International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Customer Churn Prediction in the Telecom Industry Using Machine Learning

Pappala Shanmukha Rao
Reg. No. 24Q71F0042
pappalashanmukh@gmail.com
Department of Master of Computer Applications
Avanthi Institute of Engineering and Technology (Autonomous)
Vizianagaram, Andhra Pradesh, India
Under the guidance of Mrs. T. Varalakshmi, MCA, Assistant Professor
laxmivara588@gmail.com

Abstract*—Customer churn means customers leaving a company or stopping the use of its services. This project focuses on identifying such customers at an early stage so that businesses can take necessary actions. It is very useful in industries like telecom, banking, and online platforms where customer retention is important. In this project, machine-learning techniques are used to analyse customer data such as usage details, billing information, and customer behaviour. First, the data is cleaned by handling missing values and converting it into a proper format; then, important features that influence customer decisions are selected. After preprocessing, the dataset is divided into training and testing data, and machine-learning algorithms such as Logistic Regression and Decision Tree are applied to build the prediction model. The model is trained using past customer data and learns patterns from it; once training is complete, it predicts whether a customer is likely to leave or continue using the service. The performance of the model is evaluated using accuracy and other metrics. This project helps companies understand customer behaviour and identify the main reasons for churn, so that businesses can take preventive steps such as providing better services, offering discounts, or improving customer support. The system was validated through ten functional and validation test cases that all passed. Overall, this system helps increase customer satisfaction, reduce customer loss, and improve business profit.

Keywords*—Customer Churn; Telecom Industry; Machine Learning; Logistic Regression; Decision Tree; Customer Retention; Predictive Analytics; Classification.

I. INTRODUCTION

The rapid growth of digital technologies and the widespread use of mobile communication services have significantly transformed the telecom industry. Telecom companies provide a wide range of services such as voice calls, internet data, messaging, and value-added services to millions of customers worldwide. However, with increasing competition and the availability of multiple service providers, customer retention has become a major challenge: customers can easily switch from one provider to another, leading to a phenomenon known as customer churn. High churn rates directly impact the revenue and profitability of telecom companies, so understanding customer behaviour and predicting churn has become a critical task.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Traditional methods of analysing customer churn rely on manual analysis and basic statistical techniques, which are time-consuming and less effective in handling large volumes of data. Telecom companies collect vast amounts of customer data—call records, billing information, internet usage, customer complaints, service plans, and demographic details—that contain valuable insights into customer behaviour, but without advanced analytical tools it is difficult to extract meaningful information and identify customers likely to churn.

Machine Learning has emerged as a powerful solution for predicting customer churn. By analysing historical data, machine-learning models can identify patterns and relationships between customer attributes and churn behaviour, classifying customers as "Churn" or "Non-Churn" so companies can take proactive actions. Algorithms such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting are widely used, analysing features such as customer tenure, monthly charges, contract type, service usage, and customer-support interactions. This project develops a machine-learning-based system that analyses customer data, identifies patterns, and predicts churn behaviour so that telecom companies can apply effective retention strategies and improve customer satisfaction.

II. LITERATURE SURVEY

Customer-churn prediction has been widely studied in the telecom domain because of its direct impact on revenue and profitability. Early approaches relied on manual reporting, basic statistical methods, and rule-based segmentation that analysed usage patterns and billing history to identify potential churn; while simple and interpretable, these methods are limited in handling large-scale, complex, and dynamic datasets and cannot capture hidden patterns in customer behaviour, leading to low prediction accuracy and limited suitability for real-time decision-making.

With the growth of machine learning, data-driven churn prediction has become the dominant approach. Supervised classification algorithms such as Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting learn patterns from historical customer data and classify customers into churn or non-churn categories, and studies consistently show that ensemble and tree-based methods improve accuracy over basic statistical models. The literature also emphasises the importance of data preprocessing—cleaning, encoding, normalisation, and feature selection—to improve model accuracy, as well as scalability for the millions of customers generating data continuously, and data security and privacy when handling sensitive customer information. These findings motivate the proposed machine-learning-based churn-prediction system.

