Customer Churn Prediction System
Final Project Report

Project Team

Aiman Osama Mohamed - Team Leader

Rana Hazem Hassan - Data Analyst

Motaz Ahmed Ismail - ML Model Developer

Rofida Mohamed Ahmed - MLOps & Deployment Engineer

Menna Khalid Mohamed - Documentation & Presentation Specialist

Report Date: November 25, 2025

1. Executive Summary

This project successfully delivered a complete customer churn prediction system for Orange Telecom, addressing the critical business challenge of customer retention. Our team developed and deployed a highly accurate machine learning model that identifies at-risk customers with 97% accuracy, enabling proactive retention strategies.

Key achievements include:

- Processed and analyzed 3,197 customer records with 20+ features
- Identified key churn drivers through comprehensive statistical analysis
- Developed and evaluated seven machine learning models
- Deployed XGBoost model achieving 97% test accuracy
- Implemented production-ready Streamlit web application
- Established MLFlow system for model tracking and management

The system is now operational, providing business teams with actionable insights to reduce customer churn and improve retention rates. This project demonstrates the practical application of machine learning to solve real-world business problems.

2. Project Overview

Customer churn represents a significant financial challenge in the telecommunications industry, where acquiring new customers costs 5-7 times more than retaining existing ones. This project addresses this challenge by building a predictive system that identifies customers likely to churn.

2.1 Project Objectives

- Build accurate churn prediction models using Orange Telecom dataset
- Identify key factors driving customer churn through comprehensive analysis
- Deploy interactive prediction tool for business teams
- Provide actionable insights for customer retention strategies
- Establish reproducible ML pipeline for ongoing model improvement

2.2 Project Timeline

- Milestone 1 - Data Exploration & Preprocessing: October 18, 2025
- Milestone 2 - Advanced Data Analysis: November 1, 2025
- Milestone 3 - Model Development: November 15, 2025
- Milestone 4 - Deployment & MLOps: November 27, 2025
- Final Documentation & Presentation: December 3, 2025

3. Data Collection & Preprocessing

The project began with two primary datasets containing customer information from Orange Telecom. We combined and cleaned these datasets to create a unified foundation for analysis.

3.1 Data Sources

- churn-bigml-80.csv - Primary training dataset
- churn-bigml-20.csv - Supplementary dataset
- Total combined records: 3,197 customers
- Features: 20 variables including demographics, usage patterns, and service plans

3.2 Data Cleaning Process

- Merged multiple data sources into unified structure
- Standardized state abbreviations to full state names
- Encoded categorical variables (Yes/No → 1/0)
- Removed outliers using Interquartile Range (IQR) method
- Handled missing values and validated data integrity
- Converted data types for optimal memory usage

4. Exploratory Data Analysis

We conducted comprehensive exploratory analysis to understand customer behavior patterns and identify initial churn indicators.

4.1 Key Findings

- Clear class imbalance between churn and non-churn customers
- Geographic variations in churn rates across different states
- Service usage patterns differ significantly between customer groups
- International plan subscribers show higher churn propensity
- Customer service interactions correlate with churn behavior

4.2 Visualization Approach

- Distribution analysis of numerical features
- Categorical variable analysis using count plots
- Correlation heatmaps for feature relationships
- State-level churn analysis with comparative charts
- Multivariate exploration with pair plots and conditioned histograms

5. Statistical Analysis & Feature Engineering

Advanced statistical methods helped validate initial findings and identify the most predictive features for model development.

5.1 Statistical Testing

- Chi-square tests for categorical feature significance
- T-tests for numerical variable differences between groups
- Correlation analysis to identify multicollinearity
- Hypothesis validation for key churn drivers

5.2 Feature Engineering

- Total national minutes and charges aggregation
- Average minutes per call calculations
- Tenure categories based on account length
- High service calls indicator (>3 calls)
- Service plan combination flags
- Cost per minute metrics

6. Model Development & Evaluation

We trained and optimized seven machine learning models to find the best solution for churn prediction, following rigorous evaluation protocols.

6.1 Methodology

- Applied ADASYN oversampling to handle class imbalance
- Implemented GridSearchCV for hyperparameter tuning
- Used 5-fold cross-validation for robust evaluation
- Evaluated models on multiple metrics: accuracy, precision, recall, F1, AUC

6.2 Model Performance

Final test accuracy results:

- XGBoost: 97.0% - Selected for deployment
- Random Forest: 96.0%
- Decision Tree: 92.0%
- Support Vector Machine: 88.0%
- K-Nearest Neighbors: 85.0%
- Logistic Regression: 84.0%
- Gaussian Naive Bayes: 84.0%

XGBoost demonstrated superior performance with the highest accuracy and excellent generalization capabilities, making it the ideal choice for production deployment.

7. System Deployment

The churn prediction system is deployed through multiple components that support both business users and technical teams.

7.1 Streamlit Web Application

- User-friendly interface for customer service teams
- Real-time churn probability calculations
- Interactive input forms with validation
- Clear visual indicators of churn risk
- Mobile-responsive design

7.2 MLFlow Implementation

- Comprehensive model versioning and tracking
- Performance metrics logging across all models
- Confusion matrix visualization and storage
- Experiment reproducibility management
- Centralized model registry

7.3 Production Architecture

- Model serialization using joblib for efficient loading
- Consistent preprocessing pipeline between training and inference
- Error handling and input validation mechanisms
- Scalable design supporting future enhancements

8. Business Impact & Value Proposition

The churn prediction system delivers significant business value across multiple organizational functions, transforming customer retention strategies.

8.1 Operational Benefits

- Customer service: Targeted outreach to high-risk accounts
- Marketing: Data-driven retention campaign planning
- Management: Visibility into churn trends and risk factors
- Product development: Insights into service feature impact

8.2 Financial Impact

- Reduced customer acquisition costs through improved retention
- Optimized allocation of retention resources and incentives
- Protected recurring revenue from existing customer base
- Enhanced customer lifetime value through proactive management

Based on industry analysis, reducing monthly churn by 5% could yield annual savings exceeding $2 million for a medium-sized telecom company, demonstrating substantial ROI from this predictive system.

9. Project Deliverables

The project successfully delivered all planned components:

9.1 Documentation

- EDA Report - Comprehensive data exploration and preprocessing
- Data Analysis Report - Statistical analysis and feature engineering
- Model Evaluation Report - Model development and performance analysis
- MLOps Report - Deployment architecture and implementation
- Final Project Report - Complete project summary and business impact

9.2 Technical Assets

- Cleaned and processed dataset ready for analysis
- Seven trained machine learning models with optimal parameters
- Production-ready Streamlit web application
- MLFlow experiment tracking system
- Complete codebase with documentation and deployment scripts

10. Conclusion

This project successfully delivers a robust, production-ready churn prediction system that addresses real business challenges in the telecommunications industry. Through comprehensive data analysis, rigorous model development, and practical deployment, our team has created a solution that provides immediate business value.

The system demonstrates how machine learning can transform customer relationship management by enabling proactive, data-driven decision making. With 97% prediction accuracy and user-friendly interfaces, the solution empowers business teams to effectively combat customer churn and improve retention rates.

All project objectives have been met, delivering a complete analytics solution that combines technical sophistication with practical business application. The system establishes a strong foundation for ongoing customer analytics development and represents a significant step forward in data-driven customer management.