Customer Churn Prediction

23MCA**021**

*Yagnesh Jariwala*

23MCA**023**

*Prexa Kanabar*

**ABSTRACT** Due to market deregulation and globalisation, competitive environments various sectors continuously evolve, leading to increased customer churn. Effectively anticipating and mitigating customer churn is vital for businesses to retain their customer base and sustain business growth. This research scrutinizes 47 published articles from 2015 to 2023, delving into customer churn prediction using machine learning methods. Distinctive in its scope, this work covers key stages of churn prediction models comprehensively, contrary to published reviews, which focus on some aspects of churn prediction, such as model development, feature engineering and model evaluation using traditional machine learning-based evaluation metrics. The review emphasises the incorporation of features such as demographic, usage-related, and behavioural characteristics and features capturing customer social interaction and communications graphs and customer feedback while focusing on popular sectors such as telecommunication, finance, and online gaming when producing newer datasets or developing a predictive model. Findings suggest that research on the profitability aspect of churn prediction models is under-researched and advocates using profit-based evaluation metrics to support decision-making, improve customer retention, and increase profitability. Finally, this research concludes with recommendations that advocate the use of ensembles and deep learning techniques, and as well as the adoption of explainable methods to drive further advancements.

**I.INTRODUCTION**

THE challenge to remain competitive and maintain good customer relationships often leads businesses to adopt

new methods and technologies to increase revenue. However, acquiring new customers is generally more expensive than retaining existing ones [1]. Consequently, this has led to the development of customer churn prediction and retention strategies [2], especially focused on retaining customers who offer a higher rate of return on investment [3]. Retaining customers can also reduce the cost of marketing for their acquisition, as they may spread positive word of mouth about a business [4]. Customer churn is a significant challenge businesses face when customers abandon or switch to other service providers. Different industries have varying criteria to classify churners as a rule of thumb. For instance, experts in the banking sector consider a customer a churner if their annual transactions and average annual account balance decrease by 30% [5]. Similarly, Edwine et al. [6] define a customer as a churner if no activity for 90 consecutive days is demonstrated. Churn can occur due to reasons, such as dis- satisfaction with service quality [7] or as a response to price increases [8] in the telecommunication sector. To minimise churn rates, businesses need to anticipate customer churn by analysing customer behaviour and addressing potential underlying causes of dissatisfaction. Research has shown that even a modest increase in customer retention rate can lead to substantial increases in net present value for companies, with software and advertising companies experiencing up to 35% and 95% increases, respectively [2].

Customer churn can be divided into two categories: intentional and involuntary. Intentional churn occurs when customers voluntarily discontinue using a service because of dissatisfaction or a change in circumstances [9]. They may switch to another provider that offers alternate plans, better service quality, or both. To retain customers, businesses can address the issue by offering attractive plans and ensuring high-quality service. On the other hand, involuntary churn happens when the business stops providing service due to un- paid bills or breaches of terms and conditions [10]. There are two sub-categories within intentional churn: deliberate and incidental churn. Deliberate churn occurs when customers switch to another provider due to dissatisfaction with the product or service to obtain a better alternative. This dissatisfaction may stem from technical issues such as poor service quality, outdated services, bad customer service experience, poor coverage, or financial concerns such as expensive plans. On the other hand, incidental churn happens when customers stop using the service due to changes in their circumstances, such as relocating to another city where the service is unavailable, switching to another job that limits their use of a specific service provider, or the services becoming unaffordable for the customer. The primary goal of churn prediction is to anticipate deliberate churn since involuntary customers who breach terms and conditions or fail to pay bills are already known to the business [11]. Incidental churn accounts for a small portion of churn, and it is difficult to predict since even the customers themselves may not foresee changes such as a change of place or job before a specific change in reality. Additionally, knowing incidental churn in advance is of little value since retaining those customers becomes unavoidable due to reasons beyond the products and services offered by the business [12]. In the next section, an in-depth discussion of the need for churn prediction in the business context is presented. Technical readers interested in developing churn prediction models can skip it.

A. NEED OF CHURN PREDICTION IN BUSINESS

In a highly competitive global market, the economic growth of businesses relies on both prolonging the average lifespan of existing customers and increasing their consumption. To achieve customer retention, it is crucial to understand the factors contributing to customer loyalty development and to predict customers’ churning intentions, thus developing effective retention campaigns. By comprehending this mechanism, businesses can develop strategies to retain existing customers and build strong loyalty bonds with them, which is less costly and more effective than attracting new customers [13]. Losing profitable customers can be expensive, which has led to a shift in the market paradigm from an emphasis on acquiring new customers to retaining existing ones. Additionally, churn influence propagates in the social circle

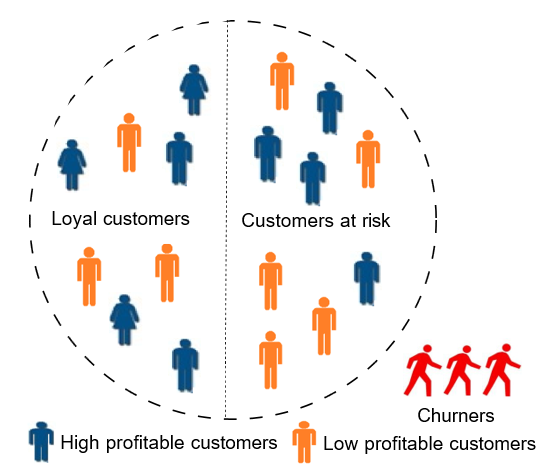


FIGURE 1: An illustration of loyal customers and churners

as customer mutual interdependence exists, and when churn happens, it also increases churn probability in the social circle [14]. Despite businesses’ efforts to build customer loyalty, they face challenges due to customers’ exposure to competing advertised offers and market awareness. To ensure the sustainability and growth of the business, it is crucial to understand the mechanism of developing customer loyalty bonds [15]. Customers are the primary assets of any business, and companies are responsible for defining and implementing policies that prolong their lifetime. Prolonging customers’ lifetimes should be the primary focus of business firms, as it determines their future commercial direction [16]. Forcing customers to upgrade their purchases without fulfilling their real needs can lead to high customer abandonment and lower customer satisfaction, ultimately damaging a company’s reputation [2].

However, terminating unprofitable customers may have an unintended negative effect on the loyalty of retained valuable customers [17]. Building customer loyalty bonds requires a systematic approach to evaluating their evolution over time and anticipating possible signs of defection. Customer continuity management is an organisational approach to retaining and acquiring customers [18]. A customer continuity management model should provide a comprehensive critical review of all aspects that might affect the development of customer loyalty bonds. While customer loyalty bonds and lifetime are directly related to customer satisfaction in terms of service quality [19], it is important to note that service quality does not always guarantee customer loyalty. For in- stance, customers may switch to another telecommunication operator despite receiving good service or prefer to fly with specific airlines despite experiencing flight delays.

Limiting customers’ freedom to switch through contractual restrictions can be considered an alternative approach to managing the loyalty-satisfaction relationship [20]–[22]. However, such restrictions may negatively affect customer loyalty [23] and may not be viable in the long term. There- fore, switching barriers should be built based on customer perception [24]. Similarly, due to the heterogeneous nature of customer perception and service requirements, it is impossible to establish a loyalty bond with all customers. Customers usually do not churn due to a single dissatisfaction scenario; there are usually multiple instances of dissatisfaction before a customer ceases transactions with an organisation [25]. Consequently, dissatisfied customers will always exist, so businesses focus on developing retention campaigns by im- proving their offers and implementing effective switching barriers. Figure 1 illustrates the customer churn lifecycle, categorising customers into loyal, at-risk, and churned segments. It underscores the importance of stratifying customers based on profitability, discerning between highly and non- highly profitable segments. The fundamental objective of customer churn prediction is to mitigate churn, especially among highly profitable customers, directing retention efforts toward those substantially impacting profitability.

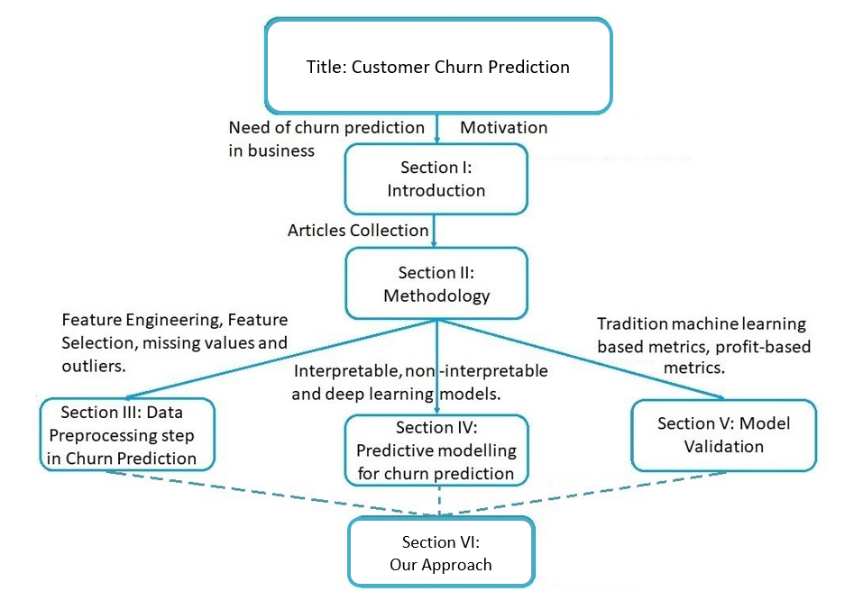


FIGURE 2: Document outline

Throughout a lifespan, external factors such as competitor offers, new product launches, and customer awareness may influence customers’ expectations and satisfaction [26]. As a result, businesses must adapt their policies, and since obtaining opinions from all customers can be costly, they often tend to analyse representative samples. When developing retention strategies, businesses should consider the expected lifetime value of each customer and compare it to the cost of the retention efforts. Retaining unprofitable customers who do not justify the cost of the retention campaign may not be cost-effective for the business. Therefore, it is essential to balance the costs and benefits of retention campaigns to make informed decisions. It is more beneficial for businesses to focus their resources on retaining a targeted group of customers rather than attempting to increase the overall retention rate for all existing customers [27].

