SCHOOL OF ECONOMICS AND MANAGEMENT

Customer Segmentation

for Targeted Churn Intervention in the Telecom Industry

Gunel Aliyeva

Department of Economics

MSc in Data Analytics and Business Economics

Lund University School of Economics and Management

Supervisor: Michal Kos
DABN01: Master Thesis in Economics, 15 ECTS
Seminar date: August 30, 2024

Contents

Customer Segmentation for Targeted Churn Intervention in the Telecom Industry

Gunel Aliyeva

2024-08-30

Abstract

The telecom industry falls among one of the most progressive markets that operates at the level of advanced technologies and has the fastest rate of growth all over the world, including from the perspective of 5G technology, the integration of digital products, and the constant increase in demand for broadband internet. In such a competitive environment, customer attrition has been known to be a major problem, which results in a loss of revenues apart from the increased costs of acquiring new customers. Most of the conventional churn prediction models proved to be unsuitable for addressing the complex issue of different customer expectations, and therefore poor retention strategies were developed.

This research focuses on integrating churn prediction models, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, K nearest neighbors, Neural Network, and XG boost, with segmentation techniques including K-means clustering, Hierarchical clustering, DBSCAN clustering, and Gaussian mixture modeling. These models are then compared, and the effect of clustering on CHURN analysis is then measured using a telco dataset. Thus, ROC-AUC, accuracy, F1 score as well as the measures for clustering efficiency, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index are applied.

Assuming that intervention opportunities are aligned with identified segments, approaches to dealing with key factors affecting churn across each segment are suggested. The findings of this research may be useful for telecom companies to formulate better retention models, minimize costs, and increase relevant metrics.

Keywords: churn prediction models, clustering algorithms, churn intervention strategies, customer churn.

Chapter 1: Introduction

1.1 Background

The telecom industry is one of the universally acknowledged service sectors that helps connect the world and, at the same time, offers employment to billions of the global population. Telecom companies face stiff competition, and this has been occasioned by new technologies as well as the entry of new players into the market.

5G has been recently implemented, while IoT has been on the rise, making consumers more demanding by using digital services such as streaming and cloud solutions. This is because, for the same reasons as mentioned above, about the prevalence of choices and the increased degree of customer knowledge, businesses are now focusing on more specific and efficient segmentation techniques; in other words, the activity of organizing customer groups based on their similarity. It has been employed frequently in the context of marketing analytics (Yang et al. , 2016) and is a great way to detect high-risk customers so that the business can create a strategy for handling them in advance.

Moreover, with the help of customer segmentation, it is possible to forecast and regulate the process of customer turnover, referred to as customer defection . Organizations can categorize their consumers into different baskets, including the ones potentially to churn, with the help of machine learning techniques such as K-means clustering (Seo et al., 2022).

High churn rates have become problematic again. Currently, customers are inspired to get the best services, the best prices, and customer support. Other factors that lead to churn include issues to do with the quality of the network, billing issues, and customer service. Huda (2023) reports that a survey published in the International Journal of Advanced Research in Computer Science reveals that the cost incurred in gaining new clients is five to sixfold higher than that of retaining the existing ones.

Acknowledgement: *I would like to express my deepest gratitude to my supervisor, Michal Kos, whose guidance, expertise, and unwavering support have been invaluable throughout this challenging journey. His willingness to help, contribute, and always be available has been instrumental in the completion of this thesis.*

I am also immensely grateful to my family and friends for their constant encouragement and support during my master's education. Their belief in me and their unwavering support have been a source of strength and motivation throughout this process.

Although customer attrition should be fought as it costs the business money, it is likewise important to think about the company image and customer loyalty. Thus, in the contemporary global environment, firms are investing in AI and machine learning to advance customer service and the relevant analytics to determine consumer demand.

1.2 Research Problem

Earlier churn prediction models lack effectiveness because they smoothly talk about the customer base without looking at the individuality of customer segments. This oversight can result in ineffective retention strategies that do not address the specific reasons for customer attrition. Service quality may decline to the point where some clients are dissatisfied and switch to other service providers, even if the quality is good, but others may do so because competitors offer more competitive pricing.

Hence, segmentation of customers is an effective way that provides an organization with an opportunity to develop unique strategies that would address problems shared among customers of similar characteristics. For example, the company can provide more specialized outreach and offers, such as promotions, to high-value clients, but to reduce early churn, newer clients may need more onboarding support.

The questions, when answered appropriately, direct us to personalized interventions for each customer.

- How to divide the large ocean of customers using advanced clustering techniques?
- Which of the different clustering techniques is most meaningful in terms of predicting churn?
- Are the clustering algorithms actually helpful to prevent churn and predict churn?

In this master's thesis, among others, we answer the above questions. We also found interesting discrepancies in customer behavior and demographic makeup across different groups as we were mapping segments. Every segment has a story on its own about hidden patterns and preferences.

Further discussion focuses on how the incorporation of customer segmentation into churning prediction models affects organizations. Such integration not only enables telcos to make more precise predictions but also changes their knowledge of customer dynamics.

1.3 Aims and Objectives

To combat the issues of customers' attitude towards a change in their telecommunication provider, telcos go on a search in a bid to understand and dissect the causes of churning customers. The company goes down to the root cause to understand why consumers left, whether due to service issues, pricing, or competition.

When deciding on promotion campaigns, using sophisticated clustering algorithms, the telco company divides its numerous and varied clients into groups. Every cluster is full of specific attributes and tendencies, indicating a detailed picture of the customer situation.

The process goes on as the company tries to determine whether the kinds of churn prediction models are effective in the different segments. When they know which models are most effective for distinct customers, they refine their ability to predict and can make their forecasts more accurate and effective.

Based on these findings, the company produces ideas on how best to fight churn in each segment. The fun is that each of the strategies suits the characteristics of the segments, so interventions are not only efficient but also meaningful to the customers.

By achieving these objectives, our study contributes to the development of new knowledge on churn prediction and identifying the best segments of customers in the telecom environment. The findings of this study present operational tactics for telecom firms to diminish customers' attrition rates and enhance their consumer loyalty while effectively utilizing its resources.

The thesis compares and discusses the effectiveness and appropriateness of such clustering techniques as K-means, DBSCAN, Hierarchical Clustering, GMM, and churn prediction models including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, KNN, Neural Network, and XGBoost for analysing the churn and segmentation issues.

Chapter 2: Literature Review

Customer churn has been a thorn in the flesh of practically all industries, among them telecommunications, subscription services, and particularly e-commerce service providers. Often, churning results in a huge cost, as it has been proven that it takes more resources to acquire a new customer than to retain an existing one.

A large number of academic papers lie in the area of churn prediction models, the effectiveness of the models, ways to improve on them, and the rationale for their use in strategy formulation. This chapter, in general, gives an overview of the current literature on churn prediction models and clustering and their use in the creation of customer retention strategies.

2.1 Churn Prediction Models

Churn prediction models are used to predict the customers who are likely to be churning through a study of behavioral correspondents from historical data. Various models have been discussed in the literature, with different strengths and weaknesses.

2.1.1 Traditional Models

Due to its simplicity and interpretability, logistic regression has become the basic approach to churn prediction. It helps businesses know the relationship between several different characteristics of customers and the possibility of churning. For instance, Jain, Khunteta & Srivastava (2019) used logistic regression as one of the algorithms alongside the Logit Boost technique to predict customer churn in the telecommunications industry. The models were evaluated using several performance metrics, including accuracy, the Kappa statistic, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), True Positive Rate (TPR), and the Area Under the ROC Curve (AUC). The results of both models were satisfactory, with Logistic Regression achieving an accuracy of 85.24% and Logit Boost providing an accuracy of 85.18%. The authors claimed that both models were effective in predicting the churn of customers, and more research should be conducted on the combination of the models to make the predictions even more accurate.

However, logistic regression may struggle with complex, non-linear relationships, prompting researchers to explore more sophisticated machine-learning techniques.

2.1.2 Machine Learning Techniques

According to the research, among the newest and most efficient methods that were created within the framework of machine learning are decision trees, random forests, and gradient boosting. Decision trees, Random forests, Boosting technique like AdaBoost and XGBoost are stated to be effective in customer churn prediction, according to Lalwani (2022). These algorithms are noted for their ability to handle large datasets and capture complex relationships within the data. Decision trees, on the other hand, give a direct, interpretable depiction of the decision-making process, while ensemble methods like random forests are able to give better predictive outcomes by integrating a number of models (Lalwani et al., 2021). Conversely, these models may suffer from overfitting problems, especially with small datasets.

Vafeiadis et al. (2015) presented a comparison of different machine learning approaches for a case of customer churn in telecommunications. The research involved two phases: the testing of several models with cross-validation on an open dataset and the validation of boosting techniques impact on performance. The implication was that models that had been boosted were used with greater efficacy than models that were not boosted. The highest accuracy belongs to the Support Vector Machine with Polynomial Kernel (SVM-POLY) using AdaBoost, which has nearly 97% accuracy while at the same time recording an 84% F-measure.

Other relevant research by Kavitha et al. (2020) focused on the analysis of different machine learning models, namely Random Forest, Logistic Regression, and XGBoost, for modeling and predicting customer churn. Thus, it is possible to conclude that such an algorithm as Random Forest was the most efficient compared to the others because of the highest accuracy rates. Thus, the research in relation to it emphasizes the significance of data preprocessing processes, feature selection, and other suitable methods of employing machine learning for finding reliable predictions.

From the studies that were reviewed by Manzoor (2023), the performance of tree-based ensemble techniques was shown to be superior to the traditional classifiers, including logistic regression, SVM, and random forest, in terms of accuracy as well as AUC. That is why ensemble methods such as bagging and boosting, and even better when used in hybrid models, revealed higher accuracy compared to conventional methods. It also reveals that Bidirectional LSTM and RNN in deep learning yield a hefty rise in predictive accuracy compared to the well-known machine learning models. The findings reassure that the potential use of machine learning is highly promising when it comes to the matter of customer churn and the subsequent application of preventive measures.

2.1.3 Deep Learning Approaches

Churn prediction has been equally revolutionized by the emergence of deep learning. Deep learning models, particularly Artificial Neural Networks (ANNs), have been reported to produce the highest accuracy rates in churn prediction, followed closely by XGBoost compared to traditional methods. At the same time, such models automatically extract features from large datasets, which reduces the need for manual intervention (Fujo, Subramanian, & Khder, 2022).

But typically, they are computationally expensive and data-intensive models; these conditions cannot be met by small companies in most cases. Moreover, customer churn data sets are normally imbalanced, that is, the number of customers who do not churn is larger than those who churn. Random Oversampling and SMOTE have therefore been used in adjusting the target class in order to improve the training of these models. (Fujo, Subramanian, & Khder, 2022).

In many cases, research has focused on integrating deep learning with other approaches like ensemble learning and transfer learning as a means of enhancing the level of accuracy in the prediction. In this regard, the combination of the highly diverse presumptions of big data analysis and new trends in machine learning is a highly promising path for studies that examine the issue of customer churn prediction in the future.

2.2 Clustering in Customer Segmentation

There is no doubt that clustering methods are very significant in carrying out customer segmentation, whereby they group customers according to similarity in behavior, preference, or demographics. This section discusses the most commonly used clustering techniques and their use in the analysis of customer behavior.

2.2.1. Clustering Techniques

Among the available methods, K-means is one of the most frequently utilized because of its simplicity and efficiency. It divides the data into K different clusters in a way that minimizes the variance of each cluster. Many studies have confirmed this view, including the study by Bahmani et al. in 2012. The thesis discusses k-means clustering by highlighting the simplicity and popularity of the method in data mining. It stresses the necessity of initializing well so as to obtain good clustering results; it also proposes the k-means++ initialization method, which is efficient but sequential. The results suggest that a new parallel version, k-means||, has reduced

the number of passes required for initialization, and it has been observed that the new algorithm is faster at yielding optimal solutions than k-means++ in both sequential and parallel cases.

However, in general, the K-means techniques depend on the selected initial centroids and the predefined values of the number of clusters, which may lead to poor solutions if not selected carefully.
The literature also provides a detailed comparison between various clustering techniques for customer segmentation, showing the pros and cons of these techniques. For instance, although hierarchical clustering does not anticipate any predefined number of clusters and, at the same time, can capture nested clusters, it is very computationally expensive and not very suitable for large datasets. On the other hand, even though DBSCAN breaks clusters of different shapes and sizes and also captures noise and outliers, it is difficult to optimize proper performance and the parameters involved.

Therefore, customer segmentation cannot have an absolute solution; the performance of any clustering method relies on the attributes of the dataset at hand as well as the business objective.

2.3 Integration of Clustering and Churn Prediction

In recent years, a lot of research has been conducted to integrate the use of clustering techniques and churn prediction models. Therefore, the work consists of the fact that the customers' data was divided into several groups based on their previous behavior and approach, and a churn prediction solution was provided for each group. Better knowledge can be gained about the behavior of the customer, which will lead to an accurate prediction.

2.3.1 Insights through Clustering

Comparison of churn prediction results can be complimented by cluster analysis that clusters customers. A study by Lee et al. (2024) developed a hybrid approach by combining statistical modeling with machine learning for the customer churn prediction process. The customers are clustered into four clusters: new, short-term, and high-value customers who are likely to churn. It uses churn boundary determination of BTYD models and K-Means clustering for customer segmentation for improved comprehension of customer behavior. Such customers enabled the construction of customer retention programs by identifying the high-value customers most likely to defect.

The hybrid methodology turned out to be very effective, with recall rates ranging significantly between 0.56 and 0.95, depending on the dataset. This success highlights its ability to find out potentially churned customers and allow remedial actions to prevent them, thereby improving revenues while fostering customer loyalty.

Another study by Bose, I., and Chen, X. (2009) explores combining various unsupervised clustering techniques—like K-means, K-medoid, SOM, and FCM—with C5.0 decision trees and boosting for customer churn prediction. Two hybridization methods were tested: one using cluster labels to build decision trees, and another predicting churn within clusters. The new hybrid models increased churn prediction and defined the most significant indicators that may define churn, including service usage and revenue. Findings of the study indicate that future studies might focus on the other methods of supervised learning, the hybrid models for developing effective solutions, and the set of valuable heuristics for mobile service providers to manage churnout rates.

2.3.2 Limitations of Integration

Though there are some possible benefits, challenges also exist in integrating clustering with churn prediction. Analyzing the literature, it can be concluded that clustering has not always provided very high predictive accuracy. For instance, at times, the addition of one more cluster increases the complications, and there could be a trade-off in performance.

Also, it may be critical to underline the choice of clustering algorithm and the selection of features for clustering as critical for providing high results, which again underlines the importance of the proper approach towards the model design. In a related study, Dalli (2022) also pointed out the time-consuming process of selecting hyperparameters for deep learning models in churn prediction. Moreover, the study of Rodan et al. (2015) also emphasized the challenges that are associated with churn prediction, especially for small classes, and the significance of creating models that can help various stakeholders. The data gathered for churn prediction is normally unbalanced in a situation where normal customers are more than the churned ones. This is considered one of the most frequent and difficult because the standard classification algorithms appear to offer a good degree of accuracy to the bigger classes while giving no consideration to the smaller ones.

2.4 Strategic Implications of Clustering

While it is possible that clustering may not contribute value by itself to the enhancement of the accuracy of the churn prediction, the literature boasts about the effectiveness of clustering in the formulation of targeted retention strategies. This way enables the telecom companies to understand the various segments better and hence assist in changing the marketing tries to reach the highest possible customers, hence increasing the retention rates. The research of Praseeda & Shivakumar (2021) has raised an effort to discuss the combination of clustering in churn prediction models and demonstrated that clustering plays a significant role in improving businesses' marketing strategies in the telecom industry. The study explored combining unsupervised clustering techniques with C5.0 decision trees and boosting for customer churn prediction. Five clustering methods were evaluated, and two methods of hybridization were

compared. The study revealed that the hybrid models enhanced the churn prediction understanding and found out significant customer's attributes that were associated with the churn. As well, the authors provide recommendations for further studies that could consider other forms of supervised learning models, other approaches to hybridization, and other measures, features, and characteristics of customers to improve the accuracy of churn prediction.

Conclusion

From the literature, it is summarized that the churn prediction model for customer attrition can never be overemphasized in its importance, and cluster techniques can further be combined to gain insight. Although this may not have any material increase in terms of prediction accuracy, the information that may be added could offer much influence on the development of effective retention strategies.

Chapter 3: Methodology

This chapter will explain the data collection method, data analysis procedure, and research design that impacted the study. We will first talk about the research approaches used to look into customer churn and telecom industry segmentation. After that, we meet the data variables as a whole and observe the issues with the data that will be affected. Following that, churn prediction models, clustering techniques, and their assessment methods will be covered. Understanding each of these is crucial because it is important to grasp the principles of methods and techniques before starting to apply models. This chapter helps establish the validity and trustworthiness of the research findings by providing an overview of the techniques that were employed.

3.1 Research Methodologies

3.1.1 Research Design

The type of research design chosen for this study is the **correlation design.** Correlational design will enable examining the association between various factors linked to customer churn and segmentation in the telecom sector. This approach is used to determine if the observed relationship between the variables is statistically significant without modifying any of them, which is why a correlational design is employed. Correlational designs are valuable to determine if there is a relationship that can help with the planning of intercessory measures.

It is especially appropriate in this study because it allows for an assessment of the relationships between multiple variables in a natural setting. It is a non-experimental design, so the variables are simply observed as they occur. This will help us study customer behavior in the telecom industry without having to manipulate any conditions.

3.1.2 Research Methods and Instruments

Quantitative research methods are used in this study. Using this approach is suitable because the analysis of numerical data can be looked at to see if they have relationships and if they are in a particular pattern. The reason quantitative methodology is opted for is its applicability in analyzing large datasets and the statistical tendency of customers to churn. Due to the research problem of customer churn analysis, the quantitative analysis provides a foundation for analyzing the customer data and deriving actionable insights.

The major research tools adopted in this research work are customer database and data mining tools. Customer databases contain information regarding the customers' transactions and interactions, demographics, and service usage, among others, all of which are useful when it

comes to churn analysis. Data mining tools that include cluster analysis and predictive analysis will be used in the evaluation of the customers' segments and predictions of churn rate. The application of these instruments is explained, as these tools allow for obtaining comprehensive and specific data needed for the analysis process.

The database used for this research incorporates all customer transactions, interactions, demographic profiles, and the facilities of services. This becomes a vast amount of data needed to understand the different variables that describe customer churn. It has a layout from which data extraction and analysis will be an easy task.

To better understand customer behavior, clustering techniques will be used in this study so as to classify the customers based on their behaviour patterns. Clustering algorithms such as K-means, Hierarchical Clustering, Gaussian Mixture Models (GMM), and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) will be used. These algorithms are mathematically based on organizing data into distinct groups, helping to pick out patterns and behaviors unique to each cluster.

Besides clustering, predictive models will be used to estimate the likelihood of customer churn. These models included Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM, and KNN. Each model will be trained on historical data and validated using established techniques to ensure their accuracy and generalizability.

A number of steps are involved in the data extraction process to get the data ready for analysis. This includes eliminating anomalies, combining related variables into single variables, and formatting the data appropriately. Specifically, it needs to be rescaled and/or normalized, meaning the data needs to be brought into the correct range, which is usually between 0 and 1.

3.2. Data Collection

3.2.1 Dataset Description

The source of data put to use in the current research is the Telco Customer Churn dataset, which is widely utilized as a source for any kind of telecom data analysis. It includes general descriptive information about customers' accounts, with details of services subscribed, demographic details, and their churn status. There are 21 features and 7043 observations in this dataset.

Below is the table that describes the variables in the dataset, the definitions, data types, and unique values or counts. This acts as a reference for how data is structured and if there's a need for preprocessing in any way.

