Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

Project Title:

A predictive analytics study of ABC Bank uses machine learning to forecast banking customer churn.

Introduction

The banking sector encounters severe challenges because of customer churn problems. Modern banks must prioritize customer retention since competition grows and customer demands shift. Current market saturation makes it difficult for banks to lose existing customers because this results in both monetary loss and damage to their reputation, together with higher costs of attracting new customers (Reichheld and Kenny,1990). Through predictive modeling with machine learning, banks can identify customers who might leave to implement prevention strategies for customer retention. A predictive model for ABC Bank customer churn prediction will be created through this project using a model-based system that produces useful business findings and supports data-driven choices.

Problem Statement

ABC Bank has suffered significant revenue loss due to its reactive approach to customer retention. Traditional methods, such as generic marketing campaigns and standard customer service protocols, fail to identify early indicators of dissatisfaction. Furthermore, the lack of real-time predictive capability prevents timely intervention. To address this, the bank requires a model that not only forecasts churn with high accuracy but also provides interpretability for actionable business decisions.

Methodology and Implementation

Data Acquisition

The project utilized the publicly available "Bank Customer Churn" dataset from Kaggle, which includes 10,000 customer records with demographic, behavioral, and transactional attributes. Data cleaning confirmed the absence of missing values. Categorical variables (e.g., Gender, Geography) were encoded using One-Hot and Label Encoding techniques.

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

Numerical variables were standardized to facilitate convergence and ensure comparability across models.

Exploratory Data Analysis (EDA)

Descriptive statistics, visualizations (Appendix A.1), and correlation analysis revealed strong relationships between churn and features such as Age, Account Balance, and Product Holding (Appendix A.7). A class imbalance (20% churners vs. 80% non-churners) necessitated the Synthetic Minority Over-sampling Technique (SMOTE) to rebalance the training data (Appendix A.2 and A.6). A new feature, *Balance per Product*, was engineered to capture nuanced spending behavior that may indicate churn intent.

Model Selection and Training

A range of models we implemented to explore both linear and non-linear relationships:

- Logistic Regression served as a baseline for interpretability and benchmarking.
- Random Forest Classifier was selected for its robustness to overfitting, ability to handle non-linearities, and capacity to model feature interactions.
- Multilayer Perceptron (MLP) neural network with two hidden layers was used to model complex, non-linear decision boundaries.
- K-Means Clustering (unsupervised) was applied post hoc to discover natural customer segments for targeted interventions.

The dataset was split 80/20 into training and test sets. To avoid overfitting, hyperparameters for Random Forest and MLP were tuned using grid search and stratified cross-validation.

Model Evaluation

The evaluation of models depended on Accuracy together with Precision, Recall, F1-Score and AUC-ROC metrics to achieve full performance assessment. The initial model evaluation used frequency analysis with Logistic Regression before applying Random Forest to resampled data to achieve the best results.

The comparison of model performances used visual tools to present ROC curves and confusion matrices (Appendix A.3).

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

Results and Analysis

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	76%	80%	73%	76%	77%
Random Forest	**85%**	**90%**	**91%**	**91%**	**85%**
MLP Neural Network	82%	88%	90%	89%	84%

Confusion Matrices

The Random Forest model demonstrated superior performance across nearly all metrics, as shown in the confusion matrix above. Notably, it exhibited the highest recall, critical in churn prediction scenarios, failing to identify at-risk customers has high business costs. While slightly trailing the Random Forest in overall accuracy, the MLP model exhibited strong performance in precision and recall, suggesting its effectiveness in identifying churn patterns. However, the increased computational complexity and training time associated with neural networks should be considered when deploying in real-time environments (Nai et al., 2023).

Serving as a baseline, the Logistic Regression model underperformed relative to the other models, particularly in recall and F1-score. This suggests limitations in its ability to capture non-linear relationships and interactions among features, which are crucial in accurately predicting churn (Kruschel et al., 2025).

Clustering Insights

K-Means clustering divided the customer base into three meaningful segments: Loyal, At-Risk, and New/Uncertain. This segmentation, visualized in Appendix A.5, allowed for customized intervention strategies, such as exclusive offers for at-risk customers or onboarding support for new clients. While this unsupervised approach did not directly enhance predictive accuracy, it provided valuable insights into customer segments, enabling more targeted retention strategies.

Model Interpretability

SHAP (values were employed to interpret model outputs globally and locally (Appendix A.4). Age emerged as the strongest predictor, followed by the Number of Products and Balance. Surprisingly, Credit Scores had limited predictive value, which is contrary to common banking assumptions. When paired with low product engagement, SHAP beeswarm plots

revealed that high balances were associated with churn, which could be a potential signal of customer dissatisfaction or unmet needs.

Demographic analysis revealed that 46.1% of customers in Germany, 33% of females, and 16% of active members were more likely to churn, indicating the need for region and gender-specific retention strategies.

Business Impact

The predictive model enables ABC Bank to transition from reactive to proactive customer management. By accurately identifying high-risk customers, the bank can deploy targeted retention strategies such as:

- Personalized loyalty programs based on customer segment.
- Automated churn alerts to customer success teams.
- Customized product bundling and financial planning offers.

