**Hoàng Lan Anh-7123112071**

***Đề tài: Khách hàng rời bỏ dịch vụ***

|  |  |  |
| --- | --- | --- |
| iid | Cite | Abstract |
|  | @article{Xiahou2022B2CEC,  title={B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM},  author={Xiancheng Xiahou and Yoshio Harada},  journal={J. Theor. Appl. Electron. Commer. Res.},  year={2022},  volume={17},  pages={458-475},  url={https://api.semanticscholar.org/CorpusID:248009771}  } | Customer churn prediction is very important for e-commerce enterprises to formulate effective customer retention measures and implement successful marketing strategies. According to the characteristics of longitudinal timelines and multidimensional data variables of B2C e-commerce customers’ shopping behaviors, this paper proposes a loss prediction model based on the combination of k-means customer segmentation and support vector machine (SVM) prediction. The method divides customers into three categories and determines the core customer groups. The support vector machine and logistic regression were compared to predict customer churn. The results show that each prediction index after customer segmentation was significantly improved, which proves that k-means clustering segmentation is necessary. The accuracy of the SVM prediction was higher than that of the logistic regression prediction. These research results have significance for customer relationship management of B2C e-commerce enterprises. |
|  | @article{Fujo2022CustomerCP,  title={Customer Churn Prediction in Telecommunication Industry Using Deep Learning},  author={Samah Wael Fujo and Suresh Subramanian and Moaiad Ahmad Khder},  journal={Information Sciences Letters},  year={2022},  url={https://api.semanticscholar.org/CorpusID:246394582}  } | Without proper analysis and forecasting, industries will find themselves repeatedly churning customers, which the telecom industry in particular cannot afford. A predictable model for customers will allow companies to retain current customers and to obtain new ones. Deep-BP-ANN implemented in this study using two feature selection methods, Variance Thresholding and Lasso Regression, in addition, our model strengthened by early stopping technique to stop training at right time and prevent overfitting. We compared the efficiency of minimizing overfitting between dropout and activity regularization strategies for two real datasets: IBM Telco and Cell2cell. Different evaluation approaches used: Holdout, and 10-fold cross-validation to evaluate the model’s efficiency. To solve unbalanced issue, the Random Oversampling technique was used to balance both datasets. The results show that the model implemented performs well with lasso regression for feature selection, early stopping technique to pick the epochs, and large numbers of neurons (250) into the input and hidden layers, and activity regularization to minimize overfitting for both datasets. In predicting customer churn, our findings outperform ML techniques: XG\_Boost, Logistic\_Regression, Naïve\_Bayes, and KNN. Moreover, our Deep-BP-ANN model’s accuracy outperforms the existing deep learning techniques that use holdout or 10fold CV for the same datasets. |
|  | @article{DangTran2023CustomerCP,  title={Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models},  author={Hoang Dang Tran and Ngoc Toan Le and Van-Ho Nguyen},  journal={Interdisciplinary Journal of Information, Knowledge, and Management},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257262441}  } | Previous research has generally concentrated on identifying the variables that most significantly influence customer churn or has used customer segmentation to identify a subset of potential consumers, excluding its effects on forecast accuracy. Consequently, there are two primary research goals in this work. The initial goal was to examine the impact of customer segmentation on the accuracy of customer churn prediction in the banking sector using machine learning models. The second objective is to experiment, contrast, and assess which machine learning approaches are most effective in predicting customer churn. Background: This paper reviews the theoretical basis of customer churn, and customer segmentation, and suggests using supervised machine-learning techniques for customer attrition prediction. Methodology: In this study, we use different machine learning models such as k-means clustering to segment customers, k-nearest neighbors, logistic regression, decision tree, random forest, and support vector machine to apply to the dataset to predict customer churn. Contribution: The results demonstrate that the dataset performs well with the random forest model, with an accuracy of about 97%, and that, following customer segmentation, the mean accuracy of each model performed well, with logistic regression having the lowest accuracy (87.27%) and random forest having the best (97.25%). Findings: Customer segmentation does not have much impact on the precision of predictions. It is dependent on the dataset and the models we choose. Recommendations for Practitioners: The practitioners can apply the proposed solutions to build a predictive system or apply them in other fields such as education, tourism, marketing, and human resources. Recommendation for Researchers: The research paradigm is also applicable in other areas such as artificial intelligence, machine learning, and churn prediction. Impact on Society: Customer churn will cause the value flowing from customers to enterprises to decrease. If customer churn continues to occur, the enterprise will gradually lose its competitive advantage. Future Research: Build a real-time or near real-time application to provide close information to make good decisions. Furthermore, handle the imbalanced data using new techniques.  Collapse |
|  | @article{Vasudevan2022CustomerCA,  title={Customer churn analysis using XGBoosted decision trees},  author={Muthupriya Vasudevan and Revathi Sathya Narayanan and Sabiyath Fatima Nakeeb and Abhishek Abhishek},  journal={Indonesian Journal of Electrical Engineering and Computer Science},  year={2022},  url={https://api.semanticscholar.org/CorpusID:245821262}  } | Customer relationship management (CRM) is an important element in all forms of industry. This process involves ensuring that the customers of a business are satisfied with the product or services that they are paying for. Since most businesses collect and store large volumes of data about their customers; it is easy for the data analysts to use that data and perform predictive analysis. One aspect of this includes customer retention and customer churn. Customer churn is defined as the concept of understanding whether or not a customer of the company will stop using the product or service in future. In this paper a supervised machine learning algorithm has been implemented using Python to perform customer churn analysis on a given data-set of Telco, a mobile telecommunication company. This is achieved by building a decision tree model based on historical data provided by the company on the platform of Kaggle. This report also investigates the utility of extreme gradient boosting (XGBoost) library in the gradient boosting framework (XGB) of Python for its portable and flexible functionality which can be used to solve many data science related problems highly efficiently. The implementation result shows the accuracy is comparatively improved in XGBoost than other learning models. |