Table 1: Dataset Description

Variable Name	Definition	Data Type	Unique Values/Count
customerID	A unique identifier is assigned to each customer.	Categorical	7043 unique IDs
gender	Gender of the customer, indicating whether the customer is male or female.	Categorical	Male, Female
SeniorCitizen	Indicates if the customer is a senior citizen (1 for yes, 0 for no).	Binary	0, 1
Partner	Indicates if the customer has a partner (Yes or No).	Binary	Yes, No
Dependents	Indicates if the customer has dependents (Yes or No).	Binary	Yes, No
tenure	Number of months the customer has been with the company.	Numerical	0 to 72
PhoneService	Indicates if the customer has a phone service (Yes or No).	Binary	Yes, No
MultipleLines	Indicates if the customer has multiple phone lines (Yes, No, or No phone service).	Categorical	Yes, No, No phone service
InternetService	Type of internet service provided to the customer (DSL, Fiber optic, or No).	Categorical	DSL, Fiber optic, No
OnlineSecurity	Indicates if the customer has online security service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service
OnlineBackup	Indicates if the customer has online backup service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service
DeviceProtection	Indicates if the customer has device protection service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service

TechSupport	Indicates if the customer has tech support service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service
StreamingTV	Indicates if the customer has streaming TV service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service
StreamingMovies	Indicates if the customer has a streaming movie service (Yes, No, or No internet service).	Categorical	Yes, No, No internet service
Contract	The contract term of the customer (Month-to-month, One year, Two years).	Categorical	Month-to-month, One year, Two year
PaperlessBilling	Indicates if the customer has paperless billing (Yes or No).	Binary	Yes, No
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer, Credit card).	Categorical	Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
MonthlyCharges	The amount charged to the customer monthly.	Numerical	18.25 to 118.75
TotalCharges	The total amount charged to the customer.	Numerical	18.80 to 8684.80
Churn	Indicates whether the customer churned (Yes or No).	Binary	Yes, No

Categorical variables have to be encoded because they represent different classes or categories when used in a single machine-learning model. On the other hand, binary variables can either be applicable directly in analyses or be translated into binary numerical values. Numerical variables include Tenure, MonthlyCharges, and TotalCharges.

There are categories that will tell us about the presence or absence of a particular service, like MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. To deal with missing or unclear values, they might need special handling. Some of the categories, like "No internet service," have missing values for some of the variables. The categorical variables are to be encoded by either label encoding or one-hot encoding.

3.2.2 Data Preprocessing

Data preprocessing is an important step that makes the data ready and ensures that it is clean and prepared for analysis. This involves standardizing numerical features, addressing missing values, and encoding categorical variables.

Data Cleaning

While cleaning the data, some modifications were performed in order to have a consistent dataset and one ready for analysis. In the process, each variable was checked, and some adjustments were made in order to prepare it for accurate analysis.

Those transformations are applied to these variables to ensure that the dataset is suitable for further analysis. A step-by-step overview of the transformations follows:

1. Variables Converted to Numeric Format

- *TotalCharges:* This variable would appear to be categorical, and it has missing values and has been further changed into a numeric data type. Instead of excluding vertices with null values, it would eliminate them from the analysis; the median of the column was used to impute the null value.

2. Binary Variables

- *SeniorCitizen:* This variable was in binary form (0 for all persons other than senior citizens and 1 for senior citizens, respectively). Hence, no transformation was required.
- *Partner, Dependents, and Gender:* These variables were converted into binary formats:
    - *Partner* and *Dependents*: Yes = 1, No = 0.
    - *Gender*: Female = 1, Male = 0.
- *PhoneService*: This variable was also transformed into a binary format (Yes = 1, No = 0).
- *MultipleLines*: The entry “No phone service” was changed to “No,” and the variable was dichotomized to indicate if a customer had multiple lines, yes being coded as 1, no as 0.
- *InternetService*: Same technique as above applied here by standardizing "No internet service" to "No" and then encoded into two binary variables:
    - *HasDSL*: DSL = 1, No = 0.
    - *HasFiberOptic*: Fiber optic = 1, No = 0.
- *PaperlessBilling*: Converted to binary format (Yes = 1, No = 0).
- *OnlineSecurity*, *OnlineBackup*, *DeviceProtection*, *TechSupport*, *StreamingTV*, *StreamingMovies*: These variables were standardized by converting "No internet service" to "No," then encoded as binary variables (Yes = 1, No = 0).

3. Categorical Variables with One-Hot Encoding

- *PaymentMethod*: This variable was analyzed and transposed into a categorical format with one hot encoding. Four binary variables were created: ElectronicCheck, MailedCheck, BankTransfer, and CreditCard
- *Contract*: This was also converted into a categorical format with one hot encoding to give three binary variables: MonthToMonth, OneYear, and TwoYear.

4. Variables Validated and Normalized

- *Tenure*: It is the number of months since the time the customer was activated, and the values range from 0 to 72. It was checked to make it reliable, and it was not significantly altered during the process.
- *MonthlyCharges and TotalCharges*: These variables were tested for outliers. The ranges were checked, and normalization was made wherever required to make all the values returned by the various ranges in the same range.

Through these cleaning processes, the dataset was prepared for robust analysis, ensuring accuracy and consistency in the results.

3.3 Clustering, Segmentation, and Prediction

To understand customer behavior and thus prevent churn, different clustering methods and churn prediction models will be used to compare the results of different models and show how the comparisons were made using various evaluation metrics. Before proceeding, it is critical to understand certain terms and topics in order to build models and make decisions based on them.

Customer segmentation in telecommunications is the process of dividing customers into clearly identifiable groups based on their common characteristics, demographics, behavior, or usage patterns. They find it to be very useful in managing the customers' needs and, therefore, decreasing churn rates and increasing the overall business performance in the telecom sector. It can also be seen as a way that a company can coordinate its services and marketing to meet customers' requirements and choices in different ways. (Fraihat et al., 2022).

In contrast, customer churn, also referred to as customer attrition, is the measure of the number and rate of customers who ceased to use the company's product or service within a specific period. The churn rate is normally obtained by dividing the number of customers who have left the company in a particular time span by the total number of customers that the company served at the beginning of the time span (Rudd et al., 2022). It is calculated as:

$$\left(\frac{customers\ at\ the\ beginning\ of\ the\ time\ period - customers\ at\ the\ end\ of\ the\ time\ period}{customers\ at\ the\ beginning\ of\ the\ time\ period}\right) \times 100$$

Despite the enormous pool of potential consumers in the consumer market, on which foundation do the telecom companies reach the conclusion that this topic is of such significance to the industry?

Qin et al.,(2022) proposed that organizational customer loss is depreciated by the fact that for every 5% reduction, a company's profits may increase by up to 3–80%. Despite the fact that it is 16 times more difficult to retain customers than it is to gain new ones, the cost of doing the former is approximately one-fifth to one-seventh of the cost of the latter (Qin et al. , 2022). That is why it is appropriate to consider the customer journey map to identify the client's behavior pattern and prevent churn. The idea of consumer behavior can help organizations keep consumers for longer, thus translating into customer lifetime value, which is the total worth of a customer to an organization until the time he or she is associated with the brand and enhancing the quality of products and services, which gives them a competitive edge.

Traditional methodologies may fail to capture modern consumers' varied behaviors and preferences. They include demographic segmentation, which divides customers based on age, gender, income, and so on; geographic segmentation, which segments customers based on their location; behavioral segmentation, which analyzes prior behaviors such as purchase history, brand loyalty, and usage patterns; and psychographic segmentation, which segments customers based on lifestyle, beliefs, and personality traits.

Modern approaches to customer segmentation and churn analysis, on the other hand, make use of new technology and methodologies such as machine learning algorithms, a focus on additional behavioral data such as social media activity, click-through rate, or transaction history, and advanced segmentation techniques, all of which result in personalized marketing campaigns that outweigh the traditional approaches by analyzing a wider range of variables at once, predicting future behavior and churn risk with higher accuracy, real-time segmenting, and real-time insight that give customers a chance to immediately take action for the changes in their behavior.

Therefore, rather than the traditional approaches used earlier, in this paper, we shall observe enhanced clustering and machine learning methods to enhance the overall process of churn intervention in the telecommunications sector. Furthermore, the results of various measures of the segment profiles and miscellaneous quantitative evaluations of the segmentation techniques will be used for comparison.

3.3.1 Clustering Algorithms and Evaluation Metrics

Clustering algorithms are used for customer segmentation, wherein a large dataset is divided into several smaller groups known as clusters. The aim is to get objects in the same cluster as similar as possible while trying to make them dissimilar from those from different clusters. (Kashwan et al., 2013)

To better understand the behavior of different segments, we will use four clustering algorithms: K-means, Hierarchical Clustering, Gaussian Mixture Models (GMM), and DBSCAN (Dense Based Spatial Clustering of Application with Noise).

K-means

To elaborate on each of these clustering methods, it is important to clarify what they are dealing with. The initial technique that will be used is K-means. It then uses this to assign data points to the nearest centroid, updating the centroids accordingly (as averages of all points assigned), and iterating through clusters until variance within a cluster is minimized. The centroids can either be initialized through 'k-means++' for faster convergence or by random selection (Scikit-learn). Mathematically, this means that we wish to reduce the variance within each cluster.

$$Objective\ Function\ = \sum_{k=1}^{K} \sum_{x_i \in C_k} \left|\left| x_i - \mu_k \right|\right|^2$$

Where:

- K denotes the number of clusters.
- x_i represents any data point.
- C_k is the set of data points that belong to a cluster k.
- μ_k is the centroid (mean) of the cluster k.
- $// \cdot //$ denotes the Euclidean distance.

In this case, the elbow method will be used to identify the K value from this dataset. It is the appropriate method for K-means clustering because it helps to find the optimal number of clusters in data mining and machine learning algorithms (Syakur et al., 2018). The "elbow" point is defined as the place on the plot of explained variance or inertia versus the number of clusters where the rate of decrease rapidly changes. So, this point is usually the best compromise between performance and complexity and, thus, the optimal number of clusters.

Algorithm:

1. **Step 1:** Dividing the dataset into K clusters and then randomly selecting K sample points that will be the centers.

2. **Step 2:** Each datapoint picked in this way is assigned to the cluster closest by the Euclidean distance computation.
3. **Step 3:** The algorithm then keeps computing the new centroids by inaverging all the data points in each cluster.

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

4. **Step 4**: Continue repeating steps 2 and 3 until the algorithm reaches convergence, which occurs when the centroids no longer change significantly.

K-means clustering has been extensively utilized in different research areas, and together with some other machine learning techniques, it has resulted in a successful customer segmentation and churn prediction process. Recently, the methods for K-means clustering and random forest, AdaBoost, XGBoost, and ensemble learning methods have been the main algorithms that have been used to predict customer churn accurately, according to Shu-li et al. (2021), Zhang (2022), and Tran et al. (2023).

Moreover, when a telecom company applies K-means clustering for customer churn prediction, it helps in the identification of unhappy customers. As a result, the company can use the retention strategy to secure them (Sam, 2024). These models will help in predicting the number of customers who are likely to defect and segmenting the customers who will be approached for targeted marketing, which will result in the company retaining more customers.

Benefits of Using K-Means

K-means clustering has the following advantages for customer segmentation in churning prediction:

- Because of its simplicity, K-Means applies the method to different clustering sets and doesn't require much time to apply. This way, it can deal with large sets of data at a very fast rate, thus being a very convenient alternative for various clustering jobs. (Larasati et al., 2019).
- Sequentially, this is a computationally efficient process where each point is assigned to the closest centroid, and the centroid is recalculated by using the average of all points within the given cluster (Bahmani et al., 2012).
- Furthermore, the K-means criterion is also supported by other metrics to varying degrees of versatility and efficiency, depending on the problem at hand.

Challenges of using K-means

Nevertheless, K-means does have some significant limitations.

- The problem with K-means is that the algorithm usually finds the local optimum instead of the global optimal one, which can bring about lower-quality clustering (Gu & Qian, 2015).
- Choosing the right number of clusters K is a big challenge since you cannot choose the right one quickly and simply (Zhao et al., 2010). The technique begins with the assumption that clusters are spheres and are all the same size, although this may not be apparent from the field data, which results in the segmentation quality deteriorating.
- The algorithm also works less well if K-means is provided with data that is dominated by noise and outliers, which can result in erroneous customer segmentation or incorrect customer-churn result prediction. Similarly, this method does not consider variation in the densities of the cluster (Muhima et al., 2020).

Hierarchical Clustering

The second type of clustering method to be utilized is hierarchical clustering, which expresses data in a tree-like structure where clusters are merged into bigger clusters or decomposed (divided) into smaller clusters. Unlike k-means, hierarchical clustering does not require the number of clusters to be preset in advance but rather gives a hierarchy from which the clusters can be chosen. (Shetty & Singh, 2021) Hierarchical clustering typically starts by considering all points as separate clusters, and then, through either a stopping criteria or continuous merging, the two closest clusters are merged into one based on some distance metrics in the data space. The results are represented on a tree, or dendrogram, with each merging serving as a node position on that tree. The procedures behind hierarchical clustering include:

Algorithm:

1. **Step 1**: Start with each data point working as a single cluster.
2. **Step 2**: The distance or similarity matrices are computed for all the pairs of clusters.
3. **Step 3**: Recognize and then merge the two closest clusters based on the distance matrix.
4. **Step 4**: Modify the distance matrix that will reflect the merged clusters.
5. **Step 5**: Repeat steps 2-4 until the desired clustering condition is met: either only one cluster remains (for agglomerative clustering) or all data points are in separate clusters (for divisive clustering).

Benefits of using Hierarchical Clustering

Hierarchical clustering exhibits these advantages in customer segmentation:

- The method produces a dendrogram, a tree-like diagram that visually represents nested clusters. This visualization is instrumental in uncovering the hierarchical relationships and underlying structure within the data, allowing for a clear understanding of how different customer segments are related to one another (Jain & Dubes, 1988).
- Hierarchical clustering is a perfect method for analyzing complicated data sets because it can discover different patterns and connections among the objects at different levels of granularity. This significant benefit of segmenting is used to discover the fine and subtle distinctions between the different customer segments. In return, organizations will obtain a better understanding and estimation of customer needs, which will surely lead to a more customized product presentation (Wang, 2024).
- As opposed to K-means clustering, which needs the number of clusters already known when using hierarchical clustering, the clusters will not be specified in the beginning. It is only the analysis of the dendrogram as a means to decide on who to "cut" the tree that acts as a way of determining the number of clusters. The adaptability feature of segmentation is that it can be more flexible than other segmentation methods and also be more context-sensitive (Mylonas et al., 2004).
- Through the algorithm's hierarchy of procedures, it allows a deeper and more detailed segmentation of customers. It is particularly advantageous in datasets where naturally inherent hierarchies or nested relationships are involved, for instance, customer segments within broader categories. By making these hierarchical relationships available, organizations can make decisions with more knowledge and strategic thought (Rokach & Maimon, 2005).

Challenges of using Hierarchical Clustering

- Hierarchical clustering has several notable disadvantages. One of the major issues is that it requires a very intensive computational process as a result of the necessity of calculating and updating distance matrices. This becomes particularly difficult for large datasets and thus leads to a surge in computational needs and processing time. Moreover, once the decisions are made in the clustering procedure (like merging or splitting the clusters), they cannot be modified. Hence, if the primary decisions are not suitable, it can lead to unsatisfactory clustering results (Murtagh, 1983).
- The sensitivity of the method to noise and outliers, which can mislead the clustering and thus make the clusters false, is another serious disadvantage. Such unusual instances can obscure the distance calculations and mislead or create inaccurate cluster formations. Besides, the complexity of time in hierarchical clustering restricts its scalability, making it unworthy of very large datasets that need fast processing efficiency (Rokach & Maimon, 2005). This mix of computational complexity and sensitivity to quality issues can limit the algorithm's ability in large-scale or noisy environments.

Gaussian Mixture Models (GMM)

As a third clustering method, the Gaussian Mixture Models (GMM) will be applied, which supposes that the data is a combination of many Gaussian distributions with unknown parameters, each having its own mean and variance. The Expectation-Maximization (EM) method will be applied to find the best fit. This allows the clusters to have more complex structures and overlap with other areas. (Scikit-learn, 2023) The Probability Density Function (PDF) for the whole mixture is given by:

Probability Density Function:

$$p(x \mid \Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

where:

- π_k is the weight of the k-th Gaussian component.
- $N(x|\mu_k, \sum k)$ is the multivariate Gaussian distribution with mean μk and covariance $\sum k$

Algorithm:

1. Step 1: Randomly initialize the parameters of the Gaussian distributions (standing for the means, covariances, and mixing coefficients) or use methods like K-means.
2. Step 2: The probability of each data point belonging to each Gaussian distribution, given that it is derived from the parameters, can thus be calculated using the current estimates of the parameters.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

3. Step 3: Revise the parameters of the Gaussian distributions based on the probabilities computed in the expectation stage.

$$\mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k)\right)$$

4. Step 4: Repeating the Expectation and Maximization steps till convergence is reached, i.e., the changes in parameters are below a specified threshold or the maximum number of iterations is achieved.

Benefits of using GMMs

- Probably the most important advantage associated with GMMs is that they work effectively in carrying out clustering tasks through the estimation of cluster means, covariance matrices, and weights. These capabilities can model complicated distributions of data; hence, GMMs can be applied to problems related to pattern recognition, image processing, and data clustering. (Ahmed, 2024).
- In addition, GMMs are very attractive because of their flexibility in terms of model structures and their unique ability to incorporate uncertainties in data, enabling models to best reflect real-world situations. They build the probability of data points from a number of clusters, and hence they can model complex structures with different certainty levels in the data, as Nimy (2023) is saying. This feature has been critical to the wide use of GMMs in decision-making processes, where uncertainty is the most important thing to understand.

Challenges of using GMMs

- The biggest obstacle is that GMMs are noise-sensitive and thus would affect segmentation accuracy (Ren & Hu 2020). The presence of noisy data in the customer data can lead to poor segmentation, thus limiting the recognition of the desired customer segments and reducing the effectiveness of the targeted marketing scenarios.
- GMMs depend on the assumption that the data within clusters are distributed according to the Gaussian distribution; if this one is incorrect, then the usage of GMMs will be less powerful. (Murphy, 2012).
- In addition, the necessity to specify the count of Gaussian units in the model is an issue in customer segmentation tasks (Khansari-Zadeh & Billard, 2010). Proper component selection can be a serious problem for GMMs, especially when the customer data exhibits intricate patterns and varies in structure.
- Moreover, customer clustering problems can be even more difficult in some large customer databases due to the high computational complexity of GMMs. This could be the main reason for GMMs' limited usage in real-time segmentation applications (Lee, Scott, 2012).

DBSCAN

The last algorithm is DBSCAN, which calls clusters zones of high density where only a few zones belong to outliers that are of low density. In contrast to k-means clusters, the DBSCAN

clusters may be of any shape, not just convex, and require only two parameters: ϵ-the maximum distance between points in a cluster, and MinPts-the minimum number of points needed to constitute a dense region. (Scikit-learn, 2023)

Algorithm:

1. **Step 1:** The procedure is started with a random, unvisited point selected from the dataset.
2. **Step 2:** All points within a distance ε from the selected point are located to construct the neighborhood around it.
3. **Step 3:** The point is scrutinized by checking if the number of points in its neighborhood is greater than or equal to MinPts to ascertain if it is the core point.
4. **Step 4:** In the case of the point being a core point, the new cluster is being created, and the point is being added to this cluster. If the point is not a core point and has fewer than MinPts neighbors, it is temporarily defined as noise. However, it might later be included in a cluster if it is found to be reachable from another core point.
5. **Step 5:** The cluster is extended by repeatedly adding all density-reachable points from the core points. Density-reachable points are those within the neighborhood that are both core points by themselves or, in another way, connected to core points.
6. **Step 6**: For each and every point in the neighborhood, the point is reported as visited, and its neighborhood is loaded. If it contains at least MinPts points, it is added to the cluster.
7. **Step 7:** The data processing continues with the next unvisited point in the dataset. Carry out Steps 2 to 5 in order until all points have been worked on.
8. S**tep 8:** Points that have been unallocated to any cluster after visiting all points are categorized as noise.

