Customer Churn Prediction Using Machine Learning

An end-to-end machine learning project predicting telecom customer churn through comprehensive preprocessing, exploratory data analysis, multiple model training, hyperparameter tuning, and final pipeline selection—delivering a complete industry-standard workflow.

Understanding the Challenge

Why Churn Prediction Matters

Customer churn represents a critical business challenge for telecom companies. Predicting which customers are likely to leave enables proactive retention strategies, reducing acquisition costs and preserving revenue streams.

This project implements a full machine learning lifecycle, transforming raw customer data into actionable predictions through systematic analysis and model development.

Project Workflow Overview

01

Data Cleaning

Standardising formats, handling missing values, and preparing raw data for analysis

02

Exploratory Analysis

Uncovering patterns and relationships that influence customer churn behaviour

03

Feature Engineering

Transforming raw features into optimal model inputs using sophisticated preprocessing

04

Model Training

Evaluating multiple algorithms to identify the best-performing approach

05

Hyperparameter Tuning

Optimising model performance through GridSearchCV and RandomizedSearchCV

06

Pipeline Deployment

Packaging the final solution for production use

Dataset Characteristics

7,043

Total Customer Records

Comprehensive dataset providing robust sample size for model training

20+

Feature Variables

Rich attribute set spanning demographics, services, and billing patterns

2

Target Classes

Binary classification: churned versus retained customers

Key Feature Categories

Demographics

Customer age, gender, and household composition

Service Details

Phone and internet service configurations

Contract Terms

Agreement types and payment methods

Financial Metrics

Monthly charges and total spending patterns

Data Preparation & Cleaning

Identifier Removal

Eliminated customerID and irrelevant columns to focus on predictive features

Type Conversion

Converted total_charges to numeric format with proper coercion handling

Missing Value Treatment

Applied systematic imputation strategies for incomplete records

Naming Standardisation

Unified column naming conventions for consistency across the pipeline

Exploratory Data Analysis Insights

Key Patterns Discovered

Churn Distribution

Analysed the balance between churned and retained customers to understand class distribution and potential imbalances

Contract Type Impact

Month-to-month contracts showed significantly higher churn rates compared to longer-term commitments

Tenure Relationships

Newer customers demonstrated elevated churn risk, with retention improving substantially after initial months

Charge Patterns

Higher monthly charges correlated with increased churn probability, suggesting price sensitivity

Visualisations created using Matplotlib and Seaborn revealed critical relationships between customer characteristics and churn behaviour, informing feature engineering decisions.

Preprocessing Pipeline Architecture

Numerical Features

StandardScaler normalises continuous variables to zero mean and unit variance

Categorical Features

OneHotEncoder and OrdinalEncoder transform categorical variables appropriately

Target Variable

LabelEncoder converts binary churn labels into numerical format

A ColumnTransformer orchestrates all transformations within a unified pipeline, ensuring consistent preprocessing across training and prediction phases whilst preventing data leakage.

Model Training & Evaluation

Algorithms Evaluated

- **Logistic Regression**

  Baseline linear model providing interpretable coefficients

- **K-Nearest Neighbours**

  Instance-based learning capturing local patterns

- **Support Vector Classifier**

  Margin-based approach for complex decision boundaries

- **Decision Tree Classifier**

  Interpretable tree-based model with feature splits

- **Random Forest Classifier**

  Ensemble method reducing overfitting through bagging

- **Gradient Boosting Classifier**

  Sequential ensemble building on previous errors

- **XGBoost Classifier**

  Optimised gradient boosting with regularisation

Each algorithm was trained and evaluated using consistent cross-validation procedures, enabling fair performance comparisons across diverse modelling approaches.

Hyperparameter Optimisation Strategy

GridSearchCV

Exhaustive search across specified parameter grids, evaluating every combination to identify optimal settings with guaranteed thoroughness

