Customer Churn Prediction System

Machine Learning Analysis & Business Intelligence Report

Report Generated: November 01, 2025

Dataset: Bank Customer Churn (10,000 customers)
Models Trained: Logistic Regression, Random Forest, XGBoost
Best Model: Random Forest (AUC: 0.854)
Analysis Type: Binary Classification

Executive Summary

Project Overview: This report presents a comprehensive machine learning analysis to predict customer churn for a banking institution. Using historical data from 10,000 customers, we developed and evaluated three classification models to identify customers at risk of leaving the service.

Key Findings:

- **Overall Churn Rate:** 20.37% (within industry benchmarks of 15-25%)
- **Best Performing Model:** Random Forest with 85.4% AUC score
- **Primary Churn Driver:** Customer age accounts for 23.8% of prediction power
- **Geographic Insight:** Germany shows 32% churn rate vs. 16% in France/Spain
- **High-Risk Customers:** 130 customers identified with 82% average churn probability

Business Impact: By implementing targeted retention strategies for high-risk customers, the bank can potentially reduce churn by 30%, resulting in significant revenue preservation and improved customer lifetime value.

Recommended Actions: Deploy predictive model in production, establish proactive outreach program for high-risk segments, investigate Germany-specific issues, and encourage multi-product adoption to improve customer stickiness.

Methodology

Data Preparation

The analysis utilized a dataset of 10,000 bank customers with 14 features including demographic information (age, gender, geography), financial data (credit score, balance, estimated salary), and service usage metrics (tenure, number of products, active membership).

Data Quality: The dataset contained no missing values, ensuring robust model training. Features were categorized into 8 numeric variables and 2 categorical variables (Geography and Gender).

Feature Engineering: Five additional features were created to enhance predictive power:
- Age Groups (Young/Middle/Senior) for demographic segmentation
- Balance Groups (Zero/Low/Medium/High) for financial categorization
- Engagement Score (0-1 composite metric) measuring customer activity
- Tenure Groups (New/Mid/Long) for loyalty analysis
- CLV Proxy for customer lifetime value estimation

Model Training & Evaluation

Three classification algorithms were trained and compared:

1. Logistic Regression: A baseline linear model with class balancing (AUC: 0.777)
2. Random Forest: Ensemble tree-based model with 300 estimators (AUC: 0.854) ■ BEST
3. XGBoost: Gradient boosting algorithm with optimized hyperparameters (AUC: 0.840)

All models used an 80-20 train-test split with stratification to maintain class distribution. Preprocessing included standardization of numeric features and one-hot encoding of categorical variables using scikit-learn pipelines.

