Pejić Bach, Mirjana; Pivar, Jasmina; Jaković, Božidar

Article

Churn management in telecommunications: Hybrid approach using cluster analysis and decision trees

Journal of Risk and Financial Management

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Pejić Bach, Mirjana; Pivar, Jasmina; Jaković, Božidar (2021) : Churn management in telecommunications: Hybrid approach using cluster analysis and decision trees, Journal of Risk and Financial Management, ISSN 1911-8074, MDPI, Basel, Vol. 14, Iss. 11, pp. 1-25,
https://doi.org/10.3390/jrfm14110544

This Version is available at:
https://hdl.handle.net/10419/258647

Journal of
Risk and Financial Management

MDPI

Article

Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees

Mirjana Pejić Bach *, Jasmina Pivar and Božidar Jaković

Faculty of Economics & Business, University of Zagreb, 10000 Zagreb, Croatia; jpivar@efzg.hr (J.P.); bjakovic@efzg.hr (B.J.)
* Correspondence: mpejic@efzg.hr

Abstract: 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.

Keywords: churn; telecommunications; clustering; k-means; market segmentation; prediction; decision trees; CHAID

Citation: Pejić Bach, Mirjana, Jasmina Pivar, and Božidar Jaković. 2021. Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees. *Journal of Risk and Financial Management* 14: 544. https://doi.org/10.3390/jrfm14110544

Academic Editors: Aleksandra Kuzior, Oleksii Lyulyov and Aleksy Kwilinski

Received: 24 August 2021
Accepted: 8 November 2021
Published: 11 November 2021

1. Introduction

Churn, also known as customer attrition, represents a situation when the customer stops buying products or using services from a company. In telecommunications, churn refers to a loss of customers when the customer leaves a telecommunication service provider and switches to another operator (Chouiekh and Haj 2020).

The telecommunication companies operate in a saturated market since telecommunication services have become widespread globally (Calzada-Infante et al. 2020). Companies compete by delivering marketing campaigns to acquire new customers and retain existing ones, considering that the costs of retaining the existing customer are usually much lower than attracting a new customer (Kim et al. 2020). Hence, customer churn prevention is crucial in the telecommunication industry (Droftina et al. 2015).

Churn management aims to minimize the churn using various retention strategies to prevent customers from cancelling subscriptions, such as offering new devices or services. For retention strategies to be successful, companies obtain insights into customers' characteristics and behavior to predict those likely to churn (Ahmad et al. 2019). Predicting which customers are about to leave allows the company to lower customer churn using specific marketing campaigns that could impact the customers to change their minds on leaving (Bell and Mgbemena 2017). Furthermore, the results of churn prediction are used to identify the leading causes of attrition (Cheng et al. 2019).

Predictive churn modelling is used to discover and analyze patterns in data so that past customer behavior can be used to forecast churn behavior. These analyses mostly use a data mining approach and focus on developing the forecasting model for predicting

churn. Most of these studies are developed using the database of all company customers, thus neglecting that not all customer segments have the same level of churn (Bayer 2010). A model that can develop a different churn prediction for different market segments is needed. It would allow companies to develop distinct marketing strategies for various market segments, such as promotions designed considering the churn causes in a specific market segment.

This paper contributes to the field of churn management in telecommunications with the development of the hybrid approach to churn management, which could support companies in the detection of market segments that are more likely to churn and target them with specific marketing campaigns. The three-stage hybrid approach that was developed combined cluster analysis and decision tree analysis. In the first stage, the customer database was prepared. In the second stage, a k-means cluster analysis was used to identify clusters with the highest level of churn. In the third stage, chi-squared automatic interaction detector (CHAID) decision tree analysis was used to determine the most significant characteristics of customers in the clusters with the highest churn levels. We applied the hybrid approach to the Telco Customer Churn open dataset.

The paper is organized as follows. After the introduction in the first section, the second section presents the overview of the literature focusing on market segmentation and churn prediction in telecommunications. In the third section, the data and procedures used for the data analysis are described. The fourth section is dedicated to interpreting the empirical results obtained using cluster analysis, variable selection and CHAID decision trees. The last section discusses the results, implications, limitations and future research directions.

2. Literature Review

2.1. Market Segmentation Using Cluster Analysis

Market segmentation is a process for dividing a given service or product into homogenous sub-groups of customers with similar characteristics (Dolnicar et al. 2018). Customers who belong to the same market segment are similar to one another concerning chosen characteristics such as demographic characteristics and purchasing behavior. Market segmentation helps companies detect niche markets and develop targeted marketing strategies (Dolnicar et al. 2018). Market segmentation is often used in customer-centric industries such as telecommunications (Hwang et al. 2004; Mu and Lee 2005).

Cluster analysis is commonly applied for market segmentation (Tuma et al. 2011). Cluster analysis groups a set of observations, such as customers, into homogenous groups, thus identifying the customers who are similar to one another within the same cluster and showing how they are different compared to customers in other clusters. When performing cluster analysis, a researcher can choose from various methods, such as hierarchical and non-hierarchical clustering (Marini and Amigo 2020). In hierarchical clustering, clusters are derived using the top-down (divisive) or bottom-up (agglomerative) approach, depending on the fashion in which the hierarchical decomposition is formed. Non-hierarchical clustering groups observations to minimize the evaluation criteria, such as the error function (Jin and Han 2011). In non-hierarchical clustering, a user needs to determine the number of clusters in advance.

Partitional clustering methods are a type of non-hierarchical clustering method. In partitional clustering, the dataset is decomposed into a set of disjoint clusters. Given a dataset of N observations, a partitioning method constructs K partitions of the data. A partition represents a cluster. In other words, by using partitional clustering, data are classified into K groups. Each group contains at least one observation, and each observation belongs to exactly one group. An example of non-hierarchical partitional clustering is k-means clustering.

To perform market segmentation in telecommunications, a variety of cluster approaches have been used in previous research. Zhou et al. (2020) integrated the recency, frequency and monetary approach with the sparse k-means clustering algorithm to conduct segmentation of the Chinese mobile telecommunications market based on extensive

consumer data. Khalili-Damghani et al. (2018) used a hybrid soft computing approach based on clustering, rule mining, and decision tree analysis. Bose and Chen (2015) used fuzzy c-means clustering to build customer profiles and study mobile services customers' migratory behavior. They discovered usage and revenue patterns for two migratory groups of customers. Wang (2018) analyzed different groups of omnichannel buyers in telecommunications. The author used k-means particle swarm optimization for cluster analysis and the C5.0 classification method to formulate classification rules. Swarm optimization was also used by Li and Marikannan (2019). Qiu et al. (2020) used the k-means-HF method for silent customer segmentation, to identify a segment of customers that a company is likely to lose. The authors concluded it is necessary to analyze such customers' features and make appropriate market decisions to improve the telecommunication industry's revenue. Zheng and Liu (2020) used a hierarchical locality sensitive hashing-local outlier factor scheme for anomalous customer behavior detection and k-means clustering analysis on the real telecommunication operating data provided by one of China's major Internet service providers. Lin et al. (2019) used a parallel large sum submatrix bi-clustering algorithm based on Spark MapReduce to identify and segment highly profitable telecommunication customers who share similar upscale purchasing behavior on a small fraction of attributes. Golubev et al. (2017) used clustering techniques to determine the clusters with typical user behavior based on call detail records data. Al-Refaie (2017) examined the factors affecting customer churn in the Jordanian telecommunication industry using cluster analysis.

Most of the researchers used cluster analysis as the sole method for determining market segments in telecommunications. Only a few authors investigated the combination of various methods, such as cluster analysis and decision trees, for market segmentation in telecommunications (Khalili-Damghani et al. 2018; Wang 2018). K-means is the most often used clustering algorithm for market segmentation.

2.2. Predicting Churn in Telecommunications

Various approaches have been used for churn prediction in telecommunication. The research could be divided into two groups. Appendix A provides a summary of the bellow-mentioned churn modelling approaches in the telecommunication industry.

The first group includes research comparing the performance of algorithms applied on the various datasets to develop the forecasting model with the highest accuracy of predicting churn using one of the competing algorithms. This group of research claims that its modelling approach is successful compared with other state-of-the-art classifiers. A comparison study performed by Pamina et al. (2019) shows that the XG boost classifier performs better than K-NN and random forest in improving the accuracy of customer churn prediction. The comparison study was performed using a publicly available telecommunication dataset. Furthermore, the research shows that fiber optic customers with more significant monthly charges have a more substantial influence on churn. Ahmad et al. (2019) experimented with the decision tree, random forest, gradient boosted machine tree "GBM", and extreme gradient boosting (XgBOOST) to predict churn based on big raw data provided by SyraTel telecommunication company. The best results were obtained by applying the XgBOOST algorithm, which was then used for classification in the churn predictive model. The same methodology was used by Swetha and Dayananda (2020). Manďák and Hančlová (2019) performed logistic regression analysis on demographic and service usage variables to predict customer churn in European telecommunications providers. Ahmed and Maheswari (2017) presented metaheuristic-based churn prediction techniques applied to a massive Orange telecommunication dataset. A hybridized form of the Firefly algorithm was used as the classifier. Ahmed et al. (2020) developed the churn prediction model using hybrid firefly-based classification. AlOmari and Hassan (2016) tested the capability of the RULES Family algorithm-6 prediction data-mining technique to predict telecommunications customers' churn. Höppner et al. (2020) applied the ProfTree decision tree to construct churn prediction models based on real-life datasets from various telecommunications providers. Faris (2018) proposed a hybrid model based

on particle swarm optimization and feedforward neural networks for churn prediction. Sjarif et al. (2019) used the Pearson correlation and k-nearest neighbor (KNN) algorithm for churn prediction, using the public Telco Customer Churn dataset available on the Kaggle platform. Azeem et al. (2017) compared neural networks, linear regression, C4.5, SVM, AdaBoost, and gradient boosting, random, and fuzzy classifiers for churn prediction. Almufadi et al. (2019) used convolutional neural networks (CNN) for the churn prediction of mobile telecom subscribers. Li and Marikannan (2019) used particle swarm optimization and an extreme learning machine to predict churn in telecommunication.

