International Journal of Engineering & Technology, 7 (4.19) (2018) 657-664

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predicting Customer Churn in Telecom Sector based on Penalization Techniques and Ensemble Machine Learning

Asia Mahdi Naser*[1], Eman al-shamery[2]

[1]*Computer Science Department, College of Science, KerbalaUniversity, Karbala, Iraq.*
[2]*Software Department, College of Information Technology, BabylonUniversity, Babel, Iraq.*
**Corresponding Author Email: asia.m@uokerbala.edu.iq*

Abstract

Customer Churn Prediction model (CCP)aims to detect customers with a high propensity to leave. The target of this research is to handle a large scale Telecommunication Company and identify potential churn. In the proposed research, Predictive Mean Matching (PMM) algorithm used to handle missing values, instead of removing features or observations with high missing data.
First Ensemble Machine learning classifieris offered to investigate and compare the combining of an Ensemble learner based on Generalized Linear Model (GLM) and the prediction values based on tree model using a Random Forest classifier. The suggested CCP model employed the Weighted Accuracy and Diversity (WAD) as an algorithm to find the optimal weights for the proposed Ensemble classifier.
The second Ensemble learner based on the generalized linear model is incorporated of penalized methods (Ridge, Lasso,and ElasticNet) with a Logistic Regression method on the binomial family. Randomly generate values between [0, 1] became the weights for this classifier. The Weights are selected according to the principle that weights of highervalue are assigned for great performance classifier to ensure the highestaccuracy of Churn Prediction model. 10-fold, based on five times repeated Cross-Validation (CV) performance technique used to enable efficient and automatic search for the optimal value of lambda λ parameter for penalization methods.
The two Ensemble classifiers incorporated within a customer churn prediction model to handle a large scale dataset, time-dependent features label, and an imbalance data distribution in the Telecommunication industry.
Experimental results show an increasein predictive performance. In addition, the results depicted that using of ensemble learning has brought a significant improvement for individual base learners in terms of performance indicators such as Area under Curve (AUC), sensitivity, specificity, Accuracy, and Mean Square Error(MSE), Accuracy is the best candidates for churn prediction tasks.

Keywords: *Customer Churn Prediction, Random Forests, Ensemble Machine Learning, Weighted accuracy and diversity, Telecommunication Industry, Boosting, Penalization Method, Regularization Techniques.*

1. Introduction

In an era of developed markets and intense competitive pressure, it is fundamental for companies to manage relationships with their customers to increase their revenues. In business economics, this concept is known as the "Customer Relationship Management" (CRM), which is a business strategy that aims to ensurecustomer's satisfaction[1]. The companies which successfully apply CRM to their business nearly always improve their retentionpower which represents the probability that a customer will not leave. In fact, a high retention power avoids useless waste of money, since acquiring new customers can cost five to six times more than satisfying and retaining existing customers[2]. This led companies to put a lot of effort in understanding and analyzing their customer behaviors and invest significant resources to provide a product or service that stands out from the competitors. The phenomenon related to the customer that ceases his relationship with a company. Besides losing the customer, it is also likely that the customer will join a competitor company is commonly called Customer Churn[3].Customer churn prediction (CCP) aims to detect customers with a high tendency to cut ties with the provider services or a Telecom company. An accurate prediction learner allows the companies to take actions for targeting the customers who are most likely to attrite. The CCP model can improve the efficient use of limited resources and results in a significant impact on businesses.
CCP has been raised as a key issue in many fields such as Telecommunication, Credit Card, Internet Service Providers, Electronic Commerce Retail Marketing, Newspaper publishing companies, banking and financial services[4].
Customer churn prediction 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 past information, prior behaviors and offering some services to persuade them to stay. In the other hand, Long-terms customersare more profitable for the service providers, since they are more dependency to buy additional products and spread the word of customers satisfaction, thus procedure will indirectly attract more and more customers. The degree of CCP model quality can be determined by both datasets and Ensemble learning algorithm. Therefore, current studies focus on these two aspects in order to optimize the accuracy of the CCP model performance[5].
The structure of this research is arranged as follows: SectionII presented a literature survey about current customer churn prediction models. Methodology, CCP model building, Data

preprocessing, execution of proposed methods is depicted in section III. In section IV, the practical experimental and results are discussed. Conclusions are considered in section V.

2. Related Works

CCP is an essential problem for telecom companies and it is defined as a customer's dropping off since they move out to other competitors. The ability to classify the churn customer in advance provides the company with high valuable insight in order to retain and increase their customer database. A wide range of customer churn predictive systems have been developed in the last years.

Authors in[6] suggested Multiple imputation strategies for an explanation of uncertainty in missing values called a Predictive Mean Matching method (PMM).They investigated the concert of PMM as well as dedicated a practical applicability of PMM on semi-continuous classes by performing simulation studies under univariate and multivariate missing data tools. They found that the performance of PMM is flexible as compared with other missing data prediction methods and it only yields reasonable imputations and salvages the distributions of original data.

Paper [7] systematically, compare the effectiveness of popular Ensemble machine learning approaches performance, i.e., Boosting, Bagging, Stacking, and Voting based on C4.5 decision tree, Support Vector Machine, neural network,and reduced incremental pruning to Produce Error Reduction based learners with different sampling techniques and performance evaluators such as AUC, sensitivity, and specificity. Experimental results indicate that the Ensemble methods are more dominated as compared with base learner methods inCCP and, the boosting Ensemble method can be the best candidate for churn prediction tasks.

