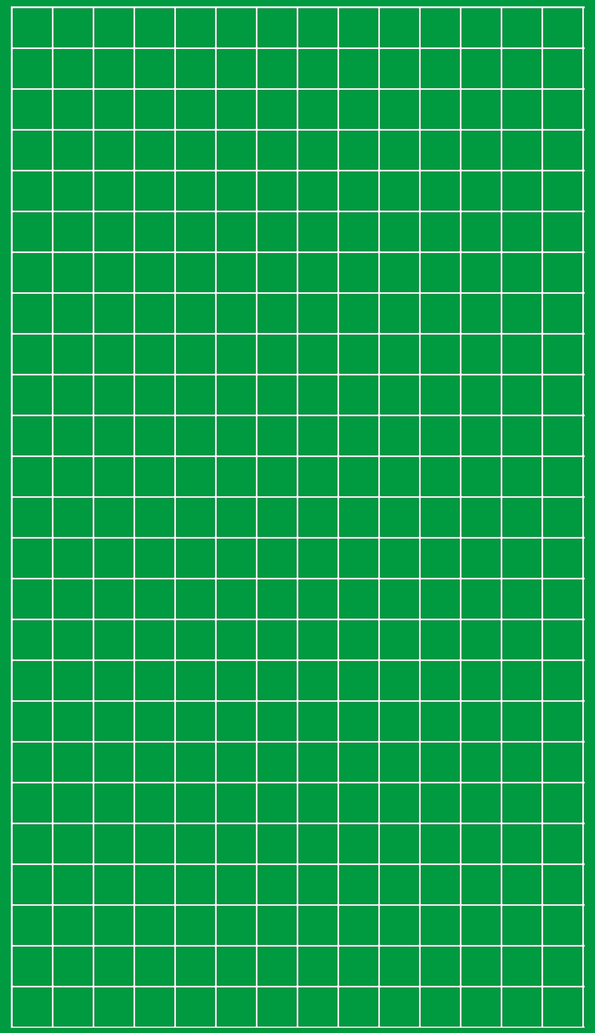
ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Bank Customer Churn Prediction

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ



Шаг первый

Сбор команды, выбор датасета, EDA

Описание датасета

IPII HSExMTS

Exited - целевая переменная (1 - ушёл, 0 - остался)
Решаем задачу бинарной классификации: предсказываем вероятность ухода клиента для тестового набора данных

Список фичей:

Название	Описание
Customer ID	Уникальный идентификатор клиента
Surname	Фамилия клиента
Credit Score	Кредитный рейтинг
Geography	Страна проживания (Франция, Испания, Германия)
Gender	Пол (Мужской, Женский)
Age	Возраст клиента
Tenure	Срок обслуживания в банке (в годах)
Balance	Баланс на счёте
NumOfProducts	Количество используемых банковских продуктов
HasCrCard	Наличие кредитной карты (1/0)
IsActiveMember	Активный клиент (1/0)
EstimatedSalary	Предполагаемая заработная плата

EDA

Данные у нас
достаточно чистые

Dataset statistics

Number of variables	11
Number of observations	15000
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%

Наш пространственный признак:

Geography

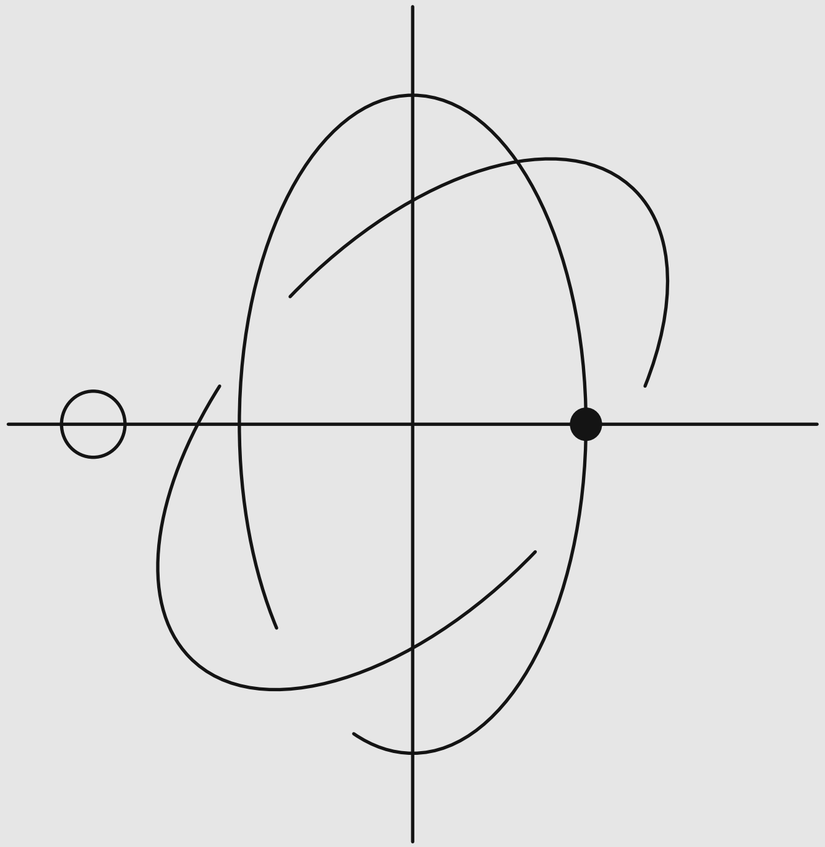
Categorical

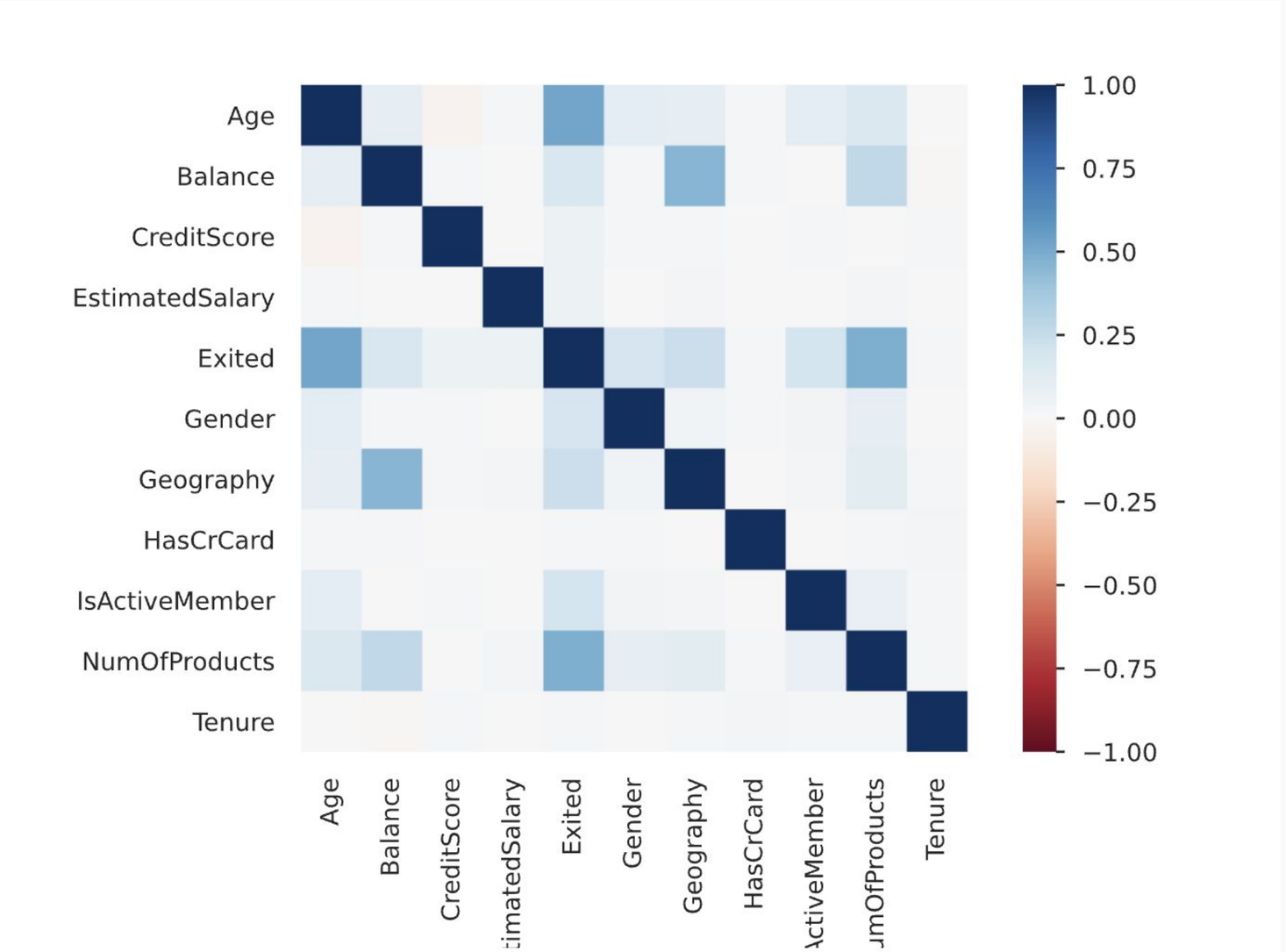
Distinct	3
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	117.3 KiB

More details

EDA

Корреляционная матрица показывает, что большинство признаков слабо связаны между собой, что снижает риск мультиколлинеарности