The second group includes research that used a hybrid approach combining several methods, claiming that the sequential usage of various methods yields the best results in churn prediction. Ullah et al. (2019) used a combined approach for churn prediction in the telecommunication sector. They used the random forest algorithm to classify the churn and non-churn customers and identify factors used in the k-means clustering. The features that were used were related to calls duration, free calls, charges and others. The authors did not provide a discussion of their results in terms of marketing strategy. Choudhari and Potey (2018) performed predictive analysis for customer churn in the telecommunication industry using a hybrid decision tree and logistic regression classifier. The authors also proposed a hybrid Fuzzy unordered rule induction algorithm with fuzzy c-means clustering for the prediction of customer churn. Olle and Cai (2014) performed customer churn analysis with a combined model using logistic regression for classification and the Voted Perceptron for churn probability estimation. Preetha and Rayapeddi (2018) used logistic regression, random forests and k-means clustering to predict customer churn in the telecommunication industry.

Based on the analysis of the data mining approaches for predicting churn in telecommunications, it can be concluded that the researchers focused primarily on the improvement of churn prediction, while they rarely focused on the development of different retention strategies for various market segments.

3. Methodology

We proposed a three-stage hybrid approach for churn prediction that combines cluster analysis and decision tree analysis (Figure 1).

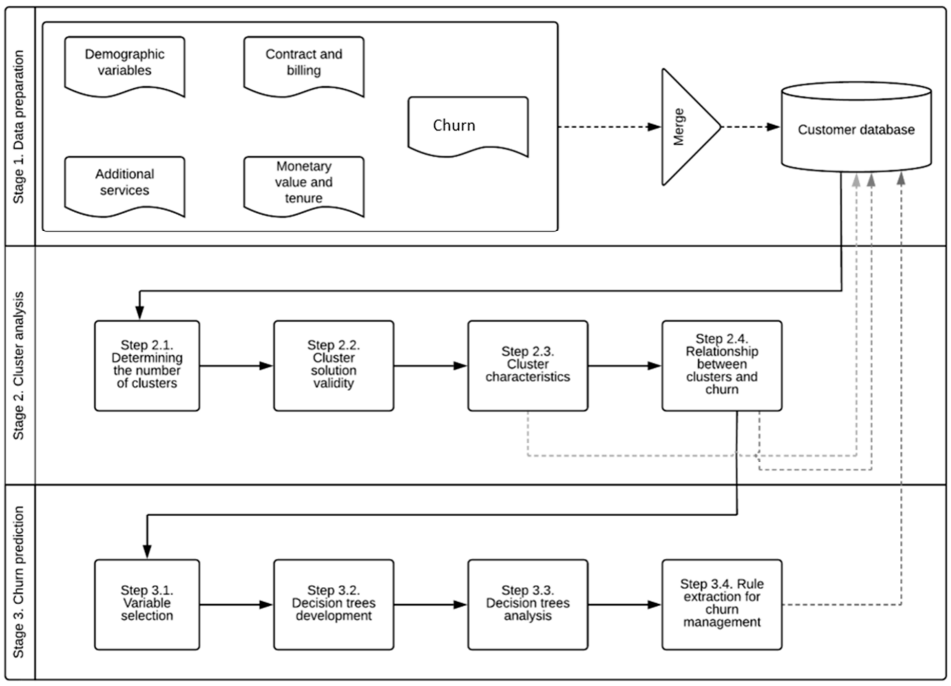

Figure 1. Hybrid methodology for churn analysis. Source: authors' work using Statistica Software.

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.

3.1. Stage 1. Data Preparation

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 characteristics, 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).

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.

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

3.2. Stage 2. Cluster Analysis

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 between 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 x_a and x_b for all j characteristics (Boehmke and Greenwell 2020):

$$\sqrt{\sum_{j=1}^{P}\left(x_{aj}-x_{bj}\right)^2} \tag{1}$$

Many partitional clustering algorithms attempt to minimize an objective function when making clusters. In k-means, the objective is to minimize squared error function, which represents intra-cluster variance. It is often called the distortion function:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}||x_i^{(j)}-c_j||^2 \tag{2}$$

in which k represents the number of clusters, n is the number of observations, i stands for observation, and c_j is a centroid for cluster j.

Step 2.1. Determining the number of clusters

Due to the vagueness of the criteria used for selecting the correct numbers of clusters and initial variables, cluster analysis is often criticized due to the possibility of deriving random solutions (Ernst and Dolnicar 2018; Vazirgiannis 2009). However, in a data mining approach, the v-fold methodology allows the automatic selection of the number of clusters (Kodinariya and Makwana 2013; Kassambara 2017). We determined the number of clusters using the v-fold cross-validation approach and by observing the cost sequence graph that showed the clustering solution's error functions for different clusters.

Step 2.2. Cluster solution validity

After finding the optimal cluster solution, the derived clusters were observed and interpreted. The analysis of variance (ANOVA) was applied to check whether numeric variables were statistically significant across clusters. Furthermore, chi-square analysis for nominal variables was performed to detect statistically significant differences between clusters.

Step 2.3. Cluster characteristics

Ultimately, the clusters were described according to the characteristics of the customers classified in a particular cluster. The distribution of variables across clusters was observed.

Step 2.4. Relationship between clusters and churn

Chi-square analysis was performed to estimate whether there was a significant difference in churn occurrence across clusters. This procedure identified the clusters in which the highest proportion of customers were likely to churn.

3.3. Stage 3. Churn Prediction

In the second stage, churn prediction was performed for the clusters that demonstrated the highest churn rate. First, the feature selection was conducted. Second, CHAID decision tree modelling was applied to the clusters with the highest churn rate using the variables selected by the feature selection and variable screening analysis.

Step 3.1. Variable selection

In this step, the variables that were most likely to be good predictors of churn were selected. In the first step, churn prediction variables were selected using chi-square statistics as the variable importance measures. This type of analysis does not assume any particular type or shape of the relationship between the predictors and the dependent variables of interest. Instead, it applies a generalized "notion of relationship" while screening the predictors, one by one, for regression or classification problems (TIBCO 2020). For continuous predictors, each predictor's range of values is divided into ten intervals by default to "fine-tune" the algorithm's sensitivity to different types of monotone and non-monotone relationships. Since our research's dependent variable of interest was categorical, a chi-square statistic for each predictor variable was calculated. Those variables with chi-square values larger than ten were selected for the decision tree analysis using a rule of thumb.

Step 3.2. Decision trees' development

The CHAID algorithm was used to build decision trees to determine how predictor variables best explain churn behavior.

The CHAID decision tree is based on the chi-square test, which is used to find the relationship between variables. It finds the most significant variable and selects the best split at each step. Each pair of predictor categories are assessed to determine what is least significantly different concerning the dependent variable. Due to these merging steps, a Bonferroni adjusted p-value is calculated for the merged cross-tabulation. In the CHAID decision tree, the root node contains the dependent variable, which is split into two or more categories, called initial or parent nodes. These categories have a significant influence

on the dependent variable. The root node is split into child nodes. The child nodes that are not further split are called the terminal nodes.

In the proposed approach, churn was the dependent variable, and the variables selected in the previous step were predictors. First, decision trees were developed for the whole database, and then separately for each cluster identified in the second stage. Second, decision trees were compared according to their accuracy. Only those decision trees with the highest ratio of churn and the best churn predictions were retained in further steps.

Step 3.3. Decision tree analysis

In the third step, decision trees developed for the clusters that contain the most significant churn ratio were analyzed to determine the customers' characteristics according to their churn behavior.

Step 3.4. Rule extraction for churn management

In the final step, the rules were extracted for churn management that could be useful for developing a relevant marketing strategy to retain the customers likely to churn.

4. Results

4.1. Stage 1. Data Preparation

The dataset used to demonstrate the proposed hybrid approach combining cluster analysis and decision trees was the Telco Customer Churn open dataset, which is available on the Kaggle platform—https://www.kaggle.com/blastchar/telco-customer-churn (accessed on 9 November 2021). This dataset was selected since it contains heterogeneous variables that reflect customer demographic characteristics, usage of services and billing behavior.

Table 1 shows the variables used in the analysis. We divided the data into five groups: demographic variables, contracts and billing, additional services used, customer monetary value and tenure, and churn behavior. The prepared dataset includes 7032 customers with 20 variables.

The first group of variables contains demographic variables, and they are all binomial. The Gender variable is represented by two modalities, Female and Male, while the SeniorCitizen, Dependents and Partners variables are represented by two modalities, No and Yes.

The variables that describe the features of contracts and billing are Contract, PaperlessBilling and PaymentMethod. The Contract variable is related to the contract term of the customer. It is of a nominal type, and it has three modalities: month-to-month, one year and two years. The PaperlessBilling variable describes whether the customer has paperless billing. The PaymentMethod variable is a nominal-type variable with four modalities that show a customer's payment method: Bank transfer, Credit card, Electronic check or mail.

The third group of variables describes additional services that the customers use. The InternetService variable is a nominal variable with three modalities, DSL, Fiber Optic and No. The DeviceProtection, OnlineBackup, OnlineSecurity, StreamingMovies, StreamingTV and TechSupport variables are all nominal-type variables with three possible modalities: No stands for not having contracted the service, No Internet service represents cases in which a customer does not use Internet service, and Yes refers to cases when customers use some of these services. The MultipleLines variable has three modalities: No describes cases when a customer does not use multiple phone lines, No phone service refers to instances in which a customer does not use phone services at all, and Yes is used when the customer has multiple phone lines.

The last group of variables describes customer monetary value and tenure. The Tenure variable is a numeric variable representing customer lifespan in months—the number of months the customer stayed with the company. The variables MonthlyCharges and TotalCharges are numeric numbers containing data on the amount charged to the customer, either monthly or in total (in USD).

