



Тема проекта: Прогнозирование оттока клиентов


Проект подготовлен в рамках курса "Прикладная наука о данных"

Выполнили:

Бобров Валерий, Гришкина Анна, Милицына Полина,

Мороз Екатерина, Соболева Татьяна, Шпилёва Лина

БАСБ 252



Актуальность: модель оттока позволяет заранее выявлять клиентов, склонных к уходу, и предпринимать оперативные меры для его удержания. В результате, применение модели позволяет сократить расходы на привлечение, увеличить LTV и достичь более стабильных пассивов.

Цель: построение модели машинного обучения для прогнозирования вероятности ухода клиента из банка на основе анализа его профиля, истории взаимодействия и финансовых показателей.

Задачи:

1. Провести разведочный анализ данных (EDA).
2. Предобработать данные для использования в модели.
3. Сгенерировать новые признаки для улучшения предсказательной способности моделей.
4. Построить несколько моделей машинного обучения.
5. Оценить полученные модели по метрикам качества, выбрав итоговую модель.
6. **Сформулировать выводы и рекомендации для бизнеса.**

За основу взят датасет из Kaggle соревнования Bank Churn Competition by IPII HSExMTS

Описание датасета

- Customer ID: Уникальный идентификатор каждого клиента.
- Surname: Фамилия клиента.
- Credit Score: Числовое значение, представляющее кредитный рейтинг клиента
- Geography: Страна проживания клиента (Франция, Испания или Германия).
- Gender: Пол клиента (Мужской или Женский).
- Age: Возраст клиента.
- Tenure: Количество лет, которое клиент обслуживается в банке.
- Balance: Баланс на счёте клиента.
- NumOfProducts: Количество банковских продуктов, которыми пользуется клиент (например, сберегательный счёт, кредитная карта).
- HasCrCard: Наличие кредитной карты у клиента (1 = да, 0 = нет).
- IsActiveMember: Является ли клиент активным членом банка (1 = да, 0 = нет).
- EstimatedSalary: Предполагаемая заработная плата клиента.
- Exited: Ушёл ли клиент (1 = да, 0 = нет).

Исследовательский анализ данных (EDA)

	id	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	0	15752626.0	Martin	675.0	France	Female	48.0	7.0	143582.89	2.0	0.0	0.0	93844.82	1.0
1	1	15797960.0	Pagnotto	673.0	France	Female	37.0	7.0	0.00	2.0	0.0	0.0	170980.86	0.0
2	2	15672056.0	T'an	607.0	France	Male	29.0	4.0	0.00	2.0	0.0	1.0	61290.99	0.0

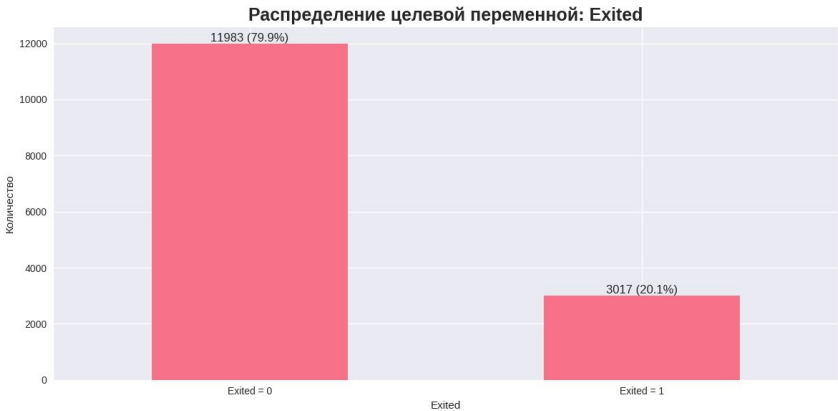
	id	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited		Surname	Geography	Gender
count	15000.000000	1.500000e+04	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000	15000.000000				
mean	7499.500000	1.569136e+07	658.223400	37.850067	5.03080	42636.168488	1.585733	0.779933	0.494400	117924.109015	0.201133				
std	4330.271354	1.355762e+05	72.851215	8.185959	2.80209	59570.493235	0.529026	0.414305	0.499985	45698.606770	0.400861				
min	0.000000	1.574723e+06	437.000000	18.000000	0.000000	0.000000	1.000000	0.000000	0.000000	11.570000	0.000000				
25%	3749.750000	1.563450e+07	601.000000	32.000000	3.000000	0.000000	1.000000	1.000000	0.000000	83419.440000	0.000000				
50%	7499.500000	1.568965e+07	660.000000	37.000000	5.000000	0.000000	2.000000	1.000000	0.000000	123475.880000	0.000000				
75%	11249.250000	1.575781e+07	708.000000	42.000000	7.000000	109079.755000	2.000000	1.000000	1.000000	157564.750000	0.000000				
max	14999.000000	1.581567e+07	850.000000	72.000000	20.000000	187911.550000	4.000000	1.000000	1.000000	199953.330000	1.000000				

