مجلة الشروق للعلوم التجارية
ISSN: 1687/8523
Online :2682-356X
2007/12870
sjcs@sha.edu.eg
https://sjcs.sha.edu.eg/index.php : موقع المجلة
الترقيم الدولي
ترقيم دولي الكتروني
رقم الايداع بدار الكتب المصرية
البريد الالكتروني
SH. A.
إكاديمية الشروق
عضو اتحاد الجامعات العربية
المعهد العالي للحاسبات وتكنولوجيا المعلومات

"Proposed Integrated Customer Churn Predication Model using Several Statistical Feature Scaling and Machine Learning Algorithms"

أستاذ دكتور/ زهدي محمد نوفل
أستاذ بقسم الإحصاء و الرياضة و التأمين
كلية التجارة - جامعة بنها
dr.zohdynofal@fcom.bu.edu.eg

الدكتور/ إيمان محمود
مدرس بقسم الإحصاء و الرياضة و التأمين
كلية التجارة - جامعة بنها
eman.abdelghani@fcom.bu.edu.eg

نهي نبوي
مدرس مساعد بقسم الإحصاء والرياضة والتأمين
كلية تجارة - جامعة بنها
noha.bahi@fcom.bu.edu.eg

Keywords:
Customer churn, Exploratory Data Analytics, Feature Scaling technique, Classification algorithms and Clustering algorithms.

التوثيق المقترح وفقا لنظام APA :

Nofal, Zohdy Mohamed, Mahmoud, Eman, Nabwy, Noha (2024), "Proposed Integrated Customer Churn Predication Model using Several Statistical Feature Scaling and Machine Learning Algorithms", Al-Shorouk Journal of Commercial Sciences, special volume, The Higher institute for Computers and Information Technology, Al-Shorouk Academy, page 621 – 638

"Proposed Integrated Customer Churn Predication Model using Several Statistical Feature Scaling and Machine Learning Algorithms"

Abstract

Data science is crucial for analytics and prediction in the telecommunication industry. Customer churn prediction is becoming progressively important. While machine learning methods are regularly utilized for predicting churn, their performance can be improved due to the complexity of consumer data structures. Managers lose trust when findings are difficult to interpret .This study utilizes data preprocessing techniques. The various elements of benchmarked data collecting can impact interpretability since imbalanced and feature scaling issues. Therefore, this study develops customer churn prediction model for those complexity issues. After training the model, the operator analyzes the data to understand its performance. To maximize interpretability, consumers are clustered based on behavioral factors. Clustering is grouping data points with similar features to maximize similarity between members. Additionally, they share few similarities with members of other groups. Using homogeneous group members improves classification algorithm prediction performance. Various algorithms, including logistic regression, support vector machine, random forest, Ada-boost, and multilayer perceptron, were tested before and after hyperparameter adjustment to achieve optimal prediction performance.

Keywords: Customer churn, Exploratory Data Analytics, Feature Scaling technique, Classification algorithms and Clustering algorithms.

1. Introduction

Progressively increasing number of telecommunication users, corporations are now offering a variety of services to maintain customers. Customer churn can happen for several different reasons. The most significant of these are call or package rates that are unsuitable for the customer (Tiwari, Sam, & Shaikh, 2017; Petkovski et al., 2016). It occurs when a client moves service providers to acquire better services and advantages. When a customer transfers from one service provider to another, the company's revenue suffers. In the telecom industry an enormous volume of data with missing values is generated.

To avoid this issue, the operator must understand the reason for the customer's choice of changing to another telecom company. Prediction remains the most effective method for analyzing churn behavior. Because of the enormous amount and challenges nature of the data set, predicting customer attrition in the telecommunications business has traditionally been a difficult challenge. Attrition prediction is used to discover which consumers are most likely to churn. Churn prediction and analysis can assist a company in improving the sustainability of their customer satisfaction strategy.