Benefits of using DBSCAN

- DBSCAN is effective in handling datasets with noise and outliers. It distinguishes between core points, border points, and noise, allowing it to identify and exclude points that do not fit into any cluster. Such a capability is especially needed in real-life situations where the data can have errors or be irregular, which makes it suitable for applications that require strong outlier resistance (Cao et al., 2006).
- Unlike K-means clustering, which is run based on the number of clusters fixed before running the algorithm, DBSCAN does not require prior knowledge regarding the number of clusters before processing a dataset. Instead, it identifies clusters based on the density parameters depicted by epsilon, ε, which represents the radius for search in the neighborhood, and MinPts, the minimum points, which suggests the minimum number of points required to form a cluster. In this way, cluster formation becomes automatic, and

hence the manual cluster number specification is eliminated, making the clustering process easy (Li, 2020).

- DBSCAN is particularly effective for handling massive data sets. The algorithm can effectively handle such large volumes of data by incorporating spatial indexing structures like KD-trees or R-trees to expedite the process of the neighborhood search. This effectiveness of DBSCAN in particular makes it a good fit for customer segmentation and data analysis tasks that involve large datasets, where such timely data processing and handling are essential. (Mai et al., 2014).

Challenges of using DBSCAN

However, DBSCAN itself has limitations:

- The performance of DBSCAN is highly sensitive to the choice of ε (epsilon) and MinPts (minimum points). It is important to correctly predict these parameters, as misprediction may result in poor clustering or characterizing noise as a cluster. According to Wang et al. (2019), the determination of ε and MinPts values is very important in obtaining the correct clustering results. When the ε value is incorrect, it can lead to too many or too few segments, which in turn causes the wrong segmentation of customers.
- DBSCAN can struggle with clusters that have varying densities. Its density-based method is very successful for clusters with similar densities, but it may not be that effective if the clusters are of different densities. This can compromise the accuracy of the algorithm when there is a high density of data (Zhang et al., 2017).
- Its computational complexity, especially because of the pairwise distance calculations, can be an issue with large datasets. One of the main problems with big datasets may arise from the computational complexity related to distance computation on multiple dimensions. One of the biggest drawbacks of this method, where distance is computed predominantly, is that it proves the algorithm to be less scalable and efficient with very large datasets (Sa'ada et al., 2019). There exists a linear relationship between the dataset size and the required number of computational resources. This can lead to issues with scalability and cause DBSCAN to not be suitable for real-time customer segmentation applications.

3.3.2 Clustering Evaluation Metrics

In order to identify which clustering algorithm is more effective compared to others, we need to see the three most-used evaluation scores, namely **the Silhouette, Davies-Bouldin Index,** and **Calinski-Harabasz Index**. Let's examine what these ratings mean and how to evaluate clustering using these techniques.

The silhouette score

The silhouette score is a measure of how similar a given object is to its cluster as opposed to other clusters, with values ranging from -1 to 1, and it contributes to the establishment of cluster compactness and separation. The higher the score, the more well-defined the clusters are. (Wang & Xu, 2019)

- +1: the object is very well aligned with its own cluster and is hardly detected in the neighboring ones.
- 0: the object is any one of the two points located around the line that splits the two clusters.
- -1: the object may have been allocated to a different cluster.

The silhouette score for a single observation s(i) is calculated as:

$$s(i) = \frac{b(i) - a(i)}{max(a(i),b(i))}$$

Where:

- $a(i)$ is expressed as the average distance between the i-th observation and all other points that are in the same cluster.
- $b(i)$ is the smallest average distance between the i-th observation and all points in the next cluster.

The comprehensive silhouette score for a dataset is computed by averaging the scores of all observations. This average silhouette score serves as a comprehensive measure of the patient's trouble with balance and posture.

- **High silhouette score**: Clusters are well-separated and internally cohesive.
- **Low silhouette score**: Clusters may overlap or are not clearly distinct.

It tells us about how well-separated and compact clusters are, with a clear scale. However, it may not handle varying cluster densities well.

Davies-Bouldin Index

The Davies-Bouldin Index (DBI) is an internal evaluation metric that is determined as the mean similarity ratio of every cluster relative to its most similar cluster, including the dispersion within the cluster and the distance between clusters. The objective is to reduce the DBI, which would mean that the clusters are well-separated and compact. (Davies and Bouldin, 2023)

The Davies-Bouldin index for a given clustering solution is the following:

$$DBI = \frac{1}{K}\sum_{i=1}^{K} max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

where:

- K is the number of clusters.
- σ_i is the average distance between all points in cluster i and the centroid of cluster j.
- σ_j is the average distance between all points in cluster j and the centroid of cluster j.
- $d(c_i, c_j)$ denotes the distance between the centroids of clusters i and j.

Intra-cluster Dispersion (σ_i):

$$\sigma_i = \frac{1}{|C_i|}\sum_{x \in C_i} \left|\left| x - c_i \right|\right|$$

where C_i denotes the set of points in cluster i, x is a data point in cluster i, and c_i is the centroid of cluster i.

Inter-cluster Distance $(d(c_i, c_j))$:

$$d(c_i, c_j) = \left|\left| c_i - c_j \right|\right|$$

where c_i and c_j are the centroids of clusters i and j, respectively.

The smaller the value, the more compact and well-separated the clusters are from one another. The DBI values show whether clusters are either not that separated or are dispersed more; thus, it would be a less good clustering solution.

Calinski-Harabasz Index

The Calinski-Harabasz Index (CHI), also called the Variance Ratio Criterion, compares the total between-cluster dispersions to the within-cluster dispersions for all clusters as well. It equalizes the ratio of the sum of between-cluster dispersions to the sum of within-cluster dispersions. Higher values of CHI show that these clusters are better defined and well separated. (Wang & Xu, 2019).

$$CHI = \frac{Tr(B_k) / (k-1)}{Tr(W_k) / (N-k)}$$

Where:

- $Tr(B_k)$ denotes the trace of the between-group dispersion matrix..
- $Tr(W_k)$ depicts the trace of the within-group dispersion matrix
- N is the number of samples in total..
- k is the number of clusters.

Between-group Dispersion Matrix (B_k):

To measure, B_k let C_i be the points in cluster i, c_i let be the centroid of cluster i, and finally c be the overall centroid of the dataset. Given this, the between-group dispersion matrix B_k is defined as:

$$B_k = \sum_{i=1}^{k} n_i (c_i - c)(c_i - c)^T$$

where n_i is the number of points in cluster i, ci is the centroid of cluster i, and c is the overall centroid of the dataset. Since the trace of B_k (noted as $Tr(B_k)$) is equal to the sum of squares of distances between each cluster centroid and the overall centroid, weighted by the number of points in each cluster,.

Within-group Dispersion Matrix (W_k):

To calculate, W_k let C_i be the set of all points in cluster i, and c_i be the centroid of cluster i. The within-group dispersion matrix (W_k) is defined as:

$$W_k = \sum_{i=1}^{k} \sum_{x \in C_i} (x - c_i)\,(x - c_i)^T$$

where x is a point in cluster i and c_i is the centroid of cluster i. The trace of W_k (denoted as Tr(W_k)) denotes the sum of the squared distances between each point and its centroid.

The higher the CHI values, the better the clustering performance. The result is that the clusters are best described as highly separated with great distribution between the clusters and very compact with low deviation within the clusters. In contrast, low CHI values indicate that clusters are overlapping or too dispersed, which signifies an inefficient clustering of the data.

3.3.3 Churn Prediction Models

After segmenting clients into different clusters based on common traits such as behavior, demographics, or other characteristics, businesses can adjust their strategy to each segment. As a result, segmenting provides a method for predicting churn that can be applied to segmented groups, allowing telecom companies to better deploy their resources. High-risk segments can benefit from more rigorous retention efforts, while low-risk segments can be managed using regular engagement strategies.

In this study, to predict churn, we experimented with several models. Our objective was to predict the likelihood of churn for each customer segment accurately by using models such as **Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVM,** and **KNN.**

The thesis uses the churn prediction models at two different stages: before segmentation and after. We first implemented six models and compared the results using the metrics, then did it for each cluster to see which one made the most suitable and logical result. Not to forget the "no free lunch theorem," which reminds us that there is no best overall approach and that efficiency depends on the context.

Logistic Regression

One of the appropriate models for regression analysis in binary classification, such as churn prediction, is logistic regression, which expresses the relationship between a target variable and

independent variables. (Lampinen, 2022). A logistic regression model predicts the probability that the input point will belong to one class. The formula is:

$$P(y = 1|x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

- $P(y = 1|x)$ represents the likelihood of a customer churn based on the predictor variables. x_1, x_2 ⋯, $x_{n.}$
- $\beta_0$is the intercept term.
- β_1, β_2…, β_n are the coefficients of the predictor variables.

Modeling Probability : A logistic regression model is a function that maps the input X to a class y that is either 1 or 0. A logistic or sigmoid function is utilized to link a real number to a value between the intervals 0 and 1. The probability of the positive class is the model output that can be interpreted as.

$$P(y = 1 \mid X) = \sigma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

Where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

Training the Model: The coefficients in the logistic regression model are estimated by maximizing the likelihood function. The logistic regression model computes its coefficients by making as large as possible. Usually, this consists of solving for a set of parameters that maximize the chance of observing the data. This most likely will be done through an iterative optimization method, such as gradient descent.

Making Predictions: After a model has been trained, it calculates the likelihood that a new data point will belong to the positive class, or the probability for the positive class. The class of the provided data point is then determined by comparing its probability with a predefined decision value, often 0.5, for one of the two classes.

Algorithm:

1. **Step 1: Model Initialization:** Logistic regression initializes a model with some initial parameters, weights, and bias.
2. **Step 2: Model Training:** The model iteratively gets optimized using a gradient descent algorithm, mostly to minimize the cost function. In the end, it will return the optimal parameters that have a minimum cost function.
3. **Step 3: Logit Calculation:** The logit model, logit(z), will form a linear combination of features with their respective weights for every observation.

4. **Step 4: Probability Prediction:** This logit z is passed through the sigmoid function to return the probability p of the event being 1:

$$p = \frac{1}{1+e^{-z}}$$

5. **Step 5: Binary Classification:** The calculated probability is compared to a threshold value, usually 0.5. If the probability is greater than or equal to the threshold, the outcome is 1; otherwise, it is 0.
6. **Step 6: Model Evaluation:** The model will be evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve.
7. **Step 7: Model Tuning:** This is tuning the hyper-parameters, such as the strength of regularization, that help optimize the model's performance.
8. **Step 8: Model Deployment:** Finally, the last step would be to deploy the final model for prediction over new, unseen data.

Benefits of Using Logistic Regression

- One key advantage is the interpretability of the results provided by Logistic Regression models (Wei, 2024).
- Nevertheless, Logistic Regression is highly recommended because of its simplicity and ease of implementation, making it a practical choice for churn prediction in the telecom sector (Jain et al., 2020). Logistic Regression, whose algorithms are rather shallow and transparent, might be a way to come up with a quick model that can be made and deployed when this information must be available very quickly in case a certain decision is to be taken quickly by business.
- Additionally, Logistic Regression is suitable for binary classification tasks like churn prediction, where the main point is to see if a customer will churn or not (Perišić & Pahor, 2020). The probabilistic nature of Logistic Regression enables the estimation of the likelihood of churn for each customer and provides information that can be optimally used for the prioritization of retention efforts and the allocation of resources.

Challenges of Using Logistic Regression

However, logistic regression also suffers from limitations:

- Predicting the customer's propensity for churn is usually universally accepted by means of Logistic Regression which nonetheless presents some problems with the implementation of the model that is discussed in this paper. One major problem is the linearity relationship assumption between the independent variables and the log-odds of the dependent variable. (Tianyuan et al., 2022).

If the relationships between determinants and churn are nonlinear, then Logistic Regression may not be that effective in such cases because it can't probably capture this nonlinear information flow, hence losing some value in churn prediction.

- Another problem is that the model may suffer from overfitting, especially when dealing with high-dimensional data or predictor variables that are numerous (Cui et al., 2022).
- Further, Logistic Regression can see that it underperforms on the imbalanced datasets, which are the usual in churn prediction tasks. (Mir et al., 2021). The imbalanced data, where the number of churners is way lower than that of non-churners, can lead to biased models that are more likely to predict the class of the majority of the time in their favor. The imbalance of classes in Logistic Regression is an important factor to consider in order to provide fair and accurate churn predictions.

Decision trees

Decision trees make no assumptions regarding the nature of the probability distributions satisfied by the class or other attributes when building the classification models. Robustness to noise and its interpretability nature, inherent in decision trees, imply that it can be used in predicting churn.

The idea of recursive partitioning basically brings into creation the decision tree models. The value of a certain attribute splits the data inside the nodes of the tree. The choice should then maximize information gain in every case, or rather minimize the impurity according to Gini impurity or entropy, for example.

For a group of items, the Gini impurity is defined as:

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$$

Where:

- D is a dataset.
- p_k is the proportion of items labeled with class k in the dataset D.
- K is the number of classes.

The algorithm iteratively applies this process to the subsets until the stopping criteria are satisfied, choosing the characteristic and value that yield the largest reduction in Gini impurity.

Benefits of Using Decision Trees

- Decision Trees provide a clear and non-technical illustration of a process for making a decision. In fact, the factors behind churn prediction in Decision Trees' tree-like

arrangement is brought out clearly and simply, which telecom companies can interpret and act on relative to the results (Shabankareh et al. 2021).

- Furthermore, they can capture the nonlinear relationships between predictors and churn—flexibility that linear methods, like Logistic Regression, might not model very well. This flexibility permits Decision Trees to handle complex patterns in customer behavior and thus results in more accurate predictiveness of churn (Höppner et al., 2020).
- Decision Trees can segment customers into different groups with ease based on their characteristics. And segmentation through characteristics puts effective retention strategies for specific customer segments within the easy reach of a telecom company. Thus, improvement of effectiveness of the efforts for preventing churn.

Challenges of using Decision Trees

However, some poor properties are associated with decision trees:

- They overfit on the training data, especially when grown too deep; hence, they generalize poorly on unseen data. (Karapinar et al., 2016)
- Ullah et al. (2019) argue that the overfitting problem can be handled by tree pruning and the use of ensemble methods like Random Forests.

Random Forest

Random forests merge the predictions of many decision trees to minimize overfitting and ensure improved resilience. Tree-based methods emphasize the significance of features. This will assist in detecting the main causes of churn. (Lampinen, 2022)

A single random forest consists of multiple decision trees, each of which is trained on a different subset of the training data. Each subset is created by bootstrapping, and then trees are selected based on the random subset of features. For the regression, the prediction is obtained by calculating the mean of the predictions given by all trees; for classification, the majority votes are used.

$$\widehat{y}_{RF} = argmax_c(\sum_{i=1}^{N} 1(\widehat{y_i} = c))$$

Where:

- c is each of the possible class labels, which are all unique.
- $1(\widehat{y_i} = c)$ is an indicator function that equals 1 if the prediction $\widehat{y_i}$ from the i-th tree is equal to class c, and 0 otherwise.
- $argmax_c$ selects the class label c that maximizes the sum of indicator functions, i.e., the class that gets the largest number of votes.

It is a technique that works on decision trees through ensemble learning. Bootstrapping aggregation is achieved using a set of decision trees. The word bootstrap aggregation itself gives a hint that the used decision tree is trained on a set of data or sample that is picked from the original data set with replacement, which means the original data set could be duplicated in some cases. Many decision trees are trained, and the final result is represented by the average of the predictions of individual trees.

Algorithm:

1. Initialization of Parameters: The number of trees N and the number of features M that will be considered at each split are defined.

2. **Creating Data Subsets:** N bootstrap samples (random samples with replacement) are formed from the primary dataset. So, any subset may include repeated instances.

3. **Training Decision Trees:** For each bootstrap sample, an unpruned decision tree is built:

1. A subset of features (M) is selected randomly at each node.
2. Only the most optimal split is selected among these features M.
3. Both steps 1 and 2 are repeated until the stopping rule is violated (minimum node size or maximum tree depth).

Prediction Aggregation:

- For regression tasks, the predictions from all trees are averaged.
- For classification tasks, the final decision is based on the majority vote from all trees.

Model Evaluation: The model's performance is evaluated using metrics such as accuracy, precision, recall, F1-score, mean squared error, etc.

Benefits of using Random Forest

- Random forest models are highly robust, and they can deal with large datasets with higher dimensionality successfully.
- They are able to capture non-linear connections between predictors and churn. Such freedom gives the model the capability to capture highly irregular patterns in customer behavior, which may not be grasped by linear models. (Zhao et al., 2021).
- They have the ability to rank the importance of features in predicting churn. In this way, telecom companies, through feature importance analysis, can identify the key drivers of churn and concentrate their retention efforts on the most important factors (Xie et al., 2009).

- It is robust to overfitting, especially compared with individual decision trees. According to Xie et al. (2009),the ensemble of Random Forest is the sum of at least half of the trees, the latter will decrease the variations and lead to better generalizations for models.
- For Random Forest, class imbalance, in particular customer churn prediction cases, can be dealt with most of the time. Thus, it is very good at class imbalance handling and offers better accuracy through the integration of sampling techniques and cost-sensitive learning. (Xie et al., 2009).

Challenges of using Random Forest

Despite its advantages, Random Forest also has limitations:

- A random forest model may become too complex when many trees are used to make up the ensemble, and thus model complexity itself might be a rather difficult thing to understand the factors that drive churn (Zhao et al., 2021).
- Training of Random Forest models may be computationally intensive, especially in the presence of large datasets or a large number of trees, and hence could increase training time and resource usage (Ahmad et al., 2019).
- Moreover, the interpretability of the Random Forest model may be at stake since they are mostly regarded as black-box models; thus, it becomes hard to comprehend the causes behind individual prediction (Qin et al., 2022).

Gradient Boosting

In this research, XGradient Boosting will be used as a powerful tool of machine learning since it can not only deal with weighted data but is also equipped with a very solid theoretical basis.. This algorithm would help, especially in situations where complex and non-linear relationships between the features are involved. These cases often occur in churn prediction models, as customer behavior patterns can often be really intricate. (Hanif, 2024). A key parameter in XGradient Boosting is the learning rate, η, which controls the contribution of each tree to the final model. This parameter scales the predictions of the individual trees and, as a result, prevents overfitting of the model so that generalization and accurate predictions are ensured.

Such flexibility should be reflected in the formula for the predicted value $\widehat{y}$ in the gradient boosting model by having a unique learning rate η_m for each tree m:

$$\widehat{y} = \sum_{m=1}^{M} \eta_m f_m(x)$$

Where :

- M is the total number of trees.
- η_m is the learning rate of m-th tree.
- $f_m(x)$ is the prediction from the m-th tree.

Every tree f_m(x) is trained to minimize the residual errors of the previous model. Furthermore, its contribution is adjusted by the related learning rate η_m. This method provides more fine-tuned controls over the boosting process to the point that, when various trees have different learning rates, those trees can outperform all the other trees in the final model.

Algorithm:

1. **Step 1: Initialization of Parameters:** Parameters may include the number of iterations, learning rate, and depth of the tree.
2. **Step 2. Model Initialization:** At the very start, a weak learner is initialized to predict the mean of the target variable for regression and log-odds for a classification problem.
3. **Step 3: Iterative Boosting Process:** At each iteration:
    1. Residuals are calculated as the difference between the predictions returned by the model so far and the target variables' actual values.
    2. Train a new weak learner on the residuals.
    3. Add the weak learner to the ensemble.
    4. Predictions are updated, adding the weak learner's predictions with the given learning rate.

4. **Step 4: Model Evaluation:** The model's performance is measured based on appropriate metrics such as accuracy, ROC-AUC, or F1-score.

Benefits of Using Gradient Boosting

Gradient boosting techniques—Gradient Boosting, XGBoost, LightGBM, and CatBoost—are greatly endowed with an advantage in the prediction of customers' churn due to their superior performance in metrics.

- Gradient Boosting Machine (GBM) classifiers can very effectively identify the potential churners, which are especially common in imbalanced datasets that are frequently encountered in churn prediction scenarios.
- Extreme Gradient Boosting has a higher level of accuracy, specificity, and sensitivity when compared to traditional algorithms like logistic regression. (Iqbal Hanif, 2020).