|  | @article{Mena2023ExploitingTR,  title={Exploiting time-varying RFM measures for customer churn prediction with deep neural networks},  author={Gary Mena and Kristof Coussement and Koen W. De Bock and Arno De Caigny and Stefan Lessmann},  journal={Annals of Operations Research},  year={2023},  pages={1-23},  url={https://api.semanticscholar.org/CorpusID:257791561}  } | Deep neural network (DNN) architectures such as recurrent neural networks and transformers display outstanding performance in modeling sequential unstructured data. However, little is known about their merit to model customer churn with time-varying data. The paper provides a comprehensive evaluation of the ability of recurrent neural networks and transformers for customer churn prediction (CCP) using time-varying behavioral features in the form of recency, frequency, and monetary value (RFM). RFM variables are the backbone of CCP and, more generally, customer behavior forecasting. We examine alternative strategies for integrating time-varying and non-variant customer features in one network architecture. In this scope, we also assess hybrid approaches that incorporate the outputs of DNNs in conventional CCP models. Using a comprehensive panel data set from a large financial services company, we find recurrent neural networks to outperform transformer architectures when focusing on time-varying RFM features. This finding is confirmed when time-invariant customer features are included, independent of the specific form of feature integration. Finally, we find no statistical evidence that hybrid approaches (based on regularized logistic regression and extreme gradient boosting) improve predictive performance—highlighting that DNNs and especially recurrent neural networks are suitable standalone classifiers for CCP using time-varying RFM measures. |
|  | @article{Charandabi2023PredictionOC,  title={Prediction of Customer Churn in Banking Industry},  author={Sina Esmaeilpour Charandabi},  journal={ArXiv},  year={2023},  volume={abs/2301.13099},  url={https://api.semanticscholar.org/CorpusID:256389525}  } | With the growing competition in banking industry, banks are required to follow customer retention strategies while they are trying to increase their market share by acquiring new customers. This study compares the performance of six supervised classification techniques to suggest an efficient model to predict customer churn in banking industry, given 10 demographic and personal attributes from 10000 customers of European banks. The effect of feature selection, class imbalance, and outliers will be discussed for ANN and random forest as the two competing models. As shown, unlike random forest, ANN doesn’t reveal any serious concern regarding overfitting and is also robust to noise. Therefore, ANN structure with five nodes in a single hidden layer is recognized as the best performing classifier. |
|  | @article{Bogaert2023EnsembleMI,  title={Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art},  author={Matthias Bogaert and Lex Delaere},  journal={Mathematics},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257251359}  } | In the past several single classifiers, homogeneous and heterogeneous ensembles have been proposed to detect the customers who are most likely to churn. Despite the popularity and accuracy of heterogeneous ensembles in various domains, customer churn prediction models have not yet been picked up. Moreover, there are other developments in the performance evaluation and model comparison level that have not been introduced in a systematic way. Therefore, the aim of this study is to perform a large scale benchmark study in customer churn prediction implementing these novel methods. To do so, we benchmark 33 classifiers, including 6 single classifiers, 14 homogeneous, and 13 heterogeneous ensembles across 11 datasets. Our findings indicate that heterogeneous ensembles are consistently ranked higher than homogeneous ensembles and single classifiers. It is observed that a heterogeneous ensemble with simulated annealing classifier selection is ranked the highest in terms of AUC and expected maximum profits. For accuracy, F1 measure and top-decile lift, a heterogenous ensemble optimized by non-negative binomial likelihood, and a stacked heterogeneous ensemble are, respectively, the top ranked classifiers. Our study contributes to the literature by being the first to include such an extensive set of classifiers, performance metrics, and statistical tests in a benchmark study of customer churn. |
|  | @article{Suh2023MachineLB,  title={Machine learning based customer churn prediction in home appliance rental business},  author={Youngjung Suh},  journal={Journal of Big Data},  year={2023},  volume={10},  url={https://api.semanticscholar.org/CorpusID:257928264}  } | Customer churn is a major issue for large enterprises. In particular, in the rental business sector, companies are looking for ways to retain their customers because they are their main source of revenue. The main contribution of our work is to analyze the customer behavior information of actual water purifier rental company, where customer churn occurs very frequently, and to develop and verify the churn prediction model. A machine learning algorithm was applied to a large-capacity operating dataset of rental care service in an electronics company in Korea, to learn meaningful features. To measure the performance of the model, the F-measure and area under curve (AUC) were adopted whereby an F1 value of 93% and an AUC of 88% were achieved. The dataset containing approximately 84,000 customers was used for training and testing. Another contribution was to evaluate the inference performance of the predictive model using the contract status of about 250,000 customer data currently in operation, confirming a hit rate of about 80%. Finally, this study identified and calculated the influence of key variables on individual customer churn to enable a business person (rental care customer management staff) to carry out customer-tailored marketing to address the cause of the churn. |
|  | @article{Liu2023AML,  title={A Machine Learning Approach to Predict Customer Churn of a Delivery Platform},  author={Qing Liu and Qiuying Chen and Sang-Joon Lee},  journal={2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)},  year={2023},  pages={733-735},  url={https://api.semanticscholar.org/CorpusID:257721020}  } | The use of delivery platforms has become widespread due to the impact of the Covid-19 and the O2O industry. However, the ELEME delivery platform, a subsidiary of Alibaba Group, which represents China, has recently been losing market share. This means that companies need to constantly look at strategies to attract new customers and maintain existing ones. In general, it costs at least five times more to attract new customers than it does to manage existing customers. This paper attempts to predict customer churn using the ELEME customer dataset to develop strategies to identify and prevent churn in advance. The results of the analysis using machine learning approach found that the most influential feature that can predict churn is the number of clicks made by the user. This paper presents the process and explanation of applying various algorithms for predicting customer churn on a distribution platform. It also proposes strategies for dealing with customer churn. |
|  | @article{Sari2023AnalysisIO,  title={Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study},  author={Renny Puspita Sari and Ferdy Febriyanto and Ahmad Cahyono Adi},  journal={Informatica (Slovenia)},  year={2023},  volume={47},  url={https://api.semanticscholar.org/CorpusID:259653180}  } | Globalization and technological advancements in the telecommunication industry have led to a significant rise in the number of operators, leading to intense market competition. This sector has become crucial in developed countries, and companies strive to increase profits by acquiring new customers, up-selling existing ones, and extending the retention period of current clients. In the traditional method of defect prediction, a single classifier is used to build a model on a pre-labeled dataset. However, this approach has limitations in predicting defects accurately under certain circumstances. To overcome these limitations, boosting is applied to combine multiple weak classifiers and create a robust classification model. Among many algorithms used for churn prediction, ensemble techniques have demonstrated greater accuracy than simpler approaches. This study aims to overcome these limitations by experimenting with five ensemble algorithms, including Adaboost, Gradient Boost, XGBoost, CatBoost, and LightGBM. The results indicate that XGBoost outperforms other techniques and is the most suitable algorithm to build the predictive model. Additionally, the study achieves higher accuracy by performing a Grid Search CV hyper-parameter setting with XGBoost, resulting in an accuracy of 81.2%. Povzetek: Študija je primerjala pet ansambelskih algoritmov za napovedovanje prekinitve naročniškega razmerja. Rezultati kažejo, da je XGBoost najboljši algoritem z natančnostjo 81,2 %. |
