Predicting Bank Account Balance from Customer Data

Greyson, Lan, Long

Framing

- Why is predicting bank account balance important?
  - **Insights** for banks to track deposit engagement and keep customers.
  - Banks are **interested in clients with high balances**

- Analyzed customer account balance using data from 10,000 customers
  - ...from several anonymous banking institutions

- 2 approaches: regression with ANOVA, and XGBoost with SHAP

Background: Data set

- **Source:** Kaggle
- **Prediction Goal:** Customer's account balance — a continuous financial outcome of direct relevance to banking strategy.
- **Synthetic data:** The data describe customers of a bank and were synthesized from a real dataset
  - Unavoidable in banking scenarios
  - ⩾95% utility in ML tasks compared to raw data

Objectives

1. Which customer characteristics are most strongly associated with account balance at the bank?

2. How well can variables such as credit score, age, tenure, geography, product usage, and active membership explain differences in customer account balance?

3. Do demographic and account-related factors—geography, gender, number of products—significantly affect customer balance?

4. Are more engaged customers (active members, credit card holders) likely to have higher balancesthan less engaged customers?

Background: Variables

Variable	Definition	Importance
CreditScore	The customer's credit score.	Higher credit scores suggest better financial reliability.
Age	The customer's age.	Compare balance across age groups.
Tenure	How many years the customer has been with the bank.	Show whether longer-term customers have different balance patterns.
Balance	The amount of money in the customer's bank account.	This is the response variable.
EstimatedSalary	The customer's estimated annual salary.	Test whether income level is related to account balance.

Table 1: Quantitative variables used in the analysis.

Background: Variables

Variable	Definition	Importance
Geography	The customer's country: France, Germany, or Spain.	Compare balance across countries.
Gender	The customer's gender: male or female.	Compare balance across gender groups.
NumOfProducts	The number of banking products or services the customer has with the bank.	Show whether using more services is related to account balance.
HasCrCard	Whether the customer has a credit card with the bank.	Compare customers with and without a credit card.
IsActiveMember	Whether the customer is an active member.	Show whether more active customers have different balances.
Exited	Whether the customer left the bank or stayed.	Compare past balance patterns between customers who stayed and left.

Table 2: Categorical variables used in the analysis.

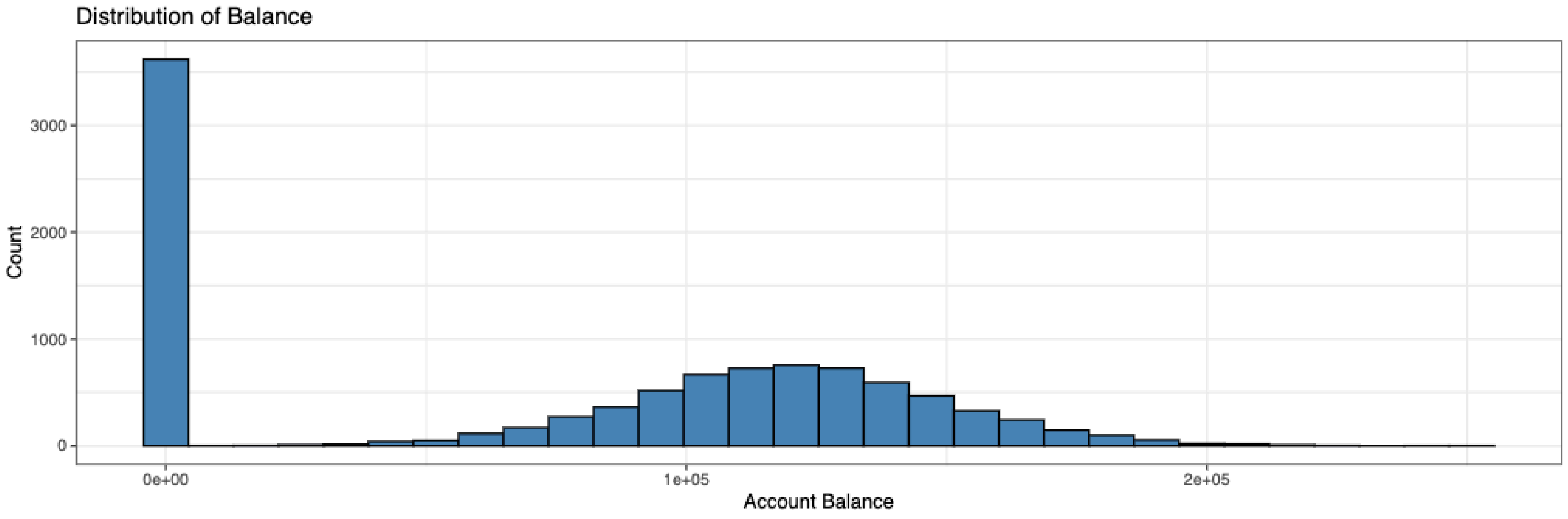
Exploratory Data Analysis

Let's first look at the summary statistics for the main quantitative variables in the dataset.

Parameter	Min	Q1	Median	Mean	Q3	Max
Balance	0.00	0.00	97198.54	76485.89	127644.24	250898.09
CreditScore	350.00	584.00	652.00	650.53	718.00	850.00
Age	18.00	32.00	37.00	38.92	44.00	92.00
Tenure	0.00	3.00	5.00	5.01	7.00	10.00
EstimatedSalary	11.58	51002.11	100193.91	100090.24	149388.25	199992.48

Table 3: Numerical summaries for Balance and quantitative predictors.

- The data is widespread.
- Customer characteristics vary



- 6884 customers have their account balance = 0.
- Without 0, it's a normal distribution.

Balance × Categorical

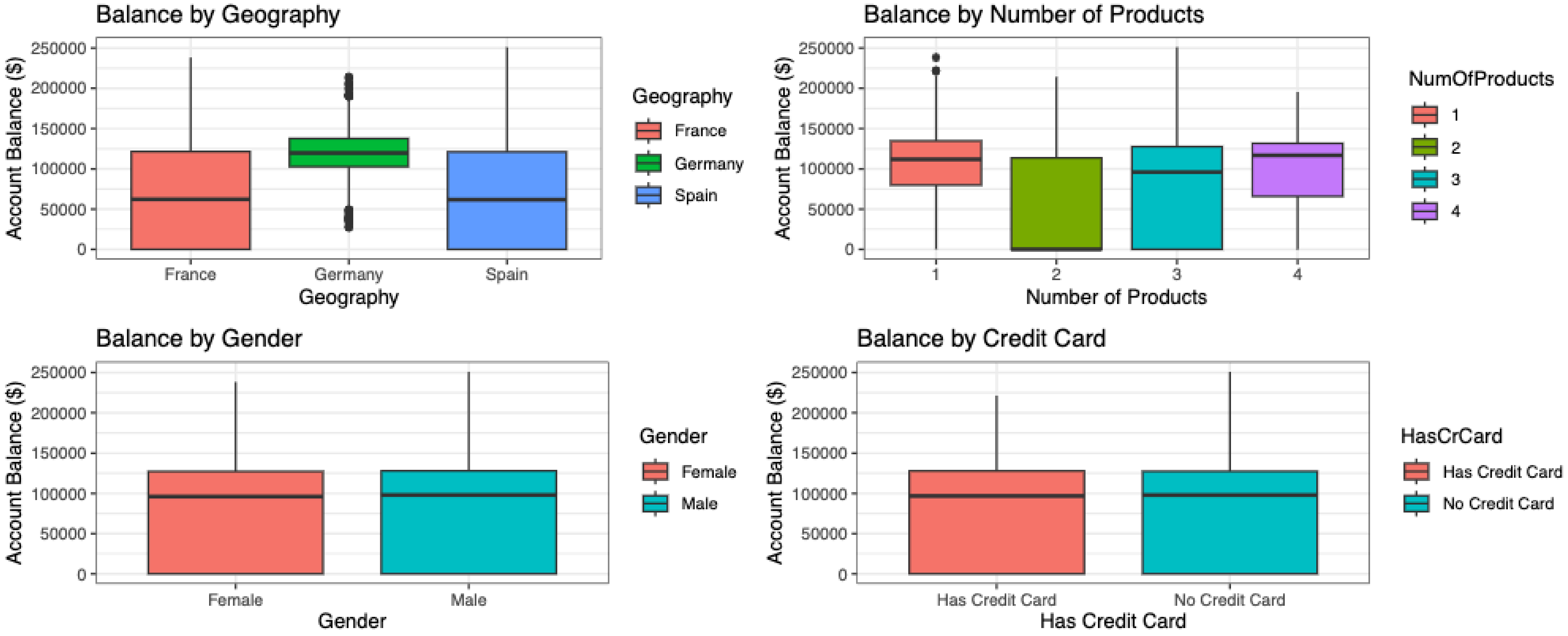


Figure 4: Account balance by geography, number of products, gender, and credit card status.

- **Geography:** Germany customers tend to have higher balances.
- **NumOfProducts:** important but non-linear relationship with balance.
- **Gender & HasCrCard:** weaker differences, less predictive.

Balance × Quantitative

- Correlations between balance and credit score, tenure, and estimated salary are all close to 0.
- Age has the largest correlation with Balance but still very small.