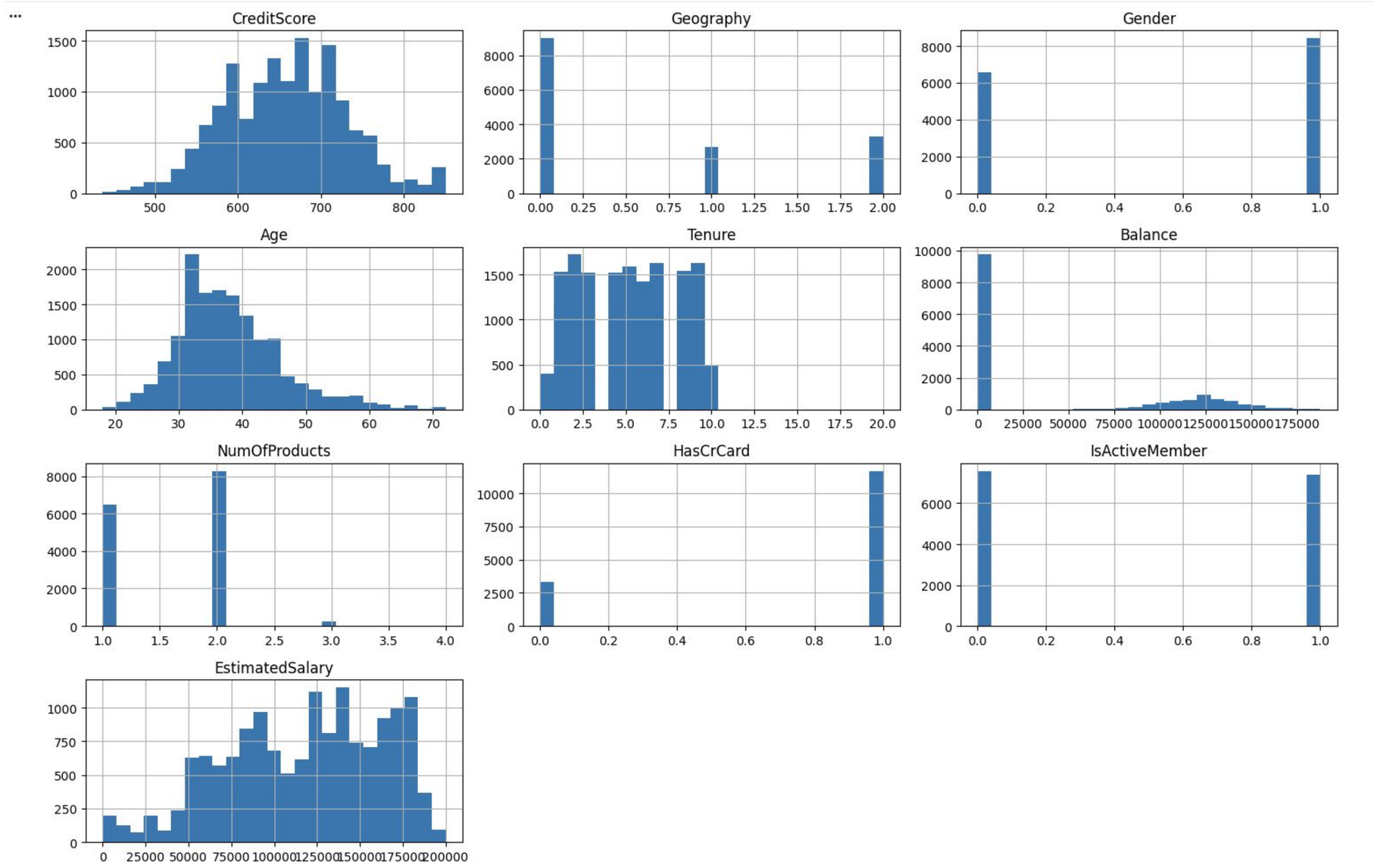


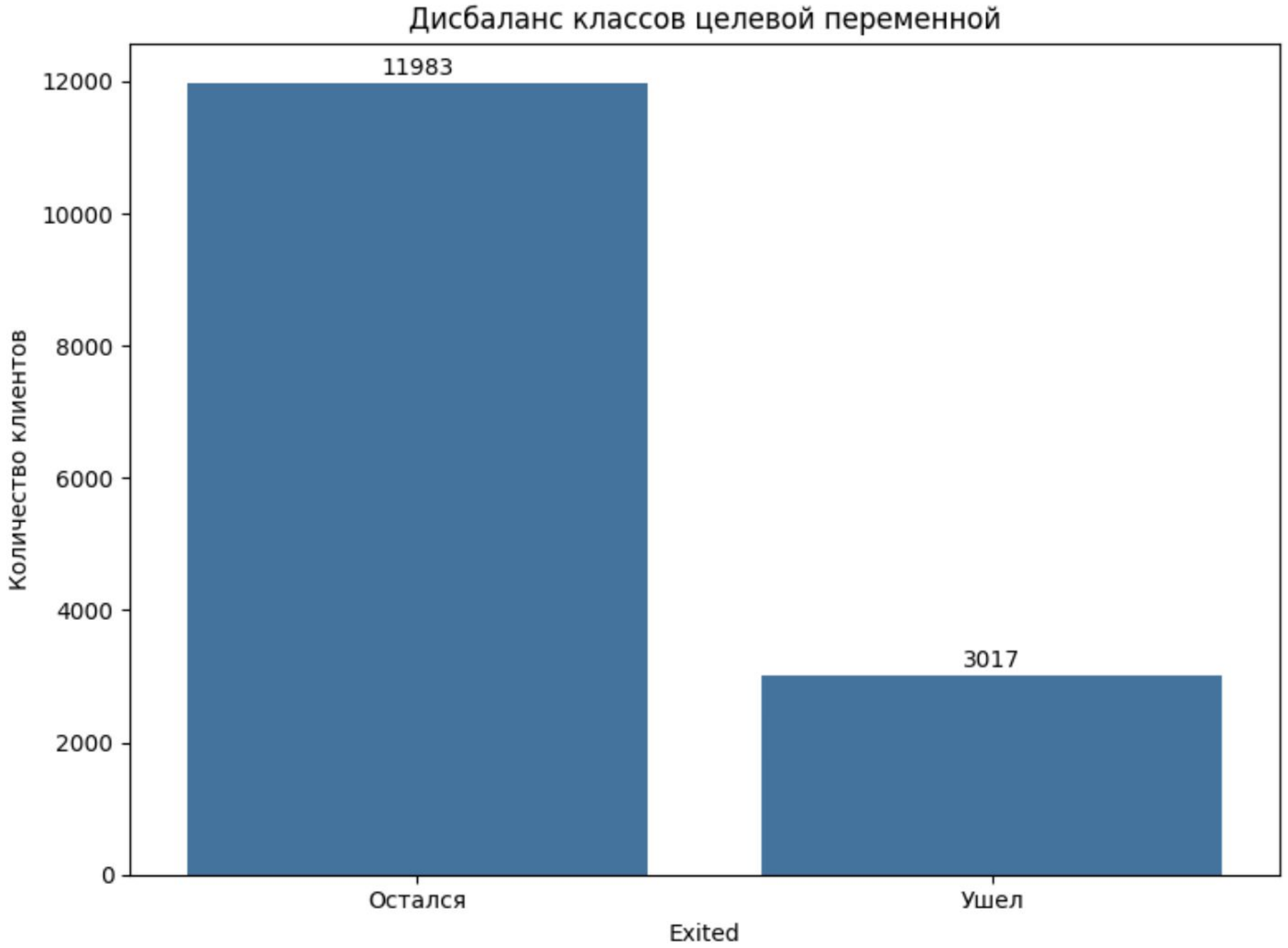


Невооружённым взглядом наблюдается высокая корреляция целевой переменной с NumOfProducts и Age - обратим на это внимание в дальнейшем

EDA

- **Числовые признаки** имеют гладкие распределения: CreditScore и Age сосредоточены вокруг средних значений, Balance — асимметричен из-за большого числа нулей, EstimatedSalary распределён относительно равномерно
- **Категориальные признаки** представлены ограниченным числом значений и в основном сбалансированы, за исключением Geography
- **Дискретные признаки** имеют компактные распределения: Tenure распределён достаточно равномерно, NumOfProducts в основном принимает значения 1–2





EDA

В целевой переменной наблюдается выраженный дисбаланс классов: большинство клиентов остаются, меньшая доля - уходит. Доля ушедших клиентов существенно ниже, что необходимо учитывать при обучении и оценке моделей

В качестве безлайна на этом этапе был построен sample_submission.csv, в котором вероятность ухода каждого клиента - ровно 0.5