Researchers in[8] introduced a thesis that covered Random Forests in terms of predictive and theory, to asses new sight on its learning capabilities, working of inner trees and it is interpreted abilities. From theoretical analysis showed Mean Decrease of Impurity variable important Measure, data analysis illustrates that variable importance as computed based Random Forest method suffered from a set of defects, resulted from masking effects, miss-classification of features impurity or due to the inner structure of Decision Trees. The thesis also depicted a Random Forest performance in the context of large datasets and showed that using the samplingmethodprovided an increase in prediction modelperformance and, lowering the memory requirements.

paper[9] aimed to build an Ensemble Machine learning that able for accurately predicate type II diabetes patients on African Americans dataset using Regression methods i.e., Logistic regression, Lasso, the Ridge and ElasticNet with a classifiertree-based method such as decision tree. They found the contribution of each learning algorithm in the Ensemble super learning equation that used as an input in Regression strategy to predict the actual outcomes, optimizing minimum MSE, error rate, and the risk and coefficients for each prediction methods.

The author in[10] introduced histhesis and using primary data collection method from customers to build a customer churn predictive model and evaluated the churn rate of six telecommunication companies in Ghana. CCP model includesthe using of machine learning methods such as the C5.0 decision tree algorithm, the Logistic Regression method,and the Discriminant Analysis algorithm. A comparative assessment is performed to discover the optimal prediction system based on accurate, reliable, and consistent outcomes. The cluster analytic output produced concerns of customers, interesting areas and churn decision with details of targeted advertising and artifact development.

paper[11] provided an Introduction to statistical techniques that can be used to solve some commonly encountered problems in data analysis and predictive tasks. The authordiscussed the theoretical framework and advantages and limitations for shrinkage methods i.e., Ridge regression, Lasso Shrinkage and variable selection, the general penalties,and ElasticNet approaches. The critical point of these methods is how Selecting good values for λ parameter, usually done numerically via cross-validation

researcher in [12]his main intention is to present a different kind of Shrinkage methods and how to apply them in financial analyses and finance researches with reasonably a huge number of variables as matched to sample size.He focused on how to Penalize Ordinary least squares, Ridge, Lasso and ElasticNet generalization methods, after using root mean square errors as an evaluator of model performance with previously mentioned predictive methods, they founded that the performance of penalization methods exceeded the ordinary least square, with the ElasticNet being the best performing method.

3. Methodology Set Up

The proposed system for CCP in Telecom industry using machine learning analysis techniques is depicted in Figure 1. This figure illustrates the methods and algorithms that used in this research. The main goal is to introduce an accurate approach for customer churn classification in Telecommunication domain which consisted from two classifiers of CCP based on Ensemble machine Learning and investigate how it improved the performance of churn model to assist the CCP.

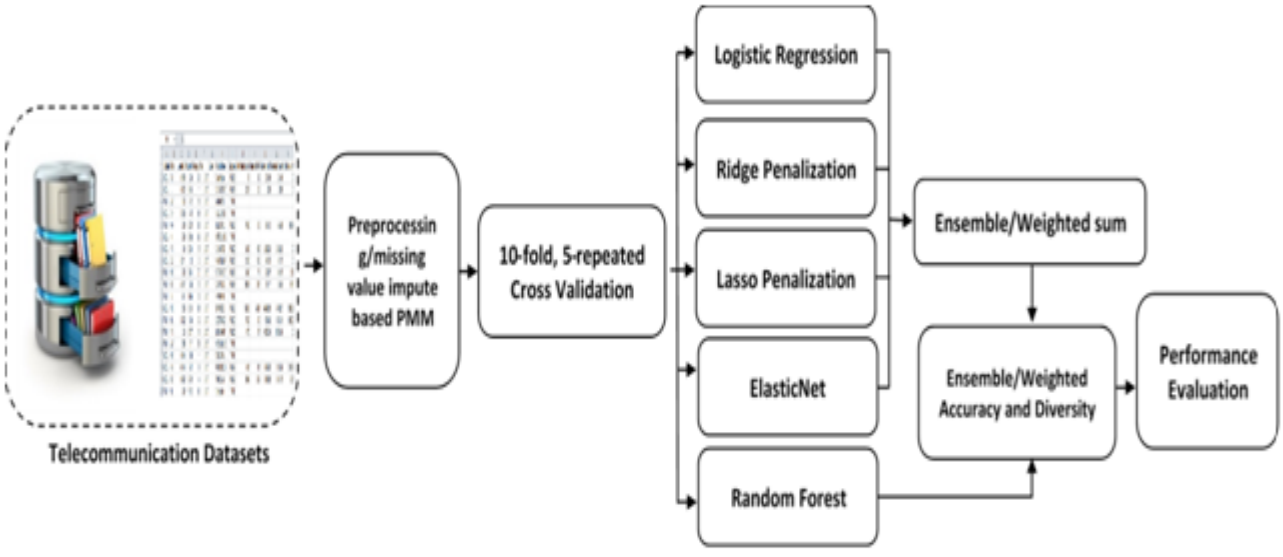

Fig. 1: Suggested Customer Chum Prediction Model

3.1. Datasets Preprocessing

The paper is performed on datasets provided by the Teradata Center at the University of Duke, involved the dataset of a major U.S. Telecom providers of 71047 observations and 78 features in a CSV file format. The prediction model based on whether a customer will churn during the period of 31-60 days, knowing that the actual monthly churn rate was stated to be approximately 3.45 %. The churn dependent is coded as a variable with y = 1 if customer churn, and y = 0otherwise. figure 2. Show the Churn percentage in used Dataset.

