Journal on Education
Volume 05, No. 01, September-November 2022, pp. 1189-1197
E-ISSN: 2654-5497, P-ISSN: 2655-1365
Website: http://jonedu.org/index.php/joe

Implementation of Data Mining for Churn Prediction in Music Streaming Company Using 2020 Dataset

Christian Hardjono[1], Sani Muhamad Isa[2]

Bina Nusantara University, Jl. K H. Syahdan No. 9, Kemanggisan Village, Palmerah District, West Jakarta, Indonesia
christian.hardjono@binus.ac.id

Abstract

Customer is an important asset in a company as it is the lifeline of a company. For a company to get a new customer, it will cost a lot of money for campaigns. On the other hand, maintaining old customer tend to be cheaper than acquiring a new one. Because of that, it is important to be able to prevent the loss of customers from the products we have. Therefore, customer churn prediction is important in retaining customers. This paper discusses data mining techniques using XGBoost, Deep Neural Network, and Logistic Regression to compare the performance generated using data from a company that develops a song streaming application. The company suffers from the churn rate of the customer. Uninstall rate of the customers reaching 90% compared to the customer's installs. The data will come from Google Analytics, a service from Google that will track the customer's activity in the music streaming application. After finding out the method that will give the highest accuracy on the churn prediction, the attribute of data that most influence on the churn prediction will be determined.
Keywords: Churn Prediction, XGBoost, Deep Neural Network, Logistic Regression, Data Mining.

Abstrak

Pelanggan merupakan aset penting dalam sebuah perusahaan karena merupakan nyawa dari sebuah perusahaan. Bagi perusahaan untuk mendapatkan pelanggan baru, itu akan menghabiskan banyak uang untuk kampanye. Di sisi lain, mempertahankan pelanggan lama cenderung lebih murah daripada mendapatkan pelanggan baru. Karena itu, penting untuk dapat mencegah hilangnya pelanggan dari produk yang kita miliki. Oleh karena itu, prediksi churn pelanggan penting dalam mempertahankan pelanggan. Makalah ini membahas teknik data mining menggunakan XGBoost, Deep Neural Network, dan Logistic Regression untuk membandingkan performa yang dihasilkan menggunakan data dari perusahaan pengembang aplikasi streaming lagu. Perusahaan menderita dari tingkat churn pelanggan. Tingkat uninstall pelanggan mencapai 90% dibandingkan dengan instalasi pelanggan. Data tersebut akan berasal dari Google Analytics, sebuah layanan dari Google yang akan melacak aktivitas pelanggan di aplikasi streaming musik tersebut. Setelah mengetahui metode yang memberikan akurasi tertinggi pada prediksi churn, akan ditentukan atribut data yang paling berpengaruh terhadap prediksi churn.
Kata kunci: Prediksi Churn, XGBoost, Deep Neural Network, Regresi Logistik, Data Mining

Corresponding author: Christian Hardjono
Email Address: christian.hardjono@binus.ac.id (Bina Nusantara University, West Jakarta, Indonesia)
Received 20 December 2022, Accepted 26 December 2022, Published 30 December 2022

INTRODUCTION

Customers are one of the important aspects of a company. A company costs a lot of money just to get a customer. Compared to getting new customers, retaining existing customers costs much less. A song streaming company that has been around since 2010 and has over five million installs on the Google Play Store has high number of installs but accompanied by a high number of users who are not active, becoming churn, and then uninstall the app. The company define users who churn as users who last login on more than 30 days from current date. The average monthly install of the application is 369,917 installs, but on the same time there will be 330,051 uninstall each month. The high number of users who uninstall causes the company's revenue to decrease. Realizing this, the company is looking for ways to retain existing customers through data mining technology.

Company that generates a lot of data can utilize its data to a lot of uses with data mining. One of its uses is customer churn prediction. Previously, data mining to detect customer churn was mostly done by telecommunication companies (Keramati et al., 2014). By using methods like Decision Tree, Artificial Neural Network, K-Nearest Neighbour, and Support Vector Machine, researchers predict customer churn rates. The results of this study can be used to make business decisions to be able to retain customers who are going into churn status. Data mining also can identify which factor is the main reason for customer to go into churn (Ullah et al., 2019).

This research is intended to detect signs of customer churn in music streaming company using data from Google Analytics, a service that will track behaviour of users in the application. The data is user level data from January 1st, 2020 – December 31st, 2020. The experiment will be conducted using three methods, XGBoost, a method used SyriaTel reaching 93.301% AUC (Ahmad et al., 2019), Deep Neural Network, a popular classifier algorithm (Yu et al., 2017), and Logistic Regression, another popular algorithm in customer churn prediction with strong predictive performance and good comprehensibility (De Caigny et al., 2018). It is hoped that with this research, we can find the right data mining method on existing customer datasets. After obtaining the method with the best performance on this research, attributes from the data that most affect the churn rate from customers will be weighted.