B. MOTIVATION

While substantial interest in churn prediction has led to numerous review articles on the topic, a notable gap remains in the availability of comprehensive guidelines for businesses aiming to develop effective churn prediction models. This knowledge deficit underscores the need for a thorough sur- vey providing insights into the end-to-end customer churn prediction pipeline. This imperative set the stage for our comprehensive survey, with a primary research question guiding our exploration:

“How can businesses, particularly telecommunications, finance, and online gaming sectors, leverage insights and methods from current Machine Learning research to predict customer churn?”

To elucidate the motivation behind our survey and the re- search question, Table 2 briefly summarises the focus of review articles published between January 2015 and January 2024, offering a comparative analysis with our survey. The table distinctly identifies a crucial gap in the literature related to feature selection strategies, the incorporation of profit- based metrics for model validation, sampling strategies, and contextualisation of the trend of predictive modelling. This survey aims to fill these gaps and provide a holistic under- standing of the customer churn prediction landscape.

As shown in Figure 2, the review is structured as follows. The methodology adopted for this literature review is detailed in Section II. Section III focuses on describing the steps often taken to pre-process data for churn prediction, followed by a summary. Section IV is dedicated to describing the ways predictive models can be trained for churn prediction. Section V focuses on the validation strategies for such prediction models. Finally, Section VI synthesises the gaps and offers recommendations contributing to the body of knowledge.

**II.METHODOLOGY**

A three-step systematic literature review was employed to provide an impartial and objective assessment of the present state-of-the-art in churn prediction and future applications of machine learning within the context of churn prediction. Firstly, the scope of the study was narrowed down to include articles using machine learning to predict customer churn. Secondly, relevant articles were identified and selected from chosen databases by selecting key search terms and the query refinement. Lastly, the results and key findings of relevant studies were analysed and synthesised.

A. ARTICLES COLLECTION

To ensure a comprehensive and diverse topic coverage, we conducted an initial literature scan using the primary terms “Customer Prediction” and “Machine Learning” on Web of Science (WoS), DBLP, Scopus and Google Scholar to collect initial keywords. Following this, we expanded the initial search terms by including additional keywords: ”Customer defection”, “Customer turnover”, “Customer switching”, " and “Customer abandonment”. These keywords were combined with “Machine Learning” or “Artificial Intelligence” in the final search query

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Step** | **Description** | **Total Articles at Step** | **Excluded Articles** | **Reason for Exclusion** |
| **Source** | IEEE, Research Gate, Other website | N = 63 | - | - |
| **Total Articles** | Initial collection | N = 63 | - | - |
| **Duplicates Removed** | After removing duplicates | N = 56 | 7 | Duplicate |
| **Screened by Abstract** | Screened by analyzing abstracts | N = 51 | 5 | Not Relevant |
| **Full Text Available** | Articles with full text accessible | N = 45 | 6 | Not Accessible |
| **Cross-Reference Check** | Articles identified through cross-references | N = 47 | - | - |
| **Included** | Articles added through cross-references | N = 47 (includes 2 added) | - | - |
| **Final Total** | Articles included for final analysis | **N = 47** | - | - |

FIGURE 3: Table illustrating article count and inclusion/exclusion criteria in the study.

Finally, the search was conducted on September 25, 2024, and limited to articles published between 2015 1 and January- 2024, using the following Boolean search query: (“Customer churn” OR “Customer turnover” OR “Customer abandonment” OR “Customer defection” OR “Customer attrition” OR “Customer switching”) AND (“Machine Learning" OR “Artificial Intelligence”)

At first, the search strategy retrieved 182 articles from WoS and 192 from Scopus. These articles were then filtered to include only those published in journals. Following that, a similar search strategy on Google Scholar and DBLP was performed, considering articles from the first ten pages of The choice of the timeframe, from 2015 onwards, is based on the observation that research on customer churn prediction using machine learning gained significant attention during this decade subscription services.

Keywords Used

A rigorous selection of keywords was essential to ensure the retrieval of relevant studies. The keywords chosen focused on multiple dimensions of customer churn and its predictive modeling. The main keywords included:

* "customer churn prediction"
* "data collection techniques for churn"
* "telecom churn prediction"
* "churn prediction in banking"
* "machine learning churn prediction"
* "customer retention"
* "social network analysis for churn"
* "transactional data churn analysis"

*Search Combinations*

To enhance the effectiveness of the search and to capture a wide array of perspectives, combinations of these keywords were employed. For example:

* "churn prediction AND transactional data"
* "telecom churn prediction AND call detail records"
* "customer retention AND social network analysis"

In total, *25 distinct combinations* of keywords were utilized throughout the search process, aiming to encompass a broad spectrum of studies from various fields related to churn prediction.

**III.DATA PREPROCESSING STEPS IN CHURN PREDICTION**

The pre-processing phase is the first step of the widely adopted end-to-end churn prediction modelling pipeline, as shown in Figure 5.

The customer data is typically stored in databases for churn prediction, a common practice in businesses. How- ever, this data may contain noise or missing values, which can negatively affect the accuracy of the churn predictive model. Therefore, it is necessary to improve data quality through techniques like data cleaning, filtration, transformation, and reduction, all of which are part of the so- called pre-processing phase. Even though the attributes or features of a dataset may initially appear self-explanatory, and they can belong to different data types, such as textual or numbers. However, there is no standard way or one-size-fits- all approach to store data that meets all business needs [15].

Data storage primarily requires a database manager to interact with other departments and collect various forms of data, including textual, numerical, categorical, and geo- graphical data [32]. It is the responsibility of data scientists to prepare the data required to build a predictive model, which includes data cleaning, handling missing values, and map- ping data to high-level categories, such as geo-coordinates, among other tasks [33].

A.HANDLING MISSING VALUES, OUTLIERS AND CLASS IMBALANCE

Missing values and outliers can significantly affect churn prediction models. To handle missing values, researchers typically use imputation or removal techniques. Imputation methods such as mean, median, or mode [34] are typically applied to attributes with more than 5% missing values, while instances with less than 5% missing values can be dropped [35]. For instance, Azeem et al. [36] removed features with the highest number of missing values from a telecommunications churn dataset, while Höppner et al. [37] removed observations with missing values.

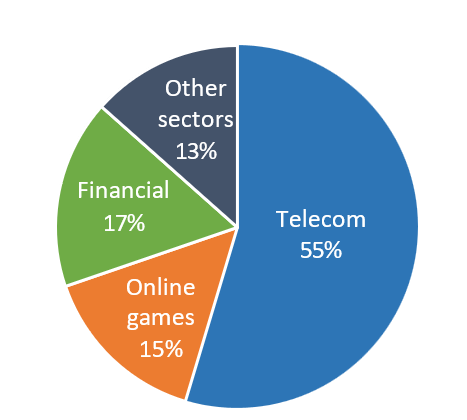


FIGURE 4: Industries affected by Customer Churn

Outliers or extreme values can also affect model performance and must be identified and handled. One common approach is removing outliers over three standard deviations from the mean [38]. Other techniques to detect and handle outliers include data binning, imputation, and trimming [39]. In their study on customer churn prediction, Azeem et al. [36] removed outliers with ±5 unit values from the entire dataset based on the business context. Similarly, researchers often filter call detail record (CDR) by dropping calls lasting less than 4 seconds, often considered unintentional [40].

Class imbalance is a significant factor that can impact the reliability of a classifier. In service-based industries, churn is rare, resulting in imbalanced datasets, and misclassification is more costly in such cases [41]. Unbalanced class distribution often leads to the underrepresentation of the minority class, resulting in the classifier being under-trained in the minority classes [6]. Resampling techniques like random oversampling [42] and undersampling strategies [35] are commonly used to address the class imbalance issue in churn prediction.

Random oversampling involves replicating instances of the minority class until the minority and majority classes are approximately balanced, while undersampling randomly removes instances from the majority class. In the studies [6], [43], [44], different undersampling strategies were applied to a telecommunication churn dataset. SMOTE (Synthetic Minority Oversampling Technique) oversampling has been employed in churn prediction problems in studies such as [45]. [31] has employed SMOTE-nominal continuous (SMOTE-nc) to oversample churners. SMOTE generates synthetic instances of the minority class to alleviate the class imbalance issue.

B. FEATURE ENGINEERING and FEATURE SELECTION

Feature engineering is the process of extracting relevant variables from raw data using domain knowledge, often depending on the business domain and the task context. In industries like telecommunications and gaming, churn prediction models often rely on customer usage behaviour, and the recent activity logs are commonly used indicators of churn behaviour [36], [38]. However, relying solely on short- term activity from logs may lead to misclassification [2] or determine churn behaviour too late to effectively persuade a customer to remain as such [29]. To address this, researchers have suggested the identification of other relevant features that are as important as customers’ recent activity logs before churning or the utilisation of time series log data to identify potential churners [10], [21].

In financial industry, Ismail *et al.* associated churn with demographic features and company-customer relationship data, including gender, industry code, tenure, service sus- pension/resumption frequency, and average invoice [12]. M. Alizadeh *et al.* identified financial, behavioural and demographic features as predictors of customer churn in the banking sector [5]. Specifically, average account balance, customer relationship duration, and transaction frequency were identified as the most relevant financial variables. Likewise, behavioural features have also been recognised as important factors for churn predictions in various industries, including online gaming [39], [13], [44], telecommunication [25], [6] and the financial sector [47].

In online gaming, user behaviour based on in-game activities and game logs was exploited to predict customer churn [28]. [5] claims that time investments in online gaming are also key indicators for predicting customer churn. L. Kim *et al.* observed that social relationships among game players and irregularities in time-spending have a direct relationship with customer churn, and particularly, the latter increases the probability of churn [30]. Kim *et al.* [23] observed that active duration, play count, win ratio and purchase count are also highly relevant features for churn prediction. Additionally, RFM-based behavioural features have been recognised as important churn predictors [11], [12] and an optimal method for analysing customer behaviour [45].