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Шаг второй

Работа с аномалиями и выбросами

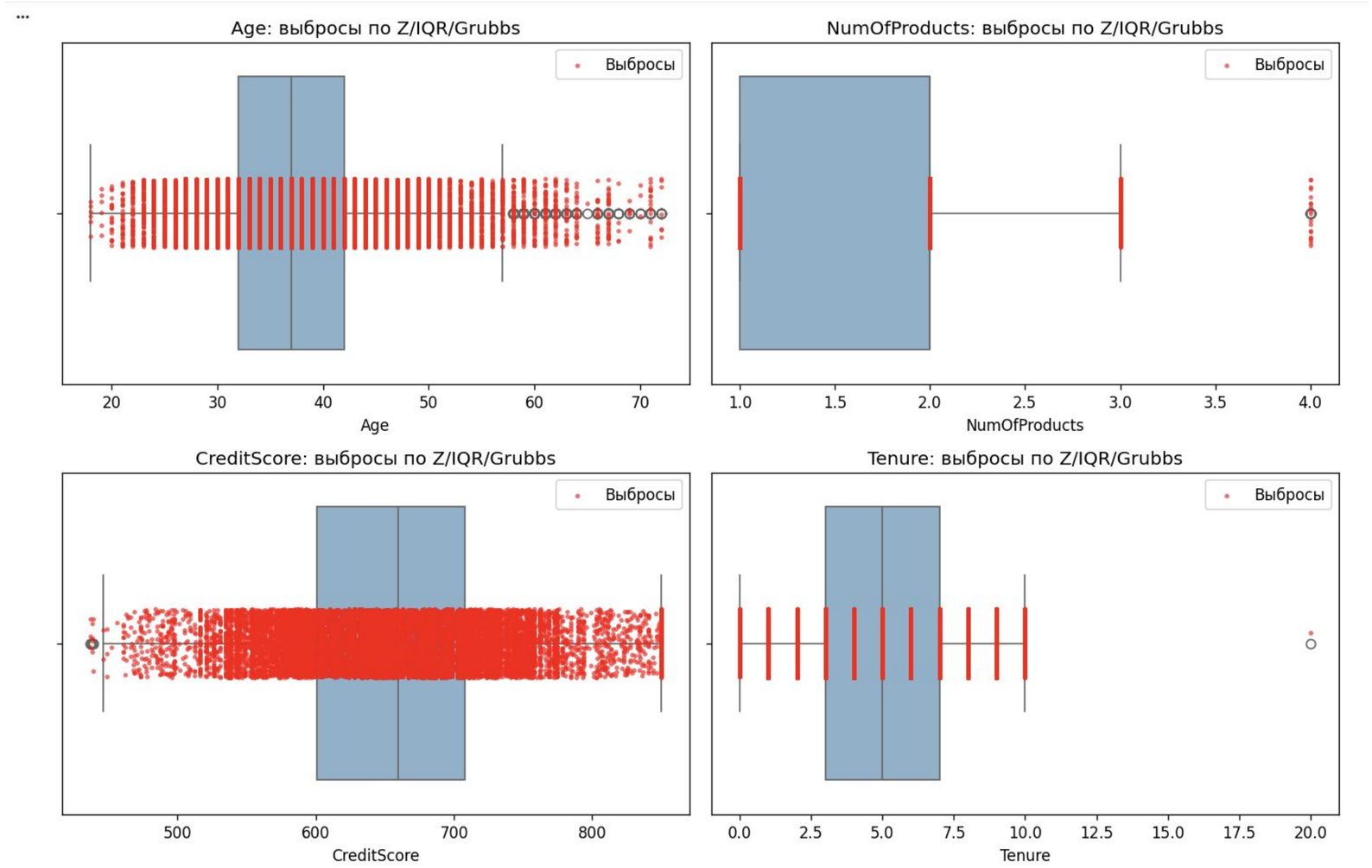
Аномалии

Для выявления аномальных значений использованы три стандартных статистических подхода:

- **Z-score** - выявляет значения, сильно отклоняющиеся от среднего
- **IQR** - находит крайние значения за пределами межквартильного размаха
- **Тест Граббса** - определяет одиночные экстремальные наблюдения

Использование нескольких методов позволяет снизить риск ложных выбросов

<u>Выбросы мы пометили флагами</u>

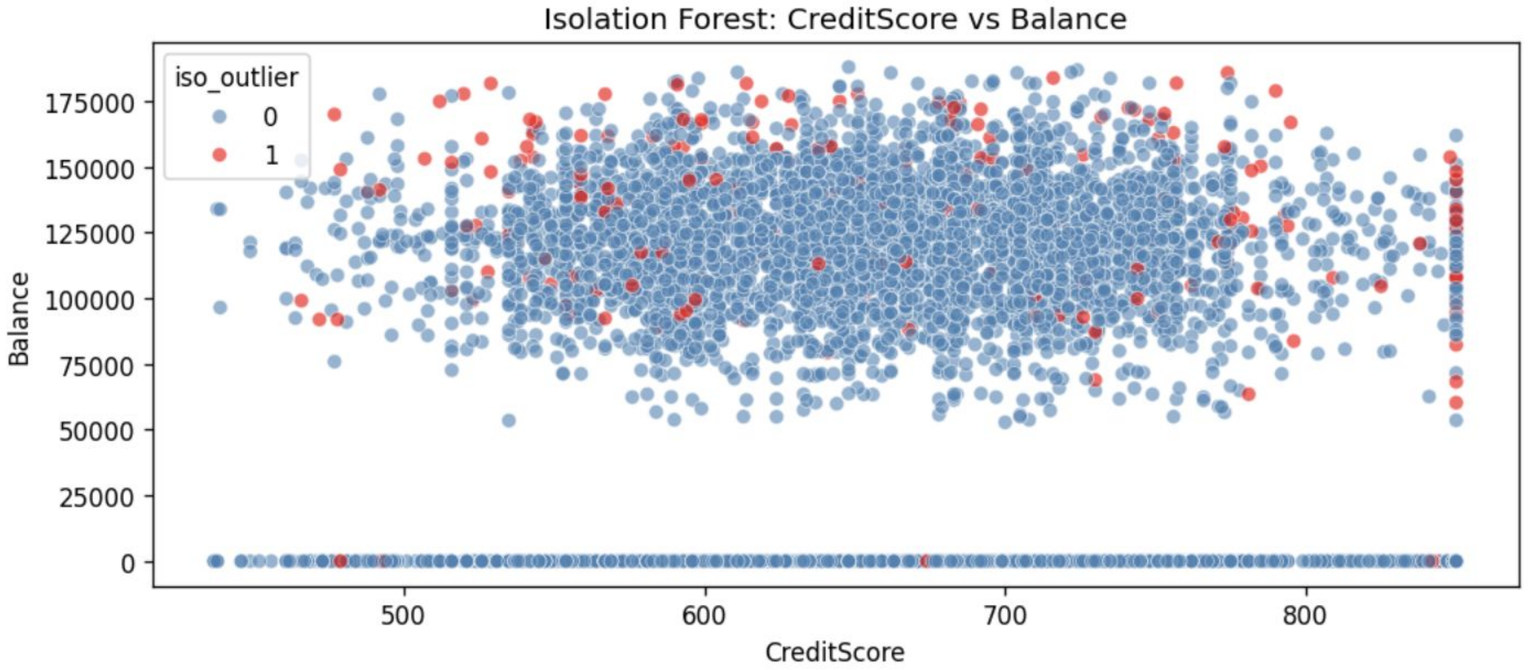


- **Age**: выбросы с краю (58+ лет). Их немного и, в целом, бизнес-логика допускает пожилых клиентов
- **NumOfProducts**: выбросы на значении 4 (редкий пакет из 4 продуктов). Край редкого сегмента, но бизнес-возможный
- **CreditScore**: выбросы в районе низких (<450) и высоких (>850) значений. Это разумные экстремумы
- **Tenure**: выброс вообще всего один

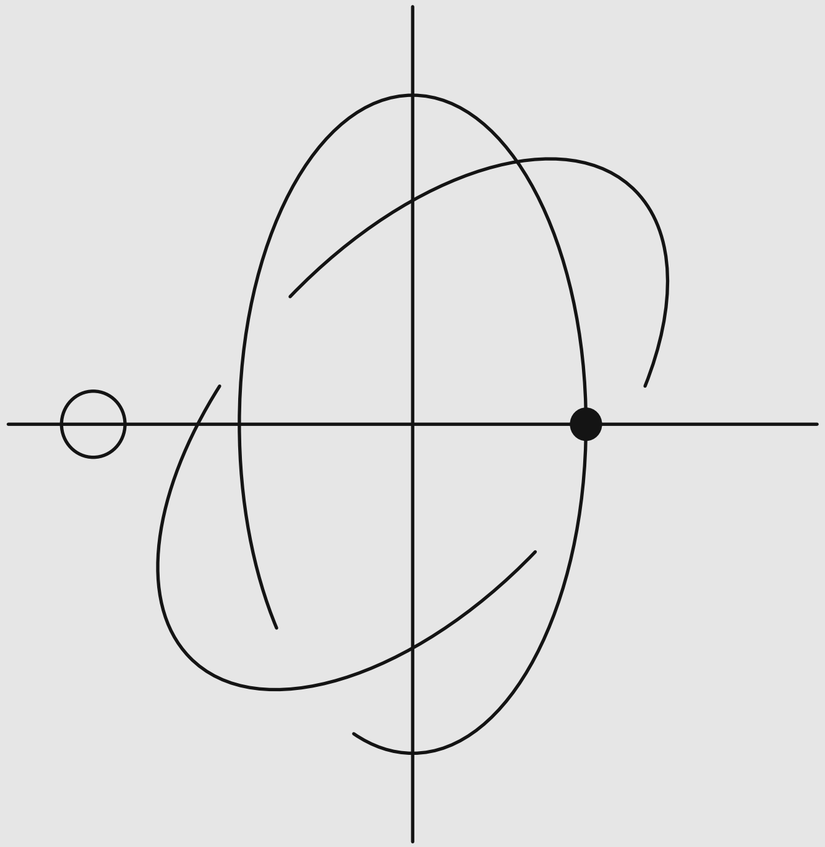
Аномалии

Для выявления нетипичных клиентов использовали методы, основанные на плотности данных и машинном обучении:

- **kNN-плотность**
  Рассчитывалось среднее расстояние до ближайших соседей: чем больше расстояние - тем менее типичен клиент
- **Isolation Forest**
  Аномалии определись как точки, которые легко изолируются случайными разбиениями пространства признаков
- **Local Outlier Factor**
  Искались объекты, находящиеся в зонах пониженной локальной плотности по сравнению с окружением



Аномалии



Аномалии по **Isolation Forest** чаще соответствуют реальному churn: **Isolation Forest** лучше выявляет клиентов с уходом, ибо выше precision и ROC-AUC

LOF в текущих настройках дает ограниченный дополнительный сигнал: **LOF** показывает слабую связь с целевой переменной

Совпадение **IF** и **LOF** можно трактовать как жесткие, редкие аномалии с повышенным риском ухода (вообще пересечение методов минимально, так что алгоритмы находят разные типы аномалий)

Генерация признаков и отбор переменных

1

Для Geography используем target encoding, потому что мало категорий, но связь с таргетом может быть сложной

2

После кодировки категориальных признаков и обучения базовой логистической регрессии с учетом дисбаланса классов модель показала: ROC-AUC ≈ 0.89, accuracy ≈ 0.81. Модель хорошо различает склонных к уходу клиентов и позволяет находить около 80% реально уходящих клиентов, хотя часть оставшихся помечается как потенциально уходящие

3

Перед построением признаков на основе ближайших соседей были заполнены пропуски в числовых переменных. Для каждого признака использовалась медиана по обучающей выборке, которая затем применялась как к train, так и к test. Это позволило корректно использовать методы, не поддерживающие значения NaN и при этом не использовать информацию из тестовой выборки.