TABLE I. REPRESENTATIVE TECHNIQUES FOR CHURN PREDICTION

S.No	Approach	Technique	Note
1	Traditional analysis	Manual reporting, statistics	Low accuracy; not real-time
2	Rule-based segmentation	Predefined rules	Misses hidden patterns

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

S.No	Approach	Technique	Note
3	Logistic Regression	Linear classification	Simple, interpretable baseline
4	Decision Tree	Rule-based ML	Interpretable, non-linear
5	Random Forest / Gradient Boosting	Ensemble ML	Higher accuracy
6	Support Vector Machine	Margin-based ML	Effective on complex data

III. EXISTING SYSTEM AND PROPOSED SYSTEM

A. Existing System

Traditional customer-churn analysis systems in the telecom industry rely on manual reporting, basic statistical methods, and rule-based approaches that analyse customer data such as usage patterns and billing history to identify potential churn. However, they are limited in handling large-scale and complex datasets, and most use simple models or basic segmentation techniques that cannot capture hidden patterns in customer behaviour, so prediction accuracy is low and not suitable for real-time decision-making. Because real-world telecom data is highly dynamic and continuously changing, traditional systems fail to update predictions in real time, making them ineffective for proactive retention.

Limitations of the existing system:

- Manual analysis required; depends on human intervention.
- Low prediction accuracy; basic models miss complex patterns.
- No real-time processing; cannot predict churn instantly.
- Poor scalability for large, dynamic telecom datasets.
- No actionable insights for timely customer retention.

B. Proposed System

The proposed system introduces a Machine-Learning-based Customer Churn Prediction System for the telecom industry. Customer data is collected, preprocessed (cleaning, encoding, normalisation), and used to train machine-learning models that learn patterns from historical data and classify customers as churn or non-churn. Several algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Gradient Boosting—are applied, and the best-performing model is selected based on evaluation metrics. The system provides accurate prediction and actionable insights so companies can take timely retention measures.

Advantages of the proposed system:

- High accuracy through machine-learning models.
- Real-time prediction for instant churn detection.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

- Scalable system that handles large telecom datasets efficiently.
- Automated analysis that reduces manual effort.
- Actionable insights for effective retention strategies.
- Improved customer satisfaction and business profitability.

IV. SYSTEM ANALYSIS AND DESIGN

A. Working of the Proposed System

The system operates as a pipeline: collect telecom customer data from multiple sources; preprocess the data (cleaning, encoding, normalisation); extract relevant features from customer records; train machine-learning models on historical data; classify customers as churn or non-churn; evaluate and select the best model; and present predictions and actionable insights. This enables proactive retention by identifying high-risk customers early.

B. Requirements and Feasibility

Functionally, the system must load a customer dataset, preprocess and validate it, extract features, train and evaluate models, predict churn for new customer data, display results, and generate/export reports. Non-functional considerations include accuracy, scalability to large datasets, real-time responsiveness, usability, and data security and privacy. The system is economically feasible because many machine-learning tools such as Python, scikit-learn, and pandas are open-source, significantly reducing development cost, with infrastructure cost mainly for data storage and model training.

C. System Architecture

The architecture is modular: a data layer storing telecom customer records; a preprocessing component performing cleaning, encoding, and normalisation; a feature-extraction component selecting attributes that influence churn (tenure, monthly charges, contract type, service usage, support interactions); a modelling component hosting the classification algorithms; an evaluation component computing metrics and selecting the best model; and a presentation component where users load data, run predictions, view results, and export reports. UML diagrams describe the interactions among these components.

V. SYSTEM IMPLEMENTATION

A. Technology Stack

TABLE II. TECHNOLOGY STACK

Component	Technology / Tool
Programming Language	Python
Machine-Learning Library	scikit-learn
Data Handling	pandas, NumPy

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Component	Technology / Tool
Algorithms	Logistic Regression, Decision Tree, Random Forest, SVM, Gradient Boosting
Additional Tooling	TensorFlow (where applicable)
Evaluation	Accuracy and related classification metrics
Deployment	Local or cloud; report export

B. Implementation Details

The implementation follows a structured pipeline. Telecom customer data including usage details, billing information, and customer behaviour is collected and cleaned by handling missing values and converting it into a proper format. Important features that influence customer decisions are selected, and the dataset is divided into training and testing sets. Machine-learning algorithms—Logistic Regression and Decision Tree, along with Random Forest, SVM, and Gradient Boosting—are trained on historical data so the model learns patterns and classifies customers into churn or non-churn categories. The trained model predicts whether a customer is likely to leave or continue using the service, the best-performing model is selected based on evaluation metrics, and results can be displayed and exported as reports.

