Telecommunications Churn Prediction

Sachin Dimri (24N0083)

Guided by: Prof. Pritam Singh Negi

Hemvati Nandan Bahuguna Garhwal University (HNBGU), Srinagar

October 4, 2024

ACKNOWLEDGEMENT

I dedicate this report with heartfelt appreciation and deep respect to my mentor, **Prof. Pritam Singh Negi**, whose unwavering support, exceptional guidance, and constant encouragement have been pivotal in shaping this project. His insightful feedback and patient mentorship guided me at every critical juncture—from conceptualisation to execution— ensuring that I not only met our academic objectives but also deepened my understanding of data science and **Telecommunication Churn Prediction**. His mentorship inspired me to think beyond conventional approaches, engage deeply with the problem, and strive for both clarity and innovation in my work.

I, **Sachin Dimri**, also take this opportunity to recognize my own journey throughout this project. The countless hours of brainstorming, the sleepless nights, and the many moments of learning and resilience have made this experience truly rewarding. This project is not just an academic milestone—it is a reflection of personal dedication, perseverance, and growth. This report represents more than just an project submission—it reflects the dedicated effort, continuous learning, and the invaluable support of an exceptional mentor. It serves as a reminder that with the right guidance and collective commitment, even the most complex challenges can be trans- formed into meaningful achievements. I am truly grateful for his guidance and for the opportunity to undertake this enriching journey under his supervision.

ABSTRACT

The telecommunications industry is highly competitive, with consumer churn being a significant concern due to market saturation. Churn, or customer attrition, refers to customers leaving a service provider due to dissatisfaction or better offers from competitors. This phenomenon is particularly impactful in the telecom sector, where retaining customers is more cost-effective than acquiring new ones. To address this issue, telecom companies increasingly rely on machine learning techniques to predict customer churn and implement strategies to retain at-risk customers.

The study highlights the importance of understanding the factors contributing to customer churn, including service quality, pricing, customer service interactions, and competition. It emphasizes that technological advancements and data analytics can be crucial in identifying customers likely to churn, allowing telecom companies to take proactive measures to prevent it.

The research involved applying machine learning algorithms—such as k-Nearest Neighbors (KNN), Random Forest, Multilayer Perceptron (MLP), Gradient Boosting, and XGBoost—on a dataset from a telecommunication company. The objective was to develop a predictive model that accurately identifies customers at high risk of churning. The study also aimed to analyze customer behaviour, demographics, usage patterns, and other relevant data to determine the primary drivers of churn.

The project aims to help telecom companies reduce churn rates, enhance customer satisfaction, and improve financial stability and growth by developing a robust predictive model. This research provides valuable insights into customer retention strategies, emphasizing the need for telecom companies to focus on retaining existing customers in an increasingly competitive market.

Keywords: *Consumer churn, Machine learning, Artificial intelligence, Telecommunication, churn prediction*

Table of Contents

List of Figures

List of Tables

Abbreviation

ML	Machine Learning
FN	False Negative
FP	False Positive
TN	True negative
TP	True Positive
Acc	Accuracy
KNN	K-Nearest Neighbor
RF	Random Forest
MLP	Multilayer Perceptron
NN	Neural Naive
XGBoost	Extreme Gradient Boosting
GBCC	Gradient Boosting Classifier
Telecom	Telecommunication
ROC	Receiver Operating Characteristic Curve
AUC	Area Under Curve
kNN	k-nearest neighbors
DT	Decision Tree
CC	Consumer Churn

Chapter 1:

1.1 Introduction:

Today, communication technology is highly competitive. Churn is a common concern in practically every sector. However, it is crucial to the telecoms sector, which many say has reached saturation. Competition is at an all-time high, and many telcos recognize the need to improve consumer experience and service to reduce consumer turnover and compete more effectively, therefore, the understanding of the causes of telecom churn is necessary.

Consumer churn is a big concern in today's telecommunications industry [1]. In the telecommunications paradigm, churn is described as consumers leaving the organization and rejecting its services, owing to dissatisfaction with the services or superior offerings from other network providers within the consumer's reasonable price range. This results in a potential loss of revenue for the firm. It is also becoming increasingly difficult to keep clients. As a result, firms will employ cutting-edge applications and technology to deliver the best possible services and retain consumers. Before doing so, it is vital to identify which clients are likely to leave the company soon, as losing them would result in a significant loss of profits for the corporation. This method is referred to as Churn Prediction.

In a commercial setting, consumer attrition simply refers to switching from one company offering to another. Consumer or subscriber churn is like attrition, where consumers anonymously transfer to one another. In this competitive environment, business is getting oversaturated. The telecommunications industry faces severe issues because of the proliferation of active and competitive service suppliers. As a result, they have found it incredibly difficult to keep existing consumers. Due to the expense reason, acquiring new consumers is much higher than the expense spent on sustaining current consumers; the telecom industry must make significant efforts to retain clients to preserve their market value. Over the last decade, various data mining strategies have been developed in the literature for predicting churners using heterogeneous consumer datasets.

Consumer churn is a typical metric of client loss. Telecommunications firms frequently lose key consumers and, consequently, profits to competitors. The telecommunications business has seen considerable changes over the previous few decades, including introducing new services, technical breakthroughs, and increasing competition due to deregulation [2]. Client churn prediction in telecommunications has become vital to industry participants to safeguard their loyal client base and organization growth and improve consumer relationship management (CRM). Retaining clients with significant churn risk is one of the most difficult tasks in the telecommunications sector today [3]. Consumers now have a choice of churn alternatives due to the increased number of service providers and more fierce competition. Thus, telecommunications sector operators realize the value of maintaining existing consumers rather than obtaining new ones. Several variables cause consumers to churn. Unlike post-paid consumers, prepaid clients are not bound by service contracts and frequently leave for the simplest of reasons. Thus, predicting their turnover rate is difficult. Another factor is client loyalty, which may be impacted by the quality of consumer service and products given by vendors. Consumers may move to a competitor with a larger reach and better reception quality if they experience network coverage issues. Slow or inadequate responses to complaints and invoicing problems enhance consumers' likelihood of defecting to the competition. Packaging costs, insufficient features, and outdated technology may also lead consumers to switch to competitors. Consumers frequently compare their suppliers with others and switch to whichever delivers better overall value [4].

Technological improvements have helped businesses realize that their competitive strategies must include strong client retention rates to compete in the industry. This is especially relevant to the telecommunications business. As a result, much research is currently being conducted to identify consumers likely to defect to competitors [5]. The telecom industry's liberalization has boosted competition, exacerbated by consumers now having more options. Thus, telecommunications businesses should better understand and satisfy their consumers' demands to avoid losing them to competitors [6]. The enormous quantity of research that considers controlling client turnover a critical CRM component further emphasizes its importance. CRM necessitates a thorough understanding of the organization's consumers and markets. CRM entails understanding the consumer's performance to retain the most lucrative clients while identifying those whose turnover no longer matters. CRM also organizes the evolution of offers and discounts: which products to sell to consumers, via which media, and which products require advertising.

In this paper, the use of machine learning algorithms is applied to assess the variables influencing consumer turnover in telecommunications. The study examined user churn factors using various models (KNN, Radnome Forest, MLP, Gradient Boosting, and XGBoost). Additionally, the study carried out an evaluation of a data set from Kaggle and applied all the models to the data set. Finally, based on the results the determination process of the best-anticipated method is chosen to define consumer churn and retention factors.

1.2 Background

With the rapid growth of the telecommunications industry, service providers are keener to increase their consumer base. To satisfy the demands of survival in a competitive market, retaining existing clients has become a major concern. According to a telecom poll, the expense of acquiring a new client is substantially more than that of sustaining an existing one. As a result, acquiring information from the telecom business can help predict consumer behaviour, such as whether they would leave the company. [7]. To maintain a stable market value, the telecom industry must take the required steps to begin recruiting associated clients.

Today's most difficult and crucial concern for telecommunications is managing churning consumers. Churn Consumers define the number of existing consumers who might leave the company within a given time frame. These clients are referred to as "churners." The fundamental purpose of churning is to find churnable consumers in their first instance feasible and understand why they churn. Consumer churn in the telecom business is a significant problem that machine-learning approaches can efficiently handle. Telecom businesses may use data analytics and ML models to determine whether customers will probably churn and take aggressive measures to retain.

In current research, we have proposed a best-predicted model that evaluates the data and gives us the basic reason to know about consumer churn in the industry. We have considered a dataset of the telecommunication industry, which we have taken from Kaggle, and evaluated the data using machine learning algorithms.

1.3 Project Goals:

Consumer attrition prediction in the telecom business has been one of the most popular study areas in recent years. It is the process of identifying clients likely to discontinue their service subscription. The mobile telecommunications sector has recently transitioned from Fast expansion to saturation and intense competition. As a result, telecom corporations' priorities have shifted from increasing their client base to maintaining existing consumers [8]. As a result, it is important to identify which clients will likely migrate to a rival shortly. Data acquired from the telecom business can assist in analyzing the causes of consumer churn and using that knowledge to retain consumers.

Telecommunications is one of the most dynamic areas in the market, and the consumer base is a crucial component in generating stable income; thus, it is critical to emphasize keeping them engaged [9]. Consumers' migration from one network to another varies per telecommunications company based on call quality, price plan, minute usage, data, SMS capabilities, consumer billing concerns, etc.

This research project aims to develop a comprehensive understanding of the factors driving consumer churn in telecommunication companies and to create a predictive model using machine learning techniques. This model will enable telecom companies to identify consumers at high risk of churning, allowing them to implement targeted retention strategies. By leveraging data analytics, the project seeks to enhance client retention, decrease churn rates, and enhance overall consumer satisfaction, contributing to telecom companies' financial stability and growth. By achieving these objectives, the project aims to empower telecommunication companies with the tools and knowledge needed to effectively combat consumer churn, ultimately leading to a more loyal consumer base and sustainable business growth.

1.4 Aims and Objectives:

Develop and refine a Machine Learning model to identify the factors of consumer churn in the telecommunication industry.

Assess the model's accuracy and reliability using metrics such as precision, recall, F1-score, and ROC-AUC, ensuring its effectiveness in real-world applications.

Conduct a thorough analysis of consumer demographics, use habits, billing information, consumer service interactions, and other pertinent data used to determine the primary drivers of consumer churn.

Employ advanced machine learning algorithms, kNN, Radnome Forest, MLP, Gradient Boosting, and XGBoost to create a predictive model capable of accurately forecasting which consumers are at a high risk of churning.

Design and implement systems capable of identifying factors affecting consumer churn in the telecommunication industry and how we can retain the consumers.

Research Questions:

What are the common characteristics of consumers who are most likely to churn, and how can these be addressed proactively?

What are the primary factors contributing to consumer churn in the telecommunication industry?

How can consumer demographics and behavioural data be used to predict churn?

Which machine learning algorithms are most effective for predicting consumer churn in the telecom sector?

What is the accuracy and reliability of the predictive churn model in identifying at-risk consumers?

How can the insights derived from the churn prediction model be used to design effective retention strategies?

How do service quality improvements and consumer support enhancements influence churn rates?

1.5 Research Methodology:

Machine learning is a computer science topic that may be defined as tackling a practical problem by gathering data and algorithmically developing a statistical model to extract information from it [21]. Learning can be semi-supervised, unsupervised, or reinforced. This study employed a supervised learning strategy since each dataset sample was labeled to train a model that accepts feature vectors as input and provides information that can be used to infer the label of newly acquired data.

The methodology for this project describes a data-driven approach to analyzing, forecasting, and managing consumer churn in telecommunications corporations. It includes data collection, preprocessing, model creation, assessment, and deployment phases, all to develop a robust predictive model capable of reliably identifying at-risk clients and informing targeted retention initiatives.

In this project, we will use machine learning algorithms to evaluate data. We get a dataset of a telecommunication company from Kaggle. We have considered five models: kNN, Random Forest, MLP, Gradient Boosting, and XGBoost. We evaluated the data on each model and got a confusion matrix as a result. We assess the accuracy of each model confusion matrix and select the best model to identify the factors of consumer churn in the telecommunication industry and the factors that will help retain consumers in the company. We have compared the performance of different models based on the evaluation metrics and selected the best-performing model for deployment. Use the model's predictions to generate actionable insights into consumer behaviour and churn risk. These insights will guide the development of personalized retention strategies, such as targeted marketing campaigns, loyalty programs, and service quality improvements. To assess the outcome, ROC, AUC, F1-score metrics, and cross-validation approaches have been taken into consideration.

This methodology provides a structured approach to understanding and mitigating consumer churn in telecommunication companies. The project aims to create a predictive model that identifies at-risk consumers and informs actionable strategies to retain them by combining data-driven analysis with advanced machine learning techniques. The goal is to reduce churn rates, enhance consumer satisfaction, and improve telecom companies' financial performance.

