Optimizing Telecom Operations with Segmentation and Churn Prediction

MUKKA DHEERAJ 21951A0538

Optimizing Telecom Operations with Segmentation and Churn Prediction

A Project Report

Submitted in partial fulfillment of the

requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By

Mukka Dheeraj 21951A0538

Under the Esteemed Guidance of

Ms. K. Sangeeta

Assistant Professor

Department of Computer Science Engineering

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043, Telangana

MAY 2025

DECLARATION

I certify that.

a. The work contained in this report is original and has been done by me under the guidance of my supervisor(s).

b. the work has not been submitted to any other Institute for any degree or diploma.

c. I have followed the guidelines provided by the Institute in preparing the report.

d. I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

e. whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the report and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Place:

Date:

Signature of the Student

Mukka Dheeraj 21951A0538

CERTIFICATE

This is to certify that the project report entitled "**Optimizing Telecom Operations with Segmentation and Churn Prediction**" submitted by **Mr. Mukka Dheeraj (21951A0538)** to the Institute of Aeronautical Engineering, Hyderabad, in partial in partial fulfillment of the requirements for the reward of the Degree Bachelor of Technology in Computer Science and Engineering is a bonafide record of work carried out by him/her under my/our guidance and supervision. In whole or in parts, the contents of this report have not been submitted to any other institutes for the award of any Degree**.**

Supervisor:

Ms. K. Sangeeta
Assistant Professor

Head of the Department:

Dr. M. Lakshmi Prasad
Professor and HOD, CSE

Date:

APPROVAL SHEET

This project report entitled **Optimizing Telecom Operations with Segmentation and Churn Prediction** by **Mr. Mukka Dheeraj (21951A0538)** is approved for the award of the Degree Bachelor of Technology in **Computer Science and Engineering**.

Examiners **Supervisor(s)**

Ms. K. Sangeeta

Principal

Dr. L. V. Narasimha Prasad

Date:

Place:

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of the people who made it possible and whose constant guidance and encouragement crown all the efforts with success. I think out college management and respected Sri **M. Rajashekar Reddy**, Chairman, IARE, Dundigal for providing me with the necessary infrastructure to conduct the project work.

I express my sincere thanks to **Dr. L. V. Narasimha Prasad**, Professor and Principal who has been a great source of information for my work, and **Dr. M. Lakshmi Prasad,** Professor and Head, Department of CSE, for extending his support to carry on this project work.

I am especially thankful to our supervisor **Ms. K. Sangeeta**, Assistant Professor, Department of CSE, for her internal support and professionalism who helped me in shaping the project into a successful one. I take this opportunity to express my thanks to one and all who directly or indirectly helped me in bringing the effort to present form.

ABSTRACT

Client turnover is a significant issue and one of the top concerns for large businesses, particularly in the telecom sector, due to its direct impact on company profits. Companies are striving to develop methods to forecast probable outcomes to mitigate this impact. Identifying the factors contributing to customer churn is crucial for preventing it. Our work's key contribution is the creation of a churn prediction model that aids telecom providers in identifying customers who are most likely to churn. We analyze various scenarios of data analysis and present the results in graphical format. The model developed employs machine learning methods in a large data environment and introduces a novel approach for designing and selecting features. The datasets used to train, evaluate, and assess the algorithm include all customer data from previous months. We tested four different algorithms: Logistic Regression, Extreme Gradient Boosting (XGBoost), Gradient Boosted Machine (GBM) Trees, Random Forests, and Decision Trees. To increase the model's accuracy, the algorithms are trained with parameter tuning. This work focuses on optimizing telecom operations with segmentation and churn prediction, ultimately enhancing operational efficiency and customer retention.

Keywords**:** Customer churn, churn prediction, telecom operations, machine learning, feature selection, customer segmentation, operational efficiency.

CONTENTS

Title Page		I
Cover Page		II
Declaration		III
Certificate by Supervisor		IV
Approval Sheet		V
Acknowledgement		VI
Abstract		VII
Contents		VIII
List of Figures		IX
List of Abbreviations		X
Chapter 1	Introduction	1-8
	1.1 Introduction	1-2
	1.2 Problem Statement	2
	1.3 Motivation	2-3
	1.4 Objectives	3-4
	1.5 Requirements Specification	4-8
Chapter 2	Literature Review	9-11
Chapter 3	Existing Mechanisms and Proposed System	12-16
	3.1 Limitations of Existing Mechanisms	13-14
	3.2 Proposed System	15-16
Chapter 4	Churn Prediction Model	17-36
	4.1 Advantages of Churn Prediction Model	18
	4.2 Methodology	19-20
	4.3 System Architecture	20-24
	4.4 Algorithms	24-33
	4.5 Sample Code	33-36
Chapter 5	Results and Discussion	37-40
Chapter 6	Conclusion and Future Scope	41-43
	6.1 Conclusion	41-42
	6.2 Future Scope	43
Chapter 7	References	44-46
List of Publications		47

LIST OF FIGURES

FIG.NO	FIG.NAME	PG.NO
4.2.1	Telecom customer churn data set	19
4.3.1	System Architecture	20
4.4.1	Random Forest Model	26
5.1	Registration Page	37
5.2	Login Page	38
5.3	Prediction Page	38
5.4	Results Page	39

LIST OF ABBREVIATIONS

ML	Machine Learning
AI	Artificial Intelligence
RF	Random Forest
LR	Logistic Regression
XGBoost	Extreme Gradient Boosting
GBM	Gradient Boosted Machine Trees
CRM	Customer Relationship Management
SQL	Structural Query language

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In the fiercely competitive telecommunication industry, customer churn—when customers discontinue their service subscriptions—poses a significant challenge. Acquiring new customers often costs more than retaining existing ones, making churn reduction crucial for telecom operators. High churn rates can erode profitability and market share, highlighting the need for effective churn prediction and management strategies. Machine learning (ML) provides a transformative solution to this challenge. By leveraging vast amounts of historical customer data, such as service usage patterns, demographic details, and interaction histories, ML models can predict the likelihood of a customer discontinuing their service. This predictive capability enables telecom companies to proactively engage with at-risk customers, offering personalized incentives or improving service quality to retain them.

Developing a churn prediction model involves several key steps: data collection and preprocessing, feature selection and engineering, model training, and evaluation. Data from various sources such as call records, billing information, customer service interactions, and social media can be integrated to form a comprehensive view of each customer's behavior. Advanced ML algorithms, including logistic regression, decision trees, random forests, and neural networks, can analyze these data points and identify patterns indicative of churn. Effective churn prediction not only helps in retaining customers but also provides insights into the underlying reasons for churn. These insights can drive strategic decisions in product development, customer service enhancements, and targeted marketing campaigns.

Moreover, reducing churn rates translates directly into increased customer lifetime value and a more stable revenue base. This project aims to build a scalable and accurate churn prediction system that can be seamlessly integrated into a telecom operator's existing infrastructure. By doing so, telecom companies can shift from reactive to proactive

customer relationship management, enhancing customer satisfaction and loyalty in a highly competitive market.

1.2 PROBLEM STATEMENT

The telecommunication industry faces significant challenges due to customer churn, where subscribers discontinue their services, leading to revenue loss and increased customer acquisition costs. This project aims to develop a machine learning model to predict customer churn by analyzing user behavior, service usage patterns, and demographic data. Accurate churn prediction enables telecom companies to proactively engage at-risk customers with retention strategies, such as personalized offers or service improvements. The goal is to enhance customer satisfaction, reduce churn rates, and ultimately increase profitability. The solution involves using various algorithms to process historical data and identify key indicators of churn.

1.3 MOTIVATION

Customer churn presents a significant challenge in the telecommunications sector, mainly because it incurs substantial financial and operational costs. Retaining a current customer is generally far more economical than acquiring a new one. The process of attracting new customers involves substantial expenses, from advertising and promotions to competitive pricing strategies, making it a costly alternative to retention. Due to these high costs, companies place great importance on customer retention, seeking ways to sustain and nurture existing customer relationships.

Loyal customers are invaluable because they contribute to consistent revenue streams and often serve as informal brand ambassadors. Over time, satisfied customers may increase their service usage, upgrade plans, or add additional services, all of which contribute to stable revenue growth. Additionally, they often refer others to the company, creating organic growth without the need for intensive marketing campaigns. These combined benefits make loyal customers a crucial asset to any telecom provider.

Churn prediction has therefore become an essential component of customer relationship management (CRM) strategies in telecom. By anticipating potential churn, companies can proactively engage with at-risk customers through tailored interventions, resolving issues or providing incentives that encourage them to stay. Churn prediction models leverage machine learning to analyze customer data, identifying patterns and behaviors that indicate a likelihood of leaving. Armed with these insights, companies can implement specific retention measures, such as personalized offers or enhanced customer support, improving retention rates and overall customer satisfaction.

The motivation behind this project lies in addressing churn by implementing machine learning models for precise customer segmentation and accurate churn prediction. Segmentation allows telecom providers to classify customers into distinct groups based on behavioral and demographic similarities. This categorization provides an opportunity to understand each segment's unique preferences and needs, enabling more targeted retention efforts. By integrating segmentation with churn prediction, this project aims to equip telecom companies with actionable insights that allow them to address customer needs effectively and retain their competitive standing.

1.4 OBJECTIVES

The central goal of this project is to create a sophisticated machine learning model that can reliably identify customers at a high risk of leaving. By examining a variety of historical customer data—such as interaction records, usage trends, and demographic profiles—the model aims to detect critical signs that suggest a customer might churn. This capability will empower telecom companies to take preemptive action, engaging with vulnerable customers early enough to implement strategies that can effectively deter them from discontinuing their service.

In addition to predicting churn, this project aims to enhance retention initiatives through personalized interventions that cater to the unique preferences and behaviors of different customer segments. By employing advanced segmentation techniques, the model will facilitate the design of targeted retention strategies that resonate with specific groups,

ultimately improving customer engagement and satisfaction. This approach not only ensures a more personalized experience for customers but also allows telecom operators to allocate their marketing and support resources more strategically, directing efforts toward those customers who stand to benefit most from tailored engagement.

Moreover, the project intends to boost the overall efficiency of customer retention operations. With precise churn predictions, telecom companies can better manage their resources by focusing on high-risk cases rather than resorting to broad, one-size-fits-all retention campaigns. This focused strategy is expected to reduce unnecessary expenditures while enhancing the effectiveness of customer service efforts and increasing customer loyalty.

The overarching aim is to drive revenue growth by minimizing customer churn and cultivating long-term relationships with satisfied customers. By incorporating predictive analytics into their retention strategies, telecom providers can not only decrease attrition rates but also enhance their capacity to generate consistent revenue through loyal customer bases. This, in turn, reinforces their competitive advantage within the industry.

In conclusion, this project seeks to develop a powerful machine learning model that enables telecom companies to predict customer churn accurately, devise effective retention strategies through tailored interventions, enhance operational efficiency, and ultimately maximize revenue generation. By taking a proactive, data-driven approach to churn management, the project aspires to strengthen customer relationships and significantly improve the overall performance of telecom operations.

1.5 REQUIREMENTS SPECIFICATIONS

1.5.1 SOFTWARE REQUIREMENTS

Software requirements delineate the essential software resources and dependencies that must be in place for an application to operate optimally. These requirements often include specific libraries, frameworks, and tools that need to be installed separately prior to the application installation. Such dependencies are not typically bundled within the installation

package and must be managed independently to ensure a smooth installation and functioning of the software.

- **Platform Specifications:** A platform in computing refers to a foundational framework—either in hardware or software—that supports the execution and development of software applications. This encompasses various elements, including:
- **Computer Architecture:** The design and organization of the components of a computer system, which can affect how software interacts with hardware.
- **Operating Systems:** The system software that manages computer hardware and provides services for computer programs. Key examples include Microsoft Windows, Linux distributions, and macOS.
- **Programming Languages:** The languages in which the software is developed, alongside their respective runtime libraries that provide essential functionalities during execution.

When defining software requirements, compatibility with the operating system is of paramount importance. Each application must align with the intended operating system, considering factors such as the version being used. While many modern operating systems strive for backward compatibility, it is not always guaranteed. For example, software designed for Windows XP may not function on Windows 98, although applications running on newer versions might exhibit some backward compatibility with older versions. Similarly, software that leverages features exclusive to newer Linux kernel versions (like v2.6) typically fails to operate correctly on older kernel versions (like v2.2 or v2.4).

APIs and Drivers

Applications that heavily interact with specialized hardware devices, such as high-performance graphics cards or sound devices, often require specific Application Programming Interfaces (APIs) or updated device drivers. APIs provide the necessary protocols and tools for building software applications, enabling communication between software and hardware.

A notable example is DirectX, which encompasses a collection of APIs designed for multimedia tasks, particularly within gaming on Microsoft platforms. DirectX facilitates

the handling of tasks related to graphics, sound, and input, allowing developers to create rich multimedia experiences without needing to write device-specific code.

Web Browser Considerations

In the context of web applications and software that relies on internet technologies, the choice of web browser installed on the system can significantly influence functionality. Many applications default to utilizing the system's pre-installed web browser, which can affect compatibility and security.

For instance, Microsoft Internet Explorer h as historically been the go-to browser for software operating on Windows systems, despite its vulnerabilities associated with features like ActiveX controls, which have been known to introduce security risks. With the evolution of web standards and security practices, many applications now strive for compatibility with multiple browsers, including more modern options like Google Chrome, Mozilla Firefox, and Microsoft Edge, to enhance user experience and security.

