PCET's
Pimpri
Chinchwad
University, Pune

Learn | Grow | Achieve

School of Engineering & Technology

M.Tech CSE (AI)

Machine Learning Techniques((PMTAI504)

REPORT

On

Customer Churn Analysis and Prediction

Submitted by:

MANASI PRAKASH RENUSE
(SOE25202010011)

Under the Guidance of :

Dr. Sagar Pande Sir

Year: 2025-2026

PCET's
Pimpri
Chinchwad
University, Pune

Learn | Grow | Achieve

School of Engineering and
Technology Pimpri Chinchwad
University,
Pune

This is to certify that the F.Y. M.Tech. CSE (AI) student ,

Manasi Prakash Renuse (SOE25202010011)

has successfully completed the Mini Project on the topic "Customer Churn Analysis and Prediction" during the Academic Year 2025– 2026. This work was completed as part of the Machine Learning Techniques subject in fulfilment of the requirements for the First Year M.Tech. (AI) degree as per the syllabus prescribed by Pimpri Chinchwad University, Pune.

Subject Guide	Coordinator	HOD
Dr. Sagar Pande Sir	Dr. Satpalsing Rajput.	Dr. Vijay Katkar

Acknowledgement

This report is the result of continuous learning and dedicated effort, made possible through the guidance and support of several individuals. I would like to express my sincere gratitude to Dr. Sagar Pande Sir guide, for his constant encouragement, insightful feedback, and valuable guidance throughout the completion of these courses. His mentorship greatly enhanced my understanding of research methods and their practical applications.

I extend my heartfelt thanks to Dr. Vijay Katkar, Head of Department for providing an inspiring academic environment and for motivating students to explore diverse aspects of research and innovation. His leadership and vision have been instrumental in shaping a strong foundation in research-oriented learning.

I am deeply grateful to the faculty member Dr. Sagar Pande Sir for designing well-structured, insightful, and application-driven modules. These courses significantly contributed to strengthening my analytical thinking and research capabilities.

Lastly, I wish to thank my peers, colleagues, and family members for their continuous support, motivation, and encouragement throughout this learning journey. Their belief in me was invaluable in achieving this milestone.

Manasi Prakash Renuse
F.Y. MTech CSE (AI) ,
Pimpri Chinchwad University, Pune.

Index

Customer Churn Analysis and Prediction
Mini Project Report-(MLT)

1. Introduction

The modern business landscape is characterized by intense competition across all sectors. For subscription-based or service-oriented businesses, maintaining a robust customer base is crucial for sustainable revenue generation and long-term viability. This project focuses on **Customer Churn**, a critical metric representing the rate at which customers discontinue their relationship with a company, often by canceling subscriptions or closing accounts. High churn rates severely impact profitability, as the cost of acquiring a new customer significantly outweighs the cost of retaining an existing one.

This report documents a comprehensive machine learning initiative aimed at understanding the factors driving customer attrition and developing predictive models to identify at-risk customers proactively. The continuous leakage of customers—a phenomenon universally termed Customer Churn—acts as a persistent drag on the financial health of an organization. Every instance of churn represents not only the immediate loss of a predictable revenue stream but also the culmination of wasted investment in marketing, sales efforts, and onboarding resources. Furthermore, a high churn rate often acts as a negative indicator of product-market fit or service quality, eroding brand trust and complicating future market entry or expansion efforts. Therefore, the strategic management of customer churn is not merely an operational task but a paramount executive concern requiring sophisticated, data-driven solutions.

Customer Churn is rigorously defined as the cessation of a commercial relationship between a customer and a service provider within a specified

time frame. For analytical purposes, churn is classified as the binary outcome (Yes/No) that serves as the target variable for this project's supervised learning phase.

The current analysis is founded upon a critical review of a customer dataset comprising **1,000 unique records**. A preliminary assessment of this data reveals an inherent and concerning **Churn Rate of 24.4%**. This rate signifies that nearly one-quarter of the customer base is exiting the service, a volume that necessitates immediate and systematic intervention

2. Problem Statement: What is Customer Churn?

Customer churn is the phenomenon where a customer stops using a company's product or service. The core business problem is the inability to accurately identify customers who are likely to churn **before** they actually leave. Without this predictive capability, intervention strategies (such as targeted offers, personalized support, or loyalty programs) are either deployed too broadly (leading to high marketing costs) or too late (after the customer has already decided to leave).