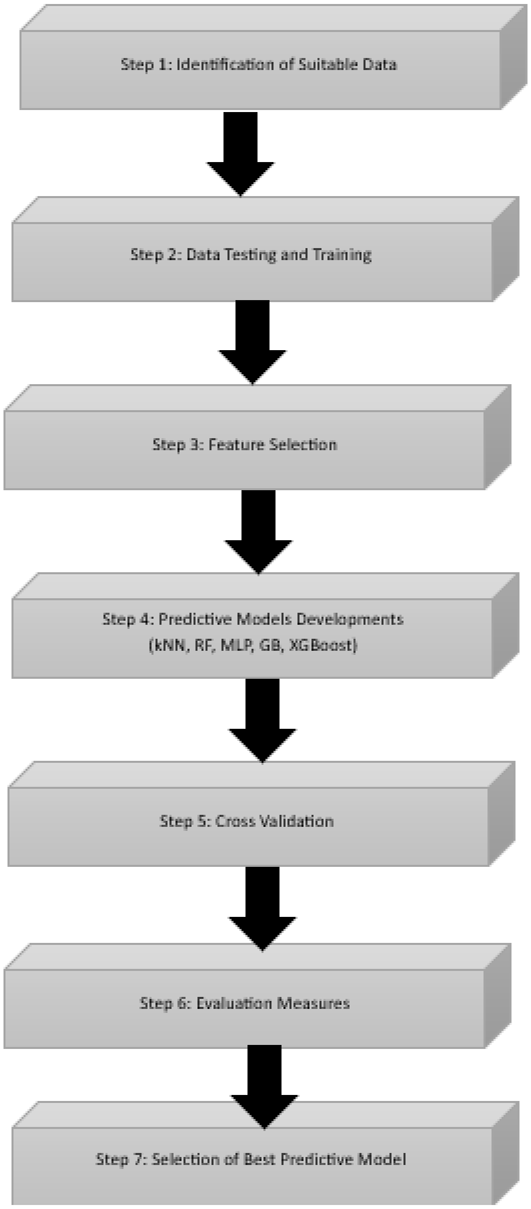

Step 1: Identification of Suitable Data
Step 2: Data Testing and Training
Step 3: Feature Selection
Step 4: Predictive Models Developments
(kNN, RF, MLP, GB, XGBoost)
Step 5: Cross Validation
Step 6: Evaluation Measures
Step 7: Selection of Best Predictive Model

Figure 1: Multi-phase model for developing a consumer churn predicted model.

1.5.1 Receiver Operating Characteristic:

PR and ROC curves are useful metrics for classifier performance [23]. The ROC analysis is a graphical tool for assessing classifier efficiency. It assesses a classifier's performance based on two statistics: true positive and false positive rates. The data are plotted in a two-dimensional graph, with the FPR on the x-axis and the real positive rate on the y-axis. Plotting sensitivity (true positive rate) against 1-specificity (false positive rate) on the y-axis for each tabular value results in the graphical ROC curve.

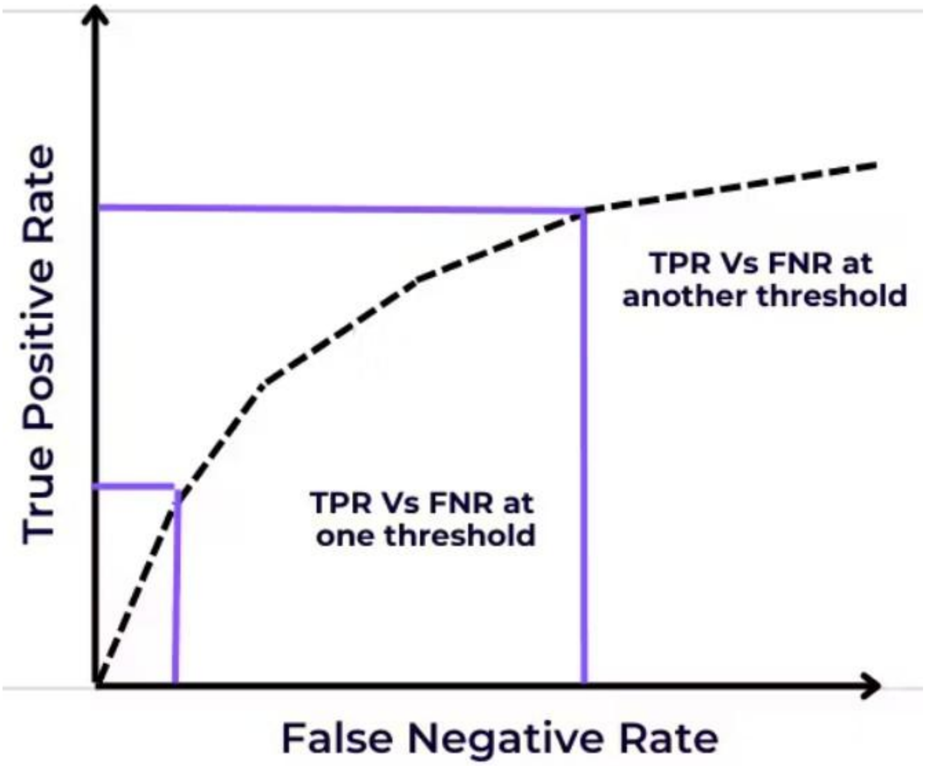


Figure 2: ROC Curve.

An ROC curve is a graph that depicts the efficacy of a classification model. It helps you comprehend the model's decision-making procedure at different confidence levels.

1.5.2 Area Under the Curve Precision Recall

The area under the precision-recall curve (AUC-PR) is a model performance metric for binary responses that applies to rare events and is unaffected by model specificity [22]. The acronym AUC stands for "Area Under the ROC Curve." That is, AUC determines the full two-dimensional region beneath the entire ROC curve.

It is a threshold-independent metric that calculates the area under a curve, which shows a trade-off between several aspects of performance when the model's prediction threshold increases.

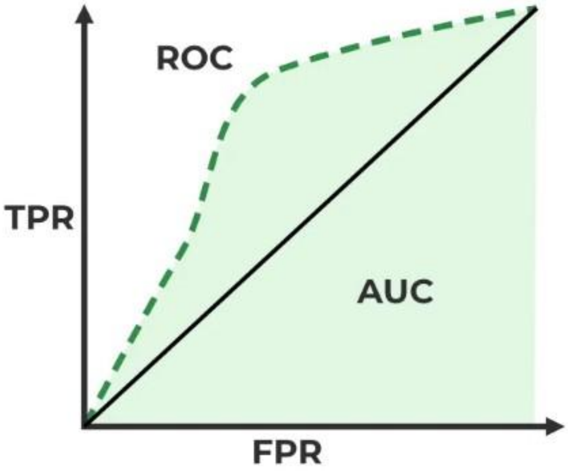


Figure 3: AUC-PR Curve.

1.5.3 Confusion Matrix:

A confusion matrix (also known as an error matrix) visualizes the results of a classification system. In simpler terms, it is a table that contrasts the total number of actual occurrences of a certain class to the expected number of occurrences. Confusion matrices are one of the evaluation methods used to determine the efficacy of classification models. They may assess a range of model performance measures, including accuracy and recall.

Columns indicate a class's expected values, while rows represent actual values (ground truth) or vice versa. It's worth noting that the reverse holds in research. This grid framework is a useful instrument for displaying model classification accuracy because it shows the number of right and wrong predictions for all adjacent classes [31].

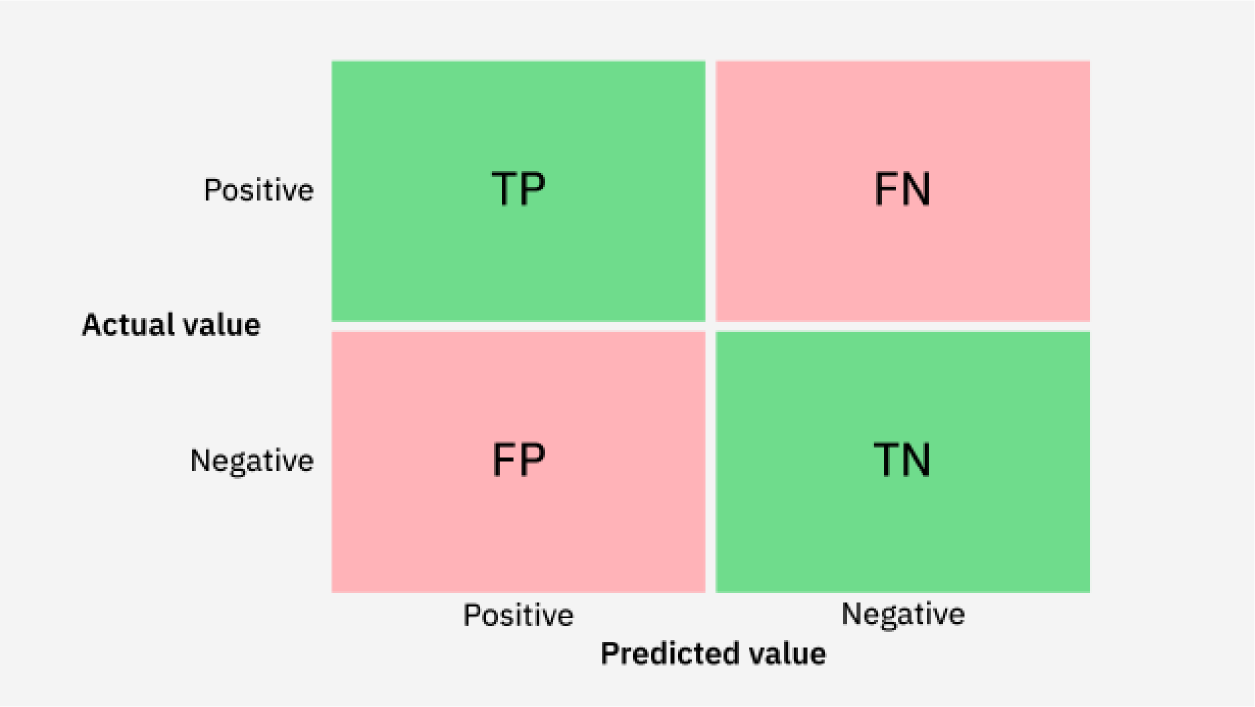


Figure 4: Confusion Matrix.

The top-left box displays the number of true positives (TP), which are accurate predictions for the positive class. The box underneath contains false positives (FP), or occurrences of the negative class that were wrongly categorized as positive. These are frequently known as type I mistakes in statistics. The top-right box shows the number of false negatives (FN), which are actual positive occurrences that were incorrectly projected as negatives. Finally, the bottom-right box shows the number of true negatives (TN), which are negative class instances that were correctly predicted as negative. The model's total number of predictions may be computed by combining all these figures.

1.5.4 Evaluation Measures:

Evaluation metrics are important in machine learning because they explain how well a model works.

1.5.5 True Positive:

You expected a positive value, which is right when the number of times our real positive values match the expected positive.

1.5.6 False Positive:

You expected a positive value, which is the number of times our real positive values match the expected positive.

1.5.7 True Negative:

A negative number can be expected when the actual negative value matches the expected negative value, and it is indeed negative.

1.5.8 False Negative:

You projected negative value, but it was positive when the number of times our model incorrectly forecasts negative values as positive.

1.5.9 Accuracy:

It is used to determine the proportion of properly categorized values. It informs us how frequently our classifier is correct. It is calculated by summing all true values and dividing them by total values. It is determined by using Equation (1) below.

$$\text{Accuracy} = (TP+TN)/\text{total} = (TP+TN)/(TP+FP+FN+TN) \tag{1}$$

1.5.10 Precision:

The efficacy of the model is assessed in terms of accuracy. It is calculated as the ratio of all cases to all accurate ones.

$$\text{Precision} = TP/(TP+FP) = TP/\text{Predicted positive} \tag{2}$$

1.5.11 Recall:

Recall, computed as given in Equation, is the fraction of each item in the test set that the recognition system correctly classifies as false.

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) = \text{TP}/\text{Actual positive} \quad (3)$$

1.5.12 F1 Score:

It is a perfect balance of recall and accuracy. It combines scores of accuracies and precision of a model.

$$\text{F1 Score} = 2*(\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

The higher the value, the better the performance for all these measures.

1.5.13 Cross Validation:

Cross-validation is one of the most used data resampling techniques for model selection and assessment. Cross-validation may be used to fine-tune the hyperparameters of statistical and machine learning models, avoid overfitting, compare learning methods, and assess prediction model generalization error. This article [46] discuss the most prevalent cross-validation varieties, including k-fold, nested, and leave-one-out cross-validation, and how they relate to other data resampling procedures.

1.6 Limitation of the study

The effectiveness of ML models heavily relies on the quality and completeness of the data. In telecom companies, consumer data may contain missing values, inconsistencies, or inaccuracies due to manual entry errors, system glitches, or outdated records. Limited access to comprehensive datasets can hinder the model's ability to make accurate predictions.

Consumer churning datasets are often imbalanced, with a smaller proportion of consumers churning than those who stay. This imbalance can lead to models biased towards predicting consumer retention, resulting in poor performance in identifying churners. Special techniques such as oversampling, under-sampling, or synthetic data generation may be required, but these methods have limitations. In addition, consumer behaviour and preferences can change over time due to various factors such as new market trends, competitor actions, or changes in consumer needs. A model trained on historical data might not capture these evolving patterns, leading to decreased prediction accuracy. Continuous monitoring and updating of the model are necessary but can be resource-intensive.

Machine learning models typically focus on internal company data such as consumer demographics, usage patterns, and billing information. However, external factors such as economic conditions, regulatory changes, or competitor actions can influence consumer churn. Integrating external data sources into the model can be complex and only sometimes feasible. Moreover, handling large volumes of consumer data raises significant privacy and security concerns. Ensuring compliance with data protection regulations such as GDPR and safeguarding sensitive consumer information is critical. These concerns may limit the extent of data collection, sharing, and processing, potentially affecting the model's performance.

