Indian Journal of Data Mining (IJDM)
ISSN: 2582-9246 (Online), Volume-2 Issue-2, November 2022

Bank Customer Churn Prediction

Jufin P. A, Amrutha N

Abstract: ***In the current challenging era, there is a stiff competition happening between the banking industries. To strengthen the grade and level of services they provide, banks focus on customer retention as well as the customer churning. Customer churning becomes one of the duties of corporate intelligences to speculate the number of customers leaving from the bank or presumed to be churned. It also helps in predicting the number of customers retained. The primary objective of this paper is "Bank customer churn prediction" is to build a model that can distinguish and visualize which factors or attributes contribute to customer churn. In addition to that, this paper also discusses a comparison between various classification algorithms. Machine learning is a modern technology that has the potential to solve classification problems. Using supervised machine learning techniques, a best model is chosen that will assign a probability to the churn to simplify customer service to prevent customer churn. Few methodologies are compared in order to accomplish different accuracy levels. XGBoost is considered in order to check if a better model can be obtained that provides best result in terms of accuracy. The other three machine learning algorithms compared are Logistic regression, Support vector machine [SVM], and Random Forest.***

Keywords: Customer Churning, Machine Learning, XG Boost, Logistic Regression, SVM, Random Forest.

I. INTRODUCTION

Bank customer churn prediction is a crucial task for banks and other financial corporations looking to retain or engage their customers and improve customer loyalty. In today's hard-fought battle in market, where customers can choose from wide varieties of option, it is necessary for banks to retain their existing customers and prevent them from leaving the bank and relocating to their competitors. Churn prediction helps banks to recognise the customers that are expected to be churned or the potential churners so that they can bring out the new strategies, improve the level of services or take proactive measures that are required to keep their customers from churning. Customer churn refers to the situation where the customers stop dealing business with a corporation and refrain from using their services. For banks, customer churn can occur due to many reasons such as poor customer service, high fees, better offers from competitors, or migrating to another nation. Losing customers from the banks not only affect the banks revenue but also troubles their reputation in the market. Predicting the customer churn is a difficult task as it requires an enormous amount of customer data such as customer behavior, their estimated salary, demography etc. However, with the help of big data technologies and machine learning algorithms, banks can now make predictions using customer data and take suitable measures to maintain their customers [5][6]. Churn prediction also helps banks to advance their resources and minimize their costs. Obtaining new customers can be expensive and difficult. Retaining the existing customers can be a cost-effective strategy. By predicting the customer churn banks can concentrate on their resources on retaining customers who are most likely to be churned or stop using the resources provided by the bank, rather than gaining the new customers. Moreover, predicting the churn benefit the bank in improving their customer relations. By identifying the reason of churn, banks can advance their resources as well as the services to meet the customer requirements and expectations. This can lead to advanced customer contentment and loyalty. Predictive modelling is one of the most popularly used machine learning techniques for churn prediction. It involves analysing customer data to recognise patterns and trends that indicate how likely a customer will exit or churn out. Using this information, a predictive model is built that is capable of identifying potential churners. The system is designed to be user-friendly and can be easily accessible to bank executives. It will provide a summarized report about the customer churn prediction and provide insights useful for retention. Overall, the system enables banks to strengthen profit and customer satisfaction by reducing the customer churn and rationalizing resource deployment. This paper will be comparing the performance of XGBoost over other machine learning algorithms like logistic regression, support vector machine and random forest.

II. RELATED WORKS

It is very clear that, for a company customer retention or customer support is very important for its corporate strategies and business development. Customer churning becomes one of the business intelligences to predict the number of customers leaving the company or which customer is most likely to be churned and to predict the number of customers will get retained. In this section, a few methodologies are compared in relation with customer churn prediction. In churning of bank customers using supervised Learning by authors Hemlata Dalmia, Ch V S S Nikil and Sandeep Kumar [1]. discusses the classification problem of banking industry that is to predict whether the customer depart from a bank. Their whole focus is to build a model using the flexible technique in supervised learning to boost the accuracy in the process of customer churning.