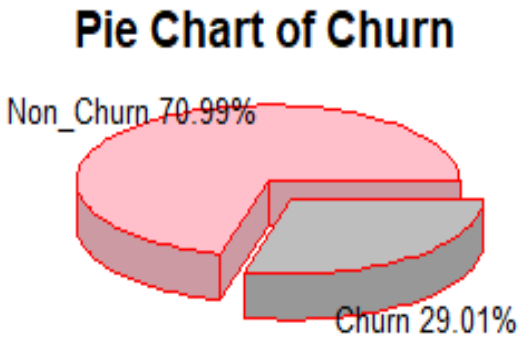

Fig. 2: Percentage of customer churn& non-churn in the dataset

Figure 3. Illustrates the percentage of missing values in the used dataset. Notice how variables like RECHARGE, REVENUE, MOU, MODELS, Phones, CHANGER, ROAM, DIRECTAS, AGE2,AGE1,and CHURNDEP have a very high number of missing values.

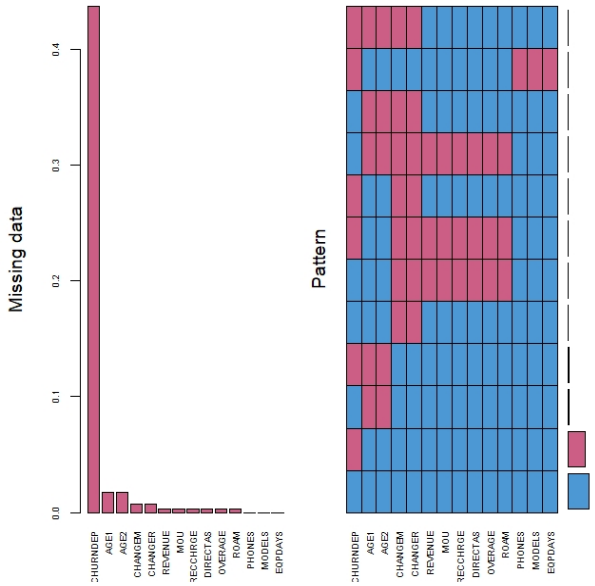

Fig. 3: Missing valuevariables inthe dataset

In this paper, Predictive Mean Matching (PMM) used to impute the feature with the missing value. Rather than removing features or observations with missing values, another approach is to fill in missing value variables. A variety of imputation approaches can be used that range from extremely simple to rather complex [6]. PMM is a well-known and widely used for creating variables imputation and is an attractive way to do multiple imputations for missing data, especially for filling up the quantitative variables which are suffering from irregular distribution. PMM can be implemented in two steps[13]. First, the predictive mean function is estimated. Second, the missing value of data are imputed by finding the similar records in the dataset, this done by using nearest-neighbor technique then, the observed outcome value of the nearest neighbor used for imputation. Because the predictive Mean function is estimated prior to matching, it absolutely encounters the missing values uncertainty due to the parameter estimation [14][6].

3.2. Predictive Mean Matching Algorithm (PMM)

1. For variables with no missing data, can be estimated using a Linear Regression model and producing a set of coefficients b
- Use the variable to impute as Y
- Use a set of good predictors as X
- Use only the observed values of X and Y to estimate the model.

2. Make a Random drawing from the posterior distribution of b coefficients and produce a new set of coefficients called b*.
- A random drawing from a multivariate normal distribution of set b and the predicted covariance matrix of set b to create a random variability in the imputed values

3. Calculate predicted values for observed and missing Y.
- Use the b coefficient for computing the predicted values of observed Y.
- Use b* coefficient for computing the predicted values of Missing Y.
- For each case where Y is missing data, calculate the closest predicted values among cases where Y is observed

4. Among close cases, randomly drawing one of those cases then impute the missing value in Y with the observed value of this close case.
5. In the case of multiple imputations, steps 2 through 5 are repeated several times for each completed dataset.
- Each repetition of steps 2-5 will create a new imputing data.
- In the case of multiple imputations, missing data are typically imputed 5 times.

PMM reflects the structure of the observed values at almost perfectly and heavily outperforming all the other imputation methods for missing data.

3.3. Generalized Linear Models (GLM)

The standard Linear Regression model is performed poorly in cases where there is a large scale ofmultivariate data that containing a number of variables superior to the number of samples. An alternative best methods, by extension the traditional linear models, called regularization approaches that gained popularity in statistical data analysis due to the flexibility of the model structure, unifying typical Regression methods i.e., Linear Regression and Logistic Regression, and the availability of model-fitting software with the ability to scale well with large datasets[15].