The Churn variable is the dependent variable that is binomial and takes on two values, No and Yes, referring to customers who did not churn and customers who did churn, respectively.

Table 1. Variables used in the analysis.

Variable Name	Variable Type	Modalities/Min–Max
Demographic variables		
Gender	Binomial	Female; Male
SeniorCitizen	Binomial	No; Yes
Dependents	Binomial	No; Yes
Partner	Binomial	No; Yes
Contracts and billing		
Contract	Nominal	Month-to-month; One year; Two year
PaperlessBilling	Binomial	No; Yes
PaymentMethod	Nominal	Bank transfer; Credit card; Other
Additional services used		
InternetService	Nominal	DLS; Fiber Optic; No
DeviceProtection	Nominal	No; No Internet service; Yes
MultipleLines	Nominal	No; No phone service; Yes
OnlineBackup	Nominal	No; No Internet service; Yes
OnlineSecurity	Nominal	No; No Internet service; Yes
StreamingMovies	Nominal	No; No Internet service; Yes
StreamingTV	Nominal	No; No Internet service; Yes
TechSupport	Nominal	No; No Internet service; Yes
PhoneService	Binomial	No; Yes
Customer monetary value and tenure		
Tenure	Numeric (Months)	[1; 72]
TotalCharges	Numeric (USD)	[18,80; 118,75]
MonthlyCharges	Numeric (USD)	[18,25; 8684,80]
Churn behavior		
Churn	Binomial	No; Yes

Source: authors' work.

4.2. Stage 2. Cluster Analysis

The k-means clustering algorithm was applied to group the customers according to the 19 observed variables' values. All of the variables were included except for churn. Customers were assigned to a particular cluster based on the Euclidian distances as a distance measure. The maximum initial distance approach was used to estimate initial centroids (Lee and Han 2012). A 10-fold cross-validation approach was applied to find the optimal clustering solution or the number of clusters with the lowest estimated error rate (intra-cluster variance), as successfully applied by the previous research (Thomassey and Fiordaliso 2006). The Statistica (Version 13.05) software was used for the cluster analysis.

Step 2.1. Determining the number of clusters

The optimal number of clusters was determined iteratively by utilizing 10-fold cross-validation. The cost sequence graph shows the error function for different numbers of clusters. As seen in the cost sequence graph (Figure 2), the best number of clusters was six since the error function decreased up to the cluster solution with six clusters.

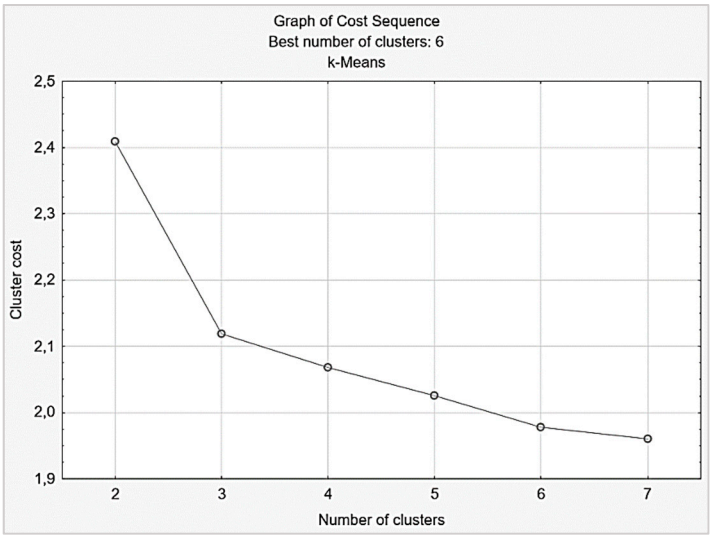

Figure 2. Graph of cost sequence. Source: authors' work using Statistica Software.

Step 2.2. Cluster solution validity

In Table 2, the results of the ANOVA of the numeric variables used in the cluster analysis are shown for the solution with six clusters. All results of the ANOVA suggest that the null hypothesis, where it is stated that the means between the analyzed variables are equal, can be rejected.

Table 2. ANOVA analysis for numeric variables.

	between SS	df	within SS	df	F	*p*-Value
Monetary value and tenure						
Tenure	1,665,336	5	2,570,629	7026	910.334	0.000 *
MonthlyCharges	5,141,226	5	1,222,995	7026	5907.181	0.000 *
TotalCharges	22,700,920,000	5	13,426,130,000	7026	2375.914	0.000 *

Source: authors' work; * statistically significant correlations at the 1% significance level.

Table 3 shows the values of chi-square statistics for the testing of differences between clusters according to nominal variables for the solution with six clusters. The values of the chi-squared test statistics for all variables and the associated *p*-values indicate that the null hypothesis for each of the variables can be rejected. In other words, a conclusion can be made that differences between the clusters exist for all nominal variables.

Table 3. Chi-square analysis for nominal variables.

Variable Name	Variable Type	df	Chi-Square	*p*-Value	G-Square	*p*-Value
Demographic variables						
Gender	Binomial	5	654.813	0.000 *	673.119	0.000 *
SeniorCitizen	Binomial	5	388.854	0.000 *	439.816	0.000 *
Dependents	Binomial	5	786.502	0.000 *	831.807	0.000 *
Partner	Binomial	5	1225.976	0.000 *	1285.330	0.000 *
Contracts and billing						
Contract	Nominal	10	3354.462	0.000 *	3680.463	0.000 *
PaperlessBilling	Nominal	5	1126.370	0.000 *	1137.811	0.000 *
PaymentMethod	Nominal	15	2171.842	0.000 *	2197.449	0.000 *

Table 3. *Cont.*

Variable Name	Variable Type	df	Chi-Square	*p*-Value	G-Square	*p*-Value
Additional services						
InternetService	Nominal	10	9689.710	0.000 *	9759.398	0.000 *
DeviceProtection	Nominal	10	8715.597	0.000 *	8737.442	0.000 *
MultipleLines	Nominal	10	4229.285	0.000 *	3740.198	0.000 *
OnlineBackup	Nominal	10	8411.848	0.000 *	8459.180	0.000 *
OnlineSecurity	Nominal	10	8699.201	0.000 *	8663.085	0.000 *
StreamingMovies	Nominal	10	8852.166	0.000 *	8854.783	0.000 *
StreamingTV	Nominal	10	8840.241	0.000 *	8841.567	0.000 *
TechSupport	Nominal	10	8863.749	0.000 *	8808.242	0.000 *
PhoneService	Nominal	5	2215.395	0.000 *	1620.087	0.000 *

Source: authors' work; * statistically significant correlations at the 1% significance level.

Both the ANOVA and the chi-square analysis support the decision to use six clusters in the cluster analysis.

Step 2.3. Cluster characteristics

Table 4 presents the cluster characteristics. The largest cluster contained 1520 observations, whereas the smallest cluster consisted of 516 observations.

Table 4. Cluster values.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Demographic variables						
Gender	Male	Female	Male	Male	Male	Female
SeniorCitizen	No	No	No	No	No	No
Partner	No	No	No	Yes	No	Yes
Dependents	No	No	No	No	No	Yes
Contracts and billing						
Contract	One year	Month-to-month	Month-to-month	Two year	Two year	Month-to-month
PaperlessBilling	No	Yes	Yes	Yes	No	Yes
PaymentMethod	Credit card	Electronic check	Electronic check	Bank transfer	Mailed check	Electronic check
Additional services						
DeviceProtection	No	No	No	Yes	No internet service	No
InternetService	DSL	DSL	Fiber optic	Fiber optic	No	Fiber optic
MultipleLines	No phone service	No	Yes	Yes	No	No
OnlineBackup	No	No	No	Yes	No internet service	Yes
OnlineSecurity	Yes	No	No	Yes	No internet service	No
StreamingMovies	No	No	No	Yes	No internet service	Yes
StreamingTV	No	No	No	Yes	No internet service	Yes
TechSupport	Yes	No	No	Yes	No internet service	No
PhoneService	No	Yes	Yes	Yes	Yes	Yes
Customer monetary value and tenure						
Tenure	37.85	14.28	21.55	59.47	30.67	36.28
MonthlyCharges	49.85	57.68	82.95	93.07	21.08	89.19
TotalCharges	1970.03	836.81	1833.46	5552.52	665.22	3223.34
Cluster members						
Number of cases	516	1410	1378	1376	1520	832
Percentage (%)	7.34	20.05	19.60	19.57	21.62	11.83

Source: authors' work, 2021.

Demographic variables. The most common Gender of customers was male for Cluster 1, Cluster 3, Cluster 4 and Cluster 5, and female for Cluster 2 and Cluster 6. Non-senior citizens were the most common in all of the clusters. For Clusters 1, 2, 3 and 5, the customers without partners were in the majority, while the customers in Cluster 4 and 6 mostly had partners. In Cluster 6, most customers had children or partners they financially supported, while the customers in all other clusters mostly did not have dependents.

Contracts and billing. The month-to-month type of contract was the most common for Cluster 2, Cluster 3 and Cluster 6. One-year contracts prevailed in Cluster 1 while the Two-year contract constituted the majority for Cluster 4 and 5. The most common payment method was electronic check for Cluster 2, Cluster 3 and Cluster 6. Credit card as a payment method was found to be related explicitly to Cluster 1, bank transfer was identified as the most common payment method for Cluster 4, and mailed check was related explicitly to Cluster 5.

Additional services. Most customers from Cluster 4 used device protection. In other clusters, that was not the case; in other words, customers from Cluster 1, 2, 3 and 6 did not use device protection. The type of internet service that prevailed in Cluster 1 and 2 was DSL internet, and fiber optic was most common in Clusters 3, 4 and 6. Cluster 5 is unique insofar as most of the customers did not use internet service at all. When it comes to using multiple phone lines, it is noticeable that in Cluster 1, customers usually did not use phone service at all, Clusters 2, 5 and 6 contained customers who usually did not use multiple lines, and Clusters 3 and 4 contained those who used multiple lines. Usage of online backups was common for Cluster 4 and Cluster 6.