|  | @article{Alves2022TelcoCT,  title={Telco customer top‐ups: Stream‐based multi‐target regression},  author={Pedro Miguel Alves and Ricardo {\^A}ngelo Filipe and Benedita Malheiro},  journal={Expert Systems},  year={2022},  volume={40},  url={https://api.semanticscholar.org/CorpusID:251194931}  } | Telecommunication operators compete not only for new clients, but, above all, to maintain current ones. The modelling and prediction of the top‐up behaviour of prepaid mobile subscribers allows operators to anticipate customer intentions and implement measures to strengthen customer relationship. This research explores a data set from a Portuguese operator, comprising 30 months of top‐up events, to predict the top‐up monthly frequency and average value of prepaid subscribers using offline and online multi‐target regression algorithms. The offline techniques adopt a monthly sliding window, whereas the online techniques use an event sliding window. Experiments were performed to determine the most promising set of features, analyse the accuracy of the offline and online regressors and the impact of sliding window dimension. The results show that online regression outperforms the offline counterparts. The best accuracy was achieved with adaptive model rules and a sliding window of 500,000 events (approximately 5 months). Finally, the predicted top‐up monthly frequency and average value of each subscriber were converted to individual date and value intervals, which can be used by the operator to identify early signs of subscriber disengagement and immediately take pre‐emptive measures. |
|  | @article{Fu2023CustomerCP,  title={Customer churn prediction for a webcast platform via a voting-based ensemble learning model with Nelder-Mead optimizer},  author={Kani Fu and Guiyang Zheng and Wei Xie},  journal={Journal of Intelligent Information Systems},  year={2023},  volume={61},  pages={859-879},  url={https://api.semanticscholar.org/CorpusID:259312879}  } | This article studies the application for customer churn prediction on webcast. Predicting churn customers become an urgent need in webcast industry because the market is getting saturated and identifying potential churn customers and developing recall marketing strategies can save companies significant costs. Despite the importance of customer churn prediction in many fields, little prior academic attention has been attached to the webcast area. To address this gap, We apply an ensemble learning method to build a binary classification model for customer churn prediction. Our proposed model uses a weighted voting ensemble method and the Nelder-Mead optimal algorithm with a specific focus on the speed of Internet customers’ mobility, extracting high-dimensional features from time series data to incorporate more detailed customer behavior information. In addition, a new customer churn indicator based on time decline is introduced to more accurately define churned customers in the training data. The experimental data is collected from a webcast application developed by a Chinese Internet company. Experimental evaluations show that compared to the traditional ensemble models, our proposed model is operationally efficient and outperforms other approaches, providing valuable insights for companies to intervene with churned customers and adopt targeting retention interventions. |
|  | @article{Khan2023PrivacyPreservingBT,  title={Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry},  author={Gul Zaman Khan and Ikram Ulhaq and Ihsan Adil and Sajad Ulhaq and Inam Ullah},  journal={VFAST Transactions on Software Engineering},  year={2023},  url={https://api.semanticscholar.org/CorpusID:267025008}  } | In recent years, customer churn has been one of the most prominent topics, especially in the telecom industry. The telecommunications industry is producing massive amounts of data every minute. Thus, the telecom industry is seeking more ways to analyze and predict their potential and churn customers. According to telecom analysis, acquiring a new customer is costlier than keeping a current one. To lessen customer churn, it is very compulsory for industries to detect an increase in customer churn factors. The number of service suppliers is increasing daily, especially in the telecom industry. Phishing attacks and fraud are crucial points in customer churn. The aim of this study is to predict customer churn with profitable churn models for retention campaigns to satisfy the business requirement of profit maximization. The proposed research used the BAT-ANN classification model with the BigML dataset to predict customer churn in the telecom industry. The proposed model achieved 89.2% testing accuracy. |
|  | @article{Zhou2023EarlyWO,  title={Early warning of telecom enterprise customer churn based on ensemble learning},  author={Yancong Zhou and Wenyue Chen and Xiaochen Sun and Dandan Yang},  journal={PLOS ONE},  year={2023},  volume={18},  url={https://api.semanticscholar.org/CorpusID:263902791}  } | Analyzing customers’ characteristics and giving the early warning of customer churn based on machine learning algorithms, can help enterprises provide targeted marketing strategies and personalized services, and save a lot of operating costs. Data cleaning, oversampling, data standardization and other preprocessing operations are done on 900,000 telecom customer personal characteristics and historical behavior data set based on Python language. Appropriate model parameters were selected to build BPNN (Back Propagation Neural Network). Random Forest (RF) and Adaboost, the two classic ensemble learning models were introduced, and the Adaboost dual-ensemble learning model with RF as the base learner was put forward. The four models and the other four classical machine learning models-decision tree, naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM) were utilized respectively to analyze the customer churn data. The results show that the four models have better performance in terms of recall rate, precision rate, F1 score and other indicators, and the RF-Adaboost dual-ensemble model has the best performance. Among them, the recall rates of BPNN, RF, Adaboost and RF-Adaboost dual-ensemble model on positive samples are respectively 79%, 90%, 89%,93%, the precision rates are 97%, 99%, 98%, 99%, and the F1 scores are 87%, 95%, 94%, 96%. The RF-Adaboost dual-ensemble model has the best performance, and the three indicators are 10%, 1%, and 6% higher than the reference. The prediction results of customer churn provide strong data support for telecom companies to adopt appropriate retention strategies for pre-churn customers and reduce customer churn. |