Manuscript received on 25 October 2022 | Revised Manuscript received on 05 November 2022 | Manuscript Accepted on 15 November 2022 | Manuscript published on 30 December 2023.
**Correspondence Author(s)*
Jufin P A*, Department of Computer Science, St. Albert's College (Autonomous), Ernakulam, India. Email: jufin.pa@gmail.com, ORCID ID: 0009-0009-4478-8438
Amrutha N, Department of Computer Science, St. Albert's College (Autonomous), Ernakulam, India. Email: amruthannandakumar@gmail.com, ORCID ID: 0009-0004-7367-6850

© The Authors. Published by Lattice Science Publication (LSP). This is an open access article under the CC-BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

A combination of KNN and XGBoost algorithm is used to enhance the accuracy of the model. XGBoost obtained the best result in terms of accuracy of 86.85% with low error rate, high sensitivity and specificity. Whereas KNN gives 83.85% percentage of accuracy.

Authors Dana AL-Najjar, Nadia Al-Rousan and Hazem AL-Najjar in machine learning to develop credit card customer churn prediction [2], aimed to investigate the capability of machine learning to predict the credit card customer churn prediction by using a feature-selection method and five different machine learning models with the collected dataset containing two types of data: categorical and continuous variables. In order to develop the prediction model three models were proposed to change the independent variable. In addition to that five machine learning models such as Bayesian network, C5 tree, CHAID, CR-Tree, and a neural network were suggested. The result indicated that C5 tree models are capable enough to outperform other functional model.

In Customer churn prediction in the banking sector using machine learning based classification models by authors Hoang Tran, Ngoc Le and Van-Ho Nguyen [3]. aims to examine the impact of customer segmentation on the accuracy of customer churn prediction using machine learning models and to experiment, contrast and assess which machine learning model are most effective in prediction analysis of customer churn. To achieve the outcome different machine learning models such as k-means clustering to segment customers, k-nearest neighbors, support vector machine, decision tree, random forest, and support vector machine are implemented. The experimental findings point out that the two best training methods are random forest and support vector machine. The end result demonstrated that the dataset obtained performs finer with the random forest model in terms of accuracy of 97.25%. Logistic regression obtains the lowest accuracy of 87.27%.

In the article B2C e-commerce customer churn prediction based on K-Means and SVM by Xiancheng Xiahou and Yoshio Harada[4][7][8][9][10][11], a loss prediction model is proposed based on the combination of k-means customer segmentation and support vector machine (SVM). Using the data of customer behavior of a B2C e-commerce enterprise prediction ability of the support vector machine and Logistic regression models are tested. This study aims to assess the effectiveness of customer segmentation in predicting customer shopping behavior based on multivariate variables over time. The results showed that customer segmentation significantly improved prediction accuracy, and k-means clustering was found to be a necessary implementation here. Additionally, the study compared the predictive performance of traditional statistical methods using logistic regression and machine learning using SVM. The final result obtained shows that the prediction accuracy of the SVM model is higher than that of LR model.

III. METHODOLOGY

A. Data collection

Data collection is the important step in the process of bank customer churn prediction. Based on the quality and quantity of data collected, the accuracy and effectiveness of churn prediction model is determined. The dataset is collected from an online source named Bank Customer Churn Prediction Dataset which consist of a csv file that has over 14 columns namely - CustomerId, CreditScore, Geography, Gender, Age, Tenure, Balance, Estimated salary etc. The csv file consists of over 10000 entries or rows [5]. There are numerous primary sources of data used by banks for the purpose of predicting customer churn, these include:

(1) Customer Transactional Data- These mainly consist of data that are related to bank related transactions such as deposits, withdrawal, and transfer history

(2) Demographic Data- They consist of data regarding the age, income, educational qualification, geography etc of the customers.

(3) Credit Score - These consist of data regarding credit scores and credit history of customers of the bank.

There are several key considerations that banks should keep in mind while collecting and managing data for bank customer churn prediction.

(1) Data quality- For churn prediction accuracy and completeness of data are crucial. In order to avoid inconsistencies and errors in the data, banks should ensure that the collected data is stored in a standardized and consistent manner.