count	15000	15000	15000
unique	778	3	2
top	Ch'iu	France	Male
freq	278	9002	8424

Датасет содержит 15 000 строк

Пропущенные значения отсутствуют

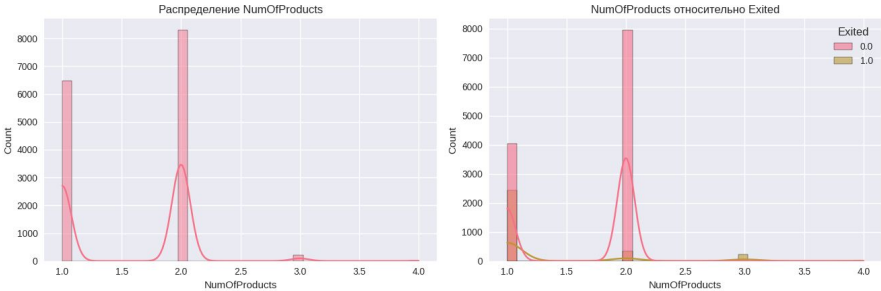
Количество выбросов незначительное



EDA. Анализ переменных

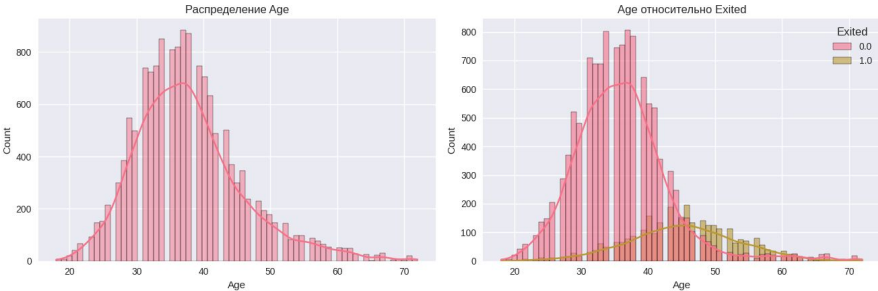
Анализ признака: NumOfProducts

Среднее: 1.59 | Медиана: 2.00 | Ст. отклонение: 0.53 | Асимметрия: 0.08 | Эксцесс: -0.84

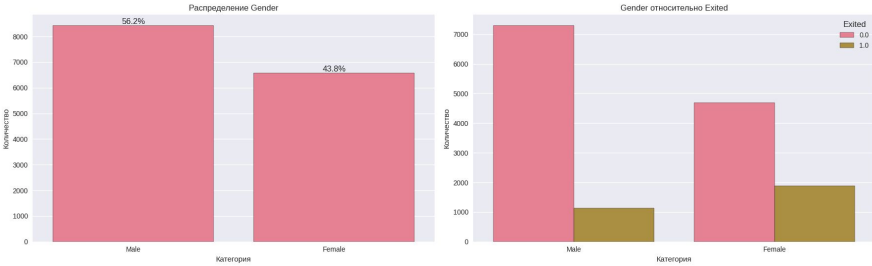


Анализ признака: Age

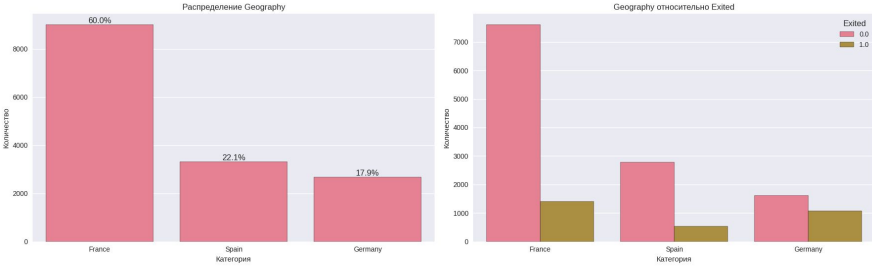
Среднее: 37.85 | Медиана: 37.00 | Ст. отклонение: 8.19 | Асимметрия: 0.86 | Эксцесс: 1.16



