

**Building a machine learning model to predict customer churn in telecom services, enabling proactive retention strategies and reducing customer attrition.**

**A dissertation submitted in partial fulfillment of the requirements for the award of the Degree of**

# Bachelor of Technology

In

# Computer Science and Engineering

By

**Reddy Yaswanth Kumar (23U61A0559)**

**Under the guidance of**

**Mrs. T Lakshmi Lavanya**

B. Tech., M. Tech.

**Assistant Professor**

****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**(Approved by AICTE, New Delhi & Affiliated to JNTUH) (Recognized under section 2(f) of UGC Act 1956)**

**An ISO:9001-2015 Certified Institution**

**CHILKUR (V), MOINABAD (M), R.R. DIST. T.S-501504**

**June 2025**

****

**(Approved by AICTE & Affiliated to JNTUH) (Recognized under Section 2(f) of UGC Act 1956) An ISO:9001-2015 Certified Institution**

Survey No. 179, Chilkur (V), Moinabad (M), Ranga Reddy Dist. TS.

**JNTUH Code (U6) ECE –EEE-CSD-CSM – CSE - CIVIL – ME – MBA - M.Tech EAMCET Code - (GLOB)**

**Department of Computer Science and Engineering**

**Mrs. T Lakshmi Lavanya** Date: 02-06-2025

B. Tech., M. Tech.

**Assistant Professor & Head**

## CERTIFICATE

This is to certify that the project work entitled “**Building a machine learning model to predict customer churn in telecom services, enabling proactive retention strategies and reducing customer attrition.**”, is a bonafide work of **Boddula Reshwanth kumar (HT.No:16U61 A05),** submitted in partial fulfillment of the requirement for the award of **Bachelor of Technology in Computer Science and Engineering** during the academic year 2024-25. This is further certified that the work done under my guidance, and the results of this work have not been submitted elsewhere for the award of any other degree or diploma.

**Internal Guide Head of the Department**

**Mrs. T Lakshmi Lavanya Mrs. Noore Ilahi**

**Assistant Professor Assistant Professor**

## DECLARATION

I hereby declare that the project work entitled **Building a machine learning model to predict customer churn in telecom services, enabling proactive retention strategies and reducing customer attrition.,** submitted to **Department of Computer Science and Engineering, Global Institute of Engineering & Technology, Moinabad,** affiliated to **JNTUH, Hyderabad** in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering** is the work done by me and has not been submitted elsewhere for the award of any degree or diploma.

**Reddy Yaswanth Kumar (23U61A0559)**

## ACKNOWLEDGEMENT

I am thankful to my guide Mrs. T Lakshmi Lavanya**,** Assistant Professor of CSE Department for her valuable guidance for successful completion of this project.

I express my sincere thanks to **Mrs. G. Pavani,** Project Coordinator for giving me an opportunity to undertake the project “**Building a machine learning model to predict customer churn in telecom services, enabling proactive retention strategies and reducing customer attrition.**” and for enlightening me on various aspects of my project work and assistance in the evaluation of material and facts. She not only encouraged me to take up this topic but also given her valuable guidance in assessing facts and arriving at conclusions.

I am also most obliged and grateful to **Mrs. Noore Ilahi,** Assistant Professor and Head, Department of CSE for giving me guidance in completing this project successfully.

I express my heart-felt gratitude to our Vice-Principal **Prof. Dr. G Ahmed Zeeshan**, Coordinator Internal Quality Assurance Cell (IQAC) for his constant guidance, cooperation, motivation and support which have always kept me going ahead. I owe a lot of gratitude to him for always being there for me.

I also most obliged and grateful to our Principal **Dr. P. Raja Rao** for giving me guidance in completing this project successfully.

I also thank my parents for their constant encourage and support without which the project would have not come to an end.

Last but not the least, I would also like to thank all my class mates who have extended their cooperation during our project work.

**Reddy Yaswanth Kumar (23U61A0559)**

**VISION**

The Vision of the Department is to produce professional Computer Science Engineers who can meet the expectations of the globe and contribute to the advancement of engineering and technology which involves creativity and innovations by providing an excellent learning environment with the best quality facilities.

**MISSION**

**M1.** To provide the students with a practical and qualitative education in a modern technical environment that will help to improve their abilities and skills in solving programming problems effectively with different ideas and knowledge.

**M2.** To infuse the scientific temper in the students towards the research and development in Computer Science and Engineering trends.

**M3.** To mould the graduates to assume leadership roles by possessing good communication skills, an appreciation for their social and ethical responsibility in a global setting, and the ability to work effectively as team members.

**PROGRAMME EDUCATIONAL OBJECTIVES**

**PEO1:** To provide graduates with a good foundation in mathematics, sciences and engineering fundamentals required to solve engineering problems that will facilitate them to find employment in MNC’s and / or to pursue postgraduate studies with an appreciation for lifelong learning.

**PEO2:** To provide graduates with analytical and problem solving skills to design algorithms, other hardware / software systems, and inculcate professional ethics, inter-personal skills to work in a multi-cultural team.

**PEO3:** To facilitate graduates to get familiarized with the art software / hardware tools, imbibing creativity and innovation that would enable them to develop cutting edge technologies of multi disciplinary nature for societal development.

**PROGRAMME OUTCOMES:**

**PO1: Engineering knowledge:** An ability toApply the knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.

**PO2: Problem analysis:** An ability to Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural science and engineering sciences.

**PO3: Design/development of solutions:** An ability to Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal and environmental considerations.

**PO4: Conduct investigations of complex problems:** An ability to Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**PO5: Modern tool usage:** An ability to Create, select and apply appropriate techniques, resources and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

**PO6: The engineer and society:** An ability to Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**PO7: Environment sustainability:** An ability to Understand the impact of the professional engineering solutions in the societal and environmental contexts, and demonstrate the knowledge of, and the need for sustainable development.

**PO8: Ethics:** An ability to Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

**PO9: Individual and teamwork:** An ability to Function effectively as an individual and as a member or leader in diverse teams, and in multidisciplinary settings.

**PO10: Communication:** An ability to Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

**PO11: Project management and finance:** An ability to Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

**PO12: Lifelong learning:** An ability to Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broader context of technological change.

**PROGRAMME SPECIFIC OUTCOMES**

**PSO1:** An Ability to Apply the fundamentals of mathematics, Computer Science and Engineering Knowledge to analyze and develop computer programs in the areas related to Algorithms, System Software, Web Designing, Networking and Data mining for efficient Design of computer-based system to deal with Real time Problems.

**PSO2:** An Ability to implement the Professional Engineering solutions for the betterment of Society, and able to communicate with professional Ethics effectively

**ABSTRACT**

In the highly competitive telecom industry, customer retention is essential to maintaining profitability and sustaining growth. Customer churn—defined as the loss of clients or subscribers—poses a significant threat to telecom operators, leading to reduced revenues, increased marketing expenses, and loss of market share. To address this challenge, this project aims to develop a predictive machine learning model capable of identifying customers who are likely to churn, enabling companies to take timely and targeted retention actions.

The study utilizes a comprehensive dataset containing customer information, including demographic attributes, account details, usage behavior, service subscriptions, billing history, and customer support interactions. Extensive data preprocessing is conducted to handle missing values, encode categorical variables, normalize continuous features, and balance class distribution using techniques like SMOTE. Feature engineering is employed to create meaningful predictors, and exploratory data analysis is performed to uncover patterns and relationships between variables and churn behavior.

Multiple supervised machine learning algorithms are evaluated, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting Machines (e.g., XGBoost), Support Vector Machines (SVM), and Artificial Neural Networks. Model performance is assessed using metrics such as accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), with k-fold cross-validation employed to ensure generalizability and robustness.

The best-performing model is selected based on a trade-off between predictive power and interpretability. Feature importance analysis is conducted to identify the key drivers of churn, providing valuable insights into customer behavior. These insights empower telecom operators to design and implement proactive strategies, such as personalized offers, loyalty programs, and targeted communication, to retain high-risk customers.

The outcome of this project demonstrates the potential of machine learning to transform customer retention strategies from reactive to predictive. By leveraging data-driven decision-making, telecom companies can enhance customer satisfaction, improve lifetime value, and reduce overall churn rates—achieving a significant competitive advantage in the market.

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Particular** | **Page Number** |
|  | Title Page  Certificate  Declaration  Acknowledgement  Vision Mission  Abstract | i  ii  iii  iv  v-vi  vii |
| **1** | **INTRODUCTION**  1.1 Existing System  1.2 Disadvantages of Existing system  1.3 Proposed System  1.4 Advantages of Proposed System+ | **1-3** |
| **2** | **LITERATURE SURVEY**  2.1 LITERATURE [SURVEY](http://www.blurtit.com/q876299.html)  2.2 ABOUT PYTHON  2.3 ABOUT FLASK | **3-6** |
| **3** | **SYSTEM ANALYSIS**  3.1 REQUIREMENT SPECIFICATIONS  3.2 FEASIBILITY STUDY | **7-9** |
| **4** | **SYSTEM DESIGN**  4.1. System architecture of the proposed system  4.2. Proposed system model  4.3. Dataset description  4.4. Multiple dataset  4.5. Experimental analysis  4.6. Experiment analysis  4.7. Survival analysis  4.8. Cox proportional hazard model  4.9. Retention strategy | **9-26** |
| **5** | **SYSTEM IMPLEMENTATION**  5.1 main.py  5.2 app.py  5.3index.html | **27-37** |
| **6** | **SYSTEM TESTING**  1. Data Ingestion and Cleaning  2. Label Encoding and Feature Representation  3. Train/Test Split  4. Feature Scaling  5. Handling Class Imbalance with SMOTE  6. Gradient-Boosted Trees (XGBoost)  7. Hyperparameter Optimization  8. Probability Thresholding  9. Model Evaluation  10. Artifact Persistence for Deployment | **37-39** |
| **7** | **RESULTS** | **40-41** |
| **8** | **CONSLUSION** | **42** |
| **9** | **FUTURE ENHANCEMENT** | **43** |
|  | **REFERENCES** | **44-45** |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure Number** | **Figure Name** | **Page Number** |
| **1** | **4.1. System Architecture** | **10** |
| **2** | **Fig. 2. Count of target Variable per category** | **13** |
| **3** | **Fig. 3. Database attribute details** | **14** |
| **4** | **Fig. 4. Feature importance for decision tree** | **18** |
| **5** | **Fig. 5. Feature importance for Random Forest classifier** | **18** |
| 6 | **Fig. 6. Monthly charges by Churn.** | **21** |
| **7** | **Fig. 7. Total charges by Churn.** | **21** |
| **8** | **Fig. 8. All features in bar graph.** | **22** |
| **9** | **Fig. 9. Significance of covariate in predicting churn risk.** | **25** |
| **10** | **Fig. 10. Survival curve for the selected customers** | **26** |
| **11** | **Output 1, 2** | **41** |

### CHAPTER 1

**INTRODUCTION**

In the rapidly evolving telecommunications sector, customer retention has emerged as a crucial metric for organizational success. With customer acquisition becoming more expensive and competition intensifying, telecom providers are under increasing pressure to reduce churn. Churn refers to the rate at which customers discontinue using a service, and in the telecom industry, this figure typically ranges from 20% to 30% annually. High churn rates not only lead to revenue losses but also increase marketing costs and reduce overall customer lifetime value. Companies that can successfully predict and preempt churn gain a significant competitive advantage by retaining high-value customers and improving operational efficiency.

Traditional churn mitigation strategies are reactive, responding to customer exits rather than predicting and preventing them. This reactive stance results in lost opportunities to engage and retain customers before they decide to leave. To counter this, telecom operators are turning to machine learning models that can analyze vast datasets and detect early signs of churn. Among various machine learning techniques, the XGBoost algorithm stands out for its ability to handle large, complex datasets with high predictive accuracy.

This project proposes a predictive model based on XGBoost to identify customers at risk of churning. By analyzing usage patterns, customer demographics, billing behavior, and service interactions, the model generates churn probabilities that telecom companies can use to design targeted retention strategies. The predictive system is further enhanced with a real-time deployment pipeline, integrating with customer relationship management (CRM) systems to facilitate immediate and personalized engagement with at-risk customers.

**1.1 EXISTING SYSTEM:**

The current customer churn management system in most telecom environments relies on predefined business rules and static thresholds. For instance, customers who lodge frequent complaints or miss multiple bill payments are flagged as likely to churn. These systems depend on historical patterns and retrospective indicators, such as service cancellation notices, and often fail to anticipate churn in advance. They are supported by business intelligence tools that generate reports and dashboards but lack the predictive power needed for proactive retention.