4

Контекстные признаки:

- balance_to_salary
- is_high_value_client
- is_new_client
- is_senior

После добавления признаков на основе kNN и контекстных бизнес-гипотез качество почти не выросло. Однако новые признаки могут оказаться более полезными для нелинейных моделей, а также для интерпретации поведения разных сегментов клиентов

Генерация признаков и отбор переменных

Корреляция

Age	0.453925
iso_score	0.293567
NumOfProducts	0.291174
knn_max_dist	0.290327
knn_mean_dist	0.279617
knn_min_dist	0.243812
Geography_te	0.232232
IsActiveMember	0.200789
Gender	0.188824
Balance	0.149047
iso_outlier	0.143339
stat_outlier_count	0.125753
stat_outlier_any	0.124683
Age_outlier	0.114445
lof_score	0.095775
is_senior	0.076258
lof_outlier	0.072809
NumOfProducts_outlier	0.070974
Geography	0.056549
CreditScore	0.049174

Хи-квадрат

Geography_te	569.549101
IsActiveMember	305.757625
iso_outlier	302.027881
Gender	234.463676
stat_outlier_any	226.349312
NumOfProducts	202.534429
Age	193.202811
Age_outlier	191.040957
Balance	147.584666
stat_outlier_count	116.707787
is_senior	85.512620
lof_outlier	78.281112
NumOfProducts_outlier	75.464700
knn_max_dist	57.231452
iso_score	56.270150
knn_mean_dist	50.680652
knn_min_dist	34.674461
Geography	26.167244
is_new_client	8.093096
lof_score	4.086720

ANOVA

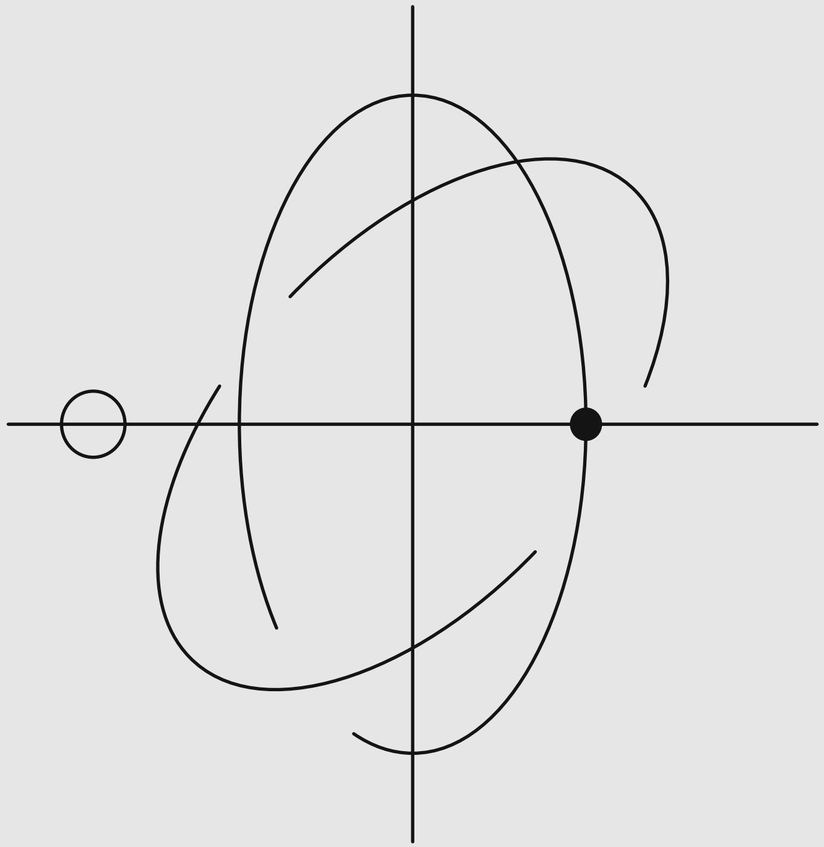
Age	3892.303500
iso_score	1414.446939
NumOfProducts	1389.353025
knn_max_dist	1380.546146
knn_mean_dist	1272.090445
knn_min_dist	947.891830
Geography_te	854.980340
IsActiveMember	630.063225
Gender	554.516452
Balance	340.751837
iso_outlier	314.614733
stat_outlier_count	240.987700
stat_outlier_any	236.840359
Age_outlier	199.044142
lof_score	138.849033
is_senior	87.726625
lof_outlier	79.929375
NumOfProducts_outlier	75.932836
Geography	48.114930
CreditScore	36.353835

Все три фильтра говорят одно и то же:

Очень важные признаки: Age, NumOfProducts, Geography_te, IsActiveMember, iso_score и iso_outlier, knn (mean/max/min dist), бинарки аномалий (stat_outlier_any, stat_outlier_count, Age_outlier, lof_outlier), Gender, Balance

Менее важные: сырой Geography, CreditScore, более слабые флаги вроде is_new_client, lof_score

Генерация признаков и отбор переменных



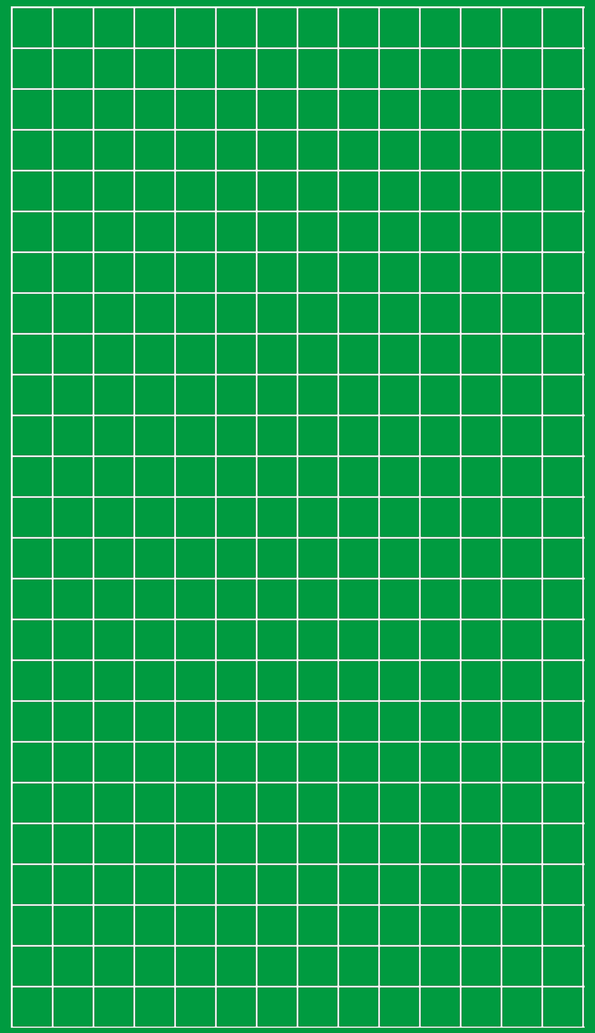
RFECV на логистической регрессии отобрал 25 признаков, включающих основные параметры клиента, географию после TE, признаки аномальности, kNN-фичи и контекстные переменные вроде balance_to_salary и индикаторов разных групп клиентов

Встроенные методы (L1 и RandomForest) в целом подтвердили важность тех же признаков. Некоторые редкие бинарные признаки оказались менее стабильными

Анализ нестабильности через разные фолды показал, что сильнее всего меняются редкие флаги и аномальные признаки, тогда как ключевые фичи остаются устойчивыми.