Development Tools and Environment

- **Coding Language:** Python is the primary programming language, known for its readability, versatility, and robust community support. It is widely used in data science, web development, and machine learning applications.
- **Tool:** Anaconda serves as the development environment, providing a comprehensive package management system and environment management capabilities tailored for scientific computing and data analysis. Anaconda simplifies package installation and dependency management, making it an ideal choice for projects that require multiple libraries and tools.
- **Interface:** The project will utilize Flask and Jupyter Notebook as interfaces. Flask is a micro web framework for Python that allows developers to build web applications quickly and efficiently. Jupyter Notebook, on the other hand, is an open-source web application that enables the creation and sharing of documents containing live code, equations, visualizations, and narrative text. This combination allows for a seamless development experience, particularly for data-driven applications.

1.5.2 HARDWARE REQUIREMENTS

In the context of software applications and operating systems, hardware requirements refer to the essential physical resources necessary for optimal performance and functionality. Alongside these hardware specifications, it is common to provide a Hardware Compatibility List (HCL), particularly for operating systems. This list enumerates devices that have been tested for compatibility, as well as those that may not function properly with a specific operating system or application. Below, we explore the various aspects of hardware requirements in greater detail.

Architecture

Computer operating systems are specifically designed to function on certain computer architectures. While there are applications that can operate independently of the underlying platform, the majority of software is closely tied to particular operating systems and their corresponding architectures. This means that even though some systems can be architecture-independent, they often require recompilation or modification to run on a different architecture. Understanding the architecture is essential for ensuring that the software can leverage the capabilities of the hardware it is intended to run on.

Processing Power

The processing power of the Central Processing Unit (CPU) is a critical requirement for any software application. For programs running on x86 architecture, processing power is typically characterized by the CPU's model and clock speed. A more powerful CPU can significantly enhance performance, allowing for quicker data processing and improved responsiveness of applications. In this context, the Intel® Core™ i3-7020U CPU, operating at a clock speed of 2.30 GHz, represents a capable choice for handling basic software tasks and applications, making it suitable for various computing needs.

Memory

Random Access Memory (RAM) is another crucial component that impacts the execution of software. Every application consumes a portion of the system's RAM when it is running, which is essential for maintaining smooth performance. Memory requirements are

determined by the software's demands, the operating system, any additional supporting applications, and the files being processed simultaneously. A system equipped with 4 GB of RAM can effectively manage lightweight applications and basic multitasking; however, resource-intensive software may require additional memory for optimal performance.

Secondary Storage

The requirements for secondary storage, typically involving hard disk space, can vary widely based on multiple factors. These include the total size of the software installation, the temporary files generated during the installation or operation of applications, and any swap space usage that may be necessary if the available RAM is insufficient. A hard disk with a capacity of 1 TB provides ample space for software installations, data storage, and temporary files, making it a robust choice for most users. This capacity ensures that the system can handle a significant amount of data without running into storage limitations.

CHAPTER 2

LITERATURE REVIEW

In the competitive landscape of the telecommunications industry, customer churn—defined as the rate at which customers stop doing business with a company—poses significant challenges and risks. Customer churn prediction is a significant area of focus in the telecom industry due to the direct impact of customer retention on profitability and competitive advantage. Understanding the dynamics of customer behavior, satisfaction, and loyalty is vital for telecom operators aiming to reduce churn rates. A considerable body of research has emerged, employing various methodologies and techniques to analyze the factors influencing customer attrition.

Early studies laid the groundwork for understanding the relationship between customer satisfaction and retention, revealing that high customer satisfaction is linked to increased loyalty and reduced churn [1]. These findings emphasized that telecom companies must prioritize customer satisfaction to mitigate churn effectively. This foundational work has been widely cited as a basis for further research into churn prediction methodologies [2]. Innovative approaches to churn prediction have been introduced by applying data mining techniques to telecommunications call detail records [3]. Significant insights can be gleaned from call data, effectively transforming raw data into actionable intelligence for predicting churn. This shift towards data-driven strategies in churn management highlights the potential of leveraging customer behavior data [4].

Recent developments have seen the emergence of machine learning-based churn prediction models that utilize various algorithms to identify customers at risk of leaving [5]. This research showcases the effectiveness of machine learning in enhancing the accuracy of churn predictions, enabling telecom operators to implement targeted retention strategies [6]. Such advancements underline the growing reliance on sophisticated analytical methods to address customer retention challenges [7].

The need for proactive measures in preventing churn has also been examined, with studies investigating the effectiveness of tailored customer interactions [8]. Experimental findings indicate that proactive engagement with customers can significantly lower churn rates, emphasizing the importance of personalized marketing in the telecom, where customized offerings enhance customer loyalty [9].The application of multilayer perceptron neural networks for predicting customer churn has demonstrated the potential of neural network architectures in accurately modeling churn behavior [10]. This research offers insights into the underlying patterns that drive customer decisions and has contributed to the growing interest in artificial intelligence techniques for churn prediction [11]. Comprehensive surveys have reviewed various datasets, methodologies, and performance metrics used in churn prediction, serving as valuable resources for researchers and practitioners [12]. These reviews provide an overview of the state-of-the-art techniques available for analyzing churn and highlight the importance of continually evolving methodologies to keep pace with changing customer behaviors and market dynamics [13].

Addressing class imbalance in churn prediction has been a crucial topic, as traditional predictive models often struggle with imbalanced datasets. Effective strategies for handling this issue are particularly relevant in the telecom sector, where the number of churned customers is often lower than retained ones. Utilizing specialized techniques for imbalanced data has been shown to enhance model performance and reliability [14][15].The influence of big data analytics on churn prediction has underscored how integrating vast amounts of customer data can significantly improve predictive capabilities [16]. Advanced analytical techniques enable telecom companies to derive insights from complex datasets, ultimately leading to better decision-making processes and enhanced retention strategies. This perspective aligns with the growing importance of big data in various industries, including telecommunications [17].

Several studies have investigated the comparative effectiveness of different predictive models for churn prediction, offering insights into which models yield the best performance in churn scenarios [18]. The findings emphasize the necessity for telecom companies to evaluate multiple modeling approaches to determine the most effective strategy for their specific customer base [19].The role of big data analytics in churn prediction highlights its

transformative impact on model accuracy and predictive capabilities[20]. By effectively analyzing customer behavior, companies can enhance their retention strategies and improve overall customer satisfaction [21]. Customer feedback plays a critical role in churn prediction, as understanding customer sentiment can inform more effective retention strategies[22]. Integrating customer insights into churn prediction models enhances the relevance and applicability of the findings, ensuring that retention strategies align with customer expectations [23].

Ethical considerations surrounding data usage have become increasingly relevant, particularly in the context of customer churn prediction. The implications of data privacy and the ethical challenges that arise from leveraging customer data for predictive analytics must be addressed [24]. Striking a balance between the benefits of data-driven strategies and the need for ethical data management practices is essential for maintaining customer trust [25].In conclusion, the literature on customer churn prediction in the telecom industry reveals a rich tapestry of research that spans various methodologies, technologies, and insights [26]. The integration of machine learning techniques, the focus on handling class imbalance, and the importance of customer feedback are central themes that emerge from this body of work [27]. The ongoing advancements in big data analytics continue to shape the landscape of churn prediction, providing telecom operators with the tools necessary to mitigate customer attrition effectively [28]. As the industry evolves, further research will be essential to address emerging challenges and refine predictive models to adapt to the ever-changing dynamics of customer behavior [29].

In summary, managing customer churn in the telecommunications sector necessitates a comprehensive approach that integrates advanced analytics, customer segmentation, and proactive retention strategies. By leveraging machine learning and big data technologies, telecom companies can build robust churn prediction models that yield valuable insights into customer behavior [30]. Nonetheless, challenges related to data integration, privacy regulations, and the ever-changing technological landscape must be addressed to unlock the full potential of churn prediction initiatives. As the industry evolves, ongoing investment in technology, collaboration, and ethical data practices will be paramount for telecom operators aiming to enhance customer loyalty and drive sustainable revenue growth.

CHAPTER 3
EXISTING MECHANISMS AND PROPOSED SYSTEM

Rule-Based Systems:

Rule-based systems serve as a foundational approach in predicting customer churn by utilizing a set of predefined rules and thresholds developed through industry expertise to identify at-risk customers. For instance, a rule may categorize a customer as likely to churn if they have not engaged with the service for a certain number of days. This methodology offers a clear and interpretable mechanism for flagging customers who may need intervention, allowing companies to concentrate their retention efforts more effectively.

However, rule-based systems exhibit several significant drawbacks. A major limitation is their static nature; once established, the rules do not readily adapt to shifts in customer behavior or changes in market dynamics. As customer preferences evolve or competitive landscapes shift, the original rules may quickly become outdated. Furthermore, these systems often struggle to account for the complexity inherent in customer behavior, as they typically overlook intricate, non-linear relationships within the data that play a crucial role in influencing churn. Additionally, maintaining the relevance of these rules requires ongoing input from domain experts, resulting in a resource-intensive process that demands regular updates and revisions to ensure effectiveness over time.

Statistical Methods:

In contrast, statistical methods, including logistic regression and survival analysis, are commonly employed in churn management to model the probability of customer attrition based on historical data. These techniques leverage statistical principles to establish connections between customer characteristics and the likelihood of churn, providing a quantitative basis for predicting behavior.

Despite their strengths, statistical methods also encounter significant challenges. Many models operate under the assumption of linear relationships, which can lead to oversimplified interpretations of the data. This reliance on simplifications often fails to

capture the more complex interactions that characterize customer behavior. Additionally, scalability presents a notable issue when applying these methods to large, high-dimensional datasets typical of the telecommunications industry. As the volume and intricacy of data increase, traditional statistical approaches may struggle to deliver accurate predictions.

In summary, while both rule-based systems and statistical methods offer valuable frameworks for understanding and predicting customer churn, their limitations underscore the necessity for more advanced analytical techniques. By addressing the inflexibility associated with rule-based approaches and the simplifying assumptions of statistical models, telecom companies can enhance their predictive capabilities and improve customer retention strategies. Embracing more sophisticated methods, such as machine learning algorithms, can enable organizations to better capture the complexities of customer behavior and generate actionable insights that drive effective retention efforts.

3.1 LIMITATIONS OF EXISTING MECHANISMS

- **Limited Predictive Accuracy:**

  Current systems for managing customer churn in the telecommunications industry face several notable limitations that impede their overall effectiveness. A primary concern is the limited predictive accuracy of these traditional methods. Many of them struggle to effectively forecast churn due to their challenges in processing large datasets and capturing the intricate relationships among various customer characteristics. This often results in either overlooking potential churners or mistakenly classifying loyal customers as being at risk.

- **Reactive Rather Than Proactive:**

  A significant number of existing systems operate in a reactive manner. They typically respond to churn signals only after customers exhibit clear signs of dissatisfaction or disengagement. This reactive stance severely restricts the opportunities for companies to intervene proactively and address issues before customers decide to leave.

- **High False Positive and False Negative Rates:**

  Another critical drawback lies in the tendency of rule-based and basic statistical models to produce high rates of false positives and false negatives. False positives occur when customers who are not at risk of churning are incorrectly identified as potential churners. Conversely, false negatives happen when actual churn risks are overlooked. These inaccuracies can lead to poor allocation of resources in retention efforts, causing companies to expend time and resources targeting customers who do not need intervention while neglecting those who truly require attention.

- **Inability to Personalize Interventions:**

  Many existing systems lack the ability to provide personalized interventions. Without detailed insights into individual customer behaviors and preferences, traditional approaches cannot effectively customize retention strategies to suit the unique needs of each customer. This often leads to generic interventions that fail to resonate with customers and are frequently ineffective.

- **Time and Resource Intensive:**

  The manual processes involved in setting up and maintaining rule-based systems, combined with the computational demands of traditional statistical methods, contribute to inefficiencies. These processes are often resource-intensive and time-consuming, ultimately reducing operational efficiency.

- **Poor Adaptability:**

  Adaptability is a crucial issue for existing systems. Many of these systems struggle to adjust to rapid shifts in customer behavior, market dynamics, or the introduction of new products. This lack of flexibility can diminish their effectiveness over time, as the industry landscape continues to evolve.

In conclusion, these limitations underscore the pressing need for more advanced analytical methodologies that can improve predictive accuracy, enable early interventions, and facilitate personalized retention strategies within the telecommunications sector.

3.2 PROPOSED SYSTEM

To address the critical issue of client turnover in the telecommunications sector, we propose an advanced churn prediction model. This model is designed to enable telecom providers to proactively identify customers who are at high risk of churning, thereby facilitating effective retention strategies. Our approach leverages advanced machine learning techniques within a robust data environment to ensure accurate and reliable predictions. The system involves a thorough analysis of customer data and employs an innovative method for feature design and selection.

This approach helps to pinpoint the most significant factors influencing churn and incorporates them into the predictive model. By analyzing various data scenarios and visualizing the results, we offer valuable insights into customer behavior and the primary drivers of churn. The model is trained, evaluated, and validated using a comprehensive dataset that includes detailed customer information from recent months. We assess the effectiveness of different predictive methods by testing four algorithms: Random Forest, Logistic Regression, Extreme Gradient Boosting (XGBoost), and Gradient Boosted Machine Trees (GBM). Parameter tuning is applied to these algorithms to further enhance the model's performance.

The resulting churn prediction model provides telecom providers with a powerful tool to anticipate and mitigate customer churn. This proactive approach allows companies to gain deeper insights into customer behavior, implement targeted retention strategies, and ultimately improve profitability.