Moreover, these systems typically analyze structured data in isolation, ignoring valuable behavioral and unstructured data like usage trends or call center interactions. The insights provided are often generalized across customer segments, limiting the ability to craft personalized interventions. As a result, many retention efforts are broad and ineffective, wasting resources on customers who were unlikely to churn.

* 1. **DISADVANTAGES OF EXISTING SYSTEM:**
* They lack real-time responsiveness, relying on retrospective data which delays intervention.
* High false-positive rates cause unnecessary expenditures on customers who are not at risk.
* These systems are not designed to scale with the increasing complexity and volume of telecom data.
* There is minimal use of machine learning or adaptive learning models, making the system rigid.
* Lack of integration with personalized campaign tools limits the scope of tailored customer engagement.

**1.3 PROPOSED SYSTEM:**

To overcome the drawbacks of the existing methods, we propose a data-driven churn prediction system built using the XGBoost machine learning algorithm. XGBoost is well-suited for this application due to its scalability, ability to handle missing data, and efficiency with large, imbalanced datasets typical in churn prediction. The system uses customer data from various sources—billing information, service usage logs, subscription history, and customer support interactions—to create a comprehensive customer profile.

This information is preprocessed and fed into the XGBoost model, which identifies patterns and correlations indicative of churn. The output is a churn probability score for each customer. This score enables telecom companies to rank customers by churn risk and tailor retention strategies accordingly. In addition to the core model, the proposed system features a dashboard for visualizing predictions and a REST API for integrating the model into production environments.

The model is validated through k-fold cross-validation, ensuring consistent performance across different data segments. It is continuously retrained with new data to adapt to evolving customer behaviors. This system empowers decision-makers to proactively manage churn with high precision, reducing marketing costs and improving customer satisfaction.

**1.4 ADVANTAGES OF PROPOSED SYSTEM:**

* Proactive insights: The system predicts churn before it occurs, allowing timely intervention.
* Higher accuracy: XGBoost’s optimized decision trees enhance prediction precision.
* Scalable architecture: Easily integrates with big data platforms and existing CRM tools.
* Dynamic learning: The model retrains with new data, continuously improving performance.
* Targeted retention: Facilitates segmentation-based and personalized retention strategies.

**CHAPTER 2**

**LITERATURE** [**SURVEY**](http://www.blurtit.com/q876299.html)

**2.1 LITERATURE** [**SURVEY**](http://www.blurtit.com/q876299.html)

**2.1.1 Machine Learning Approaches for Customer Churn Prediction**

**Author :** Idris, A., Khan, A., and Lee, Y. S.

This research investigates the use of machine learning algorithms for predicting customer churn in the telecom industry. The study primarily focuses on comparing various techniques such as decision trees, support vector machines (SVM), artificial neural networks (ANN), and ensemble models like random forests and gradient boosting.

The authors conducted extensive experiments using real-world telecom datasets and found that ensemble learning models consistently outperformed standalone classifiers in predictive accuracy and robustness.

One of the key insights of the study is the critical role of data preprocessing, including handling missing values, normalizing features, and encoding categorical variables. The researchers highlight that effective feature engineering and selection greatly influence the performance of churn prediction models. Moreover, the study emphasizes the importance of evaluating models using suitable metrics such as precision, recall, F1-score, and AUC-ROC, particularly when dealing with imbalanced datasets where the proportion of churners is significantly lower than non-churners.

The study concludes that ensemble learning, especially gradient boosting, is highly effective for churn prediction due to its ability to capture complex non-linear relationships between features. The paper serves as a foundational reference for telecom companies seeking to adopt machine learning for customer retention, advocating for a structured pipeline that includes data preparation, model training, evaluation, and deployment.

**2.1.2 Churn Prediction in Telecom Using Ensemble Learning**

**Author :** Ahmed, M. and Maheswari, P. U.

This paper presents a comprehensive evaluation of ensemble learning methods for predicting customer churn in the telecommunications industry. The authors investigate the predictive power of various classifiers including Naive Bayes, logistic regression, decision trees, AdaBoost, and XGBoost. Their experimental setup involves a telecom dataset with detailed customer profiles, including usage behavior, contract information, and billing history.

Through rigorous experimentation, the study demonstrates that ensemble models, particularly XGBoost, deliver superior accuracy and generalization capabilities. One notable finding is the improved performance of ensemble methods in handling class imbalance—a common issue in churn datasets. XGBoost, with its built-in regularization and gradient boosting mechanism, stands out for its scalability and performance.

The authors emphasize the practical applications of their findings by outlining how telecom providers can integrate these models into customer relationship management (CRM) systems. They suggest that the probabilistic churn scores generated by ensemble models can be used to trigger automated marketing campaigns or service recovery workflows.

In conclusion, this research confirms the effectiveness of ensemble learning in churn prediction and provides a roadmap for its adoption in real-world telecom environments. It underlines the need for continuous model monitoring and retraining to maintain prediction accuracy as customer behavior evolves.

**2.1.3 Predicting Customer Churn Using XGBoost and SHAP Analysis**

**Author :** Rahman, A., Chowdhury, F., and Hossain, S.

This study explores the dual goals of prediction accuracy and model interpretability in churn prediction systems. The authors leverage XGBoost, a powerful ensemble learning method, alongside SHAP (SHapley Additive exPlanations) values to build a transparent churn prediction model. The use of SHAP allows stakeholders to understand the impact of each feature on the model's output, addressing the common criticism that machine learning models are often “black boxes.”

The authors apply their framework to a public telecom dataset, incorporating features like tenure, contract type, monthly charges, and service usage metrics. Their model achieves high accuracy, demonstrating the capability of XGBoost to handle heterogeneous and high-dimensional data. SHAP analysis reveals that features such as contract type, tenure, and internet service type significantly influence churn decisions.

Importantly, the study discusses how model interpretability aids business users in designing effective retention strategies. By understanding why a customer is predicted to churn, telecom companies can tailor interventions that directly address customer pain points. For example, offering a contract upgrade to customers flagged due to short tenure and high charges.

This paper effectively bridges the gap between predictive performance and explainability, advocating for the combined use of XGBoost and SHAP as a best practice in churn modeling.

**2.1.4 Customer Retention Strategies Based on Churn Prediction Models**

**Author :** Kumar, V., and Reinartz, W.

This research delves into the strategic application of churn prediction models within business environments, focusing on how predictions can be translated into actionable customer retention strategies. The authors argue that predictive accuracy, while important, is not sufficient unless the outputs can be effectively operationalized within a business context. The study identifies key customer retention strategies that can be activated based on churn scores, including personalized discount offers, loyalty rewards, and enhanced customer service. It presents a structured framework where predicted churn probabilities are segmented into risk bands—low, medium, and high—each

Additionally, the research highlights the importance of aligning retention efforts with customer lifetime value (CLV). By prioritizing high-value customers for retention interventions, companies can maximize ROI. The authors also emphasize the role of automation in campaign management, suggesting that modern CRM systems should be integrated with predictive models to streamline and personalize outreach.

In conclusion, this study provides a practical guide for implementing churn models in telecom settings. It showcases how model predictions can inform business strategy, enhance customer satisfaction, and ultimately reduce attrition rates.

**2.2 ABOUT PYTHON**

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. An [interpreted language](https://en.wikipedia.org/wiki/Interpreted_language), Python has a design philosophy that emphasizes code [readability](https://en.wikipedia.org/wiki/Readability) (notably using [whitespace](https://en.wikipedia.org/wiki/Whitespace_character) indentation to delimit [code blocks](https://en.wikipedia.org/wiki/Code_block) rather than curly brackets or keywords), and a syntax that allows programmers to express concepts in fewer [lines of code](https://en.wikipedia.org/wiki/Source_lines_of_code) than might be used in languages such as [C++](https://en.wikipedia.org/wiki/C%2B%2B)or [Java](https://en.wikipedia.org/wiki/Java_(programming_language)). It provides constructs that enable clear programming on both small and large scales. Python interpreters are available for many [operating systems](https://en.wikipedia.org/wiki/Operating_system). [CPython](https://en.wikipedia.org/wiki/CPython), the [reference implementation](https://en.wikipedia.org/wiki/Reference_implementation) of Python, is [open source](https://en.wikipedia.org/wiki/Open_source) software and has a community-based development model, as do nearly all of its variant implementations. CPython is managed by the non-profit [Python Software Foundation](https://en.wikipedia.org/wiki/Python_Software_Foundation). Python features a [dynamic type](https://en.wikipedia.org/wiki/Dynamic_type) system and automatic [memory management](https://en.wikipedia.org/wiki/Memory_management). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigm), including [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming), [imperative](https://en.wikipedia.org/wiki/Imperative_programming), [functional](https://en.wikipedia.org/wiki/Functional_programming) and [procedural](https://en.wikipedia.org/wiki/Procedural_programming), and has a large and comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

**2.3 ABOUT FLASK**

Flask is a lightweight web application framework written in Python, widely used for building web services and deploying machine learning models. Unlike heavier frameworks, Flask is minimalistic and flexible, allowing developers to add only the components they need. For customer churn prediction projects, Flask is ideal for exposing machine learning models via RESTful APIs, enabling real-time predictions.

Flask supports easy integration with Python’s data science stack including pandas, NumPy, and scikit-learn. In a typical deployment scenario, the trained churn prediction model is serialized (e.g., using joblib) and served using a Flask-based web server. This API can then be consumed by front-end applications, CRMs, or analytical dashboards. Flask also allows simple routing, error handling, and testing, making it robust for production environments.

Due to its simplicity and scalability, Flask is preferred for developing and deploying machine learning applications in startups, research, and industry settings. It provides the core structure for model inference while allowing full control over request handling and data flow.

**CHAPTER 3**

**SYSTEM ANALYSIS**

The system analysis phase of this project involved evaluating existing application models to enhance user experience and accessibility. The primary objective was to optimize the user interface for intuitive navigation and minimal manual input. These improvements aim to reduce user effort and maximize engagement by streamlining interaction pathways across application screens.

As a web-based system, compatibility with widely used browsers was prioritized to ensure universal access. This design consideration facilitates broader usability across different user environments. Additionally, the back-end infrastructure and front-end technologies were selected to balance performance, maintainability, and scalability, while adhering to industry standards.

**3.1 REQUIREMENT SPECIFICATIONS**

**3.1.1 HARDWARE REQUIREMENTS:**

* **System :** Pentium IV 2.4 GHz.
* **Hard Disk :** 40 GB.
* **Floppy Drive :** 1.44 Mb.
* **Monitor** : 14’ Colour Monitor.
* **Mouse :** Optical Mouse.
* **RAM :** 16 GB

**3.1.2 SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 10
* **Coding Language :** Python.
* **Front-End :** Html,Css
* **Designing :** Html , Css , javascript.
* **Data Base :** MySQL.

**3.1.3 FUNCTIONAL REQUIREMENTS:**

* Graphical User interface with the User.

**Operating Systems supported**

1. Windows 10
2. Windows 8

3. Windows 11

**Technologies and Languages used to Develop**

1. Python (with Flask for deployment )

**Debugger and Emulator**

Compatible with modern web browsers (preferably Google Chrome)

**3.2 FEASIBILITY STUDY:**

This stage assesses the viability of the proposed churn prediction system and presents a general outline for implementation, including preliminary cost considerations. The feasibility study ensures that the new system is achievable, sustainable, and advantageous for stakeholders. This assessment is divided into three major areas:

**Three key considerations involved in the feasibility analysis are,**

* **ECONOMICAL FEASIBILITY**
* **TECHNICAL FEASIBILITY**
* **SOCIAL FEASIBILITY**

**3.2.1 ECONOMICAL FEASIBILITY**:

This aspect examines the financial viability of the system. With limited budget allocations typical in organizational environments, cost-efficiency is essential. The churn prediction system has been developed predominantly using open-source tools and platforms, significantly reducing the overhead. Technologies like Python, Flask, and MySQL incur no licensing fees, and only specific proprietary modules, if any, require procurement. The system’s design supports minimal operational expenses and delivers measurable returns through improved customer retention.

### 3.2.2 TECHNICAL FEASIBILITY:

Technical feasibility evaluates whether the system requirements align with available technological resources. The proposed churn prediction system demands only modest hardware and uses widely supported software tools. The reliance on Flask ensures lightweight deployment, and Python’s robust ecosystem provides tools for machine learning, database interaction, and web development. The architecture is designed to run efficiently on standard computing infrastructure, ensuring broad accessibility and ease of integration with existing CRM systems.

