TELECOM CUSTOMER CHURN PREDICTION WITH DATA MINING AND MACHINE LEARNING

Nguyen Quoc Hung[1], Nguyen Tan Khang[1],Nguyen Chanh Nghia[1] and Bui Ha Khanh[1]

[1]*Department of Computer Science,University of Information Technology*
Neighborhood 6,Linh Trung District,Thu Duc City,Ho Chi Minh City,Vietnam

E–mail: *21520253@gm.uit.edu.vn, 21520281@gm.uit.edu.vn, 21520362@gm.uit.edu.vn, 21520282@gm.uit.edu.vn*

Abstract

Customer churn poses a formidable challenge within the Telecom industry, as it can result in significant revenue losses. In this research, we conducted an extensive study aimed at gaining insight into the customers trend for telecom industry and providing guideline for businesses to adapt their operations, services, and products. Our method utilizes the data visualization strength to identify the key causes of customer churn. Furthermore, we utilized a spectrum of machine learning techniques that include Logistic Regression, KNN, Random Forrest, and XGBoost to predict if a customer is a churner or not. The XGBoost and Random Forrest yielded an accuracy rate of 0.87, underscoring the effectiveness of our approach in addressing the critical issue of customer churn in the Telecom industry.

Index Terms—Telecommunication Industry, Customer Churn, Classification, Association rules.

1. INTRODUCTION

As population grows, the demand for telecom services also grows exponentially, thus calling for comprehensive research into users' interest and need. Failing to answer the common need can lead to the loss of market share and decline in growth, revenue and profit for telecom companies.

The identification of clients that are potentially at possibility of churning assumes fundamental relevance inside this complex ecosystem. With this foresight, businesses may proactively engage in client retention measures like personalised service improvements. This tactical move has the potential to reduce client attrition, protecting revenue sources and preserving corporate viability.

Research done on the field has yielded many positive results on customer base's retention and general service satisfication.In those research, machine learning approaches and association rule mining approaches are employed on processed dataset provided by leading corporations of the field to gain knowledge on the trend among their subscribers.Garbacz and Jr (2007)

This paper presents our research into analysing the main contributors to service subscription among telecom serivce users via IBM telecom dataset. We conduct our research in two phases.

Phase 1 is processing IBM dataset and visualising the distribution to understand the relationship among custormers churning and other features of customers. Phase 2 is deploying machine learning models(Random forest, Decision Tree, XGBoost and KNN) and association rule mining algorithm(Apriori) to mine knowledges of features and establish relationship between features and customers churning.

2. Methodology

The visualization technique and Apriori algorithm are used in the initial stage to select the main features in a considered manner and the key point of churned customer, thus shaping the foundation of our analysis. Our research approach is based on comparing many machine learning algorithms to choose the best model for prediction accuracy.

2.1. Dataset and Feature Analysis

In this study, we use the IBM Telcom Customer Churn dataset that is available by the IBM Watson

Analytics community. The data encompasses 7,043 client records. The dataset contains 5 main features **demographics**: customerID, gender, age..., **location**: state, city, zip code..., **population**: population, id..., **services**: online security, streaming tv, streaming movies..., and **status**: churn label, churn score, churn value. IBM (2024)

In our research, feature analyis by visuliazation played a pivotal role to gain more insights on the dataset. We were able to evaluate relationships between the total charges, monthly charges and tenure. We found that although current customer's total charges is much higher than churner, the price churner pay monthly is more than current customer **Fig**3. Specially, when the customer pay more than 60 dollars a month, the rate of churn is increase steadily. As monthly charges increase, the probability of customer churn increases **Fig**4. Beside that, the visualization also show that the current customers have longer tenure than the churned customers. This is as expected, since the current customers had larger total charges, while maintaining lower monthly charges. As tenure increases, probability of churn decreases **Fig**5. Summarily, to reduce the rate of churn, we should provide promotion encouraging customer using services more usually and in a long time to minimize a monthly cost.