В итоге выбрано 28 признаков как объединение RFECV (25), L1 (25), Random Forest (20)

ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

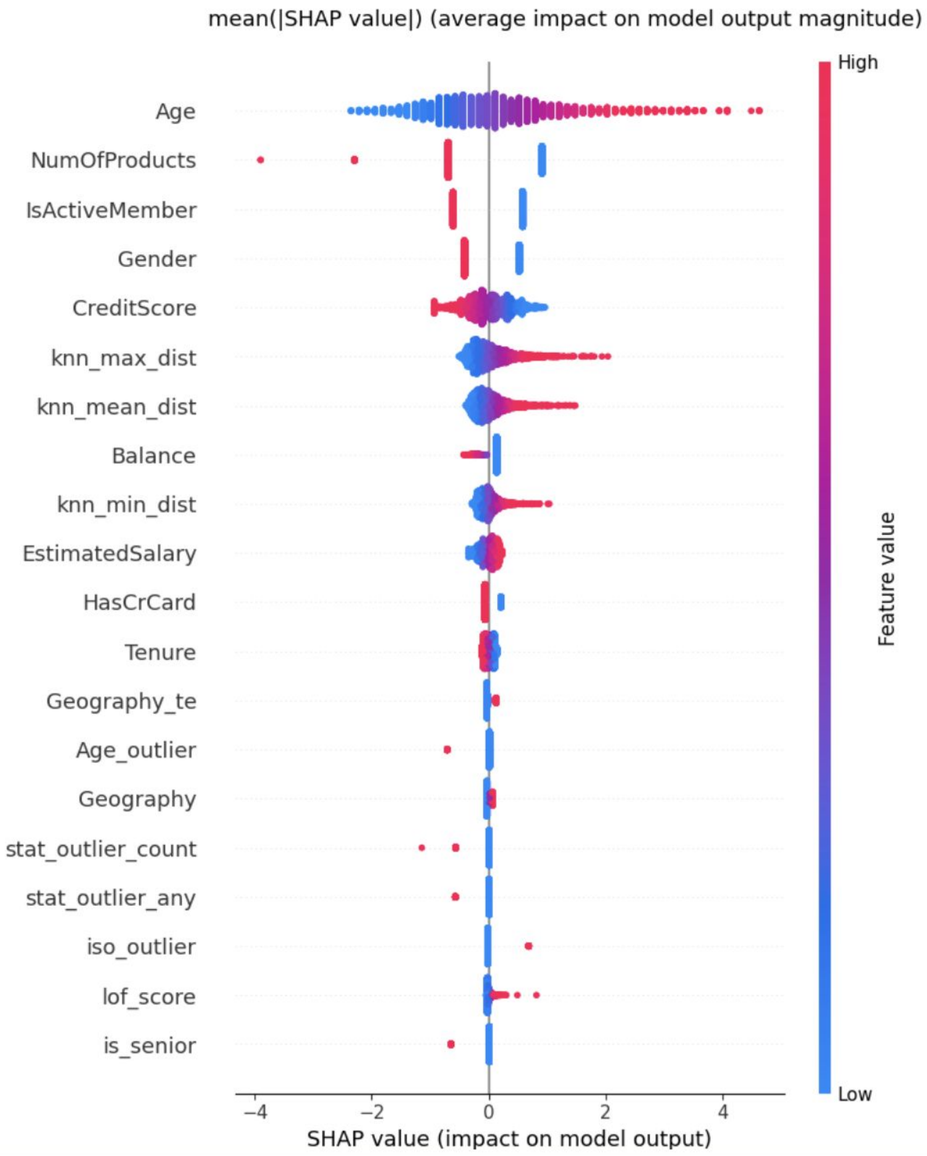


Шаг третий

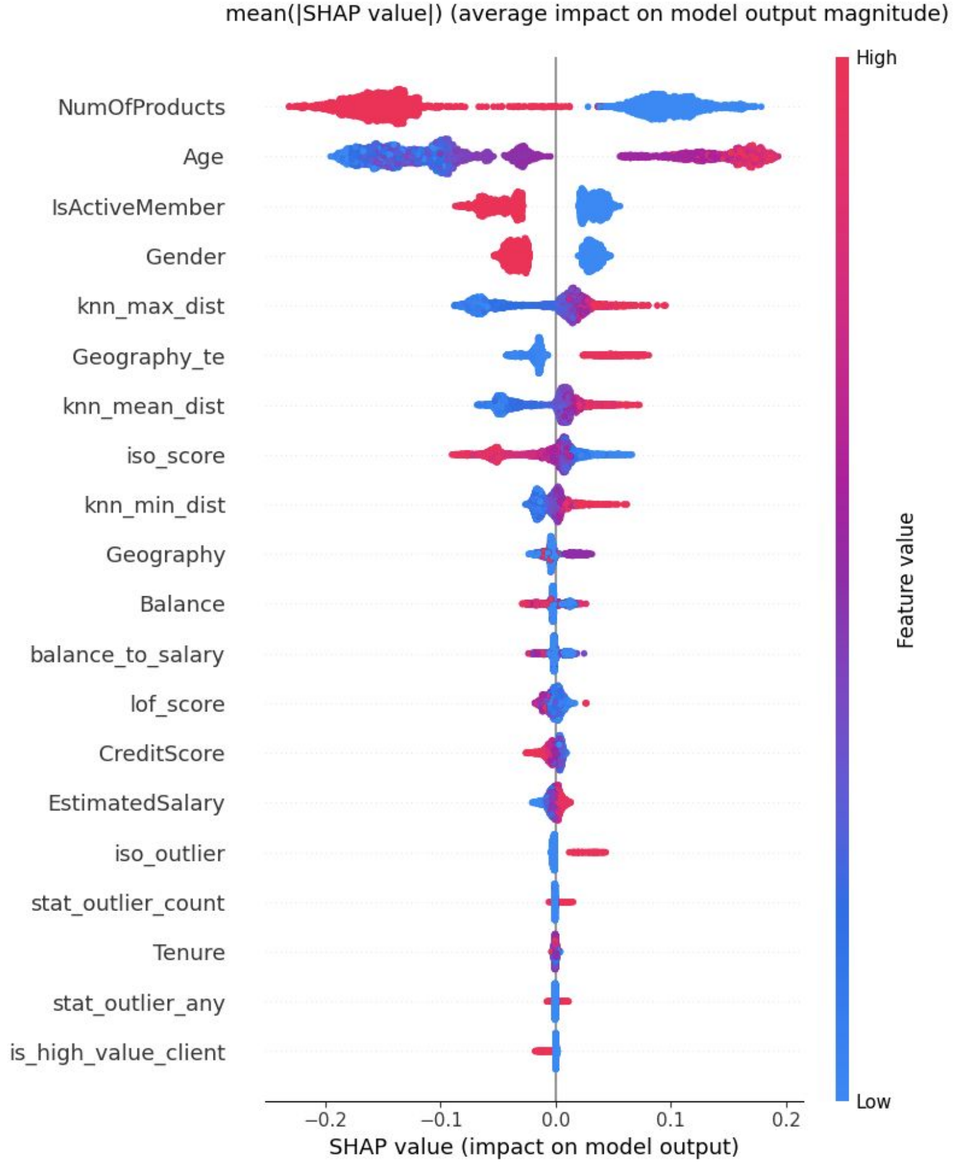
Интерпретация и диагностика моделей

Интерпретация моделей

SHAP для логистической регрессии



SHAP для Random Forest

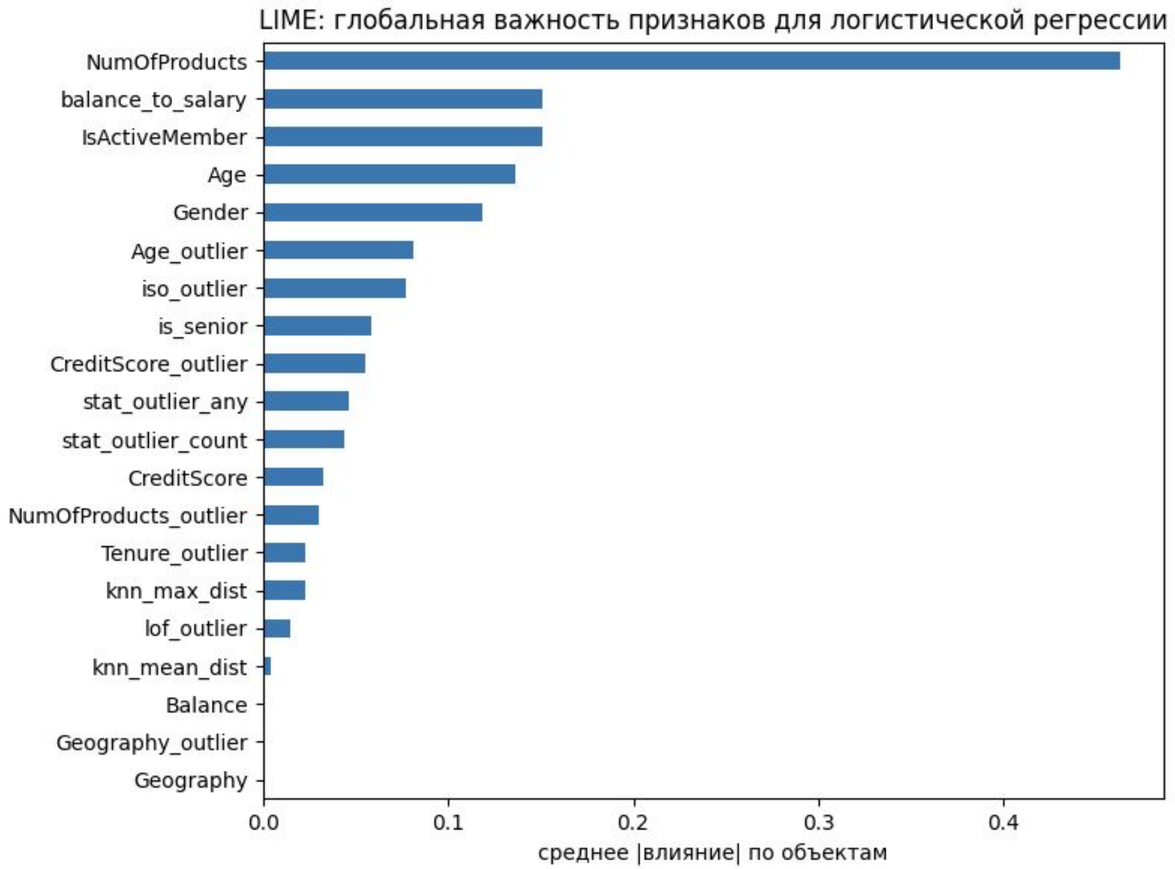


Тут все довольно ожидаемо. Самые важные признаки: Age, NumOfProducts и IsActiveMember. По второму графику видно, что большой возраст увеличивает вероятность ухода, а большое число продуктов и активность клиента наоборот уменьшают вероятность ухода. Признаки вроде Gender, CreditScore и kNN-фичи тоже что-то дают модели, но заметно слабее. Остальные признаки почти не влияют.