2. Literature review

Data pruning and cleansing are done during the pre-processing stage. Based on previous behaviors and historical customer data, it is possible to identify users or customers who are subject to churning. The Synthetic Minority Over-Sampling Technique (SMOTE) is used to optimize the model **(Ahmad, Jafar, & Aljoumaa, 2019)**. However, local optimal solution problems in feature selection strategies of this kind require large-scale data feature extraction with high accuracy in feature classification. In general, the dataset distribution for churn prediction is skewed, counting many cases in a single class relative to other classes. The class with more instances is called the majority class, and the class with fewer samples is called the minority class. The distribution of imbalanced instances within the dataset is shown by the imbalance ratio between classes. The number of non-churners in the churn-prediction

model was more than the number of churners **(Dwiyanti, & Ardiyanti, 2016)**. Pre-processing is used to balance the imbalanced dataset, and RUS (Random Under sampling) and SMOTE sampling techniques are used to extract the features **(Gui, 2017)**. There are two types of methodologies commonly used in dealing with unbalanced data inside the churn prediction model: resampling approaches and cost-sensitive learning techniques. During the model-training phase, cost-sensitive learning approaches alter the relative mistake costs. Resampling data performs effectively in managing unbalanced data and sorting out with balanced information before the model's training stages **(Nguyen, & Duong, 2021)**. Customer analysis using exploratory data analysis (EDA) for visualizing data and the use of machine learning for the classification of customer churn are often used by past analysts. The Synthetic Minority Over-Sampling Technique (SMOTE) method is a popular method applied to deal with class imbalances in datasets **(Nurhidayat, M. M. S., & Anggraini, D. ,2023)**.

3. Methodology

- The first phase of proposed model pre-processes the uploaded dataset and merge Benchmarked Data Sets. This process assists in cleaning the input data by removing any redundant items, replacing String elements, or removing non-processing data entries.
- The feature selection assists in improved data comprehension. It reduces computing time and storage requirements. Which in this study is combination of statistical techniques (Filters) and machine learning technique (wrapper) that chooses relevant features.

The real data set has more than two variables, it is still possible to glean useful information from analyzing every potential bivariate heat map matrix between all pairs of variables of the two variables given by the row and column. The squares maintaining the variable names additionally contain the variable's minimum and maximum values. Although, we lose some information about the distribution, it is important to build bivariate statistical indices that further summaries the frequency distribution, increasing data interpretation. These

indexes allow us to summarize the distribution of each data variable in the bivariate situation, and more broadly in the multivariate case, as well as learn about the relationship between the variables (corresponding to the columns of the data matrix). Independence between two variables, x and y , holds when

$$n_{ij} = \frac{n_i + n_j}{n} \qquad (1)$$

Where, ∀i = 1,2,...,I; ∀j = 1,2,...,J

For all joint frequencies of the contingency data frequencies actually observed n_{ij} and those expected in the hypothesis of independence between the two variables $\frac{n_i+n_j}{n}$. Karl Pearson is the most widely used measure for verifying the hypothesis of independence between x and y . It is defined as following

$$z^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*} \qquad (2)$$

Where,

$$n_{ij}^* = \frac{n_i + n_{+j}}{n}$$

Note that z^2= 0 if the variables x and y are independent. This reveals a serious inconvenience the value of z^2 is an increasing function of the sample size n. To overcome such inconvenience, some alternative measures have been proposed, all functions of the previous statistic.

The Cramer index is equal to

$$w^2 = \frac{z^2}{n \min\big((I-1),(J-1)\big)} \qquad (3)$$

If $0 \leq w^2 \leq 1$ for any I×J contingency data, and $w^2 = 0$ if and only if x and y are independent.

w^2 = 1 for maximum dependency between the two variables.

Furthermore, w^2 has an asymptotic probabilistic (theoretical) distribution, so it can also be used to assess an inferential threshold to evaluate inductively whether the examined variables are significantly dependent.

- High Correlation Filter which plotted the correlation between each variable and another and dropped the ones that had a very high correlation with each other, indicating that we had redundant features. For categorical data, Label encoding has been applied for binary categories data and one hot encoding has been applied for multi categories data.
- Wrapper approaches the topic as a search problem, where numerous combinations are tested and assessed to find the best possibilities and eliminate the remainder. This is like the recursive feature removal algorithm. This proposed technique could reduce attribute size while also reducing misclassification errors.
- Clustering evaluation is performed, and results are obtained to select the optimal clustering technique in terms of accuracy. K-mean, hierarchical, and DBCAN (Density-Based Spatial Clustering of Applications with Noise) algorithms.