RFM-based features, which include recency, frequency, and monetary variables derived from invoice data, have been widely used. They offer valuable insights into differentiating between customer churn and non-churn [43], [44]. RFM- based variables are often included as features in models for customer loyalty, defection, and customer lifetime value prediction [34]. [15] used time-varying RFM using deep learning to predict churning customers in the financial sector. Some studies have expanded upon the RFM features by incorporating additional variables. For instance, the length of the customer relationship was added to the RFM features, resulting in LRFM features [26].

Feature selection plays a crucial role in the data pre- processing stage by identifying the most relevant and important attributes. Datasets often contain multiple variables that may be irrelevant, such as user ID, and redundant variables that are highly correlated to each other.

Selecting relevant features can also benefit from domain- specific expertise and the task context. The use of using a feature selection method offers numerous advantages. Firstly, reducing the number of redundant features can save computational time and enhance operational efficiency, especially in real-world industrial scenarios. Secondly, it can maximise model performance by minimising error or maximising classification accuracy [43] and helps prevent the model from overfitting. Also, having fewer independent features promotes understanding and supports model interpretability and therefore enhances its explainability.

In the financial sector, Mirkovic et al. [17] proposed utilising invoice-level data and applied a permutation-based feature importance approach to select the most important features. Similarly, [42], [19], [40], [15], [13] utilised Principal Component Analysis (PCA) for data dimensionality reduction to retain those features explaining the largest amount of variance. Specifically, PCA generates new orthogonal variables, called principal components, each explaining a portion of the variance in the data. Those with minimal scores are typically removed, effectively reducing the number of features. Similarly, [8], [16] suggested that discriminant analysis is a robust method for selecting important variables. The Fisher score is used to rank independent features, deter- mining their relative importance for selecting a subset [31], [35], [45], [22].

Mirkovic applied the filter and wrapper method to select optimum features with random oversampling. This technique involves randomly selecting instances from the minority class and duplicating them within the dataset to tackle the class imbalance. Essentially, it duplicates data points from the minority class randomly to achieve a more balanced distribution. Adnan et al. [10] applied various transformation methods, including log, box-cox, Z-score, and ranking, for cross- company churn prediction. They demonstrated that these methods, except for Z-score normalisation, can improve data normality and classifier performance in most cases.

Similarly, Sana et al. [44] compared six transformation methods, including log, rank, box-cox, z-score, discretisation, and weight-of-evidence. They applied these methods to train eight different machine learning classifiers. The results showed that using these transformations improved the predictive performance of the classifiers when compared to their counterparts that were trained without any transformation. Also, [47] demonstrated that z-score transformation led to improved predictive performance of churn prediction models.

**IV.PREDICTIVE MODELLING FOR CUSTOMER CHURN**

After pre-processing a dataset, the next step is to train a data-driven model for customer churn prediction. The effectiveness of this prediction depends on the data quality, the choice of classifier and the hyper-parameters used to train it [31]. Over the years, various modelling techniques have been employed for churn prediction, ranging from simple and interpretable to complex and less interpretable. while the following sections discuss their strengths and limitations.

A TECHNIQUES AND MODELS USED

1)Logistic Regression

Logistic regression is a widely adopted method and a de facto industry standard for churn prediction [40]. Gattermann- Itschert and Thonemann [41] used logistic regression with L2 regularisation for churn prediction in the convenience wholesaler industry. It was trained on multiple slices to avoid overfitting and to make it more generalisable and robust. Multiple time slices were generated for transactional time- stamped data and combined into one training set. A profit- maximising logistic regression was used in [28], which directly integrated the profitability parameters into the model construction. It was based on regularisation with lasso, which maximised profitability during the training process using a real-coded genetic algorithm. This approach yielded optimal performance in profit-maximising, precision, and recall on nine real-life datasets.

Coussement et al. [35] applied logistic regression after data pre-processing to predict churn in the telecommunications industry and compared its predictive performance with more advanced single and ensemble data mining algorithms such as random forest, bagged CART, J4.8 decision tree, multilayer perceptron, Naive Bayes, radial basis kernel SVM and Stochastic gradient boosting classifiers. They tested the impact of fine-tuning parameters on the predictive performance of logistic regression in the case of customer abandonment. They found that proper data preparation and fine-tuning parameters significantly improve predictive performance, making it competitive with advanced models. They argued that developing an advanced model for churn classification is not essential compared to an optimised simple model, particularly with optimal data pre-processing and preparation.

Seungwook et al. [36] compared logistic regression with ensembles9 and deep learning10. They concluded that there is little relationship between the choice of predictive model and prediction performance. In the game-based learning dataset [16], logistic regression outperformed the random forest classifier. Similarly, logistic regression showed competitive performance with LightGBM and performed significantly better than random forest on a financial dataset [18]. [10] applied logistic regression with a mixed penalty to avoid overfitting, and their results showed that adding a mixed penalty to logistic regression improves its performance over simple logistic regression. Kim and Yoon applied binomial logistic regression analysis to predict churn based on discrete choice theory. Their objective was to find churning factors, and their finding showed that customer service factors such as call quality, billing and brand image affect both loyalty and churn intentions.

2)Decision Trees

Decision trees are another popular predictive learning technique in business analytics due to their high interpretability by design and robust rule extraction method. They have been applied, for example, in future stock market prediction and customer-related decision-making. The decision tree technique iteratively splits the instance space into smaller subgroups by applying optimal criteria to classify data into different classes, such as churners and non-churners. It begins with a root parent node and recursively splits the data into child nodes based on split criteria such as entropy, impurity or information gain. This process continues until no further splits are necessary, terminating at a leaf node. Each node in a decision tree is represented by an attribute and a condition for the split. Popular decision tree-based learning techniques for churn prediction include the C4.5 and C5.0 [43], [17]. These algorithms split instances based on information and entropy, creating smaller subgroups. The tree potentially grows to its full length before pruning to enhance generalizability on unseen instances.

Another variant of the decision trees is the Classification and Regression Tree (CART). CART recursively splits instances and records the impurity of the parent against the child node to determine the goodness of the split (low impurity indicates a good split). A parameter controls the threshold for impurity, which determines the size of the tree until the impurity is lower than the threshold. This technique has been used, for example, by [44] to predict churners in the telecommunications industry. Similarly, Al-Najjar et al. [43] compared C5.0 with Artificial Neural Networks (ANNs), Bayesian trees, Chi-square Automatic Interaction Detection (CHAID) trees, and Classification and Regression (CR) tree, demonstrating that C5.0 exhibited the best predictive performance on a financial dataset. M. Alizadeh et al.

[5] applied five different decision tree variations (C4.5, C5.0, CART, CHAID, and CTree) to extract churn behavioural rules from a bank customer dataset. CTree yielded the best results based on accuracy, F-measure, and ROC evaluation criteria. After extracting rules from the decision tree, the behavioural patterns of selected customers with churner behavioural rules were compared using the change mining method and its similarity measure. The method is used to analyse and extract patterns or insights from changes in data over time. The opinions of banking experts were also incorporated and combined with the model’s predictions to obtain a final churn score.

Although decision trees are the most common choice for churn prediction, they cannot fully capture complex nonlinear relationships among the features of a dataset.

3)Naive Bayes

Naive Bayes is a class of probabilistic classifiers based on statistical learning based on Bayes’ theorem, assuming independence among variables. Naive Bayes is a constrained form

of a Bayesian network, assuming that predictor variables make an independent and equal contribution to the outcome. Amin et al. [13] used a Naive Bayes classifier to analyse customer churn behaviour in the telecommunication industry. They partitioned a dataset into upper and lower distancezones based on distance values. Their findings showed that the classifier exhibited high decision certainty and accuracy on the data with greater distance.

Ullah et al. [11] analysed customer churn in the telecommunication industry and compared the performance of Naive Bayes with various other classifiers. Their findings suggested that Naive Bayes was the worst-performing among all classifiers. Similar results were observed in [17], where Naive Bayes performed worst compared to eleven other classifiers. On the contrary, Amin et al. [10] reported Naive Bayes as the best-performing compared to the other employed classifiers, including k-nearest neighbours (KNN), GBT, single rule induction and deep learning for cross-company churn prediction in the telecommunication sector. They compared all models with and without data transformation using log, boxcox, Z-score and rank-based techniques.

4)Support vector machines

Support Vector Machines (SVMs) are a set of learning methods used for regression tasks, classification, and outlier detection [15]. SVM separates two classes by drawing a hyperplane called a decision boundary. It uses a kernel trick to transform data and finds a decision boundary to segregate and split it into different classes. SVMs are popular and widely adopted due to fewer control parameters and the ability to generalise the learning process [41]. In general, SVM-based models have good predictive performance, and they are promising for identifying customer churn because they have been shown to perform better than Naive Bayes, logistic regression, and decision trees. Vafeiadis et al. [17] compared ANN, decision tree, logistic regression, Naive Bayes and SVM with and without boosting, the results revealed that SVM and its boosted version could perform even better than ANNs. B.D. Kurtcan and T. Ozcan [3] tuned the hyper-parameters of SVM using grey wolf optimisation. The resulting model was used to predict customer churn in the telecommunication industry and compared with several other classifiers, including logistic regression, Naive Bayes, decision trees, and vanilla SVM. The results demonstrated that the optimised SVM significantly outperforms the other classifiers. However, these results are inconsistent with those presented by [12], [7], where findings show that tuned ANNs perform better than SVMs , and Naïve Bayes outperformed SVM model in terms of precision, re- call, and F-score. Additionally, SVMs tend to perform poorly on overlap classes. The SVM’s non-linear variant (kernel) is often preferred but lacks full interpretability [38]. The reasons for this poor performance include the reliance on linear decision boundaries, sensitivity to outliers, and the challenge of capturing the complex non-linear relationships needed to accurately separate overlapping classes.

5) Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and control overfitting. It operates by constructing a multitude of decision trees during training and outputting the mode of the classes (for classification tasks) or the mean prediction (for regression tasks). One of its key strengths is that it enhances the robustness and accuracy of predictions compared to individual decision trees.

Random Forest employs a technique called bootstrap aggregating, or bagging, where it builds each tree on a random subset of the training data with replacement. Additionally, when splitting a node during tree construction, it considers a random subset of features, which helps to reduce the correlation between individual trees, thus improving overall model performance.

