested CCP model based on Ensemble PPForest algorithm.

Table 1. The literature review

No.	Authors	Title	Journal	Year	Objective	Techniques	Dataset	Performance Metrics and Outcomes
1	Lee, Yoon Dong Cook, Dianne Park, Ji-won Lee, Eun-Kyung [13]	PPtree: Projection pursuit classification tree	Electronic Journal of Statistics	2013	Proposed new classification tree, the projection pursuit classification tree (PPtree).	Combines tree-structured methods with projection pursuit dimension reduction. The PPtree uses LDA, Lr or PDA as indices.	Iris data	Projection coefficients can be used to extract the variable importance. This information is very helpful in classification problems.
2	Abbasimehetakand Tarokh [14]	A comparative assessment of the performance of ensemble learning in customer churn prediction.	Int. Arab J. Inf. Technol.	2014	Performed a comparative assessment of the performance of four popular ensemble methods. Also, it investigated the effectiveness of two different sampling techniques, i.e., oversampling as a representative of basic sampling techniques and the Synthetic Minority Oversampling Technique.	Bagging, Boosting, Stacking, and Voting based on four known base learners, i.e., C4.5 Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Reduced Incremental Pruning to Produce Error Reduction (RIPPER).	Larose	AUC, sensitivity, and specificity. Conclude that Boosting RIPPER and Boosting C4.5 are the two best methods and these results indicate that ensemble methods can be the best candidate for the CCP model.
3	Idris and Khan [15]	Ensemble Based Efficient Churn Prediction Model for Telecom	Proc. - 12th Int. Conf. Front. Inf. Technol. FIT 2014	2014	Exploits the discriminative feature selection capabilities of minimum redundancy and maximum relevance in the first step, leading to an enhanced feature-label association and reduced feature set.	Diverse Ensemble is constructed using majority voting then feature selection used as the second step. Final decision made using Ensembling of Random Forest, Rotation Forest, and KNN.	Orange Telecom, Cell2cell	AUC, Sensitivity, Specificity. Q-Statistics, The proposed Ensemble approach has the best performance.
4	Natalia da Silva [16]	Bagged projection methods for supervised classification in big data	Iowa State University, Digital Repository	2017	Develops new classification methods, and visual tools for random forest built on trees using linear combinations of variables.	Process of bagging and combining results from different PPtree.	Australian crab dataset	The algorithm implemented in the R package and design a small web app.
5	Natalia da Silva, Cook and Lee [17]	A Projection Pursuit Forest Algorithm for Supervised Classification,	arXiv:1807.07207v2 [stat.ML] 25 Jul 2018	2018	New ensemble learning method for classification problems called projection pursuit random forest (PPF).	PPtrees are constructed by splitting on linear combinations of randomly chosen variables. Projection pursuit is used to choose a projection of the variables that best separates the classes.	Crab, fish catch, leukemia, lymphoma, olive, and wine.	Performance comparison graphically between RF, PPtree, PPForest and CART on used datasets and found that PPF performs best as compared to other methods.

Table 1. The literature review (*continue*)

No.	Authors	Title	Journal	Year	Objective	Techniques	Dataset	Performance Metrics and Outcomes
6	Ahmad, Jafar, and. Aljoumaa [18]	Customer churn prediction in telecom using machine learning in big data platform	Journal of Big Data	2019	Develop a churn prediction model that assists telecom operators to predict customers churn in big data by extracting SNA features based on cloud computing.	The model is prepared and tested through the Spark environment using cloud computing. The model used Decision Tree, Random Forest, Gradient Boosted Machine Tree "GBM" and Extreme Gradient Boosting "XGBOOST".	SyriaTel, MTN telecom companies	AUC, the best results were obtained by applying the XGBOOST algorithm.
7	Selvaraj and Sruthi [19]	An Effective Classifier for Predicting Churn in Telecommunication	Journal of Advanced Research in Dynamical and Control Systems 11(01-special issue):221	2019	The suggested model aims to find the features that highly influence of customer churn operation.	Machine-learning algorithms like KNN, Random Forest and XG Boost.	IBM Watson	F-Score, Accuracy. Fiber Optic customers with greater monthly charges attributes have higher influence for churn. XG boost classifier performs outperform the other methods.

2. RESEARCH METHOD

The objective of the suggested scheme consists of building a classification model for indicating each individual client to be a potential churner or non-churner in Telecom datasets. This procedure will assist customer relationship management (CRM), by adopting the crucial retention policies that are likely to attract customers and attract who have the most tendency to churner and pursuit them to remain. The input for suggesting customer churn prediction (CCP) model includes information from past calls for each mobile subscriber, together with all the individual and business information preserved by the telecom service provider. After the prediction model entirely trained with the training dataset. Then, the model must be able to predict churners from the test dataset. The recommended methodology for churners prediction has been denoted as a schematic diagram as mentioned in Figure1 and the detailed explanation of the steps followed in given subsections.

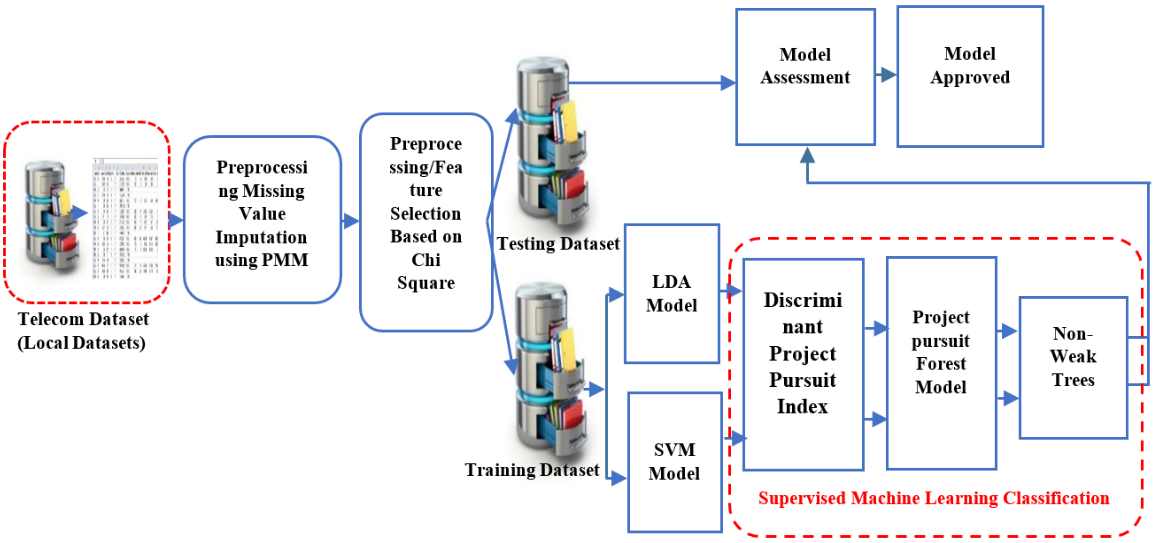

Telecom Dataset (Local Datasets)
Preprocessing Missing Value Imputation using PMM
Preprocessing/Feature Selection Based on Chi Square
Testing Dataset
Training Dataset
Model Assessment
Model Approved
LDA Model
SVM Model
Discriminant Project Pursuit Index
Project pursuit Forest Model
Non-Weak Trees
Supervised Machine Learning Classification

Figure1. Customer churn prediction using Ensemble PPForest model

2.1. Datasets

The practical part of the research is running on different Telecom datasets provided by various wireless Telco operators around the world. Table 2 summarizes the main characteristics of these datasets i.e. name, number of observations, number of attributes and Churn Rates and missing value percentage.