Challenges of using Gradient Boosting

However, despite the strengths of these gradient-boosting methods, limitations, especially in imbalanced datasets where accuracy alone would mislead, can happen.

- Researchers are suggested to prefer alternative metrics like PR plots and ROC curves for a better evaluation of model performance (Manoj and Dharmendra, 2022).
- Moreover, even though up-sampling techniques such as SMOTE and ADASYN have given some good results, it is necessary to investigate the robustness of these algorithms more, especially with regard to the complex interdependencies among features (Imani et al., (2024), Imani, Ghaderpour, & Joudaki (2024).

Support Vector Machine (SVM)

In classification tasks, the SVM is a classifier that searches for the best possible separating hyperplane of the data points, maximizes its margin between the classes, and reduces the chances of misclassification of both the training and unseen test samples (Farquad, Ravi, & Raju, 2014) In this case, the classification task will be to predict whether a customer churns or not.

The SVM optimization problem can be defined as:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Where:

- w is the weight vector.
- b is the bias term.
- ξ_i are the slack variables for misclassification.
- C is the regularization parameter.

Benefits of using SVM

- SVMs normally perform very well on noisy data. With proper methods for choosing optimal parameters, SVM often outperforms traditional logistic regression. (Coussement & Poel, 2008)

Challenges of using SVM

There are some disadvantages to Support Vector Machines. Major concerns, however, lie in their interpretability and computational cost. Though SVM models are built with high accuracy, it are often criticized as "black box" models.

- Farquad, Ravi, and Bapi Raju (2014) further postulate that the model is based on SVM and runs in high-dimensional feature space; hence, it's hard to understand or interpret how decisions are made. Sometimes, this transparency may be very harmful, to the point where it is necessary for it to be clear how the model made the decision.
- According to Farquad, Ravi, and Bapi Raju (2014), some methods, such as prototype-based approaches and some hyper-rectangle rules, used in the extraction of interpretable rules from the SVMs have certain limitations. They may be incomplete, non-exclusive, or very computationally expensive, and they do not scale very well.
- Besides, the extraction of rules is resource-intensive and hence less pragmatic for large data sets and real-time applications.
- Simplifying SVM decision boundaries into rules can sometimes reduce the model's ability to generalize well and can hurt accuracy.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor is another technique used for predicting the instance of churn; it falls into the category of direct use of existing cases in making classification decisions without assuming any prior knowledge about data distribution. KNN classifies new cases by directly comparing them to the closest existing cases that are based on some similarity metric (Pamina et al., 2019).

In classification tasks, the KNN algorithm doesn't predict the class of a new instance x by averaging. It does so by majority voting among the k nearest neighbors. The prediction y for a new instance x is given by the following function:

$$\widehat{y} = mode\left\{y_i : i \in N_k(x)\right\}$$

Where :

- $N_k(x)$ is the set of the k nearest neighbors x.
- y_i is the class label of the i-th nearest neighbor.

Although KNN is simple, it often works reasonably well with small datasets for which the distance metric, like Euclidean distance, can effectively capture the similarity between instances.

Benefits of using K-Nearest Neighbor

- KNN is easy to implement and interpret. Its simplicity also makes implementing and analyzing initial results in churn prediction tasks relatively fast (Gao et al., 2014).
- Since KNN is a nonparametric algorithm, it does not make any assumptions about the underlying distribution. It is this flexibility of KNN that empowers the modeling of complex trends in data without strong assumptions and allows it to work with many types of datasets (Koren, M., 2023).
- Most importantly, the simplicity of KNN makes it very suitable for rapid prototyping, which enables practitioners to test and iterate on their models quickly for preliminary analysis.

Challenges of using K-Nearest Neighbor

However, there are major limitations to its use in classification tasks.

- KNN overfits on high-dimensional or complex datasets. This makes the accuracy of the prediction lower because the model is too connected to the training data. (Srisuradetchai and Suksrikran, 2024)
- It may not be able to identify intricate patterns in the dataset. Numerous empirical studies reveal that significantly more sophisticated models, such as Random Forest and Gradient Boosting, outperform KNN in terms of achieving accurate and, in general, stable models (Aggarwal and Vijayakumar, 2024).
- In addition, KNN does not work well with datasets that contain many imbalanced classes. The model's efficiency could be diminished, and it will become unable to perform up to acceptable levels. Techniques such as the Synthetic Minority Oversampling Technique would need to be implemented to make this model wielded effectively (Opara et al., 2024).

3.4 Performance Evaluation Measures

Since it is previously known that various methodologies are used for churn prediction, we intend to compare the different models and determine which is optimal for both the entire training data and each segment. We will look at the results and compare the performance of several models to several known metrics such as **ROC-AUC, Precision-Recall Curve, F1-Score, Accuracy, and Classification Report** in the next chapter.

In our scenario of churn prediction with imbalanced data, the most significant metrics are **ROC-AUC and F1-score,** which will aid in accurately and robustly identifying the minority class. **The ROC-AUC** is an overall performance metric, but the **F1-score** balances the trade-offs between precision and recall, which are crucial for efficiently identifying churners.

ROC-AUC

The ROC-AUC (Receiver Operating Characteristic—Area Under Curve) gives a view of the overall performance of the model in classification and plots the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.

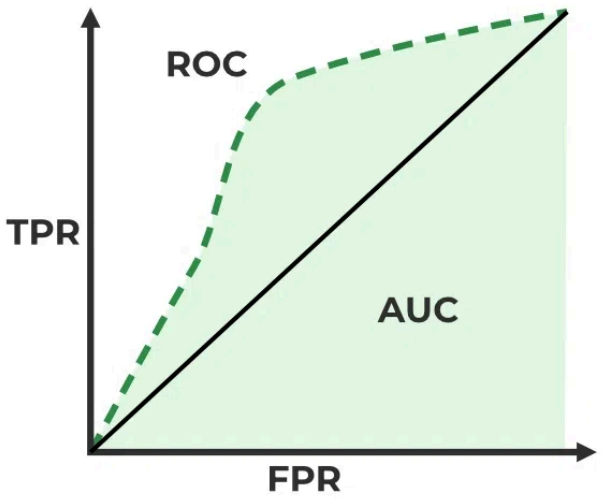


Figure 1. ROC-AUC curve *(Image taken from GeeksforGeeks (2024)).*

The ROC curve will plot TPR against FPR for different threshold values. Since TPR is equal to recall, this metric gives direct insight into the model's sensitivity—classifying the customers who will churn correctly.

The AUC (Area Under Curve) measures how easily the classes are separable; the closer to 1, the better the model, which means that it can effectively distinguish between customers who

churn and those who do not. For example, an AUC of 0.9 means a model has a 90% chance of correctly classifying a randomly chosen positive instance over a randomly chosen negative instance.

F1-Score

The F1-score is the harmonic mean of both precision and recall (where recall is the same as TPR), hence the balance between the two metrics. This would be very useful in cases where the class distribution is imbalanced.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Recall, also known as TPR, is the percentage of actual positive cases, for instance, customers that do actually churn, which the model gets right. This exemplifies the need for recall in those many cases where getting all positives is very important, such as churn prediction. The formula will return an F1-score ranging from 0 to 1; the higher, the better. The higher the F1-score, the better the performance in classifying positive instances correctly while reducing false positives and negatives.

Discussion

Of the essential evaluation metrics for model performance, the major related ones for our dataset in churn prediction are ROC-AUC and F1-score. The ROC-AUC metric is used very frequently in churn prediction because it contains information on the trade-off between a true positive rate and a false positive rate (1-specificity) (Fareniuk et al., 2022).The greater the ROC-AUC score, the better the model is at separating the two classes—churning and non-churning (Khoh, 2023). Conversely, the F1-score, which is the harmonic mean of precision and recall, is particularly useful for handling imbalanced datasets—a common occurrence in this particular scenario of predicting churn. (Shu-li et al., 2021). The F1-score considers both false positives and false negatives; hence, it is an appropriate metric when trying to evaluate model performance with respect to the identification of those customers likely to churn (Shu-li et al., 2021).

It has been shown that models optimized for better ROC-AUC and F1-score produce better precision-recall trade-offs, which means higher robustness and predictive accuracy related to identifying potential churners (Sedighimanesh, 2024). It is, therefore, emphasized that the F1-score is important in assessing models on imbalanced datasets and aligns with the churn management objective of identifying customers at risk of churning. (Shu-li et al., 2021). Accordingly, the ROC-AUC metric is preferred since it combines high accuracy and sensitivity to real situations while minimizing model specificity. This makes it a comprehensive criterion for evaluating churn prediction models (Fareniuk et al., 2022).

In this case of churn prediction, ROC-AUC and F1-score will be the most important criteria in the comparison of predictive models because they perform better at handling class imbalance. ROC-AUC gives the general performance of the model, while the F1 score presents a balanced view of precision and recall. This will help robustly and precisely identify the churning population and ensure efficient resource use in the retention of high-risk customers.

3.5 Handling Class Imbalance

We applied SMOTE (Synthetic Minority Oversampling Technique) to balance classes within the training set since the dataset is imbalanced, with fewer cases of churning as opposed to no-churn.

SMOTE thus synthesizes the minority class by forming new instances from them. The principle that was the origin of the technique was to minimize the risk of overfitting by avoiding the exclusive use of oversampling and sample duplication of minority classes. A comprehensive mathematical description of how SMOTE actually synthesizes the artificial samples is provided further below. (Jiang, Lu, & Xia, 2016)

- SMOTE generates artificial samples for the minority classes rather than replicating the existing ones. This is carried out by selecting a sample from the minority class and then tracing its k-nearest neighbors.
- Synthetic examples for each minority class sample are created on the line segments connecting any or all of the k nearest neighbors. This way, we obtain new synthetic samples that are similar but not identical to the minority-class samples.

Algorithm Steps:

1. **Step 1:** A minority class member x is randomly selected.
2. **Step 2:** The k nearest minority-class neighbors that are close to x are identified.
3. **Step 3:** One of the k-nearest neighbors, $x_{nearest}$, is chosen, and a synthetic sample, x_{new}, is generated by combining x and xnearest, as below:

$$x_{new} = x + \delta \times (x_{nearest} - x)$$

   where δ is a random number between 0 and 1.

4. This procedure is repeated until the desired number of synthetic samples has been achieved.
5. The synthetic samples were accounted for at the start of the analysis, therefore, the resulting data set was evenly divided into churn and non-churn customers in the new data.

Benefits of Using SMOTE

SMOTE is the class imbalance problem solution, which aims to oversample minority-size classes—the churned customers—later moving toward a balanced new class. As measured by most metrics, including F1-score, precision, and recall, it primarily enhances performance for the minority class. (Nurhidayat, 2023; Imani, 2023). This has the further effect of pushing class imbalance more towards the majority class and can end up in poor churn prediction. This is mitigated by SMOTE, which exposes the model to more examples of churned customers, hence making the predictions fair and balanced. (Muneer et al., 2022).

Challenges and Considerations

However, overfitting can result with synthetic samples created by SMOTE, especially if such samples do not truly represent the actual minority class distribution. (Muneer et al., 2022; Peng, 2023). Since meaningful synthetic samples are hard to generate in high-dimensional spaces, they are probably adding noise. One usually may use pre-processing methods like - Tomek links or ENN, along with other data cleaning techniques, to improve models. (Muneer et al., 2022; Arshad et al., 2022).

Balancing the training set using SMOTE improves the model's identification ability regarding churning and non-churning customers for better predictive performance.

3.6 Tools and Software

The data analysis for this research was conducted using Python, a very versatile and powerful programming language that is extremely useful when conducting data science and machine learning. Anaconda was used to manage the Python environment by installing and making all libraries required in this study compatible to ensure a smooth workflow.

- **Data Loading and Preprocessing:** The dataset was loaded and preprocessed using Pandas and NumPy. These libraries contributed to dealing with missing values, encoding categorical variables, and doing exploratory analysis preliminary to the data.
- **Data Visualization:** Matplotlib and Seaborn were used for data visualization of distributions, correlations, and clustering results. These visualization tools gave insightful graphical representations of data that aided in the interpretation of the trends and relationships within the data.

- **Machine Learning Models:** The core machine learning models, including Random Forest, Gradient Boosting, Logistic Regression, and various clustering algorithms, were implemented using the Scikit-learn library. Due to the sheer comprehensiveness of the Scikit-Learn suite of tools, it was quite efficient to realize the building, training, and assessment of these models.
- **Handling Imbalanced Data:** In handling class imbalance for this dataset, the SMOTE function was called from the Imbalanced-Learn library. This way, it ensured that models were trained on a balanced representation of classes for them to perform better in their predictions.
- **Interactive Development:** The analysis was conducted in an interactive and iterative manner using Jupyter Notebook. This provided a platform to write, test, and improve code in real-time, enabling continuous development and instant feedback.
- **Dataset Source:** The dataset used in this research project was sourced from Kaggle, one of the most well-known and trusted platforms hosting data science competitions and datasets.

Chapter 4: Data Analysis and Results

In this chapter, we will provide an in-depth analysis of the data and share our research results on customer churn prediction and segmentation within the telecommunications industry. Firstly, we will describe Exploratory Data Analysis (EDA) which was performed to understand the distribution of data and relationships. Secondly, findings from such customer segmentation and churn prediction models will be presented. Lastly, we will talk about performance metrics and the outcomes of these churn prediction models, as well as whether or not the clustering algorithm enhances the prediction outcomes.

4.1 Exploratory Data Analysis (EDA)

4.1.1 Data Distribution

First, we are going to examine a thorough examination of numerical variables, for which churn is clearly visible and correlated. Next, we will go over each feature's relationship to churn in more detail.

Customer Tenure

The distribution of customer tenure on Figure 1’s left side indicates a bimodal pattern, whereby high frequencies are at very short and very long tenures. This would seem to suggest that our customer base includes both new customers and those who have been around with the company for a decent period of time. This high frequency of short tenures points out an initial churning risk right after boarding. However, once customers reach the threshold period, they stay with you, as reflected in the peak at maximum tenure. This generally speaks to the fact that there may be a need for selective retention initiatives that would decrease early churning and result in the long-term stickiness of new customers.

The peaks at very short and very long tenures give two very important insights related to the concepts of churn risk and customer loyalty. Firstly, there is a high risk of churning, as

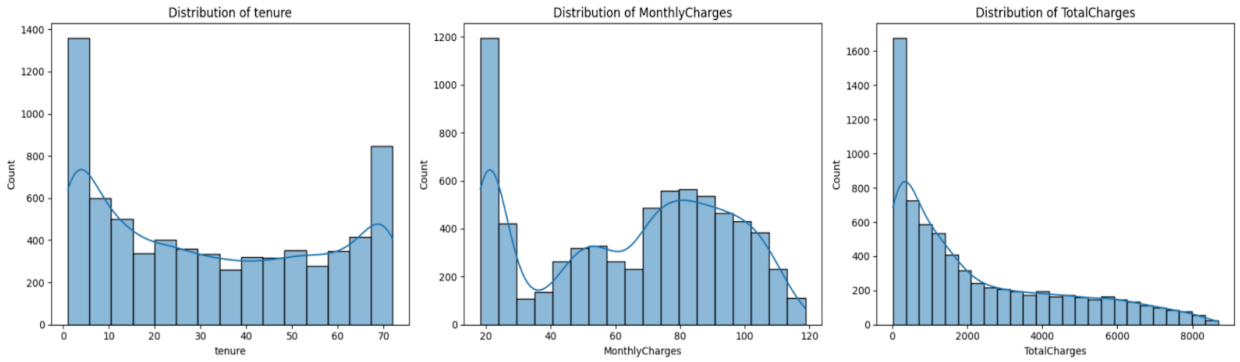
evidenced by the large number of customers with short tenures; however, once they have passed this initial period, they normally become more loyal and have longer tenures.
Thereafter, this insight drives home the need to have proper onboarding and early engagement strategies that reduce this initial churning and successfully convert new customers into long-term loyal clients.

Spend Patterns

The distribution of monthly charges displays a multimodal distribution, a clear indication of variation in the behavior of how much customers spend. There is a clear grouping of customers in the payment of different amounts, with large numbers at both ends—that is, at the lower and higher charge ends. This variation suggests a need for pricing strategies that are tailored to different segments. Those who have higher monthly charges may need more value or incentives to retain their loyalty, and lower-paying customers may be given opportunities to upsell service packages.

The right-skewed distribution for the total charge entails that most of the customers usually run a low total charge, with a long tail of higher values. This parallels the distribution of tenure since the accumulation of time is what generates the total charges. The trend strengthens the requirement of customer retention in the long run since customer retention goes directly to increased revenues.

Figure 2: Distribution of Numerical Features



Analysis of Numerical Values and Churn Rates

Tenure and Churn

The relationship between tenure and the likelihood of churning throws up some interesting patterns. As is evident from the first chart in Figure 2, customers who have very short tenures have a higher propensity to churn, a tendency that drops dramatically the longer the tenure. Clearly, the longer customers stay with the service, the less likely they are to leave.

We can draw two important inferences from the above analysis:

- **New Customers:** The heightened churning risk for customers with short tenures brings into focus the requirement for robust onboarding and early customer engagement strategies. Companies can drive better retention during these important early stages by focusing on the improvement of their early customer experience.
- **Long-Term Customers:** The customers who have been in the service for a longer time are more stable and have a lower propensity to churn. This means that once customers stay beyond a threshold, the possibility of them churning is drastically reduced.

Monthly Charges and Churn

The second graph in Figure 2 shows a decidedly non-linear relationship between churn rates and monthly charges. In particular, churn rates are relatively low for customers, with very low monthly charges. As monthly charges rise, however, the churn rate climbs to a peak at about $100 of the monthly charge.

These observations let us draw two major implications:

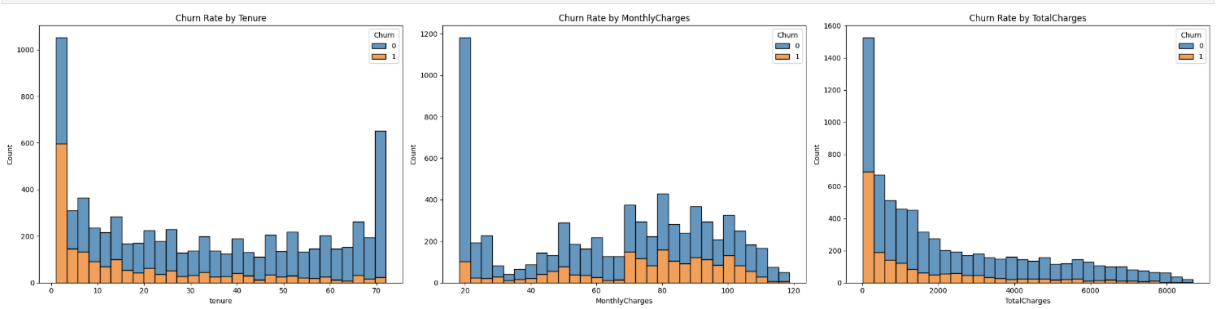
- **Higher Charges and Dissatisfaction:** Increased monthly charges can lead to customer dissatisfaction if the value they get does not meet their expectations. Customers who pay higher charges will expect premium services; when such expectations are unmet, they will seek value elsewhere.
- **Retention Strategies:** The problem can be addressed only if the telecom companies take care to provide high-value services to the higher-paying customers. In restructuring high-charge plans or revising charges to be more competitive, the company retains such customers.

Total Charges and Churn

The total charges are very strongly right-skewed, so most customers have relatively low total charges. This will be due to either a short tenure or a low monthly charge. There is a negative correlation between total charges and churn: the higher a customer's total charges, the less likely he or she is to churn.

The importance of the interrelationships between tenure, monthly charges, and total charges gives a sense of tangibility to customer behavior and churn dynamics. Clearly, according to these findings, there should be retention strategies in line with customers' tenure and pricing structures so as to ensure heightened satisfaction and reduce instances of churn.