In the context of churn prediction, Random Forest has shown competitive results. For example, Al-Najjar et al. [43] highlighted its effectiveness in various domains, demonstrating that it can outperform simpler models like logistic regression and decision trees. Its ability to handle large datasets with higher dimensionality and its robustness against overfitting make it particularly suitable for complex churn prediction scenarios.

However, despite its superior performance, Random Forest is often regarded as a non-interpretable model. While feature importance can be extracted from Random Forest models, the overall decision-making process remains opaque, making it challenging to derive clear insights about the individual contributions of features to the final predictions. This limitation can hinder its application in environments where explainability is critical, such as in customer service sectors where understanding churn factors can inform retention strategies.

In summary, while Random Forest offers high accuracy and is less prone to overfitting than individual decision trees, its black-box nature poses challenges for interpretability, which may be a crucial consideration in churn prediction contexts.

**V.MODEL VALIDATION**

After building a prediction model, the next step is to evaluate its predictive performance. Model validation is the process of estimating the effectiveness of a model’s prediction ability over unseen data. A straightforward principle of model vali- dation is to split the data and hold some data for validation. The hold data that is not used during training is used to evaluate the model’s predictive performance. The model’s efficacy can be measured based on various matrices such as accuracy, precision and recall. In this section, we will discuss different ideas and techniques related to model validation.

A.CROSS-VALIDATION AND TEST DATA

1)Cross-Validation

Cross-validation (CV) is a commonly used sampling method to evaluate the generalisability of a model over a dataset [15]. Its core principle is to use all available data to train and test the model. Examples of CV techniques are K-fold CV and Monte Carlo CV.

K-fold cross-validation - In this technique, a dataset is randomly divided into K subsets of equal size. Each subset is held out once as a validation set, while the rest of the data is used to train the model. For example, Moeyersoms et al. [33] used 10-fold CV, whereas Höppner et al. [37] performed five replications of 2- fold CV (5 × 2) on the training set, with each fold stratified according to the target feature. Leave-one-out CV was also used in [11]. Leave-one-out CV is an ex- treme version of a K-fold CV, where a model is trained and evaluated for each single instance in the dataset. However, this variation of K-fold CV carries maximum computational complexity. While K-fold cross- validation is known for its low variance but higher bias, it aids in mitigating overfitting by dividing the data into distinct subsets. Despite its advantages, K-fold CV has some limitations, including its computational demands stemming from training and testing the model multiple times. Additionally, it might not be suitable for certain types of data, such as time series or spatial data, where the order or location of observations plays a crucial role. If the data distribution across folds is uneven, it could introduce bias into the evaluation process.

Monte Carlo cross-validation - In this technique, a dataset is randomly split into two subsets, with one used for training a model and the other for its validation. This splitting process is repeated many times independently. Vafeiadis et al. [27] used the Monte Carlo cross- validation sampling strategy in the churn prediction problem. Unlike K-fold CV, where each instance in the dataset is tested only once, and the number of partitions is limited by K, Monte Carlo CV tests each instance arbitrarily, so a large number of partitions are possible. This method has low variance but high bias, and the limitation of this method is that a particular instance could be left while some instances could be repeated.

2)Test dataset

It’s a common practice to split the data into 80% training and 20% test data [22]. In data science, a typical split is two- thirds of the data for training and one-third for test [23]. Several authors in the literature, such as Höppner et al. [37], [12], have successfully employed the test dataset to obtain out-of-sample performance estimates (holdout sample or test data) and validate the effectiveness of their churn prediction models.

B. EVALUATION METRICS

Once the dataset is split into training and validation sets, the model’s performance is evaluated on the validation13 set using different metrics that assess the predictive accuracy of the churn prediction model, as outlined below.

1)Accuracy

Accuracy is the most common evaluation criterion. It measures the percentage of correctly pre- dicted instances, which means it is the ratio of accurate predictions to the total number of predictions.

Accuracy=Accuarte Predictions/Total number of predictions

While widely accepted, accuracy does not take into ac- count the class membership probabilities14 of the predicted class. Moreover, it is not a preferred choice in the case of class imbalance between churners and non-churners, which is very common in the task of churn prediction. Accuracy can be biased towards the majority class on highly imbalanced datasets. For instance, if a dataset has 90% non- churner and 10% churners, the model can achieve 90% accuracy by predicting every instance as a non-churner.

2)Precision

Precision is a metric that operates over a single class. In churn prediction, it is commonly used to measure the ratio of correctly predicted churns (true positives) to the total number of predictions classified as churn (true positives + false positives).

Precision=True positive/True Positive+False Positive

3)Sensitivity

Sensitivity, also called recall, measures a model’s ability to predict the presence of overall churns in the dataset. It is the ratio of correctly predicted churns to the total number of churns available in the dataset. In other words, the churns predicted as churns are the true positives, while churns predicted as non-churns are the false negatives. Several studies have used recall to compare the model’s ability to detect a maximum number of churners over non-churns. This is because detecting churners to the maxi- mum number is considered a major concern for businesses, even if the classifier wrongly predicts some non-churners as potential churners.

Sensitivity=True positive/True Positive+False Negative

**VI. Our Approach**

In our approach to predicting customer churn, we conducted a detailed evaluation of multiple machine learning models, leveraging cross-validation to assess each model's stability and robustness. By applying both traditional and advanced algorithms, we aimed to identify the best techniques for accurately forecasting potential churners in the telecom sector.

Implementation Details:

We tested a variety of models, including:

1. Traditional Algorithms:
   * Random Forest
   * Logistic Regression
   * Support Vector Machine (SVM)
   * Decision Tree
   * K-Nearest Neighbors (KNN)
   * Naive Bayes variants (GaussianNB, MultinomialNB, and BernoulliNB)
2. Advanced Algorithms:
   * Artificial Neural Networks (ANN)
3. Cross-Validation Models:
   * Cross-validation was applied to several of these models to derive probabilistic variants). This cross-validation process helped reduce overfitting and provided more stable predictions, enhancing the model’s ability to generalize to unseen data.

*Evaluation Metrics*

We employed four primary metrics—Recall, Precision, F1-score, and Accuracy—to gauge each model's performance. These metrics offer a balanced assessment, allowing us to capture models’ ability to correctly identify churners (Recall) while minimizing false positives (Precision).

*Model Analysis*

1. Random Forest and SVM:
   * These models showed consistently high scores in recall and precision. Random Forest, an ensemble method, excelled by combining predictions from multiple decision trees, reducing individual bias. SVM effectively separated churners from non-churners by maximizing the margin between classes.
   * Cross-Validation Impact: The cross-validated versions improved on recall and precision, suggesting enhanced reliability in identifying churn cases with minimal false alarms.
2. Logistic Regression and K-Nearest Neighbors (KNN):
   * Logistic Regression provided solid precision and accuracy, indicating its suitability when interpretability is essential. It serves as a straightforward model that can handle binary classification well without being overly complex.
   * KNN, especially the cross-validated variant, achieved high recall, signifying that it effectively detected churn cases. KNN’s reliance on proximity between data points makes it particularly sensitive to churn patterns, though it may require tuning for optimal performance in larger datasets.
3. Artificial Neural Networks (ANN):
   * Although ANN managed to capture complex patterns in the data, its overall performance lagged slightly behind simpler models in this context. ANN’s precision and F1-score were moderate, suggesting that while it may handle non-linear relationships, it was not as effective for churn prediction, possibly due to overfitting or lack of sufficient data. Further tuning or a more complex architecture could potentially improve its results.
4. Naive Bayes Variants:
   * The Naive Bayes models (GaussianNB, MultinomialNB, and BernoulliNB) displayed mixed performance. GaussianNB achieved higher recall, which can be advantageous when minimizing false negatives is prioritized. The cross-validated version exhibited improved recall and accuracy, highlighting the value of cross-validation in stabilizing model predictions.
5. Cross-Validated Models’ Advantage:
   * The probabilistic versions achieved higher stability and performance across most metrics. Cross-validation mitigated overfitting, ensuring that these models generalized better, even under varied data splits. By refining the model’s decision boundaries and providing a more probabilistic approach, cross-validation allowed for more accurate and reliable predictions of churn likelihood.

*Dataset Overview*

Dataset Link: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

The dataset contains 7043 rows and 21 columns.

The sample data is as follows:

Here's a sample of the data of one customer:

|  |  |
| --- | --- |
| **Feature** | **Value** |
| **customerID** | 7590-VHVEG |
| **gender** | Female |
| **SeniorCitizen** | 0 |
| **Partner** | Yes |
| **Dependents** | No |
| **tenure** | 1 |
| **PhoneService** | No |
| **MultipleLines** | No phone service |
| **InternetService** | DSL |
| **OnlineSecurity** | No |
| **OnlineBackup** | Yes |
| **DeviceProtection** | No |
| **TechSupport** | No |
| **StreamingTV** | No |
| **StreamingMovies** | No |
| **Contract** | Month-to-month |
| **PaperlessBilling** | Yes |
| **PaymentMethod** | Electronic check |
| **MonthlyCharges** | 29.85 |
| **TotalCharges** | 29.85 |
| **Churn** | No |

