

# Customer Churn Prediction – ML Project Report

## Author

Dhusyanth R S

GitHub Repository: [GitHub](#)

## 1. Introduction

This project predicts customer churn for a telecom company using machine learning techniques. It demonstrates a full ML workflow including data cleaning, preprocessing, EDA, model comparison, hyperparameter tuning, pipeline creation, and model saving.

## 2. Dataset Overview

- **Dataset:** Telco Customer Churn
- **Total Rows:** 7043
- **Target Variable:** Churn (Yes / No)
- **Problem Type:** Binary Classification

### 2.1 Features

- Demographics
- Phone & internet services
- Contract and payment details
- Billing and usage features

### 2.2 Data Cleaning

- Removed customerID
- Converted total\_charges to numeric
- Filled missing values
- Standardized column names

### 3. Exploratory Data Analysis (EDA)

Key findings:

- Month-to-month contract users show highest churn
- Higher monthly charges correlate with churn
- Longer-tenure customers show lower churn
- Payment method influences retention

Visualizations included numeric distributions, churn counts, and churn relationships with major features.

### 4. Preprocessing

A **ColumnTransformer** was used to apply:

Feature Type	Method
Numerical	StandardScaler
Categorical	OneHotEncoder / OrdinalEncoder
Target	LabelEncoder

All transformations were combined into a single ML pipeline.

### 5. Model Development

Several models were trained and compared:

- Logistic Regression
- KNN
- SVC
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost

## 5.1 Hyperparameter Tuning

Used:

- **GridSearchCV**
- **RandomizedSearchCV**
- 5-fold cross-validation

## 5.2 Best Model

The **RandomForestClassifier (tuned version)** showed the best overall performance and generalization.

This tuned RF pipeline was selected as the final model.

## 6. Model Evaluation

Metrics evaluated:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Churn class (minority) performance was moderate, which is expected due to dataset imbalance.

Top features affecting churn included monthly charges, contract type, tenure, and payment method.

## 7. Final Model Saving

The complete preprocessing + model pipeline was saved as:

`models/churn_model.pkl`

This allows immediate prediction without re-running preprocessing steps.

## 8. Project Structure

```
customer_churn_prediction/
|
|   -- data/
|   -- notebook/
|   -- models/
|   -- requirements.txt
|   -- README.md
```

## 9. Key Insights

- High-risk groups include month-to-month contract and high-charge customers
- Tenure is a strong indicator of churn
- Model tuning significantly improved performance
- Pipelines ensure reproducibility and ease of deployment

## 10. Future Work

- Handle class imbalance (SMOTE / class weights)
- Use advanced models like LightGBM or CatBoost
- Deploy via Streamlit or FastAPI
- Add SHAP explainability
- Create dashboards for business teams

## 11. Conclusion

This project demonstrates a complete ML workflow for churn prediction.

The final tuned Random Forest pipeline provides a stable baseline and a strong foundation for future deployment or enhancement.