PREPARED BY: DARIUS ARDALES

CREDIT CARD ATTRITION

A Comprehensive Analysis

BUSINESS CONTEXT

A major bank is experiencing customer churn in its credit card division. Understanding the factors that lead to customer attrition is critical for:

- Designing effective retention strategies
- Improving customer satisfaction

OBJECTIVE

Build a **predictive model** to identify customers likely to close their credit card accounts.

DATASET OVERVIEW

The synthetic dataset contains customer information, transaction history, and demographic details.

DATASET INFORMATION

Dataset Shape

101000 customers x 63 features

57 Numerical Features

(including the anonymous features, income, credit limit, total spend, age, total transactions, and tenure)

6 Categorical Features

(customerID, gender, marital status, education level, card type, and country)

PRELIMINARY ANALYSIS

- **Class Imbalance**: The AttritionFlag has a mean of ~0.05, indicating only about 5% customer attrition rate
- **Standardized Anonymous Features**: All 50 anonymous features (Feature_0 through Feature_49) appear to be pre-processed and standardized
- **Missing Data**: Business features show missing values
- **Feature Scale Disparities**: Business features exhibit vastly different scales
- **Anomalous Negative Values**: CreditLimit show a negative minimum value

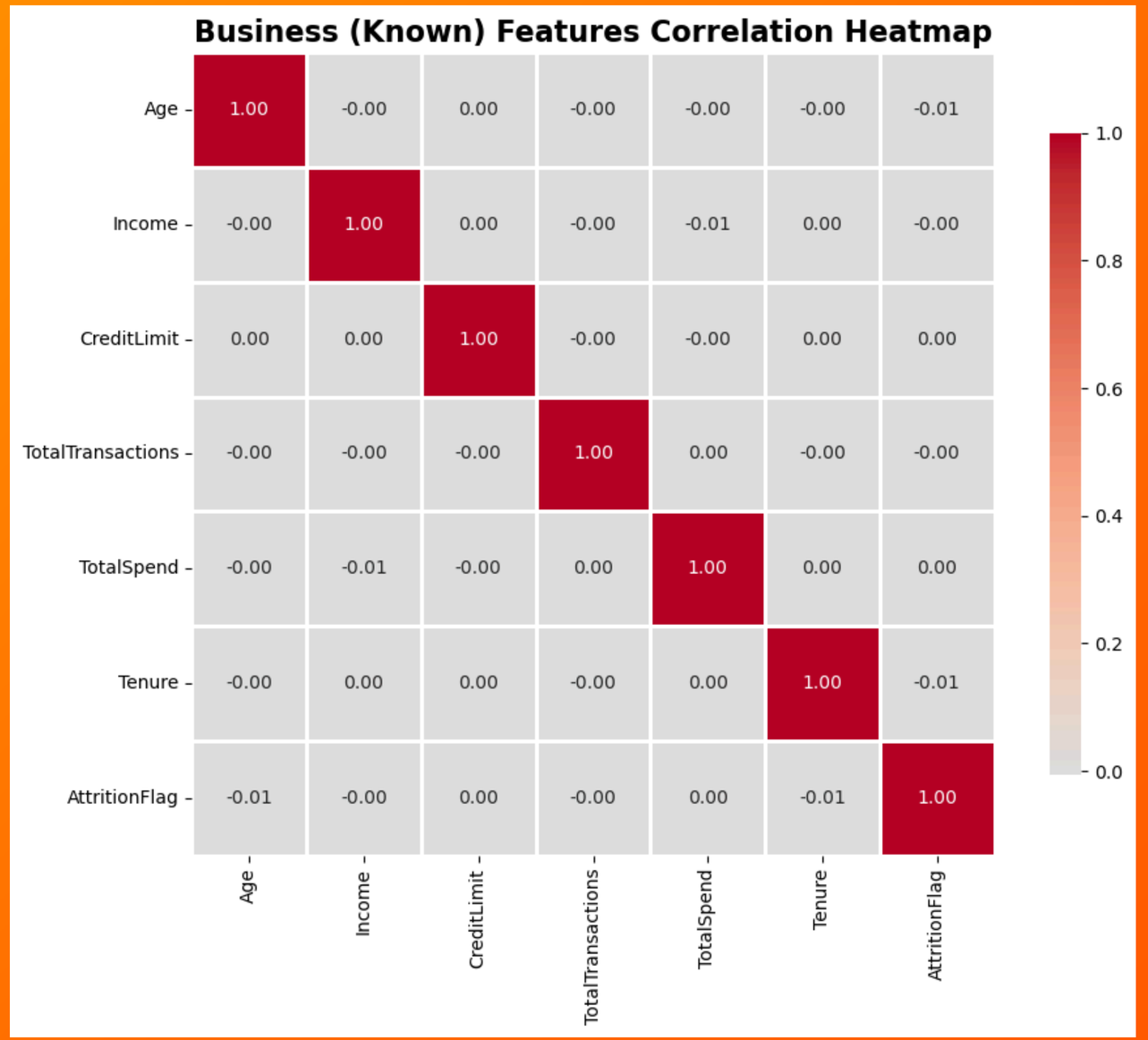
BUSINESS FEATURES

Numerical: Age, Income, CreditLimit, TotalTransactions, TotalSpend, Tenure

Categorical: Gender, MaritalStatus, EducationLevel

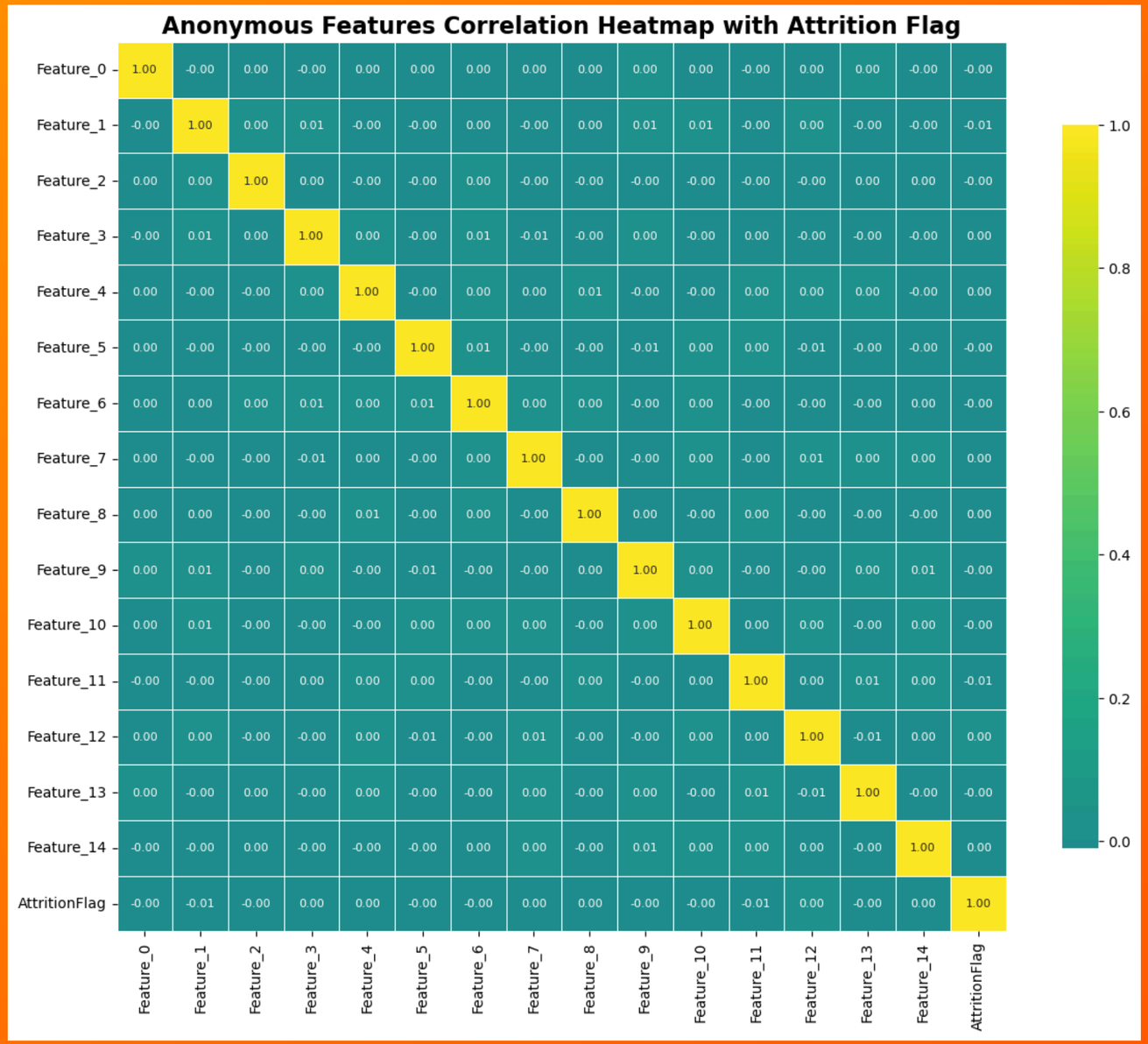
Target (Binary): AttritionFlag

- 0 = No Attrition (Customer retained)
- 1 = Attrition (Customer closed credit card)