Members of Cluster 5 mostly did not use internet service. Therefore, they usually did not use additional online backups, online security, movie streaming, TV streaming, or technical support. However, they did use phone service.

Usage of online backups, movie streaming, and TV streaming was common in Cluster 4 and Cluster 6. However, those clusters differed in terms of technical support. In Cluster 4, most customers used technical support, while in Cluster 6, the opposite was the case. When it comes to other clusters, it is noticeable that online backup and streaming services were not used by the majority of customers in Clusters 1, 2 and 4. However, in Cluster 1, the majority of customers did use online security and technical support. Additionally, in Cluster 1, customers usually did not use phone service.

Customer monetary value and tenure. The length of a customer's stay with the telecommunication service provider or tenure was measured in months. First, the tenure was the longest on average for customers in Cluster 4 (59.47 months). Clusters 1 and 7 were similar in terms of the tenure of their customers. Since Cluster 2 contained the customers with the shortest average tenure (14.28 months), telecommunication companies should pay more attention to them and possibly offer them additional services and one- or two-year contracts. New customers were found to be particularly vulnerable; therefore, if their experiences are not satisfactory, the relationship is likely to be short. Customers who were satisfied with the service provider and had high cumulative satisfaction tended to stay for longer durations.

The highest average monthly charges were in Cluster 4, related to the number of additional services. The majority of customers in that cluster also used many additional services. Interestingly, relatively high average monthly charges were related to Cluster 6. Those customers usually had month-to-month contracts and did not use some of the additional services. Telecommunication companies should consider additional offers or discuss signing long term contracts to ensure that those customers stay. Customers from Cluster 3 had relatively high average monthly charges, even though the majority of them did not use additional services. Telecommunications providers should take some precautions regarding those customers because they may become unsatisfied by high charges.

Total charges were highest on average in Cluster 4, which was expected due to the contract's duration and monthly charges. On average, Cluster 5 had the lowest total charges due to the low monthly charges and the usage of only essential services such as phone service and one phone line.

Step 2.4. Relationship between clusters and churn

Table 5 presents results of chi-square analysis. They suggest that clusters were determined to be statistically different according to churn occurrence across the clusters at the 1% significance level (Pearson chi-square 1080.666; *p*-value 0.000).

Table 5. Relationship between clusters and churn; chi-square analysis.

	No Churn	Churn	Total	Pearson Chi-Square	df	*p*-Value
Cluster 1	461	55	516	1080.666	5	0.000 *
Cluster 2	858	552	1410			
Cluster 3	670	708	1378			
Cluster 4	1216	160	1376			
Cluster 5	1407	113	1520			
Cluster 6	551	281	832			
Total	5163	1869	7032			

Source: authors' work; * statistically significant at the 1% significance level.

In total, almost a third of the customers (26.58%) were found to be churners. The highest absolute and the relative number of churn cases was identified in Cluster 3 (708 cases, 51.38%). Cluster 2 was the second for churn occurrence with 552 cases and 39.15% of a total of 1410 customers in that cluster. Cluster 5, which was the largest (1520 customers), had the lowest occurrence of churn—113 customers or 7.43% of customers in that cluster. Cluster 6 was third according to churn occurrence (281 or 33.77% of a total of 832 customers in that cluster). In all of these clusters, the majority of customers had a month-to-month contract.

It can be concluded that the Cluster 2 and Cluster 3 customers were those with the highest churn occurrence (Figure 3). These clusters were selected to be used in further analyses, with decision tree analysis being utilized for churn prediction.

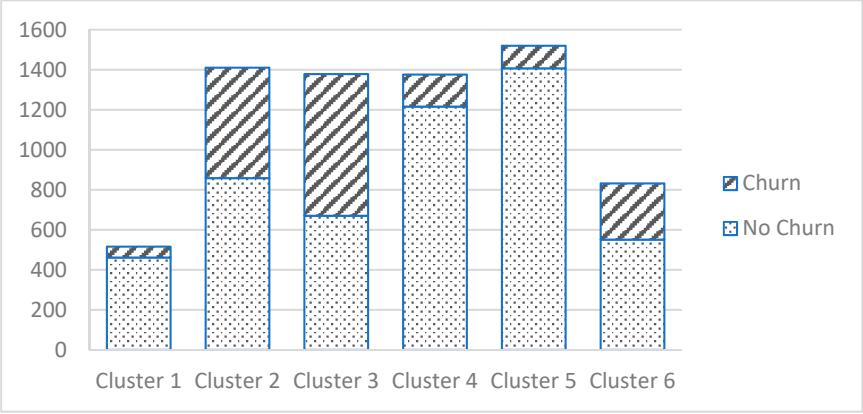

Figure 3. Structures of clusters according to churn. Source: authors' work.

4.3. Stage 3. Churn Prediction

In the third stage, the variables were selected using the Feature Selection and Variable Screening features delivered by the Statistica software (Version 13.05). Decision trees were developed for the clusters with the highest churn rate using the SPSS software (Version 22).

Step 3.1. Variable selection

In this step, the variables were screened to select the variables used for the churn prediction step. Table 6 contains the Feature Selection and Variable Screening results.

Table 6. Variable importance.

Variable	Chi-Square	*p*-Value
Contract	1179.546	0.000 *
Tenure	873.717	0.000 *
OnlineSecurity	846.677	0.000 *
TechSupport	824.926	0.000 *
InternetService	728.696	0.000 *
PaymentMethod	645.430	0.000 *
OnlineBackup	599.175	0.000 *
MonthlyCharges	563.636	0.000 *
DeviceProtection	555.880	0.000 *
TotalCharges	387.330	0.000 *
StreamingMovies	374.268	0.000 *
StreamingTV	372.457	0.000 *
PaperlessBilling	257.756	0.000 *
Dependents	187.128	0.000 *
SeniorCitizen	159.364	0.000 *
Partner	158.182	0.000 *
MultipleLines	11.272	0.004 *
PhoneService	0.961	0.327
Gender	0.513	0.474

Source: authors' work, 2021; * statistically significant at 1% level.

The chi-square and *p*-values indicated that the most important predictor of churn was the Contract variable (chi-square 1179.543; *p*-value 0.000), followed by Tenure (chi-square 873.717; *p*-value 0.00). The least important predictors of churn were the Gender and PhoneService variables. Therefore, they were not selected to be used in further analyses. All of the other 17 variables were selected to construct the CHAID decision trees.

Step 3.2. Decision trees' development

CHAID decision trees were generated at the level of the entire dataset, as well as separate clusters. Independent variables were selected in the previous step. The dependent variable was the churn binary variable.

Table 7 presents the percentages of correct classifications for the decision trees. The decision tree of Cluster 3 was identified as the most successful in predicting churners as compared to the decision trees of other clusters. It correctly classified 81.4% of all churners. The decision tree for Cluster 2 was found to be also successful in its prediction, with 70.7% of churners correctly classified.

Table 7. Percentage of correct churn classifications for CHAID decision trees.

Correct Predictions	Full Database	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Churn—No	90.4%	100.0%	71.7%	55.4%	97.2%	100.0%	97.3%
Churn—Yes	49.5%	0.0%	70.7%	81.4%	23.1%	0.0%	18.9%
Overall	79.5%	89.3%	71.3%	68.7%	88.6%	92.6%	70.8%
Rank Yes	(3)	(6)	(2)	(1)	(4)	(6)	(5)

Source: authors' work, 2021.

Since Cluster 2 and Cluster 3 had the highest prediction accuracy for churn and were in the same time clusters with the highest ratio of churn customers, it is recommended that the marketing department conduct the churn analysis solely for these groups of customers.

Step 3.3. Decision tree analysis

Figure 4 presents the decision tree developed for the customers in Cluster 2 with three levels and 23 nodes, of which 13 were considered terminal (leaf nodes). Figure 3 reveals that the Tenure, Internet service, Contract, Multiple Lines, Monthly Charges, Paperless Billing and TechSupport variables were statistically significant and used the classification tree provided by the CHAID algorithm.

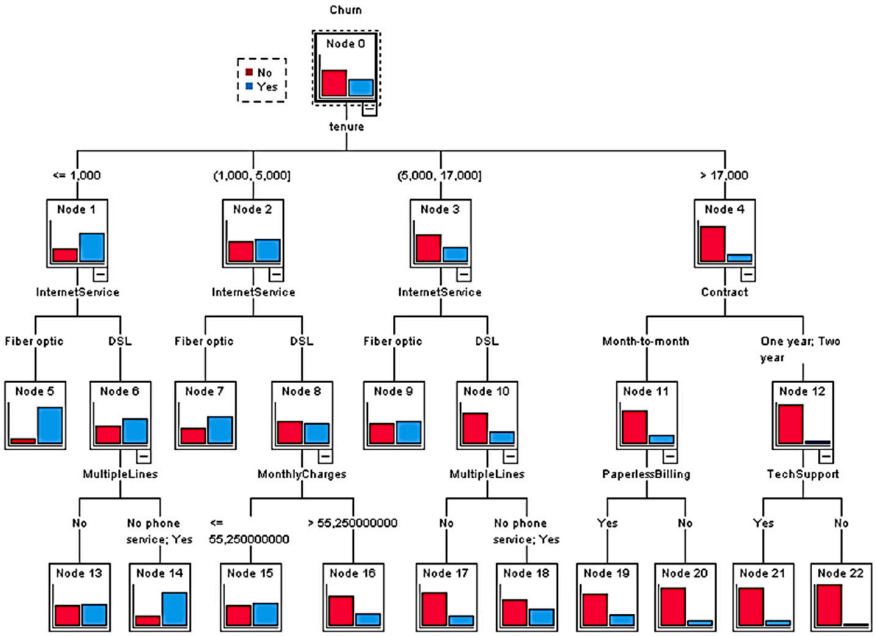

Figure 4. CHAID decision trees for Cluster 2. Source: authors' work, 2021.