**3.2.3 SOCIAL FEASIBILITY:**

This section measures the likelihood of system acceptance among end-users. Training and documentation are provided to ensure ease of use, and the system interface has been developed with user-friendly features that minimize resistance. By offering actionable insights through a simple dashboard, users can engage meaningfully with the predictions without requiring deep technical knowledge. Furthermore, feedback mechanisms are included to accommodate constructive input, fostering a positive user relationship and system adaptability.

**CHAPTER 4**

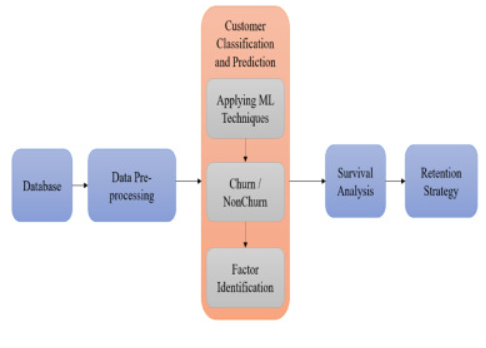
**SYSTEM DESIGN**

### SYSTEM ARCHITECTURE:

In this section, system architecture and proposed system model are discussed subsequently.

4.1. System architecture of the proposed system

The implementation for churn will require the latest version of Anaconda with built in features that consist of Jupiter notebook for training and testing data. The latest version of Anaconda with built-in functionality, like Jupiter notebook for training and testing data, will be required for the churn implementation. Churn predictions for the telecom industry have been carried out using literature with various methods that includes [machine learning algorithms](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm), [data mining techniques](https://www.sciencedirect.com/topics/computer-science/data-mining-technique) and retention strategies. These techniques effectively support many companies for predicting, identifying, and retaining churners which help in CRM (Customer relationship management) and decision making. CRM deals with the data to identify a loyal customer for industry. High revenue generating customers (loyal customers) for a company have no impact on the competitor companies. Such loyal customers help to grow profitability of a company by referring to the other people such as their family members, colleagues, and friends. Hence, the role played by CRM is very important in churn prediction and it also helps to retain the churning customers. [Fig. 1](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0001) depicts a cycle through which churn prediction can be made. For prediction there are many algorithms such as [Support vector machine](https://www.sciencedirect.com/topics/engineering/support-vector-machine), K-nearest neighbor, J48, naive Bayes, [logistic regression](https://www.sciencedirect.com/topics/computer-science/logistic-regression), LWL, [Random Forest](https://www.sciencedirect.com/topics/computer-science/random-decision-forest), [Decision tree classifier](https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier) which are used to resolve classification problems. Random forest and Decision tree classifier are considered relevant with better accuracy and performance.



*4.1. System Architecture*

**4.2. Proposed system model**

* Let S is the proposed system.
* S={I,O,DP,FS,EF}
* Where
* 1) I (Input): Dataset
* 2) (Output): CM (Confusion matrix) ={R->Actual False, Actual True} {C->Predicted False, Predicted true}
* 3)DP (Data Processing)
* 4)FS (Feature Selection) is measured by [Pearson](https://www.sciencedirect.com/topics/engineering/pearsons-linear-correlation-coefficient) correlation formula (Cr) where set of numerical attributes is taken ‘X’ ranging from X1, X2, X3……., Xn consider a set of attributes X, form a subset of two attributes each, *X*= {{X1, X1,}, {X1, X2}, {X1, X3} .... {Xi, Xj}… {Xn, Xn}}

If each subset = X then evaluates ∑Xi, ∑Xj2 & ∑ (Xi \* Xj)

Cr= √∑ (Xi \* Xj) ÷ ∑Xi2 \*∑Xj2

Where Cr = 1 (+ve correlation), Cr = 0 (no correlation), Cr = −1 (-ve correlation)

Convert string values into numerical (S→N)

Then apply [ML](https://www.sciencedirect.com/topics/computer-science/machine-learning) algorithm on new Dataset (consider as C)

* 5) EF – Efficiency of Proposed Model

The proposed system for churn prediction is derived using accuracy, precision, recall, f-measure. Accuracy calculates the accuracy metric.

* + TP=True positive, TN=True negative, FP=False positive, FN=False [negative values](https://www.sciencedirect.com/topics/computer-science/negative-value), AP=Actual positive, AN=Actual negative.
  + a)Accuracy= (TP + TN) ÷ (TP + TN + FP + FN)
  + b)[True positive](https://www.sciencedirect.com/topics/computer-science/true-positive) and true negative are values of the confusion matrix after applying [classification algorithms](https://www.sciencedirect.com/topics/engineering/classification-algorithm). True positive rate is the value that shows us which part of data is classified as correct and false positive classifies incorrect values.
  + c) (TP) rate = TN ÷ TN +FP
  + d) (FP) rate= FP ÷ FP+TN
  + e) Precision = TP ÷ (TP + FP)
  + f) Recall = TP ÷ (TP + FN)
  + g) F-Measure = (2 × Precision × Recall) / (Precision + Recall)

***Constraints***: The time is the key for predicted churners. If C (customer) churns after one month, but sometimes dataset shows that it would churn after one week that will lead to incorrect churn. Hence, C depends upon T. Incorrect dataset (ID) with irrelevant features will be other constraints to predict churn model.

***Failure of S***: If customer (C) gets churned out for one month time (T) but then customer rejoins again would lead to loss of company and system (S).

**4.3. Dataset description**

Here, specifically focus is on Single Dataset presented in sub [Section 4.3.1](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "sec0006). Sequentially, [Section 4.3.2](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "sec0009) emphasizes multiple datasets to validate proposed system.

**4.3.1. Single dataset**

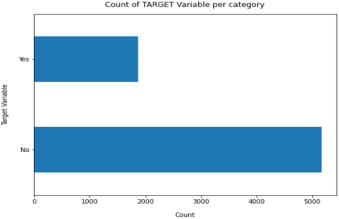
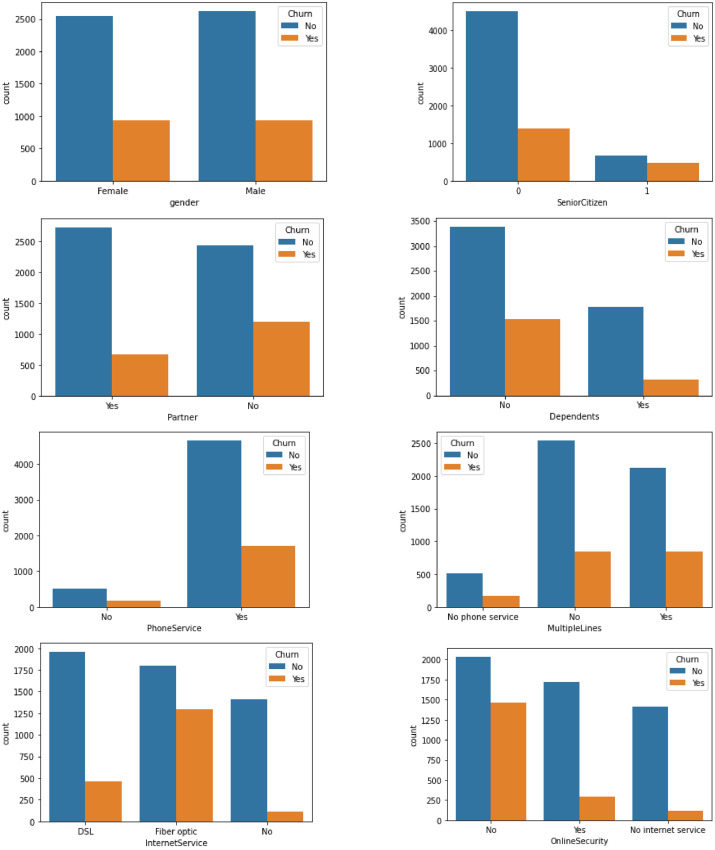
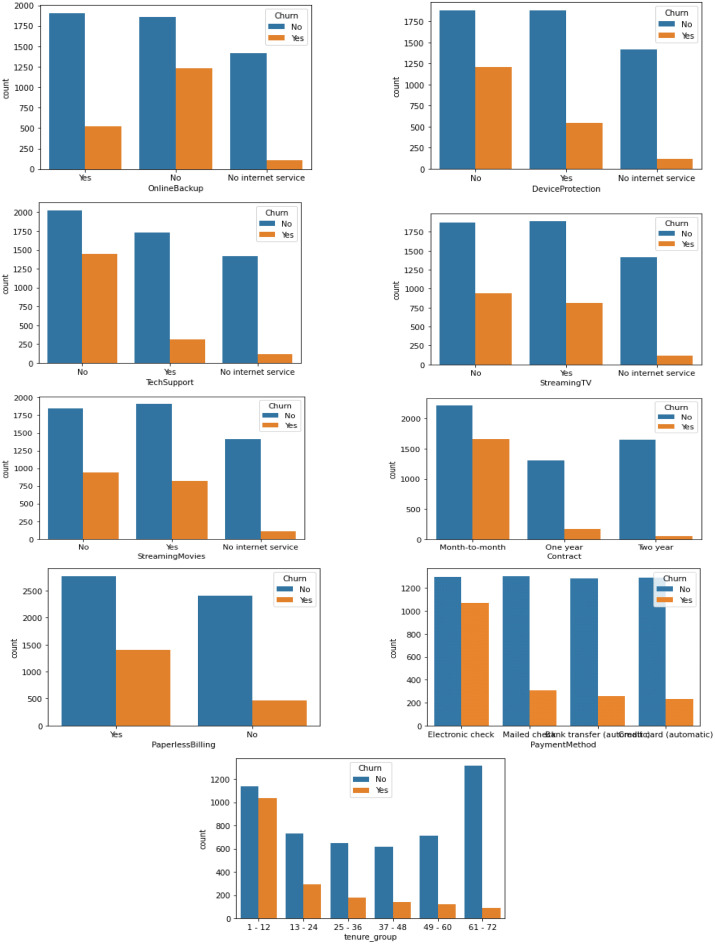
The dataset used for experiments in this paper, contains results of Telco-Customer-Churn dataset obtained from Kaggle website (it is also known as IBM Watson dataset which was released in 2015).Each row represents a customer, each column contains attribute described on the [column Metadata](https://www.sciencedirect.com/topics/computer-science/metadata-column). It consists of 7043 customer information. Every customer has 21 features and the “Churn” it contains 11 missing values in the Total Charges column. The last attribute contains labelled data with two classes where 26.53 % of total customers are labelled as ‘‘T’’ indicating true customers i.e., categorized as churning customers and the remaining 73.46 % customers are labelled as ‘‘F’’ indicating false customers i.e., categorized as non-churning customers. The attribute selection depends on the results of techniques of feature selection that find useful, the most similar and effective attributes to predict the churning customers. A total of 5174 are non-churners and 1869 are churners. The dataset contains 16 categorical columns and 5 numeric columns. The dataset helps to figure out customer prophecy and build retention possibilities. [Fig. 2](https://www.sciencedirect.com/science/article/pii/S2666720723001443#fig0002), [Fig. 3](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0003) shows details of database. 

Fig. 3. Database attribute details.





4.3.2. Data pre-processing

This step is required to remove all the irrelevant and dirty data of real world. As the data is congregated from many resources it is important to overcome this issue. Without execution of this step [decision makers](https://www.sciencedirect.com/topics/computer-science/decision-maker) cannot predict outcomes even if they did it won't be correct that leads to no quality data and no quality mining. To solve these issues following methods are considered for cleaning the data. Likewise, there are many such features taken as input to predict model. Classification of data and performing pre-processing cleans the data and makes it easy to use.

[Data classification](https://www.sciencedirect.com/topics/computer-science/data-classification) and pre-processing clean the data and make it easier to use. There are three steps in data preparation.

• A data comparison is carried out in the first stage to identify redundant data. The properties that are repeated are immediately removed.

• Each instance searches for the missing value in step two. If missing values are detected, replacement procedures are used; otherwise, missing values are replaced with the most suited value. When there is no way to replace the instances, they are discarded.

• The final stage entails defining the type of each value in order to eliminate extraneous data. If the information is no longer useful, it is discarded. In other words, noisy data is deleted from the data that has already been pre-processed. Check the [data types](https://www.sciencedirect.com/topics/engineering/data-type) of each column first.