To enhance telecom operations, a comprehensive approach was employed, emphasizing precise churn prediction and operational efficiency. The process starts with gathering an extensive dataset that includes a wide range of customer information, such as demographic details, usage patterns, and historical churn data. This approach ensures that the churn prediction model is robust, reliable, and capable of providing actionable insights for enhancing telecom operations.

The proposed churn prediction system establishes a robust foundation for future

improvements that can greatly enhance telecom providers' capabilities in managing and reducing customer churn. A significant opportunity for advancement is the integration of real-time data streams, which would facilitate immediate analysis and prediction of churn. This capability would enable telecom companies to respond promptly to churn signals and implement effective measures as they arise.

CHAPTER 4
CHURN PREDICTION MODEL

To tackle the pressing issue of client turnover in the telecom industry, our proposed system aims to develop an advanced churn prediction model. This model is designed to help telecom providers proactively identify customers who are most likely to churn, thereby enabling effective retention strategies.

Our approach utilizes sophisticated machine learning techniques within a large data environment to ensure high prediction accuracy and reliability.The system includes a comprehensive analysis of customer data, incorporating a novel method for feature design and selection. This ensures that the most relevant factors contributing to churn are effectively identified and used. By examining various data scenarios and presenting the results graphically, we provide clear insights into customer behavior and the key drivers of churn.

We train, evaluate, and validate the model using an extensive dataset that includes detailed customer information from the past months. To determine the most effective predictive approach, we test four different algorithms: Random Forest, Logistic Regression, Extreme Gradient Boosting (XGBoost), and Gradient Boosted Machine Trees (GBM). Additionally, we apply parameter tuning to these algorithms during the training process to further enhance the model's accuracy.

The outcome is a sophisticated churn prediction model that significantly improves telecom providers' ability to anticipate and address customer churn. This proactive strategy enables telecom companies to better understand their customers, implement targeted retention measures, and ultimately enhance their profitability.

4.1 ADVANTAGES OF CHURN PREDICTION MODEL:

The proposed system offers several significant advantages over existing systems in managing customer churn in the telecom industry. Firstly, the use of advanced machine learning techniques provides a more accurate and reliable prediction of customer churn compared to traditional statistical models. These techniques allow for the dynamic adaptation to changing customer behaviors, which static rule-based systems fail to address.

Secondly, the novel approach to feature design and selection in our proposed system ensures that the most relevant factors contributing to churn are identified and utilized effectively. This contrasts with the often limited and predefined features used in existing systems, leading to more precise churn predictions.

Thirdly, the comprehensive analysis of customer data, combined with graphical representation of various data scenarios, offers clear insights into customer behavior. This level of detail is not typically provided by current systems, which rely heavily on basic analytics and subjective customer feedback.

Moreover, the proposed system addresses the issue of data silos by integrating data across different platforms, providing a holistic view of customer behavior. This integration is crucial for accurate churn prediction and is a notable improvement over the fragmented data handling seen in existing systems.Our model's use of extensive datasets and parameter tuning during the training process enhances its accuracy and effectiveness. This proactive approach contrasts with the reactive nature of current systems, which often identify churn risks only after significant changes have occurred. In summary, the proposed system's advanced machine learning methods, novel feature design, comprehensive data analysis, integrated data handling, and proactive approach offer substantial improvements over existing systems.

4.2 METHODOLOGY

In this project, we aim to predict customer churn by analyzing comprehensive customer data, including usage patterns, service interactions, and demographic factors. We start by collecting, cleaning, and preprocessing the data to ensure its quality. Next, we apply feature engineering to extract and transform relevant features. To handle data imbalance, we use methods such as under-sampling and tree-based algorithms. We split the data into 80% for training and 20% for validation, optimizing model performance through hyperparameter tuning. Our system incorporates ensemble learning and deep learning techniques to improve predictive accuracy and adaptability. Continuous learning capabilities enable the model to update dynamically with evolving data trends. This scalable and robust system allows telecom companies to implement personalized retention strategies, enhancing customer satisfaction and reducing churn rates.

	A	B	C	D	E	F	G	H	I	J
1	Dependent	tenure	OnlineSecu	OnlineBack	DeviceProt	TechSuppo	Contract	PaperlessB	MonthlyCh	TotalCharges
2	No	1	No	Yes	No	No	Month-to-r	Yes	29.85	29.85
3	No	34	Yes	No	Yes	No	One year	No	56.95	1889.5
4	No	2	Yes	Yes	No	No	Month-to-r	Yes	53.85	108.15
5	No	45	Yes	No	Yes	Yes	One year	No	42.3	1840.75
6	No	2	No	No	No	No	Month-to-r	Yes	70.7	151.65
7	No	8	No	No	Yes	No	Month-to-r	Yes	99.65	820.5
8	Yes	22	No	Yes	No	No	Month-to-r	Yes	89.1	1949.4
9	No	10	Yes	No	No	No	Month-to-r	No	29.75	301.9
10	No	28	No	No	Yes	Yes	Month-to-r	Yes	104.8	3046.05
11	Yes	62	Yes	Yes	No	No	One year	No	56.15	3487.95
12	Yes	13	Yes	No	No	No	Month-to-r	Yes	49.95	587.45
13	No	16	No interne	No interne	No interne	No interne	Two year	No	18.95	326.8
14	No	58	No	No	Yes	No	One year	No	100.35	5681.1
15	No	49	No	Yes	Yes	No	Month-to-r	Yes	103.7	5036.3
16	No	25	Yes	No	Yes	Yes	Month-to-r	Yes	105.5	2686.05
17	Yes	69	Yes	Yes	Yes	Yes	Two year	No	113.25	7895.15
18	No	52	No interne	No interne	No interne	No interne	One year	No	20.65	1022.95
19	Yes	71	Yes	No	Yes	No	Two year	No	106.7	7382.25
20	Yes	10	No	No	Yes	Yes	Month-to-r	No	55.2	528.35
21	No	21	No	Yes	Yes	No	Month-to-r	Yes	90.05	1862.9

Fig.4.2.1 Telecom customer churn dataset

To implement this project, we have designed following modules:

Data Collection

In this project, we utilize a telecom customer dataset sourced from Kaggle. This dataset includes various features related to customer connections and services, with labels indicating churn (0 for no churn and 1 for churn).

Pre-processing

Our dataset comprises 21 features, and we aim to evaluate and select the most impactful ones for training. Feature engineering and selection methods are used to identify key features, ensuring all 21 are utilized effectively. The data is then split into 80% for training and 20% for testing. The training data includes both features and labels, while the testing data is reserved for evaluating the machine learning model.

Train-Test Split and Model Fitting

We split our dataset into training and testing sets to assess the model's performance on unseen data and to evaluate its generalization capabilities. This is followed by model fitting, a crucial step in the model-building process, where the machine learning model is trained on the training data and then evaluated using the testing data.

4.3 SYSTEM ARCHITECTURE

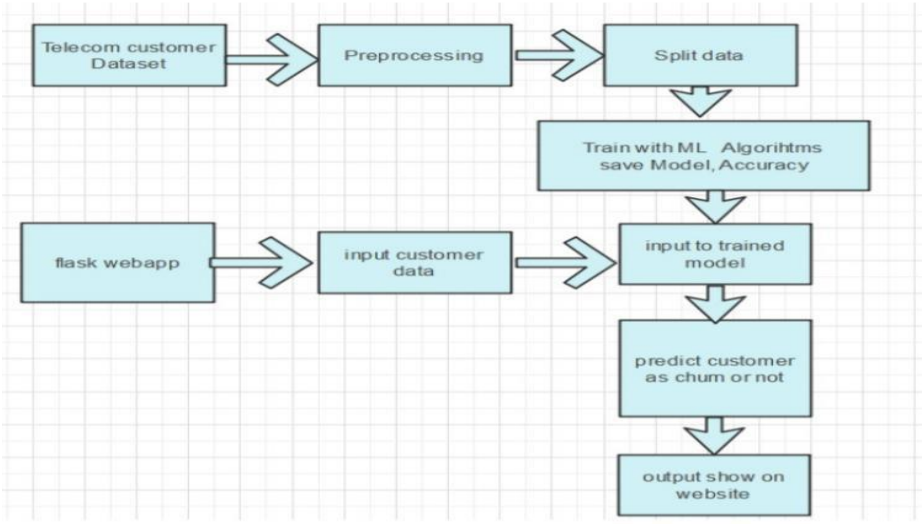

Fig 4.3.1 System Architecture

1. Preprocess the Dataset

The first step in the data preparation process is to clean and prepare the telecom customer dataset, which is crucial for ensuring that the data is suitable for machine learning applications. This preprocessing phase involves several key activities:

- **Handling Missing Values:** Identifying and addressing any missing values in the dataset is essential, as these can adversely affect the performance of machine learning models. Common strategies include removing records with missing values, imputing missing data using mean, median, or mode, or employing more complex methods such as K-nearest neighbors (KNN) imputation, depending on the context and the extent of the missing data.
- **Adjusting Data Types:** Data types must be checked and adjusted as necessary to ensure compatibility with machine learning algorithms. For example, categorical variables might need to be converted from string format to numerical format, which is often achieved through label encoding or one-hot encoding. Numerical features may also require normalization or standardization to ensure that they are on a comparable scale.
- **Encoding Categorical Variables:** Machine learning models typically require numerical input. Therefore, categorical variables must be encoded. Label encoding assigns each category a unique integer value, while one-hot encoding creates binary columns for each category, ensuring that the model can interpret these features appropriately without introducing bias related to the order of categories.

2.Split Data

Once the dataset has been preprocessed, it is essential to split the data into training and testing subsets. This division is crucial for effectively evaluating the performance of the machine learning model.

- **Training Set:** This subset is used to train the machine learning model, allowing it to learn patterns and relationships in the data. Typically, a larger portion of the dataset (often around 70-80%) is allocated for training to ensure that the model has enough data to generalize well.

- **Testing Set:** The remaining portion of the data (usually 20-30%) is reserved for testing the model. This testing set is crucial for evaluating how well the model can predict outcomes on unseen data, which is a strong indicator of its performance in real-world scenarios. This split helps to mitigate overfitting, where the model performs well on training data but poorly on new, unseen data.

3.Train with ML Algorithms and Save Model

With the data split into training and testing sets, the next step involves training the machine learning model using suitable algorithms. One commonly used algorithm in this context is the Random Forest Classifier.

- **Model Training:** During training, the Random Forest algorithm constructs multiple decision trees based on different subsets of the training data. It aggregates the predictions from these trees to improve accuracy and reduce the risk of overfitting. This ensemble learning method is particularly effective for handling large datasets with complex interactions between features.
- **Model Evaluation:** After training, the model's accuracy is evaluated using the testing dataset. Key performance metrics such as accuracy, precision, recall, and the F1 score can be computed to assess how well the model performs in predicting churn. This evaluation phase helps identify any adjustments needed to improve model performance.
- **Saving the Model:** Once the model is trained and evaluated, it is saved for future use. Using libraries such as joblib, the trained model can be serialized and stored on disk. This allows for easy retrieval and deployment in a web application without the need to retrain the model each time it is used.

4.Flask Web Application

To create an interactive user experience, a Flask-based web application is developed. This application allows users to input customer data through a web form.

- **Web Application Development:** The Flask framework is chosen for its simplicity and ease of use in building web applications. The application consists of a front-end interface where users can enter their data, and a back-end that processes the input and returns predictions.

5.User Input Handling

After users submit their data via the web form, the application must handle the input efficiently.

- **Data Processing:** The Flask application processes the input data to ensure that it aligns with the model's input requirements. This includes validating the data types, checking for missing values, and applying any necessary transformations to the data to prepare it for analysis.

6.Model Integration

Once the user input has been processed, the next step is to integrate the pre-trained machine learning model into the web application.

- **Analyzing Input Data:** The integrated model receives the processed input data and analyzes it based on the patterns it learned during training. This step is critical for predicting the likelihood of churn effectively.

7.Prediction

Based on the input data and the learned patterns from the training phase, the model makes a prediction regarding whether the customer is likely to churn.

- **Churn Prediction:** The model utilizes its internal decision-making process to evaluate the input data and generate a prediction. This prediction indicates the probability of churn, providing valuable insights into the customer's risk status.

8.Output Display

Finally, the results of the prediction are presented to the user through the web interface.

- **User Feedback:** The application displays the predicted churn likelihood or classification outcome in a clear and comprehensible manner. This output not only informs the user about the churn risk associated with the input data but can also provide recommendations for retention strategies based on the model's insight.

4.3.1 Functional Requirements and Non-Functional Requirements

Functional Requirements:

1. User Input
2. Data Processing
3. Model Integration
4. Prediction Output

Non-Functional Requirements:

1. Usability requirement
2. Data Integrity requirement
3. Maintainability requirement
4. Security requirement
5. Reliability requirement
6. Service ability requirement

4.4 ALGORITHMS:

The algorithm includes all customer data from previous months. We tested four different algorithms: Logistic Regression, Extreme Gradient Boosting (XGBoost), Gradient Boosted Machine (GBM) Trees, Random Forests, and Decision Trees. To increase the model's accuracy, the algorithms are trained with parameter tuning. This work focuses on optimizing telecom operations with segmentation and churn prediction, ultimately enhancing operational efficiency and customer retention.

