Customer Churn Prediction based on Ensemble Machine Learning

Sridhanuja.R [1], Sri hari.K V [2]

1 II year M.Sc., Decision and Computing Sciences ,Coimbatore Institute of Technology.

2 II year M.Sc., Decision and Computing Sciences ,Coimbatore Institute of Technology.

Abstract

Customer Churn Prediction (CCP) is a challenging experiments for decision makers. CCP aims to detect customers with a high affinity to leave. The objective of this model is to handle a large scale Telecommunication Company and identify potential churn. In the proposed research, instead of removing features or observations with high missing data, Mean Imputation, Data wrangling is used to handle missing values. First Ensemble Machine learning classifiers used to scrutinize 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 model used the Weighted Accuracy and Diversity (WAD) as an algorithm to find the optimal weights for the proposed Ensemble classifier. The second Ensemble learner is incorporated of penalized methods (Ridge, Lasso and ElasticNet) with a Logistic Regression method on the binomial family as based on the generalized linear model. Randomly generate values between [0, 1] became the weights for this classifier. The Weights are selected according to the principle that weights of higher value are assigned for great performance classifier to ensure the highest accuracy of Churn Prediction model. To achieve efficient and automatic search for the optimal value of lambda parameter for penalization methods , 10-fold based on five times repeated Cross-Validation (CV) performance technique should be used. These two Ensemble classifiers incorporated within a churn prediction model to handle a dataset, an imbalance data distribution, time-dependent features label in the Telecommunication industry. Experimental results show an increase in predictive performance. In addition, the results depicted that using of ensemble learning has brought a remarkable improvement for individual base learners in terms of performance indicators such as Area under Curve (AUC), specificity, sensitivity Mean Square Error(MSE), and Accuracy. Accuracy is the best candidates for churn prediction tasks.

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

INTRODUCTION

Customer churn is the percentage of customers that stopped using company's product or service during a certain time frame. All businesses in the consumer market and enterprise sectors have to deal with churn as it could end up affecting the company revenue numbers and thereby influence policy decisions . 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 ensure customer's satisfaction. The companies which successfully apply CRM to their business nearly always improve their retention power which represents the probability that a customer will not leave. Customer churn prediction in Telecommunication companies has become an increasingly well known research issue in recent years and therefore, Telecom providers using widely strategies to identify the potential churn customers based on past information, prior behaviors and offering some services to persuade them to stay.

CUSTOMER CHURN PREDICTION MODEL (CCP)

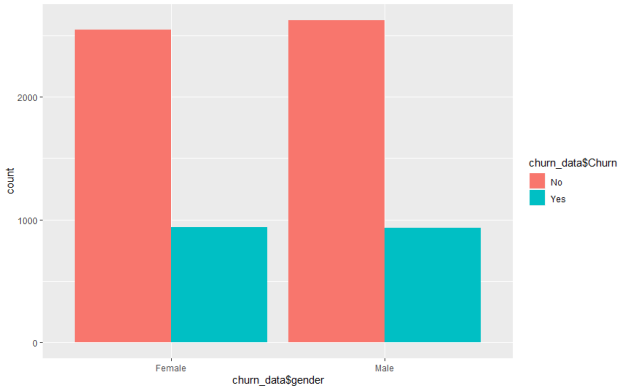
Predictive churn model gives you a quantifiable metrics and awareness to fight against in your retention efforts. This gives the ability to pattern habits of customers who leave, and step in before they make that decision. CCP model is used to predict each customer's likelihood of stopping usage of services and analyze customer data and develop customer focused retention plans. By knowing which customers are of high churn risk, can act to proactively retain those customers. Without a strong understanding of customers and their behaviors, it's hard to retain them so the first step in creating this model is understanding your customer behavior from customer data points.