Table 2. Summarize Some of Research Datasets

Datasets	#Observations	# Features	Target Churn Variable			Missing Value Rate
			Non-churn	Churn	Churn Rate	
Larose Telco	5000	21	0.859	0.145	7.07	0.000
Telecom1 Telco	12499	20	0.607	0.393	2.54	0.441
WA_Fn_use_Telco	7032	21	0.735	0.265	3.77	0.007
South Asian Telco	2000	14	0.500	0.500	1.00	1.618
Cell2cell Telco	71000	78	0.710	0.290	3.45	0.647
Telcom2 Telco	50000	163	0.494	0.506	1.98	0.000

As can be seen from the table, the smallest data set contains 2000 observations, and the largest up to 71000 observations. To implement CCP methodology this characteristic allows us to split each dataset randomly into 0.8 training set and 0.2 test set. The datasets also differ substantially regarding the number of attributes, in a range from 14 up to 163. However, more attributes do not guarantee a better classification model it means heavily increases in the computational complexity required to run the empirical codes of research. The final performance of a classifier mainly depends on the feature engineering of the attributes, and not on the number of attributes available. Most of the data, however, are collected over a period of three to six months, with a churn flag indicating whether a customer churned in the month after the month following the period when the data was collected. The table also indicates the class distribution, which is for all datasets heavily skewed. The percentage of churners typically lies within a range of 1% to 5% of the entire customer base, depending on the length of the period in which churn is measured. The table also shows the missing value rate, the presence of the ambiguity of these values has a significant influence on the low predictive accuracy of the CCP model.

2.2. Datasets pre-processing

Preprocessing is a data mining approach involved converting raw data into a comprehensible format. The actual information in the world often incomplete, inconsistent missing in certain behaviors and patterns and may have many mistakes. Pre-processing is proved for solving these problems. Most of the telecom datasets come with high missing values. Instead of removing variables and observations that have high missing values, another approach is to fill up in missing value variables. A diversity approach can be used in missing features imputation that ranges from extremely simple to relatively complex. This paper used the main method for exploring and fill with missing values called Predictive Mean Matching (PMM) [20]. PMM technique is widely used as an outstanding method for variables imputation and has an attractive way to do multiple imputations especially for filling up the quantitative variables that are suffering from irregular distribution [21]. PMM can be applied in two steps. First, the approximating mean function is predicting. Second, the data with missing value imputed by finding the similar fields in the dataset, this done by means of a nearest-neighbor technique then, the observed outcome value of the nearest neighbor can be used for imputation.

2.3. Features selection based on chi-square test

The most important step in data pre-processing is to identify attributes that are certainly relevant to the target variable. However, not all attributes are well contributed to the classifier learner model. Due to the wide-scale datasets in telecom provider services, the feature selection process became essential to improve the performance and make the CCP model easier to interpret, decrease overfitting, eliminating variables that are redundant and do not provide any information or contribution in the output of the model. Moreover, it reduces the size of the prediction problem and enables classification algorithms as possible to yield outcomes in a faster manner [22]. The Chi-Square test is a nonparametric statistical analysis method commonly used to determine the significant relationship between dataset features [23]. The methodology of measuring the independence between qualitative statistic values based on the Chi-Square test depicted in the following algorithm steps.

Chi-square independent test pseudocode:

1. State the hypotheses: The statistical test for independence can be applied to categorical variables. The null hypothesis states wither the variables are independent. The alternative hypothesis states wither the variables are dependent.

2. Formulate analysis plan: The analysis strategy describes how to use samples of data to accept or reject the null hypothesis.
3. Analyze sample data: Using data sample, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the statistic test.
   a. The degrees of freedom (DF) are computed and equal to Equation 1.

$$\mathrm{DF} = (\mathrm{r} - 1) * (\mathrm{c} - 1) \tag{1}$$

   where r is the number of levels for one categorical variable, and c is the number of levels for the other categorical variable.

   b. The expected frequency count can be computed separately for each level of one categorical variable at each level of the other categorical variable as in Equation 2.

$$\mathrm{E}_{\mathrm{r,c}} = \frac{\mathrm{nr}*\mathrm{nc}}{\mathrm{n}} \tag{2}$$

   The Chi-Square random test of the variable (X_2) defined by the following Equation 3.

$$\mathrm{X}_2 = \sum \left[\left(\mathrm{O}_{\mathrm{r,c}} - \mathrm{E}_{\mathrm{r,c}}\right)^2 / \mathrm{E}_{\mathrm{r,c}} \right] \tag{3}$$

   where Or,c is the observed frequency count at level r of variable A and level c of variable B, and Er,c is the expected frequency count at level r of variable A and level c of variable B.

   c. The P-value is the probability of observing a statistic sample as extreme as the test statistic. Since the statistic test is a Chi-Square, the distribution calculator to assess the likelihood related to the statistic test.
4. Interpret results: If the output samples are improbable means given the null hypothesis, the procedure rejects the null hypothesis. Typically, this involves comparing the P-value to the consequence level, and rejecting the null hypothesis when the P-value is smaller than the significance level [24]. Table 3 show represents the Telecom datasets after imputation of the missing values and Feature Selection.

Table 3. Telecom datasets after apply PPM and chi-square

Datasets	Original Datasets		Datasets/chi-square test	
	#Observations	#Features	#Observations	# Features
Larose Telco	5000	21	5000	7
Telecom1	12499	20	12499	18
Cell2cell Telco	71000	78	71	44
WA Fn use Telco	7032	20	7032	12
Telcom2 Telco	50000	163	50000	136
South Asian Telco	2000	14	2000	10

2.4. Project pursuit random forest (PPForest)

A Random Forest is an Ensemble-learning model built on bagging multiple oblique trees that represent independent decision trees with feature selection and generate the result of classification by feeding the input to these internal trees and collecting their outcomes based on voting technique [16]. Most of the available traditional Random forests are vulnerable to overfitting in some Telecom datasets and do not handle huge numbers of redundant features. It is more efficient to choose a random decision boundary than using the available techniques, thus making larger ensemble methods are more achievable. Although this may seem to be a benefit it has the consequence of shifting the computation complexity from training time to assessment time, which is actually a disadvantage for most machine learning implementation [25]. The most available random forest are separate features space by hyperplanes that are orthogonal to single feature axes when the data are collinear with correlated features, hyperplanes that are oblique to the axis do the better class separation. Trees that use linear combinations of variables in a node splitting procedure that included in the random coefficient generation known in the literature as oblique trees [26].

PPForest involves structured tree approaches with projection pursuit indices, for dimensionality reduction, they defined hyperplanes that are oblique to the feature axes in the decision tree that trained independently and has its unique structure and properties. In other words, PPtree optimizes a projection pursuit index to obtain a low-dimensional projection to separate classes and its classification problems where the response variable is categorical and the method is described to use quantitative feature variables [13]. At each split, a random sample of predictors are selected and then an optimal projection pursuit random forest classification adapts random variables to utilize an optimal linear association between variables instead of only one variable for each split in the construction of the tree to build the PPtree, this order may lead to a diversity of decision tracks to achieve the final forest prediction, it is desired to understand and compare all decision tree tracks in the context of all trees structure [17].