The variable that was used for branching on the first level was the Tenure variable. This branching resulted in four new nodes (Node 1, Node 2, Node 3 and Node 4). Node 1 included category $\leq$1.000 and consisted of customers who were mostly suspected of churning. Node 2 included customers who stayed with the company for between one and five months. It had slightly more customers suspected of churning than not suspected of churning. Node 3 included customers who had been with the company for between five and seventeen months. Node 4 included those who had been with the company for more than seventeen months. In both of them, the share of suspected churners was found to be greater than the share of non-suspected churners.

The variable Internet service was used for branching Node 1, Node 2 and Node 3 on the second level, while the variable contract was used for branching Node 4. Node 5 and Node 6 were derived from Node 1. Node 5 was related to the category of customers who used fiber optic. It is noticeable that majority of them were connected with suspected churn. Regarding Node 6, which was related to the category of customers who used DSL, it was shown that there was a more significant share of customers suspected of churning than the share not suspected of churning. However, the ratio between suspected churners and non-churners was much higher for fiber optic users. Similar results were shown for Node 7 and Node 8, which were derived from Node 2. Node 7 had a high share of suspected churners. Node 8 contained customers who used DSL, and it also contained a higher share of suspected non-churners than suspected churners. Node 3 branched to Node 9 and Node 10. For Node 9, which contained fiber optic customers, there was a slightly more significant share of suspected churners than non-churners. Node 10 showed a significantly larger share of suspected non-churners compared to churners. Node 4 had

two branches, Node 11 and Node 12. Node 11 contained customers with a month-to-month contract, and Node 12 contained customers with One-year and Two-year contracts. In both of the nodes, there was a larger share of suspected non-churners compared to suspected churners. However, the share of non-churners for Node 11 was more significant compared to Node 12.

The third level branching variables were MultipleLines, MonthlyCharges, PaperlessBilling and TechSupport. The MultipleLines variable was used to further branch Node 6 into two nodes (Node 13 and Node 14). Node 13 consisted of customers who did not have multiple lines. Node 14 consisted of customers who had multiple lines or who did not have phone service. For Node 14, there was a significantly larger share of suspected churners compared to non-churners.

The MultipleLines variable was also used to branch Node 10 into two nodes (Node 19 and Node 20) in which, for both cases, there was a larger share of suspected non-churners than churners. Node 9 was branched using the Monthly Charges variable (Node 15 and Node 16). Node 15 contained customers who were charged for service at a price of 55.25 dollars or less per month. In that node, more than half of the customers were suspected of churning. In Node 16, which contained customers who paid more than 55.25 dollars per month for telecommunication services, there was a significantly larger share of not churners than churners. The PaperlessBilling variable was used to branch Node 11 into two nodes (Node 19 and Node 20). In both of those nodes, there was a larger share of non-churners compared to churners. There was a slightly higher share of churners in Node 19 compared to Node 20. Finally, Node 12 was branched to two nodes (Node 21 and 22) using the TechSupport variable. Node 21 contained customers who used technical support as an additional service. There was also a higher share of churners in Node 21 compared to Node 22. In both of the nodes, there was a higher percentage of non-churners compared to churners.

Figure 5 presents the decision tree developed for the customers in Cluster 3. The CHAID decision tree developed using the data on customers from Cluster 3 had three levels and 22 nodes, of which 13 were considered terminal. Figure 4 reveals that Tenure, Internet service, Multiple Lines, Monthly Charges, PaymentMethod, StreamingMovie, Senior Citizen, and TechSupport were statistically significant, and these were used to build the classification tree using the CHAID algorithm.

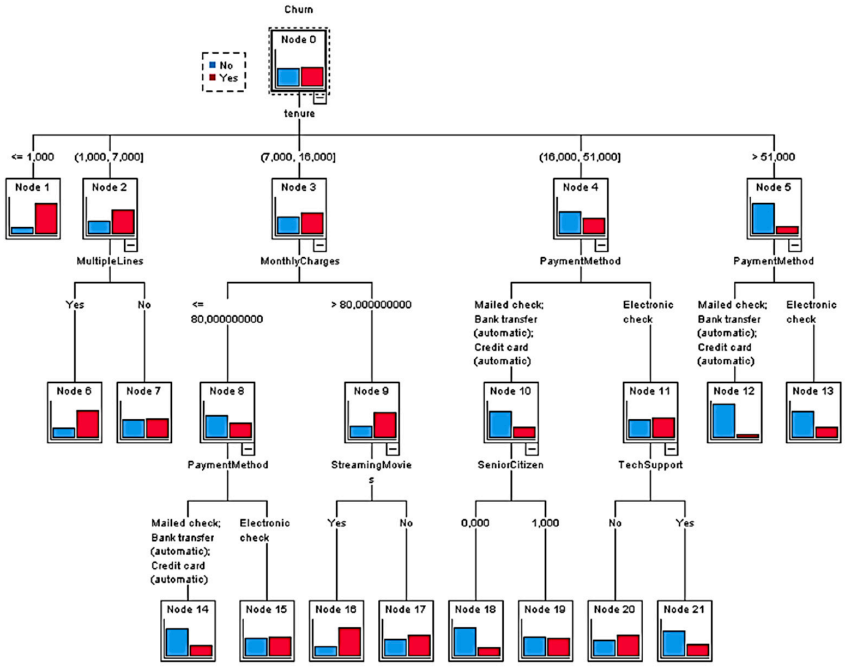

Figure 5. CHAID decision trees for Cluster 3. Source: authors' work, 2021.

The variable that was used for branching on the first level was the Tenure variable. This branching resulted in four new nodes (Node 1, Node 2, Node 3, Node 4 and Node 5). Node 1 included category $\leq$1.000 and consisted of customers who were mostly suspected of churning. Node 2 included customers who stayed with the company for between one and seven months. It had significantly more customers suspected of churning than not suspected of churning. Node 3 included customers who had been with the company for between seven and fifteen months. It had more customers suspected of churning than not suspected of churning. Node 4 included those who had been with the company for between and fifteen months, and Node 5 included customers who had been with the company for more than 51 months. In both of them, the share of suspected churners was greater than the share of non-suspected churners. However, in Node 5, the share of non-churners was significantly larger.

The second level branching variables included the MultipleLines variable for Node 2, the MonthlyCharges variable for Node 3 and the PaymentMethod variable for Node 4 and Node 5. According to Figure 3, branching resulted in eight nodes. Node 6 and Node 7 were derived from Node 2. Node 6 was related to the category of customers who used multiple phone lines. It is noticeable that a majority of them were associated with suspected churn.

Regarding Node 7, which was related to the category of customers who did not use multiple lines, it was shown that there was a slightly more significant share of customers suspected of churn than the share who were not suspected of churn. However, the ratio between suspected churners and non-churners was much higher for customers who used multiple lines. Furthermore, Node 8 and Node 9 were derived from Node 3, which was branched using the Monthly Charges variable. Node 9, in which the monthly charges were higher than 80 dollars per month, had a higher share of suspected churners.

Node 4 was branched into Node 10 and Node 11 using the PaymentMethod variable. Node 10 contained customers who paid for their service by mailed check, bank transfer and credit card, whereas Node 11 contained customers who paid by electronic check. Node 10 contained a significantly larger share of non-churners than churners. Node 11 contained a slightly more significant share of churners. Additionally, Node 5 was branched to Node 12 and Node 13 using the Payment Methods variable. Node 13, which contained customers who paid for their service by electronic check, had a larger share of churners than Node 12.

The third level branching variables included the PaymentMethod variable for Node 8, the StreamingMovies variable for Node 9, the SeniorCitizen variable for Node 10 and the TechSupport variable for Node 11. Node 8 was branched further into two nodes (Node 15 and Node 15). Node 14 consisted of customers who paid for their service by mailed check, bank transfer or credit card. For Node 14, there was a significantly larger share of non-churners compared to churners. Node 15 consisted of customers who paid for their service by electronic check. More than 50% of customers from Node 15 were churners. The StreamingMovies variable was used to branch Node 9 into two nodes (Node 16 and Node 17). There was a larger share of suspected non-churners compared to churners, but the share was more significant for Node 16, which contained customers who used movie streaming services. Node 10 was branched using the SeniorCitizen variable (Node 18 and Node 19). For the SeniorCitizen variable the value 0,000 represented non-senior customers, and 1,000 represented senior citizens. Therefore, node 18 contained non-senior customers. A significantly larger share of these customers were not suspected of churning. Node 19, which contained senior citizens, contained a larger percentage of churners than Node 18. The TechSupport variable was used to branch Node 11 into two nodes (Node 20 and Node 21). Node 20 was related to customers who did not use technical support. It was shown that more than 50% of them were suspected of churning. Finally, in Node 21, which contained customers who used technical support as an additional service, a significantly higher percentage of customers were suspected of being non-churners than churners.

Step 3.4. Rule extraction for customer relationship management

Decision trees result in the rules that can identify particular groups of customers who require special attention. In this research, the rules were used to identify groups of customers who are likely or unlikely to churn. These rules could be fed back to the customer database and be used for targeted marketing campaigns, e.g., providing incentives to the customers who were likely to churn.

The rules that predicted that customers in Cluster 2 and Cluster 3 were likely to churn for are provided in Appendix B. The rules described the specific characteristics of customers who belonged to one of the terminal nodes. For example, the rule indicated that the customers in Node 6 of Cluster 2 had a tenure shorter than one year, used fiber optic, and had an 88.17% probability of churning. In such cases, the company may decide to offer incentives to this specific group of customers. Furthermore, since these customers were using the fiber optic to connect to the Internet, the quality of the fiber optic connection could be further investigated for these specific customers it could be determined whether there were systematic problems with their internet connections.

