Customer Churn Prediction in Telecom Industry Using ML Algorithm

Muhammed Shibil C V
CB.SC.P2DSC23007
I-MTech-DSC 23-25
21DS602 (22-24((Odd))

Sudharshan R Mohan
CB.SC.P2DSC23022
I-MTech-DSC 23-25
21DS602 (22-24((Odd))

Abstract

Customers are the most important part in every business industry, especially in Telecom industry. So, this project mainly focusses on predicting the customer will retain the product or not using machine learning models. Businesses must compete fiercely to win over new consumers from suppliers. The dataset we used contains customer data collected from the respective telecom company and the datasetwas used to train and test the proposed model. We implemented machine learning models such as Decision Tree Classifier and Random Forest Classifier in this dataset. Both gave good performance with an overall precision of 0.93 and 0.94 respectively.

I. INTRODUCTION

The customer's concentration on the providers has prompted many new telecom associations to emerge. These new firms usually specialize in providing a specific service or product that the customer cannot find from the incumbent providers. In all industries or businesses Customers are the main pillar. so every business firm is always trying to make more consumers for the existence of their business. But in a highly competitive market, the customer count is saturated in nature so instead of expecting new customers it is better to take care of and maintain the available customers in an effective way. In marketing terms churning can be defined as the loss of customers from a firm also mentioned as customer turnover or attrition it is a common phenomenon in Every firm and they always try to reduce the rate of churning. Usually, the customer churns when they face any difficulties or find any good alternatives. So, it is important to analyse

and take action in this matter. If businesses have clear-cut forecasting about customer turnover they can take required action plans like campaigning, new service schemes to maximise the profit mong few methodologies for forecasting customer churn, Machine Learning is the most feasible and widely explored method. ML algorithms such as KNN, Logistic regression, Decision trees, Naïve bays, Support vector machines (SVM) and various other techniques can be used effectively to anticipate customer attrition. The implementation of model effectively predicts the customer churn rate so the organization can implement the recovery strategies from a lot of ML algorithms, the selection of a suitable one is always a big issue because each algorithm has unique accuracy and model parameters. So, in this report, we focus on a comparative analysis of different ML algorithms for churn prediction on the same dataset and the best one suited for churning is determined based on the accuracy and evaluation measures of these algorithms. Some potential churn prediction algorithms are utilized in this study: Random Forest, K-Nearest Neighbors and Decision tree. The study also focuses on single classification explores Ensemble-based classification algorithms and for class imbalance issue, studies have mainly applied the Synthetic Minority Over-sampling Technique (SMOTE) Recently, hybrid resampling methods have been proposed as a more effective method to tackle imbalanced data. Apart from algorithms this report also delves into the idea of data exploration using the idea of correlation matrix and univariate, bivariate analysis which helps to how much the target is dependent on features. The rest of the sections in the report provide ideas about, dataset description, methodologies and model description, results and conclusion of our work.

II. DATASET DESCRIPTION

The raw dataset obtained for this work is popular Telco customer churn dataset. It totally consists of 7043 customers telco data with 21 features such as customer id, gender (Whether the customer is a male or a female), Senior Citizen, Partner (Whether the customer has a partner or not), Dependents (Whether the customer has dependents or not), tenure (Number of months the customer has stayed with the company), phone service (Whether the customer has a phone service or not), multiple lines (Whether the customer has multiple lines or not), internet service (Customer's internet service provider), online security (Whether the customer has online security or not) so on. Hot encoding and label encoders were used to transform the categorical labels to numerical labels and for normalizing the labels.

customer	tenure	Online Backup	Contract	Churn
gender	Phone Service	Device Protection	Paperless Billing	
Senior Citizen	Multiple Lines	Tec Support	Payment Method	
Partner	Internet Service	Streaming TV	Monthly Charges	
Dependents	Online Security	Streaming Movies	Total Charges	

7043 customers telco data with 21 features

III. METHODOLOGY

The most well-liked prophetical models are applied within the prediction method: KNN, random forest classifier, and decision tree. The evaluation of the built model is performed by ROC curves, confusion matrix etc. And There is an over-sampling technique called SMOTE, which stands for Synthetic Minority Over-sampling Technique, which is a technique used in machine learning to address the class imbalance problem. Class imbalance occurs when the number of instances of one class (the minority class) is significantly lower than the number of instances of another class (the majority class) in a classification dataset. So we can achieve a better accuracy.