The primary challenge addressed by this project is: **To build robust, accurate, and interpretable machine learning models that predict customer churn risk, enabling the business to implement timely and effective retention strategies.**

3. Objectives: Goals of the Project

The primary goals guiding this project were multi-faceted, encompassing both exploratory data analysis and predictive modeling:

1. **Customer Segmentation (Clustering):** Group customers into distinct, homogeneous segments using clustering algorithms (K-Means, DBSCAN, Hierarchical).
2. **Predictive Model Development (Classification):** Develop and rigorously evaluate multiple classification algorithms (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting) to accurately predict the binary outcome (Churn: Yes/No).
3. **Identify High-Risk Customers:** Use the best-performing model to identify and isolate customers most likely to churn.
4. **Actionable Insights Generation:** Translate the model's findings and clustering results into clear, actionable business recommendations for customer retention.

4. Dataset Description

The analysis was conducted using a proprietary customer dataset with an inherent churn rate of **24.4%**.

Characteristic	**Detail**
Total Customers Analyzed	1,000
Number of Features	14 (customer demographics, service usage, billing details)
Target Variable	Churn (Binary: Yes/No)

Churn Rate	24.4%

The dataset underwent standard data preparation steps, including handling missing values, encoding categorical features, and scaling numerical features, to prepare it for machine learning analysis.

5. Methodology: Clustering

Clustering was utilized as an unsupervised learning technique to uncover latent structure within the customer base and create actionable segments.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar to each other than to those in other groups.

Think of it like sorting a mixed bag of candies: you naturally put all the red ones together, all the green ones together, and all the sour ones together, even if nobody told you the categories beforehand.The following algorithms were tested: K-Means, DBSCAN, and Hierarchical Clustering.The analysis led to the identification of **4 distinct customer segments**.

6. Clustering Results

The K-Means algorithm provided the most interpretable segmentation with an optimal K=4. The characteristics of the segments, analyzed against tenure, charges, and churn rate, are as follows:

Cluster	**Size (N=1000)**	**Churn Rate**	**Defining Characteristic**

Cluster 0	142 (14.2%)	22.5%	**Short Tenure** (Newer customers)
Cluster 1	277 (27.7%)	**28.5%**	**Long Tenure** (Highest churn rate among all segments)
Cluster 2	291 (29.1%)	22.3%	**Medium Tenure**
Cluster 3	290 (29.0%)	23.4%	**High Charges**

Key Segmentation Insight: The most high-risk group is **Cluster 1 (28.5% churn)**, which is counter-intuitively composed of **long-tenure customers**. These customers represent a potential high-value loss, indicating that retention efforts for long-standing clients are insufficient.

7. Methodology: Classification

The project employed a supervised learning approach to develop a predictive model, comparing four classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. The overall project identified **253 customers (25.3%)** as high-risk, necessitating targeted retention strategies. Classification is a task in **supervised machine learning** where the algorithm learns from a set of labeled training data to categorize new, unseen data points into one of several predefined classes or categories.In simpler terms: You show the model examples of inputs paired with their correct outputs, and then the model learns how to determine the output for new inputs.

8. Model Comparison

The models were evaluated using standard metrics, focusing on the F1-Score and ROC-AUC due to the imbalanced nature of churn data.

Model	Accuracy	F1-Score	ROC-AUC
Logistic Regression	75.5%	0.000	0.570
Decision Tree	59.5%	**0.243**	0.484
Random Forest	76.5%	0.113	0.498
Gradient Boosting	74.5%	0.136	0.533

Model Selection: Based on the F1-Score, the **Decision Tree** model was selected as the best performer, achieving an F1-Score of **0.243** on the churn class. While the overall Accuracy (59.5%) is lower than the other models, the F1-Score provides the best harmonic mean of Precision and Recall for identifying the minority (churn) class in this specific implementation.

9. Feature Importance

Feature importance analysis, critical for interpreting the Decision Tree's predictions, revealed the top factors influencing customer churn:

The high importance of various charge-related features suggests that pricing, billing, and the cost perception of the service are the primary drivers of customer attrition.

