

# Customer Attrition Detection Using the LGBM Model

Jie Huang\*

School of Management, Guangzhou University, Guangzhou, Guangdong, 510006, China

**Abstract.** In internet service industries, such as competitive industries, it costs more to attract new consumers to become customers of the company than saving the consumers who already are customers. Therefore, detecting the running off customers and finding a way to keep the customers from leaving is extremely important. This study addresses the problem of customer attrition in the internet service industry by choosing the best-performing model to detect the customers who are going to run off in advance. To select the most suitable model for accurately detecting customer churn, this study performs preprocessing, including data cleaning, feature engineering, and feature selection. The dataset is then split into training, testing, and validation sets. Various models are built and evaluated based on their performance, measured by calculating the mean and standardized values of the detection rate. The result is that the Light Gradient Boosting Machine (LGBM) model has superior performance in detection rate scoring.

## 1 Introduction

Customer attrition refers to the situation where customers either entirely stop paying for the service the company applies or select another internet service company that applies the same kind of service. In highly competitive markets, retaining existing customers is often more cost-effective and strategically important than acquiring new ones. Consequently, Internet service companies must detect customer attrition in advance so that these companies can react promptly to prevent customers from stopping paying for the service or choosing other Internet service companies.

In Luis's study, the models Light Gradient Boosting Machine (LGBM) and Extreme Gradient Boosting (XGB) are the best-performing individual models both without feature selection and with feature selection circumstances. According to Luis's study, the accuracy of the LGBM model without feature selection is 97.7%, and the accuracy of the LGBM model with feature selection is 97% [1]. In Temesgen's study, testing with the churn modeling data set from Kaggle, the LGBM model gets 0.955 in accuracy, 0.481 in recall, and 0.993 in AUC which is better than other models and gets 0.893 in precision which is slightly lower than the Random Forest (RF) model's 0.904 when dealing with unbalanced data set. In the class-balanced data set, the LGBM and the XGB models perform the best, the LGBM model shows its superior performance and reliability [2]. In Linda's study, dealing with a data set

---

\* Corresponding author: 3226560009@e.gzhu.edu.cn

containing data on bank customer attrition, which includes 1,750,036 customer data, the LGBM model gets 0.8789 in accuracy, 0.8978 in precision, 0.8553 in recall, 0.8758 in f1 score, and 0.9694 in AUC, which shows excellent performance [3].

In summary, the study builds different models after doing preprocessing and assesses the model's performance by calculating their detective rate. The purpose of the study is to find a model that performs well in detecting running-off customers and inform the companies in advance so that they have time to react.

## 2 Research method

### 2.1 Data

The name of the used data set is Telco Customer Churn. The data set is from IBM (International Business Corporation) and posted in Kaggle by Blast Char. The column called churn describes the customers who left. The data set contains 21 rows and 7043 columns, and there is no null value in the data set. The data set consists of numeric data such as tenure, monthly charges, total charges, and senior citizen and categorical data such as customer ID, gender, partner, and so on.

The article did target encoding to change categorical variables that have only two types to a numeric binary value and one-hot encoding to create all possible combinations of 2D categorical variables to build variables from categorical fields. The article uses a Univariate filter and calculates the KS score and CDR to select the top 20% of features, which is 73 features. Then, it uses a Multivariate wrapper LGBM Classifier and SFS to do Stepwise Selection to select the top 10 features to do feature selection and visualize it using PCA and TSNE.

### 2.2 Model

Principal Component Analysis (PCA) is a method to reduce the dimension curse of the dataset. It will promote interpretability and won't lose very much information. To achieve this, before doing PCA, people will create new covariates that are not related to each other [4]. TSNE technique visualizes high-dimensional data by reducing the high-dimensional data into two- or three-dimensional data. TSNE performs well in creating a single map to show structure at many different scales [5].

To solve the problem of customer churn detection, the author explores different models, including Logistic Regression (LR), Decision Tree (DT), RF, LGBM, LGBM with smote, Neural Network, Cat boost, and XGB to find out the best-performing one. LR is a statistical analysis method that constructs a statistical model to describe the relationship between a binary or dichotomous (yes/no type) outcome (dependent or response variable) and a set of independent predictors or explanatory variables [6].