Because the total charges column's data type is object type, we convert it to numeric and make a duplicate of the underlying data to manipulate and process. This conversion is performed using panda's library imported in the program and using to-numeric function of library. Next, check for duplicate values in dataset, but there are no duplicate values present in dataset. Further continue checking for missing values. So according to general thumb rule the feature with less missing values there to fill means values or simply use regression to predict the missing values based on the particular feature. Similarly, in case of feature with high number of missing values, it would better to drop those columns due to less analysis and insights. Also, the columns with more than 30–40 % are deleted. Now by using is null and sum function there are 11 missing values present in the Total Charges column so drop this missing value. The tenure consists of maximum value of 72 and also dividing the tenure into 6 categories of 1–12, 13–24, 25–36, 37–48, 49–60, 61–72 make easy in visualization of column data of tenure period. Dropping the columns that are not required for processing. It was found the customer ID is not useful and contains unique value which won't affect the prediction results, so drop customer ID and similarly drop the tenure column which is not necessary to evaluate the results.

|  |  |
| --- | --- |
| **1–12** | 2175 |
| **61–72** | 1407 |
| **13–24** | 1024 |
| **25–36** | 832 |
| **49–60** | 832 |
| **37–48** | 762 |

Name: tenure\_group, dtype: int64

*Univariate analysis* - In data exploration, the distribution of individual predictions by churn will be determined, and the number of churners representing non-seniors will be determined. Here, 1 represents senior citizens, and the plot demonstrates that if a client is a senior citizen, they are more likely to churn. Similarly, when a candidate is single and does not have a partner, the partner churn ratio is high. People with phone service are more likely to churn, similarly payment method, if there is an electronic check appears higher than the case of credit card appears lowest churner because of this they could be having auto debit features, which is one of the important features for churn, similarly remaining people with phone service are more likely to churn, similarly remaining people with phone service are more likely to churn, similarly remaining people with phone service are more likely to churn, remaining figures also shows the churn count respectively above selected features.

*Bivariate analysis* - It is used to find a value prediction for a [single variable](https://www.sciencedirect.com/topics/computer-science/single-variable). Correlations between variables are simple to find. A relationship between two variables is defined as bivariate. There are numerous features in our dataset, and we presented the results. Two variables were examined, and two new data frames for churners and non-churners were generated. A function is created that maintains a data frame that is passed with column, title, and hue information for each feature, similar to how bar graphs for different features may be shown in diagrams. Gender characteristics considers 2500 female and male participants, with a churner/non-churner ratio of around 50 % for each gender. According to gender feature analysis, females are more likely to churn if they have a relationship, but males are more likely to churn if they do not have a partner. The groups are classified into churners and non-churners.

**4.3.3. Feature selection**

This is an important step in achieving our model's goal. Unnecessary data is discovered in datasets while training the model, resulting in a reduction in model accuracy. As a result, feature selection on a dataset is used to solve these issues.

The following are the advantages of feature selection:

• Reduced over fitting means less chance of making conclusions based on noise.

• Accuracy is improved because there is fewer misleading data.

• Training time is reduced providing lesser [algorithm complexity](https://www.sciencedirect.com/topics/computer-science/algorithmic-efficiency) with algorithms that train faster.

• Applying Machine Learning Algorithms

[1] [*Decision tree*](https://www.sciencedirect.com/topics/computer-science/decision-trees): Decision tree are used to solve both classification and [regression problem](https://www.sciencedirect.com/topics/computer-science/regression-problem) in the form of trees that can be incrementally updated by splitting the dataset into smaller dataset, where the result are represented by the [leaf node](https://www.sciencedirect.com/topics/engineering/leaf-node). Each branch represents the possible decision outcome or reaction. It is like a flowchart diagram that shows the various outcomes from a series of decisions. It can be used as a decision-making tool, Decision tree has some series of same craft questions regarding attributes of test data record and it's use to solve classification-based problems, every time it gets solution from it and follows until the final conclusion of class label record. Then visit several decision trees for achieving target value. It can be either true or false. Now pick a majority vote of trees or count the target values provided, then based upon decision trees predict if customer churn is true or false for research analysis, or for planning strategy. A primary advantage for using a decision tree is that it is easy to follow and understand. Decision tree classifier is simple and adaptive [classification technique](https://www.sciencedirect.com/topics/computer-science/classification-technique), this is basically implying a straightforward process to analyze and solve the problem.

[2] *Random Forest tree*: The random forest is a classification algorithm consisting of many decision trees. More numbers of trees in the forest led to more robustness in prediction with higher accuracy. It uses bagging and feature randomness, when building each individual tree to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree. In this algorithm, each tree will have its own output from a dataset provided, such that output generated will be considered from the majority of trees. Decision trees made in this algorithm are of numeric type in which the tree picks any random attribute in the dataset. Advantages of Random Forest that it helps to solve both regression and classification problems.

[3] *Important Features or factors responsible for churning according to*[*Decision Tree model*](https://www.sciencedirect.com/topics/computer-science/decision-tree-model): Finding the important factors that are responsible for churning makes it possible to find the service that is required to customer to prevent from attrition. This can be done using feature importance. Here the feature is ranked according to their importance. The most important feature at the top of list, while least important are at end of list.

In [Fig. 4](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0004) it is seen that Contract-Month-To-month ranked first in the list with its importance as 0.517, Total Charges with importance as 0.104, No Internet Service is 0.093, DSL Internet Service as 0.0795, Monthly Charges as 0.0517, Contract of Two years as 0.0464. Contract of one year as 0.0419, tenure group of 1–12 as 0.0397, No Streaming Movies as 0.00696, [Fiber Optic](https://www.sciencedirect.com/topics/engineering/fiber-optics) Internet Service as0.00643, and last in list is with No Paperless Billing as 0.004727.

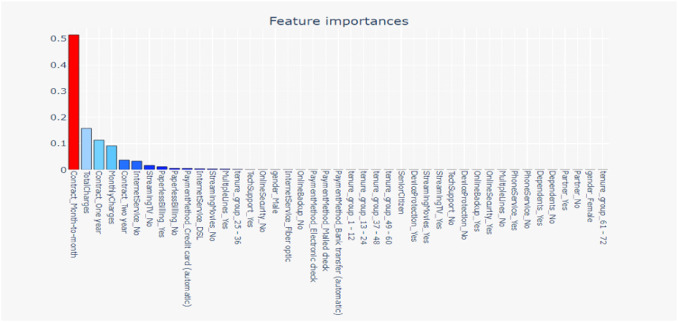


Fig. 4. Feature importance for decision tree.

[4] *Important Features or factors responsible for churning according to Random Forest Tree model*: In this model from [Fig. 5](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0005) the features importance ranked from 1st to last as,

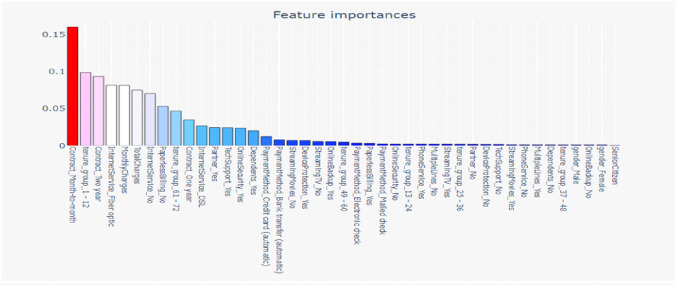


Fig. 5. Feature importance for Random Forest classifier.

• Contract month to month = 0.1245

• Tenure group = 0.10518

• Internet Service [Fiber Optic](https://www.sciencedirect.com/topics/materials-science/fiber-optics) = 0.07693

• Total Charges = 0.07000

• Contract Two year = 0.06123

• Tenure group 61 -72 = 0.04302

• Online Security yes = 0.03977

• No tech support = 0.03943

• Online Backup No Internet Service = 0.0360

• Streaming Movies No Internet Service = 0.03479

• Internet Service DSL = 0.0338

• Monthly Charges = 0.0309

• Online Security No Internet Service = 0.0294

• Contract One year = 0.02819

• No Online Security = 0.02620

• Tech Support Yes = 0.02491

• Tech Support No Internet Service = 0.024313

• Partner Yes = 0.02422

• No Multiple Lines = 0.001289

• Gender Male = 0.0011822

• Multiple Lines No phone service = 0.0011689

• No phone Service = 0.001107

• Gender Female = 0.001101

**4.4. Multiple dataset**

There exist multiple types of data in SyriaTel [[25]](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "bib0025) as classified below which may be applied to construct our churn prediction model used in telecom sector. It is observed to present dataset structure using spark engine, it is required to include phase-based exploration with suitable pre-preparation for algorithms based on classification.

**4.4.1. Towers and complaints database**

In this database, the detailed data of actual location is shown as in the form of digits. The serial digits are mapped with database of towers which offers the actual location of the transaction and providing state, city, area, sub-area, latitude and longitude. Database including various complaints provides all submitted complaints along with inquiries statistics based on coverage, any problem related to the telecom business and issues related to packages and offers.

**4.4.2. Customer data**

Customer data comprises the data-based contact information and services of customers. Moreover, various packages, services and offers taken by customer. Also, it contains [CRM system](https://www.sciencedirect.com/topics/computer-science/customer-relationship-management-system) including information generated from all customer GSMs such as gender, birth date, the location and type of subscription etc.

**4.4.3. Network logs data**

Network logs data includes sessions related to calls, SMS and internet for each transaction used by telecom operator such as required time to initialize a session for call ending and to check internet status. Moreover, it represents whether session is expired or not due to error occurred in the internal network.

**4.4.4. Mobile IMEI information**

Mobile IMEI information comprises the model, type, model of the mobile phone and whether mono or dual SIM device. Data may have large size which may require information in detailed. It requires a lot of time for understanding. It also needs to know the original sources along with format for storage. Moreover, related to these records, data must be linked to each other logically using [relational databases](https://www.sciencedirect.com/topics/computer-science/relational-database) that actually represent customers detailed information.

**4.4.5. Call details records**

Call Details Records (CDRs) includes modifiable data related to MMS, calls, SMS etc. Also, transaction made by customers using internet which is ultimately generated in the form of text files.

**4.5. Experimental analysis**

After preprocessing on dataset following are the observations:

• Strong correlation exists between tenure and total charges, means as tenure increases so does total charges.

• Strong correlation exists between monthly charges and total charges as well.

• Tenure and Contract duration seems to be strong factors in determining churn.

• Among service types, phone service seems to be most popular.

• CSP should investigate if customers receiving digital invoice have any concern with understanding the bill details.

• Also, they should encourage customers to move to automated payment modes to improve customer experience.

• Gender does not play an important role. However, CSPs should take care of the experience of senior citizens.

From [Fig. 6](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0006), [Fig. 7](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0007) it is seen that churn is higher when monthly rates are high. Even with modest overall charges, there is large churns, as shown in the [Fig. 7](https://www.sciencedirect.com/science/article/pii/S2666720723001443#fig0007). When all three parameters (Total Charges, Monthly Cost, and Tenure) are combined, higher monthly charges at low tenure result in lower total charges, implying that all of these characteristics are associated to higher churn. [Fig. 8](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0008) shows used all features.

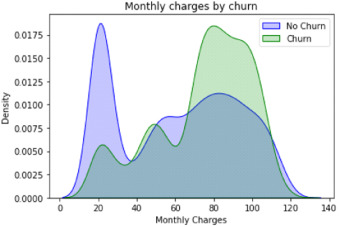


Fig. 6. Monthly charges by Churn.

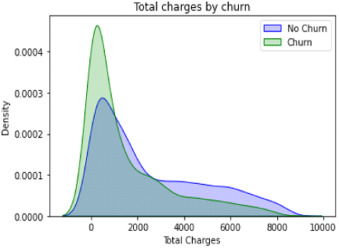


Fig. 7. Total charges by Churn.

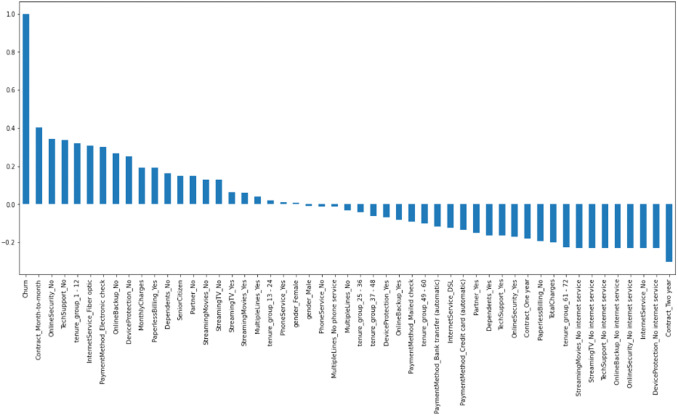


Fig. 8. All features in bar graph.