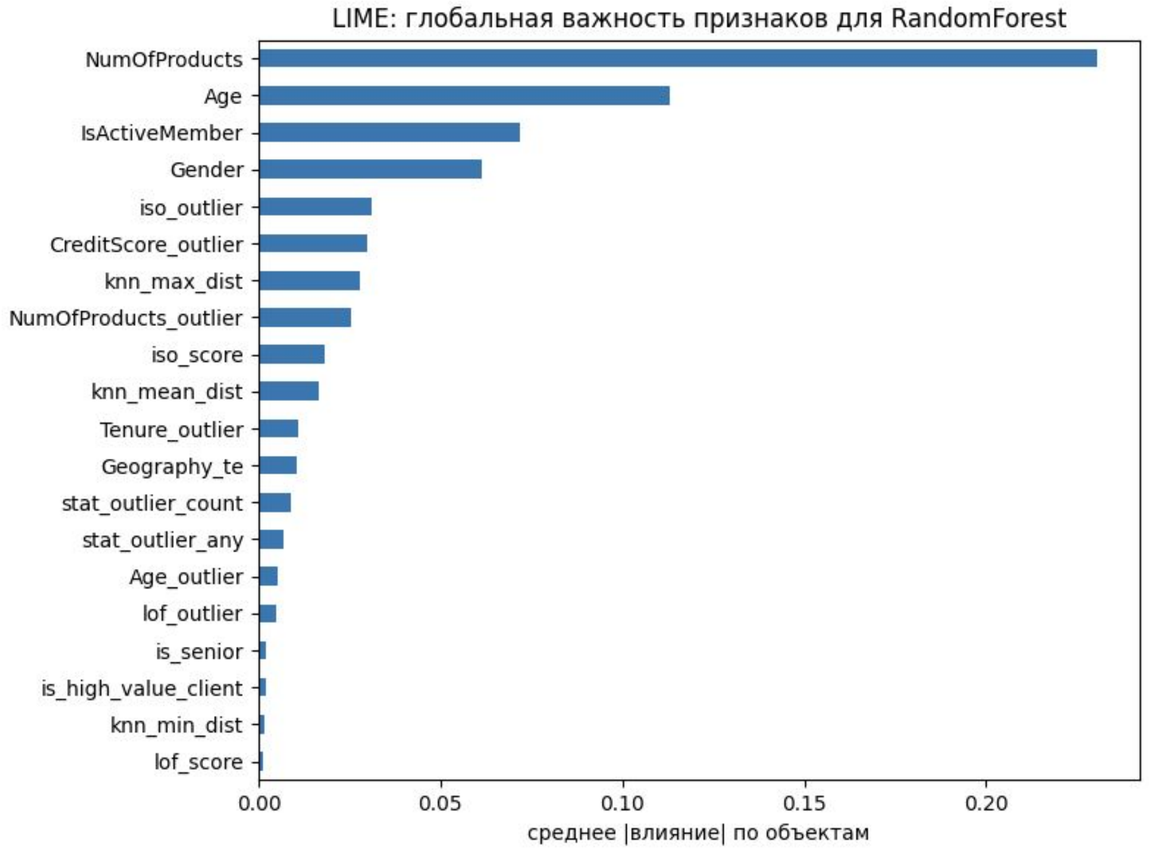
SHAP показывает, что модель живет довольно простой логикой

Интерпретация моделей

LIME для логистической регрессии



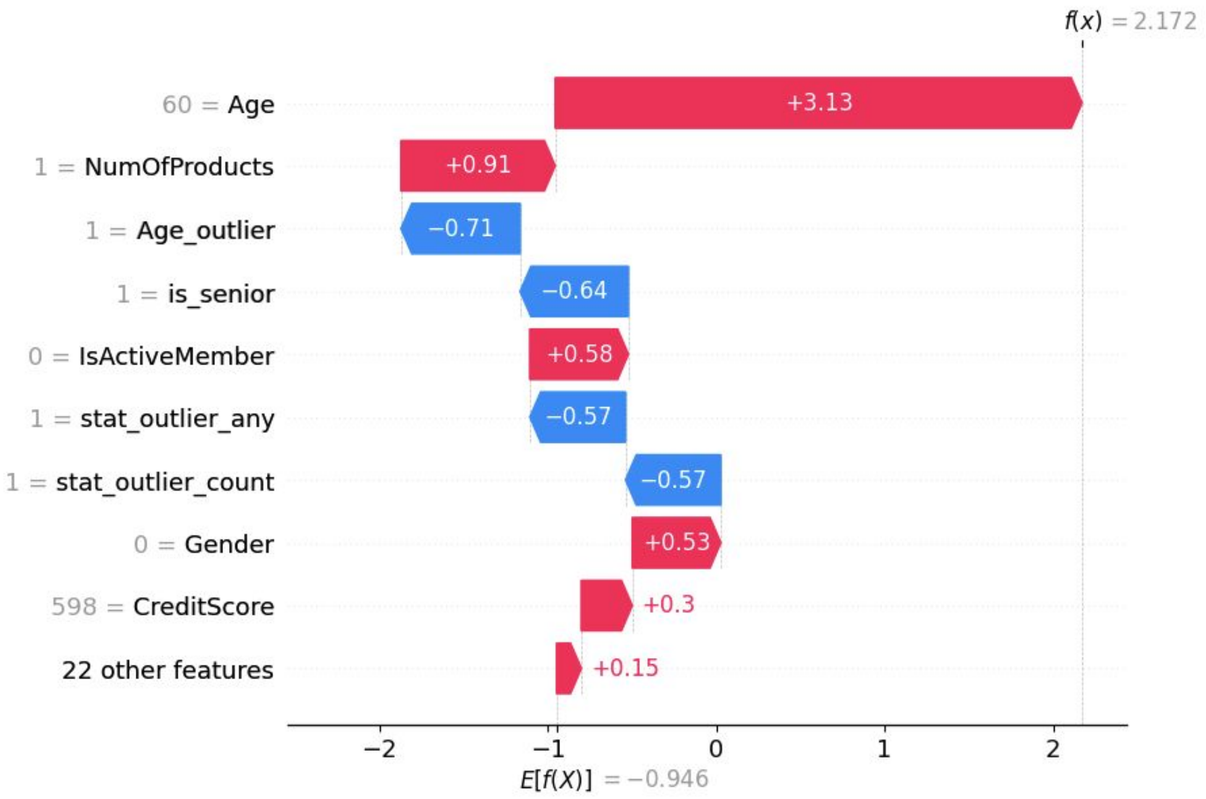
LIME для Random Forest



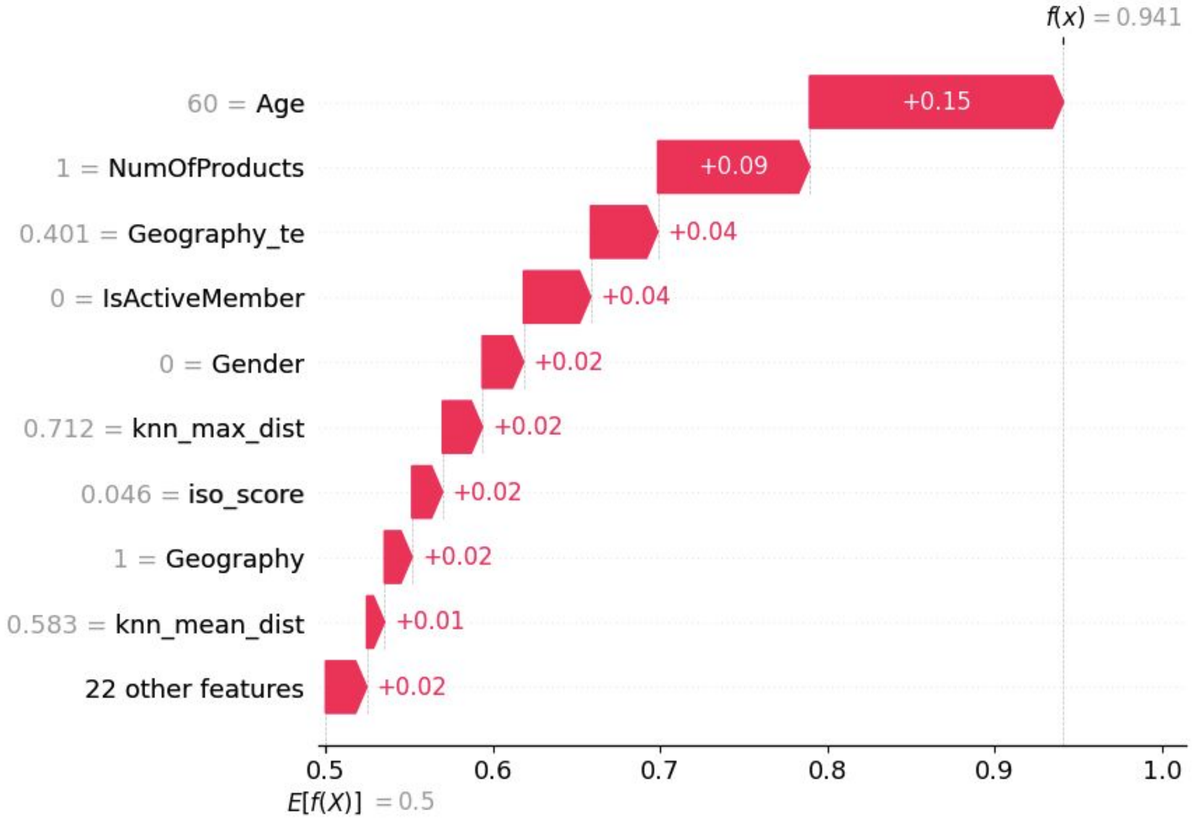
LIME подтверждает SHAP. По результатам SHAP и LIME видно, что ключевые признаки у линейной и ансамблевой моделей в целом совпадают. В обоих случаях в топе стабильно находятся Age, NumOfProducts и IsActiveMember. Направления влияния тоже согласуются: большой возраст увеличивает вероятность ухода, большое число продуктов снижает вероятность ухода, активность клиента увеличивает вероятность ухода