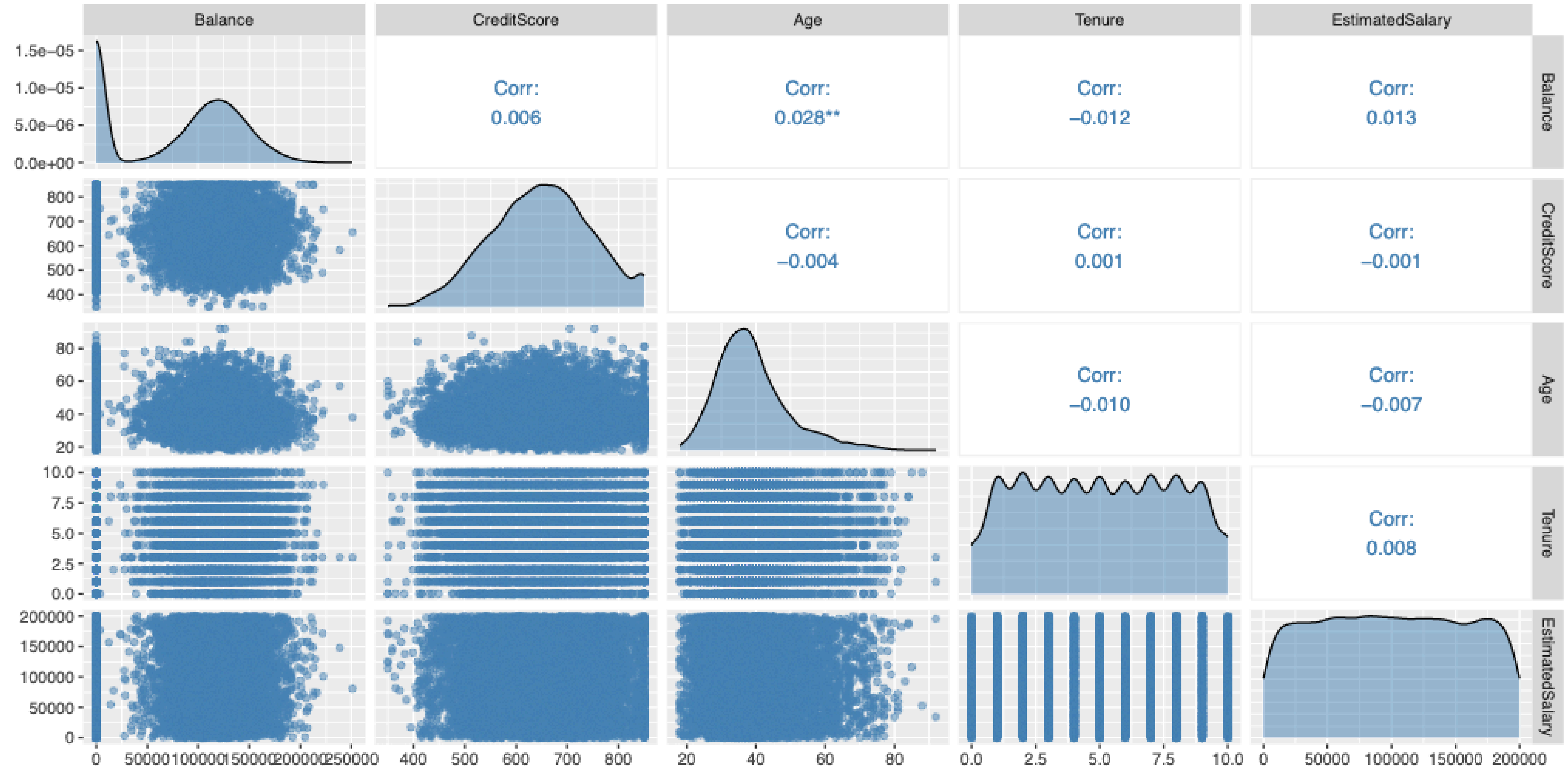


Figure 5: Correlogram of quantitative variables. Age has the strongest correlation with Balance (r 0.29); all others are near zero.