DATASET

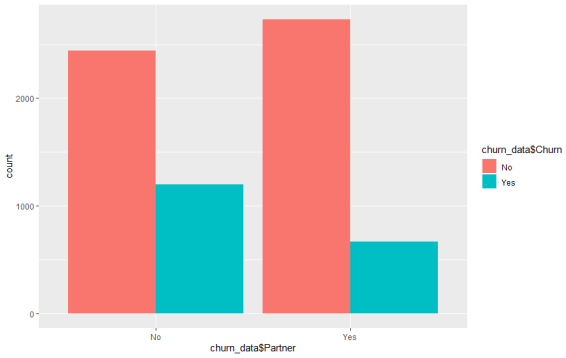
Customer Information

- Customer ID
- Gender
- Senior citizen : whether the customer is Senior citizen or not?(0,1)
- Partner: whether the customer has a partner or not (Yes, No),
- Dependents: whether the customer has dependents or not (Yes, No),
- Online Backup: whether the customer has an online backup or not (Yes, No, No internet service)
- Tenure: number of months the customer has stayed with the company,
- Monthly Charges: the amount charged to the customer monthly,
- Total Charges: the total amount charged to the customer.
- Churn : whether the customer is churned or not? (yes or no)

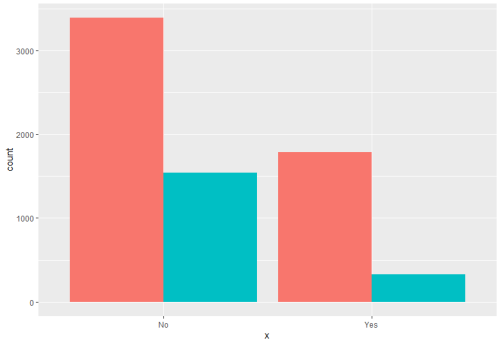
DATA VISUALIZATION



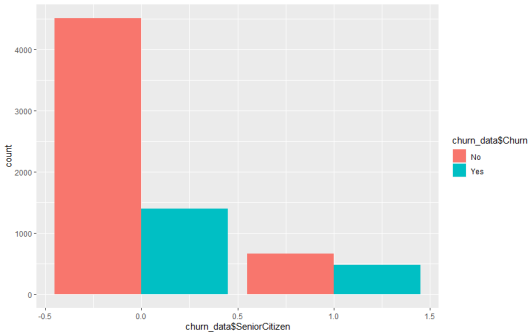
Gender vs churn



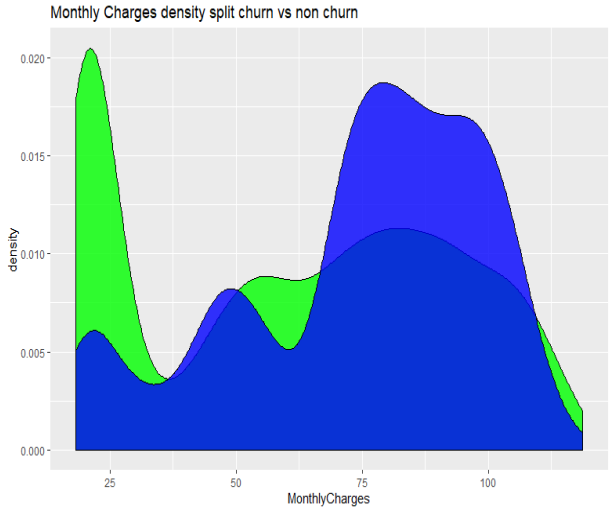
Partner vs churn

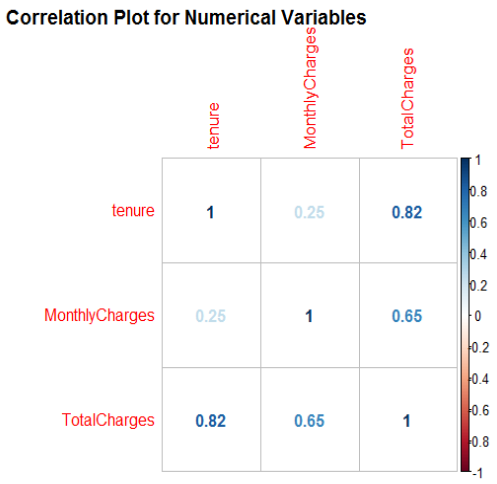


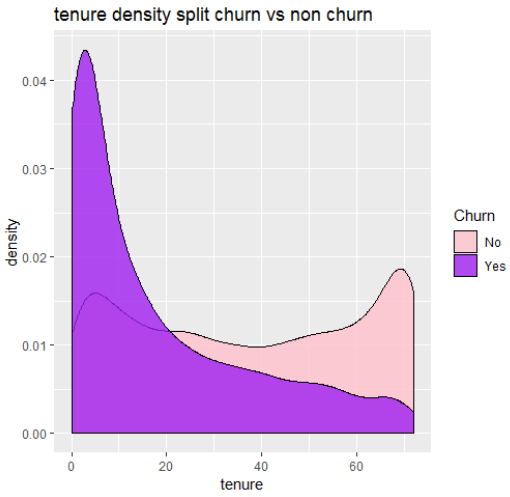
Senior citizens vs churn

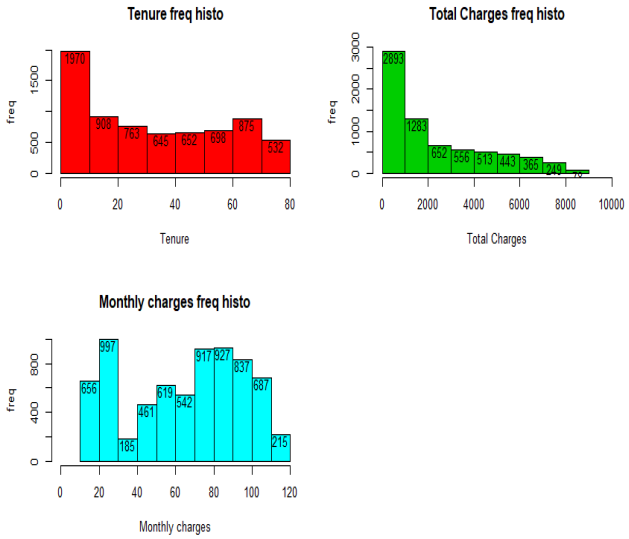


Dependents vs churn









Tenure density vs churn

Histogram for Tenure , Total charges and monthly charges

DATA PREPOCESSING

As the data contains many missing values which will affect the result if the missing values are taken as null values. If the rows containing missing values are eliminated it will reduce the accuracy percentage therefore some measures has to be taken to handle missing values like mean imputation. Some more data preprocessing techniques like Data Wrangling, Eliminating the fields which are highly correlated are also done to clean the data.

- **Mean imputation** is a **method** in which the missing value on a certain variable is replaced by the **mean** of the available cases. This **method** maintains the sample size and is easy to use
- **Data wrangling**, sometimes referred to as **data munging**, is the process of transforming and mapping **data** from one "raw" **data** form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
- When variables as super highly correlated, they can introduce instability in the errors on the betas and are, in fact, measuring the same thing, roughly speaking. So,we can eliminate the variable.

RANDOM FOREST

Ensemble classifier is a type of Supervised Learning technique which 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, no over fitting and balancing the Bias-Variance Trade-off as compared with the individual classifiers . Churn prediction based on single classifier only might be regarded as a complex Model, due to a great variance which yields over fitting or might be excessively simple

with a high bias which yields under fitting. RF algorithm is a popular Ensemble Machine Learning technique developed to support the classification and regression.

LOGISTIC REGRESSION

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. The Logistic Regression modeled the probability of observation that belonging to an output category for a given data, Pr (y= 1/x). 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. 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.