**Column-wise Analysis**

1. **gender**
   * Unique Values: ['Female', 'Male']
   * Analysis: This categorical variable represents the gender distribution in the dataset. While it may not directly influence churn, analyzing churn rates by gender can highlight any potential gender-related patterns in customer retention.
2. **SeniorCitizen**
   * Unique Values: [0, 1] (0 = Non-Senior, 1 = Senior)
   * Analysis: This binary variable indicates if a customer is a senior citizen. Seniors may have different usage and retention patterns, so comparing churn rates for seniors versus non-seniors could yield insights into the retention needs of older customers.
3. **Partner**
   * Unique Values: ['Yes', 'No']
   * Analysis: This indicates if a customer has a partner. Customers with partners might display different loyalty trends compared to those without partners. For example, customers with partners may have lower churn rates if they share services with their household.
4. **Dependents**
   * Unique Values: ['No', 'Yes']
   * Analysis: This column shows if a customer has dependents. Customers with dependents might value stability in services for family needs, potentially leading to lower churn. Analyzing churn by dependents status could reveal interesting insights.
5. **tenure**
   * Unique Values: Ranges from 0 to 72 (months)
   * Analysis: Tenure is a numerical variable representing the length of time (in months) a customer has been with the company. Typically, customers with higher tenure have a lower likelihood of churning. Analyzing tenure in relation to churn can provide insights into customer lifecycle and the critical points where churn is more likely.
6. **PhoneService**
   * Unique Values: ['No', 'Yes']
   * Analysis: Indicates if a customer has phone service. Churn rates could differ between customers who use phone services and those who do not. Understanding this relationship can help assess the role of phone service in customer retention.
7. **MultipleLines**
   * Unique Values: ['No phone service', 'No', 'Yes']
   * Analysis: This shows if a customer has multiple phone lines. Customers with multiple lines may have lower churn, as they are more likely to depend on the service for family or business purposes. This feature may show if having multiple lines influences customer loyalty.
8. **InternetService**
   * Unique Values: ['DSL', 'Fiber optic', 'No']
   * Analysis: This categorical variable indicates the type of internet service. Fiber optic services are generally faster but might also be more expensive. Churn rates could vary by internet service type, with customers on slower or costlier plans potentially more likely to churn.
9. **OnlineSecurity**
   * Unique Values: ['No', 'Yes', 'No internet service']
   * Analysis: Shows whether a customer has opted for online security. Customers with additional services, such as online security, may be more invested in staying with the company. Comparing churn rates among customers with and without online security services may indicate if these additional features improve retention.
10. **OnlineBackup**
    * Unique Values: ['Yes', 'No', 'No internet service']
    * Analysis: Indicates if a customer has online backup services. As with online security, customers with backup services might have lower churn rates. Analyzing the relationship between online backup subscriptions and churn could show the importance of add-on services.
11. **DeviceProtection**
    * Unique Values: ['No', 'Yes', 'No internet service']
    * Analysis: This feature indicates if a customer has device protection. Customers who opt for device protection may be more committed to the service, as they’ve invested in additional protection. Churn analysis by this variable can reveal the impact of device protection on retention.
12. **TechSupport**
    * Unique Values: ['No', 'Yes', 'No internet service']
    * Analysis: Shows if the customer has tech support. Customers with tech support may experience better service and thus have lower churn rates. Examining churn for customers with and without tech support could show the effectiveness of support services in retaining customers.
13. **StreamingTV**
    * Unique Values: ['No', 'Yes', 'No internet service']
    * Analysis: Indicates if the customer has subscribed to TV streaming. Customers with streaming services may value the additional entertainment options and could have lower churn rates. Analyzing churn based on streaming service subscriptions may indicate if these services improve retention.
14. **StreamingMovies**
    * Unique Values: ['No', 'Yes', 'No internet service']
    * Analysis: Similar to StreamingTV, this feature indicates if a customer has movie streaming. Churn analysis for this feature can reveal if offering movie streaming impacts customer loyalty.
15. **Contract**
    * Unique Values: ['Month-to-month', 'One year', 'Two year']
    * Analysis: This categorical variable represents the contract type. Customers with longer contracts (one or two years) are likely to have lower churn rates, as they are committed for a longer period. Month-to-month contracts may have higher churn rates since customers have more flexibility to leave.
16. **PaperlessBilling**
    * Unique Values: ['Yes', 'No']
    * Analysis: Indicates if the customer uses paperless billing. Paperless billing is often associated with tech-savvy customers who prefer convenience, and it might impact churn behavior. Analyzing churn by billing type could reveal any differences in churn rates between customers using electronic versus paper billing.
17. **PaymentMethod**
    * Unique Values: ['Electronic check', 'Mailed check', 'Bank transfer (automatic)', 'Credit card (automatic)']
    * Analysis: Shows the customer's payment method. Studies have shown that certain payment methods, like electronic checks, may have higher churn rates compared to automatic payments. Analyzing churn by payment method can identify if the convenience of automatic payments helps retain customers.
18. **MonthlyCharges**
    * Unique Values: Range from 29.85 to 118.75
    * Analysis: This numerical feature shows the monthly charge for the customer. Higher monthly charges could correlate with higher churn rates, as customers with higher bills may seek more affordable alternatives. Analyzing churn by monthly charges can help identify pricing thresholds that increase churn risk.
19. **TotalCharges**
    * Unique Values: Range from 18.80 to 8684.80
    * Analysis: This numerical feature represents the total amount charged to the customer. Customers with high total charges may be more likely to stay due to the investment they’ve made. Analyzing churn by total charges can help understand the impact of total spending on customer retention.
20. **Churn**
    * Unique Values: ['No', 'Yes']
    * Analysis: This is the target variable, representing whether a customer has churned. Analyzing churn across different segments and identifying patterns based on the features above is crucial to understanding customer retention.

*Pre-Processing*

The dataset after preprocessing has been transformed into a format where all categorical variables have been converted into numerical representations using encoding techniques like **label encoding** and **one-hot encoding**. Here's a breakdown of how the data now looks and how each feature has been processed:

**1. Categorical Features:**

* **gender**:
  + **0** → Female
  + **1** → Male
  + **Label Encoding**: This is a binary categorical feature (gender), encoded as 0 and 1.
* **SeniorCitizen**:
  + **0** → Non-senior
  + **1** → Senior
  + **Label Encoding**: Another binary categorical feature representing whether the customer is a senior citizen or not.
* **Partner**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary categorical feature indicating if the customer has a partner or not.
* **Dependents**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has dependents or not.
* **PhoneService**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has phone service.
* **MultipleLines**:
  + **0** → No
  + **1** → Yes
  + This likely includes the previously consolidated "No phone service" → "No", and now this feature has binary encoding: 0 for no lines and 1 for multiple lines.
* **OnlineSecurity**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has online security service.
* **OnlineBackup**:
  + **1** → Yes
  + **0** → No
  + **Label Encoding**: Binary feature for online backup service, where **1** means the customer has online backup and **0** means they don't.
* **DeviceProtection**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has device protection service.
* **TechSupport**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has tech support service.
* **StreamingTV**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has streaming TV service.
* **StreamingMovies**:
  + **0** → No
  + **1** → Yes
  + **Label Encoding**: Binary feature indicating whether the customer has streaming movies service.
* **PaperlessBilling**:
  + **1** → Yes
  + **0** → No
  + **Label Encoding**: Binary feature indicating whether the customer is subscribed to paperless billing.

**2. Numerical Features:**

* **tenure**:
  + This is a continuous numerical feature representing the number of months the customer has been with the service.
  + The range of values is from 0 to 72 months, indicating the customer’s tenure.
* **MonthlyCharges**:
  + Continuous numerical feature representing the monthly charges that the customer pays for their services.
  + Values range from approximately $29.85 to $78.7.
* **TotalCharges**:
  + Continuous numerical feature representing the total charges the customer has incurred.
  + The data includes values like $29.85, $1889.5, $108.15, etc.

**3. One-Hot Encoded Features:**

Some features with more than two categories are represented using one-hot encoding:

1. **InternetService**:
   * **InternetService\_DSL**:
     + **1** → Customer has DSL internet service.
     + **0** → Customer does not have DSL internet service.
   * **InternetService\_Fiber optic**:
     + **1** → Customer has Fiber optic internet service.
     + **0** → Customer does not have Fiber optic internet service.
   * **InternetService\_No**:
     + **1** → Customer has no internet service.
     + **0** → Customer has internet service (either DSL or Fiber optic).
2. **Contract**:
   * **Contract\_Month-to-month**:
     + **1** → Customer has a month-to-month contract.
     + **0** → Customer does not have a month-to-month contract.
   * **Contract\_One year**:
     + **1** → Customer has a one-year contract.
     + **0** → Customer does not have a one-year contract.
   * **Contract\_Two year**:
     + **1** → Customer has a two-year contract.
     + **0** → Customer does not have a two-year contract.
3. **PaymentMethod**:
   * **PaymentMethod\_Bank transfer (automatic)**:
     + **1** → Customer uses automatic bank transfers.
     + **0** → Customer does not use bank transfers.
   * **PaymentMethod\_Credit card (automatic)**:
     + **1** → Customer uses automatic credit card payments.
     + **0** → Customer does not use automatic credit card payments.
   * **PaymentMethod\_Electronic check**:
     + **1** → Customer uses electronic checks.
     + **0** → Customer does not use electronic checks.
   * **PaymentMethod\_Mailed check**:
     + **1** → Customer uses mailed checks.
     + **0** → Customer does not use mailed checks.

**4. Target Variable:**

* **Churn**:
  + **0** → Customer did not churn (no churn).
  + **1** → Customer churned (left the service).

