Customer Churn Prediction In Telecommunication Using Supervised Machine Learning Algorithms

Laksika Tharmalingam
Department of Computer Science and Engineering,
University of Moratuwa
Colombo, Sri Lanka
laksika.19@cse.mrt.ac.lk

***Abstract*—In each and every industry, the number of service providers is rapidly increasing. In the modern telecommunications industry, there is an abundance of options available to consumers. Consequently, customer churn and engagement have become one of the most pressing concerns for the majority of telecommunications industries[1]. Using machine learning techniques, this paper proposes a method for predicting customer churn in the telecommunications industry. The research promotes the exploration of the likelihood of churn by analyzing customer behavior. The KNN, SVM, Decision Tree, Gradient Boosting Classifiers, Naive Bayes, XGB Classifier, Logistic Regression and Random Forest Classifier are used in this study. Also, some feature engineering methods have been done to find the more relevant features and to verify system performance. The experimentation was conducted on the dataset from Chatterbox Telco Pvt Ltd. The results are compared to find an appropriate model with higher precision and predictability. As a result, the use of the XGB Classifier model after oversampling is better compared to other models in terms of accuracy and F1 score.**

***Index Terms*—Customer churn, Telecommunication ,Random forest, XGB Classifier, KNN**

I. Introduction

Customer churn is the loss of customers by a business for different reasons such as poor service and better price somewhere else. It is one of the most critical and challenging problems for telecommunication companies. Since acquiring new customers costs more than retaining existing ones, analysing customer churn is vital for businesses.

Rapid improvements and dynamics in technology market place make customer retention a competitive effort. Especially in saturated telecommunications market, there are incumbent service providers and newcomers offering deals and packages for consumers who would like to churn to their services. On the defending end, strategies and counter offers have to be made for potential churners as it is more expensive earning a customer back once s/he churns.

In this study, we concentrate on evaluation and analysis of performance of different machine learning methods for accurate churn prediction. In this approach, a model is built by training it with a set of churns and not churns. Then it is tested and its performance is analyzed through various machine learning classification algorithms. Some popular machine learning classification algorithms are selected for the training and testing phase. The following evaluation metrics are considered for the performance of algorithms: Classification Accuracy, Confusion Matrix, Precision, Recall and F1 score.

II. Methodology

The objective of this study is to predict customer churn in a telecommunication industry as early as feasible using effective machine learning techniques.A diagrammatic representation of the proposed model is given in Fig 1

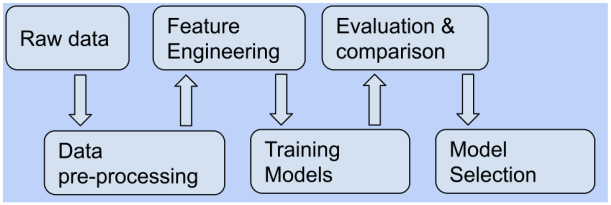

Fig. 1. Proposed model

A. Dataset Description

The dataset used in the research is supervised, obtained from Chatterbox Telco Pvt Ltd. The dataset includes information of 2312 customers, and the target parameter is a binary variable that represents whether the customer has left the chatterbox or still a customer. Of this, 1737 were positive class (Churn) samples and 575 were negative class(Not Churn) samples. The target variable reflects the binary flag 1 when the client has left from chatterbox, and 0 when the client is not left from chatterbox.This is given below in fig 2.

Churn: If the customer left or not
DataType: Categorical Nominal
Value =Yes , No
Yes: 1737 No: 575

The dataset contains 19 feature predictors. The details of these features based on theirs data type are given below:

- Categorical Nominal- customer_id, location_code, international_plan, voice_mail_plan.
- Metric Discrete- account_length, total_day_calls, number_vm_messages, total_eve_calls, total_night_calls, total_intl_calls, customer_service_calls.
- Metric Continuous- total_day_min, total_day_charge, total_eve_min, total_eve_charge, total_night_minutes, total_night_charge, total_intl_minutes, total_intl_charge.