Интерпретация моделей

SHAP локально для логистической регрессии

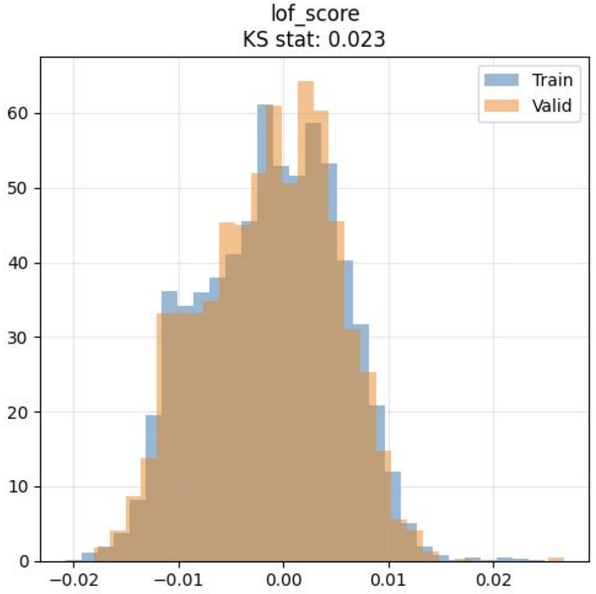


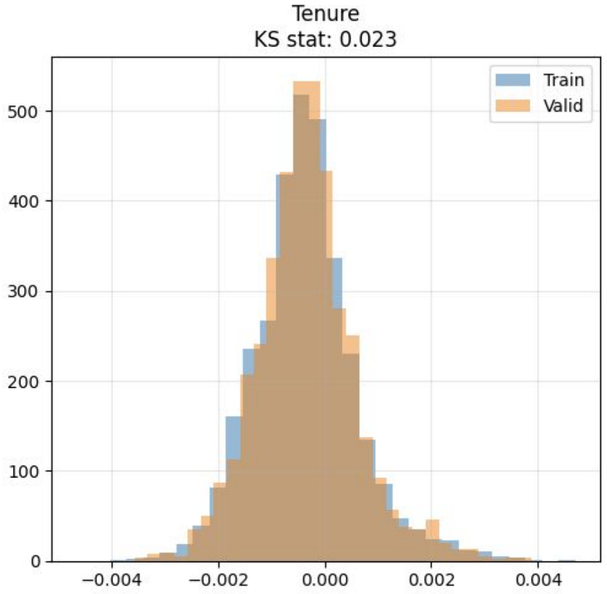
SHAP локально для Random Forest

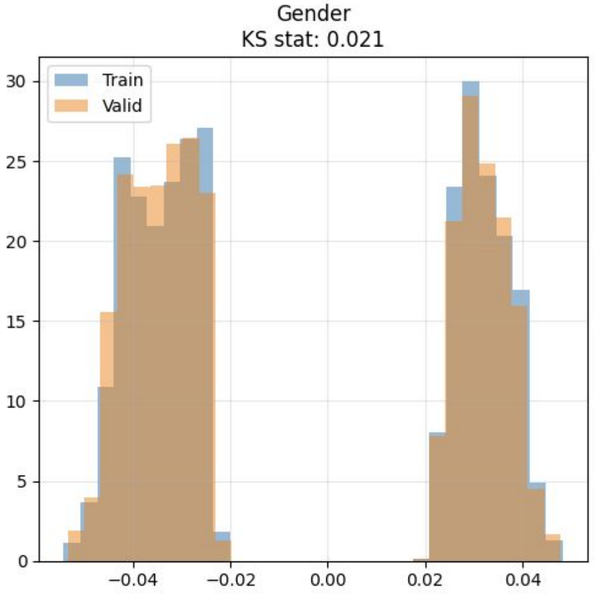


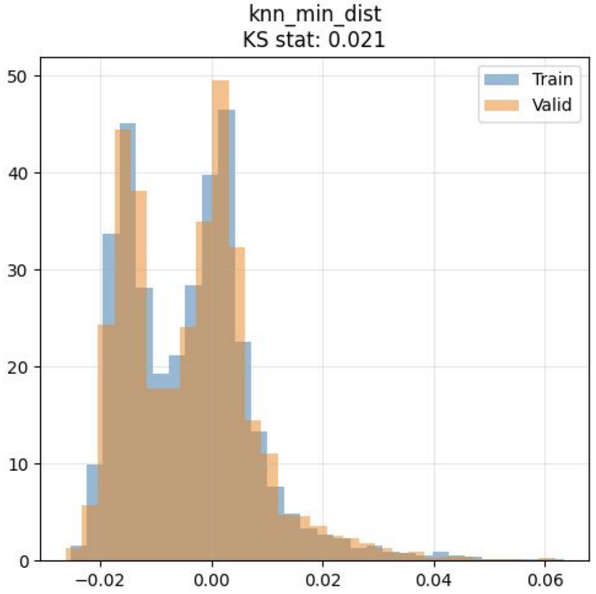
Для клиента LIME и SHAP дают похожую картину для логистической регрессии и для RandomForest. В обоих методах ключевыми факторами стали большой возраст, малое число продуктов и неактивность клиента, то есть основные причины высокого риска ухода совпадают

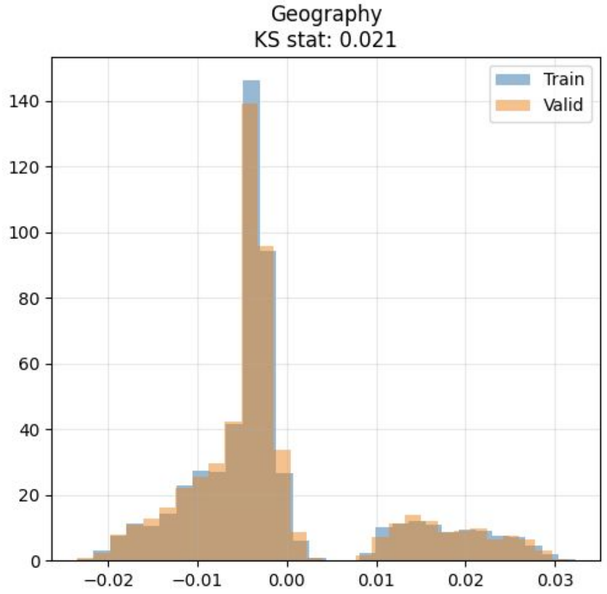
Построение SHAP-эмбеддингов и анализ сдвигов