For example, the distance is calculated for K-mean algorithm using the formula as follows:

$$d(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\|x_i - x_j\right\|\right)^2 = (\|x_i\|^2 - 2x_i * x_j + \left\|x_j\right\|^2)^{\frac{1}{2}} \qquad (4)$$

Where, $\left\|x_i - x_j\right\|$is the Euclidean distance between two clusters.

- Following clustering, a single classifier-based classification is done yielding AUC (Area Under Curve) and F1-score then choose the best performance of them.
- Sequentially, a hybrid model of the best clustering approach and each individual classifier is constructed as following:

1) Logistic Regression (LR): The sigmoid function is used in logistic regression. It produces binary results, such as yes or no. A sigmoid which unlike linear

regression. It does not appear as a straight line on a graph; instead, it displays a sigmoid. It is true for a category variable. A categorical variable represents a finite set of values. A logistic model's output is probability. Its formula is as follows:

$$p(y = 1|x) = \frac{1}{(1+\exp(-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)))} \quad (5)$$

When the calculated probability exceeds a predefined threshold (usually 0.5), the observation is classed as positive $p(y = 1|x)$= 1 and otherwise as negative $p(y = 1|x)$= 0. the parameter β determines the rates at which the curve grows or increases; the sign of β indicates whether the curve increases or decreases; and the magnitude of determines the rate at which the curve grows or decreases. When β > 0, then p(x) increases as x increases. When β < 0, then $p(y = 1|x)$ decreases as x increases. In addition, for $\beta \rightarrow 0$ the curve tends to become a horizontal straight line. Specifically, when β = 0, y is independent of x.

2) Support vector Machine (SVM): It is expert supervised learning technique that may be used for classification, regression, and outlier detection. It finds a hyperplane that optimally divides various classes in a dataset while maximizing the margin between them for better generalization.

i. Let x be an instance (a d-dimensional numerical vector), y be its true label, and f(x) be a prediction. Assume the prediction has a true value. If we must convert it to a binary forecast, we will set the threshold at zero:

$$\hat{y} = 2(I(f(x) \geq 0)) - 1 \quad (6)$$

where I() is an indicator function that returns 1 if its input is true and 0 otherwise.

ii. A loss function measures how accurate the forecast is. Typically, loss functions do not directly depend on x, and a loss function of 0 corresponds to a perfect forecast, but loss function is positive otherwise. The most noticeable

loss function is squared error is a loss function that is great in numerous ways:

$$E(f(x), y) = (f(x) - y)^2 \quad (7)$$

It does not discriminate between f(x) predictions that are nearly correct and those that are completely incorrect. Then, its derivative regarding the value of f(x) is either undefined or 0 in mathematics. As a result, the 0/1 loss function is challenging to utilize in training algorithms that attempt to minimize loss by modifying parameter values using derivatives.

When the prediction f(x) is real valued, this function is endlessly differentiable everywhere and does not lose information. When the real label is y = 1, it says that the prediction f(x) = 1.5 is just as bad as f(x) = 0.5. If the true label is +1, an incorrect prediction with a correct sign larger than 1 should not be judged incorrect.

iii. The following loss function, known as hinge loss, meets the previously mentioned intuition is more mathematically convenient but less obviously written as

$$E\,(f(x), y) = \max\,\{0, 1 - y\,f(x)\} \quad (8)$$

Indeed,

$$E\,(f\,(x), -1) = \max\,\{0, 1 + f(x)\} \quad (9)$$

$$E\,(f\,(x), +1) = \max\,\{0, 1 - f(x)\} \quad (10)$$

To better understand this loss function, consider the loss E(f(x), -1) for a negative example as a function of the classifier output f(x). If the prediction f(x) is 1, the loss is zero, which is the optimal situation. The greater the loss, the further f(x) is from 1 in the wrong direction, i.e. the greater f(x) is greater than 1. When the true label is y = +1, there is a similar but opposite pattern. The hinge loss function is first significant idea behind SVM. To minimize hinge loss, an SVM classifier f is trained. The training procedure seeks predictions f(x) for all training examples x with

true label y = +1, and predictions f(x) ≤-1 for all training instances x with y = -1. It is worth noting. We have yet to discuss the space of candidate functions f(x). Overall, training aims to correctly identify points, but it does not aim to produce predictions that are exactly +1 or 1. In this view, the training process intuitively attempts to identify the best possible classifier while avoiding any needless secondary goals.