5. Discussion

5.1. Theorethical Implications

This research explored the possibilities of using a hybrid data mining approach for churn prediction in telecommunications. A three-stage approach was used.

First, the database was prepared with the following variable groups: demographic characteristics, usage of additional services, contracts and billing, monetary value and failure, and churn. This research used the Kaggle churn dataset, which contained all the variable groups that were considered relevant to market segmentation.

Second, k-means cluster analysis was conducted using 20 independent variables, resulting in six cluster solutions. The identified clusters were compared according to the percentage of churning customers. Two clusters with the highest churn ratios were identified. Cluster 2 was the second-largest cluster, which was composed of mostly female, non-senior citizens, without dependents and partners, with month-to-month contracts with paperless billing enabled, who used using DSL internet services and phone services, and had an average tenure of 14 months which was the shortest time relative to the other clusters. Cluster 3 was the third-largest cluster, with members who were primarily male, non-senior citizens without partners and dependents, who used month-to-month contracts with paperless billing, paid relatively high monthly charges for telecommunication services by electronic check, most commonly used fiber optic for their internet service, had multiple phone lines, and were with the company for less than two years. These clusters indicate that churn in telecommunications is mainly related to tenure or to the number of months a customer has stayed with a company. Based on this, the reasons for churn can be explained in terms of a bad initial experience, low satisfaction with the services, the use of trial periods or prepaid accounts that expire automatically, or perhaps loyalties to multiple companies.

Third, decision trees, with Churn as a dependent variable, were developed using the whole database, and were developed separately for each cluster. It was discovered that the churn forecasts obtained for different clusters, i.e., marketing segments, could be drastically different. This result indicates that the use of cluster analysis as the starting point of decision tree analysis can make it possible to better understand the elements that influence client behavior. In our case study, the decision trees for the clusters with the highest churn rates (Cluster 2 and Cluster 3) also performed the best in terms of forecasting the churn behavior. Decision tree rules were extracted to identify specific groups of customers who were likely to churn in these clusters. Tailor-made marketing campaigns can be designed to incite these specific customers to stay loyal to the telecommunication company.

A theoretical implication of our findings is that a hybrid approach using k-means clustering and CHAID decision trees is appropriate for churn modelling in the telecommunications industry. Additionally, the study brings added value to the literature on customer churn prediction. Understanding the predictors of customer churn behavior

would help telecommunication companies in terms of their churn management by making it possible to prevent customers from engaging in churn behavior. The proposed approach to churn management is novel, and it differs from other studies on churn modelling in the telecommunications industry. First, most of the researchers have used accuracy-oriented approaches to churn modelling in telecommunications. They focus on the performances of the algorithms used for churn modelling while lacking an explanation of the variables identified as relevant for churn behavior or the practical recommendations regarding churn management systems. For example, Pamina et al. (2019) compared the k-nearest neighbor, random forest and XG Boost algorithms, and Ahmad et al. (2019) compared the performances of the random Forest, gradient boosted machine tree and XG Boost algorithms. Second, hybrid approaches to churn modelling in telecommunications are scarce. In terms of clustering methods and decision trees, Ullah et al. (2019) used classification methods and k-means clustering, and Preetha and Rayapeddi (2018) used logistic regression, random forest and k-means clustering for customer churn prediction. Choudhari and Potey (2018) used the decision tree, logistic regression and fuzzy unordered rule induction algorithm (FURIA) with fuzzy c-means algorithms. Our hybrid approach includes k-means clustering for customer segmentation and CHAID decision trees, which explain customer churn. Although other researchers have previously proposed the combined usage of clustering and classification methods to improve churn prediction, this paper focuses on the marketing aspects of the churn analysis and prediction process.

5.2. Practical Implications

Based on the presented hybrid approach, we developed the following recommendations for improving churn management.

First, the development of a continuous churn management process (Figure 6) is recommended. Churn management should not be a one-time action. Instead, the inclusion of the market segmentation in the customer database is recommended for use in future customer relationship programs. In addition, the effectiveness of churn rules can be measured, and rules can be further used or discarded based on their effectiveness.

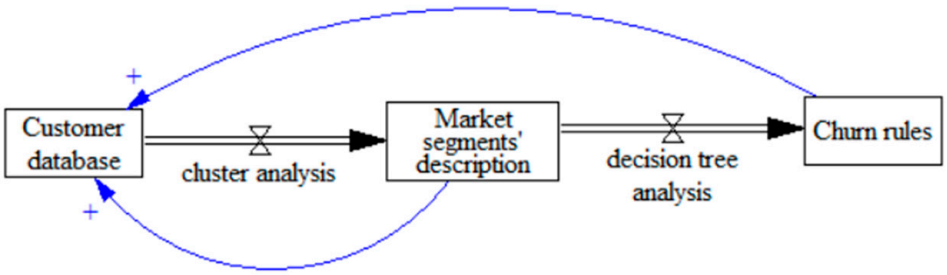

Figure 6. The continuous process of churn management. Source: authors' work, 2021.

Second, a precise approach to design incentives is recommended for the customers who are likely to churn, thus expanding the base of loyal customers who are less likely to churn. The proposed hybrid approach that segments customers into similar groups and predicts churn separately for each segment is excellent in terms of tailored incentives. Companies can also identify the customers classified as false positives and focus particular attention in their churn marketing campaigns on the customers who have similar characteristics.

Third, although our research is based on a customer base from one telecommunications provider, we strongly suggest other telecommunications companies develop their churn prediction models using their own databases, thus tailoring the proposed approach to their specific situation.

6. Conclusions

This study aimed to explain customer churn in the telecommunications industry by following a hybrid methodology for churn analysis. Four groups of variables were used to explain customer churn behavior: demographic variables, additional services, contracts and billing, and monetary value and failure. Cluster analysis based on the k-means algorithm was used to detect clusters with the highest churn occurrence. Cluster analysis as the starting point of decision tree analysis also helped to better understand the variables that influenced client behavior. The CHAID algorithm was used to generate decision trees for two clusters with the highest churn occurrence.

The research contributes to the literature on churn management using machine learning. Contrary to the majority of research, in which one method is applied for churn prediction, we propose a hybrid approach that combines clustering and classification. The combined approach resulted in the increased accuracy of classification for the clusters with the highest churn ratio, and at the same time provided in-depth knowledge about the market segment containing the customers who were the most likely to churn. Our results reveal that churn in telecommunications is mainly related to the Tenure variable, with the customers most likely to churn being those who have been with the company for the shortest period. Characteristics of the clusters with the highest level of churn indicate that churn can be explained in terms of a bad initial experience and low satisfaction with the services, and, in this study, churn was mostly associated with customers using trial periods or prepaid accounts. However, these results were tested only on the churn case study dataset, and tailor-made analyses should be planned for each company, considering that each telecommunication company has specific circumstances.

The research limitations are as follows. First, the research was based on public data from only one telecommunication company from one country. The suggested approach showed its usefulness in the case of Telco Customer Churn data from Kaggle. However, the specific rules and clusters identified by our analysis may not be relevant for other telecommunication companies when using their customer databases. The analysis of such databases could be relevant for the generalization of the proposed hybrid churn management approach. Therefore, its performance and applicability should be tested based on other data. Second, this research used only one method for clustering (k-means) and one method for classifying customers according to churn (CHAID decision trees). Although these methods are well-known and often used for churn management, future research should conduct analyses using the other research from the same group of analyses (e.g., EM clustering instead of k-means).

The proposed framework should be applied to customer downward migration, which refers to the decreasing of customer value over time due to the decreased usage of services, considering that significantly more value may be lost over time through downward migration than is lost through churn (Bayer 2010). Since customer migration and churn have similar causes, the proposed approach is likely to be useful for the prediction of downward migration. Furthermore, the proposed hybrid approach could be applied to other customer-centric industries, such as financial services, retail and health care, in future research.

Author Contributions: Conceptualization, M.P.B.; methodology, M.P.B.; software, B.J.; validation, M.P.B. and B.J.; formal analysis, M.P.B.; investigation, M.P.B.; writing—original draft preparation, M.P.B., J.P. and B.J.; writing—review and editing, M.P.B.; visualization, M.P.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly archived Telco Customer Churn open dataset, hosted on Kaggle, was used in this research, and is available at: https://www.kaggle.com/blastchar/telco-customer-churn (accessed on 9 November 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Churn Modelling in the Telecommunications Industry

Author (Year)	Methodology	Dataset	Identified Variables Relevant for Churn
Accuracy-oriented approaches to churn modelling in telecommunications			
Pamina et al. (2019)	Classifiers: k-nearest neighbor, random forest and XG Boost	IBM Watson dataset, released in 2015; the dataset contains 7043 instances and 21 attributes	Type of internet service, monthly charges
Ahmad et al. (2019)	Decision tree, random forest, gradient boosted machine tree, XG boost	Syriatel telecommunication company; customers' information over nine months	Days of last outgoing transactions, total balance, percentage of transactions with other operators, customer age, signal error and dropped calls, GSM age
Mand'ák and Hančlová (2019)	Logistic regression	Two real datasets of approximately 50,000 customers from European Telecommunications	Younger customers, customers with shorter lifetime (tenure), customers who use mobile data and SMS more than traditional calls, customers who have problems with paying bills, student accounts, contracts ending soon
Ahmed and Maheswari (2017)	Hybrid firefly-based classification	Orange dataset	N.A.
AlOmari and Hassan (2016)	Decision tree, neural network, RULES family algorithm-6	Mobile telecommunication company in Saudi Arabia; 10,000 customers, six variables	N.A.
Höppner et al. (2020)	A new classifier that integrates the EMPC metric directly into the model construction	Real-life datasets from various telecommunication service providers	N.A.
Faris (2018)	Particle swarm optimization and feedforward neural network	Unknown US mobile operator; contains 20 variables (features) and 3333 customers	Days when calls were made, voicemail messages, customer service calls, cost of international calls, local SMS fees, total consumption, total minutes of use for outgoing calls, total minutes of use for online outgoing calls
Swetha and Dayananda (2020)	Improved XgBoost	South Asia GSM (Global System for Mobile) dataset consists of over 64,000 instances and 29 variables	N.A.
Sjarif et al. (2019)	Pearson correlation, k nearest neighbor (KNN) algorithm	Public dataset Telco Customer churn available on the Kaggle platform; 7042 rows with 21 attributes	N.A.