|  | @article{Khattak2023CustomerCP,  title={Customer churn prediction using composite deep learning technique},  author={Asad Masood Khattak and Zartashia Mehak and Hussain Ahmad and Muhammad Usama Asghar and Muhammad Zubair Asghar and Aurangzeb Khan},  journal={Scientific Reports},  year={2023},  volume={13},  url={https://api.semanticscholar.org/CorpusID:263969473}  } | Customer churn, a phenomenon that causes large financial losses when customers leave a business, makes it difficult for modern organizations to retain customers. When dissatisfied customers find their present company's services inadequate, they frequently migrate to another service provider. Machine learning and deep learning (ML/DL) approaches have already been used to successfully identify customer churn. In some circumstances, however, ML/DL-based algorithms lacks in delivering promising results for detecting client churn. Previous research on estimating customer churn revealed unexpected forecasts when utilizing machine learning classifiers and traditional feature encoding methodologies. Deep neural networks were also used in these efforts to extract features without taking into account the sequence information. In view of these issues, the current study provides an effective method for predicting customer churn based on a hybrid deep learning model termed BiLSTM-CNN. The goal is to effectively estimate customer churn using benchmark data and increase the churn prediction process's accuracy. The experimental results show that when trained, tested, and validated on the benchmark dataset, the proposed BiLSTM-CNN model attained a remarkable accuracy of 81%. |
|  | @article{Babatunde2023ClassificationOC,  title={Classification of customer churn prediction model for telecommunication industry using analysis of variance},  author={Ronke Seyi Babatunde and Sulaiman Olaniyi Abdulsalam and Olanshile Abdulkabir Abdulsalam and Micheal Olaolu Arowolo},  journal={IAES International Journal of Artificial Intelligence (IJ-AI)},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257793735}  } | Customer predictive analytics has shown great potential for effective churn models. Thriving in today's telecommunications industry, discerning between consumers who are likely to migrate to a competitor is enormous. Having reliable predictive client behavior in the future is required. Machine learning algorithms are essential to predict customer turnovers, and researchers have proposed various techniques. Churn prediction is a problem due to the unequal dispersal of classes. Most traditional machine learning algorithms are ineffective in classifying data. Client cluster with a higher risk has been discovered. A support vector machine is employed as the foundational learner, and a churn prediction model is constructed based on each analysis of variance. The separation of churn data revealed by experimental assessment is recommended for churn prediction analysis. Customer attrition is high, but an instantaneous support can ensure that customer needs are addressed and assess an employee's capacity to achieve customer satisfaction. This study uses an analysis of variance (ANOVA) with a support vector machine (SVM), classification in analyzing risks in telecom systems It may be determined that SVM provides the most accurate forecast of customer turnover (95%). The projected outcomes will allow other organizations to assess possible client turnover and collect customer feedback. |
|  | @inproceedings{Che2022ThePO,  title={The Performance of Classiﬁcation Method in Telco Customer Trouble Ticket Dataset},  author={Fauzy Che and Yayah and K. R. Mohamed Imran and Ghauth and Choo-Yee Ting},  year={2022},  url={https://api.semanticscholar.org/CorpusID:260329863}  } | A customer trouble ticketing system (CTT) is an organization’s tool to track the detection, reporting, and resolution of tickets submitted by customers. It also comprises a summary of the issue reported, the status of the ticket, the incident information, and the approach that was previously utilized to resolve the problems. The technician’s skill set and experience rely solely on completing the task without the right direction on which area to focus on ﬁrst. As a result of this manual classiﬁcation of a trouble ticket, it will be necessary to build methodologies for predicting future resolution codes. The research for this report is mainly focused on one of the telco companies in Malaysia. This study result assists the telco engineer, and the specialists resolve each issue in a very short amount of time. Additionally, the classiﬁcation of the trouble ticket resolution code method used in this study will indicate the characteristics of each issue that is being investigated. The relationship between events is feasible to discover by exploring the root cause. It is critical to establish a link between recent events and events in the previous. Because of current data mining limitations, the study needs to be more comprehensive. Data processing methods are being implemented within big data platforms to overcome the limitation of data scalability, enhance classiﬁcation accuracy, and increase computation speed. The research work will continue to progress in the direction of big data centricity. Some of the most effective approaches for big data integration and machine learning will be discussed in this paper. Throughout the experiment, any problems will be explained, as well as the solutions to each situation. A wide range of research subjects will be discussed, including construction classiﬁcation models for trouble tickets. To achieve reasonable accuracy, a few customized transformations are required. The data set’s custom parameter optimization process will further increase the classiﬁcation trouble ticket’s efﬁciency. However, greater processing capacity is necessitated to use multiple parallel classiﬁers such as Bayes, Decision-Tree, and Rule-Based with help of bigdata framewrks such as Spark. |
|  | @article{AnithaM2023AnEH,  title={An Efficient Hybrid Classifier Model for Customer Churn Prediction},  author={A AnithaM.},  journal={International Journal of Electronics and Telecommunications},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260040190}  } | Customer churn prediction is used to retain customers at the highest risk of churn by proactively engaging with them. Many machine learning-based data mining approaches have been previously used to predict client churn. Although, single model classifiers increase the scattering of prediction with a low model performance which degrades reliability of the model. Hence, Bag of learners based Classification is used in which learners with high performance are selected to estimate wrongly and correctly classified instances thereby increasing the robustness of model performance. Furthermore, loss of interpretability in the model during prediction leads to insufficient prediction accuracy. Hence, an Associative classifier with Apriori Algorithm is introduced as a booster that integrates classification and association rule mining to build a strong classification model in which frequent items are obtained using Apriori Algorithm. Also, accurate prediction is provided by testing wrongly classified instances from the bagging phase using generated rules in an associative classifier. The proposed models are then simulated in Python platform and the results achieved high accuracy, ROC score, precision, specificity, F-measure, and recall. |