The Apriori technique, which has gained popularity for its association rule mining capabilities Agrawal and Srikant (1994), is used in this study to efficiently identify frequent item sets and noteworthy patterns in our dataset. We utilized this method to find the best service combinations for selling more products and enhancing customer experiences. In Table, we showcase the association rules, confidence, and lift metrics. This is to provide insights into item relationships and patterns. The relationship between services indicates that customer often open phone services, device protection, streaming movies and streaming TV services together.

antecedents	consequents	confidence	lift
(Streaming TV_Yes, Device Protection_Yes)	(Streaming Movies_Yes)	0.800255	1.616035
(Streaming TV_Yes, Device Protection_Yes, Phon...	(Streaming Movies_Yes)	0.799427	1.614363
(Streaming TV_Yes, Device Protection_Yes)	(Streaming Movies_Yes, Phone Service_Yes)	0.712189	1.614275
(Device Protection_Yes, Streaming Movies_Yes, ...	(Streaming TV_Yes)	0.792051	1.614239
(Device Protection_Yes, Streaming Movies_Yes)	(Streaming TV_Yes)	0.786207	1.602329

Fig. 1: Association rule

Based on Churn Reason and Satisfaction Score **Fig**2, we determined the main reason of churn because of the attitude of service supporters and the better products from competitor. To increase customer's loyalty, we should train support staff to be more friendly and considerate. Specially, improve the quality of services is the major concern to compete with others.

2.2. Data processing

In the course of our research, we applied a range of techniques designed to uncover concealed insights within a customer churn dataset. This step, known as data preprocessing, plays a pivotal role in analysing the dataset used in our work. The data preprocessing step is primarily performed to refine raw data, ensuring it undergoes a transformation into a well-structured format that lends itself to subsequent analysis with maximum effectiveness.Hastie et al. (2009)

- **Feature Extraction**: involves selecting and transforming relevant attributes from the dataset to form a subset of features that capture the essential information needed for predictive modeling. This process helps in reducing dimensionality and improving model performance by focusing on the most impactful variables.

- **Transform Data**: Using {`pd.get_dummies()`} function in Pandas python is a key technique for converting categorical data into numerical format, essential for machine learning preprocessing Pandas. This function transforms each category into a separate binary column, allowing algorithms to interpret and analyze the data effectively.

- **Oversampling**: Oversampling is a technique in signal processing and statistics used to handle missing or reduced-resolution digital signals. In Machine Learning and Data Mining, it is utilized to address the issue of data imbalance by increasing the number of observations in minority classes. This helps machine learning models learn more effectively from the minority classes, reducing bias and improving prediction performance on imbalanced data.In this project we used **SMOTE** (Synthetic Minority Oversampling Technique)P (2002) to generate new synthetic samples by combining nearest neighbors of the minority class.

- **Splitting**: To evaluate the performance of our predictive models accurately, we divided the dataset into two subsets, namely, the training set and the testing set. The training set, which comprised 80% of the data, was used to train our machine learning models. The models used this set as a starting point to discover patterns and connections in the data. The testing set, which was unaltered during the model training phase, was made up of the final 20% of the data.