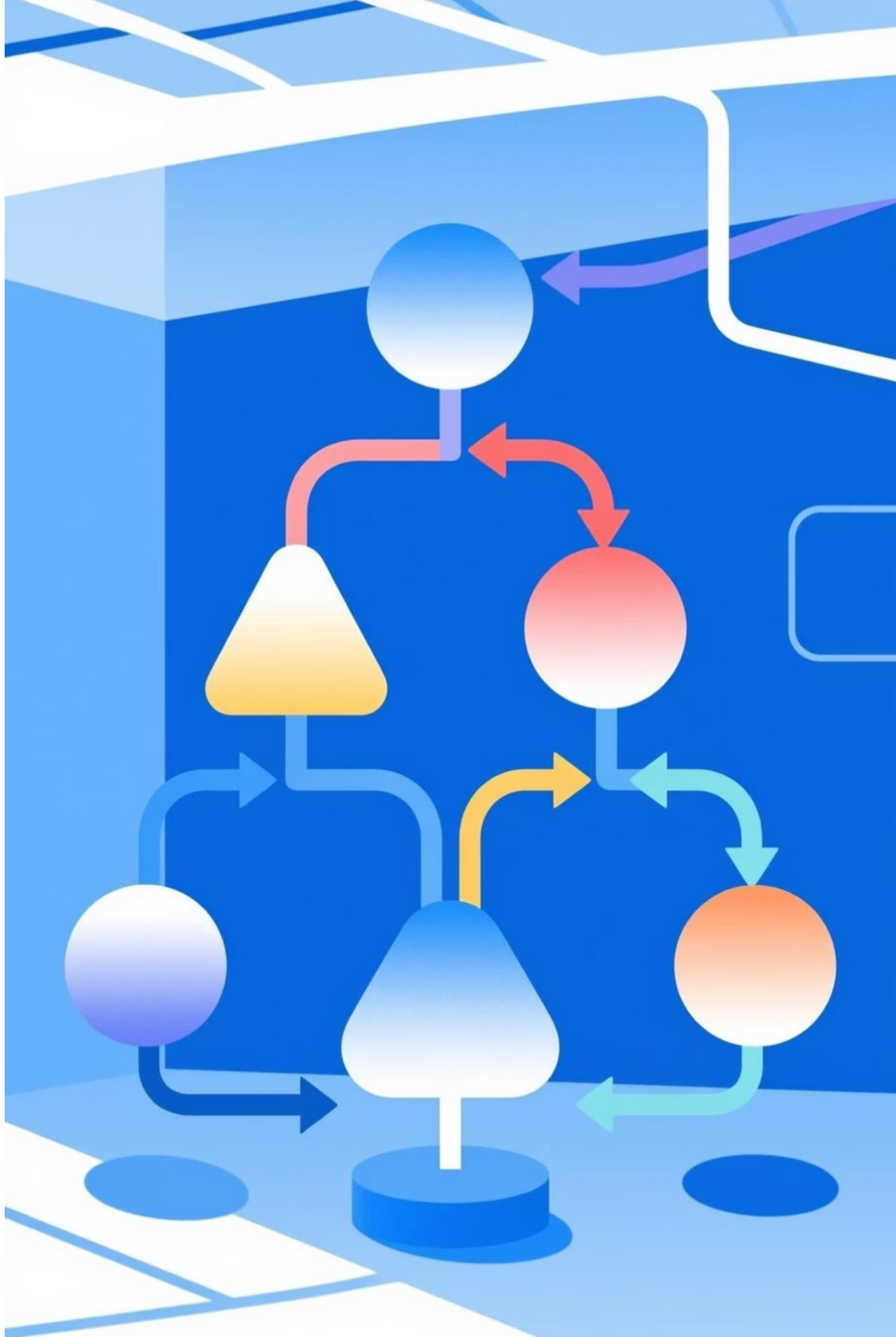
- Systematic parameter exploration
- 5-fold cross-validation
- Comprehensive coverage

RandomizedSearchCV

Efficient exploration of wider parameter spaces through random sampling, balancing computational cost with search breadth

- Rapid evaluation
- Broader parameter ranges
- Computational efficiency

Scoring metrics included accuracy, precision, and recall, ensuring models balanced overall performance with specific business requirements for churn detection.



Final Model Selection

Champion: Tuned Random Forest Classifier

After comprehensive evaluation across all candidates, the hyperparameter-tuned Random Forest Classifier emerged as the optimal solution, demonstrating superior performance characteristics:

Consistent Performance

Reliable predictions across validation folds with minimal variance

Balanced Generalisation

Strong performance on both training and test sets, avoiding overfitting

Robust Stability

Resilient to minor data variations and edge cases

The final pipeline, incorporating all preprocessing steps and the optimised model, has been serialised as a .pkl file for production deployment.

By :
Dhusyanth R S
dhusyanthrs3@gmail.com
GitHub