**Tài liệu tham khảo**

|  |  |  |
| --- | --- | --- |
|  | Cite | Abstract |
|  | @article{Lemos2022PropensionTC, title={Propension to customer churn in a financial institution: a machine learning approach}, author={Renato Alexandre de Lima Lemos and Thiago Christiano Silva and Benjamin Miranda Tabak}, journal={Neural Computing \& Applications}, year={2022}, volume={34}, pages={11751 - 11768}, url={https://api.semanticscholar.org/CorpusID:247272933} } | This paper examines churn prediction of customers in the banking sector using a unique customer-level dataset from a large Brazilian bank. Our main contribution is in exploring this rich dataset, which contains prior client behavior traits that enable us to document new insights into the main determinants predicting future client churn. We conduct a horserace of many supervised machine learning algorithms under the same cross-validation and evaluation setup, enabling a fair comparison across algorithms. We find that the random forests technique outperforms decision trees, k-nearest neighbors, elastic net, logistic regression, and support vector machines models in several metrics. Our investigation reveals that customers with a stronger relationship with the institution, who have more products and services, who borrow more from the bank, are less likely to close their checking accounts. Using a back-of-the-envelope estimation, we find that our model has the potential to forecast potential losses of up to 10% of the operating result reported by the largest Brazilian banks in 2019, suggesting the model has a significant economic impact. Our results corroborate the importance of investing in cross-selling and upselling strategies focused on their current customers. These strategies can have positive side effects on customer retention.e strategies can have positive side effects on customer retention. |
|  | @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{Matuszelaski2022CustomerCI, title={Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach}, author={Kamil Matuszelański and Katarzyna Kopczewska}, journal={J. Theor. Appl. Electron. Commer. Res.}, year={2022}, volume={17}, pages={165-198}, url={https://api.semanticscholar.org/CorpusID:246019532} } | This study is a comprehensive and modern approach to predict customer churn in the example of an e-commerce retail store operating in Brazil. Our approach consists of three stages in which we combine and use three different datasets: numerical data on orders, textual after-purchase reviews and socio-geo-demographic data from the census. At the pre-processing stage, we find topics from text reviews using Latent Dirichlet Allocation, Dirichlet Multinomial Mixture and Gibbs sampling. In the spatial analysis, we apply DBSCAN to get rural/urban locations and analyse neighbourhoods of customers located with zip codes. At the modelling stage, we apply machine learning extreme gradient boosting and logistic regression. The quality of models is verified with area-under-curve and lift metrics. Explainable artificial intelligence represented with a permutation-based variable importance and a partial dependence profile help to discover the determinants of churn. We show that customers’ propensity to churn depends on: (i) payment value for the first order, number of items bought and shipping cost; (ii) categories of the products bought; (iii) demographic environment of the customer; and (iv) customer location. At the same time, customers’ propensity to churn is not influenced by: (i) population density in the customer’s area and division into rural and urban areas; (ii) quantitative review of the first purchase; and (iii) qualitative review summarised as a topic. |
|  | @article{Sudharsan2022ASR, title={A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy}, author={R. Sudharsan and E. N. Ganesh}, journal={Connection Science}, year={2022}, volume={34}, pages={1855 - 1876}, url={https://api.semanticscholar.org/CorpusID:249940935} } | Owing to saturated markets, fierce competition, dynamic criteria, along with introduction of new attractive offers, the considerable issue of customer churn was faced by the telecommunication industry. Thus, an efficient Churn Prediction (CP) model is required for monitoring customer churn. Therefore, this work proposes a novel framework to predict customer churn through a deep learning model namely Swish Recurrent Neural Network (S-RNN). Finally, SRNN is adapted to classify the Churn Customer (CC) and a normal customer. If the result is a churn