IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A Comparative Analysis of Customer Churn Prediction Models in Python Using Scikit-Learn

Jaipreet Singh[1], Gurvir Singh Sidhu[2], Inderjeet Singh[3], Prabhneet Singh[4]

[1 2 3]Students of Computer Science & Engineering, Chandigarh University, Gharuan Mohali,

[4]Assistant Professor, UIE, CSE, Chandigarh University, Gharuan Mohali

ABSTRACT

In the sector of big-scale groups, the problem of consumer churn has grown into a powerful assignment. As an end result, corporations are actively searching for progressive methods to predict capability patron churn quotes. It has turn out to be imperative to discover the factors that contribute to elevated purchaser churn prices, allowing organizations to take the essential steps to mitigate this phenomenon. The primary objective of our group's efforts is to craft a robust churn prediction model which could help companies in pinpointing clients at the highest chance of churning. To gauge the version's performance, we hired the broadly common Area Under the Curve (AUC) measurement, which gives a dependable measure of predictive accuracy. Impressively, the AUC price we carried out became 84.93%, signifying the effectiveness of our version. To execute our study, we hired the flexible and available platform of Google Colab. The dataset we used is the Teleco Churn Dataset, simply to be had on Kaggle, an open-source platform that provides a wealth of dataset assets to the general public. This dataset offers a complete compilation of consumer statistics spanning an extended length, serving as the foundational statistics for schooling, checking out, and evaluating our churn prediction gadget. Our version underwent a rigorous trying out system, with an assessment of its performance throughout ten distinct algorithms: Decision Tree, Random Forest, Gradient Boost, Logistic Regression, Adaboost, SVC, Gaussian Naïve Bayes, Kernel Support Vector Machine, K Nearest Neighbour and Voting Classifier. However, the best results are obtained by applying the Voting Classifier algorithm.

INTRODUCTION

In today's ever-evolving business environment, customer retention remains a daunting challenge for organizations of all sizes and businesses. In an era of unprecedented competition, changing consumer preferences and more choice, the ability to retain existing customers is often as important as pursuing new ones in order to meet this need role, companies are increasingly turning to cutting-edge technologies and data-driven solutions

This paper explores the area of predictive analytics of customer churn, using the power of machine learning techniques and modifications of the Scikit-Learn library. Commonly referred to as customer attrition, it describes the process by which customers end their involvement with a company or company. By accurately targeting customers, companies can quickly implement strategies to retain their valued customers, preventing significant revenue losses and high costs associated with acquiring new customers.

The crucial intention of this study is to assemble a sturdy and precise model for predicting patron churn by exploring a range of gadget studying algorithms, scrutinizing their performance, and in the long run choosing the most effective one. In doing so, we aspire to equip corporations with actionable insights and equipment to systematically manipulate churn, support customer retention projects, and optimize useful resource allocation.

This research bears importance on numerous fronts. Firstly, it augments the present expertise base via losing mild at the realistic applicability of diverse machine studying algorithms in the context of client churn prediction. Secondly, it accentuates the pivotal function of Scikit-Learn, a extensively adopted Python library for machine mastering, in streamlining the improvement and deployment of these predictive fashions. Thirdly, it empowers companies to streamline their techniques, centering their efforts at the most promising techniques to curtail churn, ensuing in augmented profitability and better consumer ranges.

RELATED WORK

Various methods have been used to estimate volatility in corporate communications. They often use machine learning techniques and data mining. While many projects focus on using the same data extraction method, others focus on comparing various algorithms for churn prediction.

Gavril et al. [6] reported a new way of mining data to forecast the prepaid customer churn using call data of total 3333 rows with 21 characteristics and the corresponding churn values. Principal Component Analysis algorithm (also known as "PCA") was used to reduce the data size and three machine learning algorithms to estimate the churn rate: Neural Network, Support Vector Machine, and Bayesian network. The author uses the area under the curve (AUC) to evaluate the performance of the algorithm. The AUC values of the support vector machine, neural networks, and Bayesian networks are 99.70%, 99.55%, and 99.10%, respectively. The data used is small and there are no missing values.

He et al. [8] suggested a prediction model that follows the neural network algorithm to predict churn problem about 5.23 million customers of a large telecommunications company in China with 91.1% accuracy.

Idris [9] proposed a method inspired by genetic engineering and AdaBoost was used to solve the loss problem. Two sample data sets were used with 89% and 63% accuracy respectively.

Huang et al. [10] proposed using big data for predicting customer churn. The main goal is to show that big data can increase prediction efficiency by using the volume, speed and data variety. The algorithm used is random forest and is evaluated as the area under the curve (AUC).