One important distinguishing of PPtree is that it deals with the variables always as a two-class system when the classes are more than two the means of each class is determined and used to make a reduction to two groups only by using the distances between the means of classes. For example, if we have five classes and their means of the projected variables in each class are 2.1, 2.3, 2.5, 3.5, and 3.7, the classes with mean 2.1, 2.3, and 2.5 are set to the first group and the classes with mean 3.5 and 3.7 are set to the second group. Also, in each node of the PPtree, the projection coefficients denote the variable importance for the class splitting. This information is very supportive to select important variables by PPtree. PPForest outperform a traditional random forest when splitting hyperplane between classes occurs in a linear and randomly combination of predictors for separating the classes that computed by searches for a low dimensional projection pursuit index such as Linear Discriminant Analysis (LDA), Penalized Linear Discriminant (PDA), GINI, ENTROPY and Support Factor Machine (SVM) [27].

In the first step of the optimization problem and based on the class information, a projection pursuit index is used to find an optimal one-dimensional hyperplane for separating all data and project the training data into the projection line. Then, using the projected data to redefine the optimization method in a two-class problem by comparing the mean of classes, and assign a new label to each observation. The next step is to find an optimal one-dimensional projection to separate the two classes of the classification problem. Repeat all the steps until each group has only one class from the original classes. Based on these steps the tree grows and the maximum depth of each tree in the forest determined [16].

Projection Pursuit Random Forest Pseudocode:

```
1. Let $d_n = \{(xi, yi)\}i=1^N$, be the training dataset where $x_i$ is a p-dimensional vector of explanatory variables
   and $y_i$ represents class information with i = 1, . . . N.
2. Bootstrap samples: samples of size n randomly taking from the original dataset with replacement to create
   k number of ensemble trees to use as the training dataset and the remaining samples reserved as a test
   dataset for evaluating of the proposed churn prediction model.
3. Grow the oblique tree (PPtree): for each bootstrap sample build the oblique tree structure without pruning
   as detailed below:
   a. Optimize a projection pursuit index to calculate an optimum one-dimensional projection plane $\alpha$ using
      LDA or SVM for splitting all classes in the current bootstrap samples and yield a projected data $z=\alpha x$.
   b. On the projected data z, repeated decrease the number of groups until produce two classes only, by
      comparing the means of data, and assign a new label G1 or G2 to each class.
   c. On the projected data z, redo Project pursuit with these new class labels (G1, G2) and finding the one-
      dimensional projection path $\alpha^*$ and assign a new group label G1* or G2* to each group which can
      contain more than one original class.
   d. Determine the decision rules c which is the best separation of G1* and G2* and keep both $\alpha$ and c to
      providing the decision boundary for the node.
   e. Split data into two sets in each node in the tree then, using the new group labels G1* and G2*.
      If $\alpha * TM_1 < c$ then allocate G1* to the left node else allocate G2* to the right node, where $M_1$ is the mean
      of G1*.
   f. For each group, stop if there is only one class else repeats the procedure, the splitting step iterated until
      the last two classes separated.
   g. One class assigned only to one final node; the depth of the tree is at most the number of classes.
4. Repeat step 3 for k = 1…, B where B count the tree in the forest.
5. Produces the ensemble oblique trees, based on the majority vote mechanism to predict the class for training
   data.
6. Predict the classes of each case not included in the bootstrap sample and compute miss-classification error
   and system accuracy.
7. The projection coefficients used to obtain the dimension reduction at each node used to measure
   the variable importance.
8. Weak tree remover (classifier): To enhance the performance accuracy of the PPForest algorithm and to
   improve the generalization of model, batter trees with high performance selected based on the lower out of
   bag error for classification (OOB error) that use to tune the model and avoid the trees with the worst
   outcome.
9. Determine the majority voting technique, and evaluate the system based on the selection of good oblique
   trees.
```

2.5. Discriminant function analysis (DFA)

This section introduces and discusses some aspects of statistical learning philosophy concern to discriminant SVM and LDA. It is a statistical procedure used to solve problems associated with the statistical separation among distinct classes with the assumption that the sample is normally distributed for the attributes along with homogeneous variance-covariance matrices [28]. The linear models are easy to understand where the final output is a weighted sum of the input attributes ‘xi’. The magnitude of the weight ‘wi’ shows the importance of input and its sign indicates if the effect is positive or negative. Most functions are additive in that the output is the sum of the effects of several attributes where the weights may be enforcing or inhibiting [29].

2.5.1. Linear discriminant analysis (LDA)

This paper introduces the oblique tree algorithm for churner classification that can simultaneously shrink the tree size, solve the problem of the curse of dimensionality, enhance class classification, and improved tree data and structure visualization. This can be achieved by predicting a linear discriminant model to the data in each node on the tree using the discriminant function. LDA is a kind of Discriminant Function Analysis (LDA) that discoveries linear functions of the associated variables that lead to maximum discrimination between the group centroids [30]. LDA is a simple and mathematically robust technique frequently used in pattern recognition applications as a dimension reduction technique, object classification into mutually exclusive and exhaustive groups and maximizes the inter-class scatter, minimizes the intra-class scatter concurrently and discoveries appropriate project pursuit directions for classification problem [31].

Three steps needed to perform the LDA calculation. The first stage is to find the distance between the mean values of different classes which are called the between-class variance (SB), while the second stage involved the calculation of the distance between the mean and the samples of each class which is likely known as within-class variance (SW). The third one is to create the lower-dimensional space which maximizes the between-class variance and minimizes the within-class variance [32]. To achieve the main goal of these steps, LDA attractive procedure that makes class assignments by formative the linear transformation of the data in feature space that maximizes the ratio of the between-class variance to minimize the within-class variance. In the two-class variable, the maximum class splitting occurs when the vector of quantities, ‘w’, and intercept with ‘y’ vector b, used to express the linear transformation as in Equation 4. The classes are well-separation, which implies that after the original data are projected the distance between the two means is large, and the distance of instances around each mean is small [33].

$$w = \Sigma^{-1}(\mu_i - \mu)$$

$$b = -0.5 * (\mu_i + \mu))^T \Sigma^{-1}(\mu_i - \mu) + \log\left(\frac{\pi_i}{\pi}\right) \tag{4}$$

where Σ-1 is a variance-covariance Matrix, and μ represents the mean vector of class k. πk is the prior probability of the kth class. To find the between-classes variance (SB), the separation distance between different classes that denoted by $(\mu_i - \mu)$ will calculate as in Equation 5.

$$(\mu_i - \mu)^2 = w^T(\mu_i - \mu)(\mu_i - \mu)^T w \tag{5}$$

where $S_{Bi} = (\mu_i - \mu)(\mu_i - \mu)^T$ represent the separation distance between the mean of the ith class μ_i and the entire mean μ. Then, the total between-class matrix is calculated by adding all the between-class matrices of all classes SBi. The total within-class matrices (Sw) are calculated as in Equation 6.

$$\mathrm{S_w} = \sum_{\mathrm{k=1}}^{\mathrm{N}} \mathrm{S_k}$$

$$\mathrm{S_k} = \sum_{\mathrm{i=1}}^{\mathrm{N_k}} (\mathrm{x_{ki}} - \bar{\mathrm{x}}_{\mathrm{k}})(\mathrm{x_{ki}} - \bar{\mathrm{x}}_{\mathrm{k}})^{\mathrm{T}} \tag{6}$$