ANONYMOUS FEATURES

Numerical: 50 features from Feature_0 to Feature_49



DATA CLEANING

DUPLICATE RECORDS

Total duplicate rows: 1000

Solution:
Remove duplicate rows,
keeping the first occurrence

Results:
Original shape: (101000, 63)
New shape: (100000, 63)
Rows removed: 1000

DATA CLEANING

PROBLEMATIC FEATURES

AGE LIMIT (BELOW 7 YEARS OLD)

Records with Age < 7: 9 customers

Solution: Remove rows

Rows remaining after removing Age < 7: 99991 rows

REFERENCE: HTTPS://WWW.UNIONBANKPH.COM/ACCOUNTS/FIRST-SAVINGS

DATA CLEANING

PROBLEMATIC FEATURES

NEGATIVE CREDIT LIMIT

Records with negative CreditLimit: 142

Solution: Remove rows

Rows remaining after removing negative CreditLimit: 99849 rows

REFERENCE: HTTPS://WWW.UNIONBANKPH.COM/ACCOUNTS/FIRST-SAVINGS

DATA CLEANING

OUTLIERS

The **Interquartile Range (IQR)** method was chosen for outlier detection because:

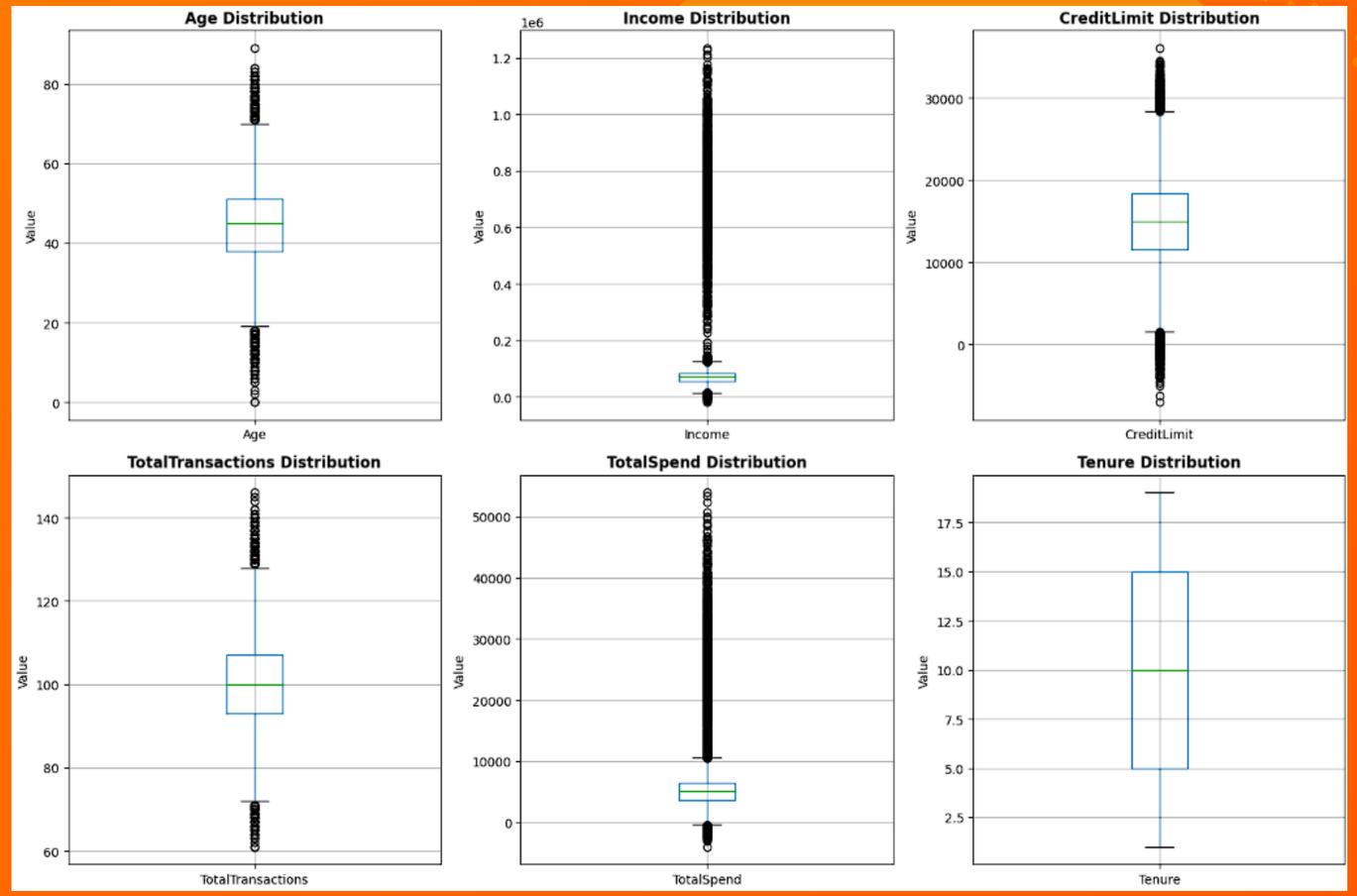
- **Non-parametric:** Doesn't assume normal distribution
- **Business-friendly:** The 1.5×IQR threshold is an industry standard that identifies truly unusual values
- **Interpretable:** Easy to explain to stakeholders (values beyond 1.5 times the middle 50% range)

DATA CLEANING

OUTLIERS

Business Features

	Feature	Outliers	Percentage	Lower_Bound	Upper_Bound
0	Age	927	0.93%	18.50	70.50
1	Income	1509	1.51%	15,640.87	124,988.64
2	CreditLimit	587	0.59%	1,555.49	28,424.14
3	TotalTransactions	474	0.47%	72.00	128.00
4	TotalSpend	1479	1.48%	-441.40	10,500.42
5	Tenure	0	0.00%	-10.00	30.00



Minimal outliers across all features

Decision: Retain All Outliers for Business Features

DATA CLEANING

OUTLIERS

Anonymous Features

Features with outliers: 50 / 50

	Feature	Outliers	Percentage	Lower_Bound	Upper_Bound
0	Feature_0	720	0.72%	-2.694	2.696
1	Feature_1	657	0.66%	-2.702	2.705
2	Feature_2	720	0.72%	-2.687	2.692
3	Feature_3	699	0.70%	-2.708	2.700
4	Feature_4	707	0.71%	-2.701	2.707
5	Feature_5	690	0.69%	-2.685	2.693
6	Feature_6	658	0.66%	-2.702	2.701
7	Feature_7	712	0.71%	-2.708	2.706
8	Feature_8	718	0.72%	-2.691	2.691
9	Feature_9	761	0.76%	-2.690	2.699

Reasons:

1. **Outlier rates:** Extremely consistent at **0.64% - 0.77%**
2. **Bounds:** Uniform at approximately **±2.7 for all** features
3. **Already Standardized:** The ±2.7 bounds correspond to ~2.7 standard deviations, confirming these features are pre-standardized (mean=0, std=1)
4. **Minimal Impact:** Less than 0.8% outliers per feature means removing them would affect <1% of the dataset