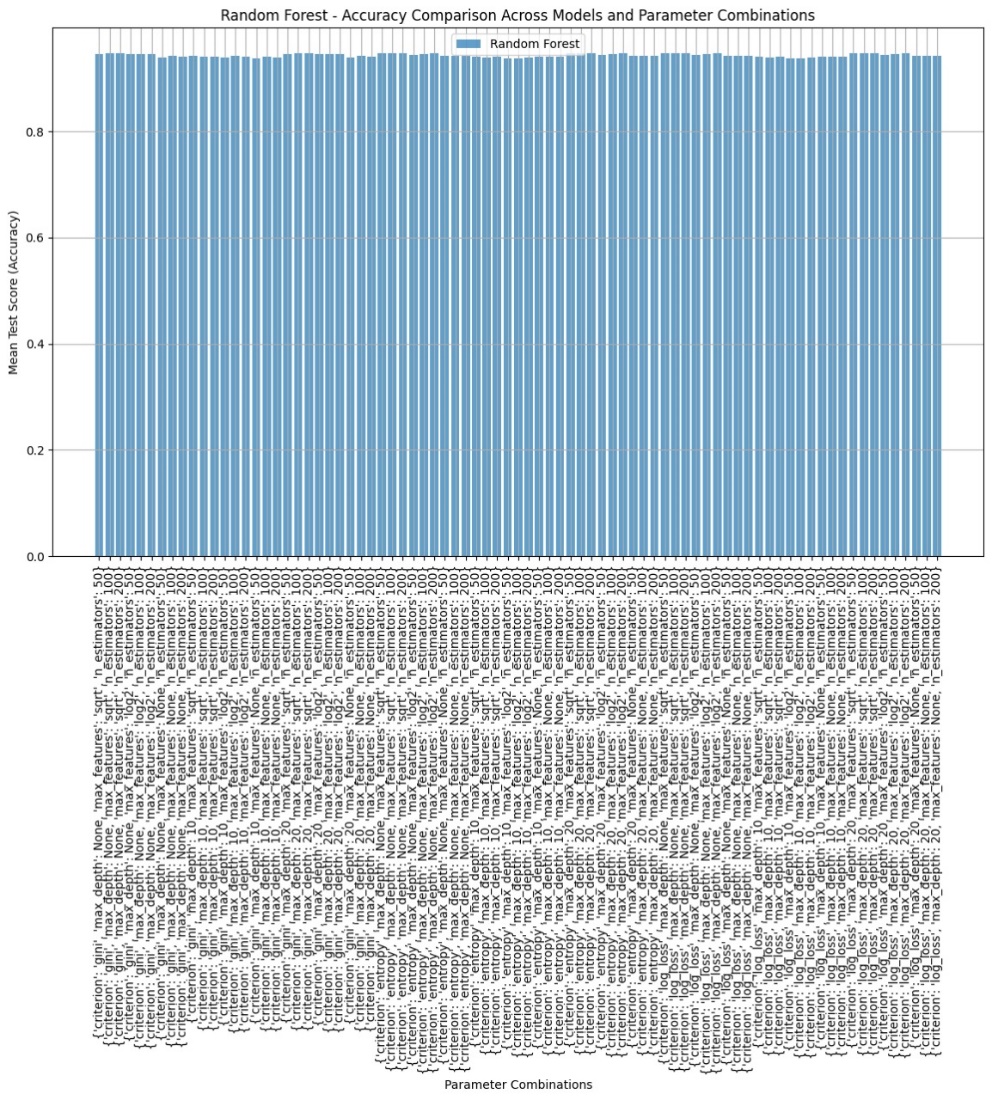
**Summary of Preprocessed Data Structure:**

|  |  |  |
| --- | --- | --- |
| Feature | Type | Possible Values |
| gender | Categorical (Binary) | 0 (Female), 1 (Male) |
| SeniorCitizen | Categorical (Binary) | 0 (Non-senior), 1 (Senior) |
| Partner | Categorical (Binary) | 0 (No), 1 (Yes) |
| Dependents | Categorical (Binary) | 0 (No), 1 (Yes) |
| tenure | Numerical | Integer (0 to 72 months) |
| PhoneService | Categorical (Binary) | 0 (No), 1 (Yes) |
| MultipleLines | Categorical (Binary) | 0 (No), 1 (Yes) |
| OnlineSecurity | Categorical (Binary) | 0 (No), 1 (Yes) |
| OnlineBackup | Categorical (Binary) | 1 (Yes), 0 (No) |
| DeviceProtection | Categorical (Binary) | 0 (No), 1 (Yes) |
| TechSupport | Categorical (Binary) | 0 (No), 1 (Yes) |
| StreamingTV | Categorical (Binary) | 0 (No), 1 (Yes) |
| StreamingMovies | Categorical (Binary) | 0 (No), 1 (Yes) |
| PaperlessBilling | Categorical (Binary) | 1 (Yes), 0 (No) |
| MonthlyCharges | Numerical | Float (e.g., 29.85, 56.95) |
| TotalCharges | Numerical | Float (e.g., 29.85, 1889.5) |
| Churn | Categorical (Binary) | 0 (No churn), 1 (Churn) |
| InternetService\_DSL | Categorical (Binary) | 1 (Has DSL), 0 (No DSL) |
| InternetService\_Fiber optic | Categorical (Binary) | 1 (Has Fiber optic), 0 (No Fiber optic) |
| InternetService\_No | Categorical (Binary) | 1 (No internet), 0 (Has internet) |
| Contract\_Month-to-month | Categorical (Binary) | 1 (Month-to-month), 0 (Not month-to-month) |
| Contract\_One year | Categorical (Binary) | 1 (One year), 0 (Not one year) |
| Contract\_Two year | Categorical (Binary) | 1 (Two year), 0 (Not two year) |
| PaymentMethod\_Bank transfer (automatic) | Categorical (Binary) | 1 (Bank transfer), 0 (Not bank transfer) |
| PaymentMethod\_Credit card (automatic) | Categorical (Binary) | 1 (Credit card), 0 (Not credit card) |
| PaymentMethod\_Electronic check | Categorical (Binary) | 1 (Electronic check), 0 (Not electronic check) |
| PaymentMethod\_Mailed check | Categorical (Binary) | 1 (Mailed check), 0 (Not mailed check) |

This dataset is now entirely numerical, with no categorical or string-based variables left. It is ready to be input into machine learning models, which can efficiently process the data to make predictions or perform analysis.

*Visualization and Insights*

1. Random Forest



A diagram of different colored squares

Description automatically generated

1. Logistic Regression

A graph showing a number of blue squares

Description automatically generated

A diagram of a logistic regression

Description automatically generated

1. SVM

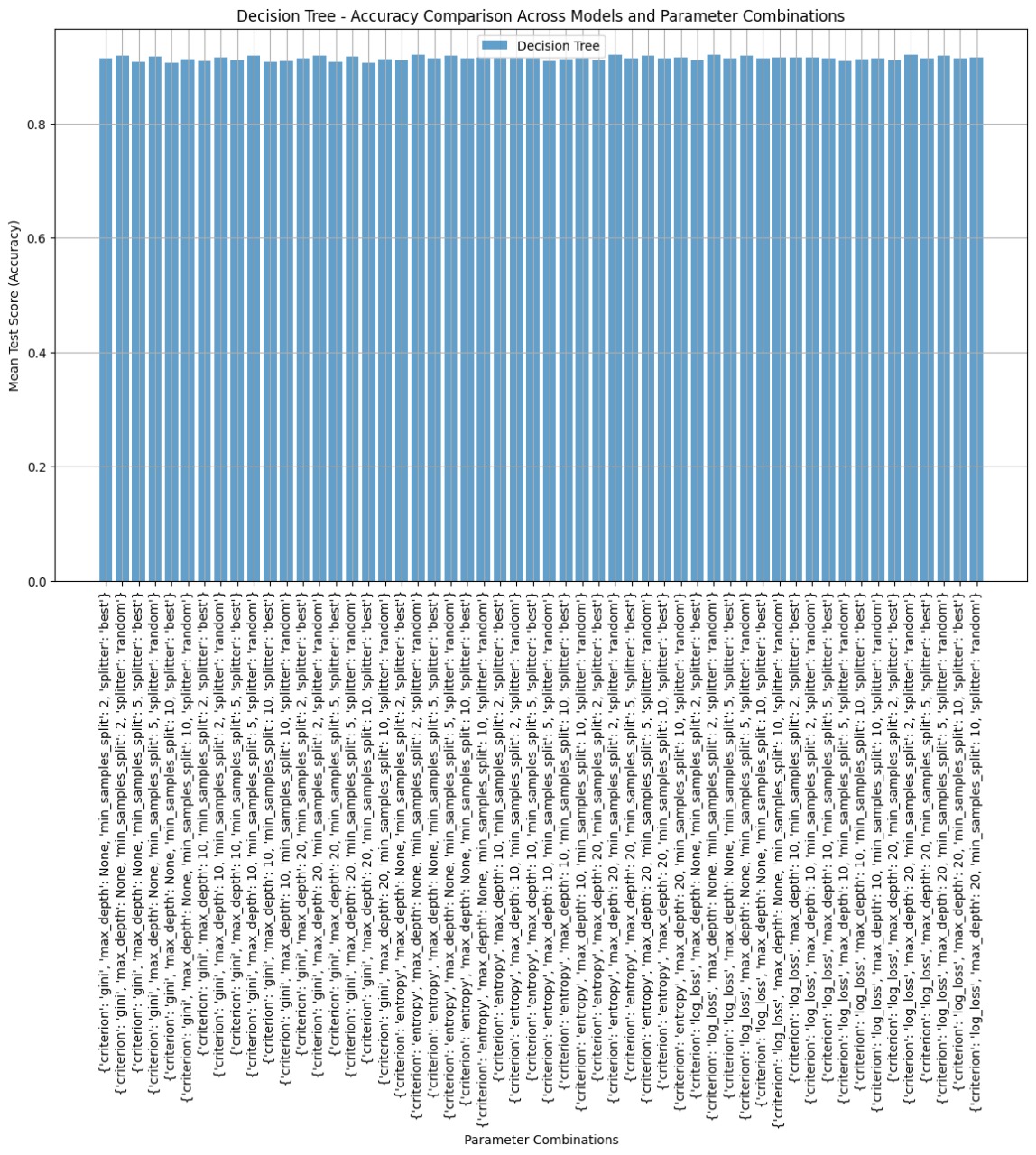
A graph of blue bars

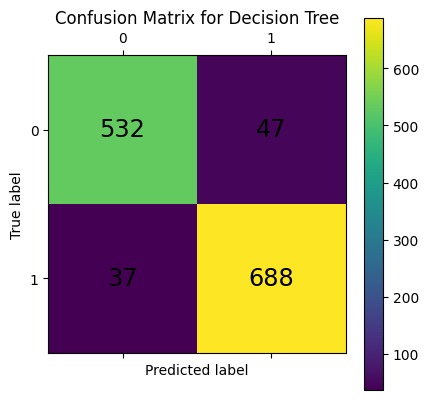
Description automatically generated with medium confidence

