Retail Bank Customer Churn Prediction

Consultant-Grade Predictive Analytics Report

(Hybrid Project – Business Sector)

Prepared by: Mo Mo Ko Win Myint, global-ad-snap

Medical AI & Healthcare Data Science Consultant

(Physician | Data Science & Explainable AI)

Date: December 2025

1. Executive Summary

1.1 Project Overview

This project delivers a **production-ready customer churn prediction system** designed to support proactive retention strategies in regulated, service-oriented industries (e.g., banking, insurance, healthcare-adjacent services).
The work is framed and executed from the perspective of a **Medical AI & Healthcare Data Science Consultant (Hybrid Project – Business Sector)**, emphasizing **interpretability, governance, and business actionability**.

Using historical customer behavior and account-level attributes, the objective was to **identify customers at elevated risk of churn with sufficient lead time for intervention**, while maintaining transparency appropriate for executive decision-making and regulatory scrutiny.

1.2 At-a-Glance Summary (Executive Box)

Item	Description
Business Objective	Early identification of customers at high churn risk
Dataset Size	~10,000 customers
Observed Churn Rate	~20%
Final Model	Class-Weighted Logistic Regression
Key Metric Focus	Recall (Churn)
Test ROC-AUC	~0.84
Test Recall (Churn)	~0.74

1.3 Business Problem

Customer churn represents a **direct revenue and continuity risk**. Traditional reactive approaches identify churn only after customer exit, limiting the effectiveness of retention efforts.

The business goals were to:

- Detect churn risk **before exit occurs**
- Prioritize customers for **targeted retention actions**
- Balance **predictive performance** with **model interpretability**
- Ensure the solution is **deployable, auditable, and defensible** in real-world operations

From a cost perspective, **false negatives (missed churners)** are substantially more expensive than **false positives (unnecessary outreach)**. Consequently, the modeling strategy intentionally prioritizes **churn recall over raw accuracy**, aligning technical decisions with real-world retention economics.

1.4 Business Value

The delivered solution enables the organization to:

- Proactively target high-risk customers
- Allocate retention resources efficiently
- Support executive decisions with **transparent, explainable risk drivers**
- Establish a foundation for advanced extensions (uplift modeling, personalization)

2. Data & Analytical Approach

2.1 Dataset Overview

- Records: ~10,000 customers
- Target variable: **Exited (Churn)**
- Feature categories:
    - Demographics (age, geography, gender)
    - Engagement behavior (activity status, tenure)
    - Product ownership
    - Financial indicators (balance, salary, credit score)

The observed churn rate (~20%) confirms a **moderately imbalanced classification problem**, necessitating metrics beyond accuracy.

2.2 Governance-Ready Pipeline

A **leakage-safe, end-to-end pipeline** was implemented:

- Exploratory Data Analysis (EDA)
- Business-driven feature engineering
- Train–test split with stratification
- Feature scaling applied to training data only
- Class imbalance handling (class weighting and SMOTE on training data only)
- Model benchmarking and explainability

All preprocessing decisions were designed to support **reproducibility, auditability, and regulatory review**.

3. Exploratory Data Analysis (EDA): Business Insights

This section focuses on **decision-relevant insights**, not statistical exhaustiveness.

