UNIVERSITY OF ESSEX

DEPARTMENT OF MATHEMATICAL SCIENCES

MACHINE LEARNING APPROACHES AND INTERPRETABILITY ANALYSIS FOR BANKING CUSTOMER CHURN PREDICTION

By: Nakarin Kungvalchokchai (2400952)

MAJOR:DATA SCIENCES AND ITS APPLICATION

OVERVIEW

- Introduction & Problem Statement
- Research Questions
- Methodology Overview
- Key Findings: Model Performance
- Key Findings: Why Customers Churn (SHAP Analysis)
- Business Implications: Customer Segmentation
- Conclusion & Recommendations

What is Customer Churn?

Definition

Customer Churn is the phenomenon where customers stop doing business with a company or service provider over a given period of time.

In Banking Context

- Customer closes their account or stops using banking services
- Switches to a competitor bank

Why is Customer Churn a Critical Problem?

- Acquiring a new customer is 5-7x more expensive than retaining an existing one.
- Digital banking enables easy switching between competitors
- Traditional retention strategies becoming less effective
- **Challenge:** Banks need to **predict** who will leave and **understand why** to act effectively.

Research Gap

- Most studies focus only on accuracy, not explainability
- Limited systematic comparison of ML algorithms for banking data
- SHAP analysis rarely applied to banking churn prediction

Research Questions

- RQ1: Effectiveness

Can machine learning accurately predict churn?

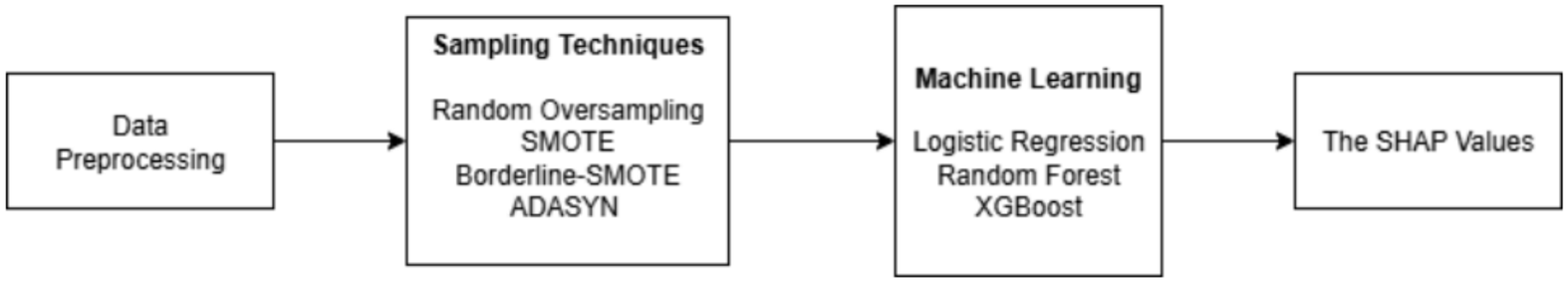
- RQ2: Comparative Performance

Which algorithm performs best? (LR, Random Forest, XGBoost)

- RQ3: Key Predictive Factors

What are the most influential factors driving churn?

Methodology Overview



- Data: Kaggle Bank Churn Dataset (165,034 customers)
- Preprocessing: Cleaned data, handled categorical variables, addressed severe class imbalance (21% churn vs. 79% stay).
- Sampling Techniques: Used 4 techniques(Random Oversampling, SMOTE, Borderline-SMOTE,ADASYN) to balance the data.
- Hyperparameter Tuning: Utilised GridSearchCV to optimise hyperparameters for each model (e.g., C for Logistic Regression, max_depth and n_estimators for Random Forest and XGBoost) to maximise performance and prevent overfitting.
- Models Trained: Logistic Regression (baseline), Random Forest, XGBoost.
- Analysis: Evaluated performance (Accuracy, Precision, Recall, F1-Score, AUC) and interpreted models using SHAP values.
- Using Python with libraries (pandas, numpy, scikit-learn, imblearn, xgboost, shap)

Customer churn data structure

Column Name	Description	Data Type	Example Values
ID	Unique identifier for each row	Integer	1, 2
Customer ID	Unique identifier for each customer	Integer	15674932, 15694510
Surname	Customer's last name	String (Text)	"Smith", "Johnson"
Credit Score	Numerical value representing creditworthiness	Integer	645, 712, 598
Geography	Country where the customer resides	String (Categorical)	"France", "Germany"
Gender	Customer's gender	String (Binary)	"Male", "Female"
Age	Customer's age in years	Integer	32, 45, 28
Tenure	Number of years with the bank	Integer	5, 3, 2
Balance	Current account balance (in currency)	Float/Decimal	125000.00, 0.00
NumOfProducts	Number of bank products used	Integer	1, 2, 3
HasCrCard	Whether the customer has a credit card (1 = yes, 0 = no)	Binary (0/1)	1, 0
IsActiveMember	Indicates active customer engagement, defined as interactions (e.g., transactions, account logins, or usage of core banking services) (1 = yes, 0 = no)	Binary (0/1)	1, 0
EstimatedSalary	Approximate annual salary (in currency)	Float/Decimal	85000.00, 65000.50
Exited	Whether the customer has churned (1 = yes, 0 = no)	Binary (0/1)	1, 0