3.4. Regularized Regression Techniques

It is well known that Ordinary Least Square (OLS) flaws with regard to prediction and interpretation accuracy. Penalization is to shrink the coefficient values towards zero, introduced to improve the performance of OLS techniques by allowing the fewer contribution variables in predictive performance to have a coefficient approximately close to or equal zero value, the model building is a process to avoid overfitting, reduce variance of prediction error, and handle the correlated classifiers [17]. The main objective of the regularization technique is to make a trade-off between accuracy and simplicity, this means the model that hasthesmallest number of coefficients, results in a good accuracy.
The two most common shrinkage approaches are Ridge, Lasso, and ElasticNet Regression methods that combine both of penalty method. Shrinkage models induce sparsity in the solution and reduce the coefficients by imposing a penalty on their size. These properties are most beneficial because they reduce the variances in

the churn prediction system and make the model more interpretableby selecting asubset of the variables[12].
In order to get the best performance of GLM models, it is necessary to fit the optimal value of the penalization parameters α with λ. The GLM model can be fitted by finding the set of parameters (α, λ) that maximizes the likelihood of the data, and minimize the number of misclassification instances while minimizing the magnitude of the parameter values. This can be done by a performed heuristic search on the training datasets based on grid search over α, λ parameters. The λ parameter controls the amount of applied regularization by usually selectit in a way that the resulting model will minimize the error of sample if it has zero value, no regularization process is applied and α parameter is ignored [18].
The efficient and automatic search for the optimal value of the λ parameter is based on the cross-validation performance. When λ search is enabled, GLM first fitting the model with maximum regularization then, keep decreasing until overfitting occurred, the resulting model yield the best λ value. When grid search over the α parameter is required in case of the ElasticNet method then, the best value can be invoked by supplying a list of interval values between zero and one instead of just single value as in Ridge and Lasso methods [11].
Moreover, Resample methods are an indispensable tool in modern statistics and it is required in the implementation of Ridge, Lasso Regression and ElasticNet regularization methods. They involved repeatedly drawing instances from a training dataset and refitting the model of interesting on each sample in order to obtain additional information about the fitted model[17]. K-folded Cross-Validation (CV) technique is enabled after fitting models for a full regularization by randomly splits the observations set into K approximately equal-sized groups, called folds, trains each of the K folds on K-1sections and calculate validation criteria on the section that was not dedicated for training datasets. In more precise, the first fold is treated as a validation fold [19]. In research the Cross-Validation procedure is repeated five times; each time, a different set of observations is treated as a validation set. Repeated Cross-Validation can be used to calculate the optimal amount of penalizing parameter λ by computing the errors on the validation dataset of the fitted model that created by using the training datasets, then select the tuning parameter value for which the Cross-Validation Error is smallest. Finally, the model is refit using all of the available observations and the selected value of the tuning parameter.

3.5. Ridge, Lasso,and ElasticNet Regression

The ElasticNet parameterα∈ [0, 1] controls the penalty distribution betweenL1-norm and L2-norm.α parameter Controls the mix of Ridge and Lasso regularization with α = 0, the L1 penalty is not used and a Ridge Regression solution with shrunken coefficients is obtained. If α = 1, the Lasso operator threshold all parameters by reducing them with a constant factor and truncated at zero value [20].
Ridge regression is obtained by shrinkage of the Regression coefficients with a penalty term called L2-norm, the sum square of the coefficients. The penalty increases so the variables with a minor contribution to the model outcome have values toward zero. In Ridge regression, the penalty term is quantitybe minimized as in equation 1.

$$\hat{\beta}(ridge) = org\ min_{\beta}\|y - x\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad (1)$$

Where
β: regression coefficients
λ: penalize parameter
X: the corresponding predictive vector
Y: Independent observation vector
λ is usually selected in a way to minimize the sampling error. The optimal value is foundedby using grid search with cross-validation technique. As λ value increased, the shrinkage of the Ridge coefficient estimated leads to a substantial reduction in the variance of predictions, at the expensive of a slight increase in bias. Consequently, the mean square error (MSE) function of variance plus square bias, drops considerably as λ increases. While, in λ = 0, the variance became a high but with no bias.
Ridge regression involved high numerical stability and is easier and faster to calculate than Lasso, the model-fitting procedure can be achieved with quite quickly, it is perfect if all various predictors have non-zero coefficients and collect from a regular distribution.
One disadvantage of the Ridge Regression, it's included all the predictors in the final model, Lasso Regression is an alternative model to overcome this drawback[11].
Lasso method, with the abbreviation of Least Absolute Shrinkage and Selection operators, it's penalized some of the Regression Coefficients to zero value by shrinking the Regression model using a penalty term called L1-norm, the sum of the absolute coefficients. The penalty term has the ability to impose many coefficients to exactly have zero value, with a small subset of non-zero coefficients, this means that lasso regression works as embedded attributes selector method that picks out the most important coefficients.
The lasso estimator uses L1-norm to penalize the least square basis to obtain sparse values as depicted in equation 2.

$$\hat{\beta}(lasso) = org\ min_{\beta}\|y - x\beta\|_2^2 + \lambda\|\beta\|_1 \quad (2)$$

The advantage of using Lasso over Ridge penalized methods, its ability to produces simpler and interpretable model that integrate a reduced set of predictors since it is robust to the highest correlations among the classifiers, and performs well when some predictors have large coefficients, while the remaining predictors have very small coefficients[16].
As a result, models that built based on Lasso are generally much easier to interpret than those computed by Ridge Regression method.
To avoid imbalance coefficients in the Lasso output vectors, the ElasticNet is projected for evaluating the high dimensionality of datasets. It yields a Regression model that is generalized by incorporating the L1-norm and L2-norm, this means that ElasticNet has improved prediction model by effectively shrinking coefficients similar as Ridge estimator and set some coefficients to zero value as in Lasso estimator. The Penalization defined in equation 3.

$$\hat{\beta}(elasticnet) = \left(1 + \frac{\lambda_2}{n}\right)\{org\ min_{\beta}\|y - x\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\} \quad (3)$$

Where:
n: number of observation
α: penalize parameter
Here the penalties parameters are α and λ arguments, α controls the ElasticNet penalty coefficients distribution between L1and L2 norms, theλ argument controls the penalty strength up until overfitting obtained [17].