|  | @article{Bilisik2023AnalysisOC,  title={Analysis of Customer Churn in Telecommunication Industry with Machine Learning Methods},  author={{\"O}zge Nalan Bilisik and Damla Tuğba Sarp},  journal={D{\"u}zce {\"U}niversitesi Bilim ve Teknoloji Dergisi},  year={2023},  url={https://api.semanticscholar.org/CorpusID:264396625}  } | In today's conditions, customer loyalty has gained importance with the increase in the competitive environment between companies, the development of marketing strategies and the improvement of companies. Therefore, it is essential to acquire customers for a company to survive. Retaining an existing customer in the telecommunication sector is less costly than gaining a new customer. Customer churn analysis is the process of predicting customers with high abandonment requests by examining the offers and utilizable behaviors. Customer churn analysis provides services to develop various campaigns aiming to increase the company’s loyalty by predicting the customers who are planning to move to another company. In this way, it gives the company a competitive advantage. This study aims to make predictions by developing models for customer churns through data mining and machine learning methods in the telecommunication sector. In addition, we believe that the application in this article will contribute to data analysts and academicians who will want to analyze customer churn with different data sets in telecommunication and other sectors in the future. The analysis in this study is carried out on a data set obtained from an open-access database, including 20 transaction records for the customer from 7043 customers and whether the customer left the company. Among the data mining methods, Random Forest (RF), Support Vector Machines (SVM) and Multilayer Artificial Neural Networks (ANN) are modeled in open-source Phyton environment. The results have shown that ANN has fared better at classifying customers than other machine learning methods. |
|  | @article{Pan2023CustomerCP,  title={Customer churn prediction in telecom based on random forest algorithm},  author={Aimin Pan and Zhendan Sun},  journal={2023 8th International Conference on Information Systems Engineering (ICISE)},  year={2023},  pages={189-192},  url={https://api.semanticscholar.org/CorpusID:265256181}  } | In recent years, with the in-depth implementation of the telcom's Number Portability work, there are more and more Number Portability users, and the Number Portability users have become a major factor of customer churn. In order to reduce customer churn, the machine learning algorithm of Python analyzes the characteristics of outbound users, establishes customer churn model prediction, locates transfer subscribers in advance, and maintains stability retention, effectively reducing the probability of outbound users. |
|  | @article{Zhao2023CustomerCP,  title={Customer Churn Prediction Based on the Decision Tree and Random Forest Model},  author={Shi Zhao},  journal={BCP Business \& Management},  year={2023},  url={https://api.semanticscholar.org/CorpusID:258543017}  } | The rate at which customers discontinue utilizing a company's services during a predetermined time period is known as the customer churn rate, also known as the attrition rate. Hence, developing a prediction model to predict the potential churn customers will generate an early alert for the company to provide them with better service. This study is divided into two main parts: dealing with a dataset about customer behaviors in a bank and building churn prediction models using machine learning algorithms. The data preprocessing part includes dataset description and some adjustments on original dataset to make it accessible for analysis, including deleting unimportant feature and adjusting feature names. Then the study apportions the modified dataset into train set and test set with an 80-20 split. Next, the study imports two kinds of machine learning algorithms, random forest classifier and decision tree classifier, to build churn prediction models. In each model, the study first performs feature selections and visualizes feature importance in bar graphs. Then the study tests each model on testing set and visualizes model performances using confusion matrices and accuracy scores. The results show that both models get most predictions correct while random forest model has a better performance due to its higher accuracy score of 91%. |
|  | @article{Maan2023CustomerCP,  title={Customer Churn Prediction Model using Explainable Machine Learning},  author={Jitendra Maan and Harsh Maan},  journal={ArXiv},  year={2023},  volume={abs/2303.00960},  url={https://api.semanticscholar.org/CorpusID:257280292}  } | It becomes a significant challenge to predict customer behavior and retain an existing customer with the rapid growth of digitization which opens up more opportunities for customers to choose from subscription-based products and services model. Since the cost of acquiring a new customer is five-times higher than retaining an existing customer, henceforth, there is a need to address the customer churn problem which is a major threat across the Industries. Considering direct impact on revenues, companies identify the factors that increases the customer churn rate. Here, key objective of the paper is to develop a unique Customer churn prediction model which can help to predict potential customers who are most likely to churn and such early warnings can help to take corrective measures to retain them. Here, we evaluated and analyzed the performance of various tree-based machine learning approaches and algorithms and identified the Extreme Gradient Boosting XGBOOST Classifier as the most optimal solution to Customer churn problem. To deal with such real-world problems, Paper emphasize the Model interpretability which is an important metric to help customers to understand how Churn Prediction Model is making predictions. In order to improve Model explainability and transparency, paper proposed a novel approach to calculate Shapley values for possible combination of features to explain which features are the most important/relevant features for a model to become highly interpretable, transparent and explainable to potential customers. |
|  | @article{Zhang2023BankCC,  title={Bank Customer Churn Analysis and Prediction},  author={Wenhui Zhang},  journal={Proceedings of the 4th Management Science Informatization and Economic Innovation Development Conference, MSIEID 2022, December 9-11, 2022, Chongqing, China},  year={2023},  url={https://api.semanticscholar.org/CorpusID:257825419}  } | The term "Customer Churn" refers to the state in which the customer stops using products or services of a company. Although the bank will inevitably lose users, which is inevitable in the process of replacing the old and new banking users, the proportion and changing trend of lost users can indicate the bank's ability to retain users and the development trend of the bank. Therefore, it is necessary for banks to know the reasons leading a client to leave the company. In order to explore the factors affecting the loss of bank users, this paper selects a dataset obtained from Kaggle, using the methods of crosstabs analysis, independent samples T test, factor analysis and one-way ANOVA respectively. In addition, logistic regression is also used to predict customer churn. From the t-test, this paper finds that those who quit have lower credit scores, are older and have larger balances than those who don't. From the factor analysis, this paper finds that the feature of country and balance are the most explanatory factors. From the logistic regression, this paper finds that the percentages of correct predictions are 69.5% and 65.3% respectively before and after selecting main components. |
|  | @article{Pan2023CustomerCP,  title={Customer churn prediction in telecom based on random forest algorithm},  author={Aimin Pan and Zhendan Sun},  journal={2023 8th International Conference on Information Systems Engineering (ICISE)},  year={2023},  pages={189-192},  url={https://api.semanticscholar.org/CorpusID:265256181}  } | In recent years, with the in-depth implementation of the telcom's Number Portability work, there are more and more Number Portability users, and the Number Portability users have become a major factor of customer churn. In order to reduce customer churn, the machine learning algorithm of Python analyzes the characteristics of outbound users, establishes customer churn model prediction, locates transfer subscribers in advance, and maintains stability retention, effectively reducing the probability of outbound users. |