Reducing churn by even 5% could significantly improve Customer Lifetime Value (CLV) and net profits. Furthermore, the model sets a foundation for real-time predictive systems and personalized marketing platforms, positioning ABC Bank as a data-driven, customer-centric institution.

Limitations

- The Kaggle dataset functions as a substitute, but it might not accurately demonstrate the complete behavioral patterns that exist in ABC Bank's actual customer base.
- The model may retain historical biases that shape operational and social elements, impacting its predictive results.
- The use of static models becomes obsolete when customer preferences and behaviors change.

These models need periodic retraining and validation to remain effective.

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

Future Work

To address limitations and expand impact, the following enhancements are recommended:

- Data Integration: Collaborate with ABC Bank's IT department to access real transaction logs and behavioral data, such as web clicks and support tickets.
- Time-Series Modeling: Incorporate sequential models such as Long-Short Term Memory to understand time-dependent churn triggers.
- Real-Time Systems: Implement a Machine Learning Operations pipeline for continuous monitoring, model retraining, and drift detection.
- Multimodal Data Fusion: Combine structured bank data with unstructured sources such as call transcripts or sentiment analysis for deeper behavioral insight.

Conclusion

This study demonstrates the viability of machine learning for customer churn prediction in the banking sector. By leveraging interpretable models like Random Forest and explainability tools like SHAP, it becomes possible to deliver predictive power and actionable insights. The strategic combination of supervised and unsupervised learning further enhances business value, offering a robust blueprint for customer-centric innovation.

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

References

Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M. and Zschech, P., (2025). Challenging the Performance-Interpretability Trade-off: An Evaluation of Interpretable Machine Learning Models. *Business & Information Systems Engineering*, pp.1-25.

Nair, N., Patel, M., Gupta, S. and Singh, P., (2023). Leveraging Neural Networks and Random Forest Algorithms for Enhanced Predictive Customer Behavior Analytics. *International Journal of AI Advancements*, *12*(8).

Reichheld, F.F. and Kenny, D.W., (1990). The hidden advantages of customer retention. *Journal of Retail Banking*, *12*(4), pp.19-24.

The Bank Customer Churn Dataset exists on the Kaggle platform. Retrieved from https://www.kaggle.com/

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

Appendix

A.1 Dataset Summary Table

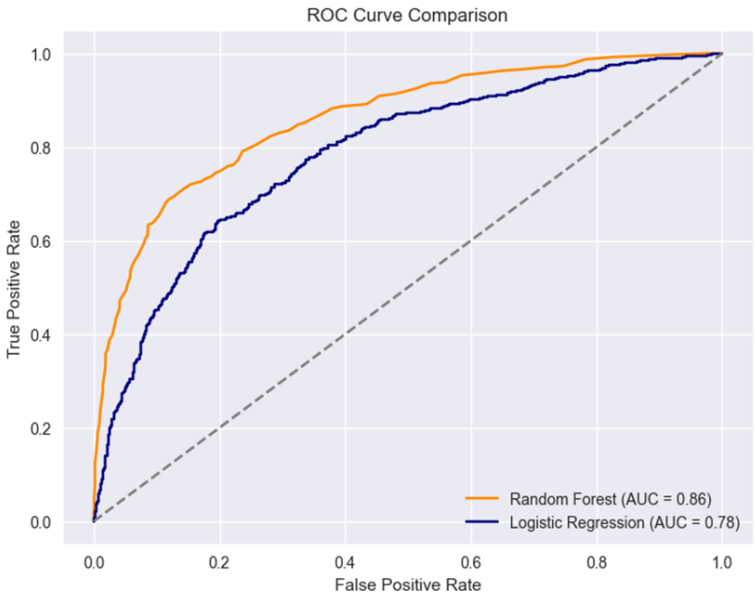
	customer_id	credit_score	age	tenure	balance	products_number	credit_card	active_member	estimated_salary	churn
count	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

A.2 SMOTE

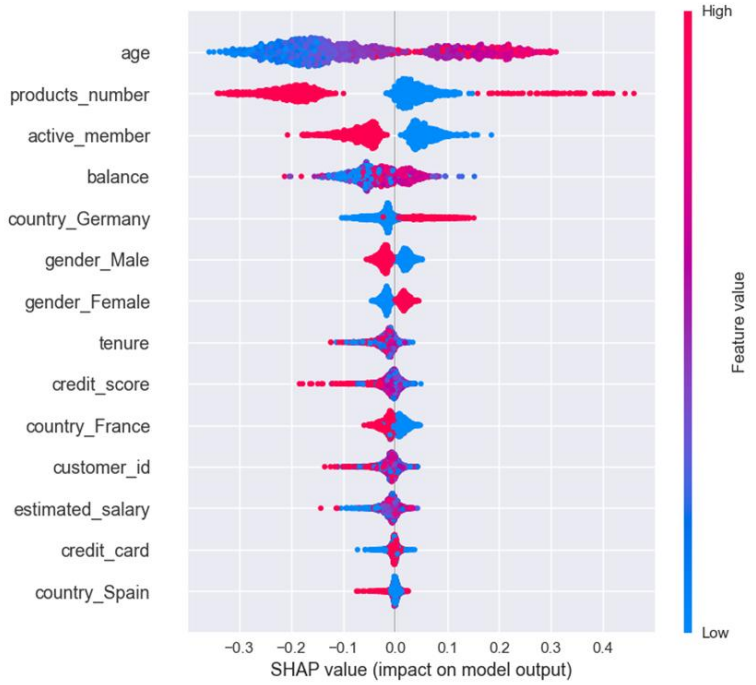
```
# Initialize SMOTE
sm = SMOTE(
    sampling_strategy='auto',  # Oversample only the minority class
    random_state=0,            # Reproducibility
    k_neighbors=5              # Number of neighbors used to create synthetic samples
)

# Fit and resample the training data
X_res, y_res = sm.fit_resample(X_train_sc, y_train)
```