In the above equations, xki and $\bar{x}_k$ denote the ith training sample of class k and the corresponding class means, respectively. After finding the between-class variance (SB) and within-class variance (SW), the index matrix (Wlda) of the LDA technique can be calculated as in equation 7.

$$W_{lda} = orgmax_W \frac{|w^T s_B w|}{|w^T s_w w|} \tag{7}$$

The solution of this equation can be calculated by finding the eigenvectors and the eigenvalues of $W_{lda} = {S_w}^{-1} S_B$. The eigenvalues are scalar values that provide information about the LDA space while the eigenvectors represent the directions of the new space [34]. The robustness of the eigenvectors reflects the ability to discriminant between different classes. The projection pursuit algorithm searches for a low dimensional projection that optimizes the LDA. The eigenvectors with the k highest eigenvalues used to construct the lower-dimensional space of LDA while the rest are negligible. The projection pursuit index $W_{lda}$is an essential in a projection pursuit LDA because it leads to achieve the purpose of the method through its optimization. The basic in projection pursuit is to find what projections pursuit is interesting [35]. The distinct benefit of the projection pursuit way over methods can avoid the curse of dimensionality by focusing on the low dimensional projections and can ignore the redundant features.

2.5.2. Discriminant SVM framework

Recently, many researchers favored using SVM as a supervised machine learning algorithm. It has been obtained well reputation in the data mining methodologies due to its promising experimental performance, reasonable memory, and time complexity with its strong mathematical basis signifying that SVM be a competitive classifier [36]. SVM regarded as one of the most influential machine learning algorithms that can be applied in large domains of real-world applications and produce many benefits over traditional classification and regression techniques. One of the greatest significant rewards is the solution of problems relates to a small subset of the original dataset, which make SVM as powerful computation, robust mathematical contextual, better generalization skill corresponding to other classification methods [37].

One remarkable characterize of SVM and other kernel-based computational methods work in multi-dimensional without significant computation cost and feature selection methods, its robustness against the error of models and has the ability to learn well with only a very small number of features. However, the main weakness of SVM that arises from it is the training phase is computationally expensive due to a good estimation of it is constant parameters such as gamma, sigma, and degree. Moreover, it is highly reliant on the size of the original dataset t [38]. This linear classifier is also known as an optimal hyperplane, the features are normally normalized to generally lie between -1 and 1 so that the samples can be divided into two distinct classes. Discriminant functions calculated by SVM are efficient ways for projecting of multi-dimensional data in a direction perpendicular to the discriminating hyperplane. Then, the projected data fitted to estimate and display the posterior probability densities and enhancement the classification accuracy of discriminant function [39]. The basic idea of classification is to try to separate different samples into different classes, for binary classification the prediction of linear hyperplane described as in Equation 8.

$$\mathrm{f(x) = w^T . x + b = 0)} \tag{8}$$

where w and b are the weight vector and a constant respectively, which have estimated from the dataset in n-dimension space, $w^T.x$ is the internal product of w ∈ R^n and x ∈ R^n vectors [40].

The dataset can be separated geometrically by a hyperplane. It should build two hyperplanes so that the hyperplanes are as far away as possible, and no samples should be between these two planes this arrangement mathematically represented by Equation 9.

$$\begin{gathered}\mathrm{w^T . x + b \geq\ +1} \\ \mathrm{w^T . x + b \leq\ -1}\end{gathered} \tag{9}$$

From this equation, it is straightforward to confirm that the normal distance between these two hyperplanes (d) is the reverse relationship to the norm ||w|| via Equation 10.

$$\mathrm{d} = \frac{2}{||\mathrm{w}||} \tag{10}$$

The hyperplane can mathematically represent by using (11).

$$\mathrm{f(x) = sgn(w^T . x + b) = sgn\big((\textstyle\sum_{i=0}^{n} \alpha_i\, y_i x_i). x + b\big)} \tag{11}$$

Where: sign () is known as a sign function, αi are non-negative Lagrange coefficients calculated by resolving a quadratic optimization function based on linear and inequity constraints. The training observations ‘xi’ with non-zero αi finds on the frontier of the margin called support vectors (SV). The transformation should be chosen in a confident way so that their dot product leads to a kernel-style function [41]. The kernel function is to use $\mathrm{k}\ (\mathrm{xi}, \mathrm{xj})$ such that its discretization $\mathrm{Kij} = \mathrm{k}\ (\mathrm{xi}, \mathrm{xj})$ is a positive certain matrix. The decision prediction can then represent as in 12.

$$\mathrm{f(x)} = \mathrm{sgn}(\textstyle\sum_{\mathrm{i}=0}^{\mathrm{n}} \alpha_{\mathrm{i}}\, \mathrm{y}_{\mathrm{i}} \mathrm{k}(\mathbf{x}\mathrm{i}, \mathbf{x}\mathrm{j}) + \mathrm{b}) \qquad (12)$$

LDA Assumes that data are normally distributed, all classes identically Gaussian distributed, in case, the classes have different covariance matrices then LDA becomes quadratic and not linear discriminant analysis [42]. However, SVM assumes that all classes are very separable; it makes use of a slack variable that permits a certain amount of overlap between the classes. SVM is a precise flexible prediction method that makes no expectations about the input datasets at all. The flexibility, on the other hand, was frequently given it more difficult to understand the outcomes from an SVM classifier, as compared to LDA. Moreover, LDA makes use of the complete input dataset to approximation covariance matrices that are somewhat prone to outliers. While SVM optimization functions over a subset of the data that locate on the separating margin [43].

3. RESULTS AND ANALYSIS

Robust practical setup and use of statistical tests and appropriate performance measures are essential to figuring a correct conclusion. The telecom industry reflects different types of measurements to assess the performance of the churn prediction model.

3.1. Accuracy

Count the correct predictions accomplished by the prediction model over all kinds of predictions made. Overall, how often the classifier model is correct

$$\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \qquad (13)$$

3.2. Precision (Confidence)

The number of positive cases that correctly recognized.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (14)$$

3.3. Sensitivity (Recall)

The amount of actual positive cases that correctly recognized.

$$\mathrm{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP+TN+FP+FN}} \qquad (15)$$

3.4. Specificity

The amount of actual negative cases that correctly recognized.

$$\mathrm{Specificity} = \frac{\mathrm{FP}}{\mathrm{FP+TN}} \qquad (16)$$

3.5. Prevalence

How often does the positive condition occur in the sample.

$$\mathrm{Prevalence} = \frac{\mathrm{TP+FN}}{\mathrm{TP+TN+FP+FN}} \qquad (17)$$

3.6. Error rate (ER)

The number of all negative predictions divided by the total number of samples, how much is the inaccurate prediction or misclassification on the predictive method

$$\mathrm{ER} = \frac{\mathrm{FP+FN}}{\mathrm{TP+TN+FP+FN}} \qquad (18)$$

3.7. F-Score

Precision is invaluable for assessing the performance of data mining classifiers, but it surely leaves out some facts and for that reason will also be complicated. The Recall is a portion of the true optimistic predictions to total positive observations in the dataset. Compute the percent of churn rate that appropriately categorized as churn/non-churn. The prediction models that have a low Recall means it miss-classifies a great amount of the positive cases [44].

$$\mathrm{F-Score} = 2.\frac{\mathrm{Precision}*\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{19}$$

3.8. Receiver operating characteristic curve (ROC)

ROC is a depiction of the relations between the benefits and costs, a plot in two-dimensional space of x- and y-axes in linear scale and used to summarize the trade-off between recall and 1-specificity. The ROC chart commonly used to visualize the performance of churners classifier over all possible thresholds for assigning observations to a given class. It generated by drawing the true Positive rate that represents the churners ratio correctly predicted as churners against the false positive rate that represents the non-churners ratio incorrectly classified as churners [45].