C. Models and Evaluation

Logistic Regression provides an interpretable linear baseline, Decision Tree captures non-linear, rule-based relationships, and ensemble methods such as Random Forest and Gradient Boosting improve robustness and accuracy, while SVM handles complex decision boundaries. The models are evaluated using accuracy and other classification metrics on the held-out test set, and the strongest model is selected. The source describes these outcomes qualitatively; no specific numeric accuracy values are asserted here, and real-world performance depends on data quality and feature relevance.

VI. SYSTEM TESTING AND RESULTS

The system was validated through ten functional and validation test cases covering application loading, dataset upload, data preprocessing, model training, churn prediction, result display, report generation, input validation, model loading, and result export, with integration testing ensuring smooth interaction between preprocessing, the machine-learning model, and the prediction modules. All test cases passed successfully and behaved as expected.

TABLE III. REPRESENTATIVE TEST CASES

ID	Description	Input	Expected Output	Status
TC-01	System loads successfully	Open application	Home page displayed	**Pass**
TC-02	Upload customer dataset	Valid dataset	Data loaded	**Pass**

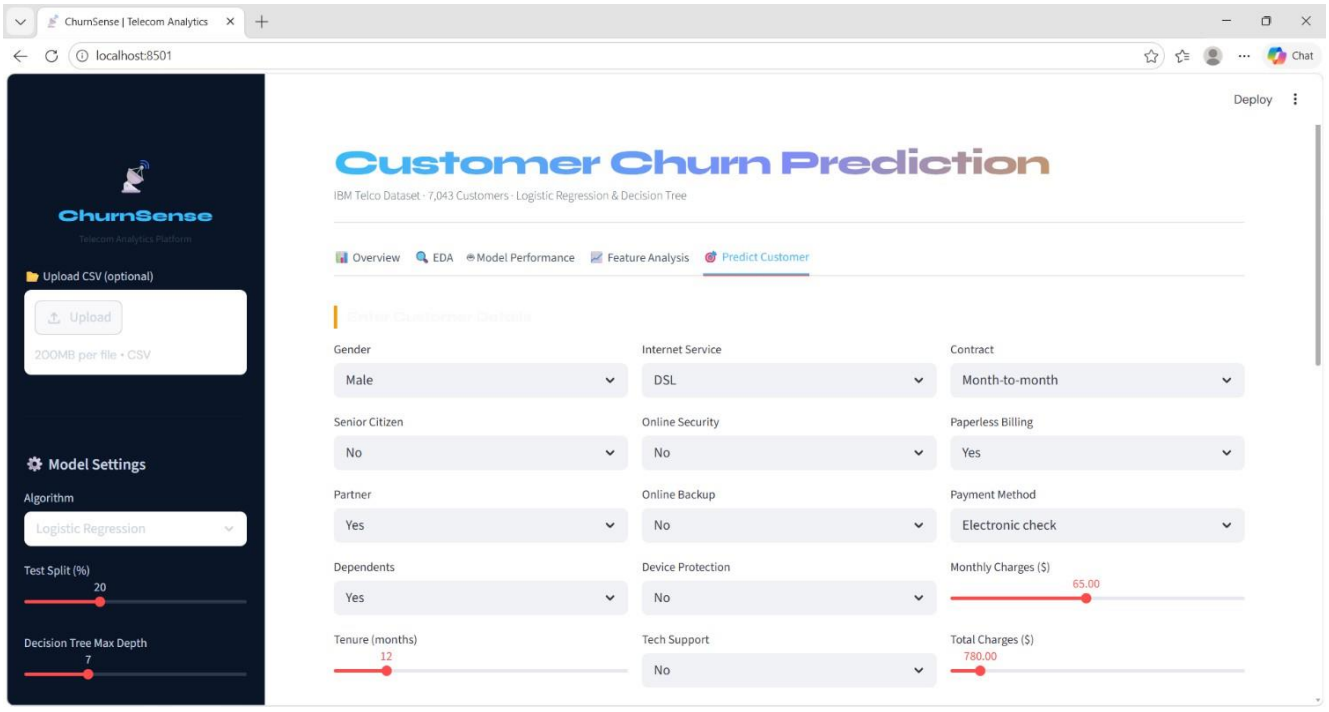
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