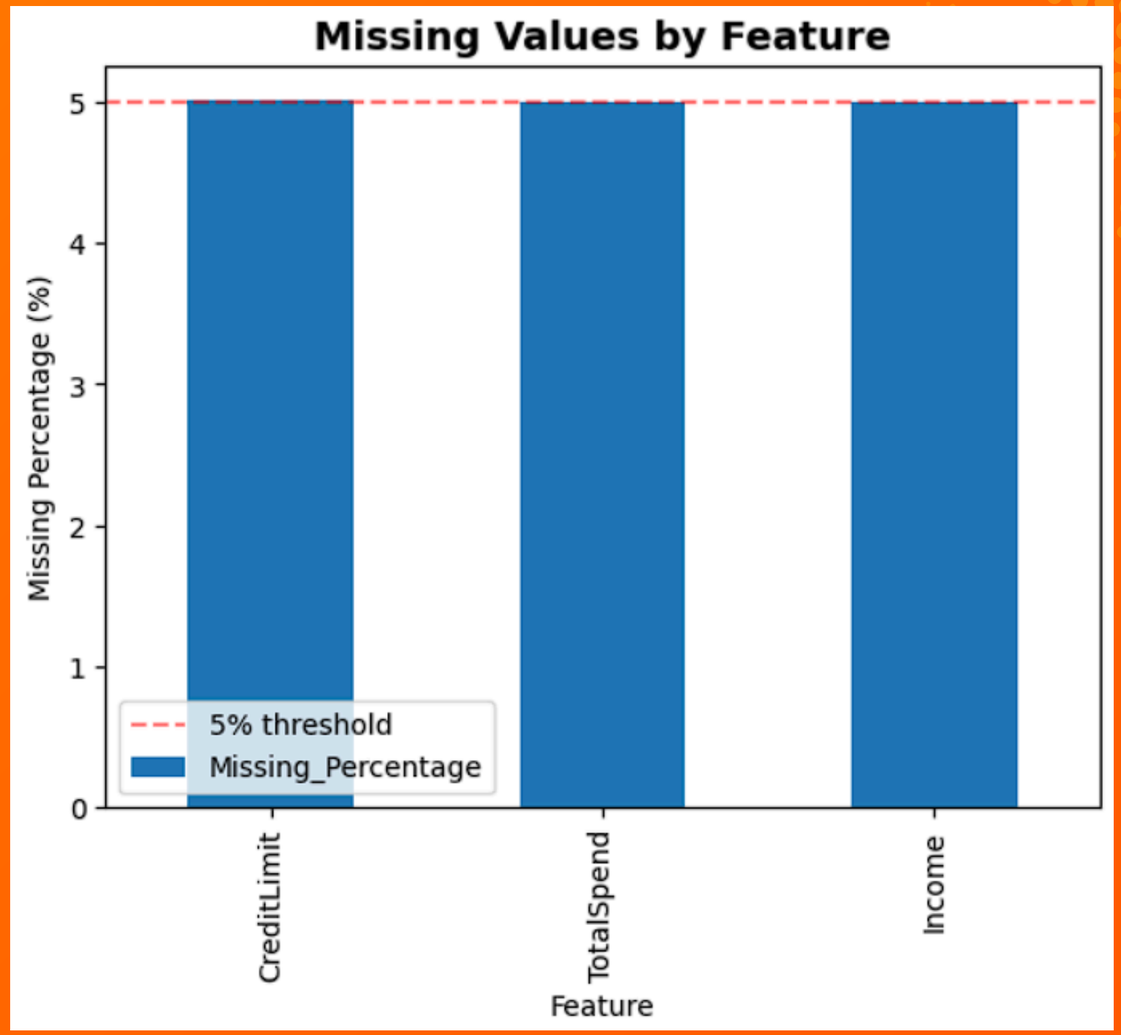
Decision: Retain All Outliers for Anonymous Features

DATA CLEANING

MISSING VALUES

All three missing: 13 rows
Income & TotalSpend missing: 262 rows
Income & CreditLimit missing: 235 rows
TotalSpend & CreditLimit missing: 233 rows
Only Income missing: 4479 rows
Only TotalSpend missing: 4483 rows
Only CreditLimit missing: 4519 rows

	Income	TotalSpend	CreditLimit
Income	1.000	0.005	0
TotalSpend	0.005	1.000	-0.001
CreditLimit	0	-0.001	1.000

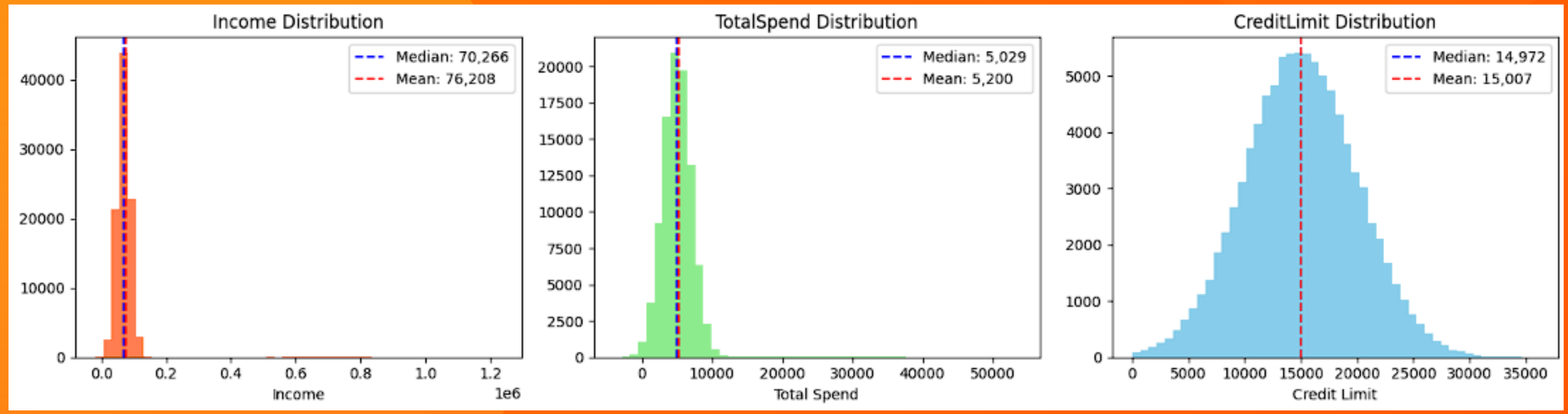


DATA CLEANING

MISSING VALUES

Decision: Apply Mean Imputation for all

- Missing values per feature is <= to 5%
- Independent missing patterns (correlation <= 0.005) tells us we can use simple imputation rather than advanced methods
- The distributions are normal (mean is best for normal data)



DATA CLEANING

CARDINALITY OF CATEGORICAL FEATURES

Distribution: All 100 countries have 924-1024 customers (uniform)
Attrition Impact: Variance of 0.00005 (negligible)

Decision: DROP the Country feature

Reasons:

- No predictive value for attrition
- Would create 100 dummy variables
- Risk of overfitting without benefit

Feature	Unique_Values
Gender	2
MaritalStatus	4
EducationLevel	4
CardType	4
Country	100

DATA CLEANING

IMBALANCED FEATURES

EducationLevel - ✓ Balanced (Ratio: 1.0:1)

Value	Count	Percentage
Master	25,037	25%
High School	24,998	25%
Bachelor	24,945	25%
PhD	24,869	25%

Gender - ✓ Balanced (Ratio: 1.0:1)

Value	Count	Percentage
Female	50,115	50%
Male	49,734	50%

CardType - ✓ Balanced (Ratio: 1.0:1)

Value	Count	Percentage
Platinum	25,061	25%
Black	25,039	25%
Gold	24,961	25%
Silver	24,788	25%

MaritalStatus - ✓ Balanced (Ratio: 1.0:1)

Value	Count	Percentage
Divorced	25,159	25%
Married	25,086	25%
Widowed	24,911	25%
Single	24,693	25%

DATA CLEANING

IMBALANCED FEATURES

AttritionFlag

Target Variable Distribution:

- **No Attrition (0):** 94,897 (95.0%)
- **Attrition (1):** 4,952 (5.0%)

Imbalance: 19:1 ratio with 5.0% minority class

Decision: Since we have nominal and continuous features, we use **SMOTE for Nominal and Continuous features (SMOTENC)**

Key Benefits

1. From 0% to 46% Recall (Simple testing with Logistic Regression)
2. Balanced Training Data
    - Changed from 19:1 to 1:1 ratio
    -

Why SMOTENC Over Other Methods?