(2) Data privacy- Banks must follow the data privacy regulations when collecting and managing the customer data.

(3) Data volume- The amount of data collected can influence the accuracy and effectiveness of churn prediction model. Having a sufficient number of samples helps to improve the model performance.

(4) Data integration- Customer data is available across different systems and platform within a bank. These data from multiple sources must be integrated to ensure that the churn prediction model is based on concrete and comprehensive view of customer behavior.

(5) Data analysis- The data must be analyzed to recognize patterns and trends that reveals customer churn. Advanced analytical tools and techniques must be used to analyze the customer data and develop predictive models.

B. Pre-Processing

Pre-processing is another critical step for bank customer churn prediction to ensure that the data collected is accurate, consistent, and free from missing values, errors and inconsistencies. It involves steps like data cleaning, data normalization, feature engineering and data selection ensuring the data is of high quality and can be analysed by machine learning models.

The dataset was completely checked for missing or null values and no such missing or noisy data was found. Next the features were carefully examined and some unnecessary columns were removed as it is irrelevant for the data analysis procedure. As an additional data pre-processing step, columns such as creditscoregiven age are added. Values such as 0 and 1 are changed to -1 and 1 respectively.

© Copyright: All rights reserved.

Min Max scaling is also performed to normalize the range of numeric data to a fixed range between 0 and 1 allowing the machine learning algorithm to learn patterns and make better accurate predictions.

C. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a process involving the use of statistical techniques, data visualizations, and data mining for analysing and summarizing the key features and patterns exhibited in a dataset. The raw dataset was examined and several comparisons were plotted in the form of pie chart, bar plot and box plot based on various attributes.

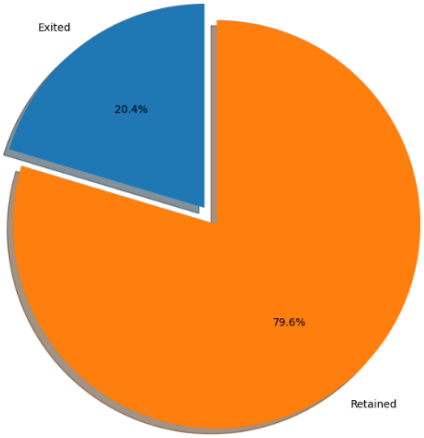

Fig. 1. Proportion of Customers Churned and Retained

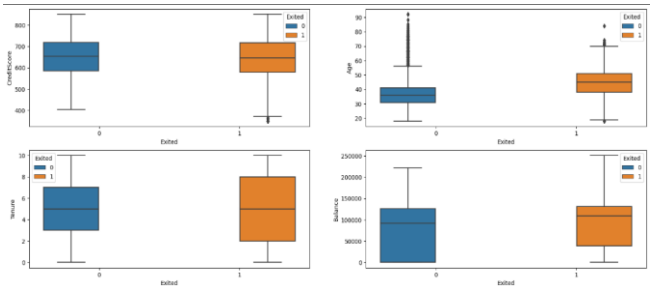

Fig. 2. Box Plots for Exit Status of Customer Based on Parameters Such as Credit Score, Age, Tenure, Balance

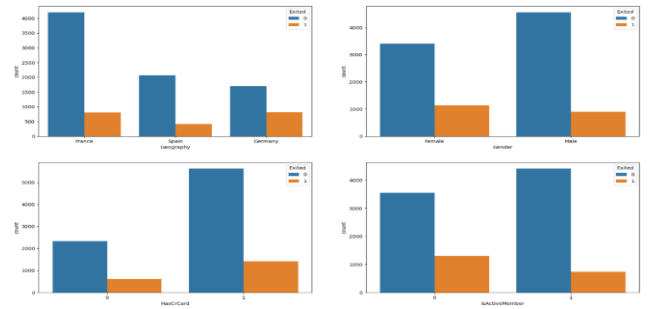

Fig. 3. Count of Distribution of Customers Retained and Exited Based on Various Attributes