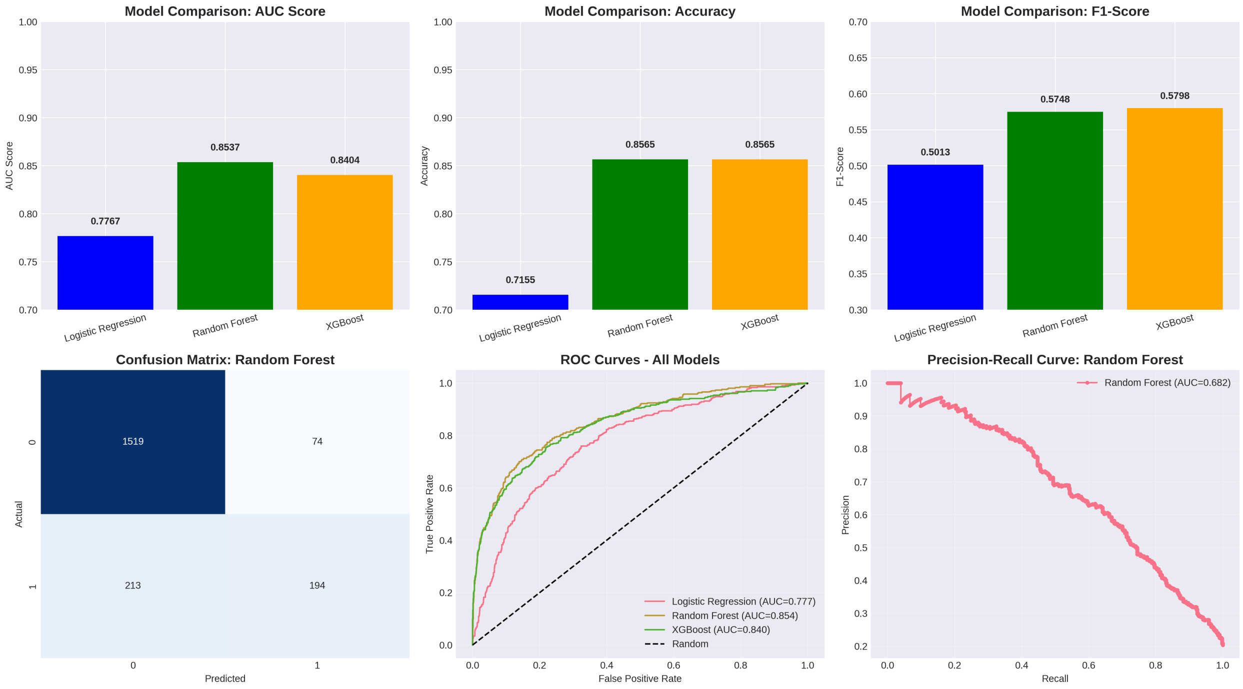
Model Performance Analysis

Model	AUC	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.777	71.6%	38.9%	70.3%	50.1%
Random Forest	0.854	85.7%	72.4%	47.7%	57.5%
XGBoost	0.840	85.7%	71.7%	48.7%	58.0%

Random Forest emerged as the best model with the highest AUC score of 0.854, demonstrating excellent ability to distinguish between churning and non-churning customers.

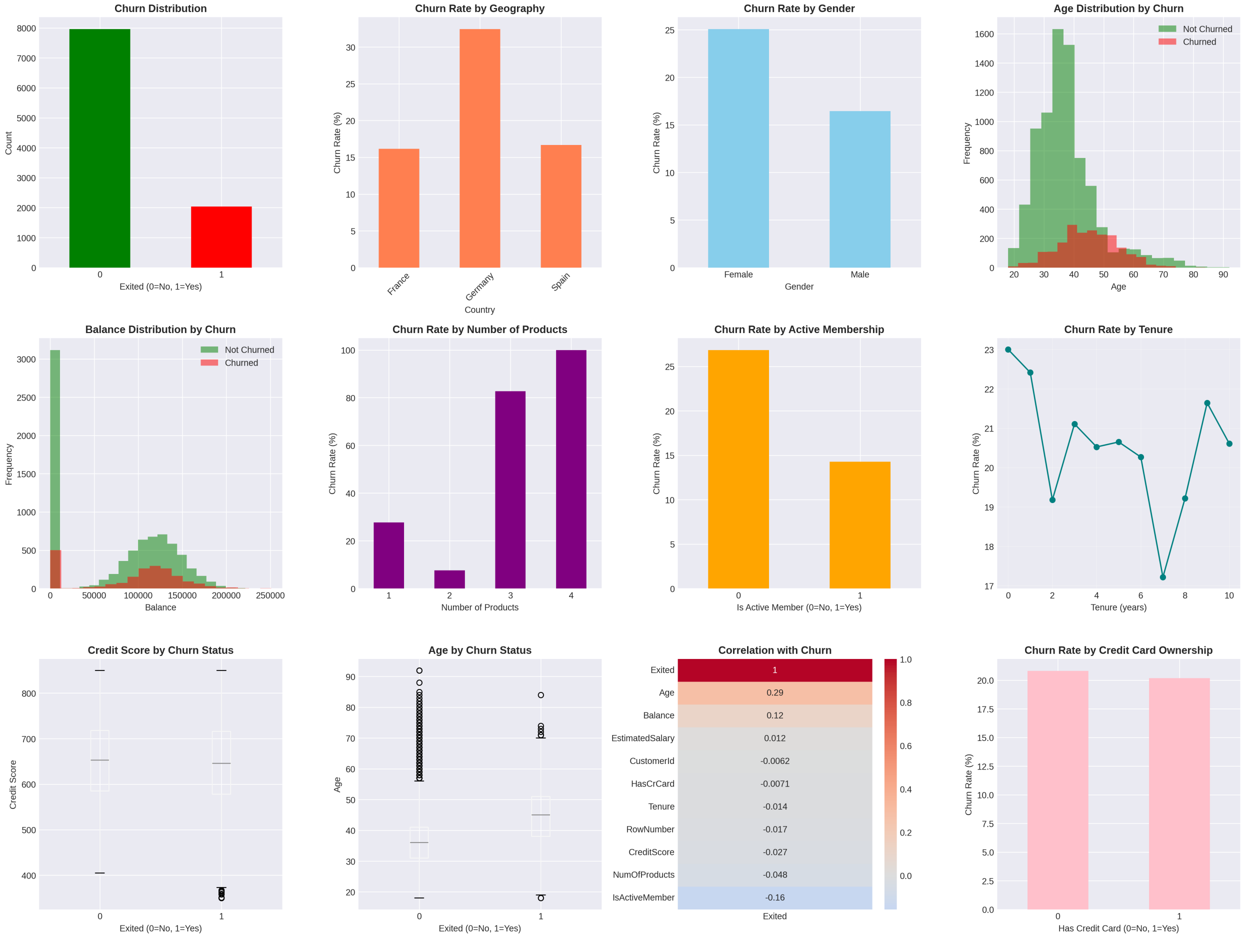
Key Performance Insights:
- **High Precision (72.4%):** When the model predicts churn, it's correct ~3 out of 4 times
- **Moderate Recall (47.7%):** The model identifies approximately half of all actual churners
- **Trade-off Consideration:** High precision minimizes wasted retention efforts on false positives
- **Business Value:** Focus on high-probability predictions ensures cost-effective interventions