Logistic Regression model is utilized to analyze the data using a training dataset while validating its performance with a testing dataset. The model is configured with specific settings, including a regularization strength (C=150) and a maximum iteration limit of 150 for optimization. Initially, the model is trained by fitting it to the training data (X_train and y_train), enabling it to learn patterns and relationships within the data. After training, the model predicts outcomes for the testing dataset (X_test). The accuracy of these predictions is then calculated, indicating how well the model performs, with a reported training accuracy of 80 percent. This means that the model correctly classified 80 percent

of the instances in the training data. To further evaluate the model's effectiveness, additional metrics such as the confusion matrix and classification report are generated, offering detailed insights into the model's precision, recall, and overall classification performance.

PSEUDO CODE:

```
# Initialize the Logistic Regression model with specified hyperparameters
Log_reg = LogisticRegression(
    C=150,   # Inverse of regularization strength; a smaller value specifies stronger
    regularization
    max_iter=150  # Maximum number of iterations for optimization
)

# Train (fit) the model on the training
dataset Log_reg.fit(X_train, y_train)

# Make predictions on the test dataset using the trained
model log_pred = Log_reg.predict(X_test)

# Calculate the accuracy score by comparing predictions with
actual test labels accuracy = accuracy_score(log_pred, y_test)

# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result = confusion_matrix(log_pred, y_test)

# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(log_pred, y_test)

# Print the accuracy score, confusion matrix, and classification report to
interpret results
 PRINT "Accuracy score:", accuracy
```

PRINT "Confusion matrix:", confusion_matrix_result
PRINT "Classification report:", classification_report_result

The Random Forest Classifier is an ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. In this setup, 120 trees are built (`n_estimators=120`), each using the Gini impurity (`criterion='gini'`) to decide splits. The model limits each tree's depth to 15 (`max_depth=15`) to avoid overfitting and requires at least 10 samples in each leaf (`min_samples_leaf=10`) and 5 samples to split a node (`min_samples_split=5`).

The classifier is trained on `X_train` and `y_train` using `Rfc.fit`, and predictions are made on `X_test` with `Rfc.predict`. The accuracy score, confusion matrix, and classification report are calculated to assess performance, showing how well the model predicts outcomes and providing insights into its strengths and areas for improvement. This process enables a well-balanced model that can handle complex data without becoming overly complex itself.

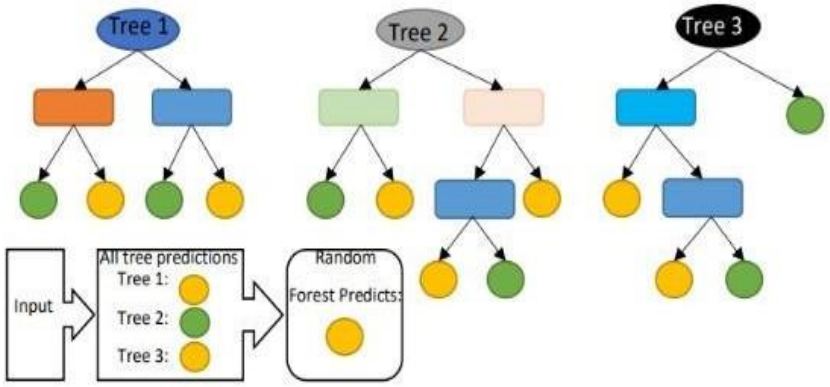

Fig.4.4.1 Random Forest Classifier

PSEUDO CODE:

```
# Initialize the Random Forest Classifier with specified hyperparameters
Rfc = RandomForestClassifier(
        n_estimators=120,          #Number of trees in the forest
```

```
    criterion='gini',           # Split criterion: Gini impurity
    max_depth=15,               # Maximum depth of each tree
    min_samples_leaf=10,        # Minimum samples required to be at a leaf node
    min_samples_split=5         # Minimum samples required to split an internal node
)

# Train (fit) the classifier on the training dataset Rfc.fit(X_train, y_train)

# Make predictions on the test dataset using the trained model
rfc_pred = Rfc.predict(X_test)

# Calculate the accuracy score by comparing predictions with
actual test labels accuracy = accuracy_score(rfc_pred, y_test)

# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result = confusion_matrix(rfc_pred, y_test)

# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(rfc_pred, y_test)

# Print the accuracy score, confusion matrix, and classification report to interpret results
    PRINT "Accuracy score:", accuracy
    PRINT "Confusion matrix:", confusion_matrix_result
    PRINT "Classification report:", classification_report_result
```

Decision Tree Classifier is employed to analyze the data, utilizing a training dataset to learn patterns and a testing dataset to evaluate its performance. The classifier is configured with specific parameters, such as using Gini impurity for splits, a random splitter method, and a requirement of at least 15 samples at each leaf node. The model is trained by fitting it to the training data, which allows it to develop rules for classification based on the input features (X_train and y_train). Once trained, the model makes predictions on the testing dataset (X_test), generating predicted labels for evaluation. The

accuracy of these predictions is then calculated, revealing that the model achieves a training accuracy of 80 percent. This indicates that the classifier correctly identifies 80 percent of the instances in the training dataset. Additional performance metrics, including the confusion matrix and classification report, are generated to provide deeper insights into the model's classification effectiveness, highlighting aspects like precision, recall, and overall accuracy.

PSEUDO CODE:

```
# Initialize the Decision Tree Classifier with specified hyperparameters
Dtc = DecisionTreeClassifier(
    criterion='gini',          # Split criterion: Gini impurity for determining best splits
    splitter='random',         # Method for choosing the split at each node: random or
    best min_samples_leaf=15  # Minimum samples required to be at a leaf node
)

# Train (fit) the classifier on the training dataset
Dtc.fit(X_train, y_train)


# Make predictions on the test dataset using the trained model
dtc_pred = Dtc.predict(X_test)


# Calculate the accuracy score by comparing predictions with actual test labels
accuracy = accuracy_score(dtc_pred, y_test)


# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result = confusion_matrix(dtc_pred, y_test)


# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(dtc_pred, y_test)
```

```
# Print the accuracy score, confusion matrix, and classification report to interpret results
    PRINT "Accuracy score:", accuracy
    PRINT "Confusion matrix:", confusion_matrix_result
    PRINT "Classification report:", classification_report_result
```

To enhance the accuracy of the model, the dataset is adjusted to ensure an equal distribution of classes using the SMOTEENN function. Before applying this function, the class distribution shows 4,129 instances for class 0 and 1,505 instances for class 1. After applying the SMOTEENN function, the class distribution becomes more balanced, with class 1 now having 2,418 instances and class 0 having 2,132 instances. This equalization helps address any potential bias in the model due to the imbalanced data.

Subsequently, the balanced dataset is divided into training and testing sets. The model is then retrained using multiple algorithms to evaluate whether there is an improvement in accuracy. This approach allows for a comprehensive comparison of the various models, aiming to determine which algorithm performs best with the newly balanced data and achieves higher predictive accuracy.

The training and testing datasets are utilized after applying the SMOTE technique to enhance the balance of classes. The Decision Tree Classifier is employed to train the model, using the balanced training data as input and the corresponding testing data for evaluation. After training, the model achieves a remarkable accuracy of 91 percent, indicating a significant improvement compared to the 80 percent accuracy observed before applying SMOTE. This increase in accuracy demonstrates the effectiveness of using SMOTE to address class imbalance, allowing the Decision Tree model to learn more robust patterns and make better predictions on the testing dataset. The results highlight the positive impact of balancing the dataset on model performance.

PSEUDO CODE:

```
# Initialize the Decision Tree Classifier with specified hyperparameters for sampling
Dtc_sampling = DecisionTreeClassifier(
    criterion="gini",  # Split criterion: Gini impurity for determining the best splits
    random_state=100,                # Random seed for reproducibility of results
    max_depth=7,                # Maximum depth of the tree to prevent overfitting
    min_samples_leaf=15      # Minimum number of samples required to be at a leaf node
  )

# Train (fit) the classifier on the balanced training dataset with sampling
Dtc_sampling.fit(X_train_sap, y_train_sap)

# Make predictions on the balanced test dataset using the trained model
dtc_sampling_pred = Dtc_sampling.predict(X_test_sap)

# Calculate the accuracy score by comparing predictions with actual test labels
accuracy = accuracy_score(dtc_sampling_pred, y_test_sap)

# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result = confusion_matrix(dtc_sampling_pred, y_test_sap)

# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(dtc_sampling_pred, y_test_sap)

# Print the accuracy score, confusion matrix, and classification report to interpret results
    PRINT "Accuracy score:", accuracy
    PRINT "Confusion matrix:", confusion_matrix_result
    PRINT "Classification report:", classification_report_result
```

The training and testing datasets, which have been adjusted using the SMOTE technique for better class balance, are utilized to train a Logistic Regression classifier. The model takes the training data as input and the corresponding testing data for evaluation. After training, the model achieves an impressive accuracy of 91 percent, indicating a notable improvement from the 80 percent accuracy observed prior to applying SMOTE. This significant increase demonstrates the effectiveness of using SMOTE to address class imbalance, allowing the Logistic Regression model to learn more effectively and make more accurate predictions on the testing dataset. The results highlight how balancing the dataset can enhance model performance and reliability.

PSEUDO CODE:

```
# Initialize the Logistic Regression model with specified hyperparameters for sampling
Log_reg_sampling = LogisticRegression(
    C=10, # Inverse of regularization strength; a smaller value specifies
    stronger regularization
    max_iter=150    # Maximum number of iterations for optimization
)

# Train (fit) the model on the balanced training dataset with sampling
Log_reg_sampling.fit(X_train_sap, y_train_sap)

# Make predictions on the balanced test dataset using the trained model
Log_sampling_pred = Log_reg_sampling.predict(X_test_sap)

# Calculate the accuracy score by comparing predictions with actual test labels
accuracy = accuracy_score(Log_sampling_pred, y_test_sap)

# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result =confusion_matrix(Log_sampling_pred, y_test_sap)
```

```
# Print the accuracy score and confusion matrix to interpret results
PRINT "Accuracy score:", accuracy
PRINT "Confusion matrix:", confusion_matrix_result

# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(Log_sampling_pred, y_test_sap)

# Print the classification report to understand the model's performance in detail
PRINT "Classification report:", classification_report_result
```

The model utilizes the training data as input and the corresponding testing data for evaluation. Following the training process, the model achieves a remarkable accuracy of 95 percent. This represents a significant enhancement compared to the 80 percent accuracy recorded before the application of SMOTE. The increase in accuracy underscores the effectiveness of the SMOTE technique in addressing class imbalance, enabling the Gradient Boosting Classifier to better capture patterns in the data and make more accurate predictions on the testing dataset. This improvement highlights the importance of data balancing in enhancing model performance.

PSEUDO CODE:

```
# Initialize the Gradient Boosting Classifier with default hyperparameters
gbc = GradientBoostingClassifier()

# Train (fit) the classifier on the balanced training dataset with sampling
gbc.fit(X_train_sap, y_train_sap)

# Make predictions on the balanced test dataset using the trained model
pred = gbc.predict(X_test_sap)

# Calculate the accuracy score by comparing predictions with actual test labels
accuracy = accuracy_score(pred, y_test_sap)
```

```
# Generate the confusion matrix to assess the performance of the classifier
confusion_matrix_result = confusion_matrix(pred, y_test_sap)

# Print the accuracy score and confusion matrix to interpret results
    PRINT "Accuracy score:", accuracy
    PRINT "Confusion matrix:", confusion_matrix_result

# Create a detailed classification report (precision, recall, f1-score)
classification_report_result = classification_report(pred, y_test_sap)

# Print the classification report to understand the model's performance in detail
PRINT "Classification report:", classification_report
```

4.5 SAMPLE CODE:

```
from flask import Flask, render_template, request
import pickle
import numpy as np
from database import *
from sklearn.preprocessing import LabelEncoder
app = Flask(__name__,static_url_path='/static')

# Load the machine learning
model @app.route('/p')
def p():
return render_template('index.html')

@app.route('/')
def m():
return render_template('main.html')

@app.route('/l')
def l():
```

```
return render_template('login.html')

@app.route('/h')
def h():
return render_template('home.html')

@app.route('/r')
def r():
return render_template('register.html')

@app.route('/m')
def menu():
return render_template('menu.html')

@app.route("/register",methods=['POST','GET'])
def signup():
if request.method=='POST':
username=request.form['username']
email=request.form['email']
password=request.form['password']
status = user_reg(username,email,password)
if status == 1:
return render_template("/login.html")
else:
Return render_template("/register.html",m1="failed")

@app.route("/login",methods=['POST','GET'])
def login():
if request.method=='POST':
username=request.form['username']
password=request.form['password']
status = user_loginact(request.form['username'],
```

```
request.form['password']) print(status)
if status == 1:
return render_template("/home.html", m1="sucess")
else:
return render_template("/login.html", m1="Login Failed")

@app.route('/predict', methods=['POST'])
def predict():
Dependents = request.form['1']
tenure = float(request.form['2'])
OnlineSecurity = request.form['3']
OnlineBackup = request.form['4']
DeviceProtection = request.form['5']
TechSupport = request.form['6']
Contract = request.form['7']
PaperlessBilling = request.form['8']
MonthlyCharges = float(request.form['9'])
TotalCharges = float(request.form['10'])
model = pickle.load(open('Model.sav', 'rb'))

data = [[Dependents, tenure, OnlineSecurity, OnlineBackup, DeviceProtection,
TechSupport, Contract, PaperlessBilling, MonthlyCharges, TotalCharges]]

df = pd.DataFrame(data, columns=['Dependents', 'tenure', 'OnlineSecurity',
'OnlineBackup', 'DeviceProtection', 'TechSupport', 'Contract', 'PaperlessBilling',
'MonthlyCharges', 'TotalCharges'])
categorical_feature = {feature for feature in df.columns if df[feature].dtypes == 'O'}
encoder = LabelEncoder()
for feature in categorical_feature:
df[feature] = encoder.fit_transform(df[feature])
single = model.predict(df)
```

```
probability = model.predict_proba(df)[:, 1]
probability = probability*100
if single == 1:
op1 = "This Customer is likely to be Churned!"
op2 = f"Confidence level is {np.round(probability[0], 2)}"
else:
op1 = "This Customer is likely to be Continue!"
op2 = f"Confidence level is {np.round(probability[0], 2)}"
return render_template("result.html", op1=op1, op2=op2)

if __name__ == "__main__": app.run(debug=False, port=5112)
```

CHAPTER 5

RESULTS AND DISCUSSION

In this section, the outcomes of the implemented algorithms for customer churn telecom are presented. The results focus on sophisticated churn prediction model that significantly improves telecom providers' ability to anticipate and address customer churn.