Figure 3. Churn Rate by Tenure, Monthly Charges and Total Charges



Analysis of Categorical Values and Churn Rates

1. Contract Duration: Compared to those customers who have one-year or two-year contracts, the monthly ones indeed suffer more from this problem. The trend of these data proves that contract length does play a role in customer loyalty. It can be that the more extended contracts are more stable and have more competitive offers, thus resulting in reduced churn. Offering longer-term contracts might be an effective way to enhance customer retention.

2. Gender: If this next dimension is considered, the analysis of the churn rate by gender does not return significant differences between male and female customers. This may, therefore, imply that gender is not that strong of a predictor of churning. Other demographic factors, such as age and geographic location, would be important for further insight, especially toward the understanding of churning behavior in refining retention strategies.

3. Type of Internet Service: In this respect, the kind of Internet service used also contributes to the Churn Rate to a great extent. In particular, Fiber Optic users churn more as compared to DSL users. This could be because of higher expectations or cost sensitivity among Fiber Optic customers. On these lines, ensuring more stable services and engaging Fiber Optic customers more aggressively will be important measures toward arresting churn.

4. Value-Added Services: Some of the more intriguing churn rate dynamics are those of value-added services. Customers who subscribe to Online Security and Tech Support have lower churn rates; thus, it can be stated that such services are essential for customer retention. On the other side of the scale, the use of streaming services brings out a higher churn rate. Probably, this means that they are also technologically savvy and would be inclined to switch providers. It means that, in such a case, targeted retention strategies should be developed for high streaming usage customers, offsetting their higher propensity to churn.

5. Payment Methods: Such an analysis of payment methods will reveal that customers who are paying through electronic checks are more likely to attrite than customers who use other

payment methods. This could indicate flaws in the electronic check process that need to be fine-tuned. Streamlining the payment process may actually help in retaining customers by removing these anomalies.

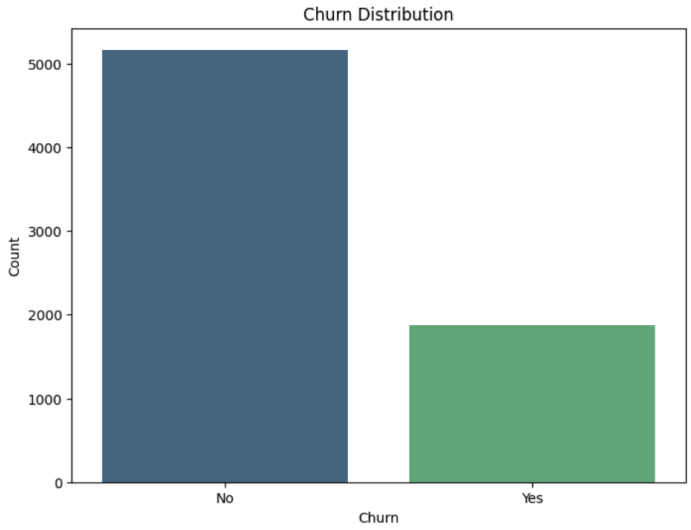
6. Contract preferences and demographic segments: Long-term contracts, especially two-year contracts, are less likely to result in customers churning. It is thus important to know which of these contract preferences different demographic segments, be it by age or income, have in order to draft in some valuable insights. This will thus ensure that the specific marketing strategies and retention methods meet their needs.

The importance of the interrelationships between tenure, monthly charges, and total charges gives a sense of tangibility to customer behavior and churn dynamics. Clearly, according to these findings, there should be retention strategies in line with customers' tenure and pricing structures so as to ensure heightened satisfaction and reduce instances of churn.

Churn Distribution

The telecom industry experiences natural churn, with a lower number of churners than non-churners, as the chart makes evident. About 26% of all customers have churned, indicating an imbalance in class in this churn distribution. Because of this disparity in class, effective predictive models need to use various techniques, such as SMOTE.

Figure 4: Churn Distribution (churn/non-churn)



The imbalance in churn also presumes the need for focused intervention strategies to retain the current clientele of the company.

Correlation Analysis

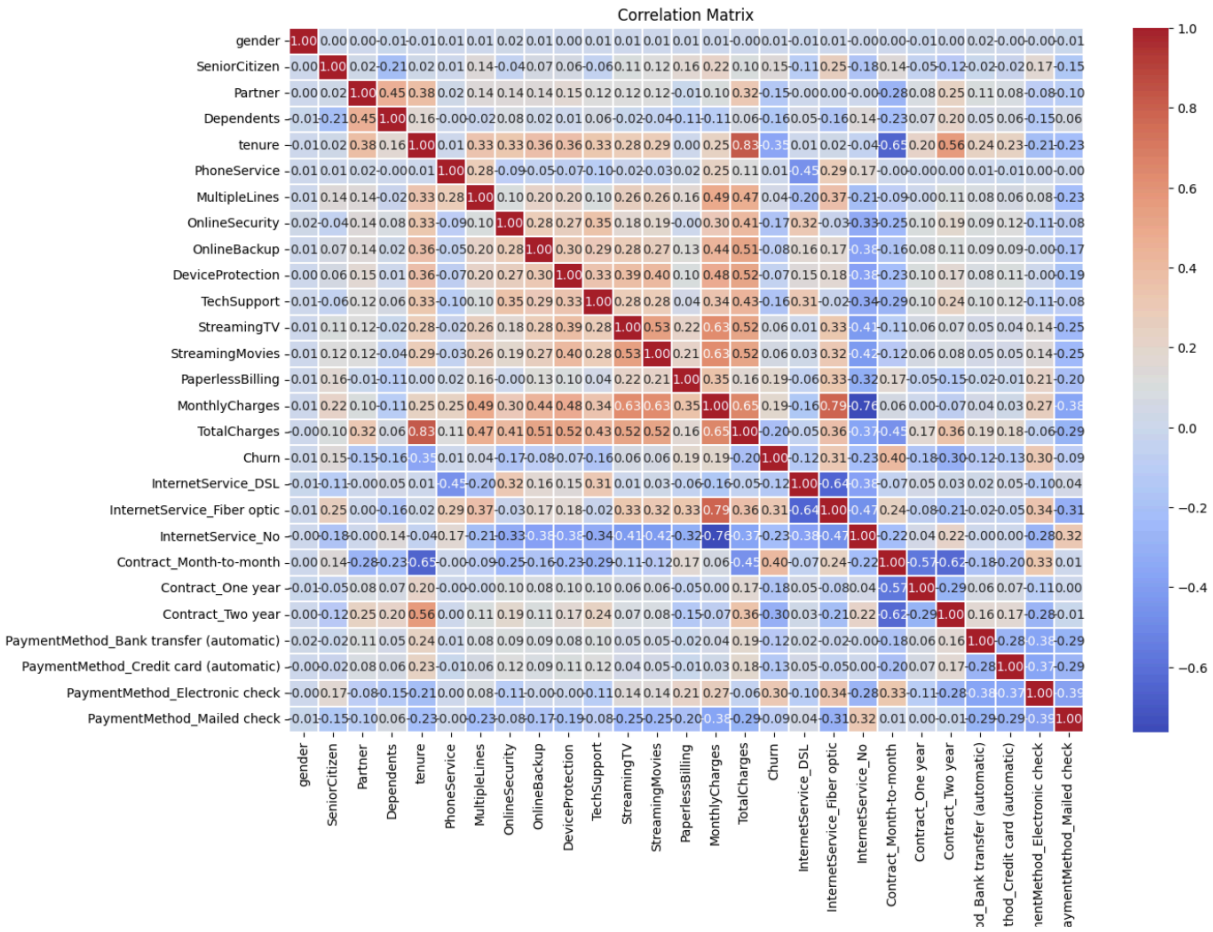
Correlation analysis helps in understanding customer churn and making predictions of the same by giving important insights into customer behavior and service attributes. Having understood deeply, we looked at each variable individually and had a discussion based on the correlation findings in Figure 4. This is because, other than numerical values, we also consider categorical variables that may have an impact.

In the process of understanding and predicting customer churn, correlation analysis reveals key insights into customer behavior and service attributes. Long-term contracts and tenure are highly negative predictors of churn, therefore proving that the more stable and long-term the engagement is, the less likely customers want to leave. In contrast, higher monthly charges, month-to-month contracts, and some payment methods like electronic checks are positively correlated with churn, thus showing the areas where customers might feel more burdened or dissatisfied.

Monthly charges are influenced by the type of internet service and additional services like streaming, while total charges naturally correlate strongly with tenure. The type of contract also indicates stability for long-term customers.

These findings underline the need for attractive long-term contracts and service cost management to arrest this propensity to churn. In improved customer engagement on flexible payment terms and month-to-month contracts lies the potential for retaining high-risk segments.

Figure 5. Correlation Matrix



4.2 Clustering Analysis of Customer Segments

4.2.1 Clustering profiles

Clustering is one of the most robust techniques for discovering different groups based on attribute similarities in a dataset. We explored the characteristics of customer segments resulting from four different clustering methods: K-Means, Hierarchical Clustering, DBSCAN, and the Gaussian Mixture Model (GMM). Each of them will give another view, and specific applications may be better suited to some of the methods than others. The results provide valuable insights into customer behavior and preferences, guiding targeted strategies for service offerings, retention efforts, and marketing campaigns.

According to the results of different clustering techniques, the cluster profiles for each model are considered below.

Table 2. Clustering profiles

Cluster Method	Cluster	General Characteristics	Services	Internet Service	Contract Type	Churn Rate
KMeans	**0**	Mix of genders, higher tenure, high monthly and total charges	Phone service, multiple lines, streaming services	Primarily fiber optic	Many on one or two-year contracts	Low (15.5%)

KMeans	**1**	More seniors and single individuals, low tenure, low charges	Few additional services like security or backup	Mostly no internet service	Mostly month-to-month contracts	High (24.8%)
KMeans	**2**	Mix of genders, medium tenure, low monthly charges but high total charges	Some have phone service, few additional services	Mostly DSL	Many on two-year contracts	Very low (5%)
KMeans	**3**	Mix of genders, lower tenure, medium monthly charges	Phone service, some streaming services	Primarily fiber optic	Many on month-to-month contracts	Very high (48.3%)
Hierarchical	**0**	Mostly non-seniors, moderate tenure, very low monthly charges	Few additional services	Mostly no internet service	Mix of contracts	Low (11.9%)

Hierarchical	**1**	Mix of seniors and non-seniors, higher tenure, moderate monthly charges	Many additional services	Mix of DSL and fiber optic	Many on month-to-month contracts	Moderate (22.9%)
Hierarchical	**2**	Mix of genders, very low tenure, moderate monthly charges	Few additional services	Mostly fiber optic	Mostly month-to-month contracts	Very high (50.2%)
Hierarchical	**3**	Mostly non-seniors, very high tenure, high monthly charges	Many additional services	Mostly fiber optic	Many on two-year contracts	Moderate (16.5%)
DBSCAN	**0**	Mix of genders, moderate tenure, moderate monthly charges	Mix of additional services	Mix of DSL, fiber optic, and no internet	Mix of contracts	Moderate (26.6%)
GMM	**0**	Mix of genders, high tenure, high monthly and total charges	Many additional services	Mostly fiber optic	Mix of contracts	Moderate (17.7%)

GMM	**1**	Mostly non-seniors, very low tenure, moderate monthly charges	Few additional services	Mix of DSL and fiber optic	Mostly month-to-month contracts	Very high (62%)
GMM	**2**	Mostly non-seniors, low tenure, very low monthly charges	Few additional services	Mostly no internet service	Mix of contracts	Very low (5.3%)
GMM	**3**	Mix of seniors and non-seniors, low tenure, moderate monthly charges	Some additional services	Mostly fiber optic	Mostly month-to-month contracts	High (39.8%)

Below, we will examine the characteristics of clusters and talk about the outcomes of each clustering technique independently.

K-Means Clustering

K-Means clustering is a popular partition-based method that divides a given dataset into various non-overlapping clusters so that post-clustering, within-cluster variance is minimized. K-means has identified four clusters with the following profiles:

Cluster 0: This segment has a balanced distribution of gender, a longer average tenure, and high monthly and total charges. Customers primarily use phone services, multiple lines, and streaming services. Fiber optic is the most used internet service, most customers have either one- or two-year contracts, and the churn rate sits relatively low at 15.5%.

Cluster 1: Characterized by a high senior population and single-person households with low tenure and charges. Additional services like security or backup are rarely used. The majority of customers do not have Internet Service, with many on month-to-month contracts. The churn rate is high at 24.8%.

Cluster 2: This cluster is quite gender-balanced, with medium tenure and low monthly but high total charges. A very low number of extra services are used, and DSL is the most common Internet service. Many customers have two-year contracts with only a 5% rate of churning, which is very low.

Cluster 3: Both genders are equally represented, medium tenure, low monthly charges. Phone services are used along with some streaming services. Fiber optic is the most common internet service, and there are a lot of month-to-month contracts. There is very high churning in this cluster, at 48.3%.

Hierarchical Clustering

Hierarchical clustering creates a tree of clusters and thus allows detailed views of the nested relationships among data points. Four clusters were identified:

Cluster 0: The majority are non-seniors, with an average tenure that is only moderate, yet at a very low monthly charge. Very few have additional services, and the majority do not have internet service. There is a mix of both contract types; the average rate of churn is very low at 11.9%.

Cluster 1: This segment consists of a mix of seniors and non-seniors with higher tenure and moderate monthly charges. They use many additional services and a mix of DSL and fiber optic internet. Many customers have month-to-month contracts, with a moderate churn rate of 22.9%.

Cluster 2: This group has an equal gender distribution with very low tenure and moderate monthly charges. Additional services used are very few, and the main internet service is fiber optic. Most of the customers are on month-to-month contracts with a very high rate of churning at 50.2%.

Cluster 3: A vast majority are not seniors. They have ultra-high tenure along with high monthly charges. There are many additional services used, the most common being Fiber Optic, Internet

Service. Two-year contracts are usual among customers, while the churning rate is moderate at 16.5%.

DBSCAN Clustering

DBSCAN is one of the algorithms for density-based spatial clustering of applications with noise. It identifies the cluster based on the density of data points, allowing the discovery of arbitrarily shaped clusters.

Cluster 0: Identifies a mixed gender, medium tenure with medium monthly charges. There is a wide variety of extra services used by customers, coupled with mixed internet services. The contract type varies with a medium churning rate of 26.6%.

On running the DBSCAN algorithm, only one cluster was returned; hence, it means either the points are spread uniformly or the chosen parameters, epsilon, and min_samples, are not bringing out the underlying density variation very well.

Gaussian Mixture Model (GMM) Clustering

A generative probabilistic model views the data as generated from a mixture of several underlying Gaussian distributions. The result provides the results of cluster membership in terms of probabilities. Four clusters were identified in this analysis:

Cluster 0: It has a balanced gender distribution, high tenure, and high monthly and total charges. Customers use many additional services, and the vast majority have a fiber-optic internet connection. There is a mix in terms of contract type and a moderate churn rate of 17.7%.

Cluster 1: Mostly non-seniors with very low tenure and moderate monthly charges. A few additional services are used, with a mix of DSL and fiber optic internet. Month-to-month contracts are common, and the churn rate is very high at 62%.

Cluster 2: Primarily non-seniors with low tenure and very low monthly charges. Additional services are rarely used, and most customers do not have internet service. Contract types vary, and the churn rate is very low at 5.3%.

Cluster 3: A mix of seniors and non-seniors with low tenure and moderate monthly charges. Some additional services are used, and fiber optic is the most common internet service. Month-to-month contracts are prevalent, with a high churn rate of 39.8%.

Churn Rates

Based on the analysis, it has been determined that in all methods, there are high churning rates found in some clusters: K-Means Cluster 3 with 48.3%, Hierarchical Cluster 2 with 50.2%, and GMM Cluster 1 with 62%. Targeted retention strategies may be formulated for such clusters to reduce the likelihood of customer loss. For example, loyalty programs, improved customer service, or targeted promotions could help in customer retention in these high-churn segments.

Service Usage

Clusters with high monthly and total charges often exhibit extensive use of additional services. For example, K-Means Cluster 0 and GMM Cluster 0 show high engagement with services and higher charges. On the contrary, these clusters, together with Hierarchical Cluster 2, such as K-Means Cluster 1 and GMM Cluster 1, have low service usage and charges. This may indicate that customers of high-charge clusters are more active and possibly better pleased with the value-added services, while those of low-charge clusters require more motivation to use the services.

Internet Service Preferences

Fiber Optic internet service is most commonly preferred in clusters with higher charges, like K-Means Cluster 0 and Hierarchical Cluster 3. In contrast, some clusters, especially those with low charges, mix internet services or do not have any internet service at all. This may suggest that high-value customers have a preference for high-speed internet services, and marketing efforts for premium services should be focused on these segments.

Contract Types

Month-to-month contracts dominate the clusters that have high rates of customer churning, like K-Means Cluster 3 and Hierarchical Cluster 2. Long-term contracts, like two-year contracts, are

found more in clusters with lower churn rates. This could be indicative of some level of customer loyalty. This basically means that the promotion of long-term contracts can become an effective means of lowering the level of churning through discounts or any other benefits offered in return for long-term commitments.

Demographic Influences

The ratios of senior citizens and dependents in various clusters vary greatly among clusters, hence affecting both service usage and churn rates. For instance, high-churn clusters often have a greater proportion of seniors or non-seniors with different service needs. It is these demographic influences that could help in fine-tuning services and communication strategies so that the different needs of the many segments of customers can be met more effectively.

4.3 Evaluation of Clustering Methods

The evaluation of clustering performance across different methods—KMeans, Hierarchical Clustering, GMM (Gaussian Mixture Model), and DBSCAN—uses several metrics, including the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. These metrics help assess the quality of clustering by measuring cluster separation, cohesion, and density. These metrics are summarized for each of the clustering algorithms in Table 2.

Table 3: Clustering Comparison Techniques

Clustering Algorithm	**Silhouette Score** *Indicates how similar an object is to its own cluster compared to other clusters.*	**Davies-Bouldin Index** *Measures cluster separation.*	**Calinski-Harabasz Index** *Evaluates cluster dispersion*
K-Means	0.472	0.707	9581.37

Hierarchical	0.386	0.859	7315.97
GMM	0.278	1.214	5066.90
DBSCAN	One cluster or noise was detected.		

Based on the metrics and profiles, the performance of the clustering algorithms can be summarized as follows:

K-Means Clustering

K-Means has been the most successful algorithm regarding general performance across almost all evaluation metrics. It returned the highest Silhouette Score of 0.472, the lowest for the Davies-Bouldin Index with 0.707, and the Calinski-Harabasz Index with the highest value of 9581.37. All these metrics indicate that the clusters K-Means is creating are very well-defined, well-separated, and dense; hence, they work most effectively on the dataset.

One has to be aware that the performance of K-Means is sensitive to the proper choice of the number of clusters, k. It has to be checked whether this k-pick is the best or not, and this might require running it with different values.

Hierarchical Clustering

Hierarchical Clustering did fairly well, but it is also slightly worse than K-Means. It scored a Silhouette Score of 0.386, a Davies-Bouldin Index of 0.859, and a Calinski-Harabasz Index of 7315.97. While these scores mean that it generally has fairly good clusters, they indicate Hierarchical Clustering to be less effective at getting really distinct and cohesive clusters.

The higher the Davies-Bouldin Index, the more overlapping the clusters are and therefore not as clearly distinct from each other compared to K-Means, which points to some weakness in cluster separation.

Gaussian Mixture Model (GMM)

GMM performs worst for the evaluation metrics. The Silhouette Score was 0.278, the Davies-Bouldin Index was 1.214, and the Calinski-Harabasz Index was 5066.90. From this evaluation, it is clearly the case that the clusters in GMM are less well separated and hence overlap a lot. This might tell a story about how GMM's assumptions for Gaussian distributions fit poorly with the structure of the dataset.

DBSCAN Clustering

DBSCAN on the dataset was not effective at all, either returning one cluster or lots of noise. This is proved by the failure to return a Silhouette Score, Davies-Bouldin Index, or Calinski-Harabasz Index. This is interpreted as a derivation that DBSCAN does not work with this dataset, and maybe because there is a problem with detecting meaningful clusters.

