|  |  |  |
| --- | --- | --- |
| iid | Cite | Abstract |
|  | @article{Wu2021IntegratedCP,  title={Integrated Churn Prediction and Customer Segmentation Framework for Telco Business},  author={Shuli Wu and Wei-Chuen Yau and Thian Song Ong and Siew-Chin Chong},  journal={IEEE Access},  year={2021},  volume={9},  pages={62118-62136}  } | [Integrated Churn Prediction and Customer Segmentation Framework for Telco Business](https://www.semanticscholar.org/paper/Integrated-Churn-Prediction-and-Customer-Framework-Wu-Yau/1526084beca5dbebbee1924df751e3b447935008)  In the telco industry, attracting new customers is no longer a good strategy since the cost of retaining existing customers is much lower. Churn management becomes instrumental in the telco industry. As there is limited study combining churn prediction and customer segmentation, this paper aims to propose an integrated customer analytics framework for churn management. There are six components in the framework, including data pre-processing, exploratory data analysis (EDA), churn prediction, factor analysis, customer segmentation, and customer behaviour analytics. This framework integrates churn prediction and customer segmentation process to provide telco operators with a complete churn analysis to better manage customer churn. Three datasets are used in the experiments with six machine learning classifiers. First, the churn status of the customers is predicted using multiple machine learning classifiers. Synthetic Minority Oversampling Technique (SMOTE) is applied to the training set to deal with the problems with imbalanced datasets. The 10-fold cross-validation is used to assess the models. Accuracy and F1-score are used for model evaluation. F1-score is considered to be an important metric to measure the models for imbalanced datasets since the premise of churn management is to be able to identify customers who will churn. Experimental analysis indicates that AdaBoost performed the best in Dataset 1, with accuracy of 77.19% and F1-score of 63.11%. Random Forest performed the best in Dataset 2, with accuracy of 93.6% and F1-score of 77.20%. Random Forest performed the best in Dataset 3 in terms of accuracy, at 63.09%, while Multi-layer Perceptron performed the best in terms of F1-score, at 42.84%. After implementing churn prediction, Bayesian Logistic Regression is used to conduct the factor analysis and to figure out some important features for churn customer segmentation. Churn customer segmentation is then carried out using K-means clustering. Customers are segmented into different groups, which allows marketers and decision makers to adopt retention strategies more precisely |
|  | @article{PejiBach2021ChurnMI,  title={Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees},  author={Mirjana Peji{\'c} Bach and Jasmina Pivar and Bo{\vz}idar Jakovi{\'c}},  journal={Journal of Risk and Financial Management},  year={2021},  url={https://api.semanticscholar.org/CorpusID:244089395}  } | [Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees](https://www.semanticscholar.org/paper/Churn-Management-in-Telecommunications%3A-Hybrid-and-Bach-Pivar/710b9636c00d70188515d493289823ee653e6059)  The goal of the paper is to present the framework for combining clustering and classification for churn management in telecommunications. Considering the value of market segmentation, we propose a three-stage approach to explain and predict the churn in telecommunications separately for different market segments using cluster analysis and decision trees. In the first stage, a case study churn dataset is prepared for the analysis, consisting of demographics, usage of telecom services, contracts and billing, monetary value, and churn. In the second stage, k-means cluster analysis is used to identify market segments for which chi-square analysis is applied to detect the clusters with the highest churn ratio. In the third stage, the chi-squared automatic interaction detector (CHAID) decision tree algorithm is used to develop classification models to identify churn determinants at the clusters with the highest churn level. The contribution of this paper resides in the development of the structured approach to churn management using clustering and classification, which was tested on the churn dataset with a rich variable structure. The proposed approach is continuous since the results of market segmentation and rules for churn prediction can be fed back to the customer database to improve the efficacy of churn management |
|  | @article{Naz2022PredictiveMO,  title={Predictive Modeling of Employee Churn Analysis for IoT-Enabled Software Industry},  author={Komal Naz and Isma Farah Siddiqui and Jahwan Koo and Mohammad Ali Khan and Nawab Muhammad Faseeh Qureshi},  journal={Applied Sciences},  year={2022},  url={https://api.semanticscholar.org/CorpusID:253030524}  } | [Predictive Modeling of Employee Churn Analysis for IoT-Enabled Software Industry](https://www.semanticscholar.org/paper/Predictive-Modeling-of-Employee-Churn-Analysis-for-Naz-Siddiqui/e351a7fbf5b9c505b01c4a8e37b99190bf460a22)  Employee churn analytics is the process of assessing employee turnover rate and predicting churners in a corporate company. Due to the rapid requirement of experts in the industries, an employee may switch workplaces, and the company then has to look for a substitute with the training to deal with the tasks. This has become a bottleneck and the corporate sector suffers with additional cost overheads to restore the work routine in the organization. In order to solve this issue in a timely manner, we identify several ML techniques that examine an employee’s record and assess factors in generalized ways to assess whether the resource will remain to continue working or may leave the workplace with the passage of time. However, sensor-based information processing is not much explored in the corporate sector. This paper presents an IoT-enabled predictive strategy to evaluate employee churn count and discusses the factors to decrease this issue in the organizations. For this, we use filter-based methods to analyze features and perform classification to identify firm future churners. The performance evaluation shows that the proposed technique efficiently identifies the future churners with 98% accuracy in the IoT-enabled corporate sector organizations |