3.6. Logistic Regression on Binomial Family

In particular, Linear Regression is a useful and a very simple tool for predicting a quantitative response[2]. The Logistic Regression model is widely used when the dependent variables are categorical, If there are two response outcomes, it is possible to use the binomial distribution, otherwise, use the multinomial [21]. The Logistic Regression modeled the probability of observation that belonging to an output category for a given data, Pr (y= 1/x)[22]. The Canonical link for the binomial family is the "logit" function. Its inverse gives the logistic function, which takes any

real number and projects it onto the desired value of [0; 1] interval to model the probability of fitting to a level. For the binomial model, assume the response variable takes value in range G = {0, 1}, then the model can be written as in the equation 4.

$$\log\left(\frac{pr(G=1|X=x)}{\Pr(G=0|X=x)}\right) = \beta^0 + \beta^T x \quad (4)$$

Logistic regression representations the probability that response variables Y belong to a specific category. In the other words, the goal is to obtain coefficient estimates that the linear model belongs to the available data well[16].

3.7. Random Forest (RF)

Ensemble classifier is a type of Supervised Learning technique uses multiple decision trees to make a prediction. The fundamental idea is to create multiple and different structures on a training dataset then simply aggregating their results to obtain accurate model performance,nooverfitting,and balancing the Bias-Variance Trade-off as compared with the individual classifiers [8]. Churn prediction based on single classifier only might be regarded as a complex Model, due to a great variance which yields overfitting or might be excessively simple with a high bias which yields underfitting. RF algorithm is a popular Ensembling Machine Learning technique developed to support the classification and regression[22]. It extends the simple idea of classification based on single Decision Tree by building for each training dataset a particular Tree, each Tree is grown on a bootstrap sample of the training dataset by using randomly attributes selection at each node. Random forests provide an enhancement over the bagged tree, very unstable, by building a forest of number of decision trees on different bootstrapped training subsets with a way of a random small pinch will leads to generate diverse classifiers and allows to reduce correlation among the trees, hence increasing the overall model performance by reducing the variance among trees using ensembling technique [5]. To classify an observation, each tree in the forest creates its response, the prediction model chooses the cases that have received the maximum weights "votes" over all the trees in the forest[4].

3.8. Random Forest Algorithm(RF) [23]

1. Let N represents the number of training sets or trees in model suggestion, Building trees until the error no longer decreases. Let M represents the number of variables in the classifier.
2. The number m of input variables is used to decide the decision at a node of the tree; m should be much less than M.
3. Choose a training set for this tree by choosing n times with replacement from all N available training sets. Use the rest cases to estimate the error of the tree.
4. At each node in the Decision Tree, randomly choose m variables based on resampling method. Calculation of the best split based on m variables in the training set using purity measures i.e., square error in regression, in classification using Gini index, deviance, and entropy.
5. Each tree is fully grown and not pruning.
6. To create a prediction, a new sample is pushed down the tree. It is assigned the label in the terminal node.
7. Steps 2 to 6 are repeatedly over all trees in the Ensemble Tree. Finally, a majority vote of all trees is reported as random forest prediction.

RF technique is able to deliver a consistently high performance, is very robust and has a reasonable computing time. The only parameter needsto adjust is the number of variables that are available for splitting at every node. On the other hand, for datasets including categorical variables with a different number of levels, RF is imposed to bias in favor of those attributes with more levels. Therefore, variable importance scores from RF are not reliable for this kind of data[24].

3.9. Ensemble Machine Learning

Ensemble-based methods have been among the most influential method on Data Mining and become very popular techniques due to their ability to deliver accurate results with the possibility of splitting the classifier into independent prediction models [25]. They are thought as a set of algorithms that are built by combining the results from different Machine learning algorithms of positive or negative type called "Base Learner Components", to boost the accuracy for a real Machine learning challenge and make the final prediction decision more robust, it incorporates the voting clue from each particular base classifier. The central issues of Ensemble learning are how much should each prediction models contribute to making the final prediction. Boosting is a sequential technique and an example of Ensemble learning that manipulates the datasets by applying different weights to classifiers [6]. Then, taking the weighted average which means giving a great or small importance to specific classifier outcome. Weights regarded as a tuning parameter,acriticalcomponent on boosting algorithm to make them able to avoid overfitting problem [26].

This paper proposed Ensemble Classifier based regularization methods. Prediction values of each regularization algorithm are used as inputs to the classifier to predict the actual outcome, randomly generate values between [0, 1] become the weights on the individual algorithm. The equation of Ensemble learner that involved the weights calculation of each prediction from each individual algorithm depicted in equation 5.

$$\Pr(Y) = 0.0294D_{logistic} + 0.4412D_{Ridge} + 0.1176D_{lasso} + 0.4118D_{elastic} \quad (5)$$

Where:
Pr(Y): prediction of the Ensemble learner basedRegularization methods.
D: predicted values from each Regularization methods.
Then Ensemble learner builds to incorporate the previously mentioned Ensemble based regularization algorithms andpredicted values based Random Forest algorithm as shown in equation 6.

$$\text{pr(Y)}_{Ensemble} = 0.2712\, pr(Y) + 0.7287 pr(RF) \ldots (6)$$

The Weights for Ensemble learner are calculated based on Weighted Accuracy and Diversity (WAD) algorithm that invoked to estimate the weights for the Random Forest classifier and Ensemble based Regularization learner.