Studies for churn prediction can use customer data from various company like telecom (Hung et al., 2006), landline (Huang et al., 2012), internet service provider (Liao & Chueh, 2011), or even a bank (Shirazi & Mohammadi, 2019). Even another data source like online media (E.-B. Lee et al., 2017) or game log from a game company (E. Lee et al., 2018) that rarely contain personal data of the customer can be used to predict customer behaviour that’s starting to churn. Aside from customer’s likelihood of churning, employee’s likelihood of churning in a company can also be predicted (Yiğit & Shourabizadeh, 2017).

Based on those various datasets, studies for churn prediction used a lot of algorithms to use as a comparison. For example, Dolatabadi used decision tree, naïve bayes, SVM and neural network which resulted in 99.83% accuracy for SVM (Dolatabadi & Keynia, 2017), just like Karvana’s research on a private bank in Indonesia which resulted in SVM with a comparison of 50:50 class sampling data is the best method (Karvana et al., 2019). Osowski also comparing SVM with Multilayer Perceptron with 99.6% accuracy on SVM (Osowski & Sierenski, 2020). On another studies, random forest algorithm reaching 94.4% accuracy (Preetha & Rayapeddi, 2018). XGBoost gives the highest performance and seems to become a favourite in many machine learning challenges (Chen & Guestrin, 2016; Do et al., 2017). Jain and Dalvi used Logistic Regression on their research comparing it with Logit Boost and Decision Tree, which resulted in Logistic Regression giving higher accuracy (Dalvi et al., 2016; Jain et al., 2020). Yanfang also uses Logistic Regression in ecommerce using user's online duration, number of logins, attentions, and other user's behaviour (Yanfang & Chen, 2017). On research for Deep Neural Network there is research using three model architectures with data from telecom company (Umayaparvathi & Iyakutti, 2017) and research using twitter of telecom company (Gridach et al., 2017). Artificial Neural Network significantly outperformed K-Nearest Neighbours, Decision Tree and SVM in a telecommunication industry on Keramati's research (Keramati et al., 2016). On another research, Decision Tree with three architectures is used on telecommunication company data (Odusami et al., 2021). Based on these related works, this study will use user's behaviour in a music streaming application data using three methods, which is XGBoost, Logistic Regression, and Deep Neural Network.

METHOD

The research will begin by collecting data owned by the company. The data will be retrieved from the Google Analytics platform, where user activity from the application is recorded and stored. Data recorded by Google Analytics contains user's daily activities.

The data will be retrieved and processed with Google BigQuery. In this process, attributes with invalid value and users who do not have a user ID will be filtered out. After the data is processed, the data will be divided into two parts, namely training data and test data. Training data is used to train the method used, and then tested using test data.

The results of data mining from the decision tree will then be tested for the level of accuracy, precision, recall, and F1-score. Then it will look for what data attributes most influence the churn rate from customers.

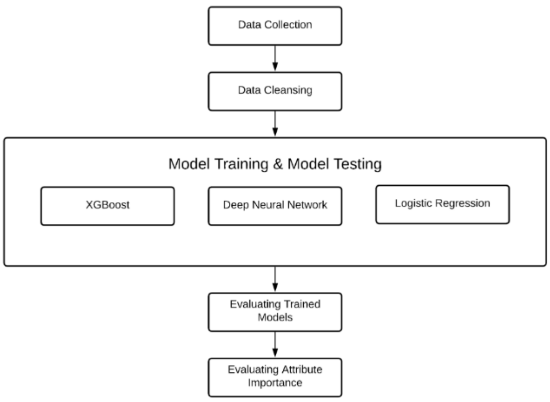

FIGURE 1. Proposed Methodukan

RESULT AND DISCUSSION

Data is collected through Google BigQuery from Google Analytics data. Data collected are daily individual user's activity from January 1st – December 31st, 2020. Data consisted of 10 attributes, which are user id, date, mobile device brand, city, app version, time on site, visits, success play, failed play, and failed login. There 38,005,447 rows of data collected for this experiment. Details of data attribute collected can be seen from Table 1.

Table 1. Data Attributes

Attribute	Description
uid	User Id of the customer
date	Date of users accessing the app
mobileDeviceBranding	Mobile device used for accessing the app
city	City the users accessing the app from
appVersion	Version of app the user is accessing from
timeOnSite	The number of time users accessing per session (seconds)
visits	Number of visits
successPlay	Number of plays the users successfully do
failedPlay	Status of users ever failed to play songs
failedLogin	Status of users ever failed to login

Data that's been collected will be filtered from attributes with invalid value and users who do not have a user ID. After that, data will be aggregated per user id activity and creating new attributes in the process. The new attributes can be seen in Table 2.