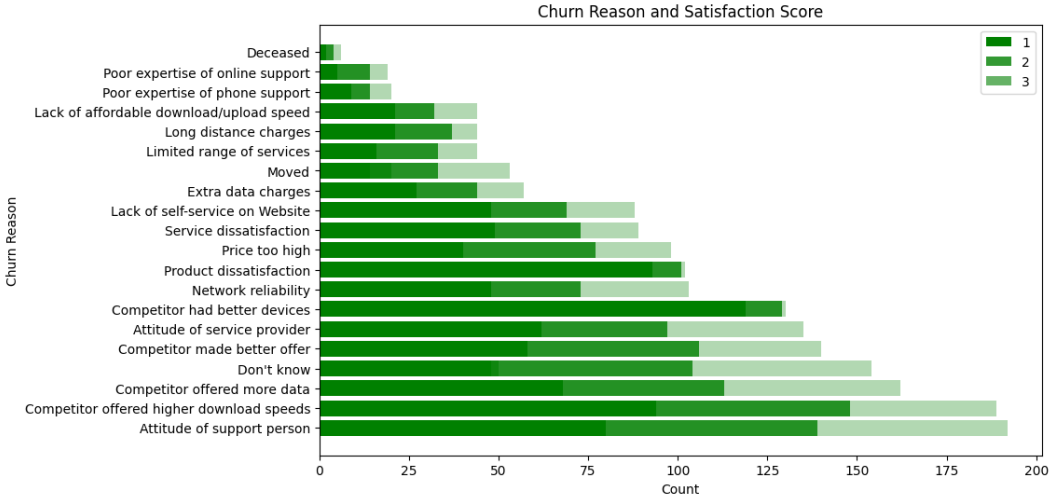

Fig. 2: Churn Reason

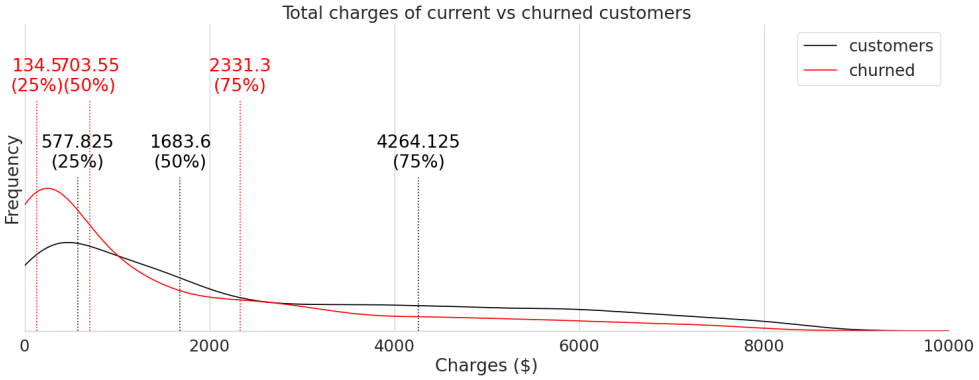

Fig. 3: Total charge distribution

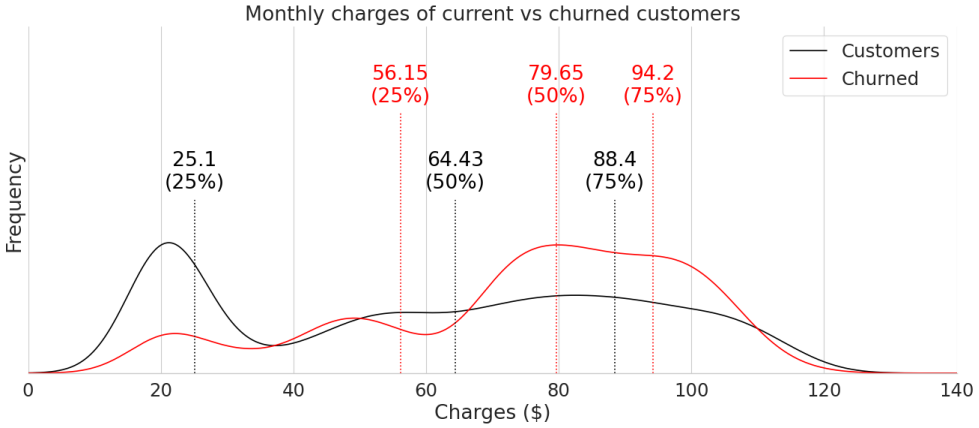

Fig. 4: Monthly Chagre distribution

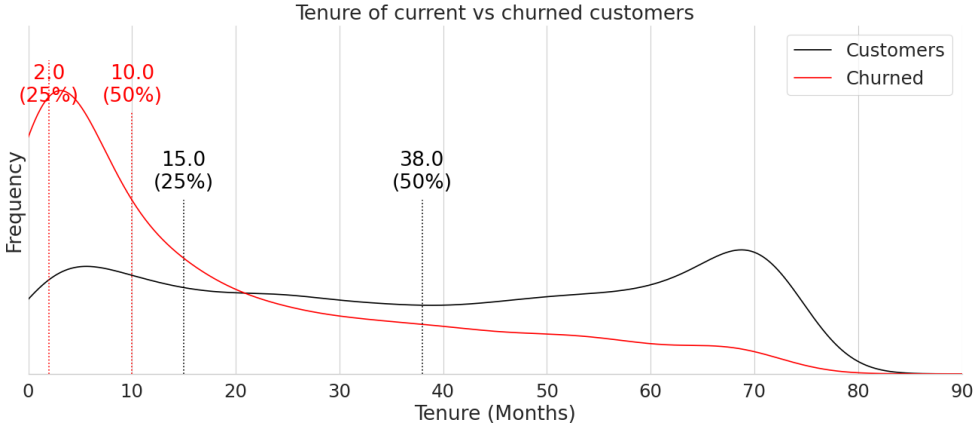

Fig. 5: Tenure distribution

2.3. Model Architecture

To predict customer churn we implemented multiple Machine Learning models. We will compare among these models to choose the best performance.