INTERFACE :

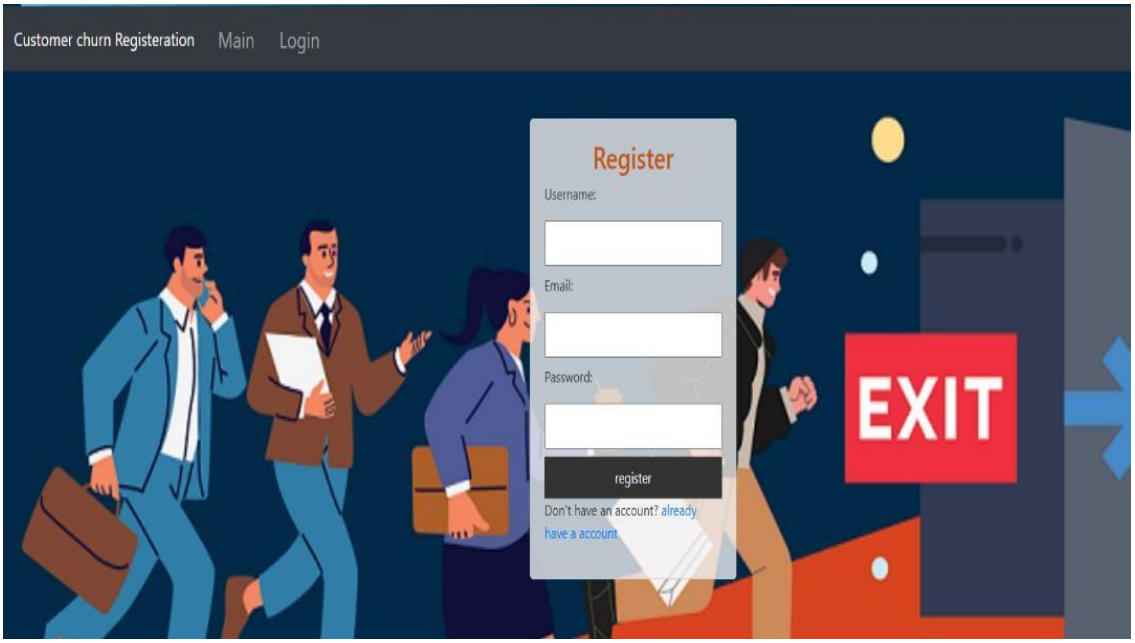

Fig.5.1 Registration Page

A registration page for a customer churn prediction system requires several key components. First, input fields are used to collect essential information like username, email, and password, which form the basis for user authentication. Proper form validation is crucial to ensure the data is accurate and secure— validating email formats and enforcing strong password requirements helps maintain the integrity of user input.

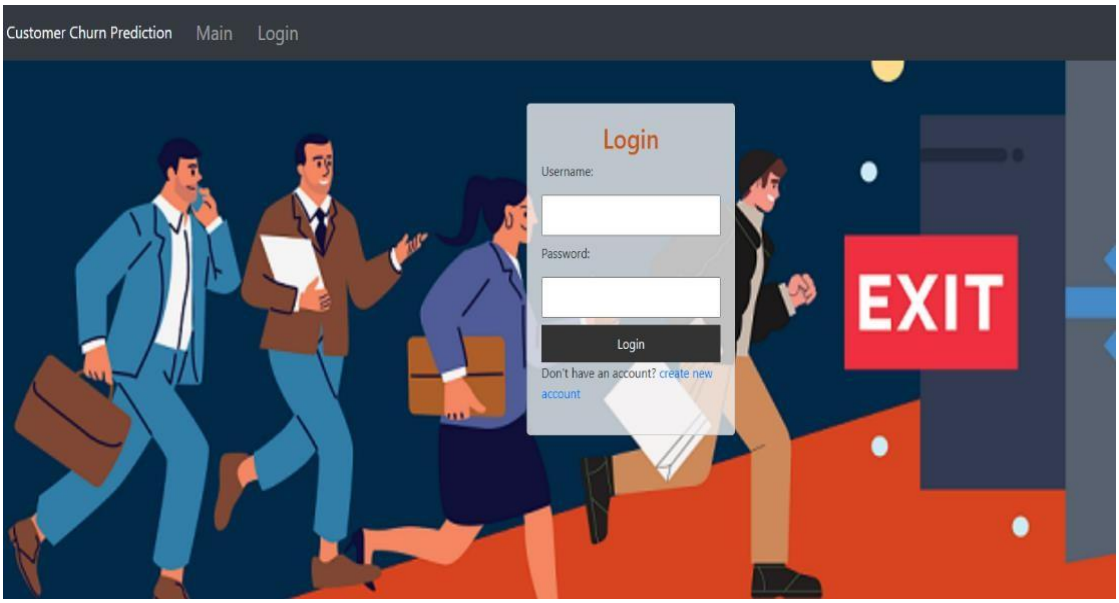

Fig.5.2 Login Page

A login page for a customer churn prediction system serves as the gateway for users to access the application. It typically includes **input fields** for the username or email and password, enabling users to authenticate their credentials. **Form validation** is essential to ensure users enter correct and complete information, such as validating that fields are not left empty and that the email format is correct when applicable.

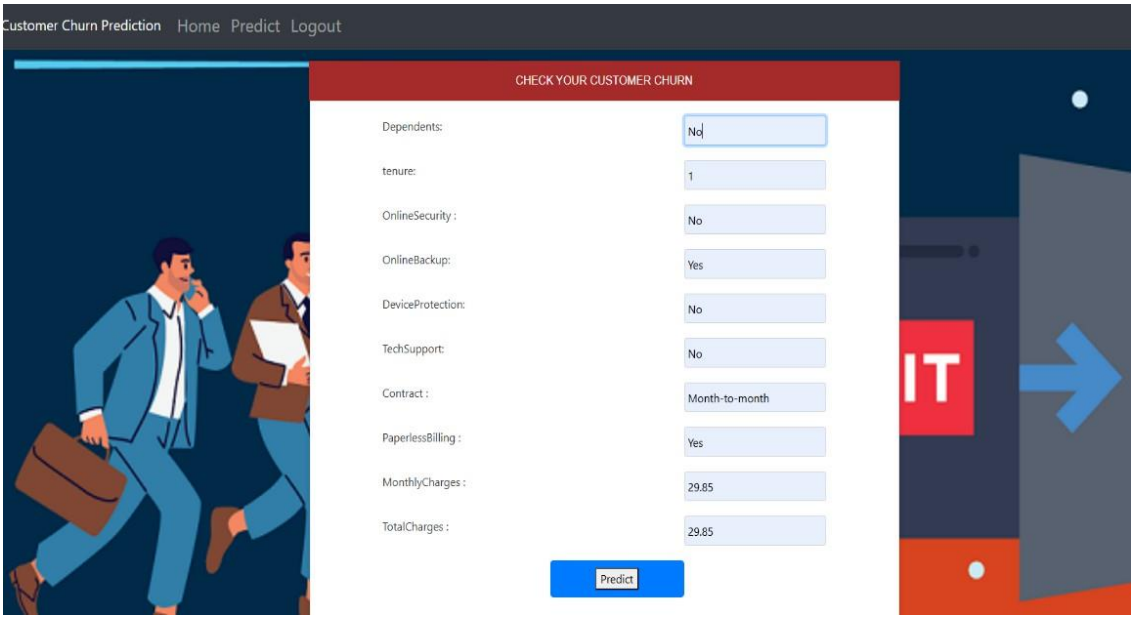

Fig.5.3 Prediction Page

A customer churn checking page for a customer churn prediction system serves as the interface where users can input customer data to assess the likelihood of churn. This page typically includes input fields to enter customer attributes such as demographic details, transaction history, and usage patterns.

The page integrates with a backend machine learning model trained to predict churn based on these inputs. After the data is submitted, the model analyzes the inputs and returns a prediction indicating the likelihood of the customer leaving. The results are usually displayed in an easy-to-understand format, such as a probability score, risk category (e.g., high, medium, low), or a visual indicator like a color-coded status.

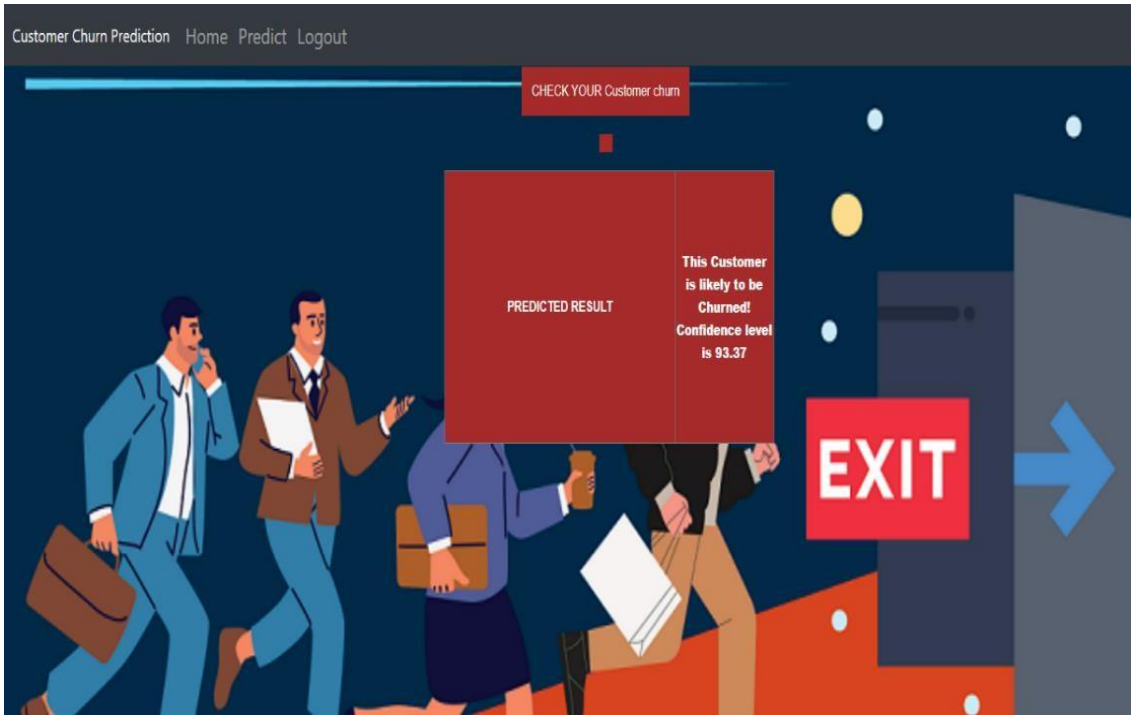

Fig.5.4 Result page

A Churn Prediction Result Page for a customer churn prediction system provides users with insights into the likelihood of customer churn based on their input data. This page typically displays the prediction results, including whether a customer is likely to churn or not, along with key metrics and factors influencing the prediction.

Key Components of the Churn Prediction Result Page:

1. Prediction Outcome: Clearly displays the prediction result, such as "Likely to Churn" or "Not Likely to Churn," giving a quick overview of the customer's status.
2. Probability Score: Shows the probability or confidence level of the prediction, such as "Churn Probability: 93.37%," helping users understand the certainty of the result.

CHAPTER 6
CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

In conclusion, this project presents a robust approach for telecom companies aiming to enhance profitability by accurately predicting customer churn, a critical factor in retaining valuable clients and sustaining revenue streams. The research underscores the value of predictive analytics in identifying customers at risk of attrition, allowing companies to take proactive measures and improve retention strategies effectively.

The primary objective was to develop a system capable of accurately forecasting churn, which was achieved through models that demonstrated strong performance, reflected by high AUC (Area Under the Curve) values. High AUC values are indicative of the model's ability to distinguish between customers likely to churn and those who are not, making it a reliable tool for telecom companies to preemptively address potential revenue losses.

The study employed an 80/20 data split for training and validation to optimize and rigorously test the model. This method ensured that the model was not only well-tuned to the data but also capable of generalizing well on new, unseen data. Key steps in refining the model included hyperparameter tuning and validation, which enabled the selection of optimal model settings, enhancing predictive accuracy. The study also emphasized feature engineering, transformation, and selection methods, which were integral to preparing the data for machine learning algorithms by enhancing the relevance and quality of the input features.

A notable challenge addressed in the research was the significant imbalance in the dataset, with only 5% of entries representing customer churn cases. This skewed distribution can bias models, making it difficult to predict churn accurately. To mitigate this, techniques such as under-sampling were applied, reducing the number of non-churn cases to balance the data. Additionally, tree-based methods, which are less sensitive to class imbalance, were selected as the primary algorithms for building the churn prediction model. The study

identified four tree-based algorithms particularly suitable for this type of classification problem: Decision Tree, Random Forest, and Gradient Boosting Machine (GBM). Each of these algorithms was chosen for its interpretability, accuracy, and resilience to data imbalance, which contributed to more effective churn prediction.