|  | @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 churn analysis using XGBoosted decision trees](https://www.semanticscholar.org/paper/Customer-churn-analysis-using-XGBoosted-decision-Vasudevan-Narayanan/1417e2d7a4100783d6f593a8f8682b4a69294022)  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{Papa2021AnalysisOF,  title={Analysis of fuzzy logic methods for forecasting customer churn},  author={Andrii Papa and Yevhen O. Shemet and Andrii Yarovyi},  journal={Technology audit and production reserves},  year={2021},  volume={1},  pages={12-14},  url={https://api.semanticscholar.org/CorpusID:233786643}  } | [Analysis of fuzzy logic methods for forecasting customer churn](https://www.semanticscholar.org/paper/Analysis-of-fuzzy-logic-methods-for-forecasting-Papa-Shemet/8a0586d8b532aa3a7912b51aa6c41389681425e7)  The object of research is the process of predicting the churn of customers of telecommunications companies based on fuzzy logic and neural networks. The research carried out is based on the application of an approach that is implemented through the combined use of fuzzy logic and neural networks. The main assumption of the study is the hypothesis that the use of a fuzzy neural network formed on the basis of fuzzy logic algorithms can improve the accuracy of predicting customer churn relative to available solutions. This result can’t be achieved neglecting the existing resource constraints and requirements, which must be determined separately for each case of research. The relevance of the problem of forecasting customer churn for companies with a large number of users is considered. A model for predicting customer churn is proposed based on the combined use of fuzzy logic and neural networks. The main feature of this approach is that a test sample of normalized data is used at the basis of fuzzy neural networks, which are processed to form the parameters of membership functions that correspond to the inference system, that is, conclusions are made on the basis of a fuzzy logic apparatus. Also, to find the parameters of the membership function, neural network algorithms are used. Such systems can use previously known information, learn, gain new knowledge, predict time series, perform image classification, and besides, they are quite visual to the user. The application of methods of fuzzy logic is considered, they make it possible to obtain a result in the form of a fuzzy inference. The expediency of choosing these methods is explained by the fact that they were previously used in fuzzy automatic control systems and showed sufficiently high quality results. The expediency and prospects of using the proposed approach in the problem of predicting the outflow of customers of telecommunications companies are shown, and the results of software implementation are presented |
|  | @inproceedings{Bagul2021RetailCC,  title={Retail Customer Churn Analysis using RFM Model and K-Means Clustering},  author={Nikita Bagul and Priya Surana and Prerana Berad and Chirag Khachane},  year={2021},  url={https://api.semanticscholar.org/CorpusID:235806270}  } | [Retail Customer Churn Analysis using RFM Model and K-Means Clustering](https://www.semanticscholar.org/paper/Retail-Customer-Churn-Analysis-using-RFM-Model-and-Bagul-Surana/f74e13df79b60ac5382463b7ef137e597adbf0ff)  In this current world of business, Customer Churn is one of the major concerns for various business owners or the organizations for maintaining existing and attracting new customers. Analysis of various types of customers can be conducted by researching customer relationship management which in turn provides strong support for business decisions. Customer churn occurs when certain customers are no longer loyal or a part of a particular business. Losing customers will not only result in losses but also develop threat to the organization. Because of multiple competitors in the same business, the re-engagement of customers who are less interested is essential rather than engaging a new one. It is observed that acquiring new buyers is costlier than retaining the present customer. Churn prediction is a new promising method in customer relationship management to analyze customer behavior by identifying customers with a high probability to discontinue the company based on analyzing their past data and also identify strategies for improvement. Once a customer becomes a churn, the loss incurred by the corporate isn't just the lost revenue but also the prices involved in additional marketing in order to attract new customer. Reducing customer churn is a key business goal. In this project, we've taken a dataset from UCI Machine Learning repository. This dataset contains records of transactions that happened between December 1, 2010 and December 1, 2011. This is recorded from a web retail gift store based in the United Kingdom. Here segmentation of customers has been done by using RFM technique and K-means algorithm. |
|  | @article{Srivastava2021IntelligentER,  title={Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach},  author={Praveen Ranjan Srivastava and Prajwal Eachempati},  journal={J. Glob. Inf. Manag.},  year={2021},  volume={29},  pages={1-29},  url={https://api.semanticscholar.org/CorpusID:237661612}  } | [Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach](https://www.semanticscholar.org/paper/Intelligent-Employee-Retention-System-for-Attrition-Srivastava-Eachempati/198f4709673da5f97b2a5ca3ae279bc1dee630a0)  The paper aims to examine the factors that influence employee attrition rate using an employee records dataset. It also aims to establish the predictive power of deep learning for employee churn prediction over ensemble machine learning techniques like random forest and gradient boosting on real-time employee data from a mid-sized fast-moving consumer goods (FMCG) company. The results are further validated through a regression model and a multi-criteria fuzzy analytical hierarchy process (AHP) model that considers the relative variable importance and computes weights. The machine learning models’ empirical results indicate that deep neural networks (91.6% accuracy) is a better predictor of churn than random forest and gradient boosting algorithm (82.5% and 85.4%, respectively). These findings provide useful insights for human resource (HR) managers in an organizational workplace context. When recalibrated by organizations’ human resource team, the model helps in better incentivization and more targeted employee retention. |
|  | @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}  } | [Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art](https://www.semanticscholar.org/paper/Ensemble-Methods-in-Customer-Churn-Prediction%3A-A-of-Bogaert-Delaere/4fbc5c03e7d6598853ce3a6b3c41e093d21db460)  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{Mahalekshmi2022AnalysisOC,  title={Analysis of customer churn prediction using machine learning and deep learning algorithms},  author={A Mahalekshmi and G. Heren Chellam},  journal={International journal of health sciences},  year={2022},  url={https://api.semanticscholar.org/CorpusID:249137845}  } | [Analysis of customer churn prediction using machine learning and deep learning algorithms](https://www.semanticscholar.org/paper/Analysis-of-customer-churn-prediction-using-machine-Mahalekshmi-Chellam/02ee073d228559a629d2958407e6900f929037e0)  -The telecommunication industry need a customer churn prediction due to many competitors. The companies also lack of churn prediction to retain the customer. This problem not only affect the growth of the business but also affect the revenues. To retain the existing customer is very crucial task for the company. A rapid increasing in technology, the various machine learning and deep learning are tools are developed which can be used by telecoms companies to monitor the churn behaviour of customers. In this study, a brief idea on the customer churn problem on various machine learning techniques such as XGBoost, Gradient Boost, AdaBoost, ANN, Logistic Regression and Random Forest are analysed. Also the various deep learning techniques such as Convolutional Neural Network, stacked auto encoders to predict the customer churn problem are analysed by comparing the models in terms of accuracy. |