3) Random Forest (RF): It constructs several decision trees and then computes the average prediction value for each distinct tree. (Quinlan, J. R. ,1996) has developed a divide-and-conquer strategy known as the decision tree. The following formulas are used to determine entropy and then information gain for each property.

$$z(s) = \sum -p(x) log_2 p(x) \quad (11)$$

Where, S is the dataset, X is set of classes in S and p(x) is probability of each class.

$$IG(S, A) = z(s) - \sum p(t) z(t) = z(s) - z(S|A) \quad (12)$$

Where, Z(S) is entropy of S, T is the subset of S, p(t) is probability of subset t and Z(t) is entropy of subset t.

4) Multilayer Perceptron (MLP): Researchers demonstrated that, given a large enough number of hidden nodes, a simple neural network structure (with two layers of weights, sigmoidal activation function for the hidden nodes, and identity activation function for the output nodes) can approximate any functional form with arbitrary accuracy.

$$E(w) = -\sum_{i=1}^{n} \sum_{k=1}^{q} (t_{i,k} \log y_{i,k} + (1 - t_{i,k}) \log(1 - y_{i,k})) \quad (13)$$

$$t_{i,k} = y_{i,k} + \varepsilon_{i,k} \quad (14)$$

Where, $t_{i,k}$ is the response vector t_i is assumed to be the sum of a deterministic component and $\varepsilon_{i,k}$ an error term. $y_{i,k}$ is the k^{th}

component of the output vector $y_{i,}$ which represents the fitted probability that the observation i belongs to the k^{th} group c^{k}.

AdaBoost (Adaptive Boosting) is a Meta algorithm that may be used in conjunction with other learning algorithms to increase their performance (Nikolaou et al., 2016). On the classifiers, weighted majority voting is used. In each iteration, each classifier is given an equal chance to draw samples. The weighted majority voting formula is as follows:

$$A_j = \sum_{w_i} \log(\frac{1}{\beta_t}) \quad , i = 1,2, \ldots . , c \qquad (15)$$

where $\beta_t$is normalized error, t represents trained classifiers and w_i represents class labels of training data.

- Following pervious the proposed prediction model, the performance of the proposed architecture is calculated using multiple performance metrics. The computed performance measures are utilized to construct comparison sets and validate the contribution of this improvement.

4. Results

The proposed methodology was implemented using the IBM telecommunications dataset. This data is suitable for machine learning models due to its various attributes. Python programming was used to perform clustering, classification, dimensionality reduction, visualization, and data preprocessing. The raw data were turned into a final dataset during the data preparation step so that we could feed it into the modelling algorithms and create models. At this phase, many tasks were completed, including data cleansing, handling missing values, selecting features, and data transformation.

This section discusses the churn analysis results, and the study work has created a script in the PYTHON programming language.

i. Data acquisition in six Data frames from Benchmarked IBM datasets.

```
Telco_customer_churn df_1 shape =  (7043, 33)
Telco_customer_churn_demographics df_2 shape =  (7043, 9)
Telco_customer_churn_location df_3 shape =  (7043, 9)
Telco_customer_churn_population df_4 shape =  (1671, 3)
Telco_customer_churn_services df_5 shape =  (7043, 30)
Telco_customer_churn_status df_6 shape =  (7043, 11)
```

ii. Data Merging: Merge all data frames df1, df2, df3, df5, df6 using Customer ID and merge all data frames with df4 population using Zip Code.

iii. Data Cleaning: Drop duplicated columns, drop noisy data (null elements) and drop missing data and Impute others.

iv. Feature selecting firstly through correlation method as follows.