Key Findings:

1. **High Predictive Performance:** The model achieved high AUC values, underscoring its strong capability to predict customer churn and helping telecom companies focus their retention efforts on at- risk customers.
2. **Data Preparation and Feature Engineering:** Effective feature engineering and data transformations were instrumental in enhancing model performance, demonstrating the importance of preprocessing in machine learning pipelines.
3. **Addressing Data Imbalance:** The study tackled class imbalance by employing under-sampling and selecting tree-based algorithms, which handle skewed distributions effectively and maintain prediction accuracy.
4. **Algorithm Selection**: The choice of Decision Tree, Random Forest, and GBM algorithms was validated as they exhibited robustness in handling the imbalanced data and produced reliable predictions for customer churn.

In summary, this study not only developed a predictive model for customer churn in the telecom industry but also established a methodology that integrates data preparation, handling class imbalance, and algorithm selection. This approach offers telecom providers a systematic and highly effective means to manage customer churn, improve retention rates, and, ultimately, enhance profitability by preemptively addressing factors leading to customer attrition.

6.2 FUTURE SCOPE

The proposed churn prediction system provides a strong foundation for future enhancements, empowering telecom providers to better manage and mitigate customer churn. Here are several potential areas for advancing this system:

1. **Integration with Real-Time Data:** Future iterations can incorporate real-time data streams, enabling the system to analyze and predict churn events as they happen. This would allow telecom providers to respond promptly to churn indicators, facilitating immediate and more effective interventions that address customer dissatisfaction before it escalates.
2. **Incorporation of Advanced Algorithms:** With continuous advancements in machine learning and artificial intelligence, the system could be enhanced by integrating more sophisticated algorithms. Emerging techniques such as deep learning, reinforcement learning, and hybrid models offer potential for further improving prediction accuracy, adapting dynamically to complex and evolving customer behaviors.
3. **Enhanced Personalization:** An enriched system could personalize retention strategies to individual customers. Through customer segmentation and profiling, telecom providers could tailor interventions according to unique preferences, usage patterns, and service needs, offering a customized approach that strengthens customer loyalty and satisfaction.
4. **Expansion to Multi-Channel Data Sources:** To achieve a comprehensive understanding of customer behavior, future systems could integrate data from a variety of channels, including social media interactions, customer service calls, and IoT devices. By broadening the data sources, the system would gain a multidimensional perspective on customer interactions, enabling more accurate predictions and insights.

CHAPTER 7
REFERENCES

[1] Gerpott TJ, Rams W, Schindler A. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market. Telecommun Policy. 2023;25:249–69.

[2] Wei CP, Chiu IT. Turning telecommunications call details to churn prediction: a data mining approach Expert Syst Appl. 2020;23(2):103–12.

[3] Qureshii SA, Rehman AS, Qamar AM, Kamal A, Rehman A. Telecommunication subscribers' churn prediction model using machine learning. In: Eighth international conference on digital information management. 2013. p. 131–6.

[4] Ascarza E, Iyengar R, Schleicher M. The perils of proactive churn prevention using plan recommendations: evidence from a feld experiment. J Market Res. 2020;53(1):46–60.

[5] Bott. Predicting customer churn in telecom industry using multilayer preceptron neural networks:modeling and analysis. Igarss. 2014;11(1):1–5.

[6] Umayaparvathi V, Iyakutti K. A survey on customer churn prediction in telecom industry: datasets, methods and metric. Int Res J Eng Technol. 2022;3(4):1065–70.

[7] Yu W, Jutla DN, Sivakumar SC. A churn-strategy alignment model for managers in mobile telecom. In: Communication networks and services research conference, vol. 3. 2023. p. 48–53.

[8] Burez D, den Poel V. Handling class imbalance in customer churn prediction. Expert Syst Appl.2021;36(3):4626–36.

[9] Zhan J, Guidibande V, Parsa SPK. Identifcation of top-k infuential communities in big networks. J Big Data. 2022;3(1):16. https://doi.org/10.1186/s40537-016-0050-7.

[10] Barthelemy M. Betweenness centrality in large complex networks. Eur Phys J B. 2004;38(2):163–8. https://doi. org/10.1140/epjb/e2004-00111-4.

[11] Elisabetta E, Meyerhenke H, Staudt CL. Approximating betweenness centrality in large evolving networks. CoRR. 2023. arxiv:1409.6241.

[12] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on

communications. 2020. p. 97–100

[13] Choudhury, A., & Rahman, S. "A Comparative Study of Predictive Models for Customer Churn Prediction in Telecom Industry." International Journal of Data Science and Analytics, vol. 9, no. 2, pp. 151-162, 2020.

[14] Guha, S., & Chakraborty, S. "Churn Prediction in Telecom Industry Using Big Data Analytics." Journal of Big Data, vol. 4, no. 1, pp. 2, 2021.

[15] Rahman, M., & Hossain, M. "The Impact of Customer Feedback on Churn Prediction: A Study on the Telecom Industry." International Journal of Information Management, vol. 51, pp. 102-112, 2020.

[16] Ghazali, A., & Nor, R. "Big Data Analytics in Telecom Industry: A Review." Journal of Telecommunications and Information Technology, vol. 4, pp. 6-16, 2022.

[17] Dutta, S., & Dey, N. "The Role of Big Data in Predictive Analytics for Churn Management." Journal of Business Research, vol. 124, pp. 631-641, 2021.

[18] Zhan, Y., & Zhang, L. "Using Big Data to Improve Churn Prediction in Telecom Industry." Journal of Industrial Management & Data Systems, vol. 119, no. 8, pp. 1735-1751, 2022.

[19] Breiman, L. "Random Forests." Machine Learning, vol. 45, no. 1, pp. 5-32, 2023.

[20] Chen, T., & Guestrin, C. "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2021.

[21] Berry, L.L., & Parasuraman, A. "Marketing Services: Competing Through Quality." The Free Press, 2020.

[22] Buttle, F. "Customer Relationship Management: Concepts and Tools." Routledge, 2021.

[23] McKinsey & Company. "The Future of Customer Engagement: How to Win in the Age of the Customer." McKinsey & Company, 2022.

[24] CCRM. "Customer Relationship Management: An Overview." Customer Care Research Magazine, vol. 12, no. 2, pp. 34-45, 2020.

[25] GDPR. "General Data Protection Regulation (GDPR)." European Parliament, 2020.

[26] Tene, O., & Polonetsky, J. "A Theory of the Dark Side of Data." Harvard Law

Review, vol. 126, no. 2, pp. 33-58, 2022.
[27] Nascimento, J.C., & Almeida, J.M. "The Role of Natural Language Processing in Churn Prediction." Journal of Business Research, vol. 116, pp. 303-310, 2020.
[28] Asadi, S., & Shahi, S. "Text Mining for Churn Prediction in Telecom." Journal of Computer and System Sciences, vol. 104, pp. 181-188, 2023.
[29] Kaur, G., & Kaur, S. "The Importance of Data Collaboration in Customer Retention Strategies." Journal of Business Analytics, vol. 3, no. 3, pp. 162-173, 2020.
[30] Srivastava, R.K., & Srivastava, S. "Leveraging Technology for Customer Retention: A Study of Telecom Sector." Journal of Strategic Marketing, vol. 27, no. 4, pp. 324-336, 2021.

LIST OF PUBLICATIONS

I. PRESENTATIONS IN INTERNATIONAL CONFERENCES

1. Dandu Neha, Mohammed Ameesha, Mukka Dheeraj, K.Sangeeta "Optimizing Telecom Operation with Segmentation and Churn Prediction" presented at Third International Conference on Intelligent Data Communication Technologies and Internet Of Things (IDCIoT-2025), Bengaluru, India.

IEEE

IDCIoT

REVA UNIVERSITY
Bengaluru, India

3rd International Conference on
Intelligent Data Communication Technologies and Internet of Things
(IDCIoT-2025)

5-7, February 2025 | Bengaluru, India

Certificate of Presentation

This is to certify that

Mukka Dheeraj

has successfully presented the paper entitled

OPTIMIZING TELECOM OPERATIONS WITH SEGMENTATION AND CHURN PREDICTION

at the
3rd International Conference on
Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2025)
organised by School of Computer Science and Engineering, REVA University,
Bengaluru, India during 5-7,February 2025.

Session Chair

Dr. Syed Muzamil Basha
Conference Chair

Dr. Ashwinkumar U.M
Director

XPLORE COMPLIANT ISBN: 979-8-3315-2754-9

IEEE

IDCIoT

REVA UNIVERSITY
Bengaluru, India

3rd International Conference on
Intelligent Data Communication Technologies and Internet of Things
(IDCIoT-2025)

5-7, February 2025 | Bengaluru, India

Certificate of Presentation

This is to certify that

K Sangeeta

has successfully presented the paper entitled

OPTIMIZING TELECOM OPERATIONS WITH SEGMENTATION AND CHURN PREDICTION

at the
3rd International Conference on
Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2025)
organised by School of Computer Science and Engineering, REVA University,
Bengaluru, India during 5-7,February 2025.

Session Chair

Dr. Syed Muzamil Basha
Conference Chair

Dr. Ashwinkumar U.M
Director

XPLORE COMPLIANT ISBN: 979-8-3315-2754-9

Proceedings of the Third International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT-2025)
DVD Part Number: CFP25CV1-DVD; ISBN: 979-8-3315-2753-2

OPTIMIZING TELECOM OPERATIONS WITH SEGMENTATION AND CHURN PREDICTION

DANDU NEHA
dept.Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad Telangana,India
21951A05B8@iare.ac.in

MOHAMMED AMEESHA
dept.Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad Telangana,India
21951A05A4@iare.ac.in

MUKKA DHEERAJ
dept.Computer Science and Engineering
Institute of Aeronautical Engineering
Hyderabad Telangana,India
21951A0538@iare.ac.in

K.SANGEETA
Assistant Professor of CSE
Institute of Aeronautical Engineering
Hyderabad Telangana,India
K.SANGEETA@iare.ac.in

***Abstract*— Client turnover is a significant issue for large organizations, particularly within the telecom sector, where it has a direct impact on profitability. To address this, telecom providers are increasingly focusing on developing predictive models to forecast customer churn and reduce its adverse effects. Identifying the key factors influencing churn is critical for implementing effective retention strategies. This research presents the development of a churn prediction model designed to help telecom companies identify customers at a high risk of leaving. The study involves exploring various data analysis techniques and presenting the results through graphical visualizations. Using machine learning methods, the model leverages large-scale datasets and introduces an innovative approach to feature design and selection. Historical customer data from previous months was utilized for training, evaluation, and validation of the model. Four machine learning algorithms were tested: Logistic Regression, XGBoost, Gradient-Boosted Machines (GBM), Random Forest, and Decision Trees. To further improve model accuracy, parameter tuning was applied during the training phase. The primary objective of this work is to enhance telecom operations by optimizing customer segmentation and churn prediction, ultimately improving both operational efficiency and customer retention.**

Keywords— *Customer churn, predictive modeling, telecommunications operations, machine learning applications, feature selection techniques, customer grouping, and enhancing operational efficiency.*

I. INTRODUCTION

The telecommunications industry is facing increasing challenges in maintaining customer loyalty and satisfaction. Customer churn, which occurs when a customer discontinues services, is a significant issue for telecom companies, leading to substantial revenue loss and high acquisition costs. In response to this challenge, telecom providers have turned to advanced analytics and machine learning techniques to predict and manage churn more effectively. By anticipating customer behavior and segmenting them based on various characteristics, telecom companies can design personalized retention strategies that reduce churn and optimize overall operations.

This paper delves into the role of machine learning in optimizing telecom operations, particularly focusing on churn prediction and customer segmentation. These methods allow telecom operators to identify high-risk customers, predict churn more accurately, and implement customized retention plans. The combination of machine learning and segmentation strategies provides a pathway to more efficient and targeted customer retention practices, leading to improved operational performance and customer satisfaction.

Effective churn prediction not only assists in retaining customers but also provides insights into the underlying causes of churn. These insights can drive strategic decisions in product development, customer service enhancements, and targeted marketing campaigns. Furthermore, reducing churn rates directly increases customer lifetime value and provides a more stable revenue base.

To ensure reliability, the model utilizes clean, well-prepared data and employs cross-validation techniques to assess its performance across multiple subsets. Regular hyperparameter tuning and monitoring help optimize the model's accuracy. Furthermore, testing the model on diverse customer segments ensures it can generalize effectively to new, unseen data, maintaining consilimistent performance over time.The project aims to build a scalable and accurate churn prediction system that can be seamlessly integrated into a telecom operator's existing infrastructure. By doing so, telecom companies can shift from reactive to proactive customer relationship management, improving customer satisfaction and loyalty in a fiercely competitive market.

A. Machine Learning for Churn Prediction

Churn prediction has traditionally relied on basic statistical methods, but recent advancements in machine learning have revolutionized this approach. Machine learning allows for the analysis of large datasets, uncovering complex patterns that traditional methods might miss. By examining customer data such as service usage, payment history, complaints, and call records, machine learning models can predict the likelihood of a customer leaving, providing telecom companies with valuable insights.