Table 2. Data Attributes After Cleansing

Attribute	Description
uid	User Id of the customer
firstDate	First date of users accessing the app
lastDate	Latest date of users accessing the app
dayDuration	Number of days from firstDate to lastDate
sessionPerDay	Average number of visits user do in a day
mobileDeviceBranding	Mobile device used for accessing the app
city	City the users accessing the app from
appVersion	Version of app the user is accessing from
timeOnSite	The total number of time users accessing (seconds)
visits	Number of visits
avgSessionDuration	Average number of time users visiting the app in a visit (seconds)
successPlay	Number of plays the users successfully do
failedPlay	Status of users ever failed to play songs
failedLogin	Status of users ever failed to login
churnStatus	Status churn of a user (1 and 0 where 1 is churn and 0 is not churn)

Numerical attribute data then standardized. Data that has been processed is shrunk to 3,941,713 rows. The standardized Table then divided to two parts. 80% training data consisting of 3,154,069 rows of data and 20% testing data consisting of 787,644 rows of data.

This research will use three machine learning methods as comparison. The first method is XGBoost, Deep Neural Network, and Logistic Regression. The models will be trained using the training data, and then we will evaluate the model using the testing data.

After we evaluate the model, we will calculate each the performance metrics of every method. We will calculate the accuracy, precision, recall, and F1-score. After that we will calculate attribute importance of the winning method.

The performance of each method can be shown in confusion matrix below.

Table 3. Confusion Matrix by Values

Method	TP	FN	FP	TN
XGBoost	619,914	60,206	9,553	97,971
XGBoost with Tuning	621,515	58,605	9,682	97,842
DNN	628,553	51,567	13,363	94,161
DNN with Tuning	640,877	39,243	17,157	90,367
Logistic Regression	636,453	43,667	15,372	92,152
Logistic Regression with Tuning	673,870	6,250	34,976	72,548

Table 4. Confusion Matrix by Percentage

Method	TP	FN	FP	TN
XGBoost	91.15%	8.85%	8.88%	91.12%
XGBoost with Tuning	91.38%	8.62%	9.00%	91.00%
DNN	92.42%	7.58%	12.43%	87.57%
DNN with Tuning	94.23%	5.77%	15.96%	84.04%
Logistic Regression	93.58%	6.42%	14.30%	85.70%
Logistic Regression with Tuning	99.08%	0.92%	32.53%	67.47%

Table 5. Experiment Result

Method	Accuracy	Precision	Recall	F1
XGBoost	91.14%	98.48%	91.15%	94.67%
XGBoost with Tuning	91.33%	98.47%	91.38%	94.79%
DNN	91.76%	97.92%	92.42%	95.09%
DNN with Tuning	92.84%	97.39%	94.23%	95.79%
Logistic Regression	92.50%	97.64%	93.58%	95.57%
Logistic Regression with Tuning	94.77%	95.07%	99.08%	97.03%

After evaluating the three methods, we got the above results. Using those confusion matrixes, we can calculate the performance metrics of the methods. Based on Table 3 and Table 4, Logistic Regression with Tuning gives the highest True Positives with 673,870 users and 99.08% of actual positives. But on the other hand, it also gives highest False Positives with 32.53% of actual negatives. If we want to get the highest number of True Negatives, XGBoost gives 97,971 users or 91.12% of all actual negatives.

From the confusion matrix, we can get the calculation of the four important metrics. The highest accuracy with 94.77% is Logistic Regression with Tuning. On precision, XGBoost gives the highest number of precisions with 98.48%. As for recall and F1-score, Logistic Regression with Tuning gives the highest result with 99.08% and 97.03% respectively. Hyperparameter Tuning greatly affects Logistic Regression with increase of 2.27% of accuracy, 5.5% of recall, and 1.46% of F1 score. Based on the condition of the company and the results of the experiment, Logistic Regression with Tuning is the best method for the company to get the highest number of churn customers.

After the best method is decided, the most affecting attribute is calculated, and the result is shown on the Table below.

