**🧾 Formal Combined Feature Analysis Report**

**Dataset Purpose:**

To model customer churn (exited) using available demographic, behavioral, and financial features.

**📌 Target Variable:**

* exited: Binary classification target (0 = stayed, 1 = churned).
* Imbalanced distribution → requires **balancing techniques** like SMOTE, class weights, or stratified sampling.

**📊 Feature Summary Table**

| **Feature Name** | **Type** | **Distribution Insights** | **Correlation with Churn** | **Model Recommendation** |
| --- | --- | --- | --- | --- |
| age\_skewed | Numerical | Right-skewed; older customers churn more | Moderate (0.29) | ✅ Keep – important for all models |
| isactivemember | Categorical | Slight imbalance; active users churn less | Moderate (−0.16) | ✅ Keep – strong predictor |
| geography\_germany | Categorical | Smaller group, higher churn | Slight (0.17) | ✅ Keep – useful location flag |
| geography\_spain | Categorical | Smaller group, lower churn | Weak | ✅ Keep – for diversity |
| geography\_france | Categorical | Largest group, neutral behavior | Negative with Germany | ❌ Drop – avoid dummy trap |
| balancerange | Numerical | Bimodal (high & low), skewed | Weak (0.11) | ✅ Keep – useful in trees |
| creditscorerange | Numerical | Fairly uniform | ≈ 0 | 🔹 Optional – try in tree models |
| tenurerange | Numerical | Bimodal distribution | Weak | ✅ Keep – might help churn |
| numofproducts | Numerical | Discrete (1–4) | Weak | ✅ Keep – good for tree models |
| estimatedsalaryrange | Numerical | Uniform with gaps | Very weak | 🔹 Optional – tree models only |
| gender\_label | Categorical | Balanced | Very weak (−0.10) | 🔹 Optional – drop if needed |
| hascrcard | Categorical | Slight majority | Almost zero | 🔹 Optional – test & remove |

**⚙️ Feature Engineering Suggestions:**

* Apply **log transformation** to balancerange for linear models.
* **Bin** age\_skewed and tenurerange for robustness in simpler models.
* One-hot encode geography\_\* and drop one (preferably france).
* Scale features for linear models; not required for tree-based ones.

**🧠 Machine Learning Recommendations**

**🏆 Best-Suited Model Types:**

| **Model Type** | **Reason** |
| --- | --- |
| **Tree-Based Models** (Random Forest, XGBoost, LightGBM) | Handle skewed features well, no scaling required, robust to irrelevant features |
| **Logistic Regression** | Good baseline, but requires scaling, binning, and dropping weak features |
| **Gradient Boosting (e.g., CatBoost)** | Works well with categorical and numerical mix, handles imbalance |

**✅ Final Feature Set for ML Training**

| **Feature Name** | **Use in ML** |
| --- | --- |
| age\_skewed | ✅ Keep |
| isactivemember | ✅ Keep |
| geography\_germany | ✅ Keep |
| geography\_spain | ✅ Keep |
| numofproducts | ✅ Keep |
| balancerange | ✅ Keep |
| tenurerange | ✅ Keep |
| creditscorerange | 🔹 Optional |
| estimatedsalaryrange | 🔹 Optional |
| gender\_label | 🔹 Optional |
| hascrcard | 🔹 Optional |
| geography\_france | ❌ Drop |