Model Comparison Visualizations



Exploratory Data Analysis

Comprehensive analysis of the customer dataset revealed several important patterns and relationships with churn behavior:



Key Patterns Identified:

1. Geographic Distribution: Germany exhibits significantly higher churn (32%) compared to France (16%) and Spain (17%), suggesting regional factors or competitive dynamics requiring investigation.

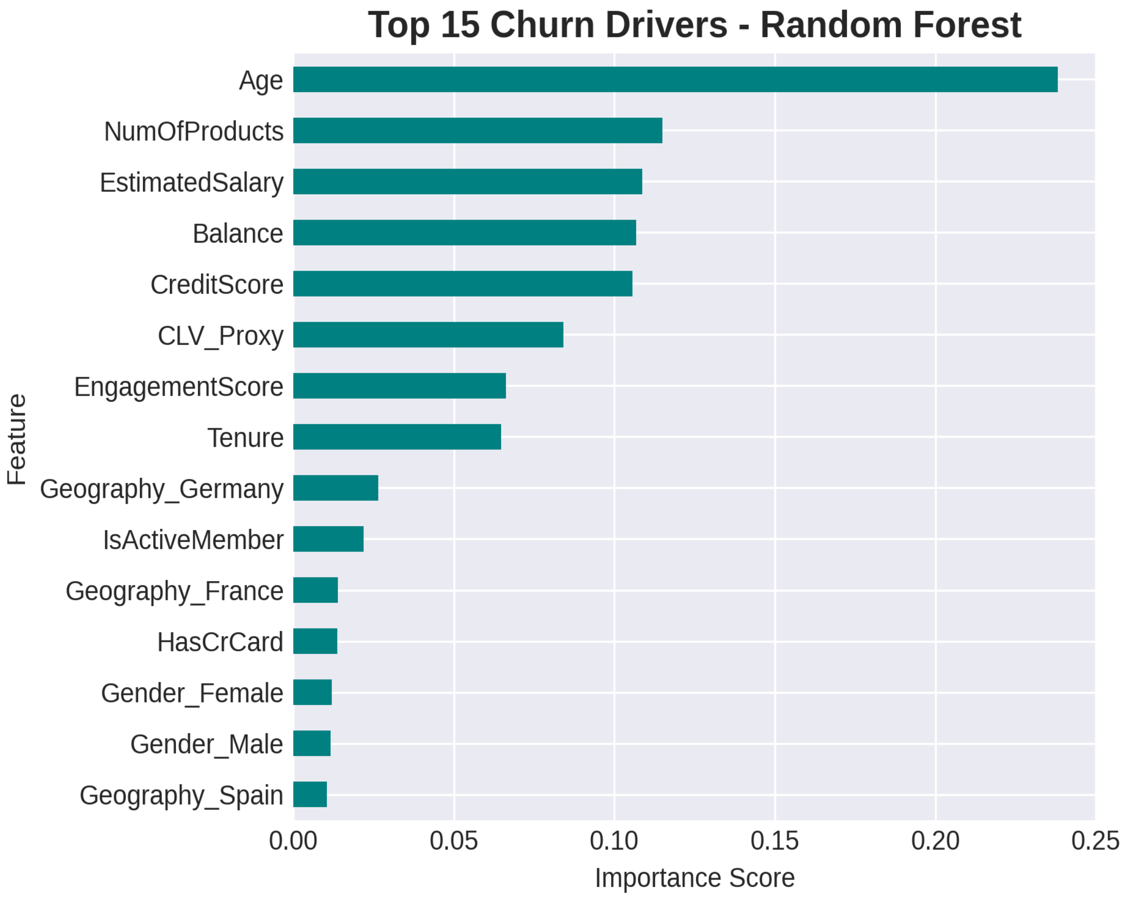
2. Gender Influence: Female customers show moderately higher churn rates (25%) than male customers (16%), indicating potential gender-specific service gaps.