Figures Included

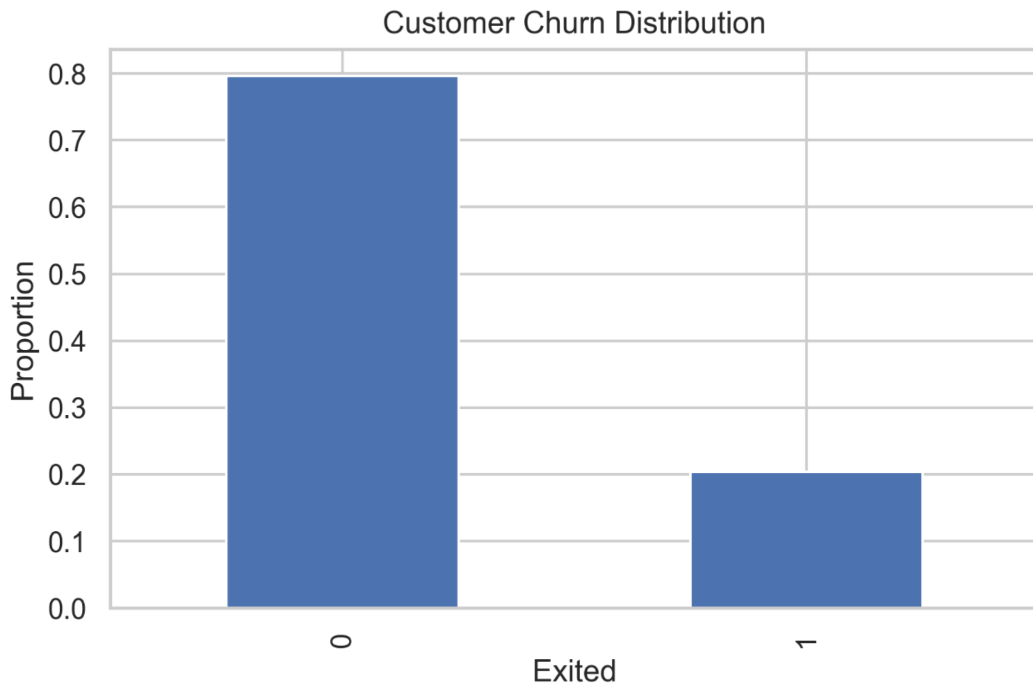
- Figure 1: Churn Distribution
- Figure 2: Age Distribution by Churn Status
- Figure 3: Churn by Activity Status
- Figure 4: Churn by Number of Products
- Figure 5: Feature Correlation Heatmap
- Figure 6: SHAP Global Feature Importance
- Figure 7: SHAP Beeswarm Plot