GENERALIZED LINEAR MODEL (GLM)

The standard Linear Regression model is performed poorly in cases where there is a large scale of multivariate 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.

RIDGE, LASSO ANDELASTICNET REGRESSION

The ElasticNet parameter $\alpha \in [0, 1]$ controls the penalty distribution betweenL1-norm and L2-norm.α parameter Controls the mix of Ridge and Lasso regularization with $\alpha = 0$, the L1 penalty is not used and a Ridge Regression solution with shrunken coefficients is obtained. If $\alpha = 1$, 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.

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 . Then, taking the weighted

average which means giving a great or small importance to specific classifier outcome. Weights regarded as a tuning parameter, a critical component on boosting algorithm to make them able to avoid over fitting problem . 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.

WEIGHTED ACCURACY AND DIVERSITY

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.

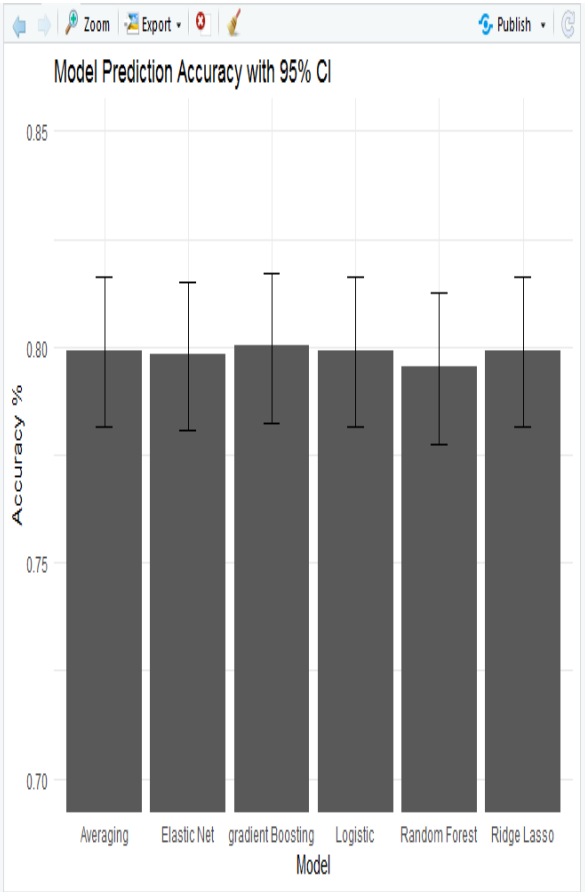
(P) Denotes the positive prediction and (N) denotes negative prediction given by the classifier.

ACCURACY

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

RESULTS

```
149:21  Ridge and lasso Penalization                    R Script
Console ~/

    intersect, setdiff, setequal, union


> accuracysummary <- bind_rows(Logistic = lraccuracy, RandomForest = rf_accuracy,El
asticNet = glmnetaccuracy, Ridge_Lasso = glm_accuracy , Gradient_Bo .... [TRUNCATED
]

> library(tibble)

> accuracysummary2 <- add_column(accuracysummary, "Model" = models, .before = "Accu
racy")

> accuracysummary2
# A tibble: 6 x 4
  Model            Accuracy AccuracyLower AccuracyUpper
  <chr>               <dbl>         <dbl>         <dbl>
1 Logistic            0.799         0.782         0.816
2 Random Forest       0.795         0.778         0.812
3 Elastic Net         0.798         0.781         0.815
4 Ridge Lasso         0.799         0.782         0.816
5 gradient Boosting   0.800         0.782         0.817
6 Averaging           0.799         0.782         0.816

> library(ggthemes)

> library(ggplot2)

> ggplot(accuracysummary2, aes(x = Model, y = Accuracy)) + geom_bar(stat = "identit
y") +
```



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. A vital 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 to compare 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. Random forest is one of most available Machine learning algorithms that produces a highly accurate prediction classifier with 0.95 as accuracy and 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 0.95 .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.