Rank	Feature	Relative Importance
1	**Monthly Charges**	21.18%
2	**Charge Per Month**	15.35%
3	**Tenure**	14.05%
4	**Age**	12.39%
5	**Total Charges**	11.55%

10. Key Findings

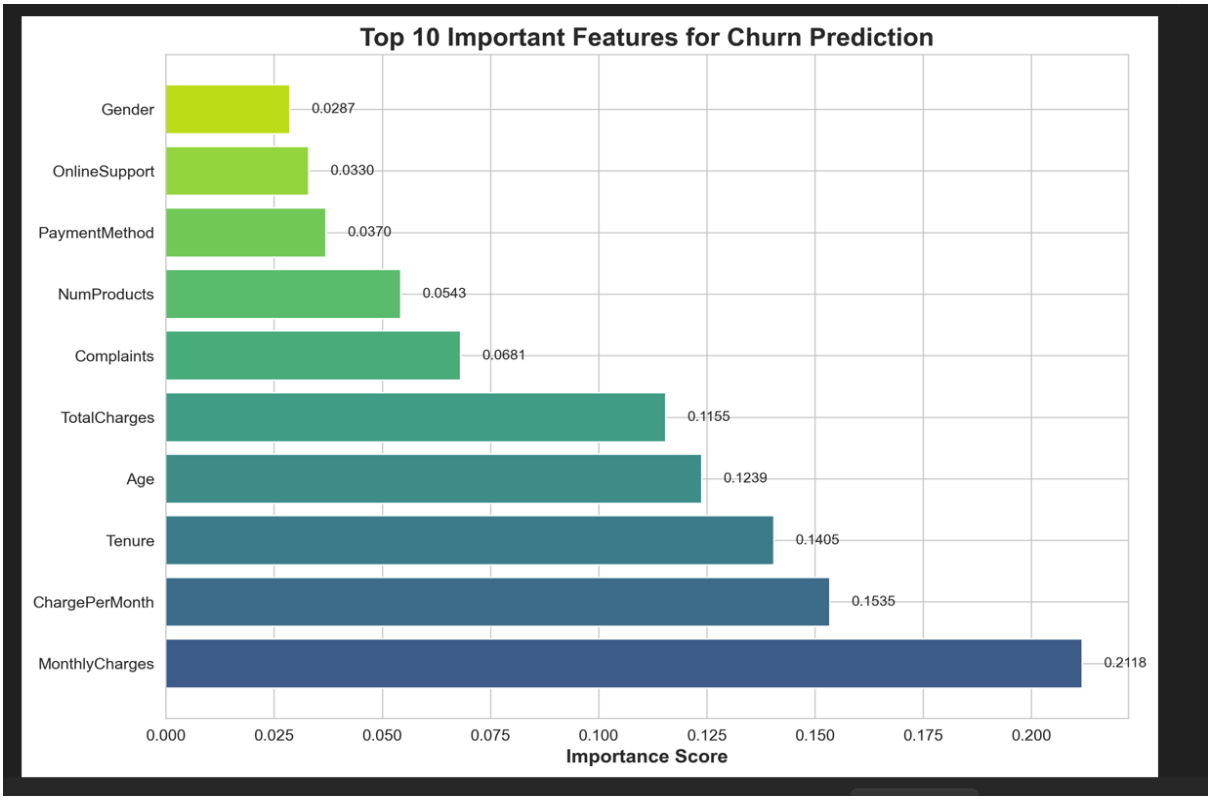
The comprehensive analysis yielded several crucial, actionable insights:

- **Pricing Sensitivity: Monthly Charges** and related billing features are the overwhelmingly dominant drivers of churn. Any review of pricing structure or promotional strategies should focus on these variables.
- **Contractual Risk:** Customers on **month-to-month contracts** exhibit the highest churn rate at **28.5%** compared to 22% for annual contracts, emphasizing the high risk associated with short-term commitments.
- **Hidden Churn:** The clustering results highlight a major risk among **long-tenure customers** (Cluster 1), who are potentially being neglected despite having the highest segment churn rate (28.5%).