3.1 Churn Distribution

Figure 1: Churn Distribution

- Churn rate ≈ 20%
- Confirms the need to **prioritize recall for churners**

- Accuracy alone would be misleading in this context

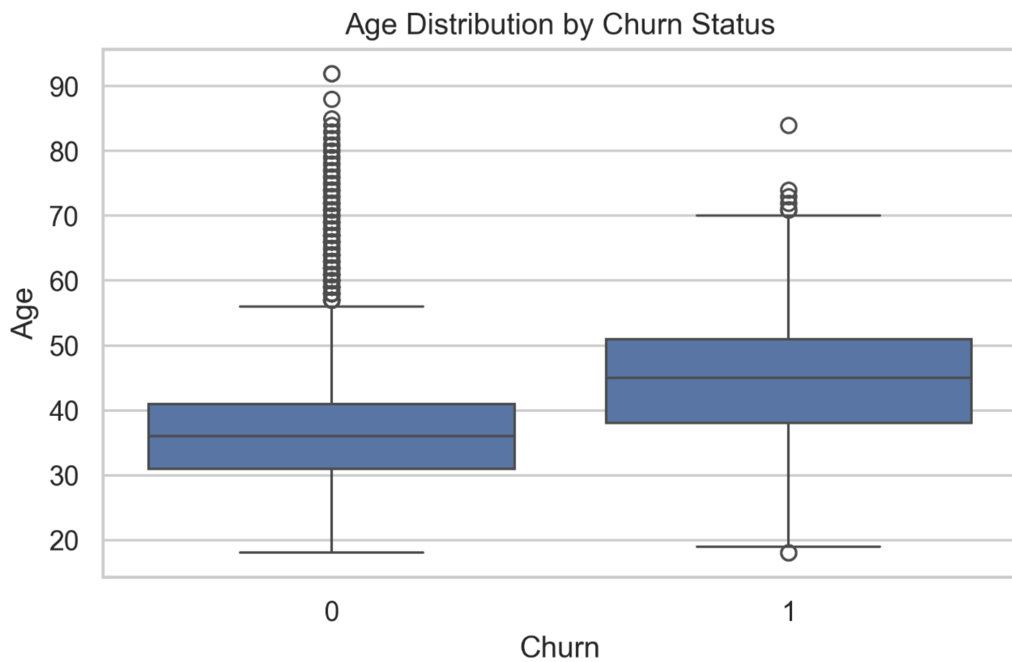


3.2 Demographic & Lifecycle Effects

Figure 2: Age Distribution by Churn Status

- Age shows a **positive relationship with churn risk**
- Older customers exhibit higher exit propensity
- Consistent with lifecycle transitions and changing service needs

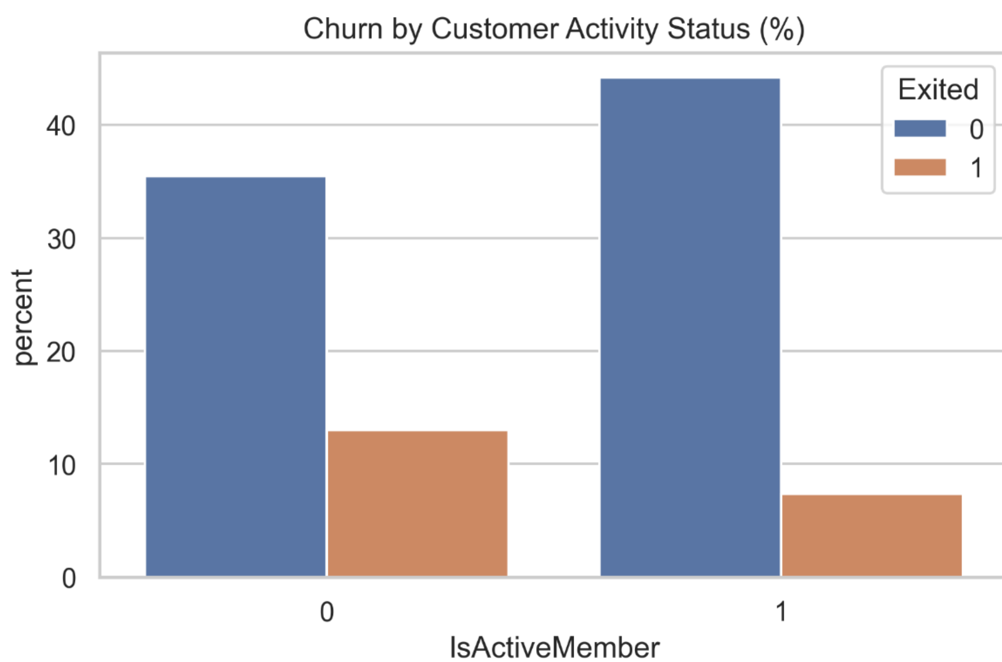
Tenure alone shows weak linear association, but becomes informative when combined with engagement and product variables.