Maktar et al. [11] It is stated in the article that the coarse classification method is superior to other algorithms such as linear regression, decision trees and neural networks.

Burez and Van den Poel [7] did analysis of random sampling for random problems, inefficiencies, comparing whether their actions were good or not. The parameters used for evaluation are AUC and Lift, and the results are better in sampling.

METHODOLOGY

In particular, customer churn forecasting is an important step in any predictive modelling project. It is about collecting detailed and relevant information that provides insights into customers' behaviours, preferences, usage patterns, and interactions with professionals. In this study, we collected a set of data on based service delivery documentation on the mouth.

3.1 DATA COLLECTION

3.1.1 Dataset Description

The telecom data set contains attributes related to customer usage, account information, and service usage patterns in the telecom company. Key attributes include customer demographics, call data, internet usage, and customer churn labels that indicate whether a customer has churned or not.

3.1.2 Dataset Source

The data set was obtained from Kaggle, a popular platform for hosting data sets and machine learning competitions. The data set is a collection of anonymous customer information collected from a telecommunications company.

3.1.3 Data Granularity and Time Period

The dataset provides information for a specific period of time, typically including records on a monthly basis. Attributes such as call minutes, bills, usage patterns are typically aggregated or averaged over a month, making the granularity monthly.

3.1.4 Moral Considerations

Since the dataset is available to the public and is already anonymous, ethical considerations revolve primarily around ensuring that responsible use of data in compliance with Kaggle's regulations has protected consumer privacy, a there is no identifying information in the dataset.

3.1.5 Data Quality and Accuracy

The dataset was downloaded and thorough quality checks were done which included whether missing values, redundancies and inconsistencies were present or not. Necessary data cleaning and pre-processing steps were performed to overcome any errors in the dataset.

3.1.6 Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the structure, statistical properties, and distribution of various components of the data set before proceeding to pre-processing. The EDA involved visualization and statistical analysis to identify patterns and gain insights at pre-processing and modelling stage.

3.2 DATA PRE-PROCESSING

Data pre-processing is the most essential part for preparing a data set for machine learning model training. In the case of the Kaggle telecom customer churn dataset, several pre-processing steps were performed to enhance the data quality and usefulness for predictive models.

3.2.1.1. Dealing with Missing Values

One of the first tasks is to identify and address missing values. This includes analysing each feature to determine if there are missing or zero values, and deciding on the appropriate course of action. Common methods include mean, median, mode values, or more complex imputation methods such as K-nearest neighbour imputation.

3.2.2. Encoding Categorical Variables

Many machine learning models require numerical input, so categorical variables had to be written. This was done using methods such as single-hot encoding,