|  | @article{AnithaM2023AnEH,  title={An Efficient Hybrid Classifier Model for Customer Churn Prediction},  author={A AnithaM.},  journal={International Journal of Electronics and Telecommunications},  year={2023},  url={https://api.semanticscholar.org/CorpusID:260040190}  } | Customer churn prediction is used to retain customers at the highest risk of churn by proactively engaging with them. Many machine learning-based data mining approaches have been previously used to predict client churn. Although, single model classifiers increase the scattering of prediction with a low model performance which degrades reliability of the model. Hence, Bag of learners based Classification is used in which learners with high performance are selected to estimate wrongly and correctly classified instances thereby increasing the robustness of model performance. Furthermore, loss of interpretability in the model during prediction leads to insufficient prediction accuracy. Hence, an Associative classifier with Apriori Algorithm is introduced as a booster that integrates classification and association rule mining to build a strong classification model in which frequent items are obtained using Apriori Algorithm. Also, accurate prediction is provided by testing wrongly classified instances from the bagging phase using generated rules in an associative classifier. The proposed models are then simulated in Python platform and the results achieved high accuracy, ROC score, precision, specificity, F-measure, and recall. |
|  | @article{Wu2023CustomerCP,  title={Customer Churn Prediction in the Telecommunication Industry},  author={Shaohua Wu},  journal={Advances in Economics, Management and Political Sciences},  year={2023},  url={https://api.semanticscholar.org/CorpusID:258160101}  } | Customer churn is essential for telecom fields because it reduces income when a customer switches from one service provider to another. It is important for managers to help them find the factors influencing customer churn so that they can make decisions and optimize services. In this essay, SPSS will be used to analyze the factors that influence customer churn. We use contrastive analysis to find out the factors of the customer churn and use logistic regression to analyze the degree of influence of different factors. The result indicates phone charge, quality and diversity of services have impacts on customer churn. |
|  | @article{Saxena2023AnalyzingCC,  title={Analyzing customer churn in banking: A data mining framework},  author={Aishwarya Saxena and Anushi Singh and Govindaraj M.},  journal={Multidisciplinary Science Journal},  year={2023},  url={https://api.semanticscholar.org/CorpusID:262100819}  } | Customer churn, the loss of customers to a business, is a significant challenge in the banking industry. Retaining existing customers is crucial for banks to maintain profitability and sustain growth. This paper focuses on analyzing customer churn in the banking sector. The study utilizes data mining and predictive analytics techniques to analyse customer behaviour, identify churn patterns, and develop predictive models. This research uses a data mining technique called Gaussian mixture model clustering-based adaptive support vector machine (GMM-ASVM) to forecast customer loss in the banking industry. By analyzing consumer competency and loyalty to the banking industry using GMM, this study predicts customer behaviour using a clustering approach. An accuracy of 98% was attained while classifying the clustering results using ASVM. This study gives bank administrators the ability to analyse the behaviour of their clients, which may trigger appropriate tactics based on engaging quality and increase appropriate actions of administrator capacities in interactions with customers. |
|  | @article{Li2023CounterfactualLI,  title={Counterfactual learning in customer churn prediction under class imbalance},  author={Yuanyuan Li and Xue Song and Taicheng Wei and Bing Zhu},  journal={Proceedings of the 2023 6th International Conference on Big Data Technologies},  year={2023},  url={https://api.semanticscholar.org/CorpusID:265608315}  } | Nowadays, in churn prediction, many decision-makers are trying to obtain knowledge from models through interpretation techniques due to the non-transparency of black-box. In these techniques, the counterfactual explanation generated by the counterfactual learning method is an easy-to-understand and quantitative explanation of a single instance in black-box model prediction. Counterfactual learning relies on the transition of instances across the decision boundary, while the impact of class imbalance issue and instance position in customer-related data is insufficiently considered in recent studies. In this case, this research innovatively proposes that when generating counterfactual explanations, the impact of class imbalance issue and the instance location in customer data needs to be considered. And through comparative experiments we prove that there are obvious differences in the success rate of finding a counterfactual explanation, the distance between the counterfactual explanation and the original instance (i.e. proximity), the proportion of feature change (i.e. sparsity), and the degree of proximity support (i.e. credibility) with the original instance in different instance locations and unbalanced scenarios. In addition, in our experiments, the impact of class imbalance and instance positions vary among counterfactual methods. This research provides a reference for the application of counterfactual learning in the field of customer churn prediction and elaborates that when counterfactual learning is used, the influence of class imbalance and instance location needs to be considered. |
|  | @inproceedings{Kingly2023ComparativeAO,  title={Comparative Analysis of Customer Churn Prediction},  author={Anithaa Kingly},  year={2023},  url={https://api.semanticscholar.org/CorpusID:259953329}  } | The wide variety of service providers are being elevated very swiftly in each business enterprise. For service providers, a rapidly expanding market in each location is leading to a larger subscriber base. Customer acquisition costs are rising as a result of increased competition, new and innovative business models, and better-suited goods. Service providers have learned the value of keeping existing clients happy in such a short period of time. Because of this, it is critical that service providers reduce churn, the occurrence when consumers of a business stop buying from or engaging with the business. Not just in banking and telecommunications, but also in other industries that are heavily reliant on customer engagement, this study examines the most well-known machine learning methods for churn prediction. |
|  | @article{Li2023IdentificationOC,  title={Identification of Customer Churn Considering Difficult Case Mining},  author={Jianfeng Li and Xuepeng Bai and Qian Xu and Dexiang Yang},  journal={Syst.},  year={2023},  volume={11},  pages={325},  url={https://api.semanticscholar.org/CorpusID:259671340}  } | In the process of user churn modeling, due to the imbalance between lost users and retained users, the use of traditional classification models often cannot accurately and comprehensively identify users with churn tendency. To address this issue, it is not sufficient to simply increase the misclassification cost of minority class samples in cost-sensitive methods. This paper proposes using the Focal Loss hard example mining technique to add the class weight α and the focus parameter γ to the cross-entropy loss function of LightGBM. In addition, it emphasizes the identification of customers at risk of churning and raises the cost of misclassification for minority and difficult-to-classify samples. On the basis of the preceding ideas, the FocalLoss\_LightGBM model is proposed, along with random forests, SVM, XGBoost, and LightGBM. Empirical analysis based on a dataset of credit card users publicly available on the Kaggle website. The AUC, TPR, and G-mean index values were superior to the existing model, which can effectively improve the accuracy and stability of potential lost users. |