Author (Year)	Methodology	Dataset	Identified Variables Relevant for Churn
Azeem et al. (2017)	Neural network, linear regression, C4.5, SVM, AdaBoost, gradient boosting and random forest are compared with fuzzy classifiers	Dataset from a telecommunication company operating in South Asia; it contains 600,000 instances with 722 attributes, all extracted from customer service usage patterns	N.A.
Li and Marikannan (2019)	Particle swarm optimization (PSO) as well as extreme learning machine (ELM)	Telecommunication churn dataset obtained from Kaggle; it consists of 3333 records and 21 features	The number of customer service calls, the number of international calls, the number of voicemail messages, night charges and international charges
Ahmed et al. (2020)	Boosted-stacked learners and bagged-stacked learners	UCI Churn dataset with 5000 samples and 20 attributes, most of which are related to the call detail records	N.A.
Almufadi et al. (2019)	Convolutional neural networks (CNN)	The Mobile Telephony Churn Prediction Dataset contains data for around 100,000 individuals	N.A.
Hybrid approaches to churn modelling in telecommunications			
Ullah et al. (2019)	Classification (various methods), k-means clustering	Two datasets with behavioral variables measuring the number of calls and minutes	Calls and minutes within and outside of network, free and charged minutes
Choudhari and Potey (2018)	Hybrid decision tree and logistic regression classifier; fuzzy unordered rule induction algorithm (FURIA) with fuzzy c-means algorithms	Telecommunication dataset with 20 attributes and 2666 entities, with 2278 non-churners and 388 churners	N.A.
Olle and Cai (2014)	Logistic regression, voted perceptron	Dataset from Asian mobile telecommunication operator; it recapitulates the 6-month activity of 2000 subscribers, over 23 different data variables	Length of contract, age and total revenue
Preetha and Rayapeddi (2018)	Logistic regression, random forests and k-means clustering	Dataset consisting of 3400 entities; 19 attributes are selected out of 22 attributes	N.A.

Appendix B. Decision Tree Rules Extracted

Appendix B.1. Cluster 2 Rules

/* Node 13 */ (IF tenure <= 1) AND (InternetService = “Fiber optic”) AND (MultipleLines = “Yes” AND MultipleLines = “No phone service”) THEN Prediction = ‘Yes’; Probability = 51.45%

/* Node 14 */ IF ((tenure <= 1) AND (InternetService = “Fiber optic”) AND (MultipleLines = “Yes” OR MultipleLines = “No phone service”) THEN Prediction = ‘Yes’; Probability = 78.00%

/* Node 6 */ IF (tenure <= 1) AND (InternetService = “Fiber optic”) THEN Prediction = ‘Yes’; Probability = 88.17%

/* Node 15 */ IF (tenure <= 5) AND (InternetService = “Fiber optic”) AND (MonthlyCharges <= 55.25) THEN Prediction = ‘Yes’; Probability = 53.50%

/* Node 16 */ IF (tenure > 1 AND tenure <= 5) AND (InternetService = “Fiber optic”) AND (MonthlyCharges > 55.25) THEN Prediction = ‘No’; Probability = 71.93%

/* Node 8 */ IF (tenure > 1 AND tenure <= 5) AND (InternetService = “Fiber optic”) THEN Prediction = ‘Yes’; Probability = 64.89%

/* Node 17 */ IF (tenure > 5 AND tenure <= 17) AND (InternetService = “Fiber optic”) AND (MultipleLines = “Yes” AND MultipleLines = “No phone service”) THEN Prediction = ‘No’; Probability = 78.24%

/* Node 18 */ IF (tenure > 5 AND tenure <= 17) AND (InternetService = “Fiber optic”) AND (MultipleLines = “Yes” OR MultipleLines = “No phone service”) THEN Prediction = ‘No’; Probability = 61.29%

/* Node 10 */ IF (tenure > 5 AND tenure <= 17)) AND (InternetService = “Fiber optic”) THEN Prediction = ‘Yes’; Probability = 52.48%

/* Node 19 */ IF (tenure > 17) AND (Contract = “One year” AND Contract = “Two year”) AND (PaperlessBilling = “No”) THEN Prediction = ‘No’; Probability = 75.71%

/* Node 20 */ IF (tenure > 17) AND (Contract = “One year” AND Contract = “Two year”) AND (PaperlessBilling = “No”) THEN Node = 20; Probability = 90.67%

/* Node 21 */ IF (tenure > 17) AND (Contract = “One year” OR Contract = “Two year”) AND (TechSupport = “Yes”) THEN Prediction = ‘No’; Probability = 89.40%

/* Node 22 */ IF (tenure > 17) AND (Contract = “One year” OR Contract = “Two year”) AND (TechSupport = “Yes”) THEN Prediction = ‘No’; Probability = 98.53%

Appendix B.2. Cluster 3 Rules

/* Node 1 */ IF (tenure <= 1)) THEN Prediction = ‘Yes’; Probability = 83.33%

/* Node 6 */ IF (tenure > 1 AND tenure <= 7)) AND (MultipleLines = “No”) THEN Prediction = ‘Yes’; Probability = 74.30%

/* Node 7 */ IF (tenure > 1 AND tenure <= 7)) AND (MultipleLines = “No”) THEN Prediction = ‘Yes’; Probability = 51.07%

/* Node 14 */ IF (tenure > 7 AND tenure <= 16) AND (AND (MonthlyCharges <= 80)) AND (PaymentMethod = “Mailed check” OR PaymentMethod = “Bank transfer (automatic)” OR PaymentMethod = “Credit card (automatic)”) THEN Prediction = ‘No’; Probability = 73.21%

/* Node 15 */ IF (tenure > 7 AND tenure <= 16) AND ((MonthlyCharges <= 80) AND (PaymentMethod = “Mailed check” AND PaymentMethod = “Bank transfer (automatic)” AND PaymentMethod = “Credit card (automatic)”) THEN Prediction = ‘Yes’; Probability = 51.42%

/* Node 16 */ IF (tenure > 7 AND tenure <= 16) AND (MonthlyCharges > 80) AND (StreamingMovies = “No”) THEN Prediction = ‘Yes’; Probability = 75.73%

/* Node 17 */ IF (tenure > 7 AND tenure <= 16)) AND (OR (MonthlyCharges > 80) AND (StreamingMovies = “No”) THEN Prediction = ‘Yes’; Probability = 56.36%

/* Node 18 */ IF (tenure > 16 AND tenure <= 51) AND (PaymentMethod = “Mailed check” OR PaymentMethod = “Bank transfer (automatic)” OR PaymentMethod = “Credit card (automatic)”) AND (SeniorCitizen= 1) THEN Node = 18; Probability = 76.80%

/* Node 19 */ IF (tenure > 16 AND tenure <= 51) AND (PaymentMethod = “Mailed check” OR PaymentMethod = “Bank transfer (automatic)” OR PaymentMethod = “Credit card (automatic)”) AND (SeniorCitizen = 1) THEN Prediction = ‘No’; Probability = 51.67%

/* Node 20 */ IF (tenure > 16 AND tenure <= 51) AND (PaymentMethod = “Mailed check” AND PaymentMethod = “Bank transfer (automatic)” AND PaymentMethod = “Credit card (automatic)”) AND (TechSupport = “Yes”) THEN Prediction = ‘Yes’; Probability = 57.20%

/* Node 21 */ IF (tenure > 16 AND tenure <= 51) AND (PaymentMethod = "Mailed check" AND PaymentMethod = "Bank transfer (automatic)" AND PaymentMethod = "Credit card (automatic)") AND (TechSupport = "Yes") THEN Prediction = 'No'; Probability = 67.86%

/* Node 12 */ IF (tenure > 51) AND (PaymentMethod = "Electronic check") THEN Prediction = 'No'; Probability = 91.67%

/* Node 13 */ IF (tenure > 51) AND (PaymentMethod = "Electronic check") THEN Prediction = 'No'; Probability = 70.76%

References

Ahmad, Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa. 2019. Customer churn prediction in telecommunication using machine learning in big data platform. *Journal of Big Data* 6: 28. [CrossRef]

Ahmed, Ammar A.Q., and D. Maheswari. 2017. Churn prediction on huge telecommunication data using hybrid firefly-based classification. *Egyptian Informatics Journal* 18: 215–20. [CrossRef]

Ahmed, Mahreen, Hammad Afzal, Imran Siddiqi, Muhammad Faisal Amjad, and Khawar Khurshid. 2020. Exploring nested ensemble learners using overproduction and choose approach for churn prediction in telecommunication industry. *Neural Computing and Applications* 3: 3237–51. [CrossRef]

Almufadi, Nasebah, Ali Mustafa Qamar, Rehan Ullah Khan, Mohamed Tahar, and Ben Othman. 2019. Deep learning-based churn prediction of telecommunication subscribers. *International Journal of Engineering Research and Technology* 12: 2743–48.