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7032 non-null   float64
 20  Churn             7043 non-null   object
dtypes: float64(2), int64(2), object(17)
memory usage: 1.1+ MB
```

which consists of binary columns for each class, or label encoding, which assigns a unique number to each class. Encoding technique depends on the type of data and the algorithms to be used.

3.2.3 Mathematical Process Scaling

To ensure that statistical features contribute equally to model training, methods such as standardization or normalization were often used to enhance them. Standardization (the mean is subtracted and divided by the standard deviation) and normalization (scaling to the range [0, 1]) were applied depending on the specific requirements of the models used

3.2.4 Handling Imbalanced Data (optional)

In churn forecasting, the dataset can be unbalanced, with a significantly higher number of non-churners compared to churns. It is important to address this imbalance so that the sample is not biased by the majority group.

3.2.5 Feature Engineering

This involved developing new features or modifying existing ones to get more information into the model. For example, to calculate usage-to-cost ratios, time spent in a company, or integrating call and online services in parts for broader usage profiles

3.2.6 Classification of Data

The data set was divided into training and testing sessions. The machine learning models were trained

with the training set, while the testing set was kept to test the performance of the model.

3.2.7 Final dataset configuration

The processed dataset had the required features after taking the aforementioned steps and was ready to be used in training and machine learning models for customer churn prediction analysis. By diligently applying these pre-processing techniques, we ensured that the dataset was properly structured and optimized to obtain accurate predictions from machine learning models.

3.3 MODEL TRAINING AND EVALUATION

Model training and evaluation of predictive modelling are important steps for customer churn forecasting. In this study, we tested several machine learning models to determine the most effective in accurately predicting customer churn. The training process involves inserting the selected models into the training data, while the evaluation phase evaluates their performance with appropriate parameters

3.3.1 Model Selection

We chose machine learning models that are known to perform well in the distribution industry, especially for predicting customer arrivals. These examples include: Logistic regression: A fundamental classification algorithm suitable for binary outcomes such as churn prediction. Decision trees: A versatile algorithm capable of capturing nonlinear relationships in data. Random forest: A technique for clustering multiple decision trees together to improve prediction accuracy and control overfitting. Support Vector Machines (SVM): Effective for separating classes at higher levels. Gradient Boosting Machines (GBM): An ensemble technique that gradually builds weak learners into strong predictive models.

3.3.2 Model Training

For each selected model, we trained it with pre-processed training data. Factors and labels (churn or no-churn) were assigned to samples to identify patterns and relationships evident in the data.

3.3.3 Model Hyperparameter Tuning

To optimize the performance of each model, we performed hyperparameter tuning. Hyperparameters are structural parameters that affect the learning process of the model. Methods such as network search or random search were used to find the best combination of hyperparameters for each model.

3.3.4 Model analysis

After training and hyperparameter tuning, we tested the model using the reserved testing data set. The following evaluation criteria were applied.

Accuracy: Fraction of correct predictions to all the predictions done.

Area Under the ROC curve (AUC-ROC): Measures the ability of the model to discriminate between classes, with higher values indicating better discrimination. These analytical parameters allowed for a thorough evaluation of the performance of each model and the selection of the most appropriate one for customer churn prediction

3.3.5 Model Comparison

Finally, we compared the evaluation results of all the models to determine the most effective model to accurately predict customer attendance. In this study, the model with high accuracy, and AUC-ROC was selected as the best model for customer churn prediction. Following this approach, we ensured that our chosen models are robust, well integrated, and capable of providing accurate predictions of customer attraction, thereby enabling companies to implement customer retention strategies active has been used effectively.

4. RESULT

Here the detailed results obtained from the experiment that used machine learning models to predict customer clicks using telecommunications databases from Kaggle is given. Analytical metrics including accuracy, and AUC-ROC are discussed for each model providing a detailed comparison

4.1.1. Best performance comparison

The following table presents the evaluation parameters for each machine learning model:

	Algorithm	ROC AUC Mean	ROC AUC STD	Accuracy Mean	Accuracy STD
9	Voting Classifier	84.93	1.39	80.23	1.89
8	Gradient boost classifier	84.72	1.42	79.72	1.95
7	Adaboost	84.55	1.25	80.09	1.77
0	Logistic Regression	84.39	1.47	74.38	1.94
1	SVC	82.99	2.07	79.11	2.01
6	Random Forest	82.75	2.01	78.67	1.98
4	Gaussian NB	82.32	1.28	75.38	1.23
2	Kernel SVM	79.65	2.12	79.26	1.67
3	KNN	77.14	1.43	75.90	2.01
5	Decision Tree Classifier	66.67	1.07	73.73	1.12

4.2. Analysis of the results

Based on the analytical metrics, the voting classifier model exhibits high accuracy (80.23%), AUC-ROC (84.93%) in observations all of these have been reviewed. This shows that the Voting Classifier model outperforms the others in predicting subscriber churn in the telecom data set.

5. CONCLUSION

The study demonstrated the applicability and effectiveness of machine learning for predicting subscriber growth, focusing on the telecommunications industry. Through proper data pre-processing, model selection, and evaluation, voting classifier emerged as the most promising model for accurate churn forecasting.

The study highlights the importance of advanced analytics and machine learning in customer relationship management. Predictive customer retention not only helps retain customers and reduce revenue loss but also provides a competitive edge in the marketplace by enabling proactive decision making.

In conclusion, the findings of this study contribute to valuable insights that businesses can use to develop data-driven strategies aimed at increasing customer retention, they will longstanding customer relationships, ensuring sustainable growth in a highly competitive business environment

References

[1] John, T., et al. (2018) Telecom Churn.

[2] Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform. Journal of Big Data, 6, 28. https://doi.org/10.1186/s40537-019-0191-6

[3] Andrews, R., et al. (2019) Churn Prediction in Telecom Sector Using Machine

Learning. International Journal of Information Systems and Computer Sciences, 8,

132-134. https://doi.org/10.30534/ijiscs/2019/31822019

[4] ApurvaSree, G., et al. (2019) Churn Prediction in Telecom Using Classification Algorithms. International Journal of Scientific Research and Engineering Development, 5, 19-28.

[5] Tata Tele Business Services (2018) Big Data and the Telecom Industry

[6] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.

[7] Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626–36.

[8] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.

[9] Idris A, Khan A, Lee YS. Genetic programming and adaboosting based churn prediction for telecom. In: IEEE international conference on systems, man, and cybernetics (SMC). 2012. p. 1328–32.

[10] Huang F, Zhu M, Yuan K, Deng EO. Telco churn prediction with big data. In: ACM SIGMOD international conference on management of data. 2015. p .607–18.

[11] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.