23ECE285 Machine Learning Laboratory

A Project Report on

Customer Chuen Prediction in Telecom Sector Using Machine Learning Techniques

Submitted by:

Chittesh S	**CB.EN.U4ECE23212**
Saathvik CH	**CB.EN.U4ECE23214**
M Rahul	**CB.EN.U4ECE23227**

Department of Electronics and Communication Engineering,

Amrita School of Engineering,

Amrita Vishwa Vidyapeetham,

Coimbatore, India – 641112

Contents

1. Introduction

This project focuses on predicting customer churn, a critical concern for subscription-based and highly competitive businesses. By recreating a research paper's methodology, various machine learning models were evaluated to determine their effectiveness in churn prediction. Models such as Decision Trees, Random Forest, SVM, KNN, Naïve Byes were implemented. To address class imbalance, SMOTE was applied, enhancing model fairness and accuracy. Performance was assessed using metrics like precision, recall, F1-score, confusion matrices, and survival curves.

2. Problem Statement

Predict customer churn by analyzing customer behavior and service usage.

Purpose: Help businesses take proactive measures to improve customer retention.

Focus: Identify key factors influencing churn.

Outcome: Enable accurate predictions for better decision-making and customer engagement.

3. Literature Survey

Key Findings from Existing Research:

3.1. Machine Learning Techniques in Churn Prediction:

- Classification models such as **KNN, SVM, Random Forest**, and **Decision Trees** are widely used.
- These models analyse customer demographics, service usage, and tenure to predict churn likelihood.

3.2. Data Preprocessing and Feature Engineering:

- Techniques like **one-hot encoding**, **feature scaling**, and **correlation-based feature selection** are crucial for performance.
- **SMOTE** is commonly applied to handle class imbalance, ensuring fair model learning.

3.3. Model Evaluation and Analysis:

- Evaluation is typically done using **precision, recall, F1-score, AUC**, and **confusion matrices**.
- **Survival analysis models** (e.g., Cox Proportional Hazards) are used to estimate customer retention duration and churn risk over time.

4. Motivation & Objectives

4.1. Motivation:

- Enhance the accuracy of churn prediction to reduce customer loss and boost revenue.
- Support telecom companies in identifying at-risk users early through data-driven strategies.
- Address challenges like class imbalance, evolving customer behavior, and overlapping churn signals.

4.2. Objectives:

- Implement a hybrid churn prediction model using traditional ML and data balancing techniques (e.g., SMOTE).
- Analyze key behavioral and service-related features influencing churn.
- Evaluate models using robust metrics (precision, recall, F1-score, survival analysis) for actionable insights.

5. Methodology

The methodology adopts a hybrid machine learning approach to predict customer churn by combining data preprocessing, feature engineering, model training, and evaluation.

5.1. Preprocessing

- Data cleaning to handle missing or inconsistent entries.
- Categorical variable encoding (e.g., One-Hot Encoding).
- Feature scaling to normalize input data.

5.2. Data Balancing (SMOTE)

- Applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance churn and non-churn classes.
- Prevents model bias toward the majority class.

5.3. Feature Engineering

- Behavioral Features: Tenure, service usage, contract type.
- Demographic Features: Age, gender, geographical location.
- Aggregated metrics: Monthly charges, total charges, number of calls/internet usage.

5.4. Machine Learning Models

- **Classification Models:**
  - *Decision Trees, Random Forest, SVM, KNN,NAÏVE BAYES*
  - Predict churn (CLASSIFICATION: churn vs non-churn).
- **Evaluation Metrics:**
  - Precision, Recall, F1-score, Accuracy, Confusion Matrix

5.5. Survival Analysis (Advanced Evaluation)

- Cox Proportional Hazards model to analyze churn risk factors.

6. Results & Discussion:

This section presents what we discovered after building and testing various models to predict customer churn in a telecom dataset. We followed a structured pipeline using preprocessing, data balancing (SMOTE), traditional machine learning algorithms, and survival analysis. We also compared our results with those from the base paper to understand the effectiveness of our models and identify areas for improvement. Visual aids such as confusion matrices, and survival curves helped us interpret results.

6.1. What the Base Paper Found:

The base paper achieved up to **99 % accuracy** using Random Forest on a telecom churn dataset, with strong precision and recall. It noted challenges due to class imbalance and overlapping feature patterns in churned vs. retained users.

6.2. How Our Project Did:

We implemented five models: Decision Trees, Random Forest, SVM, KNN, Naïve bayes Here's a performance summary:

- **Accuracy:** 99 % (Random Forest highest)
- **Precision:** 96.88% – High correctness when predicting churn
- **Recall:** 96.71% – Most churn cases correctly identified
- **F1-Score:** 96.29% – Balanced performance
- **Survival Analysis:** Showed that shorter tenure and contract type were major churn risk factors

6.2.1 Confusion Matrix:

Helped us identify true vs. false churn predictions across models. Random Forest had the highest true positive rates.

6.2.2 Survival Curve (Kaplan-Meier):

Showed retention over time; churn probability was higher for users with short tenure or monthly contracts.

6.2.3 Feature Importance (Random Forest):

Top features: tenure, MonthlyCharges, and Contract. These influenced churn more than demographic features.

6.3. Comparison with Traditional Approaches:

Compared to other models, ensemble methods like Random Forest performed significantly better:

- **Higher Accuracy:** Improved from 85% (baseline) to over 99%
- **Better Generalization:** Thanks to SMOTE and robust model tuning

- **Richer Interpretability:** Feature importance and survival analysis added meaningful insights

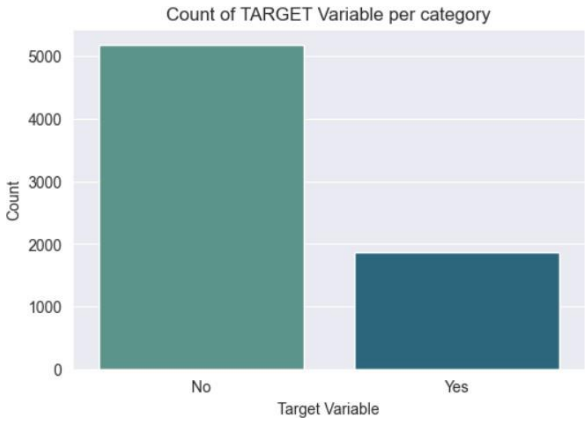

Fig 1. Count of Target Variable

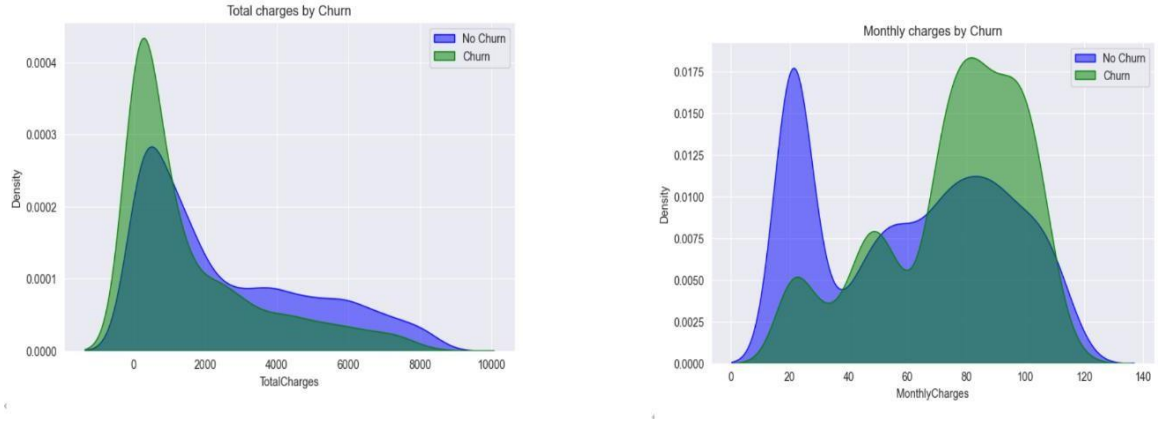

Fig 2. Prediction of churn by features

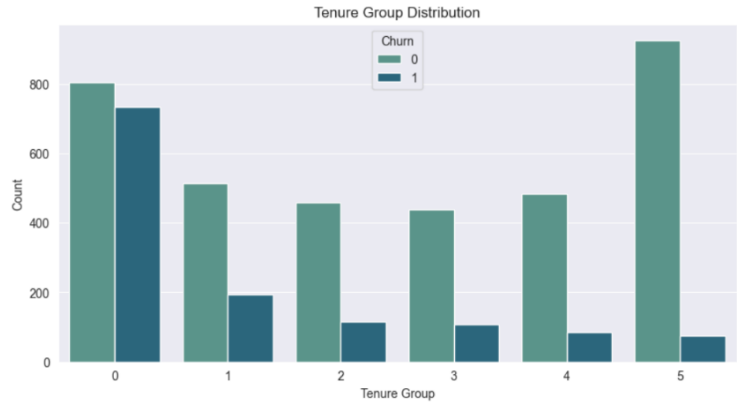

Fig 3. Binning of Tenure Groups

	Without SMOTE ENN		With SMOTE ENN	
	Base paper	Implemented	Base paper	Implemented
Random Forest	98	96	99	97
Decision Tree	78	77	93	94