customer, network utilisation history is analysed for retention process. Whereas, the number of churn customers based on the area network usage is not recognised in this frameworkOwing to saturated markets, fierce competition, dynamic criteria, along with introduction of new attractive offers, the considerable issue of customer churn was faced by the telecommunication industry. Thus, an efficient Churn Prediction (CP) model is required for monitoring customer churn. Therefore, this work proposes a novel framework to predict customer churn through a deep learning model namely Swish Recurrent Neural Network (S-RNN). Finally, S-RNN is adapted to classify the Churn Customer (CC) and a normal customer. If the result is a churn customer, network utilisation history is analysed for retention process. |
|  | @article{Rahmaty2022CustomerCM, title={Customer Churn Modeling via the Grey Wolf Optimizer and Ensemble Neural Networks}, author={Maryam Rahmaty and Amir Daneshvar and Fariba Salahi and Maryam Ebrahimi and Adel Pourghader Chobar}, journal={Discrete Dynamics in Nature and Society}, year={2022}, url={https://api.semanticscholar.org/CorpusID:248899754} } | The customer churn is one of the key challenges for enterprises, and market saturation and increased competition to maintain business position has caused companies to make all attempts to identify customers who are likely to leave and end their relationship with a company in a particular period to become the customer of another company. In recent years, many methods have been developed including data mining for predicting the customer churn and manners that customers are likely to behave in the future and therefore, taking action early to prevent their leaving. This study proposes a hybrid system based on fuzzy entropy criterion selection algorithm with similar classifiers, grey wolf optimization algorithm, and artificial neural network to predict the customer churn of those companies that suffer losses from losing customers over time. The research results are evaluated by other methods in the criteria of accuracy, recall, precision, and F\_measure, and it is declared that the proposed method is superior over other methods. |
|  | @article{Zhang2022ADA, title={A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation}, author={Tianyuan Zhang and S{\'e}rgio Moro and Ricardo F. Ramos}, journal={Future Internet}, year={2022}, volume={14}, pages={94}, url={https://api.semanticscholar.org/CorpusID:247532720} } | Numerous valuable clients can be lost to competitors in the telecommunication industry, leading to profit loss. Thus, understanding the reasons for client churn is vital for telecommunication companies. This study aimed to develop a churn prediction model to predict telecom client churn through customer segmentation. Data were collected from three major Chinese telecom companies, and Fisher discriminant equations and logistic regression analysis were used to build a telecom customer churn prediction model. According to the results, it can be concluded that the telecom customer churn model constructed by regression analysis had higher prediction accuracy (93.94%) and better results. This study will help telecom companies efficiently predict the possibility of and take targeted measures to avoid customer churn, thereby increasing their profits. |
|  | @article{Kostic2020SocialNA, title={Social Network Analysis and Churn Prediction in Telecommunications Using Graph Theory}, author={Stefan M. Kostic and Mirjana Simi{\'c} and Miroljub V. Kostic}, journal={Entropy}, year={2020}, volume={22}, url={https://api.semanticscholar.org/CorpusID:221141679} } | Due to telecommunications market saturation, it is very important for telco operators to always have fresh insights into their customer’s dynamics. In that regard, social network analytics and its application with graph theory can be very useful. In this paper we analyze a social network that is represented by a large telco network graph and perform clustering of its nodes by studying a broad set of metrics, e.g., node in/out degree, first and second order influence, eigenvector, authority and hub values. This paper demonstrates that it is possible to identify some important nodes in our social network (graph) that are vital regarding churn prediction. We show that if such a node leaves a monitored telco operator, customers that frequently interact with that specific node will be more prone to leave the monitored telco operator network as well; thus, by analyzing existing churn and previous call patterns, we proactively predict new customers that will probably churn. The churn prediction results are quantified by using top decile lift metrics. The proposed method is general enough to be readily adopted in any field where homophilic or friendship connections can be assumed as a potential churn driver. |