- 13 features: demographic, financial, behavioural

Data preprocessing & feature engineering

1. Removed: Non-predictive columns (IDs, Surname)
2. Encoded: Categorical variables (Geography, Gender) into binary features
3. Scaled: Numerical features using Min-Max normalisation (CreditScore, Age, Tenure, Balance, EstimatedSalary, NumOfProducts)
4. Split: Dataset into 80% training and 20% testing sets

Handling Class Imbalance

Four Sampling Techniques Applied:

- Random Oversampling - Duplicates minority class samples
- SMOTE - Generates synthetic samples between neighbours
- Borderline-SMOTE - Focuses on boundary cases
- ADASYN - Adaptive synthetic sampling for difficult regions

Three Algorithms Evaluated

1. Logistic Regression
   - White box interpretable model
   - Linear relationship assumptions
   - Fast training, regulatory-friendly
2. Random Forest
   - Ensemble of decision trees
   - Captures non-linear relationships and feature interactions
   - Less interpretable than linear models, but explainable with feature importance or SHAP
3. XGBoost
   - Advanced gradient boosting
   - Strong performance, especially on imbalanced or complex datasets
   - A black-box model requires post-hoc explainability methods like SHAP

Classification Metrics

Confusion matrix

- True Positives(TP): correctly predicted positives
- False Positives(FP): predicted positive but actually negative
- False Negatives(FN): predicted negative but actually positive
- True Negatives(TN): correctly predicted negatives

Metrics

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
  - The proportion of all correct predictions, but it can be misleading with imbalanced data.
- Precision = TP / (TP + FP)
  - How many of our positive predictions were correct.
- Recall = TP / (TP + FN)
  - How many of actual positive cases we found.
- F1-score = 2 × (Precision × Recall) / (Precision + Recall)
  - Balances precision and recall in one number.
- AUC (Area Under ROC Curve) = Measures ability to separate classes (0.5 = random, 1 = perfect)

Results & Performance Analysis

Sampling Technique	Machine Learning	Accuracy	Precision	Recall	F1	AUC	Running time (Sec)
Without Sampling Technique	Logistic Regression	0.833	0.695	0.379	0.491	0.814	26.35
	Random Forest	0.863	0.765	0.509	0.611	0.888	2169.88
	XGBoost	0.866	0.748	0.550	0.634	0.890	305.81
Random Oversampling	Logistic Regression	0.754	0.451	0.734	0.559	0.816	40.67
	Random Forest	0.851	0.663	0.603	0.632	0.872	3405.48
	XGBoost	0.822	0.558	0.756	0.642	0.881	459.8
SMOTE	Logistic Regression	0.757	0.453	0.733	0.560	0.816	43.32
	Random Forest	0.846	0.637	0.635	0.636	0.872	4561.51
	XGBoost	0.858	0.677	0.632	0.654	0.884	437.6
BorderlineSMOTE	Logistic Regression	0.741	0.436	0.759	0.554	0.811	38.44
	Random Forest	0.842	0.618	0.663	0.640	0.872	4719.83
	XGBoost	0.851	0.642	0.664	0.653	0.886	434.76
ADASYN	Logistic Regression	0.740	0.435	0.761	0.553	0.813	35.66
	Random Forest	0.842	0.620	0.649	0.634	0.870	4524.56
	XGBoost	0.857	0.676	0.620	0.647	0.883	420.3

Best Performing Model: XGBoost

- **Best F1-Score: 0.654** (with SMOTE)
- **Best AUC: 0.89**
- Consistently outperformed Logistic Regression and Random Forest