DT is a classifier that splits data by selected features and builds a tree-like model. The DT model's tree-like structure makes its decision logic easily explained and can be trained efficiently and find nonlinear relationships through its simple principle. The RF model constructs more than one DT and assembles the different predicted results. The RF model lowers the risk of overfitting by getting its result from several DTs and keeps the advantages of the model DT, as mentioned above.

The LGBM model is based on the Gradient Boosting Framework and achieves efficient training and high-accuracy predictions by combining histogram algorithm, Gradient-based One-Side Sampling (GOSS), Exclusive Feature Bundling (EFB), and leaf-wise growth strategy. The LGBM model supports large-scale data processing and high prediction

accuracy, especially performing well in High-dimensional Sparse Data Scenarios. The LGBM model with Smote uses Synthetic Minority Over-sampling Technique (SMOTE) to create samples of minorities to balance the unbalanced data set and combine it with the model LGBM. The LGBM with smote model performs great in detecting minorities and keeps the advantages of the LGBM model mentioned above. The Neural Network model constructs nonlinear mapping by multiple layers of neurons, uses the Backpropagation Algorithm to optimize weight learn data features, and achieves prediction by minimizing the Loss Function.

The Cat Boost Model deals with Categorical Features by Symmetric Tree and Ordered Target Encoding and combines with a Gradient Boosting Framework. The Cat Boost model doesn't need to preprocess the categorical data and lowers the risk of overfitting. Based on the gradient boosting framework, XGBoost optimizes the objective function using second-order Taylor expansion, incorporates regularization terms (L1/L2) to control model complexity, and employs the Weighted Quantile Sketch to accelerate feature splitting, achieving efficient training and high-accuracy predictions. XGBoost model supports efficient computation and sparse data handling, which means the XGBoost model performs well in training large-scale and high-dimension data and prediction.

Finally, the article uses the ROC curve and calculates the KS score of the training set, testing set, and validation set to evaluate the chosen model's performance. To get the ROC curve, the article separates the data sets into 100 bins in a positive sequence of its churn score and calculates the TPR and FPR of each bin. Additionally, the article also calculates the KS score and makes the ROC curve, which estimates the finally selected model's performance.

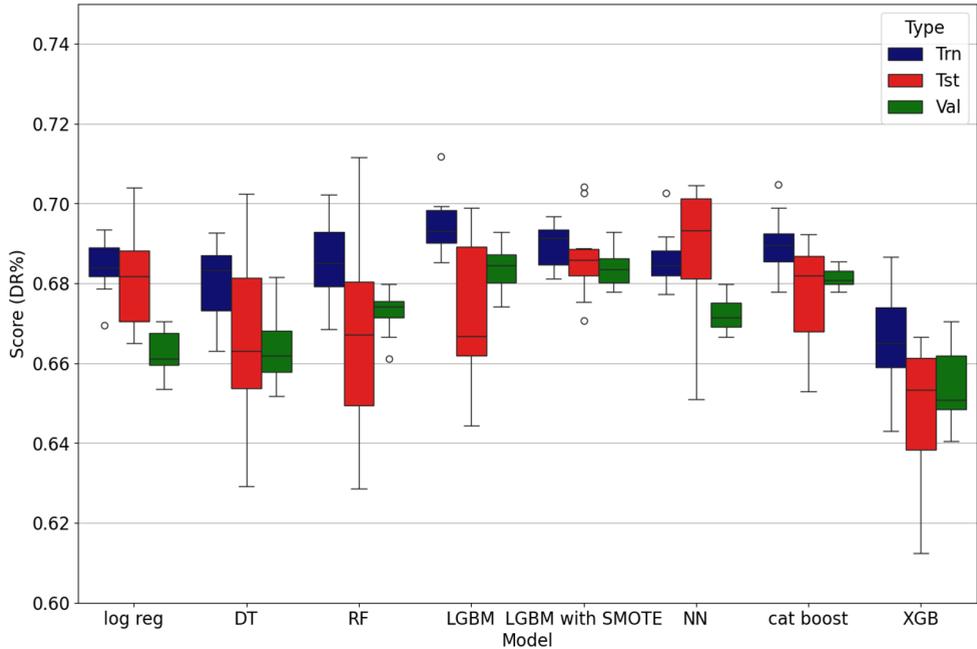
### 3 Result

After building different model algorithms, tuning their hyper-parameters, doing iteration training, and calculating the mean and standardization value's detection rate of customer attrition of the training set, testing set, and validation set, the article selects the LGBM model. As shown in Table 1 and Fig. 1, the LGBM model is relatively high in the mean value and relatively low in the standardization value, meaning that the LGBM model has high generalization ability, high stability, fast operation speed, free of unnecessary calculates and perform well enough in the detection rate of customer attrition.