This provides a way of reflecting on why it occurs. First, many datasets often have a mix of numerical and categorical features. DBSCAN is not that good at mixing these appropriately since it really is designed for continuous data. Second, high dimensionality makes it very hard for DBSCAN to recognize clear clusters. Other challenges in applying this algorithm to such contexts are that the algorithm itself is sensitive to its parameters, especially 'eps' and 'min_samples' which are really hard to tune correctly for complex and high-dimensional data. If not set optimally, these could lead to cases where a single cluster is formed or many points are misclassified as noise.

Moreover, in most cases, there are undefined density differences between possible clusters in the telecom dataset, which is a requirement underlying the good functionality of DBSCAN. Hence, meaningful clusters may be totally missed, or noisy results may be produced.

The results of clustering methods

K-Means is best suited to this dataset; it returns very distinct and well-defined clusters, evidenced by its performance in all key metrics. Hierarchical Clustering did well but failed to match up to K-Means in terms of cluster separation and cohesion. GMM did not fare well for this data set because the methods rely on Gaussian assumptions, which causes less striking separation. Theoretically strong in handling noise and irregular cluster shapes, DBSCAN was nevertheless inefficient, probably due to its high sensitivity to parameters and the characteristics of this particular dataset.

In this analysis, for a given dataset, K-Means should be the recommended clustering technique. Of course, careful consideration in terms of the number of clusters is called for to ensure meaningful segmentation. Hierarchical Clustering could then be the second option, and GMM and DBSCAN would need a large amount of tuning of their parameters or maybe not even be appropriate for this dataset.

In the next subsection, we will apply churn prediction models to each cluster defined by the K-means clustering method for the identified customer segments. Specific models being set for this goal will emphasize the uniqueness of every particular customer segment. Thus, this implementation is expected to increase accuracy. The combination of segmentation and churn prediction in businesses may boost retention ability by far due to targeted, best-matched, customer-friendly, and customer-churning-reduced initiatives.

4.4 Churn Prediction Models

In our analysis, we followed best machine learning practices by including a train-test split to ensure robust model evaluation. We had a split of the dataset parts in the ratio of 80% for training and 20% for testing. Given this size of a dataset, such an approach didn't bring an overfitting problem to provide a reliable judgment on how well generalization over unseen data will be held by the model.

4.4.1 Model Training and Evaluation

Each model was trained using training data and tested using test data. As such, the class imbalance at the training phase was mitigated by applying SMOTE to ensure that both models learned from a balanced representation. Accuracy, ROC-AUC, and F1 scores were used for the test set evaluation. These metrics, therefore, had been run against the predictions that the model made on the test set to give a fairly high degree of reliability in terms of how it would work in real-life scenarios.

In this respect, the performance of the churn prediction model was supposed to be measured before and after segmentation to prove gains from segmentation. It will be possible to tell whether the clustering helps in actually differentiating between churning and non-churning cases, and if so, whether a step in the model prediction process is necessary. Creating specialized strategies for each of these distinct customer segments could help make it feasible to meet the unique needs and behaviors of both and thus realize increased customer satisfaction and reduced churning rates. This holistic model integrates the strength of segmentation and bespoke churn prediction to address the issue of retention in the very competitive sector of the telecoms industry. Model evaluation was performed within clusters.

4.4.2 Model Evaluation within Clusters

After segmenting the data into clusters, a number of classification models were trained and evaluated within each cluster. The models were considered preferred based on their certain abilities.

- **Logistic Regression:** The model is interpretable and acts as a baseline for how each feature is related to the outcome variable.
- **Decision Tree:** It works well in the case of capturing nonlinear relationships and variable interactions.
- **Random Forest:** Use decision trees to further enhance their robustness by averaging predictions across different trees.
- **Gradient Boosting:** A strong predictor is made by combining weak learners; generally, complex datasets often realize high performance.
- **Support Vector Machine (SVM):** It performs pretty well in high-dimensional spaces and mostly nonlinear decision boundaries.

- **K-Nearest Neighbors (KNN):** It is a non-parametric method for classifying an instance based on the majority of its nearest neighbors.
- **Neural Network:** It can model complicated patterns and interactions in data; hence, it very often churns out state-of-the-art performance in different tasks.
- **XGBoosting:** This model is very suitable for structured data, such as telecom customer churn. It models complicated interactions and nonlinear relationships but is robust to overfitting owing to regularization techniques.

It gives insight into how model performance may differ across different segments. This helps in understanding and assessing how models may work more effectively against others within clusters and, therefore, shows the unique characteristics of a segment.

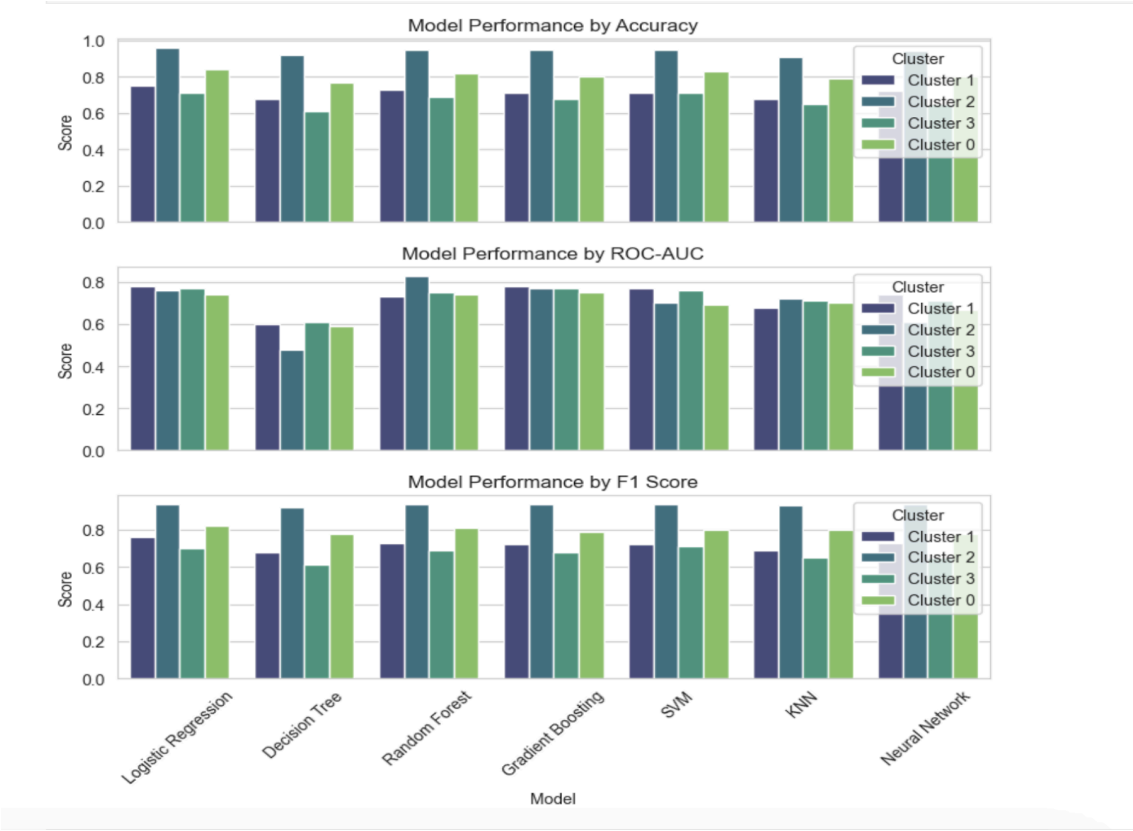
Next is the table showing the models applied to each cluster and their respective evaluation metrics.

Table 4: Churn Prediction Metrics after Segmentation

Model	Cluster	Accuracy	ROC-AUC	F1 score weighted average
Logistic Regression	Cluster 1	0.75	0.78	0.76
Decision Tree	Cluster 1	0.68	0.60	0.68
Random Forest	Cluster 1	0.73	0.73	0.73
Gradient Boosting	Cluster 1	0.71	0.78	0.72
SVM	Cluster 1	0.71	0.77	0.72
KNN	Cluster 1	0.68	0.68	0.69
Neural Network	Cluster 1	0.72	0.74	0.73
Logistic Regression	Cluster 2	0.96	0.76	0.94
Decision Tree	Cluster 2	0.92	0.48	0.92
Random Forest	Cluster 2	0.95	0.83	0.94
Gradient Boosting	Cluster 2	0.95	0.77	0.94
SVM	Cluster 2	0.95	0.70	0.94
KNN	Cluster 2	0.91	0.72	0.93
Neural Network	Cluster 2	0.94	0.61	0.94
Logistic Regression	Cluster 3	0.71	0.77	0.7

Decision Tree	Cluster 3	0.61	0.61	0.61
Random Forest	Cluster 3	0.69	0.75	0.69
Gradient Boosting	Cluster 3	0.68	0.77	0.68
SVM	Cluster 3	0.71	0.76	0.71
KNN	Cluster 3	0.65	0.71	0.65
Neural Network	Cluster 3	0.65	0.71	0.65
Logistic Regression	Cluster 0	0.84	0.74	0.82
Decision Tree	Cluster 0	0.77	0.59	0.78
Random Forest	Cluster 0	0.82	0.74	0.81
Gradient Boosting	Cluster 0	0.80	0.75	0.79
SVM	Cluster 0	0.83	0.69	0.8
KNN	Cluster 0	0.79	0.70	0.8
Neural Network	Cluster 0	0.80	0.67	0.78

Figure 6: Model Performance Comparison Bar Chart (after segmentation)



It is obvious from both Table 3 and Figure 5 that each segment presents unique characteristics, influencing the effectiveness of predictive models.

Performance Analysis of Churn Prediction Models Across Different Clusters

To see how efficient various machine learning models were at predicting the likelihood of churning, their performance was assessed across different customer clusters. The primary goal was to determine if models are most likely to contribute to successful churn intervention strategies.

Cluster 1 Analysis

Cluster 1 demonstrated diverse performance across several models with moderate complexity in churn prediction. An accuracy rate of 0.75, ROC-AUC score of 0.779, and F1 weighted score of 0.756 were all registered by Logistic Regression. This means logistic regression is a good model for predicting churn in this cluster since it handles precision and recall trade-offs well.

In contrast, Decision Tree performed less favorably, showing an accuracy rate of 0.67, ROC-AUC score of 0.575, and F1 weighted score of 0.671. All this shows that interpreting decision trees may not capture the intricacy of data as well as other models in this cluster.

The use of Random Forest and Gradient Boosting models, including XGBoost, demonstrated better results. The Random Forest model had an accuracy of 0.73, a ROC-AUC score of 0.726, and an F1 weighted score of 0.732. Similarly, Gradient Boosting reached an accuracy of 0.71, a ROC-AUC score of 0.779, and an F1 weighted score of 0.723. On competitive performance, XGBoost was very efficient and accurate, showing a high accuracy of 0.74, with ROC-AUC of 0.775 and F1 weighted of 0.735. These findings stress the strength of ensemble methods in dealing with different data patterns, thus making them viable for churn prediction within this cluster.

Support Vector Machines (SVM) performed competitively with an accuracy of 0.71, an ROC-AUC score of 0.767, and an F1 weighted score of 0.720 which implies that SVMs are useful in distinguishing churn from non-churn cases in this cluster, specifically if the data is not linearly separable.

On the other side, K-Nearest Neighbors (KNN) exhibited lower performance, with an accuracy of 0.68, a ROC-AUC score of 0.676, and an F1 weighted score of 0.690. This indicates that KNN

might not handle high-dimensional data or noisy features in this cluster. On the other hand, Neural Networks, with an accuracy of 0.71, a ROC-AUC score of 0.751, and an F1 weighted score of 0.723, performed similarly to Gradient Boosting and SVM, which makes them suitable for capturing complex data patterns.

Cluster 2 Analysis

Cluster 2 presents a high level of precision and recall for class 0 but exhibits challenges with class 1. Logistic Regression turned in an accuracy of 0.957, a ROC-AUC score of 0.760, and an F1 weighted score of 0.940. Though the performance is extremely good in general, logistic regression does poorly on class 1 predictions, which indicates a bias toward the majority class.

Even though the Decision Tree model was very accurate, 0.935, with a similarly high F1 weighted score of 0.929, it had an abnormally low ROC-AUC score of 0.486, which probably means that despite their ability to perform well in terms of accuracy and F1 score, decision trees fail to significantly tell between the classes.

Within this cluster, both Random Forest and Gradient Boosting models, including XGBoost, did quite well. Random Forest has an accuracy of 0.952, a ROC-AUC score of 0.827, an F1 weighted score of 0.943. Gradient Boosting holds an accuracy of 0.948, ROC-AUC score of 0.766, with an F1 weighted score of 0.941. On its part, XGBoost turned out to be good: accuracy—0.955, ROC-AUC score—0.825, F1 weighted score—0.950. Class balancing and capturing of complex patterns from data are very strong in these models..

One of the models returning good performance was Support Vector Machines, which had an accuracy of 0.952 with an ROC-AUC score of 0.700 and an F1 weighted score of 0.938. Its AUC score might stand on the low side in comparison with Random Forest, Gradient Boosting, and even XGBoost, but within this context, SVM also works fine. K-Nearest Neighbors has an accuracy of 0.913, with a ROC-AUC score of 0.722 and an F1-weighted score of 0.926. KNN alone seems to do quite well but does not turn out such good performances in high-dimensional spaces.

Neural networks had an accuracy of 0.944 and a ROC-AUC score of 0.598, with an F1 weighted score of 0.933. Although they work very well in general, with a lower class distinction when compared to other models, their ROC-AUC score is not as high.

Cluster 3 Analysis

Models in Cluster 3 performed much more balancedly between classes 0 and 1. Logistic Regression resulted in an accuracy of 0.705 with a resulting ROC-AUC score of 0.769, and an F1-weighted score of 0.705—sufficient performance with a balanced class distinction approach.

The performance of the Decision Tree model was relatively poor, where accuracy stood at 0.610, with a corresponding ROC-AUC score of 0.613 and an F1-weighted score of 0.610. The results obtained indicated that decision trees could not handle such complexities regarding data in this particular cluster.

Evidently, the two most compelling models were Random Forest and Gradient Boosting, along with XGBoost. For Random Forest, there was an accuracy of 0.692, a ROC-AUC score of 0.749, and a weighted F1 score of 0.692. Gradient Boosting had an accuracy of 0.681 and a ROC-AUC score of 0.769, with an F1 weighted score of 0.680. The performance for XGBoost was an accuracy of 0.700, with a ROC-AUC score of 0.755 and an F1 weighted score of 0.695. These models, however, are designed to handle the most complex patterns of data and provide a balanced approach toward predicting churn.

One of the methods that turned in good performance was SVM, with an accuracy of 0.714 and a weighted F1 score of 0.712, making it quite suitable for this cluster. K-Nearest Neighbors had poorer performance: its accuracy was 0.648, its ROC-AUC score was 0.705, and its weighted F1 measure was 0.648, hence showing difficulty in handling high-dimensional data.

Class accuracy by Neural Networks is 0.654, with a resulting ROC-AUC score of 0.707 and an F1 weighted score of 0.655. While it does capture complex patterns, the results are not significantly better compared to other models in this cluster.

Cluster 0 Analysis

Class imbalance is more prevalent in Cluster 0, and this greatly impacts the performance of the model. Logistic Regression realized an accuracy of 0.841 with a ROC-AUC score of 0.739 and an F1-weighted score of 0.822. While this model is rather good in general, its class 1 F1 score shows some difficulties in handling the minority class.

The Decision Tree model resulted in an accuracy of 0.765, a ROC-AUC score of 0.571, and an F1 weighted score of 0.771. While this accuracy and the F1 score might be considered reasonable, the low ROC-AUC is indicative of class imbalance or issues with feature interactions.

Random Forest and Gradient Boosting models, including XGBoost, performed strongly. Random Forest achieved an accuracy of 0.817, a ROC-AUC score of 0.735, and an F1 weighted score of 0.808. Gradient Boosting reached an accuracy of 0.796, a ROC-AUC score of 0.746, and an F1 weighted score of 0.790. XGBoost also demonstrated effective performance with an accuracy of 0.810, a ROC-AUC score of 0.742, and an F1 weighted score of 0.800. These models are powerful to handle class imbalances and make robust predictions related to churn.

The SVMs achieved an accuracy of 0.828, a ROC-AUC score of 0.688, and an F1 weighted score of 0.804. SVM performed well in terms of high accuracy, with quite a well-balanced F1 score. It generated a lower test ROC-AUC score than Random Forest, Gradient Boosting, and XGBoost. K-Nearest Neighbors showed an accuracy of 0.791, a ROC-AUC score of 0.697, and an F1 weighted score of 0.802, adequate in performance but poor on high-dimensional data.

Neural Networks are accurate to the tune of 0.802 and have an F1 weighted score of 0.788, but on the other hand, they scored only 0.684 in the ROC-AUC. Their performance was balanced, although they were less effective as compared to Random Forest, Gradient Boosting, and XGBoost on the basis of the ROC-AUC.

Class Imbalance

Class imbalance was one of the big problems faced, particularly in clusters with a very low count of positive samples, such as Cluster 2 and some models from Cluster 0. This could be observed by models often performing poorly on the minority class and, therefore, returning low recall and precision for the positive class. These challenges describe certain inherent complexities involved in the prediction of churning accuracy in segments with imbalanced class distributions.

SMOTE Application

To address class imbalance, balancing classes in the training data was done through the application of SMOTE: Synthetic Minority Over-sampling Technique. Despite this, the performance of the model was still constrained by the inherent difficulties associated with the prediction of rare classes.

Clustering results

In the performance of churn prediction, across all clusters analyzed, Random Forest, Gradient Boosting, and XGBoost were very prominent as a result of their flexibility to handle different data pattern and class imbalances. Logistic Regression can be helpful in some of the clusters, but it may not have good performance with class imbalance. Among all scenarios with complex data, Decision Trees and K-Nearest Neighbors performed worst. Neural Networks provide balanced performance, but no significant overfitting of ensemble methods is noted. Of these, however, XGBoost stands out in terms of accuracy and efficiency, proving to be a robust choice for churn prediction. Of these, however, there is a clear winner in XGBoost in terms of accuracy and efficiency, hence proving this to be a robust model for churn prediction. The appropriate model to be used for intervention will be determined by cluster characteristics and needs, with preference given to XGBoost and other ensemble methods due to their robustness and effectiveness. However, considering our goal to know if clustering is necessary or not in the churn prediction models, we will bring them into the same format and compare the results. Weighted averages will be used to compare each model's output with the nasal result in order to achieve that.