- Undersampling: Would discard 95% of data, losing valuable information
- Basic SMOTE: Cannot handle categorical features properly
- Class Weights Alone: Insufficient for extreme 19:1 imbalance

DATA CLEANING

IMBALANCED FEATURES

AttritionFlag

Dataset Transformation:

- Original samples: 99,849
- Synthetic samples added: 89,945
- Total samples in df_clean: 189,794

Sample New CustomerIDs:

- First: ['CUST100000', 'CUST100001', 'CUST100002']
- Last: ['CUST189942', 'CUST189943', 'CUST189944']

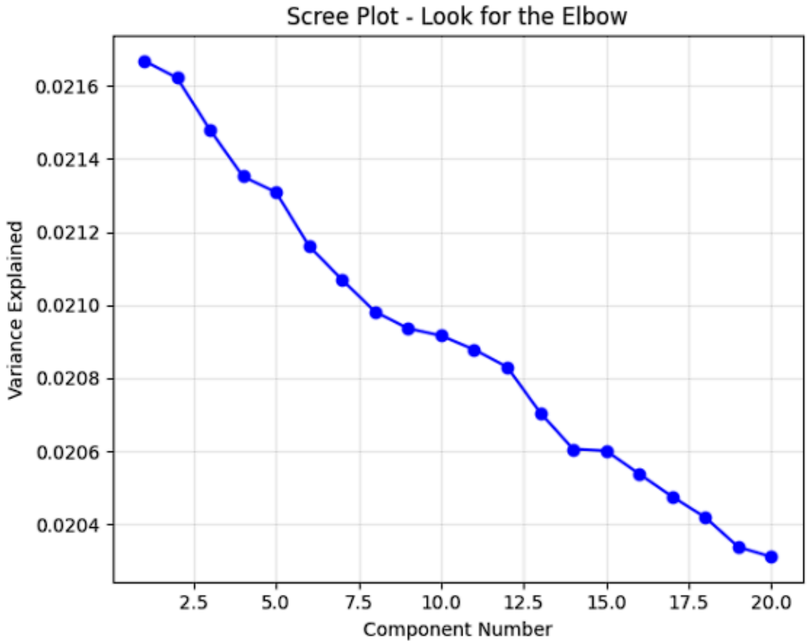
Final Class Distribution:

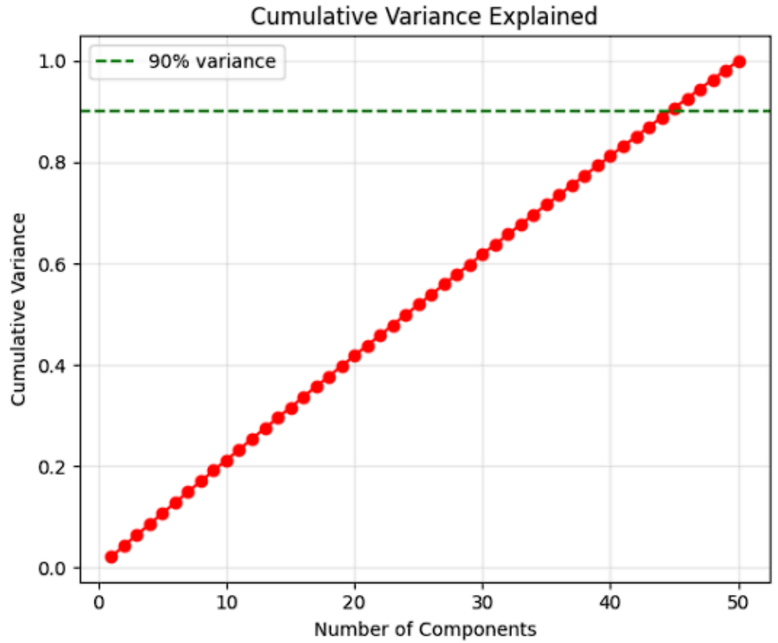
- Class 0 (No Attrition): 94,897
- Class 1 (Attrition): 94,897

FEATURE ENGINEERING

HIGH DIMENSIONALITY

PRINCIPAL COMPONENT ANALYSIS (PCA)





The cumulative variance plot displays an almost linear increase, indicating that each component contributes relatively equally to the total variance.

Decision: we use the **90% variance threshold.** 90% is a widely accepted benchmark for retaining sufficient information

The scree plot shows a gradual, continuous decline without a clear elbow point. The variance explained decreases smoothly from component to component, making it difficult to identify an obvious cutoff point using the elbow method.

FEATURE ENGINEERING

HIGH DIMENSIONALITY

PRINCIPAL COMPONENT ANALYSIS (PCA)

Components needed for 90% variance:

45

Variance retained:

90.7%

Dimensionality reduction:

50 → 45 features

New DataFrame Shape:

(189794, 57)

FEATURE ENGINEERING

SPENDING RATE:

AVERAGE SPENDING PER TRANSACTION

Description: TotalSpend ÷ TotalTransactions

Why: Identifies high-value vs. low-value spenders. Customers who spend more per transaction may be more engaged and less likely to churn.

FEATURE ENGINEERING

MONTHLY SPENDING RATE:

AVERAGE MONTHLY SPENDING

Description: TotalSpend ÷ Tenure (in months)

Why: Normalizes spending by time to compare customers fairly. Higher monthly spending indicates better engagement.

FEATURE ENGINEERING

TRANSACTION FREQUENCY:

AVERAGE TRANSACTIONS PER MONTH OF TENURE

Description: TotalTransaction ÷ Tenure (in months)

Why: Measures how often customers use their card. Frequent users are typically more loyal and less likely to leave.

FEATURE ENGINEERING

TENURE GROUPS:

CUSTOMER LOYALTY SEGMENTS

Description: Categorizes tenure into groups

Why: Captures non-linear tenure effects. Different customer lifecycle stages have different churn patterns.

FEATURE ENGINEERING

CREDIT LIMIT TIER:

CATEGORIZE CREDIT LIMITS

Description: Groups continuous credit limits into tiers

Why: Different credit segments behave differently. Helps target retention strategies by customer value tier.

FEATURE ENGINEERING

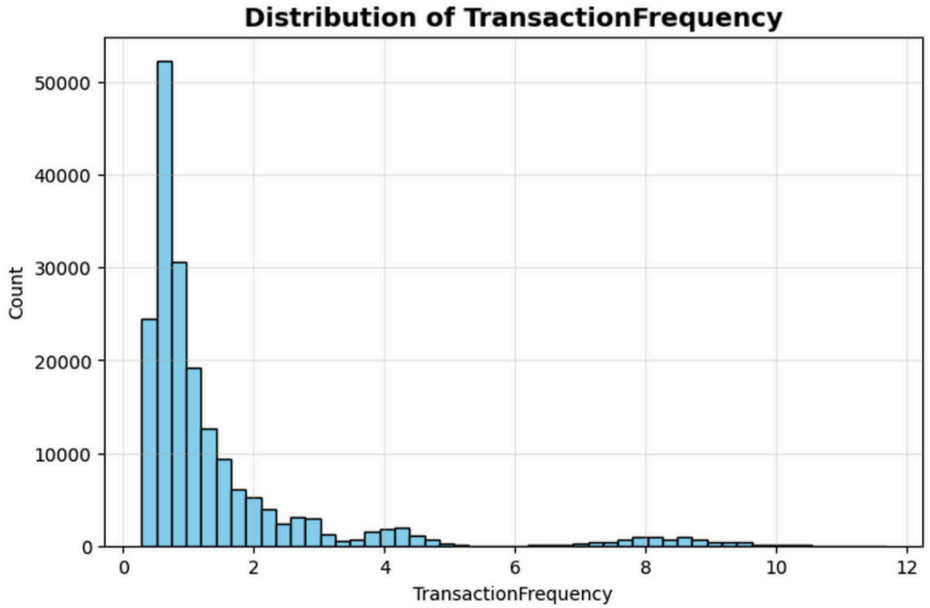
LOG TRANSFORMATION

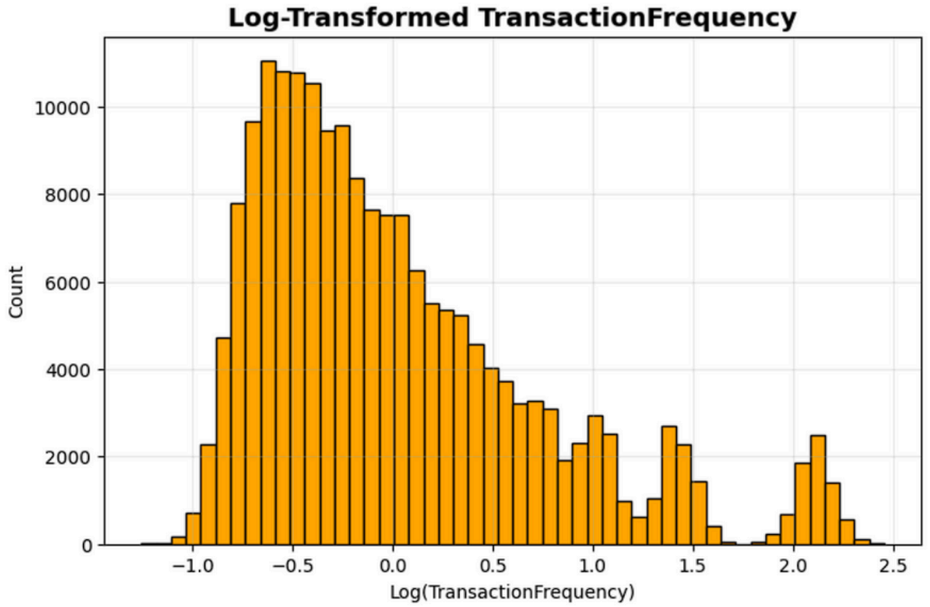
TRANSACTIONFREQUENCY FEATURE

TransactionFrequency has:

- skewness > 1.0
- a positive min value

	Feature	Skewness	Min_Value	Apply_Log
1	Income	10.317593	-18277.719329	False
7	MonthlySpendingRate	7.618697	-159.057085	False
4	TotalSpend	5.234408	-4005.132671	False
6	SpendingRate	5.208767	-44.501474	False
8	TransactionFrequency	3.103723	0.285088	True
3	TotalTransactions	0.123237	61.000000	False
2	CreditLimit	0.041732	12.273649	False
5	Tenure	0.039382	1.000000	False
0	Age	0.011689	7.000000	False





FEATURE ENGINEERING

LABEL ENCODER

NOMINAL FEATURES

- Gender
- MaritalStatus
- EducationLevel

ORDINAL ENCODER

ORDINAL FEATURES

- Card Type: ['Platinum' 'Gold' 'Silver' 'Black']
- Tenure Group: ['New', 'Established', 'Loyal', 'Growing']
- Credit Tier: ['High', 'Low', 'Medium', 'Premium']

FEATURE ENGINEERING

SUMMARY

Removed Features (54 total):

- 50 anonymous features (Feature_0 to Feature_49)
- 5 categorical features (Gender, MaritalStatus, EducationLevel, CardType, Country)

Added Features (54 total):

- 45 PCA components (replacing anonymous features)
- 3 encoded categoricals (Gender_encoded, MaritalStatus_encoded, EducationLevel_encoded)
- 3 ordinal features (CardType_ordinal, TenureGroup_ordinal, CreditTier_ordinal)
- 3 ratio features (SpendingRate, MonthlySpendingRate, TransactionFrequency)

1 log-transformed feature (Log_TransactionFrequency)

FINAL DATASET

Dropped AttritionFlag and CustomerID:

- Features shape: (189794, 60)

Only AttritionFlag:

- Target shape: (189794,)
- Number of features: 60

Performing Train-Test Split (80/20)

Training set shape:
X_train = (151835, 60)
y_train = (151835,)

Test set shape:
X_test = (37959, 60)
y_test = (37959,)

Class distribution in training set:
AttritionFlag
1 0.500003
0 0.499997
Name: proportion, dtype: float64

Class distribution in test set:
AttritionFlag
0 0.500013
1 0.499987

MODEL DEVELOPMENT

Why No Standardization?
Tree-based models (LightGBM, XGBoost) make splits based on value thresholds, not distances. They only care about the relative ordering of values, not their scale.

Selected Models:

- **LightGBM:** Fast training, memory efficient, handles large datasets well
- **XGBoost:** Often best performance, extensive hyperparameter tuning options

Our main metric is **Recall**. Why? Because missing a churner (False Negative) is expensive:

- You lose the customer permanently
- No chance for intervention
- High replacement cost

MODEL DEVELOPMENT

WHY TREE-BASED MODELS?

1. Low Feature Correlations

- Our analysis shows correlations between variables are nearly 0
- Tree-based models excel with independent features, as they can capture non-linear patterns without multicollinearity issues

2. High-Dimensional Data Without Feature Selection

- We have 60+ features after engineering (including 45 PCA components)
- Tree-based models automatically perform implicit feature selection through splitting
- No need for manual feature selection - the models naturally use only informative features
- Can handle hundreds of features without overfitting through ensemble averaging

3. Feature Importance

- These models provide clear feature importance scores
- Helps identify which factors drive customer attrition for business insights

4. Robust to Outliers

- Unlike linear models, tree-based methods are not sensitive to outliers
- Important for financial data, which often has extreme values that we can't remove

MODEL DEVELOPMENT

LIGHTGBM CLASSIFIER

```
param_lgbm = {
    'n_estimators': [100, 200, 500],
    'learning_rate': [0.01, 0.05, 0.1],
    'max_depth': [3, 5, 7],
    'num_leaves': [15, 31, 63],
    'feature_fraction': [0.6, 0.8, 1.0],
    'bagging_fraction': [0.6, 0.8, 1.0],
}
```

Using RandomizedSearchCV (Due to time constraints):

- Cross Validation: 5 folds
- Scoring: Recall
- Number of iterations: 200

LIGHTGBM HYPERPARAMETER REFERENCE

MODEL DEVELOPMENT

XGBOOST CLASSIFIER

```
param_xgb = {
    'n_estimators': [50, 100, 200],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.05, 0.1],
    'subsample': [0.8, 1.0],
    'colsample_bytree': [0.8, 1.0]
}
```

Using RandomizedSearchCV (Due to time constraints):

- Cross Validation: 5 folds
- Scoring: Recall
- Number of iterations: 200

XGBOOST HYPERPARAMETER REFERENCE

MODEL EVALUATION

RESULTS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LightGBM	0.9272	0.9037	0.9564	0.9293	0.9791
XGBoost	0.8946	0.8688	0.9295	0.8981	0.9611