- KNN(classfication) is an algorithm whose inputs of the k closest training examples in a dataset and outputs are class memberships. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. Cover (1967)

- Logistic Regression Classifier: The machine learning statistical model is employed in this work to perform binary classification tasks. The model's primary objective is to categorize clients based on attributes that have been previously mined from the dataset. By utilizing this approach, the study aims to effectively distinguish and assign clients to specific categories, enhancing the understanding and decision-making process in the given context.Juliana Tolles (2016)

- XGBoost Decision Tree and Random Forest are ensemble of shallower decision trees to help the performance of decision tree algorithm.Random forest "bagging" minimizes the variance and overfitting, while GBDT "boosting" minimizes the bias and underfitting.NVIDIA NVIDIA

3. EXPERIMENTAL RESULTS

To predict customer churn we implemented multiple Machine learning models as shown in Fig [6]

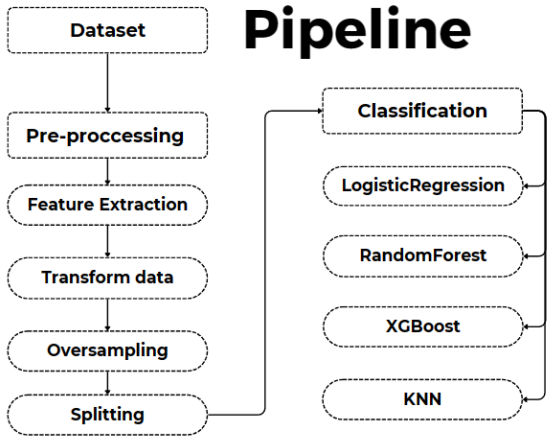

Fig. 6: Pipline

Method	Churn	Precision	Recall	F1-score	Accuracy
LogisticRegression	No	0.81	0.75	0.78	0.79
	Yes	0.77	0.83	0.80	
RandomForest	No	0.87	0.86	0.86	0.87
	Yes	0.86	0.87	0.87	
XGBoost	No	0.86	0.89	0.87	0.87
	Yes	0.89	0.86	0.87	
KNN	No	0.84	0.70	0.76	0.79
	Yes	0.75	0.87	0.81	

Fig. 7: Results

4. CONCLUSION

In conclusion, based on churn reason and satisfaction score, to minimize the rate of churn, business have to increase the friendliness and enthusiasm in supporting services, and improve services to compete with others. Moreover, business need to maximize the tenure of customer as tenure increases, probability of churn decreases and lower the monthly charges as monthly charges increase, the probability of customer churn increases.

Due to Apriori algorithm, combination of phone services, device protection, streaming movies and streaming TV services may encourage more customer using services.

Beside that, various methods were investigated to build a model that aims to predict customer churn on the IBM dataset. In this work, Random Forest and XGBoost achieved the best results with an accuracy of 0.87.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. Proceedings of the 20th International Conference on Very Large Data Bases. In *Fast Algorithms for mining association rules*, pages 487 – 499, Santiago,Chile, 1994. VLDB.

[2] Peter E. Cover, Thomas M.; Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[3] Christopher Garbacz and Herbert G Thompson Jr. Demand for telecommunication services in developing countries. *Telecommunications policy*, 31.5, 2007.

[4] Hastie, Trevor, Tibshirani, Robert, Friedman, and Jerome H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer*, 2009.

[5] IBM. Telecom Churn dataset. https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset, 2024.

[6] William J. Meurer Juliana Tolles. Logistic Regression Relating Patient Characteristics to Outcomes. *JAMA*, 316, 2016.

[7] NVIDIA. RandomForest. https://www.nvidia.com/en-us/glossary/random-forest/, .

[8] NVIDIA. XGBoost. https://www.nvidia.com/en-us/glossary/xgboost/, .

[9] Chawla N. V.; Bowyer K. W.; Hall L. O.; Kegelmeyer W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[10] Pandas. pandas.get_dummies. https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html.