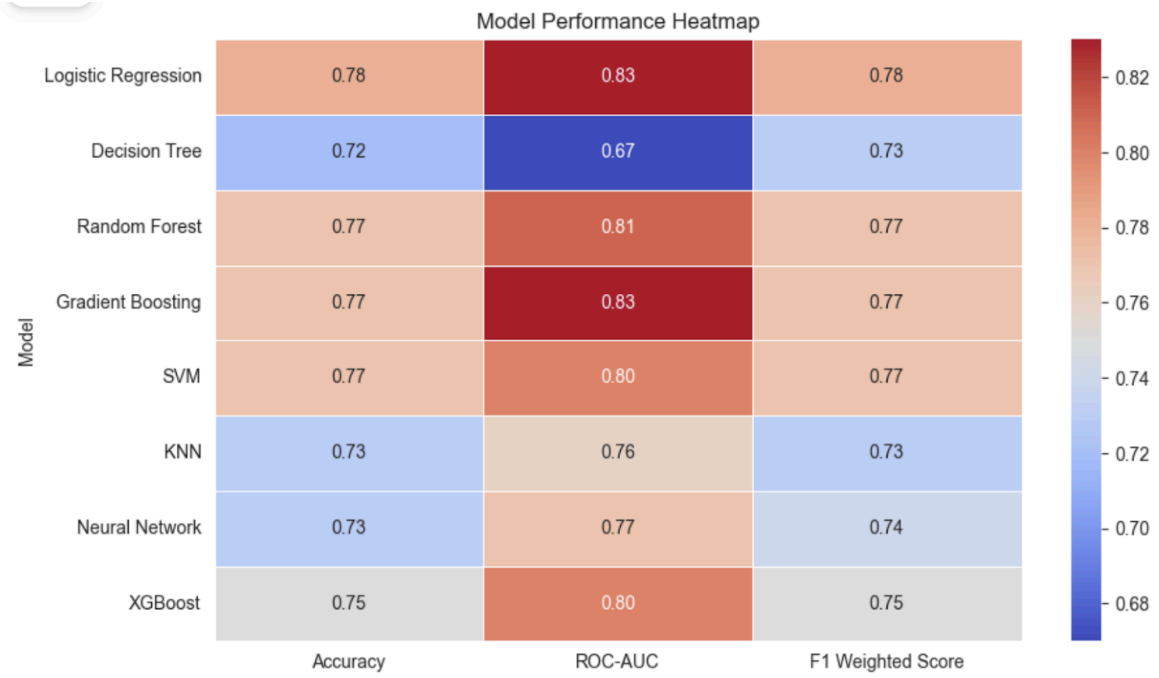
Let us now examine the actual outcome of the churn prediction metric without segmentation.

Table 5: Churn Prediction Metrics Before Segmentation:

Model	Accuracy	ROC-AUC	F1 Weighted Score	F1-Score (Class 0)	F1-Score (Class 1)
Logistic Regression	0.78	0.83	0.78	0.85	0.60
Decision Tree	0.72	0.67	0.73	0.80	0.52
Random Forest	0.77	0.81	0.77	0.84	0.56
Gradient Boosting	0.77	0.83	0.77	0.84	0.60
SVM	0.77	0.80	0.77	0.84	0.58

KNN	0.73	0.76	0.73	0.81	0.53
Neural Network	0.73	0.77	0.74	0.81	0.52
XGBoost	0.75	0.80	0.75	0.83	0.53

Figure 7. Model Performance Heatmap (before segmentation)



Overall Accuracy

From the above, it is evident that the Logistic Regression and Gradient Boosting models show the highest accuracy among the other assessed models, leading with 0.78 and 0.77, respectively. Again, it very clearly states that these two models are able to correctly classify a large part of this dataset. In contrast, the Decision Tree and Artificial Neural Network models are categorized as poor, with accuracies of 0.72 and 0.73, respectively, thus showing how ineffective the models are at making correct predictions. On the other hand, XGBoost has an accuracy of 0.75, making for a performance that is very balanced and stays very close to Gradient Boosting's, which is equally effective.

ROC-AUC

ROC-AUC scores describe how well the model is in terms of distinguishing a customer who has churned from one who hasn't. Logistic Regression and Gradient Boosting lead in this metric with

a score of 0.83, evidencing strong ability in distinguishing between the two classes. On the other hand, the Decision Tree model is the poorest, with an ROC-AUC score of 0.67, indicating that it struggles to effectively separate churned customers from non-churned ones. XGBoost also performs well, with a ROC-AUC of 0.80, comparable to other ensemble methods like Random Forest, hence further solidifying its reliability in classification tasks.

F1 Weighted Score

A further indication of just how good some of these models are is the F1 weighted score below, a measure representing the harmonic mean of precision and recall across both classes, adjusted for class imbalance. gain, logistic regression performed best, at 0.78, with gradient boosting and random forest both performing a little worse, at 0.77. XGBoost, with a respectable F1 weighted score of 0.75, indicates decent overall performance; hence, it may be useful in scenarios requiring balanced precision and recall.

Class-Specific F1-Scores

As was observed for the class-specific F1-scores, marked differences between performance on Class 0, the non-churned, and Class 1, the churned, exist. Across all models, the F1 scores for Class 0 are really very high, led by logistic regression and gradient boosting at 0.85 and 0.84, respectively. Again, this consistency demonstrates how well the models fit the majority class. By contrast, most of the F1 scores for Class 1 lie significantly lower, indicative of challenges in the fitting of the minority class of churn. Gradient Boosting and Logistic Regression both show the best performance for Class 1, with scores of 0.60, while XGBoost achieves an F1-Score of 0.83 for Class 0 and 0.53 for Class 1. This means it performed slightly better in non-churn case prediction than in churn prediction.

In a nutshell, Gradient Boosting, Random Forest, and XGBoost all perform quite similarly and well across most of the metrics. The models balance well in terms of accuracy, ROC-AUC, and F1 scores; they thus are reliable choices that can be used in any churn prediction tasks. Further, most of these models performed well on the majority class, Class 0—non-churned, compared to that of the minority class, Class 1—churned. This is evident in all models' lower F1 scores for Class 1, which may require further techniques such as cost-sensitive learning or more sophisticated sampling methods.

If interpretability or fast deployment are prior requirements in model selection, then logistic regression would turn out to be very good, considering its accuracy, ROC-AUC, and F1 scores obtained. Gradient Boosting and XGBoost should be considered when a slightly better performance is required and, most importantly, when computational resources would allow the training of more complex models.

From the viewpoint of accuracy and ROC-AUC, the decision tree model is not very good; hence, while it will not be that bad a model to be solely deployed for this problem, it still has some value during an ensemble method. Neural Network and SVM do pretty well but can't outperform simple models like Logistic Regression or ensembling methods, so they likely would not benefit from this particular dataset that much.

4.5 Churn Prediction Results Comparison

Finally, let’s discuss the results of the prediction models before and after clustering and determine if clustering is a beneficial step in the churn prediction process.

4.5.1 Comparing Clustering and Non-Clustering Results

In order to have an effective comparison of clustering and non-clustering approaches based on results from the models for churn prediction, we did a calculation for the **weighted average** of each performance metric from the different models. This creates a balanced comparison, especially where the clusters are of different sizes.

Mathematical Function for Weighted Average

The weighted average W for a metric (e.g., Accuracy, ROC-AUC, F1-Score) is calculated as

$$W = \sum_{i=1}^{n} (w_i \times M_i)$$

Where:

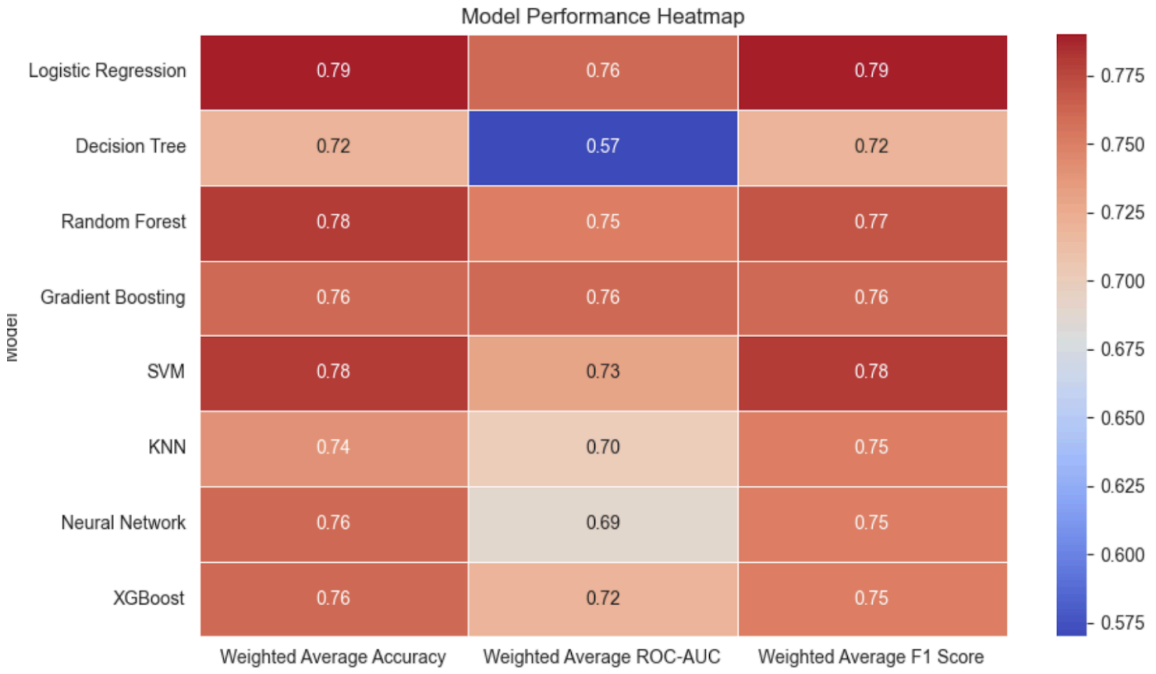
- n is the total number of clusters.
- w_i is the weight for cluster i, which is typically the proportion of observations in that cluster ($e.g., w_i = \frac{size\ of\ cluster\ i}{total\ size\ of\ all\ clusters}$)
- M_i is the performance metric (e.g., Accuracy, ROC-AUC, F1-Score) for the model on cluster i.

The metrics for each model are displayed as weighted averages in the table below.

Table 6: Churn Prediction Metrics Before Segmentation in Weighted Averages across Clusters

Model	**Weighted Average Accuracy**	**Weighted Average ROC-AUC**	**Weighted Average F1 Score**
Logistic Regression	0.79	0.76	0.79
Decision Tree	0.72	0.57	0.72
Random Forest	0.78	0.75	0.77
Gradient Boosting	0.76	0.76	0.76
SVM	0.78	0.73	0.78
KNN	0.74	0.70	0.75
Neural Network	0.76	0.69	0.75
XGBoost	0.76	0.72	0.75

Figure 8: Model Performance Heatmap for Weighted Averages (Before Segmentation)



4.5.2 Discussion and Comparison

Let us examine each metric (accuracy, ROC-AUC, and F1 weighted score) in turn, basing our discussion on the pre-and post-clustering data.

1. Accuracy:

Before Clustering: The accuracy of the models ranged from 0.72 for the Decision Tree to 0.78 for Logistic Regression. From this range, it can be seen that different models perform variedly, and Logistic Regression has the highest accuracy.

After Clustering (Weighted Average): There was a slightly better improvement in some of the models toward their accuracy, as in the case of SVM, which went from 0.77 to 0.78, and in some others, a slight worsening, such as Logistic Regression, from 0.79 to 0.78.

Thus, it suggests that clustering retained or increased accuracy across models just slightly and that again, its effect on accuracy generally was not significant, meaning that it did not significantly enhance the performance of the model's prediction accuracy.

2. ROC-AUC:

Before Clustering: ROC-AUC varied between 0.67 for the Decision Tree and 0.83 for both Logistic Regression and Gradient Boosting. For the ROC-AUC, higher values mean better separation between churners and non-churners.

After Clustering (weighted average), ROC-AUC scores showed minor fluctuations but remained relatively stable. The clustering approach did not result in significant changes to the ROC-AUC scores.

So, from the perspective of ROC-AUC scores, there was no significant advantage obtained by clustering. The models retained their discriminative power in churning versus non-churning cases, which remained the same with or without clustering.

3. F1 Weighted Score:

Before Clustering: The F1 weighted scores varied from 0.73 for the Decision Tree model to 0.78 for the Logistic Regression model. Since the F1 score is a balanced measure of how well the model is performing in classes, it will give information on how the model is performing in classes.

After Clustering (Weighted Average): The F1 scores generally remained consistent, with slight improvements observed in some models, such as SVM, where they went from 0.77 to 0.78.

The F1 scores at the end of clustering still tell us that this clustering methodology did not improve the performance of the models in handling class imbalances. Minimal improvements in some models may indicate that, while clustering does provide marginal benefits, it does not significantly improve any gains in handling imbalanced classes.

4.5.2 Evaluating the Utility of Clustering for Churn Prediction and Strategy Development

Clustering for Strategy Development:

While one of the major reasons for clustering in customer segmentation is to improve predictive accuracy, clustering here has more to do with gaining insight into distinct customer groups. This thereby enables telecommunication operators, by recognizing and interpreting these segments, to structure their marketing strategies, product offers, and quality of customer service with an increased focus on specific needs and behaviors in these different segments.

Clustering allows for targeted interventions to be carried out. For example, premium services or special discounts could be offered to high-value customers at risk of churn to retain their loyalty, while low engagement by customers means for personal engagement plans. Such strategic segmentation may lead to more effective use of resources and probably higher retention rates.

Predictive Modeling Perspective:

Results indicate that clustering does not improve the models significantly, either in terms of accuracy, ROC-AUC, or F1 scores; it even leads to a bit lower performance in some cases. It might imply that, though clustering is interesting for gaining insight into customer segments, it does not add up to the effectiveness of churn prediction models.

This adds additional complexity to the modeling process due to a clustering step. If predictive accuracy is at least as good or even a little lower, considering only the predictive task of churning, the extra work in clustering may not be warranted.

4.5.3 Conclusion: Balancing Strategy and Predictive Accuracy

Clustering can still be a valuable tool in terms of gaining strategic insight by understanding these different customer segments and developing focused strategies to reduce the likelihood of churn more effectively than a one-size-fits-all policy.

Unless it is solely to provide a prediction on churning with the best possible accuracy, clustering does not seem to add much additional value. It could even be an unnecessary complication. Companies would have to weigh the benefits of strategic insights gained against additional clustering complications in predictive modeling.

That is to say, at the strategic level of segmentation, it is recommended that clustering be applied by telecommunication firms to segment their customers with the view of developing appropriate customer retention strategies. For instance, for the question of churn prediction, directly applying models across the board to all customers can ensure the highest accuracy with the least complexity.

In these regards, companies can adopt a balanced approach: clustering as a means for the development of strategy and predictive modeling as a means for identifying customers at risk. In this way, companies are able to segment customers and design interventions targeted at these segments while at the same time predicting at-risk customers. This dual strategy ensures that not only is effective churn prediction achieved but also segment-specific activities that bring about improved customer retention and higher profitability are executed. The best results in customer retention efforts would integrate both strategic segmentation and predictive modeling. This should, however, be driven by broader business goals, considering the strategic insights and predictive performance of the models.

Chapter 5. Discussion, Intervention Strategies, Insights, and Recommendations

Customer cluster analysis and comparing the performance of ML models across these clusters can give us a clear idea of how telecom companies could frame effective strategies to reduce churn. Differences between the clusters indicate that retention efforts must be customized to maximize effectiveness. Here are the recommended strategies for each cluster:.

Cluster 1: Moderately Complex Customers

A second cluster of customers (cluster 1) is composed of those at moderate risk, meaning that a broader mix of determinants is associated with their likelihood to stay or leave. Looking into ensemble methods, particularly Random Forests, Gradient Boosting, and XGBoost, turned out to be surprisingly potent churn predictors within this cluster. These models were able to capture interesting and complex patterns in a very robust manner.

Recommended Strategy: For Cluster 1, it is advisable to employ a proactive engagement strategy. As the churn risk is moderate, companies should focus on the early identification of potential churners using the aforementioned ensemble models. More personalized-related communication and offers, like specialized discounts tailored to their usage or even complimentary service upgrades, should be integrated into the system in order for a customer's great satisfaction to turn it on as loyalty. Further, using predictive analytics to highlight the most important drivers of churn among this cluster can sharpen intervention tactics for developing retention efforts so that endeavors are targeted toward the more causal levers.

Cluster 2: High Precision, Challenging Recall

In cluster 2, it is difficult to predict class 1, even for the best classifiers (low recall), but there are also quite well-predicted non-churners in this group. This segment often encompasses those customers who look loyal but could also leave for some other, not very obvious, reason. This cluster was characterized by high XGBoost and Random Forest model performance, also suggesting a good feasibility for detecting patterns that could cause churn.

Recommended Strategy. For cluster 2, a strategy that concentrates on two aspects is recommended. Initially, companies should work on the development of more accurate predictive models to increase the recall of class 1 customers. Such a strategy would help to ensure that the customers who are at risk of churn are correctly identified. To achieve this, companies have to consider more data sources, such as customer satisfaction regarding the service in the restaurant or the way the customers utilize the service. In such a way, they would be able to identify some of the key early information to churn. Once identified, the class 1 customers should be offered extremely tailored infographic retention in the form. They can vary, beginning with informational leaflets and ending with promotions that are exclusive to these customers. The main idea is to incorporate the factors of churn the customer faces. Moreover, the company should closely monitor customers' feedback to re-develop and improve infographic offers over time**.**

Cluster 3: Balanced Churn Prediction

Cluster 3 consists of customers with more even churn behavior and less focus on extreme factors. The performance of most models, including Logistic Regression, SVM, and ensemble methods, was relatively consistent in this cluster, indicating that customer decisions seem to be distributed across a wide set of influencing factors without any particular driving factor.

Recommended Strategy: For Cluster 3, a comprehensive engagement strategy is recommended. Telcos have to look at a cross-subscriber-type customer retention strategy. This could be attested to via customer satisfaction surveys conducted every now and then for its customers, loyalty programs run by it, or guidelines issued to retain quality in each of the touchpoints with their esteemed buyers. Furthermore, businesses might consider ways to maximize cross-sell and up-sell opportunities and offer corresponding services or enhancements that meet a customer's desires and needs.

Cluster 0: High Imbalance, Specialized Intervention Needed

Cluster 0 is particularly challenging, with a high degree of class imbalance, where the proportion of non-churners is large compared to that of the churners. The models did not perform that well

in this cluster, and class 1 was poorly predicted, although ensemble methods like random forest and XGBoost still turned out to be the best performers.

Recommended Strategy: For Cluster 0, a highly targeted retention strategy is necessary. First of all, telcos should focus on ensemble models to identify the small but critical segment of customers who are at risk of churning. Given the class imbalance, this would focus efforts on high-impact interventions such as bespoke solutions or direct customer service outreach to address specific complaints or service issues. Predictive modeling efforts will have to be continually improved, possibly adding cutting-edge approaches such as anomaly detection or deep learning to capture the very weak signals of this cluster much more accurately. A feedback loop should be part of the customer response strategy to the retention effort for further fine-tuning of the predictive models and intervention tactics.

In summary, the strategic recommendations mentioned above are based on the unique characteristics of each customer cluster in view of the analysis of churn prediction models. The telecom companies could, therefore, adopt a cluster-based approach that would enable them to effectively allocate resources, enhance customer satisfaction, and ultimately reduce churn rates. For a successful implementation of these strategies, they rely very strongly on advanced predictive models involving ensemble methods such as Random Forest, Gradient Boosting, and XGBoost. Accurate identification of at-risk customers will help firms intervene on a proactive level.

Clustering Results: Conclusion

We probe to explore the constraints of traditional churn prediction models that study customers in aggregate without considering their diversity: behavior and attribute heterogeneity, variability across sales channels, and income brackets. When it comes to the telecommunications sector, an industry leveled by customer churn as a top business challenge, using one-size-fits-all methods may miss out on individual nuances at play that lead all of us across different segments to not vacate or stay. Thus, we decided to see if grouping customers into meaningful segments could increase customer insights and improve churn prevention strategies prior to running this algorithm.

Addressing the research problem

The research problem centered around the need for telecom companies to move beyond traditional churn prediction methods and instead use clustering to create targeted strategies for churn prevention. By clustering customers based on specific characteristics and behaviors, telecom companies can develop tailored retention strategies that meet the unique needs and expectations of each segment. This approach contrasts with the traditional method of applying predictive models to the entire customer base as a homogeneous group, which can lead to generalized strategies that may not effectively address the diverse reasons why customers churn.

Contribution to Existing Research

The study focuses on how telecom companies should move from old churn prediction methods to using clustering for better churn prevention strategies. By grouping customers by their traits and actions, telecom companies can make special plans to keep them happy and meet their different wants. This method is different from the usual way of using predictions for all customers as one group. This can cause simple plans that might not work well for all the reasons why customers leave.