As the telecom company grows and accumulates more data, the model's scalability becomes a concern. Ensuring that the model can handle increasing data volumes and complexity without a significant drop in performance or speed is essential but can be technically challenging.

Chapter 2: Literature Review

2.1 Introduction:

The section literature review represents the previous studies and research related to the topic of Predicting consumer churn. It also highlights the significance and role of artificial intelligence in predicting consumer churn. It also highlights the limitations of previous studies and identifies the gaps in knowledge.

2.2 Literature Review:

(Wanchai, P., 2017): The telecoms business is changing tremendously due to information technology and global competitiveness. Consumers may now use their purchasing power by selecting various suppliers to meet their communication demands. As a result, the primary challenge facing the telecoms business is delivering good facilitation and higher quality at a reduced cost. Despite the highly competitive industry, corporations have worked actively with campaigns, marketing programs, and other ways to attain the most excellent DM immersion. As a result, it is helpful to understand which clients are likely to migrate to a rival soon. These clients are referred to as churned-consumers. Churn has various reasons, including unhappiness with services and expensive bills. Consumers frequently receive appealing discounts when joining up with a new telecom company. CC is an essential activity in the fast-increasing and competitive telecommunications industry because of the expansion of recruiting new consumers

(Naz, N. A., Shoaib, U., & Sarfraz, M. S., 2018): They used DT and multilayer sensory neural network evaluations to create a model for churn prediction in telecom companies. The researchers evaluated the strategies using a dataset from a Taiwanese telecommunications business that included different categories of parameters such as consumer, billing information, demographics, call detail records, a service change log, and contract/service status. According to the study, age, gender, tenure, billing amount, number of missed payments, in-net call time, and number of changes in account information are significant factors in distinguishing churners from non-churners. Researchers use the k-means clustering approach to classify clients based on billing amount, tenure, and consumption. Every cluster has its decision tree. A basic DT is also generated for all clients, with no segmentation. The findings present no considerable difference in performance across decision trees with and without customer fragmentation. A comparison of DT with multilayer sensory NN reveals that NN is more productive on this dataset. This research discovered that data mining tools were beneficial in accurately detecting future churners.

(Patro CS, 2020): They stated that technological factors, which include a lack of plan upgrades, roaming difficulties, voice quality issues, inadequate internet access, network coverage concerns, and new technologies launched by competing service providers, have no significant influence on the switching intentions of consumers.

(Mbarek R, Baeshen Y, 2019): Most studies that employed the service quality construct discovered that it indirectly impacted client satisfaction. The study found a direct impact of service quality on telecom companies' churning rate. The researcher discovered that more consumers want to switch to another operator if they are unsatisfied with the service provided by their present supplier.

(Shirazi, F., Mohammadi, M., 2019): The authors examined the implications of this predicament for the Canadian banking sector. Their primary purpose is constructing a projected churn model by combining structured archival data with unstructured information such as online websites, website traffic, and phone call logs. They also investigated how different consumer habits influence churn decisions.

(Stripling, E.; Vanden Broucke, S.; Antonio, K.; Baesens, B.; Snoeck, M., 2015): described ProfLogit, which combined a machine learning classifier with a genetic algorithm to leverage the EMPC during the training stage.

(Mohammad NI, SA, I., MN, K., OM, Y., A, A., 2019): conducted a study to discover the variables influencing consumer turnover, develop an effective churn prediction model, and offer the best interpretation of data visualization findings. Kaggle, an open data platform, created the dataset.

(Jain H., Khunteta A., & Srivastava S., 2020): Due to the significant expense of acquiring new consumers, the focus of research has shifted to client retention. Furthermore, it has been understood that a high turnover rate may damage a sector by creating poor services in client acquisition and higher operational expenses. A strategy is essential to assess progress toward significant improvements enabling telecom business development based on retention. Due to the expanding number of publications on the issue, there is a greater need to map the emerging consumer churn study domain and recommended areas for future research emphasis.

(Mustafa, Sook Ling, & Abdul Razak ,2021): This study demonstrated that customer satisfaction with helpdesk services influences NPS scores. Every grade fulfils client expectations. Analyzing the provider's prospective churners will elucidate the company's operational dynamics, indicating if it offers a superior product accompanied by exceptional customer service or requires substantial enhancements to remain competitive. They used 33 variables, including Linear Discriminant Analysis, Logistic Regression, Classification, the k-nearest-neighbors Classifier, Regression Trees (CART), Gaussian Naïve Bayes, and Support Vector Machine, to predict the influence of customer satisfaction with helpdesk service on NPS ratings. Each grade meets customer requirements.

(Srivastava & Sinha, 2024): This study aims to develop a churn prediction model for an internet service provider to identify at-risk customers by analyzing their usage patterns. The model facilitates targeted retention efforts, allowing the company to focus special offers on high-risk customers, thereby reducing churn rates, enhancing customer satisfaction, and increasing profitability. Evaluating machine learning models for churn prediction, Gradient Boosting emerged as the top performer with the highest AUC and AP scores. AdaBoost and Logistic Regression also showed strong performance. Models like Multi-Layer Perceptron, Light Gradient Boosting, and XGBoost were reliable, while K-Nearest Neighbours and Decision Tree showed lower precision and consistency. Based on these findings, boosting techniques and Logistic Regression are recommended for tasks requiring high recall, with a focus on promoting additional services to at-risk customers to effectively reduce churn.

(Ahmad, Jafar, & Aljoumaa, 2019): This model utilizes machine learning methodologies on an extensive collection of data to provide a novel approach to feature engineering and selection. The model's effectiveness is assessed using the Area Under Curve (AUC) metric. The model was developed and evaluated within the Spark environment using an extensive dataset generated by SyriaTel, a telecommunications provider. The model evaluated four algorithms: Gradient Boosted Machine Tree (GBM), Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBOOST). They opt to validate and optimize hyperparameters by tenfold cross-validation. They utilized effective feature transformation, feature engineering, and a selection methodology to produce features for machine learning algorithms.

(Bhuse, Gandhi, Meswani, Muni & Katre, 2020): In this study, ML and deep learning approaches are used to forecast telecom consumer attrition. Traditional techniques, like Random Forest Classifiers and Support Vector Machines, are contrasted with modern models such as XGBoost and Deep Neural Networks to forecast customer attrition. The effectiveness of these models is also evaluated by grid search. The experiment indicated that the RF model is the most suitable option for this application, with an accuracy rate for predictions of 90.96% on the testing data before grid search. Their research focused on using the following methodologies: SVM, RF, XGBoost, KNN, Ridge classifier, and Deep Neural Network. The dataset used was sourced from a customer retention campaign.

(Saheed & Hambali, 2021): This study utilizes several machine learning methods, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and Naive Bayes (NB). This paper presents an innovative technique for feature selection. This method integrates the Information Gain and Ranker methodologies. Standard metrics such as accuracy, precision, and F-measure, along with 10-fold cross-validation, are employed to evaluate model performance.

(Mishra & Rani, 2017): Based on the findings of this research, when dealing with datasets that are not evenly distributed, it is necessary to do proper preprocessing and data balancing to enhance the effectiveness of the classifiers that are used. The performance criterion for classifiers that are assessed in terms of area under the curve (AUC) is raised by the SMOTE-based classifier. Additionally, for the purpose of evaluating the effectiveness of distinguishing features in classifier training, suitable feature extraction methods are used. This may lead to a decline in performance and a reduction in the amount of model learning that occurs. By SMOTE analysis, correlation-based feature extraction, and classifier ensembles using CART, Bagged CART, and PART as foundation classifiers, the suggested approach is a promising combination of these three techniques. To produce the greatest results in terms of prediction performance indices (area under the curve, sensitivity, and specificity), the AdaBoost classifier is the best option. As a result, the technique that has been provided might be considered a viable alternative for precisely determining the number of customers who have left the telecoms firm.

(Chowdhury, A., Kaisar, S., Rashid, M. M., Shafin, S. S., & Kamruzzaman, J., 2021): They offer a model that amalgamates complex oversampling approaches with ensemble methods, such as Random Forest, Gradient Boosting, XGBoost, and AdaBoost. A variety of alterations were applied to the hyperparameters of the baseline ensemble and excessive sampling algorithms to evaluate their effect on prediction performance. The predictive performance of the proposed model was evaluated using metrics including accuracy, precision, recall, F1 score, and area under the ROC curve. This assessment was conducted using a publicly available and commonly utilized customer turnover dataset. Compared to previous techniques, their model was better as it significantly decreased the incidence of false positive and false negative predictions.

(Shumaly, Neysaryan & Guo, 2020): The challenge in this study was the segmentation of consumers into two groups: customers who want to discontinue use of the organization's services and loyal customers. Initially, 5 ML methods were applied in an imbalanced dataset, and the results were unsatisfactory. As a result, three strategies for data balancing were applied, all of which increased the algorithms' performance as measured by the AUC index.

(Jain, H., Yadav, G., & Manoov, R., 2020): Consumer churn avoidance is one of the significant variables in increasing an organization's revenue. Customer attrition happens when a company's products or services are no longer used by its customers. In this paper, they predict customer attrition in advance to enable the implementation of effective customer retention strategies through the use of preliminary data analysis and personalized offers specific to the target audience. A comparative analysis of four algorithmic models across three domains—banking, telecommunications, and information technology—comprises our implementation for attrition prediction. The rationale for conducting this comparative study is the scarcity of research studies that analyses the performance of a variety of algorithms in distinct disciplines. They employ exploratory data analysis to develop a variety of retention methods.

(Labhsetwar, S. R. 2020): This study uses machine learning algorithms to detect prospective churn consumers, divided them based on their usage habits, and visually represent the evaluation outcomes. Machine learning algorithms could properly forecast customer turnover and help develop methods to retain clients, according to the study. The article's opening highlights the significant risk of client attrition to the telecommunications industry, supported by statistical proof.

A comprehensive analysis of pertinent scientific literature is presented. The study examined controlled machine learning approaches, which were evaluated using the dataset.

(Ebrah & Elnasir, 2019): This study used orange software to visualize two datasets—the cell2cell dataset and the IBM Watson dataset—and evaluated them. This article aims to determine the main elements that contribute to customer churn and to build the best accurate model for telecom churn prediction. The study determined that telecommunications operators can achieve optimal predictive models by examining their comprehensive data and monitoring customer behavior, enabling the formulation of diverse marketing strategies to retain churners based on the identified predictors from historical customer data analysis. Client acquisition, cross-selling, and up-selling are just a few examples of the many customer response models that may benefit from the attrition prediction models offered here.

(Singh, M., Singh, S., Seen, Kaushal & Kumar, 2018): This study was performed in a churn prediction modelling context, comparing four distinct machine learning algorithms against a publicly available dataset in the telecoms sector. Two significant conclusions may be drawn from the results: i) The Random Forest method surpasses other fundamental classification models; and ii) Feature engineering is an essential component of model effectiveness. This study will try to tackle this problem with four traditional machine learning methodologies. The efficacy of several classification methods in delivering precise predictions was examined. The used data set remains unaltered and unbalanced. A decision has been made to eliminate a set of attributes deemed unimportant and with little potential for qualitative prediction. Individuals are chosen, based on the characteristics most strongly correlated with the outcome variable. Models are trained using the optimal parameter set identified by the Grid Search process. Traditional assessment techniques, including confusion matrices and the ROC curve, are used to determine the classification efficacy in issue.

Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & devidas Chincholkar, Y. (2023): This paper proposes a framework for detecting potential customer attrition using machine learning approaches, especially support vector machines. Companies in the e-commerce industry have access to a wealth of data that might be used to discover norms of behaviour; any deviation from these patterns or occurrences may be seen as potential client churn. Another use of hybrid recommendation algorithms is the development of individualized retention strategies. Evaluating the efficacy of suggestions and giving automatic feedback are the two steps suggested for enhancing the SVM model. Also, to see whether ensemble methods may enhance churn detection, they are applied to the SVM model.

(Nalatissifa & Pardede, 2021): This investigation is being undertaken with the objective of creating a new design for the DNN algorithm that employs a hard modulator to obtain optimal hyperparameter supplements. The number of nodes that are included in each concealed layer is determined by the tuning hyperparameters, which include dropout, learning rate, and random search. Furthermore, this investigation implements two distinct activation functions and three distinct concealed layer counts. The aim of this study is to create a model for predicting customer attrition in the telecoms industry. The secondary data that was incorporated into the dataset was obtained from Kaggle. A DNN was implemented during the modelling phase of this project. The accuracy performance of DNN modelling was demonstrated to be preferable to that of KNN modelling, RF modelling, and DT modelling. This was a result of the utilization of three concealed layers in conjunction with hyperparameters that were suitable.

(Sai, B. K., & Sasikala, 2019): This study performed an exploratory data analysis including statistical testing for feature selection, data mining, and visualization approaches to forecast potential churners using a logistic regression model. Prior to beginning the modeling procedure, the dataset was analyzed using data visualization tools. We assessed the model using the Receiver Operating Characteristic (ROC), which quantifies the proportion of genuine positives and false positives. The curve's area is consistently between zero and one. They used several statistical tests to initiate the feature selection process and created a logistic regression model.

