# Customer Churn Prediction Analysis Report

## Table of Contents

1. [Executive Summary & Project Overview](https://claude.ai/chat/1e206703-7478-4cc7-8bb1-581a92f75860" \l "executive-summary--project-overview)
2. [Technical Implementation](https://claude.ai/chat/1e206703-7478-4cc7-8bb1-581a92f75860" \l "technical-implementation)
3. [Results & Business Insights](https://claude.ai/chat/1e206703-7478-4cc7-8bb1-581a92f75860" \l "results--business-insights)

## Executive Summary & Project Overview

This report analyzes a customer churn prediction project that employed multiple machine learning algorithms to identify at-risk customers. The project utilized a comprehensive approach combining exploratory data analysis, advanced preprocessing techniques, and ensemble learning methods across four different models.

**Key Results:** Logistic Regression achieved the highest F1 score (0.60), while Random Forest demonstrated the most balanced performance. The analysis identified tenure months, monthly charges, and customer age as the strongest predictive factors for customer churn.

### Dataset Overview

The dataset contains customer information including demographics (age, gender), service characteristics (subscription type, payment method, auto-renewal), usage patterns (tenure, monthly charges), and support interactions. The target variable represents binary churn status with approximately 5,200 non-churned versus 4,800 churned customers, initially showing a 75:25 churn ratio that indicated moderate class imbalance requiring correction.

## Technical Implementation

### Libraries and Technologies

**Core Libraries:**

* **pandas & numpy**: Data manipulation and numerical operations
* **matplotlib & seaborn**: Data visualization and statistical plotting
* **scikit-learn**: Machine learning framework including preprocessing, algorithms, and evaluation metrics
* **XGBoost & LightGBM**: Advanced gradient boosting implementations
* **imbalanced-learn**: SMOTE implementation for class imbalance handling
* **joblib**: Model serialization and persistence

### Data Preprocessing Pipeline

**Categorical Encoding:**

* **Label Encoding**: Applied to binary variables (Gender, AutoRenew, Churn target)
* **One-Hot Encoding**: Implemented for multi-class variables (SubscriptionType, PaymentMethod) using pd.get\_dummies(drop\_first=True)

**Feature Scaling:**

* **StandardScaler**: Applied to numerical features (Age, MonthlyCharges, TenureMonths, NumSupportTickets) achieving mean=0, std=1

**Class Imbalance Handling:**

* **SMOTE**: Created synthetic samples transforming the initial 75:25 churn ratio to a balanced 52:48 distribution (~4,200 samples per class in training set)

### Machine Learning Models

Four algorithms were implemented and compared:

1. **Logistic Regression**: max\_iter=1000, class\_weight='balanced' - Linear baseline model providing interpretable coefficients
2. **Random Forest**: n\_estimators=100 - Ensemble method robust to overfitting with feature importance metrics
3. **XGBoost**: eval\_metric='logloss' - Advanced gradient boosting for complex pattern recognition
4. **LightGBM**: Default configuration - Efficient gradient boosting implementation

**Training Process:** 80/20 train-test split with stratification, SMOTE applied only to training data, models evaluated using accuracy, F1 score, and ROC AUC.

## Results & Business Insights

### Model Performance Comparison

| **Model** | **Accuracy** | **F1 Score** | **ROC AUC** | **Recommendation** |
| --- | --- | --- | --- | --- |
| **Logistic Regression** | 0.51 | **0.60** | 0.54 | **Primary Choice** |
| **Random Forest** | 0.53 | 0.51 | 0.53 | Most Balanced |
| **XGBoost** | 0.51 | 0.50 | 0.51 | Alternative |
| **LightGBM** | 0.52 | 0.53 | 0.52 | Alternative |

**Analysis:** Logistic Regression achieved the highest F1 score (0.60), indicating superior precision-recall balance for churn prediction. Random Forest showed the most consistent performance across metrics. The uniform performance range (accuracy: 0.51-0.53) suggests inherent dataset classification challenges.

### Feature Importance Analysis (Random Forest)

**Primary Predictors:**

1. **TenureMonths** (0.24) - Customer relationship duration is strongest predictor
2. **MonthlyCharges** (0.23) - Pricing directly impacts churn probability
3. **Age** (0.23) - Significant demographic influence
4. **NumSupportTickets** (0.18) - Service quality indicator

**Secondary Factors:** AutoRenew (0.03), Gender (0.02), Subscription/Payment types (0.01-0.02)

### Business Insights & Recommendations

**Key Findings:**

* **Tenure Risk**: Newer customers (shorter tenure) show higher churn propensity
* **Price Sensitivity**: Higher monthly charges correlate with increased churn risk
* **Age Demographics**: Specific age groups demonstrate higher churn rates
* **Service Quality**: Multiple support tickets indicate higher churn risk

**Strategic Recommendations:**

1. **Retention Focus**: Prioritize customers with tenure < 12 months for proactive campaigns
2. **Pricing Strategy**: Review pricing for high-charge customers, implement loyalty discounts
3. **Support Enhancement**: Proactive support for customers with multiple tickets
4. **Demographic Targeting**: Develop age-specific retention programs
5. **Auto-renewal Promotion**: Incentivize auto-renewal adoption

**Model Deployment:** Recommend Logistic Regression for production due to highest F1 score (0.60) and interpretability for business stakeholders.

### Future Improvements

* **Feature Engineering**: Create interaction features (tenure × charges)
* **Advanced Methods**: Experiment with neural networks and deep ensemble techniques
* **Temporal Analysis**: Incorporate customer behavior time-series patterns
* **Cost-sensitive Learning**: Implement business cost considerations in optimization

**Conclusion:** The project successfully demonstrates a complete ML pipeline for churn prediction, achieving moderate but actionable predictive performance while providing clear business insights for customer retention strategies. The balanced approach between model performance and interpretability makes it suitable for business deployment and decision-making.