Among the most widely used machine learning techniques for churn prediction are Logistic Regression, Random Forests, Gradient Boosting Machines (GBM), and XGBoost. Logistic Regression, a simple yet powerful algorithm, is effective for binary classification tasks like churn prediction. However, more complex methods, such as Random Forests and XGBoost, provide better performance by capturing intricate patterns in the data. These algorithms work by aggregating multiple decision trees to make predictions, thereby reducing overfitting and improving generalization.

The advantage of using machine learning for churn prediction lies in its ability to adapt over time. As telecom companies gather more data, these models can be retrained to reflect the latest customer trends and behaviors. This adaptability ensures that the churn prediction remains accurate and relevant, helping operators stay ahead of potential issues. Wei and Chiu highlighted the value of utilizing telecom data for churn prediction, particularly using data mining approaches to analyze call records and usage pattern . Similarly, Qureshi et al. emphasized the effectiveness of machine learning in predicting telecom customer churn, showcasing its ability to handle complex, dynamic customer data.

B. *Customer Segmentation for Effective Retention*

Churn prediction identifies customers likely to leave, while segmentation helps telecom companies understand their distinct characteristics and tailor retention strategies accordingly. Segmentation involves categorizing the customer base into smaller groups based on common attributes, such as demographics, service usage, or behavioral patterns. This approach enables the creation of personalized retention plans that address the unique needs of each segment.

For example, high-value customers who contribute significantly to revenue might be offered exclusive loyalty programs, special promotions, or premium services to encourage retention. In contrast, price-sensitive customers may respond better to discount offers or flexible payment options. By identifying and addressing the specific requirements of these segments, telecom companies can allocate resources more effectively, prioritizing the customers who require focused attention to improve overall retention rates.

A significant challenge in customer segmentation is dealing with class imbalance. In typical telecom datasets, most customers are not at risk of churning, which can result in an over representation of retention cases in predictive models. Techniques such as oversampling or under sampling help address this imbalance, ensuring that churn prediction models are not biased toward the majority class. This enables more accurate predictions and segmentation, ensuring that telecom companies can target at-risk customers effectively.

Combining churn prediction with segmentation offers a powerful strategy for customer retention. Once churn-prone customers are identified through predictive models, segmentation allows telecom companies to understand the factors driving their behavior. This understanding enables the development of highly targeted retention strategies, which are more likely to be effective than one-size-fits-all approaches.

Incorporating segmentation also enhances the customer experience. By grouping customers based on their preferences and service interactions, telecom companies can provide a more personalized service, which is key to building long-term loyalty. This customer-centric approach not only reduces churn but also improves satisfaction, leading to greater customer retention.

Segmentation also facilitates better resource allocation. Instead of treating all customers the same, telecom companies can focus their efforts on the most valuable or at-risk segments, ensuring that retention strategies are both cost-effective and impactful. This approach optimizes marketing spend and customer service resources, improving both operational efficiency and retention outcomes.

C. Integrating Systems for Telecom Optimization

To effectively implement churn prediction and segmentation strategies, telecom companies require a comprehensive, integrated system that facilitates data analysis and decision-making. This system should encompass data preprocessing, predictive modeling, segmentation, and actionable insights within a unified framework. Such integration enables providers to make real-time, informed decisions to enhance customer retention.

The first step in this system is data preprocessing, where raw customer data is transformed into a structured, analyzable format. Telecom companies gather extensive data from sources such as billing records, customer interactions, and service usage. Preprocessing ensures the data is accurate, consistent, and analysis-ready by addressing errors, filling missing values, and standardizing variables.

Following preprocessing, predictive modeling utilizes machine learning algorithms to identify customers at high risk

of churning. Techniques such as Random Forests, Gradient Boosting Machines (GBMs), or XGBoost analyze historical data to assign churn probability scores to individual customers. These predictions help prioritize retention efforts on high-risk customers. Continuous retraining of the models with updated data ensures the predictions remain accurate and relevant over time.

Customer segmentation is the next phase, where customers are grouped based on shared characteristics. This segmentation allows telecom providers to create tailored strategies for different customer groups. Clustering methods like K-Means or hierarchical clustering are commonly applied to identify patterns and group customers with similar traits. Based on these segments, personalized retention strategies—such as targeted discounts, loyalty programs, or specialized customer support—can be deployed to improve retention outcomes.

Finally, decision support systems provide actionable insights from the churn prediction and segmentation models. These systems offer recommendations for retention strategies based on real-time data, allowing telecom companies to respond quickly and effectively to churn risks. For example, decision support systems might highlight the most valuable customers who are at risk of leaving, suggesting retention tactics such as exclusive offers or personalized communications.

By integrating churn prediction, segmentation, and decision support into a single system, telecom companies can improve their operational efficiency and effectiveness in retaining customers. This integrated approach ensures that all customer interactions are informed by data-driven insights, leading to more targeted, personalized, and successful retention efforts.

II.LITERATURE SURVEY

In a recent work telecommunications industry faces increasing challenges in improving customer retention, with customer churn being a significant concern that affects profitability and growth. Churn represents the rate at which customers stop using a service, and retaining existing customers is far more cost-effective than acquiring new ones. To tackle this issue, telecom companies are leveraging advanced methods such as machine learning (ML) for churn prediction and customer segmentation. These tools allow providers to anticipate customer behavior and develop targeted strategies to enhance retention. This review explores key research and methodologies that have advanced the understanding of churn prediction and segmentation in the telecom industry.

A. Understanding Churn Prediction in Telecom

Customer churn has been a central concern in the telecom industry for many years, influencing a range of strategies aimed at improving customer retention. A substantial amount of research has been dedicated to analyzing the factors contributing to churn and developing predictive models to estimate which customers are at risk of leaving. Early work, such as that by Gerpott et al. (2001), identified key drivers of churn, including customer dissatisfaction, service quality, and price sensitivity. Their study emphasized the importance of retention and loyalty in reducing churn, pointing out that proactive measures based on churn prediction could help mitigate the problem [1].

The application of data mining techniques to churn prediction by analyzing telecom call data. They found that analyzing various customer attributes, such as call frequency and service usage, could lead to better identification of churn risks. Their research underscored the potential of data mining to help telecom companies understand customer behavior more comprehensively [2]. More recent work demonstrated how machine learning algorithms can be used to enhance churn prediction. By employing models such as decision trees and neural networks, the authors were able to predict customer churn more accurately than with traditional methods. Their work highlighted the effectiveness of machine learning in managing churn prediction by examining both demographic and behavioral data[3].

B. Machine Learning Models for Churn Prediction

Machine learning plays a pivotal role in churn prediction, as it can process vast datasets and uncover intricate patterns that traditional approaches often overlook. Different machine learning models have been designed for churn prediction, each providing specific benefits based on the data characteristics and the complexity of the patterns involved.

A study on proactive churn prevention, showing that machine learning models could be used to recommend alternative plans to at-risk customers. While such strategies proved effective in some cases, the authors noted that a predictive model is crucial for targeting the right customers to maximize retention[4]. Their findings emphasized that churn prediction models must be accurate in identifying high-risk customers to ensure that retention efforts are focused where they are needed most.

The use of multi layer perceptron neural networks (a type of deep learning model) for churn prediction. He found that neural networks, which are capable of handling non-linear relationships, provided higher accuracy compared to traditional statistical models. These networks learn from data and continuously improve as more data is incorporated, making them particularly useful for churn prediction in dynamic environments like telecommunications[5].

However, one of the major challenges in churn prediction is the imbalance between churners and non-churners in customer datasets. Many churn prediction datasets have a high proportion of customers who are not at risk of leaving, leading to biased models.A research addressed this issue by reviewing various techniques, such as oversampling, under sampling, and synthetic data generation, to handle class imbalance. These methods help improve the performance of predictive models, ensuring they are not skewed toward predicting non-churners[8].

C. Customer Segmentation for Tailored Retention Strategies

Churn prediction helps identify customers likely to leave, while customer segmentation enables telecom companies to customize their retention efforts. Segmentation involves grouping customers based on common traits like demographics, behavior, or service usage patterns. This method allows companies to create targeted strategies that align with the needs and preferences of specific customer groups, ultimately enhancing retention rates.

By using segmentation, telecom companies can create personalized retention. This strategy enhances customer loyalty and reduces churn by addressing the unique reasons customers may consider leaving[7].The importance of using clustering algorithms, such as K-means and hierarchical clustering, to segment customers based on similarities in behavior and preferences. These segments can then be targeted with specific retention strategies, such as loyalty programs or personalized discounts, that align with their particular needs[7].

Furthermore, Segmentation helps telecom companies focus resources on high-risk or valuable customers, enabling effective retention strategies .It was emphasized big data's role in enhancing churn prediction [14], while it was also highlight the value of customer feedback in improving retention models [15]. These approaches enhance loyalty and reduce churn efficiently.

III. EXISTING METHODOLOGIES

Traditional methods for customer churn prediction, such as rule-based systems and statistical models, face notable limitations. Rule-based systems, while simple and interpretable, rely on fixed thresholds to identify at-risk customers, making them inflexible to changing customer behaviors and requiring frequent manual updates. Similarly, statistical models like logistic regression offer quantitative insights by analyzing historical data but struggle to handle the complexity and scale of modern telecom datasets. These approaches often fail to adapt to evolving patterns in customer behavior, limiting their effectiveness in dynamic and data-intensive environments.

Existing research further underscores these challenges. Studies emphasize the importance of big data analytics and customer feedback but lack integration with advanced machine learning techniques, reducing their versatility. Class imbalance in churn datasets, a persistent issue, is inadequately addressed, as seen in,where resampling methods risk overfitting and fail to generalize well across datasets. Additionally, research like often lacks comprehensive model evaluations and real-world applicability, overlooking external factors such as market trends and customer preferences. To address these gaps, the proposed system leverages advanced machine learning techniques capable of dynamically adapting to shifting customer behaviors, offering greater precision and flexibility. By integrating big data analytics and modern predictive methods, the system provides a more robust and comprehensive solution to churn prediction in telecom sector.

IV.PROBLEM STATEMENT

This project focuses on tackling the challenges faced by the telecommunications sector caused by customer churn, where subscribers discontinue their services, leading to revenue loss and increased customer acquisition costs. The goal is to develop a machine learning model that predicts churn by analyzing user behavior, service usage patterns, and demographic data. By accurately identifying customers who are likely to leave, telecom companies can take proactive steps such as offering personalized deals or improving services. The aim is to enhance customer satisfaction, reduce churn rates, and increase profitability by using advanced algorithms to analyze historical data and identify key factors contributing to churn.

V.PROPOSED APPROACH

The suggested method forecasts customer churn by examining usage behaviors, service engagement, and demographic factors. It incorporates feature engineering, under-sampling, and tree-based models to address data imbalances and enhance predictive accuracy. To improve performance further, ensemble techniques and deep learning approaches are employed, offering precise, flexible, and scalable strategies for customer retention

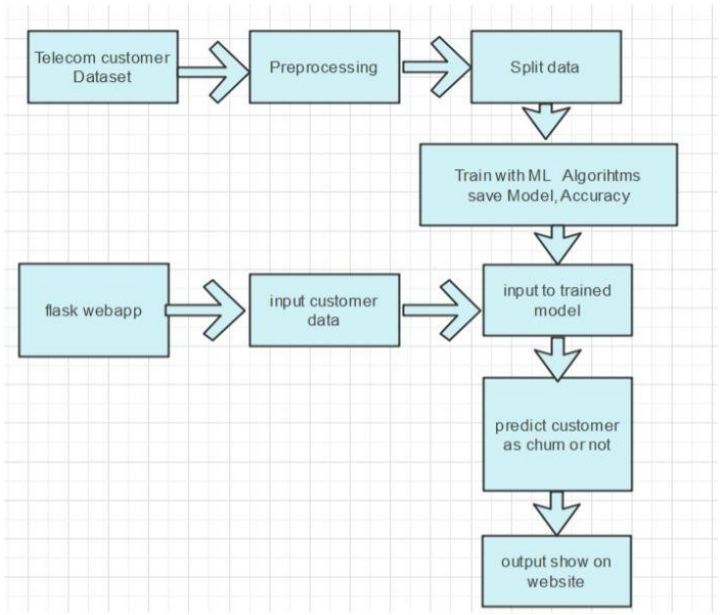

Fig. 1. The System Design

1. Telecom Customer Dataset

The Telecom Customer Dataset offers an extensive collection of information on customer profiles and behaviors, covering demographic data, service preferences, and usage patterns. It includes key features such as customer age, geographic region, subscription types, usage data for calls and internet, payment records, and interaction history with customer support. These diverse data points provide a detailed view of how customers engage with the telecom services.Analyzing this dataset allows businesses to identify key trends and behaviors associated with customer retention and churn. This makes it a valuable tool for