3.9. The area under receiver operating (AUC)

Measures the area below the ROC curve, the diagonal line represents a random process, it has an AUC of 0.5, thus the AUC of a reputable churn classifier should be a lot higher, preferably virtually 1, as a worth of 1 represents ultimate classifier. It represents a tradeoff between specificity and sensitivity of the model. The field-specific by means of AUC represents the chance that a random pair of churning and non-churning customers are properly identified, i.e. A positive instance receives a greater rating than a negative instance. Although AUC is well known and widely used it has been shown to be incoherent when comparing different methods [46].

3.10. Kolmogorov-smirnov test (KS)

The KS test measures the performance of classification models by match a sample with reference likelihood. Measure the amount of separation between desirable and undesirable distributions. The KS statistic test gives the maximum distance between the ROC curve and the sloping at a specific cut-off point. In most prediction models, the KS test falls in the range of zero and one, the higher value means better model in separating the positive from negative classes.

3.11. H- measure

Some researchers have been proven that the problem of the AUC is that it depends totally on the use of the data mining method and differed based on the classification method. H-measure is successfully overcoming the variance of AUC, so it captures the performance advantage of AUC but not its flaw i.e. incoherent and potentially misleading yields when the ROC curve is cross [47].

3.12. Lift measure

The effectiveness of the prediction model is expressed in the lift curve, which shows the fraction of all churners that may be caught when a designated fraction of subscribers used to be contacted. This is equal to the ratio between the sensitivity and the ratio of predicted churners after applying the churn model to the testing dataset. Formula 20 represents the lift value [48].

$$\mathrm{Lift} = \frac{\mathrm{precision}}{\mathrm{p}/(\mathrm{p}+\mathrm{N})} \tag{20}$$

3.13. Gini coefficient

Evaluator that is carefully concern with the AUC chart is Gini coefficients are equal to double of the area between the ROC arc and the baseline means *Gini =2*AUC-1*. The Gini coefficient differs between zero when the ROC curve locates on the diagonal then, the classification model does not achieve better than a random classification model. While one value means the maximum ROC curve and perfect classification.

3.14. Out-of-bag error

For each oblique tree model, in the bagging forest, some cases of the original data are not used. Predicting the response for these cases gives a better estimate for the error of the model with future data. The OOB error rate is a measure for each bagging model and used to provide the overall error of the ensemble [49].

3.15. Cost

Many of the above metrics attempt to take a balanced view between FP and FN. A principled method to acquire this is to introduce the suggestion of misclassification expenses. Let c in [0, 1] denote the ‘price’ of misclassifying a category zero object as category one (FP), and 1–c the fee of misclassifying a category one object as category zero (FN) based on the calculated confusion matrix. It's implicitly assumed that the two misclassification expenses sum is 1. Minimal Error Rate (MER), Minimum Weighted Loss (MWL).

3.16. Youden index

Youden index is one of the well-known measures of diagnostic measure of accuracy. It is a global measure of test performance, used in the evaluation of the overall discriminative power of a diagnostic procedure and comparison of one test with other tests. It can be calculated via sensitivity- (1- specificity). It ranges from 0 for poor diagnostic accuracy and to a value of 1.0 for a perfect diagnostic test [50]. Different prediction modeling techniques will result in different performance based on the evaluation criteria using different data and different telecom scenarios. The best performance of the SVM model using the Radial Basis Function (RBF) kernel can be accomplished when choosing the constant parameters of SVM as shown in Table 4.

Table 4. Parametres estimation for best SVM performance

Datasets	Cost	Gamma
Larose Telco	10	1.5
Telecom1	10	1.5
Cell2cell Telco	1	1.5
WA_Fn_Use_Telco	250	250
South Asian Telco	8	3
Telcom2 Telco	1	1

After completing the constructing of proposed PPForest model performance evaluators will help in evaluating model accuracy, significant functions of evaluation metrics are used to assist the skill to discriminate among model outcomes. In this section, the churners/non-churners prediction methodology is assessing by comparing the performance of two discriminant functions used in the construct of PPtree. The first method used SVM as a project pursuit index in the construction of PPForest to differentiate between churners and non-churners customer classes. The second method is LDA to achieve linear splitting of variables node during oblique PPtree construction to produce individual classifiers that are an ensemble, robust and more diverse than classical Random Forest. Tables 5 and 6 depicted the performance of the proposed churner classification framework using ensemble PPForest using two techniques, LDA and the other is SVM. In order to prove that PPForest makes important feature selection, the performance of the churn prediction model evaluated using the comparison of datasets after applying feature selection strategy based on the Chi-Square statistical test and the whole features of telecom datasets.

Depended on the performance of the proposed PPForest method in this research and the exploration studies by other researchers, it is predicted that the LDA with linear project pursuit index could be used as the classifier of customer churners in telecom datasets and it produced uplifted outcomes than SVM in most of the performance measures in terms of class discrimination by using telecom datasets where the discriminatory information not aligned with the direction of maximum variance. The respectable accuracy value depicts that the LDA has the best performance in some telecom datasets. Moreover, better AUC, ROC coefficients and KS statistical tests show that the prediction model can retain more covert churner customers with less cost in the Telecom sectors and diverse churn rates. The reason for reasonable outcomes is acquired based on the proposed models, the LDA and SVM models, which have appropriate kernel function and constant values parameters that make churn prediction model on structural risk minimization which includes empirical risk and confidence minimization.

Table 5. PPForest based on LDA, SVM with whole features of Telcom datasets

	Larose	Telcom1	Cell2cell	WA-FN-USE	South Asian	Telcom2
PPForest Based on LDA with the Whole Feature						
Accuracy	0.7959	0.7139	**0.9914**	0.7246	0.6853	0.59596
RMSE	0.4671	1.1500	**0.9958**	0.5267	0.7169	0.49432
KS	0.1362	0.6068	**0.7099**	0.1753	0.0315	0.19092
OOB	0.2198	0.2874	**0.0110**	0.2803	0.3115	0.40590
AUC	0.7520	0.7030	**0.9850**	0.7500	0.6855	0.59546
Prevalence	0.7267	0.5936	**0.7215**	0.5605	0.4654	0.50568
Precision	0.9395	0.7690	**0.9839**	0.9049	0.7025	0.59921
Sensitivity	0.7953	0.7523	**0.9999**	0.6909	0.6540	0.63938
Specificity	0.6888	0.6512	**0.9600**	0.7994	0.7230	0.55155
Gini	0.5037	0.4064	**0.9704**	0.4996	0.3710	0.19092
F-score	0.8614	0.7605	**0.9918**	0.7835	0.6774	0.57439

Table 5. PPForest based on LDA, SVM with whole features of Telcom datasets (*continue*)