Impact of Sampling Techniques

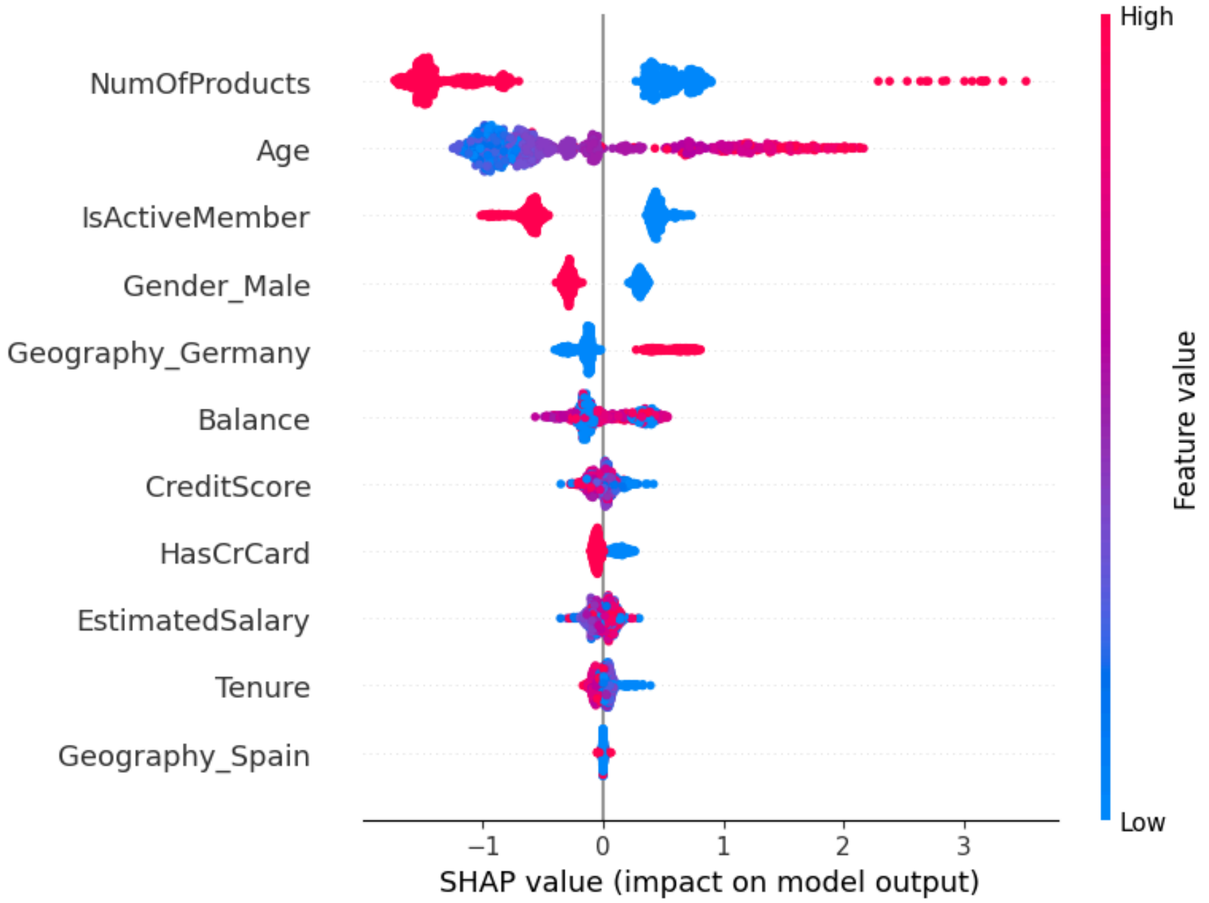
- Sampling methods (SMOTE, etc.) significantly improved F1-scores across all models.
- They effectively addressed the class imbalance problem.

Training Time Comparison

- **Logistic Regression:** 26-43 sec
  - Fastest, ideal for real-time use
- **XGBoost:** 306-460 sec
  - Best performance/speed balance
- **Random Forest:** 2,170-4,720 sec
  - Too slow for practical deployment

✅ Conclusion: XGBoost + SMOTE is the recommended optimal solution.