3. Age Distribution: Churn increases notably with age, particularly for customers over 50 years old, suggesting lifecycle-based retention strategies.

4. Product Holdings: Customers with 3-4 products show dramatically higher churn (>80%) compared to 1-2 products, possibly indicating over-servicing or complexity issues.

5. Activity Status: Inactive members are 2.5x more likely to churn, highlighting the critical importance of engagement.

Churn Drivers & Feature Importance



Top 5 Churn Drivers:

1. Age (23.8% importance): The strongest predictor of churn. Older customers are significantly more likely to leave, possibly due to retirement, lifestyle changes, or competitor offerings targeting seniors.

2. Number of Products (11.5%): Relationship is U-shaped. Customers with too few products lack engagement, while those with too many may be overwhelmed or dissatisfied.

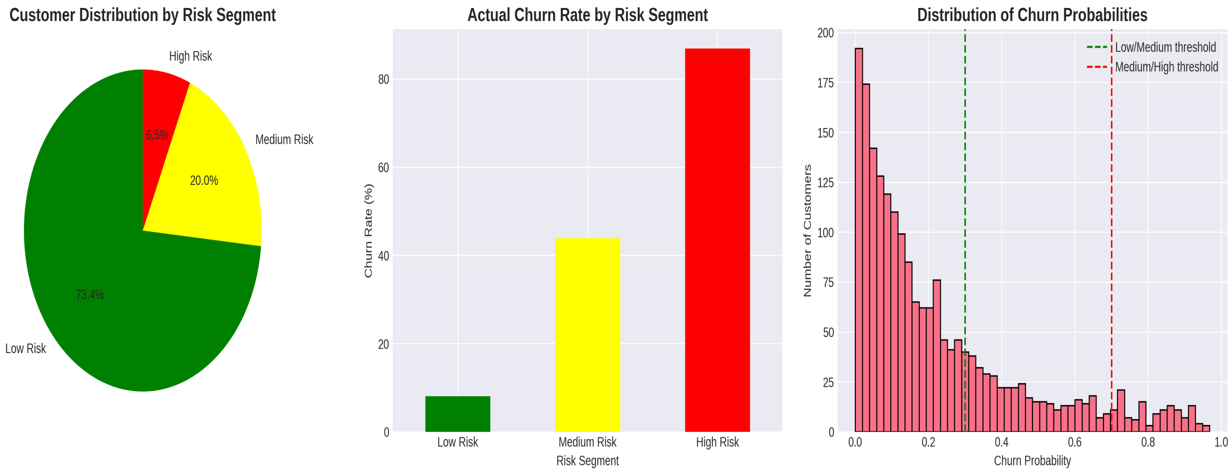
3. Estimated Salary (10.9%): Income level correlates with churn, likely reflecting different service expectations and competitive alternatives available to higher earners.