	Larose	Telcom1	Cell2cell	WA-FN-USE	South Asian	Telcom2
PPForest Based on SVM with the Whole Feature						
Accuracy	0.6085	0.6776	**0.7336**	0.6705	0.6868	0.6064
RMSE	0.6318	0.5681	0.5125	0.5797	0.5568	**0.3946**
KS	**0.3520**	0.0631	0.2286	0.2675	0.0880	0.2100
OOB	0.3914	0.3227	**0.2076**	0.3295	0.3130	0.3936
AUC	0.6980	0.6760	**0.7977**	0.7300	0.6900	0.6050
Prevalence	0.5195	0.5555	**0.5628**	0.4846	0.4159	0.5050
Precision	**0.9496**	0.6549	0.8941	0.9176	0.7248	0.6018
Sensitivity	0.5747	0.6918	**0.7088**	0.6057	0.6030	0.6455
Specificity	0.8147	0.7557	0.7945	**0.8497**	0.7710	0.5645
Gini	0.3956	0.3519	**0.5954**	0.4601	0.3800	0.2100
F-score	0.7160	0.7223	**0.7907**	0.7297	0.6583	0.6229

Table 6. PPForest based on LDA, SVM with whole features of Telcom datasets

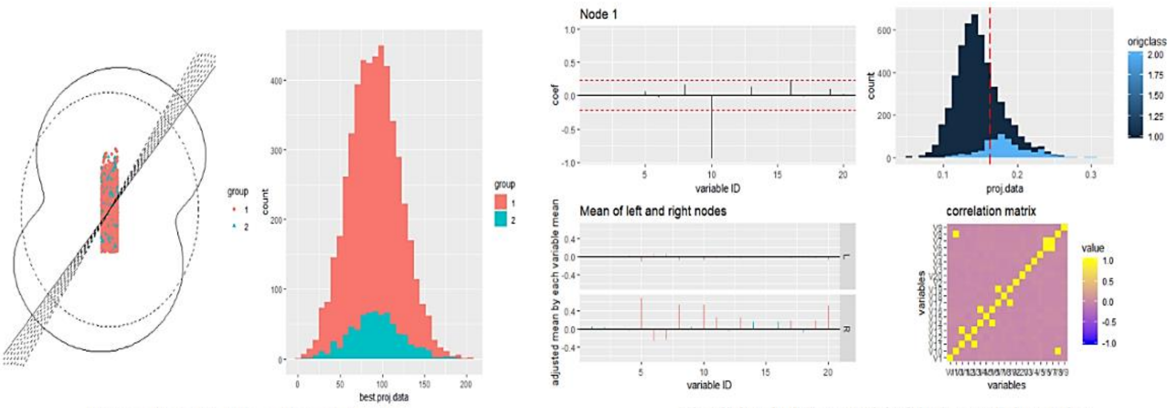
	Larose	Telcom1	Cell2cell	WA-FN-USE	South As	Telcom2
PPForest Based on LDA with the Whole Feature						
Accuracy	0.7959	0.7139	**0.9914**	0.7246	0.6853	0.59596
RMSE	**0.4671**	1.1500	0.9958	0.5267	0.7169	0.49432
KS	0.1362	0.6068	**0.7099**	0.1753	0.0315	0.19092
OOB	0.2198	0.2874	**0.0110**	0.2803	0.3115	0.40590
AUC	0.7520	0.7030	**0.9850**	0.7500	0.6855	0.59546
Prevalence	**0.7267**	0.5936	0.7215	0.5605	0.4654	0.50568
Precision	0.9395	0.7690	**0.9839**	0.9049	0.7025	0.59921
Sensitivity	0.7953	0.7523	**0.9999**	0.6909	0.6540	0.63938
Specificity	0.6888	0.6512	**0.9600**	0.7994	0.7230	0.55155
Gini	0.5037	0.4064	**0.9704**	0.4996	0.3710	0.19092
F-score	0.8614	0.7605	**0.9918**	0.7835	0.6774	0.57439
PPForest Based on SVM with the Whole Feature						
Accuracy	0.6085	0.6776	**0.7336**	0.6705	0.6868	0.6064
RMSE	0.6318	0.5681	0.5125	0.5797	0.5568	**0.3946**
KS	**0.3520**	0.0631	0.2286	0.2675	0.0880	0.2100
OOB	0.3914	0.3227	**0.2076**	0.3295	0.3130	0.3936
AUC	0.6980	0.6760	**0.7977**	0.7300	0.6900	0.6050
Prevalence	0.5195	0.5555	**0.5628**	0.4846	0.4159	0.5050
Precision	**0.9496**	0.6549	0.8941	0.9176	0.7248	0.6018
Sensitivity	0.5747	0.6918	**0.7088**	0.6057	0.6030	0.6455
Specificity	0.8147	0.7557	0.7945	**0.8497**	0.7710	0.5645
Gini	0.3956	0.3519	**0.5954**	0.4601	0.3800	0.2100
F-score	0.7160	0.7223	**0.7907**	0.7297	0.6583	0.6229

4. PPFOREST GRAPH

A conceptual framework for comparing the PPForest structure tree model and traditional random forest for the cell2cell Telecom dataset can summarize in Figure 2 (refer Appendix). See the project pursuit model is simpler than a regular classification tree, due to a combination of features that mostly separate the churn classes; just one projection needed to see the differences between the two classes.

5. HUBER PLOT AND PPTREE VISUALIZE

Huber’s is a plot of the various projection pursuit indices with class. Figure 3(left). Depicted the LDA index with important variables of Larose dataset and the histogram of the projected data on the best projection, see the separation of the churn class from the other classes. Figure3(right) depicted the nodes of the tree based on projections of the data, the coefficients of which form the building block to calculate the variable importance. The density plot displays the Larose data projection at each node, and the mosaic plot depicts the confusion matrix PPtree. Having a better visualization tool provides a selection of interactive plots to diagnose PPForest models will provide a better understanding of the dataset attributes, the model strengths, and weaknesses, model results for analysis and visualization of future data. The philosophy underlying the collection of display tools is to show the CCP model in the data space it is not easy to do this and completely take this on would require plotting the model in the dimensional data space.



Huber's plot of LDA indices with class information

Visualizing the PPtree model of the Larose Telecom data.

Figure 3. Huber's plot and the visualization of one tree of Larose telecom data

6. CONCLUSION

PPForest is a new methodology to construct a decision tree in which LDA or SVM employed to separate dataset features. Using PPForest as a forest procedure will reduce the decision tree size without sacrificing of prediction accuracy; this could be done by attractive the full benefit of the visual influence of multi-dimensional graphical displays and the predictive influence of LDA and SVM. One of the main advantages of an oblique decision tree is that it successfully uses an association between variables to find class separations, and it has a visual illustration of the variances between classes in feature space that can be used to understand model outcomes. PPForest uses the correlation between predictor variables to find the best separation between classes. It has shown that PPForest achieves better predictive performance than even a random forest where the correlation between predictors is large. Projection pursuit solves the problem with the original random forest algorithm, where oblique projections were an option but effectively useless because it simply used arbitrary projections. The space of projections is very big, so the random forest rarely can find good oblique projections. Additionally, the tree structure produced by PPForest are tested by weak tree remover procedure to ignore it, this step improved the accuracy of PPForest and subsequent one-dimensional projections of the data made for convenient visualizations of the group separations, especially for multiclass classification problems. The running of the proposed system illustrates, that the decision PPtree is not essentially easy to understand. Easy of interpretation diminutions rapidly with increasing tree size. While tree size naturally grows with the number of features. The experimental results have shown that PPForest using LDA with weak tree remover is better than PPForest using SVM in many Telecom datasets based on the evaluation measures.