(Wu, S., Yau, W. C., Ong, T. S., & Chong, 2021): This study contributes to current literature in three ways. For starters, only a small portion of the available research considers both attrition prediction and customer segmentation in the telecom business. This study addresses this gap by presenting an integrated consumer analytics platform that easily connects these two components. Second, Bayesian Analysis is used in very restricted studies. This study uses it for factor analysis, allowing it to function as an intermediate between churn prediction and client segmentation. Third, this study gives operators an overall likelihood of churning for each cluster, allowing them to better comprehend the churning condition in each cluster.

(Li, W., & Zhou, C. 2020): The concept of user categorization and sequential regression

Modelling was developed by this research. This approach involves the identification of fields that are highly associated with consumer Congestion and consumer behavior, the division of the total customer base into distinct groups based on these fields using clustering methods, and the subsequent development of personalized regression prediction models for each group using regression analysis. A notion of user segmentation and bit by bit regression modelling is created by all of this. To improve the precision of their forecasts, they check the authenticity of the client data and standardize the linear segmentation of the initial variable. To further understand how the system impacts consumer attrition decisions, the model was again tested with the updated dataset. The model was then put into production after achieving good results.

(Ali, Rehman, , Hafeez, & Ashraf, 2018): It is well acknowledged that acquiring new clients is more expensive than maintaining existing ones. There is an existing issue in which clients quit the firm for unclear reasons. In their research, they used numerous data mining approaches to forecast client attrition. It will eventually aid in evaluating consumer behavior and determining if a client is churning or not. Multiple classifiers, including Support Vector Machine (SVM), Bagging, Stacking, Naïve Bayes, Bayesian Network, and Adaboost, were used to assess customer turnover behaviour. The online data set provided by Kaggle was crucial in this endeavor. In their research, they seek to find noteworthy patterns for predicting customer churn behavior in the telecom sector.

(Al-Mashraie, Chung, & Jeon, 2020): In this study, they examine the efficacy of several churn prediction models using real data gathered from a partner firm. This research aims to identify the factors that most significantly influence churn, enabling targeted improvement efforts in those areas. The methods and conclusions presented in this study might assist telecommunications businesses create more efficient and successful marketing campaigns and customer retention incentives. Prior to doing the PPM analysis, this study employs statistical techniques and models such as logistic regression, SVM, random forest, and decision tree.

(Coussement, Lessmann, & Verstraeten, 2017): This research, conducted within a churn prediction modelling framework, evaluates an enhanced logit model against eight advanced data mining algorithms using basic input data, including empirical cross-sectional data from a prominent European telecom provider. The study concludes with many managerial implications and recommendations for further study, including evidence of the generalizability of the results in different business contexts. Although this study adds value to existing literature, it is not without limits, which might serve as the foundation for fascinating future research routes. First, they employ a case study technique using a single churn data set to assess the influence of numerous DPTs on prediction performance, shedding fresh insight on an understudied area. However, more study is required to improve the generalizability of their findings by evaluating them against data sets from various application areas and business settings.

(Cenggoro, Wirastari, Rudianto, Mohadi, Ratj & Pardamean, 2021): The goal of this work was to evaluate the utilization of deep learning models to produce embedded vectors that possess discriminative properties to forecast customer attrition. They demonstrated that the discriminative vectors can be represented in two-dimensional space to distinguish between two categories of rotating consumers. The likelihood of one of the categories departing is high, while the other category may be maintained. Organizations that offer telecommunications services may benefit from this data to make informed decisions regarding the most effective targeting of their marketing campaigns, thereby retaining the majority of their consumers. Although the results of this research were favourable, it is feasible that the vector embedding model could be improved through the implementation of more advanced methodologies. Cosine-distance-based loss functions, which have been demonstrated to be beneficial in the context of face recognition, are one approach that could be investigated in the future. This approach has the capacity to enhance the differentiation of embedding vectors across a diverse array of classes.

Previous research and studies have devoted only infrequent attention to synthesizing considerable studies on consumer turnover behaviour, particularly in a particular industry. The consumer churn behaviour study draws on various disciplines, including marketing and other domains like marketing, social science, etc. As stated in the literature, the purpose of the numerous research streams is to provide a holistic answer to consumer turnover and thereby increase company performance.

2.2.1 Key Takeaways:

- The telecoms business is changing tremendously due to information technology and global competitiveness.
- Acquiring new consumers is more costly than retaining existing ones.
- Data mining tools were beneficial in accurately detecting future churners.

- More consumers want to switch to another operator if they are unsatisfied with the service provided by their present supplier.
- It is necessary to do proper preprocessing and data balancing to enhance the effectiveness of the classifiers that are used.
- With so many datasets available, the need for machine learning is growing.
- A strategy is essential to assess progress toward significant improvements enabling telecom business development based on retention.

Chapter 3: Project Description

3.1 Data Description

3.1.1 Data Collection

The data set selected[1] contains consumer-level details for a Telecommunication company. It includes different attributes such as the number of weeks the consumer has had an active account, contract renewal details, monthly data usage by the consumer, etc. It also includes demographics. Consumer demographics include geography, income level, age, and gender. Use patterns include call frequency, internet usage, and peak usage periods. Billing details include contract terms, payment history, and monthly bill amounts. Every row represents a consumer, and every column has consumer-related attributes as specified in the description.

3.2 Data Preprocessing:

We split the dataset for training and testing before making models (kNN, Random Forest, MLP, Gradient Boosting, XGBoost).

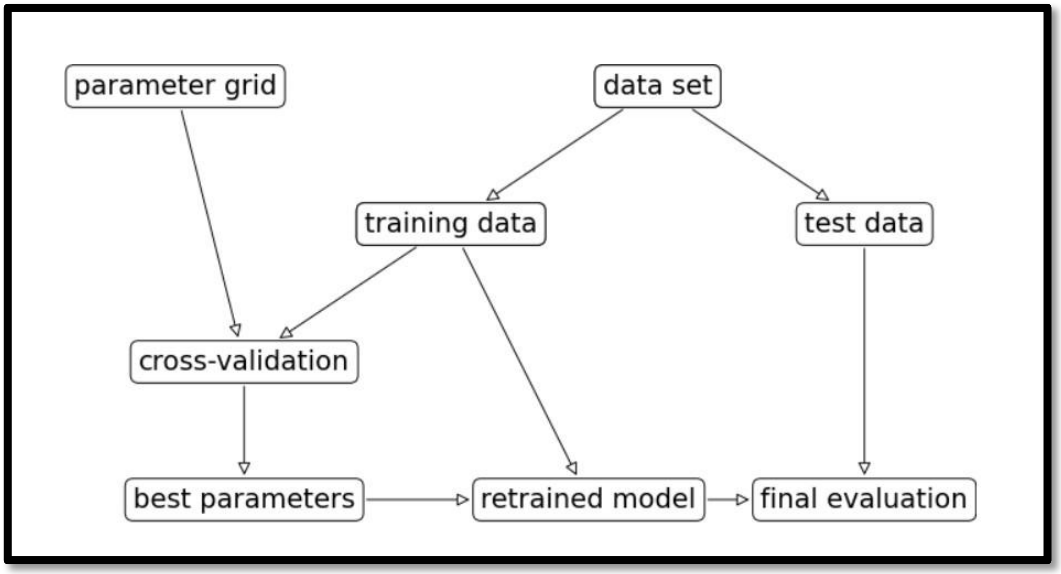


Figure 5: Data Classification.

[1] https://www.kaggle.com/datasets/barun2104/telecom-churn/data

3.3 Featured Engineering

3.3.1 Possible Scenarios

```
Out[7]: Text(0.5, 1.0, '% of Churn')
```

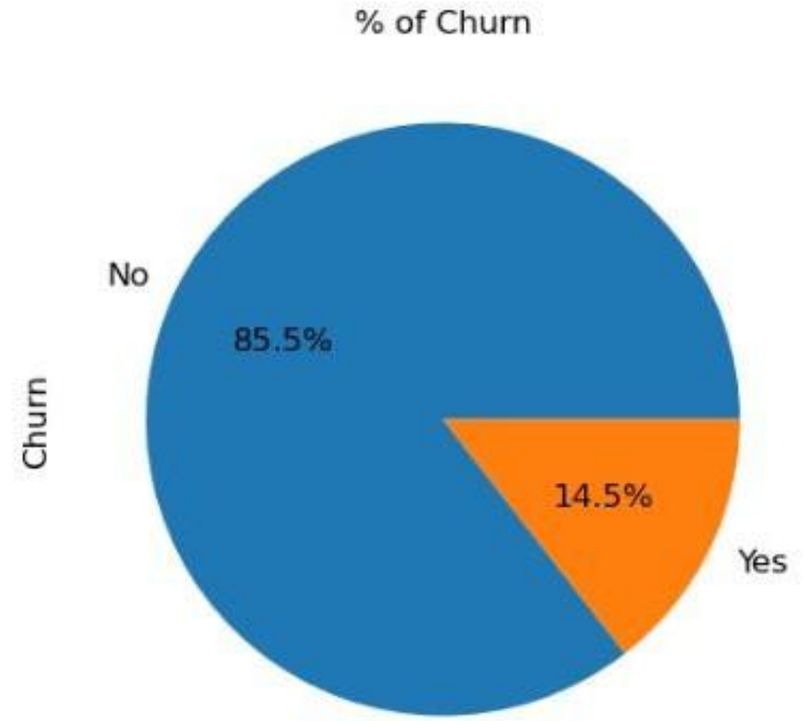


Figure 6: Churn Rate.

- The statistics indicate that 85.5% of users did not churn and remained to use Telecom services.
- 14.5% of consumers discontinued utilizing Telecom services.

3.3.2 Possible Scenarios:

The graphic might be applicable to a variety of events, including:

- Analyze consumer turnover to optimize retention efforts.
- Understanding the percentage of subscription cancellations.
- Tracking product usage, including the percentage of people that cease using it.

3.3.3 Boxplots

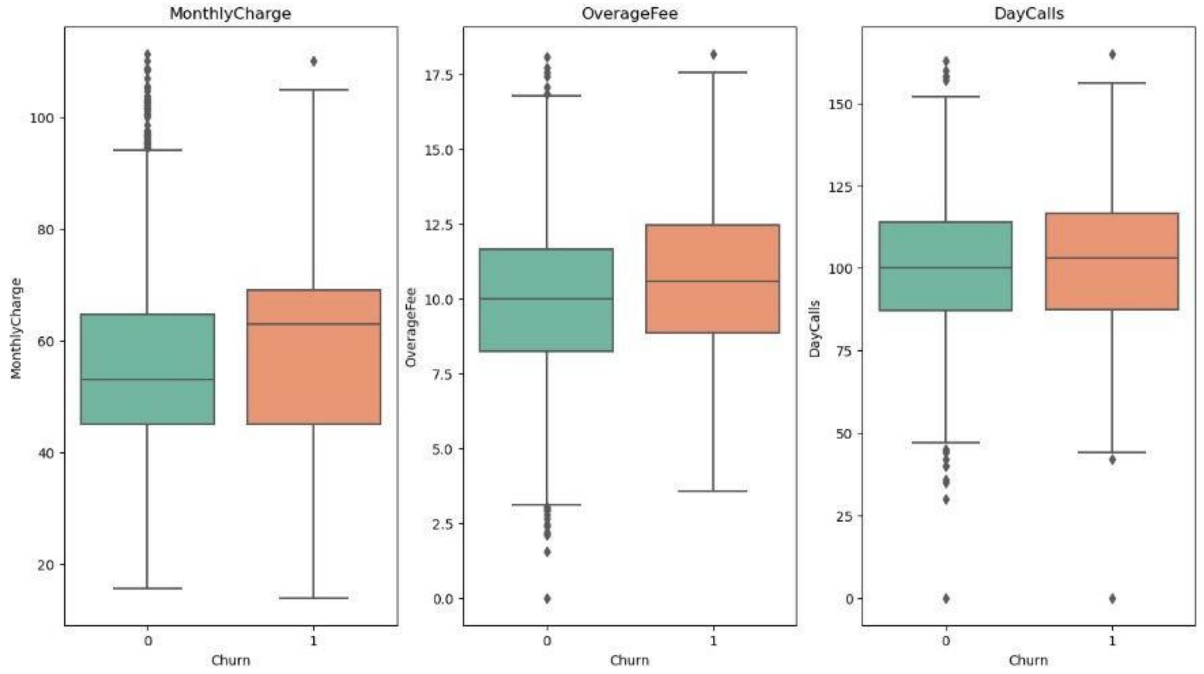


Figure 7: Box Plots.

3.3.4 Monthly Charge:

- Consumers in group "1" generally have higher monthly charges than group "0."
- The distribution of monthly charges is more widespread (larger IQR) for group "1."
- There are some outliers with exceptionally high monthly charges in both groups.

3.3.5 Average Fee:

- Consumers in group "1" tend to have higher overage fees than those in group "0."
- The distribution of overage fees is similar in shape for both groups.
- There are some outliers with extremely high overage fees in both groups.

3.3.6 Day Calls:

- Consumers in group "1" make fewer day calls than group "0."
- The distribution of day calls is relatively similar for both groups, with group "1" having a slightly lower median.

- There are some outliers with exceptionally high numbers of daily calls in both groups.