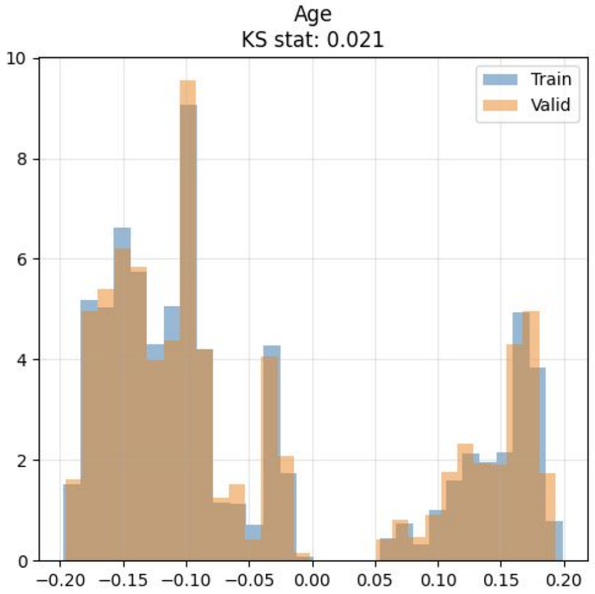






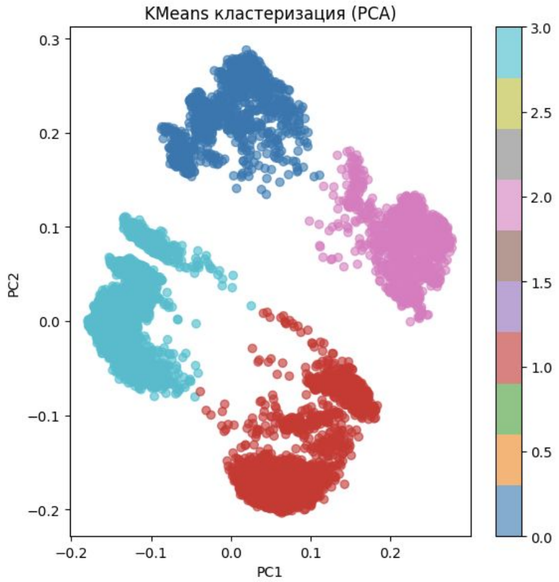


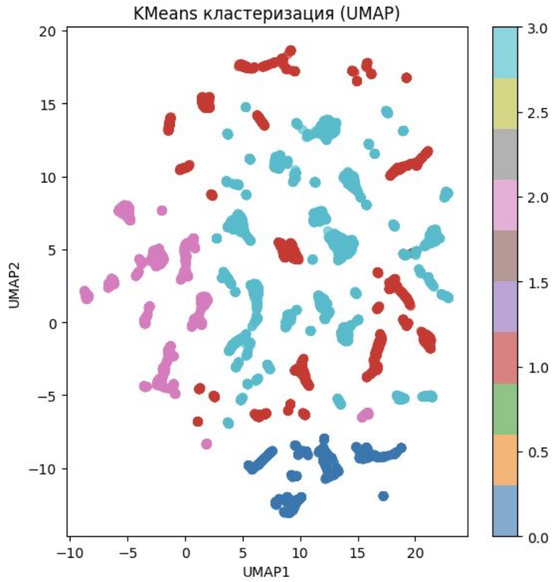


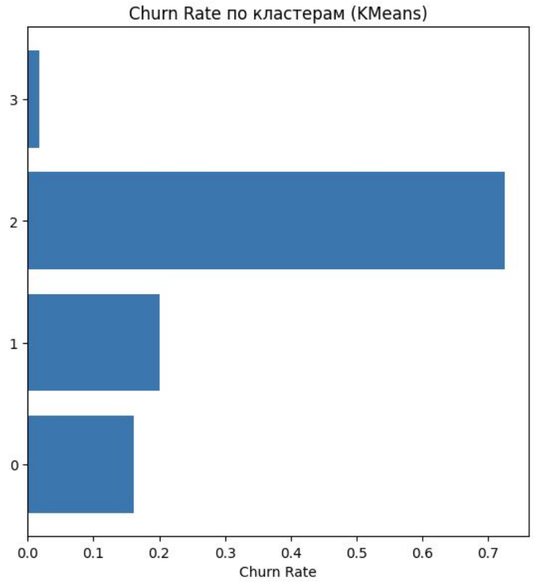


Была проведена очистка данных на основе анализа сдвигов, но она не дала прироста ROC-AUC

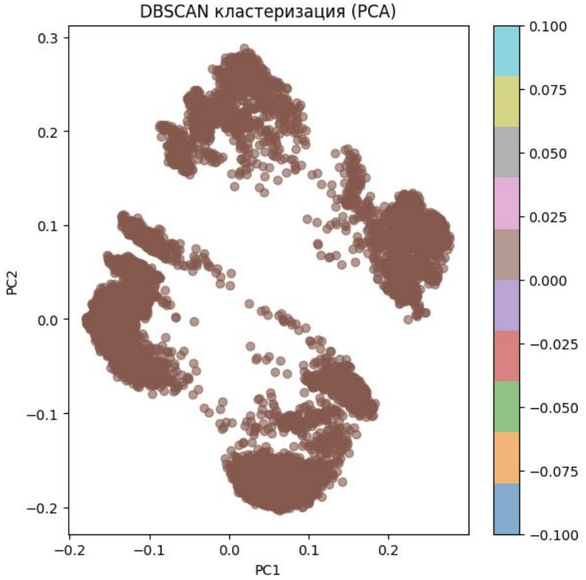
Построение SHAP-эмбеддингов и анализ сдвигов

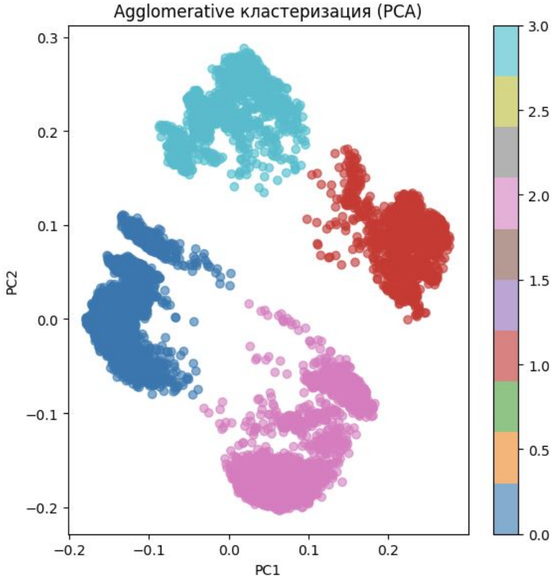


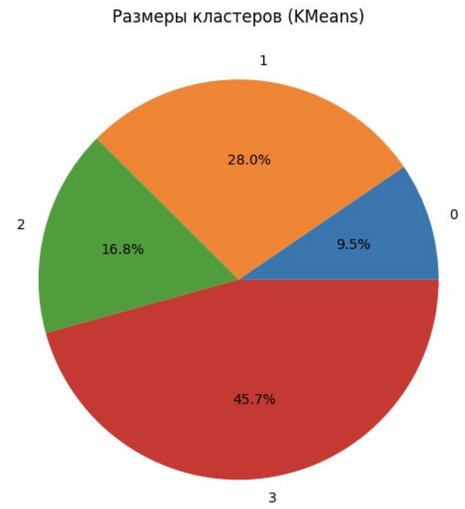




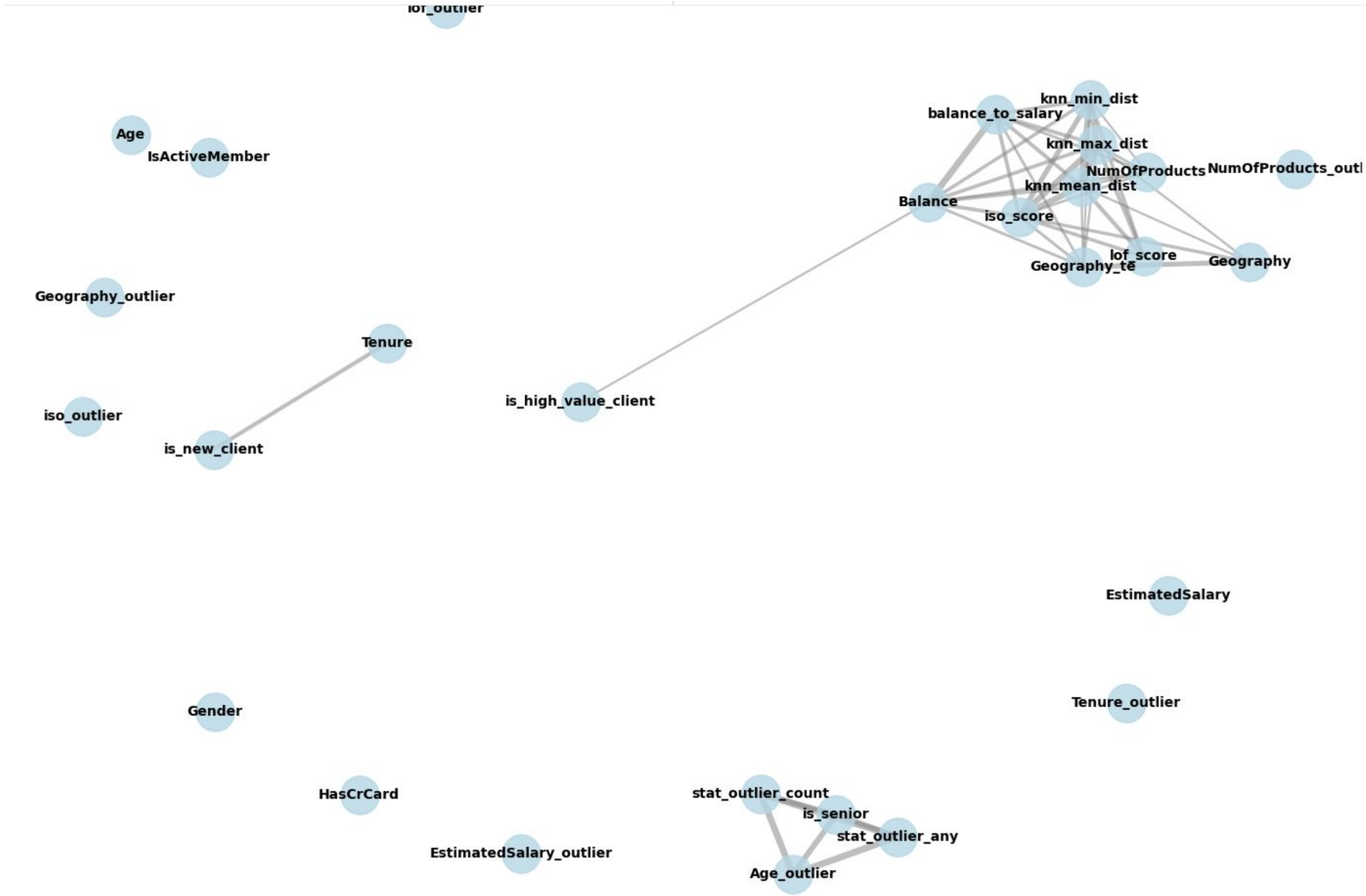
Кластеризация SHAP-эмбеддингов выявила 4 кластера с разной долей оттока клиентов







Валидация и Snaplry Flow



Кросс-валидация ROC-AUC:
Только исходные признаки: 0.9221
Только SHAP-эмбеддинги: **0.9481**
Комбинация: 0.9463

!Обучение на SHAP-эмбеддингах дало наилучший результат на втором же месте - комбинация исходных данных