D. Predictive Modeling

In bank customer churn prediction, predictive modelling is an integral part that involves the use of statistical algorithms and machine learning techniques to evaluate large amount of customer data and recognize patterns and trends that helps to predict the customer behaviour. The predictive modelling techniques used in bank customer churn prediction is logistic regression, random forests, Support Vector Machine and XGBoost. Logistic Regression, Random Forests, Support Vector Machine were machine learning algorithms used in the existing system, while XGBoost algorithm has been included in this to obtain a better result than previously used algorithms and compare them based on performance. The train-test split is important in machine learning. The dataset is split into 80-20, where 80% of data is used for training and rest 20% of data for testing.

i) Logistic Regression

It is a statistical technique used to predict the probability of binary outcome. The goal of logistic regression is to find the relationship between dependent and independent variables.

ii) Support Vector Machine

SVM can be used for both classification and regression problem. But preferably it used for classification problem. What SVM does is, it plots the data objects in an n-dimensional space. The objective of SVM is to find the hyperplane which divides the data points in such a way that it divides the data points into 2 different classes.

iii) Random Forest

It is a classifier that contains multiple decision trees on various subset of given dataset. Random forest work by taking predictions from multiple trees and based on the majority votes, the final output is predicted. It is an ensemble method as it contains number of decision trees. Greater the number of trees in forest, better the accuracy.

iv) XGBoost

XGBoost is extreme gradient boosting. It is an ensemble method as it uses gradient boosting and decision trees. Gradient boosting is an algorithm which create a new model by eliminating the errors of previously built model. XGBoost is popularly used for its better performance and high speed.

E. Model Evaluation

The evaluation metrics used in bank customer churn prediction are Accuracy, Precision, Recall and F1 score. In addition to that, robustness and generalization ability of the churn prediction model are also evaluated.

a. Accuracy

The ratio of total correct predictions to total forecasts is known as accuracy.

Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives).

b. Precision

Precision allows us to measure the accuracy of predictions that were true.

Precision = True Positives / (True Positives + False Positives).

c. Recall

It is basically a measure that tells us how correctly the true positives were identified.

Recall = True Positives / (True Positives + False Negatives)

© Copyright: All rights reserved.

d. F1 score

It is a single metric that utilizes the harmonic mean to combine recall and precision.

F1 Score = 2* ((precision * recall) / (precision + recall))

e. Support

The number of instances of the true response that fall into every category of target values can be used to determine support.

f. AUC-ROC curve

AUC stands for Area under the Curve, and ROC stands for Receiver operating characteristic curve. It is a graph that displays a classification model's performance across all conceivable thresholds. The curve between the two parameters True Positive Rate (TPR) and False Positive Rate (FPR) is displayed.

g. Confusion Matrix

The effectiveness of a model developed using machine learning on a set of test data is summarized by a confusion matrix, which is a matrix. It is frequently used as a metric to assess the effectiveness of categorization models, whose goal is to forecast a categorical label for every input occurrence. In the matrix, the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) generated by the model on the test data is shown.

IV. RESULT

Table 1. Metrics Evaluation of Logistic Regression

1. Metrics Evaluation of Logistic Regression

	precision	recall	f1-score	support
0	0.83	0.97	0.89	6353
1	0.64	0.24	0.35	1647
accuracy			0.82	8000
macro avg	0.73	0.60	0.62	8000
weighted avg	0.79	0.82	0.78	8000

Table 2. Metrics Evaluation of Random Forest

2. Metrics Evaluation of Random Forest

	precision	recall	f1-score	support
0	0.87	0.98	0.92	6353
1	0.83	0.42	0.56	1647
accuracy			0.86	8000
macro avg	0.85	0.70	0.74	8000
weighted avg	0.86	0.86	0.84	8000

Table 3. Metrics Evaluation of Support Vector Machine

3. Metrics Evaluation of Support Vector Machine

	precision	recall	f1-score	support
0	0.86	0.98	0.92	6353
1	0.85	0.40	0.54	1647
accuracy			0.86	8000
macro avg	0.86	0.69	0.73	8000
weighted avg	0.86	0.86	0.84	8000