Below mentioned flow chart below gives an overview of the implementation

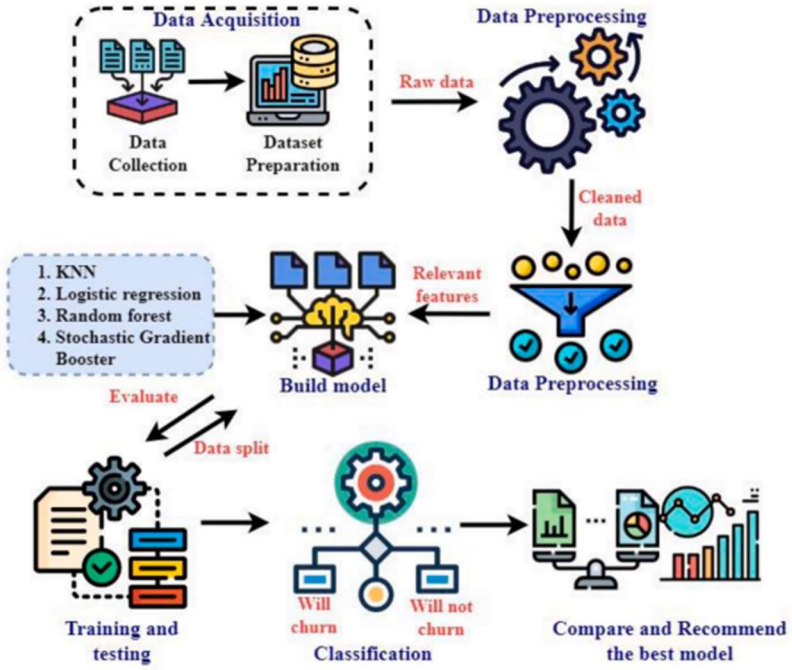


Figure 1: System layout

Data preprocessing and exploratory data analysis:

Data preprocessing and exploratory data analysis (EDA) are crucial steps in the data analysis that help ensure the quality of data and provide valuable insights into the dataset. Let's explore each concept.

Data preprocessing usually includes the process of cleaning the data and make the data fit for analysis without proper data preprocessing it is difficult to get a desired output

It usually involves the procedures such as

1. Data cleaning
2. Handling of missing values
3. Data transformation

4. Handling the outliers

Exploratory Data Analysis Also known as EDA gives insight into how the features are related to each other by both statistical and visual interpretation of data, so it helps to understand the trend, relationships, and anomalies within the data. Below mentioned procedures are usually involved in EDA

1. Univariate analysis
2. Bivariate analysis
3. Multivariate analysis
4. Descriptive statistics
5. Identifying Patterns and the trends

Implementation of ML algorithms

These are the 3

1. K-Nearest Neighbors
2. Decision tree
3. Random Forest model

K-Nearest Neighbors:

In the KNN model we initialize the K value and find the distance to each data point and sort the arranged data in ascending order, after checking for the labels of K instances we finalize the target (figure1). Usually, it is very easy to implement because of few hyperparameters and it give a satisfied accuracy.

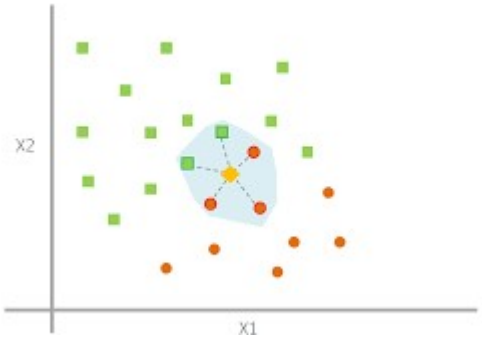


Figure 2: Working of KNN

Decision tree:

Decision trees are another supervised learning algorithm and it is a graphical representation of getting all possible solutions to a problem based on the given condition. Compared to KNN it gives much more accuracy to the given problem and the logic behind the decision tree can be easily understood because it can refer to the decision-making capability of humans.