|  | @article{Dalli2022ImpactOH, title={Impact of Hyperparameters on Deep Learning Model for Customer Churn Prediction in Telecommunication Sector}, author={Anouar Dalli}, journal={Mathematical Problems in Engineering}, year={2022}, url={https://api.semanticscholar.org/CorpusID:246723633} } | In this paper, in order to predict a customer churn in the telecommunication sector, we have analysed several published articles that had used machine learning (ML) techniques. Significant predictive performance had been seen by utilising deep learning techniques. However, we have seen a tremendous lack of empirically derived heuristic information where we had to influence the hyperparameters consequently. Here, we had demonstrated three experimental findings, where a Relu activation function was embedded and utilised successfully in the hidden layers of the deep network. We can also see that the output layer had the service ability of a sigmoid function, in which we had seen a significant performance of the neural network model and obviously it was improved. Furthermore, we had also seen that the model's performance was noticed to be even better, but it was only considered better though when the batch size in the model was taken less than the test dataset’s size, respectively. In terms of accuracy, the RemsProp optimizer beat out the other algorithms such as stochastic gradient descent (SGD). RemsProp was seen even better from the Adadelta algorithm, the Adam algorithm, the AdaGrad algorithm, and AdaMax algorithm as well. |
|  | @article{Alzubaidi2020ProjectionPR, title={Projection pursuit random forest using discriminant feature analysis model for churners prediction in telecom industry}, author={Asia Mahdi Naser Alzubaidi and Eman Salih Al-Shamery}, journal={International Journal of Electrical and Computer Engineering (IJECE)}, year={2020}, url={https://api.semanticscholar.org/CorpusID:209049251} } | A major and demanding issue in the telecommunications industry is the prediction of churn customers. Churn describes the customer who is attrite from one Telecom service provider to competitors searching for better services offers. Companies from the Telco sector frequently have customer relationship management offices it is the main objective in how to win back defecting clients because preserve long-term customers can be much more beneficial to a company than gain newly recruited customers. Researchers and practitioners are paying great attention and investing more in developing a robust customer churn prediction model, especially in the telecommunication business by proposed numerous machine learning approaches. Many approaches of Classification are established, but the most effective in recent times is a tree-based method. The main contribution of this research is to predict churners/non-churners in the Telecom sector based on project pursuit Random Forest (PPForest) that uses discriminant feature analysis as a novelty extension of the conventional Random Forest approach for learning oblique Project Pursuit tree (PPtree). The proposed methodology leverages the advantage of two discriminant analysis methods to calculate the project index used in the construction of PPtree. The first method used Support Vector Machines (SVM) as a classifier in the construction of PPForest to differentiate between churners and non-churners customers. The second method is a Linear Discriminant Analysis (LDA) to achieve linear splitting of variables node during oblique PPtree construction to produce individual classifiers that are robust and more diverse than classical Random Forest. It found that the proposed methods enjoy the best performance measurements e.g. Accuracy, hit rate, ROC curve, Gini coefficient, Kolmogorov-Smirnov statistic and lift coefficient, H-measure, AUC. Moreover, PPForest based on direct applied of LDA on the raw data delivers an effective evaluator for the customer churn prediction model. |
|  | @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} } | Aim/Purpose: 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. |