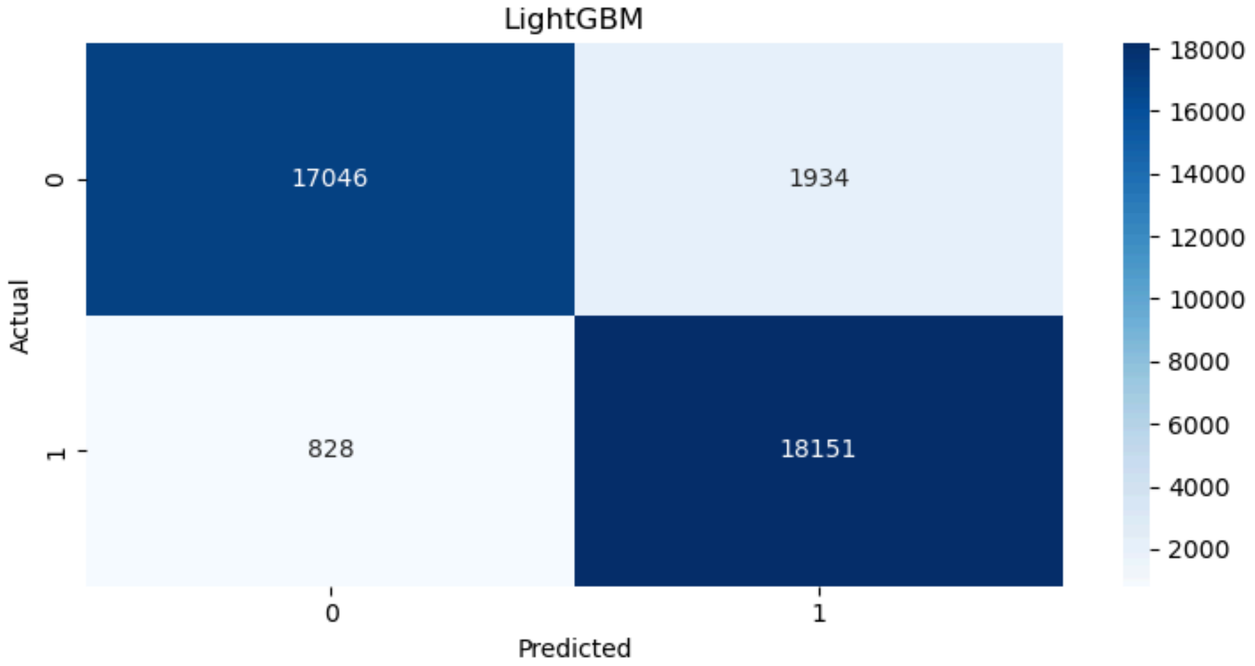
Best Model: LightGBM

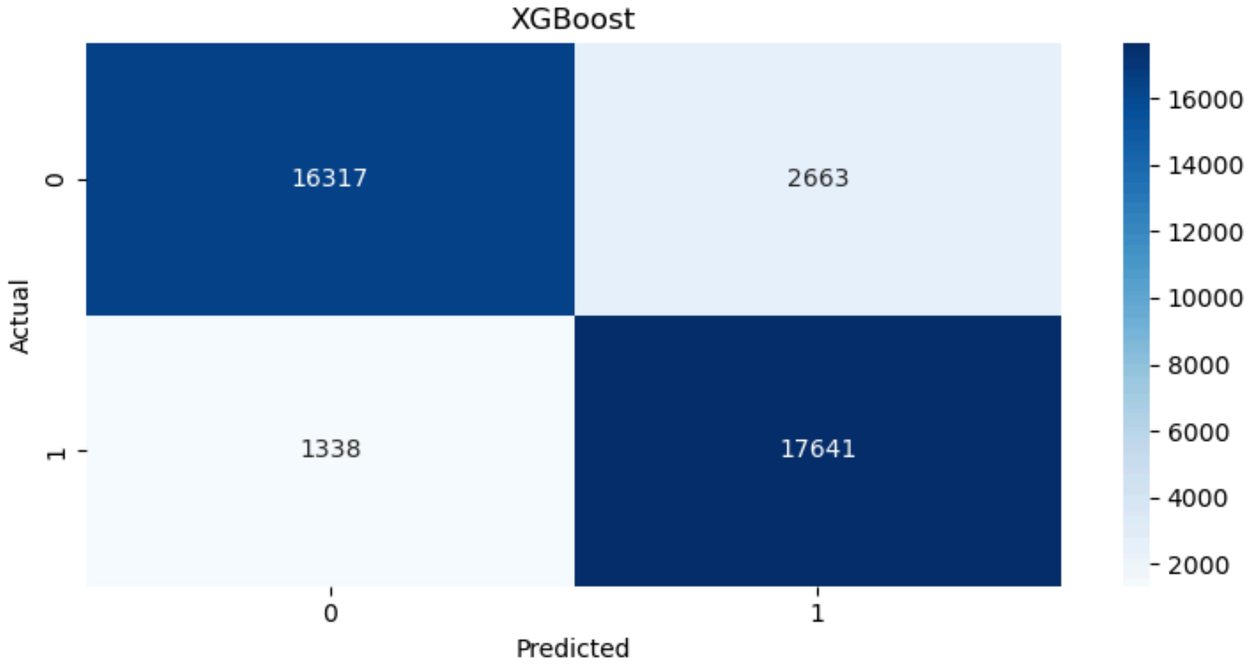
- Recall: 95.64% (catches 96% of churners)
- Precision: 90.37%

Conclusion: Using LightGBM for maximum credit card attrition detection

MODEL EVALUATION

CONFUSION MATRICES



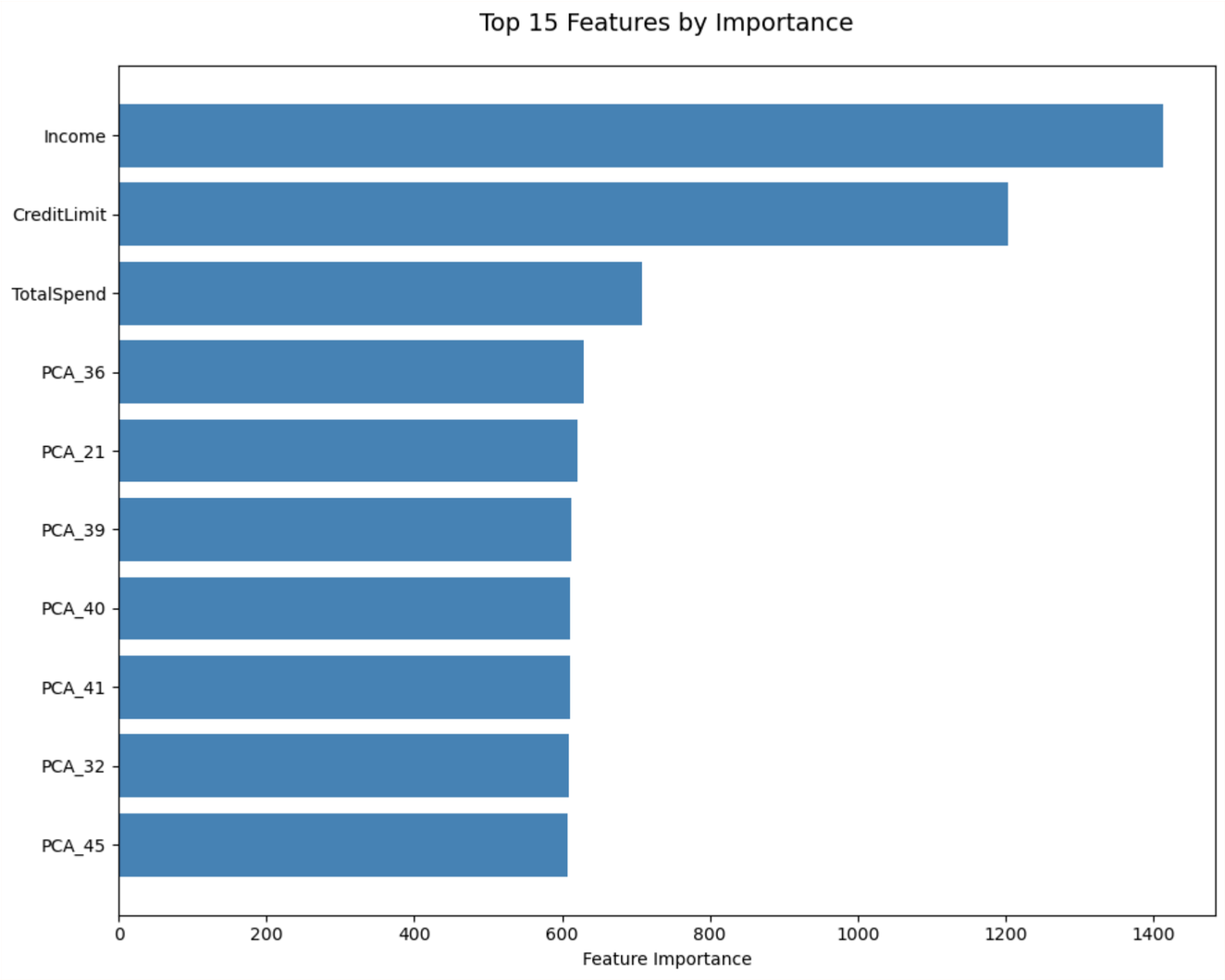


LightGBM Confusion Matrix Details:

- True Negatives: 17,046 (correctly identified non-churners)
- False Positives: 1,934 (false alarms)
- False Negatives: 828 (missed churners)
- True Positives: 18,151 (correctly identified churners)

INSIGHTS & RECOMMENDATIONS

FEATURE IMPORTANCE FROM LIGHTGBM



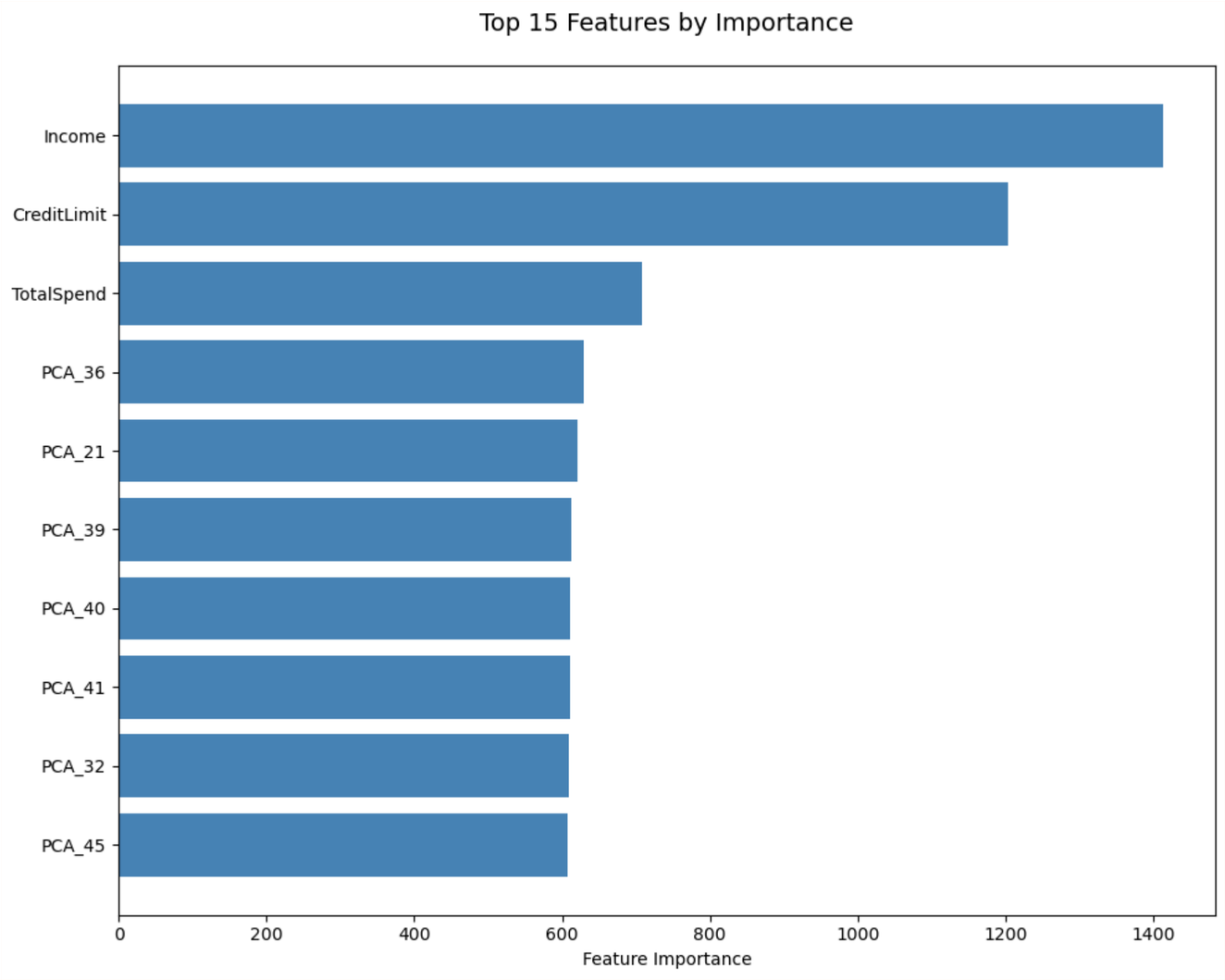
Three financial factors:

1. **Income** - Customer's annual income
2. **Credit Limit** - Maximum credit available on the card
3. **Total Spend** - Customer's total spending amount

These three features alone contribute most of the model's predictive power, indicating that financial behavior and capacity are the strongest indicators of whether a customer will close their credit card account.

INSIGHTS & RECOMMENDATIONS

FEATURE IMPORTANCE FROM LIGHTGBM

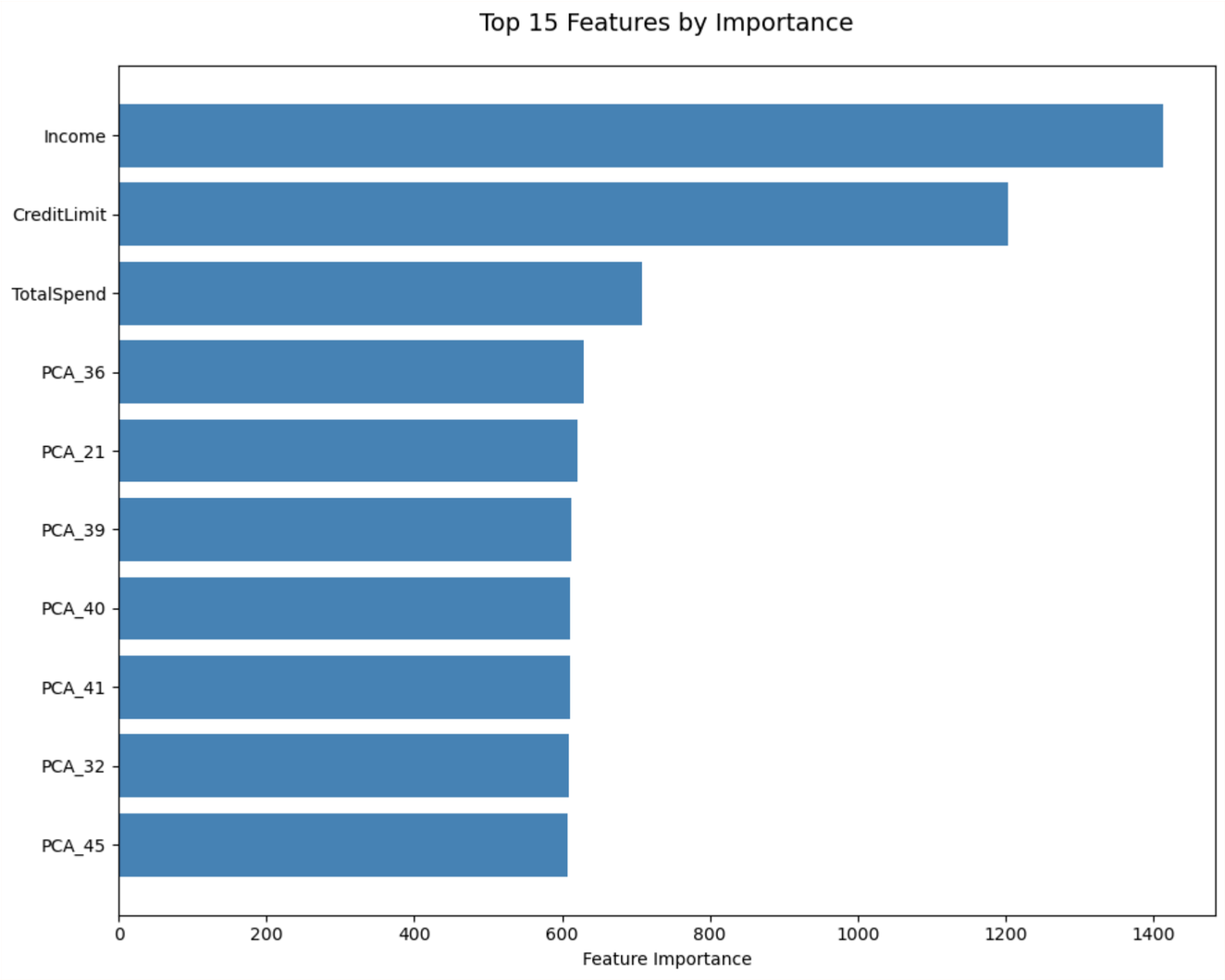


“Financial Profile Matters Most”

- **Higher income customers** tend to have different attrition patterns, possibly due to having more options or different expectations
- **Credit limit** reflects the bank's assessment of the customer's creditworthiness and their relationship value
- **Spending patterns** directly show customer engagement with the credit card

INSIGHTS & RECOMMENDATIONS

FEATURE IMPORTANCE FROM LIGHTGBM

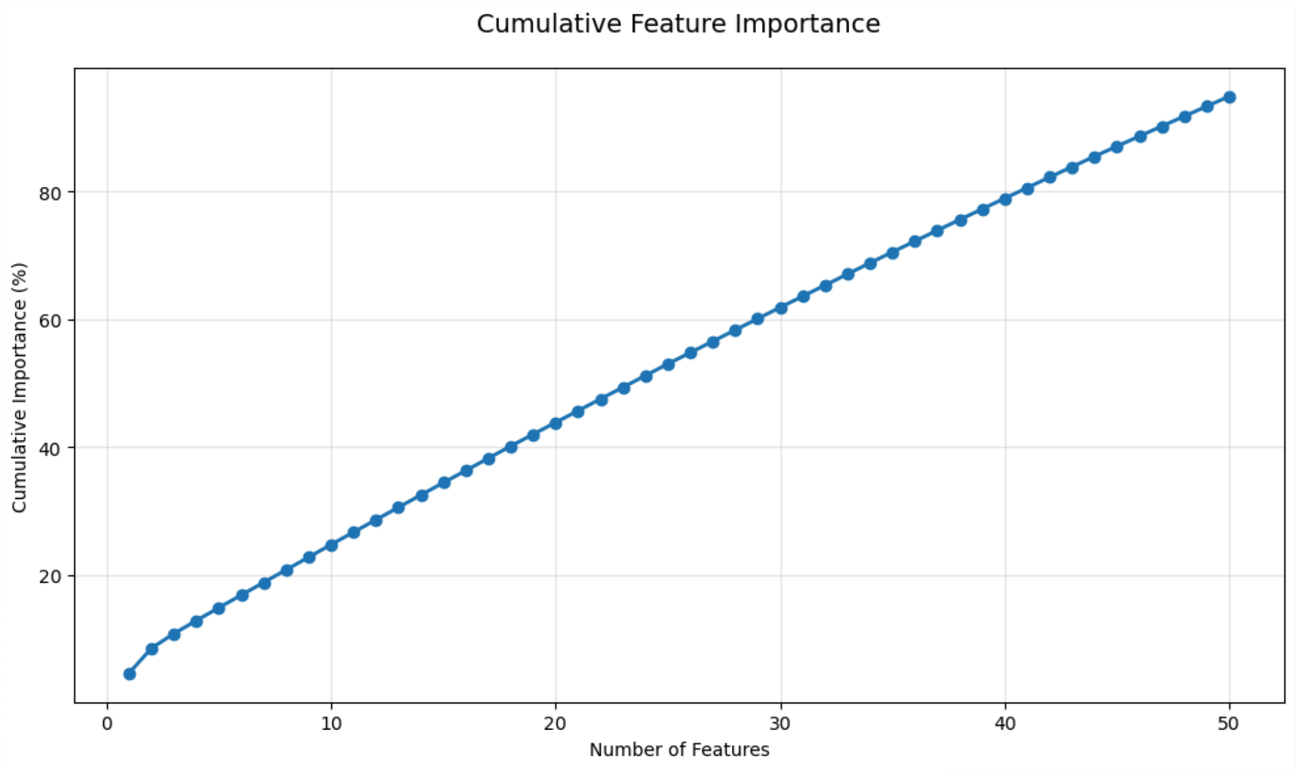


Customers who are likely to leave typically show:

- Mismatch between their income level and credit limit offered
- Changes in spending behavior over time
- Lower overall transaction amounts compared to their credit limit

INSIGHTS & RECOMMENDATIONS

CUMULATIVE IMPORTANCE



The **linear trend** in cumulative importance indicates that:

- Each additional feature contributes roughly equal incremental value
- There's no clear "elbow point" where adding more features stops being useful
- The model benefits from the full feature set rather than just a few dominant variables

INSIGHTS & RECOMMENDATIONS

RECOMMENDATIONS BASED ON THESE FINDINGS

1. **Focus retention efforts** on customers showing unusual patterns in income-to-credit-limit ratios

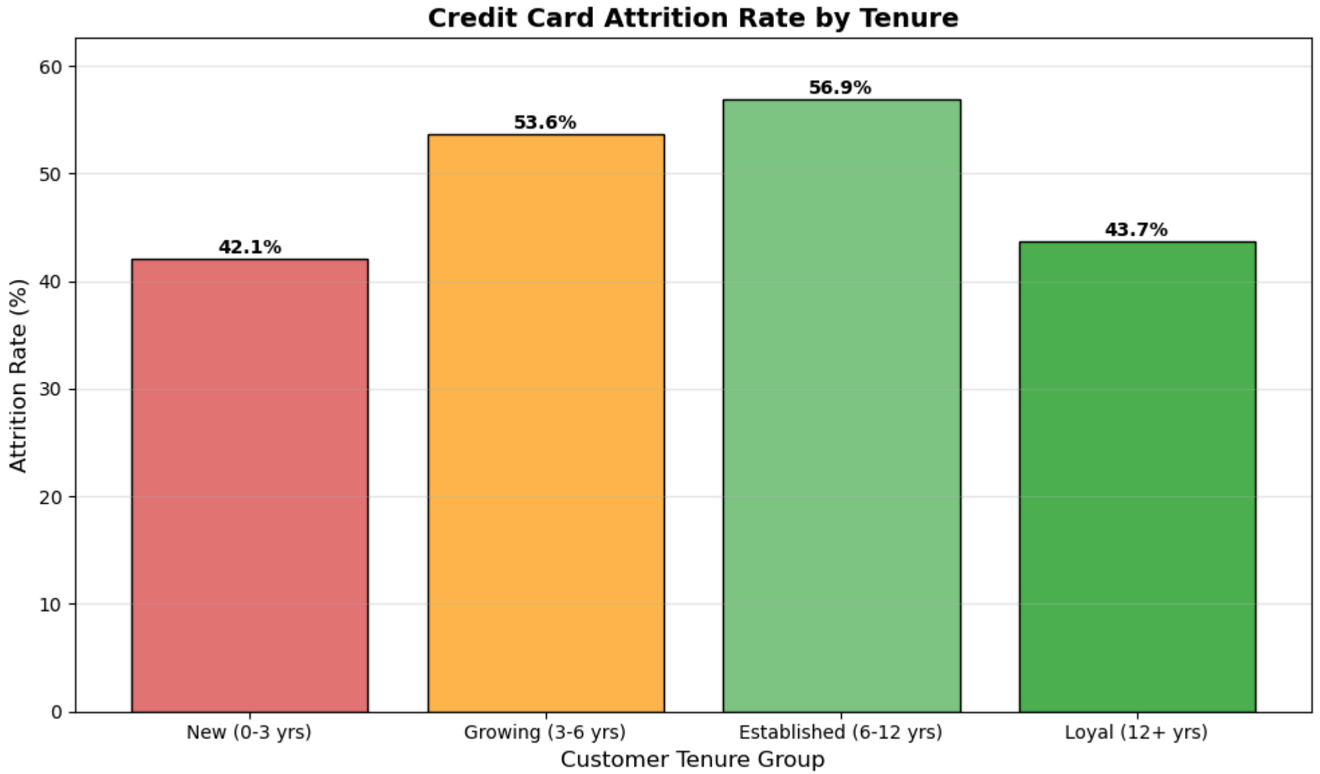
2. **Monitor spending trends** as early warning signals for potential attrition

3. **Consider all features** as much as possible when building predictive models, as the linear cumulative pattern shows that each adds value

4. **Develop targeted strategies** for different financial segments based on income and credit limit combinations

INSIGHTS & RECOMMENDATIONS

CUSTOMER TENURE ANALYSIS



Risk Pattern: Highest risk in Established phase (56.9%)

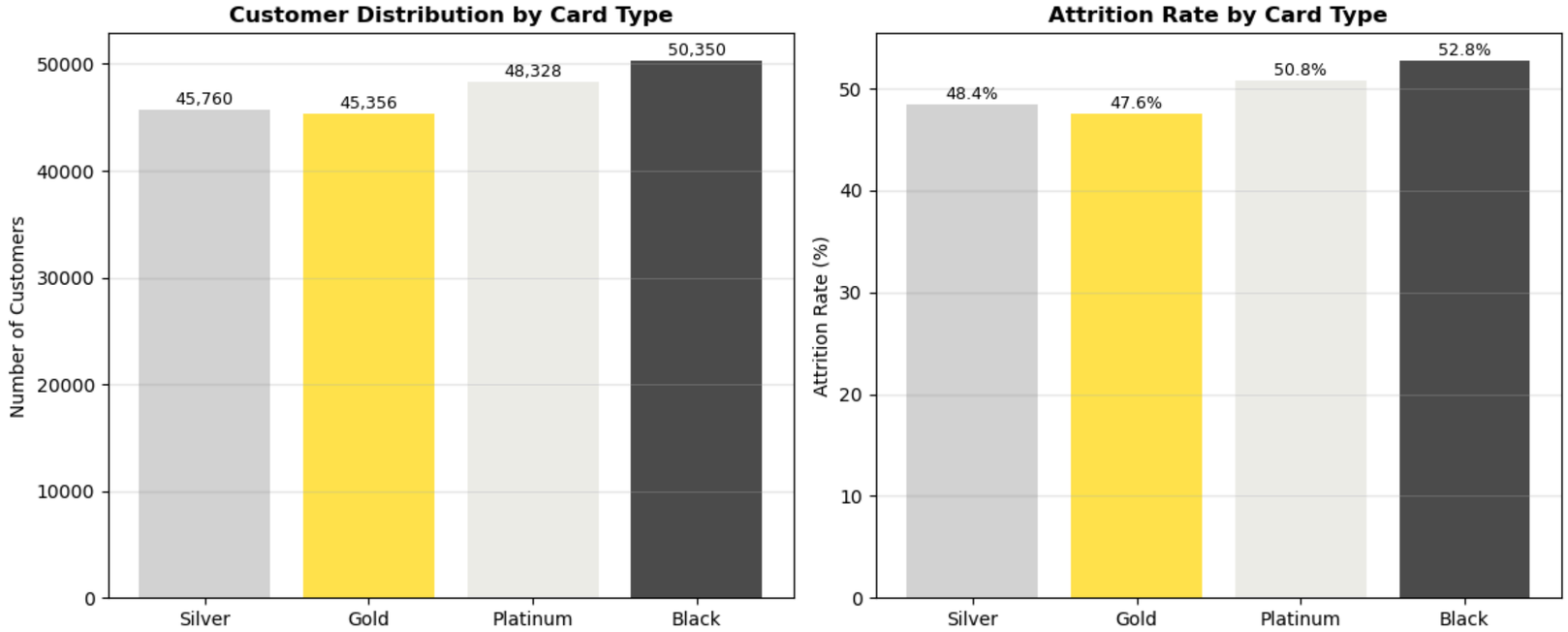
Recommendations:

- **Milestone Rewards** - Special benefits at 6, 9, and 12-year anniversaries
- **Re-engagement Campaign** - Personalized offers for 6-12 year customers showing reduced activity

Why it will work: Customers in the 6-12 year range have likely paid off initial purchases/debts and are reassessing their card portfolio. Intervention before year 6 is critical.

INSIGHTS & RECOMMENDATIONS

CARD TYPE ANALYSIS



Premium card holders show higher attrition - they may have more alternatives and higher expectations.

Recommendations:

- **Premium Perks Audit -** Survey Black/Platinum holders about unmet expectations
- **Exclusive Experiences -** Add unique benefits competitors can't match (concert access, concierge upgrades)

Why it will work: Premium customers have more options and higher expectations. They're leaving because the premium doesn't feel premium enough compared to competitors.

THANK YOU!

PREPARED BY: DARIUS ARDALES