LGBM is a tree-based algorithm that is faster in responding than another model algorithm as it proceeds in a vertical level. LGBM performs well when we are dealing with a large data set, and it can provide results accurately, so it requires very little memory to compute thousands of rows [7]. Using the LGBM model to deal with the customer attrition data set, the article finds that the company should lower service costs for about the top 30% of the accounts, and the max possible saving is 64,953.073974\$.

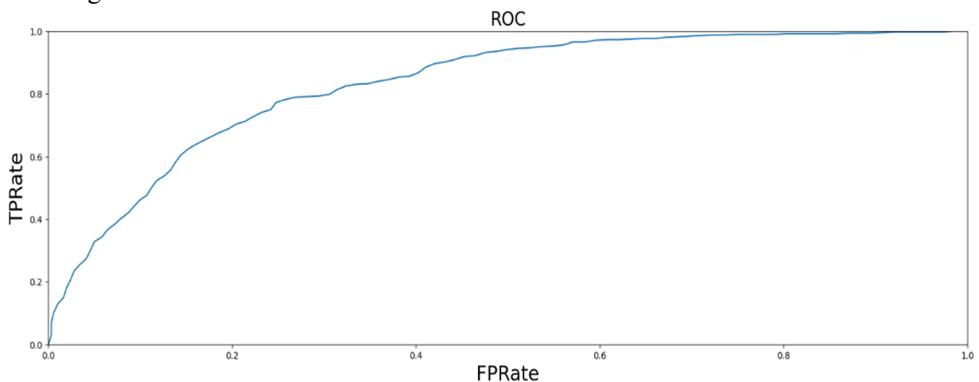
**Table 1.** Detective rate of different models

model	Trn		Tst		Val	
	mean	std	mean	std	mean	std
DT	0.680057	0.009619	0.666548	0.021535	0.663483	0.008697
LGBM	0.694587	0.007693	0.671904	0.018815	0.68427	0.005666
LGBM with smot	0.6896	0.005523	0.68655	0.010500	0.683708	0.004453
NN	0.686156	0.007101	0.688859	0.016652	0.672097	0.004089
RF	0.685829	0.011217	0.667933	0.024890	0.672846	0.005516
XGB	0.665352	0.012840	0.648751	0.017410	0.653933	0.010019
Cat Boost	0.689919	0.007772	0.67738	0.013283	0.681461	0.002566
Log Reg	0.684019	0.006774	0.680894	0.012068	0.66236	0.005656



**Fig. 1** Box plot of detective rate of different models (Photo/Picture credit: Original).

To detect the model LGBM model's performance, the article calculates the KS score and uses the ROC curve of the model to estimate the LGBM model's performance, and the KS score of the model comes to 54.46 for the training set, 51.41 for the testing set, and 52.39 for the validation set. According to the ROC curve shown in Fig. 2, it can be seen that the curve is far from the diagonal line of Fig. 2, which shows the model does well in the problem of detecting customer attrition.



**Fig. 2** ROC curve of LGBM model (Photo/Picture credit: Original).

## 4 Discussion

However, there are still some problems with the article. Firstly, when building different model algorithms, the article only did LGBM with smote but didn't do RF with smote, DT with smote, and so on, so the article can't promise the model LGBM is the best-performing one without considering other model algorithms with smote. Additionally, although smote can deal with the unbalanced data effectively, smote is sensitive to outliers and may cause

the problem of overfitting. According to these problems, the article can use the SMOTE Tomek or ADASYN and combine it with other models to weaken the influence of the outliers.