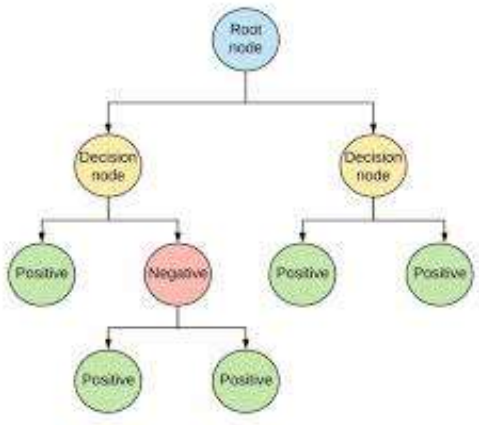


Figure3: Decision tree

Random Forest:

The Random Forest approach is actually an Ensemble learning approach it includes several classifiers to build the solution by getting predictions from every model (combining a lot of decision tree models) and making a final decision on majority voting or averaging.it is usually efficient for large datasets. Since it is a much time-consuming process however it can provide better accuracy compared to the other implemented models.

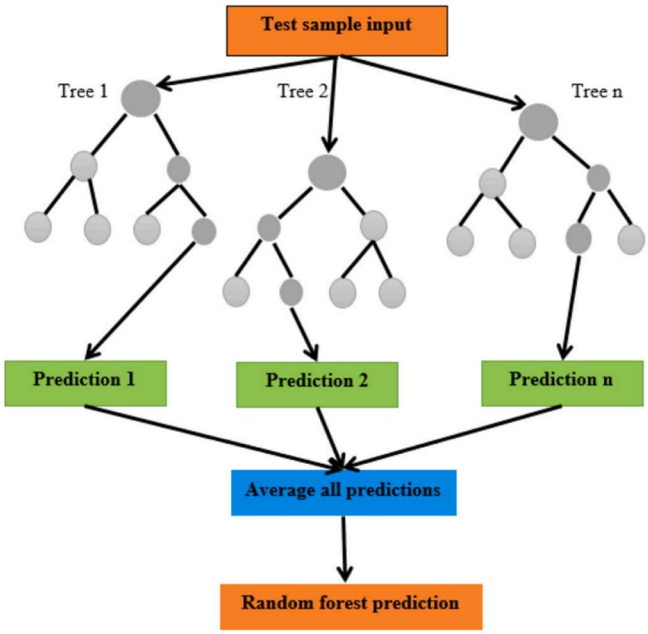


Figure3: Random forest

IV. RESULT AND DISCUSSION

- The KNN Algorithm give a model with an accuracy of 78.6 after applying SMOTE and the AUROC obtained is 0.75

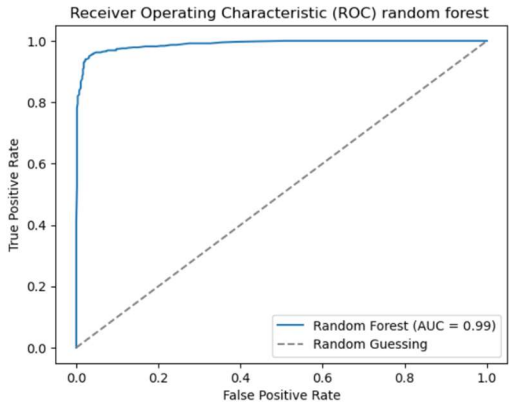
```
Confusion Matrix:
[[748 301]
 [146 871]]

Accuracy: 0.7836398838334947

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.71      0.77      1049
           1       0.74      0.86      0.80      1017

    accuracy                           0.78      2066
   macro avg       0.79      0.78      0.78      2066
weighted avg       0.79      0.78      0.78      2066
```