|  | @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{Jajam2023ArithmeticOW, title={Arithmetic Optimization With Ensemble Deep Learning SBLSTM-RNN-IGSA Model for Customer Churn Prediction}, author={Nagaraju Jajam and Nagendra Panini Challa and Kamepalli SL Prasanna and Venkata Sasi Deepthi Ch}, journal={IEEE Access}, year={2023}, volume={11}, pages={93111-93128}, url={https://api.semanticscholar.org/CorpusID:260911095} } | Companies in a wide variety of industries use the customer churn prediction (CCP) process to keep their current clientele happy. Insurance companies need to be able to forecast churn to enhance the potency and functionality of deep learning methods. Deep learning techniques have a significant impact on improving and forecasting customer retention. Numerous studies employ standard machine learning and Deep Learning strategies to enhance customer retention, despite the fact that these strategies have a number of accuracy issues. In light of this need, this piece is dedicated to the development of a stacked bidirectional long short-term memory (SBLSTM) and RNN model for Arithmetic Optimisation Algorithm (AOA) in CCP. The proposed AOA-SBLSTM-RNN model intends to proficiently forecast the occurrence of Customer Churn in the Insurance industry. Initially, the AOA model performs pre-processing to transform the original data into a useful format. In addition, the SBLSTM-RNN model is used to distinguish between churning and non-churning customers. To improve the CCP outcomes of the SBLSTM-RNN model, an optimal Hyperparameters tuning process using Improved Gravitational Search Optimization Algorithm (IGSA) is used in this study. In this work, Three Health Insurance datasets were used to evaluate performance, and four sets of experiments were conducted. The Measures of true churn, false churn, specificity, precision, and accuracy are employed to assess the efficacy of the proposed approach. Experimental result shows that the Ensemble Deep Learning model AOA-SBLSTM-RNN with IGSA produces accuracy value of 97.89 and 97.67 on dataset 2 and dataset 1. which is better and had higher predictability levels in compared with all other models. |
|  | @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{Alizadeh2023DevelopmentOA, title={Development of a Customer Churn Model for Banking Industry Based on Hard and Soft Data Fusion}, author={Masoud Alizadeh and Danial Sadrian Zadeh and Behzad Moshiri and Allahyar Montazeri}, journal={IEEE Access}, year={2023}, volume={11}, pages={29759-29768}, url={https://api.semanticscholar.org/CorpusID:257574015} } | There has been an increase in customer churn over the past few years—customers decide not to continue purchasing products or services from an organization. Customers’ data lie in two categories: soft and hard. The term “hard data” refers to the records generated by various devices and programs, including but not limited to smartphones, computers, sensors, smart meters, fleet management systems, call detail records (CDRs), and consumer bank transaction data. On the other hand, information that is subject to interpretation and viewpoint is known as “soft data.” Fusing these two types of data leads to better customer’s behavior analysis. This paper uses a supervised machine learning algorithm, namely a decision tree (DT), and the change mining method to model hard data. K-means clustering, an unsupervised machine learning algorithm, is also used along with the data preprocessing techniques. This paper also considers the Dempster-Shafer theory and other steps for soft data modeling. By fusing soft and hard data, the churn rate of customers compared with each other can be calculated. Besides, the customers’ banking data are leveraged for data modeling. The results show that the banking industry will gain a more dynamic and efficient customer relationship management system by using this model.  Collapse |
|  | @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{Nalatissifa2021CustomerDP, title={Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data}, author={Hiya Nalatissifa and Hilman Ferdinandus Pardede}, journal={Jurnal Elektronika dan Telekomunikasi}, year={2021}, url={https://api.semanticscholar.org/CorpusID:245630188} } | Customer churn is the most important problem in the business world, especially in the telecommunications industry, because it greatly influences company profits. Getting new customers for a company is much more difficult and expensive than retaining existing customers. Machine learning, part of data mining, is a sub-field of artificial intelligence widely used to make predictions, including predicting customer churn. Deep