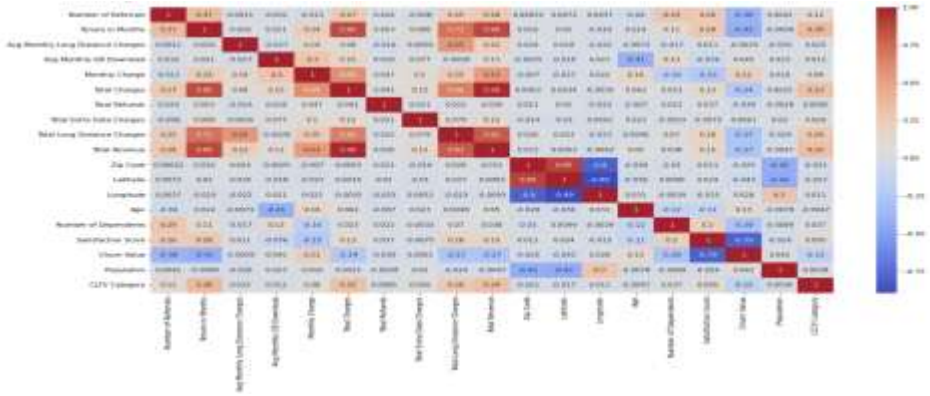

Figure1. Heat map correlation of features.

v. Proposed selection feature by splitting Data into numerical and categorized data.

A. Numerical columns selected by applying (Pearson and ANOVA)

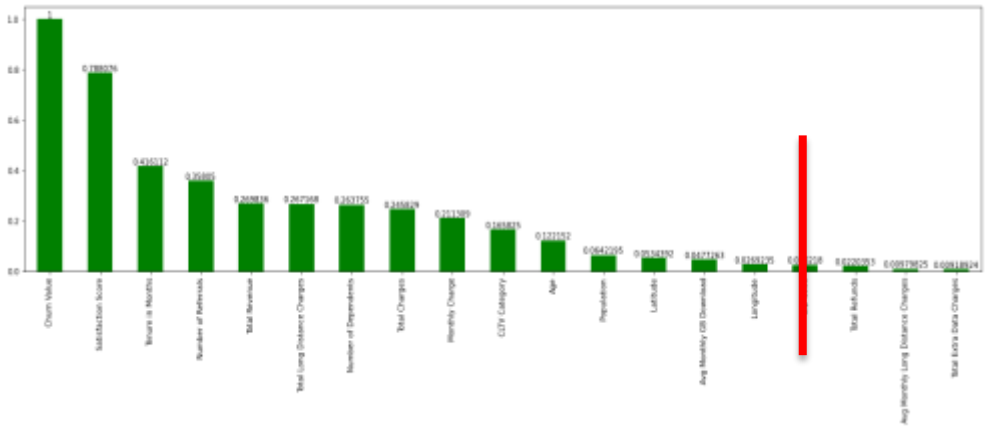

Figure2. persons with threshold 0.1

Pearson selected column (10 columns) and ANOVA feature selection (11 columns). Add ANOVA and persons selected column to Data frame then select common between two techniques.

Selected_numerical_columns = ['Satisfaction Score', 'Tenure in Months', 'Number of Dependents', 'Number of Referrals', 'Age', 'Monthly Charge', 'Total Revenue', 'CLTV Category', 'Total Charges', 'Total Long-Distance Charges'].

B. Categorical columns selected by applying three feature selection techniques (Chi-Square test, ANOVA, and Mutual_info_classif).

The categorical data has split to binary and multi category. Common columns have selected from 3 techniques, 23 columns have remained after hyper tuning use ANOVA and chi-Square. Encoding is the process of transforming category data into numerical data. most common type of data in the sample was categorical data, and the values for these variables were typically kept as text. However, because machine learning algorithms are built on mathematical equations, they can only be used with numerical data. As a result, leaving the categorical variables in their current state was impossible, and they had to be transformed into a numerical format. Therefore, apply label encoding on binary category columns and apply one hot

encoding on Multi category columns then category columns became 35 columns.

vi. Merge selected numerical and categorical Columns by wrapper. Became 33 column and apply wrapper technique (backward and forward). Set K = 28 to wrapper models then select common features between backward and forward became 23 columns.

vii. By deeply analytics for the final features from Data found the following Churn Reason