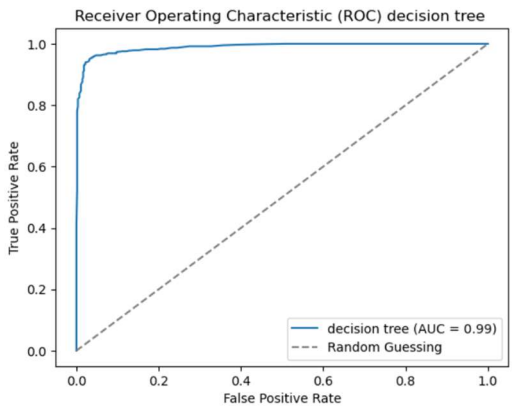
- The prediction results of Random Forest algorithms give a reasonable accuracy of 95.76 The corresponding area under the ROC curve is obtained as 0.99

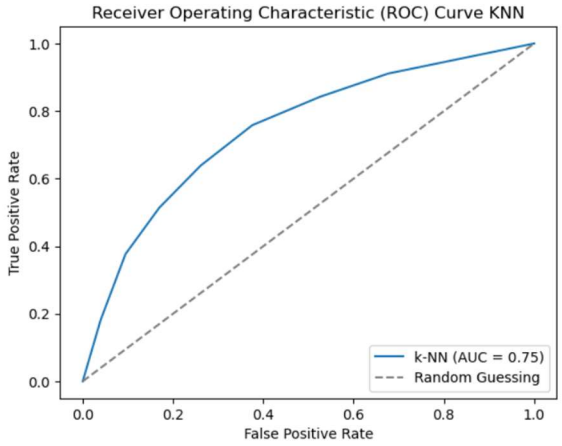
```
accuracy 0.9576490924805532
classification report
              precision    recall  f1-score   support

           0       0.95      0.96      0.96       545
           1       0.96      0.96      0.96       612

    accuracy                           0.96      1157
   macro avg       0.96      0.96      0.96      1157
weighted avg       0.96      0.96      0.96      1157

confusion matrix
[[522  23]
 [ 26 586]]
```

- The prediction results of the decision tree also give an almost near accuracy of 93.6 is almost equal to the accuracy of the random forest model and the AUROC is 99

```
Acciracy: 0.934819897084048
              precision    recall  f1-score   support

           0       0.92      0.93      0.93       525
           1       0.94      0.94      0.94       641

    accuracy                           0.93      1166
   macro avg       0.93      0.93      0.93      1166
weighted avg       0.93      0.93      0.93      1166

confusion matrix
[[489  36]
 [ 40 601]]
```







Here after analyzing the classification report and ROC curve, it is clear that the Random Forest model shows an accuracy of 95% which is greater than the rest of two models.

V. CONCLUSION

This study focuses on finding the most suitable Machine Learning algorithm selection for the given problem which is churn prediction. During the study we selected 3 algorithms to do a comparative study they are K-N earest Neighbors, Decision tree, and Random Forest classifier. The application of SMOTE played a crucial r ole in mitigating the impact of class imbalance within the dataset. By oversampling the minority class throug h the generation of synthetic samples, SMOTE effectively balanced the class distribution, enhancing the mod el's ability to learn and generalize patterns from both classes. So as a result, out of 3different machine learnin g models the random forest Classifiers have the largest performance measures and AUROC of 95.76 and 0.9 9 respectively

REFERENCES

[1]. R. Shalini, B.R. Kavitha Customer churning analysis using machine learning algorithms B. Prabadevi International Journal of Intelligent Networks 4(2023) 145–15

[2]. Youngjung Suh Suh learning-based customer churn prediction in-home appliance rental business, Journal of Big Data (2023) 10:41

[3]. Kimura T. Customer churn prediction with hybrid resampling and ensemble learning. J Manage Inform Decis Sci. 2022;25(1):1–2