Table 4. Metrics Evaluation of XGBoost

4. Metrics Evaluation of XGBoost

	precision	recall	f1-score	support
0	0.89	0.97	0.93	6353
1	0.83	0.53	0.64	1647
accuracy			0.88	8000
macro avg	0.86	0.75	0.79	8000
weighted avg	0.88	0.88	0.87	8000

5. AUC-ROC Curve

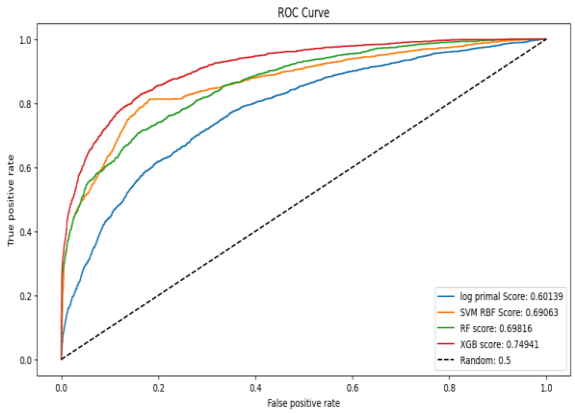

Fig. 2. AUC-ROC Curve

6. Confusion Matrix

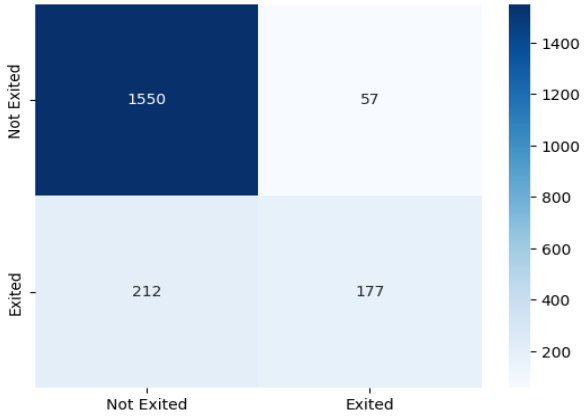

Fig. 3. Confusion Matrix of XGBoost (Test Data)

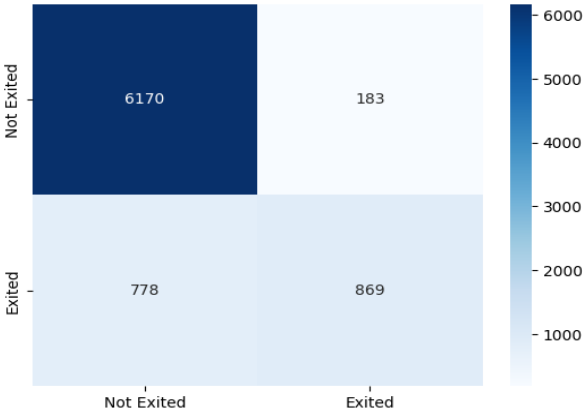

Fig. 4. Confusion Matrix of XG Boost (Train Data)

V. FUTURE ENHANCEMENT

Natural language processing (NLP) approaches integration is one possible improvement for bank customer turnover prediction. Banks can get important insights into consumer behaviour and sentiment by applying NLP approaches to analyse unstructured information about consumers such as customer comments, social media posts, and online reviews. Banks may utilise NLP to pinpoint the main reasons for client turnover and create focused measures to stop it.

Retrieval Number:100.1/ijdm.B1628113223
DOI:10.54105/ijdm.B1628.112222
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)
© Copyright: All rights reserved.

For instance, banks can discover frequent problems that consumers are having and take proactive measures to solve them by analysing customer feedback. Utilising modern methods for machine learning like reinforcement learning and deep learning will be another potential improvement. A particular category of machine learning called deep learning algorithms may automatically train to represent intricate connections and patterns in data. By creating tailored retention plans for each customer, banks may improve their churn prediction models. Banks may considerably improve the efficacy of retention tactics by analysing customer information and preferences to create customised offers and rewards that are suited to each client's needs and preferences.