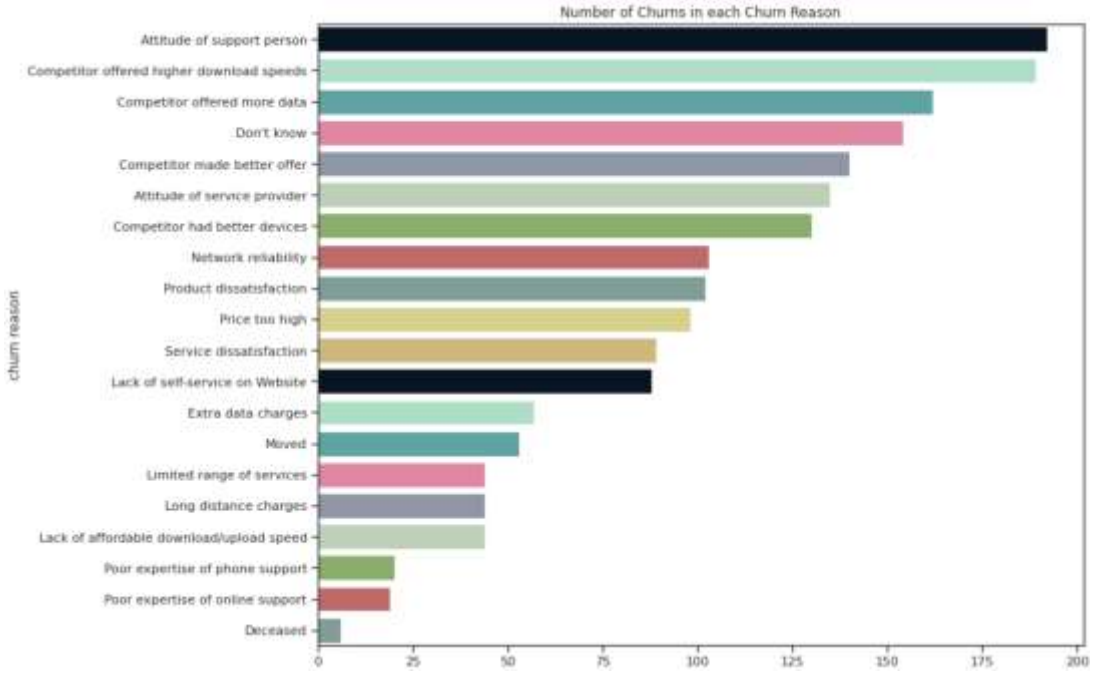

Figure 3. Bar chart for counting number of churns in each churn reason

It's clear from pervious chart that competitor offers, and attitude of support person is the most important columns or customer churn reason. Also, internet service and technical support and the contract type.

viii. As previously stated, the datasets from literature were employed for 80% of the training and 20% of the testing sets. The tables below show the performance parameters for ordinary and set-

based machine learning methods, including precision, F1 score, and recall. Additional result breakdowns in terms of imbalanced and balanced datasets are also offered for each dataset. The classification performance will be displayed in the next sections utilizing measures such as recall, precision, F1-scores, and ROC-AUC values. These metrics were calculated separately for each algorithm and resampling approach. Each classification algorithm was also applied to the imbalanced dataset.

- **Clustering algorithms**: Unsupervised learning has applied several algorithms K-mean, hierarchical, and DBCAN. Select K-mean best model performance. Through the elbow method and the silhouette method. To get its performance as following.