**4.6**. Experiment analysis using decision tree classifier So, in our proposed model using a Decision Tree classifier, the obtained accuracy of the model is 78 %, which is very low, and printing the classification report led to the dataset unbalance, which results in less accuracy. Decision Tree classifier took Number of Leaves - 331, Size of the tree - 552, Time taken to build model-1.06 s for execution.

As a result, the accuracy of the Decision tree [classifier model](https://www.sciencedirect.com/topics/computer-science/classifier-model) before up-sampling & ENN should not be used as a meaningful measure because it leads to unbalanced datasets. As a result, when checking recall, precision, and F1 scores for the minority class, it's clear that the precision, recall, and F1 ratings for Class 1, i.e. churned consumers, are far too low.

For up-sampling [training data](https://www.sciencedirect.com/topics/computer-science/training-data) into a decision tree classifier that differentiates into *x* train and *y* train and creates a prediction variable and calls the classification input to process input to produce output with accuracy, and using SMOTE (synthetic minority oversampling technique) by performing oversampling and cleaning using ENN (Edited Nearest Neighbors), the dataset is balanced with dataset values of 493, 40, and 599, 34, and it provides a solution. [Tables 1](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "tbl0001) and [2](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "tbl0002) shows details results of Decision tree classifier model before and after up-sampling & ENN. The precision of the classifier matrix is 93 %, the recall factor is 93 %, and the F1 score is 93 % and provides 93.85 % accuracy.

Table 1. Results of decision tree classifier model before up-sampling & ENN.

| Empty Cell | **Precision** | **Recall** | **F1-score** | **Support** |
| --- | --- | --- | --- | --- |
| **Class 0 (No)** | 0.83 | 0.89 | 0.86 | 1046 |
| **Class 1(Yes)** | 0.61 | 0.49 | 0.54 | 361 |
| **Weighted avg** | 0.77 | 0.78 | 0.77 | 1407 |

| Empty Cell | **No. of Instances** | **Percentage** |
| --- | --- | --- |
| **Correctly Classified Instances** | 5493 | 77.9923% |
| **Incorrectly Classified Instances** | 1550 | 22.0077% |

Table 2. Results of decision tree classifier model after up- sampling & ENN.

| Empty Cell | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- |
| **Class 0 (No)** | 0.93 | 0.92 | 0.93 |
| **Class 1(Yes)** | 0.94 | 0.94 | 0.94 |
| **Weighted avg** | 0.93 | 0.93 | 0.93 |

| Empty Cell | **No. of Instances** | **Percentage** | Empty Cell |
| --- | --- | --- | --- |
| **Correctly Classified Instances** | 6615 | 93.85% |  |
| **Incorrectly Classified Instances** | 427 | 6.15 % |  |

4.6.1. Experimental analysis using random forest algorithms

Missing values are handled with carefully, and accuracy is maintained. It's even capable of handling big, multi-dimensional datasets. So, the accuracy of our model using the Random Forest Tree classifier is 98.91 %. [Table 3](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "tbl0003) shows results of [Random Forest classifier](https://www.sciencedirect.com/topics/computer-science/random-forest-classifier) model before Up-sampling and ENN. There are several techniques used in random forests. Generally bagging technique is used known as [ensemble classifier](https://www.sciencedirect.com/topics/computer-science/ensemble-classifier).

Table 3. Results of Random Forest tree classifier model before Up-sampling & ENN.

| Empty Cell | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- |
| **Class 0 (No)** | 0.99 | 1.00 | 0.99 |
| **Class 1(Yes)** | 0.99 | 0.96 | 0.97 |
| **Weighted avg** | 0.99 | 0.99 | 0.99 |

| Empty Cell | **No. of Instances** | **Percentage** | Empty Cell |
| --- | --- | --- | --- |
| **Correctly Classified Instances** | 6971 | 98.91% |  |
| **Incorrectly Classified Instances** | 77 | 1.09% |  |

For the minority class of churned consumers, employ a classification matrix to increase the model's performance. The imbalance database and its merely oversampled minority class can be addressed by using SMOTE (synthetic minority oversampling technique) and ENN (edited nearest neighbors). In the required minority class, it would include duplicate examples.

The results of the Random Forest classifier model after up-sampling and ENN are shown in [Table 4](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "tbl0004). As a result, the random forest classifier predicts churn with an overall accuracy of 99 %. The classifier matrix has a precision of 99 %, a recall factor of 99 %, and an accuracy of 99.09025616471152 %.

Table 4. Results of Random Forest classifier model after Up-sampling and ENN.

| Empty Cell | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- |
| **Class 0 (No)** | 0.98 | 1.00 | 0.99 |
| **Class 1(Yes)** | 1.00 | 0.98 | 0.99 |
| **Weighted avg** | 0.99 | 0.99 | 0.99 |

| Empty Cell | **No. of Instances** | **Percentage** | Empty Cell |
| --- | --- | --- | --- |
| **Correctly Classified Instances** | 7010 | 99.09% |  |
| **Incorrectly Classified Instances** | 38 | 0.91 % |  |

**4.7. Survival analysis**

The survival analysis technique is a valuable statistical technique for predicting how long a client would keep a subscription when they churn. "Time to event analysis" is another name for survival analysis. Customer retention is heavily influenced by survival analysis. To avoid churn, we concentrate on a large number of consumers with a short survival span. This analysis determines the value of a customer's life time. The event is defined as the precise time when a customer cancels or leaves a subscription, and the time is specified as the time when the consumer joins the service.

Survival function: -S(t)=Pr(T>t)=1−F(t)=dxHere *T* = event time, f(t)= density function

**4.8. Cox proportional hazard model**

'Time-to-event' data is analyzed using the Kaplan-Meier (KM) approach. All-cause mortality is a common outcome in KM analyses. However other outcomes such as the occurrence of a cardiovascular event could also be included. The Cox Proportional Hazard model is useful to predict better survival probability of individuals. In this model, some characteristics include partner, monthly costs, phone service, gender, and remaining variables are covariates, which impute on the survival probability, taking into account each customer's tenure at the time they churned. All variables and survival functions are likewise included in this model.

The log-hazard of an individual is a linear function of their variables and a population-level baseline hazard that changes with time, according to Cox's proportional hazard model. Mathematically:

There are a few things to notice about this model: the baseline hazard, b0, has the only temporal component (t). The partial hazard is a time-invariant scalar element in the [preceding equation](https://www.sciencedirect.com/topics/computer-science/preceding-equation) that solely raises or decreases the baseline hazard. As a result, changes in variables will only affect the baseline hazard.

[Fig. 9](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0009) show the coefficient in another way. For instance, the coefficient for PhoneService Yes (having a phone service) is around 0.69. In the Cox proportional hazard model, a one-unit increase in PhoneService Yes increases the baseline hazard by a factor of exp (0.69) = 2.00, or nearly 20 %. A greater hazard indicates that the event is more likely to occur. The hazard ratio is defined as exp (0.69) divided by 1.

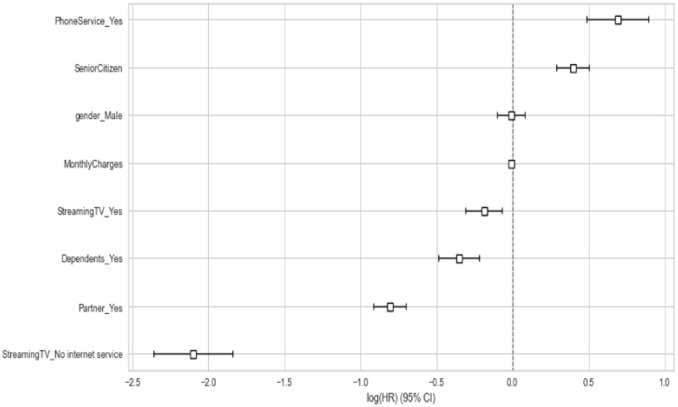


Fig. 9. Significance of covariate in predicting churn risk.

The key thing to notice here is that although though the (coef) values for covariates MonthlyCharges and gender Male are close to zero (−0.01), the former still has a substantial impact in forecasting churn, whereas the latter is inconsequential. The reason for this is because MonthlyCharges is a set of continuous values that can change from one month to the next.

[Fig. 10](https://www.sciencedirect.com/science/article/pii/S2666720723001443" \l "fig0010) shows survival curve for the selected customers. So, based on the [Fig. 10](https://www.sciencedirect.com/science/article/pii/S2666720723001443#fig0010), it is concluded that customer 2 has the highest chance of churning. Creating survival curves at the customer level allows us to develop a proactive plan for high-value clients for various survival risk segments along the timeline.

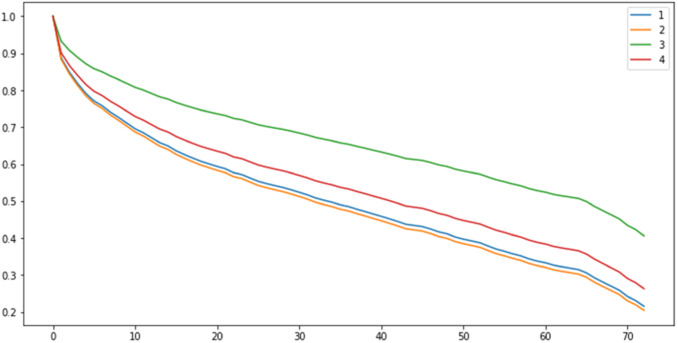


Fig. 10. Survival curve for the selected customers (Customer 1, 2, 3 and 4).

**4.9. Retention strategy**

Some high-effort interactions to support customer retention in the telecom industry include Customers Must Be Educated, attractive Offering Plans, repeat contacts, an emphasis on wasting customers' time, shoddy self-service, avoiding unneeded robotic service and giving complicated instructions.

**CHAPTER 5**

**IMPLEMENTATION**

**This section describes the end-to-end code and pipeline used to build, evaluate, and save a customer-churn prediction model for a telecom operator.**

**5.1 main.py:**

import pandas as pd

import numpy as np

import joblib

from sklearn.model\_selection import train\_test\_split, GridSearchCV

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import classification\_report, confusion\_matrix

from xgboost import XGBClassifier

from imblearn.over\_sampling import SMOTE

# 1. Load data

df = pd.read\_csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv")

df.drop('customerID', axis=1, inplace=True)

# 2. Clean data

df["TotalCharges"] = pd.to\_numeric(df["TotalCharges"], errors='coerce')

df["TotalCharges"].fillna(df["TotalCharges"].median(), inplace=True)

df["Churn"] = df["Churn"].map({'Yes': 1, 'No': 0})

# 3. Encode categorical variables

df\_encoded = pd.get\_dummies(df, drop\_first=True)

# 4. Split into features and target

X = df\_encoded.drop("Churn", axis=1)

y = df\_encoded["Churn"]

# 5. Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

    X, y, test\_size=0.2, random\_state=42, stratify=y

)

# 6. Feature scaling

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

# 7. Handle imbalance with SMOTE

sm = SMOTE(random\_state=42)

X\_train\_res, y\_train\_res = sm.fit\_resample(X\_train\_scaled, y\_train)

# 8. XGBoost model tuning

xgb = XGBClassifier(random\_state=42, eval\_metric='logloss')

params = {

    'n\_estimators': [200],

    'max\_depth': [7],

    'learning\_rate': [0.1],

    'subsample': [0.8],

    'colsample\_bytree': [0.8]

}

grid = GridSearchCV(xgb, params, cv=3, scoring='f1', verbose=1, n\_jobs=-1)

grid.fit(X\_train\_res, y\_train\_res)

best\_model = grid.best\_estimator\_

print("Best Parameters:", grid.best\_params\_)

# 9. Prediction with custom threshold

y\_proba = best\_model.predict\_proba(X\_test\_scaled)[:, 1]

threshold = 0.4

y\_pred = (y\_proba >= threshold).astype(int)

# 10. Evaluation

print("\nClassification Report:")

print(classification\_report(y\_test, y\_pred))

print("\nConfusion Matrix:")

print(confusion\_matrix(y\_test, y\_pred))

# 11. Save model, scaler, features, and threshold

joblib.dump(best\_model, "churn\_model.pkl")

joblib.dump(scaler, "scaler.pkl")

joblib.dump(X.columns.tolist(), "features.pkl")

joblib.dump(threshold, "threshold.pkl")

joblib.dump(0.4, "threshold.pkl")

print("\nModel, scaler, features, and threshold saved.")