|  | @article{Batta2023CrossPlatformAO,  title={Cross-Platform Analysis of Seller Performance and Churn for Ecommerce Using Artificial Intelligence},  author={Anuj Batta and Arpan Kumar Kar and Shyamali Satpathy},  journal={J. Glob. Inf. Manag.},  year={2023},  volume={31},  pages={1-21},  url={https://api.semanticscholar.org/CorpusID:258466882}  } | [Cross-Platform Analysis of Seller Performance and Churn for Ecommerce Using Artificial Intelligence](https://www.semanticscholar.org/paper/Cross-Platform-Analysis-of-Seller-Performance-and-Batta-Kar/4513264d39b55aa8808f6b9bfb9878678e3dabef)  Suppliers and sellers play a crucial role in the ecommerce ecosystem. Sellers and ecommerce firms use social media to increase user engagement, visibility, and sales. Seller ratings are as important as the product ratings on ecommerce platforms to drive buying decisions. Based on sellers’ actions on social media, this study examines seller turnover and disengagement on e-commerce platforms. The study has been supported by the justice theory. Seller reviews and ratings from e-commerce platforms and conversations from social media platforms have been gathered. Using natural language processing, machine learning, partial least squares (PLS) path analysis, and statistical inferences, objectives of the study are met. The study offers recommendations for both practitioners and researchers. The sellers must focus more on interaction and communication than marketing. Through a longitudinal analysis, the study also establishes that ecommerce organizations can use seller social media performance as a predictor of future seller churn and disengagement so they can take the necessary remedial action. |
|  | @article{Groll2022ChurnMO,  title={Churn modeling of life insurance policies via statistical and machine learning methods - Analysis of important features},  author={Andreas Groll and Carsten Wasserfuhr and Leonid Zeldin},  journal={ArXiv},  year={2022},  volume={abs/2202.09182},  url={https://api.semanticscholar.org/CorpusID:246996810}  } | [Churn modeling of life insurance policies via statistical and machine learning methods - Analysis of important features](https://www.semanticscholar.org/paper/Churn-modeling-of-life-insurance-policies-via-and-Groll-Wasserfuhr/f5238bdaad92a2a9a309bb1af89f63114c914af0)  Life assurance companies typically possess a wealth of data covering multiple systems and databases. These data are often used for ana- lyzing the past and for describing the present. Taking account of the past, the future is mostly forecasted by traditional statistical methods. So far, only a few attempts were undertaken to perform estimations by means of machine learning approaches. In this work, the individual contract cancellation behavior of customers within two partial stocks is modeled by the aid of various classification methods. Partial stocks of private pension and endowment policy are considered. We describe the data used for the modeling, their structured and in which way they are cleansed. The utilized models are calibrated on the basis of an extensive tuning process, then graphically evaluated regarding their goodness-of-fit and with the help of a variable relevance concept, we investigate which features notably affect the individual contract cancellation behavior. |
|  | @article{Jiao2021AnalysisAC,  title={Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn},  author={Guie Jiao and Hong Xu},  journal={Journal of Physics: Conference Series},  year={2021},  volume={1881},  url={https://api.semanticscholar.org/CorpusID:235282172}  } | [Analysis and Comparison of Forecasting Algorithms for Telecom Customer Churn](https://www.semanticscholar.org/paper/Analysis-and-Comparison-of-Forecasting-Algorithms-Jiao-Xu/bef4e26b4f9d6194432ea6d158110174d266b231)  The integrated algorithm is a highly flexible data analysis and prediction algorithm. In many big data competitions at home and abroad, the winning teams basically adopt the idea of integrated algorithms such as random forest, GBDT, XGBoost and other algorithms. This shows that accuracy of ensemble algorithms is still very advantageous in terms of predictive classification. The main task of this article is to predict the loss of telecom customers. Under the current situation of saturation of the telecom market, how to retain the original customers is the main task of each telecom operator. This article mainly compares the four prediction models on the telecom data set. Predictive performance, the final performance evaluation index also shows that the random forest model and XGBoost model of integrated thought have better predictive models. |
|  | @article{Sarac2021AHT,  title={A Hybrid Two-Level Support Vector Machine-Based Method for Churn Analysis},  author={Ferdi Sarac and Huseyin Seker and Marcin Lisowski and Alan Timothy},  journal={Proceedings of the 2021 5th International Conference on Cloud and Big Data Computing},  year={2021},  url={https://api.semanticscholar.org/CorpusID:244661765}  } | [A Hybrid Two-Level Support Vector Machine-Based Method for Churn Analysis](https://www.semanticscholar.org/paper/A-Hybrid-Two-Level-Support-Vector-Machine-Based-for-Sarac-Seker/dc7bebd0fd2059cd8ff0f50ae3d6ec206210fea7)  Customer churn is a central problem in almost every sector. Due to the diversity of the customers, products and services, and a massive amount of data being generated as a result of e-commerce tools and services, (big) data analytics and artificial intelligence-based methods have been developed and used for churn analysis in order to develop a strategy that is expected to understand the reasons behind the customer churn and subsequently to develop an effective and profitable customer retention programme. The analysis based on the data analytics and artificial intelligence methods focuses more on the profiling of customers, the classification of customer churn and identification of features that affect the churn. However, there doesn’t seem many studies that would be able to help understand how much a potential customer is likely to (or not likely to) pay for the products or services when churned or not, and to predict how much a particular customer or group of customers may have paid for the products or services. Therefore, in this study, a two-level churn analysis is proposed to (1) classify the