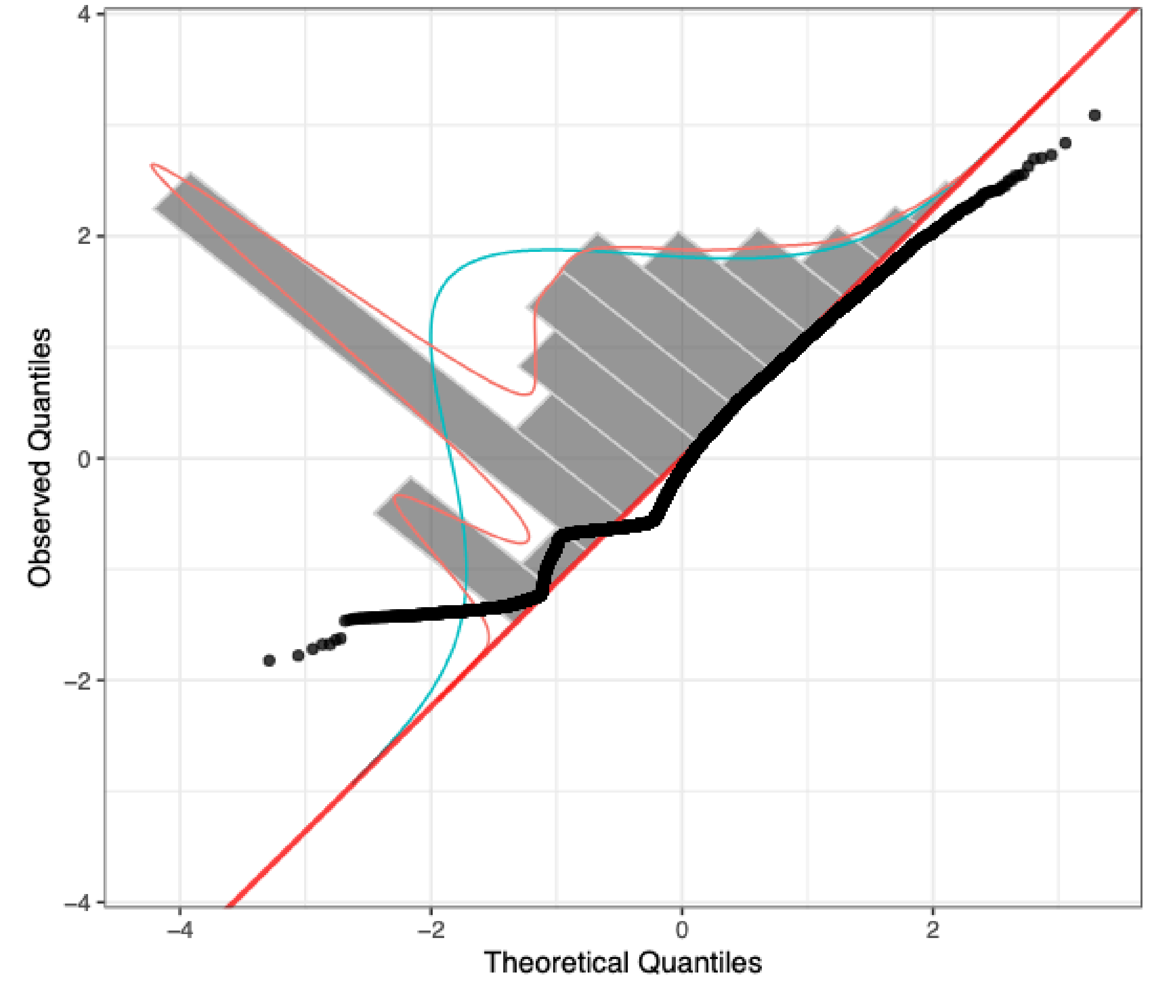
Correlogram

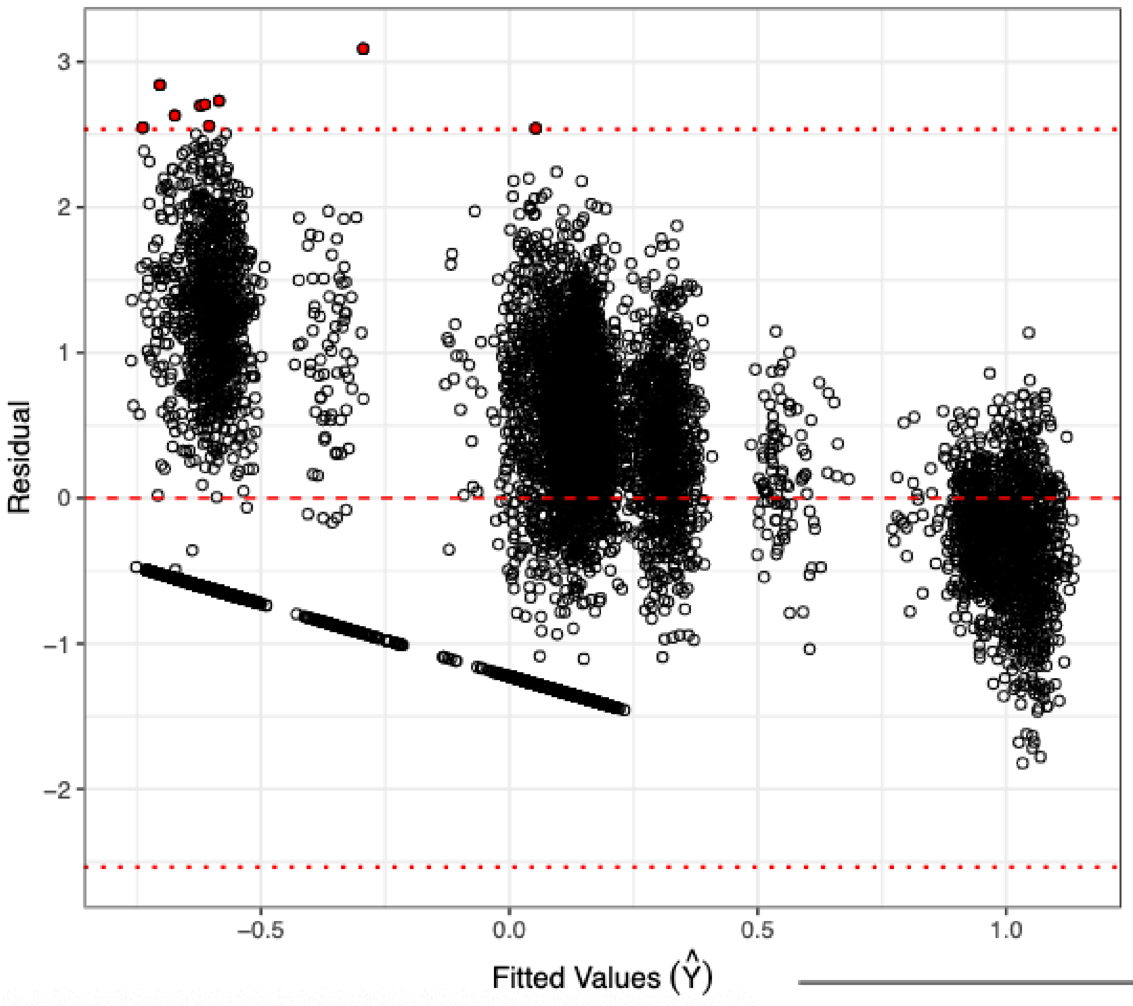
First-Order Model

	Estimate	Std. Error	t value	p-value	2.5%	97.5%
(Intercept)	0.0278	0.0280	0.9911	0.3217	-0.0271	0.0827
CreditScore_z	0.0086	0.0085	1.0171	0.3091	-0.0080	0.0252
GeographyGermany	0.9046	0.0210	43.1422	< 0.0001	0.8635	0.9457
GeographySpain	0.0016	0.0208	0.0762	0.9393	-0.0391	0.0423
GenderMale	0.0390	0.0171	2.2806	0.0226	0.0055	0.0725
Age_z	-0.0135	0.0089	-1.5124	0.1305	-0.0309	0.0040
Tenure_z	-0.0079	0.0085	-0.9373	0.3486	-0.0245	0.0087
NumOfProducts2	-0.7317	0.0178	-41.0477	< 0.0001	-0.7666	-0.6967
NumOfProducts3	-0.4045	0.0546	-7.4019	< 0.0001	-0.5116	-0.2974
NumOfProducts4	-0.1285	0.1111	-1.1566	0.2475	-0.3462	0.0893
HasCrCardNo Credit Card	0.0388	0.0186	2.0917	0.0365	0.0024	0.0752
IsActiveMemberInactive Member	-0.0063	0.0173	-0.3649	0.7152	-0.0402	0.0276
EstimatedSalary_z	0.0121	0.0085	1.4316	0.1523	-0.0045	0.0287
ExitedStayed	0.0787	0.0244	3.2237	0.0013	0.0308	0.1265

Table 4: Inferential results for the first-order model of standardized Balance.

ANOVA





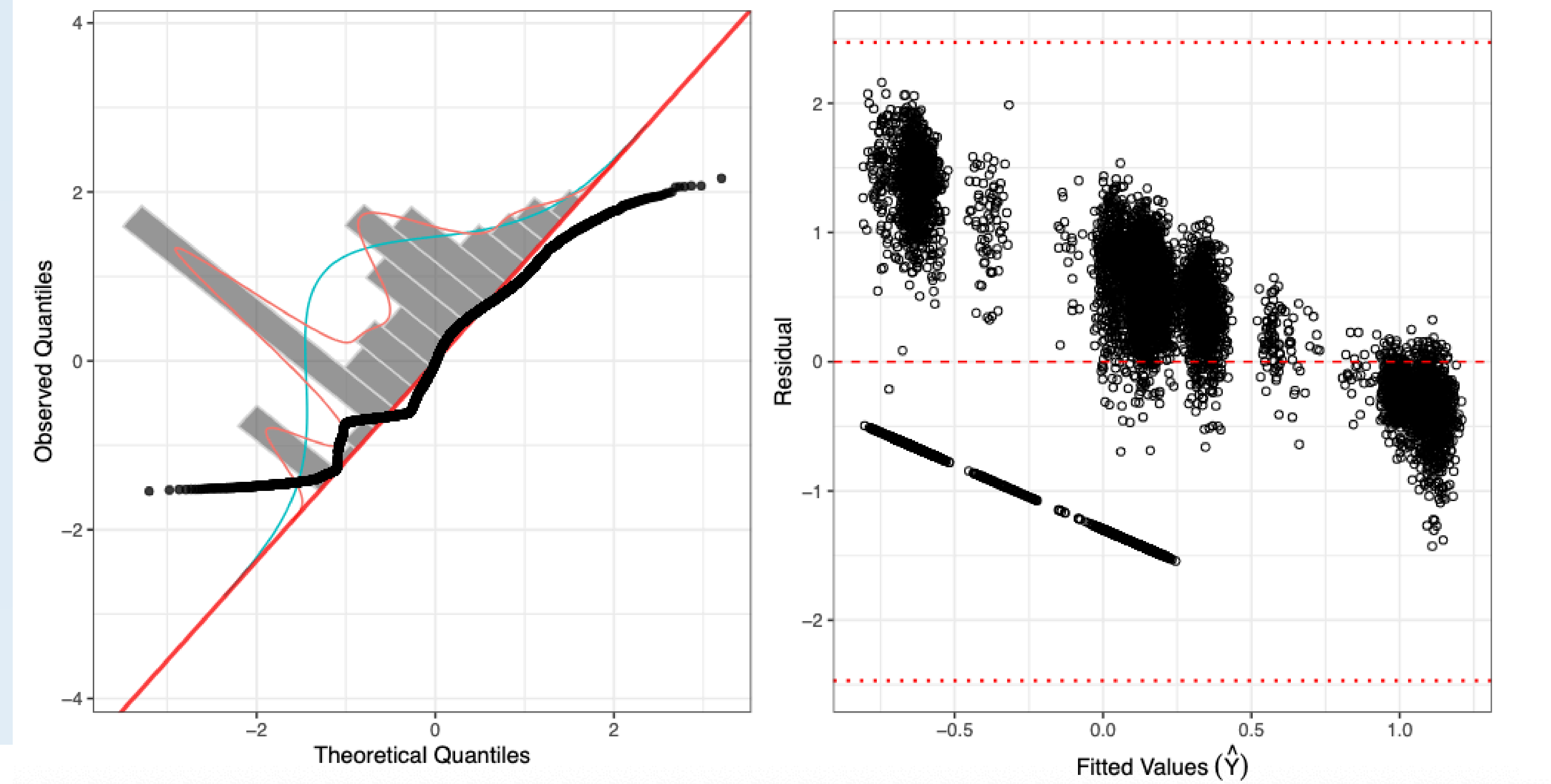
- Both normality and constant variance conditions of the residuals are violated due to the excessive amount of 0-balance customers.

- Member Status is least important to predict Balance.

