Journal of Computer Engineering & Information Technology

Rapid Communication

Machine Learning Models Developed For Telecom Churn Prediction

Hemlata Jain*

Summary

Data Analytics has become an important topic in the area of Telecom Churn Prediction. Researchers have come out with very efficient experiments for Churn Prediction and have given a new direction to the telecommunication Industry to save their customers. Companies are eagerly developing the models for predicting churn and putting their efforts to save the potential churners. Therefore, for a better churn prediction model, finding the factors of churn is very important. This study is aiming to find the factors of user's churn by evaluating their past service usage details. For this purpose, study is taking the advantage of feature importance, feature normalisation, feature correlation and feature extraction. After feature selection and extraction this study performing seven different experiments on the dataset to bring out the best results and compared the techniques.

Keywords

Machine Learning; Churn Prediction; Random Forest; Feature Importance

Introduction

First Experiment includes a hybrid model of Decision tree and Logistic Regression, second experiment include PCA with Logistic Regression and Logit Boost, third experiment using a Deep Learning Technique that is CNN-VAE (Convolutional Neural Network with Variational Autoencoder), Fourth, fifth, sixth and seventh experiments was done on Logistic Regression, Logit Boost, XGBoost and Random Forest respectively. First four experiments are hybrid models and rest are using standalone techniques. The Orange dataset was used in this technique which has 3333 subscriber's entries and 21 features. On the other hand, these experiments are compared with already existing models that have been developed in literature studies. The performance was evaluated using Accuracy, Precision, Recall rate, F-measure, Confusion Matrix, Marco Average and Weighted Average.

This study proved to get better results as compared to old models. Random Forest outperformed in this study by achieving 95% Accuracy and all other experiments also produced very good results. The study states the importance of data mining techniques for a churn prediction model and proposes a very good comparison model where all machine Learning Standalone techniques,

Deep Learning Technique and hybrid models with Feature Extraction tasks are being used and compared on the same dataset to evaluate the techniques performance better.

Telecom Churn Prediction

In the present world, digital media has become a powerful tool for managing large data especially in the telecom industry where there is an essential need to store large dataset. A huge volume of data is being generated by telecom companies at an exceedingly fast rate [1]. The large data generated in these companies are bulky and managing and accessing the information out of this data is a main challenging task.

Data mining resolves this issue. Data mining is the process of analyzing data from various aspects and summarizes it into valuable information [2]. Since the early 1960 Data Mining techniques have been considered to be an area of applied artificial intelligence [3].

A large number of data mining techniques available to find out the hidden knowledge about the customer data. Some of them are clustering, classification, attribute selection, Association etc. A churn prediction model is purely based on the customers past service usages behavior data. Telecom companies develop churn prediction models to increase their client share, maximize profit and stay active in a competitive environment. A consumer churn is switching from one service provider to another. In today's competitive environment customers have multiple options for better services and prices.

There are multiple reasons for customer churn. Unlike post-paid customers, prepaid customers are not bound to a service provider and may churn at any time [1]. Customer churn normally happens due to lack of engagement, lack of promotions or new offers, lack of customer service support, high call rate or SMS charges, non-payment bills, fraud or miss usages of services and change of location. When the total number of customers are dropping down it causes major revenue loss. Churn Prediction model uses a telecom database for prediction. It analyses customer's behavior and predicts the future churners.

Telecom databases are running into terabytes and petabytes having large numbers of attributes and hence to model these complex datasets it needs advanced data sciences models to be developed. There is a huge advancement in the field of big data and machine learning. Due to that many models have been developed widely. Researchers have developed and compared different machine learning techniques in their models. Research [3] contributed to develop a churn prediction model to assist telecom companies for predicting customers who are near to churn.

This research compared the machine learning techniques that are XGBoost, Decision Tree, Random Forest and Gradient Boosted Machine Tree. This research analyzed the factors which played an important role in customer churn by feature engineering and selection [1]. Also, it is identified by the factors "WHICH LEAD TO CUSTOMER CHURN" by selecting features using information gain, correlation and ranking attributes. These researches proved factor identification is useful for churn prediction models. Research [4] proved that data preparation techniques they choose affects the churn prediction model performance and enhances Logistic Regression is competitive with advanced single ensemble data mining techniques [5] have

shown that customer misclassification, the amount of service they used and some demographic attributes plays an important role in customer churn. This research used binomial Logistic Regression for the prediction.

Research [6] used different data mining techniques for churn prediction and compared them. For comparing the different techniques this research used different evaluation metrics and also worked on extracting datasets features. DT handles interaction effects between variables very well but has difficulties to handle linear relations between variables. For LR the opposite is true: it handles linear relations between variables very well but it does not detect and accommodate interaction effects between variables [7].

This study added multiple functionalities of feature engineering and selection at one place and worked on improving the model performance. This study used literature models and identified some new work and applied multiple feature analysis tasks to improve performance and at last compare them with each other and with literature works. This study using correlation matrix, feature engineering, feature importance, handling categorical feature, handling continuous features, normalizing features and giving this altered and informative dataset to four different hybrid models and to five standalone techniques.

Some hybrid models already have some feature extraction functionality in it therefore it added the double feature extraction capability to the model. The idea of comparing hybrid techniques and standalone techniques is very helpful for future research and the double feature selection process really worked on improving performance.

This paper is organized in following sections: section II: Literature review highlighting work already done by researchers; Section III briefly describes methodologies leveraged in this study. In Section IV Proposed work and database are detailed while in section V Results and discussion are discussed, section VI is the conclusion of this paper detailing what the author has accomplished and what is planned in future.

References

1. Irfan U, Basit R, Ahmad KM, Muhammad I, Saif U, et.al. (2019) A Churn Prediction Model Using Random Forest Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector.
2. Nadeem AN, Umar S, Shahzad S (2018) A Review on Customer Churn Prediction Data Mining Modeling Techniques. Indian J Sci Tec 11(27):1-7.
3. Abdelrahim K, Assef J, Kadan A (2019) Customer churn prediction in telecom using machine learning in big data platform", journal of BigData.
4. Kristof C, Stefan L, Geert V (2017) A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry.
5. Abbas K (2011) Churn analysis for an Iranian mobile operator. Telecommunications Policy 35(4)344-356.
6. Adnan A (2017) Customer Churn Prediction in Telecommunication Sector using Rough Set Approach.
7. Arno D (2018) A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269 (2) 760-772.

Author Affiliation   **Top**

Department of Computer Science, Poornima University, Jaipur, India.