customer churn or not, and (1) predict how much the customer has paid for the service. In order to achieve it, a machine learning method, namely support vector machine (SVM), was used for the classification part whereas a monthly service charge was predicted by using support vector regression (SVR) method. In order to select the most appropriate feature subset for both analyses, an unsupervised feature selection method, namely the multi-cluster feature selection method was utilized. The same feature selection method was used for both analyses for the sake of consistency to understand its performance over both analyses. The proposed hybrid approach was then applied in IBM’s Telcom data set with over 7000 customers in order to demonstrate the applicability and generalization ability of the proposed two-level approach. The SVMbased classification method has yielded AUC 85.6 and total classification accuracy of 81.5% being higher than those of a recent study where an aggressive set of the supervised classification methods was performed. The SVR-based prediction of the monthly charge has resulted in RMSE of 1.27, which is a reasonably acceptable outcome in the sector given the diversity of the ranges of charges as evidenced in its standard deviation. The approach presented in the study demonstrates that both the churn classification and charge prediction can be performed at the same time with a higher degree of accuracy. As the approach is open for further improvement, future analysis will be carried out to improve the accuracy for both analyses over other data sets to demonstrate its robustness and generalization ability. |
|  | @article{Elyusufi2022ChurnPA,  title={Churn Prediction Analysis by Combining Machine Learning Algorithms and Best Features Exploration},  author={Yasyn Elyusufi and M. Ait Kbir},  journal={International Journal of Advanced Computer Science and Applications},  year={2022},  url={https://api.semanticscholar.org/CorpusID:251407251}  } | [Churn Prediction Analysis by Combining Machine Learning Algorithms and Best Features Exploration](https://www.semanticscholar.org/paper/Churn-Prediction-Analysis-by-Combining-Machine-and-Elyusufi-Kbir/04a8a97e0663cc0f990562a1fd2537e0b71bf0e7)  The market competition and the high cost of acquiring new customers have led financial organizations to focus more and more on effective customer retention strategies. Although the banking and financial sectors have low churn rates compared to other sectors, the impact on profitability related to losing a customer is comparatively high. Thereby, customer turnover management and analysis play an essential part for financial organizations in order to improve their long-term profitability. Recently, it appears that using machine learning to predict churning improves customer retention strategies. In this work, we discuss some specific machine learning models proposed in the literature that deal with this problem and compare them with some emerging models, based on Ensemble learning algorithms. As a result, we build a predictive churn approaches that look at the customer history data, check to see who is active after a certain time and then create models that identify stages where a customer can leave the concerned company service. Ensemble learning algorithms are also used to find relevant features in order to reduce their number which is of great importance when performing the training step with some classical models such us Multi-Layer Perception Neural networks. The proposed approaches can achieve up to 89% in accuracy when other research works, dealing with the same dataset, can achieve less than 86%. |
|  | @article{Rudd2021CausalAO,  title={Causal Analysis of Customer Churn Using Deep Learning},  author={David Hason Rudd and Huan Huo and Guandong Xu},  journal={2021 International Conference on Digital Society and Intelligent Systems (DSInS)},  year={2021},  pages={319-324},  url={https://api.semanticscholar.org/CorpusID:245866432}  } | [Causal Analysis of Customer Churn Using Deep Learning](https://www.semanticscholar.org/paper/Causal-Analysis-of-Customer-Churn-Using-Deep-Rudd-Huo/e9af0d9933c1fb0fd6a84a372c8f651a6d4a7860)  Customer churn describes terminating a relationship with a business or reducing customer engagement over a specific period. Two main business marketing strategies play vital roles to increase market share dollar- value: gaining new and preserving existing customers. Customer acquisition cost can be five to six times that focus tomer retention, hence investing in customers with churn risk is smart. Causal analysis of the churn model can predict whether a customer will churn in the foreseeable future and assist enterprises to identify effects and possible causes for churn and subsequently use that knowledge to apply tailored incentives. This paper proposes a framework using a deep feedforward neural network for classification accompanied by a sequential pattern mining method on high-dimensional sparse data. We also propose a causal Bayesian network to predict cause probabilities that lead to customer churn. Evaluation metrics on test data confirm the XGBoost and our deep learning model outperformed previous techniques. Experimental analysis confirms that some independent causal variables representing the level of super guarantee contribution, account growth, and customer tenure were identified as confounding factors for customer churn with a high degree of belief. This paper provides a real-world customer churn analysis from current status inference to future directions in local superannuation funds.  Most previous studies focused on determining churn variables for a particular dataset rather than customer churn causation analysis. CHAMP (Churn analysis, modeling, and prediction) is an integrated system for forecasting consumers canceling their cellular phone service [3]. Alyuda Neuro-Intelligence employs neural networks (NNs) for data mining to forecast customer churn at banks [4]. Integrating textual data using customer churn prediction (CCP) algorithms adds value [5], and combining different classifiers, e.g., gradient boost, oversampling, and contrast sequential pattern mining on single year observation windows, has been shown to be a practical strategy to deal with highly skewed data collected from superannuation funds [6]. Different churn prediction techniques have been evaluated to identify optimal approaches [7]. Hidden churn is a common problem for superannuation funds, where customer accounts become dormant once mandatory employer payments cease. Various remedies for insufficient consumer interaction have been proposed, addressing unbalanced and fully leveraged data problems, and multiple classifiers have been developed from sampled datasets [6]. Deep learning techniques can handle very large datasets compared with standard machine learning (ML) approaches, and combining deep learning and convolutional neural networks (CNNs) have successfully forecast churn [8]. |