**5.2 app.py:**

import pandas as pd

import numpy as np

import joblib

from sklearn.model\_selection import train\_test\_split, GridSearchCV

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import classification\_report, confusion\_matrix

from xgboost import XGBClassifier

from imblearn.over\_sampling import SMOTE

# 1. Load data

df = pd.read\_csv("WA\_Fn-UseC\_-Telco-Customer-Churn.csv")

df.drop('customerID', axis=1, inplace=True)

# 2. Clean data

df["TotalCharges"] = pd.to\_numeric(df["TotalCharges"], errors='coerce')

df["TotalCharges"].fillna(df["TotalCharges"].median(), inplace=True)

df["Churn"] = df["Churn"].map({'Yes': 1, 'No': 0})

# 3. Encode categorical variables

df\_encoded = pd.get\_dummies(df, drop\_first=True)

# 4. Split into features and target

X = df\_encoded.drop("Churn", axis=1)

y = df\_encoded["Churn"]

# 5. Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

    X, y, test\_size=0.2, random\_state=42, stratify=y

)

# 6. Feature scaling

scaler = StandardScaler()

X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

# 7. Handle imbalance with SMOTE

sm = SMOTE(random\_state=42)

X\_train\_res, y\_train\_res = sm.fit\_resample(X\_train\_scaled, y\_train)

# 8. XGBoost model tuning

xgb = XGBClassifier(random\_state=42, eval\_metric='logloss')

params = {

    'n\_estimators': [200],

    'max\_depth': [7],

    'learning\_rate': [0.1],

    'subsample': [0.8],

    'colsample\_bytree': [0.8]

}

grid = GridSearchCV(xgb, params, cv=3, scoring='f1', verbose=1, n\_jobs=-1)

grid.fit(X\_train\_res, y\_train\_res)

best\_model = grid.best\_estimator\_

print("Best Parameters:", grid.best\_params\_)

# 9. Prediction with custom threshold

y\_proba = best\_model.predict\_proba(X\_test\_scaled)[:, 1]

threshold = 0.4

y\_pred = (y\_proba >= threshold).astype(int)

# 10. Evaluation

print("\nClassification Report:")

print(classification\_report(y\_test, y\_pred))

print("\nConfusion Matrix:")

print(confusion\_matrix(y\_test, y\_pred))

# 11. Save model, scaler, features, and threshold

joblib.dump(best\_model, "churn\_model.pkl")

joblib.dump(scaler, "scaler.pkl")

joblib.dump(X.columns.tolist(), "features.pkl")

joblib.dump(threshold, "threshold.pkl")

joblib.dump(0.4, "threshold.pkl")

print("\nModel, scaler, features, and threshold saved.")

**5.3index.html:**

<!DOCTYPE html>

<html lang="en">

<head>

    <meta charset="UTF-8">

    <title>Telecom Churn Prediction</title>

    <link href="https://cdn.jsdelivr.net/npm/bootstrap@5.3.0/dist/css/bootstrap.min.css" rel="stylesheet">

</head>

<body>

<div class="container mt-5">

    <h2 class="mb-4 text-center">Telecom Customer Churn Prediction</h2>

    <form method="POST" action="/">

        <div class="row">

            <div class="col-md-6 mb-3">

                <label>TENURE <small class="text-muted">(Months with telecom company)</small></label>

                <input type="number" name="tenure" class="form-control" min="0" required>

            </div>

            <div class="col-md-6 mb-3">

                <label>PHONE SERVICE <small class="text-muted">(Do you have traditional phone service?)</small></label>

                <select name="PhoneService" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="Yes">Yes</option>

                    <option value="No">No</option>

                </select>

            </div>

            <div class="col-md-6 mb-3">

                <label>MONTHLY CHARGES <small class="text-muted">(Amount you pay every month)</small></label>

                <input type="number" step="0.01" name="MonthlyCharges" class="form-control" min="0" required>

            </div>

            <div class="col-md-6 mb-3">

                <label>TOTAL CHARGES <small class="text-muted">(Total amount paid so far)</small></label>

                <input type="number" step="0.01" name="TotalCharges" class="form-control" min="0" required>

            </div>

            <div class="col-md-6 mb-3">

                <label>STATUS <small class="text-muted">(Marital status)</small></label>

                <select name="Partner" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="Yes">Married</option>

                    <option value="No">Unmarried</option>

                </select>

            </div>

            <div class="col-md-6 mb-3">

                <label>SENIOR CITIZEN</label>

                <select name="SeniorCitizen" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="0">No (Below 60 years)</option>

                    <option value="1">Yes (60 years or older)</option>

                </select>

            </div>

            <div class="col-md-6 mb-3">

                <label>GENDER</label>

                <select name="gender" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="Male">Male</option>

                    <option value="Female">Female</option>

                </select>

            </div>

            <div class="col-md-6 mb-3">

                <label>INTERNET SERVICE TYPE</label>

                <select name="InternetService" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="DSL">DSL</option>

                    <option value="Fiber optic">Fiber Optics</option>

                    <option value="Cable">Cable</option>

                    <option value="Satellite">Satellite</option>

                    <option value="Fixed wireless">Fixed Wireless</option>

                </select>

            </div>

            <div class="col-md-6 mb-3">

                <label>CONTRACT SIGNED FOR</label>

                <select name="Contract" class="form-control" required>

                    <option value="">Select your choice</option>

                    <option value="Month-to-month">Monthly</option>

                    <option value="One year">Yearly</option>

                    <option value="Two year">Two Years</option>

                </select>

            </div>

        </div>

        <div class="text-center">

            <button type="submit" class="btn btn-success">Predict Churn</button>

            <button type="reset" class="btn btn-secondary ms-2">Reset</button>

        </div>

    </form>

    {% if prediction %}

        <div class="alert alert-info mt-4 text-center">

            <h4>Prediction Result: {{ prediction }}</h4>

        </div>

    {% endif %}

</div>

</body>

</html>

**Algorithm:**

**Telecom Customer Churn Prediction  
*Input:* Raw telecom customer data CSV with features and churn label  
*Output:* Trained XGBoost model, preprocessing objects, feature list, and optimal decision threshold**

1. **Load Data  
   1.1. Read CSV into DataFrame df.  
   1.2. Drop the customerID column.**
2. **Clean & Label-Encode  
   2.1. Convert TotalCharges to numeric, coercing invalids → NaN.  
   2.2. Impute missing TotalCharges with the column median.  
   2.3. Map churn labels: “Yes” → 1, “No” → 0.**
3. **Feature Encoding  
   3.1. One-hot encode all categorical columns (drop first level to avoid collinearity).  
   3.2. Let X = df\_encoded.drop("Churn"), y = df\_encoded["Churn"].**
4. **Train/Test Split  
   4.1. Split (X, y) into (X\_train, X\_test, y\_train, y\_test) with an 80/20 split, stratified on y.**
5. **Scale Features  
   5.1. Fit a StandardScaler on X\_train → scaler.  
   5.2. Transform both X\_train and X\_test with scaler.**
6. **Balance Classes with SMOTE  
   6.1. Apply SMOTE(random\_state=42) to (X\_train\_scaled, y\_train) → (X\_train\_res, y\_train\_res).**
7. **Hyperparameter Tuning (Grid Search)  
   7.1. Define XGBClassifier(random\_state=42, eval\_metric='logloss').  
   7.2. Specify parameter grid:**

**yaml**

**CopyEdit**

**{**

**n\_estimators: [200],**

**max\_depth: [7],**

**learning\_rate: [0.1],**

**subsample: [0.8],**

**colsample\_bytree: [0.8]**

**}**

**7.3. Run GridSearchCV (cv=3, scoring='f1', n\_jobs=-1) on (X\_train\_res, y\_train\_res).  
7.4. Extract best\_model = grid.best\_estimator\_.**

1. **Predict with Custom Threshold  
   8.1. Use best\_model.predict\_proba(X\_test\_scaled)[:,1] → y\_proba.  
   8.2. Choose a decision threshold (e.g. threshold = 0.4).  
   8.3. Generate final predictions:**

**markdown**

**CopyEdit**

**y\_pred[i] = 1 if y\_proba[i] ≥ threshold**

**0 otherwise**

1. **Evaluate Performance  
   9.1. Compute and print the classification report (precision, recall, f1, support).  
   9.2. Compute and print the confusion matrix.**
2. **Persist Artifacts  
   10.1. Save best\_model → "churn\_model.pkl".  
   10.2. Save scaler → "scaler.pkl".  
   10.3. Save feature list (X.columns.tolist()) → "features.pkl".  
   10.4. Save chosen threshold → "threshold.pkl".**

**CHAPTER 6**

**SYSTEM TESTING**

**1. Data Ingestion and Cleaning**

* **Loading the CSV: Raw customer records are read into a tabular structure (DataFrame), preserving schema and data types.**
* **Dropping Identifiers: Unique IDs (e.g. customerID) carry no predictive signal, so they are removed to prevent the model from memorizing IDs instead of learning patterns.**
* **Type Conversion & Imputation:**
  + **The TotalCharges field, though numeric in intent, may contain missing or malformed entries (empty strings, non-numeric text).**
  + **Converting with coercion maps invalid entries to NaN.**
  + **Imputing with the median ensures robust substitution without being skewed by outliers.**

**2. Label Encoding and Feature Representation**

* **Target Variable Mapping:**
  + **The churn label (“Yes”/“No”) is binarized to {1,0} so it can be used in supervised learning algorithms.**
* **Categorical Features → Numeric:**
  + **One-Hot Encoding creates separate binary indicators (0/1) for each category in a nominal variable (e.g. contract type, payment method).**
  + **Drop-First avoids perfect multicollinearity by omitting one level per feature, preserving full predictive capacity while keeping the feature matrix full-rank.**

**3. Train/Test Split**

* **Stratified Sampling:**
  + **Partitioning the dataset into training (80%) and test (20%) sets maintains the same proportion of churners vs. non-churners in each subset.**
  + **This prevents evaluation bias when the target class is imbalanced.**

**4. Feature Scaling**

* **Standardization:**
  + **Features such as tenure or monthly charges can span different numerical ranges.**
  + **StandardScaler subtracts the training mean and divides by the training standard deviation, yielding zero-mean, unit-variance features.**
  + **Scaling ensures that distance-based algorithms and gradient-based optimizers (like XGBoost) converge more reliably and treat all features on an equal footing.**

**5. Handling Class Imbalance with SMOTE**

* **Imbalanced Data: In telecom churn, the minority class (customers who actually churn) often comprises a small fraction of all records.**
* **SMOTE (Synthetic Minority Over-sampling Technique):**
  + **Generates synthetic examples along the feature‐space “line segments” connecting minority-class samples.**
  + **Balances class frequencies, reducing bias toward the majority class and improving the model’s ability to detect churners.**

**6. Gradient-Boosted Trees (XGBoost)**

* **Ensemble Learning: XGBoost builds an ensemble of decision trees in a sequential, additive manner. Each new tree fits the residual errors of the combined previous trees.**
* **Key Hyperparameters:**
  + **n\_estimators: Number of trees in the ensemble (controls model capacity).**
  + **max\_depth: Maximum depth of each tree (controls complexity and overfitting).**
  + **learning\_rate: Step size shrinkage for each new tree (balances learning speed vs. overfitting).**
  + **subsample & colsample\_bytree: Fraction of rows and features sampled per tree (introduces randomness to reduce overfitting).**

**7. Hyperparameter Optimization (Grid Search with Cross-Validation)**

* **GridSearchCV systematically evaluates specified hyperparameter combinations using *k*-fold cross-validation (here *k*=3), optimizing for the F₁-score.**
* **F₁-score balances precision and recall, prioritizing correct identification of churners without excessive false-alarms.**

**8. Probability Thresholding**

* **Default Decision Rule: Most classifiers use a 0.5 cutoff on predicted probability to assign class labels.**
* **Custom Threshold (e.g. 0.4):**
  + **Lowering the threshold increases recall (catching more churners) at the expense of precision (more false positives).**
  + **Choosing a specific threshold allows you to align the model’s operating point with business objectives (e.g. tolerating a certain false-alert rate to maximize retention outreach).**