3.3 Engagement & Behavioral Signals

Figure 3: Churn by Activity Status

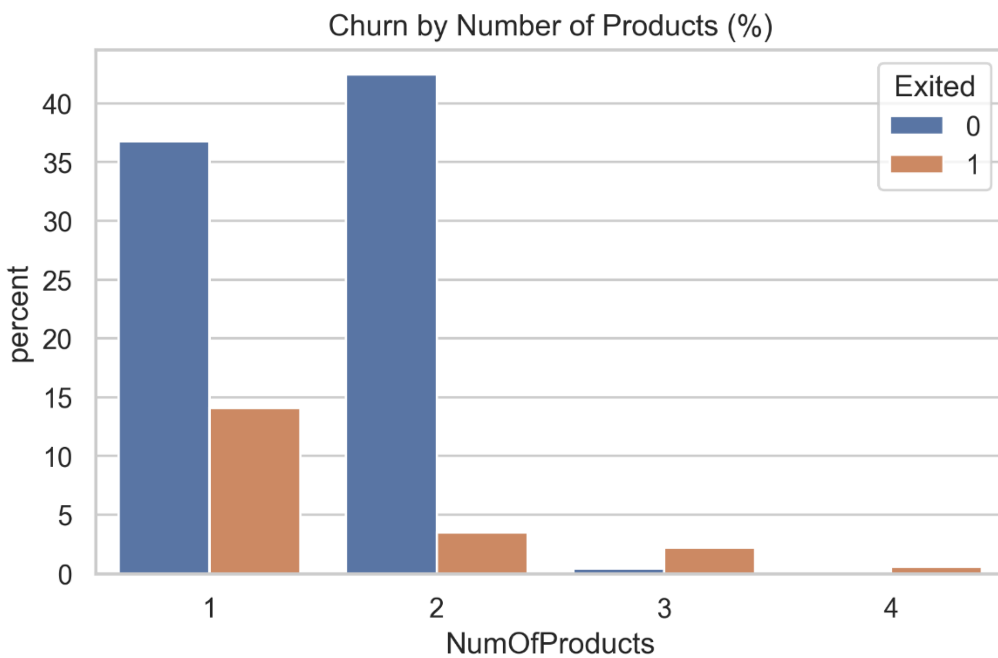
- Inactive members churn at significantly higher rates
- Engagement status emerges as a **primary behavioral risk signal**



3.4 Product Complexity & Risk

Figure 4: Churn by Number of Products

- Non-linear relationship observed
- Customers holding **3–4 products** exhibit elevated churn risk
- Suggests product overload or service friction rather than loyalty



3.5 Financial Characteristics

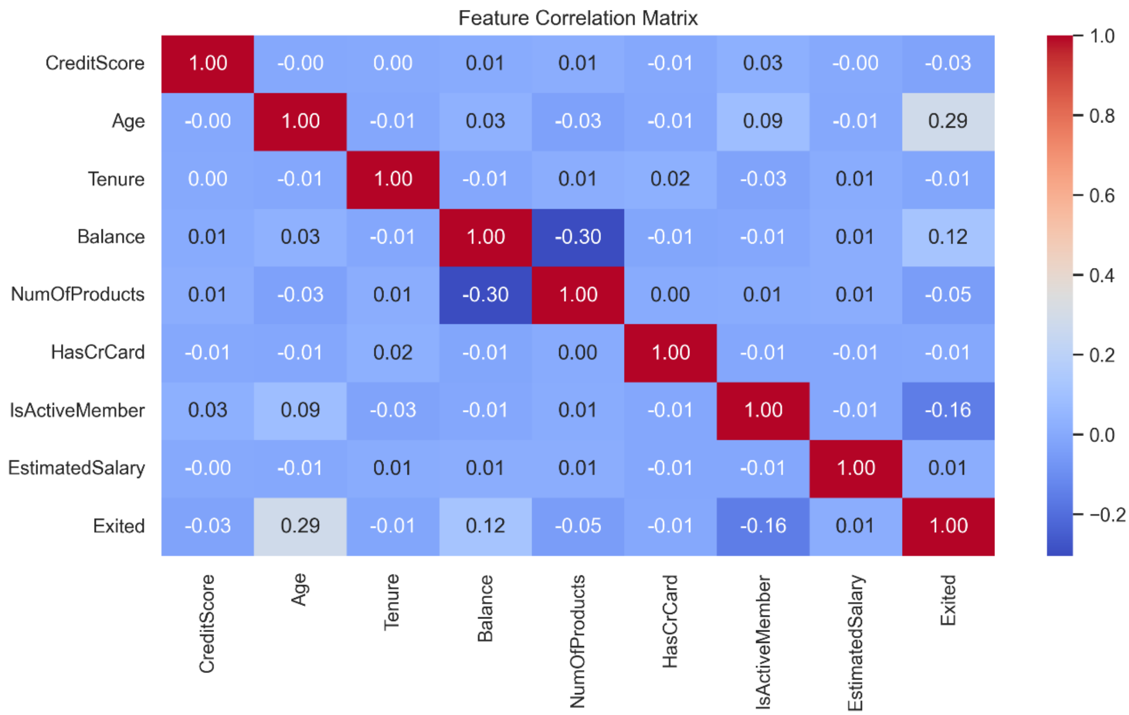
- Zero-balance and non-zero-balance customers behave differently
- Absolute balance values alone show limited linear relationship with churn
- Financial features contribute meaningfully when combined with engagement context