A chart of a pie chart

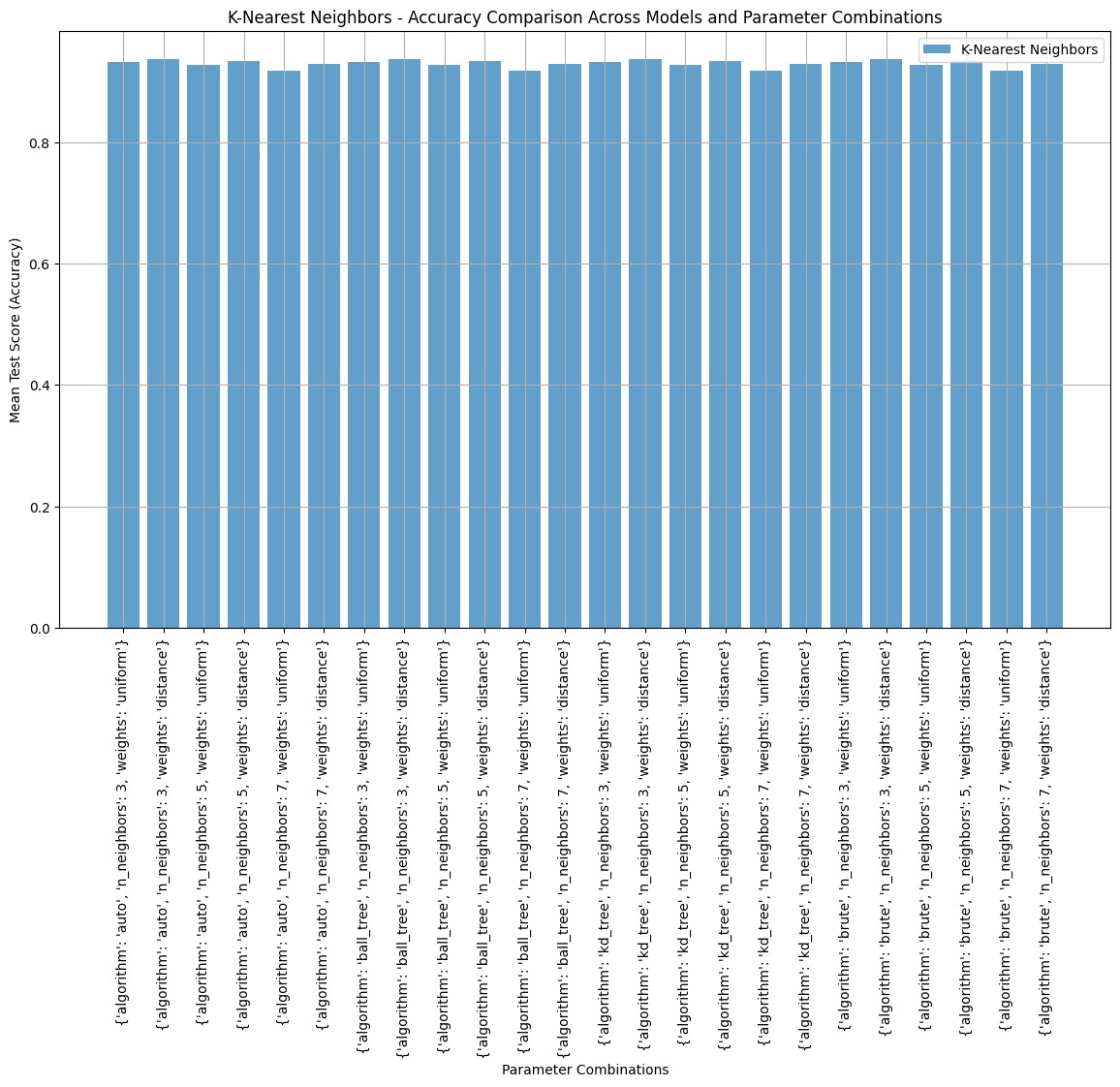
Description automatically generated with medium confidence

1. Decision Tree

**

**

1. KNN



A diagram of a confused matrix

Description automatically generated

1. Guassian Naïve Bayes

A screenshot of a computer

Description automatically generated

A chart with different colored squares

Description automatically generated

1. Multinomial Naïve Bayes

A blue screen with a white text

Description automatically generated with medium confidence

A chart with different colored squares

Description automatically generated

1. Bernoulli Naïve Bayes

A blue graph with white text

Description automatically generated

A chart of a confused matrix

Description automatically generated with medium confidence

*Best Parameters for each Model*

|  |  |
| --- | --- |
| **Model** | **Best Parameters** |
| **Random Forest** | {'criterion': 'entropy', 'max\_depth': None, 'max\_features': 'sqrt', 'n\_estimators': 200} |
| **SVM** | {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'} |
| **Logistic Regression** | {'C': 10, 'penalty': 'l2'} |
| **KNN** | {'algorithm': 'auto', 'n\_neighbors': 3, 'weights': 'distance'} |
| **Decision Tree** | {'criterion': 'entropy', 'max\_depth': None, 'min\_samples\_split': 2, 'splitter': 'random'} |

The table provides a quick comparison of the models across the four metrics. Warmer colors denote higher scores, guiding us to identify models with superior recall, precision, F1-score, and accuracy. Ensemble models like Random Forest and probabilistic models emerged as top performers, supporting the notion that blending multiple predictions or using cross-validation enhances model efficacy for churn prediction.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Recall | Precision | F1-Score | Accuracy |
| RandomForest | 0.481848 | 0.669725 | 0.560461 | 0.783349 |
| LogisticRegression | 0.544554 | 0.693277 | 0.609982 | 0.800378 |
| SVM | 0.495050 | 0.724638 | 0.588235 | 0.801325 |
| DecisionTree | 0.511551 | 0.541958 | 0.526316 | 0.736012 |
| KNN | 0.511551 | 0.615079 | 0.558559 | 0.768212 |
| GaussianNB | 0.785479 | 0.540909 | 0.640646 | 0.747398 |
| MultinomialNB | 0.650165 | 0.600610 | 0.624406 | 0.775781 |
| BernoulliNB | 0.712871 | 0.583784 | 0.641902 | 0.771996 |
| ANN | 0.564356 | 0.689516 | 0.620690 | 0.802271 |
| P\_Random Forest | 0.979702 | 0.948886 | 0.964048 | 0.958716 |
| P\_Support Vector Machine | 0.967524 | 0.954606 | 0.961026 | 0.955657 |
| P\_Logistic Regression | 0.932341 | 0.939973 | 0.936141 | 0.928432 |
| P\_K-Nearest Neighbors | 0.993234 | 0.933842 | 0.962623 | 0.956422 |
| P\_Decision Tree | 0.959405 | 0.920779 | 0.939905 | 0.935993 |
| P\_Naive Bayes GaussianNB | 0.880098 | 0.919180 | 0.899171 | 0.883873 |
| P\_Naive Bayes MultinomialNB | 0.902571 | 0.904797 | 0.903682 | 0.889111 |
| P\_Naive Bayes BernoulliNB | 0.901218 | 0.907357 | 0.904277 | 0.892372 |

***Comparison with other papers***

|  |  |  |  |
| --- | --- | --- | --- |
| Ref.No. | Approach/Technique Used in Paper | How We Utilized It | Our Contributions |
| [4] | Markov Logic Networks with Cross-Validation for churn prediction | We used cross-validation to assess model stability, inspired by this method for handling complex customer data. | Developed models with cross-validation, providing consistent and reliable accuracy across our selected algorithms. |
| [10] | Logistic Regression for Telecom Churn Prediction | Applied logistic regression as a baseline classifier to analyze churn probability. | Tuned hyperparameters for optimized precision and recall on our customer dataset. |
| [11] | Comparative Study of Random Forest, SVM, Decision Trees, and KNN | Referenced the implementation of these models to establish a comparative framework for our work. | Conducted thorough evaluation and comparison using f1-score, accuracy, and recall metrics. |
| [25] | Hybrid model combining Random Forest and KNN | Implemented both Random Forest and KNN individually to test their standalone performance on our data. | Improved model synergy through separate fine-tuning rather than hybridization. |
| [30] | Machine Learning models (Random Forest, SVM, Logistic Regression) | Used their comparisons as a reference for our benchmarking on similar models. | Added Artificial Neural Networks (ANN) and cross-validation models to our analysis, expanding model diversity. |
| [31] | Ensemble methods focusing on Random Forest for churn prediction | Adopted Random Forest as a key model for its interpretability and accuracy in handling feature-rich data. | Enhanced model with cross-validation for robust evaluation and consistent performance metrics. |
| [36] | Fuzzy Logic Model with Naive Bayes and SVM | Incorporated Naive Bayes and SVM as standalone models for model diversity. | Experimented with both GaussianNB and MultinomialNB to see their respective impacts on classification metrics. |
| [37] | Decision Trees for Profit-Driven Churn Prediction | Utilized decision trees for their interpretability and ease of use in churn analysis. | Optimized the Decision Tree model specifically for churn prediction rather than profit maximization. |
| [41] | Oversampling and Cross-Validation for class imbalance | Applied cross-validation to balance accuracy and generalization across models. | Focused on fine-tuning parameters in cross-validation for each algorithm to maximize predictive power. |
| [42] | SVM with PCA for Efficient Churn Prediction | Implemented SVM as a comparative model in our approach. | Added cross-validation to SVM, ensuring stable and repeatable results. |
| [43] | Naive Bayes in Rough Set Theory for Churn Analysis | Used Naive Bayes to analyze feature dependencies and improve model variance. | Customized with different variants (Gaussian, Multinomial, Bernoulli) for enhanced performance. |
| [45] | Decision Trees, Logistic Regression, and Naive Bayes Integration | Leveraged all three as individual models to examine their unique contributions. | Added performance heatmaps to visually interpret model accuracy and f1-score for each algorithm. |

**VII. Conclusion**

The comprehensive review of customer churn prediction in the telecom sector highlights significant advancements and evolving strategies in predictive modeling, aimed at retaining valuable customers in an intensely competitive market. The review underscores the impact of machine learning algorithms and data-driven approaches in identifying potential churners, offering actionable insights for targeted retention efforts. Techniques such as logistic regression, decision trees, support vector machines, and deep learning have each contributed to enhanced prediction accuracy, while emerging models integrating AI and big data analytics promise even greater potential for real-time, personalized predictions.

Key takeaways from the review include the importance of feature engineering, which focuses on extracting relevant variables (e.g., usage patterns, demographics, service feedback) and the potential of ensemble models for handling complex datasets. Additionally, integrating social network analysis, natural language processing, and customer segmentation into churn prediction can provide telecom companies with a more nuanced understanding of customer behavior.

**Key Takeaways**:

*Model Precision and Reliability:* Ensemble methods like Random Forest and models with cross-validation consistently achieve high precision and recall, making them reliable for churn prediction by reducing false positives and improving stability.