ID	Description	Input	Expected Output	Status
TC-03	Data preprocessing	Raw data	Cleaned data	**Pass**
TC-04	Train model	Training dataset	Model trained	**Pass**
TC-05	Predict churn	Customer data	Churn / No Churn	**Pass**
TC-08	Validate input data	Invalid data	Error message	**Pass**
TC-10	Export results	Download request	File downloaded	**Pass**

A. Observed Results

The implemented system loads and preprocesses telecom customer data, trains the classification models, and predicts whether a customer is likely to churn or stay, presenting results and exportable reports. By replacing manual statistical analysis with a machine-learning pipeline, the system identifies high-risk customers earlier and provides actionable insights for retention, handling larger datasets more efficiently than traditional approaches. The source reports these outcomes qualitatively; no specific numeric metrics are claimed here, and performance depends on the quality and representativeness of the customer dataset.

Representative screenshots from the prototype implementation:



International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

F Fig. 1. Input Screen1 For Customer Chunk Prediction.

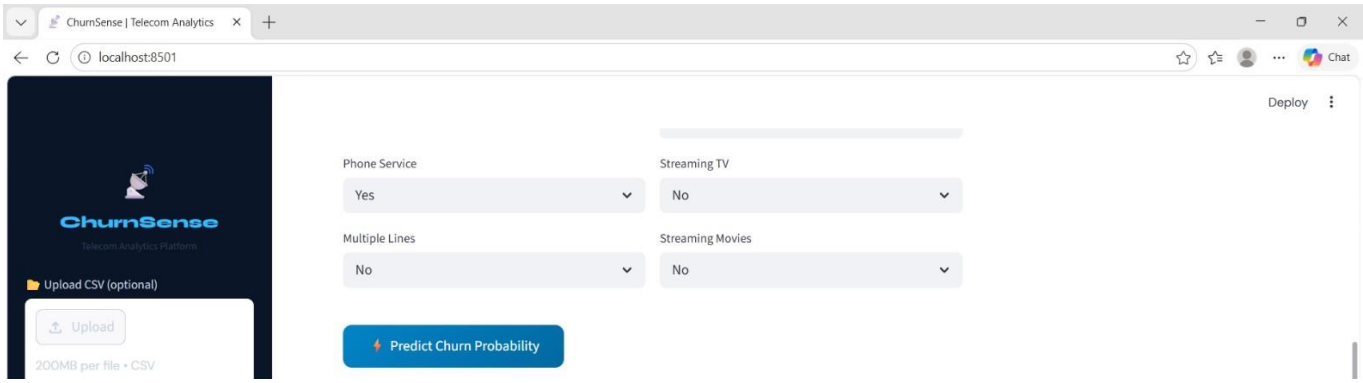


Fig.1 Input Screen2 for the Customer Chunk Prediction .