REFERENCES

[1] S. Gupta, "Telecommunications at the Crossroads in India," *IIMB Manag. Rev.*, vol. 27, no. 3, pp. 196–208, Sep. 2015.

[2] P. K. Banda and S. Tembo, "Factors Leading to Mobile Telecommunications Customer Churn in Zambia," *Int. J. Eng. Res. Africa*, vol. 31, pp. 143–154, Jul. 2017.

[3] S. O. Adebiyi, E. O. Oyatoye, and B. B. Amole, "Relevant Drivers for Customers Churn and Retention Decision in the Nigerian Mobile Telecommunication Industry," *J. Compet.*, vol. 6, no. 3, pp. 52–67, Sep. 2016.

[4] A. Rodan, H. Faris, J. Alsakran, and O. Al-Kadi, "A Support Vector Machine approach for Churn Prediction in the Telecom Industry," *Inf.*, vol. 17, no. 8, pp. 3961–3970, 2014.

[5] B. T. G. S. Kumara, "Customer Churn Analysis and Prediction in Telecommunication for Decision Making," in *International Conference On Business Innovation (ICOBI), 25-26 August 2018, NSBM, Colombo, Sri Lanka*, pp. 7, August 2018.

[6] N. Imbug, S. N. A. Ambad, and I. Bujang, "The Influence of Customer Experience on Customer Loyalty in Telecommunication Industry," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 8, no. 3, pp. 103–116, 2018.

[7] W. Verbeke, "Profit-Driven Data Mining in massive Customer Networks: New Insights and Algorithms.," 2012.

[8] C. Dullaghan and E. Rozaki, "Integration of Machine Learning Techniques to Evaluate Dynamic Customer Segmentation Analysis for Mobile Customers," *Int. J. Data Min. Knowl. Manag. Process*, vol. 7, no. 1, pp. 13–24, 2017.

[9] B. I. B. and T. G., "Predicting Churn in Mobile Telecommunications Industry," *Electron. Telecommun. Predict. CHURN Mob.*, vol. 4, no. 4, pp. 271–292, 2003.

[10] J. C. Obi, "A Comparative Study of the Fisher's Discriminant Analysis and Support Vector Machines," *Eur. J. Eng. Res. Sci.*, vol. 2, no. 8, p. 35, 2017.

[11] A. J. Izenman, "Linear Discriminant Analysis," in *Modern Multivariate Statistical Techniques. Springer Texts in Statistics. Springer*, vol. 30, no. 2, pp. 237–280, 2013.

[12] O. S. Bachri, Kusnadi, M. Hatta, and O. D. Nurhayati, "Feature Selection based on CHI Square in the Artificial Neural Network to Predict the Accuracy of the Student Study Period," *Int. J. Civ. Eng. Technol.*, vol. 8, no. 8, pp. 731–739, 2017.

[13] Y. D. Lee, D. Cook, J. Park, and E.-K. Lee, "PPtree: Projection Pursuit Classification Tree," *Electron. J. Stat.*, vol. 7, no. 1, pp. 1369–1386, 2013.

[14] H. Abbasimehr, M. Setak, and M. J. Tarokh, "A Comparative Assessment of the Performance of Ensemble Learning in Customer Churn Prediction," *Int. Arab J. Inf. Technol.*, vol. 11, no. 6, pp. 599–606, 2014.

[15] A. Idris and A. Khan, "Ensemble Based Efficient Churn Prediction Model for Telecom," *Proc.-12th Int. Conf. Front. Inf. Technol. FIT 2014*, pp. 238–244, 2014.

[16] N. Da Silva Cousillas, "Bagged Projection Methods for Supervised Classification in Big Data," *Iowa State University, Digital Repository*, Ames, 2017.

[17] N. da Silva., D. Cook, and E.-K. Lee, "A Projection Pursuit Forest Algorithm for Supervised Classification," *arXiv1807.07207v2 [stat.ML] 25 Jul 2018*, pp. 1–25, Jul. 2018.

[18] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer Churn Prediction in Telecom using Machine Learning in Big Data Platform," *J. Big Data*, vol. 6, no. 1, p. 28, Dec. 2019.

[19] S. Selvaraj and M. Sruthi, "An Effective Classifier for Predicting Churn in Telecommunication," *J. Adv. Res. Dyn. Control Syst. 11(01-special issue)221*, vol. 11, pp. 10, 2019.

[20] R. Armina, A. Mohd Zain, N. A. Ali, and R. Sallehuddin, "A Review on Missing Value Estimation using Imputation Algorithm," *J. Phys. Conf. Ser.*, vol. 892, no. 1, pp. 012004, Sep. 2017.

[21] G. Vink, L. E. Frank, J. Pannekoek, and S. van Buuren, "Predictive Mean Matching Imputation of Semicontinuous Variables," *Stat. Neerl.*, vol. 68, no. 1, pp. 61–90, Feb. 2014.

[22] L. G. G. K. Paul Nesselroade Jr, *Statistical Applications for the Behavioral and Social Sciences, 2nd Edition*, Second edi. John Wiley & Sons, Inc., 2019.

[23] M. L. McHugh, "The Chi-Square Test of Independence," *Biochem. Medica*, vol. 23, no. 2, pp. 143–149, 2013.

[24] A. Q. Miah, *Applied Statistics for Social and Management Sciences*. Singapore: Springer Singapore, 2016.

[25] G. Gandhi and R. Srivastava, "Analysis and Implementation of Modified K-Medoids Algorithm to Increase Scalability and Efficiency for Large Dataset," *Int. J. Res. Eng. Technol.*, vol. 03, no. 06, pp. 150–153, Jun. 2014.

[26] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, and J. Vanthienen, "Social Network Analytics for Churn Prediction in Telco: Model Building, Evaluation, and Network Architecture," *Expert Syst. Appl.*, vol. 85, pp. 204–220, Nov. 2017.

[27] N. da Silva, D. Cook, and E.-K. Lee, "Interactive Graphics for Visually Diagnosing Forest Classifiers in R," *Electron. J. Stat.*, vol. 7, pp. 1369–1386, Apr. 2017.

[28] E. M.N and B. D.C, "A Discriminant Function Analysis Approach to Country's Economy Status," *J. Adv. Stat.*, vol. 2, no. 4, pp. 125–136, Dec. 2017.

[29] G. A. Giraldi, P. S. Rodrigues, E. C. Kitani, J. R. Sato, and C. E. Thomaz, "Statistical Learning Approaches for Discriminant Features Selection," *J. Brazilian Comput. Soc.*, vol. 14, no. 2, pp. 7–22, 2008.

[30] H. Zhao, Z. Wang, and F. Nie, "A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 4, pp. 629–640, Apr. 2019.

[31] D. Lu, C. Ding, J. Xu, and S. Wang, "Hierarchical Discriminant Analysis," *Sensors*, vol. 18, no. 1, pp. 279, Jan. 2018.

[32] B. Ghojogh and M. Crowley, "Linear and Quadratic Discriminant Analysis: Tutorial," *ArXiv 2019*, no. 4, pp. 1–16, Jun. 2019.

[33] D. Pregibon and T. J. Hastie, "Generalized Linear Models," in *Statistical Models in S*, pp. 271–291, 2019.

[34] A. J. Izenman, *Modern Multivariate Statistical Techniques*. New York, NY: Springer New York, 2008.