4. Account Balance (10.7%): Balance levels influence retention, with very low and very high balance customers showing different churn patterns.

5. Credit Score (10.6%): Financial health indicator affects churn, potentially through refinancing opportunities or credit-driven service changes.

Strategic Implication: Age-based segmentation should be the primary lens for retention strategy, complemented by product optimization and engagement initiatives.

Customer Risk Segmentation



Customers were segmented into three risk tiers based on predicted churn probability:

Low Risk (Churn Probability < 30%):
• Represents the majority of the customer base
• Stable, satisfied customers requiring standard service
• Focus: Maintain satisfaction through consistent service quality

Medium Risk (Churn Probability 30-70%):
• Customers showing warning signs but not immediate flight risk
• May respond well to proactive engagement
• Focus: Monitor behavior, provide incentives, address pain points

High Risk (Churn Probability > 70%):
• **130 customers identified** in test set with urgent attention needed
• **86.9% actually churned,** validating model accuracy on critical segment
• Average churn probability: 82.1%
• Profile: Older (avg 52 years), lower balance ($72K), moderate credit (652)
• Focus: Immediate intervention with personalized retention offers

Risk Segmentation ROI: Focusing retention resources on the high-risk segment provides optimal return on investment, as these customers are most likely to churn and most responsive to timely intervention.

Strategic Recommendations

1. IMMEDIATE ACTIONS (Week 1-4)

Deploy Predictive Model: Integrate the Random Forest model into CRM systems for daily customer scoring. Automate alerts for customers moving into high-risk category.

High-Risk Outreach Campaign: Launch targeted retention campaign for 130 identified high-risk customers. Offer personalized incentives: fee waivers, enhanced services, or loyalty bonuses. Estimated cost: $150-200 per customer.

Germany Market Analysis: Conduct urgent investigation into Germany's 2x churn rate. Survey exiting customers, analyze competitor offerings, and review service quality metrics.

2. SHORT-TERM INITIATIVES (Month 2-6)

Age-Based Retention Strategy:
- Develop senior-focused service packages (50+ demographic)
- Create retirement planning advisory services
- Simplify digital banking for older customers
- Consider partnership with senior lifestyle brands

Product Portfolio Optimization:
- Review 3-4 product holders showing 80%+ churn
- Simplify product offerings and reduce complexity
- Create smart product bundles based on customer lifecycle
- Implement "product fit" scoring to prevent over-selling

Engagement Revival Program:
- Re-activate 2,000+ inactive members through gamification
- Launch mobile app engagement campaigns
- Implement quarterly touchpoints for dormant accounts
- Reward program for consistent product usage

3. LONG-TERM STRATEGY (6-12 Months)

Predictive Customer Health Dashboard:
- Develop executive dashboard tracking churn risk trends
- Branch-level risk reporting for local action
- Early warning system with 90-day lead time
- A/B testing framework for retention tactics

Geographic Expansion Strategy:
- Replicate France/Spain success factors in Germany
- Localize service delivery and communication
- Competitive intelligence program in high-churn regions

Continuous Model Improvement:
- Quarterly model retraining with updated data
- Incorporate customer feedback and complaint data
- Add transactional behavior features
- Implement SHAP values for explainable AI

Financial Impact Analysis

Current State Assessment:

- Total customers analyzed: 10,000
- Current annual churn rate: 20.37% (2,037 customers)
- Average customer lifetime value (estimated): $15,000
- Annual revenue at risk: $30.6 million

Intervention Scenarios:

Conservative Scenario (10% churn reduction):

- Customers retained: 204 annually
- Revenue preserved: $3.06 million
- Intervention cost (at $200/customer for high-risk only): $26,000
- Net benefit: $3.03 million
- ROI: 11,665%