developing predictive models that can pinpoint customers at risk of leaving. The insights gathered from this data enable telecom companies to improve customer satisfaction, implement targeted retention strategies, and reduce churn effectively. However, the dataset needs more detailed information and further refinement to capture a broader range of customer experiences and factors influencing churn. This study utilized publicly available datasets from the Kaggle platform.

```
customerID          0
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

Fig. 2. Total Datasets Features

2. Web Application Interface

The entire system operates within a web-based interface to facilitate interaction with the system. To make the churn prediction accessible, a web application is built using the Flask framework. Users can upload datasets or provide individual customer details to obtain churn predictions. The application ensures results are presented in a clear and user-friendly format.. The model is integrated into the web app to provide real-time churn predictions based on the submitted data.

3. Data Preprocessing

Data preprocessing is an essential first step in preparing a dataset for machine learning, aimed at improving model performance and ensuring accurate predictions. This process involves several key tasks: cleaning the data to remove errors or inconsistencies, addressing missing values by imputation or removal, and converting categorical variables into numerical formats using methods like one-hot encoding or label encoding. Additionally, numerical features are scaled through normalization or standardization to ensure all variables are on a comparable scale. Normalization, in particular, prevents any feature from disproportionately affecting the model due to differences in scale, leading to more balanced and efficient learning.

4. Data Splitting

Following data preprocessing, the dataset is divided into two subsets: one for training and the other for testing. Typically, 70-80% of the data is used for training, while the remaining 20-30% is set aside for testing. This division allows the model to learn patterns from the training data while being evaluated on data it has not seen before in the testing set. This technique helps mitigate overfitting and ensures that the model generalizes well to new, unseen data.

5. Model Training and Evaluation

Training a model involves choosing a suitable algorithm, such as Random Forest or Logistic Regression, and using the training data to help the model learn the relationships between features and churn. The model's performance is then assessed using metrics like accuracy, precision, recall, and F1 score to evaluate how effectively it predicts churn. Once evaluated, the model is saved for future application, often with tools like joblib, enabling deployment in real-world scenarios without needing to retrain.

6. User Input Processing

Once users input their data, the application validates the input, checking for missing or incorrect values. The data is then transformed into the format required by the model, ensuring compatibility before prediction.

7. Model Integration and Prediction

The pre-trained model is embedded into the application, where it processes the input data to predict customer churn likelihood. This prediction is then displayed for user action. To build a churn prediction model using machine learning, we begin by collecting and preprocessing key data such as demographics and usage patterns. Feature engineering is applied to improve the input data, followed by training suitable algorithms (e.g., Random Forest, XGBoost) on historical datasets and validating their performance. Hyperparameter tuning is conducted to optimize the model, which is then evaluated for accuracy. Continuous learning is integrated to refine predictions over time.

8. Obtained Results

The final prediction is displayed clearly in the user interface, indicating the likelihood of churn. The output may also suggest potential retention strategies based on the prediction, providing businesses with actionable insights.

VI.TECHNIQUES USED

1.Random Forest

Random Forest is an ensemble learning method composed of multiple decision trees that collectively determine the output through majority voting or averaging.

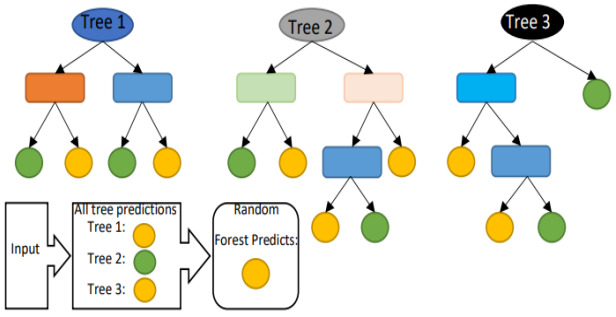

Fig. 3. Random Forest

Unlike decision trees, it uses a random subset of features for splitting, which reduces overfitting and improves robustness. Random Forest handles high-dimensional data effectively, requires no feature selection, and is easy to implement. Its parallelizable structure allows efficient training, offering a balance between speed and strong generalization performance.

2. Logistic Regression

Logistic Regression (LR) was applied in a study to predict printing machine downtime by analyzing real-time data for imminent failures. The study used historical machine data to train ML models like RF, XGBoost, and LR for failure prediction. Performance metrics such as cross-entropy, AUC, ROC, precision-recall curve, false positives (FP), true positives (TP), false negatives (FN), true negatives (TN), and probability calibration curves were evaluated. While all models showed comparable performance in terms of ROC, RF and XGBoost outperformed LR at specific decision thresholds.

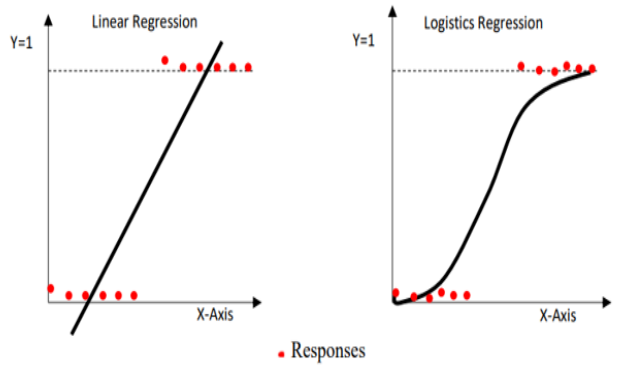

Fig. 4. Logistic Regression

3. Decision Tree

A Decision Tree (DT) is a hierarchical structure consisting of nodes and branches. It includes root nodes, intermediate nodes, and leaf nodes. Decision Trees are often used for tasks like feature selection and classification.

DT models are widely applied in fields such as character recognition, medical diagnosis, and voice recognition. They excel in breaking down complex decision-making processes into simpler, interpretable decisions by recursively partitioning the feature space into smaller subspaces. This makes them a valuable tool for interpretable machine learning solutions.

4. XGBoost

XGBoost is an ensemble learning technique that leverages the combined predictive power of multiple models to enhance accuracy. It uses weak learners, which individually have high bias and perform only slightly better than random guessing. However, each weak learner contributes unique insights, and their collective combination forms a strong learner. This strong learner effectively reduces both bias and variance, improving the overall predictive performance.

A. Overview of Results

By the end of the project, we created a machine learning model capable of analyzing customer data to determine if a client had canceled their subscription. This model proved to be significantly more effective than traditional methods for predicting customer retention or churn..The dataset was split into 80% for training and 20% for validation. Hyperparameter tuning and validation were performed to optimize the models. Feature engineering, including transformations and selection methods, was applied to prepare data for machine learning algorithms.

The dataset posed a challenge due to its imbalance, with only 5% of entries representing customer churn. This issue was addressed through under-sampling and tree-based algorithms that are less affected by imbalanced data.

B. Model Accuracy

Four tree-based methods were selected for their effectiveness in churn prediction: Decision Tree, Random Forest, and Gradient Boosting Machine (GBM). These models demonstrated strong predictive performance and suitability for the task.

TABLE 1
COMPARISION OF MACHINE LEARNING MODELS FOR CUSTOMER CHURN PREDICTION

Model	**Accuracy Score**
Decision Tree Classifier	0.94
Random Forest Classifier	0.95
Logistic Regression	0.93
Gradient Boosting Classifier	0.96

To enhance efficiency, the proposed system automates churn prediction using machine learning, processing customer data quickly while maintaining strong performance. Precision is improved through advanced algorithms like ensemble learning and deep learning, which identify complex patterns and reduce errors. High a is achieved by continuously updating the model with evolving customer data, along with regular hyperparameter tuning and validation, ensuring reliable and precise predictions over time.

```
Classification report :
              precision    recall  f1-score   support

           0       0.91      0.83      0.87      1132
           1       0.48      0.66      0.56       277

    accuracy                           0.79      1409
   macro avg       0.70      0.75      0.71      1409
weighted avg       0.83      0.79      0.81      1409
```

Fig. 5. Random Forest Score

B.Evalution

The customer churn prediction system provides a user-friendly interface where users can input customer details such as demographics, transaction history, and usage patterns to assess the likelihood of churn. The system is powered by a backend machine learning model trained to analyze these inputs and predict customer churn with high accuracy. After submission, the model processes the data and delivers the results on a

dedicated prediction results page. This page presents the outcome in a clear format, such as a probability score, risk category (e.g., high, medium, low), or visual indicators like color-coded statuses. Additionally, it highlights key metrics and factors influencing the prediction, enabling users to gain actionable insights and make informed decisions to address potential churn risk.

VII. Conclusion and Future Work

The primary goal of this study was to create a churn prediction system for the telecom industry, aiming to improve customer retention and increase revenue. To ensure accurate predictions, the system relied on achieving high AUC (Area Under the Curve) values. The dataset was split into 80% for training and 20% for validation, with hyperparameter tuning applied to optimize model performance. Feature engineering techniques were used to enhance the data for machine learning algorithms.

One of the main challenges encountered was class imbalance, with only 5% of customers identified as churned. This issue was addressed through under-sampling and the use of tree-based algorithms like Decision Tree, Random Forest, and Gradient Boosting Machine (GBM), which are less sensitive to data imbalance. These algorithms were chosen for their effectiveness in handling complex data and providing reliable churn predictions.

However, the study revealed areas for improvement. While tree-based models performed well, they may not be the best solution in all cases, suggesting the need for exploring alternative machine learning techniques. Additionally, external factors such as market dynamics, competitor actions, and customer sentiment were not incorporated into the model, limiting its scope. Future work should consider including these variables and experimenting with advanced algorithms like deep learning to improve model accuracy and applicability.

A. Future Work

Future work could focus on several key areas. First, The proposed churn prediction system establishes a robust foundation for future improvements that can greatly enhance telecom providers' capabilities in managing and reducing customer churn. A significant opportunity for advancement is the integration of real-time data streams, which would facilitate immediate analysis and prediction of churn. This capability would enable telecom companies to respond promptly to churn signals and implement effective measuresas they arise. Another promising area for development is the adoption of advanced algorithms. As machine learning and artificial intelligence technologies progress, incorporating cutting- edge techniques such as deep learning, reinforcement learning, and hybrid models could further refine prediction accuracy and better address complex customer behaviors.

References

[1] Gerpott, T.J., Rams, W., & Schindler, A. (2001). Analysis of customer retention, loyalty, and satisfaction within the German mobile cellular telecommunications market. Telecommunications Policy, 25, 249–269.

[2] Wei, C.P., & Chiu, I.T. (2002). Utilizing telecommunications call data for churn prediction through data mining techniques. Expert Systems with Applications, 23(2), 103–112.

[3] Qureshi, S.A., Rehman, A.S., Qamar, A.M., Kamal, A., & Rehman, A. (2013). A machine learning approach for predicting telecommunications subscribers' churn. In Proceedings of the Eighth International Conference on Digital Information Management, 131–136.

[4] Ascarza, E., Iyengar, R., & Schleicher, M. (2016). Risks associated with proactive churn prevention via plan recommendations: Insights from a field experiment. Journal of Marketing Research, 53(1), 46–60.

[5] Bott, R. (2014). Using multilayer perceptron neural networks for customer churn prediction in the telecommunications industry: A modeling and analysis approach. IGARSS, 11(1), 1–5.

[6] Umayaparvathi, V., & Iyakutti, K. (2016). A review of customer churn prediction in the telecommunications sector: Datasets, methodologies, and metrics. International Research Journal of Engineering and Technology, 3(4), 1065–1070.

[7] Yu, W., Jutla, D.N., & Sivakumar, S.C. (2005). A model for aligning churn strategies with management in mobile telecom. In Proceedings of the Communication Networks and Services Research Conference (Vol. 3, pp. 48–53).

[8] Burez, D., & den Poel, D.V. (2009). Addressing class imbalance in customer churn prediction. Expert Systems with Applications, 36(3), 4626–4636.

[9] Zhan, J., Guidibande, V., & Parsa, S.P.K. (2016). Identification of influential communities in large networks. Journal of Big Data, 3(1), 16. https://doi.org/10.1186/s40537-016-0050-7.

[10] Barthelemy, M. (2004). Analysis of betweenness centrality in complex networks. European Physical Journal B, 38(2), 163–168. https://doi.org/10.1140/epjb/e2004- 00111-4.

[11] Elisabetta, E., Meyerhenke, H., & Staudt, C.L. (2014). Approximation techniques for betweenness centrality in evolving large networks. CoRR. arXiv:1409.6241.

[12] Brandusoiu, I., Toderean, G., & Ha, B. (2016). Approaches for predicting churn in the prepaid mobile telecommunications sector. In Proceedings of the International Conference on Communications, 97–100.

[13] Choudhury, A., & Rahman, S. "A Comparative Study of Predictive Models for Customer Churn Prediction in Telecom Industry." International Journal of Data Science and Analytics, vol. 9, no. 2, pp. 151-162, 2020.

[14] Guha, S., & Chakraborty, S. "Churn Prediction in Telecom Industry Using Big Data Analytics." Journal of Big Data, vol. 4, no. 1, pp. 2, 2017.

[15] Rahman, M., & Hossain, M. "The Impact of Customer Feedback on Churn Prediction: A Study on the Telecom Industry." International Journal of Information Management, vol. 51, pp. 102-112, 2020.

19% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Exclusions

- 3 Excluded Sources

Match Groups

171 Not Cited or Quoted 18%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

11 Missing Citation 1%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
14% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

1	Name of the Student	Mukka Dheeraj			
2	Email ID and Phone Number	21951A0538@iare.ac.in 7815855613			
3	Roll Number	21951A0538			
4	Date of submission				
5	Name of the Guide	Ms.K.Sangeeta			
6	Title of the project work/ research article	Optimizing Telecom Operations with Segmentation and Churn Prediction			
7	Department	Computer Science and Engineering			
8	Details of the payment				
9	No. of times submitted	**First / Second / Third** (First time – Free; Second time – Rs 200/-; Third – Rs 500/- ;There after multiple of third)			
10	Similarity Content (%) (up to 25%acceptable)	**1st**	**2nd**	**3rd**	**4th**
For R & D Centre Use					
Date of plagiarism check					
Similarity report percentage					
R&D staff Name and Signature					

I / We hereby declare that, the above mentioned research work is original & it doesn't contain any plagiarized contents. The similarity index of this research work is……..

Justification for similarity index:

……...……...……..

……...……...……..

……...……...……..

……...……...……..

……...……...……..

……...……...……..

……...……...……..

Signature of Student Signature of the Guide