📊 Machine Learning Feature Summary & Recommendations

This document provides a comprehensive and unified analysis of features from correlation, histogram, and pair plot perspectives to guide machine learning model development.

# 🎯 Target Variable

• exited: Binary classification target (0 = stayed, 1 = churned).

• Imbalanced target distribution – handle using SMOTE, class weighting, or stratified sampling.

# 📌 Feature Analysis Summary with Reasons

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Name | Type | Correlation with Target | Distribution Insight | ML Recommendation | Reason for Inclusion or Exclusion |
| age\_skewed | Numerical | Moderate (0.29) | Right-skewed; older customers churn more | ✅ Keep | Age is a reliable predictor of churn behavior. |
| isactivemember | Categorical | Moderate (−0.16) | Slight imbalance (more active) | ✅ Keep | Strong indicator; inactive members are more likely to churn. |
| geography\_germany | Categorical | Slight (0.17) | Smaller group; higher churn rate | ✅ Keep | Germany is more churn-prone. |
| geography\_spain | Categorical | Very Weak | Small group; less churn | ✅ Keep | Useful contrast with Germany. |
| geography\_france | Categorical | Negative vs. Germany | Majority group; neutral churn behavior | ❌ Drop | Avoid dummy variable trap. |
| balancerange | Numerical | Weak (0.11) | Bimodal; right-skewed | ✅ Keep | Financial indicator; helpful in trees. |
| creditscorerange | Numerical | ≈ 0 | Fairly uniform | 🔹 Optional | Weak signal alone; useful in tree models. |
| tenurerange | Numerical | Weak | Bimodal | ✅ Keep | Could indicate loyalty. |
| numofproducts | Numerical | Weak | Discrete (mostly 1–2) | ✅ Keep | Product count may relate to churn. |
| estimatedsalaryrange | Numerical | Very Weak | Fairly uniform; scattered | 🔹 Optional | No direct link; test for non-linear impact. |
| gender\_label | Categorical | Very Weak (−0.10) | Balanced (0/1) | 🔹 Optional | Minimal predictive power. |
| hascrcard | Categorical | Almost zero | Slightly more have cards | 🔹 Optional | Weak predictor. |
| exited | Categorical | — | Imbalanced | 🎯 Target | Classification target. |

# ✅ Final Feature Set for Modeling

Include the following features based on strong signal or modeling relevance:

• age\_skewed  
• isactivemember  
• geography\_germany  
• geography\_spain  
• balancerange  
• numofproducts  
• tenurerange

# 🔍 Optional Features

Test these features during model training and drop if they show low importance:

• creditscorerange  
• estimatedsalaryrange  
• gender\_label  
• hascrcard

# 🧠 Model Recommendations

• Use all features (final + optional) for tree-based models like Random Forest, XGBoost, or LightGBM.

• For logistic regression:

– Drop weak features like gender\_label, hascrcard, estimatedsalaryrange.

– Apply transformations (log/binned), and scale inputs.