Supervised Machine Learning Project Report

Telco Customer Churn Classification

Author: Dhatri Shree Podugu

Date: October 2025

Dataset: IBM Telco Customer Churn Dataset

Table of Contents

1. Main Objective of the Analysis

The main objective of this project is to develop a supervised machine learning classification model that predicts whether a telecom customer will churn (leave the service). The analysis focuses on improving prediction accuracy while maintaining interpretability for business stakeholders. Identifying at-risk customers early helps the business launch retention campaigns, enhance customer satisfaction, and minimize revenue loss.

2. Description of the Dataset

The IBM Telco Customer Churn dataset contains 7,043 customer records and 21 features describing account, billing, and service details. The target variable 'Churn' indicates whether a customer has left the company.

Key attributes include:
- Customer demographics: gender, senior citizen, partner, dependents
- Account information: tenure, contract type, payment method, billing type
- Service features: internet service, online security, tech support, streaming services

Data preparation steps:
- Converted 'TotalCharges' to numeric format and handled missing values.
- Dropped irrelevant identifiers like 'customerID'.
- Encoded categorical variables using OneHotEncoder with handle_unknown=ignore.
- Scaled numeric features using StandardScaler.
- Split data into training (80%) and testing (20%) subsets with stratified sampling.

3. Data Exploration and Feature Engineering

Exploratory data analysis (EDA) revealed that approximately 26% of customers churned. Churn rates were significantly higher among customers with month-to-month contracts, higher monthly charges, and no add-on services like tech support or online security. Conversely, customers with longer tenure and two-year contracts showed strong loyalty.

No new features were engineered beyond encoding, but transformations improved model interpretability.

4. Model Training and Evaluation

Three classification models were trained using the same training-test split and preprocessing pipeline:
1. Logistic Regression – baseline model focused on interpretability.
2. Random Forest – ensemble method that handles nonlinear relationships.
3. Gradient Boosting – boosting-based ensemble optimized for predictive power.

Performance Summary:
- Logistic Regression: ROC-AUC ≈ 0.80
- Random Forest: ROC-AUC ≈ 0.84

- Gradient Boosting: ROC-AUC ≈ 0.85

All models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

5. Final Model Recommendation

The Gradient Boosting model is recommended as the final choice due to its highest ROC-AUC and F1-score, indicating a strong balance between predictive power and reliability. Although Logistic Regression offers more transparency, Gradient Boosting captures complex relationships and outperforms in recall — a crucial metric for churn prevention.

6. Key Findings and Insights

Key Insights:
- Customers with month-to-month contracts and high monthly charges are more likely to churn.
- Lack of tech support or online security correlates with higher churn.
- Customers with longer tenure and annual contracts tend to stay.

Business Recommendations:
- Encourage customers to switch to annual contracts.
- Offer loyalty discounts or bundled tech support.
- Monitor high-charge customers closely and offer retention incentives.

7. Model Limitations and Next Steps

Limitations:
- Moderate class imbalance in churn data may cause prediction bias.
- The model does not account for cost-sensitive business metrics.
- Absence of behavioral or competitor data limits accuracy.

Next Steps:
- Apply SMOTE or class weighting to handle imbalance.
- Perform probability calibration and optimize decision thresholds.
- Add new behavioral features (usage patterns, complaints).
- Test XGBoost and LightGBM models for improved generalization.

8. Conclusion

This project demonstrates the value of supervised ML in customer retention. The Gradient Boosting model successfully predicts churn and provides actionable insights for targeted retention strategies. Implementing these recommendations can improve retention rates and profitability, making ML-driven analytics a vital component of strategic decision-making in telecom businesses.