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

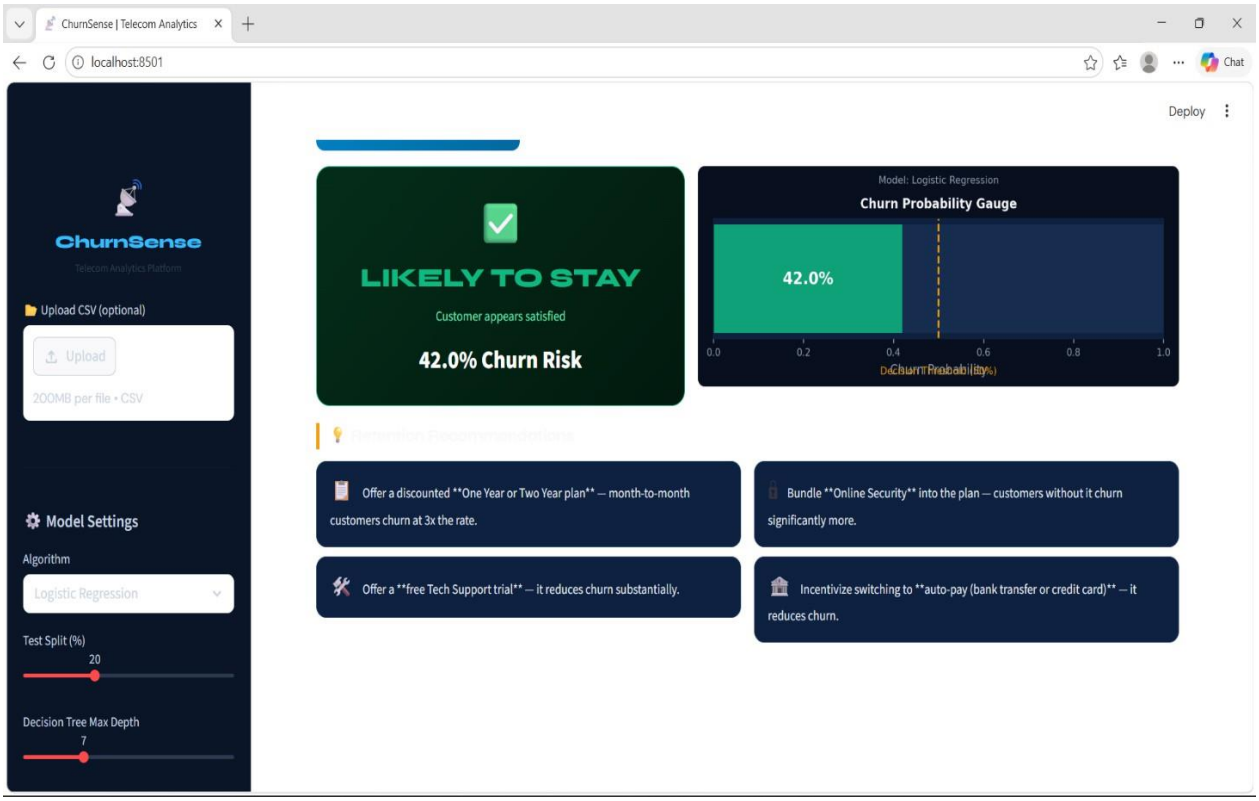


Fig. 3. Output Screen For Customer Chunk Prediction.

VII. CONCLUSION AND FUTURE SCOPE

The proposed project, Customer Churn Prediction in the Telecom Industry, successfully demonstrates the application of Machine Learning techniques to analyse customer behaviour and predict whether a telecom customer is likely to leave (churn) or stay with the service provider. In the highly competitive telecom sector, customer retention is a major challenge, and identifying potential churners in advance helps companies take proactive retention measures. By cleaning and preprocessing customer data, selecting important features, and training classification models such as Logistic Regression and Decision Tree (along with Random Forest, SVM, and Gradient Boosting), the system learns patterns from historical data and classifies customers accurately, providing actionable insights that help companies offer better services, discounts, or improved support. Compared with traditional manual and statistical methods, the system is more accurate, scalable, and automated, ultimately helping increase customer satisfaction, reduce customer loss, and improve business profit.

Although the system provides accurate and useful predictions, several opportunities exist for further improvement. Future versions can incorporate sentiment analysis from customer feedback, call records, and social-media data to improve prediction accuracy, and integrate with Customer Relationship Management (CRM) systems to make the solution more industry-ready. Real-time data processing, more advanced

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

models, larger and richer datasets, and cloud-based deployment can further enhance accuracy, scalability, and responsiveness, providing better customer-retention strategies for telecom companies.

REFERENCES

[1] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Sebastopol, CA, USA: O'Reilly Media, 2019.

[2] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[3] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[4] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT Press, 2012.

[5] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[6] Python Software Foundation, "Python Language Documentation," 2023. [Online]. Available: https://docs.python.org/

[7] Scikit-learn Developers, "Scikit-learn Machine Learning Library," 2023. [Online]. Available: https://scikit-learn.org/

[8] TensorFlow Team, "TensorFlow Documentation," 2023. [Online]. Available: https://www.tensorflow.org/

[9] IBM Research, "AI and Machine Learning for Business Analytics," 2022.

[10] Telecom Industry Reports, "Customer Retention and Churn Analysis," 2022.