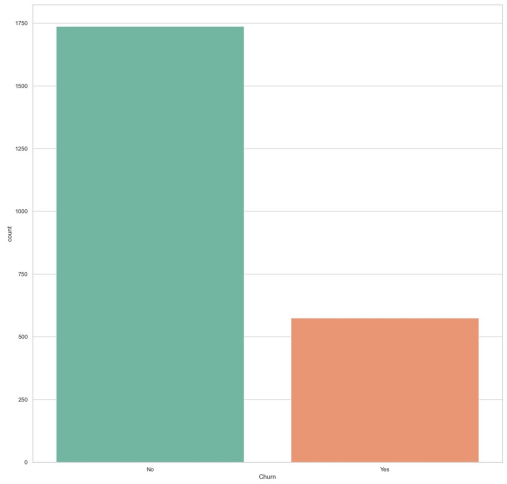

Fig. 2. Churn

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2321 entries, 0 to 2320
Data columns (total 21 columns):
customer_id               2321 non-null int64
account_length            2319 non-null float64
location_code             2321 non-null int64
intertiol_plan            2318 non-null object
voice_mail_plan           2315 non-null object
number_vm_messages        2318 non-null float64
total_day_min             2320 non-null float64
total_day_calls           2318 non-null float64
total_day_charge          2316 non-null float64
total_eve_min             2318 non-null float64
total_eve_calls           2317 non-null float64
total_eve_charge          2313 non-null float64
total_night_minutes       2319 non-null float64
total_night_calls         2316 non-null float64
total_night_charge        2316 non-null float64
total_intl_minutes        2319 non-null float64
total_intl_calls          2318 non-null float64
total_intl_charge         2316 non-null float64
customer_service_calls    2320 non-null float64
Churn                     2316 non-null object
Unnamed: 20               0 non-null float64
dtypes: float64(16), int64(2), object(3)
memory usage: 380.9+ KB
```

Fig. 3. Missing value

B. Data Preprocessing

Data preprocessing is the most important phase in prediction models as the data consists of missing values,ambiguities, errors, redundancy and transformation which needs to be cleaned beforehand.

1) Data Cleaning:

- **Detect and Remove Duplicates**: When the customer_id field is excluded from the duplicate check, four rows are duplicated. Therefore, remove those four rows.
- **Handling Missing Values**:When verified for missing values, the majority of columns have null values. It is given in the fig 2
  If the output column (Churn) has missing values, remove the row entirely. Otherwise, populate these values relative to other column values.This is given below in fig 3.
  Techniques used to fill the missing values are given below.
  - account_length,customer_service_calls: filled with median values
  - international_plan: filled with ”no”
  - Voice_mail_plan: If number_vm_messages = 0 then filled with “no”. Otherwise filled with “yes”.
  - number_vm_messages: If Voice_mail_plan = “no” then filled with 0. Otherwise filled with median value based on Voice_mail plan = “yes” rows.
  - total_day_min,total_eve_min,total_night_minutes,total_intl_minutes: First sort the dataset with location code and recpective charge. After that fill the null values with the forward fill method.
  - total_day_calls,total_eve_calls,total_night_calls,total_intl_calls: First sort the dataset with respective minutes and charge. After that fill the null values with the forward fill method.
  - total_day_charge,total_eve_charge,total_night_charge,total_intl_charge: First sort the dataset with location code and respective total minutes. After that fill the null values with the forward fill method.
- **Handling Out-of-range values**: In the dataset, some columns has negative values. Therefore, replace those values with the absolute values.
- **Handling Outliers** Using pairplot,identified the Outliers and fill those with corrected values.

2) Data Transformation: The dataset contains four category columns. For location_code, one hot encoding are used. And for the remaining three columns (international_plan, voice_mail_plan, and Churn), ordinal encoding was used.

C. Handling imbalance data

Not churn rows are three times larger than churn rows in the dataset. They are in the ratio 3:1. When we use this dataset to train our model, it will be biased. In this churn prediction, it is crucial to appropriately identify the minority classes. Therefore, the model should not be biased to identify only the majority class, but should also assign equal weight or significance to the minority class.Therefore using oversampling method, duplicated the minority class(Churn) until it become equal to majority class (Not churn) and after that shuffled the dataset to randamized.

D. Feature engineering

Feature engineering is important to find new features and to identify the correct features to train the model.

1) feature selection: Using a correlation table, it was determined that all model training features are essential. Consequently, only customer_id is eliminated.

2) feature creation: From the existing features, Some new features has been created. Those features are given below.