***I. Giới thiệu bài toán:***

Khách hàng rời bỏ là một thuật ngữ kinh doanh được gọi là kích động khách hàng. Trong kinh doanh, khi khách hàng không hài lòng với dịch vụ hoặc sản phẩm mà bạn cung cấp thì việc tiêu hao sẽ xảy ra, hay nói cách khác là họ sẽ ngừng kết nối hoặc hợp tác với doanh nghiệp của bạn. Hiện nay ngày càng có nhiều doanh nghiệp quan tâm đến việc giữ chân khách hàng. Việc khách hàng rời bỏ sẽ gây tổn hại cho doanh nghiệp vì nó có thể dẫn đến tổn thất lớn về phí bảo hiểm, giảm tỷ suất lợi nhuận và có thể mất cơ hội kinh doanh giới thiệu từ các khách hàng trung thành. Theo Baran và Galka (2017), “Công trình tiên phong của F. F. Reichheld và W. E. Sasser Jr. đã tìm thấy một mối quan hệ chặt chẽ giữa việc giữ chân khách hàng và lợi nhuận của công ty. Họ phát hiện ra rằng tỷ lệ giữ chân khách hàng chỉ tăng 5% sẽ mang lại lợi nhuận được cải thiện theo giá trị hiện tại ròng từ 20 lên 85% trên một loạt các doanh nghiệp”. Ngoài ra, nhiều loại nghiên cứu đã chứng minh rằng Chi phí giữ chân khách hàng hiện tại thấp hơn khoảng 5 lần so với việc có được khách hàng mới (Dawes & Swailes, 1999).

Theo Reichheld (1996), giá trị hiện tại ròng trung bình của khách hàng cải thiện 35% đối với phần mềm. Các công ty và 95% cho các đại lý quảng cáo với tỷ lệ giữ chân khách hàng tăng 5%. Để hạn chế điều này, doanh nghiệp phải lường trước những khách hàng cụ thể có nguy cơ rời bỏ để điều chỉnh chiến lược cho phù hợp, chẳng hạn như nâng cao chất lượng sản phẩm, dịch vụ hay gia tăng lợi ích cho họ. Vì vậy, nó là điều cần thiết để tạo ra các mô hình dự đoán có thể hỗ trợ việc giữ chân khách hàng. Phương pháp học máy có thể dự đoán thành công doanh thu của khách hàng.