1. Giới thiệu

Viễn thông là lĩnh vực liên quan đến việc truyền thông và truy cập thông tin từ xa, thường sử dụng các phương tiện và công nghệ để truyền tải dữ liệu, âm thanh, hình ảnh và video giữa các kết nối điểm. Lĩnh vực này đã trở thành một phần quan trọng trong cuộc sống hàng ngày của chúng ta, từ việc liên lạc và truyền đạt thông tin đến công việc, giải trí và giáo dục. Dưới đây là một bài giới thiệu tổng quan về lĩnh vực Viễn thông: Giới thiệu về Viễn thông Viễn thông là lĩnh vực kỹ thuật và công nghệ liên quan đến truyền tải thông tin, âm thanh, hình ảnh và dữ liệu từ một địa điểm đến một hoặc nhiều địa điểm khác một cách hiệu quả và nhanh chóng. Với sự phát triển của công nghệ, Viễn thông không chỉ đơn thuần là công việc truyền tải giọng nói qua dây điện mà còn mở rộng ra các phương tiện truyền thông thông tin và giao tiếp trực tuyến

Phân tích nhân quả của mô hình rời bỏ có thể dự đoán liệu khách hàng có rời bỏ trong tương lai gần hay không và hỗ trợ doanh nghiệp xác định các tác động cũng như nguyên nhân có thể gây ra tình trạng rời bỏ, sau đó sử dụng kiến ​​thức đó để áp dụng các biện pháp khuyến khích phù hợp. Bài báo này đề xuất một khung sử dụng mạng nơ-ron tiếp liệu sâu để phân loại kèm theo phương pháp khai thác mẫu tuần tự trên dữ liệu thưa nhiều chiều. Chúng tôi cũng đề xuất mạng Bayesian nhân quả để dự đoán xác suất nguyên nhân dẫn đến việc khách hàng rời bỏ. Các số liệu đánh giá trên dữ liệu thử nghiệm xác nhận XGBoost và mô hình học sâu của chúng tôi hoạt động tốt hơn các kỹ thuật trước đó. Phân tích thử nghiệm xác nhận rằng một số biến nhân quả độc lập thể hiện mức độ đóng góp siêu bảo lãnh, tăng trưởng tài khoản và thời hạn sử dụng của khách hàng được xác định là các yếu tố gây nhiễu đối với việc rời bỏ khách hàng với mức độ tin tưởng cao. Bài viết này cung cấp phân tích tỷ lệ rời bỏ khách hàng trong thế giới thực từ suy luận về tình trạng hiện tại đến hướng đi trong tương lai của các quỹ hưu bổng địa phương.