3.6 Correlation Structure

Figure 5: Feature Correlation Heatmap

- Low overall multicollinearity
- No problematic correlations detected

- Confirms churn is **multifactorial**, not driven by a single variable



4. Feature Engineering Rationale

Feature engineering followed two principles:

1. **Business interpretability**
2. **Signal amplification without leakage**

Key engineered features:

- Age groups (lifecycle stages)
- Tenure groups (customer maturity)
- HasBalance (financial engagement)
- HighProductCount
- InactiveHighProducts (explicit interaction risk)
- Credit score groupings

Each engineered feature is **explainable to non-technical stakeholders** and aligned with observed EDA patterns.

5. Modeling Strategy & Evaluation

5.1 Models Evaluated

- Logistic Regression (Class-Weighted)
- Logistic Regression (SMOTE)
- Random Forest (Class-Weighted)
- Random Forest (SMOTE)

5.2 Evaluation Metrics

Primary evaluation focused on:

- **Recall (Churn)** – minimize missed churners
- **ROC-AUC** – ranking and discrimination
- Precision–Recall tradeoff – operational feasibility

Accuracy was intentionally deprioritized.

5.3 Model Selection

Final Model: Class-Weighted Logistic Regression

Metric (Test Set)	Value
ROC-AUC	~0.84
Recall (Churn)	~0.74
Precision (Churn)	~0.47

While ensemble models achieved comparable ROC-AUC, they did not provide a meaningful recall advantage and reduced interpretability. The selected model represents the **optimal balance between performance, stability, and governance**.