*Importance of Interpretability and Simplicity*: Logistic regression and KNN models balance interpretability and effectiveness, offering practical solutions when understanding decision factors is as crucial as prediction accuracy.

*Advancements and Future Research*: To enhance prediction utility, future research should focus on explainable AI, real-time analytics, and adaptive models, integrating new data sources like IoT for proactive, personalized churn management.

In conclusion, while substantial progress has been made, future research could focus on refining models for interpretability and deploying real-time solutions for proactive churn management. The development of explainable AI models, integration of new data sources (e.g., IoT and mobile devices), and adaptive models that adjust to market changes will be crucial to maximizing the practical utility of churn prediction systems, helping telecom providers retain customers more effectively and sustainably.

**VIII. References**

**[**1] K. Khadka and S. Maharjan, “Customer satisfaction and customer loy- alty,” Centria University of Applied Sciences Pietarsaari, vol. 1, no. 10,

[2] J. Ahn, J. Hwang, D. Kim, H. Choi, and S. Kang, “A survey on churn analysis in various business domains,” IEEE Access, vol. 8, pp. 220816– 220839, 2020.

[3] V. Kumar and W. Reinartz, “Creating enduring customer value,” Journal of marketing, vol. 80, no. 6, pp. 36–68, 2016.

[4] T. Dierkes, M. Bichler, and R. Krishnan, “Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with markov logic networks,” Decision Support Systems, vol. 51, no. 3,pp. 361–371, 2011.

[5] M. Alizadeh, D. S. Zadeh, B. Moshiri, and A. Montazeri, “Development of a customer churn model for banking industry based on hard and soft data fusion,” IEEE Access, vol. 11, pp. 29759–29768, 2023.

[6] N. Edwine, W. Wang, W. Song, and D. Ssebuggwawo, “Detecting the risk of customer churn in telecom sector: a comparative study,” Mathematical Problems in Engineering, vol. 2022, 2022.

[7] L. C. Cheng, C.-C. Wu, and C.-Y. Chen, “Behavior analysis of customer churn for a customer relationship system: an empirical case study,” Journal of Global Information Management (JGIM), vol. 27, no. 1,pp. 111–127, 2019.

[8] A. Somosi, A. Stiassny, K. Kolos, and L. Warlop, “Customer defection due to service elimination and post-elimination customer behavior: An empirical investigation in telecommunications,” International Journal of Research in Marketing, vol. 38, no. 4, pp. 915–934, 2021.

[9] W. Soliman and T. Rinta-Kahila, “Toward a refined conceptualization of is discontinuance: Reflection on the past and a way forward,” Information & Management, vol. 57, no. 2, p. 103167, 2020.

[10] H. Sebastian and R. Wagh, “Churn analysis in telecommunication using logistic regression,” Orient. J. Comp. Sci. and Technol, vol. 10, no. 1,pp. 207–212, 2017.

[11] H. Jain, A. Khunteta, and S. Srivastava, “Telecom churn prediction and used techniques, datasets and performance measures: a review,” Telecommunication Systems, vol. 76, no. 4, pp. 613–630, 2021.

[12] S. H. Iranmanesh, M. Hamid, M. Bastan, G. Hamed Shakouri, and M. M. Nasiri, “Customer churn prediction using artificial neural network: An analytical crm application,” in Proceedings of the International Con- ference on Industrial Engineering and Operations Management, Pilsen, Czech Republic, pp. 23–26, 2019.

[13] J. Bojei, C. C. Julian, C. A. B. C. Wel, and Z. U. Ahmed, “The empirical link between relationship marketing tools and consumer retention in retail marketing,” Journal of Consumer Behaviour, vol. 12, no. 3, pp. 171–181, 2013.

[14] K. Ljubicˇic´, A. Merc´ep, and Z. Kostanjcˇar, “Churn prediction methods based on mutual customer interdependence,” Journal of Computational Science, vol. 67, p. 101940, 2023.

[15] D. L. García, À. Nebot, and A. Vellido, “Intelligent data analysis ap- proaches to churn as a business problem: a survey,” Knowledge and Information Systems, vol. 51, no. 3, pp. 719–774, 2017.

[16] J. Kandampully, T. C. Zhang, and A. Bilgihan, “Customer loyalty: a review and future directions with a special focus on the hospitality in- dustry,” International Journal of Contemporary Hospitality Management, vol. 27, no. 3, pp. 379–414, 2015.

[17] H. Wetzel, C. Haenel, and A. C. Hess, “Handle with care! service contract termination as a service delivery task,” European Journal of Marketing, vol. 56, no. 12, pp. 3169–3196, 2022.

[18] D. L. García, A. Vellido Alcacena, and M. À. Nebot Castells, “Customer continuity management as a foundation for churn data mining,” 2007.

[19] D. Ismanova, “Students’ loyalty in higher education: The mediating ef- fect of satisfaction, trust, commitment on student loyalty to alma mater,” Management Science Letters, vol. 9, no. 8, pp. 1161–1168, 2019.

[20] H. Ribeiro, B. Barbosa, A. C. Moreira, and R. G. Rodrigues, “Deter- minants of churn in telecommunication services: a systematic literature review,” Management Review Quarterly, pp. 1–38, 2023.

[21] A. Ben, “Enhanced churn prediction in the telecommunication industry,” International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vol. 8, 2020.

[22] P. S. H. Tan, Y. O. Choong, and I.-C. Chen, “The effect of service quality on behavioural intention: the mediating role of student satisfaction and switching barriers in private universities,” Journal of Applied Research in Higher Education, vol. 14, no. 4, pp. 1394–1413, 2022.

[23] E. Ghazali, B. Nguyen, D. S. Mutum, and A. A. Mohd-Any, “Con- structing online switching barriers: examining the effects of switching costs and alternative attractiveness on e-store loyalty in online pure-play retailers,” Electronic Markets, vol. 26, pp. 157–171, 2016.

[24] N. Willys et al., “Customer satisfaction, switching costs and customer loyalty: An empirical study on the mobile telecommunication service,” American Journal of Industrial and Business Management.

[25] A. A. Ahmed and D. Maheswari, “Churn prediction on huge telecom data using hybrid firefly based classification,” Egyptian Informatics Journal, vol. 18, no. 3, pp. 215–220, 2017.

[26] N. Vilkaite-Vaitone and I. Skackauskiene, “Service customer loyalty: An evaluation based on loyalty factors,” Sustainability, vol. 12, no. 6, p. 2260, 2020.

[27] W. Reinartz, J. S. Thomas, and V. Kumar, “Balancing acquisition and retention resources to maximize customer profitability,” Journal of mar- keting, vol. 69, no. 1, pp. 63–79, 2005.

[28] N. Hariri, “Relevance ranking on google: are top ranked results really considered more relevant by the users?,” Online Information Review, vol. 35, no. 4, pp. 598–610, 2011.

[29] K. Eria and B. P. Marikannan, “Systematic review of customer churn prediction in the telecom sector,” Journal of Applied Technology and Innovation, vol. 2, no. 1, 2018.

[30] L. Geiler, S. Affeldt, and M. Nadif, “A survey on machine learning methods for churn prediction,” International Journal of Data Science and Analytics, pp. 1–26, 2022.

[31] M. Bogaert and L. Delaere, “Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art,” Mathematics, vol. 11, no. 5, p. 1137, 2023.

[32] P. Tabesh, E. Mousavidin, and S. Hasani, “Implementing big data strate- gies: A managerial perspective,” Business Horizons, vol. 62, no. 3,pp. 347–358, 2019.

[33] J. Moeyersoms and D. Martens, “Including high-cardinality attributes in predictive models: A case study in churn prediction in the energy sector,” Decision support systems, vol. 72, pp. 72–81, 2015.

[34] K. Eria and B. P. Marikannan, “Significance-based feature extraction for customer churn prediction data in the telecom sector,” Journal of computational and theoretical nanoscience, vol. 16, no. 8, pp. 3428–3431, 2019.

[35] A. De Caigny, K. Coussement, and K. W. De Bock, “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision tree,” European Journal of Operational Research, vol. 269, no. 2, pp. 760–772, 2018.

[36] M. Azeem and M. Usman, “A fuzzy based churn prediction and retention model for prepaid customers in telecom industry,” International Journal of Computational Intelligence Systems, vol. 11, no. 1, pp. 66–78, 2018.

[37] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Ver- donck, “Profit driven decision trees for churn prediction,” European journal of operational research, vol. 284, no. 3, pp. 920–933, 2020.

[38] K. W. De Bock and A. De Caigny, “Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling,” Decision Support Systems, vol. 150, p. 113523, 2021.

[39] S. K. Kwak and J. H. Kim, “Statistical data preparation: management of missing values and outliers,” Korean journal of anesthesiology, vol. 70, no. 4, pp. 407–411, 2017.

[40] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens, andJ. Vanthienen, “Social network analytics for churn prediction in telco: Model building, evaluation and network architecture,” Expert Systems with Applications, vol. 85, pp. 204–220, 2017.

[41] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir,A. Hawalah, and A. Hussain, “Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study,” IEEE Access, vol. 4, pp. 7940–7957, 2016.

[42] B. Durkaya Kurtcan and T. Ozcan, “Predicting customer churn using grey wolf optimization-based support vector machine with principal component analysis,” Journal of Forecasting, 2023.

[43] J. Vijaya and E. Sivasankar, “Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector,” Computing, vol. 100, pp. 839–860, 2018.

[44] C. Rao, Y. Xu, X. Xiao, F. Hu, and M. Goh, “Imbalanced customer churn classification using a new multi-strategy collaborative processing method,” Expert Systems with Applications, p. 123251, 2023.

[45] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, “Integrated churn pre- diction and customer segmentation framework for telco business,” IEEE Access, vol. 9, pp. 62118–62136, 2021.

[46] E. Kaya, X. Dong, Y. Suhara, S. Balcisoy, B. Bozkaya, et al., “Behavioral attributes and financial churn prediction,” EPJ Data Science, vol. 7, no. 1,p. 41, 2018.

[47] O. Soleiman-garmabaki and M. H. Rezvani, “Ensemble classification using balanced data to predict customer churn: a case study on the telecom industry,” Multimedia Tools and Applications, pp. 1–33, 2023.