The main idea of ADASYN is to consider the difference of different minority class examples in difficulty to use a weight distribution and generate more data for minority class examples that are harder to learn so it can reduce the bias from imbalanced data [8]. In Stephen Hasson's study, the LGBM with SMOTE Tomek is the best-performing model in churn prediction and gets 97.92% in precision, 95.25% in recall, and 96.57 in f1 score [9]. In the study, the model LGBM performs better in the standardization value and mean value of detective rate than other model algorithms, which means the model algorithm LGBM would work better in customer churn detection, which is the same as Risuna Nkolele and Linda Wahyu Widiyanti's result [3, 10].

However, the model LGBM has its disadvantages; for example, the model LGBM is slightly insufficient in the aspect of preprocessing. Probably because its Built-in Processing Methods are not as robust as the Cat Boost model's Ordered Target Encoding, which means that before using the model, the data should be done one-hot encoding manually, and the cost of preprocessing increases.

## 5 Conclusion

After doing data description and data cleaning, the study does feature engineering by creating expert variables, creating variables from numeric fields, and creating variables from categorical fields, which includes target encoding and one-hot encoding. After building models including LR, DT, RF, LGBM, LGBM with smote, Cat Boost, XGBoost and NN and testing the model's performance, the study chooses the LGBM model, which is better in the mean value and standardization value of detective rate. To further test the performance of the LGBM model, the study calculates the KS score and makes the ROC curve of the model.

In the Internet service industry, retaining customers is easier and less costly than attracting new consumers, so detecting customer attrition precisely means a lot to the Internet service industry. As the result of the study shows, the performance of the model LGBM in analyzing customer attrition is brilliant, and the model algorithm can effectively detect those customers who are at high risk of attrition. With the great use of the study result, the companies offering internet service can promptly and accurately detect the customers who are going to run off and implement a customer saving scheme.

When building different model algorithms, the study only did LGBM with smote but didn't do RF with smote, DT with smote, and so on. Additionally, when selecting the best performing model, the article uses the indicator Detection rate without using the index f1 score, accuracy or recall which is more commonly used indicators to evaluate different model algorithm's performance and this may underestimate other model's performance. Considering the probability of underrating other models' performance, the study can add the index of F1 score accuracy to evaluate the performance of different models.

The data set used is unbalanced, and it may cause problems in the model's detective rate. Though the LGBM with the smote model's performance is inferior to the model LGBM, the study may use another algorithm to deal with an unbalanced data set to combine with the model LGBM and find a better solution to the detection of customer attrition.

## References

1. L. Tejada-Vicente, D. Rosado-Oliden, D. Mauricio-Santos, Prediction of telecommunications customer churn based on hybrid machine learning and deep

- learning algorithms. In: 2024 IEEE XXXI Int. Conf. Electronics, Electrical Eng. Computing (INTERCON), pp. 1-6 (2024)
2. T. Asfaw, Customer churn prediction using machine-learning techniques in the case of commercial bank of Ethiopia. *Sci. Temper* **14**(3), 618–624 (2023)
  3. L. W. Widiyanti, A. S. B. Karno, W. Hastomo, A. N. Utomo, D. Arif, I. S. K. Wardhana, D. Strydom, Improved banking customer retention prediction based on advanced machine learning models. *Indones. J. Inf. Syst.* **7**(2), 178–193 (2025)
  4. B. M. S. Hasan, A. M. Abdulazeez, A review of principal component analysis algorithm for dimensionality reduction. *J. Soft Comput. Data Mining* **2**(1), 20–30 (2021)
  5. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11) (2008)
  6. A. Das, LR. In: *Encyclopedia of Quality of Life and Well-Being Research*, pp. 3985-3986. Springer Int. Publ. (2024)
  7. R. M. Aziz, M. F. Baluch, S. Patel, A. H. Ganie, LGBM: a machine learning approach for Ethereum fraud detection. *Int. J. Inf. Technol.* **14**(7), 3321–3331 (2022)
  8. H. He, Y. Bai, E. A. Garcia, S. Li, Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE Int. Joint Conf. Neural Networks (IEEE World Congr. Comput. Intell.), pp. 1322-1328 (2008)
  9. S. Hasson, Evaluation and implementation of machine learning models to predict customer churn in the telecommunications sector (2024)
  10. R. Nkolele, H. Wang, Explainable machine learning: a manuscript on the customer churn in the telecommunications industry. In: 2021 Ethics Explainability Responsible Data Sci. (EE-RDS), pp. 1–7 (2021)