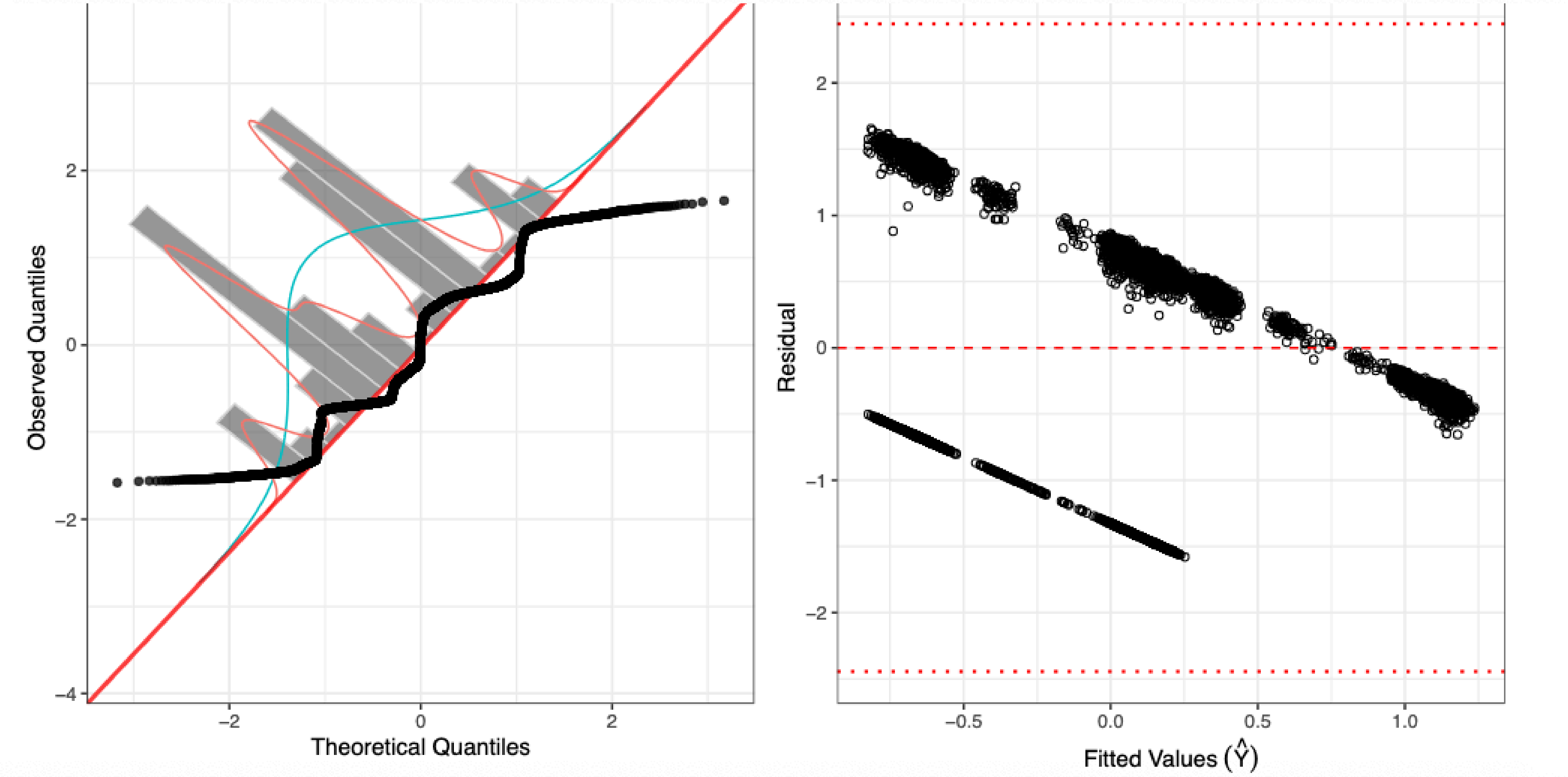
	Sum Sq	Df	F value	p-value
(Intercept)	0.70	1	0.98	0.3217
CreditScore_z	0.74	1	1.03	0.3091
Geography	1489.17	2	1041.87	< 0.0001
Gender	3.72	1	5.20	0.0226
Age_z	1.63	1	2.29	0.1305
Tenure_z	0.63	1	0.88	0.3486
NumOfProducts	1215.06	3	566.73	< 0.0001
HasCrCard	3.13	1	4.38	0.0365
IsActiveMember	0.10	1	0.13	0.7152
EstimatedSalary_z	1.46	1	2.05	0.1523
Exited	7.43	1	10.39	0.0013
Residuals	7136.63	9986		

Table 5: Type III ANOVA table for the first-order model.