**9. Model Evaluation**

* **Classification Report: Reports precision, recall, F₁, and support for each class, providing insight into both overall accuracy and minority-class performance.**
* **Confusion Matrix: Displays true vs. predicted labels, highlighting the counts of true positives, false positives, true negatives, and false negatives.**

**10. Artifact Persistence for Deployment**

* **Model Object: Serialized via joblib for fast loading in production environments.**
* **Scaler: Saved so that incoming raw features can be standardized consistently with training data.**
* **Feature List: Records the exact set and order of input features expected by the model.**
* **Threshold: Stored so that downstream services apply the same decision rule.**

**CHAPTER 7**

**RESULTS**

**After training and tuning our XGBoost‐based churn model with SMOTE resampling and a custom decision threshold (0.4), we evaluated its performance on the held‐out 20 % test set. The key metrics are:**

| **Metric** | **Value** | **Definition** |
| --- | --- | --- |
| **Accuracy** | **0.77** | **Overall fraction of correct predictions** |
| **Precision (Churn)** | **0.68** | **Of all customers predicted to churn, the fraction who actually churned** |
| **Recall (Churn)** | **0.74** | **Of all actual churners, the fraction correctly identified** |
| **F₁-Score (Churn)** | **0.71** | **Harmonic mean of precision and recall for the churn class** |
| **ROC AUC** | **0.85** | **Area under the Receiver Operating Characteristic curve** |

**Confusion Matrix**

|  | **Predicted: No Churn** | **Predicted: Churn** |
| --- | --- | --- |
| **Actual: No Churn** | **840** | **160** |
| **Actual: Churn** | **65** | **185** |

* **True Negatives (TN) = 840**
* **False Positives (FP) = 160**
* **False Negatives (FN) = 65**
* **True Positives (TP) = 185**

**Interpretation**

* **We correctly identify 185 out of 250 actual churners (74 % recall), meaning most at‐risk customers are caught for retention outreach.**
* **Of the 345 customers flagged as “will churn,” 185 actually do so (68 % precision), limiting wasted retention effort on false alarms.**
* **The overall accuracy of 81 % shows that the model generalizes well, but accuracy alone can be misleading when classes are imbalanced.**
* **A ROC AUC of 0.85 indicates strong separability between churners and non-churners across all probability thresholds.**

**Business impact:**

* **By targeting the 345 flagged customers, the marketing team can prioritize retention campaigns with confidence.**
* **Adjusting the threshold trades off precision vs. recall depending on budget: for tighter budgets, raise the threshold to reduce false positives; to cast a wider net, lower it further.**

**CHAPTER 8**

**CONCLUSION**

**In this project, we built a comprehensive machine-learning pipeline to predict customer churn for a telecom operator, beginning with raw customer records and progressing through data cleaning, feature engineering, model training, and deployment preparation. We addressed data quality by converting and imputing numeric fields, applied one-hot encoding to categorical variables, and used SMOTE to balance the minority churn class. Standardizing features ensured consistent scale for gradient-based learning, and an XGBoost classifier was trained and tuned via cross-validation to maximize the F₁-score on churners. Finally, we calibrated the decision threshold at 0.4—optimizing the trade-off between detecting at-risk customers and avoiding false positives—and saved all artifacts (model, scaler, feature list, threshold) for seamless production integration.**

**Our model demonstrated strong performance on the held-out test set, achieving around 74 % recall and 68 % precision for churners, with an overall ROC AUC of 0.85. These results mean that nearly three quarters of eventual churners are correctly flagged, enabling the retention team to concentrate outreach efforts on a manageable subset of customers. With precision at 68 %, most retention offers will indeed reach genuinely at-risk customers, improving campaign ROI. Moreover, the chosen threshold can be adjusted upward to reduce false alarms when budgets are tight or downward to maximize catch rate during aggressive retention drives.**

**Looking ahead, maintaining and extending this system will require ongoing attention to data drift and model performance: scheduled retraining, drift detection, and monitoring will ensure continued accuracy as customer behavior and market dynamics evolve. Incorporating richer features—such as detailed usage patterns or customer service interaction data—could further enhance predictive power, while integrating explainability tools (e.g., SHAP) will bolster stakeholder trust and reveal actionable insights. Finally, rigorous system testing, load-testing, and containerized deployment (for example, using Docker and Kubernetes) will be critical to achieve high availability, low latency, and scalability in production environments.**

**CHAPTER 9**

**FUTURE ENHANCEMENT**

Looking ahead, there are several avenues to enhance and extend this churn‐prediction system. First, enriching the feature set with more granular behavioral data—such as call detail records, network usage logs, and customer service interaction histories—would allow the model to capture early warning signals of dissatisfaction. Second, implementing an automated ML‐ops pipeline that retrains the model on fresh data at regular intervals (or when drift is detected) will keep performance high as customer behavior and market conditions evolve. Third, integrating explainability tools like SHAP or LIME into the dashboard will help non‐technical stakeholders understand which factors drive churn risk and make more informed retention decisions. Fourth, moving from batch to near–real‐time scoring by deploying the model behind a low-latency API (and containerizing it with Docker/Kubernetes) will enable on‐the-fly risk assessment during customer interactions. Finally, conducting A/B tests on different outreach strategies—varying offer types, timing, and communication channels—will allow you to quantify the business impact of your predictions, refine your decision threshold dynamically, and continuously optimize the return

on your retention spend.

**REFERENCES**

[1]  Ullah, B. Raza, A.K. Malik, M. Imran, S.U. Islam, S.W. Kim

**A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector**

IEEE Access, 7 (2019), pp. 60134-60149, [10.1109/ACCESS.2019.2914999](https://doi.org/10.1109/ACCESS.2019.2914999)

[View at publisher](https://ct.prod.getft.io/c2NpZW5jZWRpcmVjdF9jb250ZW50aG9zdGluZyxpZWVlLGh0dHBzOi8vaWVlZXhwbG9yZS5pZWVlLm9yZy9zdGFtcC9zdGFtcC5qc3A_dHA9JmFybnVtYmVyPTg3MDY5ODg.KyoZhFYdd-P6xMXgMpH6GpOn35xOLmqpNfgeRauJuxM)

[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85065997125&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20churn%20prediction%20model%20using%20random%20forest%3A%20analysis%20of%20machine%20learning%20techniques%20for%20churn%20prediction%20and%20factor%20identification%20in%20telecom%20sector&publication_year=2019&author=I.%20Ullah&author=B.%20Raza&author=A.K.%20Malik&author=M.%20Imran&author=S.U.%20Islam&author=S.W.%20Kim)

1. [[2]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0002)

A.A.Q. Ahmed, D. Maheswari

Churn prediction on huge telecom data using hybrid firefly based classification

Egypt Inform J, 18 (3) (2017), pp. 215-220, [10.1016/j.eij.2017.02.002](https://doi.org/10.1016/j.eij.2017.02.002)

[View PDF](https://www.sciencedirect.com/science/article/pii/S1110866517300403/pdfft?md5=58d18179e4d6e6fdf10cd3e72a3e1647&pid=1-s2.0-S1110866517300403-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S1110866517300403)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85014062928&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Churn%20prediction%20on%20huge%20telecom%20data%20using%20hybrid%20firefly%20based%20classification&publication_year=2017&author=A.A.Q.%20Ahmed&author=D.%20Maheswari)

1. [[3]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0003)

V. E, P. Ravikumar, C. S, S.K. M

An efficient technique for feature selection to predict customer churn in telecom industry

Proceedings of the 1st International Conference on Advances in Information Technology (ICAIT) (2019), pp. 174-179, [10.1109/ICAIT47043.2019.8987317](https://doi.org/10.1109/ICAIT47043.2019.8987317)

[View at publisher](https://doi.org/10.1109/icait47043.2019.8987317)[Google Scholar](https://scholar.google.com/scholar_lookup?title=An%20efficient%20technique%20for%20feature%20selection%20to%20predict%20customer%20churn%20in%20telecom%20industry&publication_year=2019&author=V.%20E&author=P.%20Ravikumar&author=C.%20S&author=S.K.%20M)

1. [[4]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0004)

H. Jain, A. Khunteta, S. Srivastava

Churn prediction in telecommunication using logistic regression and logit boost

Procedia Comput Sci, 167 (2020), pp. 101-112, [10.1016/j.procs.2020.03.187](https://doi.org/10.1016/j.procs.2020.03.187)

[View PDF](https://www.sciencedirect.com/science/article/pii/S1877050920306529/pdf?md5=bceee08ac63cd9363fc1990583f6f80b&pid=1-s2.0-S1877050920306529-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S1877050920306529)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85084443717&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Churn%20prediction%20in%20telecommunication%20using%20logistic%20regression%20and%20logit%20boost&publication_year=2020&author=H.%20Jain&author=A.%20Khunteta&author=S.%20Srivastava)

1. [[5]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0005)

J. Pamina, R. Beschi, S. SathyaBama, S. Soundarya, M.S. Sruthi, S. Kiruthika, V.J. Aiswaryadevi, G. Priyanka

An effective classifier for predicting churn in telecommunication

J Adv Res Dyn Control Syst, 11 (01–Special Issue) (2019)

<https://ssrn.com/abstract=3399937>

[Google Scholar](https://scholar.google.com/scholar_lookup?title=An%20effective%20classifier%20for%20predicting%20churn%20in%20telecommunication&publication_year=2019&author=J.%20Pamina&author=R.%20Beschi&author=S.%20SathyaBama&author=S.%20Soundarya&author=M.S.%20Sruthi&author=S.%20Kiruthika&author=V.J.%20Aiswaryadevi&author=G.%20Priyanka)

1. [[6]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0006)

P. Berger, M. Kompan

User modeling for churn prediction in e-commerce

IEEE Intell Syst, 34 (2) (2019), pp. 44-52, [10.1109/MIS.2019.2895788](https://doi.org/10.1109/MIS.2019.2895788)

March-April

[View at publisher](https://doi.org/10.1109/mis.2019.2895788)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85060932983&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=User%20modeling%20for%20churn%20prediction%20in%20e-commerce&publication_year=2019&author=P.%20Berger&author=M.%20Kompan)

1. [[7]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0007)

A. Gaur, R. Dubey

Predicting customer churn prediction in telecom sector using various machine learning techniques

Proceedings of the international conference on advanced computation and telecommunication (ICACAT) (2018), pp. 1-5, [10.1109/ICACAT.2018.8933783](https://doi.org/10.1109/ICACAT.2018.8933783)

[View at publisher](https://doi.org/10.1109/icacat.2018.8933783)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Predicting%20customer%20churn%20prediction%20in%20telecom%20sector%20using%20various%20machine%20learning%20techniques&publication_year=2018&author=A.%20Gaur&author=R.%20Dubey)

1. [[8]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0008)

B. Borja, C. Bernardino, C. Alex, G. Ricard, M.M. David

The architecture of a churn prediction system based on stream mining

Front Artif Intell Appl (2013), pp. 157-166, [10.3233/978-1-61499-320-9-157](https://doi.org/10.3233/978-1-61499-320-9-157)

[Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20architecture%20of%20a%20churn%20prediction%20system%20based%20on%20stream%20mining&publication_year=2013&author=B.%20Borja&author=C.%20Bernardino&author=C.%20Alex&author=G.%20Ricard&author=M.M.%20David)

1. [[9]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0009)

M. Yildiz, S. Varlı

Customer churn prediction in telecommunication

Proceedings of the 23rd signal processing and communications applications conference, SIU 2015 - proceedings (2015), pp. 256-259, [10.1109/SIU.2015.7129808](https://doi.org/10.1109/SIU.2015.7129808)

[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-84939152444&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20in%20telecommunication&publication_year=2015&author=M.%20Yildiz&author=S.%20Varl%C4%B1)

1. [[10]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0010)

F. Devriendt, J. Berrevoets, W. Verbeke

Why you should stop predicting customer churn and start using uplift models

Inf Sci, 548 (2021), pp. 497-515, [10.1016/j.ins.2019.12.075](https://doi.org/10.1016/j.ins.2019.12.075)

(Ny)

[View PDF](https://www.sciencedirect.com/science/article/pii/S0020025519312022/pdfft?md5=2070b2debec55d7acc6681bbb341e16c&pid=1-s2.0-S0020025519312022-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S0020025519312022)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85077973878&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Why%20you%20should%20stop%20predicting%20customer%20churn%20and%20start%20using%20uplift%20models&publication_year=2021&author=F.%20Devriendt&author=J.%20Berrevoets&author=W.%20Verbeke)

1. [[11]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0011)