Customer churn is a central problem in almost every sector. Due to the diversity of the customers, products and services, and a massive amount of data being generated as a result of e-commerce tools and services, (big) data analytics and artificial intelligence-based methods have been developed and used for churn analysis in order to develop a strategy that is expected to understand the reasons behind the customer churn and subsequently to develop an effective and profitable customer retention programme. The analysis based on the data analytics and artificial intelligence methods focuses more on the profiling of customers, the classification of customer churn and identification of features that affect the churn. However, there doesn’t seem many studies that would be able to help understand how much a potential customer is likely to (or not likely to) pay for the products or services when churned or not, and to predict how much a particular customer or group of customers may have paid for the products or services. Therefore, in this study, a two-level churn analysis is proposed to (1) classify the customer churn or not, and (1) predict how much the customer has paid for the service. In order to achieve it, a machine learning method, namely support vector machine (SVM), was used for the classification part whereas a monthly service charge was predicted by using support vector regression (SVR) method. In order to select the most appropriate feature subset for both analyses, an unsupervised feature selection method, namely the multi-cluster feature selection method was utilized. The same feature selection method was used for both analyses for the sake of consistency to understand its performance over both analyses. The proposed hybrid approach was then applied in IBM’s Telcom data set with over 7000 customers in order to demonstrate the applicability and generalization ability of the proposed two-level approach. The SVMbased classification method has yielded AUC 85.6 and total classification accuracy of 81.5% being higher than those of a recent study where an aggressive set of the supervised classification methods was performed. The SVR-based prediction of the monthly charge has resulted in RMSE of 1.27, which is a reasonably acceptable outcome in the sector given the diversity of the ranges of charges as evidenced in its standard deviation. The approach presented in the study demonstrates that both the churn classification and charge prediction can be performed at the same time with a higher degree of accuracy. As the approach is open for further improvement, future analysis will be carried out to improve the accuracy for both analyses over other data sets to demonstrate its robustness and generalization ability.

Sự rời bỏ khách hàng là một vấn đề trung tâm trong hầu hết mọi lĩnh vực. Do sự đa dạng của khách hàng, sản phẩm và dịch vụ cũng như lượng dữ liệu khổng lồ được tạo ra nhờ các công cụ và dịch vụ thương mại điện tử, phân tích dữ liệu (lớn) và các phương pháp dựa trên trí tuệ nhân tạo đã được phát triển và sử dụng để phân tích rời bỏ để phát triển một chiến lược dự kiến ​​​​sẽ hiểu được lý do đằng sau việc rời bỏ khách hàng và sau đó phát triển một chương trình giữ chân khách hàng hiệu quả và mang lại lợi nhuận. Phân tích dựa trên phương pháp phân tích dữ liệu và trí tuệ nhân tạo tập trung nhiều hơn vào việc lập hồ sơ khách hàng, phân loại khách hàng rời bỏ và xác định các tính năng ảnh hưởng đến tỷ lệ rời bỏ. Tuy nhiên, dường như không có nhiều nghiên cứu có thể giúp hiểu được một khách hàng tiềm năng có khả năng (hoặc không có khả năng) trả bao nhiêu cho các sản phẩm hoặc dịch vụ khi họ rời bỏ hay không, và dự đoán một khách hàng cụ thể hoặc bao nhiêu. nhóm khách hàng có thể đã trả tiền cho sản phẩm hoặc dịch vụ. Do đó, trong nghiên cứu này, phân tích rời bỏ hai cấp độ được đề xuất để (1) phân loại khách hàng có rời bỏ hay không và (1) dự đoán số tiền khách hàng đã trả cho dịch vụ. Để đạt được điều đó, một phương pháp học máy, cụ thể là máy vectơ hỗ trợ (SVM), đã được sử dụng cho phần phân loại trong khi phí dịch vụ hàng tháng được dự đoán bằng phương pháp hồi quy vectơ hỗ trợ (SVR). Để chọn tập hợp con tính năng phù hợp nhất cho cả hai phân tích, phương pháp lựa chọn tính năng không giám sát, cụ thể là phương pháp lựa chọn tính năng nhiều cụm đã được sử dụng. Phương pháp lựa chọn tính năng giống nhau đã được sử dụng cho cả hai phân tích nhằm mục đích thống nhất để hiểu hiệu suất của nó trên cả hai phân tích. Phương pháp kết hợp được đề xuất sau đó đã được áp dụng trong bộ dữ liệu Telcom của IBM với hơn 7000 khách hàng nhằm chứng minh khả năng ứng dụng và khả năng khái quát hóa của phương pháp tiếp cận hai cấp độ được đề xuất. Phương pháp phân loại dựa trên SVM đã mang lại AUC 85,6 và tổng độ chính xác phân loại là 81,5%, cao hơn so với nghiên cứu gần đây, nơi thực hiện một loạt các phương pháp phân loại có giám sát. Dự đoán dựa trên SVR về phí hàng tháng đã dẫn đến RMSE là 1,27, đây là một kết quả hợp lý có thể chấp nhận được trong lĩnh vực này do có sự đa dạng về phạm vi phí được chứng minh bằng độ lệch chuẩn của nó. Cách tiếp cận được trình bày trong nghiên cứu chứng minh rằng cả phân loại rời bỏ và dự đoán điện tích đều có thể được thực hiện cùng lúc với độ chính xác cao hơn. Khi cách tiếp cận này mở ra để cải tiến hơn nữa, phân tích trong tương lai sẽ được thực hiện để cải thiện độ chính xác của cả hai phân tích so với các tập dữ liệu khác nhằm chứng minh tính mạnh mẽ và khả năng khái quát hóa của nó.