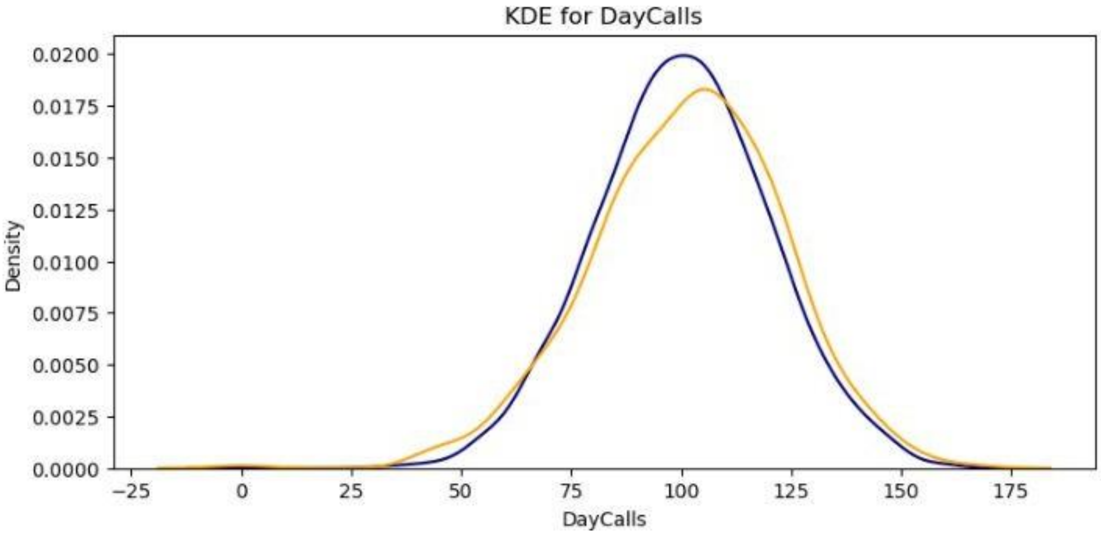


Figure 8: KDE for Day Calls.

- Recent clienteles are more prone to churn.
- Clients with frequent Day calls are more likely to churn.
- Average charge and day calls are likely essential elements.

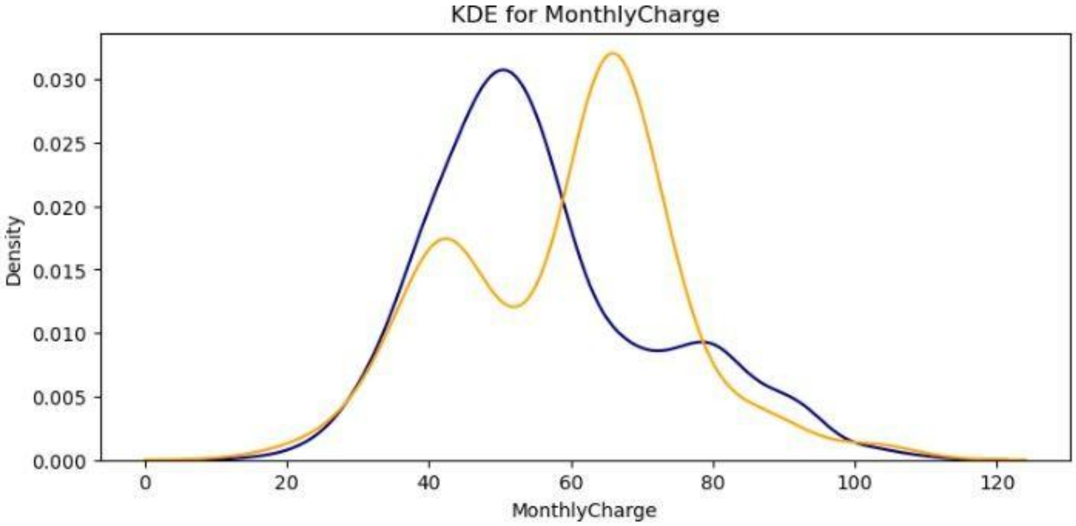


Figure 9: KDE for Monthly Charges.

- Former consumers are more inclined to churn.

- Consumers who pay greater cost per month are likewise more prone to churn.
- Average Payment and Every month Payments are likely key characteristics.

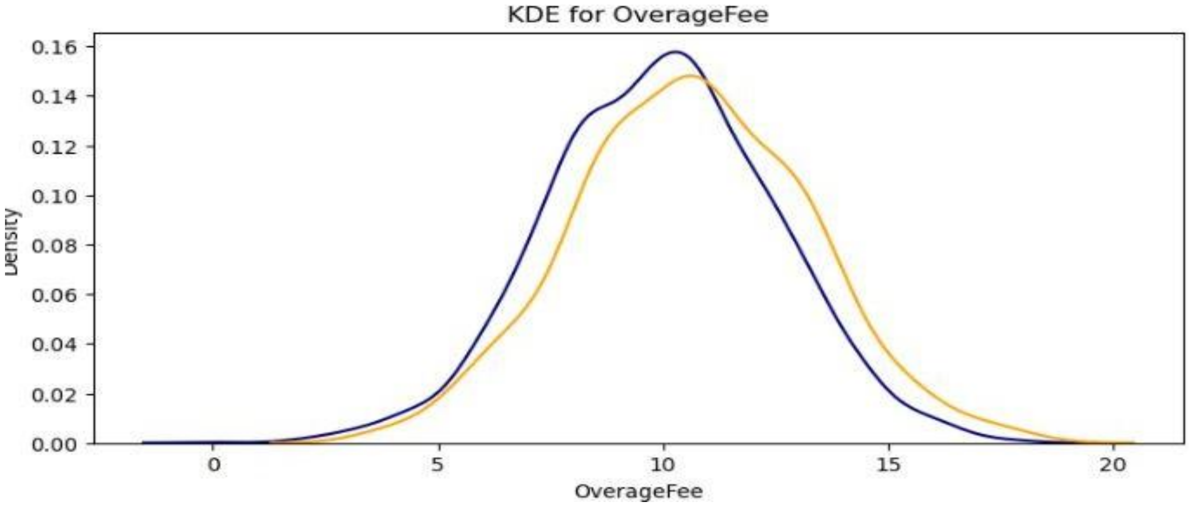


Figure 10: KDE for Overage Fee.

A Kernel Density Estimation (KDE) graphic illustrates the likelihood density of a discrete quantity. Churns with higher average fees are more probable to churn than those on the yellow line. The bar chart below in figure 11 shows that many contract renewal consumers are not to be a part of the churn for telecom services. In addition, figure 12 depicts a high proportion of users who utilize a data plan, indicating that there is no churn for telecommunications services.

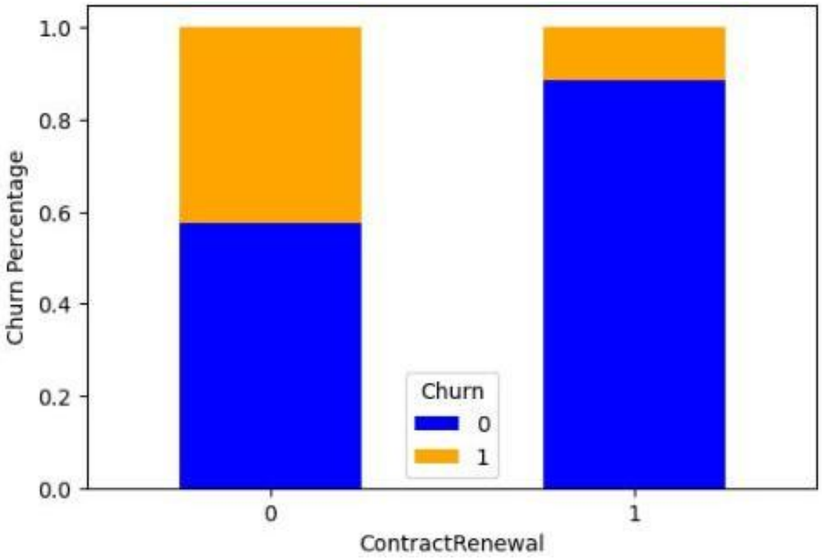


Figure 11: Contract Renewal Percentage.

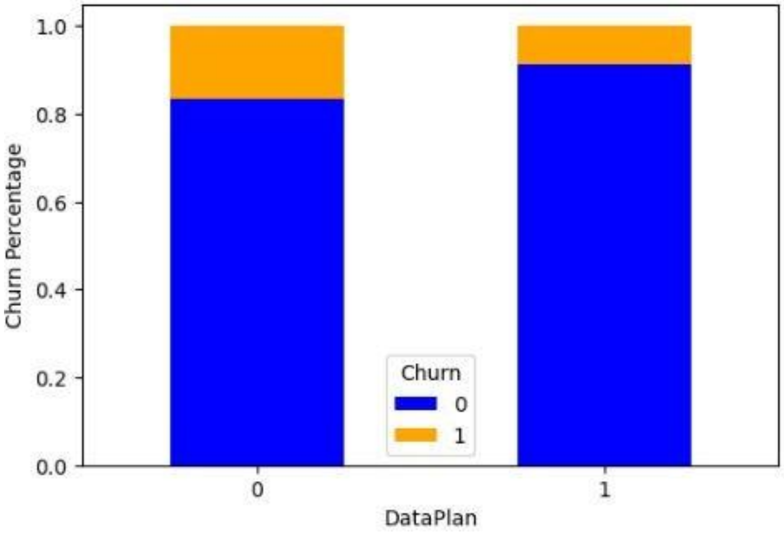


Figure 12: Data Plan Percentage.

Moreover, in figure 13 the trend of consumers making service calls have a high chance of not being a part of consumer churn.

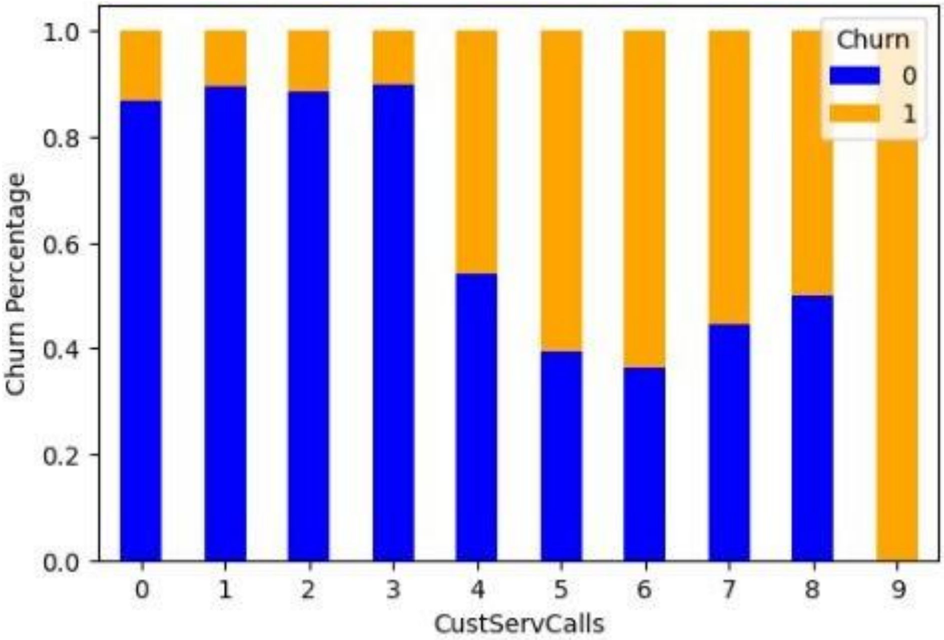


Figure 13: Consumer Service Calls Percentage.

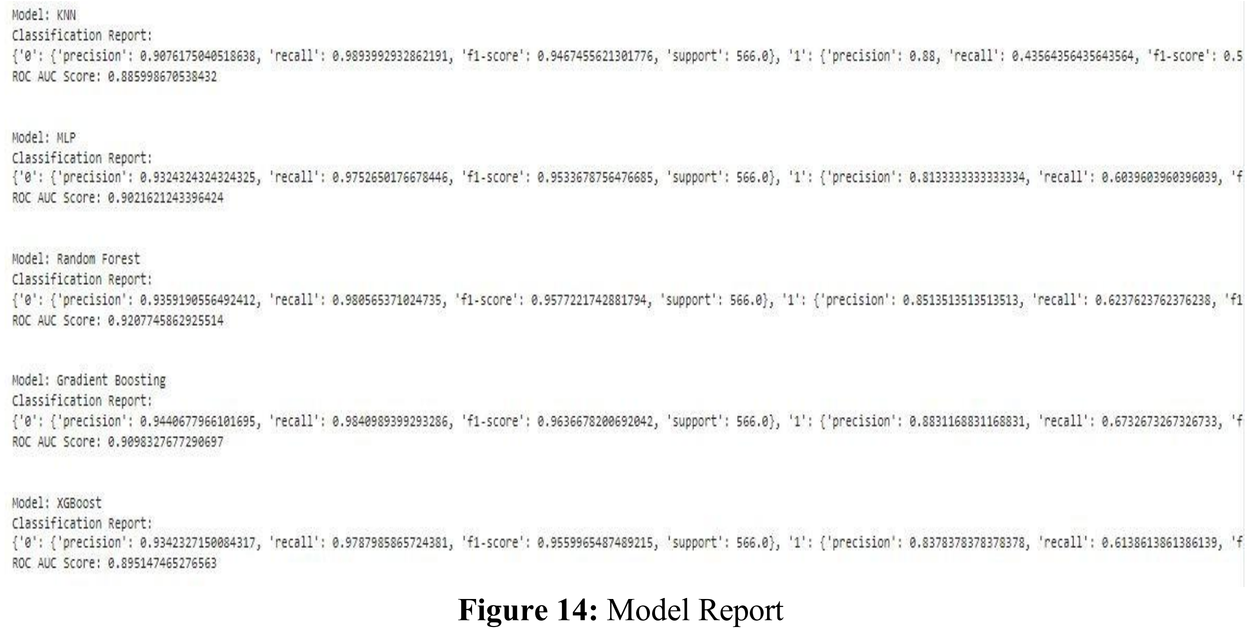
Chapter 4: Analysis

4.1 Project Description

4.1.1 Analysis and Result:

The study and results section describes the performance of ML models, designed to forecast consumer turnover in a telecom firm. It contains a full study of the data pretreatment procedures, model training, evaluation metrics, final findings, and model-derived insights. The dataset was divided into a training set and a test set to assess the model's performance. Cross-validation techniques were employed during training to improve model parameters and reduce overfitting. Several machine learning models were trained and tested on the preprocessed dataset. The investigation shows that machine learning algorithms, notably XGBoost and Gradient Classifier algorithm with an accuracy of 93.70%, are extremely successful at forecasting consumer attrition in the telecom business. Using these models, telecom businesses may get important insights into consumer behaviour, devise tailored retention tactics, and, eventually, minimize churn. The model's effective deployment and favourable impact demonstrate the value of data-driven decision-making in improving client retention and overall business performance. The model's findings were utilized to improve the consumer experience, particularly in high-risk categories. This has lowered turnover while also increasing client happiness and loyalty.