Feature Importance - SHAP Analysis

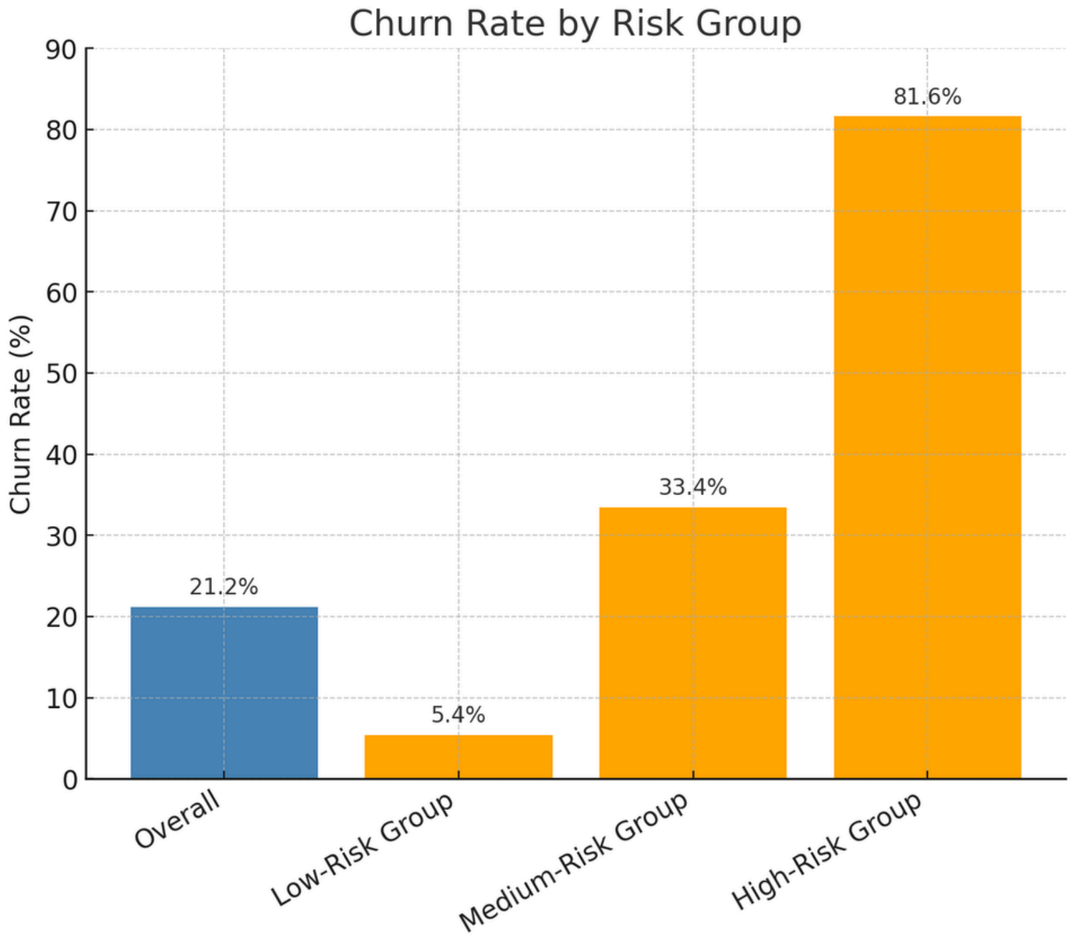


XGBoost

Top five key features

- Number of Products (⬇️ Fewer products = ⬆️ Churn risk)
- Age (⬆️ Older age = ⬆️ Churn risk)
- IsActive Member (⬇️ Inactive = ⬆️ Churn risk)
- Gender: Male (⬇️ Less likely to churn than female)
- Geography: Germany (⬆️ German customers = ⬆️ Churn risk)

Customer Risk Segmentation



Three Risk Segments Identified:

High-Risk Segment (81.6% churn rate)

- Female customers in Germany
- Age > 37 years
- Inactive members
- Only 1 product

Medium-Risk Segment (33.4% churn rate)

- Active customers > 37 years
- Single product holders

Low-Risk Segment (5.4% churn rate)

- Customers < 37 years
- Active members

Business Implications

Targeted Retention Strategies Based on Risk Profile

For High-Risk Customers:

- Action: Proactive, personalised outreach.
- Strategy: Offer product bundles, loyalty programs, or dedicated support to increase engagement and the number of products customers use.

For Medium-Risk Customers:

- Action: Engagement campaigns.
- Strategy: Encourage the use of online banking features, and notify customers of new products.

Limitations & Future Research

Study Limitations:

* Single dataset - generalisability needs validation
* Limited to 3 algorithms - could explore deep learning
* Time-point analysis - longitudinal studies needed
* European focus - other regions may differ

Future Research Directions:

* Real-time prediction systems implementation
* Dynamic model updates as customer behaviour changes
* Deep learning approaches for complex pattern recognition
* Multi-bank validation across different institutions

Conclusions & Recommendations

Research Questions Answered:

RQ1: ✅ ML algorithms effectively predict banking churn (up to 89% AUC)
RQ2: ✅ XGBoost with SMOTE achieves optimal performance (F1: 0.654, AUC: 0.884)
RQ3: ✅ Five consistent predictors identified across all models (Top 5 factors are: Products, Age, Activity, Location, Gender)

Practical Recommendations for Banks:

- XGBoost + SMOTE = Optimal solution for banking churn prediction
- Use SHAP analysis continuously to monitor and explain model decisions.
- Focus retention resources on the high-risk customer segments identified.
- Develop strategies that specifically target the key drivers: boost product adoption and customer activity.

UNIVERSITY OF ESSEX

DEPARTMENT OF MATHEMATICAL SCIENCES

THANK YOU

For your attention

MAJOR:DATA SCIENCES AND ITS APPLICATION