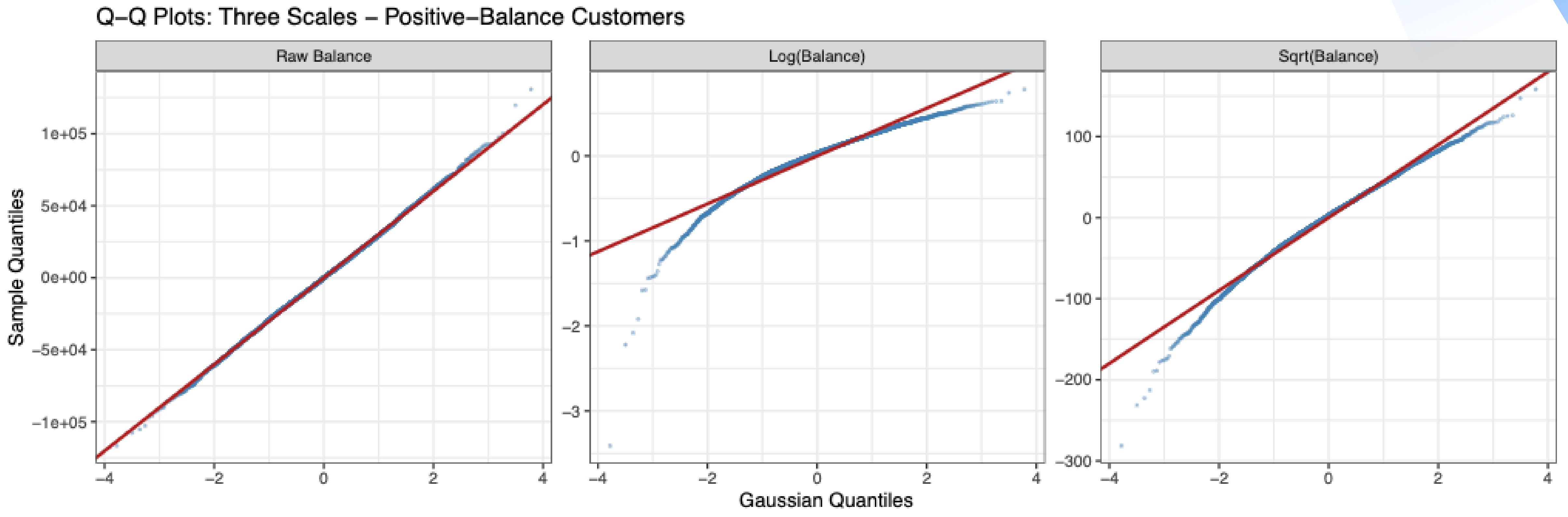
Log Transformation

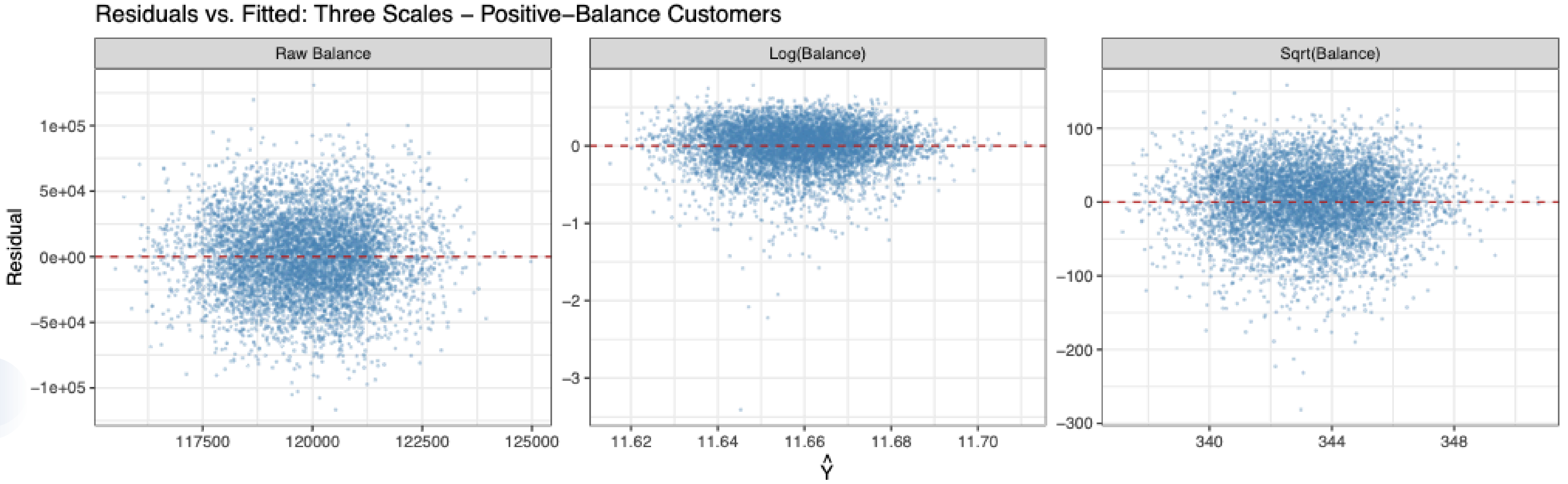


Square-root Transformation



0-Filtered Balance Transformation Residuals Plots





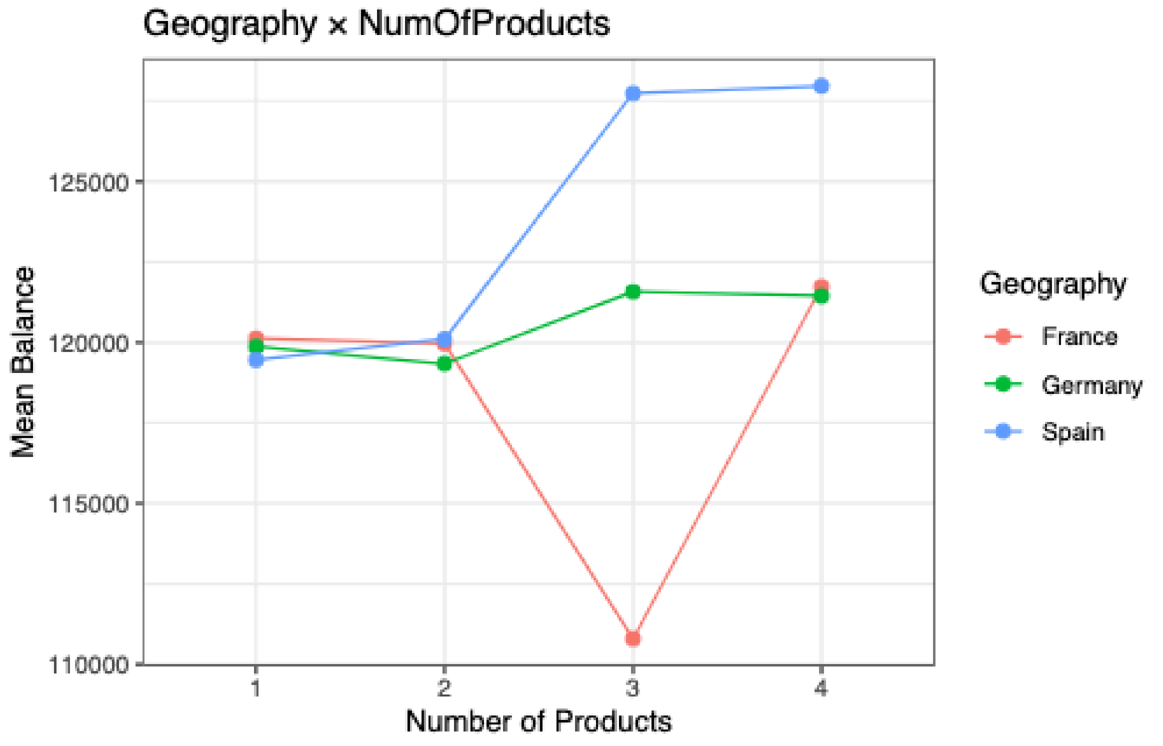
The AIC models have slightly higher R^2 and $R^2{}_{adj}$ (both are **moderate**), but they also use more parameters. The BIC models have the same fit with less parameters → **easier to interpret.**

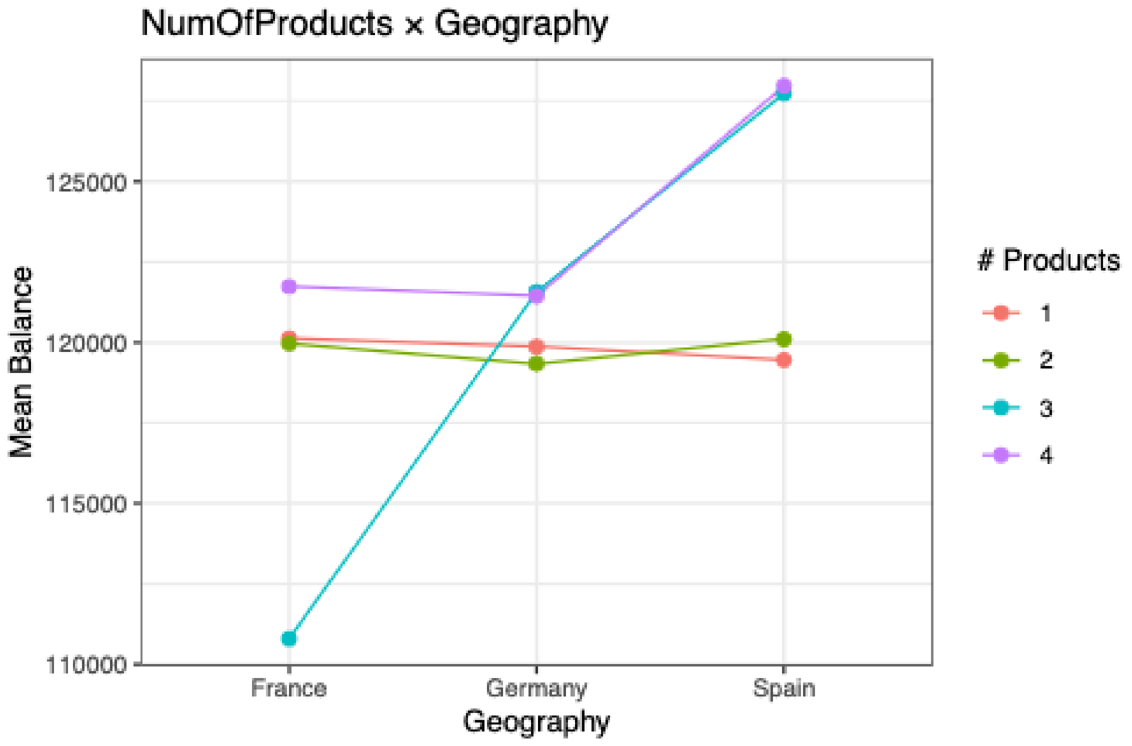
Metric	Model	R^2	R^2_{adj}	LL	Value	RMSE	MAE	Parameters
AIC	One-way Backward	0.2861	0.2854	-12503.707	25031.41	0.8449	0.7176	11
	One-way Forward	0.2861	0.2854	-12503.707	25031.41	0.8449	0.7176	11
	One-way Stepwise	0.2861	0.2854	-12503.707	25031.41	0.8449	0.7176	11
BIC	One-way Backward	0.2851	0.2847	-12510.583	25094.85	0.8455	0.7181	7
	One-way Forward	0.2851	0.2847	-12510.583	25094.85	0.8455	0.7181	7
	One-way Stepwise	0.2851	0.2847	-12510.583	25094.85	0.8455	0.7181	7

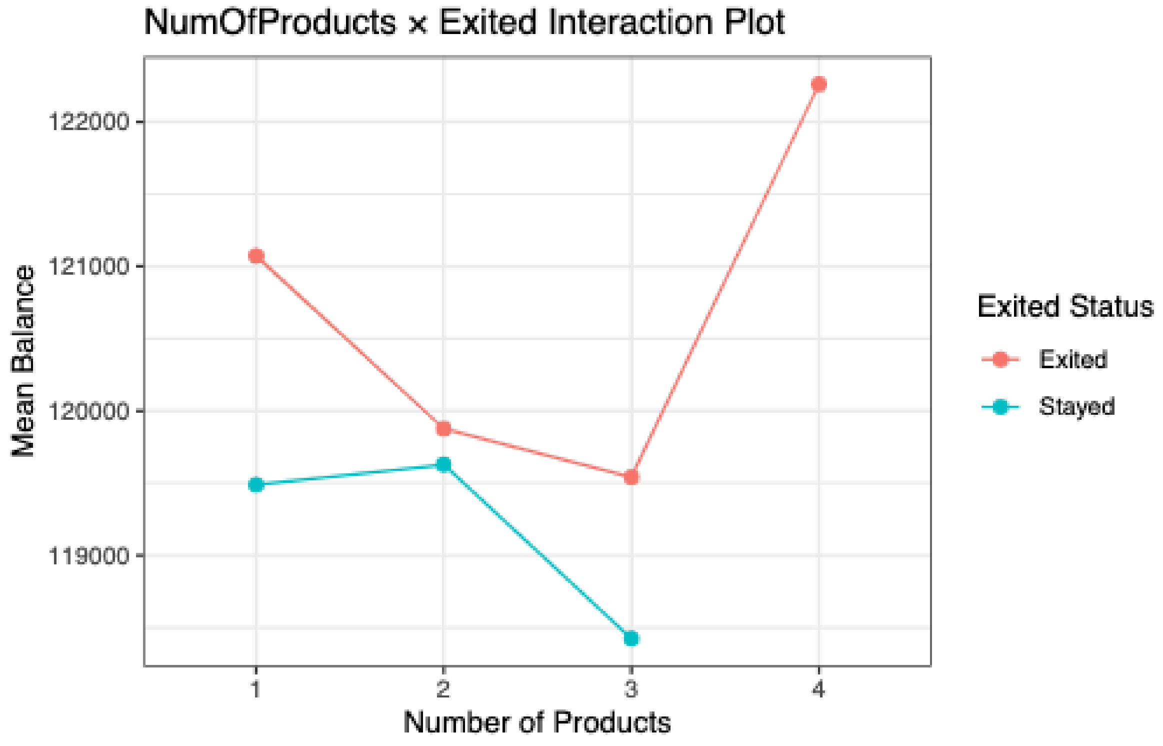
Table 8: Model comparison using AIC and BIC selection.