Losing customers is inevitable for most businesses, but churn can be managed at acceptable levels by investing in customers with churn risk. A new churn propensity model was designed and integrated with the Bayesian causal network. Unbalanced churned and non-churned classes were leveled in pre-processing, and then ac- curacy was compared between the proposed DFF NN approach and ten current best-practice churn prediction classifiers. Although XGBoost achieved superior AUC, DFF NN obtained comparable AUC with the highest accuracy of all considered models on test data. We analyzed possible customer churn causes for a particular financial dataset created at superannuation fund(s) corporations. Causal analysis results confirmed variables representing recent SG contribution, annual report preference changed, account growth, and the balance amount was identified as confounding factors for customer churn with a high degree of belief. Churn rate can be reduced by ∼ 3% for customers with active accounts> 1year, consistent with expert knowledge. A natural progression of this work is extending pattern mining techniques with smaller outcome windows that should be investigated to obtain more efficient prediction results in future studies. Also, different methods to identify causes of churn based on counterfactual causal analysis should be investigated.

Mất khách hàng là điều không thể tránh khỏi đối với hầu hết các doanh nghiệp, nhưng việc rời bỏ khách hàng có thể được quản lý ở mức chấp nhận được bằng cách đầu tư vào những khách hàng có rủi ro rời bỏ. Một mô hình xu hướng rời bỏ mới đã được thiết kế và tích hợp với mạng lưới nhân quả Bayes. Các lớp bị xáo trộn và không bị xáo trộn không cân bằng đã được cân bằng trong quá trình xử lý trước và sau đó độ chính xác được so sánh giữa phương pháp DFF NN được đề xuất và mười phân loại dự đoán tình trạng rời bỏ theo phương pháp tốt nhất hiện nay. Mặc dù XGBoost đạt được AUC vượt trội, DFF NN thu được AUC tương đương với độ chính xác cao nhất trong tất cả các mô hình được xem xét trên dữ liệu thử nghiệm. Chúng tôi đã phân tích các nguyên nhân có thể khiến khách hàng rời bỏ đối với một tập dữ liệu tài chính cụ thể được tạo tại (các) công ty quỹ hưu bổng. Kết quả phân tích nguyên nhân đã xác nhận các biến đại diện cho đóng góp SG gần đây, thay đổi ưu tiên trong báo cáo hàng năm, tăng trưởng tài khoản và số dư được xác định là các yếu tố gây nhiễu đối với việc khách hàng rời bỏ với mức độ tin tưởng cao. Tỷ lệ rời bỏ có thể giảm ∼ 3% đối với khách hàng có tài khoản hoạt động > 1 năm, phù hợp với kiến ​​thức chuyên môn. Tiến triển tự nhiên của công việc này là mở rộng các kỹ thuật khai thác mẫu với cửa sổ kết quả nhỏ hơn cần được nghiên cứu để thu được kết quả dự đoán hiệu quả hơn trong các nghiên cứu trong tương lai. Ngoài ra, cần nghiên cứu các phương pháp khác nhau để xác định nguyên nhân gây ra tình trạng rời bỏ dựa trên phân tích nguyên nhân phản thực tế.

1. Nghiên cứu liên quan

We proposed a three-stage hybrid approach for churn prediction that combines cluster

analysis and decision tree analysis (Figure 1).

In the first stage, a database for churn prediction was developed. In the second stage, we performed k-means cluster analysis to detect market segments. Additionally, we used chi-square analysis to identify the clusters with the highest level of churn. In the third stage, we used the CHAID decision trees to generate models for predicting churn behavior for each cluster separately, focusing specifically on clusters with the highest churn rate. Rules were extracted that could be used for churn management. The extracted rules and cluster descriptions could be added to the customer database to increase its value and effectiveness.