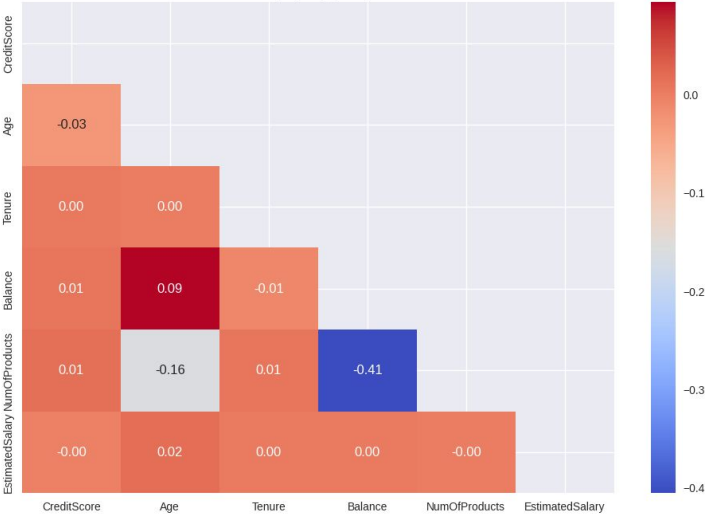
Анализ признака: Gender



Анализ признака: Geography



Матрица корреляций



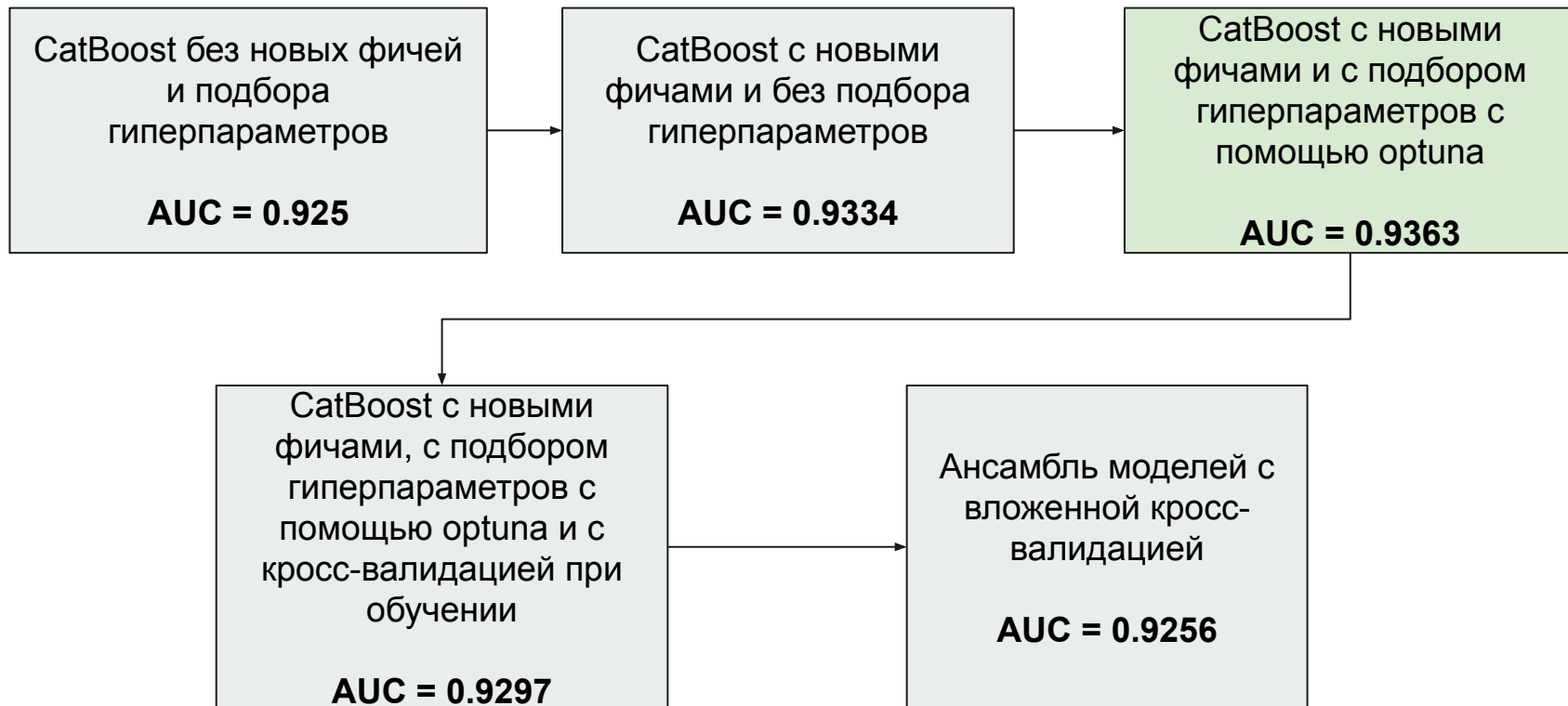
Feature engineering



Основные методы feature engineering, которые мы использовали:

- Комбинирование категориальных признаков
 - Пример: страна проживания + пол клиента
- Комбинирование числовых признаков
 - Пример: произведение возраста и кредитного рейтинга
- Логарифмирование числовых непрерывных величин
 - Пример: логарифм заработной платы
- Target encoding
 - Пример: среднее для каждой комбинации признаков страна проживания + возраст

ML моделирование



Лучшая модель

По итогам обучения было выявлено, что лучшими моделями являются:

- CatBoost с новыми фичами и с подбором гиперпараметров с помощью optuna
- Ансамбль моделей с вложенной кросс-валидацией

Таблица 1. Сравнение моделей по метрикам качества

	Accuracy	Precision	Recall	F1	AUC
CatBoost	0.8693	0.6320	0.8375	0.7204	0.9363
Ансамбль	0.8966	0.8029	0.6440	0.7147	0.9256

В данном кейсе более предпочтительно ложное предсказание оттока (больше FP -> ниже Precision), чем ложное предсказание отсутствия оттока (больше FN -> ниже Recall), поэтому выбираем CatBoost с optuna.

Таблица 2. Топ-10 признаков по вкладу в предсказательную способность модели

Признак	Важность
Количество продуктов	10.42
Кредитный скоринг x Количество продуктов	5.10
Возраст x Количество продуктов	4.95
Количество продуктов/ Год возраста	4.89
Возраст x Активность	4.65
Кредитный рейтинг	4.60
Логарифм зарплаты X Кредитный рейтинг	4.47
Target-encoded группа страна x Пол x Возрастная группа	4.36
Заработная плата	3.87

Основные выводы

1. Кредитный рейтинг и баланс менее информативны. Корреляция между CreditScore, Balance и оттоком крайне низкая, важнее фокусироваться на поведенческих и демографических признаках.
2. **Риск оттока выше для отдельных сегментов:**
 - Клиенты из Германии демонстрируют заметно больший процент оттока по сравнению с Францией и Испанией. Это связано с региональными особенностями и, возможно, продуктовой политикой банка.
 - Владельцы только одного банковского продукта гораздо чаще уходят, чем клиенты с несколькими продуктами. Т. е. лояльность выше у клиентов, использующих несколько сервисов одновременно.
 - Возраст также коррелирует с вероятностью оттока: более высокая вероятность ухода среди клиентов среднего возраста, особенно в возрасте 40–50 лет.
3. **Модель CatBoost с продвинутым feature engineering** даёт сбалансированную производительность по ключевым метрикам (Recall, Precision, ROC-AUC). **Подбор гиперпараметров (Optuna)** позволяет увеличить полноту (Recall) до 0.846, что важно: такие модели "ловят" максимальное число реально уходящих клиентов.
4. Ансамбли моделей усиливают Precision, но теряют Recall. Такой подход может быть полезен, если банк ориентирован на снижение издержек от неверных обращений к лояльным клиентам. Однако приоритетнее минимизировать пропуски уходящих (FN), а не минимизировать ложные срабатывания (FP).

Рекомендации для бизнеса



1. **Сфокусировать удержание на следующих сегментах:**

- Клиенты из Германии;
- Клиенты, имеющие 1 продукт;
- Клиенты среднего возраста (40-50 лет).

Для этих групп запускать кросс-продажи, персонализированные предложения, улучшать клиентский опыт (например, поддержка, бонусы).

2. **Использовать рекомендованные модели (модель CatBoost с применением feature engineering и подбором гиперпараметров) в скрининге клиентской базы.** Модель показывает высокую полноту и позволит банку выявить до 85% клиентов с реальным риском ухода.

3. **Автоматизировать работу с моделью:**

- Внедрить триггерные сценарии/уведомления для CRM – запускать процессы удержания для всех клиентов с вероятностью оттока выше выбранного порога (например, 0.5).
- Регулярно дообучать модель, чтобы адаптироваться к изменяющемуся поведению клиентов.

4. **Оценивать экономический эффект внедрения модели.**

Сравнить стоимость удержания с потенциальными потерями при уходе выявленных клиентов. Особое внимание уделить чувствительным сегментам "high-value customers".