3.10. Weighted Accuracy and Diversity Algorithm (WAD)

1. The output of the algorithm is to compute the weights of the Ensemble classifier
2. Compute accuracy (Acc) for the learner Suppose the Confusion matrix for two classifiers. (P) Denotes the positive prediction and (N) denotes negative prediction given by the classifier.

Confusion matrix

	P	**N**
P	TP	FP
N	FN	TN

Then accuracy is in equation 7.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

3. Compute diversity (Div) for the classifier using equation 8.

$$Div_{ij} = \frac{TN + FP}{TP + TN + FP + FN} \quad (8)$$

4. Estimateα and β values randomly correspondingly to the condition $\alpha + \beta = 1$
5. Compute Weighted Accuracy and Diversity by equation 9.

$$\text{WAD} = \frac{\text{Acc} * \text{Div}}{\beta\text{Acc} + \alpha\text{Div}} \quad (9)$$

The authors ofthe paper[27] used WAD as a criterion for evaluating the quality of the classifier ensemble and assisting in the best Ensemble selection methods.

This research, employed the WAD as a novel algorithm to calculate and find the best weights for proposed Ensemble Machine Learning algorithm.

4. Results & Discussion

The final step of the CCP model is to evaluate the predictive performance and analyzing it by using several assessment metrics. In the predictionsystem, the optimal values of α and λ parameters and the number of Cross-Validation folds for penalization methods are depicted in table 1.

Table 1: Optimal values of penalization parameter

Method \ Parameter	A	Λ	#fold
Ridge	0	0.025	3
Lasso	1	0.0001	1
ElasticNet	0.1111	0.0038	7

Figure 4. Show penalization parameters and RMSE for estimating the values of α and λ parameters on shrinkage methods based five times repeated Cross-Validation.

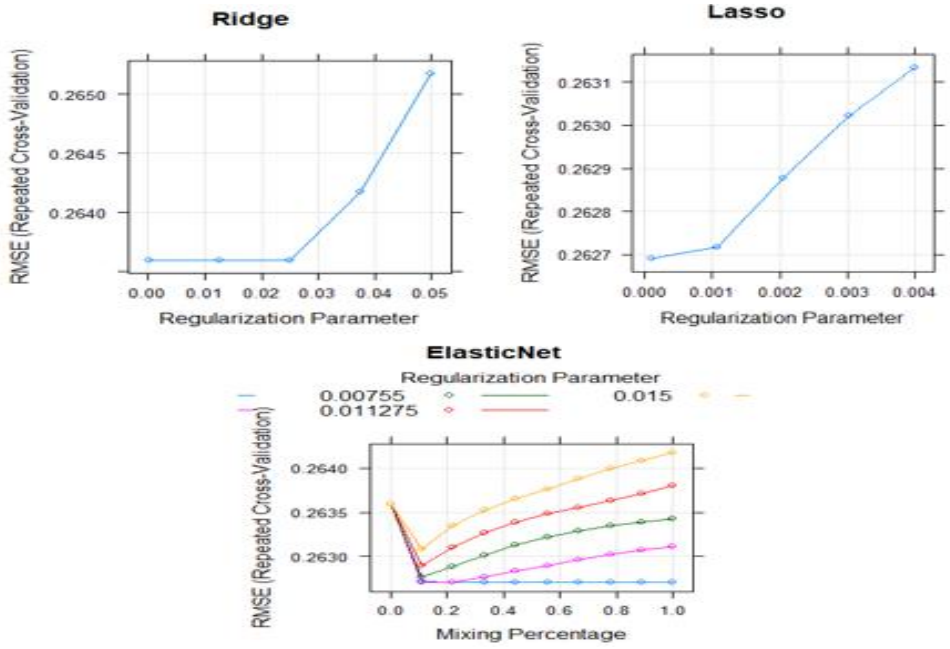

Fig. 4: Regularization Parameters vie Cross-Validation

The outcomes of churn prediction models can be compared to whether they actually churned in order to evaluate predictionaccuracy based on their performances in Cross-Validation. The main methods for evaluating the suggested churn predictionmodels based Ensemble learning is used on the tested Datasets to be able to compare them are precision, accuracy,F-Score, and AUC. Table 2. Shows the application of these evaluators inthe suggested churn prediction system.

4.1. Accuracy

The overall accuracy of predictions is the most common way to evaluate machine learning methods. In prediction classification, accuracy is defined in Equation 7 it is the sum of the number of true positive prediction and true negative predictions divided by the total amount of predictions.

4.2. Error Rate

The number of all positive predictions divided by the total number of datasets[18].

4.3. Precision

In the classification task where an action is only performed on positive predictions. In the case of customer churn, for example, for a negative prediction, one might not want to perform any action since this customer is predicted to stay with the company under the current circumstances[4]. On the other hand, for a positive prediction, some action needs to be taken, otherwise the customer is likely to leave. Precision is the rate of true positive (TP) predictions to total positive predictions (P).

4.5. F-Score

Precision is useful for evaluating binary classifiers, but it leaves out some information and thus can be misleading. The recall is a fraction of the true positive predictions (TP) to total positive observations (P) in the dataset it measures the fraction of the churn rate that correctly classified[5]. The classifier that has a low recall means it miss-classifies a large proportion of the positive cases. The F-Score is defined as in equation 10.

$$F - Score = 2.\frac{Precision * Recall}{Precision + Recall} \quad (10)$$

F-score is the harmonic mean of both precision and recall.Therefore, be a better evaluator for binary classifiers than precision.