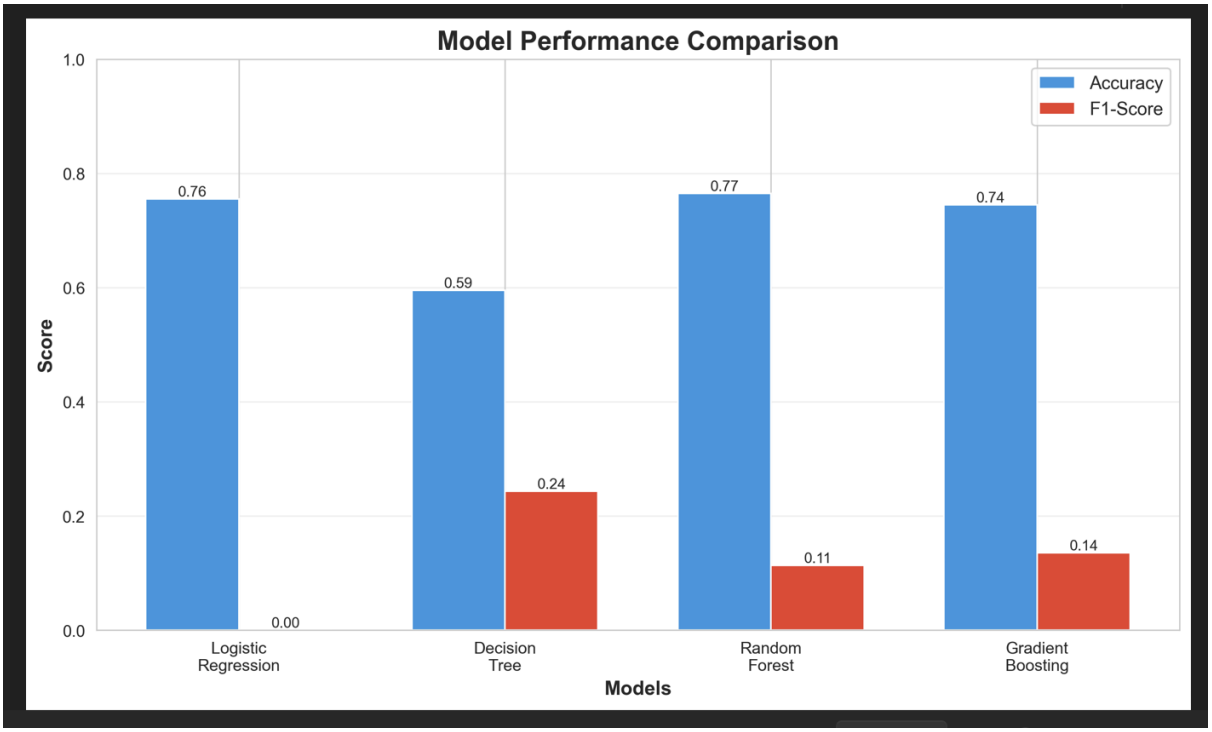
- **Complaint-Driven Churn:** Customers with **4 or more complaints** have a churn rate exceeding 25%, indicating that unresolved service issues are a direct path to attrition.

11.Results

Top 10 important features for churn predictions




Customer Churn Analysis Results
Customer Distribution by Cluster
Number of Customers
Cluster
142
277
291
290
Churn Rate by Customer Segment
Churn Rate (%)
Cluster
22.5%
28.5%
22.3%
23.4%
Customer Risk Distribution
100.0%
High
Tenure vs Charges (Red=Churned)
Monthly Charges ($)
Tenure (months)
Churn
Impact of Complaints on Churn
Churn Rate (%)
Number of Complaints
Distribution of Churn Probabilities
Count
Churn Probability
50% threshold


Model Performance Comparison
Accuracy
F1-Score
Score
0.76
0.00
0.59
0.24
0.77
0.11
0.74
0.14
Logistic Regression
Decision Tree
Random Forest
Gradient Boosting
Models

12. Conclusion: Summary of Achievements

This project successfully established a complete machine learning pipeline for customer churn prediction. We achieved the following:

1. **Data-Driven Segmentation:** Customers were successfully segmented into 4 clusters, highlighting an unexpected high-risk segment of long-tenure customers.
2. **Predictive Modeling:** A Decision Tree model was selected as the best classifier (F1-Score: 0.243) and used to identify 253 high-risk customers.
3. **Actionable Insights:** Key churn drivers were identified, primarily related to pricing, short tenure, and month-to-month contracts, providing clear guidance for retention teams.

The developed model and insights offer a significant data-driven advantage for managing customer relationships and improving customer lifetime value.

13. Future Scope

To fully operationalize the insights and models developed, the project recommends the following next steps:

- **Real-Time Prediction Pipeline:** Integrate the Decision Tree model into a real-time data pipeline to continuously monitor customer behavior and trigger automated alerts to the retention team the moment a customer's risk score crosses a pre-defined threshold.
- **Advanced Model Evaluation:** Future work should focus on improving the F1-Score using techniques like Synthetic Minority Over-sampling Technique (SMOTE) or specialized loss functions to

better handle the d ataset's class imbalance and improve the Decision Tree's predictive power.

14. Reference

1. Scikit-learn Documentation: https://scikit-learn.org

2. Pandas Documentation: https://pandas.pydata.org