Chúng tôi đã đề xuất một phương pháp kết hợp ba giai đoạn để dự đoán tỷ lệ rời bỏ kết hợp cụm phân tích và phân tích cây quyết định (Hình 1). Trong giai đoạn đầu tiên, cơ sở dữ liệu để dự đoán tỷ lệ rời bỏ đã được phát triển. Trong giai đoạn thứ hai, chúng tôi thực hiện phân tích cụm k-mean để phát hiện các phân khúc thị trường. Ngoài ra, chúng tôi đã sử dụng phân tích chi bình phương để xác định các cụm có mức độ rời bỏ cao nhất. Ở giai đoạn thứ ba, chúng tôi đã sử dụng cây quyết định CHAID để tạo mô hình dự đoán hành vi rời bỏ cho từng cụm riêng biệt, tập trung cụ thể vào các cụm có tỷ lệ rời bỏ cao nhất. Các quy tắc đã được trích xuất có thể được sử dụng để quản lý việc rời bỏ. Các quy tắc được trích xuất và mô tả cụm có thể được thêm vào cơ sở dữ liệu khách hàng để tăng giá trị và tính hiệu quả của nó.

Various data sources are used for churn prediction and analysis (Verbeke et al. 2012). The first group of data sources contains the data on the telecommunication transactions, e.g., number and duration of calls (e.g., Wei and Chiu 2002; Kisioglu and Topcu 2011). The second group of data sources contains the customers’ data, e.g., demographic character- istics, usage of additional services, contracts and billing, and monetary value and failure. We recommend using the second type of data for customer relationship management since it contains the relevant information for market segmentation (e.g., demographic characteristics).

Nhiều nguồn dữ liệu khác nhau được sử dụng để dự đoán và phân tích tỷ lệ rời bỏ (Verbeke và cộng sự 2012). Nhóm nguồn dữ liệu đầu tiên chứa dữ liệu về các giao dịch viễn thông, ví dụ: số lượng và thời lượng cuộc gọi (ví dụ: Wei và Chiu 2002; Kisioglu và Topcu 2011). Nhóm nguồn dữ liệu thứ hai chứa dữ liệu của khách hàng, ví dụ: đặc điểm nhân khẩu học, việc sử dụng các dịch vụ bổ sung, hợp đồng và thanh toán, giá trị tiền tệ và sự thất bại. Chúng tôi khuyên bạn nên sử dụng loại dữ liệu thứ hai để quản lý quan hệ khách hàng vì nó chứa thông tin liên quan đến phân khúc thị trường (ví dụ: đặc điểm nhân khẩu học).

We support our recommendation with the findings of Verbeke et al. (2012), who analyzed various datasets used in the churn prediction, noting that it is essential to consider if the data are the predictor or the symptom of the occurrence of churn. For example, at first sight, it seems that the attribute suggesting a significant decline in total minutes called may be substantially connected with churn. However, this decline is more likely to occur after the customer has already decided to leave the company—in other words, when a churn event has already occurred but has not yet been recorded in the data. More reliable data for churn prediction includes socio-demographic data, financial information and marketing-related variables.

Chúng tôi ủng hộ khuyến nghị của mình với những phát hiện của Verbeke et al. (2012), người đã phân tích các tập dữ liệu khác nhau được sử dụng trong dự đoán tỷ lệ rời bỏ, lưu ý rằng điều cần thiết là phải xem xét liệu dữ liệu đó là yếu tố dự đoán hay là triệu chứng của sự xuất hiện của tỷ lệ rời bỏ. Ví dụ: ngay từ cái nhìn đầu tiên, có vẻ như thuộc tính gợi ý sự sụt giảm đáng kể về tổng số phút được gọi có thể có mối liên hệ đáng kể với tình trạng rời bỏ. Tuy nhiên, sự suy giảm này có nhiều khả năng xảy ra sau khi khách hàng đã quyết định rời công ty—nói cách khác, khi một sự kiện rời bỏ đã xảy ra nhưng chưa được ghi lại trong dữ liệu. Dữ liệu đáng tin cậy hơn để dự đoán tỷ lệ rời bỏ bao gồm dữ liệu nhân khẩu học xã hội, thông tin tài chính và các biến liên quan đến tiếp thị.

Furthermore, a variable measuring churn should be included in the analysis. Cus-tomers terminate their relationship with the telecommunication company by ending the contract or through the cessation of the one-time payment for prepaid users.

Hơn nữa, một thước đo thay đổi nên được đưa vào phân tích. Khách hàng chấm dứt mối quan hệ với công ty viễn thông bằng cách chấm dứt hợp đồng hoặc thông qua việc ngừng thanh toán một lần đối với người dùng trả trước.

In the second stage, we applied a clustering procedure to identify the homogenous groups of customers from the telecommunication database. A non-hierarchical partitioning k-means clustering procedure was applied since most marketing research uses k-means for market segmentation in telecommunication. We used all the observed variables in the cluster analysis besides churn since clusters will be compared according to the churn ratio. Clustering requires a method for computing the distance or the (dis)similarity be- tween each pair of observations. In the clustering procedure, a distance measure is a function that quantifies the similarity between two observations. It determines how the similarity of two observations will be calculated, and it will influence the size of the clusters. Euclidean distance is a standard distance measure used in clustering, which measures the straight-line distance between observation xa and xb for all j characteristics (Boehmke and Greenwell 2020):

Nhiều thuật toán phân cụm phân vùng cố gắng giảm thiểu hàm mục tiêu khi tạo cụm. Trong k-mean, mục tiêu là giảm thiểu hàm sai số bình phương, biểu thị phương sai trong cụm. Nó thường được gọi là hàm biến dạng: