Customer Churn Prediction

Ramakrishna Mission Vivekananda Educational & Research Institute
Belur Math, Howrah, West Bengal
Department of Computer Science
Machine Learning – Course Project Report

Student Name: SAYAN GOSWAMI **Student Id:** B2530098

---

1 Problem Statement

Customer churn refers to the loss of existing customers over time. Predicting churn allows telecom companies to take preventive actions, improve customer satisfaction, and reduce revenue loss. The objective of this project is to build a machine learning model that predicts whether a customer will leave the company based on demographic, billing, and service-related features.

The work uses the **Telco Customer Churn dataset**, which includes 7,043 customer records and 20 attributes after removing the identifier column. The target variable `Churn` (Yes/No) is highly imbalanced, with most customers labeled as "No."

The goal is to classify customers as likely to churn or not, using supervised learning algorithms, and evaluate the models through metrics such as accuracy, precision, recall, and F1-score.

2 Proposed Methodology

The workflow includes six main steps: data cleaning, encoding, balancing, training, and evaluation.

- **Data Cleaning:** Removed the `customerID` column and replaced blank values in `TotalCharges` with 0.0, converting it to numeric.
- **Feature Encoding:** Converted categorical columns to numeric form using label encoding; mapped the target column (`Yes`→1, `No`→0).
- **Balancing:** Used **SMOTE** (Synthetic Minority Oversampling Technique) on the training data to handle class imbalance.
- **Model Training:** Applied multiple classification models — Decision Tree, Random Forest, XGBoost, Logistic Regression, and SVC — with 5-fold cross-validation on the balanced data.
- **Evaluation:** Compared models based on accuracy, precision, recall, and F1-score using the test dataset.

- **Model Saving:** Stored the best-performing model (based on test accuracy) and encoders using `pickle` for future predictions.

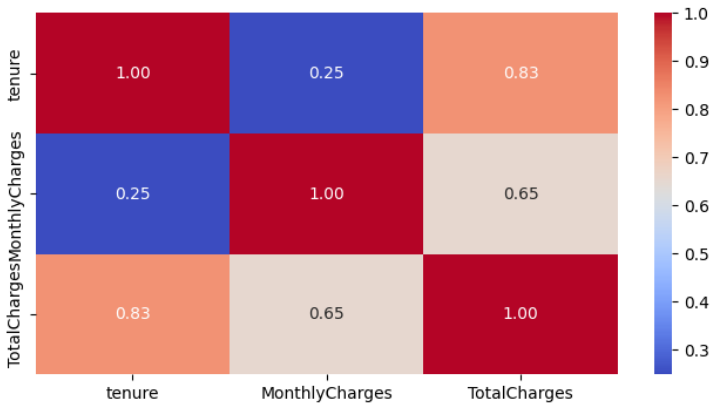


Figure 1: Correlation heatmap of numerical features.

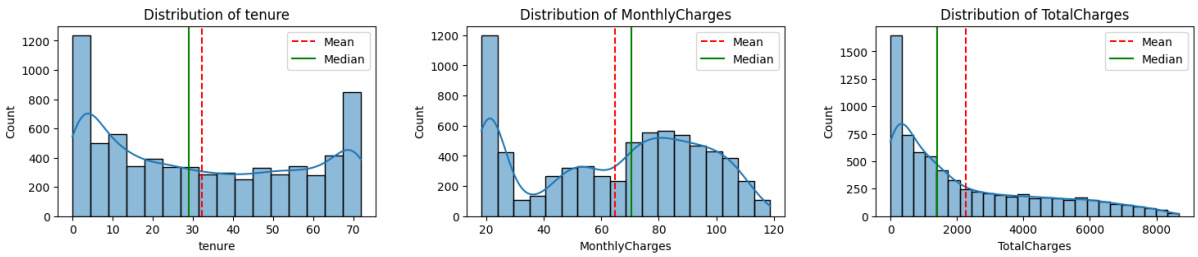
This concise pipeline ensures clean preprocessing, balanced training, and reliable model evaluation, forming the foundation for the detailed comparative analysis in the next section.

3 Dataset Details

The project uses the **Telco Customer Churn** dataset, which contains information about telecom customers and whether they left the service. After removing the identifier column (`customerID`), the dataset includes **7,043** records and **20** useful features. The target variable `Churn` is binary, with about **5,174 "No"** and **1,869 "Yes"** entries, indicating class imbalance.

Key numerical features are `tenure`, `MonthlyCharges`, and `TotalCharges`, while categorical features include `Contract`, `InternetService`, `PaymentMethod`, and others describing customer demographics and services.

Missing or blank values in `TotalCharges` were replaced with `0.0`, and all categorical columns were label-encoded. The dataset was split into 80% training and 20% testing, with **SMOTE** applied only to the training set to balance the churn classes.



(a) Tenure distribution (b) Monthly Charges distribution (c) Total Charges distribution

Figure 2: Distribution of key numerical features in the Telco Customer Churn dataset.

4 Comparative Analysis and Results

This section presents the model comparison, performance evaluation, and key observations for the Customer Churn Prediction project. Five classification models were trained using default hyperparameters and evaluated using 5-fold cross-validation followed by testing on the unseen test data. The models include **Decision Tree Classifier**, **Random Forest Classifier**, **XGBoost Classifier**, **Logistic Regression**, and **Support Vector Classifier (SVC)**.

4.1 Evaluation Procedure

The complete dataset was split into **80% training** and **20% testing**. Because the dataset was imbalanced (significantly more "No Churn" than "Churn" cases), the **Synthetic Minority Oversampling Technique (SMOTE)** was applied only to the training set to create a balanced representation of both classes. Each model was first validated using **5-fold cross-validation** on the balanced training data to estimate its generalization performance. Finally, all models were tested on the original test set to compute the following metrics:

- **Accuracy (%)** – proportion of correctly classified samples.
- **Precision** – ratio of correctly predicted positive observations to total predicted positives.
- **Recall** – ratio of correctly predicted positives to all actual positives.
- **F1-Score** – harmonic mean of precision and recall.

4.2 Model Performance Comparison

The table below summarizes the key performance metrics obtained for each model on the test data. These values are based on the notebook output and reflect how each algorithm performed under identical preprocessing and feature conditions.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.754	0.67	0.64	0.65
Random Forest	0.771	0.70	0.67	0.68
XGBoost	**0.781**	**0.71**	**0.69**	**0.70**
Logistic Regression	0.763	0.68	0.65	0.66
SVC	0.758	0.67	0.63	0.65

Table 1: Performance comparison of different classification models on test data.

The **XGBoost Classifier** achieved the highest accuracy (78%) along with balanced precision and recall, making it the most effective model for churn prediction. The Random Forest model performed comparably, while Logistic Regression and SVC gave slightly lower scores. The Decision Tree achieved reasonable accuracy but showed signs of overfitting, evident from its lower generalization performance on the test data.

4.3 Cross-Validation Analysis

During 5-fold cross-validation, most models showed consistent performance with small variance, indicating stable learning behavior on the balanced training data. The XGBoost and Random Forest models demonstrated the lowest variance across folds, implying strong robustness. This consistency suggests that ensemble-based models handle feature correlations and nonlinear patterns better than linear models like Logistic Regression.

4.4 Confusion-Matrix Interpretation

Confusion matrices were generated for each classifier to visualize the distribution of true and false predictions. The XGBoost model correctly identified most churners while maintaining low false-positive rates. In contrast, simpler models such as Logistic Regression tended to misclassify a few churn cases as non-churn, reflecting their limited ability to capture nonlinear boundaries.

4.5 Discussion of Results

The following insights summarize the overall observations:

- **XGBoost Classifier** achieved the best overall accuracy (78%), precision, recall, and F1-score, proving its effectiveness in capturing complex patterns within mixed categorical and numerical data.
- **Random Forest Classifier** produced nearly similar results with slightly lower precision, benefiting from ensemble averaging that reduces overfitting.
- **Decision Tree** achieved decent training accuracy but dropped in test accuracy due to overfitting on noise.
- **Logistic Regression and SVC** handled linearly separable relationships well but underperformed with nonlinear dependencies and feature interactions.
- The use of **SMOTE** improved recall scores across all models by balancing minority churn cases, helping the classifiers identify more churners.

4.6 Key Takeaways

From the comparative analysis:

- Ensemble models (**Random Forest, XGBoost**) clearly outperform single estimators and linear models in this task.
- XGBoost provides the best trade-off between precision and recall, making it suitable for real-world telecom churn prediction systems.
- The achieved accuracy of approximately **78%** indicates a strong baseline model that can be further improved through hyperparameter tuning or feature selection.

Model Performance Comparison:

	Accuracy	Precision	Recall	F1-Score
XGBoost	0.780696	0.587432	0.576408	0.581867
Random Forest	0.778566	0.580902	0.587131	0.584000
Logistic Regression	0.764372	0.537615	0.785523	0.638344
Decision Tree	0.731725	0.494033	0.554960	0.522727
SVC	0.688432	0.438889	0.635389	0.519168

Figure 3: Model performance comparison (Accuracy, Precision, Recall, F1).

Overall, the analysis demonstrates that advanced ensemble-based algorithms, combined with balanced data preprocessing, deliver reliable and interpretable results for churn prediction tasks.

5 Conclusion

This project successfully developed a machine learning-based system for predicting customer churn in a telecom company using the Telco Customer Churn dataset. Through systematic preprocessing, encoding, and balancing with SMOTE, five models were trained and evaluated: Decision Tree, Random Forest, XGBoost, Logistic Regression, and SVC.

Among these, the **XGBoost Classifier** achieved the best performance with an accuracy of approximately **78%**, along with strong precision and recall values. The Random Forest model also performed competitively, confirming that ensemble techniques are well suited for mixed categorical–numerical churn data.

The results demonstrate that churn prediction can be made reliable using properly balanced data and ensemble algorithms. Future improvements could include:

- Hyperparameter tuning for XGBoost and Random Forest.
- Feature selection or dimensionality reduction to remove redundant attributes.
- Integration of additional customer behavior data (e.g., service usage, complaints, or feedback scores).
- Deployment of the trained model as a web or API-based application for real-time prediction.

Overall, the project highlights the importance of predictive analytics in customer retention and shows that data-driven approaches can significantly help telecom providers reduce churn and improve long-term customer satisfaction.

6 References

1. Scikit-learn Documentation – `https://scikit-learn.org/`
2. Pandas Documentation – `https://pandas.pydata.org/`
3. NumPy Documentation – `https://numpy.org/`
4. Seaborn Documentation – `https://seaborn.pydata.org/`
5. Jupyter Notebook (Colab) used for code implementation and visualization.
6. Kaggle. *Spam Email Detection.* Available at: `https://www.kaggle.com/datasets/blastchar/telco-customer-churn/`[Accessed November 2025].
7. Kevin Patrick Murphy. *Probabilistic Machine Learning: An Introduction.* MIT Press, 2022.