4.1.2 Model Report:



```
Model: KNN
Classification Report:
{'0': {'precision': 0.9076175040518638, 'recall': 0.9893992932862191, 'f1-score': 0.9467455621301776, 'support': 566.0}, '1': {'precision': 0.88, 'recall': 0.43564356435643564, 'f1-score': 0.5
ROC AUC Score: 0.885998670538432


Model: MLP
Classification Report:
{'0': {'precision': 0.9324324324324325, 'recall': 0.9752650176678446, 'f1-score': 0.9533678756476685, 'support': 566.0}, '1': {'precision': 0.8133333333333334, 'recall': 0.6039603960396039, 'f
ROC AUC Score: 0.9021621243396424


Model: Random Forest
Classification Report:
{'0': {'precision': 0.9359190556492412, 'recall': 0.980565371024735, 'f1-score': 0.9577221742881794, 'support': 566.0}, '1': {'precision': 0.8513513513513513, 'recall': 0.6237623762376238, 'f1
ROC AUC Score: 0.9207745862925514


Model: Gradient Boosting
Classification Report:
{'0': {'precision': 0.9440677966101695, 'recall': 0.9840989399293286, 'f1-score': 0.9636678200692042, 'support': 566.0}, '1': {'precision': 0.8831168831168831, 'recall': 0.6732673267326733, 'f
ROC AUC Score: 0.9098327677290697


Model: XGBoost
Classification Report:
{'0': {'precision': 0.9342327150084317, 'recall': 0.9787985865724381, 'f1-score': 0.9559965487489215, 'support': 566.0}, '1': {'precision': 0.8378378378378378, 'recall': 0.6138613861386139, 'f
ROC AUC Score: 0.895147465276563
```

Figure 14: Model Report

4.1.2 Machine Learning Algorithm:

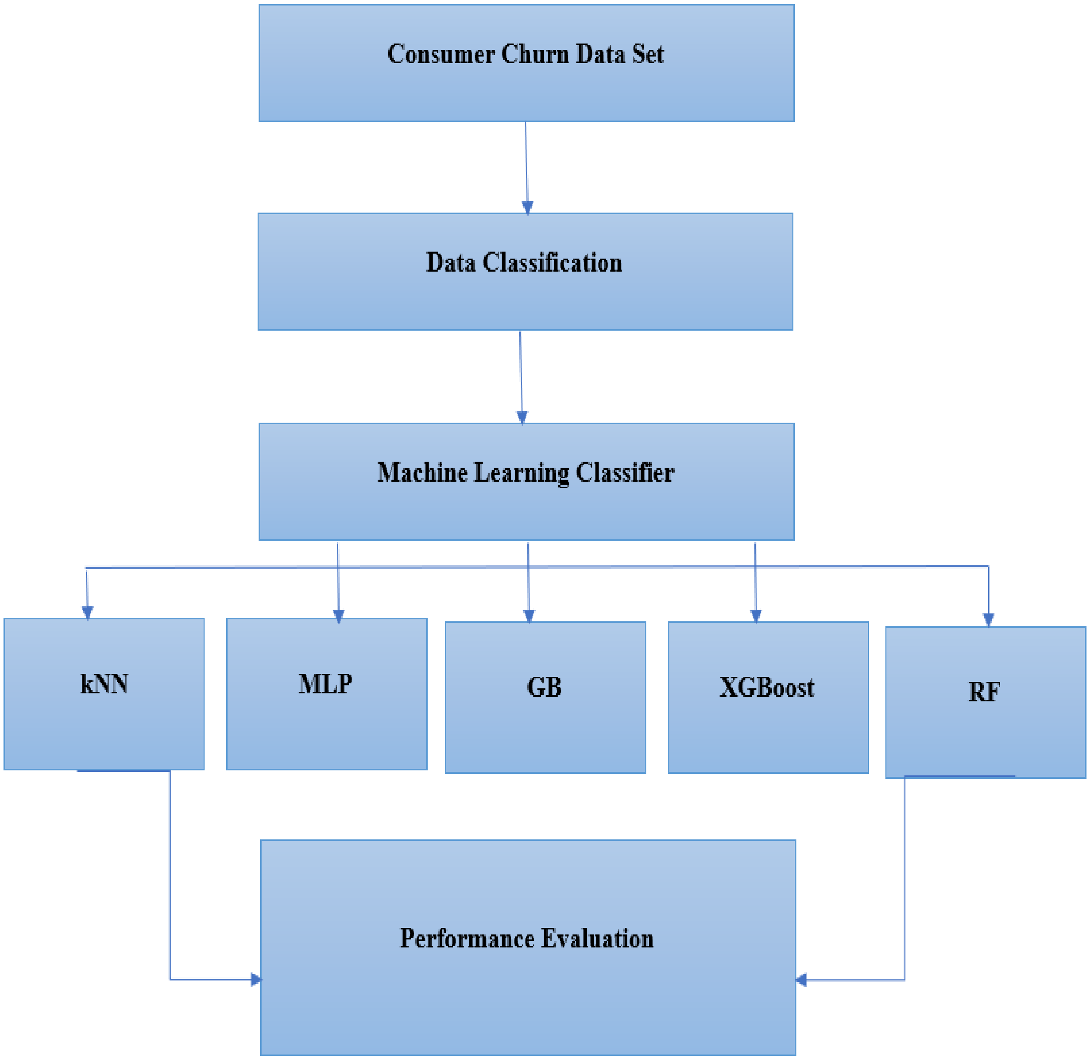


Figure 15: Consumer churn predicted model flow diagram.

Machine learning (ML) teaches machines to process data more efficiently. Sometimes, we cannot analyze the information retrieved after reviewing the data. It is the scientific study of algorithms and statistical models that computer systems use to carry out a certain activity without being explicitly programmed [25].

In such a case, we consider ML. With so many datasets available, the need for ML is growing. Many firms utilize ML to obtain valuable information. Many tests have been undertaken to see if robots can learn independently without being explicitly programmed. Many mathematicians and programmers employ various techniques to address this problem, which require massive data sets.

4.2 Data Preprocessing

4.2.1 KNN:

The K Nearest Neighbor (KNN) approach is frequently utilized in data mining and machine learning applications because of its ease of implementation and superior performance. However, the prior kNN approaches' use of the same k value for all test data has proven unworkable in real-world applications. The goal of employing a K-Nearest Neighbors (KNN) model and its associated confusion matrix is to assess the model's effectiveness in forecasting whether a client would churn (leave) or remain. The confusion matrix gives a thorough analysis of the model's categorization findings, assisting in assessing its accuracy and potential areas for improvement. The current study uses a kNN Confusion matrix to determine the causes of consumer churn in the telecommunications business. kNN classification finds a group of k items in the training set closest to the test object and labels them according to the prevailing class in this neighbourhood. This tackles the issue that, in many data sets, one item is unlikely to match another perfectly and that objects closest to an object may offer contradictory information regarding its class [26].

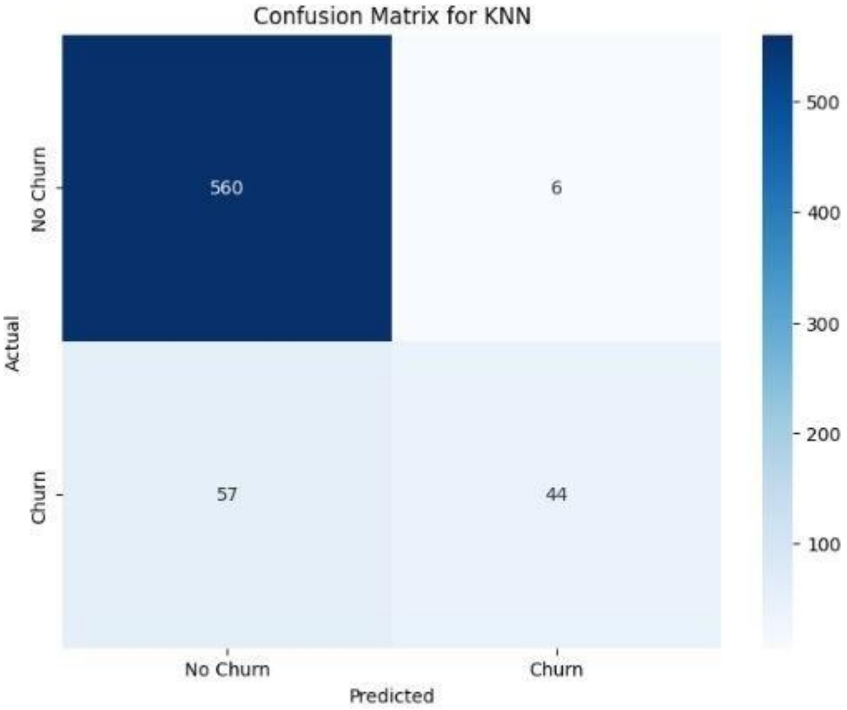


Figure 16: kNN Confusion Matrix.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Precision = 560/ (560+ 57)

Precision = 560 / 617

=0.9076

Precision = 90.76%

Recall = $\frac{TP}{TP+FN}$

Recall = 560 / (560 + 6)

Recall = 560 / 566

=0.9893

Recall = 98.93%

$$\boldsymbol{F1 - Score = 2 * \frac{Precision*Recall}{Precision+Recall}} \tag{6}$$

F1-Score = 2 * [(0.90 * 0.98) / (0.90 + 0.98)]

F1-Score = 2 * (0.882 / 1.88)

F1-Score = 2 * 0.469

=0.934

F1-Score = 93.4%

We predict the value of precision is 98% recall is 90%, and F1-score is 93.94%

$$Accuracy = \frac{\text{TP+TN}}{\text{total}} = \frac{(TP+TN)}{\text{(TP+FP+FN+TN)}} \tag{7}$$

= (560+44)/ (560+57+6+44)

Accuracy = 604/ 667

Accuracy = 0.905

The KNN accuracy is 90.55%

See Figure 16, If we examine a plot generated by the KNN Classifier, we can see that the TP value is 560, the FN value is 6, the FP value is 57, and the TN value is 44. For the '0', we anticipate a recall of 90% and a precision of 98%. The F1 score is 93.4%. The F1-score is used to assess how well our model works throughout the full dataset, and the number of instances of each given class in the true responses is known as support.

4.2.2 Random Forest Confusion Matrix:

Random Forest is an ensemble learning approach that builds numerous decision trees and combines their results to increase forecast accuracy and prevent overfitting. The Random Forest model and its accompanying confusion matrix are used to evaluate the model's ability to forecast consumer attrition. The confusion matrix gives a thorough breakdown of the classification results, allowing you to understand the model's performance better and find areas for improvement [27]. An RF model is a classifier that comprises numerous decision trees and produces the class that is the average of the classes created by individual trees.

For many data sets, it generates a very accurate classifier. It supports a huge number of input variables. It assesses the relevance of factors in categorization. As the forest grows, it generates an internal, unbiased estimate of the generalization error. It features an effective approach for guessing missing data and retains accuracy even when much is missing. It offers an experimental approach to detecting variable interactions. It can correct errors in the class population of imbalanced data sets. It calculates proximities between instances, which are important for grouping, spotting outliers, and (by scaling) displaying the data.