Дашборд

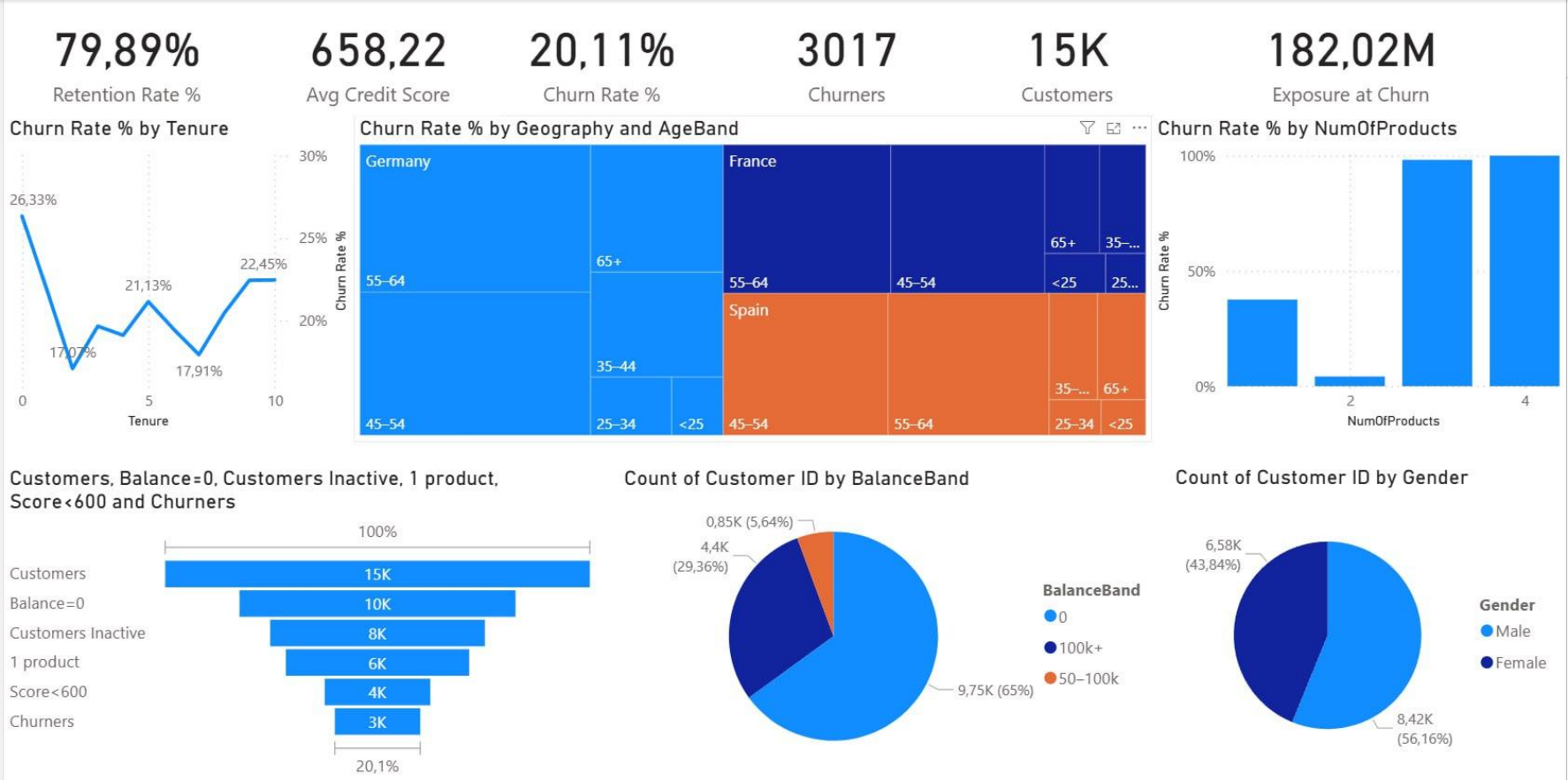


Рис. 7

ЮНИТ-ЭКОНОМИКА

- Всего клиентов: 15000
- Churn rate: 20.11%
- Средний баланс: 42636.17
- ARPU: 584.94
- LTV: 1177.08
- CAC: 350.00
- ROI: 236.31%

Какие метрики мы считали и зачем

- Churn rate
- Retention rate
- Средние показатели по клиентам
- ARPU (Average Revenue Per User)
- LTV
- ROI (Return on Investment)
- CAC — это стоимость привлечения клиента (например, расходы на рекламу, бонусы и т.д.).

(Мы задали CAC вручную (например, 350€), чтобы рассчитать окупаемость)