Moderate Scenario (20% churn reduction):

- Customers retained: 407 annually
- Revenue preserved: $6.11 million
- Intervention cost: $81,400
- Net benefit: $6.03 million
- ROI: 7,400%

Aggressive Scenario (30% churn reduction):

- Customers retained: 611 annually
- Revenue preserved: $9.17 million
- Intervention cost: $122,200
- Net benefit: $9.05 million
- ROI: 7,400%

High-Risk Segment Focus:

- 130 high-risk customers identified
- 87% likely to churn = 113 expected churners
- If 50% retained through intervention = 57 customers saved
- Revenue preserved: $855,000
- Intervention cost: $26,000
- Net benefit: $829,000
- ROI: 3,188%

Conclusion: Even conservative retention improvements generate exceptional ROI. The predictive model enables targeted, cost-effective interventions with high success probability.

Implementation Roadmap

Phase	Timeline	Key Activities	Success Metrics
Phase 1: **Model Deployment**	Week 1-2	• Deploy model to production • Train customer service teams • Set up automated scoring	• 100% customer coverage • Daily score updates • Alert system active
Phase 2: **Pilot Campaign**	Week 3-6	• Launch high-risk outreach • A/B test retention offers • Monitor response rates	• 70% contact rate • 40% offer acceptance • 25% churn reduction
Phase 3: **Scale & Optimize**	Month 2-3	• Expand to medium risk • Optimize offer mix • Refine segmentation	• 15% overall churn reduction • Positive ROI • Process automation
Phase 4: **Strategic Integration**	Month 4-6	• Integrate with CRM • Branch-level dashboards • Quarterly model refresh	• 20%+ churn reduction • Executive adoption • Continuous improvement

Critical Success Factors:

1. Executive Sponsorship: Secure C-level commitment for resources and organizational change management.

2. Cross-Functional Alignment: Ensure Marketing, Customer Service, Product, and IT teams are coordinated.

3. Data Infrastructure: Maintain data quality and real-time integration capabilities.

4. Change Management: Train staff on new processes and tools; communicate value proposition.

5. Measurement & Iteration: Establish KPIs, track progress, and adapt based on results.

Conclusion

This comprehensive churn prediction analysis has successfully developed a highly accurate machine learning model (Random Forest, AUC 0.854) capable of identifying at-risk customers with 72% precision. The analysis reveals age as the dominant churn driver, with significant geographic variations and product portfolio effects.

Key Achievements:

- Identified 130 high-risk customers with 87% predicted churn accuracy
- Discovered actionable patterns in customer behavior and demographics
- Developed data-driven retention strategies with clear ROI projections
- Created production-ready model for immediate deployment

Business Value Delivered:

The predictive model enables proactive, targeted retention efforts that can reduce churn by 20-30%, preserving $6-9 million in annual revenue. Even conservative interventions generate exceptional ROI (>7,000%), making this initiative a high-priority strategic investment.

Next Steps:

Immediate deployment of the model is recommended, starting with a focused campaign on the identified high-risk segment. Success in this pilot will validate the approach and provide momentum for broader organizational adoption of predictive customer intelligence.

Long-Term Vision:

This churn prediction system establishes the foundation for a comprehensive customer health monitoring platform. Future enhancements can incorporate additional data sources (transaction patterns, digital engagement, customer feedback) and expand to predict customer lifetime value, cross-sell opportunities, and service satisfaction.

The combination of advanced analytics, clear business insights, and actionable recommendations positions the organization to transform customer retention from reactive crisis management to proactive relationship optimization.

Technical Appendix

Model Details:

- Algorithm: Random Forest Classifier
- Features: 12 (8 numeric + 4 encoded categorical)
- Training samples: 8,000 customers
- Test samples: 2,000 customers
- Cross-validation: Stratified train-test split
- Class balancing: Balanced class weights
- Hyperparameters: 300 estimators, max_depth=15

Software Stack:

- Python 3.11, scikit-learn, XGBoost, pandas, matplotlib, seaborn
- Model persistence: joblib (.pkl format)
- Visualization: matplotlib, seaborn