Table 6. Attribute Weight

Processed_Input	Category	Weight
appVersion	0	23.7639044
appVersion	4.1.7	0.38086121
appVersion	4.1.8	0.51607379
appVersion	4.1.8.1	-2.6704006
appVersion	4.1.8.2	0.62392512
appVersion	4.1.8.3	0.40805272
appVersion	4.1.9	0.80308194
appVersion	4.1.9.1	0.34218876
appVersion	4.1.9.2	1.75882559
appVersion	4.1.9.3	1.82160327
appVersion	5.0.0	1.22410201
appVersion	5.0.1	0.80710692
appVersion	5.0.1.1	1.71648231
appVersion	5.0.2	1.58157278
appVersion	5.0.3	1.75806296
appVersion	5.0.4	1.30430462
appVersion	5.0.7	1.74658981
appVersion	5.0.8	1.30450263
appVersion	5.0.9	1.48960225
appVersion	5.1.0	1.47860932
appVersion	5.1.1	0.83799971
appVersion	5.1.2	-0.6331276
appVersion	5.1.3	-2.6670381
appVersion	5.1.4	-51.933396
appVersion	5.1.5	-3.0190852
appVersion	5.2.0	1.87683023
appVersion	5.3.0	1.87105841
appVersion	5.3.1	1.86911074
appVersion	5.4.0	1.85660549
appVersion	5.4.1	-114.66453
appVersion	5.5.0	1.78017601
appVersion	5.5.1	97.4685808
appVersion	5.6.0	1.57690299
appVersion	5.6.1	-0.1610149
appVersion	5.7.0	-5.1716695
appVersion	5.7.1	-5.1264378
appVersion	5.7.2	-5.0930795
avgSessionDuration		0.09342136
city	Jakarta	0.76734085
dayDuration		-0.3220689
failedLogin		0.00239973
failedPlay		-0.0509372
mobileDeviceBranding	Samsung	0.76734085
sessionPerDay		0.29390708
successPlay		-0.0631738
timeOnSite		-0.0446155
visits		-0.2141609

Based on Table 6, appVersion is the most affecting attribute in the model. appVersion 5.5.1 is the most affecting for the model to make the customer churn. On the other hand, the older 5.4.1 version is most affecting for the customer to not churn.

CONCLUSION

This study of data mining implementation for churn prediction on music streaming company shows that for the company use case, Logistic Regression with Hyperparameter Tuning is the best method to get the highest number of customer churn. But on different use case, XGBoost can be the method with the highest number of True Negatives and Precision. Hyperparamater Tuning can also be a solution to increase the performance of a model, but there can be a compromise on the other side. The dataset used also can affect the performance of the model.

For future work, different methods can be used to get an even better performance. Tuning different parameters can also be a solution to increase the performance of the methods used. Lastly, bigger or different dataset can be used to on the model.

REFERENCE

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1), 1–24.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1–4.

De Caigny, A., Coussement, K., & De Bock, K. W. (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.

Do, D., Huynh, P., Vo, P., & Vu, T. (2017). Customer churn prediction in an internet service provider. *2017 IEEE International Conference on Big Data (Big Data)*, 3928–3933.

Dolatabadi, S. H., & Keynia, F. (2017). Designing of customer and employee churn prediction model based on data mining method and neural predictor. *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*, 74–77.

Gridach, M., Haddad, H., & Mulki, H. (2017). Churn identification in microblogs using convolutional neural networks with structured logical knowledge. *Proceedings of the 3rd Workshop on Noisy User-Generated Text*, 21–30.

Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications*, *39*(1), 1414–1425.

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, *31*(3), 515–524.

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, 101–112.

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer churn analysis and prediction using data mining models in banking industry. *2019 International Workshop on Big Data and Information Security (IWBIS)*, 33–38.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1), 1–13.

Keramati, A., Jafari-Marandi, R., Aliannejadi, M., Ahmadian, I., Mozaffari, M., & Abbasi, U. (2014). Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, *24*, 994–1012.

Lee, E.-B., Kim, J., & Lee, S.-G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management & Data Systems*.

Lee, E., Jang, Y., Yoon, D.-M., Jeon, J., Yang, S., Lee, S.-K., Kim, D.-W., Chen, P. P., Guitart, A., & Bertens, P. (2018). Game data mining competition on churn prediction and survival analysis using commercial game log data. *IEEE Transactions on Games*, *11*(3), 215–226.

Liao, K.-H., & Chueh, H.-E. (2011). Applying fuzzy data mining to telecom churn management. *International Conference on Intelligent Computing and Information Science*, 259–264.

Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Sharma, M. M. (2021). A hybrid machine learning model for predicting customer churn in the telecommunication industry. *International Conference on Innovations in Bio-Inspired Computing and Applications*, 458–468.

Osowski, S., & Sierenski, L. (2020). Prediction of customer status in corporate banking using neural networks. *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–6.

Preetha, S., & Rayapeddi, R. (2018). Predicting Customer Churn in the Telecom Industry Using Data Analytics. *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, 38–43.

Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *International Journal of Information Management*, *48*, 238–253.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, *7*, 60134–60149.

Umayaparvathi, V., & Iyakutti, K. (2017). Automated feature selection and churn prediction using deep learning models. *International Research Journal of Engineering and Technology (IRJET)*, *4*(3), 1846–1854.

Yanfang, Q., & Chen, L. (2017). Research on E-commerce user churn prediction based on logistic regression. *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, 87–91.

Yiğit, İ. O., & Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 1–4.

Yu, H., Tan, Z.-H., Ma, Z., Martin, R., & Guo, J. (2017). Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(10), 4633–4644.