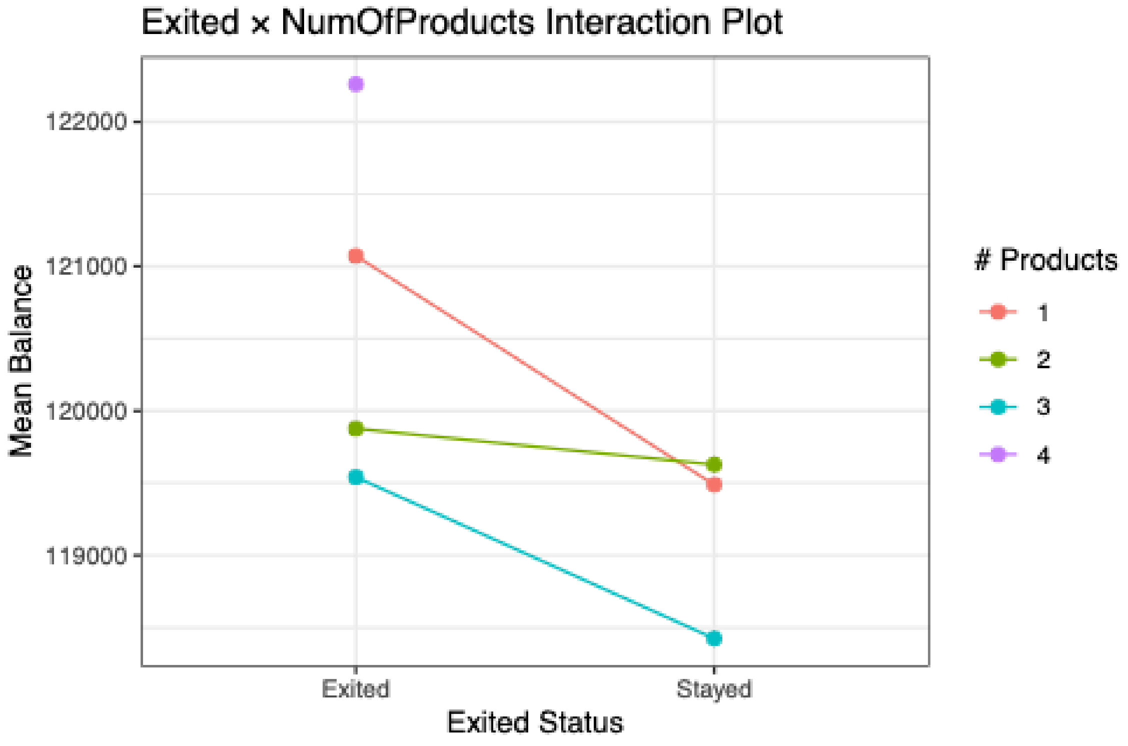
MODEL SELECTION SUMMARY

CATEGORY × CATEGORY INTERACTION



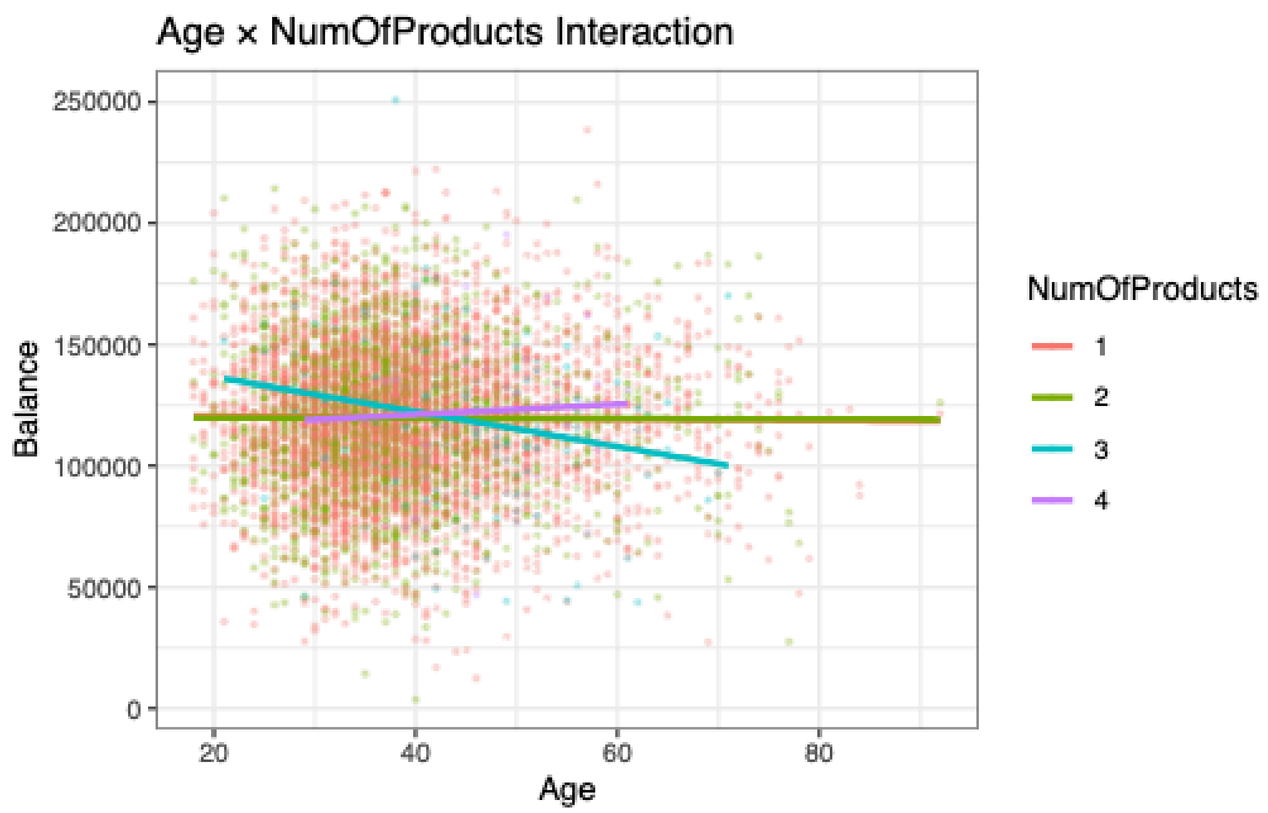


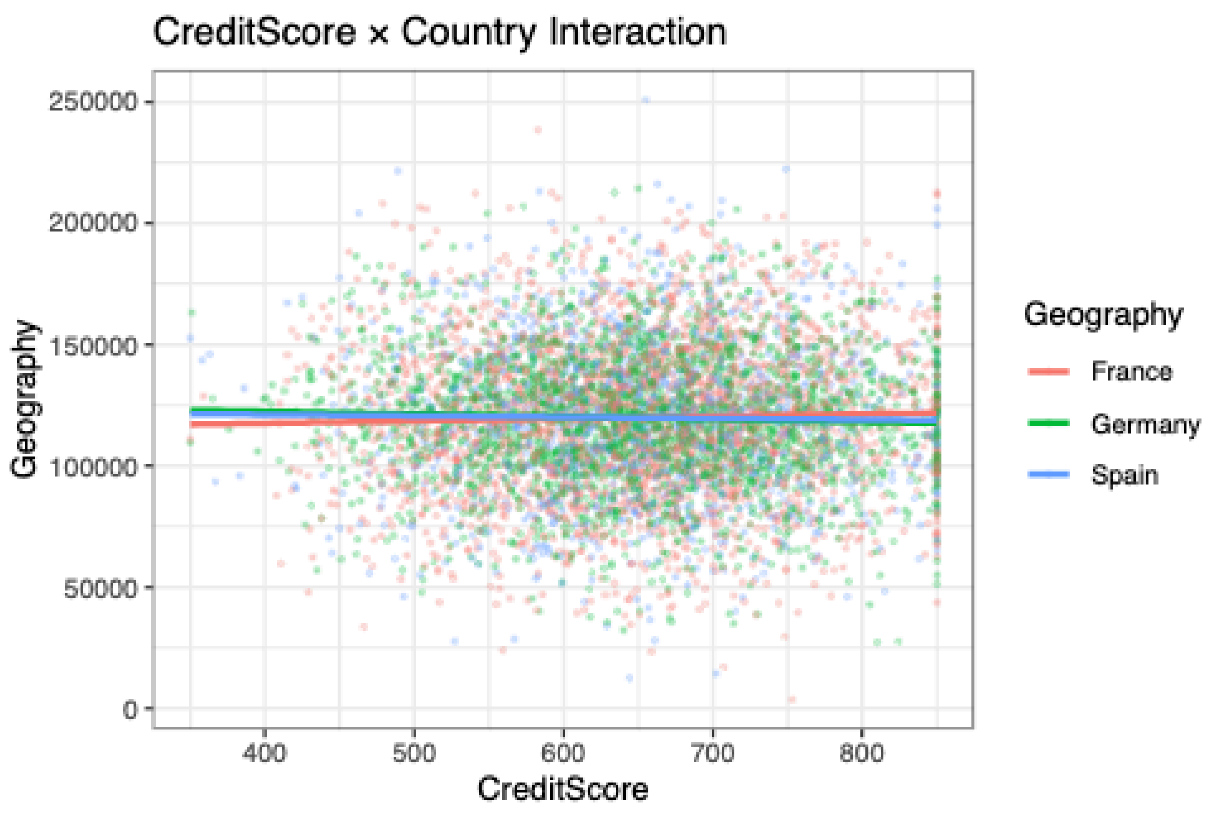




Non-parallel lines indicates interactions between **Geography** and **Number of Products** as well as between **Exit Status** and **Number of Products.**

- No clear interactions between 1 and 2 in #Products.
- Almost little to no interactions among the countries.