A.3 ROC Curve

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025

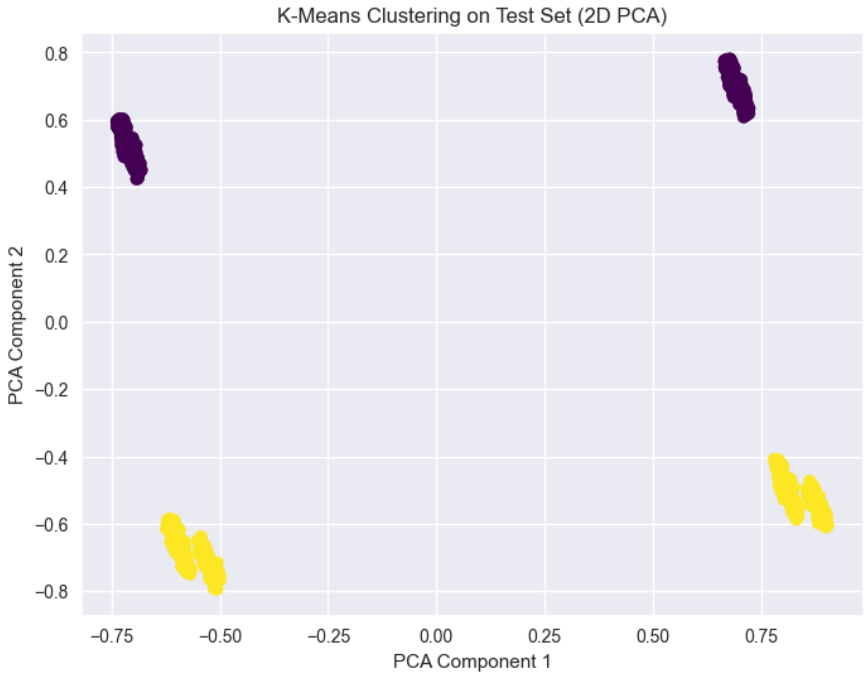


A.4 SHAP Summary Plots

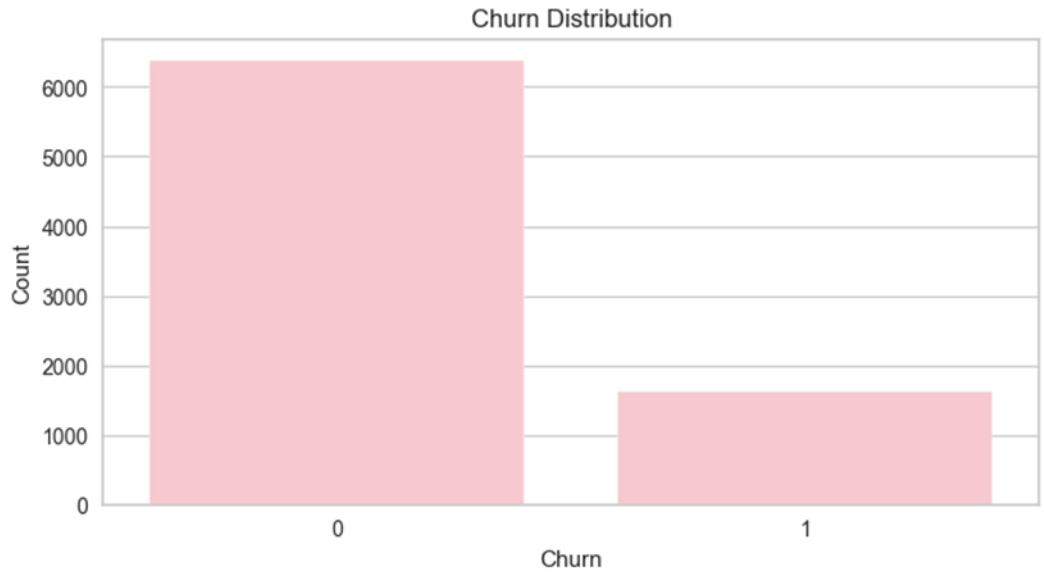


A.5 K-Means Clustering

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025



A.6 Class Distribution



A. 7

Dingolo Chikomborero
Kamoto Shingai
May 3, 2025