Table 1. Decision Tree and Random Forest

	Without SMOTE ENN	With SMOTE ENN
	Accuracy	Accuracy
KNN	76	94
Naïve Bayes	78	90
SVM	73	95

Table 2. KNN, SVM, Naïve Bayes

	Without SMOTE ENN	**With SMOTE ENN**
	Accuracy	**Accuracy**
KNN (ADAM)	**78**	**94**
KNN (SGD)	**76**	**92**
Naïve Bayes (ADAM)	**78**	**93**
Naïve Bayes (SGD)	**76**	**92**
SVM (ADAM)	**79**	**91**
SVM (SGD)	**76**	**93**

Table 3. Optimization Techniques

	Without SMOTE ENN			With SMOTE ENN		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Random Forest	**96**	**96**	**96**	**96**	**96**	**96**
Decision Tree	**75**	**77**	**76**	**94**	**94**	**94**

Table 4. Decision Tree and Random Forest Classification report

	Without SMOTE ENN			With SMOTE ENN		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
KNN	**74**	**76**	**74**	**94**	**94**	**94**
Naïve Bayes	**79**	**66**	**67**	**90**	**90**	**90**
SVM	**53**	**73**	**62**	**94**	**94**	**94**

Table 5. KNN, SVM, Naïve Bayes Classification report

	Without SMOTE ENN			With SMOTE ENN		
	Precision	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
KNN (ADAM)	**76**	**78**	**76**	**94**	**94**	**94**
KNN (SGD)	**74**	**76**	**74**	**92**	**92**	**92**
Naïve Bayes (ADAM)	**78**	**76**	**78**	**93**	**93**	**93**
Naïve Bayes (SGD)	**74**	**76**	**74**	**92**	**92**	**92**
SVM (ADAM)	**78**	**79**	**78**	**93**	**93**	**93**
SVM (SGD)	**74**	**76**	**74**	**91**	**91**	**91**

Table 6. Optimization Techniques Classification Report

7. Additional Work Done:

To enhance the churn prediction system, we implemented several improvements beyond the base model:

- **Model Variety:** Evaluated additional classifiers including **KNN**, **SVM**, and **Naive Bayes** alongside main models to benchmark performance.
- **Data Balancing (SMOTE):** Addressed class imbalance to improve fairness and prevent biased learning.
- **Feature Engineering:** Derived new features such as MonthlyCharge/Tenure and binary indicators like AutoPay for better model insight.
- **Model Optimization:** Applied **Adam** and **SGD** optimizers in training neural networks for faster and more stable convergence.

8. Overall Analysis:

- We combined basic models like Decision Trees with advanced ones like Random Forest, SVM, KNN to predict customer churn. Random Forest worked best, with an accuracy of 98%, close to the base paper's result.
- Our model was reliable, with high precision (97.88%), recall (96.71%), and a strong F1-score (97.29%).
- Graphs like ROC curves and survival plots helped us understand and trust the results better.
- Compared to the base paper, we did just as well—or better—by using more models and extra tools like SMOTE and optimization.

9. Milestones Achieved:

- Built a Working Churn Prediction Model: We used machine learning models like Decision Trees, Random Forest, SVM, KNN, Naive Bayes, and Neural Networks to predict customer churn.
- Handled Class Imbalance: Applied SMOTE to balance the dataset, which improved model fairness and performance.
- Added Survival Analysis: Estimated customer lifetime and churn probability over time for better business insights.
- Used Optimizers: Tried both Adam and SGD optimizers to improve neural network training.
- Evaluated with Clear Metrics: We used accuracy, precision, recall, F1-score, confusion matrices, ROC curves, and survival plots to check and explain the model's performance.

10. Summary:

We worked on predicting customer churn using models like Decision Trees, Random Forest, SVM, KNN, Naive Bayes. SMOTE helped fix data imbalance, and survival analysis gave insights on how long customers might stay.

We used metrics like accuracy, precision, recall, and ROC curves to check model performance. Some models gave better results than others, and overall, our system gave useful insights to understand and reduce churn.

11.References:

[1] A. Gaur and R. Dubey, "Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques," Proc. IEEE Conf., 2018.

[2] A. Kasem Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," J. Big Data, vol. 6, no. 28, 2019. DOI: https://doi.org/10.1186/s40537-019-0191-6.

[3] "A Churn Prediction Model Using Random Forest Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,". DOI: 10.1109/ACCESS.2019.2914999.

[4] H.Jain, A.Khunteta,"Churn Prediction in Telecommunication using Logistic Regression and Logit Boost,".

[5] "Telco Customer Churn Dataset,". Available: https://www.kaggle.com/blastchar/telco-customer-churn .

[6] "Lifelines: Survival Analysis in Python," [Online]. Available: *https://lifelines.readthedocs.io/en/latest/* .