INTERACTION SUMMARY

Significance	Interaction	Estimate	Std. Error	t value	p-value
Very Strong	Germany × NumProducts(2)	0.8859	0.0434	20.412	<0.001
Very Strong	Germany × NumProducts(3)	0.4715	0.1231	3.830	<0.001
Very Strong	NumProducts(2) × Stayed	-0.7018	0.0552	-12.726	<0.001
Very Strong	NumProducts(3) × Stayed	-1.0273	0.1458	-7.046	<0.001
Strong	CreditScore × Germany	-0.0433	0.0201	-2.156	0.031
Strong	CreditScore × NumProducts(4)	0.2155	0.1067	2.020	0.043
Strong	Germany × HasCrCard	-0.0957	0.0448	-2.135	0.033
Strong	Age × NumProducts(2)	0.0064	0.0180	0.354	0.004
Borderline	NumProducts(3)	-0.2282	0.1346	-1.695	0.090
Borderline	CreditScore × Spain	-0.0284	0.0204	-1.393	0.164

QUANTITATIVE × CATEGORY INTERACTION

FINAL MODEL SELECTION

- About the same fit and adjusted R squared with AIC models.
- Convergent agreement across three selection methods.
- Easier to interpret and prevent overfitting with fewer parameters.

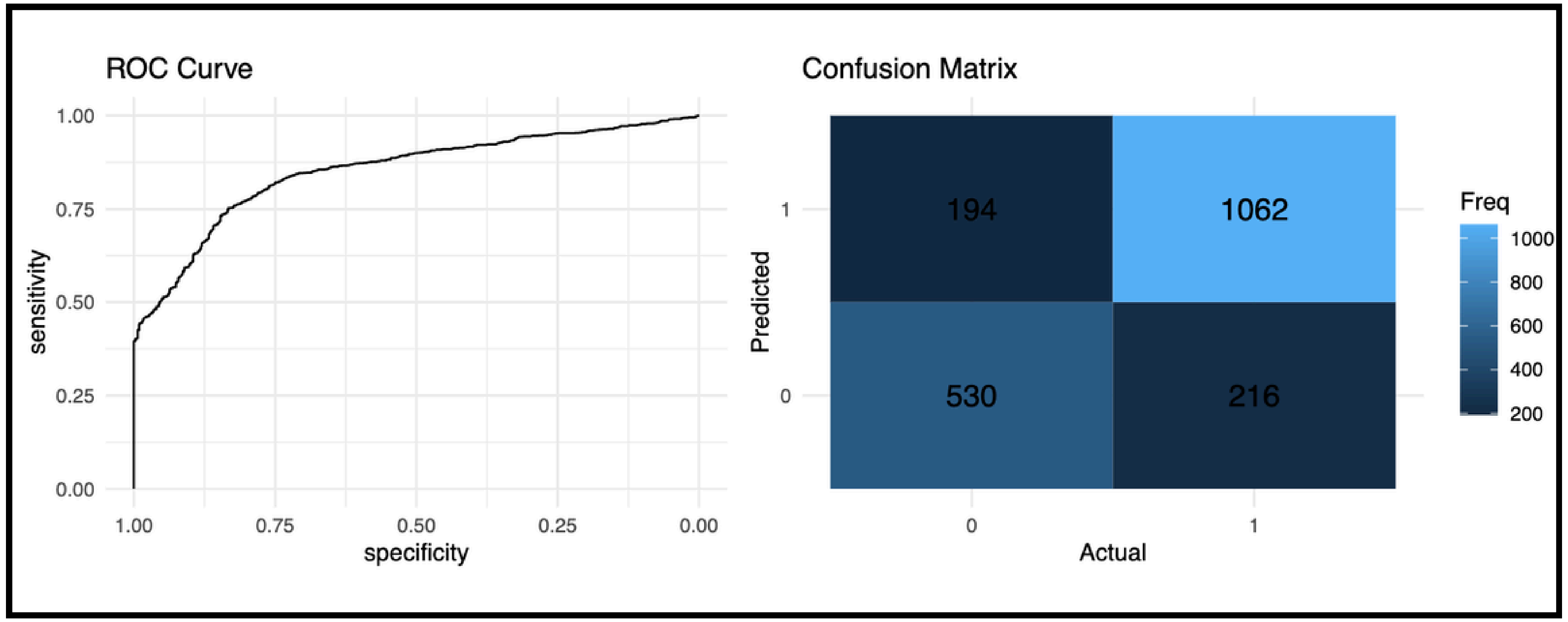
We apply backward elimination, forward selection, and stepwise selection using both AIC and BIC to the full two-way interaction model. Table 12 summarizes the results.

Metric	Model	R^2	R^2_{adj}	LL	Value	RMSE	MAE	Parameters
AIC	Interaction Backward	0.3406	0.3388	-12106.531	24271.06	0.8120	0.6516	29
	Interaction Forward	0.3404	0.3388	-12107.953	24269.91	0.8121	0.6518	27
	Interaction Stepwise	0.3404	0.3388	-12107.953	24269.91	0.8121	0.6518	27
BIC	Interaction Backward	0.3383	0.3374	-12124.334	24396.03	0.8134	0.6536	16
	Interaction Forward	0.3383	0.3374	-12124.334	24396.03	0.8134	0.6536	16
	Interaction Stepwise	0.3383	0.3374	-12124.334	24396.03	0.8134	0.6536	16

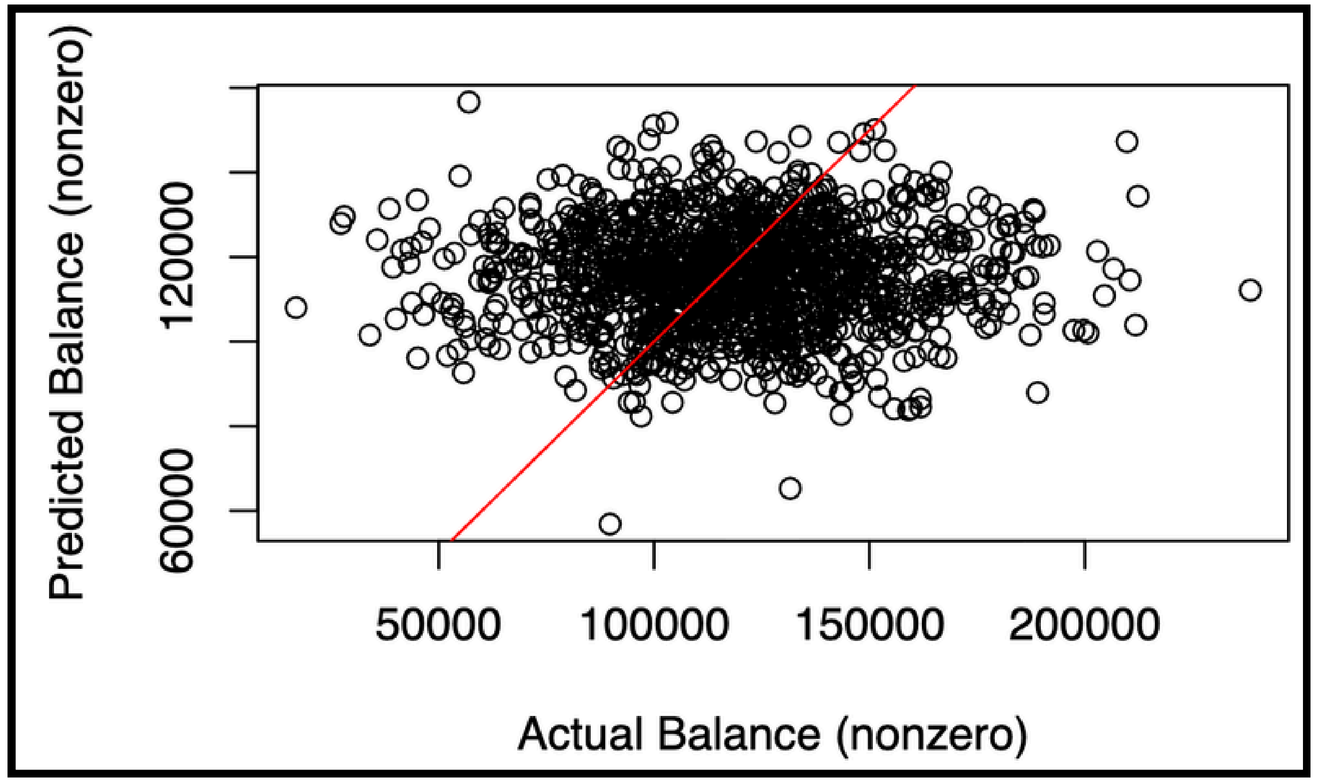
Table 12: Comparing interaction models using AIC and BIC selection