Thu thập dữ liệu và phân tích cho phép xác định những khách hàng có thể sắp rời bỏ doanh nghiệp, từ đó tập trung vào việc giữ chân khách hàng hơn là thu hút khách hàng mới. Ngoài ra, các mô hình học máy cũng được áp dụng trong lĩnh vực bán lẻ và kinh tế.

***II. Nghiên cứu liên quan:***

Việc rời bỏ có thể xảy ra một cách tự nguyện khi khách hàng chọn kết thúc mối quan hệ của họ với một người cụ thể người bán hoặc vô tình khi hành động của khách hàng buộc họ tránh tham gia vào bất kỳ loại mối quan hệ kinh doanh nào với một thực thể cụ thể. (Tsai, Lu, et al.) Sự rời bỏ mạnh mẽ xảy ra khi một khách hàng vi phạm các quy tắc truyền thống để tiến hành kinh doanh trong một địa điểm cụ thể. Trong tình huống này, có thể

là kết quả của các hoạt động bất hợp pháp như trộm cắp hoặc không thanh toán kịp thời. Một loại tự nguyện sự thay đổi được gây ra bởi các vấn đề liên quan đến người tiêu dùng, sự hài lòng hoặc mất đi lợi thế cạnh tranh của tổ chức. (Burez và Van Den Poel) Bất kỳ công ty nào giao dịch với nhiều đối tượng người tiêu dùng khác nhau nên thường xuyên thay đổi các phương pháp đánh giá của mình.

Ngăn chặn tình trạng rời bỏ đã trở thành một trong những vấn đề quan trọng nhất, mục tiêu của các tổ chức do sự cạnh tranh ngày càng tăng đối với khách hàng và khó khăn trong việc thay thế sự mất mát của doanh thu do sự rời bỏ của những khách hàng có lợi nhuận. Tuy nhiên, như đã thảo luận, việc giữ chân khách hàng hiện tại hiện là một trong những thách thức lớn nhất của hoạt động tài chính. Các tổ chức trong một thị trường bão hòa và cạnh tranh, nơi khách hàng ngày càng có khả năng chuyển sang dịch vụ khác nhà cung cấp. Trong bối cảnh này, sự phát triển của chính xác và các mô hình thống kê hiệu suất cao cho phép xác định trước những khách hàng có xu hướng rời bỏ trở thành điều kiện thiết yếu để duy trì khả năng cạnh tranh của các công ty này.

Một số nghiên cứu cũng đã tích hợp việc sử dụng kỹ thuật phân khúc khách hàng và học máy để cải thiện độ chính xác của mô hình dự đoán. Phân khúc khách hàng là quá trình phân chia khách hàng thành các nhóm dựa trên đặc điểm chung của khách hàng mục tiêu để công ty có thể tiếp thị đến từng nhóm một cách hiệu quả và phù hợp. Đây là bước khá quan trọng, là yếu tố giúp tăng tỷ lệ chuyển đổi cho doanh nghiệp. Nếu doanh nghiệp thực hiện tốt bước này có thể giúp họ phân chia ngân sách quảng cáo của họ tốt hơn và tiết kiệm nhiều hơn. Ngoài ra, việc phân khúc khách hàng sẽ giúp họ hiểu rõ hơn về người tiêu dùng, xác định nhóm khách hàng mục tiêu để doanh nghiệp tập trung vào và sau đó xem xét các yếu tố ảnh hưởng đến việc rời bỏ khách hàng đó.

Nghiên cứu trước đây thường tập trung vào việc xác định các biến số có ảnh hưởng đáng kể nhất đến việc rời bỏ khách hàng hoặc chỉ đơn giản sử dụng phân khúc khách hàng để xác định một tập hợp con người tiêu dùng tiềm năng, loại trừ ảnh hưởng của nó đến độ chính xác của dự báo. Do đó, có hai mục tiêu nghiên cứu chính

trong công việc này. Mục tiêu ban đầu là kiểm tra tác động của phân khúc khách hàng đến độ chính xác của

dự đoán tỷ lệ rời bỏ khách hàng trong lĩnh vực ngân hàng bằng cách sử dụng mô hình học máy. Mục tiêu thứ hai là thử nghiệm, đối chiếu và đánh giá phương pháp học máy nào hiệu quả nhất trong

dự đoán sự rời bỏ của khách hàng. Mối quan hệ với khách hàng có tác động lớn đến sự phát triển lâu dài,

do đó, hiểu được hành vi của họ cho phép các ngân hàng dễ dàng cải thiện các chính sách hiện có.

Nghiên cứu nhằm mục đích phát triển và so sánh các hoạt động được giám sát các thuật toán học máy như RF, LR, GBC, EGBC và LGBMC để dự đoán tỷ lệ rời bỏ khách hàng bằng cách sử dụng quảng cáo bộ dữ liệu của ngân hàng Ethiopia. Chơi thuật toán học máy, vai trò quan trọng trong việc phân tích dự báo hoạt động kinh doanh các tổ chức như ngân hàng và tổ chức tài chính Kumar và cộng sự, (2021). Thuật toán học máy có giám sát là một thuật toán học máy được sử dụng trong phân loại trường hợp các vấn đề. Phân loại là quá trình tách biệt mọi thứ thành các lớp riêng biệt. Dự đoán tỷ lệ rời bỏ khách hàng là một vấn đề phân loại nhằm xác định liệu khách hàng sẽ bỏ cuộc hay không.