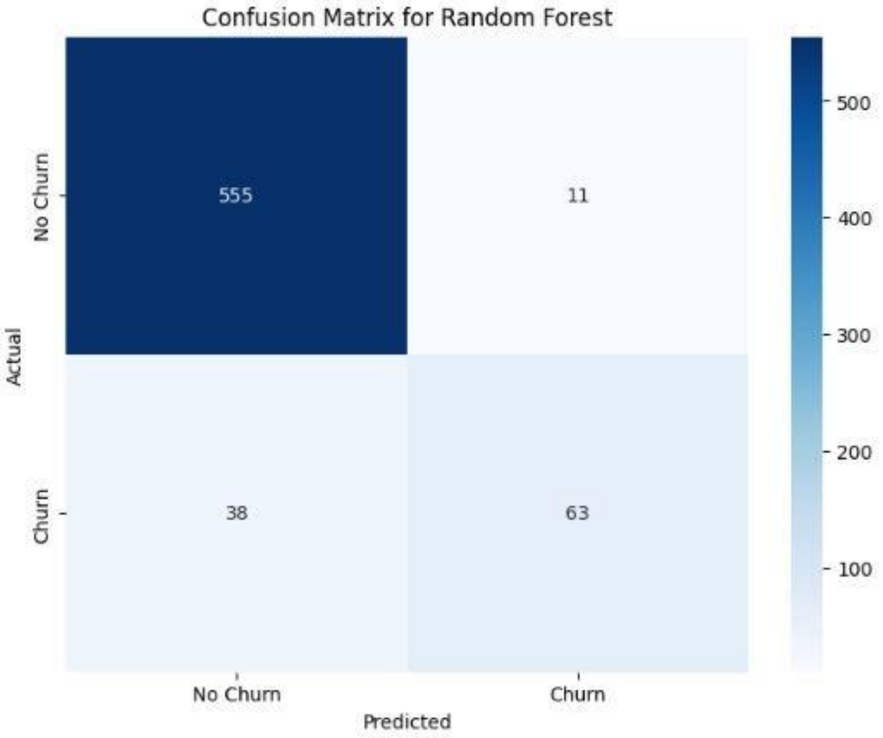


Figure 17: Random Forest Confusion Matrix

$$Precision = \frac{TP}{TP + FP}$$

Precision = 555/ (555+38)

Precision = 555 / 593

=0.9359

Precision = 93.5%

Recall = $\frac{TP}{TP+FN}$

Recall = 555 / (555 + 11)

Recall = 555/ 566

=0.980

Recall = 98%

$$\boldsymbol{F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}}$$

F1-Score = 2 * [(0.935 * 0.98) / (0.935 + 0.98)]

F1-Score = 2 * (0.9163 / 1.915)

F1-Score = 2 * 0.4784

=0.9569

F1-Score = 95.69%

We predict the value of precision is 93.5% recall is 98%, and F1-score is 95.69%

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{total}} = \frac{(TP + TN)}{(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})}$$

= (555+63)/ (555+38+63+11)

Accuracy = 618/ 667

Accuracy = 0.926

The RF accuracy is 92.6%

See Figure 17, we can see that the genuine positive value is 555, the FN is 11, the false positive is 38, and the TN is 63 after evaluating a plot generated by an RF Classifier. For the '0', we anticipate a recall of 98% and a precision of 93.5%. The F1 score is 93.4%. The F1-score is used to assess how well our model works throughout the full dataset, and the total number of instances of each given class in the true responses is known as support.

4.2.3 MLP Confusion Matrix:

When applied to an MLP model for consumer churn analysis, the confusion matrix comprehensively examines the model's performance. It enables organizations to analyze their performance, make data-driven decisions, and improve consumer retention strategies by correctly identifying probable churners and balancing precision and recall based on business requirements. It can handle huge datasets with many characteristics, making it ideal for consumer churn research in which several factors impact consumer behavior.

-

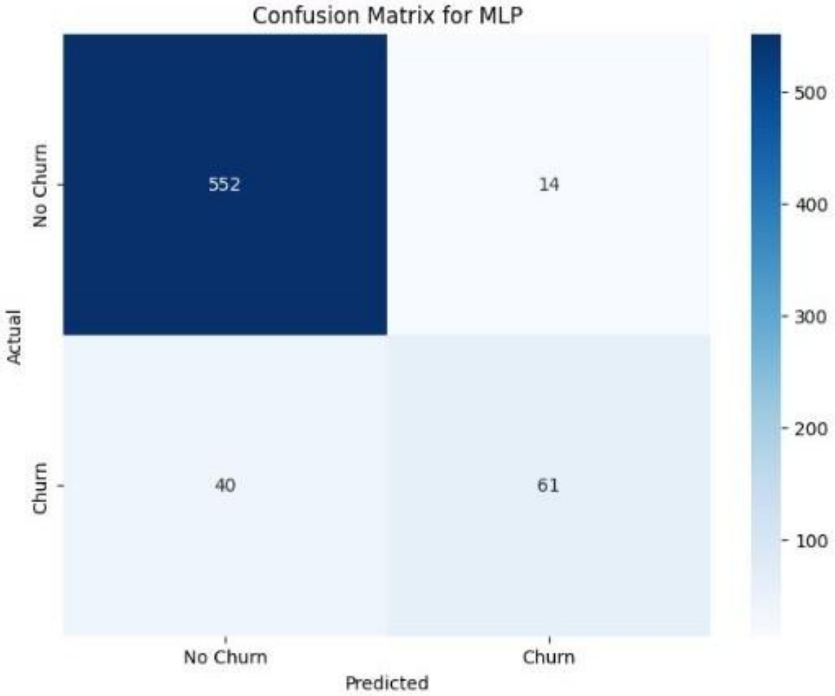


Figure 18: MLP Confusion Matrix.

$$Precision = \frac{TP}{TP + FP}$$

Precision = 552/ (552+40)

Precision = 552 / 592

=0.9324

Precision = 93.24%

Recall = $\frac{TP}{TP+FN}$

Recall = 552 / (552 + 14)

Recall = 555/ 566

=0.9752

Recall = 97.52%

$$\boldsymbol{F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}}$$

F1-Score = 2 * [(0.932 * 0.975) / (0.932 + 0.975)]

F1-Score = 2 * (0.9087 / 1.907)

F1-Score = 2 * 0.4765

=0.9530

F1-Score = 95.3%

We predict the value of precision is 93.2% recall is 97.5%, and F1-score is 95.3%

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{total}} = \frac{(TP + TN)}{(\mathrm{TP + FP + FN + TN})}$$

= (552+61)/ (552+40+14+61)

Accuracy = 613/ 667

Accuracy = 0.9190

The MLP accuracy is 91.90%

See Figure 18, If we examine a plot generated by the MLP Classifier, we can see that the TP value is 552, the FN is 14, the FP is 40, and the TN is 61. For the '0', we anticipate a recall of 97.5% and a precision of 93.2%. The F1 score is 95.3%. The F1-score is used to assess how well our model works throughout the full dataset, and the total amount of instances of each given class in the true responses is known as support.

4.2.4 Gradient Boosting Confusion Matrix:

Gradient Boosting is recognized for its high accuracy and capacity to handle complicated, non-linear connections in data, making it an excellent choice for forecasting consumer attrition.

Gradient Boosting models can give insights into feature significance, allowing you to determine which elements are most significant in forecasting churn. The confusion matrix in the context of a Gradient Boosting model for consumer churn study thoroughly evaluates the model's effectiveness. It enables organizations to analyze their performance, make data-driven decisions, and improve consumer retention strategies by correctly identifying probable churners and balancing precision and recall based on business requirements. Gradient boosting Algorithm is a method that stands out for its prediction speed and accuracy, especially on big and complicated datasets. This algorithm has provided the greatest results in Kaggle contests and business machine-learning solutions. It is a boosting strategy. It has many applications, including regression, classification, ranking, and survival analysis [29].

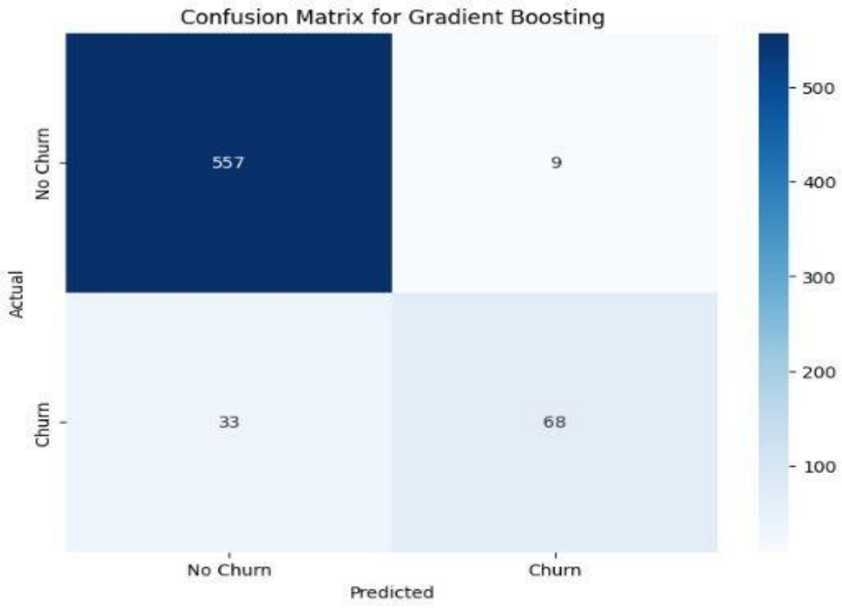


Figure 19: Gradient Boosting Confusion Matrix

$$Precision = \frac{TP}{TP + FP}$$

Precision = 557/ (557+33)

Precision = 557 / 590

=0.944

Precision = 94.40%

Recall = $\frac{TP}{TP+FN}$

Recall = 557 / (557 + 9)

Recall = 557/ 566

=0.9840

Recall = 98.40%

$$\boldsymbol{F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}}$$

F1-Score = 2 * [(0.944 * 0.9840) / (0.944 + 0.9840)]

F1-Score = 2 * (0.9288 / 1.928)

F1-Score = 2 * 0.4817

=0.9634

F1-Score = 96.34%

We predict the value of precision is 94.4% recall is 98.40%, and F1-score is 96.34%

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{total}} = \frac{(TP + TN)}{(\mathrm{TP + FP + FN + TN})}$$

= (557+68)/ (557+33+9+68)

Accuracy = 625/ 667

Accuracy = 0.9370

The GBC accuracy is 93.70%

See Figure 19, If we examine a plot generated by the MLP Classifier, we can see that the TP value is 557, the FN is 9, the FP is 33, and the true negative is 68. For the '0', we anticipate a recall value of 98.40% and a precision of 94.4%. The F1 score is 96.34 percent. The F1-score is used to assess how well our model works throughout the full dataset, and the total amount of instances of each given class in the true responses is known as support.

4.2.5 XGBoost Confusion Matrix:

XGBoost, also known as eXtreme Gradient Boosting, is a machine learning technique that has become popular after winning many Kaggle data science contests [30]. An XGBoost model and its accompanying confusion matrix are used to assess the model's accuracy in forecasting whether a client will churn (leave) or stay. The confusion matrix thoroughly breaks down the classification findings, allowing you to understand the model's accuracy better. XGBoost, is an implementation of gradient-boost and decision trees, is well-known for its excellent predictive accuracy and efficiency, making it ideal for complicated applications such as churn prediction.

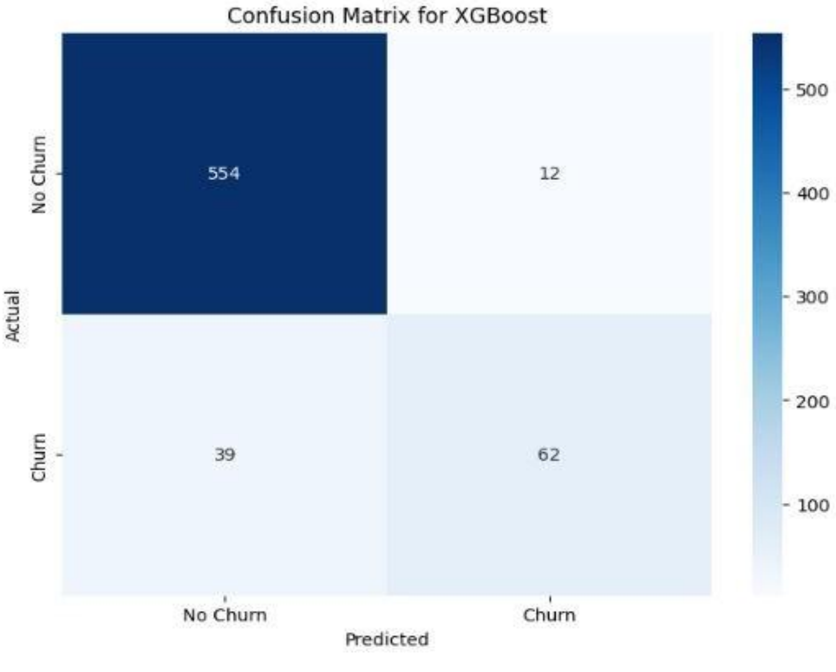


Figure 20: XGBoost Confusion Matrix.

$$Precision = \frac{TP}{TP + FP}$$

Precision = 554/ (554+39)

Precision = 557 / 593

=0.9392

Precision = 93.92%

Recall = $\frac{TP}{TP+FN}$

Recall = 554 / (554 + 12)

Recall = 554/ 566

=0.9787

Recall = 97.87%

$$\boldsymbol{F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}}$$

F1-Score = 2 * [(0.9392 * 0.9787) / (0.9392 + 0.9787)]

F1-Score = 2 * (0.919 / 1.917)

F1-Score = 2 * 0.4793

=0.9587

F1-Score = 95.87%

We predict the value of precision is 93.92% recall is 97.87%, and F1-score is 95.87%

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{total}} = \frac{(TP + TN)}{(\mathrm{TP + FP + FN + TN})}$$

= (557+68)/ (557+33+9+68)

Accuracy = 625/ 667

Accuracy = 0.9370

The XGBC accuracy is 93.70%

See Figure 20, If we examine a plot generated by the MLP Classifier, we can see that the TP value is 554, the FN is 12, the FP is 39, and the TN is 62. For the '0', we anticipate a recall value of 97.87% and a precision of 93.92%. The F1 score is 95.87 percent. The F1-score is used to assess how well our model works throughout the full dataset, and the number of instances of each given class in the true responses is known as support.