This study adds to the current research by showing how important it is to divide customers into groups for predicting who might stop using a service. Traditional models are good, but sometimes they don't work well with many different customers. The study shows that grouping didn't always make the predictions better, but it gave us important information about how customers act that can help us make better plans to keep them from leaving. This study adds to what we already know about predicting customer behavior. It shows that customers are different, and we need different plans to keep them from leaving. This text questions the idea that one predictive model can work for all customers. This text suggests that dividing customers into groups and creating specific plans are important for understanding and changing customer actions.

Practical Applications

The results of this research have significance and potential application in the real world, specifically for the telecommunications industry. This paper shows how, with the help of

intelligent clustering techniques, telecom companies could remain associated with their customers by profiling and classifying them to understand their future behavior better. While planning and distinguishing between different customer bases, clustering does help, but it is not something that can ensure better predictions. This is where the above insights can be put to usage by practitioners to facilitate practices that prevent churning by segmenting customers into clusters. The companies can, therefore, come up with special plans that are likely to work for different groups of customers based on what they want and how they act in order to avoid customer churn. High-value customers might appreciate things like loyalty programs, while price-sensitive customers might stay if given discounts or flexible pricing.

Limitations and Future Work

A couple of drawbacks of the thesis have to be considered. Although a split train-test has been conducted, overfitting cannot be completely ruled out. Since the study based the argument on one set of data, the findings could not be said to be applicable to all phone companies or any other business. Future research can confirm that grouping and forecasting work fine based on proofing different data sets and fields to see whether these findings are solid.

This paper considers various prediction approaches and grouping methodologies, without investigating each of these algorithms or their mix-and-match possibilities. A next study in more detail could investigate the use of more complex methods, such as deep learning and mixing techniques, to get better or different results regarding how to find ways to improve the prediction and gain a better-understanding customer actions by combining different models or groups.

This thesis further establishes the problem of understanding complex models, such as Random Forest, XGBoost, and Neural Networks. These models further have high accuracy but are highly incomprehensible for business people to concretely understand and put to use. Future work would look at how to balance the capacity of the models to forecast well and be understandable, ensuring the model is so that the working people in the telecoms sector can use it.

Conclusion

To sum up, findings from this thesis revealed that although clustering does not improve the accuracy levels in churn prediction models at a consistent rate, its role in understanding the behavior of customers and consequently designing concentrated strategies to avoid churn cannot be under-estimated. This is a comprehensive, successful organizational strategy for solving customer churn problems by telecom firms. All in all, this dual method is the correct answer to strike a balance. When the company sees people as different groups that have their own desires and aspirations, they can speculate better on who would leave them. They can plan how to best retain customers accordingly. In the end, this helps make customers happier, and the company makes more money. The current paper points out the importance of combining customer segmentation with future action guessing while creating full plans to retain customers.

References:

1. Yang, J., Liu, C., Teng, M., Liao, M. & Xiong, H., 2016. Buyer targeting optimization: A unified customer segmentation perspective. In: 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 2016. IEEE, pp. 1262-1271. doi: 10.1109/BigData.2016.7840730.
2. Huda, I., Suhendra, A.A. & Bijaksana, M.A., 2023. Design of prediction model using data mining for segmentation and classification of customer churn in e-commerce mall. JOIV: International Journal on Informatics Visualization, 7(4), pp. 2280-2289. DOI: 10.62527/joiv.7.4.2414.
3. Seo, D., Choi, S., Yoo, Y. et al., 2022. Prevention of customer churn due to issuance of real-time coupons based on deep learning. PREPRINT (Version 1) [online]. Available at: https://doi.org/10.21203/rs.3.rs-1384946/v1.
4. CustomerGauge, 2024. What's the average churn rate by industry? Available at: https://customergauge.com.
5. Jain, H., Khunteta, A. & Srivastava, S., 2019. Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, pp. 101-112. Available at: https://doi.org/10.1016/j.procs.2020.03.187
6. Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G. & Chatzisavvas, K.C., 2015. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory, 55, pp. 1-9. DOI: 10.1016/j.simpat.2015.03.003
7. Kavitha, V., Kumar, G. H., Mohan Kumar, S. V., & Harish, M. (2020). Churn prediction of customers in the telecom industry using machine learning algorithms. International Journal of Engineering Research & Technology (IJERT), 9(05). http://www.ijert.org(p.g 183)
8. Manzoor, A., Qureshi, M. A., Kidney, E. M., & Tam, V. (2023). Customer Churn Prediction Using Machine Learning Techniques: A Comparative Study. Journal of Big Data, 10, 77. (pg 15-17)
9. Fujo, A. T., Subramanian, M., & Khder, M. (2022). Deep learning approach for customer churn prediction in the telecommunications industry. Journal of Applied Computing and Information Technology, 21(1). (pg 185-196)
10. Bahmani, Bahman, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii, 2012. "Scalable k-means++", Proceedings of the VLDB Endowment(7), 5:622-633. https://doi.org/10.14778/2180912.2180915
11. Bose, I. and Chen, X. (2009) 'Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn', Journal of Organizational Computing and Electronic Commerce, 19(2), pp. 133–151. doi: 10.1080/10919390902821291.
12. Dalli, M. (2022). Hyperparameter tuning for deep learning models in churn prediction: Challenges and solutions. Neural Computing and Applications, 34, 17755–17768.
13. Rodan, Ali, Ayham Fayyoumi, Hossam Faris, Jamal Alsakran, and Omar S. Al-Kadi, 2015. "Negative correlation learning for customer churn prediction: a comparison study", The Scientific World Journal(1), 2015. https://doi.org/10.1155/2015/473283
14. Praseeda, C.K., Shivakumar, B.L. Fuzzy particle swarm optimization (FPSO) based feature selection and hybrid kernel distance based possibilistic fuzzy local information

C-means (HKD-PFLICM) clustering for churn prediction in telecom industry. SN Appl. Sci. 3, 613 (2021). https://doi.org/10.1007/s42452-021-04576-7

15. Fraihat, M., Fraihat, S., Awad, M., & Alkasassbeh, M. (2022). An efficient enhanced k-means clustering algorithm for best offer prediction in telecom. International Journal of Electrical and Computer Engineering (IJECE), 12(3), 2931. https://doi.org/10.11591/ijece.v12i3.pp2931-2943
16. Rudd, D., Huo, H., & Xu, G. (2022). Improved churn causal analysis through restrained high-dimensional feature space effects in financial institutions. Human-Centric Intelligent Systems, 2(3-4), 70-80. https://doi.org/10.1007/s44230-022-00006-y
17. Qin, Z., Liu, Y., & Zhang, T. (2022). Research on early warning of customer churn based on random forest. Journal on Artificial Intelligence, 4(3), 143-154. https://doi.org/10.32604/jai.2022.031843
18. Kashwan, Kishana R., and C. M. Velu. "Customer segmentation using clustering and data mining techniques." International Journal of Computer Theory and Engineering 5.6 (2013): 856.
19. Syakur, Muhammad Ali, B K Khotimah, Eka Mala Sari Rochman and Budi dwi Satoto. “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster.” IOP Conference Series: Materials Science and Engineering 336 (2018)
20. Sam, G. (2024). Customer churn prediction using machine learning models. Journal of Engineering Research and Reports, 26(2), 181-193. https://doi.org/10.9734/jerr/2024/v26i21081
21. Larasati, A., Hajji, A. M., Handayani, A. N., Azzahra, N., Farhan, M., & Rahmawati, P. (2019). Profiling academic library patrons using k-means and x-means clustering. International Journal of Technology, 10(8), 1567. https://doi.org/10.14716/ijtech.v10i8.3440
22. Gu, C. and Qian, T. (2015). Clustering algorithm combining cpso with k-means..
23. https://doi.org/10.2991/ameii-15.2015.140
24. Zhao, H., Han, Q., & Pan, H. (2010). A hierarchical clustering algorithm based on grid partition.. https://doi.org/10.1109/mediacom.2010.46
25. Muhima, R., Kurniawan, M., & Pambudi, O. (2020). A lof k-means clustering on hotspot data. International Journal of Artificial Intelligence & Robotics (Ijair), 2(1), 29-33. https://doi.org/10.25139/ijair.v2i1.2634
26. Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. International Journal of Applied Research.
27. Wang, X. (2024). A new efficient medoid based divisive hierarchical clustering method for large-scale data.. https://doi.org/10.21203/rs.3.rs-4256002/v1
28. Mylonas, P., Wallace, M., & Kollias, S. (2004). Using k-nearest neighbor and feature selection as an improvement to hierarchical clustering., 191-200. https://doi.org/10.1007/978-3-540-24674-9_21
29. Rokach, L., & Maimon, O. (2005). Clustering methods. In Data mining and knowledge discovery handbook (pp. 321-352).
30. Ahmed, H. (2024). On expectation-maximization algorithm with split and merge in r. IJDM, 1(2), 179-190. https://doi.org/10.62054/ijdm/0102.14

31. Nimy, E. (2023). Web-based clustering application for determining and understanding student engagement levels in virtual learning environments. E-Journal of Humanities Arts and Social Sciences, 4-19. https://doi.org/10.38159/ehass.20234122
32. Ren, H. and Hu, T. (2020). An adaptive feature selection algorithm for fuzzy clustering image segmentation based on embedded neighbourhood information constraints. Sensors, 20(13), 3722. https://doi.org/10.3390/s20133722
33. Lee, Gyemin and Clayton Scott, 2012. "Em algorithms for multivariate gaussian mixture models with truncated and censored data", Computational Statistics &Amp; Data Analysis(9), 56:2816-2829. https://doi.org/10.1016/j.csda.2012.03.003
34. Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). Density-based clustering over an evolving data stream with noise.. https://doi.org/10.1137/1.9781611972764.29
35. Li, S. (2020). An improved dbscan algorithm based on the neighbor similarity and fast nearest neighbor query. Ieee Access, 8, 47468-47476. https://doi.org/10.1109/access.2020.2972034
36. Mai, S., He, X., Feng, J., Plant, C., & Böhm, C. (2014). Anytime density-based clustering of complex data. Knowledge and Information Systems, 45(2), 319-355. https://doi.org/10.1007/s10115-014-0797-0
37. Wang, C., Ji, M., Wang, J., Wen, W., Li, T., & Sun, Y. (2019). An improved dbscan method for lidar data segmentation with automatic eps estimation. Sensors, 19(1), 172. https://doi.org/10.3390/s19010172
38. Zhang, W., Zhang, X., Zhao, J., Qiang, Y., Tian, Q., & Tang, X. (2017). A segmentation method for lung nodule image sequences based on superpixels and density-based spatial clustering of applications with noise. Plos One, 12(9), e0184290. https://doi.org/10.1371/journal.pone.0184290
39. Sa'ada, N., Harsono, T., & Basuki, A. (2019). Improvement of segmentation performance for feature extraction on whirlwind cloud-based satellite image using dbscan clustering algorithm. Emitter International Journal of Engineering Technology, 7(1), 301-325. https://doi.org/10.24003/emitter.v7i1.372
40. Wang, X. and Xu, Y., 2019. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. IOP Conference Series: Materials Science and Engineering, 569(5).
41. Davies, D.L. and Bouldin, D.W., 2023. DBI: A partitioning Davies-Bouldin index for clustering evaluation. Neurocomputing, 528, pp.178-199.
42. Lampinen, S. (2022). Modeling Customer Engagement with Churn and Upgrade Prediction. University of Helsinki.
43. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167, 101-112. https://doi.org/10.1016/j.procs.2020.03.187
44. Perišić, A. and Pahor, M. (2020). Extended rfm logit model for churn prediction in the mobile gaming market. Croatian Operational Research Review, 11(2), 249-261. https://doi.org/10.17535/crorr.2020.0020
45. Tianyuan, Z., Moro, S., & Ramos, R. (2022). A data-driven approach to improve customer churn prediction based on telecom customer segmentation. Future Internet, 14(3), 94. https://doi.org/10.3390/fi14030094

46. Cui, S., Lai, P., Deng, Y., & Zheng, X. (2022). Risk decision and predicting of customer churn based on principal component analysis., 693-701. https://doi.org/10.2991/978-94-6463-005-3_71
47. Mir, S.Z., Ghani, A., Yasin, S., Aftab, A. and Khuda, I.E., 2021. Software development and modelling for churn prediction using logistic regression in telecommunication industry. International Journal of Advanced Trends in Computer Science and Engineering, 10(3), pp.1050-1057. Available at: http://www.warse.org/IJATCSE/static/pdf/file/ijatcse1501032021.pdf [Accessed 19 August 2024].
48. Sam, G. (2024). Customer churn prediction using machine learning models. Journal of Engineering Research and Reports, 26(2), 181-193. https://doi.org/10.9734/jerr/2024/v26i21081
49. Shabankareh, M., Shabankareh, M., Nazarian, A., Ranjbaran, A., & Seyyedamiri, N. (2021). A stacking-based data mining solution to customer churn prediction. Journal of Relationship Marketing, 21(2), 124-147. https://doi.org/10.1080/15332667.2021.1889743
50. Höppner, S., Stripling, E., Baesens, B., Broucke, S., & Verdonck, T. (2020). Profit driven decision trees for churn prediction. European Journal of Operational Research, 284(3), 920-933. https://doi.org/10.1016/j.ejor.2018.11.072
51. Karapinar, H., Altay, A., & Kayakutlu, G. (2016). Churn detection and prediction in automotive supply industry.. https://doi.org/10.15439/2016f245
52. Zhao, Z., Zhou, W., Qiu, Z., Li, A., & Wang, J. (2021). Research on ctrip customer churn prediction model based on random forest., 511-523. https://doi.org/10.1007/978-3-030-92632-8_48
53. Xie, Y., Li, X., Ngai, E., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. Expert Systems With Applications, 36(3), 5445-5449. https://doi.org/10.1016/j.eswa.2008.06.121
54. [18] Qin, Z., Liu, Y., & Zhang, T. (2022). Research on early warning of customer churn based on random forest. Journal on Artificial Intelligence, 4(3), 143-154. https://doi.org/10.32604/jai.2022.031843
55. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1). https://doi.org/10.1186/s40537-019-0191-6
56. [8] Hanif, I. (2024). Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction. Media & Digital Department, PT. Telkom Indonesia, Jakarta, 10110, Indonesia, 436-437.
57. Manoj, K., & Dharmendra, K. Y. (2022). Customer Churn Prediction in Telecommunication Using Gradient Boosting Machine. Doi: 10.1007/978-981-16-2597-8_66
58. Imani, M., & Arabnia, H. R. (2023). Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: a comparative analysis. https://doi.org/10.20944/preprints202308.1478.v2
59. Hanif, I. (2020). Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction. Proceedings of the Proceedings of the 1st International

Conference on Statistics and Analytics, ICSA 2019, 2-3 August 2019, Bog. https://doi.org/10.4108/eai.2-8-2019.2290338

60. Manoj, K., & Dharmendra, K. Y. (2022). Customer Churn Prediction in Telecommunication Using Gradient Boosting Machine. Doi: 10.1007/978-981-16-2597-8_66
61. Imani, M., Ghaderpour, Z., Joudaki, M., & Beikmohammadi, A. (2024). The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction. Doi: 10.20944/preprints202403.0213.v2
62. Farquad, M. A. H., Ravi, V., & Raju, S. B. (2014). Churn Prediction Using Comprehensible Support Vector Machine: An Analytical CRM Application. Institute for Development and Research in Banking Technology, Hyderabad, University of Hyderabad, Hyderabad, The University of Hong Kong.
63. Coussement, K. and Van den Poel, D., 2008. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), pp.313-327.
64. Li, J., Zhang, S., Xuan, Y. and Sun, Y., 2007. Identifying Skype traffic by random forest. 2007 International Conference on Wireless Communications, Networking and Mobile Computing, Shanghai, China, pp. 2841-2844. doi: 10.1109/WICOM.2007.705.
65. Raymond, Oetama. (2023). Unveiling churn prediction at bank ivory. JITET: Jurnal Informatika dan Teknik Elektro Terapan, Available from: 10.23960/jitet.v11i3s1.3394
66. Muteb, Alotaibi., Mohd, Anul, Haq. (2024). (1) Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods. Engineering, Technology & Applied Science Research, Available from: 10.48084/etasr.7480
67. Shradhdha, Gohil., Manas, Ameria., Vergin, Raja. (2023). (5) An intelligent Prediction of Customer Churn in Telecom Industry Using Business Analytics Classification Models. Available from: 10.1109/icoac59537.2023.10249776
68. Pamina, J., Beschi Raja, S., Soundarya, S., Sruthi, M. S., Kiruthika, S., Aiswaryadevi, V. J., & Priyanka, G. (2019). An Effective Classifier for Predicting Churn in Telecommunication. Jour of Adv Research in Dynamical & Control Systems, 11(01-Special Issue).
69. Gao, Y., Yan, P., & Pan, J. (2014). Nearest neighbor classification method based on the mutual information distance measure. In Proceedings of the 11th World Congress on Intelligent Control and Automation. https://doi.org/10.1109/wcica.2014.7053251
70. Koren, M. (2023). Weighted distance classification method based on data intelligence. Expert Systems, 41(2). https://doi.org/10.1111/exsy.13486
71. Patchanok, Srisuradetchai., Korn, Suksrikran. (2024). (2) Random kernel k-nearest neighbors regression. Frontiers in big data, Available from: 10.3389/fdata.2024.1402384
72. Pallav, Aggarwal., V., Vijayakumar. (2024). (1) Customer Churn Prediction in the Telecom Sector. Available from: 10.1109/aiiot58432.2024.10574660
73. Opara, John, Ogbonna., Gilbert, I.O., Aimufua., Muhammad, Umar, Abdullahi., Suleiman, Abubakar. (2024). (5) Churn Prediction in Telecommunication Industry: A Comparative Analysis of Boosting Algorithms. Available from: 10.4314/dujopas.v10i1b.33

74. [Fareniuk, Y., Zatonatska, T., Dluhopolskyi, O., & Kovalenko, O. (2022). Customer churn prediction model: a case of the telecommunication market. Economics, 10(2), 109-130. https://doi.org/10.2478/eoik-2022-0021
75. [49] Khoh, W. H. (2023). Predictive churn modeling for sustainable business in the telecommunication industry: optimized weighted ensemble machine learning. Sustainability, 15(11), 8631. https://doi.org/10.3390/su15118631
76. Shu-li, W., Yau, W., Ong, T., & Chong, S. (2021). Integrated churn prediction and customer segmentation framework for telco business. IEEE Access, 9, 62118-62136. https://doi.org/10.1109/access.2021.3073776
77. Sedighimanesh, M. (2024). Optimizing hyperparameters for customer churn prediction with PSO-enhanced composite deep learning techniques. https://doi.org/10.20944/preprints202403.1048.v1
78. Jiang, K., Lu, J., & Xia, K. (2016). A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE. Arabian Journal of Science and Engineering, 41, 3255–3266. https://doi.org/10.1007/s13369-016-2179-2
79. Muneer, A., Ali, R. F., Alghamdi, A., Taib, S. M., Almaghthawi, A., & Ghaleb, E. A. A. (2022). Predicting customers churning in banking industry: a machine learning approach. Indonesian Journal of Electrical Engineering and Computer Science, 26(1), 539. https://doi.org/10.11591/ijeecs.v26.i1.pp539-549
80. Nurhidayat, M. M. S. (2023). Analysis and classification of customer churn using machine learning models. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 7(6), 1253-1259. https://doi.org/10.29207/resti.v7i6.4933
81. Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. PLOS ONE, 18(12), e0289724. https://doi.org/10.1371/journal.pone.0289724
82. Arshad, S., Iqbal, K., Naz, S., Yasmin, S., & Rehman, Z. (2022). A hybrid system for customer churn prediction and retention analysis via supervised learning. Computers, Materials & Continua, 72(3), 4283-4301. https://doi.org/10.32604/cmc.2022.025442
83. GeeksforGeeks, 2024. AUC - ROC Curve. [online] Available at: https://www.geeksforgeeks.org/auc-roc-curve/
84. Farquad, M.A.H., Ravi, V. and Bapi Raju, S., 2014. Churn prediction using comprehensible support vector machine: An analytical CRM application. *Applied Soft Computing*, 19, pp.31-40. Available at: https://doi.org/10.1016/j.asoc.2014.01.031