AlOmari, Diana, and Mohammad Mehedi Hassan. 2016. Predicting Telecommunication Customer Churn Using Data Mining Techniques. In *Internet and Distributed Computing Systems*. Edited by Wenfeng Li, Shawkat Ali, Gabriel Lodewijks, Giancarlo Fortino, Giuseppe Di Fatta, Zhouping Yin, Mukaddim Pathan, Antonio Guerrieri and Qiang Wang. Cham: Springer, pp. 167–78. [CrossRef]

Al-Refaie, Abbas. 2017. Cluster Analysis of Customer Churn in Telecommunication Industry. *International Journal of Business, Human and Social Sciences* 11: 1222–26. [CrossRef]

Azeem, Muhammad, Muhammad Usman, and Alvis Cheuk Fong. 2017. A churn prediction model for prepaid customers in telecommunication using fuzzy classifiers. *Telecommunication Systems* 66: 603–14. [CrossRef]

Bayer, Judy. 2010. Customer segmentation in the telecommunications industry. *Journal of Database Marketing & Customer Strategy Management* 17: 247–56. [CrossRef]

Bell, David, and Chidozie Mgbemena. 2017. Data-driven agent-based exploration of customer behavior. *Simulation: Transactions of the Society for Modeling and Simulation International* 94: 195–212. [CrossRef]

Boehmke, Bradley, and Brandon Greenwell. 2020. Hands-On Machine Learning with R. *Chapman and Hall/CRC*. Available online: https://bradleyboehmke.github.io/HOML/knn.html (accessed on 9 November 2021).

Bose, Indranil, and Xi Chen. 2015. Detecting the migration of mobile service customers using fuzzy clustering. *Information & Management* 52: 227–38. [CrossRef]

Calzada-Infante, Laura, María Óskarsdóttir, and Bart Baesens. 2020. Evaluation of customer behavior with temporal centrality metrics for churn prediction of prepaid contracts. *Expert Systems with Applications* 160: 113553. [CrossRef]

Cheng, Li Chen, Chia-Chi Wu, and Chih-Yi Chen. 2019. Behavior Analysis of Customer Churn for a Customer Relationship System: An Empirical Case Study. *Journal of Global Information Management* 27: 111–27. [CrossRef]

Choudhari, Atul Sunil, and Manish Potey. 2018. Predictive to Prescriptive Analysis for Customer Churn in Telecommunication Industry Using Hybrid Data Mining Techniques. Paper presented at 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, August 16–18.

Chouiekh, Alae, and El Hassane El Haj. 2020. Deep Convolutional Neural Networks for Customer Churn Prediction Analysis. *International Journal of Cognitive Informatics and Natural Intelligence* 14: 1–16. [CrossRef]

Dolnicar, Sara, Bettina Grün, and Friedrich Leisch. 2018. *Market Segmentation Analysis*. Singapore: Springer.

Droftina, Uroš, Mitja Štular, and Andrej Košir. 2015. Predicting Influential Mobile-Subscriber Churners using Low-level User Features. *Automatika* 56: 522–34. [CrossRef]

Ernst, Dominik, and Sara Dolnicar. 2018. How to avoid random market segmentation solutions. *Journal of Travel Research* 57: 69–82. [CrossRef]

Faris, Hossam. 2018. A Hybrid Swarm Intelligent Neural Network Model for Customer Churn Prediction and Identifying the Influencing Factors. *Information* 9: 288. [CrossRef]

Golubev, Alexey, Ngyuen Anh Tuan, Maxim Shcherbakov, and Tran Van Phu. 2017. Clustering helps to determine the changes in telecommunication subscribers' behavior. In *Advances in Computer Science Research (ACSR), Proceedings of the IV International research conference "Information technologies in Science, Management, Social sphere and Medicine" (ITSMSSM 2017), Russia, Tomsk, 5–8 December 2017*. Atlantis Press: pp. 339–43; ISBN 978-94-6252-432-3. [CrossRef]

Höppner, Sebastiaan, Eugen Stripling, Bart Baesens, Seppe vanden Broucke, and Tim Verdonck. 2020. Profit Driven Decision Trees for Churn Prediction. *European Journal of Operational Research* 284: 920–33. [CrossRef]

Hwang, Hyunseok, Taesoo Jung, and Euiho Suh. 2004. An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry. *Expert Systems with Applications* 26: 181–88. [CrossRef]

Jin, Xin, and Jiawei Han. 2011. Partitional Clustering. In *Encyclopedia of Machine Learning*. Edited by Claude Sammut and Geoffrey Webb. Boston: Springer. [CrossRef]

Kassambara, Alboukadel. 2017. *Practical Guide to Cluster Analysis in R Edition 1 sthda.com Unsupervised Machine Learning*. STHDA Online.

Khalili-Damghani, Kaveh, Farshid Abdi, and Shaghayegh Abolmakarem. 2018. Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries. *Applied Soft Computing* 73: 816–28. [CrossRef]

Kim, Seungyeon, Younghoon Chang, Siew Fan Wong, and Myeong Cheol Park. 2020. Customer resistance to churn in a mature mobile telecommunications market. *International Journal of Mobile Communications* 18: 41–66. [CrossRef]

Kisioglu, Pınar, and Y. Ilker Topcu. 2011. Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert Systems with Applications* 38: 7151–57. [CrossRef]

Kodinariya, Trupti M., and Prashant R. Makwana. 2013. Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies* 1: 90–95.

Lee, Samuel Sangkon, and Chia Y. Han. 2012. Finding Good Initial Cluster Center by Using Maximum Average Distance. In *JapTAL 2012: Advances in Natural Language Processing*. Edited by Hitoshi Isahara and Kyoko Kanzaki. Berlin: Springer, pp. 228–38. [CrossRef]

Li, Koh Guan, and Booma Poolan Marikannan. 2019. Hybrid Particle Swarm Optimization-Extreme Learning Machine Algorithm for Customer Churn Prediction. *Journal of Computational and Theoretical Nanoscience* 16: 3432–36. [CrossRef]

Lin, Qin, Huailing Zhang, Xizhao Wang, Yun Xue, Hongxin Liu, and Changwei Gong. 2019. A Novel Parallel Biclustering Approach and Its Application to Identify and Segment Highly Profitable Telecommunication Customers. *IEEE Access* 7: 28696–711. [CrossRef]

Mand'ák, Jan, and Jana Hančlová. 2019. Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications. *STATISTIKA* 99: 129–41.

Marini, Federico, and José Manuel Amigo. 2020. Unsupervised exploration of hyperspectral and multispectral images. In *Data Handling in Science and Technology*. Amsterdam: Elsevier, vol. 32, pp. 93–114. [CrossRef]

Mu, Qing, and Keun Lee. 2005. Knowledge diffusion, market segmentation and technological catch-up: The case of the telecommunication industry in China. *Research Policy* 34: 759–83. [CrossRef]

Olle, Georges D., and Shuqin Cai. 2014. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of e-Education, e-Business, e-Management and e-Learning* 4: 55–62. [CrossRef]

Pamina, Jeyakumar, Beschi Raja, S. SathyaBama, Selvaraj Soundarya, M. S. Sruthi, S. Kiruthika, V.J. Aiswaryadevi, and G. Priyanka. 2019. Effective Classifier for Predicting Churn in Telecommunication. *Journal of Advanced Research in Dynamical & Control Systems* 11. Available online: https://ssrn.com/abstract=3399937 (accessed on 9 November 2021).

Preetha, Shivanna, and Rohit Rayapeddi. 2018. Predicting Customer Churn in the Telecommunication Industry Using Data Analytics. Paper presented at 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, August 16–18, pp. 38–43. [CrossRef]

Qiu, Yuhang, Pingping Chen, Zhijian Lin, Yongcheng Yang, Lanning Zeng, and Yaqi Fan. 2020. Clustering Analysis for Silent Telecommunication Customers Based on k-means plus. Paper presented at 4th IEEE Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, June 12–14, pp. 1023–27. [CrossRef]

Sjarif, Nilam Nur Amir, Muhammad Rusydi, Mohd Yusof, Doris Hooi, Ten Wong, Suraya Ya'akob, Roslina Ibrahim, and Mohd Zamri Osman. 2019. A Customer Churn Prediction using Pearson Correlation Function and K Nearest Neighbor Algorithm for Telecommunication Industry. *International Journal of Advances in Soft Computing and Its Applications* 11: 46–59.

Swetha, P., and R.B. Dayananda. 2020. Improvised_XgBoost Machine learning Algorithm for Customer Churn Prediction. *EAI Endorsed Transactions on Energy Web* 7: 1–7. [CrossRef]

Thomassey, Sébastien, and Antonio Fiordaliso. 2006. A hybrid sales forecasting system based on clustering and decision trees. *Decisison Support Systems* 42: 408–21. [CrossRef]

TIBCO. 2020. Feature Selection and Variable Screening Overview. Available online: https://docs.tibco.com/data-science/GUID-0111ABBC-1B92-4DD0-BBA8-79374F9BE57C.html (accessed on 9 November 2021).

Tuma, Michael N., Reinhold Decker, and Sören W. Scholz. 2011. A survey of the challenges and pifalls of cluster analysis application in market segmentation. *International Journal of Market Research* 53: 391–414. [CrossRef]

Ullah, Irfan, Basit Raza, Ahmad Kamran Malik, Muhammad Imran, Saif Ul Islam, and Sung Won Kim. 2019. A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecommunication Sector. *IEEE Access* 7: 60134–49. [CrossRef]

Vazirgiannis, Michalis. 2009. Clustering Validity. In *Encyclopedia of Database Systems*. Edited by Liu Ling and Özsu Tamer. Boston: Springer.

Verbeke, Wouter, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research* 218: 211–29. [CrossRef]

Wang, Shen-Tsu. 2018. Integrating KPSO and C5.0 to analyze the omnichannel solutions for optimizing telecommunication retail. *Decision Support Systems* 109: 39–49. [CrossRef]
Wei, Chih-Ping, and I-Tang Chiu. 2002. Turning telecommunications call details to churn prediction: A data mining approach. *Expert Systems with Applications* 23: 103–12. [CrossRef]
Zheng, Feng, and Quanyun Liu. 2020. Anomalous Telecommunication Customer Behavior Detection and Clustering Analysis Based on ISP's Operating Data. *IEEE Access* 8: 42734–48. [CrossRef]
Zhou, Jian, Linli Zhai, and Athanasios A. Pantelous. 2020. Market segmentation using high-dimensional sparse consumes data. *Expert Systems with Applications* 145: 113136. [CrossRef]