[35] A. M. Pires and J. A. Branco, "Projection-Pursuit Approach to Robust Linear Discriminant Analysis," *J. Multivar. Anal.*, vol. 101, no. 10, pp. 2464–2485, Nov. 2010.

[36] R. Y. Goh and L. S. Lee, "Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches," *Adv. Oper. Res.*, pp. 1–30, Mar. 2019.

[37] C. Satisfaction, "Efficient Customer Churn Prediction Model Using Support Vector Machine with Particle Swarm Optimization," *Int. J. Pure Appl. Math.*, vol. 119, no. 10, pp. 247–254, 2018.

[38] R. Gholami and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," in *Handbook of Neural Computation*, 1st ed., Elsevier, pp. 515–535, 2017.

[39] X.-S. Yang and J. P. Papa, "Bio-Inspired Computation and its Applications in Image Processing: an Overview," in *Bio-Inspired Computation and Applications in Image Processing*, Elsevier, 2016, pp. 1–24.

[40] J. Cervantes, F. García Lamont, A. López-Chau, L. Rodríguez Mazahua, and J. Sergio Ruíz, "Data selection based on Decision Tree for SVM Classification on Large Data Sets," *Appl. Soft Comput. J.*, vol. 37, pp. 787–798, 2015.

[41] A. Muñoz, J. M. Moguerza, and G. Martos, "Support Vector Machines," *Wiley StatsRef Stat. Ref. Online*, pp. 1–13, 2019.

[42] M. A. J. Tengnah, R. Sooklall, and S. D. Nagowah, "A Predictive Model for Hypertension Diagnosis Using Machine Learning Techniques," in *Telemedicine Technologies*, Elsevier, pp. 139–152, 2019.

[43] J. Nalepa and M. Kawulok, "Selecting Training Sets for Support Vector Machines: a Review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 857–900, 2019.

[44] S. Ahmad, "26 Thoughts on 'Basic Evaluation Measures from the Confusion Matrix,'" 2016. [Online]. Available: https://classeval.wordpress.com/introduction/basic-evaluation-measures/.

[45] A. S. Halibas, A. Cherian Matthew, I. G. Pillai, J. Harold Reazol, E. G. Delvo, and L. Bonachita Reazol, “Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modeling,” *2019 4th MEC Int. Conf. Big Data Smart City, ICBDSC 2019*, no. March 2019.

[46] M. Azeem and M. Usman, “A Fuzzy-Based Churn Prediction and Retention Model for Prepaid Customers in the Telecom Industry,” *Int. J. Comput. Intell. Syst.*, vol. 11, no. 1, pp. 66, 2018.

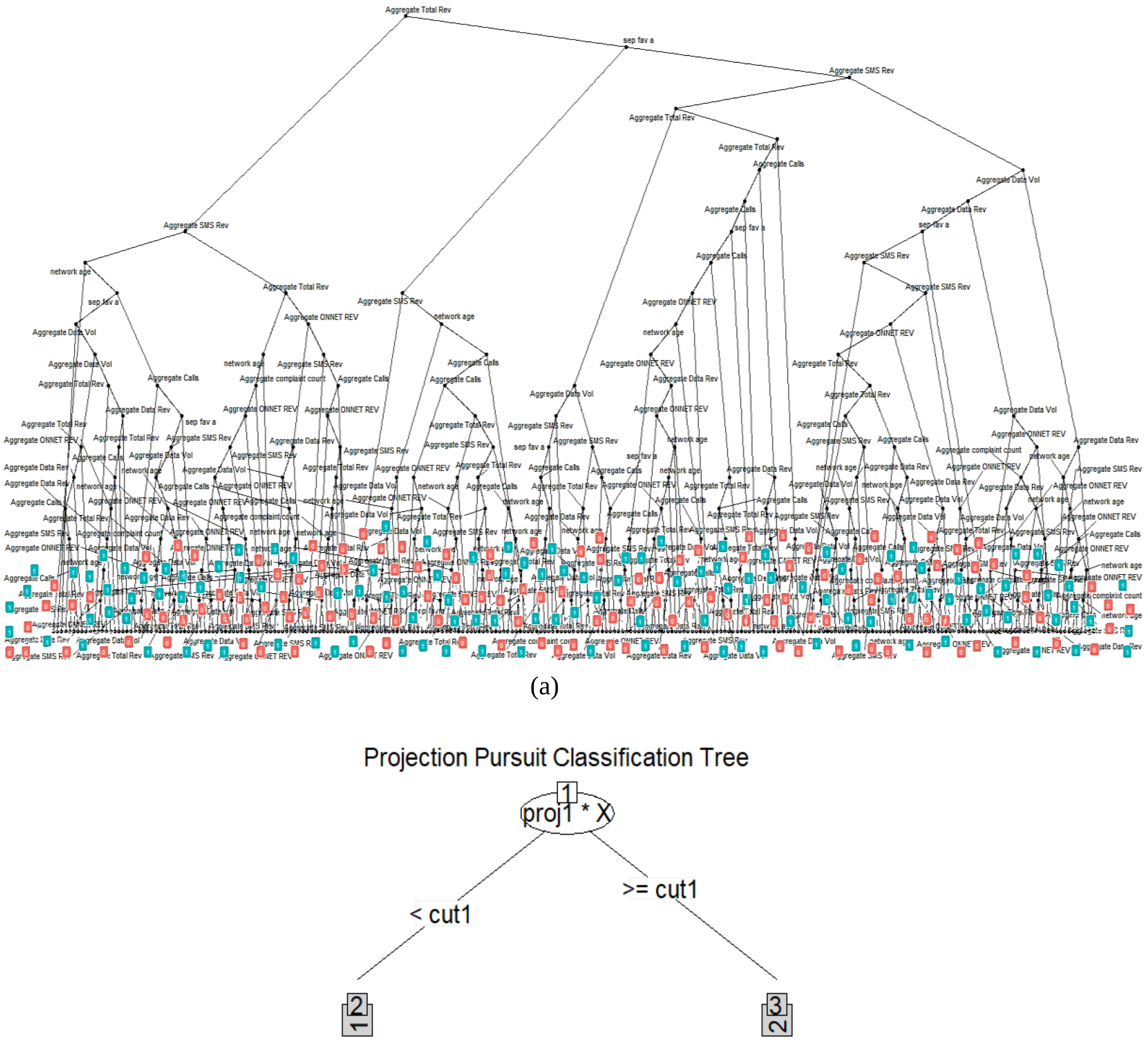
[47] D. J. Hand and C. Anagnostopoulos, “A Better Beta for the H Measure of Classification Performance,” *Pattern Recognit. Lett.*, vol. 40, no. 1, pp. 41–46, Feb. 2012.

[48] W. Verbeke, “Profit-Driven Data Mining in Massive Customer Networks: New Insights and Algorithms.,” Katholieke Universiteit Leuven, 2012.

[49] N. Adams, E. Cohen, and D. J. Hand, “Evaluating Statistical and Machine Learning Supervised Classification Methods,” *Stat. Data Sci.*, pp. 37–53, 2018.

[50] R. Fluss, D. Faraggi, and B. Reiser, “Estimation of the Youden Index and It’s Associated Cutoff Point,” no. c, pp. 1–14, 2001.

APPENDIX



(a)

(b)

Figure 2. Comparison of the Classical Decision Tree (a) and PPForest (b) Algorithms on Simulated dataset