SRM UNIVERSITY - DEPARTMENT OF COMPUTATIONAL INTELLIGENCE

SCHOOL OF COMPUTING

CASE STUDY REPORT

Course Code: 21AIC401T
Course Name: Inferential Statistics and Predictive Analytics
Assignment Type: Case Study-Based Modeling Project
Title: Machine Learning Based Customer Churn Prediction Using Logistic Regression and Decision Tree Models

Student Name & Register Number:

Student Name: **Aagam Chhajer** Register No.: **RA2211047010110**

Date: 10 November 2025

---

1. Introduction and Data Preparation

Customer churn represents a major strategic challenge for the telecommunications industry. Telecom companies face aggressive competition, easy switching options, and rapidly changing consumer expectations. When customers leave, the organization not only loses current revenue but also must spend additional funds and effort to acquire new customers. Therefore, churn prevention and proactive decision-making are more cost-effective than customer acquisition.

Predictive analytics plays a vital role in identifying churn risks early. Machine Learning models can learn patterns from historical customer records and identify behavioral trends that indicate future churn. In this study, two supervised machine learning models—Logistic Regression and Decision Tree—are built and evaluated using the Telco Customer Churn dataset. These models classify whether a customer is likely to churn and help telecom companies take data-driven retention decisions.

1.1 Objective

The goal of this case study is to:

- Predict which customers have high probability of leaving the telecom service.
- Compare model performances between Logistic Regression and Decision Tree.
- Identify key variables that strongly influence churn outcomes.
- Develop actionable business level insights based on model interpretation.
- Present recommendations that can reduce attrition and improve customer retention strategies.

1.2 Dataset Description

The dataset used is the publicly available "Telco Customer Churn Dataset".
It contains 7043 customer records and 21 independent attributes including demographic data, subscription type, billing pattern, monthly charges, tenure, payment method, service add-ons and internet service types.

Target Variable: Churn (Binary: Yes = 1, No = 0)
Data Structure: Mix of numeric, categorical, nominal and binary fields

---

1.3 Data Cleaning and Preparation

Data preprocessing steps used:

- Converted "TotalCharges" to numeric, removed invalid numeric rows post conversion.
- Standardized categorical labels such as "No Internet Service".
- Applied one-hot encoding for categorical variables.
- Scaled continuous attributes (MonthlyCharges, Tenure) for Logistic Regression stability.
- Performed 70:30 train-test split with stratification to preserve churn rate proportion.

Category	Attributes
Demographic	**Gender, SeniorCitizen, Partner, Dependents**
Service Subscription	**InternetService, StreamingTV, OnlineSecurity, OnlineBackup etc**
Billing & Contract	**MonthlyCharges, TotalCharges, PaymentMethod, Contract**
Tenure	**Length of customer relationship in months**

1.5 Exploratory Data Analysis (EDA)

Churn Distribution

~26.5% of customers churn. This indicates imbalance and business critical minority class.

[Insert Figure: churn_distribution.png]

Monthly Charges vs Churn

Customers with higher MonthlyCharges show higher churn tendency indicating cost sensitivity.

Correlation Heatmap Key Patterns

[Insert Figure: correlation_heatmap.png]

- `tenure` negatively correlated → long term customers less likely to leave.
- Contract (Two Year) negatively correlated → commitment protects churn risk.
- OnlineSecurity (Yes) negatively correlated → service add-ons increase stickiness.

2. Methods

Two supervised machine learning models were selected for experimentation and comparison: **Logistic Regression** and **Decision Tree**. These models are widely used in binary classification problems and provide interpretable insights into churn behavior.

Logistic Regression

This algorithm predicts probability of churn using the sigmoid function:

$$P(churn = 1|X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 X1 + \beta 2 X2 + ... + \beta n Xn)}}$$

Parameters are estimated by minimizing Binary Cross Entropy Loss optimized with gradient descent.

Decision Tree Classifier

Decision Tree splits dataset into branches using impurity measures like Gini:

$$Gini = 1 - \sum_{k=1}^{n} p_k^2$$

where pk is proportion of class k in the node.
Decision Trees are simple to interpret but can overfit, so depth-based pruning was applied.

3. Model Building

- Dataset split into **70% training** and **30% testing** using stratified partitioning.
- Logistic Regression trained after One Hot Encoding and Feature Scaling.
- Decision Tree trained with controlled depth and minimal split constraints.
- Both models saved using **pickle (.pkl)** format for reproducibility and deployment support.

4. Model Evaluation

Model performance was evaluated using Accuracy, Precision, Recall, F1-Score and ROC-AUC.

These metrics evaluate correctness, class sensitivity and ranking of predicted churn probability.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	0.7313	0.4947	0.5027	0.4987	0.6584
Logistic Regression	**0.8038**	**0.6494**	**0.5695**	**0.6068**	**0.8364**

Based on ROC-AUC comparison, Logistic Regression shows superior capability in differentiating churners vs non-churners.

[Insert Figure: roc_curve_comparison.png]

Interpretation

- Logistic Regression performs significantly better overall and is chosen as final model.
- Recall ~0.57 suggests more churners are correctly identified than random baseline.
- ROC-AUC 0.8364 indicates strong model discrimination capability.

5. Conclusion

This case study demonstrated the development of a predictive analytics solution to identify telecom customer churn risk. Multiple machine learning algorithms were tested and Logistic Regression achieved the highest performance. Key churn indicators include contract type, high monthly charges, lower tenure, and lack of online security services. The results support data-driven decision making for customer retention. By applying proactive retention strategies such as contract upgrade incentives, personalised billing adjustments, and targeted add-on service offers, telecom companies can significantly reduce churn, retain revenue and sustain customer loyalty.

References

- Kaggle: Telco Customer Churn Dataset
- Pedregosa et al., Scikit-Learn (2011)
- Molnar, C. Interpretable Machine Learning (2022)

Appendix

Python Libraries Used: Pandas, NumPy, Scikit-learn, Matplotlib
Project Repository: https://github.com/AagamChhajer/Customer-Churn-Prediction