VI. CONCLUSION

Predicting bank customer turnover is important for both client retention and profitability in the financial services industry. Banks must anticipate customer attrition and take preventative efforts to keep consumers given the growth of internet banking and rising competition. Customer churn can be predicted using machine learning approaches like predictive modelling, decision trees, and neural networks. Banks can create specific advertising initiatives and individualised offers to entice customers to stay by analysing customer data for trends and patterns that suggest a client is likely to leave. Banks must anticipate customer attrition and take preventative efforts to keep consumers given the growth of internet banking and rising competition. Customer churn can be predicted using machine learning approaches like predictive modelling, decision trees, and neural networks. Banks can create specific advertising initiatives and individualised offers to entice customers to stay by analysing customer data for trends and patterns that suggest a client is likely to leave.

DECLARATION STATEMENT

Funding	No, I did not receive.
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/input
Authors Contributions	All authors have equal participation in this article.

REFERENCES

1. Dalmia, Hemlata & Nikil, Ch V S S & Kumar, Sandeep. (2020). Churning of Bank Customers Using Supervised Learning. 10.1007/978-981-15-3172-9_64. https://doi.org/10.1007/978-981-15-3172-9_64
2. Hoang Dang Tran, Ngoc Le, Van-Ho Nguyen. Interdisciplinary Journal of Information, Knowledge, and Management Volume 18 2023 pp. 087-105 https://doi.org/10.28945/5086
3. Xiahou X, Harada Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research. 2022; 17(2):458-475. https://doi.org/10.3390/jtaer17020024
4. AL-Najjar D, Al-Rousan N, AL-Najjar H. Machine Learning to Develop Credit Card Customer Churn Prediction. Journal of Theoretical and Applied Electronic Commerce Research. 2022; 17(4):1529-1542. https://doi.org/10.3390/jtaer17040077
5. https://www.kaggle.com/code/kmalit/bank-customer-churn-prediction/input
6. https://www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning
7. Nikam, S. S., & Dalvi, Prof. R. (2020). Fake News Detection on SocialMedia using Machine Learning Techniques. In International Journal of Innovative Technology and Exploring Engineering (Vol. 9, Issue 7, pp. 940–943). https://doi.org/10.35940/ijitee.g5428.059720
8. Radhamani, V., & Dalin, G. (2019). Significance of Artificial Intelligence and Machine Learning Techniques in Smart Cloud Computing: A Review. In International Journal of Soft Computing and Engineering (Vol. 9, Issue 3, pp. 1–7). https://doi.org/10.35940/ijsce.c3265.099319
9. Dogra, A., & Dr. Taqdir. (2019). Detecting Intrusion with High Accuracy: using Hybrid K-Multi Layer Perceptron. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 3, pp. 4994–4999). https://doi.org/10.35940/ijrte.c5645.098319
10. Gupta, R., Gowalker, N., Patil, D. S., & Joshi, Dr. S. D. (2019). Predicting Risk in Sentiment Analysis using Machine Learning. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 1, pp. 455–460). https://doi.org/10.35940/ijeat.a9540.109119
11. Arora, A., & Basu, N. (2023). Machine Learning in Modern Healthcare. In International Journal of Advanced Medical Sciences and Technology (Vol. 3, Issue 4, pp. 12–18). https://doi.org/10.54105/ijamst.d3037.063423

AUTHORS PROFILE

Jufin P A is currently pursuing Masters of Science in Computer Science from St. Albert's College (Autonomous), Kochi. Prior to this, she completed her Bachelor of Science degree in Computer Science. She has a wide range of interests ranging from IoT, python and Machine learning.

Amrutha N joined Department of Computer Science, St. Albert's College (Autonomous), Kochi as Assistant Professor in 2021 July. She graduated in B-Tech Information Technology in 2017. She completed her post-graduation in M-Tech Computer Science & Engineering in 2019. She holds a patent in the year 2021 entitled as Intelligent IoT based smart irrigation system using cloud computing. Her major areas of Interests include Security and Cloud Computing.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© Copyright: All rights reserved.