4.6. Area Under Curve AUC

A Comparative Assessment of the performance of Ensemble learning in the Customer Churn Prediction [5]

4.7. Gini Coefficient

Gini coefficient is measure closely related to the AUC, which is equal to twice the area between the ROC curve and the diagonal, i.e., Gini =2*AUC-1. The Gini coefficient varies between zero, ROC curve lies on the diagonal and the model does not perform better than a random classification model, and one, maximum ROC curve and perfect classification.

Table 2: The results of applying evaluators on the binary classification system

Values / Method	Acc	MSE	Rsquared	Error rate	AUC	Gini	F-Score
Logistic Regression	0.8465	0.2661	0.6577	0.1534	0.98401	0.96802	0.8794
Ridge	0.8531	0.2668	0.6578	0.1468	0.98417	0.96834	0.8853
Lasso	0.8443	0.2660	0.6579	0.1556	0.98402	0.96804	0.8775
ElasticNet	0.8410	0.2659	0.6581	0.1589	0.98405	0.9681	0.8746
Ensemble Regularization	0.8473	0.2661	0.6580	0.1526	0.98411	0.96822	0.8802
Random forest	0.9901	0.0968	0.9547	0.0099	0.9918	0.9836	0.9930
Ensemble Learning	**0.9907**	0.1190	0.9372	**0.00957**	0.98448	0.9689	**0.9932**

4.8. Input Selection

Usually using a limited number of highly predictive features is favored to be included in a customer churn classification model, in order to improve the comprehensibility. Therefore a procedure can be applied in order to select the most predictive attributes and to eliminate redundant attributes and filtered by firstly, using feature selection based correlation between all features with churn factor. Secondly, applying the important of variables procedure by Generalization linear model and Random Forest since the most predictive penalized methods i.e., Lasso and ElasticNet have the abilities to perform prediction as well as make some features which are less contribution for prediction model toward zero value. Figure 5illustrates the selection procedure to compare the methods by plotting theclassifiers with a number of included attributes using boxplot instruction.The number of attributes needed by prediction methods seemed to be strongly dependent on the datasets. On average, these techniques only need a small number of attributes around 21 or 22 variables with a threshold equal to zero are sufficed to yield an effective and powerful CCP model.

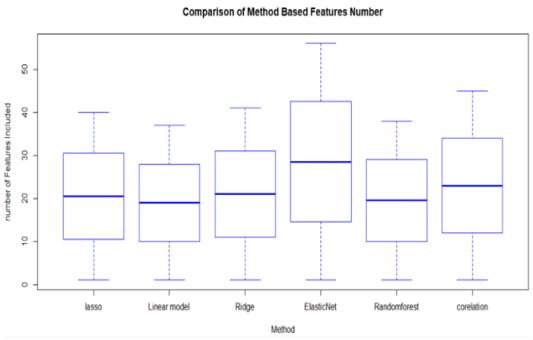

Fig. 5: Comparison of prediction methods based feature selection

The feature selection based Correlation coefficient puts the features in an ordinal list by general features like correlation with the variable to predict or variance in them. The ranked features then provide a list to make a decision of keeping or removing features based on ranks.Pearson correlationmeasures a linear dependence between two variables (x and y). It's also known as a parametric correlation test because it depends on the distribution of data. It can be used only when x and y are from a normal distribution. Figure 6 illustrates the Pearson correlation between all features of the dataset and Churn variable.

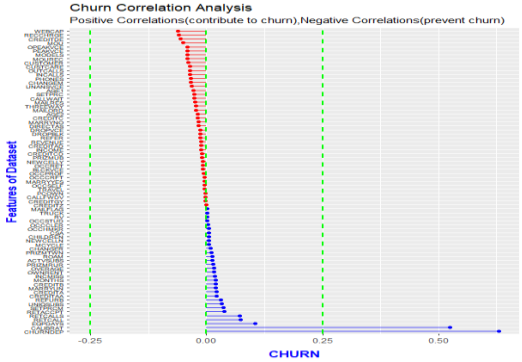

Fig. 6: Pearson Correlation

The most important algorithms that employed for Input selectionsof importance variables are Lasso and ElasticNet penalization tasks. With the large scale datasets in telecom sectors, variables selection process became necessary process to be applied in order to make the churn model easier to interpret,reduce overfitting, removing features that are redundant and do not add any information, and to reduce the size of the prediction task and enable classification algorithms to yield outcomes in faster manner as possible.

This research,has been proved that employing the PMM algorithm for predicting missing values in the dataset is essential for the purpose of increasing the efficiency of the churn Prediction model, as this algorithmis able to guess variables that have a Negative correlation with the response variable when compared with the most relevant attributesto prevent Churn from accurse such as,RECHARGE, PHONE, MODELS,CHANGE, AGE1,AGE2,…, MOU,that estimated based on aforementioned PMM algorithm.

5. Conclusion

The customer churn prediction model is one of the most important tasks for any Telecommunication company, because of the financial penalty associated with churn issue and the high cost associated with attracting new customers. Avital parts of churn model generation are data preprocessing and a split procedure of system dataset into training and test data are required. Within the training data, which are used for building the models, the churn ratio is found for the churning customers. The testing data remains unaltered for evaluation of the suggested CCP framework.Predictive mean matching algorithms work well for missing values imputation on continuous and categorical (binary & multi-level) variables. The common routine tocompare classification algorithms is to perform k-fold cross-validation experiments to estimate the accuracy of these algorithms. It has been shown that comparing algorithms using cross-validation experiments results in an increased in the churn prediction model accuracy and reduced the mean square errors. Furthermore, Regularization methods as a technique in customer churn prediction show that the Ridge Regressionoften outperforms the penalization methods. The worst case of them is the ElasticNet Regression model, which achieved an accuracy of 0.84 but, had the highest MSE value of 0.2659. The highest precision value was achieved by Ridge model, which attained an accuracy of 0.8531, with MSE of 0.2668. The ensemble model achievedan accuracy value of 0.8473 with an MSE of 0.2661.