4.2.6 Best Predicted Algorithm:

Table 1: Model Comparison.

Classifier	Accuracy	Precision (Class 0)	Recall (Class 0)	F1-Score (Class 0)
kNN	0.905	0.98	0.90	0.9394
Random Forest	0.926	0.935	0.98	0.9569
MLP	0.9190	0.932	0.975	0.953
Gradient	0.9370	0.944	0.9840	0.9634
XGBoost	0.9370	0.9392	0.9787	0.9587

Table 1, demonstrates that the Gradient Classifier and XGBoost algorithms achieved the highest accuracy of 93.70%, showcasing strong predictive capabilities. These models effectively identified critical churn factors, with customer service calls emerging as a significant driver of churn. This insight provides a basis for targeted retention strategies.

Chapter 5: Conclusion and Recommendations

Although company losses are unavoidable, churn may be controlled and kept at an acceptable level. The telecom industry has been suffering from high rates of churn and enormous churning losses. To avoid obstacles in the telecommunication business, new ways must be established and old ones improved. This study analyzed and compared the quality metrics of many prediction models, including kNN, Random Forest, XGBoost Classifier, Gradient Boosting, and MLP. The growth tendency of the twenty-first century has proven to be the most dramatic boom in history. As technology advances, so does the number of services available, and it is difficult for a corporation to forecast the number of clients who will quit. In the telecom business, researchers have shown interest in consumer churn in the telecom sector in recent years. To address these rising concerns, this paper offers a comparative analysis of consumer churn prediction utilizing well-known machine learning algorithms. Moreover, some of these algorithms show a positive future prospect for ensemble learning using Gradient classifier and XGBoost classifier, as these algorithms provide around 93.7% accuracy. They beat competing algorithms across all performance criteria, including accuracy, precision, F-measure, and recall. Churn's prediction for a firm may be a time-consuming effort, and many rising organizations and startups face stiff competition. It is challenging to forecast actual clients for the organization.

5.1 Recommendations:

Enhance the predictive model by integrating external data sources such as social media sentiment, economic indicators, and competitive landscape data. This can provide a more holistic view of the factors influencing consumer churn and improve the accuracy of predictions. Stay abreast of the latest machine learning and telecom industry trends to continuously refine and improve the churn prediction model. Consider future proofing the model by incorporating advanced techniques such as deep learning, real-time data analytics, and reinforcement learning, which can further enhance predictive accuracy and adaptability.

5.1.1 Future Work:

In the future, with new ML, using theories and structures from the reinforcement learning and deep learning sectors is one of the most successful techniques for solving problems such as forecasting churn with more precise and accurate results. The future scope of this effort will use hybrid classification techniques to emphasize the present relationship between churn prediction and client lifetime value. The retention policies must be examined by picking relevant variables from the dataset. The industry's passive and dynamic character ensures that data mining will become increasingly important in the future of the telecommunications industry. We will soon be able to broaden our idea beyond telecom to include other sectors. We may also use contemporary technologies such as Deep Learning to develop an effective system for client retention. The study might be enhanced by including AI techniques for trend analysis and consumer prediction. To keep clients from churning, we may tailor the services they receive.

References:

[1]: Umayaparvathi, V., & Iyakutti, K. (2016). A survey on consumer churn prediction in telecom industry: Datasets, methods, and metrics. *International Research Journal of Engineering and Technology (IRJET)*, *3*(04).

[2]: Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H., & Almazroi, A. A. (2022). An ensemble-based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in the telecom industry. *PeerJ Computer Science*, *8*, e854.

[3]. Liu, Y., Fan, J., Zhang, J., Yin, X., & Song, Z. (2023). Research on telecom customer churn prediction based on ensemble learning. *Journal of Intelligent Information Systems*, *60*(3), 759-775.

[4]. Wu, X., Li, P., Zhao, M., Liu, Y., Crespo, R. G., & Herrera-Viedma, E. (2022). Customer churn prediction for web browsers. *Expert Systems with Applications*, *209*, 118177.

[5]. Xi, M., Luo, Z., Wang, N., & Yin, J. (2019). A latent feelings-aware rnn model for user churn prediction with behavioral data. *arXiv preprint arXiv:1911.02224*.

[6]. Almana, A. M., Aksoy, M. S., & Alzahrani, R. (2014). A survey on data mining techniques in consumer churn analysis for the telecom industry. *International Journal of Engineering Research and Applications*, *4*(5), 165-171.

[7]. Dahiya, K., & Bhatia, S. (2015, September). Consumer churn analysis in the telecom industry. In 2015, the 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO)(Trends and Future Directions) (pp. 1-6). IEEE.

[8]: Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of consumer churn prediction in the telecom industry using decision trees and logistic regression. In *2016 symposium on colossal data analysis and networking (CDAN)* (pp. 1-4). IEEE.

[9]: Melian, D. M., Dumitrache, A., Stancu, S., & Nastu, A. (2022). Consumer churn prediction in the telecommunication industry. A data analysis techniques approach. *Postmodern Openings*, *13*(1 Sup1), 78-104.

[10]. Wanchai, P. (2017, December). Consumer churn analysis: A case study on the telecommunication industry of Thailand. In *2017 12th International Conference for Internet Technology and Secured Transactions (ICITST)* (pp. 325-331). IEEE.

[11]. Naz, N. A., Shoaib, U., & Sarfraz, M. S. (2018). A review of customer churn prediction data mining modeling techniques. *Indian Journal of Science and Technology*, *11*(27), 1-27.

[12]. Patro CS (2020) Consumer switching behavior towards mobile number portability: a study of mobile users in India. Int J Cyber Behav Psychol Learn 10(3):31–46.

[13]. Mbarek R, Baeshen Y (2019) Telecommunications consumer churn and loyalty intention. Mark Manage Innovations 4:110–117.

[14]. Shirazi, F.; Mohammadi, M. A big data analytics model for consumer churn prediction in the retiree segment. Int. J. Inf. Manag. 2019, 48, 238–253.

[15]. Stripling, E.; Vanden Broucke, S.; Antonio, K.; Baesens, B.; Snoeck, M. Profit maximizing logistic regression modeling for consumer churn prediction. In Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, 19–21 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–10.

[16]. Mohammad NI, SA, I., MN, K., OM, Y., A, A. (2019). Consumer Churn Prediction In the Telecommunication Industry Using Machine Learning Classifiers. Proceedings of the 3rd International Conference on Vision, Image and Signal Processing., 1–7, 2019.

[17]. Jain H., Khunteta A., & Srivastava S. (2020). Telecom churn prediction and used techniques, datasets and performance measures: A review. Telecommunication Systems, 76(4), 613–630.

[18].Mustafa, N., Sook Ling, L., & Abdul Razak, S. F. (2021). Customer churn prediction for telecommunication industry: A Malaysian Case Study. F1000Research, 10, 1274.

[19].Srivastava, S., & Sinha, S. (2024). Machine Learning Techniques: Predictive Modeling for Customer Churn in Telecommunications. Nanotechnology Perceptions, 1151-1166.

[20].Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1-24.

[21]. Bhuse, P., Gandhi, A., Meswani, P., Muni, R., & Katre, N. (2020, December). Machine learning based telecom-customer churn prediction. In 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS) (pp. 1297-1301). IEEE.

[22]. Saheed, Y. K., & Hambali, M. A. (2021, October). Customer churn prediction in telecom sector with machine learning and information gain filter feature selection algorithms. In 2021 International Conference on Data Analytics for Business and Industry (ICDABI) (pp. 208-213). IEEE.

[23]. Mishra, K., & Rani, R. (2017, August). Churn prediction in telecommunication using machine learning. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 2252-2257). IEEE.

[24]. Chowdhury, A., Kaisar, S., Rashid, M. M., Shafin, S. S., & Kamruzzaman, J. (2021, December). Churn prediction in telecom industry using machine learning ensembles with class balancing. In 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1-6). IEEE.

[25]. Shumaly, S., Neysaryan, P., & Guo, Y. (2020, October). Handling class imbalance in customer churn prediction in telecom sector using sampling techniques, bagging and boosting trees. In 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 082-087). IEEE.

[26].Jain, H., Yadav, G., & Manoov, R. (2020). Churn prediction and retention in banking, telecom and IT sectors using machine learning techniques. In Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019 (pp. 137-156). Singapore: Springer Singapore.

[27].Labhsetwar, S. R. (2020). Predictive analysis of customer churn in telecom industry using supervised learning. ICTACT Journal on Soft Computing, 10(2), 2054-2060.

[28]. Ebrah, K., & Elnasir, S. (2019). Churn prediction using machine learning and recommendations plans for telecoms. Journal of Computer and Communications, 7(11), 33-53.

[29].Singh, M., Singh, S., Seen, N., Kaushal, S., & Kumar, H. (2018, November). Comparison of learning techniques for prediction of customer churn in telecommunication. In 2018 28th International Telecommunication Networks and Applications Conference (ITNAC) (pp. 1-5). IEEE.

[30].Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & devidas Chincholkar, Y. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. Measurement: Sensors, 27, 100728.

[31]. Nalatissifa, H., & Pardede, H. F. (2021). Customer decision prediction using deep neural network on telco customer churn data. Jurnal Elektronika Dan Telekomunikasi, 21(2), 122-127.

[32]. Sai, B. K., & Sasikala, T. (2019, April). Predictive analysis and modeling of customer churn in telecom using machine learning technique. In 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 6-11). IEEE.

[33]. Wu, S., Yau, W. C., Ong, T. S., & Chong, S. C. (2021). Integrated churn prediction and customer segmentation framework for telco business. Ieee Access, 9, 62118-62136.

[34].Li, W., & Zhou, C. (2020, March). Customer churn prediction in telecom using big data analytics. In IOP Conference Series: Materials Science and Engineering (Vol. 768, No. 5, p. 052070). IOP Publishing.

[35]. Ali, M., Rehman, A. U., Hafeez, S., & Ashraf, M. U. (2018, August). Prediction of churning behavior of customers in telecom sector using supervised learning techniques. In 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE) (pp. 1-6). IEEE.

[36]. Amin, A., Al-Obeidat, F., Shah, B., Tae, M. A., Khan, C., Durrani, H. U. R., & Anwar, S. (2020). Just-in-time customer churn prediction in the telecommunication sector. The Journal of Supercomputing, 76, 3924-3948.

[37]. Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. Computers & Industrial Engineering, 144, 106476.

[38]. Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95, 27-36.

[39]. Cenggoro, T. W., Wirastari, R. A., Rudianto, E., Mohadi, M. I., Ratj, D., & Pardamean, B. (2021). Deep learning as a vector embedding model for customer churn. Procedia Computer Science, 179, 624-631.

[40]. Amin A., Anwar S., Adnan A., Nawaz M., Alawfi K., Hussain A., & Huang K. (2017). Consumer churn prediction in the telecommunication sector using a rough set approach. Neurocomputing, 237, 242–254.

[41]. Al-Weshah G. A., Al-Manasrah E., & Al-Qatawneh M. (2019). Consumer relationship management systems and organizational performance: Quantitative evidence from the Jordanian telecommunication industry. Journal of Marketing Communications, 25(8), 799–819.

[42]. Buttle, F., & Maklan, S. (2019). Customer relationship management: concepts and technologies. Routledge.

[43]. Burkov, A. *The Hundred-Page Machine Learning Book*, 1st ed.; Publishing Kindle Direct: Seattle, WA, USA, 2019.

[44]. Zhou, Q. M., Zhe, L., Brooke, R. J., Hudson, M. M., & Yuan, Y. (2021). A relationship between the incremental values of area under the ROC curve and of area under the precision-recall curve. *Diagnostic and Prognostic Research*, *5*, 1-15.

[45]. Grau, J., Grosse, I., & Keilwagen, J. (2015). PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, *31*(15), 2595-2597.

[46]. Berrar, D. (2019). Cross-validation.

[47]. Mahesh, B. (2020). Machine learning algorithms review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.

[48]. Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), 1-19.

[49]. Sooknunan, K., Lochner, M., Bassett, B. A., Peiris, H. V., Fender, R., Stewart, A. J., ... & Lahav, O. (2021). Classification of multiwavelength transients with machine learning. *Monthly Notices of the Royal Astronomical Society*, *502*(1), 206-224.

[50]. Peter, S., Diego, F., Hamprecht, F. A., & Nadler, B. (2017). Cost-efficient gradient boosting. *Advances in neural information processing systems*, *30*.

[51]. Aydin, Z. E., & Ozturk, Z. K. (2021, March). Performance analysis of XGBCoost classifier with missing data. In *1st Int. Conf. Comput. Mach. Intell., no*.

[52]. Flach, P. (2019, July). Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9808-9814).