neural network (DNN) has been used for churn prediction, but selecting hyperparameters in modeling requires more time and effort, making the process more challenging for the researcher. Therefore, the purpose of this study is to propose a better architecture for the DNN algorithm by using a hard tuner to obtain more optimal hyperparameters. The tuning hyperparameter used is random search in determining the number of nodes in each hidden layer, dropout, and learning rate. In addition, this study also uses three variations of the number of hidden layers, two variations of the activation function, namely rectified linear unit (ReLu) and Sigmoid, then uses five variations of the optimizer (stochastic gradient descent (SGD), adaptive moment estimation (Adam), adaptive gradient algorithm (Adagrad), Adadelta, and root mean square propagation (RMSprop)). Experiments show that the DNN algorithm using hyperparameter tuning random search produces a performance value of 83.09 % accuracy using three hidden layers, the number of nodes in each hidden layer is [20, 35, 15], using the RMSprop optimizer, dropout 0.1, the learning rate is 0.01, with the fastest tuning time of 21 seconds. Better than modeling using k-nearest neighbor (K-NN), random forest (RF), and decision tree (DT) as comparison algorithms. |
|  | @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{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{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{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{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{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{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. |
|  |  |  |

**1.Giới thiệu bài toán**

Khách hàng rời bỏ, một hiện tượng gây ra tổn thất tài chính lớn khi khách hàng rời bỏ doanh nghiệp, khiến các tổ chức hiện đại gặp khó khăn trong việc giữ chân khách hàng. Khi khách hàng không hài lòng nhận thấy dịch vụ của công ty hiện tại không phù hợp, họ thường chuyển sang nhà cung cấp dịch vụ khác. Các phương pháp học máy và học sâu (ML/DL) đã được sử dụng để xác định thành công tình trạng rời bỏ khách hàng. Tuy nhiên, trong một số trường hợp, các thuật toán dựa trên ML/DL thiếu khả năng mang lại kết quả đầy hứa hẹn trong việc phát hiện tình trạng rời bỏ khách hàng. Nghiên cứu trước đây về việc ước tính tỷ lệ rời bỏ khách hàng đã tiết lộ những dự báo bất ngờ khi sử dụng các bộ phân loại học máy và các phương pháp mã hóa tính năng truyền thống. Mạng lưới thần kinh sâu cũng được sử dụng trong nỗ lực trích xuất các đặc điểm mà không tính đến thông tin trình tự. Trước những vấn đề này, nghiên cứu hiện tại cung cấp một phương pháp hiệu quả để dự đoán tỷ lệ rời bỏ khách hàng dựa trên mô hình học sâu kết hợp có tên BiLSTM-CNN. Mục tiêu là ước tính tỷ lệ rời bỏ khách hàng một cách hiệu quả bằng cách sử dụng dữ liệu chuẩn và tăng độ chính xác của quá trình dự đoán tỷ lệ rời bỏ. Kết quả thực nghiệm cho thấy khi được huấn luyện, kiểm tra và xác thực trên tập dữ liệu benchmark, mô hình BiLSTM-CNN đề xuất đạt độ chính xác vượt trội 81%.

Động lực nghiên cứu.Việc rời bỏ khách hàng có tác động đáng kể đến doanh nghiệp, dẫn đến lợi nhuận hoặc thua lỗ có thể xảy ra và thậm chí có khả năng đóng cửa doanh nghiệp3. Khách hàng có thể chuyển nhượng nhà cung cấp vì nhiều lý do, bao gồm giá cả, sản phẩm

phương thức giao hàng và các cuộc gặp gỡ dịch vụ khách hàng. Sự mất khách hàng cũng có thể được gây ra bởi các vấn đề như chất lượng sản phẩm hoặc cảm nhận thiếu an toàn. Sự rời bỏ của khách hàng cũng có thể được gây ra bởi sự không hài lòng với hiện tại cung cấp hoặc yêu cầu chưa được đáp ứng. Khi người tiêu dùng rời đi, doanh nghiệp phải gánh chịu chi phí đáng kể, khiến khách hàng việc duy trì rất quan trọng cho khả năng tồn tại của nền kinh tế. Dự đoán tỷ lệ rời bỏ khách hàng thông qua phân tích dữ liệu đã trở nên quan trọng để thu hút và giữ chân khách hàng, vì nó cho phép các doanh nghiệp dự đoán các lý do có thể xảy ra đối với doanh thu của khách hàng và có những hành động sớm để giải quyết chúng.

Các phương pháp dự đoán tỷ lệ rời bỏ truyền thống thường có những lo ngại về quy mô. Đối với các trình phân loại học máy, một số nghiên cứu dựa vào các phương pháp kỹ thuật tính năng của con người. Gupta và cộng sự đã sử dụng KNN để phân loại trong nghiên cứu trước đây. Tuy nhiên, những mô hình này không cung cấp một phương pháp hiệu quả để xác định những khách hàng có khả năng rời khỏi tổ chức.

1. **Nghiên cứu liên quan**

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.