Random forest is one of most available Machine learning algorithms that produces a highly accurate prediction classifier with 0.9901as accuracy and 0.096 for MSE evaluator.

Actually, to achieve the goal of maximizing the prediction accuracy, it is not sufficient to only use a single prediction model, at least two models should be utilized. Ensemble-based systems have proven to be the most efficient way to construct a high predictive model with an accuracy of 0.9907 and MSE of 0.119. The added advantage is due to the use of multiple algorithms to generate better predictive performance than could be obtained using a single model-based system.

Acknowledgment

I am grateful especially to the supervisor Dr. Eman Alshamery for her time and constructive remarks that improved the quality of the paper. I would also like to thank Dr. W. Verbekefor his great help.

References

[1] J. Donald, “Predicting Attrition in Financial Data with Machine Learning Algorithms,” 2018.

[2] M. K. Sahu, R. Pandey, and S. Silakari, “ISSN NO : 0076-5131 Analysis of Customer Churn Prediction in Telecom Sector Using Logistic Regression and Decision Tree Keywords :” J. Appl. Sci. Comput., vol. 5, no. 6, pp. 62–67, 2018.

[3] P. K. Nyambane, “CHURN PREDICTION IN TELECOMMUNICATION INDUSTRY IN KENYA USING DECISION TREE,” 2017.

[4] G. C. Esteves, “Churn Prediction in the Telecom Business,” p. 96, 2016.

[5] W. Verbeke, “Profit-driven data mining in massive customer networks: new insights and algorithms,” no. 379, 2012.

[6] 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, 2014.

[7] H. Abbasimehr, M. Sestak, 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.

[8] G. Louppe, “Understanding Random Forests: From Theory to Practice,” 2014.

[9] K. Bailey, J. Miller, and Valerie Santiago-Gonzalez, “predicting diabetes diagnosis in African Americans using Ensemble machine learning.”

[10] I. Stephen Nabareseh, “Predictive analytics: a data mining technique in customer churn management for decision making Prediktivní analytika: technika data miningu pro rozhodování s využitím v řízení odchodu zákazníků,” no. February, 2017.

[11] F. Andreis, “Shrinkage methods (ridge, lasso, elastic nets),” no. November 2017.

[12] J. Vorlíčková, “Least Absolute Shrinkage and Selection Operator Method,” 2017.

[13] A. Agarwal, G. Verma, H. B. Sri, K. Mannem, and F. Hamid, “Indian Institute of Technology, Kanpur Department of Industrial and Management Engineering IME672A Data Mining and Knowledge Discovery Course Project Report,” 2016.

[14] G. Vink, G. Laserdisc, and S. Van Buuren, “Partitioned predictive mean matching as a multilevel imputation technique,” Psychol. Test Assess. Model. vol. 5, no. 4, pp. 1–16, 2015.

[15] P. Allison, “Imputation by Predictive Mean Matching: Promise & Peril,” http://statisticalhorizons.com/, 2015. [Online]. Available: http://statisticalhorizons.com/predictive-mean-matching.

[16] A. J. van der Koij, “Regularization with ridge penalties, the lasso, and the elastic net for regression with optimal scaling transformations,” Predict. Accuracy Stab. Regrets. With Optim. Scaling Transform. no. 2006, pp. 65–90, 2007.

[17] J. Lanford, T. Nykodym, A. Rao, and A. Wang, Generalized Linear Modeling with H2O’s R Package. 2015.

[18] S. Dardouri and R. Bouallegue, “Performance Analysis of Regularized Linear Regression Models For Oxazolines and Oxazoles Derivatives Descriptor Dataset,” vol. 1, no. 4, pp. 111–123, 2013.

[19] D. Dalpiaz, R for Statistical Learning. 2017.

[20] Art Owen, “Regularization : Ridge Regression and the LASSO the Bias-Variance Tradeoff,” 2007.

[21] E. Krona, “A simulation study of model fitting to high dimensional data using penalized logistic regression Mathematica institution,” Stockholm University.

[22] D. S. De Groot, “Churn prediction in telecommunication Classification problem,” 2017.

[23] A. Lemmens and C. Croux, “Bagging and Boosting Classification Trees to Predict Churn,” J. Mark. Res., vol. 43, no. 2, pp. 276–286, 2006.

[24] J. Van Haver, “Benchmarking analytical techniques for churn modeling in a B2B context,” 2016.

[25] C. Zhang and Yunqian Ma Editors, Ensemble Machine Learning. 2012.

[26] J. Vijaya and E. Sivasankar, “Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in the telecommunication sector,” Computing, vol. 100, no. 8, pp. 839–860, 2018.

[27] X. Zeng, D. F. Wong, and L. S. Chao, “Constructing better classifier ensemble based on weighted accuracy and diversity measure,” Sci. World J., vol. 2014, 2014.

[28] M. Ewing, “Teknisk-naturvetenskaplig fakultet UTH-enheten”, 2012.