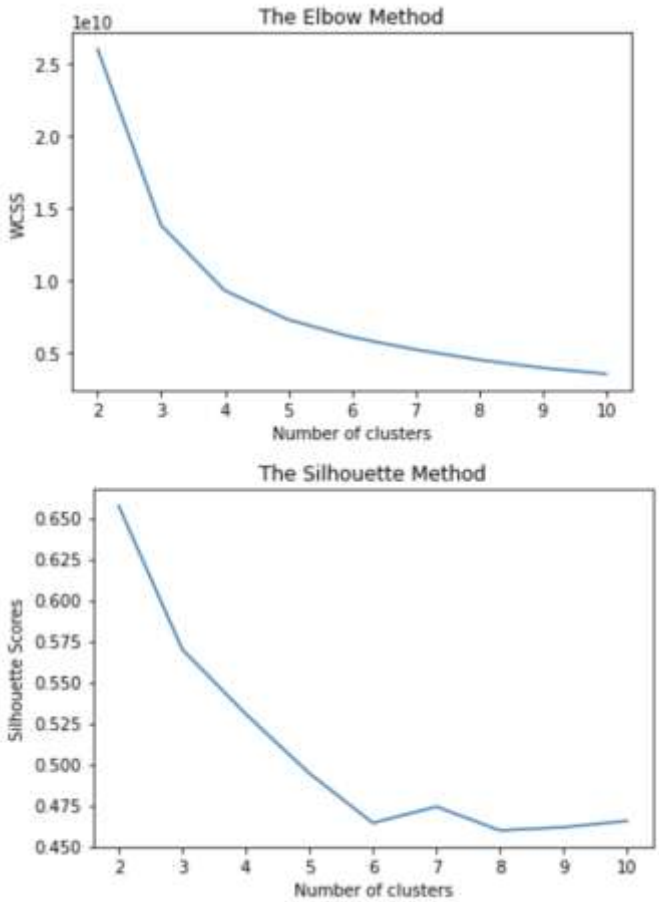

Figure 4. The elbow and the silhouette methods for K-mean algorithm

- **Classification algorithms:** Customer churn prediction is stated as a classification problem in this study, which should be conducted using supervised learning. Because the dataset comprises both predictors and labels, it was divided into two datasets: predictor (named X) and label (called y). The splitting allows machine learning algorithms to use predictors to predict the label. To evaluate generalizability and compare the performance of classification algorithms, trained models must be evaluated on unseen data. As a result, a fixed number of cases must be set aside as unseen data for testing. This study will deal with the results obtained. Thus, the performance parameters mentioned above are evaluated and discussed for the different individual and combined classifiers. In this study, the following individual classification algorithms have been implemented Logistic regression, SVM, Random Forest and MLP classifier.
- **Standardization:** Applying different scalers for feature scaling through Standard Scaler, MinMax scaler, and Robust scaler which is more efficiently which solving imbalance for basic proposed model as following.

Table 1. Evaluation for different feature scaling methods

Model	pure F1_score	MinMaxScaler F1_score	Standard scaler F1_score	Robust scaler F1_score
Logistic Regression	91.85%	92.88%	93.20%	93.20%
Support Vector Machine	88.59%	93.06%	92.86%	92.95%
Random Forest	94.26%	94.27%	94.63%	94.42%
Adaptive Boosting	93.67%	93.67%	93.67%	93.67%
MLP Classifier	93.49%	93.34%	93.76%	93.95%

5. Conclusion

Accurate proposed customer churn prediction model can drive decision-making and provide valuable insights. Data preprocessing is done for merged IBM Benchmarked datasets and the feature scaling for standardization process solving imbalance issue in datasets. The performance results demonstrated the significance of selecting a feature selection procedure to develop a higher quality customer churn prediction model. The study recommended employing feature selection to obtain relevant features, which results in improved customer churn prediction framework performance. The study allows for the monitoring of a broader range of consumer behavioral attributes. The customer behavioral attributes must be valued based on the weight criteria of those churn effects. A combination of clustering and classification techniques is used to predict customer churn on enormous data set of the telecom industry. The study includes churn prediction techniques which are useful in a variety of businesses. The analytics clearly demonstrate the high significance of the churn predictive method, particularly in the communication industry, for maintaining customer retention and producing high essentials for all service sectors.

References

- **Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019).** Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.
- **Dwiyanti, E., Adiwijaya, & Ardiyanti, A. (2017).** Handling imbalanced data in churn prediction using rusboost and feature selection (case study: Pt. telekomunikasi indonesia regional 7). In Recent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016, Proceedings Second (pp. 376-385). Springer International Publishing.
- **Gui, C. (2017**). Analysis of imbalanced data set problem: The case of churn prediction for telecommunication. Artif. Intell. Res., 6(2), 93.
- **Nguyen, N. N., & Duong, A. T. (2021).** Comparison of two main approaches for handling imbalanced data in churn prediction problem. Journal of advances in information technology, 12(1).
- **Nurhidayat, M. M. S., & Anggraini, D. (2023).** Analysis and Classification of Customer Churn Using Machine Learning Models. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 7(6), 1253-1259.
- **Petkovski, A. J., Stojkoska, B. L. R., Trivodaliev, K. V., & Kalajdziski, S. A. (2016, November).** Analysis of churn prediction: a case study on telecommunication services in Macedonia. In 2016 24th Telecommunications Forum (TELFOR) (pp. 1-4). IEEE.
- **Quinlan, J. R. (1996).** Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), 71-72.
- **Tiwari, A., Sam, R., & Shaikh, S. (2017, February).** Analysis and prediction of churn customers for telecommunication industry. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) (pp. 218-222). IEEE.