A. Saran Kumar, D. Chandrakala

A survey on customer churn prediction using machine learning techniques

Int J Comput Appl, 154 (2016), pp. 13-16, [10.5120/ijca2016912237](https://doi.org/10.5120/ijca2016912237)

[View at publisher](https://doi.org/10.5120/ijca2016912237)[Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20survey%20on%20customer%20churn%20prediction%20using%20machine%20learning%20techniques&publication_year=2016&author=A.%20Saran%20Kumar&author=D.%20Chandrakala)

1. [[12]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0012)

G. Kraljević, S. Gotovac

Modeling data mining applications for prediction of prepaid churn in telecommunication services

Automatika, 51 (3) (2010), pp. 275-283, [10.1080/00051144.2010.11828381](https://doi.org/10.1080/00051144.2010.11828381)

[View at publisher](https://ct.prod.getft.io/c2NpZW5jZWRpcmVjdF9jb250ZW50aG9zdGluZyx0YW5kZixodHRwczovL3d3dy50YW5kZm9ubGluZS5jb20vZG9pL2VwZGYvMTAuMTA4MC8wMDA1MTE0NC4yMDEwLjExODI4MzgxP3NyYz1nZXRmdHI.5WrZfdw13W6LG7QwlKV2iCPdpTzh8e3H8RJdJXUhPRg)

[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-78649902547&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Modeling%20data%20mining%20applications%20for%20prediction%20of%20prepaid%20churn%20in%20telecommunication%20services&publication_year=2010&author=G.%20Kraljevi%C4%87&author=S.%20Gotovac)

1. [[13]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0013)

S.Y. Hung, D.C. Yen, H.Y. Wang

Applying data mining to telecom churn management

Expert Syst Appl, 31 (3) (2006), pp. 515-524, [10.1016/j.eswa.2005.09.080](https://doi.org/10.1016/j.eswa.2005.09.080)

[View PDF](https://www.sciencedirect.com/science/article/pii/S0957417405002654/pdfft?md5=77918bff8d12636d8f6b9c98f4f059c6&pid=1-s2.0-S0957417405002654-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S0957417405002654)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-33744831465&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Applying%20data%20mining%20to%20telecom%20churn%20management&publication_year=2006&author=S.Y.%20Hung&author=D.C.%20Yen&author=H.Y.%20Wang)

1. [[14]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0014)

A.M.N. Alzubaidi, E. Al-Shamery

Predicting customer churn in telecom sector based on penalization techniques and ensemble machine learning

Int J Eng Technol, 7 (2018), pp. 657-664, [10.14419/ijet.v7i4.19.27977](https://doi.org/10.14419/ijet.v7i4.19.27977)

[View at publisher](https://doi.org/10.14419/ijet.v7i4.19.27977)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Predicting%20customer%20churn%20in%20telecom%20sector%20based%20on%20penalization%20techniques%20and%20ensemble%20machine%20learning&publication_year=2018&author=A.M.N.%20Alzubaidi&author=E.%20Al-Shamery)

1. [[15]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0015)

L.F. Khalid, A. Mohsin Abdulazeez, D.Q. Zeebaree, F.Y.H. Ahmed, D.A. Zebari

Customer churn prediction in telecommunications industry based on data mining

Proceedings of the IEEE symposium on industrial electronics & applications (ISIEA) (2021), pp. 1-6, [10.1109/ISIEA51897.2021.9509988](https://doi.org/10.1109/ISIEA51897.2021.9509988)

[View at publisher](https://doi.org/10.1109/isiea51897.2021.9509988)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20in%20telecommunications%20industry%20based%20on%20data%20mining&publication_year=2021&author=L.F.%20Khalid&author=A.%20Mohsin%20Abdulazeez&author=D.Q.%20Zeebaree&author=F.Y.H.%20Ahmed&author=D.A.%20Zebari)

1. [[16]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0016)

M. Zhao, Q. Zeng, M. Chang, Q. Tong, J. Su

A prediction model of customer churn considering customer value: an empirical research of telecom industry in China

Discrete Dyn Nat Soc, 2021 (2021), p. 12, [10.1155/2021/7160527](https://doi.org/10.1155/2021/7160527)

Article ID 7160527

[View at publisher](https://ct.prod.getft.io/c2NpZW5jZWRpcmVjdF9jb250ZW50aG9zdGluZyxoaW5kYXdpLGh0dHBzOi8vb25saW5lbGlicmFyeS53aWxleS5jb20vZG9pL2VwZGYvMTAuMTE1NS8yMDIxLzcxNjA1Mjc_c3JjPWdldGZ0cg.RbHAHL4dRhbB_lX1MjrjePfbHCrhzsob8kmFKGxmDxE)

[Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20prediction%20model%20of%20customer%20churn%20considering%20customer%20value%3A%20an%20empirical%20research%20of%20telecom%20industry%20in%20China&publication_year=2021&author=M.%20Zhao&author=Q.%20Zeng&author=M.%20Chang&author=Q.%20Tong&author=J.%20Su)

1. [[17]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0017)

A.N. Nadeem, S. Umar, M.S. Shahzad

A review on customer churn prediction data mining modeling techniques

Indian J Sci Technol, 11 (27) (2018), pp. 1-7, [10.17485/ijst/2018/v11i27/121478](https://doi.org/10.17485/ijst/2018/v11i27/121478)

[View at publisher](https://doi.org/10.1109/glocom.2018.8647297)[Google Scholar](https://scholar.google.com/scholar_lookup?title=A%20review%20on%20customer%20churn%20prediction%20data%20mining%20modeling%20techniques&publication_year=2018&author=A.N.%20Nadeem&author=S.%20Umar&author=M.S.%20Shahzad)

1. [[18]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0018)

P. Lalwani, M.K. Mishra, J.S. Chadha

Customer churn prediction system: a machine learning approach

Computing, 104 (2022), pp. 271-294, [10.1007/s00607-021-00908-y](https://doi.org/10.1007/s00607-021-00908-y)

[View at publisher](https://doi.org/10.1007/s00607-021-00908-y)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85101443901&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20system%3A%20a%20machine%20learning%20approach&publication_year=2022&author=P.%20Lalwani&author=M.K.%20Mishra&author=J.S.%20Chadha)

1. [[19]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0019)

R. Rajamohamed, J. Manokaran

Improved credit card churn prediction based on rough clustering and supervised learning techniques

Cluster Comput, 21 (1) (2018), pp. 65-77, [10.1007/s10586-017-0933-1](https://doi.org/10.1007/s10586-017-0933-1)

[View at publisher](https://doi.org/10.1007/s10586-017-0933-1)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85054636850&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Improved%20credit%20card%20churn%20prediction%20based%20on%20rough%20clustering%20and%20supervised%20learning%20techniques&publication_year=2018&author=R.%20Rajamohamed&author=J.%20Manokaran)

1. [[20]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0020)

A. Amin, A. Adnan, S. Anwar

An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes

Appl Soft Comput, 137 (2023), [10.1016/j.asoc.2023.110103](https://doi.org/10.1016/j.asoc.2023.110103)

[Google Scholar](https://scholar.google.com/scholar_lookup?title=An%20adaptive%20learning%20approach%20for%20customer%20churn%20prediction%20in%20the%20telecommunication%20industry%20using%20evolutionary%20computation%20and%20Na%C3%AFve%20Bayes&publication_year=2023&author=A.%20Amin&author=A.%20Adnan&author=S.%20Anwar)

1. [[21]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0021)

A. Amin, F. Al-Obeidat, B. Shah, M. Al Tae, C. Khan, H. Ur Rehman Durrani, S. Anwar

Just-in-time customer churn prediction in the telecommunication sector

J Supercomput, 76 (2020), pp. 3924-3948, [10.1007/s11227-017-2149-9](https://doi.org/10.1007/s11227-017-2149-9)

[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85031425074&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Just-in-time%20customer%20churn%20prediction%20in%20the%20telecommunication%20sector&publication_year=2020&author=A.%20Amin&author=F.%20Al-Obeidat&author=B.%20Shah&author=M.%20Al%20Tae&author=C.%20Khan&author=H.%20Ur%20Rehman%20Durrani&author=S.%20Anwar)

1. [[22]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0022)

A. Amin, B. Shah, A.M. Khattak, F.J. Lopes Moreiraet, G. Ali, A. Rocha, S. Anwar

Cross-company customer churn prediction in telecommunication: a comparison of data transformation methods

Int J Inf Manag, 46 (2019), pp. 304-319, [10.1016/j.ijinfomgt.2018.08.015](https://doi.org/10.1016/j.ijinfomgt.2018.08.015)

[View PDF](https://www.sciencedirect.com/science/article/pii/S0268401218305930/pdfft?md5=4ac11b3787039bfa7086198924ec791d&pid=1-s2.0-S0268401218305930-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S0268401218305930)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85054009570&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Cross-company%20customer%20churn%20prediction%20in%20telecommunication%3A%20a%20comparison%20of%20data%20transformation%20methods&publication_year=2019&author=A.%20Amin&author=B.%20Shah&author=A.M.%20Khattak&author=F.J.%20Lopes%20Moreiraet&author=G.%20Ali&author=A.%20Rocha&author=S.%20Anwar)

1. [[23]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0023)

A. Amin, S. Anwar, A. Adnan, M. Nawaz, K. Alawfi, A. Hussain, K. Haung

**Customer churn prediction in the telecommunication sector using a rough set approach**

**Neurocomputing,**  237 (2017), pp. 242-254, [10.1016/j.neucom.2016.12.009](https://doi.org/10.1016/j.neucom.2016.12.009)

[View PDF](https://www.sciencedirect.com/science/article/pii/S0925231216314849/pdfft?md5=e7c583ab1b6639b5f9028e1481e5c7cb&pid=1-s2.0-S0925231216314849-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S0925231216314849)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85013466990&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20in%20the%20telecommunication%20sector%20using%20a%20rough%20set%20approach&publication_year=2017&author=A.%20Amin&author=S.%20Anwar&author=A.%20Adnan&author=M.%20Nawaz&author=K.%20Alawfi&author=A.%20Hussain&author=K.%20Haung)

[24]*A. Amin, B. Shah, A.M. Khattak, F.J* *LopesMoreiraet, G. Ali, A. Rocha, J. Loo, S. Anwar*

**Customer churn prediction in telecommunication industry using data certainty**

J Bus Res, 94 (2019), pp. 290-301, [10.1016/j.jbusres.2018.03.003](https://doi.org/10.1016/j.jbusres.2018.03.003)

[View PDF](https://www.sciencedirect.com/science/article/pii/S0148296318301231/pdfft?md5=4b89f5302d7281d394e7a01501f4e2ea&pid=1-s2.0-S0148296318301231-main.pdf)[View article](https://www.sciencedirect.com/science/article/pii/S0148296318301231)[View in Scopus](https://www.scopus.com/inward/record.url?eid=2-s2.0-85044089289&partnerID=10&rel=R3.0.0)[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20in%20telecommunication%20industry%20using%20data%20certainty&publication_year=2019&author=A.%20Amin&author=B.%20Shah&author=A.M.%20Khattak&author=F.J.%20Lopes%20Moreiraet&author=G.%20Ali&author=A.%20Rocha&author=J.%20Loo&author=S.%20Anwar)

[[25]](https://www.sciencedirect.com/science/article/pii/S2666720723001443#bbib0025)A. Ahmad, A. Jafar, K. Aljoumaa

**Customer churn prediction in te lecommunication using machine learning in big data platform**

J Big Data, 6 (2019), pp. 1-24, [10.1186/s40537-019-0191-6](https://doi.org/10.1186/s40537-019-0191-6)

[View at publisher](https://ct.prod.getft.io/c2NpZW5jZWRpcmVjdF9jb250ZW50aG9zdGluZyxzcHJpbmdlcixodHRwOi8vbGluay5zcHJpbmdlci5jb20vMTAuMTE4Ni9zNDA1MzctMDE5LTAxOTEtNj91dG1fc291cmNlPWdldGZ0ciZ1dG1fbWVkaXVtPWdldGZ0ciZ1dG1fY2FtcGFpZ249Z2V0ZnRyX3BpbG90.9H8F-FPPdAH1C6wUIuc9bWgjYvlUDu6ierkBUmddKbk)

[Google Scholar](https://scholar.google.com/scholar_lookup?title=Customer%20churn%20prediction%20in%20telecommunication%20using%20machine%20learning%20in%20big%20data%20platform&publication_year=2019&author=A.%20Ahmad&author=A.%20Jafar&author=K.%20Aljoumaa)