- total_min: Sum of total_day_min,total_eve_min, total_night_minutes,total_intl_minutes
- total_calls: Sum of total_day_calls,total_eve_calls, total_night_calls,total_intl_calls
- total_charge: Sum of total_day_charge,total_eve_charge, total_night_charge,total_intl_charge
- day_charge_per_min: ratio between total_day_charge, total_day_min
- eve_charge_per_min: ratio between total_eve_charge, total_eve_min
- night_charge_per_min: ratio between total_night_charge, total_day_minutes

3) feature scaling: Often, the problem with categorized learning is the magnitude of the features. A good learning method needs dataset that be scaled, which facilitates better convergence. Normalizing data can be accomplished in a variety of ways. One of the simplest methods for normalizing a given dataset is scaling each feature's maximum value. The similar methodology is applied in this research.

III. Models experimental and Result analysis

In this paper,we experimented with some models. Mainly they are K-Nearest Neighbor(KNN), Support Vector Machine(SVM), Decision Tree, Gradient Boosting Classifiers, XGB Classifier, and Random Forest classifiers.

To select the optimal model from the above list, we used 5-fold cross-validation with f1 score to our dataset. Using cross-validation, each data point is evaluated precisely once and utilized in training k-1 times. And the f1 score balances precision and recall for the positive class.

We observed results in two cases that are with the unbalanced dataset and the balanced dataset. Results for the unbalanced dataset is given in fig 4 and Results for the balanced dataset is given in fig 5.

	Model	ScoreMean	Score Standard Deviation
5	LinearSVC	0.469460	0.056031
7	LogisticRegression	0.475760	0.056687
1	KNeighborsClassifier	0.668175	0.028765
2	GaussianNB	0.669449	0.022350
4	SVC	0.831227	0.020911
0	DecisionTreeClassifier	0.901571	0.059389
3	RandomForestClassifier	0.961239	0.009521
9	BaggingClassifier	0.963122	0.007397
6	XGBClassifier	0.963204	0.009263
8	GradientBoostingClassifier	0.965164	0.003134
10	BaggingClassifier	0.966808	0.003656

Fig. 4. Unbalanced

For the unbalanced dataset gradient boosting classifier has the highest f1 score. For this model mean score is 0.9651 and standard deviation is 0.003. To improve this model further bagging classifier is used with gradient boosting classifier.This bagging classifier improved the f1 score to 0.9668 and standard deviation for the score is 0.003.

	Model	ScoreMean	Score Standard Deviation
2	GaussianNB	0.676282	0.021307
7	LogisticRegression	0.762852	0.009437
5	LinearSVC	0.767418	0.011567
1	KNeighborsClassifier	0.874227	0.015443
4	SVC	0.931977	0.003205
8	GradientBoostingClassifier	0.973577	0.004256
0	DecisionTreeClassifier	0.981978	0.005170
3	RandomForestClassifier	0.995639	0.004871
6	XGBClassifier	0.996215	0.004187

Fig. 5. balanced

When balanced the dataset f1 score increased a lot.After balanced the dataset, XGBClassifer shows the best result with 0.996 f1 score.For this model standard deviation is 0.004

For the best three models I tried hyper parameter tuning. But f1 score is not change a lot.

From above results, XGB classifier is the best model.And after that, Gradient Boosting Classifier and Random forest Classifier shows the best result with the nearly same f1 score.

IV. Conclusion

In this research, we offer two models for predicting churn in the telecommunications industry: the XGB classifier and the Gradient Boosting Classifier. These models estimate customer churn based on the behavior of the client and are independent of data from other clients. From a service provider's database on service usage, such as call centre detail records, it is simple to retrieve the necessary information. Customers' churn in the telecommunications industry can be accurately predicted by evaluating the outcomes and considering the practical perspective of the models. In accordance with the metrics that are of utmost importance for obtaining the best performance, decision-makers can develop a variety of marketing approaches to manage and retain churners. The experimental findings indicate that the proposed method is very capable of detecting churns. The proposed method attained an accuracy of 99 percent, which is higher than other existing methods.

References

[1] CUSTOMER ANALYTICS ON TELECOM CHURN Dr. P. Ravilochan (Professor) School of management, SRM University, Kattankulathur 603203.