6. Explainability & SHAP Insights

6.1 Global Drivers

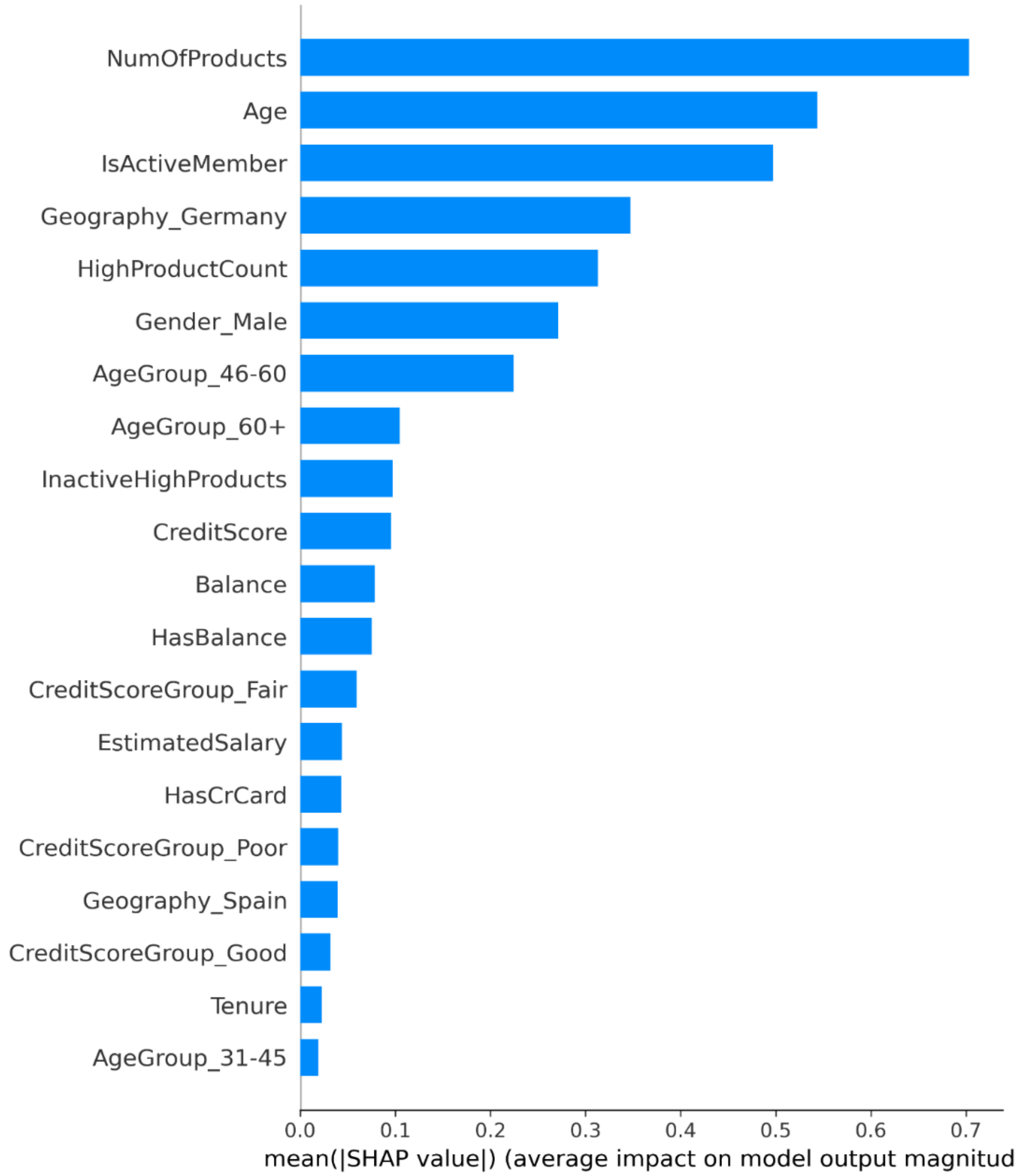


Figure 6: SHAP Global Feature Importance

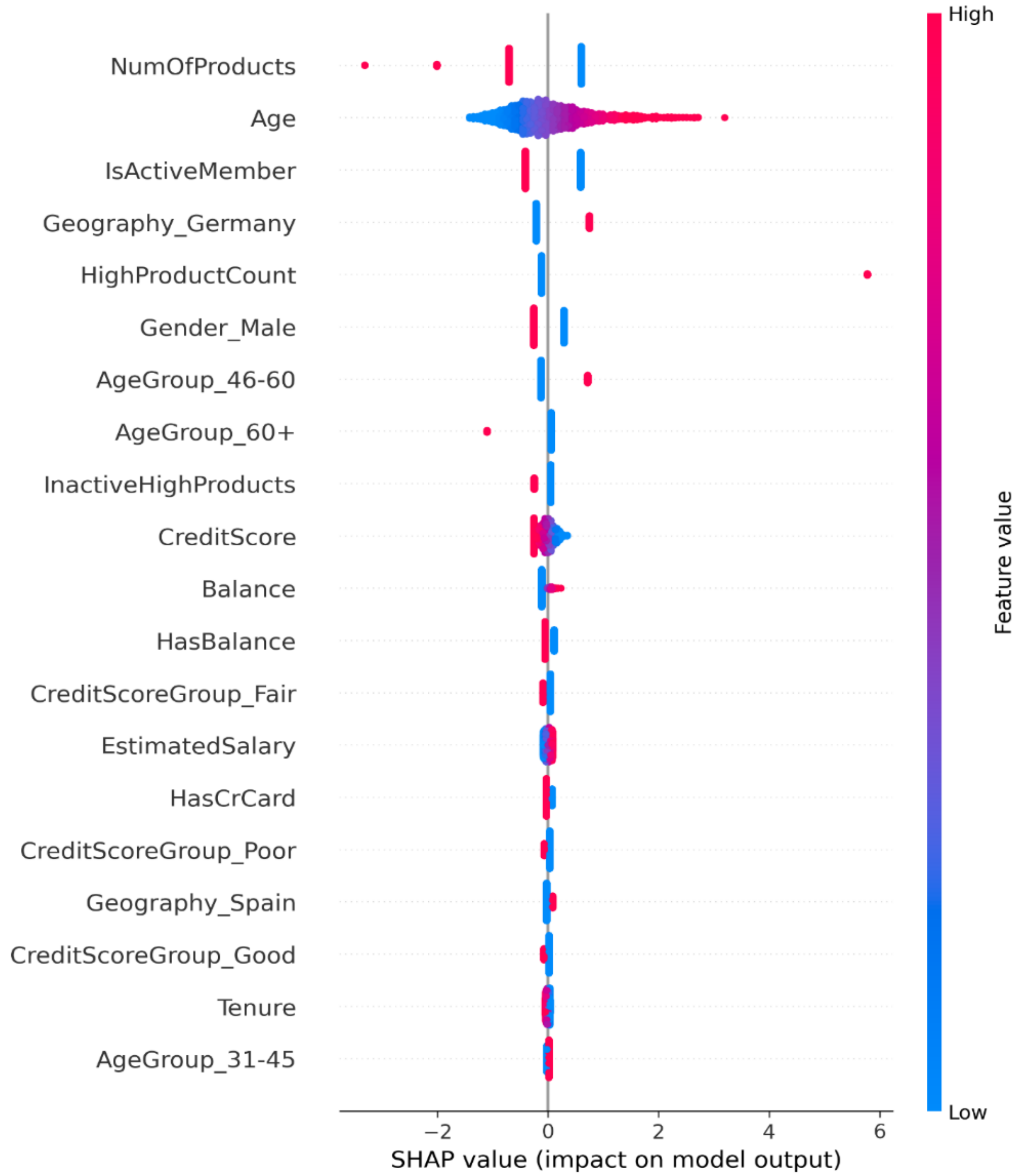


Figure 7: SHAP Beeswarm Plot

Primary churn drivers:

- Number of products
- Age

- Engagement status
- Product–engagement interactions
- Geography and selected demographics

These effects are **additive and directional rather than causal**, supporting risk prioritization without overstating behavioral drivers.

6.2 Business Interpretation

Churn risk is driven primarily by **behavioral disengagement and product complexity**, not income or credit score alone. The consistency and intuitiveness of these drivers increases executive trust and adoption.

7. Business Recommendations

1. **Proactive Retention Targeting**
   Focus outreach on high-risk customers, especially inactive multi-product holders.
2. **Engagement Re-Activation Programs**
   Intervene before exit using personalized check-ins and simplified offerings.
3. **Product Rationalization**
   Review customers with multiple products to identify friction or redundancy.
4. **Age-Sensitive Retention Design**
   Tailor strategies for older customer segments with higher churn risk.
5. **Operational Integration**
   Embed churn risk scores into CRM systems to trigger retention workflows.

8. Limitations & Next Steps

Limitations

- Snapshot-based data
- No direct observation of retention interventions
- External market factors excluded

Next Steps

- Survival or time-to-event modeling
- Uplift modeling for intervention targeting

- Ongoing performance monitoring and drift detection
- A/B testing of retention strategies

9. Conclusion

This engagement delivers a **consultant-grade, governance-ready churn prediction system** suitable for deployment in regulated, customer-centric industries. By combining disciplined modeling, transparent explainability, and business-aligned evaluation, the solution enables proactive, defensible retention decisions that protect long-term customer value.