SHAP ANALYSIS: XGB MODELS

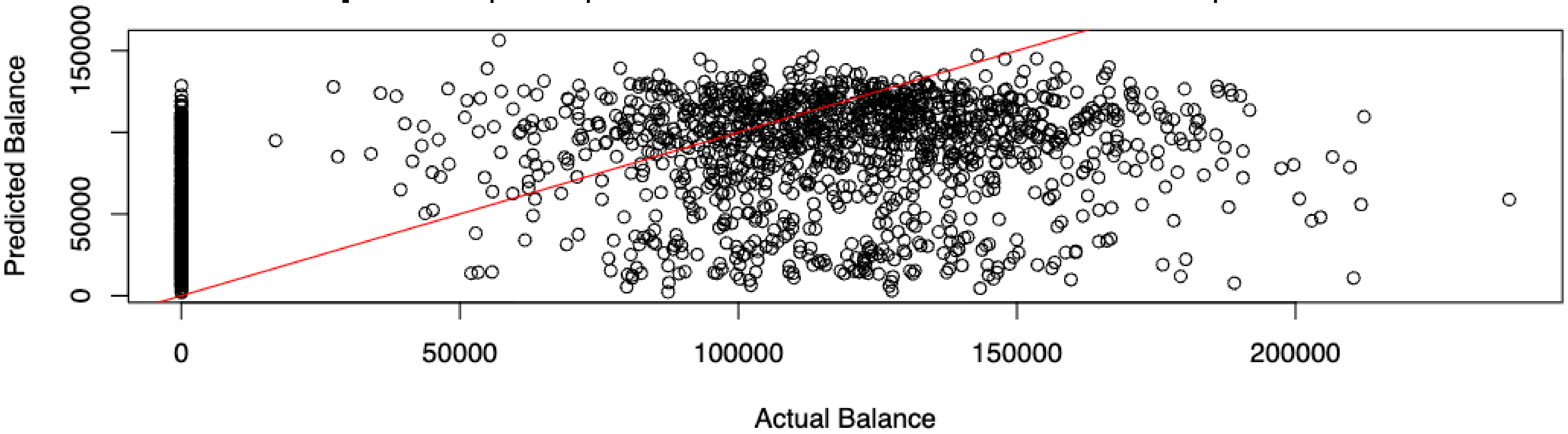
Binary classifier: good performance



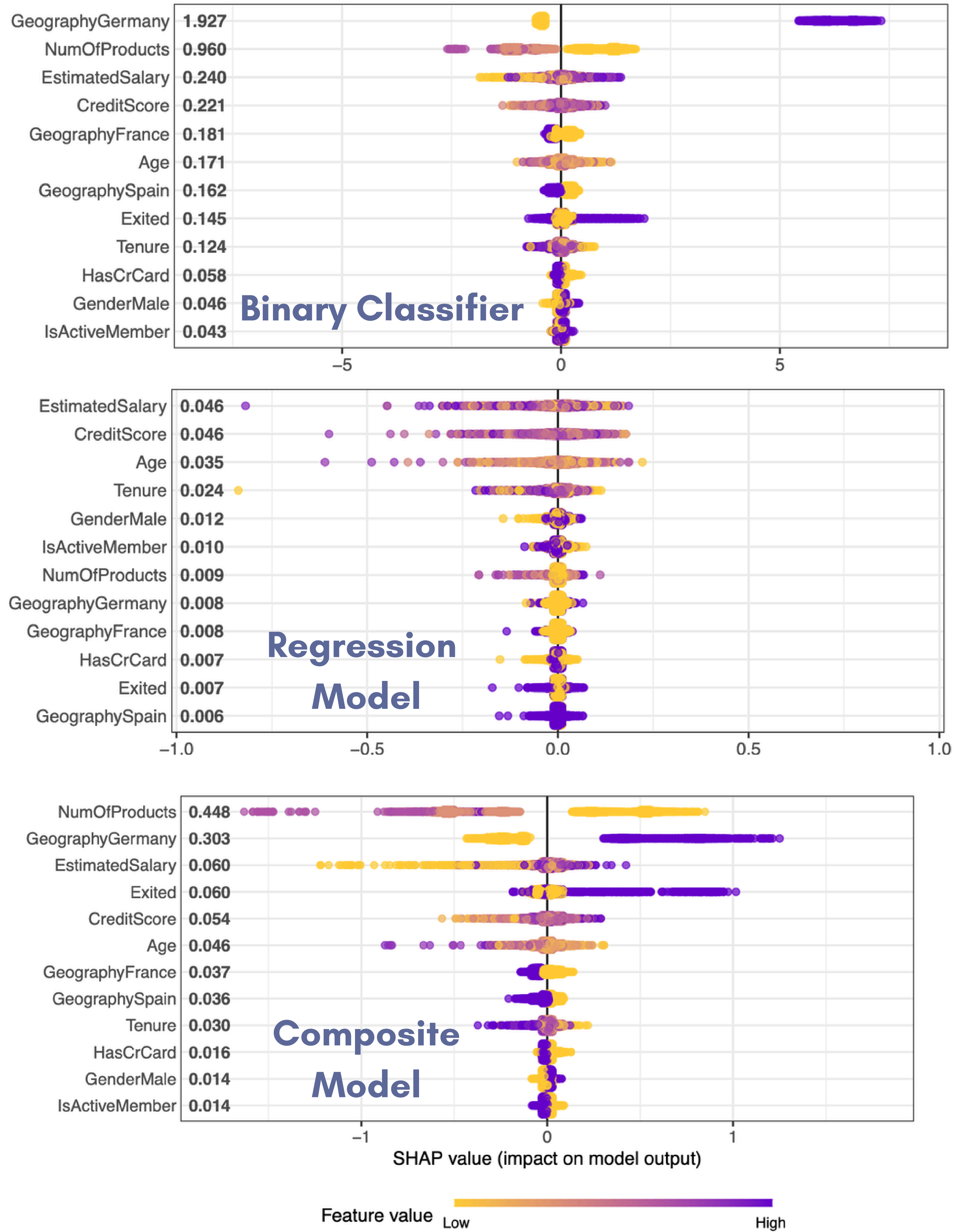
Regression: poor performance

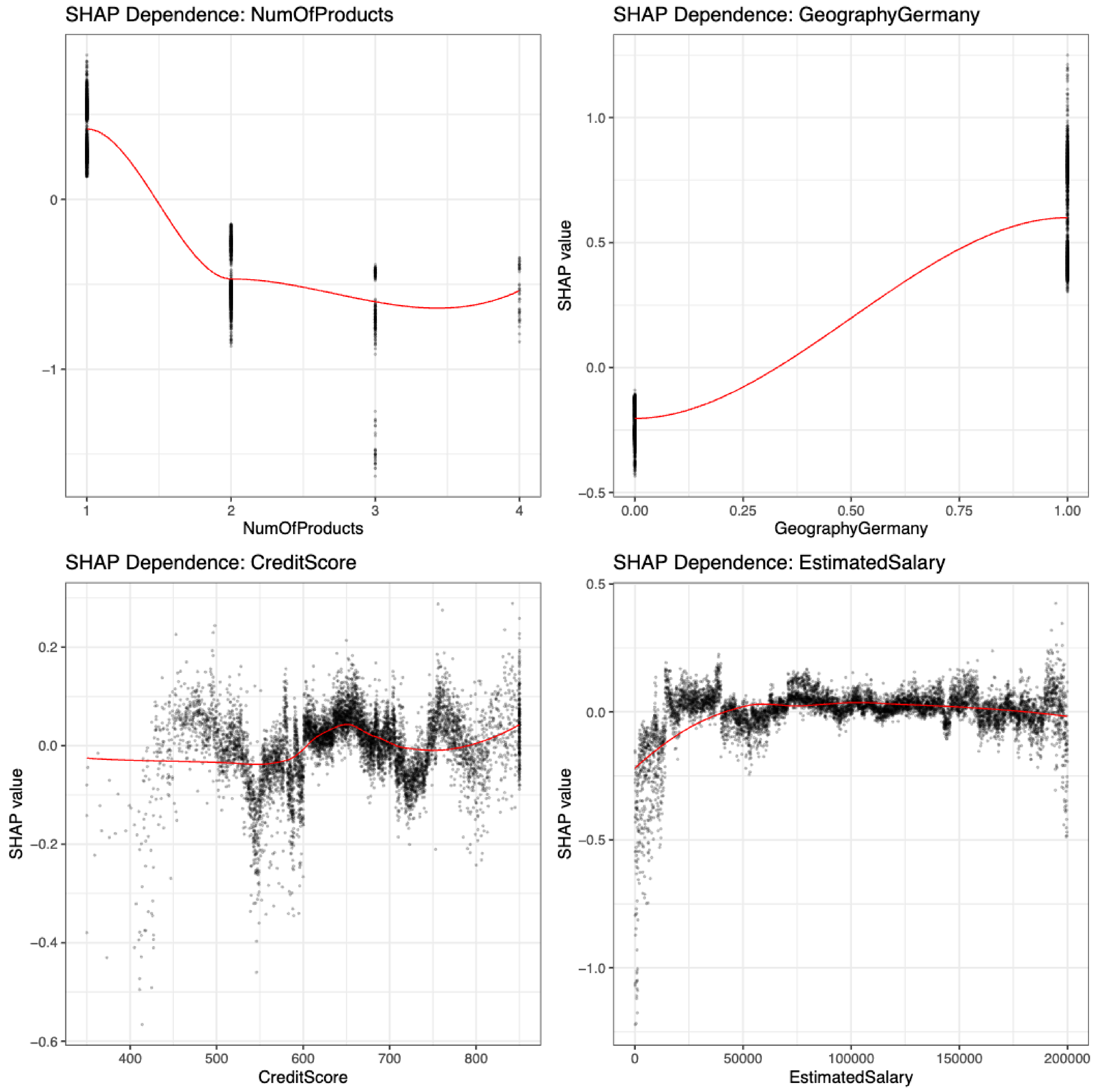


Composite: poor performance, but sensitive to real patterns



SHAP ANALYSIS: EXPLANATION





Limitations!!!

- Synthetic data
- Skew hurts prediction
- Underspecificity